



UNIVERSIDAD LA SALLE

FACULTAD DE NEGOCIOS

Con Reconocimiento de Validez Oficial de Estudios de la
Secretaría de Educación Pública según acuerdo número 974226
de fecha 25 de julio de 1997

CASO DE ESTUDIO

**“Diseño de flujo de datos automatizado con herramientas
ágiles para el análisis de información en una empresa
mexicana de procesamiento de video en tiempo real.”**

QUE PARA OBTENER EL GRADO DE
MAESTRO EN TECNOLOGÍAS DE INFORMACIÓN EN LA
DIRECCIÓN DE NEGOCIOS

PRESENTA:

Erendira Reyes Rangel

Asesor: Mtro. Ismael Villegas Ochoa

Ciudad de México, a 09 de mayo de 2025

Ciudad de México, a 11 de abril de 2024.

MTRO. GABRIEL NUÑEZ GONZÁLEZ
DIRECTOR DE GESTIÓN ESCOLAR
UNIVERSIDAD LA SALLE
P R E S E N T E

Le informo que el (la) C.

REYES RANGEL ERENDIRA

Egresada (o) de la Facultad de: Negocios

de la **UNIVERSIDAD LA SALLE**, de la Maestría en:

TECNOLOGÍAS DE INFORMACIÓN EN LA DIRECCIÓN DE NEGOCIOS

Con reconocimiento de validez oficial de estudios de la Secretaría de Educación

Pública según acuerdo número 974226 de fecha 25 de julio de 1997

Ha elaborado el Estudio de Caso titulado: "Diseño de flujo de datos automatizado con herramientas ágiles para el análisis de información en una empresa mexicana de procesamiento de video entiempos real."

De conformidad con la modalidad para la obtención de grado aprobada para esta Maestría de acuerdo a lo establecido en el Reglamento General de las Universidades La Salle Integrantes del Sistema Educativo de las Universidades la Salle.

Cumplió con todos los requisitos y el trabajo que fue elaborado bajo la conducción del Mtro. Ismael Villegas Ochoa que fungió como asesor, tiene la calidad suficiente para ser la base de sustentación de su Examen de Grado por lo que se le autoriza presentarlo.


Mtro. Markel Israel Lehman Elizondo
Director Facultad de Negocios

Dedicatoria

Erandi Sofia:

Hija te dedico el presente caso de estudio ya que tú eres la motivación más grande que me impulsa día a día a ser la mejor versión de mí. Todos los días ruego a Dios me de vida y sabiduría para guiarte y cuidarte siempre.

Todo el esfuerzo y sacrificio vale la pena cuando te veo sonreír, nunca olvides que te amo y que siempre puedes conseguir lo que desees con trabajo, esfuerzo y perseverancia.

Agradecimiento

A mis Tíos:

Agradezco a todos ustedes su apoyo, tiempo, cariño y paciencia, pero sobre todo el amor que le tienen a mi hija. Sin ustedes no hubiera sido posible este logro, es por eso por lo que lo comparto con ustedes.

Mil gracias, Angelica, Rodolfo, Víctor, Carlos, José y Guadalupe

Director de tesis:

Mtro. Ismael Villegas Ochoa agradezco por todos sus consejos y apoyo, pero sobre todo por compartir su conocimiento inspirándonos siempre con temas innovadores y disruptivos.

Sinodales:

Agradezco a todos su consejo, guía y paciencia para la realización de este trabajo ya que cada uno de sus aportes fue muy valioso para mí.

Muchas gracias Mtra. Alejandra Rivera Alfaro, Mtro. Marco Antonio Siller Jiménez

Equipo:

A todo mi equipo de trabajo, por creer en mí y por arriesgarse a aprender algo nuevo, nunca pierdan las ganas de seguir aprendiendo e innovar nuevas formas de trabajo. Sean mejores cada día. Muchas gracias: Humberto, Gustavo, Pablo, Kevyn, Luis, Abel, Ramón, Liz.

Contenido

Contenido	1
Introducción	3
Objetivo general	4
Objetivos específicos	4
Descripción del problema	5
Marco Conceptual	6
Antecedentes	6
Datos disponibles	6
Tipos de Datos	7
Tiempo de generación y procesamiento del dato	9
Análisis de Datos	11
Tubería de Datos	14
Fases de una Tubería de Datos	16
Data Driven	20
DataOps	21
Definición	21
Origen	22
Desarrollo Agile	24
Lean Thinking	26
DevOps	29
Alcance DataOps	31
Principios DataOps	31
Ciclo de Vida DataOps	33
Componentes DataOps	35
Procedimiento Metodológico	39
Metodología de estudio	39
Situación Actual	46
Arquitectura de procesos actual As - Is	49
Arquitectura tecnológica actual As - Is	52
Diagnóstico	54
Alternativa de solución	57
Arquitectura de procesos To-Be	57
Arquitectura tecnológica To-Be	64
Propuesta para monitoreo de la tubería de datos	74
Propuesta estructura Gerencia de Datos	76
Gestión del Proyecto	78

Herramienta de gestión y seguimiento del proyecto	79
Cronograma de Tareas	80
Conclusiones	96
Bibliografía	98
Bibliografía Web	99
Glosario	101
Relación de Tablas	102
Relación de Ilustraciones	104
Anexos	106

Introducción

El presente caso de estudio se enfocó en una empresa mexicana de procesamiento de video en tiempo real con una problemática específica en la complejidad de su operación de datos.

A lo largo de este documento se desarrolla el análisis, diagnóstico y propuesta de solución de mejora a sus procesos operativos y tecnológicos buscando mejorar el tiempo de respuesta desde que se genera el dato hasta que las diferentes áreas de negocio tienen disponible la información para la toma de decisiones informada.

Particularmente esta compañía es parte de un sector de entretenimiento donde sus clientes que consumen sus servicios generan datos constantemente las 24 horas los 365 días del año siendo sus patrones de consumo de video su principal fuente para conocer a sus clientes y poder mejorar el servicio que presta a los mismos.

La empresa busca colocarse al nivel de sus principales competidores proveedores de contenido de video los cuales hoy día ya toman decisiones y manejan su negocio creando y adecuando sus productos basados en datos.

El diseño detallado en la sección de alternativa de solución es una propuesta innovadora que busca resolver la problemática actual reduciendo el tiempo desde que sus usuarios generan datos interactuando con la plataforma de video de la compañía hasta que las personas de negocio puedan hacer uso de estos datos para poder tomar decisiones oportunas que les permitan manejar su negocio.

Con esta propuesta se busca liberar el tiempo que hoy se consume en tratar de actualizar y entregar información a las diferentes áreas de la empresa por parte del área de datos de la compañía recuperando así capacidad de tiempo y recursos para invertirlos en nuevos proyectos que busquen la mejora continua y dar mayor valor agregado al negocio con nuevas ideas e incluso dedicarse a generar innovación de mercado.

Objetivo general

Ofrecer una mejora en el monitoreo y control del flujo de datos para la toma de decisiones de negocio y operacionales de una compañía mexicana que se dedica al procesamiento de video en tiempo real que opera en servidores virtuales sobre Internet.

Objetivos específicos

- Identificar y mapear cada uno de los diferentes Flujos de Datos de Carga que se controlan durante el proceso operativo del equipo de datos que permita identificar duplicidad de procesos, procesos no actualizados o procesos innecesarios.
- Diseñar la homologación de procesos de validación y monitoreo operativo que nos permita medir la eficiencia de cada fase del Flujo de Datos diseñado en términos cualitativos y cuantitativos.
- Definir indicadores que permitan la medición de la eficiencia en cada fase del flujo diseñado para poder identificar cuellos de botella para generar una mejora continua.
- Definir indicadores línea base para medir el tiempo de integración de nuevos flujos de datos en cada fase del Flujo de Datos a fin de generar una mejora continua.
- Definir indicadores línea base para medir los tiempos fuera de operación por errores en la carga de datos y procesos a fin de generar una mejora continua.
- Definir valores mínimos y máximos sobre los indicadores principales de negocio que permita detectar comportamientos o patrones fuera de estos rangos para generar alertas automatizadas para la toma de decisiones de negocio de manera oportuna.
- Proponer una arquitectura informática que permita la automatización del Flujo de Datos definido para la operación e integración constante de flujos de datos.

Descripción del problema

La problemática surge al tener que controlar y gestionar diferentes flujos de datos complejos, no homogéneos, con mantenimientos extensos y factibles de error debido a procesos manuales ya sea por errores humanos o del propio sistema que vuelve lenta y difícil su operación diaria generando demora en la entrega y disponibilidad de información hacia el negocio. Lo cual se ve reflejado al invertir el mayor tiempo del equipo de datos en mantener y actualizar la data de negocio quitando así la posibilidad de dedicar estos esfuerzos a nuevos desarrollos y generar nuevos productos.

Para ello este trabajo se centrará en el diseño de un flujo ágil, automatizado, escalable y medible desde la obtención de datos, tratamiento, almacenamiento y disponibilidad de la información permitiendo el monitoreo y detección de errores a lo largo del proceso con el fin de generar un ciclo de mejora continua garantizando la integración de nuevos orígenes de datos.

Marco Conceptual

Antecedentes

Hoy día existe una gran variedad de empresas que ofrecen la opción de visualización de contenido de video por medio de una suscripción o un modelo de compra/renta de películas, series o canales en vivo para el mercado mexicano.

Datos disponibles

Los usuarios de las plataformas de video hoy pueden realizar reproducciones de video desde un sitio web, televisores, dispositivos móviles, tabletas digitales, decodificador o receptor de video. prácticamente desde cualquier dispositivo de video como se muestra en la Figura 1.

Cada canal de comunicación va generando y recopilando información de la interacción del cliente con la aplicación desde cada dispositivo. Generando una gran cantidad de datos que pueden ser analizados para conocer comportamientos de compra, preferencias, tendencias de visualización que pueden servir para la generación de estrategias comerciales dirigidas o creación de nuevos productos con base en las preferencias de sus clientes más rentables. En pocas palabras conocer su mercado y el comportamiento de sus productos.



Figura 1. Datos generados o disponibles del usuario de una plataforma de vídeo.

Tipos de Datos

Ahora bien, estos dispositivos generan datos en diferentes formatos estructurados y no estructurados.

¿Pero qué son datos estructurados y no estructurados?

Los datos estructurados de acuerdo TIBCO¹ (empresa estadounidense de software de inteligencia empresarial), Se llama *datos estructurados* cuando los datos están en un formato estandarizado, tienen una estructura bien definida, cumplen con un modelo de datos, siguen un orden persistente y son de fácil acceso para humanos y programas.

Este tipo de datos generalmente se almacena en una base de datos que contienen datos relacionados y organizados en tablas que están compuestas de filas (valores) y columnas (atributos) como una hoja de cálculo.

Los datos estructurados tienen los mismos atributos para todos sus valores por ejemplo en la Tabla 1 se muestra un ejemplo de una tabla con datos de las frutas inventariadas en un sistema de supermercado donde se puede tener un id numérico, descripción, color y cantidad disponible donde todos los datos que se guarden en esta estructura tendrán estos atributos.

Id	Descripción	Color	Cantidad Disponible
1	Manzana	rojo	100
2	Plátano	amarillo	50
3	Pera	verde	75
4	Fresa	rojo	150

Tabla 1. Ejemplo de Tabla de datos de una base de datos relacional

Estas tablas tendrán valores en común, siguiendo el mismo ejemplo podemos agrupar las frutas por color.

También otra característica de estos datos es que al tenerlos organizados por tipo de dato se pueden realizar fácilmente operaciones matemáticas o cálculos en atributos cuantitativos por ejemplo sumar el total de frutas en el inventario por la columna de cantidad disponible.

¹ TIBCO. (2023, Enero). ¿Qué son los datos estructurados?. <https://www.tibco.com/es/reference-center/what-is-structured-data>. Recuperado el 16 junio 2023 de <https://www.tibco.com/es/reference-center/what-is-structured-data>

Algunos otros ejemplos de estos datos pueden ser:

- Archivos de Excel
- Bases de datos SQL
- Datos de puntos de venta
- Resultados de formularios web
- Etiquetas de optimización para motores de búsqueda
- Directorios de productos
- Control de inventarios
- Sistema de Nómina

Mientras que los *datos no estructurados* TIBCO 1 los define como aquellos que no se ajustan a ningún otro modelo y no tienen una estructura fácilmente identificable. No tiene organización y no se puede almacenar de ninguna manera lógica. Los datos no estructurados no encajan en ninguna estructura de base de datos, no tienen reglas ni formato, y los programas no pueden utilizarlos fácilmente. Algunos ejemplos de estos datos son el video, audio, archivos de texto, documentos e imágenes.

Normalmente estos datos necesitan diferentes tecnologías y estrategias para llevar a cabo un análisis eficiente de los mismos.

Existe también una combinación llamada *datos semi estructurados* que son aquellos datos que no están estructurados en filas y columnas, pero si están bien definidos, por lo que su procesamiento no podrá resolverse de manera estándar. Estos datos poseen etiquetas internas o datos sobre ellos mismos que describen su contenido, así como la relación entre ellos y pueden ser aceptados bajo convención. Algunos orígenes de este tipo de datos pueden ser los emails e información proveniente de una red social los cuales pueden almacenarse en bases de datos llamadas no relacionales.

Estas bases de datos no relacionales son un conjunto de datos que a diferencia de las bases de datos relacionales están guardadas en una organización de claves (etiquetas) y valores, pero no guardan una relación entre un identificador físico que sirva para relacionar dos o más conjuntos de datos.

El siguiente es un ejemplo bajo un formato llamado JSON (JavaScript Object Notation), el cual fue inspirado por la sintaxis de JavaScript (un lenguaje de programación usado para desarrollo web).

El formato JSON es un formato de texto sencillo para el intercambio de datos conformado por etiquetas y valores.

```
{  
  
  "tamano": "grande",  
  
  "precio": 15.67,  
  
  "toppings": ["champiñones", "pepperoni", "queso"],  
  
  "queso_extra": false,  
  
  "delivery": true,  
  
  "cliente": {  
  
    "nombre": "Erendira Reyes",  
  
    "telefono": null,  
  
    "correo": "erendira_rr@email.com"  
  
  }  
  
}
```

Tiempo de generación y procesamiento del dato

Al sistema formado por la combinación clientes, servicio(s), infraestructura y aplicaciones. Debemos agregar el factor tiempo de cómo se generan estos datos a través de los procesos de negocio o de sistemas ya que algunos datos se generan en tiempo real como lo es la navegación o la propia visualización ya que se generan y procesan datos de manera continua de un contenido a través de la plataforma.

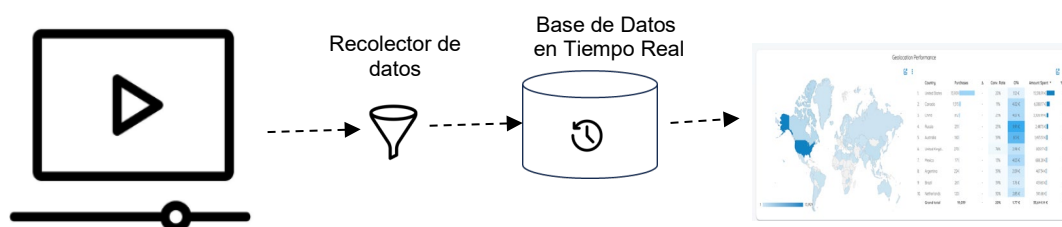


Figura 2. Ejemplo de diagrama de Flujo de Información en tiempo real

En la Figura 2, se muestra un con líneas punteadas para indicar que constantemente está pasando información por periodos muy cortos de tiempo a un recolector de datos que es un programa informático que se encarga de estar monitoreando cada que se reciben datos y los envía a la base de datos de tiempo real la cual puede alimentar a un tablero de datos de la compañía por ejemplo para monitorear la audiencia de un evento de transmisión en vivo como podría ser un partido de fútbol.

De un extremo el usuario está visualizando el contenido y durante el evento se está revisando cuantos usuarios se encuentran reproduciendo el evento de tal manera que le puede permitir a la empresa medir su audiencia del partido para ver si lo vieron completo o dejaron de visualizarlo luego del evento de medio tiempo.

Otros datos son generados por procesos que pueden necesitar más tiempo para su procesamiento como lo puede ser información proveniente de un tercero o de otro sistema como el de facturación o el proceso contable que corre de manera nocturna al cortar el día de venta.

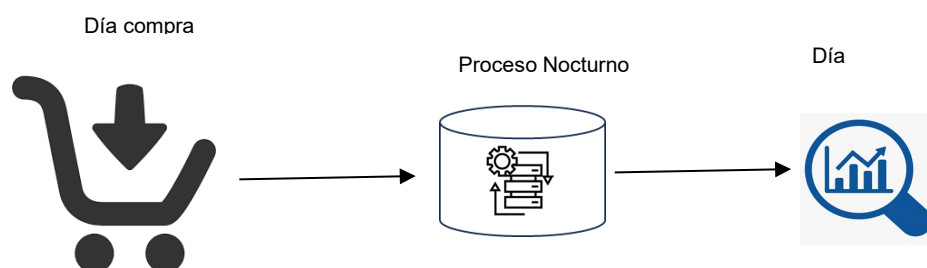


Figura 3. Ejemplo de un procesamiento por lotes.

En la Figura 3 se muestra un ejemplo de un flujo del proceso contable el cual se procesa de manera nocturna y se visualiza en la reportería luego de terminar este proceso.

Este tipo de procesos que requieren un poco más de procesamiento para estar disponibles se conocen como procesos por lotes, llamados así porque se procesan por lotes o bloques de información.

Análisis de Datos

¿De qué sirve tener una gran cantidad de datos disponibles?, ¿Qué hacer con estos datos? y ¿Cómo podemos obtener valor?

El análisis de datos es una práctica que nos ayuda a generar información útil que se pueda utilizar para tomar decisiones con conocimiento de causa. Siendo información un conjunto organizado de datos relevantes.

Los datos se convierten en información agregando valor cuando se categorizan, se añade contexto, cálculos, correcciones y/o se concentran.

Estos análisis pueden dar origen a nuevos productos, generar nuevos nichos de mercado o incluso apoyar al cambio o definición de procesos empresariales basados en el resultado del análisis de los datos.

Un ejemplo de esto en una empresa comercial podría ser la segmentación de clientes la cual se puede generar realizando un análisis de la frecuencia de compra, tendencia en cierto producto(s), monto gastado en cada producto y si sumamos además las características del propio consumidor como edad, género, ubicación o domicilio de la compra e incluso la época del año de la compra.

Podríamos encontrar algún patrón de consumo que permita clasificar o agrupar a los clientes para realizar campañas de venta dirigida y poder medir el impacto de estas.

Otro ejemplo puede ser la creación de nuevos productos por ejemplo en el año 2013 la empresa Netflix luego de analizar el éxito obtenido en la oferta de algunas series y películas de determinados actores, así como las tramas de los mismos contenidos. Inicio a generar contenido basado en el análisis de sus datos. En este caso produjo la serie titulada “House of Cards” inspirada en una serie británica retransmitida por la BBC en el año 1990.

Esta producción se basó en el análisis de los datos de preferencia de los actores de reparto con mayor éxito en visualizaciones e incluso el contenido de los capítulos el guion y detalles de la grabación se desarrolló con base en la preferencia de los usuarios de la plataforma con ayuda de un algoritmo que pudo predecir el éxito de cada capítulo incluso antes de ser grabado. ²

² Rodríguez, S., & Balmaceda, A. (n.d.). *House Of Cards: el éxito que Netflix creó gracias a la inteligencia artificial*. Big Data Magazine. Recuperado el 15 noviembre 2024 de <https://bigdatamagazine.es/house-of-cards-el-exito-que-netflix-creo-gracias-a-la-inteligencia-artificial>

El análisis de datos pasa en general por varias fases iterativas:

- 1) Identificar el caso de negocio a resolver, hipótesis a validar o qué problema se desea resolver.
- 2) Recopilación de datos de las diferentes fuentes que se necesitan para atender el punto anterior.
- 3) Limpieza de datos para su análisis con el fin no tener datos incompletos, duplicados, datos atípicos, sin categorizar o no homogéneos.
- 4) Análisis de datos utilizando diversas herramientas técnicas matemáticas, gráficas y computacionales.
- 5) Interpretación de los resultados para ver en qué medida los datos responden a tu pregunta del punto 1 como por ejemplo ¿Si se comprueba la hipótesis? ¿Que ocasiona el problema? ¿Qué recomendaciones se puede realizar como plan de acción para atender el problema? o si existe algún pronóstico de que vuelva a suceder.

Los puntos 4 y 5 dependen del tipo de análisis a realizar.

Existen diferentes tipos de análisis de datos con técnicas estadísticas y/o lógicas que apoyen para la toma de decisiones, a continuación, se mencionan algunos tipos:

Análisis descriptivo

El objetivo de este análisis es realizar un estudio de datos basado en los hechos del pasado. Su finalidad no es explicar el ¿por qué? han sucedido, ni la explicación de causa - efecto.

Busca solo describir de una manera fácil de digerir incluyendo mediciones y patrones describiendo estos como una serie de sucesos o variables recurrentes.

Técnicas utilizadas:

- 1) Agregación de datos que incluye la recolección y presentación en formato resumido de los datos.
- 2) Minería de datos para el análisis y exploración de datos en busca de patrones o tendencias.

Análisis de datos exploratorio

El objetivo de este tipo de análisis es realizar un descubrimiento de datos para descubrir patrones, detectar desviaciones o datos atípicos para probar hipótesis con la ayuda de estadísticas resumidas y representaciones gráficas.

Este tipo de análisis ayudan a descubrir relaciones entre mediciones, así como la identificación de conexiones entre atributos o eventos.

Técnicas utilizadas: Herramientas visuales para mostrar datos como mapas de calor, Histogramas, Gráficas de líneas, Pictogramas y Diagramas de dispersión

Análisis de datos de diagnóstico

Este tipo de diagnóstico busca entender ¿Por qué? ha ocurrido algo. Su principal objetivo es identificar y responder a las anomalías en los datos.

También sirve para identificar que impulsa los resultados positivos o esperados.

Técnicas utilizadas: Teoría de la probabilidad, el análisis de regresión, el filtrado y el análisis de series temporales.

Análisis predictivo

El análisis predictivo consiste en estudiar los datos históricos y actuales para hacer predicciones sobre el futuro en función de patrones encontrados en datos.

Técnicas utilizadas: Utiliza una mezcla de técnicas matemáticas, estadísticas y de aprendizaje automatizado.

Análisis prescriptivo

El análisis prescriptivo está basado en la recopilación y análisis de datos actuales e históricos para estimar la probabilidad de que algo suceda permitiendo evaluar diferentes caminos de acción y decisión buscando la mejor solución para la toma de decisiones.

Técnicas utilizadas: Técnicas estadísticas para cuantificar el efecto de decisiones futuras, así como técnicas de simulación.

Tubería de Datos

Los datos no siempre se generan en condiciones para poder tomar decisiones es por ello por lo que se debe de procesar para convertir estos datos en información.

Para ello se debe llevar a cabo procesos que incluyen una serie de herramientas y tecnologías informáticas que permiten llevar a cabo las tareas de recolección, codificación, procesamiento, limpieza y validación de datos, así como procesos estadísticos o modelos matemáticos para encontrar tendencias y poder detectar y/o resolver problemas o casos de negocio por medio del análisis de información.

Las Tuberías de Datos o tuberías por su nombre en español analógicamente funcionan como una tubería de agua que al abrir el grifo sale agua limpia (Figura 4). Pero para que eso suceda se requiere una instalación y procesamiento del agua desde su origen que puede ser un lago hasta que se vuelve potable y llega a su destino en este caso nuestros hogares.

Esto es lo mismo que pasa con los datos en las empresas. Donde normalmente las personas de negocio toman los datos ya transformados en información para la toma de decisiones, pero no todos piensan, se enteran o están conscientes de los procesos que deben limpiar, homologar, transformar, categorizar y dejar disponibles estos datos garantizando la confiabilidad y consistencia necesaria para realizar distintos análisis que permitan generar información para la toma de decisiones.



Figura 4. Una tubería de agua es una analogía de un Pipeline de Datos

Hewlett Packard Enterprise, define a Las Tuberías de Datos como el proceso para mover datos de un origen a destino que puede ser un lago de datos o almacén de datos.³

Podemos definir a una tubería de datos como una construcción lógica de fases que mueven o transportan datos normalmente con ayuda de herramientas tecnológicas.

3

¿Qué son Las Tuberías de Datos?.Glosario.(n.d.).Hewlett Packard Enterprise (HPE). Recuperado 15 noviembre 2024 de <https://www.hpe.com/mx/es/what-is/data-pipelines.html>

Beneficios de las Tuberías de Datos

Las Tuberías de Datos permiten consolidar datos desde diferentes orígenes y transformarlos para su análisis.

Mejoran la calidad de los datos integrando procesos repetibles de limpieza, refinamiento, estandarización de formatos y homologación tipos de datos que garanticen la integridad y consistencia de datos.

Entendiendo como integridad de datos como el aseguramiento de la no alteración o pérdida de datos durante su transporte.

Mientras que la consistencia de datos garantiza la ausencia de diferencias entre la fuente y destino.

Mejora la trazabilidad, ya que al tener identificados todos los procesos que son requeridos para tener información se puede incluir controles para poder dar seguimiento y detectar cuellos de botella que pueden ser trabajados en un ciclo de mejora continua.

Tipos de Tuberías de Datos

A continuación, se explican los dos tipos de tuberías definidas por su tipo de procesamiento.

Tuberías de procesamiento de flujos

El flujo de datos es una secuencia progresiva y continua de mensajes o eventos. Este tipo de flujos procesan estos eventos para un análisis en tiempo real.

Este tipo de flujos requieren tiempo de respuesta rápida entre el emisor y el receptor, pero deben tener una alta tolerancia a errores en la secuencia o pérdida de paquetes y estos errores deben ser monitoreados para ser detectados.

Tuberías de procesamiento por lotes

Este tipo de procesamiento permite generar estrategias como procesos en paralelo, una mejor gestión en tiempos y capacidad de procesamiento. En caso de que se produzca algún error es posible volver a procesar el lote completo.

Fases de una Tubería de Datos

Las tuberías de datos normalmente tienen los siguientes pasos o fases ilustradas en la Figura 5.



Figura 5. Fases principales de una tubería de datos

Origen de Datos

Antes de analizar datos es importante identificar y definir en dónde, cómo y cuándo se generan los datos que se requiere analizar, así como el tipo de dato ya que esto nos permitirá generar la mejor estrategia para obtener y analizar los mismos.

Como ya se ha mencionado en este documento, podemos encontrar diferentes fuentes de datos que guardan o generan información relacionada con el cliente y su interacción con la compañía y sus servicios.

Estos datos pueden provenir de Sistemas Gerenciales, Sistemas de Nómina, Sistemas de Compra/Venta, Archivos de Excel, Archivos de Texto, Información de redes sociales, Emails, Imágenes, Archivos PDF, Audios, entre otros.

En esta fase es primordial identificar si los datos radican en la Nube de algún proveedor o bien se encuentran en sistemas con infraestructura tecnológica propia.

Un entorno propio o como algunos lo conocen como “On Premise”, este entorno se encuentra en un centro de datos físico propio de la empresa o arrendado exclusivamente por y para la misma empresa. Tiene la particularidad que si la arquitectura actual no está siendo utilizada por ejemplo para el 1% de uso tanto el costo de adquisición como de mantenimiento es el mismo que si se utilizara el 100%. Otro dato para analizar es que si el negocio cambia y requiere una mayor necesidad de procesamiento debe comprar o adquirir más infraestructura y mantener sus costos fijos todo el tiempo esté o no en uso.

