Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Profesional de Ingeniería de Sistemas

Implementación de un Data Lake para la centralización de datos analíticos y transaccionales en la empresa Belcorp

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Michael Antonio MARTÍN BALBOA

ASESOR

Mg. Rosa MENÉNDEZ MUERAS

Lima, Perú

2022

Referencia bibliográfica

Martín, M. (2022). *Implementación de un Data Lake para la centralización de datos analíticos y transaccionales en la empresa Belcorp*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Michael Antonio Martín Balboa
Tipo de documento de identidad	DNI
Número de documento de identidad	47559826
URL de ORCID	https://orcid.org/0000-0002-0072-3319
Datos de asesor	
Nombres y apellidos	Rosa Menéndez Mueras
Tipo de documento de identidad	DNI
Número de documento de identidad	10246770
URL de ORCID	https://orcid.org/0000-0003-2403-7679
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Santiago Domingo Moquillaza Henríquez
Tipo de documento	DNI
Número de documento de identidad	08280889
Miembro del jurado 1	
Nombres y apellidos	Frank Edmundo Escobedo Bailón
Tipo de documento	DNI
Número de documento de identidad	41671087
Datos de investigación	
Línea de investigación	No aplica
Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento

Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Lima Distrito: Cercado de Lima Jr. Carlos Amezaga No. 375 Universidad Nacional Mayor de San Marcos Latitud: -12.0564232 Longitud: -77.0843327
Año o rango de años en que se realizó la investigación	2022
URL de disciplinas OCDE	2.02.04 -- Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación
del Trabajo de Suficiencia Profesional

Siendo las 19:00 horas del día 17 de agosto del año 2022, se reunieron virtualmente los docentes designados como Miembros de Jurado del Trabajo de Suficiencia Profesional, presidido por el Mg. Moquillaza Henríquez Santiago Domingo (Presidente), Dr. Escobedo Bailón Frank Edmundo (Miembro) y la Mg. Menéndez Mueras Rosa (Miembro Asesor), usando la plataforma Meet (https://meet.google.com/tfs-rxrv-fza), para la sustentación virtual del Trabajo de Suficiencia Profesional intitulado: ***"IMPLEMENTACIÓN DE UN DATA LAKE PARA LA CENTRALIZACIÓN DE DATOS ANALÍTICOS Y TRANSACCIONALES EN LA EMPRESA BELCORP",*** *por el Bachiller* ***Martín Balboa Michael Antonio;*** *para obtener el Título Profesional de Ingeniero de Sistemas.*

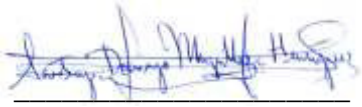
Acto seguido de la exposición del Trabajo de Suficiencia Profesional, el Presidente invitó al Bachiller a dar las respuestas a las preguntas establecidas por los miembros del Jurado.

El Bachiller en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el Bachiller obtuvo la nota de ***17 (DIECISIETE).***

A continuación el Presidente de Jurado el Mg. Moquillaza Henríquez Santiago Domingo, declara al Bachiller ***Ingeniero de Sistemas****.*

Siendo las ***19:55*** *horas, se levantó la sesión.*



Presidente
Mg. Moquillaza Henríquez Santiago Domingo

Miembro
Dr. Escobedo Bailón Frank Edmundo

Miembro Asesor
Mg. Menéndez Mueras Rosa

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

INFORME DE EVALUACIÓN DE ORIGINALIDAD

1. Facultad de Ingeniería de Sistemas e Informática
2. Escuela Profesional de Ingeniería de Sistemas
3. Autoridad académica que emite el informe de originalidad
   Directora (e) de la EPIS
4. Apellidos y Nombres de la autoridad académica
   Dra. Luzmila Elisa Pró Concepción
5. Operador del programa informático de similitudes
   Dra. Luzmila Elisa Pró Concepción
6. Documento evaluado
   Título de pregrado: **"Implementación de un Data lake para la centralización de datos analíticos y transaccionales en la empresa Belcorp"**
7. Autor del documento
   Bach. **Martín Balboa, Michael Antonio**
8. Fecha de recepción del documento **10/09/**2022
9. Fecha de aplicación del programa informático de similitudes **10/09/**2022
10. Software utilizado
    - Turnitin
11. Configuración del programa detector de similitudes
    - Excluye textos entrecomillados
    - Excluye bibliografía
    - Excluye cadenas menores a 40 palabras
12. Porcentaje de similitudes según programa detector de similitudes **7 (siete)%**
13. Fuentes originales de las similitudes encontradas
    Se adjunta en el anexo 1
14. Observaciones
15. Calificación de originalidad
    - Documento cumple criterios de originalidad, sin observaciones [x]
    - Documento cumple criterios de originalidad, con observaciones [ ]
    - Documento no cumple criterios de originalidad [ ]
16. Fecha de informe **16/09/2022**

UNMSM
Firmado digitalmente por PRO CONCEPCION Luzmila Elisa FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 14.11.2022 13:24:36 -05:00

Firma de evaluador
Dra. Luzmila E. Pró Concepción
Directora (e) de la EPIS

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

ANEXO 1

Fuentes originales de las similitudes encontradas

1. hdl.handle.net: 5%
2. cybertesis.unmsm.edu.pe: 1%
3. www.empiricsystems.com: 1%

UNMSM
Firmado digitalmente por PRO CONCEPCION Luzmila Elisa FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 14.11.2022 13:24:12 -05:00

Firma de evaluador
Dra. Luzmila E. Pró Concepción
Director (e) de la EPIS

Dedicatoria

El presente trabajo es dedicado a mi familia y especialmente a mis padres que me apoyaron en toda mi formación académica

Agradecimiento

Agradezco a la Universidad Nacional Mayor de San Marcos que me brindo los conocimientos para enfrentar mi vida laborar

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Implementación de un Data lake para la centralización de datos analíticos y transaccionales en la empresa Belcorp

Autor: Martin Balboa, Michael Antonio

Asesor: Menéndez Mueras, Rosa

Título: Trabajo de Suficiencia Profesional para optar el Título Profesional de Ingeniero de Sistemas

Fecha: Agosto de 2022

RESUMEN

En el presente trabajo de suficiencia profesional describe el proyecto de implementación de un data lake en la empresa Belcorp, el principal objetivo de la implementación es la centralización de los datos para la creación de una vista holística de toda la data de la empresa y de sus clientes con el fin de tener los datos homologados y estandarizados, ya que hasta en ese momento la data de los diferentes sistemas que manejaba la empresa se encontraban aislados causando duplicidad y desactualización en los datos haciendo que se cree barreras para el intercambio de información y la colaboración entre las distintas áreas de la empresa. La implementación de esta plataforma significó un gran pilar para convertir a Belcorp en una empresa data driven, ya que la data procesada serviría como input para los diferentes equipos analíticos que crearían métricas y modelos de predicción, los cuales ayudarían a la mejor toma de decisiones de la empresa y así Belcorp se diferencie entre sus otros competidores.

Palabras claves: Data Lake, Scrum, Big Data, Data Driven, Spark

MAJOR NATIONAL UNIVERSITY OF SAN MARCOS

FACULTY OF SYSTEMS AND INFORMATICS ENGINEERING

PROFESSIONAL SCHOOL OF SYSTEMS ENGINEERING

Implementation of a Data Lake for the centralization of analytical and transactional data in Belcorp company.

Author: Martin Balboa, Michael Antonio

Advisor: Menéndez Mueras, Rosa

Título: Professional Sufficiency Work to opt for the Professional Title of Systems Engineer

Date: August 2022

ABSTRACT

The main objective of the implementation is the centralization of data for the creation of a holistic view of all the data of the company and its customers in order to have homologated and standardized data, since until then the data of the different systems that the company managed were isolated causing duplicity and outdated data causing barriers to the exchange of information and collaboration between different areas of the company. The implementation of this platform meant a great pillar to turn Belcorp into a data driven company, since the processed data would serve as input for the different analytical teams that would create metrics and predictive models, which would help the company to make better decisions and thus differentiate Belcorp among its other competitors.

Keywords: Data Lake, Scrum, Big Data, Data Driven, Spark

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

INTRODUCCIÓN

En el presente trabajo de suficiencia se aborda un problema que las empresas actuales están comenzando a afrontar al momento de usar los datos para la toma de decisiones, el cual es la falta de un repositorio unificado y centralizado de todos los datos generados por los diferentes sistemas de la empresa.

La empresa expuesta en el presente trabajo es líder de ventas de productos de belleza en latinoamérica, y en la actualidad busca diferenciarse de sus competidores para lo cual tiene pensado usar tecnologías innovadoras como el Big Data.

El autor del trabajo de suficiencia profesional explicará el proceso de implementación del data lake en la empresa Belcorp lo cual es un hito importante para la empresa ya que será el repositorio unificados de datos, el cual servirá como fuente de datos para la generación de procesos analíticos.

En el presente trabajo de suficiencia se organiza de manera siguiente:

En el capítulo I, se detalla la experiencia laboral desempeñada del autor durante toda su vida profesional.

En el capítulo II, se presentan las características principales de la empresa como su misión, visión y estructura, además de la experiencia del autor en la empresa.

En el capítulo III, se abordará el problema que enfrenta la empresa y la solución a esta, explicando a detalle la metodología e implementación de la solución planteada, además de la evaluación financiera del proyecto a implementar.

En el capítulo IV, se realiza una crítica a partir de la experiencia aprendida en la implementación del proyecto.

En el capítulo V, se exponen las conclusiones y sugerencias al proyecto aplicado, y la bibliografía para la realización del informe.

CAPÍTULO I: TRAYECTORIA PROFESIONAL

El autor del presente trabajo es un bachiller de la carrera de Ingeniería de Sistemas e Informática que cuenta con experiencia de 6 años en el campo de los datos.

Por lo cual cuenta con alta capacidad de análisis e integración de datos, que cuenta con experiencia de implementación de plataformas de datos en organizaciones desde el inicio, para los cual maneja herramientas de programación, sistemas operativos, base de datos y cloud computing, con experiencia trabajando para empresas consultoras, transnacionales y startups.

1. B89

   Periodo: Julio 2020 - Abril 2022

   Área: Technology

   Rol: Data Platform Technical Lead

   Proyecto: Implementación de la plataforma de datos de la empresa.

