



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 en AWS para
mejorar la gestión de datos en una universidad**

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Ingeniería de Sistemas

AUTOR

Winy Yaribel MOSQUEDA DÁVILA

ASESOR

Arturo Alejandro BARTRA MORE

Lima, Perú

2024



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Mosqueda, W. (2024). *Implementación de un Data Lake en AWS para mejorar la gestión de datos en una universidad*. [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	Winy Yaribel Mosqueda Dávila
Tipo de documento de identidad	DNI
Número de documento de identidad	72251434
URL de ORCID	https://orcid.org/0009-0006-3073-6440
Datos de asesor	
Nombres y apellidos	Arturo Alejandro Bartra More
Tipo de documento de identidad	DNI
Número de documento de identidad	40233946
URL de ORCID	https://orcid.org/0000-0002-3411-7456
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Frank Edmundo Escobedo Bailón
Tipo de documento	DNI
Número de documento de identidad	41671087
Miembro del jurado 1	
Nombres y apellidos	Norberto Antonio Osorio Beltrán
Tipo de documento	DNI
Número de documento de identidad	08799230
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	2023
URL de disciplinas OCDE	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 la 20:00 horas del día 30 de enero del año 2024, se reunieron virtualmente los docentes designados como Miembros de Jurado del Trabajo de Suficiencia Profesional, presidido por el Dr. Escobedo Bailón Frank Edmundo (Presidente), Ing. Osorio Beltrán Norberto Antonio (Miembro) y el Ing. Bartra More Arturo Alejandro (Miembro Asesor), usando la plataforma Meet (<https://meet.google.com/byr-ntdg-dks?hs=224>), para la sustentación virtual del Trabajo de Suficiencia Profesional intitulado: **“IMPLEMENTACIÓN DE UN DATA LAKE EN AWS PARA MEJORAR LA GESTIÓN DE DATOS EN UNA UNIVERSIDAD”**, por la Bachiller Mosqueda Dávila Winy Yaribel; para obtener el Título Profesional de Ingeniería de Sistemas.

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

La 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, la Bachiller obtuvo la nota de 18 (**DIECIOCHO**).

A continuación el Presidente de Jurados el Dr. Escobedo Bailón Frank Edmundo, declara a la Bachiller **Ingeniería de Sistemas**.

Siendo las 20:50 horas, se levantó la sesión.

Presidente

Dr. Escobedo Bailón Frank Edmundo

Miembro

Ing. Osorio Beltrán Norberto Antonio

Miembro Asesor

Ing. Bartra More Arturo Alejandro



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Vicerrectorado de Investigación y Posgrado



CERTIFICADO DE SIMILITUD

Yo ARTURO ALEJANDRO BARTRA MORE en mi condición de asesor acreditado con la Resolución Decanal N° 001246-2023-D-FISI y Oficio N° 005-PTTSP-FISI-2023 del Trabajo de Suficiencia Profesional cuyo título es IMPLEMENTACIÓN DE UN DATA LAKE EN AWS PARA MEJORAR LA GESTIÓN DE DATOS EN UNA UNIVERSIDAD, presentado por el bachiller MOSQUEDA DÁVILA WINY YARIBEL para optar el título de Ingeniero de Sistemas CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de 12% de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional.**

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título correspondiente.

Firma del Asesor _____

DNI: 40233946

Nombres y apellidos del asesor: ARTURO A. BARTRA MORE



DEDICATORIA

El presente es dedicado a mi papá, quien me inculco que la educación es lo más importante para cumplir tus objetivos.

A mi mamá y mi hermana que todavía están conmigo y me apoyan en todo lo que quiero realizar.

AGRADECIMIENTO

Agradezco a la Universidad Nacional Mayor de San Marcos que me brindó los conocimientos base que hicieron que pueda crecer como profesional.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

**IMPLEMENTACIÓN DE UN DATA LAKE EN AWS PARA MEJORAR LA
GESTIÓN DE DATOS EN UNA UNIVERSIDAD**

Autor: Mosqueda Dávila, Winy Yaribel
Asesor: Bartra More Arturo Alejandro
Título: Trabajo de Suficiencia Profesional
Fecha: octubre 2023

RESUMEN

Los datos son actualmente parte fundamental para la toma de decisiones en las organizaciones, y su análisis eficiente proporciona una ventaja competitiva. Las universidades, al igual que muchas otras organizaciones, generan y procesan conjuntos de datos amplios y diversos procedentes de diversas fuentes. Los sistemas tradicionales de gestión de datos ya no son suficientes para manejar este volumen y diversidad de datos.

La universidad reconoció la necesidad de mejora, ya que los sistemas de bases de datos locales y los procesos luchan por hacerle frente a la diversidad de información en constante aumento. Los retrasos, los bloqueos del sistema y las limitaciones de almacenamiento se volvieron habituales. Además, los datos a menudo permanecían aislados, limitando su utilidad. Para abordar estos desafíos, se optó por implementar un Data Lake en la nube de AWS, que transformará a la universidad en una organización data-driven. El Data Lake sirve como un repositorio centralizado para diversas fuentes de datos, lo que permite una integración y análisis de datos sin problemas. Estos datos integrados empoderarán a diversos departamentos para desarrollar y aplicar modelos predictivos, mejorando en última instancia el proceso de toma de decisiones en toda la organización.

Al adoptar este enfoque orientado al futuro para la gestión de datos, la universidad tiene como objetivo lograr una mayor agilidad, flexibilidad y competitividad en un mundo cada vez más impulsado por los datos.

Palabras claves: data lake, cloud computing, integración de datos, gestión de datos, AWS

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

**IMPLEMENTATION OF A DATA LAKE IN AWS TO IMPROVE
INFORMATION MANAGEMENT AT A UNIVERSITY**

Autor: Mosqueda Dávila, Winy Yaribel
Asesor: Bartra More Arturo Alejandro
Título: Trabajo de Suficiencia Profesional
Fecha: Octubre 2023

ABSTRACT

Data is now a fundamental part of decision making in organizations, and its efficient analysis provides a competitive advantage. Universities, like many other organizations, generate and process large and diverse data sets from a variety of sources. Traditional data management systems are no longer sufficient to handle this volume and diversity of data.

The university recognized the need for improvement as local database systems and processes struggle to cope with the ever-increasing diversity of information. Delays, system crashes and storage limitations became commonplace. In addition, data often remained isolated, limiting its usefulness. To address these challenges, they chose to implement a Data Lake in the AWS cloud, which will transform the university into a data-driven organization. The Data Lake serves as a centralized repository for diverse data sources, allowing for seamless data integration and analysis. This integrated data will empower various departments to develop and apply predictive models, ultimately improving the decision-making process across the organization.

By adopting this forward-looking approach to data management, the university aims to achieve greater agility, flexibility and competitiveness in an increasingly data-driven world.

Keywords: data lake, cloud computing, data integration, data management, AWS

ÍNDICE GENERAL

RESUMEN	v
ABSTRACT	vi
INTRODUCCIÓN	1
CAPÍTULO I: TRAYECTORIA PROFESIONAL	3
CAPÍTULO II: CONTEXTO EN EL QUE SE DESARROLLÓ LA EXPERIENCIA.....	10
2.1. Empresa – Actividad que realiza	10
2.2. Visión	10
2.3. Misión	10
2.4. Organización de la Empresa	10
2.5. Área, cargo y funciones desempeñadas.....	13
2.6. Experiencia profesional realizada en la organización	14
CAPÍTULO III: ACTIVIDADES DESARROLLADAS	16
3.1. Situación Problemática	16
3.1.1. Definición del Problema	16
3.2. Solución	17
3.2.1. Objetivos	19
3.2.2. Alcance	19
3.2.3. Etapas y Metodología.....	20
3.2.4. Fundamentos utilizados	23
3.2.5. Implementación de las áreas, procesos y sistemas	40
3.3. Evaluación	57
3.3.1. Evaluación económica.....	57
3.3.2. Beneficios para la organización.....	58
CAPÍTULO IV: REFLEXIÓN CRÍTICA DE LA EXPERIENCIA.....	60
CAPITULO V: CONCLUSIONES Y RECOMENDACIONES	61
5.1. Conclusiones.....	61
5.2. Recomendaciones	61
5.3. FUENTES DE INFORMACIÓN.....	62
ANEXOS	64
Anexo 1: Mapeo de tablas de la fuente de base de datos Oracle	64
Anexo 2: Mapeo de módulos de Zoho CRM.....	67
Anexo 3: Flujo de orquestación.....	72

ÍNDICE DE FIGURAS

Figura 1. <i>Organigrama corporativo de la universidad</i>	10
Figura 2. <i>Organigrama académico de la universidad</i>	12
Figura 3. <i>Arquitectura del Data Lake</i>	17
Figura 4. <i>Arquitectura DWH</i>	24
Figura 5. <i>Componentes de un Data Lake</i>	26
Figura 6. <i>Cuadrante Gartner. Fuente: Gartner Inc</i>	32
Figura 7. <i>Funcionamiento de una máquina de estados en Step Functions</i>	35
Figura 8. <i>AWS Database Migration Service</i>	38
Figura 9. <i>Gantt referencial del caso de uso de ingesta desde Oracle</i>	42
Figura 10. <i>Gantt referencial del caso de uso Zoho CRM y Canvas</i>	43
Figura 11. <i>Muestra previa de los datos del módulo Leads</i>	44
Figura 12. <i>Mapeo de esquemas de Canvas</i>	44
Figura 13. <i>Arquitectura propuesta</i>	47
Figura 14. <i>Política establecida para el grupo de usuarios Administradores</i>	48
Figura 15. <i>Política establecida para el grupo de usuarios Desarrolladores</i>	49
Figura 16. <i>Política establecida para el grupo de usuarios Analistas</i>	50
Figura 17. <i>Nomenclatura de los nombre de los buckets</i>	50
Figura 18. <i>VPC configurado</i>	52
Figura 19. <i>VPN configurada</i>	53
Figura 20. <i>Customer Gateway configurado</i>	53
Figura 21. <i>Transit Gateway configurado</i>	54
Figura 22. <i>Vista ejemplo de la visualización de los datos en Amazon Athena</i>	56
Figura 23. <i>Vista ejemplo de los correo enviados de las validaciones realizadas</i>	57

ÍNDICE DE TABLAS

Tabla 1. <i>Experiencia profesional</i>	3
Tabla 2. <i>Formación Académica</i>	8
Tabla 3. <i>Formación Académica Complementaria</i>	8
Tabla 4. <i>Etapas para la implementación del Data Lake</i>	20
Tabla 5. <i>Diferencias ente Data Warehouse y Data Lake</i>	28
Tabla 6. <i>Roles involucrados en el proyecto</i>	41
Tabla 7. <i>Casos de uso</i>	42
Tabla 8. <i>Servicios usados.</i>	45
Tabla 9. <i>Descripción de las columnas de la tabla en DynamoDB</i>	51
Tabla 10. <i>Distribución mensual de costos</i>	58

INTRODUCCIÓN

En la actualidad, las organizaciones data-driven, tratan de extraer el conocimiento de todos sus datos para optimizar los aspectos de su negocio y servir mejor a sus clientes. La capacidad para aprovechar los datos se ha convertido en un factor crítico para optimizar procesos o lograr una ventaja competitiva. Sin embargo, manejar los volúmenes de datos generados y extraer el valor de los datos supone un reto para las organizaciones y requiere nuevas estrategias y conceptos que van más allá, por ejemplo, de las tradicionales soluciones de inteligencia de negocios o data warehousing.

Las organizaciones, como las universidades, recopilan y analizan cada vez más datos de fuentes tradicionales, como ERP, CRM, redes y sistemas de gestión de aprendizaje en línea, encuestas, así como fuentes de datos recientes como las redes sociales, para enriquecer sus datos y análisis. Además, están integrando fuentes modernas como las redes sociales para enriquecer sus análisis. Sin embargo, la organización en la que se desarrolla el presente trabajo, enfrenta un desafío significativo cargar todos sus datos en un almacén de datos tradicional, debido a que los datos de algunas fuentes suelen llegar en formatos semiestructurados que requieren que los datos se transformen y procesen antes de cargarlos. Además, el costo y la complejidad de almacenar grandes cantidades de datos en bruto son significativos.

Este trabajo de suficiencia profesional se centra en una solución fundamental para enfrentar los desafíos de la gestión de datos en una universidad, mediante la implementación de un Data Lake en la nube de Amazon Web Services (AWS), que se trata de una plataforma de gestión de datos diseñada para incorporar datos a escala, procedentes de fuentes heterogéneas, de estructura variable y en su formato bruto, con el propósito de mejorar la eficiencia, la accesibilidad y la capacidad de análisis de la información clave para la toma de decisiones estratégicas.

El presente trabajo de suficiencia profesional, se estructura de la siguiente manera:

- En el CAPÍTULO I, se detalla la trayectoria profesional donde se demuestra la experiencia lograda.

- En el CAPÍTULO II, se describe el contexto en el que se desarrolló la experiencia, todo lo relacionado a la organización, así como la experiencia profesional realizada en esta.
- En el CAPÍTULO III, se detalla la actividades desarrolladas en el presente trabajo, como la situación problemática, los objetivos, la metodología e implementación y los resultados de la solución.
- En el CAPÍTULO IV, se realiza una crítica a partir de la experiencia obtenida durante el proyecto.
- En el CAPÍTULO V, se exponen las conclusiones y sugerencias al proyecto aplicado.

CAPÍTULO I: TRAYECTORIA PROFESIONAL

La autora del presente trabajo de suficiencia profesional cuenta con el grado de Bachiller de la carrera de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos, con experiencia en el análisis de datos, desarrollo de soluciones de inteligencia de negocios, y actualmente enfocada en el diseño y procesamiento de datos en entornos cloud.

Actualmente, tiene más de 6 años de experiencia laboral, dentro de los cuales ha trabajado en sectores como telecomunicaciones, banca, seguros y educación. Adicional a ello, ha llevado diversos cursos de especialización en distintas entidades educativas.

Tabla 1. *Experiencia profesional*

Universidad Norbert Wiener	
Período	Abril 2022 – Actualmente
Cargo	Desarrollador de Business Intelligence
Funciones	- Desarrollar pipelines en Python y SQL Server para automatizar la generación de reportes. - Desarrollo de reportes en Power BI. - Gestión de información usando lenguajes Transact SQL y SQL Standard. - Desarrollo de procesos automáticos de carga y transaccionales usando Python. - Desarrollo de pipelines en Oracle Cloud usando servicios como Autonomous Databases, OCI y Analytics Cloud. - Desarrollo de pipelines en AWS Cloud.
ZAT Consulting	
Período	Octubre 2021 – Actualmente
Cargo	Ingeniero de datos

Funciones	- Desarrollo de ETL usando PowerCenter. - Desarrollo de pipelines y procesos de orquestación de datos en AWS Cloud. Mediante el manejo servicios como S3, Glue, Athena, Step Functions, Dynamo DB, Lambdas y EventBridge.
------------------	--

Zoluxiones S.A

Período **Agosto 2021 – Marzo 2022**

Cargo **Ingeniero de software**

Funciones	- Desarrollo de reportes y dashboards en Power BI y DataStudio. - Creación y programación de consultas avanzadas en BigQuery. - Desarrollo de automatizaciones en marketing cloud. - Análisis de información de marketing cloud, con respecto al envío de diferentes campañas, y así medir la variaciones en indicadores como open rate,click rate y el impacto de estas en el negocio.
------------------	--

Banco Ripley S.A

Período **Marzo 2021 – Agosto 2021**

Cargo **Coordinador de Contact Center**

Funciones	- Automatización de procesos y reportes se seguimiento en la gestión de cobranzas mediante el uso de herramientas de BI ETL (SSIS), SQL Server, PLSQL. - Elaboración el informe de indicadores para el área de riesgos y comercial. - Optimizar el proceso de generación de información gerencial y reportes mediante Python y Microsoft Flow. - Generación y desarrollo de reportes y estadísticas de planificación de cobranza (Evaluación de Riesgo, Mora Proyectada, Indicadores, Segmentaciones de Cartera).
------------------	--

Indra Perú S.A	
Período	Agosto 2020 – Diciembre 2020
Cargo	Analista de Datos
Funciones	- Elaboración de consultas usando lenguaje Transact SQL. - Extracción y procesamiento de información mediante el desarrollo de cargas ETL (creación de Jobs) y procesos SQL. - Análisis y construcción de dashboards e indicadores usando la herramienta Power BI. - Apoyar en la construcción modelos de datos alineados a las necesidades del negocio. - Levantar requerimientos funcionales alineados a BI de las áreas internas de la compañía - Brindar soporte al cliente y validar la información entregada en los reportes. - Analizar los modelos y procesos existentes proponiendo acciones de mejora.