Mientras que una arquitectura en Nube se refiere a un entorno tecnológico arrendado a un tercero(s) que corre sobre internet y que mediante servicios virtuales ofrece procesamiento, memoria, almacenamiento y herramientas que físicamente puede vivir en muchos centros de datos distribuidos en el mundo, pero a diferencia del entorno propio no pertenece a la empresa. En este esquema la empresa paga por el uso y tiempo que utilice esa infraestructura y una de las principales ventajas es que se puede crecer o decrecer según sea la necesidad de la empresa en cualquier momento que lo desee y solo pagará lo que utilice.

Actualmente existen varios proveedores de servicios en la Nube como Amazon con su plataforma AWS o Google con su plataforma de GCP.

Cabe mencionar que existe la posibilidad de tener servicios contratados en varias nubes conviviendo en una misma arquitectura con varios Proveedores a esto se le llama Nube Híbrida.

Extracción de datos

En esta fase involucra todos los procesos necesarios para extraer los datos para su tratamiento y posterior análisis. Dicho procesamiento dependerá del origen, tipo de dato, así como de la velocidad de actualización del dato y la frecuencia con que se desea analizar estos datos.

Aquí es importante identificar si se requiere analizar datos en tiempo real o bien si conviene la carga de datos por lotes. Con el fin de definir la arquitectura tecnológica más adecuada y generar una estrategia de consumo acorde a la necesidad de la empresa y su infraestructura tecnológica.

Procesamiento y Transformación de datos

En la fase de procesamiento y transformación se limpian, agregan, modelan y optimizan los datos para su análisis e incluso puede que se creen nuevos datos provenientes de agregaciones o cálculos generados con los datos insumo, para ello es primordial la definición de reglas de negocio para estos cálculos, agrupaciones o transformaciones.

En esta fase normalmente se incluyen tareas de limpieza como eliminación de espacios, eliminación de datos duplicados, atípicos e identificación de datos incoherentes por ejemplo si en el atributo de teléfono se espera solo caracteres numéricos no podrían incluir letras.

Otra tarea que se incluye en esta fase es la estandarización y normalización de las estructuras formatos de datos por ejemplo Año/Mes/Día para las fechas.

Es importante definir el modelo de datos y estructura(s) de datos acorde al tipo de análisis requerido, modelo de negocio y al tipo de dato para analizar.

Modelo de Datos

El modelado de datos permite entender los datos para tomar la mejor decisión tecnológica para su almacenamiento y gestión.

Así como un arquitecto diseña planos antes de tener una casa, también se diseña un modelo de datos para crear una solución de negocio.

Normalmente se inicia generando una representación conceptual de la solución y después una representación que incluye las tecnologías propuestas para cada tarea.

Existen varios tipos de modelos que pueden ser utilizados, a continuación, se mencionan algunos:

- Modelo conceptual

Estos modelos ofrecen una visión general de los datos. Es una representación gráfica simple que ayude a las partes interesadas, ya sean técnicas o no, a compartir una visión común y a ponerse de acuerdo acerca del propósito, el alcance y el diseño de su proyecto de datos.

- Modelo de datos lógico

Son modelos donde se asignan más detalles sobre los conceptos de datos y las relaciones de datos complejas que se identificaron en el modelo conceptual como tipos de datos, relaciones entre entidades o atributos primarios o campos clave.

Estos modelos lógicos sirven de puente entre el modelo de datos conceptual y la tecnología. Deben ser independientes a alguna tecnología o herramienta, pero si son la base para la toma de decisión de las herramientas tecnológicas a utilizar.

- Modelo físico

Los modelos físicos son una fusión entre los modelos de datos lógicos y la tecnología a utilizar.

Incluyen también los detalles técnicos como tipo de datos, relaciones y otros datos como variables de rendimiento. AWS (2024)

Almacenamiento de datos

El almacenamiento de datos indica dónde se estarán guardando o depositando los datos de la fase anterior listo para su consumo y análisis.

Así como por cuánto tiempo se guardarán estos datos.

En caso de estar implementando la tubería de datos sobre alguna nube con un proveedor. Es importante revisar las políticas de los diferentes esquemas de almacenamiento, así como los costos de estos y el tiempo para dejar disponibles los mismos en caso de ser requeridos.

Otro factor para tomar en cuenta es la regulación legal del país donde se opera la compañía que posee los datos desde temas de protección de datos personales hasta el tiempo que por resolución deberá ser resguardado por la compañía, así como las propias políticas que dictamine el área de gobierno de datos de la compañía.

Análisis y Visualización de datos

Este es el producto final y es en esta fase donde se genera realmente los diferentes tipos de análisis que se desea ya con la información procesada y optimizada por medio de modelos matemáticos o estadísticos con ayuda de visualizaciones como gráficos que nos ayuden a detectar tendencias o comportamientos por medio de un tablero de mando o un reporte para la toma decisiones.

Las Tuberías de Datos son necesarios ya que no se recomienda analizar los datos en los mismos sistemas origen donde se crean debido a que son procesos que suelen ser muy costosos computacionalmente, por lo que se separa para evitar perjudicar el rendimiento del sistema donde se están generando los datos. De esta forma, tenemos sistemas encargados de capturar y crear datos, y sistemas encargados de analizar los datos.

Data Driven

Después de analizar el concepto de Tubería de Datos nos damos una idea de todos los procesos que deben aplicar las empresas para generar información para la toma de decisiones o generación de estrategias comerciales.

Es fácil imaginar que por su naturaleza tecnológica una empresa de procesamiento de video en tiempo real tiende a ser una empresa con un modelo de toma de decisiones basado en el análisis e interpretación de datos, es decir que sus procesos de negocio y operativos se rigen por la toma de decisiones estratégicas basadas en los datos por lo que se podría definir como una empresa Data Driven.

Lo anterior requiere una gran capacidad de compartir, capturar y organizar información de distintas áreas del negocio, lo que facilita el trabajo colaborativo e incrementa la agilidad de sus procesos (más rápidos y eficientes).

Para el desarrollo del presente caso de estudio tomaremos el caso de una empresa mexicana de procesamiento de video en tiempo real que quiere convertirse en una empresa Data Driven para garantizar su permanencia en el mercado con respecto a sus competidores en México.

En la Figura 6. Se muestra el gráfico del Top 8 de las Empresas de procesamiento de video en tiempo real que ofrecieron sus servicios en el año 2022 liderado por Netflix, Prime video y Disney.

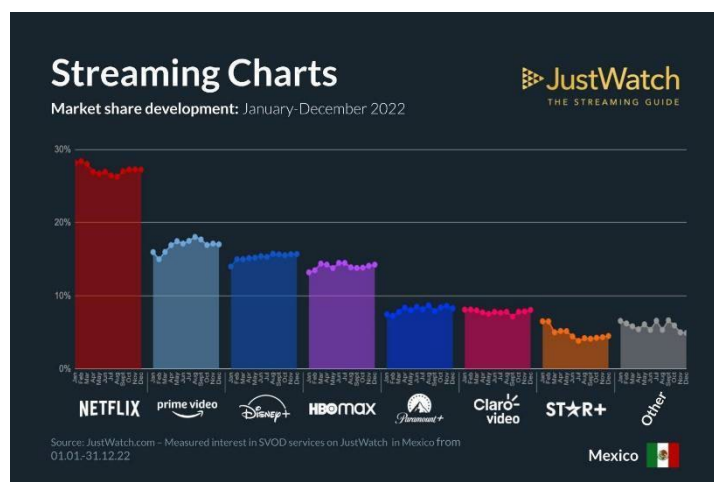


Figura 6. “Top de Empresas de procesamiento de video en tiempo real que ofrecieron servicios al mercado mexicano en 2022”, imagen tomada de Just Watch empresa que proporciona un servicio que sirve para encontrar dónde se puede ver de forma legal cada película o serie en 11 países como Alemania, EE. UU., México y España⁴

⁴ Gunn, J. (2023, February 13). *Revisión de desempeño de Netflix, Prime Video, etc. en México*. Robotto.mx. Recuperado 15 febrero 2024 de <https://robotto.mx/2023/02/revision-de-desempeno-de-netflix-prime-video-etc-en-mexico/>

DataOps

Las estrategias y productos comerciales basados en el análisis de la información clave de negocio, así como el uso de tecnologías basadas en el aprendizaje automático puede generar una gran ventaja competitiva para las empresas ya que permite conocer el comportamiento y tendencias de compra de nuestros clientes, así como la detección del éxito de una campaña, producto o estrategia comercial.

Sin embargo, el proceso de recopilación, limpieza, preparación y procesamiento de datos muchas veces suele ser manual, no optimizado y propenso a errores. Esto aumenta el tiempo que tardan los datos en generar valor para el negocio. Para reducir este tiempo, las empresas están buscando la automatización y validación de los procesos de datos.

A medida que los datos siguen creciendo tanto en cantidad como en complejidad, crece la necesidad de tener un proceso eficaz y eficiente para aprovechar el valor de esos datos.

De acuerdo con el libro “DataOps Cook Book” los analistas, científicos de datos e ingenieros de datos pueden dedicar el 80 % de su tiempo laboral a actualizar, mantener y asegurar la calidad de la canalización de datos siendo este trabajo no percibido por el área de negocio es decir se asume que existe este proceso y se da por hecho que sucede pero las peticiones de nuevos requerimientos sigue creciendo con igual o mayor proporción que las necesidades de mantener los engranajes operativos de estos procesos que hacen posible que los datos se puedan utilizar para dichos requerimientos de negocio.

Definición

DataOps es la contracción de los términos “Data” & “Operaciones”, se podría definir según el artículo “DataOps, una alternativa para la gestión de Data Pipelines” de la revista Ibérica Risti como un conjunto de prácticas, procesos, métodos de ingeniería de software ágil y tecnologías que combina una perspectiva integrada y orientada a los procesos sobre datos con automatización (Erath, 2018) , el cual se enfoca en la gestión de datos, comunicación, colaboración, integración, automatización y la medición continua de la cooperación entre los ingenieros de datos con los científicos de datos, sin descuidar a los demás miembros de los equipos (Palmer, 2015); dicha integración se basa en crear y mantener una relación bidireccional entre los objetivos de la gestión de datos con el consumo de datos (Whitmore, 2017).⁵

⁵ Rueda, L. V., Díaz, S. o., & Sanabria, J. S. G. (2023, June 16). YouTube. Recuperado 10 octubre 2024 de <https://www.risti.xyz/issues/ristie38.pdf>

Gartner lo define como una práctica de gestión de datos colaborativa enfocada en la integración y automatización de flujo de Datos entre los gestores y los consumidores de datos en una organización. El objetivo de DataOps es crear entregas predecibles y una gestión de cambio en los datos, modelos de datos y artefactos relacionados. DataOps utiliza tecnología para automatizar la entrega de datos con un nivel apropiado de seguridad, calidad y metadatos para mejorar el uso y el valor de los datos en un entorno dinámico.⁶

En pocas palabras el objetivo de DataOps, es proporcionar nuevos modelos de trabajo y colaboración para generar valor a las empresas más rápidamente utilizando la tecnología para automatizar el diseño como la gestión y la distribución de los datos con los niveles adecuados de gobernanza y emplea los metadatos para mejorar la usabilidad y el valor de la información en entornos dinámicos.

Origen

El 19 de junio de 2014, Lenny Liebmann, editor colaborador de Information Week, presentó el término DataOps en una publicación de blog en IBM Big Data & Analytics Hub titulada "Tres razones por las que DataOps es esencial para el éxito de Big Data" para explicar la necesidad de una práctica que permitiera mayor probabilidad de éxito a las iniciativas de Big Data.

Es Andy Palmer de Tamr y Steph Locke quienes lo popularizaron y comenzó a ganar terreno en 2017 al incorporar con éxito un enfoque DevOps a los datos, este último término lo explicaremos más adelante, pero es un enfoque que fusiona el ambiente de desarrollo con el de operaciones en busca del desarrollo, integración y publicación continua de código dentro de un ambiente productivo.

Además, hay que recordar que en esta línea de tiempo el desarrollo ágil de sistemas ya llevaba 10 años en el mercado de una manera exitosa ya que incorporó la iteración y el desarrollo evolutivo.

6

DigitalIberia369. (2021, June 24). Francisco Gaitán Díaz. Recuperado 5 diciembre 2023 de

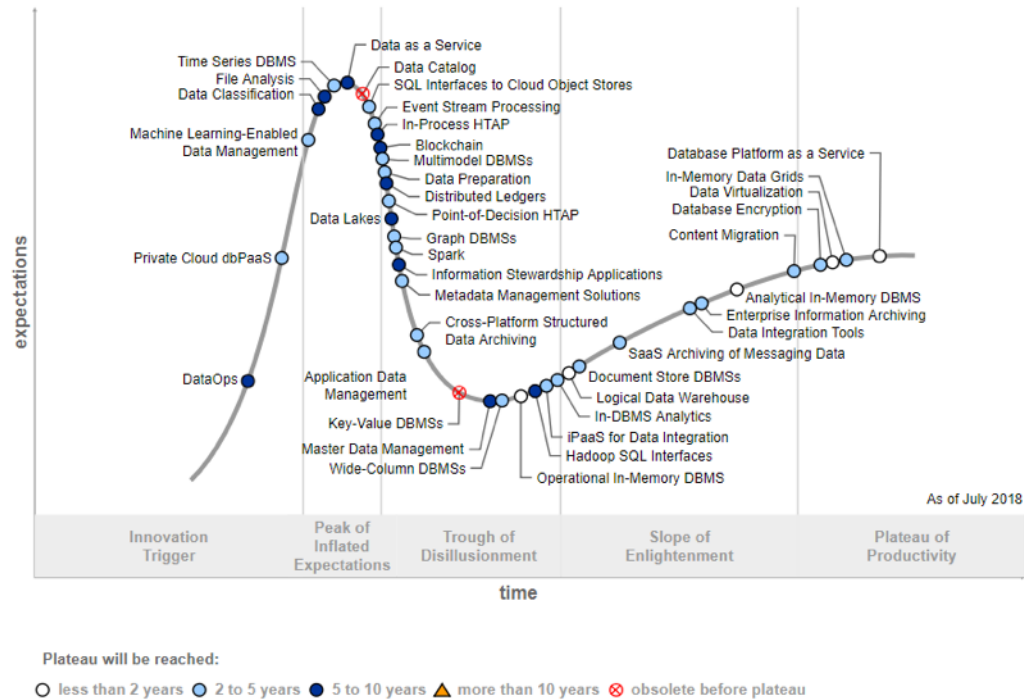


Figura 7. Gartner nombró a DataOps en su gráfico de madurez y adopción de tecnologías para la gestión de datos en 2018.⁷

DataOps termina combinando Desarrollo Agile, Lean Thinking y DevOps al desarrollo y operación del análisis de datos por lo que a continuación se explica un poco más sobre estos conceptos a fin de entender cómo se forma DataOps.

⁷ Prieto, M. (2018, October 10). *Las 5 tendencias en el Hype Cycle de tecnologías emergentes 2018 de Gartner*. smartlighting. Recuperado 15 noviembre 2024 de <https://smart-lighting.es/tendencias-tecnologias-emergentes-2018-gartner/>

Desarrollo Agile

El Desarrollo Agile es una forma o una manera de trabajar y organizarse que permite al equipo de trabajo descomponer el producto final en pequeñas partes que deben completarse y entregarse en un corto tiempo.

Lo anterior con el objetivo de desarrollar productos y servicios de calidad que satisfagan las necesidades de clientes cuyas prioridades cambian con la velocidad cada vez más rápida.

Esta metodología se utiliza para el desarrollo de sistemas o para desarrollo de proyectos de alto rendimiento ya que es flexible y se centra en la implementación rápida para planear el flujo de trabajo brindando la capacidad de elegir la mejor opción en cada situación sin comprometer el proyecto final.

Esta forma de trabajar nació en el sector de desarrollo de sistemas ya que es un sector muy cambiante y en la forma de trabajo y administración tradicional al generar un contrato cerrado con escasa comunicación entre las personas del proyecto conducía a un entregable de mala calidad y sobre todo que en el tiempo ya no era funcional o requerido por el usuario dado que su necesidad ya había cambiado.

En el año 2001, se reunieron los altos ejecutivos de las principales empresas de 'software' en Utah. Allí pusieron en común las mejores prácticas de cada compañía y crearon el 'Manifiesto Agile'. Un modelo de mejora continua en el que se planifica se crea, se comprueba el resultado y se mejora. Algo que es constante y rápido, con plazos de entregas reducidos que buscan evitar la dispersión y centrar toda la atención en una tarea encomendada.⁸

En la reunión llegaron a cuatro valores, los cuáles fueron pauta para el desarrollo del Manifiesto Ágil los cuales son los siguientes:

1. Individuos e interacciones sobre procesos y herramientas
2. Software funcionando sobre documentación extensiva
3. Colaboración con el cliente sobre negociación contractual
4. Respuesta ante el cambio sobre seguir un plan

Bajo estos valores es como surgen los principios del Manifiesto Ágil.

Principios del Manifiesto Ágil⁹

- Nuestra mayor prioridad es satisfacer al cliente mediante la entrega temprana y continua de software con valor.
- Aceptamos que los requisitos cambien, incluso en etapas tardías del desarrollo. Los procesos ágiles aprovechan el cambio para proporcionar ventaja competitiva al cliente.
- Entregamos software funcional frecuentemente, entre dos semanas y dos meses, con preferencia al periodo de tiempo más corto posible.
- Los responsables de negocio y los desarrolladores trabajamos juntos de forma cotidiana durante todo el proyecto.

- Los proyectos se desarrollan en torno a individuos motivados. Hay que darles el entorno y el apoyo que necesitan, y confiables la ejecución del trabajo.
- El método más eficiente y efectivo de comunicar información al equipo de desarrollo y entre sus miembros es la conversación cara a cara.
- El software funcional, es la medida principal de progreso.
- Los procesos ágiles promueven el desarrollo sostenible. Los promotores, desarrolladores y usuarios debemos ser capaces de mantener un ritmo constante de forma indefinida.
- La atención continua a la Excelencia técnica y al buen diseño mejora la agilidad.
- La simplicidad, o el arte de maximizar la cantidad de trabajo no realizado, es esencial.
- Las mejores arquitecturas, requisitos y diseños emergen de equipos autoorganizados.
- A intervalos regulares el equipo reflexiona sobre cómo ser más efectivo para a continuación ajustar y perfeccionar su comportamiento en consecuencia.

Finalmente, tras la reunión mencionada se creó “The Agile Alliance”¹⁰ dedicada a promover el desarrollo ágil de software y ayudar a las empresas que lo adoptaran.

⁸ La reunión se celebró en febrero de 2001, entre el 11 y el 13 de ese mes, en Snowbird, una estación de ski muy conocida, situada en las montañas Wasatch. El manifiesto lo firmaron 17 autores, (Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, Dave Thomas) a partir de entonces pasaría a conocerse como «ágiles», como eXtreme Programming, Crystal Clear, Pragmatic Programming, DSDM o ASD.

⁹ (n.d.). Agile Alliance. Recuperado 15 noviembre 2024 de <http://www.agilealliance.org/>

¹⁰ Tena, M. (n.d.). ¿Qué es la metodología 'agile'? ¿Revolución de las formas de trabajo? BBVA. Recuperado septiembre 2024 de <https://www.bbva.com/es/innovacion/metodologia-agile-la-revolucion-las-formas-trabajo/>

Lean Thinking

Antes de hablar de *Lean thinking*, es importante mencionar *Lean Manufacturing* o *Lean Production* la cuál es una metodología de gestión de trabajo enfocado en mejorar la comunicación y el trabajo en equipo para poder entregar a los clientes un producto o servicio de valor.

En español, se puede traducir como fabricación esbelta, fabricación ajustada, fabricación ágil, pero lo más común es el término anglosajón lean manufacturing.

El principio central del Lean Manufacturing es la reducción y eliminación del desperdicio, tratando de eliminar o reducir todas las actividades que no añadan valor dentro en el proceso de producción. Se basa en los sistemas de producción como Calidad Total, Justo a tiempo, Mejora Continua, Teoría de las restricciones y reingeniería de procesos.

En 1913, Henry Ford inició a combinar las piezas intercambiables con el trabajo estándar y el transporte en movimiento para crear lo que denominó producción en cadena. Con esto era capaz de hacer girar los inventarios de toda la empresa en pocos días, pero era incapaz de ofrecer variedad.

Así prácticamente todas las máquinas de la Ford Motor Company trabajaban con un único número de piezas, y prácticamente no había cambios.

El problema surgió cuando llegó la necesidad de la variedad y sobre todo que los clientes exigían periodos más cortos entre modelos.

Y es en este contexto en la década de 1930 que Kiichiro Toyoda, Taiichi Ohno y otros miembros de Toyota analizaron el pensamiento original de Ford justo después de la Segunda Guerra Mundial y se les ocurrió que una serie de sencillas innovaciones podrían hacer más posible la continuidad del flujo de procesos y una amplia variedad de ofertas de productos e inventaron el Sistema de Producción Toyota.

Este sistema, en esencia, desplazó el enfoque del ingeniero de fabricación de las máquinas individuales y su utilización, al flujo del producto a través del proceso total.

Toyota llegó a la conclusión de que si se ajustaba el tamaño de las máquinas al volumen real necesario, se introducían máquinas con autocontrol para garantizar la calidad, se alineaban las máquinas en la secuencia del proceso, se iniciaban las configuraciones rápidas para que cada máquina pudiera fabricar pequeños volúmenes de muchos números de piezas y se hacía que cada etapa del proceso notificara a la etapa anterior sus necesidades actuales de materiales, sería posible obtener un bajo coste, una gran variedad, una alta calidad y unos tiempos de producción muy rápidos para responder a los deseos cambiantes de los clientes.¹¹

¹¹ (n.d.). ▷ Lean Manufacturing: Qué es y cómo implementarlo en tu empresa. Recuperado 11 julio 2023 de <https://leanmanufacturing10.com/>

Principios Lean Manufacturing

- Hacerlo bien a la primera: cero defectos, esto se debe conseguir con la detección de los problemas y su solución en el origen.
- Minimizando el derroche: excluyendo las actividades que no agregan valor al producto.
- Mejora continua: teniendo como axioma el poder garantizar la calidad del producto o servicio, tratar continuamente de aumentar la productividad, y la reducción de costes.
- Procesos “pull”: las cantidades producidas se fabrican en respuesta a la demanda (para evitar sobreproducción).
- Flexibilidad: tener la capacidad de poder fabricar variedad de códigos de productos diferentes y en cantidades diferentes a petición.
- Construcción y gestión de una relación y colaboración a largo plazo con los proveedores, llegando a acuerdos para compartir el riesgo, los costes y la información.
- Cambio del enfoque principal: al cliente no se le vende un producto, si no que al cliente se le aporta una solución.

En el año 1996, seis años después de haber lanzado al mercado su primer best-seller “The Machine that changed the world”, los autores Daniel T. Jones y James P. Womack publican su segundo libro, “Lean Thinking”. Donde se utiliza por primera vez el término, siendo este el momento a partir del cual se lo adoptaría en occidente para denominar a la filosofía del sistema de producción de Toyota.

En este libro, los autores definieron 5 principios LEAN: *valor, flujo de valor, flujo, atracción y perfección*, estos 5 principios deberían guiar las acciones de las empresas interesadas en sistemas Lean.

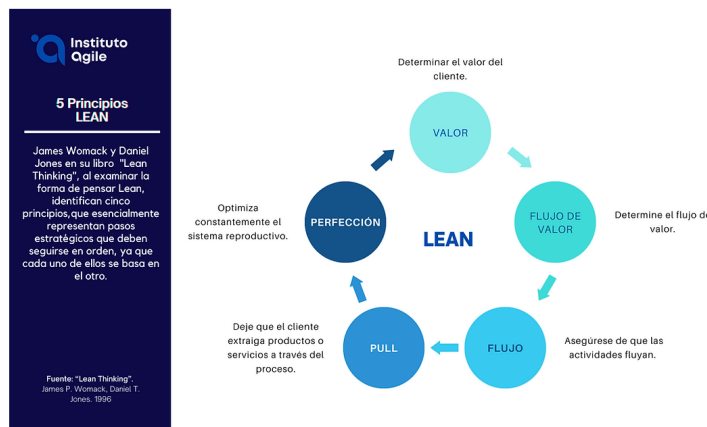


Figura 8. Se muestran los 5 principios Lean publicados en el libro Lean Thinking en 1996. (*Lean Thinking. 5 principios De Lean.*, 2022)¹¹

Principios Lean

Valor

El valor son las características o funciones de un producto o servicio que debe satisfacer las necesidades del cliente y que el cliente está dispuesto a pagar. Esta identificación de necesidades del cliente debe estar alineada con las actividades del productor.

Cadena de Valor

La cadena valor, está integrada por las actividades necesarias para transferir productos o servicios a través de la resolución del problema y la gestión de la información hasta llegar a la transformación física de la materia prima o producto final.

Para ello es importante realizar identificar y analizar los procesos para detectar mejoras en los mismos para cual normalmente se realiza un Mapeo de Flujo de Valor.

Flujo de Valor

El objetivo es desarrollar sin interrupciones y eliminando costos del trabajo produciendo productos terminados en el menor tiempo posible, garantizando la generación de valor continuo.

Para diseñar este flujo de valor, es necesario analizar con una vista sistemática a la organización completa para identificar funciones y departamentos, así como dependencias para eliminar contraflujos y paradas innecesarias.

Pull

Con base en la filosofía Justo a tiempo se adopta el proceso de extracción, solicitando la producción hasta que exista una demanda del cliente.

Este sistema hace posible reducir la cantidad de inventarios en bodegas ya que los componentes dentro del sistema solo se reemplazan cuando son consumidos por lo que el reto más grande es adaptarse a la velocidad de la demanda.

Perfección

El objetivo final es aproximarse a la perfección buscando la mejora continua siguiendo mantenimientos estandarizados, estos cambios pueden ser cambios radicales o solo un incremento para ello los líderes de la cadena de valor deben centrarse en los primeros cuatro principios para decidir qué cambio aplicar primero. (*Lean Thinking. 5 principios De Lean.*, 2022)

DevOps

Es un enfoque o forma de desarrollo de sistemas que mediante herramientas de automatización aceleran el ciclo de desarrollo de los sistemas y se centra en la integración y entrega continua de aplicaciones aprovechando y reutilizando recursos como código desarrollado y automatizando la integración, pruebas e implementación de nuevo código de una forma replicable.

El concepto DevOps surge en 2009 con la conferencia “DevOpsdays 2009” organizada por Patrick Debois en Bélgica con la idea de conocer a otras personas interesadas en la noción de poder manejar la infraestructura de TI de forma ágil. En el evento se presentó como tema emergente la combinación entre Desarrollo y Operaciones, así como la necesidad de automatizar la implementación y administración de infraestructura y aplicaciones para infraestructuras escalables.

Es por eso por lo que DevOps surge como una fusión del desarrollo y operación de sistemas que reduce el tiempo de implementación, comercialización buscando minimizar errores acortando el tiempo de detección y resolución de problemas sin tener que dar de baja el aplicativo al 100% o dejar de dar servicio a los clientes del sistema. Integrando el desarrollo ágil aplicando los principios ágiles de administración de sistemas mejorando la comunicación de las personas en operaciones y desarrollo para que colaboren en la implementación y lanzamiento de sistemas.

Las empresas que han implementado DevOps han logrado reducir el tiempo de ciclo de lanzamiento de sistemas de meses a literalmente segundos esto obviamente genera una ventaja competitiva para dichas empresas tales como Google Amazon y otras más del ramo de tecnología ya que ahora lanzan mejoras e incluso nuevas funcionalidades a sus aplicaciones muchas veces al día mejorando la calidad bajo un riesgo controlado y monitoreado dentro de un ambiente productivo que les permite tener la menor afectación posible a sus clientes por lo que DevOps ha logrado ser la base del éxito de muchas de estas empresas.