   Funciones:

   - Definir, configurar y desplegar la plataforma de datos de la empresa.
   - Integración de la base de datos TiDB del paradigma de datos NewSQL a nuestra plataforma de datos.
   - Configuración y despliegue del orquestador procesos Apache Airflow y su integración con los diferentes servicios de AWS, tales como: S3, Secret Manager y EMR.
   - Configuración y despliegue de Kafka Connect y Kafka Schema Registry para el tratamiento de datos real-time.
   - Definición, configuración y despliegue la herramienta BI, Metabase

- Configuración y despliegue del framework Apache Spark en Kubernetes para el procesamiento de datos batch.
- Despliega de la infraestructura mencionada usando código terraform.

2. Belcorp

   Periodo: Septiembre 2018 - Junio 2020

   Área: Data & Analytics

   Rol: Big Data Engineer

   Proyecto: Implementación del Data Lake de datos.

   Funciones:

   - A cargo de definir, construir y mantener el data lake de la empresa en un ambiente cloud.
   - Configuración, despliegue y administración de un ambiente hadoop con los servicios YARN, Hive, Presto, Hue y Spark.
   - Desarrollo de procesos batch de data lake usando el lenguaje de programación Scala junto al framework Spark, conectando con fuentes de datos como MongoDB y AWS Redshift.
   - Configuración de Elasticsearch, Logstash y Kibana para el almacenamiento de los del data lake.
   - Configuración de los servicios EC2, Glue, SQS, SNS, Lambda, KMS, ECR, ECS, MSK and S3.
   - Despliega de la infraestructura mencionada usando código cloudformation.

3. Everis

   Periodo: Junio 2017 - Agosto 2018

Área: Advanced Technology Solutions

Rol: Big Data Engineer

Proyecto: Enriquecimiento de tramas bancarias en real time.

Funciones:

- Desarrollo de procesos batch usando el lenguaje de programación Scala junto al framework Spark, conectando con fuentes de datos como Kafka y Cassandra.
- Diseño del modelo de datos Nosql en Apache Cassandra.
- Administración y despliegue de la base de datos Apache Cassandra.

Certificados obtenidos:

- CKAD: Certified Kubernetes Application Developer
- AWS Certified Big Data – Specialty
- AWS Certified Solutions Architect – Associate

CAPÍTULO II: CONTEXTO EN EL QUE SE DESARROLLÓ LA EXPERIENCIA

2.1 EMPRESA - ACTIVIDAD QUE REALIZA

Belcorp es una empresa multinacional peruana de venta directa de productos de belleza, para el cuidado de la piel, cuerpos y cosméticos, fue fundada en 1968 y cuenta con presencia en 14 países del continente americano. Belcorp cuenta con sus tres marcas Cyzone, Esika y L'bel, cada una especializada y dirigida a un grupo específico del mercado.

Belcorp se ha caracterizado usar el método de venta directa como principal canal de venta, en el cual existe un contacto directo que ocurre entre el consumidor y vendedor, haciendo así que no necesariamente exista un establecimiento comercial, en el caso de Belcorp la parte del vendedor es realizado por las consultoras, personas naturales que mediante el uso catálogos y aplicaciones venden los productos al cliente final.

Belcorp lanza sus nuevos productos y ofertas mediante campañas, estas campañas duran aproximadamente tres semanas, en el transcurso de las campañas las consultoras pueden realizar el pedido a la empresa y al finalizar la campaña las consultoras reciben sus pedidos para que así puedan los productos solicitados a sus clientes.

Actualmente la empresa se está encontrando en la dificultad de captar a nueva audiencia que usa las nuevas tecnologías, si bien el método de venta directa ha obtenido buenos resultados financieros para Belcorp a los largo de sus años, con el pasar del tiempo y el mayor uso de la tecnología, se está evaluando la posibilidad de usar otros métodos de venta para así diferenciarse entre sus competidores.

2.2 VISIÓN

En su página web, Belcorp (Belcorp 2021) manifiesta:

´´Ser la compañía que más contribuye a acercar a la mujer a su ideal de belleza y realización personal.´´

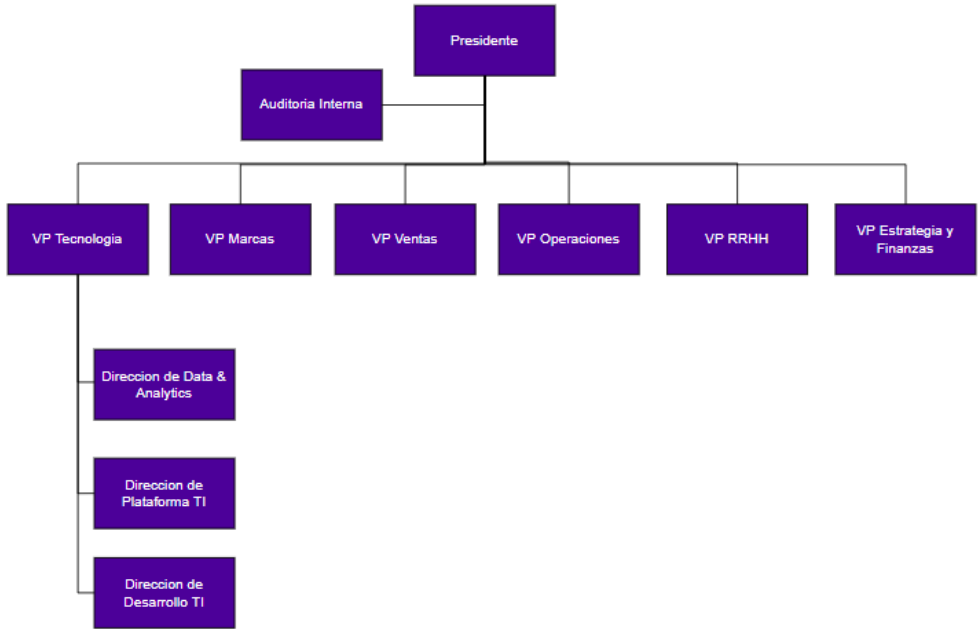
2.3 MISIÓN

En su página web, Belcorp (Belcorp 2021) manifiesta:

´´Crear oportunidades que inspiren a las personas y las empoderen para lograr un impacto positivo en la sociedad por medio del principal propósito 'impulsamos belleza para lograr la realización personal'´´

2.4 ORGANIZACIÓN DE LA EMPRESA

Figura 1: Organigrama Belcorp



2.5 ÁREA, CARGO Y FUNCIONES DESEMPEÑADAS

La empresa está implementado un proceso de transformación digital como principal estrategia organizacional para así conseguir una diferenciación con sus principales competidores, teniendo así como uno de sus principales pilares la toma de decisiones a partir de los datos obtenidos de sus clientes, por lo cual en los últimos años ha creado el área de Data & Analytics bajo la gerencia de Tecnología.

El área de Data & Analytics está compuesto por dos perfiles técnicos bien marcados, ingenieros de datos y científicos de datos, siendo la función principal del ingeniero de datos proveer datos de calidad al científico de datos para que este pueda crear modelos analíticos.

El autor de este informe de experiencia profesional trabajó como Senior Data Engineer en el área de Data & Analytics, el cargo Senior Data Engineer se encarga de liderar técnicamente al equipo de ingenieros de datos para el desarrollo, implementación, despliegue y buenas prácticas de procesos de tratamientos de datos con la finalidad de centralizar todos los datos de la empresa.

Las principales funciones realizadas como Senior Data Engineer fueron:

- Definir, desarrollar, desplegar y mantener el data lake de la empresa.
- Trabajar en conjunto con las áreas de tecnología como infraestructura, seguridad y devops al momento de implementar alguna nueva herramienta big data.
- Coordinar con las diferentes áreas de negocio para el entendimiento de las lógicas de negocio a aplicar en los distintos flujos de datos solicitados.
- Realizar pruebas de concepto para la evaluación de nuevas tecnologías que puedan ayudar a la mejora de la arquitectura de datos.
- Capacitar al equipo de ingeniero de datos y científico de datos en nuevas tecnologías big data que se implementan en la organización.

2.6 EXPERIENCIA PROFESIONAL REALIZADA EN LA ORGANIZACIÓN

El autor del presente trabajo fue parte del equipo que desarrolló e implementó el data lake de parte de la compañía, con la finalidad de centralizar todas las fuentes de datos transaccionales y analíticos en una misma repositorio de datos.

Además posteriormente a la implementación del data lake, capacitó a otros equipos para poder explotar los datos almacenados en el data lake.

Fue también parte del equipo que mejoró la primera versión del data lake, en esta segunda versión se realizó principalmente mejoras de performance e implementación de herramientas de monitoreo

Finalmente trabajó con el equipo de científicos de datos para desplegar sus modelos predictivos usando el framework de MLOPS.

CAPÍTULO III: ACTIVIDADES DESARROLLADAS
3.1 SITUACIÓN PROBLEMÁTICA

3.1.1 DEFINICIÓN DEL PROBLEMA

Los principales competidores de la empresa mencionada en el informe presentaban un gran crecimiento en el sector, por lo cual la empresa se planteó como estrategia la implementación de la transformación digital de la empresa, con la finalidad de fortalecer más los canales de venta digital, para así tener un ingreso adicional al de canal de venta directa.

Como parte de la transformación digital de la empresa, se buscaba que la organización implementara una cultura data driven para la toma de decisiones a partir de los datos.

La empresa gestiona dos tipos de datos, los datos transaccionales provenientes de las aplicaciones web, sistema ERP y sistemas de facturación, y los datos analíticos que en parte eran generados a partir de los datos transaccionales mediante procesos analíticos.

Los problemas a nivel de datos transaccionales es que en dicho momento la empresa contaba con diferentes fuentes de datos lo cual hacía que los datos transaccionales están dispersos en diferentes sistemas, además se tenía diferentes tipos datos tales como del tipo estructurados y semiestructurados, estos problemas ocasionan que los datos no estén homologados y estandarizados, causando que no se tenga una sola fuente fidedigna de datos.

Además a nivel de datos analíticos, estos se procesan usando una herramienta warehouse para la generación de datamarts, dicho warehouse necesitaba constantemente incrementos de recursos a nivel de infraestructura tales como almacenamiento y

cómputo, debido a que los volúmenes de datos incrementaron constantemente, esto conllevaba a un aumento en los costos de operación.

3.2 SOLUCIÓN

Desarrollar un data lake que permita almacenar, normalizar y enriquecer los datos transaccionales proveniente de diferentes fuentes de datos y de diferentes estructuras de datos.