Telefónica Ingeniería de Seguridad S.A.U. (TIS)	
Período	Setiembre 2019 – Julio 2020
Cargo	Analista de fraude
Funciones	- Elaboración y ejecución de consultas avanzadas PL-SQL y Transact-SQL. - Generación de reportes recurrentes de riesgos, con el uso de herramientas de BI como Power BI. - Generar automatizaciones de procesos recurrentes usando herramientas ETL como SSIS y RPA (. NET).

-
- Mejora y automatización de procesos para la generación de reportes.
 - Administrar los sistemas de gestión de riesgo del área.
 - Atender los requerimientos de información internos y externos
 - Generar nuevos escenarios de riesgos, a través de generación de perfiles y alertas (nuevas fuentes).

Manpower Perú S.A

Período **Marzo 2019 – Agosto 2019**

Cargo **Gestor de datos**

-
- Funciones**
- Elaboración de consultas usando Transact SQL.
 - Creación de dashboards con el uso de herramientas de BI (Power BI).
 - Creacion de ETL's para la automatización de reportes y carga de base de datos usando SSIS.
 - Definición y construcción de KPIs.
 - Mejora y automatización de procesos para la generación de reportes.

Gestión de Servicios Compartidos S.AC (TGESTIONA)

Período **Junio 2018 – Enero 2019**

Cargo **Técnico de base de datos**

-
- Funciones**
- Elaboración y ejecución de consultas tanto en PL-SQL y Transact-SQL.
 - Creación de dashboards o reportes con el uso de herramientas de BI como Power BI y Tableau
 - Construcción de un datamart para la centralización de la información del área.
 - Definición y construcción de KPIs.
-

-
- Mejora en los procesos de adquisición de información.
 - Automatización de procesos para la realización de reportes y gestión de información.
 - Carga de bases, evaluación y monitoreo de carteras.

Everis BPO Perú S.A

Período	Marzo 2017 – Febrero 2018
----------------	----------------------------------

Cargo	Operador de base de datos
--------------	----------------------------------

Funciones	- Elaboración y ejecución de consultas usando PL-SQL. - Extracción y validación de la data obtenida para la generación de reportes. - Ejecución de procesos automatizados mediante uso de UNIX (Programación en Shells) - Monitoreo y actualización de los informes de gestión operativa (Cuadros de Mando) y de los indicadores de rendimiento de los procesos, usando como herramienta principal Excel. - Generación de reportes automáticos usando herramientas como VBA y Excel. - Realizar carga y descarga de archivos vía FTP.
------------------	--

Nota. Elaboración propia

Tabla 2. *Formación Académica*

Periodo	Formación
2011 - 2016	Universidad Nacional Mayor de San Marcos Bachiller en Ingeniería de Sistemas

Nota. Elaboración propia

Tabla 3. *Formación Académica Complementaria*

Periodo	Formación
2018	Universidad Tecnológica del Perú Especialización en Gestión de Proyectos
2020	J & J Data Mining Consulting Curso de Python for Analytics
2020	Instituto Peruano de Asuntos Públicos, Política y Gobierno Diplomado de Alta Especialización en Power BI
2020	J & J Data Mining Consulting Programa de Especialización Analítica Data Science Fundamentals
2021	J & J Data Mining Consulting Curso de AWS Fundamentals
2021	Big Data Academy Cloud AWS Big Data Analytics Professional y Introduction AWS Big Data Architect Cloud
2021	Coursera Distributed Computing with Spark SQL
2021	Big Data Academy Spark y Scala en Databricks
2021	Big Data Academy

	Programa de Especialización en Google Cloud Professional Data Engineer
2021	Coursera Exploring and Preparing your Data with BigQuery
2022	Datapath Curso Data Engineer With GCP
2023	Big Data Academy Change Data Capture on GCP
2023	Big Data Academy PySpark Multicloud

Nota. Elaboración propia

CAPÍTULO II: CONTEXTO EN EL QUE SE DESARROLLÓ LA EXPERIENCIA

2.1. Empresa – Actividad que realiza

En su página web, la universidad, manifiesta que:

Es una institución académica que asume la misión de potenciar a las personas, formándolas profesional y humanísticamente para que logren su autorrealización. Con su renovada imagen, consecuencia y evidencia del interés en sus potenciales usuarios, brinda una educación de calidad certificada y acreditada, en línea con los intereses y aspiraciones de desarrollo de la sociedad.

2.2. Visión

En su página web, la universidad, manifiesta que:

Ser una universidad reconocida por su liderazgo en calidad educativa y servicio.

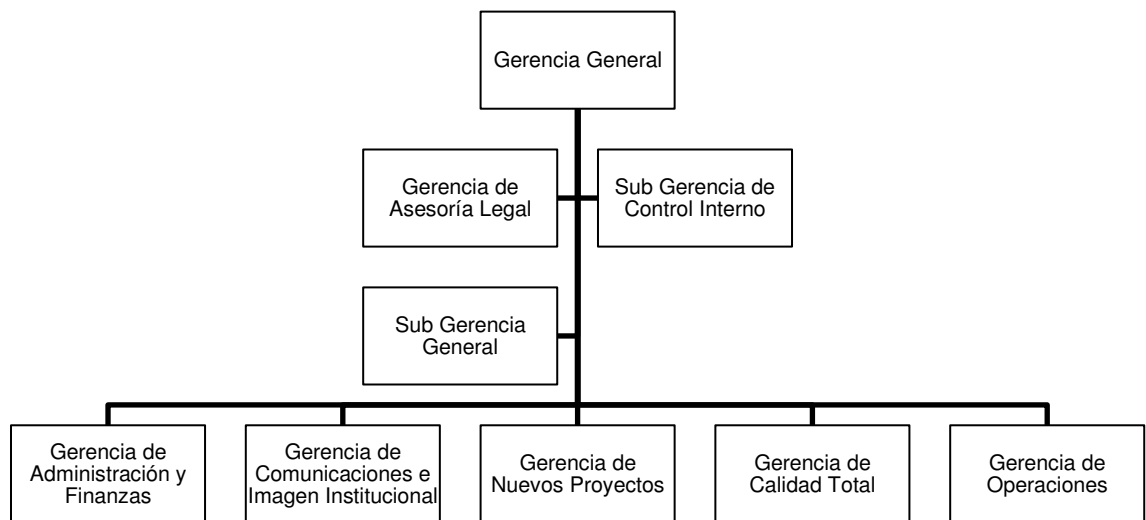
2.3. Misión

En su página web, la universidad, manifiesta que:

Transformamos vidas a través de una experiencia educativa excepcional, formando profesionales que generan valor en la sociedad.

2.4. Organización de la Empresa

Figura 1. *Organigrama corporativo de la universidad*

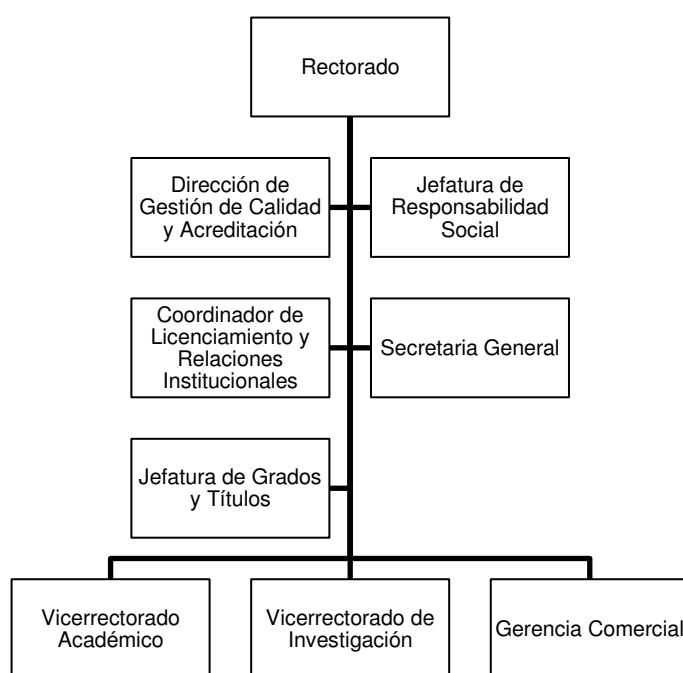


Nota. Elaboración propia

- La Gerencia General realiza la supervisión general de los asuntos académicos, administrativos y de la actividad empresarial de la universidad.
- Gerencia de Asesoría Legal, supervisa legalmente las actividades y los proyectos de la empresa; además de establecer estrategias de defensa de los intereses legales de la empresa
- Sub Gerencia de Control Interno, desarrolla importantes políticas y procedimientos respecto a los procesos financieros y de control que actualmente tiene. Adicionalmente, supervisan la implementación y monitoreo, y garantizan el cumplimiento.
- Gerencia de Administración y Finanzas, elabora y propone políticas, normas y procedimientos de administración y control para el registro de la información contable.
- Gerencia de Comunicaciones e Imagen Institucional, planifica, dirige y controla las actividades de comunicación, prensa y difusión de la gestión institucional; así como, fortalecer las relaciones interinstitucionales con las instituciones públicas o privadas.

- Gerencia de Calidad Total, se encarga de verificar que las políticas y estrategias tengan un mejoramiento continuo conducente a la calidad educativa.
- Gerencia de Operaciones, implementar los procesos y prácticas correctos en la organización.

Figura 2. *Organigrama académico de la universidad*



Nota. Elaboración propia

- Rectorado, se encarga de dirigir la actividad académica de la universidad.
- Vicerrectorado Académico, planificar, organizar, dirigir, ejecutar y controlar las políticas y actividades del vicerrectorado de acuerdo a las leyes y reglamentos.
- Vicerrectorado de Investigación, dirigir y promover el desarrollo de las investigaciones entre los miembros de la comunidad universitaria.
- Secretaría General, subscribir documentos emitidos por la universidad, a fin de dar fe que fueron emitidos por la autoridades competentes de la misma.

- Dirección de Gestión de Calidad y Acreditación, garantizar y mejorar la calidad de la educación superior que ofrece la institución, así como obtener y mantener acreditaciones que validen la excelencia de sus programas académicos.
- Jefatura de Responsabilidad Social, supervisar, coordinar y promover iniciativas y programas relacionados con la responsabilidad social de la institución.
- Coordinador de Licenciamiento y Relaciones Institucionales, cumplimiento normativo, la acreditación y las relaciones externas de la institución educativa.
- Jefatura de Grados y Títulos, gestionar y supervisar todo lo relacionado con la expedición de grados y títulos académicos a los estudiantes que han cumplido con los requisitos necesarios para obtenerlos.

2.5. Área, cargo y funciones desempeñadas

La autora del presente trabajo del presente trabajo de suficiencia profesional, ejerce el cargo como Desarrollador de Business Intelligence en el área de Arquitectura de datos en la universidad.

Dentro de las principales actividades del área de arquitectura de datos es construir crear la arquitectura tecnológica para soportar el proceso de big data, incluyendo la recopilación de datos, lectura y explotación de la información y su salida y presentación. Así como estandarizar el modelo de las estructuras de datos, gestionar y realizar el seguimiento a los proyectos de análisis de datos e inteligencia de negocios.

La autora dentro de sus funciones como Desarrollador de Business Intelligence tiene las siguientes responsabilidades:

- Realizar la ingesta de datos desde múltiples fuentes hacia el repositorio (Data Warehouse) de la universidad.
- Construir modelos de datos óptimo que permita la analítica y explotación de los datos.
- Realizar los procesos ETL de forma eficiente, para garantizar que los datos estén limpios, coherentes y listos para su uso en los reportes y análisis.

- Elaborar y diseñar dashboards ejecutivos y operativos de alto impacto.
- Construcción de consultas SQL avanzadas, desarrollo de store procedures y jobs para optimizar la explotación de los datos.
- Participar en la planificación e implementación de proyectos de inteligencia de negocios.
- Responsable del cumplimiento de las políticas de gobierno de datos.
- Participar en diseño y la implementación del proyecto de Data Lake, que implica la extracción, análisis y procesamiento de grandes volúmenes de datos.
- Disponibilizar las soluciones a los usuarios finales en el entorno Data Lake.
- Validar que los datos generados en el entorno Data Lake sean precisos y confiables.

2.6. Experiencia profesional realizada en la organización

La autora del presente trabajo de suficiencia profesional, ha obtenido las siguientes experiencias claves dentro de la organización:

- Mejora de los procesos de extracción y procesamiento de los datos, y diseño de los tableros de control existentes, que permitían el control de los ingresos y la morosidad de la universidad, brindando una comprensión más profunda del modelo de negocio de la universidad.
- Diversos proyectos para las áreas operativas y de gestión académica de la organización, que centraban en la automatización de procesos para la generación de reportes, que han optimizado la toma de decisiones y eficiencia de operaciones.
- Sin embargo, uno de los proyectos más significativos fue la implementación de un Data Lake. El cual implicaba el desarrollo de pipelines utilizando los diversos servicios de analítica que brinda la nube de Amazon Web Services (AWS), y seguir los lineamientos de seguridad de la universidad. A pesar de su complejidad, esta experiencia resultó en un profundo conocimiento de la gestión de datos y la infraestructura tecnológica de la institución.

- Mi experiencia previa como ingeniero de datos, ha contribuido para tener una visión más clara y completa al analizar y diseñar soluciones en el entorno del Data Lake.

CAPÍTULO III: ACTIVIDADES DESARROLLADAS

3.1. Situación Problemática

La universidad brinda un servicio educativo de calidad certificada a más de 20.000 estudiantes en más de 60 programas y especializaciones de pregrado y posgrado. A lo largo de los años, la universidad ha implementado aplicaciones web, ERP y sistemas de gestión académica para gestionar sus operaciones.

Sin embargo, a medida que la universidad ha seguido creciendo y diversificando sus operaciones, ha visto la necesidad de introducir nuevas plataformas como Zoho CRM, Canvas y los próximos módulos de HCM y Campus de Oracle Cloud. Estos últimos se convertirán en los sistemas principales para la gestión académica y operativa, reemplazando a los sistemas locales actuales. Esto conlleva a que la universidad enfrente varios desafíos críticos como tiempos de respuesta lento y la incapacidad para manejar la creciente cantidad de datos que la universidad genera diariamente, tanto de los sistemas tradicionales como de las nuevas plataformas que brindan datos en nuevos formatos, por parte del entorno de Data Warehouse que se está manejando, así como la falta de acceso oportuno a datos de calidad.

Sobre este contexto, la universidad decide implementar un Data Lake en la nube de Amazon Web Services (AWS), que brinda una estructura centralizada y escalable para almacenar, gestionar, integrar y analizar la creciente cantidad de información.

3.1.1. Definición del Problema

3.1.1.1. Problema Principal

La falta de un repositorio centralizado que permita una eficiente y organizada gestión de los datos en la universidad para la toma de decisiones eficiente.