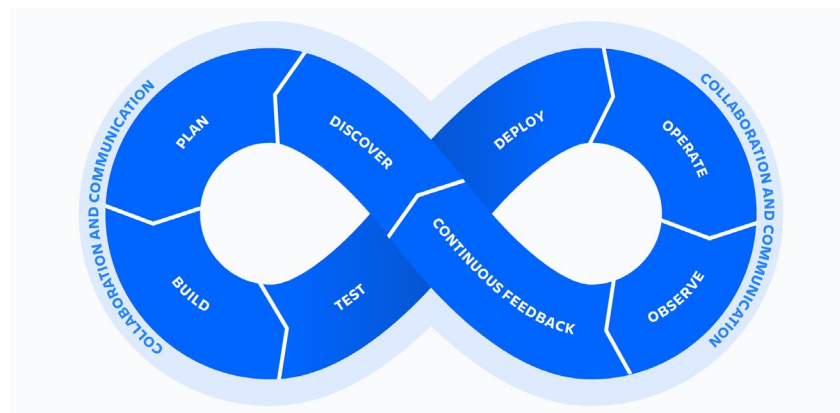


Figura 9. En la siguiente imagen se representa un flujo de desarrollo DevOps (Quijano, 2018)

Principios DevOps

Colaboración

Es uno de los principios clave que permite a desarrollo y operaciones fusionarse que se comunican colaborando durante todo el ciclo de vida de desarrollo de la aplicación.

Todos los integrantes del equipo DevOps son responsables de garantizar la entrega de calidad en cada fase del ciclo de desarrollo. Esto lleva a los desarrolladores a integrar el diseño de la experiencia del cliente en una página (front end), así como la programación y mantenimiento de la arquitectura interna del sitio (back end) por lo que la comunicación y colaboración en equipo debe ser lo primordial.

La colaboración facilita que varias personas puedan desarrollar al mismo tiempo sin interrupciones o percances en la modificación del código que se esté programando por cambios no notificados.

Automatización

La automatización es una de las características principales de DevOps ya que busca automatizar la mayor parte posible del ciclo de vida del desarrollo de aplicaciones para ahorrar tiempo en escribir código y desarrollar nuevas funciones. Siendo la automatización un elemento clave del desarrollo e integración continua ayudando a reducir los errores humanos y aumentar la productividad del equipo. Al automatizar procesos los equipos pueden lograr iteraciones más cortas en un ciclo de mejora continua llevándolos a responder cada vez más rápido a las peticiones de sus clientes.

Mejora Continua

La mejora continua es un elemento heredado de las prácticas ágiles que fomentan la fabricación ajustada y la experimentación, minimizando el desperdicio y optimizando la velocidad, costo y facilidad de entrega. La mejora continua, está relacionada con la calidad y entrega continua esto se refleja en que los equipo DevOps desarrollan y liberan continuamente versiones o actualizaciones de la aplicación que agregué nuevas funcionalidades, corrijan errores, eliminan desperdicios o aporten mayor valor al cliente, mejorando y perfeccionando en cada iteración la eficiencia de los sistemas.

Acción centrada en el cliente

Los equipos DevOps utilizan mecanismos para una retroalimentación breve con los clientes y usuarios finales ya que el enfoque de desarrollo de productos y servicios es centrarse en las necesidades de los usuarios para lo cual se requiere recopilar y dar respuesta de manera rápida mediante el monitoreo en vivo.

Siendo un sistema vivo en el cual el equipo DevOps tiene visibilidad de la interacción de los usuarios con el sistema detectando necesidades o mejoras adicionales que pueden luego convertirse en una nueva versión o actualización del sistema.

Crear con el objetivo en mente

Este principio está enfocado a desarrollar productos y servicios que realmente atiendan las necesidades o problemas reales de negocio, Para ello los equipos DevOps deben tener una comprensión completa del producto desde la creación hasta la implementación.

Lo anterior ayuda al equipo a trabajar en un producto mínimamente viable que atienda una necesidad específica o genere un valor agregado bien definido lo cual permite un mejor enfoque y seguimiento del desarrollo de este.

Alcance DataOps

El objetivo de DataOps es mejorar la calidad del análisis de datos automatizando el ciclo de procesamiento desde su extracción, preparación, agregación, gestión y desarrollo de los datos agregando la metodología agile para reducir el tiempo de desarrollo e integrando principios de lean para la reducción de desperdicios y procesamiento de acuerdo con la demanda alineándose a los objetivos de la empresa.

Principios DataOps

Satisfacer continuamente al consumidor:

La máxima prioridad es satisfacer la demanda de entrega temprana y continua de analíticas de valor.

Apreciar las soluciones de analítica eficientes:

La medida prioritaria del rendimiento del análisis de datos es el grado de detalle en el que se entregan estos análisis, incorporando información fiel, por encima de marcos y sistemas rígidos.

Abrazando el cambio:

Recibir necesidades cambiantes de los consumidores para obtener una ventaja competitiva. Mejorar la comunicación e interacción con un método más eficiente, efectivo y ágil de comunicación con los consumidores.

Es un deporte de equipo:

Los equipos analíticos deberán tener una variedad de roles, habilidades, herramientas y títulos. Una diversidad de antecedentes y opiniones aumenta la innovación y la productividad.

Interacciones diarias:

Los consumidores, los equipos de analíticas y las operaciones deben colaborar día a día a lo largo del proyecto.

Organización autónoma:

Los mejores análisis, algoritmos, arquitectura, exigencias y diseños provienen de equipos autoorganizados.

Reducir el heroísmo:

Los equipos de analítica deberían procurar reducir el heroísmo y crear equipos y procesos sostenibles y flexibles de análisis de datos.

Reflexión:

Los equipos de analíticas deben configurar su rendimiento operativo reflexionando, intervalos regulares, sobre los comentarios de sus consumidores y estadísticas de las operaciones.

Los análisis son códigos:

Al tener una gran variedad de herramientas individuales para acceder, integrar, modelar y visualizar la información. Es esencial que en estas herramientas se genere código y configuraciones que describan las acciones realizadas con los datos.

Organización:

Organización de la información, herramientas, código y entornos en todo el proceso de inicio a fin.

Realizar reproducciones:

Los resultados reproducibles son necesarios y, por lo tanto, creamos versiones de toda información, hardware de bajo nivel y configuraciones de software y el código y configuración de cada herramienta en la cadena de herramientas.

Entornos desechables:

Es importante minimizar los costos, generar entornos técnicos sencillos de crear, aislados, seguros y desechables.

Sencillez:

Un buen diseño mejora la agilidad y al mismo tiempo la sencillez. Entre más sencillo es el proceso es más sencillo controlarlo.

Analizar es producir:

Generar un proceso centrado en la eficiencia continua.

La calidad es primordial:

Se deben construir procesos capaces de detectar de forma automática los fallos y problemas de seguridad, código, configuración e información y debe proporcionar una retroalimentación continua.

Supervisar la calidad y el rendimiento:

Nuestro objetivo es tener medidas de rendimiento, seguridad y calidad que puedan ser supervisadas de forma continua para detectar patrones de comportamiento.

Reutilización:

La reutilización evita repeticiones de trabajo previo.

Mejorar la duración de los ciclos:

La idea es minimizar el tiempo y esfuerzo convirtiendo el desarrollo en un proceso repetible de producción que permita reeditar y reusar el código desarrollado.¹²

¹² *The DataOps Manifesto - Read The 18 DataOps Principles.* (n.d.). DataOps Manifesto. Recuperado 15 noviembre 2024 de <https://DataOpsmanifesto.org/es/>

Ciclo de Vida DataOps

El siguiente ciclo de vida típico de DataOps se ilustra en el documento “Good practices for the adoption of DataOps in the software industry” publicado por Manuel Rodriguez en 2020 en Journal of Physics: Conference Series.¹³

Este ciclo está basado en los principios de DevOps e ilustra el proceso iterativo por el cual se busca automatizar el flujo operativo de los datos desde su planeación, extracción, tratamiento, almacenaje, modelo (proceso de entrenamiento en caso de utilizar analítica avanzada como aprendizaje automatizado), publicación o disponibilidad de la información para consumo y por último el monitoreo de todas las fases.

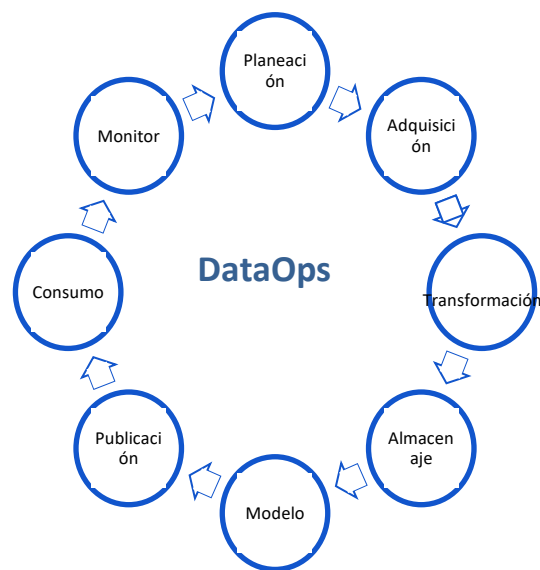


Figura 10. Ciclo de vida de DataOps, siguiendo el principio de DevOps (Rodriguez, 2020)

¹³ Manuel Rodriguez, Manuel Mazzara, & Pires de Araújo, L. J. (2020, Diciembre 18). Buenas prácticas para la adopción de DataOps en la industria del software. *Journal of Physics: Conference Series*, 1694. <https://iopscience.iop.org/article/10.1088/1742-6596/1694/1/012032/pdf>

Componentes DataOps

DataOps como parte de lo heredado por DevOps y Agile. DataOps se descompone en elementos y/o subprocesos para integrarlos en un ciclo de mejora continua y atendiendo a sus propios principios de reutilización de código automatizando el pase entre cada elemento para ello existen herramientas especializadas para gestionar cada una de las tareas.

En un flujo de DataOps existen algunos componentes principales como:

Gestión de configuraciones

La gestión de configuraciones sienta las bases para una automatización de todos los entornos de una aplicación.

La infraestructura como código y la configuración como código son parte de este componente y se definen como la creación de scripts para entornos.

Infraestructura como código: permite definir entornos, incluidas redes, servidores y otros recursos informáticos, como un archivo de texto (script o definición) que se registra en el control de versiones y se utiliza como fuente base para crear o actualizar esos entornos.

Configuración como código: define la configuración de sus servidores, código y otros recursos como un archivo de texto (script o definición) que se registra en el control de versiones y se utiliza como fuente base para crear o actualizar esas configuraciones.

Algunos de los beneficios de utilizar infraestructura como código y la configuración como código son:

- La automatización del entorno de infraestructura proporciona estandarización
- Las configuraciones están libres de errores humanos
- Se mejora la colaboración entre operaciones y desarrollo.
- Evita que las configuraciones se desvíen
- Hace que la infraestructura sea más flexible y esté lista para escalar
- Cada paso es consistente en todos los recursos.
- El control de versiones es un hecho, ya que permiten configurar y proporcionar una infraestructura totalmente versionada y plenamente documentada ejecutando un script.

Orquestación

Al tener un flujo automatizado el orquestador se vuelve primordial ya que es necesario tener un flujo de trabajo dirigido.

El orquestador es un programa de software que se encarga de administrar cada proceso y lleva igual que un director de orquesta la coordinación de ejecución de cada paso. Siendo el único componente que debe ser capaz de ejecutar cada proceso dentro de la tubería de datos.

Coordina código, datos, tecnologías e infraestructura. Las herramientas de orquestación inician trabajos, supervisan el progreso y canalizan errores y alertas a las interfaces apropiadas. Son responsables de mover datos entre diferentes etapas en una tubería aprovisionando y desaprovisionando automáticamente los entornos de desarrollo, prueba y producción.

Para lo anterior normalmente este tipo de tuberías trabaja con contenedores, siendo un contenedor “un paquete de software que incluye todos los elementos necesarios para ejecutar tus productos en cualquier entorno”.

Por lo que las herramientas de orquestación realizan las actividades de despliegue y puesta en marcha automático de servicios basados, balanceo de cargas, auto escalado y auto reinicio de los contenedores y lleva el control de la "salud" de cada contenedor.

Aseguración, control y monitoreo de la calidad

En un flujo automatizado es necesario monitorear y controlar continuamente la tubería de datos en cada proceso asegurando la calidad de los datos.

Esto ayudará al orquestador para que realice su trabajo de ejecutar cada tarea en el momento correcto. Es por ello por lo que debe integrarse la automatización de pruebas, así como la integración de alarmas desde la extracción de datos hasta el consumo de datos.

Desarrollo Continuo

Al mantener un flujo automatizado es necesario que los desarrollos se puedan integrar de una manera fácil y con riesgos controlados sin necesidad de detener el flujo de la tubería.

Para ello se recomienda el uso de Cajas de Arena o Sandbox por su nombre en inglés. Estas cajas son tuberías de desarrollo y prueba donde los ingenieros de datos pueden desarrollar y probar lo necesario sin afectar el ambiente productivo.

Modelo de ciencia de datos

Esta fase se incluye dentro del flujo de trabajo de DataOps dado que por flujo natural el siguiente paso luego de tener los datos limpios y listos para su análisis es más fácil integrar modelos matemáticos para el entendimiento y generación de nuevos indicadores de negocio.

Pruebas Automatizadas

Las pruebas automatizadas son una manera ágil de verificar que todos los cambios implementados en el proceso funcionan correctamente y no afectan alguna funcionalidad del sistema existente.

En este rubro es importante generar pruebas incrementales y que consideren un caso de prueba completo.

Sistema de Control de versiones

La tubería de datos está compuesta por varios procesos o etapas automatizadas que deben de ser configuradas, desarrolladas, mantenidas y controladas.

Para seguir el principio de generación de ambientes de DataOps es necesario tener parametrizado y resguardado todas las configuraciones de las herramientas utilizadas en cada fase esto para garantizar que se puede replicar la tubería en caso de que se requiera. Por otro lado, en el caso del propio código desarrollado para la operación, pruebas y transformación de los datos.

Una de las mejores prácticas es contar con un software o sistema que sea el repositorio, que permita la administración y control de los cambios sobre el código y las configuraciones y que a su vez permita el desarrollo colaborativo garantizando así la desactualización de código o la sobrescritura de este.

Ramificar y fusionar

El desarrollo de código en un ambiente colaborativo requiere de trabajar de una manera que un desarrollador pueda modificar y probar su código sin afectar el código de otra persona. Esto se resuelve trabajando en un modelo que hace referencia a la estructura de un árbol que tiene tronco y ramas. Donde el tronco sería el código principal o maestro y las ramas copias editables del tronco que se crean normalmente por desarrollador y que es donde cada uno trabaja, prueba y manipula su código una vez que esté seguro de aplicar ese cambio a el tronco o código maestro fusionará sus cambios a el tronco. De esta manera siempre se protege la versión del tronco y se garantiza que cada desarrollador pueda tener disponibles los últimos cambios antes de publicar sus cambios para el resto del equipo de trabajo.

Múltiples entornos

Con el fin de minimizar problemas y aislar el impacto de riesgo en los datos de la organización se debe trabajar en un ambiente que permita desarrollar, probar, analizar y explorar los datos de tal forma que tenga un subconjunto con los datos que se requiera para el análisis sin afectar los datos productivos.

Reutilizar y contener

Esta capacidad atiende al principio de reutilizar código, Tomando en cuenta que en la tubería de datos se tienen varios procesos y que los datos de salidas de un proceso es el insumo de entrada para el siguiente proceso se sugiere subdividir las tareas en componentes reutilizables y de preferencia en un contenedor siendo un contenedor una tecnología que empaqueta una aplicación con todos los archivos necesarios para un entorno de ejecución.

Métricas DataOps

La salud entre cada componente de la tubería de datos impacta a la capacidad para la toma de decisiones de los consumidores es por eso por lo que se debe monitorear cada indicador de desempeño de la tubería.

Indicadores de desempeño

- Exactitud de la ingesta: Es el número de registros correctos ingeridos como proporción de todos los registros ingeridos.
- Corrección de los datos: Es el número de registros con valores correctos luego de pasar por el proceso de corrección de datos con respecto a la proporción de datos por corregir.
- Puntualidad de los datos: Una vez fijado el umbral de tiempo con lo que se requiere los datos actualizados será la proporción de datos que están por encima o por debajo de ese rango.
- Latencia de extremo a extremo: Este dependerá del tipo de flujo e ingesta de datos ya que en un proceso por lotes será en comparación de la estimación de tiempo en el que se toma la ingesta de origen a destino mientras que en un tipo de flujo semi -tiempo real dependerá del rango de tiempo definido y que la cantidad de datos extraídos y procesados sea el esperado en un rango definido como línea base.

Procedimiento Metodológico

Metodología de estudio

La investigación del presente caso de estudio es de tipo explicativo y descriptivo, ya que busca encontrar la relación entre los factores que intervienen en la problemática del área de datos.

Para ello se abordará una metodología cualitativa para entender y comprender la situación actual desde la perspectiva del equipo de datos.

Esto permitirá encontrar áreas de oportunidad dentro del flujo operativo actual identificando la forma de operar y los tiempos dedicados en este flujo para ofrecer una alternativa de solución.

Otro objetivo de esta investigación es conocer los diferentes Flujos que son controlados y operados por el equipo de datos actualmente con el fin de identificar duplicidad de tareas, procesos no actualizados o procesos innecesarios.

Identificar si existen indicadores punto de partida que permitan tener un conocimiento sobre la eficiencia en cada fase del flujo actual.

Medir el nivel de resistencia al cambio ante la implementación de DataOps como propuesta de mejora operativa y que elementos actuales pueden ser reutilizados.

Todo ello nos permitirá tener un escenario de cómo opera actualmente el área con el fin de proponer una automatización que reduzca el tiempo de tareas operativas que hoy es el principal problema identificado por el área. Por lo cual esta investigación nos puede ayudar a saber si realmente es un problema de automatización de tareas operativas o bien el problema es una automatización de procesos.

También indagaremos para saber si el equipo conoce y cuenta con la documentación necesaria que permita una adopción más rápida, así como la manera en cómo hoy se realizan nuevos desarrollos ya que la solución deberá ser capaz de soportar los procesos actuales, pero también los nuevos procesos.

Por último, en esta sección abordaremos qué herramientas son utilizadas actualmente y si son propias o de un proveedor, así como el nivel de resistencia ante un cambio en las mismas.

Instrumento de Investigación

Como instrumentos para esta investigación se utilizaron cuestionarios a nivel exploratorio al área operativa y de gerencia y posteriormente se realizó una entrevista a los Líderes Técnicos del equipo para documentar tanto los flujos como herramientas utilizadas actualmente.

A continuación, se muestran los cuestionarios:

Cuestionario Gerencia	
Datos: Fecha del cuestionario: 01/05/2023 Nombre a quien se aplicó cuestionario: Anónimo Cargo: Gerente de Datos Área / Departamento: Datos Responsable de aplicación: Erendira Reyes Rangel Medio de aplicación: Google Forms Duración: 15 min	Objetivo: ● Explorar el nivel de automatización. ● Evaluar la disposición del equipo y la empresa para adoptar DataOps ágil. ● Identificar Indicadores Operativos.
Preguntas: 1. ¿Qué tan frecuente cambian los procesos de negocio de la compañía? 2. ¿El área de datos es responsable de las tareas de desarrollo y operación? 3. ¿Existen personas dedicadas exclusivamente a temas operativos? 4. ¿Hoy manejan algún estándar de desarrollo? 5. ¿Cuentan con indicadores operativos del área? 6. ¿Realizan monitoreo automatizado de sus procesos? 7. ¿Cómo consideran dentro de la compañía al área de datos?	

Cuestionario Gerencia	
8.	¿Conoce el término DataOps?
9.	¿Cree que puedan ofrecer algún beneficio a la compañía la implementación de DataOps?
10.	¿Si se implementara DataOps en la compañía, cuál sería el nivel de resistencia del equipo de datos a este cambio?
11.	¿Por qué?

Tabla 2. Cuestionario Gerencia

Los resultados de este cuestionario se encuentran en el Anexo 1.

Análisis de resultados, Cuestionario Gerencia

Analizando los resultados del cuestionario a nivel Gerencia, se detecta que la empresa está en constantes cambios a nivel negocio que implica cambios en las reglas de sus procesos constantemente.

El área de datos es responsable tanto de llevar nuevos desarrollos como el mantenimiento operativo de los datos de la compañía y no existe monitoreo automatizado de procesos.

En cuanto a la distribución de personal en estos dos rubros desarrollo y operativo se detecta una discrepancia por lo que se puede entender que no se tienen estas tareas bien definidas.

No existen indicadores operativos que puedan ayudar a mejorar el rendimiento del área ya que, aunque se tiene algunos manuales al no tener un seguimiento de ellos y/o un control sobre los mismos, es igual que no tenerlos.

La gerencia percibe que el área de Datos es un área operativa, aunque también la consideran como un aliado para crecer.

La gerencia conoce el término DataOps y considera que, si pudiera ser una alternativa de apoyo para mejorar su operación diaria, expresando que el equipo siente un gran interés por temas tecnológicos así que, aunque indica que consideran requerir ser capacitado, el nivel de resistencia del equipo puede ser menor.

Cuestionario Operativo	
Datos: Fecha del cuestionario: 08/05/2023 Nombre de Entrevistados: Anónimo Cargo: 1 Líder Técnico, 1 Analista de Datos, 3 Ingeniero de Datos Área / Departamento: Datos Responsable de aplicación: Erendira Reyes Rangel Medio de la entrevista: Google Forms Duración: 20 min	Objetivo: • Este conjunto de preguntas permitirá: • Explorar el nivel de automatización. • Identificar el tipo de operación y procesos operativos. • Contar con una medida base de cuánto tiempo se dedica a temas operativos. • Identificar si se han detectado procesos duplicados. • Evaluar si existen procesos de validación de datos. • Identificar si el área de desarrollo sigue un estándar de desarrollo para evaluar si se requiere ajustar. • Identificar si existe documentación sobre los procesos actuales. • Identificar áreas de oportunidad que pueden optimizarse. • Identificar si se cuenta con componentes reutilizables dentro del flujo actual. • Evaluar la disposición del equipo y la empresa para adoptar DataOps ágil.
Preguntas: 1. ¿Cuál es tu rol en el equipo de Datos? 2. ¿Tienen procesos de datos diarios? 3. ¿Tienen procesos de datos mensuales? 4. ¿Tienen procesos de datos en tiempo real? 5. ¿Tienen procesos de datos nocturnos?	

Cuestionario Operativo

6. ¿Tienen procesos de validación de datos?
7. ¿La base de datos actual donde reside el repositorio de datos de la compañía se encuentra en infraestructura propia o de algún proveedor?
8. ¿En caso de no tener infraestructura propia crees que la compañía pudiera cambiar de proveedor?
9. ¿Cuántos registros tiene la tabla más grande que hoy se tiene en su repositorio de datos de la compañía?
10. ¿Tienen algún proceso para depuración de información histórica?
11. ¿Cuánto tiempo del día dedican a temas operativos diarios?
12. ¿Cuánto tiempo del mes dedican a temas operativos mensuales?
13. ¿Hoy manejan algún estándar de desarrollo?
14. ¿Tienen documentación de sus procesos operativos?
15. ¿Sabes si existe documentación de ingesta de datos?
16. ¿Sabes si existe documentación de calidad de datos?
17. ¿Sabes si existe documentación de validación de datos?
18. ¿Tienen algún control que permita monitorear indicadores operativos?
19. ¿Cuánto tiempo tomaría desarrollar desde cero el flujo datos más sencillo que hoy tiene el área?
20. ¿Cuándo realizas un desarrollo, puedes reutilizar componentes o código?

Cuestionario Operativo
21. ¿Si tuvieras componentes reutilizables, crees que el tiempo que utilizas para desarrollar sería menor? 22. ¿Existen procesos manuales que podrían ser automatizados? 23. ¿Tienen procesos que realizan la misma tarea? 24. ¿Consideras que si se automatizan las tareas de limpieza y carga de datos se pudiera dedicar ese tiempo para nuevos desarrollos? 25. ¿Cuándo ocurre un error en la carga de datos es fácil saber en qué pasó ocurrió el error? 26. ¿De las horas laborables en un día normal cuántas horas le dedicas a la corrección de errores? 27. ¿Si se requiere realizar algún reproceso de datos que tan complicado es? 28. ¿Conoces el término DataOps? 29. ¿En caso de conocerlo, crees que podría ser una solución para automatizar el flujo de datos? 30. ¿En caso de no conocerlo te gustaría aprender sobre él y su implementación?

Tabla 3. Cuestionario Gerencia

Los resultados de este cuestionario se encuentran en el Anexo 2.

Análisis de resultados, cuestionario para el área operativa

El área de datos cuenta con procesos diarios, mensuales y datos en tiempo real con procesos manuales y duplicados.

Se tienen algunos controles operativos tanto manuales como operativos.

La validación de datos actualmente se realiza, aunque revisando con el equipo de datos el resultado del cuestionario se aclara que es por intuición del equipo de datos, pero no existe un proceso automatizado y documentado para ello.

En cuanto a documentación se reporta que sí existe a nivel de ingesta de datos, pero para calidad y validación de datos no.

Si ocurre un error en los procesos actuales es difícil detectar en qué parte de todo el flujo ocurrió.

El equipo de datos ha dedicado un gran tiempo en detección y corrección de errores desde 3 horas hasta más de 1 día en procesos diarios.

En cuanto a procesos mensuales le dedican días y cuando se requiere realizar reprocesos de datos es un proceso muy complejo y no documentado.

A nivel de desarrollo se tiene una incongruencia en las respuestas ya que algunos opinan que, si llevan un estándar de desarrollo, pero otro no, por lo que se detecta que no es algo fijo y que sea de conocimiento de todo el equipo.

Hoy día el tiempo de desarrollo para un componente sencillo va desde días a semanas por lo que no se detecta que se tenga una percepción igual en el equipo en cuanto a tamaños o complejidad de un desarrollo.

El equipo externo, que mejoraría su tiempo de desarrollo si tuviera la oportunidad de reutilización de código.

La infraestructura que se tiene es propia y de un proveedor externo el equipo operativo que cree que la empresa puede cambiar de proveedor ya que lo ha hecho en el pasado, pero fue reciente.

Se detecta que no se tiene un mecanismo para depuración de datos históricos y el tamaño de las tablas que manejan son muy grandes, aunque se visualiza que no es un dato que constantemente se revise por ello la diferencia de respuestas.

El equipo operativo conoce el término DataOps, aunque respondió que le gustaría aprender más sobre esto lo cual confirma lo mencionado por la Gerencia ya que tienen interés y comentan que puede ayudar a mejorar su operación y desarrollo.

Con base en los resultados de los cuestionarios aplicados tanto a nivel operativo como de la gerencia, se realizó una investigación más a fondo por medio una entrevista directa a los Líderes Técnicos con el objetivo de documentar los flujos actuales.