Además usar dicho data lake para la ejecución de procesos analíticos, los cuales ya tendrían una única fuente fidedigna de datos, con data de calidad y abundante para el análisis.

Finalmente para ahorrar costos se optaría por un cloud provider el cual facilita el despliegue de la infraestructura del data lake y además brindaría flexibilidad al momento de escalar la infraestructura cuando la necesidad de almacenamiento y la capacidad de cómputo incremente.

3.2.1 OBJETIVOS

OBJETIVO GENERAL

Implementar un Data lake para la centralización de datos analíticos y transaccionales en la empresa Belcorp.

OBJETIVO ESPECIFICOS

- Reducir los costos generados por el procesamiento y almacenamiento de datos.
- Enseñar a los diferentes tipos de usuarios el uso de los datos que se almacenaran en el data lake.

– Definir la arquitectura de los procesos de ingesta y procesamientos del data lake para las futuras ingestas de nuevas fuentes de datos.

3.2.2 ALCANCE

El alcance del proyecto consiste en el desarrollo del proceso de extracción de datos de diferentes fuentes y su carga al data lake, la limpieza de datos (estandarización y tipificaciones) y a su vez aplicación de lógica del negocio. Además de la implementación de la infraestructura del data lake en proveedor cloud AWS. Finalmente el despliegue de dicha infraestructura y los procesos del data lake.

3.2.3 ETAPAS Y METODOLOGÍA

3.2.4 FUNDAMENTOS UTILIZADOS

3.2.4.1 Big data

Según Gartner "big data se refiere activos de información de gran volumen, alta velocidad y/o gran variedad que exigen formas rentables e innovadoras de procesamiento de la información que permiten una mejor comprensión, toma de decisiones y automatización de procesos."

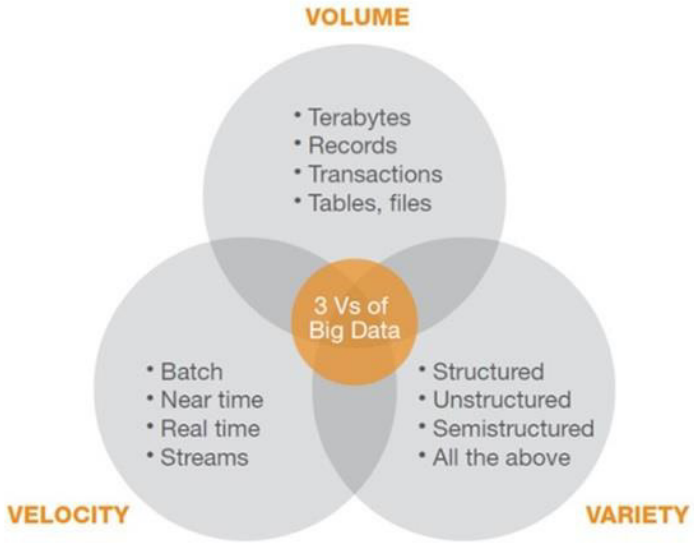
Según ese concepto se puede extraer que big data tiene tres componentes, llamados las 3 V´s del big data:

– Velocidad: Hace referencia a la rapidez de carga y procesamiento que se debe ejecutar sobre los datos mediantes los procesos de datos.
– Variedad: Se refiere a los diferentes tipos de datos que se disponen, como datos estructurados tomando como ejemplo los datos almacenados en las

bases de datos, y los datos no estructurados tales como textos, audio o video.

– Volumen: Hace mención a la gran cantidad de datos que se generan y recopilan constantemente.

Figura 2: Las V's de Big data



3.2.4.2 Data lake

Según Pasupuleti, Pradeep, & Beulah Salome Purra (2014) definen a un Data lake como un gran repositorio que almacena todo tipo de data en un formato crudo hasta que sea necesitado por cualquiera en una organización para analizar la data, estos datos pueden ser del tipo datos estructurados y no estructurados, además de data que se ingesta al data lake puede ser cargada de una forma batch (en grandes lotes de datos) y streaming (pequeños lotes de datos constantes).

Los autores (Pasupuleti, Pradeep, & Beulah Salome Purra) comentan que un data lake no es solo hadoop, para la implementación de un data lake se puede utilizar diferentes herramientas tanto de código abierto como herramientas licenciadas.

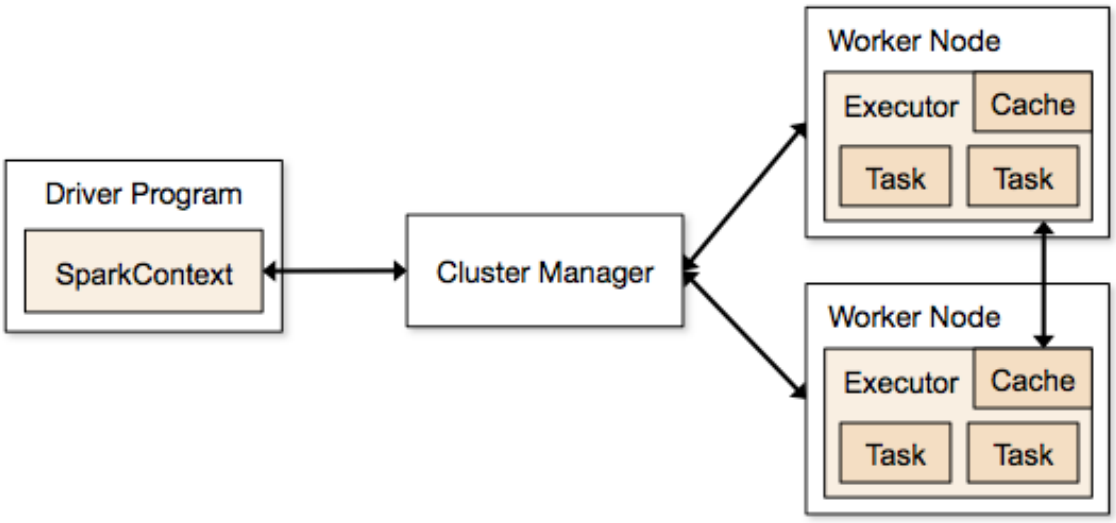
Además mencionan que sobre el data lake se puede realizar enriquecimiento de datos que ayuda al mejoramiento, clasificación, aumento y estandarización de la data de la organización.

3.2.4.3 Apache Spark

Según lo mencionado por Thottuvaikkatumana, Rajanarayanan (2016) "Apache Spark es un sistema de analítica de datos en memoria altamente escalable y distribuido, el cual prevé la habilidad de desarrollar aplicaciones en lenguajes como Java, Scala, Python y R. Teniendo así uno de los mayores grupo de contribuciones entre los proyectos de Apache."

Complementando lo mencionado, Frampton, Mik (2015) mencionan que "Spark se guía de la arquitectura maestro-esclavo, lo cual permite escalar a demanda, teniendo así dos componentes como el Programa Driver que es en donde el código de spark se ejecuta y él es responsable de programación de operaciones paralelas en un clúster de computo. Y el Ejecutor el cual es quien prevé recursos para ejecutar las tareas lanzadas por el programa driver."

Figura 3: Spark arquitectura driver y ejecutor



3.2.4.4 Scala

Se define como un lenguaje de programación que utiliza la programación orientada a objetos y la programación funcional.

El código desarrollado en Scala al ser compilado puede ser ejecutado en la Máquina Virtual de Java, lo cual lo hace muy compatible con Java, tanto así que se puede integrar librerías de Java como dependencias.

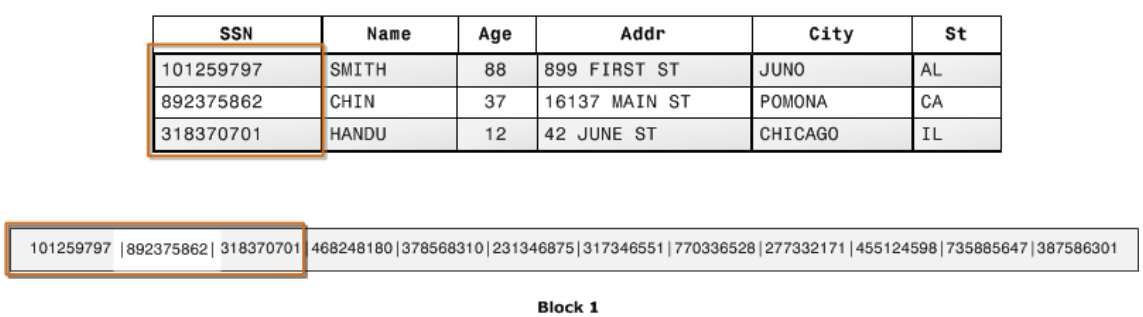
En los últimos años se ha vuelto un lenguaje popular ya es el lenguaje usado por Apache Spark que viene a ser el framework más usado en la programación distribuida para el procesamiento de datos.

3.2.4.5 Parquet

Primeramente se tiene que mencionar el concepto de almacenamiento basado en columnas el cual organiza y almacena tablas basado en columnas, en donde la data de cada columna es almacenada junta, este tipo de almacenamiento tiene una gran ventaja a nivel de desempeño, como por ejemplo si se realiza un consulta que solo necesita dos columnas, solo las columnas requeridas serán accedidas, mientras que en un enfoque tradicional se tendrían que acceder a todas las columnas por cada fila que sea requerida.

Teniendo en cuenta lo mencionado, Turkington, Garry, and Gabriele Modena (2015) mencionan que "Paquet permite almacenar complejas y anidadas estructuras de datos, ya que internamente tiene una eficiente encodificamiento a nivel de columnas."

Figura 4: Almacenamiento basado en columnas



3.2.4.6 Particiones

Para Karanth, Sandeep (2014) las particiones son “subdivisiones de tablas basados en distintos valores de columnas, cuantos las columnas particiones son especificadas, todos los registros correspondientes a valores de las columnas distintos o combinados son almacenados en un subdirectorio del directorio de la tabla.”

Ademas Karanth, Sandeep menciona que “las particiones son usadas como pre-filtros depurar los innecesarios registros de ser procesados, ayudando a disminuir la latencia de las consultas e I/O.”

3.2.4.7 Scrum:

Según (Rising & Janoff, 2000) “Scrum corresponde a un marco de trabajo usual en proyectos de software independientemente su complejidad siendo adaptable, iterativo, eficiente, permitiendo el desarrollo colaborativo generando valor al negocio. Las entregas son incrementales y son priorizadas según lo requerido por la organización.”

