# UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

*La Universidad Católica de Loja*

## FACULTAD DE INGENIERÍAS Y ARQUITECTURA

### MAESTRÍA EN CIENCIAS Y TEGNOLOGÍAS DE LA COMPUTACIÓN

# Propuesta de una Arquitectura Lago de Datos Académicos, Caso de Estudio UTPL

Tesis previo a la obtención del título de:

## MAGÍSTER EN CIENCIAS Y TECNOLOGÍAS DE LA COMPUTACIÓN

**Autor**: Guevara Rivas Elvia Digna

**Director**: Piedra Pullaguari Nelson Oswaldo

LOJA

2022



*2022*

**Aprobación del director de tesis**

Loja, 4, de marzo, de 2022

Doctor
Rommel Vicente Torres Tandazo.
**Director de la maestría de Ciencias y Tecnologías de la Computación**

Ciudad.-

De mi consideración:

Me permito comunicar que, en calidad de director de la presente tesis denominado: Propuesta Arquitectura Lago de Datos Académico realizado por Elvia Digna Guevara Rivas completos ha sido orientado y revisado durante su ejecución, así mismo ha sido verificado a través de la herramienta de similitud académica institucional, y cuenta con un porcentaje de coincidencia aceptable. En virtud de ello, y por considerar que el mismo cumple con todos los parámetros establecidos por la Universidad, doy mi aprobación a fin de continuar con el proceso académico correspondiente.

Particular que comunico para los fines pertinentes.

Atentamente,

Nelson Oswaldo Piedra Pullaguari, Doctor
C.I.: 1102809462
Correo electrónico: nopiedra@utpl.edu.ec

**Declaración de autoría y cesión de derechos**

Yo, Elvia Digna Guevara Rivas, declaro y acepto en forma expresa lo siguiente:

Ser autora de la tesis denominado: Propuesta de una Arquitectura Lago de Datos Académicos, Caso de Estudio UTPL, de la maestría de Ciencias y Tecnologías de la Información específicamente de los contenidos comprendidos en: (se debe colocar los nombres de los capítulos elaborados en la tesis), siendo (nombres y apellidos completos), director (a) del presente trabajo; también declaro que la presente investigación no vulnera derechos de terceros ni utiliza fraudulentamente obras preexistentes. Además, ratifico que las ideas, criterios, opiniones, procedimientos y resultados vertidos en el presente trabajo investigativo, son de mi exclusiva responsabilidad. Eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones judiciales o administrativas, en relación a la propiedad intelectual de este trabajo.

Que la presente obra, producto de mis actividades académicas y de investigación, forma parte del patrimonio de la Universidad Técnica Particular de Loja, de conformidad con el artículo 20, literal j), de la Ley Orgánica de Educación Superior; y, artículo 91 del Estatuto Orgánico de la UTPL, que establece: “Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se realicen a través, o con el apoyo financiero, académico o institucional (operativo) de la Universidad”, en tal virtud, cedo a favor de la Universidad Técnica Particular de Loja la titularidad de los derechos patrimoniales que me corresponden en calidad de autor/a, de forma incondicional, completa, exclusiva y por todo el tiempo de su vigencia.

La Universidad Técnica Particular de Loja queda facultada para ingresar el presente trabajo al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública, en cumplimiento del artículo 144 de la Ley Orgánica de Educación Superior.

.....

Autora: Elvia Digna Guevara Rivas

C.I.: 1103773931

Correo electrónico: edguevara@utpl.edu.ec

## Dedicatoria

Dedico con profundo amor y reconocimiento esta investigación en primer lugar a mis padres, quienes me dieron la oportunidad de ser una profesional y forjaron en mí, amor y responsabilidad al diario devenir de la vida y de mis metas.

A mis hijos, quienes han alegrado mi vida desde su existencia y diariamente me brindan su amor y comprensión incondicional, impulsándome a seguir con mis sueños en la vida profesional.