Entrevista Situación Actual	
Datos: Fecha del Entrevistado: 15 al 19 de Mayo de 2023. Nombre de Entrevistado: Anónimos. Cargo: Líder Técnico. Área / Departamento: Datos. Responsable de aplicación: Erendira Reyes Rangel. Medio de la entrevista: Presencial. Duración: 5 horas.	Objetivo: • Documentar los flujos actuales. • Documentar herramientas utilizadas actualmente. • Documentar la infraestructura actual. • Documentar procesos actuales. • Identificar fuentes de datos. • Identificar tipos de consumidores. • Documentar controles actuales.
Preguntas: 1. ¿Quiénes son sus consumidores de datos? 2. ¿Qué operaciones o procesos realizan los datos para su consumo? 3. ¿Cómo es que dispone de sus datos? 4. ¿Todos pueden ver todos los datos o cuentan con grupos de seguridad? 5. ¿Cuáles son las responsabilidades del área de Datos? 6. ¿Cómo está conformado el equipo de Datos? 7. ¿Cuáles son sus fuentes de datos? 8. ¿Cada cuando se refresca o actualiza estos Datos? 9. ¿Si tuvieras que identificar los pasos por los cuales pasan los datos desde su origen hasta su consumo final, cuáles serían? 10. ¿Existe alguna herramienta para coordinar o configurar estos pasos? 11. ¿Con que Proveedor tienen contratado sus servicios de infraestructura? 12. ¿Tienen pensado cambiar de Proveedor? 13. ¿Qué herramientas utilizan para extraer, transformar y/o cargar datos?	

Entrevista Situación Actual
14. ¿Cómo controlar o monitorear sus procesos técnicos? 15. ¿Qué Base de Datos tienen? 16. ¿Cómo se enteran o detectan errores? 17. ¿Cómo se ejecutan y controlan sus procesos diarios? 18. ¿Cómo se ejecutan y controlan sus procesos mensuales? 19. ¿Cómo realizan la validación de datos actuales?

Tabla 4. Entrevista a Líderes Técnicos.

Situación Actual

Con base a los resultados de la entrevista realizada a los líderes del equipo de Datos a continuación se muestra la situación actual del equipo, procesos y herramientas.

La empresa mexicana de procesamiento de video para quien opera el equipo de Datos ofrece sus servicios en tiempo real con una plataforma que ofrece contenidos a 16 países de América Latina y se encuentra constante en expansión y crecimiento de cartera de clientes.

La Gerencia de Datos es el área encargada de realizar la extracción, tratamiento y mantenimiento, de los datos generados por sus usuarios dentro de la plataforma de vídeo a nivel oficina central desde México por medio de archivos en formato CVS a cada país. respetando siempre que cada país solo puede tener acceso a la información generada por los usuarios de ese país.

Otra forma en que hoy se entrega información a cada país es por medio de reportes y tableros de control por la plataforma de visualización de datos llamada Locker del proveedor Google.

Estos reportes y tableros hoy se controlan por medio de filtros regidos por permisos asignados a nivel de grupo y usuarios por país.

Dentro de las responsabilidades de la Gerencia de Datos está atender las solicitudes de análisis especiales que el negocio requiera para la toma de decisiones a nivel corporativo.

El equipo de la Gerencia de Datos está conformado de la siguiente manera:

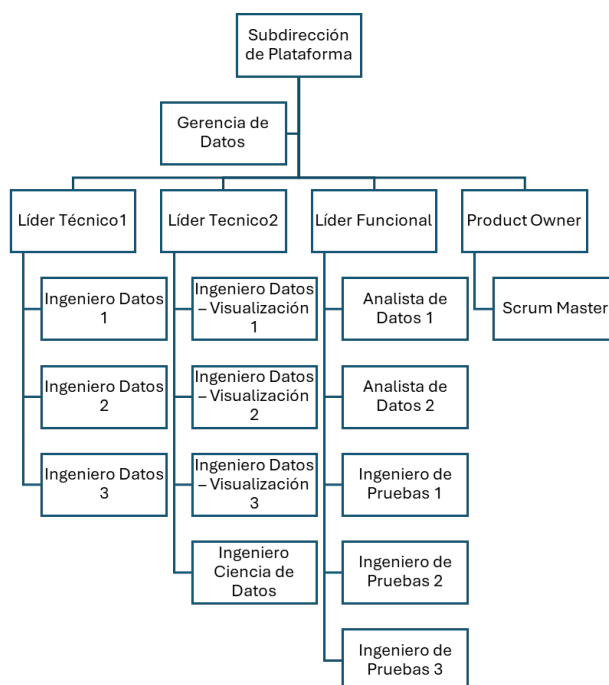


Figura 11. Organigrama de la Gerencia de Datos

En la operación diaria de esta gerencia se identifican 3 tipos de flujos para el tratamiento de la información.

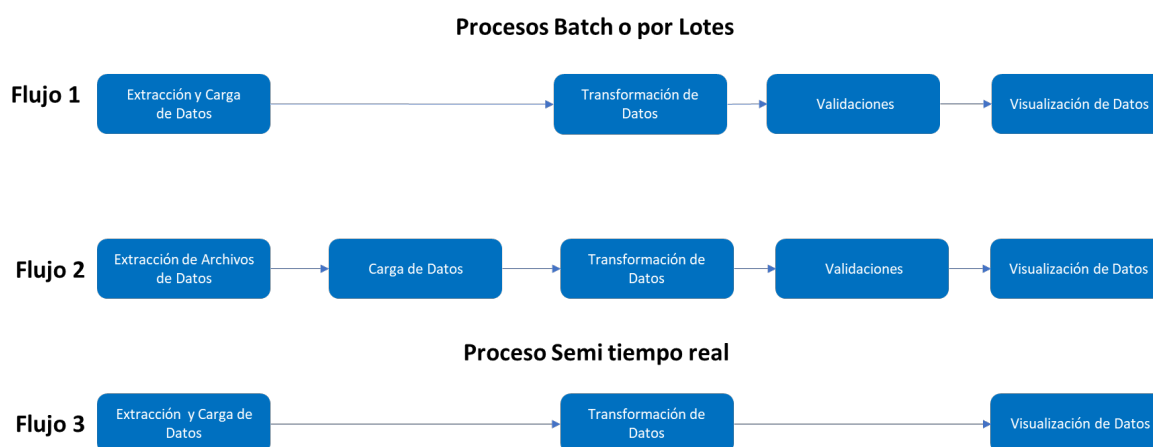


Figura 12. Diagrama que muestra los tipos de flujos actuales para el tratamiento de la información identificados en el análisis diagnóstico de este caso de estudio.

Arquitectura de procesos actual As - Is

Luego de analizar los flujos de datos actuales se puede concluir que los tres coinciden con los siguientes procesos.

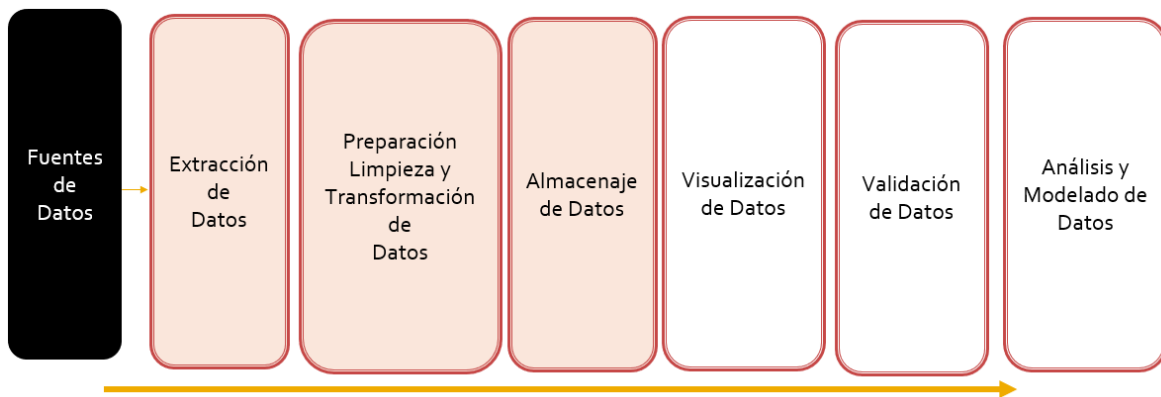


Figura 13. Flujo identificado para el tratamiento de datos para su análisis de información en el presente caso de estudio.

Extracción de datos

Como ya se mencionó, esta fase involucra todos los procesos necesarios para extraer los datos. Actualmente se tiene:

- Flujos de datos por lotes o bloques de información
- Flujos en semi tiempo real o en línea

Dichos procesos se ejecutan de la siguiente manera:

- Procesos programados de manera simultánea llamados en paralelo
- Ejecución automatizada de manera secuencial o serializados por tiempo
- Procesos dependientes controlados con una validación mediante una consulta a la base de datos origen revisando si otro proceso de la base de datos origen término para poder ejecutarse.

En cuanto a la forma de cómo es que se realiza la tarea de extracción, actualmente se leen los datos de las fuentes origen y se depositan en tablas de paso donde se guarda toda la data con tipo de dato texto y posteriormente se ejecuta un proceso que mueve los datos a otra estructura, pero regresando los datos a su tipo de dato original para que sirvan de insumo para la siguiente capa llamada de transformación.

Transformación de datos

La fase de transformación de datos está programada para ejecutarse la etapa de extracción de datos únicamente por tiempo manteniendo un espacio de 1 hora entre fases, pero si algún proceso de extracción marca error el proceso de transformación no se detiene y se ejecuta.

La transformación actualmente se ejecuta de la misma manera para los tres tipos de flujos detectados por medio de procesos almacenados en la Base de Datos BigQuery.

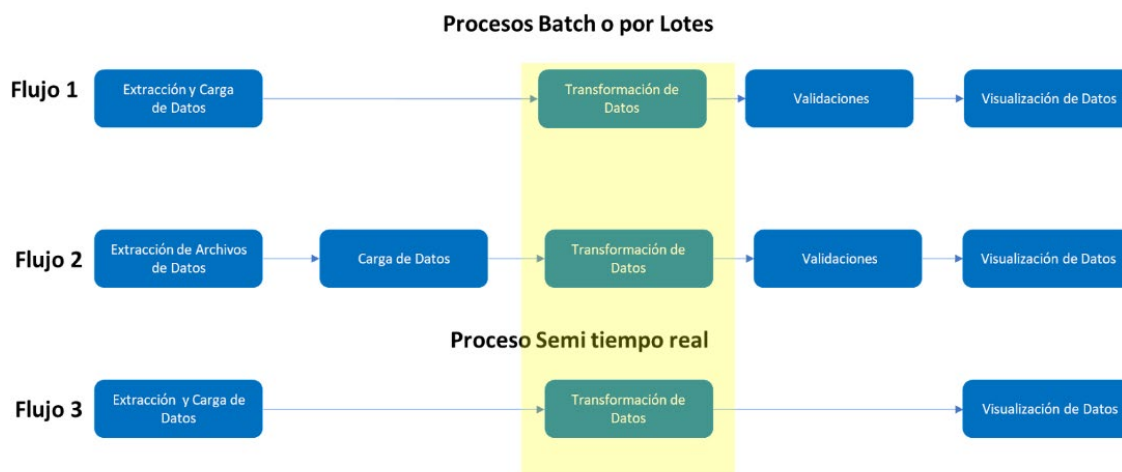


Figura 14. Tipos de flujos identificados en la situación actual

En cuanto a la programación de estas ejecuciones se realizan mediante la herramienta de programación de la misma Base de Datos en un horario fijo mediante una tarea que a su vez invoca la ejecución de cada proceso de transformación de datos de manera secuencial.

En cuanto a las validaciones de integridad y consistencia de datos se realizan de manera manual y en caso de fallos, estos errores son atendidos por una guardia diaria conformada por 3 analistas de calidad, 1 ingeniero de datos y 1 analista de datos.

Almacenaje de Datos

En esta fase se depositan los datos en una base de datos única llamada BigQuery en la nube del proveedor Google; pero no se realiza ninguna depuración es decir que todo el tiempo se escriben datos y las estructuras actuales siguen creciendo de manera exponencial, actualmente la Base de Datos tiene un tamaño de 10 Terabytes y contiene datos del negocio desde 2017.

Visualización de Datos

Una vez que se cargan los datos en las tablas de la base de datos en BigQuery. Esta información se visualiza inmediatamente en los tableros de mando productivos que consultan los consumidores de datos de la gerencia y pertenecen a los departamentos de mercadotecnia, finanzas y comercial.

Estos tableros están desarrollados en la herramienta de Google llamada Looker y están conectados directamente a las tablas de la base de datos de BigQuery.

Dichos tableros al ser consultados se envía internamente una consulta que se ejecuta en ese momento a la Base de Datos y cada que es consultado alguno de los tableros de información, mostrando en ese momento la data disponible de la base de datos con 1 día de desfase al momento de la generación de suscripciones o visualizaciones generadas por la plataforma transaccional la cual graba todos los eventos de los usuarios de la plataforma de video.

Monitoreo y seguimiento

Ningún proceso actual contiene la detección de errores, así como acciones programadas en caso de alguna falla.

En flujos por lotes, cuando estas fallas se llegan a presentar en la fase de extracción el proceso continúa hasta la fuente destino provocando una salida de datos inconsistente o no esperada, así como el consumo de recursos por procesar estos datos erróneos. Lo que se refleja en un mayor tiempo para entrega de la data a usuarios finales puesto que requerirá de reprocesos completos en todas las fases.

Mientras que en los flujos de semi tiempo real una pequeña desconexión no es un problema ya que guarda 24 horas los mensajes, pero pueden ocasionar encolamientos y si no son procesados estos mensajes en este tiempo se pueden perder los mismos.

Validación de Datos

Los analistas de calidad se encargan de controlar por medio de consultas a la base de datos o consultando reportes de la plataforma de explotación de datos existente para detectar falta de información o indicadores de negocio principales fuera de rangos normales por intuición o conocimiento del propio negocio no existe algo automatizado o una lógica más elaborada que garantice este tipo de controles lo cual hace susceptible a errores humanos así como aumentar el tiempo que se dedica a estas actividades, es importante recalcar que esta fase actualmente toma hasta 5 horas al día la cual inicia a las 8 de la mañana lo que implica que después de mediodía esta información ya está certificada.

La fase de análisis y modelado de datos actualmente es a solicitud de negocio ya que el área de datos está con su tiempo ocupado en mantener y poner a disponibilidad la información para la toma de decisiones.

En cuanto a la visualización y explotación de datos actualmente se realiza por medio de la plataforma de explotación de datos que cuenta con tableros de control que muestran indicadores de negocio agrupados o acumulados, así como algunos reportes detallados filtrados por medio de grupos de usuarios relacionados a cada país.

Arquitectura tecnológica actual As - Is

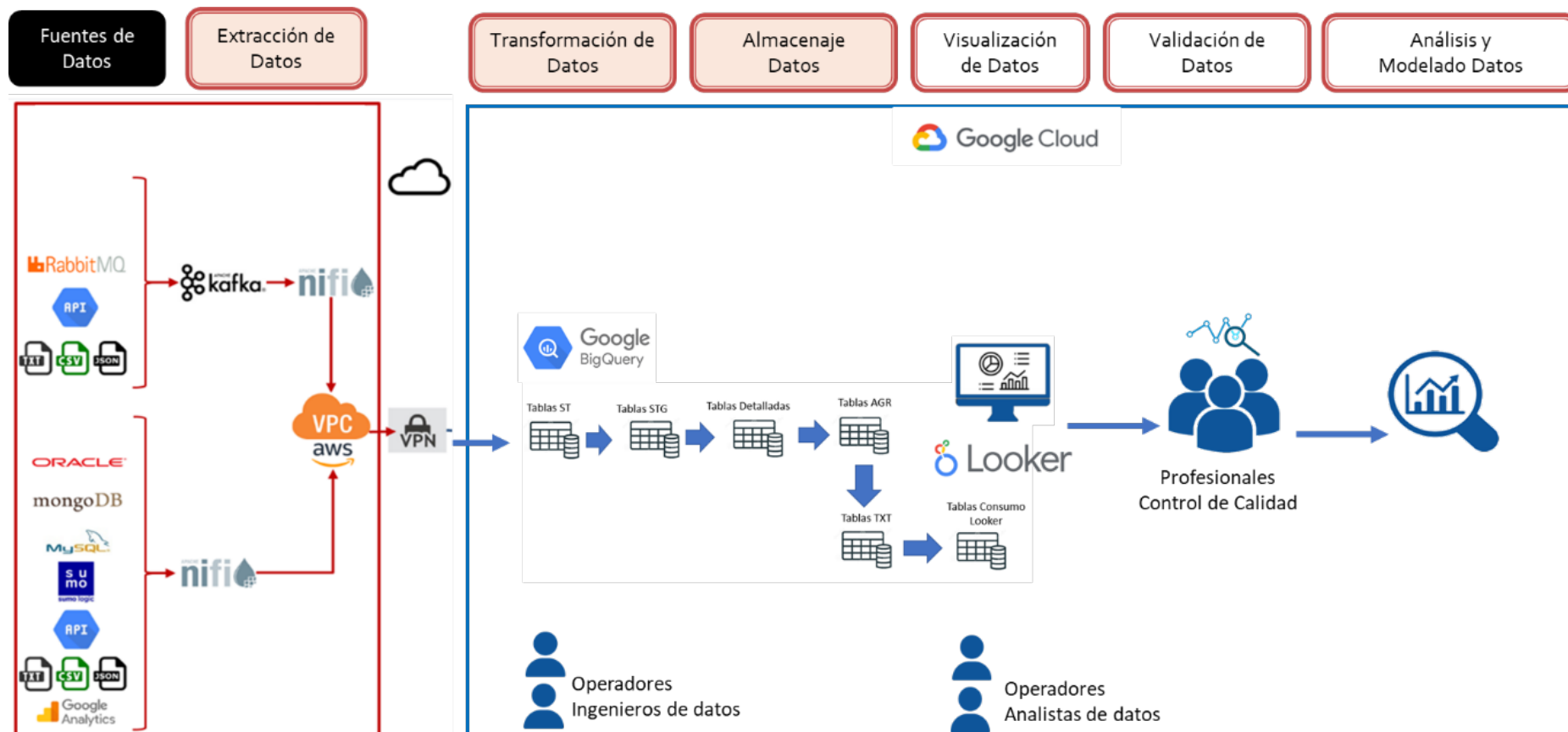


Figura 15. Diagrama de arquitectura tecnológica actual detectado en el diagnóstico del presente caso de estudio.

En el esquema operativo de la Figura 15, el proceso desde la extracción hasta la disponibilidad de información se lleva a cabo de 2 a 8 de la mañana de México.

Todo el flujo entre herramientas es programado de manera manual conforme a la experiencia del equipo de datos.

Estos procesos no están programados con dependencias entre sí es decir la extracción corre a las 2 de la mañana y termina aproximadamente a las 6 am a las 6:30 inicia el proceso de transformación de datos y las validaciones la cual se inician a las 7:30 de la mañana que inicia su día laboral del ingeniero de datos que está encargado de monitorear y revisar que todos los procesos de extracción de datos se ejecute correctamente en caso de que detecte algún error tendrá que volver a iniciar a esa hora el proceso de extracción de datos de manera manual. En caso de que sea exitoso solo monitorea que el proceso de transformación de datos se ejecute sin inconvenientes.

A las 8 de la mañana aproximadamente confirma que los últimos procesos de transformación terminaron y es ahí con ese aviso por la herramienta colaborativa empresarial de Microsoft Teams notifica al equipo de control de calidad para que inicie sus validaciones manuales y visuales conforme a consultas directas a base de datos o bien con ayuda de los tableros de mando productivos que también ya están disponibles para el usuario. Por lo que si detectan algún error deben notificar al analista de datos para que publique en estos tableros de mando un mensaje de que la información que se visualiza en el tablero está en revisión aun y no está disponible para que los usuarios finales eviten tomar esta información para su uso.

Esta validación llega a tomar 3 a 5 horas, así que si sumamos los tiempos del flujo hasta que está la disponibilidad de datos al usuario es de 10 a 12 horas y si se inicia a las 2 de la mañana implica que hasta el mediodía el usuario final puede estar seguro de que los datos están disponibles.

No así con los datos en semi-tiempo real ya que este flujo muestra de inmediato los datos, pero sin ninguna validación previa.

Todo este proceso de inicio a fin en días no laborables se realiza por una guardia que solo revisa la parte técnica las validaciones de datos se realizan hasta el día laboral siguiente por lo que el tiempo de validación de ese primer día laboral puede extenderse aún más.

Estas validaciones en cierre de mes aumentan ya que también se tienen procesos de carga de datos mensuales, al ser más extensas las validaciones por parte de negocio se da un plazo de 5 días al iniciar el mes para dar por cerrado el cierre de datos del mes anterior.

Diagnóstico

Luego de realizar un estudio de los componentes tecnológicos y procesos operativos se llegó al siguiente diagnóstico.

Falta de documentación y metodología de desarrollo

Los procesos actuales no se encuentran documentados y además en su estructura son diferentes ya que cada desarrollador lo realizó a su estilo personal, no existe algún estándar de desarrollo que permita la detección y atención rápida de una falla.

Diferentes tipos de Flujos de Datos de Carga que Controlar

Por ejemplo, algunos se extraen de archivos formato CSV y luego se cargan a la base de datos del DWH, otros se leen directamente de una base de datos a otra base de datos y adicional por medio de procesos por lotes nocturnos y proceso en semi – tiempo real.

Pasos repetidos

En la fase de extracción de datos se detecta en el flujo actual se tienen procesos que ya no se usan, así como estructuras de paso que agregar únicamente un paso adicional sin generar un valor agregado al procesamiento de datos.

Por ejemplo, como primer paso en la lectura de todas las fuentes se carga a estructuras donde se convierten a tipo texto todos los atributos para posteriormente aplicarles un proceso para regresar los datos a su dato original, agregando un paso más a controlar y provocando un posible fallo en la transformación de datos ya que si existe un fallo en el tipo de dato de entrada se detectará hasta la capa de transformación.

Falta de comunicación entre procesos

Estos procesos no se comunican entre sí, no son dependientes es decir si un proceso en la cadena de procesos falla la siguiente fase continua y si el operador a la hora que inicia su día se da cuenta que uno de los primeros procesos de la madrugada fallo debe parar todo de manera manual y relanzar de nuevo todo el flujo de datos.

Costos innecesarios

La infraestructura actual está implantada en las nubes de Google y AWS por lo que se paga por el uso de red, procesamiento y almacenamiento todo el tiempo incluso aunque el consumo de recursos no sea el mismo todo el tiempo tomando en cuenta que los procesos con mayor requerimiento de recursos de procesamiento y almacenamiento se generan en la noche diariamente teniendo un consumo mayor los fines de mes.

Toda la infraestructura está disponible las 24 horas del día se use o no, por ejemplo, los procesos de extracción se ejecutan de 2am a las 6 de la am de México por lo que solo se utiliza 4 horas diarias sin embargo se paga por su uso las 24 horas del día.

Cuando se detectan errores en el proceso la atención de estos se vuelve lenta y muy compleja.

En caso de algún error se tienen que buscar de manera manual en el historial de logs de varias herramientas.

En caso de error en la carga de datos los reprocesos pueden ser complicados y tomando encuenta que a las 7 am inicia a revisar el operador suponiendo que todo hubiera fallado en la etapa de extracción tendrá que volver a lanzar los procesos tomando otras 4 horas para su procesamiento esto sin tomar en cuenta que se deba corregir algo en el trayecto.

Tiempos de Validación y control de Cifras extenso (5 horas de validación en un día normal)

Las validaciones son manuales y se llevan a cabo con un rol de operador de las herramientas de extracción de datos, 3 profesionales de control de calidad, 1 operador para la plataforma de visualización de datos y todos validan la carga diaria que toma alrededor de las 5 a 6 horas diarias.

Integración de nuevos flujos

Para integrar nuevos datos se requiere de un análisis extenso, ya que se debe evaluar en qué parte de la malla de procesos se debe incluir, así como la integración al modelo de datos actual.

Medio día se dedica a validación de cifras y procesos operativos

El tiempo que se dedica para el análisis de datos de negocio es menor al tiempo que se dedica a la operación diaria.

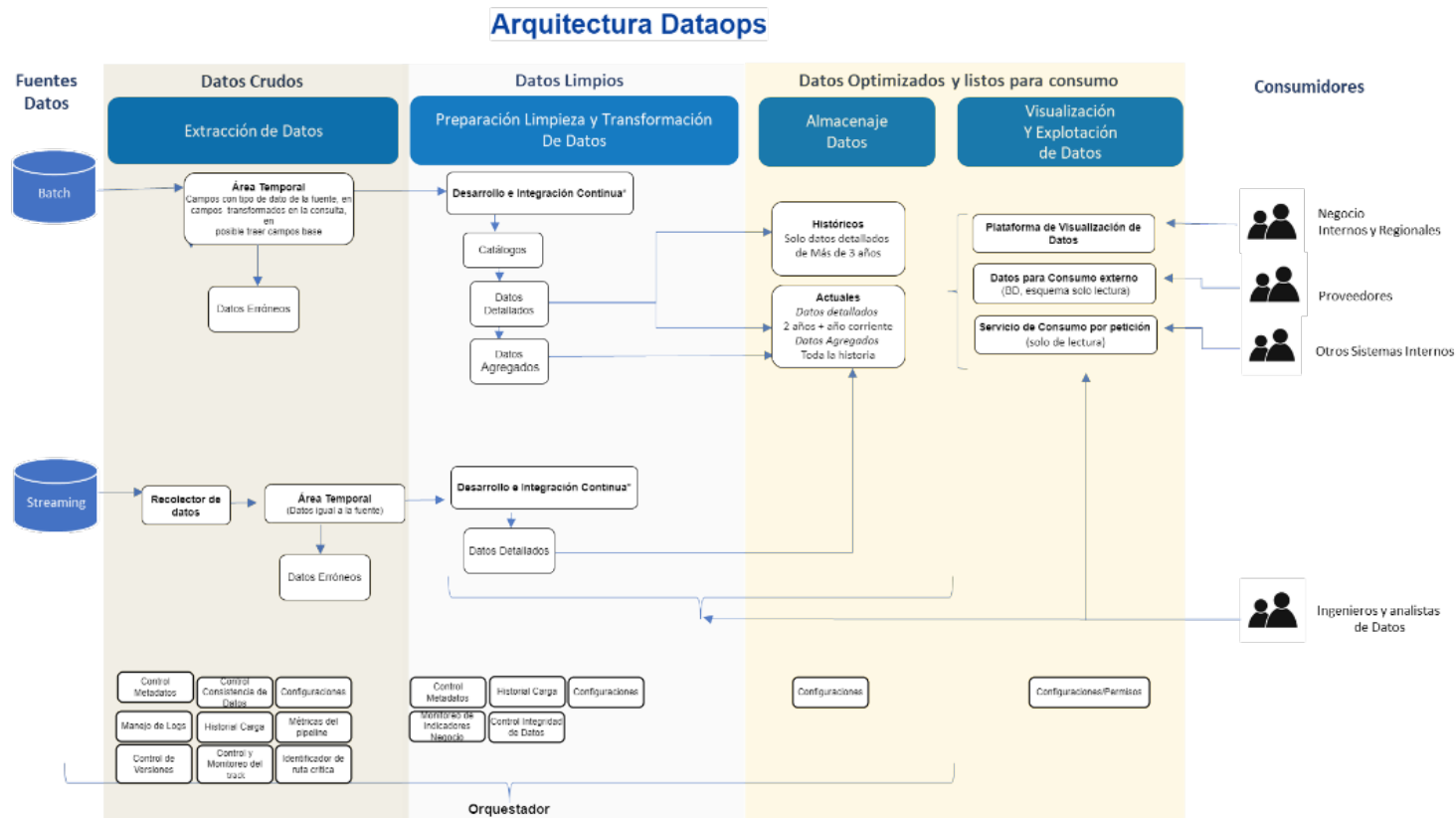
Lo anterior ocasiona que no se optimice y se aproveche el tiempo laboral ya que hoy solo se realizan análisis a solicitud de negocio cuando la operación lo permite tomando mayor tiempo de respuesta a estas peticiones. Por lo anterior se detecta una oportunidad para mejorar el tiempo de la operación diaria y que el equipo de datos pueda utilizar el tiempo ganado en generar nuevas oportunidades de negocio o a generar nuevos servicios a la compañía, detectando comportamientos o patrones en los datos lo cual requiere de un modelo especial, hoy día se tiene por parte de negocio la solicitud de esto sin embargo la operación diaria deja poco tiempo laboral para estas solicitudes por el negocio que además son sobre demanda.