Indicado eso, las historias de usuario se desarrollan en epicas, y estas se dividen en historias, las cuales son desarrolladas en lapsos de tiempo llamados sprints.

Para tener mas claro según (Menzinsky, López, Palacio, Sobrino, & Rivas, 2020), una epica es una historia de usuario que se distingue por su gran tamaño. Es como una etiqueta que asignamos a una historia cuyo esfuerzo impide completarla de una sola vez o en un solo sprint.

3.2.4.8 Cloud computing

Según (Microsoft, 2021), “Es proveer servicios informáticos a través de Internet, tales como servidores, servicios de almacenamiento, bases de datos, redes, software,

Big Data, IoT (Internet de las cosas) e inteligencia artificial. La computación en la nube ofrece una innovación más rápida, recursos flexibles y economías de escala."

3.2.4.9 DevOps

Según (Lwakatare, y otros, 2019) indican: "DevOps, un acrónimo de desarrollo y operaciones, es un enfoque en el que los desarrolladores de software y las operaciones trabajan en estrecha colaboración. El objetivo es mejorar la comunicación y la integración del desarrollo y las operaciones con 17 el fin de obtener todos los beneficios de los enfoques de desarrollo de software moderno que emplean lanzamientos rápidos de nuevas funciones de software para los usuarios finales y, posteriormente, aprender de ellos."

3.2.5 IMPLEMENTACIÓN DE LAS ÁREAS, PROCESOS, SISTEMAS Y BUENAS PRÁCTICAS

El desarrollo del presente proyecto se realizó bajo el enfoque de la metodología SCRUM, donde el autor de este informe participó como Senior Data Engineer, siguiendo el marco metodológico de SCRUM la formación del equipo encargado de cumplir el proyecto fue:

- Un Product Owner: Encargado de llevar el liderazgo y seguimiento del equipo a nivel del producto.
- Un Technical Leader (Senior Data Engineer): Encargado de difinir la arquitectura, las teconologias, las integraciones y procesos del datalake.
- Tres Team Members: Encargado de la implementacion de las logicas de negocio en los procesos del data lake.

Para la realización de proyecto se realizaron 12 sprints de dos semanas cada uno, las épicas que se abarcaron durante todo el proyecto fueron: Levantamiento de la información, definición de tablas funcionales, definición de la arquitectura, definicion de los procesos y despliegue de los procesos

3.2.5.1 Levantamiento de la información

En la primera etapa del proyecto de implementación del data lake, se encontró que la empresa contaba con una gran cantidad de fuentes de datos y a su vez cada fuente de datos manejaba gran cantidad de archivos, por lo cual se procedió a definir las fuentes de datos más relevantes para los procesos analíticos con la finalidad de agregar valor al negocio al finalizar el proyecto.

Para la definición de las fuentes de datos más importantes se realizó reuniones con los equipos de negocios en donde se llegó a la conclusión y posterior acuerdo que la información más relevante la cual debería ser cargada en el primer estadio del data lake seria:

- Información 360 del cliente final.
- Información del producto.
- Información de las ventas de los diferentes medios.

Finalmente para obtener dicha información se escogieron cinco diferentes fuentes de datos (SICC, Planit, SAP, Digital, BI), que a partir de ahora las llamaremos interfaces, además cada interfaz contiene diferentes archivos que además se encuentran en diferentes formatos.

- Sistema Comercial (SICC) - Sistema comercial que entrega la información relacionada con la facturación de la venta de las consultoras, información de la

fuerza de venta como es su evolución, aplicaciones, las reacciones, la geografía Belcorp, y el estado de las consultoras de Belcorp.

Tabla 1: Archivos de la interfaz SICC

Interfaz	Archivos	Formato	Pk
Sicc	Debelistadatosadic	Csv	Codebelista
	Fstaebeadic	Csv	Codcanalventa
	Debelista	Csv	Codebelista
	Dgeografiacampana	Csv	Aniocampana
	Dmatrizcampana	Csv	Codproducto
	Dletsrangocomision	Csv	Codprograma
	Dnrodocumento	Csv	Nrodocumento
	Fstaebecam	Csv	Codebelista
	Fvtaproebecam	Csv	Codebelista
	Dstatusf	Csv	Aniocampana
	Tstaebecam	Csv	Codebelista
	Dcontrolcierre	Csv	Aniocampana
	Dgeografia	Csv	Codterritorio

- Planit.- Sistema de planeamiento en donde se registra la táctica de los productos que serán expuestos en las campañas en curso y en campañas futuras.

Tabla 2: Archivos de la interfaz Planit

Interfaz	Archivos	Formato	Pk
Planit	Dtipooferta	Csv	Codtipooferta
	Dmatrizcampana	Csv	Codproducto, codventa
	Dcostoproductocampana	Csv	Codproducto
	Fnumpedcam	Csv	Codcanalventa
	Dcontrolcierre	Csv	Codcontrol

- SAP.- Es el sistema logístico en donde se registra el maestro del producto así como el detalle del estado del ciclo de vida del producto.

Tabla 3: Archivos de la interfaz SAP

Interfaz	Archivos	Formato	Pk
Sap	Dproducto	Csv	Codsap

- Digital.- Sistema web donde se almacenan los registros de los pedidos realizados por las consultoras, el registro de la información de la bolsa de compra, logs de accesos a sistema web y desde que parte de la página realizó el pedido de un producto.

Tabla 4: Archivos de la interfaz Digital

Interfaz	Archivos	Formato	Pk
Digital	Dorigenpedidoweb	Csv	Codorigenpedidoweb
	Flogingresoportal	Csv	Codebelista
	Fpedidowebdetalle	Csv	Pedidoid
	Fofertafinalconsultora	Csv	Codebelista

- BI.- El sistema de explotación de información, desde el entorno del modelo de Datamart (MS SQL), hacia un modelo de BI Corporativo.

Tabla 5: Archivos de la interfaz BI

Interfaz	Archivos	Formato	Pk
Bi	Dpais	Csv	Codpais
	Debelista	Csv	Codebelista
	Bdlideres_base_paises	Csv	Codseccion
	Dcatalogovehiculo	Csv	Codcatalogo
	Dstatus	Csv	Codstatus
	Fstaebecam	Csv	Codebelista

3.2.5.2 Definición de tablas funcionales

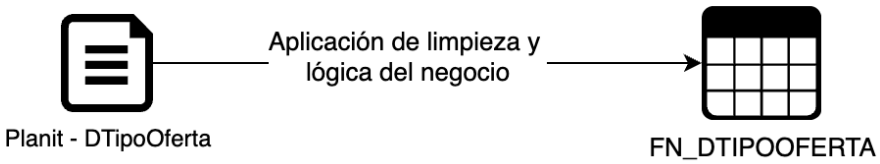
Luego de la selección de las interfaces y sus respectivos archivos a tener en cuenta, se procede a la aplicación de reglas de limpieza de datos, nomenclatura de campos y definición de tipo de datos.

Posteriormente a la estandarización de las interfaces y sus respectivos archivos, se procedió a definir las tablas funcionales, las cuales serán creadas a partir de un o más archivos de la misma o diferente interfaz de datos, teniendo así tablas simples y compuestas.

Además, sobre los archivos de las interfaces que sirven de insumo, se aplica lógica del negocio, las cuales le dan sentido a los datos en crudo, esto ayudará que el negocio pueda entender el contenido de las tablas funcionales.

– Tablas funcionales simples.- Son las tablas resultados de la aplicación de reglas de limpieza y lógica del negocio a partir de un solo archivo de una interfaz.

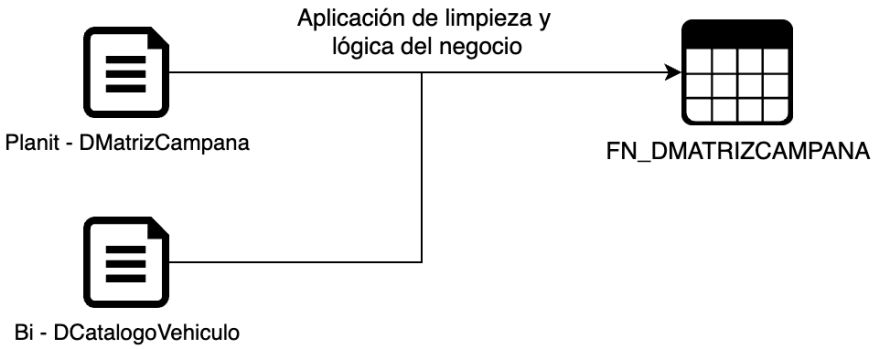
Figura 5: Tablas funcionales simples



– Tablas funcionales compuestas.- Son las tablas resultados de la aplicación de reglas de limpieza y lógica del negocio a partir de dos archivos de una misma o diferente interfaz. Además estas tablas cuentan con una lógica adicional que permite la independencia entre los archivos insumos, es decir que si un día el

archivo A no arriba al data lake, pero si arriba el archivo B, entonces la tabla funcional que usa como insumo el archivo A y B no se debería ver afectada.

Figura 6: Tablas funcionales compuestas



Se utilizó el prefijo FN_* a todas las tablas funcionales para diferenciarlas de las tablas insumo, el resultado de la definición de las tablas funcionales se pueden observar en la tabla 6.

Tabla 6. Tablas funcionales

Tablas funcionales	**Archivo insumo**	**Interface**
FN_DPRODUCTO_CORP	Dproducto	Sap
	Dpais	Bi
FN_DEBELISTA	Debelistadatosadic	Sicc
	Debelista	Sicc
	Debelista	Bi
	Dpais	Bi
FN_FSTAEBECAM	Fstaebeadic	Sicc
	Fstaebecam	Sicc
	Tstaebecam	Sicc
	Dcontrolcierre	Sicc
	Dstatus	Bi
	Fstaebecam	Bi
	Dpais	Bi
FN_DGEOGRAFIACAMPANA	Dgeografiacampana	Sicc
	Dmatrizcampana	Sicc
	Dgeografia	Sicc
	Bdlideres_base_paises	Bi

	Dpais	Bi
FN_DLETSRANGOCOMISION	Dletsrangocomision	Sicc
	Dpais	Bi
FN_FVTAPROEBECAM	Fvtaproebecam	Sicc
FN_DSTATUSFACTURACION	Dstatusf	Sicc
	Dpais	Bi
FN_DMATRIZCAMPANA	Dmatrizcampana	Planit
	Dcatalogovehiculo	Bi
	Dpais	Bi
FN_FVTAPROEBECAM	Dcostoproductocampana	Planit
	Fnumpedcam	Planit
	Dcontrolcierre	Planit
	Dpais	Bi
FN_DORIGENPEDIDOWEB	Dorigenpedidoweb	Digital
	Dpais	Bi
FN_FLOGINGRESOPORTAL	Flogingresoportal	Digital
	Dpais	Bi
FN_FPEDIDOWEBDETALLE	Fpedidowebdetalle	Digital
	Dpais	Bi
FN_FOFERTAFINALCONSULTORA	Fofertafinalconsultora	Digital
	Dpais	Bi

Finalmente se definió las lógicas de negocio que se aplicaron sobre los archivos insumo, es importante mencionar que cada tabla funcional tiene su propia regla del negocio ya que cada tabla funcional tiene diferentes insumos y además cada uno tiene diferente funcionalidad.