3.1.1.2. Problemas Secundarios

- La falta de escalabilidad en el entorno actual de Data Warehouse se manifiesta en la incapacidad para satisfacer la creciente demanda de velocidad de procesamiento y capacidad de almacenamiento, principalmente debido a la inclusión de nuevos orígenes de datos en grandes volúmenes y con una amplia variedad de formatos.

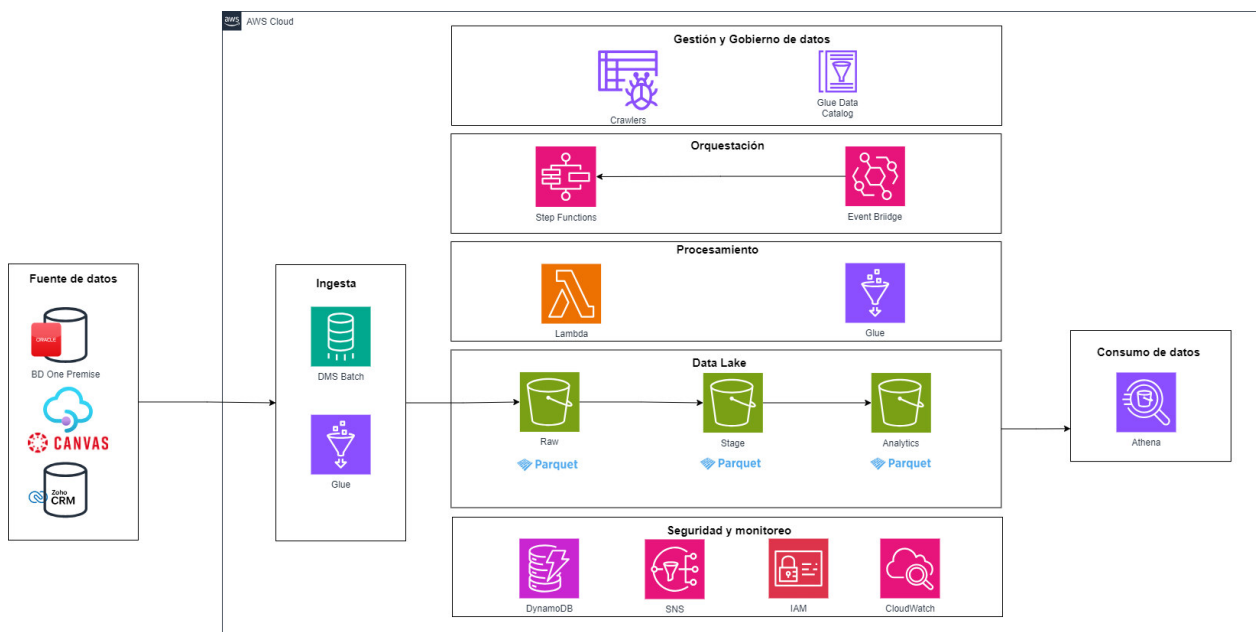
- La falta de integración de los datos de diversas fuentes hace que la universidad no pueda aprovechar todo el potencial de la información que posee.

3.2. Solución

Implementación de un Data Lake en la nube Amazon Web Services (AWS) para mejorar la gestión de datos en la universidad, para esto se realizará el uso de los diversos servicios que brinda esta nube como: Amazon S3, AWS Glue, Amazon VPC, AWS Data Migration Service (AWS DMS), AWS Lambda, AWS Dynamo DB, Amazon Athena, AWS Identity and Access Management (IAM), y otros servicios complementarios como Amazon SNS, AWS Step Functions, Amazon EventBridge y Amazon CloudWatch.

La Figura 3 muestra la arquitectura del Data Lake en AWS mediante el uso de servicios específicos de AWS.

Figura 3. *Arquitectura del Data Lake*



Nota. Fuente: Elaboración propia.

En la arquitectura de Data Lake en AWS para este proyecto, se han gestionado datos procedentes de tres fuentes distintas: una base de datos Oracle local, Zoho CRM y Canvas.

El proceso de carga de datos al Data Lake en AWS comienza con la ingesta de datos. Para la base de datos Oracle local, utilizamos el servicio AWS Data Migration Service (DMS). DMS se conecta a la base de datos Oracle local y realiza cargas completas o incrementales de las tablas según sea necesario. En el caso de Zoho CRM y Canvas, a los cuales se puede acceder mediante llamadas a sus API's respectivas, empleamos AWS Glue.

El procesamiento de datos se efectúa mediante funciones AWS Lambda, las cuales inician y supervisan las tareas de migración. Además, utilizamos AWS Glue para transformar los datos según nuestras necesidades.

Para orquestar estos flujos de ingesta y procesamiento, utilizamos AWS Step Functions. Estos flujos están programados mediante AWS EventBridge, lo que permite una ejecución programada.

Los datos ingresados y transformados se almacenan en Amazon S3 en tres capas: Raw (datos sin procesar), Stage (datos transformados) y Analytics (datos enriquecidos):

Una vez que los datos se han transformado y enriquecido, los usuarios finales pueden acceder a ellos a través de Amazon Athena para realizar consultas ad-hoc.

La gestión de los metadatos de todos los conjuntos de datos en el Data Lake se lleva a cabo mediante el AWS Glue Data Catalog como un catálogo central. Esto facilita la búsqueda de los datos en el Data Lake.

Para controlar los permisos y el acceso a los servicios de AWS, utilizamos AWS Identity and Access Management (IAM) para garantizar que los usuarios tengan el acceso adecuado según sea su rol y de esta manera poder cumplir con los principios de seguridad.

Para supervisar los flujos, utilizamos Amazon CloudWatch. Además, Amazon SNS se emplea para enviar notificaciones y alertas al equipo responsable. Estas alertas pueden indicar tanto la finalización exitosa de los flujos como la detección de errores críticos durante el proceso, lo que garantiza un monitoreo efectivo y una respuesta rápida a cualquier problema.

3.2.1. Objetivos

3.2.1.1. Objetivo General

Implementar un Data Lake en la nube Amazon Web Services (AWS) para mejorar la gestión de datos en la universidad, que facilite la toma de decisiones de las diferentes áreas académicas y operativas de la organización.

3.2.1.2. Objetivos Específicos

- Realizar una migración efectiva y el desarrollo de flujos de ingesta eficientes desde múltiples fuentes, para que los datos estén disponibles en el Data Lake de manera oportuna y precisa.
- Establecer un proceso de organización de datos en las tres capas (raw, stage y analytics) en el Data Lake.
- Facilitar el acceso a los datos almacenados en el Data Lake, permitiendo consultas ad-hoc y análisis avanzados para apoyar la toma de decisiones de las diferentes áreas de la universidad.
- Implementar una estrategia de seguridad para garantizar que los datos estén protegidos y que los usuarios tengan acceso controlado.

3.2.2. Alcance

3.2.2.1. Alcance Funcional

El alcance de este proyecto conlleva a la implementación de un Data Lake que implicara el almacenamiento de los datos de las bases de datos tradicionales actuales, así como de las nuevas plataformas, tanto de Zoho CRM y Canvas que brindan datos en formatos no estructurados.

Esta implementación incluye, la creación de los flujos de ingesta y procesamiento de datos de manera óptima, la configuración de gestión de accesos mediante políticas de acceso basadas en roles. Así como la validación de que los usuarios finales puedan acceder a los datos, para lo cual es necesario sesiones de capacitación para que puedan navegar en el Data Lake, ejecutar sus consultas y generar sus reportes.

3.2.2.2. Alcance Organizacional

El Data Lake será usado por los diferentes áreas tanto operativas y académicas de la organización para consultar datos. El área de Arquitectura de datos es la encargada de la implementación técnica, administración y monitoreo del Data Lake en AWS.

3.2.2.3. Alcance Tecnológico

Se usarán los servicios de Amazon Web Services(AWS), como:

- Amazon S3 para el almacenamiento de objetos.
- AWS Glue para el procesamiento, gobernanza y catálogo de los datos.
- AWS Data Migration Service para la migración de bases de datos relacionales.
- AWS Dynamo DB para la gestión de metadatos.
- AWS Athena para consultas SQL.
- AWS Identity and Access Management (IAM) para el control de accesos.

Otros servicios complementarios como, AWS Step Functions para la orquestación y AWS CloudWatch para la monitorización de los servicios.

3.2.3. Etapas y Metodología

Las etapas que se realizaron para la implementación del Data Lake, la cual se basa en una metodología en cascada, se describen a continuación:

Tabla 4. Etapas para la implementación del Data Lake

Planificación y análisis	
Tareas	- Definición de objetivos
	- Definición del equipo del proyecto
	- Relevamiento de datos: identificación de las fuentes, como las que ya se están trabajando (base de datos onepremise) como de las

	nuevas plataformas, describir tanto los formatos, volúmenes y estructuras.
	– Evaluación de servicios: investigación y selección de los servicios de AWS adecuados.
Entregables	– Registro de fuentes de datos y su documentación. – Selección de recursos de AWS.
Técnicas a utilizar	– Reuniones con las partes interesadas para comprender objetivos y necesidades. – Análisis de datos existentes y definición de flujos de datos necesarios
Diseño	
Tareas	– Diseño de arquitectura: creación de la arquitectura de alto nivel que incluya las estructuras de almacenamiento, flujos de procesamiento y medidas de seguridad. – Seguridad y acceso: establecimiento de las políticas de seguridad y control de accesos para proteger los datos almacenados.
Entregables	– Diseño de arquitectura.
Técnicas a utilizar	– Modelado de flujos de datos con diagramas. – Revisión y validación técnica de la arquitectura.
Implementación técnica	
Tareas	– Configuración de infraestructura : configuración de los recursos de AWS, como los buckets de S3, VPN y los otros servicios necesarios. – Desarrollo de los flujos de datos: implementación de los flujos para la extracción, transformación y carga en el Data Lake.

	– Carga de datos inicial: realización de cargas de datos inicial desde las fuentes existentes hacia el Data Lake.
Entregables	– Infraestructura de AWS configurada y provisionada. – Flujos de datos implementados. – Datos migrados y cargados en el Data Lake.
Técnicas a utilizar	– Programación y desarrollo de flujos ETL con servicios como AWS Glue. – Configuración de recursos de AWS
Pruebas y validación	
Tareas	– Pruebas de rendimiento: implica la ejecución completa de los flujos para validar el rendimiento, y garantizar que funciones según lo previsto. – Validación de datos: comparado entre los sistemas fuente y los datos cargados al Data Lake.
Entregables	– Informes de las pruebas de rendimiento. – Informes de validación de datos.
Técnicas a utilizar	– Pruebas de carga para evaluar el rendimiento. – Pruebas exhaustivas de validación de datos.
Despliegue y monitoreo	
Tareas	– Puesta en marcha: colocar el Data Lake en producción y se permite el acceso a los usuarios finales – Monitoreo continuo: establecen las alertas para garantizar el rendimiento continuo.
Entregables	– Data Lake en producción.

	– Alertas configuradas y operativas.
Técnicas a utilizar	– Despliegue del sistema en producción de acuerdo con las mejores prácticas de AWS.

Nota. Elaboración propia

3.2.4. Fundamentos utilizados

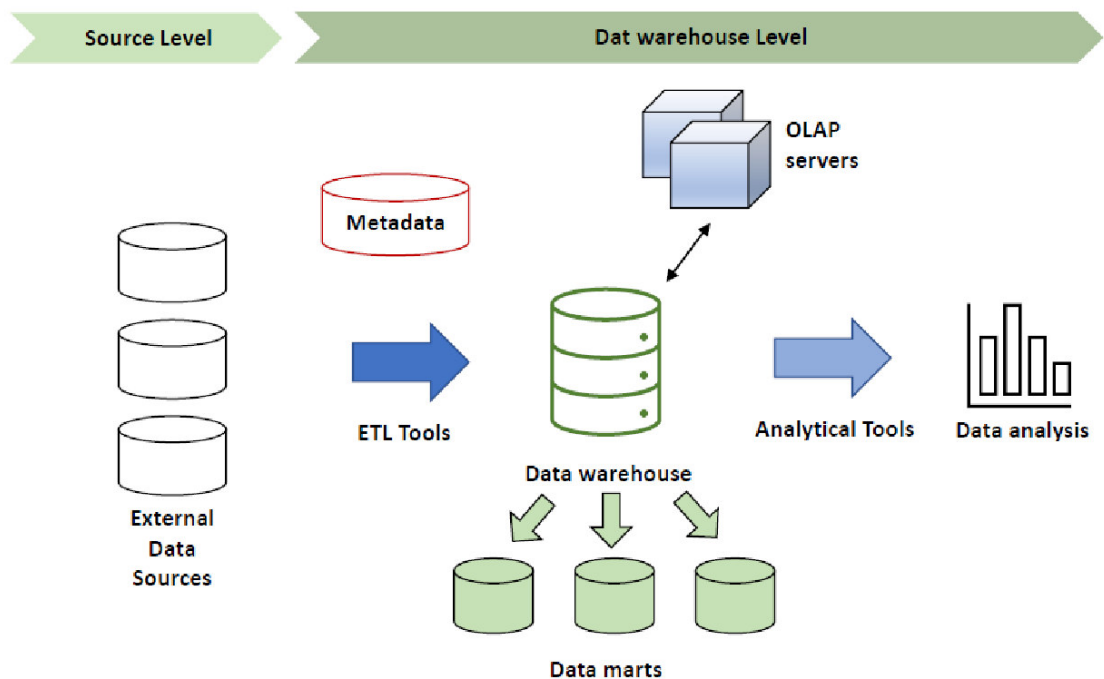
3.2.4.1. Data Warehouse (DWH)

El concepto de data warehouses (DW) o almacén de datos fue introducido a finales de los años 1980 por los investigadores de IBM, Barry Devlin y Paul Murphy con el objetivo de entregar un modelo arquitectónico para resolver el flujo de datos a los entornos de apoyo de toma de decisiones.

Según (Nambiar & Mundra, 2022), un almacén de datos es un gran repositorio de datos en el que se pueden almacenar e integrar datos de diversas fuentes y ayudar en el proceso de toma de decisiones mediante un análisis de datos adecuado.

Para (Pulidori, 2021), la tecnología de Data Warehouse tiene la capacidad de reunir y consolidar grandes volúmenes de información procedente de diversas fuentes. A través de sus funciones analíticas, brinda a las empresas la capacidad de aprovechar estos datos para tomar decisiones más informadas. Además, el Data Warehouse mantiene un registro histórico a lo largo del tiempo, que puede ser utilizado por científicos de datos y analistas de negocio. Es fundamental destacar que el enfoque principal del Data Warehouse radica en la lectura de datos, ya que estos no pueden ser modificados y generalmente solo están disponibles para su consulta y de acuerdo con lo expuesto por (Palanivel y Suresh, 2020), las actualizaciones normalmente se llevan a cabo cuando los almacenes de datos están fuera de línea.

Figura 4. *Arquitectura DWH*



Nota. La figura muestra la arquitectura de Data Warehouses. Tomado de Pulidori, M. (2021). Design and Implementation of an Application Data Lake to Support Repair Process in Baker Hughes (Master's thesis). Politecnico di Torino, Italia.

La arquitectura del Data Warehouse (DWH), como se muestra en la Figura 4, comprende varios componentes esenciales:

- Fuentes de datos externas: estas fuentes pueden ser diversas, incluyendo datos transaccionales de bases de datos relacionales OLTP, textos, y archivos de excel.
- Herramientas ETL: se utilizan para llevar a cabo las etapas de extracción, transformación y carga de datos, preparándolos para su carga en el DWH.
- Extracción: se adquieren los datos de las fuentes mencionadas.
- Transformación: los datos se someten a procesos de limpieza y transformación para adaptarlos al formato operativo del DWH.
- Carga: los datos transformados se cargan en el DWH.
- Metadatos: esta información es esencial para el funcionamiento del DWH, ya que describe la organización de los datos y controla su comportamiento.