Bajo este escenario surge la problemática de tener que controlar y gestionar diferentes flujos de datos complejos, no homogéneos, con mantenimientos extensos y factibles de error debido a procesos manuales ya sea por errores humanos o del propio sistema que vuelve lenta y difícil su operación diaria generando demora en la entrega y disponibilidad de información hacia negocio. Lo cual se ve reflejado al invertir el mayor tiempo del equipo de datos en mantener y actualizar la data

de negocio quitando así la posibilidad de dedicar estos esfuerzos a nuevos desarrollos y generar nuevos productos.

Para ello este trabajo se centrará en el diseño de un flujo ágil, automatizado, escalable y medible desde la obtención de datos, tratamiento, almacenamiento y disponibilidad de la información permitiendo el monitoreo y detección de errores a lo largo del proceso con el fin de generar un ciclo de mejora continua y garantizando la integración de nuevos orígenes de datos.

Alternativa de solución



Arquitectura de procesos To-Be

Figura 16. Figura. Diagrama de Procesos To-Be propuesto para este caso de estudio.

En la Figura 16 se presenta el flujo de datos propuesto como solución a la problemática de controlar y gestionar diferentes flujos de datos complejos, no homogéneos, con mantenimientos extensos y factibles de error debido a procesos manuales ya sea por errores humanos o del propio sistema que vuelve lenta y difícil su operación diaria generando demora en la entrega y disponibilidad de información hacia negocio. Lo cual se ve reflejado al invertir el mayor tiempo del equipo de datos en mantener y actualizar la data de negocio quitando así la posibilidad de dedicar estos esfuerzos a nuevos desarrollos y generar nuevos productos.

Este diseño se realizó pensando en que la operación de datos sea ágil, automatizada, escalable y medible desde la obtención de datos, tratamiento, almacenamiento y disponibilidad de la información permitiendo el monitoreo y detección de errores a lo largo del proceso con el fin de generar un ciclo de mejora continua y garantizando la integración de nuevos orígenes de datos.

El nuevo ecosistema de datos está pensado para también optimizar los recursos de la nube contratados de tal forma que se pague únicamente lo necesario para el buen funcionamiento de este.

Para ello se define lo siguiente:

Flujos de datos

Conforme a lo detectado en el diagnóstico la propuesta debe poder mantener dos tipos de flujos por lo que el diagrama de procesos se identifican dos tipos de flujos de datos

Procesamiento por lotes

Este flujo se encargará de los datos que se pueden cargar por unidades específicas de tiempo con una consulta predefinida fuera de línea es decir que no necesariamente se requiere su carga al momento de generarse el dato. La gran mayoría de los datos que actualmente se cargan a la base de datos se encuentran hoy en este tipo de procesamiento.

Procesamiento en semi tiempo real

Este flujo está diseñado para los datos que se generan, se leen y se escriben en semi-tiempo real y que requieren de una infraestructura que permita estar censando los datos en el momento en que se crean para poder recolectarlos y cargarlos a la base de datos.

Una vez identificados los tipos de flujos de datos que deberá soportar la presente propuesta de arquitectura se define zonas de áreas de datos, esto nos ayudará a identificar y clasificar los datos especificando las características y tratamiento que debe aplicar a los datos conforme vayan fluyendo a través de la tubería que se propone construir.

Áreas de Datos

Para este caso de estudio se definen 3 zonas o áreas de datos: Datos Crudos, Datos Limpios y Datos Optimizados.

1) Datos Crudos

En esta área se encontrarán los datos que se extraen de los diferentes orígenes sin alguna transformación y respetando el tipo de dato correspondiente desde origen, ofreciendo trazabilidad del dato desde origen.

El proceso que se define dentro del área de datos crudos es “Extracción y carga de Datos “

Flujo por lotes:

Para el caso particular de este caso de estudio, esta área estará comprendida con el resultado de las consultas de extracción actuales, pero respetando el tipo de dato de la fuente sin hacerle alguna transformación o conversión de datos. Es importante aclarar que actualmente en los procesos de extracción de datos también se crean nuevos campos calculados en el script de extracción en estos casos, para lo cual se define que se mantendrán estos campos calculados sin embargo se incorporan la carga de los campos los atributos base para el cálculo de esos campos calculados siempre y cuando él extrae estos campos base no impliquen que la consulta de extracción cambie el total de registros a extraer por cambiar la granularidad de los datos.

Una vez extraído el dato de la fuente origen se depositan los datos en un área temporal para después utilizarlos en la siguiente fase de “Preparación, limpieza y transformación de datos “

Flujo Semi tiempo real:

Para el flujo de datos de semi tiempo real se contempla igual que en el flujo por lotes un esquema de base de datos un área temporal donde se estará cargando la data de extracción sin embargo se incluye un paso de recolector de datos donde se estará censando y cargado la data cada determinado número de eventos generados por los usuarios finales de la plataforma.

Las tareas que se recomienda deben contemplarse cubrir y/o monitorear en lo posible de manera automatizada son: Consistencia de Datos, Control de Metadatos, Mapeo de Logs, Control de versiones, Administración de Configuraciones, Historial de Carga, Control y monitoreo del flujo del proceso, Identificación de ruta crítica y Métricas de la Tubería de Datos.

2) Datos Limpios

En esta área se encontrarán los datos una vez limpios y/o transformados e incluso se generan nuevos datos, será el área donde vivirán los datos finales.

En esta área se contempla una caja que indica el proceso de “Desarrollo e Integración continua” la cual se muestra en la Figura 15 también conocido como CI/CD que garantizara que se pueda realizar el desarrollo, integración, pruebas y liberación de manera continua sin afectar los datos y procesos productivos.

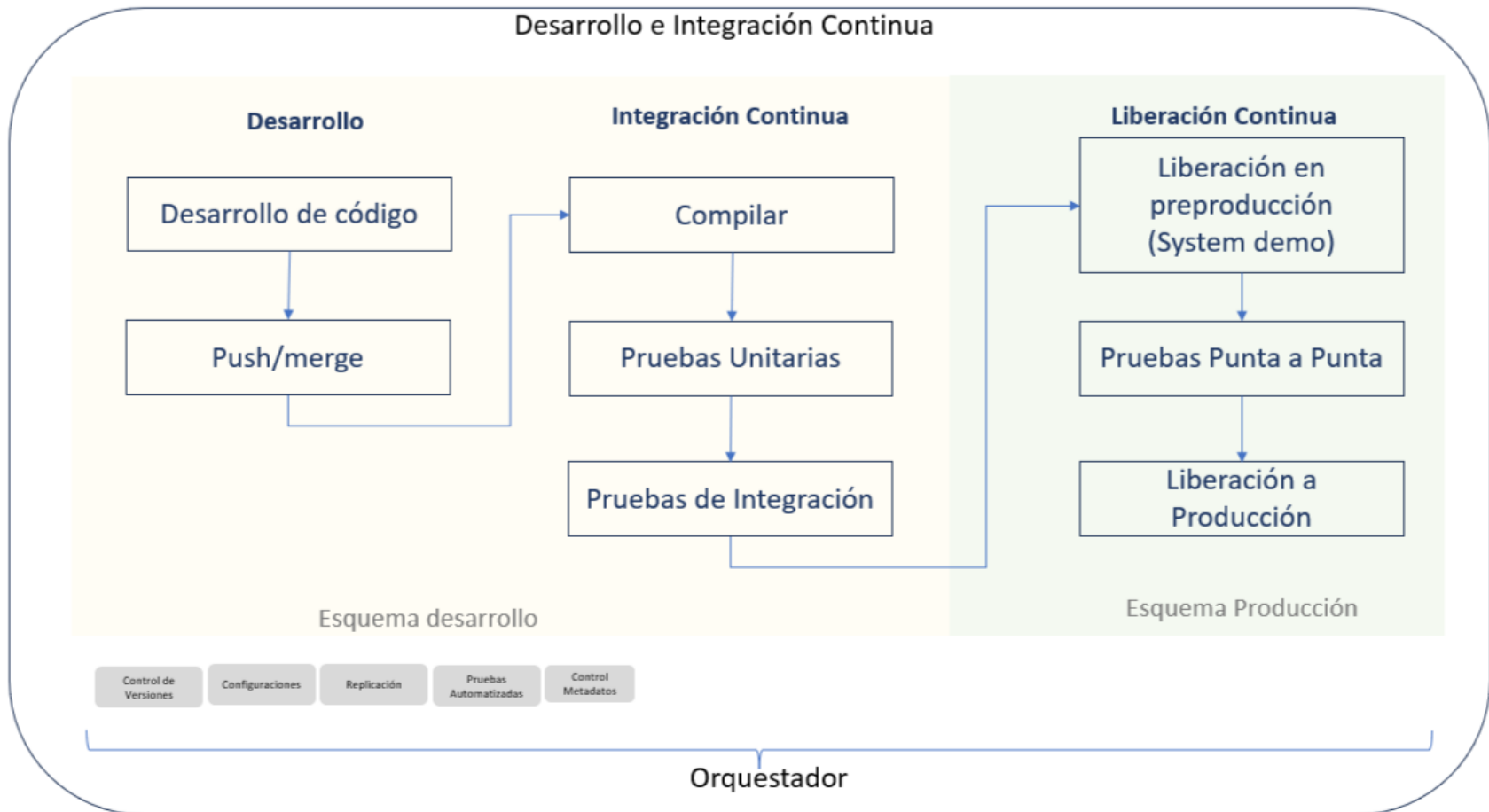


Figura 17. Desarrollo e Integración continua

Este proceso de ingeniería o desarrollo de software se integra a este flujo de datos para combinar la creación y prueba de código de los procesos de transformación y limpieza de datos automatizados con la implementación de este.

Para la creación y prueba automatizadas de código. CI (Integración Continua) permite a los desarrolladores enviar múltiples cambios a un repositorio compartido y colaborativo con una rama de código principal que mantiene el control de versiones.

CD (Entrega continua) permite el desarrollo incremental de ciclo corto y permite a los equipos de desarrollo crear y lanzar nuevas versiones de código en cualquier momento.

Con este esquema de programación colaborativa cada desarrollador tiene una rama de desarrollo que siembre puede ser integrada a la rama maestra o principal en desarrollo y posteriormente este mismo código una vez probado y aprobado se empuja a un ambiente productivo.

Bajo este concepto es que se sugiere desarrollar, testear, integrar y desplegar los procesos de limpieza y transformación de datos

En el diagrama propuesto en la hoja dos se detalla cómo debería funcionar este proceso de CI/CD, es importante declarar que dentro de este proceso “Desarrollo e Integración continua”, se deberá tener la facilidad de replicación esto con el fin de poder tener los “N” flujos que se requieran.

Una vez pasando por este proceso “Desarrollo e Integración continua “, se cargará de manera automatizada la data a las estructuras finales (catálogos, tablas detalladas y agregadas) donde vivirán los datos para consumo.

Las tareas que se recomienda deben contemplarse cubrir y/o monitorear en lo posible de manera automatizada son:

Control de integridad de datos, Control de metadatos, Historial de Carga, Administración de configuraciones y Monitoreo de indicadores de negocio.

3) Datos Optimizados y listos para su consumo

Como parte de las tareas a realizar en esta área sucederá el proceso de almacenamiento el cuál dependerá de el corte de información por antigüedad del dato esto para optimizar el consumo de datos en tablas detalladas para no ir a leer datos a tablas con toda la historia si no a estructuras con información actual o histórica de acuerdo con lo que se requiera de consumo de datos solicitados por nuestros consumidores.

Tanto para los flujos por lotes como Streaming se define para los datos detallados dos tipos de estructuras:

Histórica

Se guardarán los datos con antigüedad mayor a 3 años de acuerdo con la fecha del evento del dato

Actual

Estructuras detalladas -> 2 años atrás a la fecha corriente más el año actual.

Estructuras agregadas → Toda la historia disponible.

Dentro de esta área también se llevarán a cabo las tareas de visualización y explotación de datos

En esta sección se definen tres tipos de consumo de datos para nuestros consumidores.

1. **Plataforma de Visualización de Datos:** Se refiere al portal donde los usuarios internos y regionales por país pueden visualizar información de negocio por medio de tableros de mando, descargar archivos de texto y/o reportería.
2. **Proveedores:** Actualmente sólo se tiene contemplado a un proveedor, pero se deja definida la forma de compartir datos a través de una conexión directamente en nuestra base de datos actualmente BigQuery por medio de un esquema controlado y de solo lectura.
3. **Otros Sistemas Internos:** En este tipo de consumo se propone poner a disposición de sistemas, aplicaciones o áreas de sistemas internas por petición de tipo API para consumo de sus procesos o análisis interno.

Los consumidores del área de datos se clasificaron en 4 grupos:

Usuarios internos y regionales

Los usuarios internos se refieren por ejemplo usuarios del área comercial o finanzas, así como los regionales se refieren a usuarios de las operaciones como lo pueden ser usuarios de países específicos como Argentina, Perú.

*Nota: El área de datos que consultaron será la de datos optimizados y listos para consumo

Proveedores

Usuarios proveedores que requieren de insumo de datos provenientes del equipo de datos y para proveernos de soluciones desarrolladas a la medida por un tercero, un ejemplo de ello es el proveedor del sistema de rating con el cuál el equipo de datos puede brindar este servicio a las operaciones mediante su plataforma desarrollada por el equipo de Jump con métricas especializadas y Gestionada por el equipo de Datos.

*Nota: El área de datos que consultarán será la de datos optimizados y listos para consumo

Otros Sistemas Internos

Sistemas, aplicaciones o áreas de sistemas internas de la compañía como pueden ser el área de pagos, ventas, mercadeo o compromiso con el usuario.

*Nota: El área de datos que consultarán será la de datos optimizados y listos para consumo

Ingenieros y Analistas de Datos

De momento incluimos aquí al mismo equipo de Datos ya que ellos también serán consumidores de datos para nuevos análisis que permitan ofrecer a la Dirección un mejor entendimiento del negocio y sus clientes, detectando áreas de oportunidad o nuevos servicios por medio ejercicios de analítica avanzada, aprendizaje automatizado o detección de patrones de consumo.

*Nota: Las áreas de datos que consultarán este grupo de usuarios será la de datos limpios, datos optimizados y listos para consumo

Orquestación

Por último, se define en el diagrama un proceso de orquestación transversal que está involucrado durante todo el proceso de ingesta y transformación del dato hasta llegar a su consumo.

En este proceso se involucra y define todo el flujo del dato desde su ingesta hasta el consumo y se debe realizar los disparadores de acciones que llevan el dato a la siguiente fase del flujo de datos o lo regresan de fase de manera automatizada según se requiera.

Arquitectura tecnológica To-Be

La Arquitectura Propuesta estaba basada en la arquitectura actual de nube híbrida la cual se mantendrá por medio de la VPC actual que comunica a AWS y BQ.

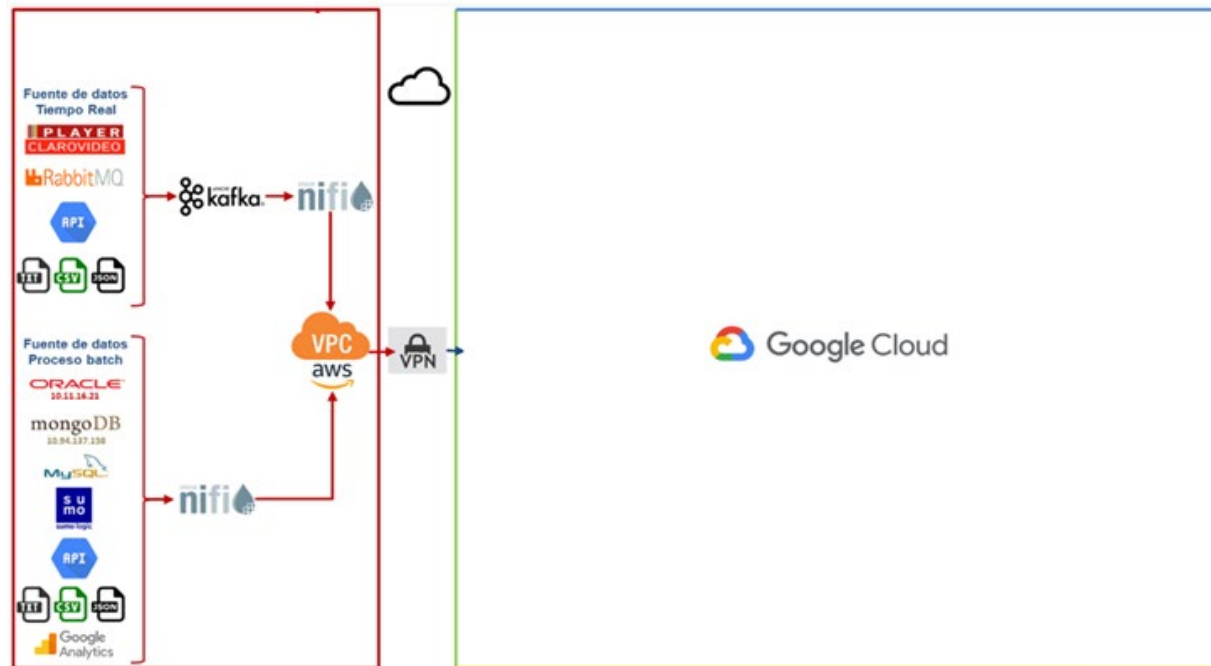


Figura 18. Arquitectura en la nube

A continuación, se muestra la Arquitectura de Procesos propuesta fusionada con las herramientas tecnológicas definidas para el procesamiento de datos.

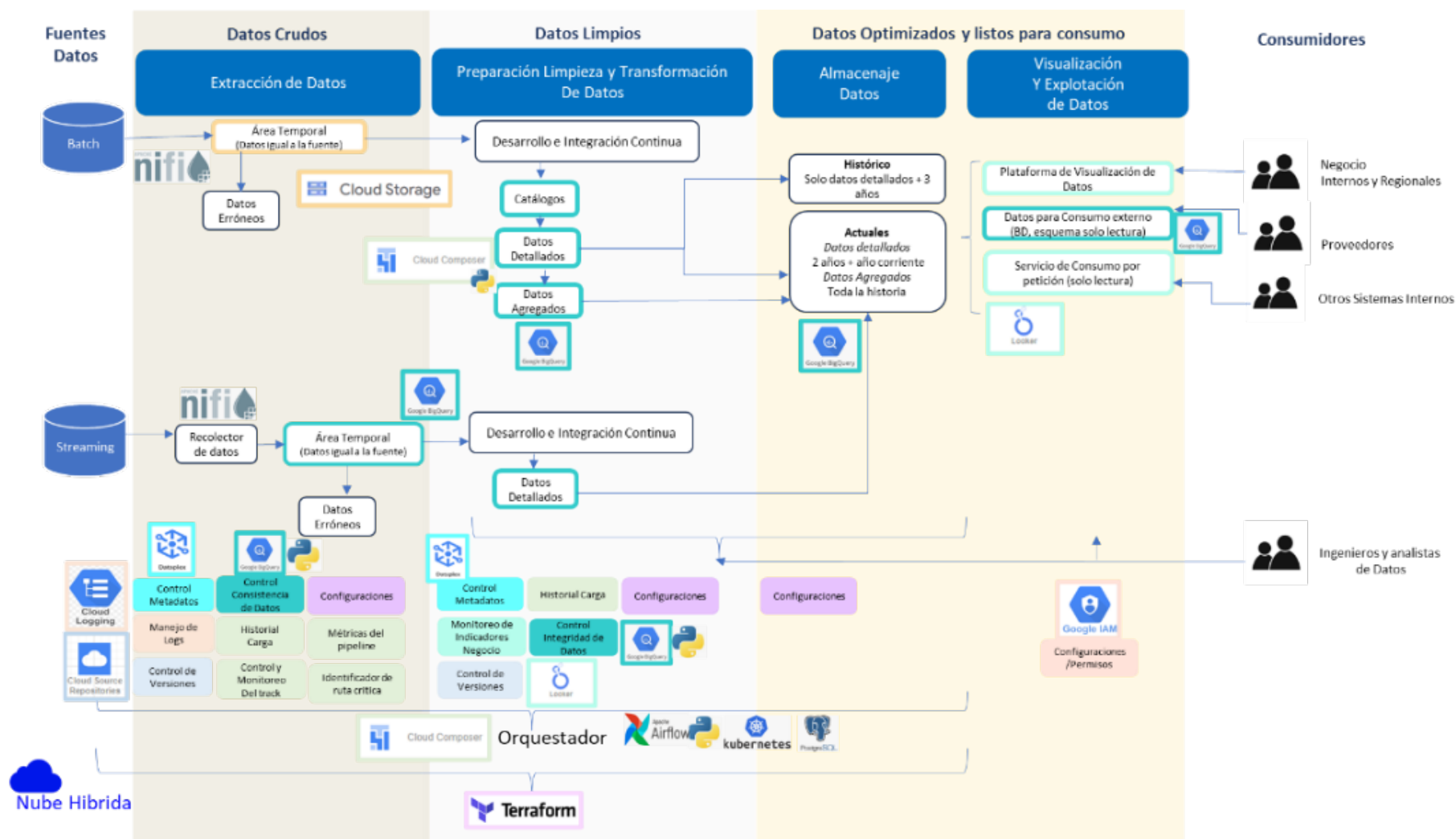


Figura 19. Arquitectura Tecnológica propuesta.

En el diagrama se colocó un código de colores que denota por el color del contorno en qué procesos se utiliza cada herramienta por ejemplo en color naranja se puede visualizar la herramienta de Cloud Storage como Área Temporal dentro de la fase de Extracción de Datos en el flujo de datos por lotes.

Herramientas tecnológicas propuestas

A continuación, se describe brevemente cada herramienta y su relación con cada proceso o tarea dentro de la arquitectura propuesta.



Administración de Configuraciones

Terraform

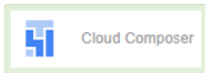
Permite definir y configurar la infraestructura de un centro de datos en un lenguaje de alto nivel, generando un plan de ejecución para desplegar la infraestructura en la nube para proporcionar una infraestructura como servicio.

Utiliza un lenguaje de codificación descriptiva de alto nivel para automatizar el suministro de la infraestructura. Esta automatización elimina la necesidad de suministro y gestión manual de servidores, sistemas operativos, conexiones de bases de datos, almacenamiento y otros elementos de la infraestructura cada vez que deseen desarrollar, probar o implementar una nueva versión de software.

Una de sus ventajas es poder recrear ambientes como se desee para pruebas de ambientación de nuevos desarrollos por medio de código.

La infraestructura se define utilizando la sintaxis de configuración HashiCorp denominada Hashi Corp Configuration Language (HCL) o, en su defecto, el formato JSON.

“Esta herramienta se utilizará para aprovisionar, crear y administrar los recursos de la arquitectura de la tubería de datos de manera automatizada y la ventaja principal por lo que se define es para poder tener un punto de control que permita crear o destruir la arquitectura completa del ecosistema de datos propuesto permitiendo tener siempre respaldo y versionada la arquitectura con ayuda de otras herramientas que veremos más adelante. Lo cual ofrece la flexibilidad en todo momento de replicar el ambiente productivo actual o alguno previo al productivo”



Orquestador

Cloud Composer

Cloud Composer es un servicio dentro de la plataforma Google Cloud Platform utilizado para definir, configurar, administrar y monitorear de manera gráfica flujos de trabajo por medio de servicios distribuidos en la nube usando una herramienta de código abierto llamada Apache Airflow.

Apache Airflow es una plataforma de gestión de flujo de trabajo de código abierto escrita en Python.

Los flujos de trabajo son una serie de tareas definidas para transferir, transformar, analizar y obtener datos.

Air Flow maneja flujos de trabajo que se crean mediante circuitos llamados DAG.

Podemos definir a un DAG como un grupo de tareas que se programan para ser ejecutadas de manera organizada y en el orden correcto.

Los DAG funcionan como circuitos acíclicos con un inicio y fin que no permiten regresos a tareas previas, por lo que trabajan bajo un principio incremental agregando tareas nuevas siempre hacia adelante con dependencias y relaciones.

Airflow cuenta con una extensa biblioteca de conectores y diversas representaciones gráficas que permitirá la mejor representación del flujo de trabajo en acción que facilita la detección de problemas.

Un DAG no debería preocuparse por la función de cada tarea constituyente, su propósito es garantizar que cada tarea se ejecute en el momento correcto, en el orden adecuado o con el control correcto para monitorear problemas.

“Esta herramienta se utilizará para las tareas de orquestación de la arquitectura definida como lo realizaría un director de orquesta en una sinfónica. Es el encargado de disparar y configurar los flujos o caminos (procesos) por donde pasarán los datos dentro de nuestra tubería.”



Repositorio Privado

Cloud Source Repositories

Son repositorios privados alojados en Google Cloud Platform y sirven para almacenar, gestionar y monitorear y desarrollar código de manera colaborativa para cualquier aplicación o servicio.

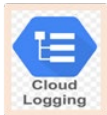
Este repositorio realmente trabaja muy parecido a su homólogo GitHub (herramienta de control de versiones de código) por lo que se eligió por las ventajas y facilidad de adopción.

Al igual que GitHub trabaja bajo un esquema de ramas y subramas por medio de cuenta con un explorador de código que permite ver los archivos del repositorio desde Cloud Source Repositories.

Mediante el explorador de código se puede definir filtros para que la vista se centre en una rama, etiqueta o confirmación determinada lo cual facilita el regreso o revisión de versiones anteriores.

En Cloud Source Repositories se pueden configurar cuentas de GitHub o Bitbucket las cuales se pueden sincronizar de manera automatizada.

“En este caso será la herramienta encargada de mantener y administrar las versiones de código proveniente de los procesos dentro de la tubería de datos “



Administración de Logs

Cloud Login

Cloud Loggin es un sistema de administración de logs de Google Cloud Platform que recopila automáticamente registros en tiempo real compatible con el almacenamiento, la búsqueda, el análisis y la supervisión de recursos locales y recursos de otros proveedores de servicios en la nube.

Permite configurar alertas para que te notifiquen si ciertos tipos de eventos se informan en tus registros y por motivos regulatorios o de seguridad, puedes determinar dónde se almacenan tus datos de registros.

El explorador de registros te permite buscar, ordenar y consultar mediante instrucciones de búsquedas flexibles, junto con visualizaciones de histogramas, un explorador de campo sencillo y la capacidad de guardar las consultas.

“Esta herramienta se define para el manejo, administración y generación de alarmas de logs generados por la tubería de datos “



Herramienta ETL

Apache NiFi

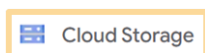
NiFi se creó para automatizar flujos de datos entre sistemas y es un proyecto de Apache basado en el concepto de ETL (Extracción, transformación y Carga de Datos) que trabaja bajo un ambiente gráfico.

Se basa en el modelo de programación de flujo y ofrece funciones que le dan la capacidad de operar dentro de clústeres.

NiFi es un programa Java que se ejecuta dentro de una máquina virtual Java.

Se propone utilizar Airflow como controlador de los flujos, pero se utilizarán todas las funciones y componentes de ETL de NiFi para la extracción y procesamiento de datos, por lo que se aprovechará la comunicación con Airflow.