Tabla 7: FN_DPRODUCTO_CORP

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	CODSAP	nvarchar	9	
2	CODPRODUCTO	nvarchar	9	
3	DESPRODUCTO	nvarchar	100	
4	CUC	nvarchar	50	
5	DESCRIPCUC	nvarchar	50	
6	CODUNIDADNEGOCIO	nvarchar	4	
7	DESUNIDADNEGOCIO	nvarchar	50	
8	CODMARCA	nvarchar	2	
9	DESMARCA	nvarchar	50	

10	CODCATEGORIA	nvarchar	2
11	DESCATEGORIA	nvarchar	50
12	CODCLASE	nvarchar	5
13	DESCLASE	nvarchar	50
14	DESNEGOCIO	nvarchar	50
15	CODIGONEGOCIO	nvarchar	4
16	CODSUBCATEGORIA	nvarchar	3
17	CODTIPO	nvarchar	3
18	DESTIPO	nvarchar	100
19	DESSUBTIPO	nvarchar	150
20	CODSUBTIPO	nvarchar	15
21	DESSUBCATEGORIA	nvarchar	100
22	DESTIPOSOLO	nvarchar	70
23	DESSUBTIPOSOLO	nvarchar	70
24	DESSUBCATEGORIASOLO	nvarchar	70
25	CODLINEA	nvarchar	3
26	DESLINEA	nvarchar	50
27	DESPRODUCTOSUPERGENERICO (DESPRODUCTOGENERICOI)	nvarchar	50
28	CODPRODUCTOSUPERGENERICO (CODPRODUCTOGENERICOI)	nvarchar	15
29	DESPRODUCTOGENERICO (DESPRODUCTOGENERICOII)	nvarchar	50
30	CODPRODUCTOGENERICO (CODPRODUCTOGENERICOII)	nvarchar	15
31	DESCRIPPEQ (DESPRODUCTOPEQ)	nvarchar	50
32	PEQ (CODPEQ)	nvarchar	15
33	CODGRUPOARTICULO	nvarchar	10
34	DESGRUPOARTICULO (DESGRUPOPRODUCTO)	nvarchar	30
35	MERCADOACCCOSMETICO	nvarchar	50
36	NOMBREGENBIJ	nvarchar	50
37	ADICIONAL	nvarchar	20
38	APLICACIONLUGARUSO	nvarchar	50
39	DETALLEBRAZO	nvarchar	20
40	GROSORBRAZO	nvarchar	20
41	NEGOCIACIONUSO	nvarchar	50
42	BENEFICIO	nvarchar	50
43	DETALLESCAJA	nvarchar	50

44	FORMACAJA	nvarchar	50
45	DETALLEBOT	nvarchar	50
46	TIPOESTUCHE	nvarchar	50
47	DETALLEPRODUCTO	nvarchar	70
48	ROPADETALLE	nvarchar	70
49	COLORTONO	nvarchar	50
50	COLORLUNA	nvarchar	50
51	TEORIACOLOR	nvarchar	50
52	INSUMOSCOMPLENT	nvarchar	50
53	DISENO	nvarchar	50
54	DETALLETOP	nvarchar	50
55	ORIGENACCESORIOS	nvarchar	50
56	COLORMARCO	nvarchar	50
57	REPORTAJEMAS	nvarchar	50
58	CARACTERISTICASBIJ	nvarchar	50
59	ACABADOACCESORIOS	nvarchar	50
60	ACABADOLUNA	nvarchar	50
61	MATERIALCAJA	nvarchar	50
62	FAMILIAFRAGANCIA	nvarchar	60
63	DATOSLENTE	nvarchar	50
64	TIPOFABRICACION	nvarchar	50
65	DESMERCADO	nvarchar	50
66	TIPOMATERIALCALZADO	nvarchar	50
67	DETALLESUELA	nvarchar	50
68	PRESENTACIONENVASE	nvarchar	50
69	TIPOPLANEACION	nvarchar	50
70	DESPOSICIONAMIENTO	nvarchar	50
71	PRESENTACION	nvarchar	50
72	PRESENTACIONFORMA	nvarchar	50
73	TIPOESTAMPADO	nvarchar	50
74	TEMPORADA	nvarchar	50
75	TAMANOMANGA	nvarchar	50
76	DETALLEESPECIFI (DESPRODUCTOCOMESPECIFICO)	nvarchar	50
77	COLORCORREA	nvarchar	50
78	TIRASASA	nvarchar	50
79	DESESTILO	nvarchar	50
80	EDADOBJETIVO	nvarchar	50
81	TEMA	nvarchar	50
82	TIPOLENTE	nvarchar	50
83	TIPOPIELCABELLO	nvarchar	60
84	CONTNETO	nvarchar	50

85	UM_CONTENIDO (CAPACIDAD)	nvarchar	50	
86	UNIDADDEMEDIDABASE (DESUNIDADMEDIDA)	nvarchar	50	
87	VERSATILIDAD	nvarchar	50	
88	TALLATAMANO	nvarchar	50	
89	CODTIPOMATERIAL	nvarchar	50	
90	DESTIPOMATERIAL	nvarchar	50	
91	TIPOPRODUCTO	nvarchar	50	
92	GAMA	nvarchar	50	
93	ROTULOSFILA	nvarchar	50	
94	INSUMOSMATPRIM	nvarchar	50	
95	TEMATICO	nvarchar	50	
96	DIAL	nvarchar	50	
97	DIALACABADO	nvarchar	50	
98	TIPOLOGIA	nvarchar	50	
99	SEXO	nvarchar	50	
100	DESCARACTERISTICABULK	nvarchar	50	
101	NOMBRERELOJ	nvarchar	50	
102	SUBTIPOPRODCOMPLEMENTOS	nvarchar	50	
103	CODJERQ02	nvarchar	30	
104	DESCJERQ02	nvarchar	90	

Tabla 8: FN_DEBELISTA

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANAINGRESO	nvarchar	12	
2	ANIOCAMPANAPRIMERPEDIDOWEB	nvarchar	12	
3	ANIOCAMPANAULTIMOPEDIDO	nvarchar	12	
4	CODEBELISTA	nvarchar	30	
5	DESESTADOCIVIL	nvarchar	30	
6	DESNSE	nvarchar	30	
7	FECHANACIMIENTO	datetime	8	
8	FLAGGERENTEZONA	tinyint	1	
9	ANIOCAMPANAPRIMERPEDIDO	nvarchar	12	Particular 01
10	CODPAIS	nvarchar	4	Común 01
11	DESAPENOM	nvarchar	90	
12	DESLIDER	nvarchar	110	Particular 02

13	TELEFONOMOVIL	nvarchar	50	

Regla Particular 1: Campaña primer pedido

Se actualiza de la interfaz del "DATAMART DEbelista (5.2)", utilizando el "código de la consultora"

Regla Particular 2: Nombre de la líder

Se actualiza tomando el valor de "DESLIDERX"

Regla Particular 3: Teléfono Móvil

Se actualiza de la interfaz del Interfaces "SICC DebelistaDatosAdic.txt (2.1)", utilizando el "código de la consultora"

Tabla 9: FN_DGEOGRAFIACAMPANA

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANA	nvarchar	12	
2	CODGERENTEREGIONAL	nvarchar	20	
3	CODGERENTEZONA	nvarchar	20	
4	CODLIDER	nvarchar	20	
5	CODREGION	nvarchar	6	
6	CODSECCION	nvarchar	14	
7	CODTERRITORIO	nvarchar	20	
8	CODZONA	nvarchar	12	
9	DESREGION	nvarchar	60	
10	DESZONA	nvarchar	60	
11	CODPAIS	nvarchar	4	Común 01
12	DESPAIS	nvarchar	30	Común 02
13	DESDEPARTAMENTO	nvarchar	30	
14	DESCIUDAD	nvarchar	60	
15	DESDISTRITO	nvarchar	50	Particular 01
16	DESNIVELSOCIA	nvarchar	20	
17	DESRENDIMIENTOSOCIA	nvarchar	24	Particular 02

Regla Particular 01: Atributos geopolíticos

Se actualiza de la interfaz del "DATAMART DGeografia.txt (2.13)", utilizando el "código del territorio"

Regla Particular 02: Rendimiento de la SE

Se actualiza "RendimientoEtapa" de la interfaz del "DATAMART BDLideres_Base_Paises (5.4)", utilizando "país + campaña + sección"

Se actualiza "DesNivel" de la interfaz del "DATAMART BDLideres_Base_Paises (5.4)", utilizando "país + campaña + sección"

Tabla 10: FN_DMATRIZCAMPANA

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANA	nvarchar	12	
2	CODCANALVENTA	nvarchar	4	
3	CODCATALOGO	nvarchar	4	
4	CODESTRATEGIA	nvarchar	6	
5	CODTIPOOFERTA	nvarchar	8	
6	CODVENTA	nvarchar	10	
7	DESCATALOGO	nvarchar	100	
8	DESTIPOOFERTA	nvarchar	100	
9	NROPAGINA	Int	4	
10	NUMOFERTA	Int	4	
11	PRECIONORMALMN	decimal	17	
12	PRECIOOFERTA	decimal	17	
13	PRECIOVTAPROPUESTOMN	decimal	17	
14	CODTIPOCATALOGO	nvarchar	4	
15	DESARGVENTA	nvarchar	160	
16	DESEXPOSICION	nvarchar	160	
17	DESLADOPAG	nvarchar	160	
18	DESTIPOCATALOGO	nvarchar	100	
19	DESUBICACIONCATALOGO	nvarchar	160	
20	FOTOMODELO	nvarchar	2	
21	FOTOPRODUCTO	nvarchar	2	
22	NROPAGINAS	Int	4	
23	PAGINACATALOGO	Int	4	Particular 01
24	DESOBSERVACIONES	nvarchar	160	Particular 02
25	VEHICULOVENTA	nvarchar	40	Común 01
26	CODPAIS	nvarchar	4	Homologa 06
27	CODSAP	Char	18	

Regla Particular 01: Atributos de Planit

Se actualiza estos valores de la interfaz de "PLANIT DMatrizCampana.txt (3.2)", utilizando "campaña + tipo de oferta y el producto".