- Data Warehouse: es el repositorio central donde se almacenan todos los datos OLTP transformados.
- Data Marts: Estos son almacenes de datos más pequeños que se centran en áreas o departamentos específicos. Pueden alimentarse desde el DWH central o directamente desde las fuentes de datos.
- Servidores OLAP: se utilizan para realizar agregaciones y análisis sobre los datos.
- Herramientas analíticas: ofrecen capacidades para el análisis estadístico, generación de informes y minería de datos.
- Herramientas de cliente: se utilizan para la visualización y presentación de datos a los usuarios del negocio.

Según (Leo, 2017), la aparición del Big Data ha planteado desafíos significativos para los Data Warehouses tradicionales. Estos desafíos engloban la incapacidad de recopilar información proveniente de diversas fuentes, como sistemas en línea, juegos y redes sociales. También implica la necesidad de optimizar tanto el diseño lógico como el diseño físico de los almacenes de datos para gestionar grandes cantidades de información. Además, se destaca que los Data Warehouses convencionales no son capaces de procesar datos no estructurados, que pueden variar en su formato, y convertirlos en conocimiento. Esto se debe a que, en su concepción original, estos sistemas se crearon con la finalidad de integrar datos estructurados provenientes de fuentes transaccionales y respaldar análisis basados en OLAP.

3.2.4.2. Data Lake

El termino Data Lake fue introducido por James Dixon in 2010. Según (Helttula, 2023), un Data Lake se refiere a un sistema escalable de almacenamiento y procesamiento de información de diversas naturalezas, conservada en su formato original, y su principal uso se centra en permitir a profesionales especializados en datos, como estadísticos o analistas, extraer valioso conocimiento a partir de dicha información.

Para (Palanivel y Suresh, 2020), el Data Lake democratiza los datos y es una forma rentable de almacenar todos los datos de una organización para su posterior

procesamiento. El Data Lake tiene una arquitectura plana. Cada elemento de datos recibe un identificador único y se etiqueta con un conjunto de información de metadatos. Los conceptos clave de un Data Lake son auditoría, descubrimiento, exploración, ingesta, gobernanza, linaje, calidad, almacenamiento y seguridad de datos. Es necesario comprender estos conceptos clave para comprender completamente la arquitectura. Uno de sus principales objetivos de un Data Lake es ayudar a promover el autoservicio desde la perspectiva del usuario y analista empresarial.

Figura 5. *Componentes de un Data Lake*



Nota. La figura muestra los componentes de un Data Lake. Tomado de Palanivel K, Suresh Joseph K (2020). Data Lake Model to Modern Educational Organizations. International Research Journal of Engineering and Technology (IRJET) Volumen 07.

Según (Palanivel y Suresh, 2020), como se ilustra en la Figura 5, un Data Lake está compuesto por los siguientes elementos:

- La ingesta de datos permite a los conectores obtener información de diversas fuentes de datos y cargarlos en el Data Lake.
- El almacenamiento de datos debe ser escalable, proporcionar un costo eficiente de almacenamiento y permitir un acceso rápido para la exploración de datos. Debe admitir múltiples formatos de datos.
- El gobierno de datos implica la gestión de la disponibilidad, usabilidad, seguridad e integridad de los datos utilizados en una organización.
- La seguridad de los datos debe estar presente en todas las capas del Data Lake. La prioridad fundamental es evitar el acceso de usuarios no autorizados, y la autenticación, la autorización y la protección de datos son aspectos clave de ello.
- La calidad de los datos es un componente esencial en la arquitectura del Data Lake. Extraer conocimiento a partir de datos de baja calidad resultará en conocimiento deficiente.
- La etapa de descubrimiento de datos es crucial antes de comenzar la preparación y el análisis de datos. En esta etapa, se utiliza la técnica de etiquetado para expresar la comprensión de los datos, organizándolos e interpretándolos después de su ingreso en el Data Lake.
- Las tareas de auditoría de datos supervisan los cambios en el conjunto de datos principal y contribuyen a evaluar los riesgos y el cumplimiento. Este componente se ocupa de los orígenes de los datos y su evolución a lo largo del tiempo, lo que facilita la corrección de errores en el proceso de análisis de datos desde el origen hasta el destino.
- La exploración de datos es la primera etapa en el análisis de datos y es fundamental para identificar el conjunto de datos adecuado antes de iniciar la exploración de datos.

Todos estos componentes deben trabajar de manera conjunta para desempeñar un papel crucial en la creación de un Data Lake que pueda evolucionar y explorar eficazmente su entorno.

Para (Nambiar & Mundra, 2022) los Data Lakes, están experimentando una creciente popularidad en las organizaciones, ya que, en contraposición a la forma tradicional de gestionar datos procedentes de distintas fuentes, los Data Lakes sustituyen esta práctica por un repositorio centralizado, altamente eficiente y altamente adaptable. Estos almacenes también permiten a las organizaciones escalar sus necesidades al separar el almacenamiento de la capacidad computacional. Esto conlleva a la eliminación de la compleja transformación y preprocesamiento de datos que suelen requerir los almacenes de datos tradicionales. Asimismo, se reduce la carga financiera inicial asociada a la ingestión de datos. Una vez que los datos se han recopilado en el Data Lake central, la organización puede llevar a cabo un análisis exhaustivo de los mismos.

3.2.4.3. Diferencias entre Data Warehouse y Data Lake

En la tabla 3, se detallan las diferencias entre Data Warehouses y Data Lakes:

Tabla 5. *Diferencias ente Data Warehouse y Data Lake*

Parámetros	Data Warehouse	Data Lakes
Data	Solo se presentan en procesos de negocio	Almacenan todo
Procesamiento	Los datos están altamente procesados	Los datos están principalmente sin procesar
Tipo de datos	Los datos están en forma y estructura tabular	Los datos pueden ser no estructurados, semiestructurados o estructurados
Agilidad	Menos ágil y tiene una configuración fija	Muy ágil y puede configurarse y reconfigurarse según sea necesario.

Usuarios	Ampliamente utilizados por profesionales y analistas de negocio	utilizados por científicos de datos, desarrolladores de datos y analistas de negocio
Almacenamiento	Se utiliza un almacenamiento caro que ofrezca tiempos de respuesta rápidos	Diseñados para un almacenamiento de bajo coste
Seguridad	Permite un mejor control de los datos	Ofrece menos control
Esquema	Esquemas predefinidos (On-write)	Sin esquemas predefinidos (On-read)
Procesamiento de datos	Lleva tiempo ingesta de nuevos datos.	Ayuda a la ingestión rápida de nuevos datos
Granularidad de datos	Datos resumidos o agregados	Datos con bajo nivel de detalle o granularidad
Herramientas	Herramientas comerciales en su mayoría	Pueden utilizarse herramientas de código abierto como Hadoop o Map Reduce

Nota. Fuente: Big Data Cogn. Comput. Tomado de Nambiar, A.; Mundra, D (2022).

3.2.4.4. Cloud Computing

La computación en la nube (o cloud computing) es la entrega bajo demanda de capacidad informática, almacenamiento de bases de datos, aplicaciones y otros recursos de TI a través de una plataforma de servicios en la nube mediante Internet con un sistema de precios de pago por uso. Para (Nishimiura,2022), al adoptar la computación en la nube, las empresas pueden disfrutar de diversas ventajas:

- Reemplazar el gasto de capital por gasto variable: en lugar de realizar una inversión inicial costosa en data centers y servidores físicos, solo se paga por los recursos informáticos que se utilizan.

- Eliminar la necesidad de gestionar y mantener data centers: al dejar esta responsabilidad en manos de proveedores de servicios en la nube como AWS, las empresas pueden enfocarse más en el desarrollo de aplicaciones y la atención a los clientes.
- Olvidarse de las estimaciones de capacidad: ya no es necesario adivinar cuánta infraestructura se necesita de antemano. Con la computación en la nube, se puede escalar los recursos según la demanda, ya sea para aumentarlos o reducirlos.
- Aprovechar las economías de escala: los proveedores de servicios en la nube, como AWS, pueden ofrecer costos más bajos gracias a su capacidad para atender a un gran número de clientes, lo que se traduce en precios más competitivos.
- Ganar en velocidad y flexibilidad: la computación en la nube simplifica el desarrollo y despliegue de aplicaciones, permitiendo acceder a recursos en cuestión de minutos, lo que acelera los procesos.
- Expandir la presencia global: la implementación de aplicaciones en la nube permite llegar a clientes en todo el mundo con baja latencia, independientemente de su ubicación geográfica, lo que amplía la presencia de la empresa de manera eficiente.

Para (Nishimiura, 2022), la computación en la nube ofrece tres modelos de servicio distintos. La primera es el Software como Servicio (SaaS), que ofrece aplicaciones administradas por el proveedor, lo que significa que los usuarios no tienen que preocuparse por la infraestructura ni la gestión de servicios, como sucede con servicios como Facebook o Dropbox. Luego, tenemos la Plataforma como Servicio (PaaS), donde los usuarios no controlan la infraestructura subyacente, pero pueden personalizar el entorno de ejecución de aplicaciones, como la personalización de las características de un blog en WordPress. Por último, encontramos la Infraestructura como Servicio (IaaS), que se refiere a la infraestructura informática física o virtual proporcionada por un proveedor de servicios en la nube. En este caso, los clientes tienen acceso a una variedad de recursos que pueden configurar y utilizar, como servicios de red, almacenamiento y servidores.

Cuando se trata de implementar la computación en la nube, existen tres modelos diferentes en función de dónde se ubique la infraestructura de TI: nube pública, nube

privada e híbrida. En el enfoque de nube pública, todas las partes de una aplicación o infraestructura se ejecutan en la nube, y los usuarios acceden a estos recursos a través de internet. Por otro lado, en la nube privada (también conocida como “on-premise”), todo se mantiene local, similar a la infraestructura tradicional de TI, utilizando servidores físicos o centros de datos, pero aprovechando la virtualización y la gestión de aplicaciones en la nube para optimizar los recursos de TI. Aunque no se obtienen todos los beneficios de la nube pública, la nube privada ofrece ventajas, como un mayor control de la seguridad, ya que los datos se mantienen en las instalaciones y no están expuestos a la internet pública. El tercer modelo es la nube híbrida, que combina elementos de los modelos anteriores al conectar infraestructura en la nube con recursos locales, como servidores físicos in situ (Nishimiura, 2022).

3.2.4.5. Amazon Web Services (AWS)

Según su página web, Amazon Web Services (AWS), fue lanzando en 2002, ofrece servicios tecnológicos bajo demanda a través de Internet con precios de pago por uso.

La nube de AWS abarca un amplio conjunto de productos globales basados en la nube que incluyen computación, almacenamiento, bases de datos, análisis, redes, dispositivos móviles, herramientas para desarrolladores, herramientas de administración, IoT, seguridad y aplicaciones empresariales.

Según el cuadrante de Gartner, como se muestra en la Figura 6 Amazon Web Services (AWS) es líder en servicios de infraestructura y plataformas (Amazon Web Services, 2023).

Figura 6. *Cuadrante Gartner.*

Figura 1: Magic Quadrant para servicios de infraestructura y plataforma en la nube



Nota. Tomado de Amazon Web Services. (2023). 2022 Gartner Magic Quadrant for Cloud Infrastructure & Platform Services. Tomado de <https://aws.amazon.com/es/resources/analyst-reports/22-global-gartner-mq-cips>

3.2.4.6. Amazon Simple Storage Service (Amazon S3)

Según lo señala (Daswani, 2021), Amazon S3 es una solución de almacenamiento de objetos en la nube que es escalable, resistente y rentable. Es adecuada para alojar una amplia variedad de datos y para diferentes casos de uso, que van desde contenido multimedia digital para sitios web, data lakes, aplicaciones móviles, datos de dispositivos IoT y análisis de big data.

De acuerdo a (Abbasi, 2021), Amazon S3 organiza la información en forma de objetos dentro de “buckets” . Un bucket es un contenedor, cuyo nombre es único a nivel global que puede albergar una cantidad ilimitada de objetos. Cada objeto tiene un identificador único o ID, y no puede existir sin estar asociado a un bucket específico. En un solo bucket se pueden almacenar diversos objetos, y cada objeto tiene un límite máximo de tamaño de 5 terabytes (TB).

3.2.4.7. AWS Identity and Access Management (IAM)

IAM es un servicio que permite controlar el acceso a los recursos de AWS de manera segura (Daswani, 2021).

Como lo describe (Abbasi, 2021), IAM permite crear y gestionar identidades como usuarios, grupos de usuarios y roles. Un usuario IAM, representan a una persona o aplicación que interactúa con AWS, y consiste de un nombre y credenciales. Un grupo IAM es una colección de usuarios, que desempeñan funciones similares o comparten roles comunes, usados para especificar los mismos permisos para múltiples usuarios. Un rol IAM es un conjunto de permisos que se pueden asignar a diferentes entidades tanto usuarios como a los mismos servicios para que puedan realizar acciones específicas en tu cuenta de AWS.

IAM también permite definir los permisos, a través de la configuración de “políticas”, que definen lo que una identidad puede o no puede hacer dentro de la cuenta de AWS.

3.2.4.8. Amazon Virtual Private Cloud (VPC)

Amazon VPC (Virtual Private Cloud) es una red privada virtual en la nube. Como lo menciona Daswani (2021), permite a los clientes construir múltiples redes privadas y seguras dentro de las cuentas de AWS. Permite configurar subredes, tablas de ruteo y configuraciones de seguridad.

Dentro de esta VPC, se pueden desplegar otros recursos de AWS, como por ejemplo una base de datos o una instancia de EC2.

3.2.4.9. Red Privada Virtual(VPN)

En una plataforma como AWS, se tiene la opción de conectar una VPC a una red corporativa, mediante una VPN, o Red Privada Virtual. Una VPN establece un túnel cifrado y seguro entre dos puntos finales a través de internet público. Utiliza una encriptación de Internet Protocol Security (IPsec) de 128 o 256 bits, lo que significa que puedes transferir datos de forma segura entre los dos puntos finales (Daswani, 2021).

Para configurar una conexión VPN entre la red corporativa y la VPC que se ha configurado en AWS, se necesita crear una puerta de enlace virtual privada (VPG) y asociarla a la VPC. Mientras que, en la red corporativa, se deberá configurar una puerta de enlace del cliente, que es un dispositivo físico o virtual que se conecta a la VPG a través de internet (Daswani, 2021).

3.2.4.10. Amazon DynamoDB

Amazon DynamoDB, es un servicio de base de datos NoSQL. DynamoDB es un servicio serverless (solución sin servidor) porque no es necesario definir ninguna configuración de instancia de base de datos, como la configuración de CPU o memoria y es AWS que se encarga de provisionamiento de recursos (almacenamiento e infraestructura), escalabilidad, respaldo, actualización de software, escalabilidad y monitoreo (Abbasi , 2021).