“Actualmente con esta herramienta se realiza la ingesta y carga de datos desde la fuente origen mediante el Scheduler del propio NiFi. Para la implementación de este proyecto se mantendrá, pero la ejecución automatizada será disparada por el orquestador a través de una petición o llamado a NiFi”



Almacenamiento

Cloud Storage

Cloud Storage es un servicio administrado de Google Cloud Platform para almacenar datos en la nube la cual permite tener diferentes tipos de almacenamiento dependiendo la antigüedad de búsqueda.

Lo anterior es muy conveniente y ventajoso ya que nos va a permitir mantener los datos del DWH en diferentes tipos de almacenamiento lo cual permitirá optimizar el costo de cada almacenamiento dependiendo del nivel de acceso a ese histórico. Siendo la solución más apegada a los requerimientos de la compañía ya que si bien el DWH por concepto mantiene información histórica no se está consultando constantemente todo el tiempo.

Otra ventaja de usar este tipo de almacenamiento en la nube es que es escalable y además se puede recuperar las veces que sean necesarias.

Con Cloud storage, se puede mover los datos de manera automatizada entre las diferentes clases de almacenamiento optimizando costos con la colocación de objetos en las diferentes clases de almacenamiento.

En el nivel del bucket, puedes habilitar el movimiento de objetos automático basado en políticas a clases de almacenamiento.

Otra ventaja que se tiene al usar esta herramienta es la facilidad de replicación.

Al ser nativa de Google nos permitirá que la comunicación e integración de todas las herramientas que utilizaremos en la tubería sea más eficiente y rápida.

“Utilizaremos Cloud Storage para el almacenaje temporal para la extracción de datos desde las fuentes origen hacia archivos para después ser procesado en la siguiente fase de transformación y preparación de datos “

*Nota: Actualmente esta área temporal se encuentra en BigQuery por medio de un esquema que contiene tablas temporales, en la arquitectura nueva se propone Cloud Storage con el fin de reducir costos por procesamiento y de servicio proveniente de BigQuery.



Base de Datos

BigQuery

BigQuery como almacén de datos de Google Cloud Platform funciona en varias nubes y es escalable.

Esta herramienta actualmente es donde se tiene alojado el DWH actualmente.

BigQuery es una base de datos columnar. Lo que significa que técnicamente contiene algoritmos que permiten que las consultas sean más rápidas puesto que cada columna de una tabla de la base de datos está optimizada para realizar consultas más eficientes y de manera nativa.

Otra de las ventajas del uso de BigQuery es que guarda las consultas realizadas de manera física permitiendo guardar persistencia que después es muy conveniente dado que podemos reutilizar esta data en caso de requerir esos datos sin volver a ejecutar otra consulta. Lo anterior es muy importante ya que trabajando en un ambiente en la nube debemos tomar en cuenta que hay detrás de cada consulta a esta Base de Datos un costo por uso de red, por memoria y uso de CPU. por lo cual una forma de mejorar los costos es reutilizar datos consultados que pueden volver a utilizar otro usuario u otro proceso.

Entre las ventajas que tenemos con esta base de datos está el retorno de información a un punto en el tiempo en término de segundos lo cual en caso de errores o por auditoría es muy funcional.

Es una base de datos que permite en todo momento tener estadísticas de consumo y costos, así como permitir el uso de cuotas de consumo por usuario o por grupo de usuarios.

Así mismo permite colocar políticas de acceso a datos no solo a nivel de tablas sino también a nivel de columnas.

“Esta será la base de datos que contendrá toda la información y estructuras de la base de datos para que después sean consumidos por medio de nuestra herramienta de visualización, servicios o por conexiones predefinidas para su consumo “



Linaje de Datos

Dataplex

Dataplex es una herramienta de Google Cloud Platform que permite estandarizar y unificar metadatos, políticas de seguridad, administración, clasificación y administración del ciclo de vida de los datos en estos datos distribuidos.

Entre las funcionalidades que ofrece esta herramienta están: explorar, seleccionar, organizar, unificar datos, así como administrar, supervisar y centralizar datos.

Dataplex registra automáticamente todos los metadatos en un almacén de metadatos unificado.

Puedes acceder a datos y metadatos mediante varios servicios y herramientas como los son: Servicios de Google Cloud, BigQuery, Dataproc Metastore, Data Catalog, Herramientas de código abierto, como Apache Spark y Presto.

El linaje de datos es otra función que permite hacer un seguimiento de cómo los datos se mueven a través de tus sistemas: de dónde provienen, a dónde se pasan y qué transformaciones se aplican a ellos. Lo cual ayuda a tener auditabilidad de los datos.

Dataplex funciona con la API de Data Lineage para identificar las entradas cuyo nombre completamente calificado coincide con las entidades reconocidas por el linaje de datos.

Una de las principales ventajas es que este rastreo del dato se puede visualizar de una manera gráfica que permite fácilmente entender el origen y destino de los datos. Lo cual ayuda a detectar áreas de oportunidad o en caso de fallos identificar la causa raíz del problema sirviendo a su vez de una auto documentación del propio DWH.

“Se utilizará Dataplex para el control, administración, políticas, organización y etiquetado de metadatos de la base de datos o repositorio de datos empresarial que estará alojado en BigQuery “



Administración y Control de Usuarios

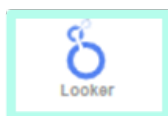
Google IAM

Con Google IAM podemos definir *quién* (identidad) tiene *qué acceso* (rol) a *qué recurso* dentro de Google Cloud Platform. Este componente de Google permite administrar toda la infraestructura de la nube por lo que nos servirá para su vez controlar y administrar la tubería de datos propuesta dado que toda ella se configurará sobre la nube de Google.

Este elemento administra desde máquinas virtuales, storage, base de datos y herramientas de la plataforma.

Nos permite definir grupos y usuarios a los cuales se les puede definir habilitar permisos y otorgar privilegios desde el uso de herramientas, uso de cuotas de servicio, monitoreo de costos y capacidades de acción dentro de la nube de Google.

“Con Google IAM se realiza la administración de accesos y políticas de acceso a recursos dentro de la tubería de datos”



Herramienta de monitoreo y visualización de datos

Looker

Looker es una plataforma de Google para BI (Inteligencia de Negocios) proveniente de las aplicaciones de datos y estadísticas. Con Looker podemos explorar, compartir y visualizar los datos de la empresa para la toma de decisiones.

Looker es una herramienta de visualización de datos que permite crear reportes y tableros de información con diferentes fuentes de datos desde archivos de Excel hasta conexiones a diferentes

bases de datos como BigQuery. En este caso la base de datos que utilizaremos para el DWH está alojada en BigQuery por lo que la conexión es una conexión nativa a Looker.

Otra ventaja que tiene Looker es la facilidad de poder encapsular páginas o aplicaciones web dentro de la plataforma de Looker utilizando el control de acceso de usuarios y permisos de la propia plataforma de Looker.

Otras ventajas que tenemos con esta herramienta es el uso de persistencia de datos y precarga de datos. Esto es muy útil también para reducir costos puesto que ya anteriormente hemos hecho énfasis en que el uso de bases de datos en la nube implica costos desde red, memoria, almacenamiento, repositorio de archivos y procesamiento.

Looker permite generar vistas con cierta persistencia que puede ser definida cada X tiempo como se requiera mientras esa vista no vuelva a regenerarse hasta ese momento lanzará una nueva consulta a BD lo que evita que si varios usuarios de compañía visualizan un tablero o reporte no esté enviando en cada ejecución esta consulta hasta BD si no que guarda en caché esta información dentro de la memoria del aplicativo de Looker.

Otra ventaja que explotaremos con esta herramienta es que es una herramienta que puede conectarse con herramientas de control de versiones de código lo que permite tener un desarrollo colaborativo, así como el manejo de diferentes ambientes de prueba y producción.

“Actualmente ya es la plataforma definida para ofrecer a nuestros clientes de negocio dashboard que le permitan analizar la información proveniente de la actividad de nuestros usuarios y se desea mantener para explotar todas sus ventajas”

Se mantendrá como plataforma de explotación y visualización de datos por las ventajas que nos ofrece gracias a sus recursos disponibles para graficar y presentar informes.

También es importante mencionar que otras de las ventajas que ofrece y que actualmente aún no se está explotando es la capacidad de generar aplicaciones propias contenidas dentro del ambiente de Looker (desarrolladas a la medida) así como la recepción y llamado de petición de servicios (API) por medio de conectores proveniente de otros servicios o aplicaciones que con la implementación de la infraestructura planteada para el análisis de datos podrá ofrecer nuevos recursos e insumos para negocio que le permitan tener mayor valor agregado a sus datos.

Esta herramienta también será utilizada para generar dashboard de control y monitoreo de la tubería para luego ser delegada a un área operativa que se dedique a estas tareas.

Propuesta para monitoreo de la tubería de datos

Se propone desarrollar un sistema de alarmas que esté comunicado con el orquestador (AirFlow) el cual deberá alimentar un tablero en Looker el cual va a monitorear un operador 7x24 donde visualizará si se requiere la notificación a otro equipo que no sea Datos por ejemplo al equipo técnico del sistema fuente o bien al equipo de datos dependiendo del tipo de error que se detecte.

Por otro lado, también se propone tener un botón de emergencia el cual tendrá la función de dar de baja la infraestructura completa de la tubería y volver a iniciarla para luego retomar los procesos donde fueron interrumpidos por algún factor. Tomando en cuenta que la tubería tiene definidos “N” intentos en cada fase de su proceso este botón se habilita en caso de haber excedido estos intentos y no poder auto estabilizarse.

Será necesario establecer un catálogo de errores conocidos con sus acciones de restablecimiento correspondientes con el fin de poder automatizar estas acciones.

Para lo anterior se definen 3 niveles de criticidad.

Emergencia: Corresponde al error arrojado luego de agotar los “N” intentos configurados en la tubería para su restablecimiento (botón de pánico) el cual deberá dar clic el operador para detener, apagar, levantar y restablecer el flujo de la tubería a partir de la fase donde se quedó detenido. Así mismo notificará por Microsoft Teams al equipo de Datos para su conocimiento.

Urgente: Notificará en pantalla al operador, así como por Microsoft Teams al equipo de datos de que debe contactar a determinado equipo ya que en este caso corresponde a errores de falta de datos en los sistemas fuente.

Warning: Se trata de un error no conocido el cual no provocó la detención del flujo de datos por no ser un error en la ruta crítica o detener un proceso al 100%, va a mostrarse en pantalla al operador y notificar por Microsoft Teams al Equipo de Datos.

Dicha propuesta se diagrama en la Figura 17 que se muestra a continuación.

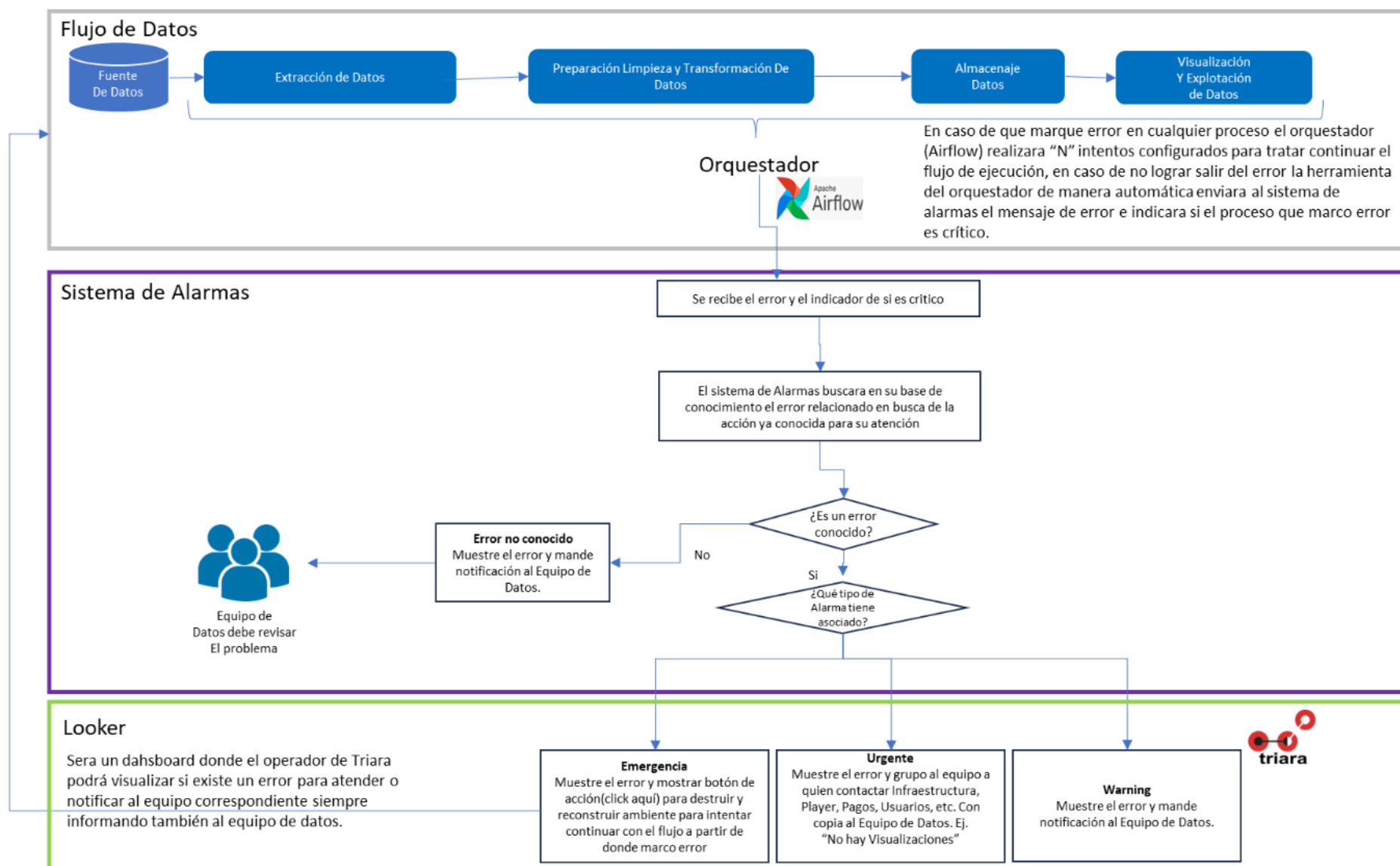


Figura 20. Diagrama de roles y responsabilidades para el monitoreo y gestión de la Tubería de Datos.

Indicadores

Se sugieren los siguientes indicadores para el seguimiento y detección de áreas de oportunidad, así como identificación de cuellos de botella dentro de la tubería de datos.

Extracción de Datos

Fase	Indicador	Valor Actual (mensual)	Valor Esperado después de la implementación	Mejora
Extracción de datos	No. de reprocesos completo por falla en la extracción de datos	15 veces	10 veces	Reducción al 40%
Extracción de datos	Tiempo promedio de carga de datos completa diaria para procesos por lote	4.5 horas	15 min	Reducción del 94% en tiempo de extracción de datos
Extracción de datos	Tiempo promedio entre la generación del dato y carga del mensaje para procesos en tiempo real	4 horas (cada 4 horas se ejecuta carga)	10 min (carga al momento de recibir el mensaje)	Reducción del 96%.
Extracción de datos	No. de procesos críticos con falla	N/A	Tendencia a cero	Identificar procesos para atención inmediata
Extracción de datos	No. de procesos no críticos con falla	N/A	Tendencia a cero	Identificar procesos a mejorar
Extracción de datos	Costo promedio Por carga	N/A	Contar con un costo promedio por carga completa	Identificar costos por carga de datos
Extracción de datos	Tiempo promedio por afectación en extracción de datos	120 horas	Tendencia a 0	

Tabla.5 Indicadores propuestos para monitorear fase de extracción

Preparación Limpieza y Transformación de Datos

Fase	Indicador	Valor Actual (mensual)	Valor Esperado después de la implementación	Mejora
Transformación de datos	No. de reprocesos completo por falla en la transformación de datos	30 veces	10 veces	Reducción al 70%
Transformación de datos	Tiempo promedio de procesamiento de datos en fase de transformación.	2 horas	1 hora	Reducción del 50% en tiempo de transformación de datos
Transformación de datos	Tiempo promedio en procesamiento de mensajes para procesos en tiempo real	2 min	2 min	Tener un control de medición para identificar demora en la actualización de mensajes.
Transformación de datos	No. de procesos críticos con falla	N/A	Tendencia a cero	Identificar procesos para atención inmediata
Transformación de datos	No. de procesos no críticos con falla	N/A	Tendencia a cero	Identificar procesos a mejorar
Transformación de datos	Costo promedio Por carga	N/A	Contar con un costo promedio por carga completa	Identificar costos por carga de datos
Transformación de datos	Tiempo promedio por afectación en extracción de datos	120 horas	Tendencia a 0	

Tabla6. Indicadores propuestos para monitorear fase de transformación de datos

Almacenamiento de Datos

Fase	Indicador	Valor Actual (mensual)	Valor Esperado después de la implementación	Mejora
Almacenamiento de datos	No. de Tablas con información mayor a 3 años	Todas tienen 8 años	Solo tablas agregadas conservaran historia completa	Reducción de costo de almacenamiento en un 37%
Almacenamiento de datos	Costo promedio por tipo de almacenamiento	N/A	Contar con el costo por almacenamiento en línea y fuera de línea	Administrar costos basado en la temporalidad del dato
Almacenamiento de datos	No. de fallas en el traspaso entre almacenamiento en línea a fuera de línea	N/A	Tendencia a cero	Identificar posibles fallas en este proceso

Tabla 7. Indicadores propuestos para monitorear fase de almacenamiento de datos

Visualización de Datos

Fase	Indicador	Valor Actual (diario)	Valor Esperado después de la implementación	Mejora
Visualización de datos	No. de visualizaciones en promedio por día y por hora	50 millones	Diferencia 10% arriba /abajo	Se sugiere implementar un modelo matemático para identificación de pico atípicos. Lo anterior para detectar si existe algún problema al momento de reproducir contenidos.
Visualización de datos	No. de Suscripciones diarias	15 mil	Diferencia 10% arriba /abajo	Identificar si se tiene algún problema en el proceso de suscripción dentro de la plataforma.
Visualización de datos	Promedio de usuarios con visualizaciones en la plataforma	20 millones	25 millones	Al mejorar la experiencia de usuario se espera una retención en clientes activos
Visualización de datos	No. de usuarios que visualizan tableros de Looker	50	100	Aumentar el uso de datos del repositorio central de datos de la compañía.
Visualización de datos	Promedio de ventanas de liberación (tableros)	0.2	2	Aumentar liberaciones a producción
Visualización de datos	Costo promedio, por día generado por Looker	N/A	Contar con un costo promedio por día	Identificar Costo por día y por usuario de Looker
Visualización de datos	No. de conexiones de aplicaciones consumidoras de datos como servicio	N/A	Contar con un indicador de consumidores por servicio	Medir el uso del repositorio de datos desde conexiones por aplicativos

Tabla 8. Indicadores propuestos para monitorear fase de visualización de datos

Operativos en general

Fase	Indicador	Valor Actual (mensual)	Valor Esperado después de la implementación	Mejora
Operativos	Tiempo promedio fuera de servicio	40 horas	8 horas	Reducción del 75%
Operativos	Tiempo promedio dedicado a la operación de datos por equipo de datos	224 horas	40 horas	Reducción del 83%
Operativo	Tiempo promedio de disponibilidad de la infraestructura para carga de datos	7x24	3 horas diarias	Reducción del 98%
Operativo	No. de Recursos humanos dedicados a revisión de temas operativos	5	2	Reducción del 60%
Operativo	Tiempo promedio a revisión y validación de datos	5 horas	1 hora	Reducción del 80%

Tabla9. Indicadores propuestos para monitorear la operación en general

Propuesta estructura Gerencia de Datos

Con la nueva arquitectura se visualiza un nuevo rol en el equipo que se sugiere sea el encargado de mantener y controlar los cambios en la arquitectura implementada llamada Arquitecto de Datos el cual además de estas funciones tendrá la tarea de mantener la arquitectura de datos conforme a las necesidades del negocio con apoyo de las mejores prácticas de la industria y de sistemas.

Se propone cambiar la visión de tener ingenieros de datos exclusivos de extracción y de visualización de datos por una visión más convergente que una en un solo perfil a ingenieros que tengan las capacidades tanto de desarrollar procesos de extracción, transformación y visualización de datos.

Aprovechando a que el equipo actual de datos es el propuesto para la implementación de esta propuesta por el conocimiento de negocio como de las herramientas se puede reducir la curva de aprendizaje para adquirir nuevas habilidades.

Con la propuesta de monitoreo se delega la actividad de operación y revisión de tareas operativas por lo que podemos enfocar al equipo en capacitación con el fin de hacerlo más robusto y esté en las mejores condiciones para proponer ideas novedosas y que aporten mayor valor agregado al negocio.

Se propone fortalecer al equipo con otro Ingeniero en Ciencias de Datos para potencializar los beneficios de la arquitectura propuesta en la búsqueda de patrones y diseño de modelos predictivos.

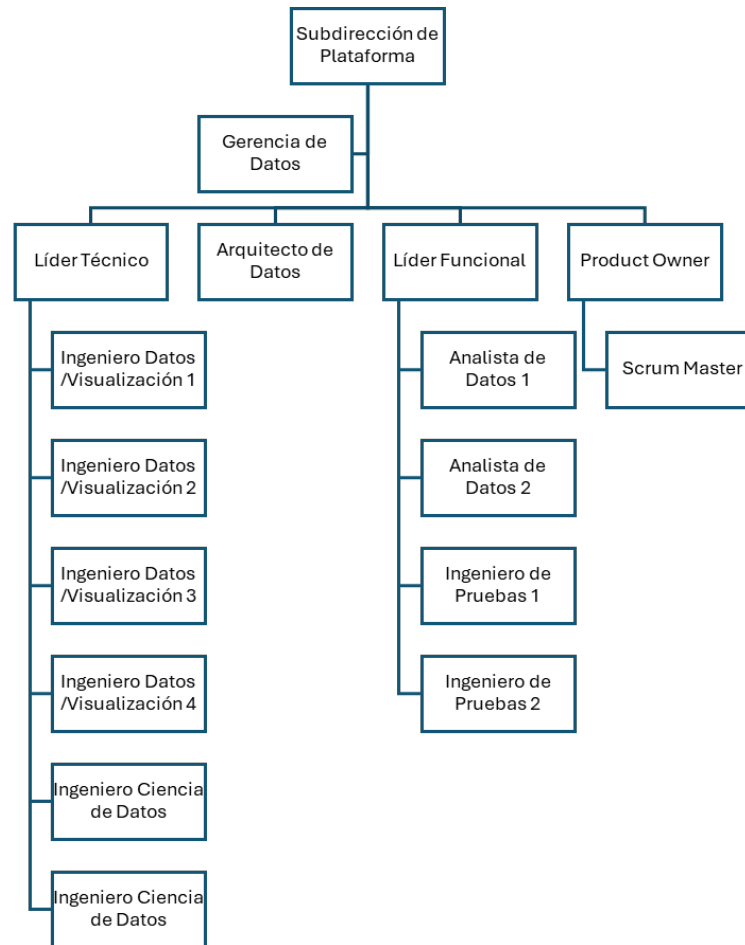


Figura 21. Organigrama Equipo de Datos propuesto

Propuesta de plan para implementación

Gestión del Proyecto

El presente caso de estudio se desarrollará en una empresa mexicana que hoy administra y gestiona sus proyectos bajo el marco de referencia ágil llamado SAFe® que es el acrónimo de Scaled Agile Framework Enterprise.

Este marco de referencia es un conjunto de patrones de organización y flujo de trabajo basado en los principios de Lean, Agile y DevOps para implementar prácticas ágiles a escala empresarial.

SAFe promueve la alineación, la colaboración y la entrega entre una gran cantidad de equipos ágiles. Este marco constituye un cúmulo de conocimientos que incluye instrucciones estructuradas sobre las funciones y responsabilidades, la forma de planificar y gestionar el trabajo sin perder los valores y principios ágiles.¹⁴

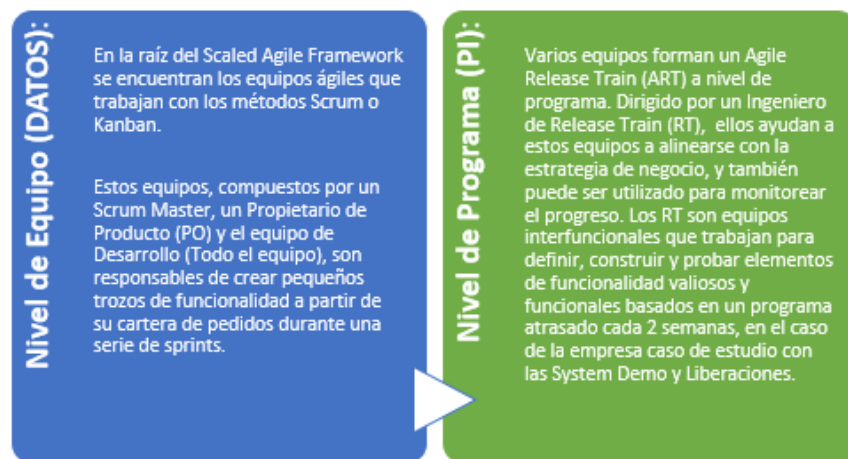


Figura 22. Proceso actual dentro de la empresa que donde se enfoca en presente caso de estudio

¹⁴ Dean Leffingwell © Scaled Agile, Inc. Include this copyright notice with the copied content. Read the FAQs on how to use SAFe content and trademarks here:

<https://www.scaledagile.com/about/about-us/permissions-faq/> © Scaled Agile, Inc. Include this copyright n. (2023). *Welcome to Scale Agile Framework*. © Scaled Agile, Inc. Include this copyright notice with the copied content. Read the FAQs on how <https://www.scaledagile.com/about/about-us/permissions-faq/>

Herramienta de gestión y seguimiento del proyecto



Jira es un producto de software propietario para la gestión de proyectos, seguimiento de errores e incidencias.

La herramienta fue desarrollada por la empresa australiana Atlassian y es la herramienta que la compañía hoy utiliza para el seguimiento de requerimientos por lo que será utilizada para dar seguimiento a cada tarea de este proyecto conforme al plan de trabajo propuesto a continuación.

Cronograma de Tareas

Plan Optimista (5 meses)

El siguiente plan está diseñado en 6 fases tomando en cuenta las 5 fases del flujo de la Tubería de Datos más 1 adicional para propuesta para el monitoreo automatizado.

Fase	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5			
	Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase 1 Extracción de Datos																				
Fase2 Preparación Limpieza y Transformación De Datos																				
Fase3 Almacenaje Datos																				
Fase4 Validación de Datos																				
Fase5 Visualización Y Explotación de Datos																				
Fase6 Monitoreo																				

Figura 23. Cronograma de Tareas - Plan Optimista

En este plan se contempla tareas en paralelo y asignar un subequipo de trabajo dedicado 100% a esta implementación, para fines de este proyecto no se tiene un presupuesto adicional para esta implementación por cuál se reduciría la capacidad instalada para nuevos requerimientos y atención de incidencias al 30% del Equipo de Datos como se muestra en la Tabla 2.