Para hallar el código de venta en la información que llega de la interfaz de PLANIT, se toma el primero sorteado de forma ascendente, utilizando "campaña + tipo de oferta + producto"

Regla Particular 02: Vehículo de venta

Se actualiza de la interfaz del "DATAMART DCatalogoVehiculo (4.4)", tomando como base el "código del catálogo".

Tabla 11: FN_DLETSRANGOSCOMISION

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	CODPAIS	Nvarchar	4	Común 01
2	CODPROGRAMA	Nvarchar	12	
3	CODNIVEL	Nvarchar	4	
4	TIPOVALOR	Nvarchar	8	
5	CODRANGO	Tinyint	1	
6	PORCOMISION	Decimal	17	
7	MONTOINI	Decimal	17	
8	MONTOFIN	Decimal	17	

Tabla 12: FN _DNROFACTURA

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANA	nvarchar	12	
2	CODACCESO	tinyint	1	
3	CODCANALVENTA	nvarchar	4	
4	CODEBELISTA	nvarchar	30	
5	CODTERRITORIO	nvarchar	20	
6	CODTIPODOCUMENTO	nvarchar	2	
7	FLAGORDENANULADO	tinyint	1	
8	FLAGPROL	tinyint	1	
9	REALVTAMNFACTURA	decimal	17	
10	SALDOBANCO	decimal	17	
11	CANALINGRESO	nvarchar	6	
12	FECHAEMISIONFACTURA	smalldatetime	4	Particular 01
13	NROFACTURA	nvarchar	22	Particular 02
14	CODPAIS	nvarchar	4	Común 01
Regla Particular 1: Fecha de emisión de la factura				
De la misma interfaz tomar el valor de “FechaEmisionFactura”				
Regla Particular 1: Número de la factura				
De la misma interfaz tomar el valor de “NroDocumento”				

Tabla 13: FN _FSTAEBECAM

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANA	nvarchar	12	
2	CODCANALVENTA	nvarchar	4	
3	CODEBELISTA	nvarchar	30	
4	CODSTATUS	tinyint	1	Particular 01

5	CODTERRITORIO	nvarchar	20	
6	FLAGPASOPEDIDO	tinyint	1	
7	REALNROORDENES	tinyint	1	
8	FLAGACTIVA	tinyint	1	
9	FLAGPASOPEDIDOCUIDADOPERSONAL	tinyint	1	
10	FLAGPASOPEDIDOMAQUILLAJE	tinyint	1	
11	FLAGPASOPEDIDOTRATAMIENTOCORPORAL	tinyint	1	
12	FLAGPASOPEDIDOTRATAMIENTOFACIAL	tinyint	1	
13	FLAGPEDIDOANULADO	tinyint	1	
14	FLAGPASOPEDIDOFRAGANCIA	tinyint	1	Particular 02
15	CODPAIS	nvarchar	4	Común 01
16	CODIGOFACTURACIONINTERNET	nvarchar	10	Particular 03
17	CODCANALORIGEN	nvarchar	2	Particular 04
18	FLAGMULTIMARCA	tinyint	1	Particular 02
19	CONSTANCIA	nvarchar	10	
20	FRECUENCIACOMPRA	int	4	
21	CODCOMPORTAMIENTOROLLING	tinyint	1	
22	DESCRIPCIONROLLING	nvarchar	40	Particular 05
23	FLAGPASOPEDIDOWEB	tinyint	1	Particular 02
24	NROLOGUEOS	int	4	Particular 06
25	FLAGIPUNICOZONA	tinyint	1	Particular 07
26	FLAGEXPUESTAODD	tinyint	1	
27	FLAGEXPUESTAOF	tinyint	1	
28	FLAGEXPUESTAFDC	tinyint	1	
29	FLAGEXPUESTASR	tinyint	1	
30	FLAGCOMPRAOPT	tinyint	1	
31	FLAGCOMPRAODD	tinyint	1	
32	FLAGCOMPRAOF	tinyint	1	
33	FLAGCOMPRAFDC	tinyint	1	
34	FLAGCOMPRASR	tinyint	1	
35	FLAGEXPUESTAOPT	tinyint	1	Particular 08

Regla Particular 1: Código de Status

Se filtran solo los registros cuyo “Status Corporativos” sean diferentes a “Retiradas” o “Registradas”.

Para hallar el Status Corporativos se obtiene relacionando el código del Status con la interfaz de “DATAMART DStatus (5.5)”.

Regla Particular 2: Indicadores de venta

(ver documento de mapeo de reglas de negocio)
Regla Particular 3: Código de internet De la misma interfaz tomar el valor de “CodigoFacturaInternet ”
Regla Particular 4: Canal de Origen Se actualiza tomando el valor de la interfaz del “SICC FStaEbeAdic.txt (2.2)”, utilizando “campaña + consultora”.
Regla Particular 5: Campos calculados en base a la información histórica Se actualiza tomando los valores de la interfaz de “DATAMART FStaEbeCam (5.6)”, utilizando “campaña + consultora”.
Regla Particular 6: Número de Logueos (ver documento de mapeo de reglas de negocio)
Regla Particular 7: IP único (ver documento de mapeo de reglas de negocio)
Regla Particular 8: Indicadores de exposición (ver documento de mapeo de reglas de negocio)
Regla Particular 9: Cierre de campaña Considerar que la información corresponde al cierre de campaña, cuando en la interfaz “SICC DControlCierre.txt 2.12”, el campo “AnioCampana” tiene la campaña que está cerrado, cuando se da el caso, el “CodStatus” toma el valor de la interfaz “SICC TStaEbeCam.txt 2.11”, utilizando la “campaña + consultora”.

Tabla 14: FN_FVTAPROEBECAM

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANA	nvarchar	12	
2	CODCANALVENTA	nvarchar	4	
3	CODEBELISTA	nvarchar	30	
4	CODTERRITORIO	nvarchar	20	
5	CODTIPODOCUMENTO	nvarchar	2	
6	CODTIPOOFERTA	nvarchar	8	
7	CODVENTA	nvarchar	10	
8	DESCUENTO	decimal	17	
9	OPORTUNIDADAHORROMN	decimal	17	
10	NROFACTURA	nvarchar	22	
11	REALANULMNNETO	decimal	17	
12	REALDEVMNNETO	decimal	17	
13	REALUUANULADAS	int	4	
14	REALUUDEVUELTAS	int	4	
15	REALUUFALTANTES	int	4	
16	REALUUVENDIDAS	int	4	
17	REALVTAMNFACTURA	decimal	17	
18	REALVTAMNFALTNETO	decimal	17	
19	REALVTAMNNETO	int	4	

20	REALANULMNCATALOGO	decimal	17	
21	REALDEVMCATALOGO	decimal	17	
22	REALVTAMNCATALOGO	decimal	17	
23	REALVTAMNFALTCATALOGO	decimal	17	Particular 01
24	COSTOREPOSICIONMN	decimal	17	Particular 02
25	REALTCPROMEDIO	decimal	17	Particular
26	ESTTCPROMEDIO	decimal	17	03
27	CODORIGENPEDIDOWEB	int	4	Particular 04
28	CODPAIS	nvarchar	4	Común 01
29	CODSAP	char	18	
30	CODPALANCAPERSONALIZACION	nvarchar	6	
31	DESPALANCAPERSONALIZACION	nvarchar	200	Particular 05
32	ANIOCAMPANAREF	nvarchar	12	Particular 06

Regla Particular 1: Venta Catálogo
(ver documento de mapeo de reglas de negocio)

Regla Particular 2: Costo de Reposición
Se obtiene relacionando la interfaz de "PLANIT DCostoProductoCampana.txt (3.3)" por campaña y producto.

Regla Particular 3: Tipo de cambio
Se obtiene relacionando la interfaz de "PLANIT FNumPedCam.txt (3.4)" por campaña.

Regla Particular 4: Indicadores Web
El valor de "CodOrigenPedidoWeb" se obtiene del campo "CanalIngreso"

Regla Particular 5: Indicadores Web
El valor de "CodPalancaPersonalizacion" se obtiene del campo "CodigoPalanca"
El valor de "DesPalancaPersonalizacion" se obtiene del campo "DesOrigenPedidoWeb" de la interfaz "DIGITAL DOrigenPedidoWeb.txt (4.1)".

Regla Particular 6: Campaña referencia
Se copia el valor de la campaña aquellos registros que llegan vacíos.

Tabla 15: FN_DSTATUSFACTURACION

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANA	char	18	
2	CODREGION	nvarchar	6	
3	CODZONA	nvarchar	12	

4	FLAGSTATUSFACTSC	tinyint	1	
5	FECHA	smalldatetime	4	Particular 01
6	CODPAIS	nvarchar	4	Común 01

Regla Particular 1: Fecha de actualización

Tomar el valor de la fecha que se está procesando

Tabla 16: FN_DTIPOOFERTA

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	CODCANALVENTA	nvarchar	4	
2	CODTIPOOFERTA	nvarchar	8	
3	ABRTIPOOFERTA	nvarchar	40	
4	DESTIPOOFERTA	nvarchar	100	
5	CODSUBGRUPOTO1	nvarchar	4	
6	DESSUBGRUPOTO1	nvarchar	40	
7	CODSUBGRUPOTO2	nvarchar	4	
8	DESSUBGRUPOTO2	nvarchar	40	
9	DESTIPOPROFIT	nvarchar	40	
10	CODTIPOPROFIT	nvarchar	4	
11	CODPAIS	nvarchar	4	Común

Tabla 17: FN_DORIGENPEDIDOWEB

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ORIGENPEDIDOWEB	int	4	
2	DESORIGENPEDIDOWEB	nvarchar	200	
3	CODPOPUP	int	4	
4	DESPOPUP	nvarchar	100	
5	FLAGPERSONALIZACION	tinyint	1	
6	CODZONA	int	4	
7	DESZONA	nvarchar	100	Particular 01
8	CODMEDIO	int	4	
9	DESMEDIO	nvarchar	40	Particular 02
10	CODSECCION	int	4	
11	DESSECCION	nvarchar	100	Particular 03
12	CODPAIS	nvarchar	4	Común 01

Regla Particular 1: Datos de la zona

El valor de “CodZona” se obtiene del campo “CodArea”

El valor de “DesZona” se obtiene del campo “DesArea”

Regla Particular 2: Medio Técnico

El valor de “CodMedio” se obtiene del campo “CodMedioTec”

El valor de “DesMedio” se obtiene del campo “DesMedioTec”
Regla Particular 2: Datos de la sección
El valor de “CodSeccion” se obtiene del campo “CodEspacio”
El valor de “DesSeccion” se obtiene del campo “DesEspacio”

Tabla 18: FN_FLOGINGRESOPORTAL

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANAWEB	Nvarchar	12	
2	CODEBELISTA	nvarchar	30	
3	IPORIGEN	nvarchar	200	
4	FECHAHORA	smalldatetime	4	
5	CODPAIS	nvarchar	4	Común
6	FechaInt	int	4	
7	HoraInt	int	4	Particular 01
Regla Particular 1: Fecha y hora				
No se consideran.				