Le dedico con mucho cariño este trabajo, a mi esposo, quien ha sido mi apoyo y visión para alcanzar nuevas habilidades en el desarrollo profesional.

## Agradecimiento

Agradezco primeramente a Dios, por darme una vida llena de bendiciones y ser mi guía para forjarme en el futuro.

A mis padres, hijos y esposo por ser una familia que han depositado en mi vida alegría y confianza en mis nuevos retos.

A mi tutor de Teisis PhD. Nelson Oswaldo Piedra Pullaguari, quien me ha brindado sus conocimientos, confianza y tiempo, para guiarme el presente trabajo de investigación.

Al Director de la Maestría de Ciencias y Tecnologías de la Computación, PhD. Rommel Vicente Torres Tandazo y profesores quienes me han compartido sus conocimientos en el transcurso de la carrera.

# Índice de Contenido

## Índice de Tablas



## Índice de Figuras

**Resumen**

El trabajo de investigación tiene como finalidad, hacer una “Propuesta de una Arquitectura de Lago de Datos Académico, Caso de Estudio UTPL”, esta Institución de Educación Superior tiene como finalidad garantizar el correcto cumplimiento de sus objetivos en sus líneas estratégicas, las fuentes de datos que maneja en la actualidad, son de origen académico en sus diferentes sedes, carreras, y modalidades. En la actualidad las Instituciones Educativas tienen diferentes tipos de datos denominados datos heterogéneos, la infraestructura de almacenamiento tradicional no es la más la adecuada para el manejo de estos datos, por lo que se necesita que sean tratados con diferentes tecnologías que se adapten a los retos que conlleve las necesidades actuales de almacenamiento. Se propuso dar respuesta a la variedad de datos Académicos de La Universidad Técnica, Particular de Loja, provenientes de diferentes fuentes, con tecnologías Big Data. Esto implica tener centralizada la información, reduciendo así los diferentes silos de información y dar un valor unificado de los datos para la gestión en la analítica como objetivo principal en la implementación de una ALDA. En la actualidad en la literatura científica no existe ninguna arquitectura definida, como ALDA, lo cual permite analizar las diferentes propuestas y considerar cual puede ser la más acorde de acuerdo a los requisitos de los Datos Académicos, como requisitos LD. El aporte final es proponer una solución para la gestión de Datos Académicos Heterogéneos en la Gestión de procesos de Ingesta, Almacenamiento, Procesamiento, Acceso, Visualización, como funciones principales, así como su gobernabilidad en el manejo de metadatos.

*Palabras clave*: Big Data Arquitectura, Lago de Datos, Arquitectura Lago de Datos

**Abstract**

The purpose of the research work is to make a "Proposal of an Academic Data Lake de Architecture, UTPL Case Study", this Higher Education Institution has the purpose of guaranteeing the correct fulfillment of its objectives in its strategic lines, the data sources that manages today, are of academic origin in their different locations, careers, and modalities. Currently Educational Institutions have different types of data called heterogeneous data, the traditional storage infrastructure is no longer adequate for handling these data, so they need to be treated with different technologies that adapt to the challenges that the current storage needs. It was proposed to respond to the variety of Academic data of UTPL, coming from different sources, with Big Data technologies. This implies having the information centralized, thus reducing the different silos of information and giving a unified value of the data for management in analytics as the main objective in the implementation of an ADLA. Currently in the scientific literature there is no defined architecture, as ADL, which allows analyzing the different proposals and considering which may be the most consistent according to the requirements of the Academic Data, as LD requirements. The final contribution is to propose a solution for the management of Heterogeneous Academic Data in the Management of processes of Intake, Storage, Processing, Access, Visualization, as main functions, as well as its governance in the management of metadata.

Keywords: Big Data Architecture, Data Lake, Data Lake Architecture

## Introducción

En la actualidad las Instituciones Educativas dentro de sus clasificaciones manejan diferentes tipos de datos, pueden provenir de diferentes fuentes y puede ser; datos estructurados, semiestructurados y no estructurados, tanto internos como externos.