Roles	Recursos reasignados proyecto	Total Equipo Datos actual	% Capacidad
Líder Técnico	1	2	50%
Líder Negocio y QA	1	1	100%
Product Owner	1	1	100%
Analista Funcional	1	2	50%
Tester	2	3	67%
Desarrollador backend	3	4	75%
Desarrollador frontend	3	3	100%
Gerente de Datos	1	1	100%
Total	13	17	76%

Tabla 10. Recursos humanos utilizados en plan optimista

Plan Optimista Detallado por fase (5 meses)



Figura 24. Fase 1 Cronograma - Plan Optimista

Roles		Documentos
TL	Líder Técnico	AT
LNQ	Líder Negocio y QA	AF
PO	Product Owner	MP
AF	Analista Funcional	
QA	Tester	
BE	Desarrollador backend	
FE	Desarrollador frontend	
GTE	Gerente de Datos	

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 1				Mes 2				Mes 3				Mes 4			
					Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase2 Preparación Limpieza y Transformación De Datos	Mapear Flujos de Datos y procesos actuales (As-is)	TL + LNQ	TL, BE	Diagrama As-is																
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, BE	Identificación de Proceso a eliminar en Diagrama As-is																
	Diseño del proceso integración entre la fase anterior y esta de transformación y limpieza de datos	TL	TL + LNQ	AT y Diagrama de Flujo																
	Método de validación de consistencia de datos	LNQ	TL + LNQ	AF, AT, MP																
	Definición de KPIs de monitoreo de esta fase	LNQ	AF + QA	AF																
	Definición de outliers para alertas	LNQ	AF + QA	AF																
	Definición de ajustes a procesos de transformación	TL	TL, AF	AT, AF																
	Definición de Arquitectura tecnológica a utilizar	TL	TL + LNQ	AT y Diagrama de Arquitectura																
	Mapear Flujos de Datos y procesos (To-be)	TL + LNQ	TL, LNQ	Diagrama To-Be																
	Plan de desarrollo e implementación para implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades																
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar																
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación																

Figura 25. Fase 2 Cronograma - Plan Optimista

Roles		Documentos	
TL	Líder Técnico	AT	Análisis Técnico
LNQ	Líder Negocio y QA	AF	Análisis Funcional
PO	Product Owner	MP	Matriz de Prueba
AF	Analista Funcional		
QA	Tester		
BE	Desarrollador backend		
FE	Desarrollador frontend		
GTE	Gerente de Datos		

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 1				Mes 2				Mes 3				Mes 4			
					Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase3 Almacenaje Datos	Mapear Flujos de Datos y procesos actuales(As -Is)	TL + LNQ	TL, BE	Diagrama As-Is																
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, BE	Identificación de Proceso a eliminar en Diagrama As-Is																
	Definición de ciclo de vida del dato (temporalidad)	TL	TL + LNQ	AT y Diagrama de Flujo																
	Definición de procesos de almacenaje	TL + LNQ	TL + LNQ	AF, AT, MP																
	Método de validación de almacenaje	LNQ	AF + QA	AF																
	Definición de KPIS de monitoreo de esta fase	LNQ	AF + QA	AF																
	Definición de outliers para alertas	TL	TL, AF	AT, AF																
	Definición de Arquitectura tecnológica a utilizar	TL	TL + LNQ	AT y Diagrama de Arquitectura																
	Mapear Flujos de Datos y procesos (To-be)	TL + LNQ	TL, LNQ	Diagrama To-Be																
	Plan de desarrollo e implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades																
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar																
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación																

Figura 26. Fase 3 Cronograma - Plan Optimista

Roles

TL	Lider Tecnico
LNQ	Lider Negocio y QA
PO	Product Owner
AF	Analista Funcional
QA	Tester
BE	Desarrollador backend
FE	Desarrollador frontend
GTE	Gerente de Datos

Documentos

AT	Análisis Técnico
AF	Análisis Funcional
MP	Matriz de Prueba

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 1				Mes 2				Mes 3				Mes 4			
					Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase4 Validación de Datos	Mapear Flujos de Datos y procesos actuales (As-is)	TL + LNQ	TL, BE	Diagrama As-is																
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, BE	Identificación de Proceso a eliminar en Diagrama As-is																
	Definición de Método de validación que garantice la calidad y confiabilidad de los datos	LNQ	AF + QA	AF, MP, AT																
	Definición de KPIS de monitoreo de esta fase	TL + LNQ	TL + LNQ	AF																
	Definición de outliers para alertas	TL + LNQ	AF + QA	AF																
	Definición de Arquitectura tecnológica a utilizar	TL + LNQ	TL, LNQ, AF, QA	AT y Diagrama de Arquitectura																
	Mapear Flujos de Datos y procesos (To-be)	TL + LNQ	TL + LNQ	AT, AF																
	Plan de desarrollo e implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades																
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar																
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación																

Figura 27. Fase 4 Cronograma - Plan Optimista

Roles		Documentos	
TL	Lider Tecnico	AT	Análisis Técnico
LNQ	Lider Negocio y QA	AF	Análisis Funcional
PO	Product Owner	MP	Matriz de Prueba
AF	Analista Funcional		
QA	Tester		
BE	Desarrollador backend		
FE	Desarrollador frontend		
GTE	Gerente de Datos		

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5			
					Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase 5 Visualización Y Explotación de Datos	Mapear Flujos de Datos y procesos actuales (As-Is)	TL + LNQ	TL, FE	Diagrama As-Is																				
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, FE	Identificación de Proceso a eliminar en Diagrama As-Is																				
	Definición de Metodología de desarrollo	TL+LNQ+GTE+PO	AF + QA+FE	Diagrama de Proceso de Desarrollo																				
	Definición de KPI'S para monitoreo de esta fase	LNQ	AF + QA	AF																				
	Definición de outliers para alertas	LNQ	AF + QA	AF																				
	Definición de Arquitectura tecnologica a utilizar	TL + LNQ	TL, LNQ, AF, QA	AT y Diagrama de Arquitectura																				
	Mapear Flujos de Datos y procesos actuales de origen a destino (To-be)	TL + LNQ	TL + LNQ	Diagrama To-Be																				
	Plan de desarrollo e implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades																				
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar																				
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación																				

Figura 28. Fase 5 Cronograma - Plan Optimista

Roles		Documentos	
TL	Líder Técnico	AT	Análisis Técnico
LNQ	Líder Negocio y QA	AF	Análisis Funcional
PO	Product Owner	MP	Matriz de Prueba
AF	Analista Funcional		
QA	Tester		
BE	Desarrollador backend		
FE	Desarrollador frontend		
GTE	Gerente de Datos		

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5			
					Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase 6 Monitoreo	Mapear Flujos de Datos y procesos actuales (As-is)	TL + LNQ	TL, LNQ, FE	Diagrama As-is																				
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, LNQ, FE	Identificación de Procesos a eliminar en Diagrama As-is																				
	Diseño de Dashboard de monitoreo operativo	TL + LNQ	TL, LNQ, AF, FE	AF, AT																				
	Definición de sistema de alertas y notas dinámicas	TL + LNQ	TL, LNQ, FE, QA	AF, AT																				
	Definición de acciones sobre outliers	TL + LNQ	TL, LNQ, FE, QA	AF, AT																				
	Plan de desarrollo e implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades																				
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar																				
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación																				

Figura 29. Fase 5 Cronograma - Plan Optimista

Plan Conservador (11 meses)

El siguiente plan está diseñado en 11 fases tomando en cuenta las 5 fases del flujo de la Tubería de Datos más 1 adicional para propuesta para el monitoreo automatizado.

Mes	Sprint	Semana	Fase					
			Fase 1 Extracción de Datos	Fase 2 Preparación Limpieza y Transformación De Datos	Fase 3 Almacenaje Datos	Fase 4 Validación de Datos	Fase 5 Visualización Y Explotación de Datos	Fase 6 Monitoreo
1	1	S1						
		S2						
	2	S1						
		S2						
2	1	S1						
		S2						
	2	S1						
		S2						
3	1	S1						
		S2						
	2	S1						
		S2						
4	1	S1						
		S2						
	2	S1						
		S2						
5	1	S1						
		S2						
	2	S1						
		S2						
6	1	S1						
		S2						
	2	S1						
		S2						
7	1	S1						
		S2						
	2	S1						
		S2						
8	1	S1						
		S2						
	2	S1						
		S2						
9	1	S1						
		S2						
	2	S1						
		S2						
10	1	S1						
		S2						
	2	S1						
		S2						
11	1	S1						
		S2						
	2	S1						
		S2						

Figura 30. Cronograma de Tareas - Plan Conservador

En este plan se contempla un equipo de trabajo dedicado 100% a esta implementación, para fines de este proyecto no se tiene un presupuesto adicional para esta implementación por cuál se reduciría la capacidad instalada al 59% del Equipo de Datos

Roles	Recursos reasignados proyecto	Total Equipo Datos actual	% Capacidad
Líder Técnico	1	2	50%
Líder Negocio y QA	1	1	100%
Product Owner	1	1	100%
Analista Funcional	1	2	50%
Tester	2	3	67%
Desarrollador backend	2	4	50%
Desarrollador frontend	1	3	33%
Gerente de Datos	1	1	100%
Total	10	17	59%

Tabla 11. Recursos humanos utilizados en plan conservador.

Plan Conservador Detallado por fase (11 meses)

Roles		Documentos	
TL	Lider Tecnico	AT	Análisis Técnico
LNQ	Lider Negocio y QA	AF	Análisis Funcional
PO	Product Owner	MP	Matriz de Prueba
AF	Analista Funcional		
QA	Tester		
BE	Desarrollador backend		
FE	Desarrollador frontend		
GTE	Gerente de Datos		

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 1				Mes 2				Mes 3			
					Sprint1		Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase 1 Extracción de Datos	Mapear Flujos de Datos y procesos actuales de origen a destino (As-is)	TL + LNQ	TL's, LNQ, BE, QA	Diagrama As-is												
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL's, BE, QA	Identificación de Proceso a eliminar en Diagrama As-is												
	Diseño de estrategia(s) de extracción de datos	TL	TL, AF	AT y Diagrama de Flujo												
	Método de validación de integridad de datos	TL + LNQ	TL + LNQ	AF, AT, MP												
	Definición de KPIs para garantizar el monitoreo de cada fase	LNQ	AF + QA	AF												
	Definición de outliers para alertas	LNQ	AF + QA	AF												
	Definición de ajustes a procesos de carga	TL	TL + AF	AT, AF												
	Definición de Arquitectura tecnologica a utilizar	TL	TL + LNQA	AT y Diagrama de Arquitectura												
	Mapear Flujos de Datos y procesos (To-be)	TL	TL + LNQA	Diagrama To-Be												
	Plan de desarrollo e implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades												
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar												
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación												

Figura 31. Fase 1 del Plan Conservador

Roles

TL	Líder Técnico
LNQ	Líder Negocio y QA
PO	Product Owner
AF	Analista Funcional
QA	Tester
BE	Desarrollador backend
FE	Desarrollador frontend
GTE	Gerente de Datos

Documentos

AT	Análisis Técnico
AF	Análisis Funcional
MP	Matriz de Prueba

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 3			Mes 4				Mes 5			
					Sprint1	Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase2 Preparación Limpieza y Transformación De Datos	Mapear Flujos de Datos y procesos actuales (As-is)	TL + LNQ	TL, BE	Diagrama As-is											
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, BE	Identificación de Proceso a eliminar en Diagrama As-is											
	Diseño del procesos integración entre la fase anterior y esta de transformación y limpieza de datos	TL	TL + LNQ	AT y Diagrama de Flujo											
	Método de validación de consistencia de datos	LNQ	TL + LNQ	AF, AT, MP											
	Definición de KPIS de monitoreo de esta fase	LNQ	AF + QA	AF											
	Definición de outliers para alertas	LNQ	AF + QA	AF											
	Definición de ajustes a procesos de transformación	TL	TL, AF	AT, AF											
	Definición de Arquitectura tecnológica a utilizar	TL	TL + LNQ	AT y Diagrama de Arquitectura											
	Mapear Flujos de Datos y procesos (To-be)	TL + LNQ	TL, LNQ	Diagrama To-Be											
	Plan de desarrollo e implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades											
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar											
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación											

Figura 32. Fase 2 del Plan Conservador

Roles

TL	Líder Técnico
LNQ	Líder Negocio y QA
PO	Product Owner
AF	Analista Funcional
QA	Tester
BE	Desarrollador backend
FE	Desarrollador frontend
GTE	Gerente de Datos

Documentos

AT	Análisis Técnico
AF	Análisis Funcional
MP	Matriz de Prueba

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 5		Mes 6				Mes 7			
					Sprint2		Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Fase3 Almacenaje Datos	Mapear Flujos de Datos y procesos actuales(As-is)	TL + LNQ	TL, BE	Diagrama As-is										
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, BE	Identificación de Proceso a eliminar en Diagrama As-is										
	Definición de ciclo de vida del dato (temporalidad)	TL	TL + LNQ	AT y Diagrama de Flujo										
	Definición de procesos de almacenaje	TL + LNQ	TL + LNQ	AF, AT, MP										
	Método de validación de almacenaje	LNQ	AF + QA	AF										
	Definición de KPIS de monitoreo de esta fase	LNQ	AF + QA	AF										
	Definición de outliers para alertas	TL	TL, AF	AT, AF										
	Definición de Arquitectura tecnológica a utilizar	TL	TL + LNQ	AT y Diagrama de Arquitectura										
	Mapear Flujos de Datos y procesos (To-be)	TL + LNQ	TL, LNQ	Diagrama To-Be										
	Plan de desarrollo e Implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades										
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar										
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación										

Figura 33. Fase 3 del Plan Conservador

Roles

TL	Líder Técnico
LNQ	Líder Negocio y QA
PO	Product Owner
AF	Analista Funcional
QA	Tester
BE	Desarrollador backend
FE	Desarrollador frontend
GTE	Gerente de Datos

Documentos

AT	Análisis Técnico
AF	Análisis Funcional
MP	Matriz de Prueba

Fase	Actividades	Responsable(s)	Equipo	Entregable				Mes 8			
						Sprint2		Sprint1		Sprint2	
					S2	S1	S2	S1	S2	S1	S2
Fase4 Validación de Datos	Mapear Flujos de Datos y procesos actuales(As –is)	TL + LNQ	TL, BE	Diagrama As-is							
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, BE	Identificación de Proceso a eliminar en Diagrama As-is							
	Definición de Método de validación que garantice la calidad y confiabilidad de los datos	LNQ	AF + QA	AF, MP, AT							
	Definición de KPIS de monitoreo de esta fase	TL + LNQ	TL + LNQ	AF							
	Definición de outliers para alertas	TL + LNQ	AF + QA	AF							
	Definición de Arquitectura tecnológica a utilizar	TL + LNQ	TL , LNQ, AF, QA	AT y Digrama de Arquitectura							
	Mapear Flujos de Datos y procesos (To-be)	TL + LNQ	TL + LNQ	AT , AF							
	Plan de desarrollo e Implementación	PO,TL,LNQ	PO, TL, LNQ	Plan de actividades							
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar							
	Plan de Capacitación	TL + LNQ	PM,TL, LNQ	Calendario y Material de Capacitación							

Figura 34. Fase 4 del Plan Conservador

Roles

TL	Líder Técnico
LNQ	Líder Negocio y QA
PO	Product Owner
AF	Analista Funcional
QA	Tester
BE	Desarrollador backend
FE	Desarrollador frontend
GTE	Gerente de Datos

Documentos

AT	Análisis Técnico
AF	Análisis Funcional
MP	Matriz de Prueba

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 9				Mes 10			
					Sprint1		Sprint2		Sprint1		Sprint2	
					S1	S2	S1	S2	S1	S2	S1	S2
Fase5 Visualización Y Explotación de Datos	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, FE	Identificación de Proceso a eliminar en Diagrama As-is								
	Definición de Metodología de desarrollo	TL+LNQ+GTE+PO	AF + QA+FE	Diagrama de Proceso de Desarrollo								
	Definición de KPIS para monitoreo de esta fase	LNQ	AF + QA	AF								
	Definición de outliers para alertas	LNQ	AF + QA	AF								
	Definición de Arquitectura tecnológica a utilizar	TL + LNQ	TL, LNQ, AF, QA	AT y Diagrama de Arquitectura								
	Mapear Flujos de Datos y procesos actuales de origen a destino (To-be)	TL + LNQ	TL + LNQ	Diagrama To-Be								
	Plan de desarrollo e Implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades								
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar								
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación								

Figura 35. Fase 5 del Plan Conservador

Roles

TL	Líder Técnico
LNQ	Líder Negocio y QA
PO	Product Owner
AF	Analista Funcional
QA	Tester
BE	Desarrollador backend
FE	Desarrollador frontend
GTE	Gerente de Datos

Documentos

AT	Análisis Técnico
AF	Análisis Funcional
MP	Matriz de Prueba

Fase	Actividades	Responsable(s)	Equipo	Entregable	Mes 10	Mes 11			
					Sprint2	Sprint1		Sprint2	
					S2	S1	S2	S1	S2
Fase6 Monitoreo	Mapear Flujos de Datos y procesos actuales(As -is)	TL + LNQ	TL, LNQ, FE	Diagrama As-is					
	Eliminar procesos duplicados o ya no utilizados	TL + LNQ	TL, LNQ, FE	Identificación de Proceso a eliminar en Diagrama As-is					
	Diseño de Dashboard de monitoreo operativo	TL + LNQ	TL , LNQ, AF, FE	AF, AT					
	Definición de sistema de alertas y notas dinámicas	TL + LNQ	TL, LNQ, FE, QA	AF, AT					
	Definición de acciones sobre outliers	TL + LNQ	TL, LNQ, FE, QA	AF, AT					
	Plan de desarrollo e Implementación	PO, TL, LNQ	PO, TL, LNQ	Plan de actividades					
	Aprobación de Plan de Trabajo y Diagrama To-be	GTE	GTE	OK para continuar					
	Plan de Capacitación	TL + LNQ	PM, TL, LNQ	Calendario y Material de Capacitación					

Figura 36. Fase 6 del Plan Conservador

Conclusiones

Se propone un Diseño de Arquitectura que permita integrar la operación diaria de la Gerencia de Datos que permite el constante monitoreo y rastreo de los datos de manera más eficiente a su operación actual.

Con el diseño propuesto se buscó mejorar el tiempo de procesamiento, pero también la detección de errores desde su origen.

Al tener una visión más completa y panorámica de la operación de la empresa, los procesos operativos del área de datos sus principales necesidades de información así como el objetivo que busca la compañía a corto y mediano plazo se logró ofrecer una propuesta innovadora que además de reducir el consumo de recursos tecnológicos y aprovecha todas las ventajas y beneficios de su arquitectura tecnológica actual ya que esta empresa se ha preocupado siempre de mantenerse actualizada en las nuevas novedades tecnológicas y ha invertido en plataformas y aplicaciones tecnológicas de punta.

Sin embargo, aquí es donde se detectó la oportunidad de aprovechar estas herramientas y plataformas que ya están siendo pagadas y presupuestadas, pero por desconocimiento debido a la falta de tiempo y recursos humanos disponibles para la investigación ya que el mayor tiempo del tiempo el área de Datos está enfocado en entregar en “El reporte de Hoy” que dicho sea de paso además se entregará con desfase de tiempo lo que provoca toma de decisiones tardías o que pierden ese tiempo de oportunidad.

El uso de recursos en la nube normalmente para este tipo de empresas es muy atractivo ya que se paga más y se obtiene un menor tiempo de respuesta por el uso de las herramientas disponibles, aunque no necesariamente sea el tiempo deseado ya que sin una reingeniería que prepare su información para ser procesada en un ambiente como este. Pocas empresas realmente analizan lo que requieren y cómo pueden hacer uso eficiente y racional de recursos antes de subirse a esa “Nube” que paradójicamente termina siendo una “Nube, soñada”, la cual funciona inmediatamente termine la migración como si fuera solo ya me subí con mis procesos y mis datos ya pagué ahora que resuelva esto.

Con el Diseño propuesto no solo se logrará reducir el tiempo operativo dedicado a limpiar y tratar datos de manera masiva si no que se prepara a la compañía para su posible crecimiento o decrecimiento ofreciendo la principal ventaja de un entorno virtual que es la elasticidad pagando únicamente lo necesario en el momento necesario.

Dicho diseño también ofrece la capacidad de oportunidad para explorar nuevas formas de generar valor agregado por parte del área de Datos, por ejemplo, generar un perfilamiento de clientes

basado en un modelo de consumo podría ayudar a la retención de clientes mejorando la experiencia de usuario en la interacción de los usuarios de la plataforma de consumo de video.

Este tipo de plataformas normalmente son muy grandes y atienden a miles de usuarios por lo que si se logra hacer sentir único al usuario puede generar una muy grata experiencia ya que puede sentir que lo conocen como si adivinaran sus necesidades y gustos generando una fidelidad del cliente.

Lo anterior da pauta al negocio de enfocarse en crecer y crear nuevos productos, campañas de uso o nuevos servicios convirtiéndola en una empresa que se maneja, opera y camina desde el conocimiento de sus datos por lo que el reducir las tareas operativas que solo distraen al área de datos sin ningún valor agregado a nivel negocio se vuelve algo vital.

Bibliografía

Atwal, H. (2019). *Practical DataOps: Delivering Agile Data Science at Scale*. Apress.

Uttamchandani, S. (2020). *The Self-service Data Roadmap: Democratize Data and Reduce Time to Insight*. O'Reilly.

Cristopher Bergh. (2020). *The DataOps Cookbook: Methodologies and Tools That Reduce Analytics Cycle Time While Improving Quality* (Inglés ed.).

Schmidt, J. G. (2019). *DataOps: The Authoritative Edition (English Edition)* (Inglés ed.).

Lakshmanan, V. (2022). *Data Science on the Google Cloud Platform: Implementing End-to-end Real-time Data Pipelines: from Ingest to Machine Learning*. O'Reilly.

Lakshmanan, V., & Tigani, J. (2019). *Google BigQuery: The Definitive Guide: Data Warehousing, Analytics, and Machine Learning at Scale*. O'Reilly Media, Incorporated.

Mucchetti, M. (2020). *BigQuery for Data Warehousing: Managed Data Analysis in the Google Cloud*. Apress.

Bibliografía Web

(n.d.). Agile Alliance. Retrieved November 15, 2024, from <http://www.agilealliance.org/>

(n.d.). Manifesto for Agile Software Development. Recuperado 15 noviembre 2024 de <http://agilemanifesto.org/>

(n.d.). ▷ Lean Manufacturing: Qué es y cómo implementarlo en tu empresa. Recuperado 11 julio 2023 de <https://leanmanufacturing10.com/>

About. (n.d.). Scaled Agile Framework. Recuperado 15 noviembre 2024 de <https://scaledagileframework.com/about/>

The DataOps Manifesto - Read The 18 DataOps Principles. (n.d.). DataOps Manifesto. Recuperado 15 noviembre 2024 de <https://DataOpsmanifesto.org/es/>

Dean Leffingwell © Scaled Agile, Inc. Include this copyright notice with the copied content. Read the FAQs on how to use SAFe content and trademarks here: <https://www.scaledagile.com/about/about-us/permissions-faq/> © Scaled Agile, Inc. Include this copyright n. (2023). *Welcome to Scale Agile Framework.* © Scaled Agile, Inc. Include this copyright notice with the copied content. Read the FAQs on how <https://www.scaledagile.com/about/about-us/permissions-faq/>

Gunn, J. (2023, February 13). *Revisión de desempeño de Netflix, Prime Video, etc. en México.* Robotto.mx. Recuperado 15 noviembre 2024 de <https://robotto.mx/2023/02/revision-de-desempeno-de-netflix-prime-video-etc-en-mexico/>

Manuel Rodriguez, Manuel Mazzara, & Pires de Araújo, L. J. (2020, Diciembre 18). Buenas prácticas para la adopción de DataOps en la industria del software. *Journal of Physics: Conference Series*, 1694. <https://iopscience.iop.org/article/10.1088/1742-6596/1694/1/012032/pdf>

Prieto, M. (2018, October 10). *Las 5 tendencias en el Hype Cycle de tecnologías emergentes 2018 de Gartner*. smartlighting. Recuperado 15 noviembre 2024 de <https://smart-lighting.es/tendencias-tecnologias-emergentes-2018-gartner/>

¿Qué son Las Tuberías de Datos? | Glosario. (n.d.). Hewlett Packard Enterprise (HPE). Recuperado 15 noviembre 2024 de <https://www.hpe.com/mx/es/what-is/data-pipelines.html>

Rodríguez, S., & Balmaceda, A. (n.d.). *House Of Cards: el éxito que Netflix creó gracias a la inteligencia artificial*. Big Data Magazine. Recuperado 15 noviembre 2024 de <https://bigdatamagazine.es/house-of-cards-el-exito-que-netflix-creo-gracias-a-la-inteligencia-artificial>

Rueda, L. V., Díaz, S. o., & Sanabria, J. S. G. (2023, June 16). YouTube. Recuperado 10 octubre 2023 de <https://www.risti.xyz/issues/ristie38.pdf>

Techedgegroup. (2023, June 16). Techedgegroup. Recuperado 5 diciembre 2023 de <https://www.techedgegroup.com/es/blog/que-es-DataOps-y-su-importancia-en-las-iniciativas-de-analitica-de-datos>

Tena, M. (n.d.). *¿Qué es la metodología 'agile'? ¿Revolución de las formas de trabajo?* BBVA. Recuperado 15 noviembre 2024 de <https://www.bbva.com/es/innovacion/metodologia-agile-la-revolucion-las-formas-trabajo/>

TIBCO. (2023, Enero). *¿Qué son los datos estructurados?* <https://www.tibco.com/es/reference-center/what-is-structured-data>. Retrieved Junio 16, 2023, from <https://www.tibco.com/es/reference-center/what-is-structured-data>

Glosario

D

Dato: Es una representación simbólica de una variable o atributo de un hecho, objeto un suceso o una entidad de la realidad.

Data Driven: Es una forma estratégica de operar y tomar decisiones basada en el conocimiento y análisis de datos.

DataOps: Es una forma de gestionar y operar los datos basada en los principios ágiles y basada en la automatización de procesos con apoyo de herramientas tecnológicas.

DevOps: Es una forma de desarrollar y operar de manera automatizada conjuntando procesos y tecnología para ofrecer un mayor valor agregado basado en un proceso de mejora continua.

E

Engagement: Significa conexión emocional, pero que hace referencia a cuán activamente se encuentra involucrado el individuo en una determinada actividad, en términos comerciales se refiere a este vínculo del cliente con la marca o el servicio que ofrece la compañía a sus clientes muchas veces está asociado a la fidelidad del cliente.

G

Google Analytics: Es una herramienta de analítica web de la empresa Google que ofrece información agrupada del tráfico que llega a los sitios web según la audiencia, la adquisición, el comportamiento y las conversiones que se llevan a cabo en el sitio web.

H

Hype Cycle: Es una representación gráfica publicada por Gartner en forma de curva que representa la madurez y adopción de las tecnologías o aplicaciones cómo son potencialmente relevantes para resolver problemas comerciales reales y aprovechar nuevas oportunidades.

La metodología del Hype Cycle (literalmente, "ciclo de sobre expectativa" en español y a menudo llamado "curva de Gartner") te permite ver cómo una tecnología o aplicación evolucionará con el tiempo, y te proporciona una fuente sólida de información para que puedas gestionar su implementación dentro del contexto de tus objetivos de negocio específicos.¹⁵

I

Información: Es un conjunto organizado de datos relevantes con un contexto que en su conjunto pueden generar un conocimiento sobre un hecho, suceso, entidad o tema específico.

¹⁵*Hype Cycle.* (n.d.). Gartner.Recuperado 15 noviembre 2024 de <https://www.gartner.mx/es/metodologias/hype-cycle>

J

JIT: Las siglas JIT (del japonés just-in-time), también denominado sistema Justo a Tiempo, es una metodología originalmente creada para empresas japonesas para la producción cuyo objetivo es el de contar únicamente con la cantidad necesaria de producto, en el momento y lugar justo, eliminar cualquier desperdicio o elemento que no aporte valor.