Tabla 19: FN_ FPEDIDOWEBDETALLE

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANAWEB	nvarchar	12	
2	CANTIDAD	int	4	
3	CODEBELISTA	nvarchar	30	
4	CODVENTA	nvarchar	10	
5	FECHACREACION	smalldatetime	4	
6	FLAGOFERTAWEB	nvarchar	2	
7	FLAGPROCESADO	int	4	
8	IMPORTETOTAL	decimal	17	
9	ORDENPEDIDOWD	int	4	
10	ORIGENPEDIDOWEB	int	4	
11	PEDIDODETALLEID	int	4	
12	PEDIDOID	int	4	
13	FECHAINT	int	4	
14	HORAINT	int	4	Particular 01
15	CODPAIS	nvarchar	4	Común
Regla Particular 1: Fecha y hora				
No se consideran.				

Tabla 20: FN _FOFERTAFINALCONSULTORA

ID	CAMPO	TIPO	TAMAÑO	ACCION
1	ANIOCAMPANAWEB	nvarchar	12	
2	CANTIDAD	int	4	
3	CODEBELISTA	nvarchar	30	
4	CODVENTA	nvarchar	10	
5	TIPOOFERTAFINAL	nvarchar	20	
6	FECHACREACION	smalldatetime	4	Particular 01
7	REALMNGAP	decimal	17	Particular 02
8	FECHAINT	int	4	
9	HORAINT	int	4	Particular 03
10	CODPAIS	nvarchar	4	Común
Regla Particular 1: Datos de la zona El valor de "FechaCreacion" se obtiene del campo "Fecha"				
Regla Particular 2: GAP El valor de "RealMNGap" se obtiene del campo "GAP"				
Regla Particular 3: Fecha y hora No se consideran.				

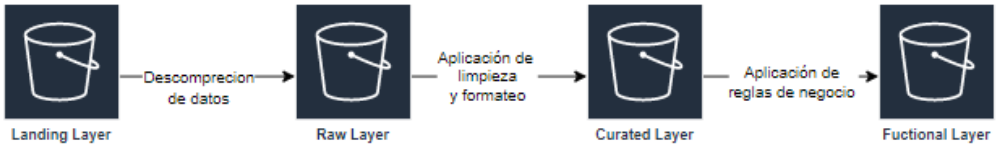
3.2.5.3 Arquitectura propuesta

La arquitectura propuesta consiste en la definición de la arquitectura lógica del data lake, la infraestructura del data lake y los procesos del data lake.

3.2.5.3.1 Arquitectura lógica del data lake

Se definió una arquitectura data lake de 3 capas, en los cuales los datos de cada capa estarían almacenados en un object storages, usando AWS S3, y el procesamiento que llevara los datos entre las diferentes capas aplicando la limpieza y formateo de datos, además de reglas de negocio, sería realizado usando el framework Spark mediante el lenguaje de programación Scala desplegado en el servicio de procesamiento AWS EMR.

Figura 7: Capas de datos

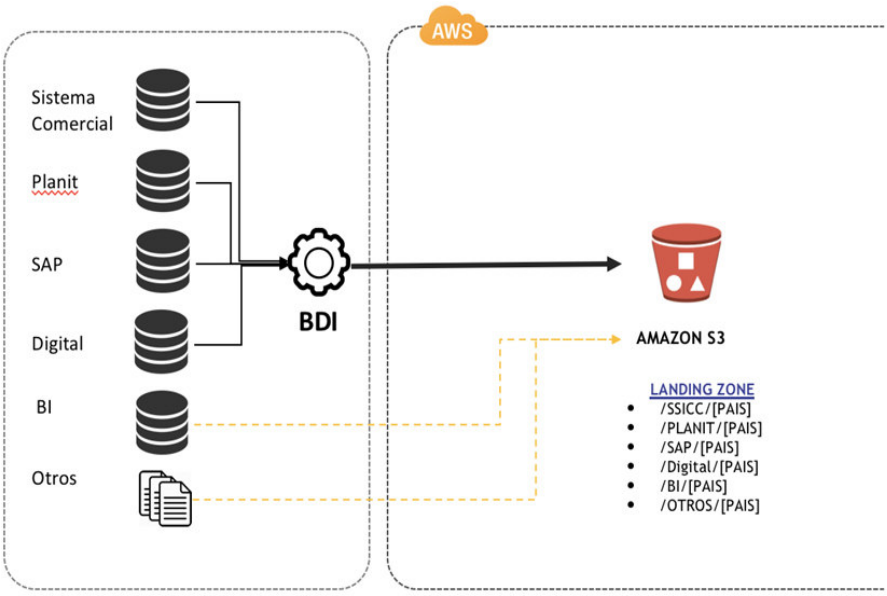


Las capas del data lake se subdividen en Landing Layer, Raw Layer, Curated Layer y Functional Layer.

- Landing Layer.- Es en donde los archivos de las diferentes fuentes arriban, en formato comprimido.
- Raw Layer.- La Raw Data almacena la información tal cual están en los distintos sistemas de origen, de ser necesario se puede aplicar una ligera transformación a los datos.
- Curated Layer.- Consume data de Raw Layer, en esta capa los datos pasan por un proceso de transformación y calidad a fin de obtener información lista para el análisis
- Functional Layer.- En esta capa es donde se aplican la lógica del negocio usando la datas de diferentes para crear las tablas funcionales, para su posterior uso para herramientas de visualización, reportes o combinación de datos por usuarios finales

A más detalle, la puerta de entrada al data lake es el Landing Layer, en donde la herramienta BDI (Belcorp data integrator) almacenara los archivos de las diferentes interfaces en un formato comprimido en carpetas con nombres referentes como se observa en Figura 8.

Figura 8: Ingesta de Interfaces de datos.



3.2.5.3.2 Infraestructura del data lake

La infraestructura del data lake viene a ser los servicios que se usan para el almacenamiento, proceamiento y orquestacion de todos los procesos del data lake.

Almacenamiento

Los layers mencionadas del data lake serán representados por Buckets S3, el cual viene a ser un servicio de Object storage de AWS el cual permite el almacenamiento de grandes volúmenes de datos. Teniendo el beneficio de solo se paga por el almacenamiento utilizado.

Procesamiento

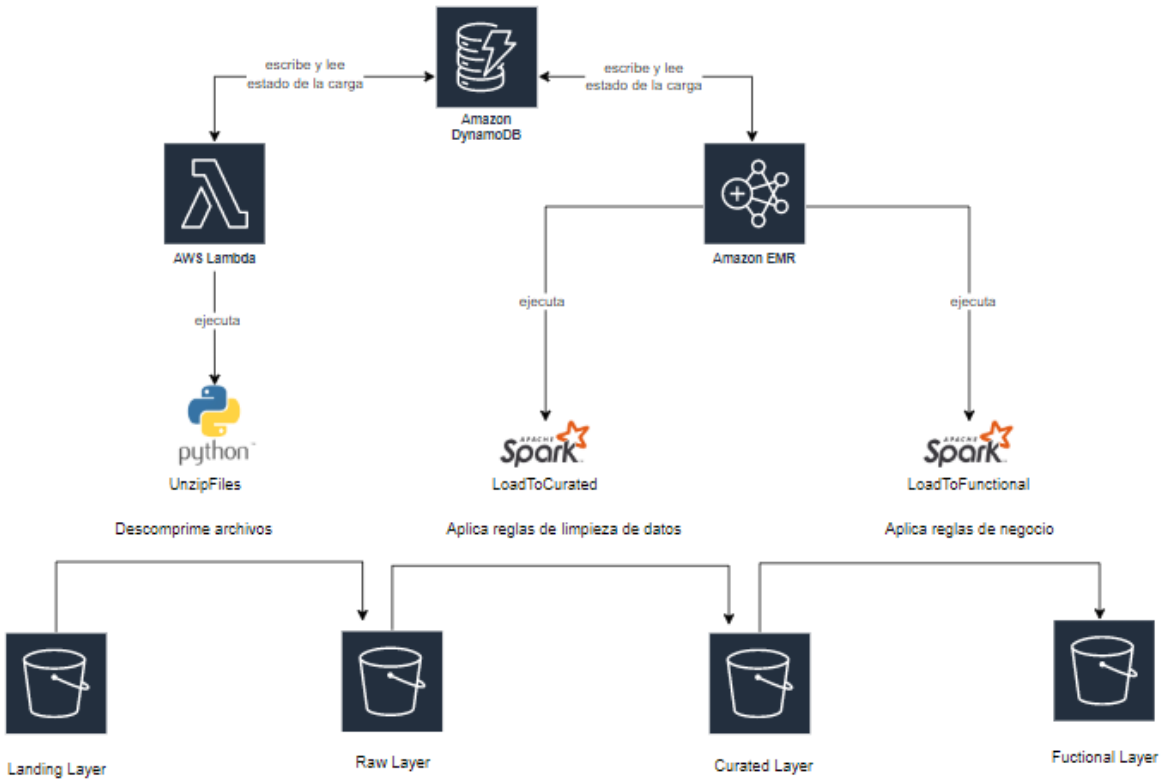
El procesamiento se dará por dos lenguajes de programacion, el primero Python que se usa para la descompresión inicial de datos comprimidos y el segundo Spark-Scala para el procesamiento de datos, mientras que la aplicación Python será ejecutada en AWS Lambda, que es un servicio de cómputo, la aplicación Spark-Scala sería

ejecutado en AWS EMR, el cual brinda un clúster de hadoop para el procesamiento distribuido en memoria.