3.2.4.11. Amazon Simple Notification Service (SNS)

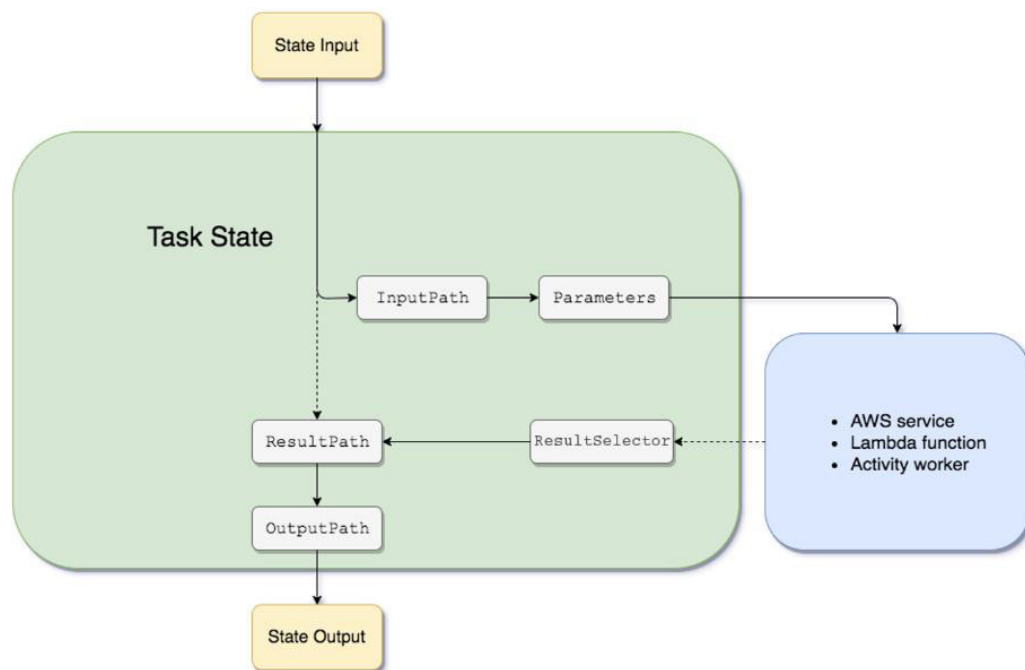
Amazon SNS es un servicio de mensajería y notificaciones basado en el envío de mensajes, que como lo menciona (Daswani, 2021), se basa en un modelo de publicación/suscripción, de modo que un componente de una aplicación, que actúa como el emisor de los mensajes, envíe mensajes a otros componentes de la aplicación o a usuarios finales, los cuales actúan como suscriptores. Pero antes de poder enviar mensajes a los suscriptores, es necesario crear un tema o tópico, que actúa como una canal de comunicación. Los editores deben saber a qué tema de SNS enviar los mensajes, y los clientes finales deben suscribirse al tema para poder recibir esos mensajes.

3.2.4.12. Amazon Step Functions

Amazon Step Functions, según la descripción de es un servicio serverless, que facilita la orquestación de múltiples servicios de AWS, usando flujos de trabajo visuales. Abbasi (2021)

Como lo describe Daswani (2021), Amazon Step Functions están escritos en el Lenguaje de Estados de Amazon (ASL), que es un lenguaje estructurado en formato JSON, que ayudan a definir las máquinas de estado o “state machine” en inglés que no es más que una serie de pasos llamados estados, cada estado realiza decisiones basados en el input, realiza alguna acción y produce un output a otros estados, tal como se muestra en la Figura 7.

Figura 7. *Funcionamiento de una máquina de estados en Step Functions*



Nota. Tomado de <https://catalog.workshops.aws/stepfunctions/en-US/introduction/input-output>

3.2.4.13. Amazon Athena

Amazon Athena, como lo define Daswani (2021), es una solución serverless que permite realizar consultas interactivas y analizar datos de Amazon S3 utilizando SQL estándar. Amazon Athena utiliza Presto, que es un motor de consultas SQL de código abierto diseñado para permitir análisis ad hoc.

Amazon Athena es compatible con una amplia variedad de formatos de datos, incluyendo CSV, JSON, ORC, Avro y Parquet, y permite presentar datos no estructurados, semiestructurados y estructurados.

3.2.4.14. Amazon Glue

Abbasi (2021) define, Amazon Glue como un servicio serverless de extracción, transformación y carga (ETL) que te permite extraer la data y metadata de múltiples fuentes.

AWS Glue proporciona para desarrollar los ETL o jobs como se le conocen en este entorno, tanto interfaces visuales y basadas en código, y ejecutarlos de manera completamente gestionada dentro de un entorno de shell de Python o en un entorno Spark serverless completamente gestionado. (Abbasi , 2021)

Amazon Glue incluye un servicio de rastreo, o “crawler” por su nombre en inglés, para explorar varios repositorios, clasificar datos e inferir información de esquema, como su formato y tipos de datos. Los metadatos se almacenan luego como tablas en el “Catálogo de Datos”, que es un repositorio central de metadatos (Daswani, 2021).

3.2.4.15. AWS Lambda

AWS Lambda es un servicio serverless, que permite ejecutar código y realizar tareas, por lo que se le conoce como una solución de Function as a Service (FaaS) (Daswani, 2021).

Para usar AWS Lambda, se debe crear una función utilizando un lenguaje de programación compatible, como Python, Node.js, Java o .NET.

Una función de AWS Lambda se puede activada en un horario programado o se debe definir un desencadenador que active la función y se ejecute el código como, por

ejemplo, se puede crear una función de Lambda que se active cuando cargue una imagen en un bucket de Amazon S3 (Daswani, 2021).

3.2.4.16. Amazon CloudWatch

Amazon CloudWatch permite monitorear los servicios y aplicaciones de AWS en tiempo real, se recopilan y rastrean datos de rendimiento y operativos (Daswani, 2021), ya que todos los recursos envían automáticamente sus métricas a CloudWatch (Piper & Clinton, 2023).

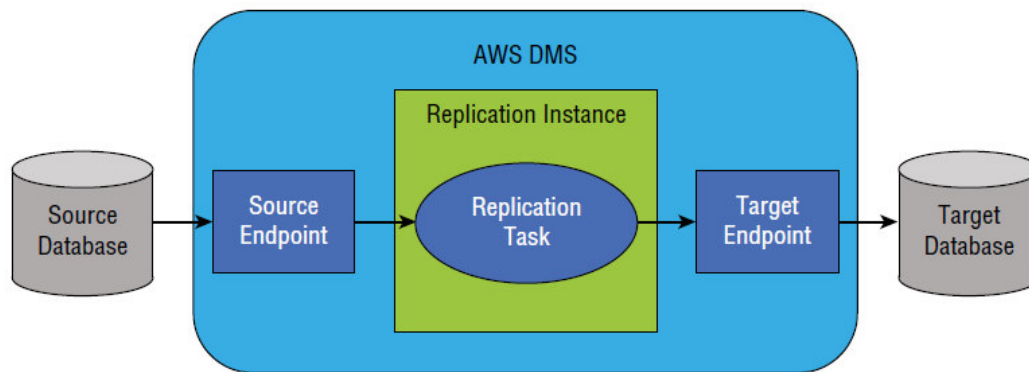
Amazon CloudWatch se puede utilizar para configurar alarmas, de modo que, si esas métricas superan ciertos umbrales durante un período especificado, se puede generar una alarma para poder tomar medidas correctivas (Daswani, 2021).

3.2.4.17. Amazon Data Migration Service (AWS DMS)

DMS es un servicio que facilita la migración de base de datos relacionales, NoSQL, entre otros. Permite realizar migraciones únicas y replicar los cambios en curso para mantener las fuente y el destino sincronizados. (Abbasi, 2021).

La Figura 8, muestra cómo funciona AWS DMS, donde es necesario definir los endpoints de origen y destino, que representan las base de datos de origen y destino respectivamente. En este proceso, se utiliza una instancia de replicación, la cual es una instancia de EC2. Esta instancia se conecta a la base de datos de origen, lee los datos y los formatea para que la base de datos de destino pueda consumirlos, además también carga los datos en la base de datos de destino. La mayoría del procesamiento ocurre en memoria, aunque transacciones de mayor volumen pueden requerir almacenamiento en disco (Amazon Web Services, 2023).

Figura 8. *AWS Database Migration Service*



Nota. Tomado de Abbasi A (2021). AWS Certified Data Analytics Study Guide Specialty (DAS-C01) Exam.

3.2.4.18. Amazon EventBridge

Amazon EventBridge es un servicio de bus de eventos *serverless* que permite recibir, transformar, enrutar y entregar eventos entre los servicios de AWS u otras aplicaciones. (Amazon Web Services, 2023). Un evento puede generarse cuando hay un cambio de estado en un recurso específico como, por ejemplo, cuando una instancia de EC2 pasa de estar en ejecución a detenida (Daswani, 2021).

Con EventBridge, es posible establecer reglas que definen patrones de coincidencia para los eventos entrantes. Cuando se produce un evento que cumple con los criterios definidos por una regla, este puede ser dirigido hacia un destino específico para un procesamiento adicional. Por ejemplo, si se detecta que un servidor crítico se detiene accidentalmente, el evento correspondiente puede enviarse automáticamente a una función de Lambda para reiniciar la instancia (Daswani, 2021).

EventBridge también permite la configuración de eventos programados, lo que implica la activación de eventos en un horario predefinido (Daswani, 2021). Por ejemplo, se puede programar el inicio de una instancia de EC2 de lunes a viernes a las 8 a.m.

3.2.4.19. Canvas LMS

Canvas LMS es un Sistema de Gestión de Aprendizaje de código abierto, desarrollada por Instructure en 2011, empleada como herramienta educativa por escuelas, colegios y universidades para administrar y recursos e interacciones de aprendizaje en línea (Instructure, 2023).

Entre las características esenciales de Canvas LMS (Instructure, 2023) se incluyen:

- Administración de cursos: los profesores pueden crear, organizar y gestionar el contenido de sus cursos, como materiales de lectura, tareas, exámenes y discusiones en línea.
- Comunicación: facilita la comunicación entre docentes y estudiantes mediante herramientas como mensajes internos, foros de discusión y notificaciones.
- Evaluación y seguimiento: los profesores pueden diseñar y calificar tareas y exámenes en línea, además de llevar un registro del progreso y desempeño de los estudiantes.
- Colaboración: permite la colaboración en línea a través de herramientas como grupos de estudiantes, videoconferencias y wikis.
- Personalización: los docentes y administradores pueden personalizar la apariencia y estructura de sus cursos para adaptarlos a sus necesidades y preferencias.
- Integraciones: admite una variedad de integraciones con otras herramientas y servicios, lo que facilita la conexión con sistemas de gestión de contenido, bibliotecas digitales, sistemas de registro y más.

3.2.4.20. Zoho CRM

Zoho CRM es un software de CRM de ventas en línea que permite gestionar las ventas, el marketing y el soporte en una sola plataforma de CRM, desarrollada por Zoho Corporation (Zoho Corporation, 2023).

CRM se refiere a un conjunto de herramientas y aplicaciones diseñadas para ayudar a las empresas a gestionar y optimizar sus relaciones con los clientes, lo que incluye la administración de contactos, ventas, marketing, servicio al cliente y otros aspectos relacionados con la interacción con los clientes (Zoho Corporation, 2023) .

Algunas de las características y funcionalidades que ofrece Zoho CRM incluyen (Zoho Corporation, 2023):

- Gestión de procesos
- Gestión de clientes potenciales y contactos.
- Automatización de ventas.
- Automatización de marketing y seguimiento de campañas.
- Gestión de casos y soporte al cliente.
- Integración con otras aplicaciones y servicios, como correo electrónico, calendario y redes sociales.
- Informes y paneles personalizables.

3.2.5. Implementación de las áreas, procesos y sistemas

La implementación del Data Lake en AWS en el área de arquitectura de datos de la universidad se llevó a cabo en colaboración con un proveedor, partner de AWS. Este proveedor fue contratado para brindar orientación sobre las mejores prácticas a seguir durante el desarrollo del proyecto. El proceso de implementación se dividió en las siguientes etapas:

3.2.5.1. Planificación y análisis

Durante la etapa de planificación del proyecto, en colaboración con el partner, se definieron los siguientes objetivos clave:

- Ingestión desde Oracle: trasladar las tablas relacionadas con los reportes de recaudación y matrícula desde la fuente de datos relacional de Oracle a la nube de

AWS. Esto proporcionará beneficios significativos en términos de rendimiento, durabilidad y escalabilidad.

- Implementación de stored procedures en AWS Glue: transformar los stored procedures en jobs ejecutables con AWS Glue. Esto nos permitirá ejecutarlos de forma serverless, bajo demanda y de manera eficiente.
- Caso de uso de Zoho CRM y Canvas: abordar los casos de uso relacionados con Zoho CRM y Canvas de manera similar a los proyectos anteriores.

El equipo responsable del proyecto estuvo conformado por los siguientes roles, descritos en la Tabla 6:

Tabla 6. *Roles involucrados en el proyecto*

Rol	Descripción
Product Owner	Encargado de la toma de decisiones sobre el alcance y características del producto y colabora con los usuarios del negocio para entender sus necesidades.
Project Manager	Encargado de la planificación y coordinación de todas las actividades del proyecto, así también se encarga de la supervisión del progreso de las actividades.
Arquitecto	Encargado de colaborar con el Product Owner para alinear la arquitectura diseñada con los objetivos del negocio.
Líder Técnico	Encargado de supervisar la ejecución de las tareas técnicas, así como de proporcionar orientación técnica al equipo de desarrollo.
Lead Developer	Encargado de liderar el equipo de desarrollo y asegura la coherencia y calidad de la implementación.
Data Engineer	Encargado de la ingesta de datos y el desarrollo de los ETL.

Nota. Fuente: Elaboración propia

Para la implementación del proyecto se estableció dividir las implementaciones por tipo de fuente generando de esta manera tres casos de usos a desarrollar.

Tabla 7. *Casos de uso*

Código	Descripción	Actor Principal
CU-01	Ingesta y procesamiento desde Oracle one-premise	Equipo de Proyecto
CU-02	Ingesta y procesamiento de Zoho CRM	Equipo de Proyecto
CU-03	Ingesta y procesamiento de Canvas	Equipo de Proyecto

Además, se definió una duración proyectada para cada uno de los casos de uso, lo que nos proporcionó una línea de tiempo clara y ayudó a gestionar eficazmente los recursos y las expectativas.

Figura 9. *Gantt referencial del caso de uso de ingesta desde Oracle*

Ingesta y Analítica	S1 (24/07-31/07)*	S2 (01/08-07/08)	S3 (08/08-14/08)	S4 (15/08-21/08)	S5 (22/08-28/08)	S6 (29/08-05/09)*
Full load histórico de Oracle						
Cargas incrementales diarias de Oracle (Ingresos y Matricula)						
Implementación de stored procedures en AWS						
Pruebas y preparación para pase a producción						
Pase a producción						

Nota. Fuente: Universidad

Figura 10. Gantt referencial del caso de uso Zoho CRM y Canvas

Ingesta & Analítica - Zoho & Canvas	S1	S2	S3	S4	S5	S6
Cargas incrementales diarias Zoho						
Implementación de stored procedures Zoho en AWS						
Validaciones Zoho						
Cargas full de Canvas data						
Implementación de stored procedures en AWS						
Validaciones canvas						
Pruebas y preparación para pase a producción						
Pase a producción						

Nota. Fuente: Universidad

Una vez que se establecieron los objetivos y los casos de usos a desarrollar, el siguiente paso implicó la creación de un mapeo detallado de las fuentes de datos identificadas: bases de datos, API's externas o archivos adicionales.

Para el caso de uso CU-01, se recopiló información detallada que abarcaba el nombre de las tablas, el esquema al que pertenecen, la estrategia de carga utilizada (total o incremental), las claves primarias aplicadas en las tablas, así como el nombre y tipo de datos de las columnas que funcionarán como filtros en caso de cargas incrementales. El Anexo 1 proporciona el detalle de esta información.

En el caso de uso CU-02, se definió detalles fundamentales que abordan la relación entre los elementos dentro de la plataforma y sus equivalentes en la API. Los detalles que incluyeron fueron el nombre y la sección del módulo, así como el nombre del campo según el registro en Zoho CRM y su correspondencia en la API, como se muestra en el Anexo 2.

En esta instancia, se llegó al acuerdo para que todas las tablas se carguen de forma incremental. La razón detrás de esta elección es que todos los módulos de Zoho CRM cuentan con columnas por defecto como la de fecha de creación (Created_Time) y fecha de modificación (Modified_Time) para cada registro. Esto se basa en las consultas previas que hemos realizado utilizando la API, como se muestra en la Figura 11.