JSON: Es un formato de archivo de texto que se utiliza para intercambiar y guardar información estructurada en una estructura de tipo clave-valor donde las claves se pueden explicar cómo etiquetas o identificadores de los valores, siendo estos últimos los datos asociados a la etiqueta o clave.

L

Lago de datos: Es un repositorio centralizado de datos que almacena todos sus datos estructurados y no estructurados sin modificarlos y sin tener que estructurarlos primero.

M

Metadato: Se denominan así a los datos que describen a los datos sobre su contenido, disponibilidad, condiciones, calidad y en general son todas sus características.

P

Pipeline de Datos: Es un grupo de tareas que permiten mover datos de un lugar a otro para transformarlos en información.

T

TIBCO: Empresa estadounidense de software de inteligencia empresarial

Relación de Tablas

Nombre de la tabla	Página
Tabla 1. Ejemplo de Tabla de datos de una base de datos relacional	7
Tabla 2. Cuestionario Gerencia	41
Tabla 3. Cuestionario Operativo	44
Tabla 4. Entrevista Líderes Técnicos	47
Tabla 5. Indicadores propuestos para monitorear fase de extracción	78
Tabla 6. Indicadores propuestos para monitorear fase de transformación de datos	79
Tabla 7. Indicadores propuestos para monitorear fase de almacenamiento de datos	80
Tabla 8. Indicadores propuestos para monitorear fase de visualización de datos	81
Tabla 9. Indicadores propuestos para monitorear la operación en general	82
Tabla 10. Recursos humanos utilizados en plan optimista	81
Tabla 11. Recursos humanos utilizados en plan conservador.	96

Relación de Ilustraciones

Nombre de la Ilustración	Página
Figura 1. Datos generados o disponibles del usuario de una plataforma de vídeo	6
Figura 2. Ejemplo de diagrama de Flujo de Información en tiempo real	9
Figura 3. Ejemplo de un procesamiento por lotes	10
Figura 4. Una tubería de agua es una analogía de un Pipeline de Datos	14
Figura 5. Fases principales de una tubería de datos	16
Figura 6. Top de Empresas de procesamiento de video en tiempo real que ofrecieron servicios al mercado mexicano en 2022	20
Figura 7. Gartner nombró a DataOps en su gráfico de madurez y adopción de tecnologías para la gestión de datos en 2018	23
Figura 8. Se muestran los 5 principios Lean publicados en el libro Lean Thinking en 1996	27
Figura 9. En la siguiente imagen se representa un flujo de desarrollo DevOps	29
Figura 10. Ciclo de vida de DataOps, siguiendo el principio de DevOps	34
Figura 11. Organigrama de la Gerencia de Datos	39
Figura 12. Diagrama que muestra los tipos de flujos actuales para el tratamiento de la información identificados en el análisis diagnóstico de este caso de estudio	40
Figura 13. Flujo identificado para el tratamiento de datos para su análisis de información en el presente caso de estudio	40
Figura 14. Diagrama de arquitectura tecnológica actual detectado en el diagnóstico del presente caso de estudio	42
Figura 15. Tipos de flujos identificados en la situación actual	43
Figura 16. Figura. Diagrama de Procesos To-Be propuesto para este caso de estudio	48
Figura 17. Desarrollo e Integración continua	50
Figura 18. Arquitectura en la nube	54
Figura 19. Arquitectura Tecnológica propuesta	55
Figura 20. Diagrama de roles y responsabilidades para el monitoreo y gestión del Pipeline	64

Figura 21. Organigrama Equipo de Datos propuesto	69
Figura 21. Proceso actual dentro de la empresa que donde se enfoca en presente caso de estudio	71
Figura 22. Cronograma de Tareas - Plan Optimista	70
Figura 23. Fase 1 Cronograma - Plan Optimista	72
Figura 24. Fase 2 Cronograma - Plan Optimista	73
Figura 25. Fase 3 Cronograma - Plan Optimista	74
Figura 26. Fase 4 Cronograma - Plan Optimista	75
Figura 27. Fase 5 Cronograma - Plan Optimista	76
Figura 28. Fase 5 Cronograma - Plan Optimista	77
Figura 29. Cronograma de Tareas - Plan Conservador	78
Figura 30. Fase 1 del Plan Conservador	80
Figura 31. Fase 2 del Plan Conservador	81
Figura 32. Fase 3 del Plan Conservador	82
Figura 33. Fase 4 del Plan Conservador	83
Figura 34. Fase 5 del Plan Conservador	84
Figura 35. Fase 6 del Plan Conservador	85

Anexos

Anexo 1

Resultados del Cuestionario para la Gerencia

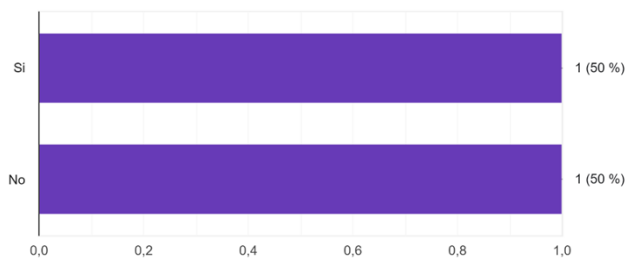
¿Qué tan frecuente cambian los procesos de negocio de la compañía?
2 respuestas



¿El área de datos es responsable de las tareas de desarrollo y operación?
2 respuestas

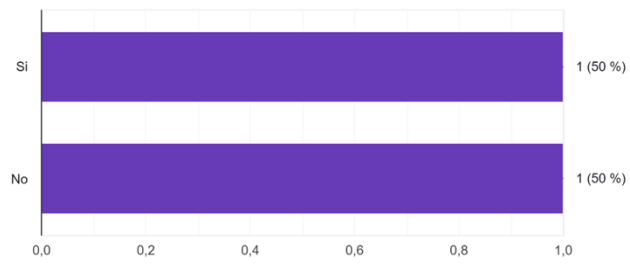


¿Existe personas dedicadas exclusivamente a temas operativos?
2 respuestas



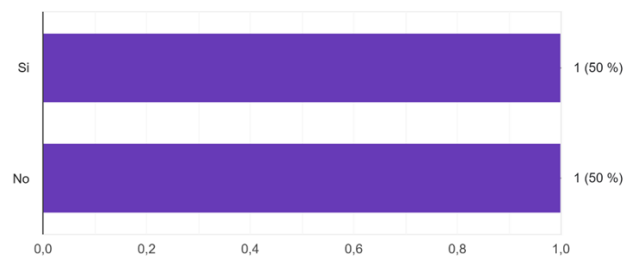
¿Hoy manejan algún estándar de desarrollo?

2 respuestas



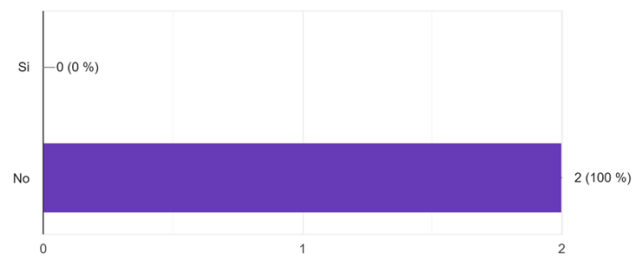
¿Cuentan con indicadores operativos del área?

2 respuestas



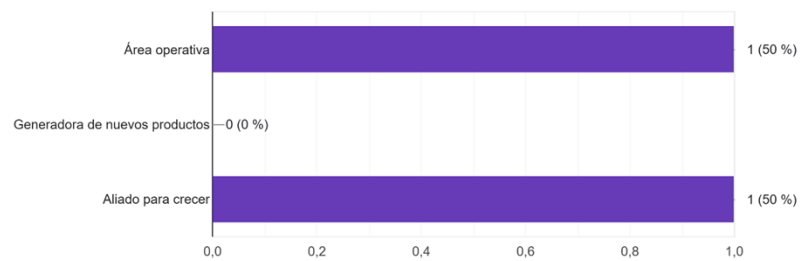
¿Realizan monitoreo automatizado de sus procesos?

2 respuestas



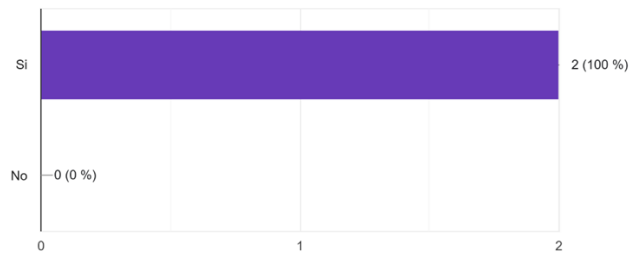
¿Como consideran dentro de la compañía al área de datos?

2 respuestas



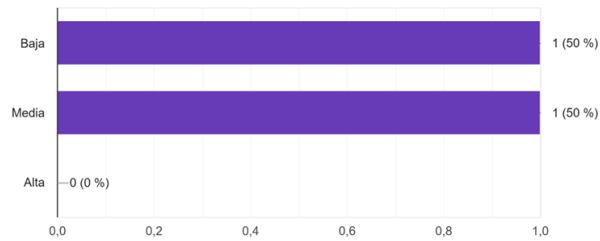
¿Conoce el termino DataOps?

2 respuestas



¿Si se implementara DataOps en la compañía, cuál sería el nivel de resistencia del equipo de datos a este cambio?

2 respuestas



¿Por qué?

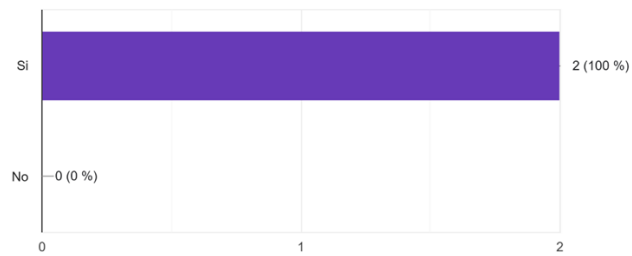
2 respuestas

El equipo tiene interés por temas tecnológicos

Se requiere capacitación al equipo técnico

¿Cree que puedan ofrecerle algún beneficio a la compañía la implementación de DataOps?

2 respuestas

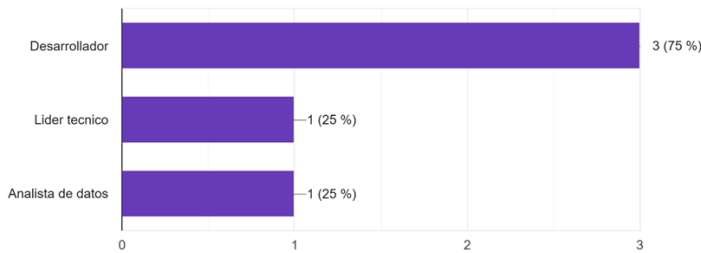


Anexo 2

Resultados del Cuestionario para el Área Operativa

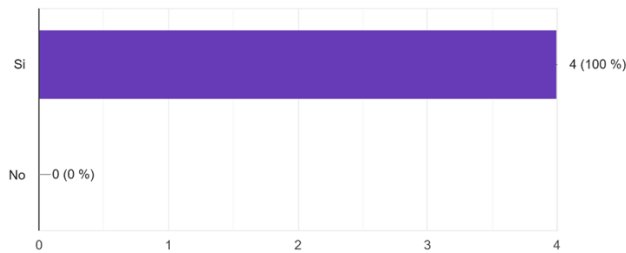
¿Cuál es tu rol en el equipo de Datos?

4 respuestas



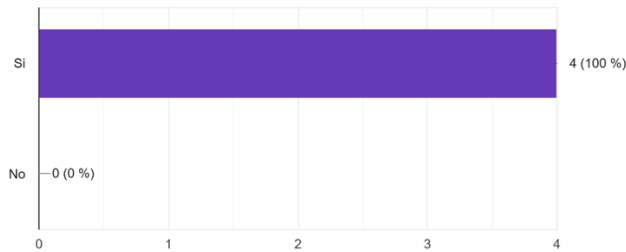
¿Tienen procesos de datos diarios?

4 respuestas



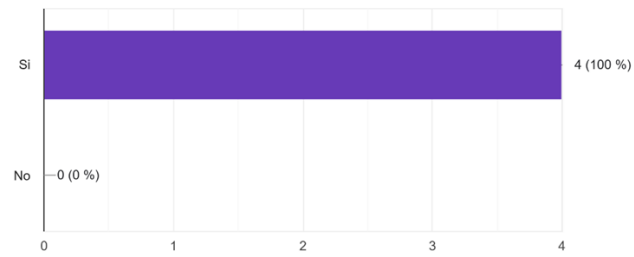
¿Tienen procesos de datos mensuales?

4 respuestas



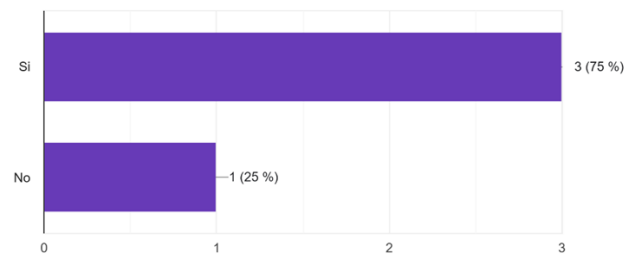
¿Tienen procesos de datos en tiempo real?

4 respuestas



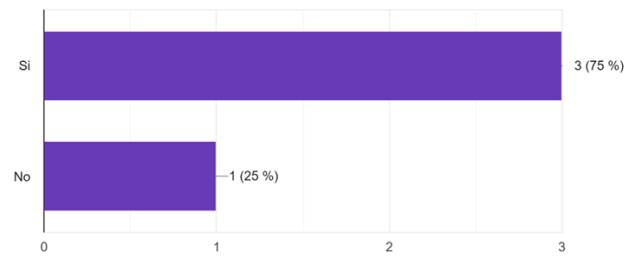
¿Tienen procesos de datos nocturnos?

4 respuestas



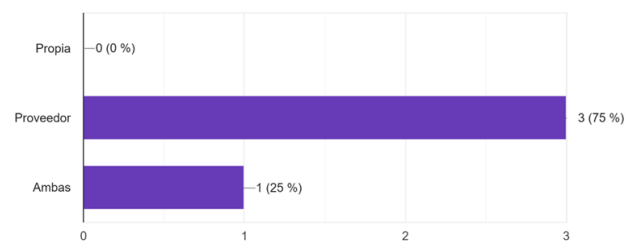
¿Tienen procesos de validación de datos?

4 respuestas

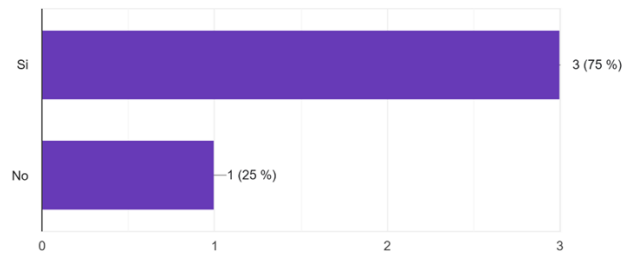


¿La base datos actual donde reside el repositorio de datos de la compañía se encuentra en infraestructura propia o de algún proveedor?

4 respuestas



¿En caso de no tener infraestructura propia crees que la compañía pudiera cambiar de proveedor?
4 respuestas



¿Cuántos registros tiene la tabla más grande que hoy se tiene en su repositorio de datos de la compañía?
4 respuestas

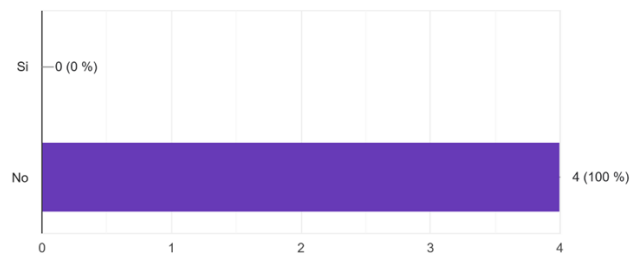
Más de 50 Millones

5 mil millones

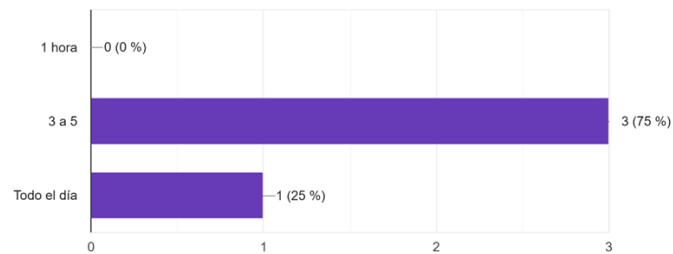
500M

10 tb

¿Tienen algún proceso para depuración de información histórica?
4 respuestas

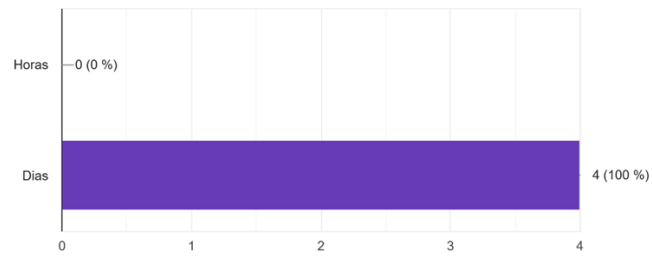


¿Cuánto tiempo del día dedican a temas operativos diarios?
4 respuestas



¿Cuánto tiempo del mes dedican a temas operativos mensuales?

4 respuestas



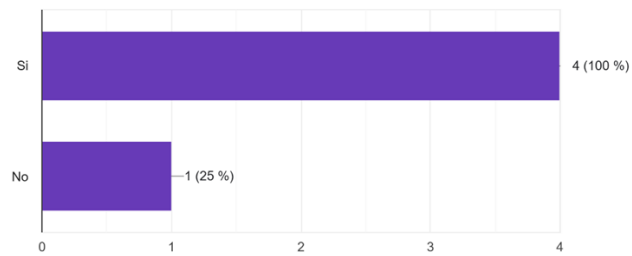
¿Hoy manejan algún estándar de desarrollo?

4 respuestas



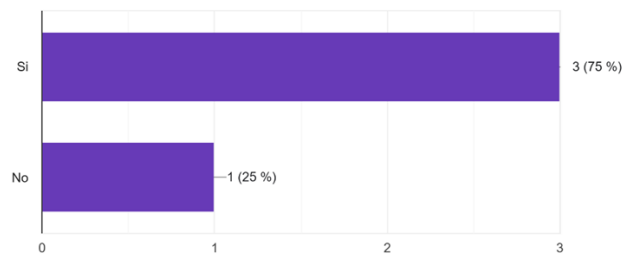
¿Tienen documentación de sus procesos operativos?

4 respuestas

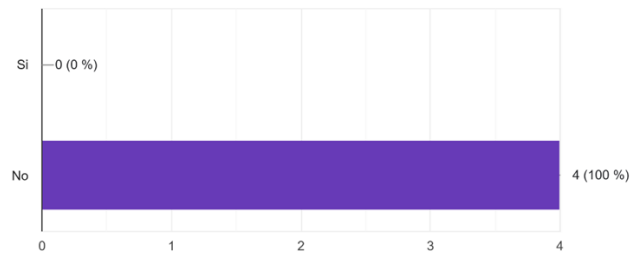


¿Sabes si existe documentación de ingesta de datos?

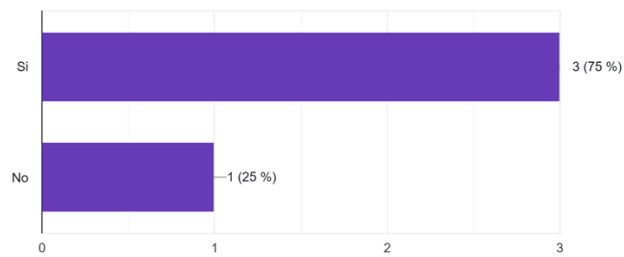
4 respuestas



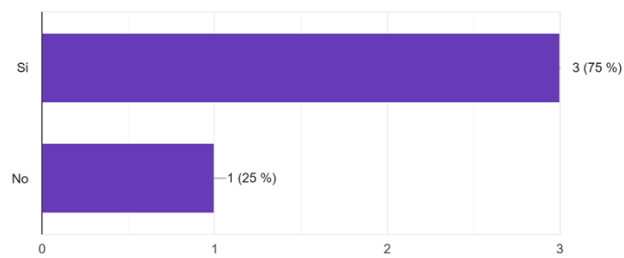
¿Sabes si existe documentación de calidad de datos?
4 respuestas



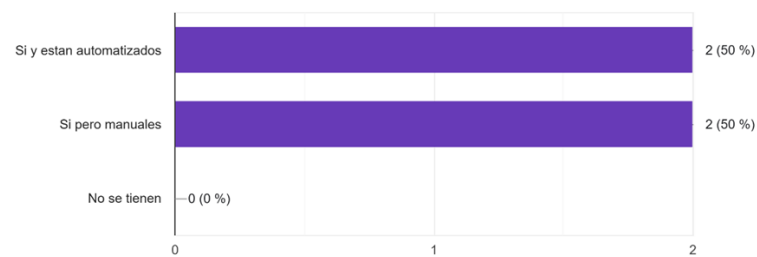
¿Sabes si existe documentación de validación de datos?
4 respuestas



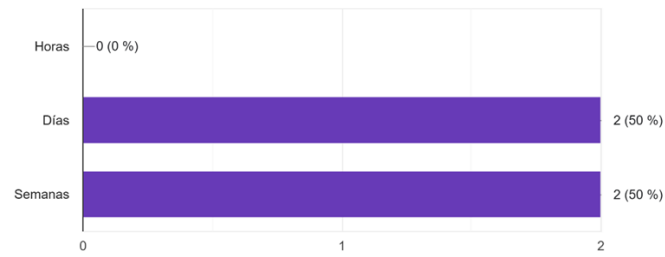
¿Sabes si existe documentación de validación de datos?
4 respuestas



¿Tienen algún control que permita monitorear indicadores operativos?
4 respuestas



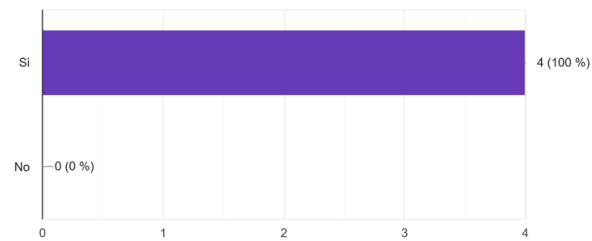
¿Cuánto tiempo tomaría desarrollar desde cero el flujo datos más sencillo que hoy tiene el área?
4 respuestas



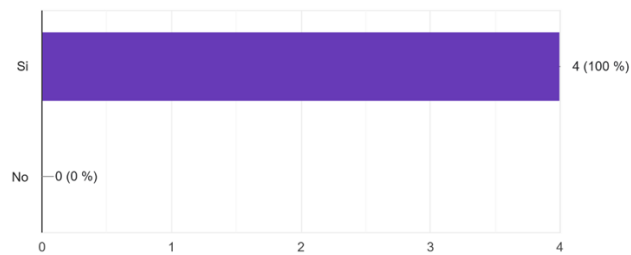
¿Cuándo realizas un desarrollo, puedes reutilizar componentes o código?
4 respuestas



¿Si tuvieras componentes reutilizables, crees que el tiempo que utilizas para desarrollar sería menor?
4 respuestas

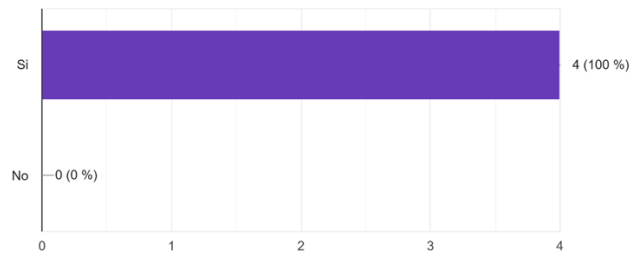


¿Existen procesos manuales que podrían ser automatizados?
4 respuestas



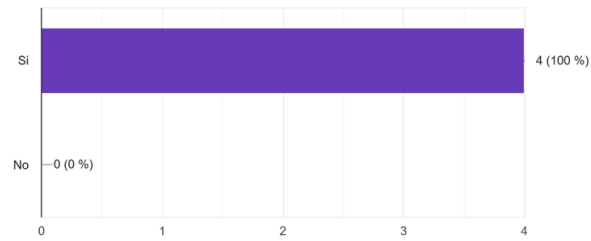
¿Tienen procesos que realizan la misma tarea?

4 respuestas



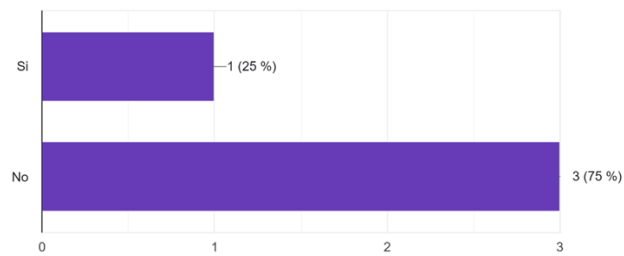
¿Consideras que si se automatizaran las tareas de limpieza y carga de datos se pudiera dedicar ese tiempo para nuevos desarrollos?

4 respuestas



¿Cuándo ocurre un error en la carga de datos es fácil saber en qué paso ocurrió el error?

4 respuestas



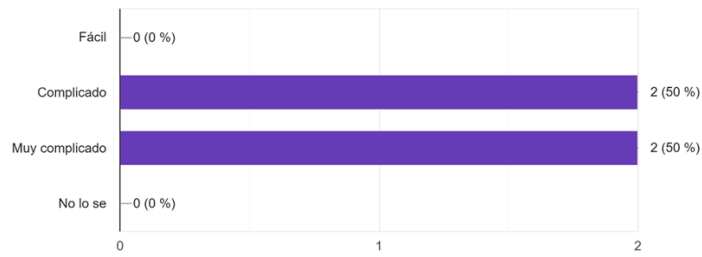
¿De las horas laborables en un día normal cuantas horas le dedicas a la corrección de errores?

4 respuestas



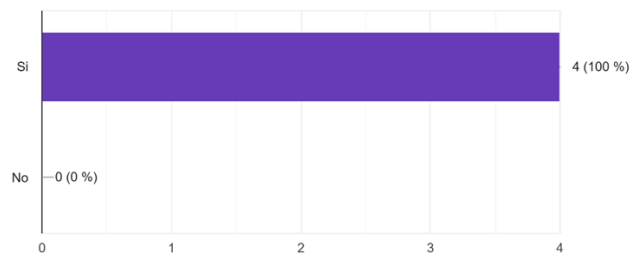
¿Si se requiere realizar algún reproceso de datos que tan complicado es?

4 respuestas



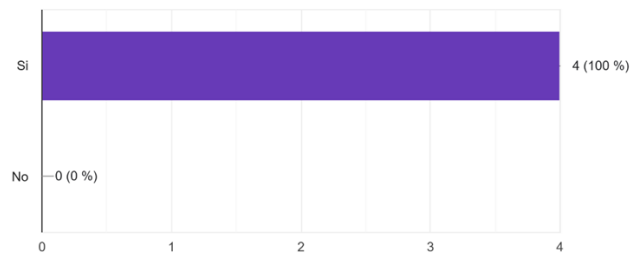
¿Conoces el termino DataOps?

4 respuestas



¿En caso de conocerlo, crees que podría ser una solución para automatizar el flujo de datos?

4 respuestas



¿En caso de no conocerlo te gustaría aprender sobre él y su implementación?

4 respuestas