Orquestación

El flujo descrito en la figura 5 es orquestado mediante eventos de AWS cuando un proceso termina inmediatamente gatilla al siguiente proceso, el estado de dichos procesos es almacenado en una tabla de estados, para lo cual usaremos AWS DynamoDB, el cual es un servicio de AWS que brinda una tabla serverless.

Figura 9: Arquitectura del data lake



3.2.5.3.2 Procesos del data lake

Los procesos data lake vienen a ser las aplicaciones Python y Spark-Scala, los cuales se despliegan en los servicios AWS Lambda y AWS EMR respectivamente, los cuales son los siguientes:

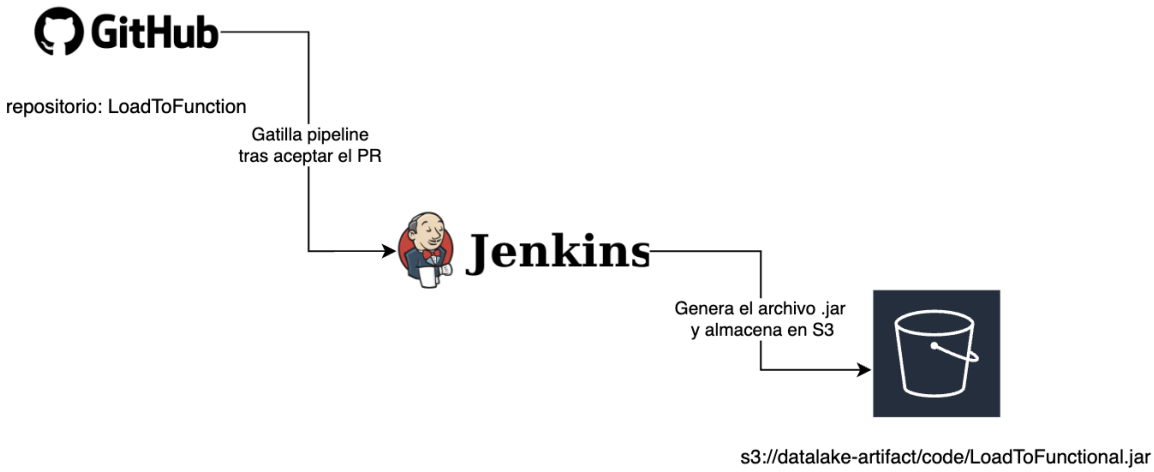
- UnzipFiles.- Proceso encargado de mover los archivos desde el bucket landing hasta el bucket Raw Layer, durante este proceso la data se descomprime ya que los archivos llegan comprimidos, finalmente este proceso se encarga de almacenar los archivos descomprimidos en folders los país, interfaz y fecha.
- LoadToCurated.- Proceso encargado de aplicar reglas de casteo, limpieza y estandarización de datos sobre la data almacenada en el bucket raw, el resultado se almacena en el bucket Curated Layer, además se cambia el formato de los datos a archivos parquet y finalmente particiona la data por los campos sistema, país y fecha.
- LoadToFunctinal - Proceso encargado de aplicar las reglas de negocio sobre los datos ya curados almacenados en el bucket Curated Layer, en este proceso se crean las tablas funcionales las cuales son alimentadas de una o más fuentes de datos, el resultado se almacena en el bucket funcional, finalmente particionada la data por el campo país.

3.2.5.4 Despliegue de la aplicación

Se propuso la implementación de un proceso de integración continua para el despliegue de las aplicaciones spark, el código de cada aplicación spark se encontraba almacenado en un repositorio, por lo cual se tuvo dos repositorios para las aplicaciones LoadToCurated y LoadToFunctional.

Cuando se realiza cambios al código y este tiene que ser desplegado en las herramientas de AWS, se gatilla un proceso en la herramienta Jenkins que compila el código y lo almacena en un bucket S3 en donde podrá ser ejecutado por el servicio AWS EMR.

Figura 10: Despliegue de aplicaciones spark.



3.3 EVALUACIÓN

3.3.1 EVALUACIÓN COSTO-BENEFICIO

En el siguiente gráfico se muestra el pago total de los roles que participaron en la implementación del data lake durante todo el proyecto.

Tabla 21: Inversión del capital humano

Rol	**N°**	**Sueldo**	**Marzo**	**Abril**	**Mayo**	**total**
Product Owner	1	5000	5000	5000	5000	15000
Technical Leader	1	9000	9000	9000	9000	27000
Team Members	3	5000	15000	15000	15000	45000

Total	**5**	**19000**	**29000**	**29000**	**29000**	**87000**

En total fueron 5 los integrantes del proyecto trabajaron durante los 3 meses de duración del proyecto, siendo asi la inversión del capital humano unos S/. 87 000.

Beneficio para la organización

Este proyecto que fue capaz la implementación del data lake de la empresa ha permitido la centralización de datos de la información, haciendo que muchos usuarios del negocio tengan acceso a dicha información, para que puedan realizar sus tareas diarias que tengan la intervención de los datos para la toma de decisiones.

Además de los usuarios del negocio, los usuarios especialistas como los científicos de datos, podrán entrenar sus modelos predictivos con la certeza que los datos será limpios y consistentes, y aparte los analistas de datos podrán crear dashboards para el negocio con una única fuente de datos, por lo cual la herramienta de visualización solo necesitara una conexión, y esta será directa al data lake.

Es importante mencionar que al estar desplegado en un ambiente cloud, en donde se paga por los recursos utilizados, el data lake permitirá ahorrar costos de cómputo al momento de hacer cálculos sobres los datos y así descontinuar el uso del data warehouse.

Otro factor importante del uso de un ambiente cloud es la facilidad de escalar el data lake al momento de ingestar más datos, lo cual conllevara a mayor uso de almacenamiento y procesamiento, dicha demanda será provista gracias a la facilidad y rapidez que un proveedor cloud ofrece.

CAPÍTULO IV: REFLEXIÓN CRÍTICA DE LA EXPERIENCIA

Esta experiencia ayuda al autor del presente trabajo a liderar proyectos de implementación de una arquitectura de datos desde su concepción que se puedan presentar posteriormente en su carrera profesional, además que ser parte del proyecto desde el inicio hace que el autor pueda que a partir de la fecha asumir el rol de arquitecto de datos para futuros proyectos.

Este proyecto se implementó con la metodología ágil Scrum lo cual permitió que el proyecto se divida en diferentes fases lo cual ayudó al autor a participar en los distintos pasos del proyecto como: la definición de la arquitectura, el levantamiento de la información, el desarrollo de los procesos big data y el despliegue de los procesos.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- Se implementó satisfactoriamente el data lake de acuerdo a las necesidades, tiempos y alcance definido en el proyecto.
- La implementacion del data lake permitio dejar de usar herramientas warehouse, la cuales resultaban costosos al momento de procesar grandes cantidades de datos.
- Posteriormente a la implementación del data lake, los equipos de analistas de datos y científicos de datos comenzaron a usar el data lake como única fuente de datos, para lo cual se capacito a ambos equipos en el uso de los datos y de las herramientas que componenl data lake.
- Se definió y elaboró procesos de carga y transformación de datos hacia el data lake, posterior a esto se procedió a crear un equipo de ingenieros de datos para que reúsen dichos procesos y los implementen en futuras cargas de nuevas fuentes de datos al data lake.

5.2 RECOMENDACIONES

- Se debería implementar una sistema de gobierno de datos transversal a las diferentes capas del data lake, teniendo en cuenta los datos que contengan información sensible del cliente, para así tomar acciones correctivas y preventivas en el acceso a los datos del data lake.
- Uno de los puntos de falla de los data lake es el ownership de los datos, ya que una fuente de datos puede ser usada por diferentes equipos y cualquier modificación sobre las data de dicha fuente de datos puede afectar a todos los equipos, por lo cual se recomendaría una mejor documentación e integración de nuevos roles como al área como el de un Product Data Owner.
- Aumentar el tipo de fuentes de datos, si bien en el proceso de levantamiento de información del proyecto se seleccionó a las interfaces más relevantes para el negocio, sería importante agregar al data lake los archivos de las interfaces no incluidas en este proyecto, ya que posiblemente puedan contener información relevante para el negocio a futuro.
- Estandarización a nivel de nomenclaturas de objetos y archivos a nivel del código Scala, el cual esta almacenados en los repositorios de Github con el fin de una fácil integración y mejora para procesos futuros del data lake.

5.3 FUENTES DE INFORMACIÓN

- Frampton, Mike. Mastering Apache Spark. Birmingham: Packt Publishing, Limited, 2015.
- Gartner (2011) Information Technology Glossary – Big data https://www.gartner.com/en/information-technology/glossary/big-data
- Karanth, Sandeep. Mastering Hadoop Generation of Hadoop Data Processing Platforms. 1st edition. Birmingham, England: Packt Publishing, 2014.
- Lwakatare, L., Kilamo, T., Karvonena, T., Sauvolaa, T., Heikkiläc, V., Itkonenc, J., Lassenius, C. (Junio de 2019). DevOps in practice: A multiple case study of five companies. Information and Software Technology.
- Menzinsky, A., López, G., Palacio, J., Sobrino, M. Á., & Rivas, R. Á. (2020). Historias de Usuario. Scrum Manager®.
- Microsoft. (2021). What is cloud computing? Retrieved from Azure Learning Path: https://docs.microsoft.com/en-us/learn/modules/intro-to-azurefundamentals/what-is-cloud-computing?ns-enrollment-type=LearningPath&nsenrollment-id=learn.az-900-describe-cloud-concepts.
- Miloslavskaya, N., & Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. Procedia Computer Science, 300-305.
- Pasupuleti, Pradeep, and Beulah Salome Purra. Data Lake Development with Big Data: Explore Architectural Approaches to Building Data Lakes That Ingest, Index, Manage, and Analyze Massive Amounts of Data Using Big Data Technologies. 1st ed. PACKT Publishing, 2015.
- Thottuvaikkatumana, Rajanarayanan. Apache Spark 2 for Beginners. Birmingham: Packt Publishing, Limited, 2016.