Figura 11. Muestra previa de los datos del módulo Leads

	Modified_Time	Email	WIN_CONTACT_SRCH	WIN_PRODUCT_SRCH	Created_Time	WIN_SUB_TIPIFICACION	WIN_DEAL_SRCH	First_Name	Lead_S
0	2023-10-09T17:49:26-05:00	abarbarandaza@gmail.com	None	{'id': '5267437000026558059'}	2023-09-02T16:47:11-05:00	-None-	None	Ana	En G
1	2023-10-15T00:30:28-05:00	miguelowada@gmail.com	None	{'id': '5267437000026606097'}	2023-09-02T17:50:11-05:00	Seguimiento	None	Miguel Angel Severo	gestio
2	2023-10-09T17:49:25-05:00	jackelinemontoya15@gmail.com	None	{'id': '5267437000026558509'}	2023-09-02T22:57:47-05:00	-None-	None	Jackeline	En G
3	2023-10-09T17:49:25-05:00	soniajulca9@gmail.com	None	{'id': '5267437000026558059'}	2023-09-03T04:53:52-05:00	-None-	None	Ceilia Sonia	En G
4	2023-10-09T17:49:25-05:00	aliagamartinezliz@gmail.com	None	{'id': '5267437000026558109'}	2023-09-03T09:31:11-05:00	-None-	None	Mar	En G

Nota. Fuente: Elaboración propia

En el casos de uso CU-03, hemos considerado la documentación disponible en su sitio web oficial como fuente de información clave el análisis. En la Figura 12, se presenta un documento de acceso público que detalla el mapeo de las tablas utilizados en la plataforma.

Figura 12. Mapeo de esquemas de Canvas

[PUBLIC]: CD1 to CD2 schema mapping ☆ ⓘ ☁					
Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda					
100% Solo ver					
A1	GO BACK TO MASTER				
	A	B	C	D	E
1	GO BACK TO MASTER				
2	CD1 table	CD1 field	CD2 table	CD2 field	Comments
3	account_dim	id	accounts	id	shard_id * 10000000000000 + accounts.id as id
4	account_dim	canvas_id	accounts	id	
5	account_dim	name	accounts	name	
6	account_dim	depth			calculated field
7	account_dim	workflow_state	accounts	workflow_state	
8	account_dim	parent_account	accounts	name	where accounts.parent_account_id = ac
9	account_dim	parent_account_id	accounts	parent_account_id	CD2: The account's parent ID, or NULL
10	account_dim	grandparent_account	accounts	name	
11	account_dim	grandparent_account_id	accounts	id	CD2 doesn't store the id of grandparent get the parent_account_ids to build the account.
12	account_dim	root_account	accounts	name	
13	account_dim	root_account_id	accounts	id	CD2 doesn't store the id of grandparent get the parent_account_ids to build the account.
14	account_dim	subaccount1			Account hierarchy is no longer stored in accounts.parent_account_id.
15	account_dim	subaccount1_id			
16	account_dim	subaccount2			
17	account_dim	subaccount2_id			
18	account_dim	subaccount3			
19	account_dim	subaccount3_id			
20	account_dim	subaccount4			

Nota. La figura muestra el mapeo de las tablas de la API, realizando una comparación con la primera versión de la API. Tomado de <https://docs.google.com/spreadsheets/d/1kqCXAD9K45L0QeEtbuuMAFp2fW8o0oC8EBzJf58SjrY/edit#gid=527545099>

Posterior al mapeo de las fuentes, se continuo con el análisis de los servicios de AWS, y el porqué de su razón de elección con respecto a otros servicios que ofrece este proveedor de servicios.

Tabla 8. *Servicios usados.*

Servicio	Descripción	Ventajas	Desventajas	Razón de elección
Amazon S3	Almacenamiento en la nube.	Escalabilidad, durabilidad	Costos de almacenamiento o.	Además de las ventajas presentadas, es elegido por su integración con otros servicios como AWS Glue y Amazon Athena.
AWS Glue	Servicio de ETL en la nube.	Automatización, integración con AWS.	Costos de ejecución.	Elegido por su capacidad de automatizar el proceso ETL y se elegio sobre AWS EMR debido a la simplicidad requerida en el proyecto.
AWS DMS	Servicio de migración de bases de datos	Simplifica la migración de datos. desde Oracle a Amazon S3.	Limitado en procesamiento de transformación de datos.	AWS DMS se prefirió debido a su integración nativa con Oracle y la necesidad de una migración simple.
Amazon VPN	Red privada virtual (VPN) para acceder de forma	Seguridad de datos, conectividad,	Configuración y costos de	Elegida para garantizar una conexión segura

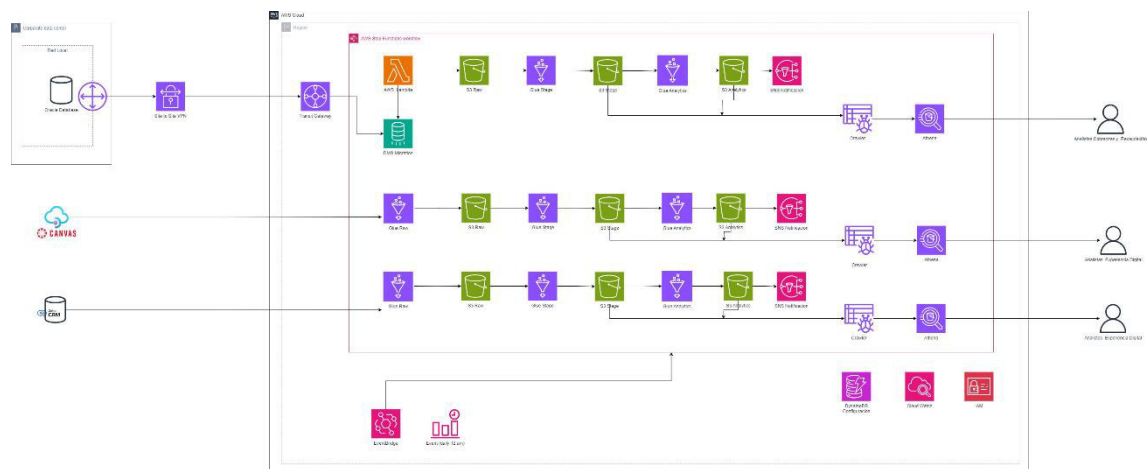
	segura a recursos en la nube.	segura, acceso remoto a recursos en la nube.	mantenimiento de VPN.	entre la red y los recursos en AWS.
Amazon DynamoDB	Base de datos NoSQL utilizada para almacenar configuraciones y metadatos específicos.	Almacenamiento de configuración, escalabilidad, sin administración de servidores.	Costos asociados al uso de DynamoDB.	Se seleccionó para gestionar y almacenar configuraciones y metadatos relacionados con la implementación del Data Lake. Tuvo un uso específico: DynamoDB se utiliza exclusivamente como base de datos de configuración.
Amazon Athena	Servicio de análisis interactivo en datos en Amazon S3.	Consultas SQL en datos almacenados en Amazon S3 sin necesidad de ETL.	Costos asociados al uso de consultas.	Elegido para consultas SQL interactivas en datos de alto rendimiento en Amazon S3.
AWS Lambda	Funciones serverless	Escalabilidad y rápida implementación	Existen límites en la cantidad de recursos (tiempo de ejecución, memoria)	El pago por uso, es decir pagas por el tiempo de ejecución. Esto es más rentable para que las llamadas a DMS son de una vez por día.

Nota. Fuente: Elaboración propia

3.2.5.2. Diseño

Utilizando el mapeo de los servicios seleccionados, hemos desarrollado un diseño de arquitectura de alto nivel que ilustra el flujo del proceso de ingesta para los tres casos de uso. Este diseño se presenta en la Figura 13.

Figura 13. *Arquitectura propuesta*



Nota. Fuente: Elaboración propia

3.2.5.3. Implementación técnica

En esta etapa se realizaron varias tareas esenciales para el funcionamiento y desarrollo exitoso del Data Lake en AWS.

A continuación, se detallan las configuraciones implementadas, aplicables a los tres casos de uso.

a) Creación de usuarios, roles y configuración de políticas

Esta etapa es crucial para la gestión de la seguridad y el acceso en la plataforma de AWS.

Para el desarrollo, se definieron usuarios de tipo administradores, de modo que estos tengan acceso a todos los servicios, así como usuarios con accesos a servicios específicos, por lo que se dividieron en 3 grupos de usuarios específicos.

- Administradores: acceso a todos los servicios

Figura 14. *Política establecida para el grupo de usuarios Administradores*

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "*",
      "Resource": "*"
    }
  ]
}
```

Nota. Fuente: Universidad

- Desarrolladores: con accesos a servicios como Amazon S3, AWS Glue, Amazon Athena, Lambda, Step Functions, SNS, DynamoDB. Pueden crear, configurar, y gestionar recursos dentro de estos servicios.

Figura 15. *Política establecida para el grupo de usuarios Desarrolladores*

```
"Version": "2012-10-17",
"Statement": [
  {
    "Effect": "Allow",
    "Action": [
      "s3:*",
      "glue:*",
      "athena:*",
      "lambda:*",
      "states:*",
      "sns:*",
      "dynamodb:*"
    ],
    "Resource": "*"
  }
]
```

Nota. Fuente: Universidad

- Analistas: con accesos solo Amazon S3 y Amazon Athena, pero no tienen permitido realizar acciones que modifiquen los datos o la estructura de las tablas.

Figura 16. Política establecida para el grupo de usuarios Analistas

```
{
  "Effect": "Allow",
  "Action": [
    "s3:GetObject",
    "s3:ListBucket"
  ],
  "Resource": [
    "arn:aws:s3::",
    "arn:aws:s3:::"
  ]
},
{
  "Effect": "Allow",
  "Action": [
    "athena:StartQueryExecution",
    "athena:GetQueryExecution",
    "athena:GetQueryResults",
    "athena:GetWorkGroup"
  ],
  "Resource": "*"
}
]
```

b) Estructuración de los buckets en S3

Una vez establecida la conectividad, los usuarios y permisos, se procedió a crear los buckets en Amazon S3. Se estableció que los buckets que se crearan estarán divididos por tipo de entorno ya se de desarrollo (dev) y producción (prd).Estos a su vez se organizaran según los estándares en tres capas raw, stage y analytics, lo que facilita la gestión eficiente de nuestros datos a lo largo de su ciclo de vida.

Los nombres que se estableció para nombrar los buckets fue como se muestra en la Figura 17.

Figura 17. Nomenclatura de los nombre de los buckets

datalakeingestion	-prd-	<organizacion>	prd	raw	bucket/	<caso de uso/fuente>
	-dev-		dev	stage analytics		

Nota. Fuente: Elaboración propia

c) Configuración de la tabla de control en DynamoDB

Se ha creado una tabla en DynamoDB que actúa como una tabla de control, la cual será usada para los procesos de ingesta para los tres casos de uso a implementar. Esta tabla es esencial para definir aspectos clave del proceso de la carga de datos, tales como la selección de las tablas específicas a cargar, la fuente de estos datos, y el tipo de carga a realizar, ya sea total o incremental. Además, esta tabla de control facilita la especificación de las columnas específicas a incluir en la carga, entre otros detalles cruciales del proceso.

La Tabla 9 describe las columnas de la tabla de DynamoDB creada.

Tabla 9. Descripción de las columnas de la tabla en DynamoDB

Columna	Descripción
	Nombre de la que describe la fuente
TARGET_TABLE_NAME	
	Cuando se establece en Y (Si), habilita la carga de datos para esta fuente, si se coloca N no se habilita su carga
ACTIVE_FLAG	
	Cuando se establece en true, se activará el crawler para esta fuente
CRAWLER	
	Fecha de fin de carga
END_VALUE	
	Cuando se genera algún error en la carga de la fuente se registra el error relacionado
FAIL_REASON	
	El nombre de la columna en la fuente que se utiliza para filtrar registros delta.
FILTER_COLUMN	
	Un parámetro puede tener uno de los siguientes valores:
	lte – carga total
	between – carga incremental
FILTER_OPERATOR	

	Lista de columnas de la fuente a considerar en la carga
ID_COLUMN	
	Fecha de la última carga realizada de la fuente
LAST_UPDATE_DATE	
	Esquema de la fuente
SOURCE_SCHEMA	
	Nombre de la tabla a cargar
SOURCE_TABLE	
	Fecha de inicio de carga
START_VALUE	

Nota. Fuente: Elaboración propia

Una vez configurado lo mencionado anteriormente, se procedió a desarrollar los componentes necesarios por cada caso de uso establecido.

a) Configuración de la VPC

Para el proyecto se utilizó la VPC que está configurada por defecto en AWS. La VPC está configurada con el bloque CIDR 172.31.0.0/16, que define el rango de direcciones IP disponibles dentro de la VPC.

Define por defecto 6 subredes creadas dentro de esta VPC, cada una en una zona de disponibilidad diferente a la región US East (Norte de Virginia), como us-east-1a, us-east-1b, etc.

Figura 18. VPC configurado



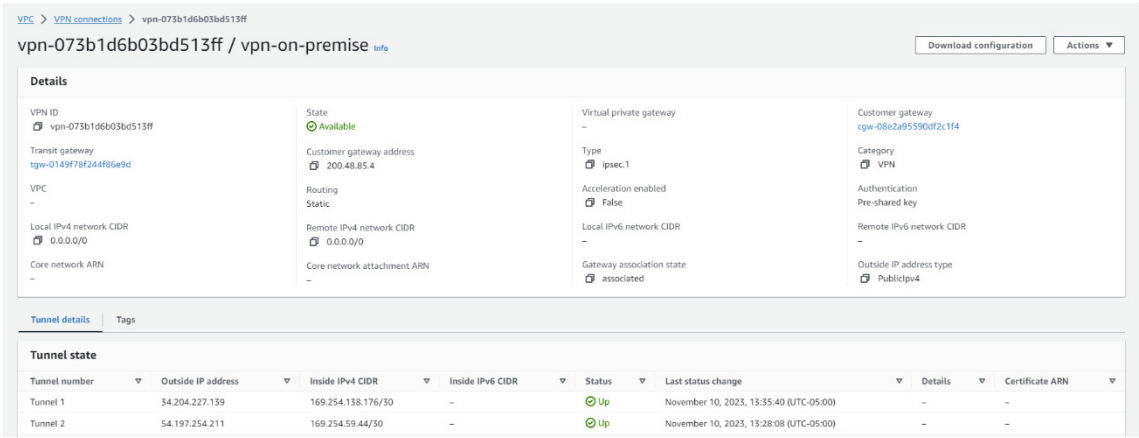
Nota. Fuente: Universidad

b) Configuración de la VPN Site to Site

Utilizamos AWS Site-to-Site VPN para establecer una conexión segura entre el datacenter de la universidad y la VPC en AWS.

Se configuro el Customer Gateway y el Transit Gateway en AWS y se conectó estos con el equipo de VPN en el datacenter de la universidad.

Figura 19. VPN configurada



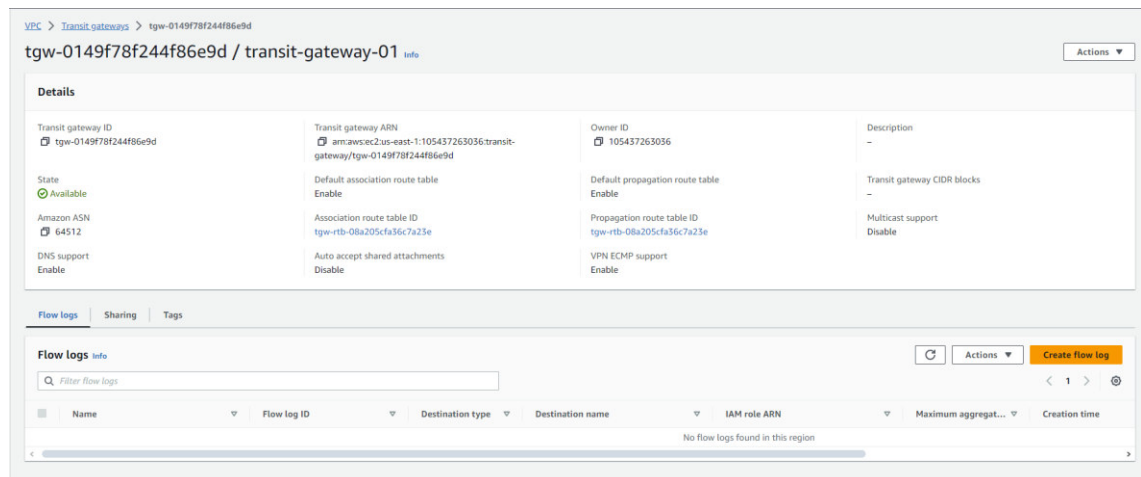
Nota. Fuente: Universidad

Figura 20. Customer Gateway configurado



Nota. Fuente: Universidad

Figura 21. *Transit Gateway configurado*



Nota. Fuente: Universidad

- c) Desarrollo del flujo de orquestación usando funciones Lambda para la ingesta mediante la creación de instancia de replicación de DMS.

Para desarrollar este flujo se utiliza la tabla de DynamoDB que contiene información de la configuración para crear dinámicamente las tareas de AWS DMS. Estas tareas se utilizan para cargar cambios incrementales basados en la columna de filtro especificada (por ejemplo, la FECHA_MODIFICACION del registro). Los valores en esta columna de filtro se utilizan para limitar el número de registros transferidos desde la fuente al destino.

El flujo de trabajo, como muestra en el Anexo 3 se muestra el detalle del flujo configurado mediante un step function, activa una serie de funciones Lambda, en lenguaje Python, para crear recursos y gestionar tareas de AWS DMS y cargar los datos a Amazon S3 en el bucket de Raw. Una vez completada la carga, se una función Lambda actualiza la tabla de DynamoDB con la estadísticas de la carga, incluyendo el último valor para la columna de filtro. Al final del flujo, los recursos de AWS DMS se eliminan.

- d) Desarrollo de los jobs de raw, stage y analytics

El siguiente paso crítico en la implementación fue el desarrollo de los jobs en AWS Glue, en lenguaje PySpark.

Para los casos de uso CU-02 y CU-03, se tuvo que realizar el desarrollo de jobs de tipo raw, que se encargaron de realizar la conexión a su respectiva API, y cargar en formato parquet la información que se puede obtener en el bucket Raw, según sea el tipo de carga.

Los jobs de Stage, se encarga 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 Stage

Los jobs de Analytics, se encarga de aplicar las reglas de negocio sobre los datos ya curados almacenados en el bucket Stage, 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 Analytics.

e) Configuración del flujo de orquestación

Se configura los componentes que se han descrito en los puntos e) y d) en un solo flujo de orquestación mediante el uso del servicio de step functions, de modo que los tres casos de uso se ejecuten de manera paralela.

Al final de cada subflujo se configura el envío de una notificación esta notificación indica si cada subflujo ha finalizado de manera correcta, o en su contrario si ha existido algún error en alguna parte del flujo.

f) Acceso a Datos a través de Amazon Athena

En nuestra configuración, los datos procesados y almacenados en los buckets Stage y Analytics han sido configurados para ser accesibles mediante Amazon Athena. Para lograr esto, se han implementado tres crawlers distintos dentro del servicio de Glue. Cada uno de estos crawlers está diseñado para interactuar con los datos correspondientes a un caso de uso específico.

La forma en que se organizan los datos en Amazon Athena se pueden ver en la Figura 22.

Figura 22. Vista ejemplo de la visualización de los datos en Amazon Athena

Query 1 : X Query 2 : X

1 SELECT * FROM "wiener-prd-stage-database-prod01"."up100_alumno" limit 10;

SQL Ln 1, Col 1

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results Query stats

Completed Time in queue: 163 ms Run time: 4.726 sec Data scanned: 14.96 MB

Results (10) Copy Download results

Search rows

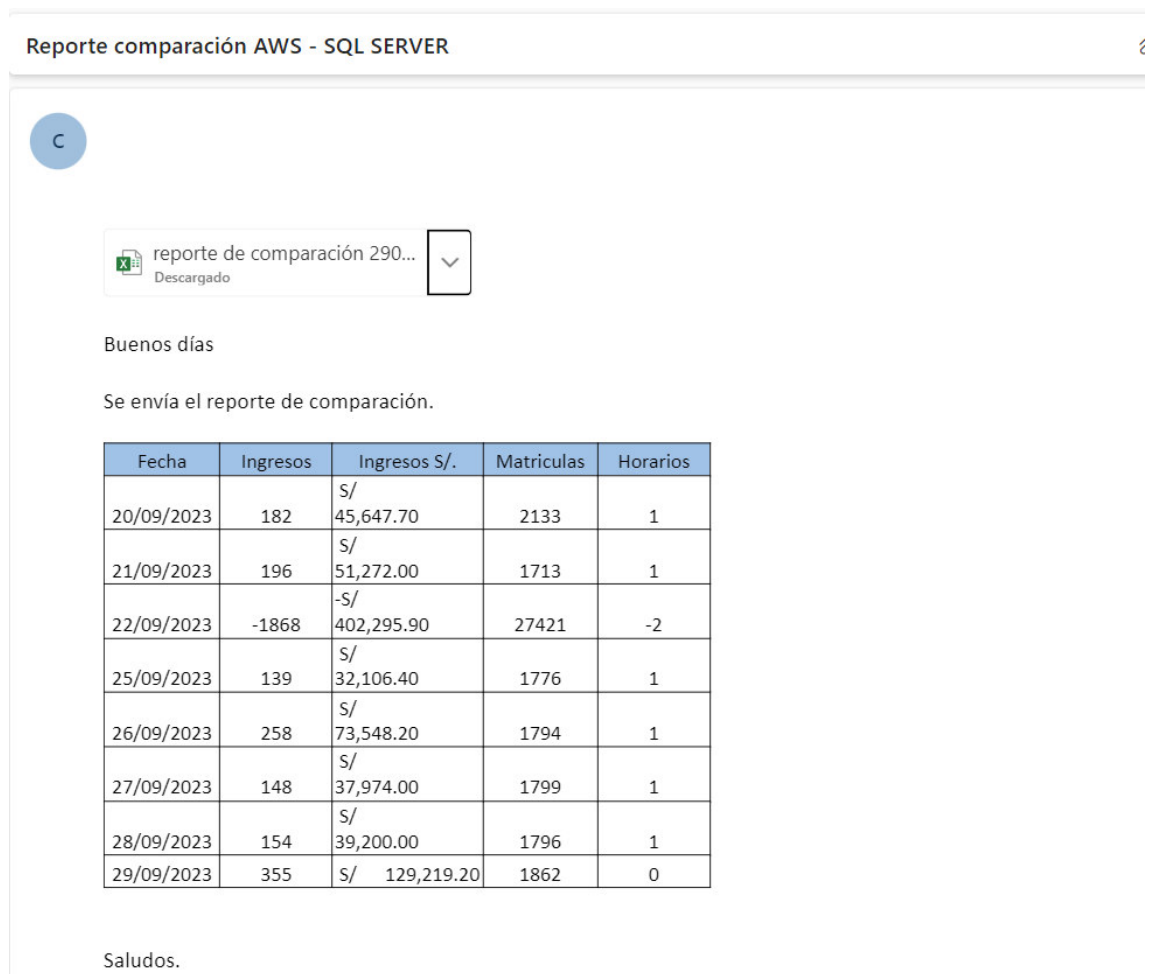
#	alumno	lugar_nacimiento	lugar_residencia	direccion_residencia	colegio	fecha_nacimiento	nacionalidad	nombres
1	7705114ET							DELIA
2	8500303SE							ROSA
3	8407869ET							MARIA
4	8500308SE							ELIZABETH
5	8315734ET							NANCY
6	C07AP03363		140137	LOS MINERALES 596	0336891	1982-11-01 00:00:00.000 UTC		JACQUELINE
7	8500296SE							MARGARITA
8	W200722242		140104	JR CHAVIN # 16 2DO. PISO	0644880	1988-02-26 00:00:00.000 UTC		LUZ ROSEMI

Nota: Fuente: Elaboración propia

3.2.5.4. Pruebas y validación

Se realizaron las pruebas de validación, las que consistieron en la comparaciones entre las fuentes originales, que son los reportes anteriores desplegados a nivel local, con lo que lo que se había cargado en AWS. Para lo cual se realizaron validaciones diarias, las cuales se comunicaban mediante correo a los implicados.

Figura 23. Vista ejemplo de los correo enviados de las validaciones realizadas



Nota: Fuente: Universidad

3.2.5.5. Despliegue y monitoreo

Después de múltiples validaciones se realizó el despliegue o pase a producción de los flujos para que los usuarios puedan usar la data, pero solo los servicios de Athena por el momento.

3.3. Evaluación

3.3.1. Evaluación económica

La evaluación económica se centra en los costos del personal involucrado en el proyecto.

La Tabla 10, se muestra la distribución mensual de costos por cada uno de los roles que participaron en el proyecto, teniendo en cuenta el sueldo mensual para cada rol y el porcentaje de tiempo que cada rol lo dedico al proyecto.

Tabla 10. *Distribución mensual de costos*

Rol	Sueldo Mensual (S/.)	% de Participación	Costo mensual atribuible (S/.)	Agosto	Setiembre	Octubre	Noviembre	Total (S/.)
Product Owner	8000	50%	4000	4000	4000	4000	4000	16000
Project Manager	7500	100%	7500	7500	7500	7500	7500	30000
Arquitecto	7000	75%	5250	5250	5250	5250	5250	21000
Líder Técnico (x2)	5000	100%	10000	10000	10000	10000	10000	40000
Lead Developer	7500	100%	7500	7500	7500	7500	7500	30000
Data Engineer	6000	100%	6000	6000	6000	6000	6000	24000

Nota. Fuente: Elaboración propia.

En total fueron 7 los integrantes del proyecto que trabajaron durante los 4 meses de duración del proyecto, siendo así la inversión del capital humano un total de S/. 161 000.

3.3.2. Beneficios para la organización

La implementación del Data Lake en AWS proporciono una serie de beneficios significativos para la universidad. En primer lugar, se logró una mayor agilidad y eficiencia en la gestión de datos. Al aprovechar la nube de AWS y las capacidades de AWS Glue, se ha reducido el tiempo y los recursos necesarios para la ingesta, transformación y almacenamiento de datos. Esto significa que podemos responder más rápidamente a las cambiantes demandas de datos y a las necesidades de análisis.

Además, la implementación de crawlers y Amazon Athena ha mejorado la accesibilidad y la capacidad de búsqueda de los datos. Ahora, los equipos pueden explorar y consultar datos de manera rápida y eficiente utilizando SQL estándar, lo que facilita la toma de decisiones informadas. Esto ha tenido un impacto directo en la capacidad para identificar oportunidades, detectar tendencias y tomar decisiones basadas en datos con mayor rapidez.

La seguridad y la escalabilidad también son factores clave. Se ha implementado políticas de seguridad, que garantizan la protección de los datos en todo momento. Además, la arquitectura de la implementación es escalable, lo que permite crecer a medida que las necesidades de datos evolucionan.

En resumen, la implementación de Data Lake en AWS ha permitido optimizar la gestión de datos, mejorar el acceso a la información y fortalecer la seguridad, lo que se traduce en una toma de decisiones más rápida y precisa, un mayor rendimiento y una mayor flexibilidad para adaptarse a los desafíos cambiantes del mercado.

CAPÍTULO IV: REFLEXIÓN CRÍTICA DE LA EXPERIENCIA

La implementación del Data Lake en AWS ha sido una experiencia enriquecedora para la universidad y para la profesional que ha presentado el siguiente informe. Durante este proceso, se ha extraído valiosas lecciones sobre la gestión de datos en la nube, la importancia de contar con una arquitectura escalable y segura, y la necesidad de estar preparados para abordar desafíos imprevistos. En particular, se ha reconocido el valor de la colaboración interdisciplinaria y la comunicación efectiva entre equipos.

Durante la etapa de implementación, surgieron retrabajos debido a actualizaciones no anticipadas en la lógica establecida, lo que extendió el tiempo y generó costos adicionales para el proyecto. Además, en la etapa de validación de lo implementado, nos enfrentamos a necesidades no satisfechas debido a la falta de directrices bien definidas. Esto resultó en reuniones repetidas para abordar estos desafíos.

Se ha observado también que la adopción de nuevas tecnologías y procesos puede implicar un cambio cultural en la organización, ya que se adoptan prácticas y enfoques de trabajo novedosos. La capacitación continua y el apoyo a los empleados son esenciales para garantizar una transición sin problemas.

CAPITULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

1. A pesar de los desafíos encontrados, la implementación del Data Lake en AWS ha sido un éxito, en la universidad. Se logro crear una infraestructura escalable para la gestión de datos, lo que nos permite aprovechar el potencial de la información.
2. La colaboración interdisciplinaria y la comunicación efectiva son clave en los proyectos. La interacción constante entre los equipos de TI y de negocio ha fortalecido nuestro enfoque y resolución de problemas.
3. La implementación de Data Lake ha otorgado una mayor agilidad en la gestión de datos. Ahora se puede adaptar rápidamente a las cambiantes demandas de datos y a las necesidades de análisis.
4. La adopción de nuevas tecnologías y procesos puede requerir un cambio cultural dentro de la organización. La capacitación y el apoyo continuo son esenciales para asegurar una transición efectiva.

5.2. Recomendaciones

1. Definir directrices y requerimientos concretos antes de la implementación es esencial. Esto evita retrabajos y garantiza que las necesidades se satisfagan desde el principio.
2. Continuar invirtiendo en la capacitación de los empleados es fundamental. Asegurarse de que todos los miembros del equipo estén familiarizados con las nuevas tecnologías y procesos es esencial.
3. La monitorización constante y la gestión proactiva de problemas son esenciales para garantizar la integridad de los datos y el rendimiento óptimo. Esto debe mantenerse como una práctica estándar.
4. Mantener una documentación detallada de la arquitectura y los procesos del Data Lake es fundamental. Esto facilita la resolución de problemas y futuros proyectos.
5. Realizar evaluaciones regulares para medir el rendimiento y la eficiencia de la implementación. Identificar áreas de mejora y oportunidades para la optimización.

5.3. FUENTES DE INFORMACIÓN

- Nambiar, A., Mundra, D. , (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. Big Data Cogn. Comput. 6, 132.
- Pulidori, M., (2021). Design and Implementation of an application Data Lake to support Repair Process in Baker Hughes. Master's Degree Thesis.
- Leo, Y., (2017). Data Warehouse with Big Data Technology for Higher Education. Procedia Computer Science 124, 93–99.
- Palanivel,K., Suresh, J., (2020). Data Lake Model to Modern Educational Organizations. International Research Journal of Engineering and Technology (IRJET), 269-276
- Nishimura, H., (2022). AWS Community Hero - AWS for Non-Engineers Prepare for the AWS Certified Cloud Practitioner Exam Version 3. Manning Publications.
- Abbasi, A., (2021).AWS Certified Data Analytics Study Guide. John Wiley & Sons, Inc.,
- Daswani, R., (2021). AWS Certified Cloud Practitioner Exam Guide. Packt Publishing.
- Helttula, M. (2023). Enabling data-driven decision-making for a Finnish SME: a data lake solution. Master of Science in Technology Thesis. University of Turku.
- Zoho Corporation. (2023). Zoho CRM. Recuperado de <https://www.zoho.com/crm>
- Zoho Corporation. (2023). Características - Zoho CRM. <https://www.zoho.com/es-xl/crm/features.html>
- Instructure, (2023). What is Canvas. Recuperado de <https://community.canvaslms.com/t5/Canvas-Basics-Guide/What-is-Canvas/ta-p/45>
- Amazon Web Services. (2023). Amazon EventBridge. <https://aws.amazon.com/es/eventbridge>

- Amazon Web Services. (2023). 2022 Gartner Magic Quadrant for Cloud Infrastructure & Platform Services. Amazon Web Services. <https://aws.amazon.com/es/resources/analyst-reports/22-global-gartner-mq-cips>
- Amazon Web Services. (2023). Input and Output - Step Functions. Amazon Web Services. <https://catalog.workshops.aws/stepfunctions/en-US/introduction/input-output>
- Amazon Web Services. (2023). Instancias de replicación. AWS Documentation. https://docs.aws.amazon.com/es_es/dms/latest/userguide/CHAP_ReplicationInstance.html

ANEXOS

Anexo 1: Mapeo de tablas de la fuente de base de datos Oracle

ESQUEMA	TABLA	LOAD STRATEGY	PRIMARY KEY COLUMN(S)	FILTER 1	TIPO DE DATO
ui100	alumno_carrera_periodo	INCREMENTAL	PERIODO,MES,CARRERA,ALUMNO	FECHA_MODIFICACION	DATETIME
ui100	seccion_periodo	INCREMENTAL	SECCION,PERIODO	FECHA_MODIFICACION	DATETIME
ui100	seccion	INCREMENTAL	SECCION	FECHA_MODIFICACION	DATETIME
up100	alumno	INCREMENTAL	ALUMNO	FECHA_MODIFICACION	DATETIME
up100	ubigeo	FULL	PAIS,UBIGEO		
up100	carrera	FULL	CARRERA		
up100	alumno_periodo	INCREMENTAL	ALUMNO,CARRERA,INSTITUCION,MES,PERIODO,PKCODIGO	FECHA_MODIFICACION	DATETIME
up100	seccion_periodo	INCREMENTAL	PERIODO,SECCION	FECHA_MODIFICACION	DATETIME
up100	seccion	INCREMENTAL	SECCION	FECHA_MODIFICACION	DATETIME
up100	ed_gradostudios	INCREMENTAL	CODIGOGRADO	FECHAACTUALIZACION	DATETIME
up100	atencion_informes	INCREMENTAL	ATENCION,CARRERA,CLIENTE,INSTITUCION,MES,MODALIDAD,PERIODO,PKCODIGO	FECHA_MODIFICACION	DATETIME
wiener	co_cobranza	INCREMENTAL	UNIDADREPLICACION,COBRANZANUMERO	ULTIMAFECHAMODIF	DATETIME
wiener	co_documento_cobranza	FULL	COMPANIASOCIO,TIPODOCUMENTO,NUMERODOCUMENTO,UNIDADREPLICACION,COBRANZANUMERO,		
wiener	co_cobranza_detalle	INCREMENTAL	UNIDADREPLICACION,COBRANZANUMERO,LINEA	ULTIMAFECHAMODIF	DATETIME

wiener	co_tipopago	INCREMENTAL	TIPOPAGO	ULTIMAFECHAMODIF	DATETIME
wiener	banco	INCREMENTAL	BANCO	ULTIMAFECHAMODIF	DATETIME
wiener	co_documento	INCREMENTAL	COMPANIASOCIO, TIPODOCUMENTO, NUMERODOCUMENTO	ULTIMAFECHAMODIF	DATETIME
wiener	co_documento_detalle	INCREMENTAL	COMPANIASOCIO, TIPODOCUMENTO, NUMERODOCUMENTO, LINEA	ULTIMAFECHAMODIF	DATETIME
wiener	ac_sucursal	INCREMENTAL	SUCURSAL	ULTIMAFECHAMODIF	DATETIME
wiener	co_servicioclasificacion	INCREMENTAL	SERVICIOCLASIFICACION	ULTIMAFECHAMODIF	DATETIME
wiener	wh_itemmast	INCREMENTAL	ITEM	ULTIMAFECHAMODIF	DATETIME
wiener	co_serviciogrupo	INCREMENTAL	GRUPO	ULTIMAFECHAMODIF	DATETIME
wiener	accountmst	INCREMENTAL	ACCOUNT, PRIME, ELEMENT,	LASTDATE	DATETIME
wiener	co_conceptofacturacion	INCREMENTAL	CONCEPTOFACTURACION	ULTIMAFECHAMODIF	DATETIME
wiener	ac_costcentermst	INCREMENTAL	COSTCENTER	LASTDATE	DATETIME
wiener	ac_costcentergroup	INCREMENTAL	COSTCENTERGROUP	LASTDATE	DATETIME
wiener	ac_costcentersubgroup	INCREMENTAL	COSTCENTERGROUP, COSTCENTERSUBGROUP	LASTDATE	DATETIME
wiener	ac_costcenterclasificacion	INCREMENTAL	COSTCENTERCLASIFICACION	LASTDATE	DATETIME
wiener	ac_costcentermajorgroup	INCREMENTAL	COSTCENTERMAJORGROUP	LASTDATE	DATETIME
up100	ed_asistencia_detalle	FULL			
up100	alumno_curso_seccion	FULL	ALUMNO, CARRERA, CURSO, INSTITUCION, MES, PERIODO, SECCION		

up100	postulante	FULL	CARRERA,INSTITUCION,MES,NUMERO_ ADMISION,PERIODO,PKCODIGO,POSTU LANTE
up100	modalidad	FULL	MODALIDAD
up100	sub_modali dad	FULL	SUB_MODALIDAD
up100	alumno_carr era	FULL	ALUMNO,CARRERA,INSTITUCION,PKCO DIGO
up100	area	FULL	AREA
up100	tabla_gener al	FULL	CODIGO,TABLA
up100	institucion_ procedencia	FULL	INSTITUCION
up100	ed_modalid adestudio	FULL	CODIGOGRADO,CODIGOMODALIDAD,C ODIGOPROGRAMA,DESCRIPCION,INSTI TUCION,PKCODIGO
up100	ed_gradotip omodalidad	FULL	CODIGOMODALIDAD
up100	carrera_grad otipomodal idad	FULL	

Anexo 2: Mapeo de módulos de Zoho CRM

MÓDULO	SECCIÓN DEL MÓDULO	CAMPO ZOHO CRM	NOMBRE DE API
Posibles Clientes	Relacionamiento	Campaña	WIN_CAMPAIGN_SRCH
		Contacto	WIN_CONTACT_SRCH
		Producto	WIN_PRODUCT_SRCH
		Ficha de Datos URL	WIN_FICHA_DATOS_URL
		Programa Académico	WIN_CARRERA_INTERES_FINAL
		Oportunidad	WIN_DEAL_SRCH
		Curso	WIN_CURSO_LKUP
		Propietario de Posible Cliente	Owner
	Datos personales	Nombre	First_Name
		Apellidos	Last_Name
		Tipo de documento	WIN_TIPO_DOC
		Fecha de nacimiento	WIN_FECHA_NACIMIENTO
		Número de documento	WIN_NUM_DOC
		Edad	WIN_EDAD
		Estado Civil	WIN_ESTADO_CIVIL
		Género	WIN_GENERO
		Grado académico	WIN_GRADO
		Categoría de pago	WIN_CATEGORIA_PAGO
		Tipo de persona pagante	WIN_TIPO_PERSONA_PAGANTE
		Ocupación	WIN_OCUPACION
	Datos de contacto	Móvil	Mobile
		Teléfono	Phone
		Correo electrónico	Email
		Otro teléfono	WIN_OTRO_TELEFONO
	Información de interés	Estado de Posible cliente	Lead_Status
		Fuente de Posible cliente	Lead_Source
		Carrera de Interés	WIN_CARRERA_INTERES
		Fase de Posible Cliente	WIN_FASE
		Carrera de Interés	WIN_CARRERA_INTERES
		Modalidad	WIN_MODALIDAD
		Código de Carrera	WIN_CODIGO_CARRERA
		Motivo	WIN_MOTIVO

	Tipo de programa	WIN_TIPO_PROGRAMA
	Motivo	WIN_MOTIVO
	Tipo de creación	Tipo_de_creaci_n
	¿Estas interesado en este inicio de clases?	Estas_interesado_en_este_inicio_de_clases
	Código de alumno	C_digo_de_alumno
	Fase Previa No Gestionable	Fase_Previa_No_Gestionable
Información de inscripción	Sección	WIN_SECCION_LKUP
	Turno	WIN_TURN0
	Periodo de inscripción	WIN_PERIOD0_SECCION
	Tipo de comprobante de pagante	WIN_TIPO_COMPROBANTE_PAGANTE
	Código SIGU Tipo persona pagante	WIN_CODIGO_SIGU_TIPO_PERSONA_PAGANTE
	Sede	WIN_SEDE
	Código SIGU Turno	WIN_CODIGO_SIGU_TURN0
	Mes de inscripción	WIN_MES_SECCION
Residencia	Departamento	Departamento
	Provincia	Provincia
	País de residencia	WIN_PAIS_RESIDENCIA
	Dirección	Direcci_n
	Dirección referencia	WIN_DIRECCION_REFERENCIA
Lugar de nacimiento	País de nacimiento	WIN_PAIS_NACIMIENTO
	Distrito de nacimiento2	WIN_DISTRITO_NACIMIENTO
	Nacionalidad	WIN_NACIONALIDAD
	Ciudad de nacimiento	WIN_CIU0AD_NACIMIENTO
Institución	Institución	Instituci_n1
	Nombre de institución	Instituci_n
Bases recicladas	Fase Original	Fase_Original
	Código de alumno bd	C_digo_de_alumno_bd
	Fuente Original	Fuente_Original
Motivos / Sub tipificación /Medio contacto	Medio de contacto	WIN_MEDIO_CONTACTO
	Motivo de No Gestionable	WIN_MOTIVO_NO_GESTIONABLE
	Motivo de Lead Perdido	WIN_MOTIVO_PERDIDO
	Sub Motivo Lead Perdido	WIN_SUBMOTIVO_PERDIDO
UTMs	Plataforma	Plataforma

	Formato Source	Formato_Source
	utm_source	WIN_UTM_SOURCE
	utm_campaign	WIN_UTM_CAMPAIGN
	utm_medium	WIN_UTM_MEDIUM
	utm_term	WIN_UTM_TERM
	utm_content	WIN_UTM_CONTENT
URLs	Landing	WIN_ORIGEN
Datos del Padre o Apoderado	Nombres	WIN_FIRST_APOD
	Apellido	WIN_LAST_APOD
	Celular	WIN_MOBILE_APOD
	Correo	WIN_EMAIL_APOD
	Tipo documento	WIN_TIP_DOC_APOD
	Número de documento	WIN_NUM_DOC
	Parentesco de apoderado	WIN_PARENTESCO_APOD
	Dirección de apoderado	WIN_DIRECCION_APOD
Políticas	Política de privacidad	WIN_POLITICA
	Acepta recibir información	WIN_RGPD
Datos del sistema	Hora de creación	Created_Time
	Creado por	Created_By
Historial de fases	Fase de Posible Cliente	WIN_HF_FASE
	Motivo/Medio/Tipificación	WIN_HF_MOTIVO
	Duración (Min)	WIN_HF_DURACION
	Tipificado por	WIN_HF_TIPIFICADO_POR
	Hora de modificación	WIN_HF_HORA_MODIFICACION
Historial Propietarios	Propietario inicial	WIN_OWNER_INICIAL_LKUP
	Propietario final	WIN_OWNER_FINAL_LKUP
	Fase	WIN_FASE
	Motivo de cambio	WIN_MOTIVO_CAMBIO
	Hora de modificación	WIN_HF_HORA_MODIFICACION
	Duración propietario inicial	WIN_DURACION_OWNER
	Modificado por	Modified_By
Ficha de datos	Completó Ficha Carrión	WIN_COMPLETO_FICHA_CARRION
	Estado Ficha Carrión	WIN_ESTADO_FICHA_CARRION
	ZCreator Ficha ID	WIN_ZCREATOR_FICHA_ID
	Última Oportunidad Perdida ID	WIN_ULTIMO_DEAL_PERDIDO_ID

Woztell	Woztell Member Id	Member_Id
	Woztell Fecha Último Mensaje	WOZTELL_LAST_MESSAGE_DATETIME
	Woztell Member Id Whatsapp	WIN_WOZTELL_MEMBER_ID_WHATSAPP
	Woztell Fecha Último Mensaje Whastapp	WIN_WOZTELL_FECHA_LTIMO_MENSAJE_WHASTAPP
	Woztell Member Id Instagram	WIN_WOZTELL_MEMBER_ID_INSTAGRAM
	Woztell Fecha Último Mensaje Instagram	WIN_WOZTELL_FECHA_LTIMO_MENSAJE_INSTAGRAM
	Woztell Member Id Facebook	WIN_WOZTELL_MEMBER_ID_FACEBOOK
	Woztell Fecha Último Mensaje Facebook	WIN_WOZTELL_FECHA_ULTIMO_MENSAJE_FACEBOOK
	Woztell Channel Type	WOZTELL_CHANNEL_TYPE
	Woztell Identificación externa	Identificaci_n_externa
	Día de creación	F_rmula_4
	Woztell Channel Id	WIN_WOZTELL_CHANNEL_ID
	Woztell Message Timestamp Inicial	WiN_WOZTELL_MESSAGE_TIMESTAMP_INICIAL
	Woztell Message Template Nuevo	WIN_WOZTELL_MESSAGE_TEMPLATE_NUEVO
Auditoría	Creado por	Created_By
	Modificado por	Modified_By
	Diseño	Layout
	Ex-propietario-concat	Ex_propietario_concat
	Ex-propietario	Ex_propietario
	Hablar con humano	Hablar_con_humano
	Channel ID	Channel_ID
	Detalle	Detalle
	Channel ID Whatsapp	WIN_CHANNEL_ID_WHATSAPP
	Regla de Asignación Ejecutada	WIN_ASSIGNMENT_RULE_EXECUTED
	Channel ID Facebook	WIN_CHANNEL_ID_FACEBOOK
	Channel ID Instagram	WIN_CHANNEL_ID_INSTAGRAM
	Ex Fase	WIN_EX_FASE
	Tipo de producto	WIN_TIPO_PRODUCTO

	Número de intentos de llamadas	WIN_NUMERO_INTENTOS_CALLS
	Equipo	Equipo
	IC - Regla de asignación nivel 1	WIN_ASSIGNMENT_RULE_LEVEL_1
	IC - Regla de asignación nivel 2	WIN_ASSIGNMENT_RULE_LEVEL_2
	Ex Fase - Fase Actual	WIN_EX_FASE_FASE_ACTUAL
	ZCRM Origen reasignación	ZCRM_ORIGEN_REASIGNACION
	ZCliq Último Chat ID	ZCLIQ_LAST_CHAT_ID
	IC - APIs consumidos	WIN_APIS_CONSUMED
	Creacion	Creacion
Cálculo de metas	Fecha - Metas	Fecha_Nuevo_Base_Reciclada
	Mes - Meta	WIN_MES_METAS
	Fase - Metas	Fase_Nuevo_Base_de_datos
Campos Performlead	Ejecutivo_pl	ejecutivo
	utm_source_pl	utm_source_pl
	utm_medium_pl	utm_medium_pl
	Campus	Campus
	Flujo_pl	Flujo_pl
	Etapa_pl	Etapa_pl
	Origin_pl	Origin
	utm_campaign_pl	utm_campaign_pl
	Distrito_pl	Distrito_pl
	ce_pl	ce
	Formulario_pl	Formulario_pl
	Estado_pl	Estado_pl
	fecha hora creacion pl	fecha_hora_creacion_pl

Anexo 3: Flujo de orquestación