La Universidad Técnica Particular de Loja en adelante UTPL, según el Plan Estratégico de Desarrollo Institucional 2020-2025(PEDI), dentro de su direccionamiento estratégico en líneas y objetivos, acoge como una condición institucional, el objetivo de potenciar en la comunidad universitaria las competencias relacionadas con la transformación digital.

La UTPL contiene diferentes sedes a nivel local regional, nacional, internacional, contiene 23 carreras en su Modalidad Presencial, 17 en su Modalidad Abierta y a Distancia, 5 posgrados vigentes, la planta docente contiene 1070 profesores, hasta la actualidad se han graduado más de 65.000 profesionales.

Los datos académicos, surgen del componente considerado como áreas de dominio, en los mecanismos de implementación de la estrategia del PEDI, los cuales requieren un mejor manejo y gestión adecuada.

Las tecnologías tradicionales de almacenamiento de la información no son la mejor solución para el manejo de grandes volúmenes de datos variados, las Arquitecturas Big Data se dan en base a las fuentes de datos, orientada a manejar volúmenes de diferentes tipos, se requiere de cambios en la infraestructura de almacenamiento de la información en la mayoría de empresas.

Se considera un concepto emergente Lago de Datos, como alternativa adicional al manejo de resultados de diferentes tipos de datos académicos ya que en la actualidad según se manifiesta en el PEDI, se maneja con un sistema informático.

El análisis de datos que conlleva a los resultados en el manejo de un Lago de Datos Académico, se lo realiza mediante profesionales del área técnica como Científicos de Datos.

El principal objetivo de esta investigación es la propuesta de una “Arquitectura Lago de Datos Académico”(ALDA).

Se propone dar una solución basándose en los requisitos de los diferentes tipos de datos Académicos que proceden de variadas fuentes; mediante la realización de las diferentes fases para identificar con claridad la mejor forma de ingerir, procesar, almacenar, presentar este tipo de datos.

La propuesta se basa en soluciones de tecnologías de Big Data que permite el manejo de datos estructurados, no estructurados y semiestructurados.

En la primera fase se realiza un estudio de Revisión Sistemática de Literatura de todos los trabajos relacionados con lo que propone la ciencia en “Arquitectura Lago de Datos”, este estudio permite dar respuesta a la pregunta de investigación que se plantea en este apartado.

En la segunda fase se plantea la problemática mediante un contexto, alcance, problema y objetivos.

La tercera fase describe la metodología que contiene los siguientes procesos: 1) Análisis de varios modelos de arquitecturas Lago de Datos. 2) Evaluación de las diversas arquitecturas en base a los diferentes casos de estudio en un caso de uso. 3) Selección de un prototipo.

En la cuarta fase se desarrolla el diseño del prototipo ALDA, como parte de la solución del problema, para la gestión de distintas fuentes, el proceso de ingestión, almacenamiento, procesamiento, acceso, presentación, gobernabilidad, protección de identidad, acceso a usuarios.

En la quinta fase se manifiesta lo que se aprendió y la importancia de dar una solución basada en el arte de investigar lo cual se espera que represente un aporte valioso para la UTPL, en el manejo de los datos heterogéneos y transformación digital que se ofrece en las conclusiones y recomendaciones.

# Capítulo uno

# Metodología

## 1.1 Revisión sistemática de la literatura

Se realiza la Revisión Sistemática de Literatura(RSL), que comprende los siguientes ítems: Realización del mentefacto, un modelo de script de búsqueda, preguntas de investigación, modelo protocolo de revisión, modelo de (RSL), listado de revistas, modelo de resultado de búsquedas, finalmente todo esto incluye el análisis del estado del arte, estos ítems de formatos desarrollado por (Torres et al., 2018) según los autores esto aplica a una adaptación del método de (Kitchenham, 2004), que divide el proceso en tres subpartes: planificación, realización y presentación de informes de resultados, para la elaboración del estado del arte.

### *1.1.1 Mentefacto*

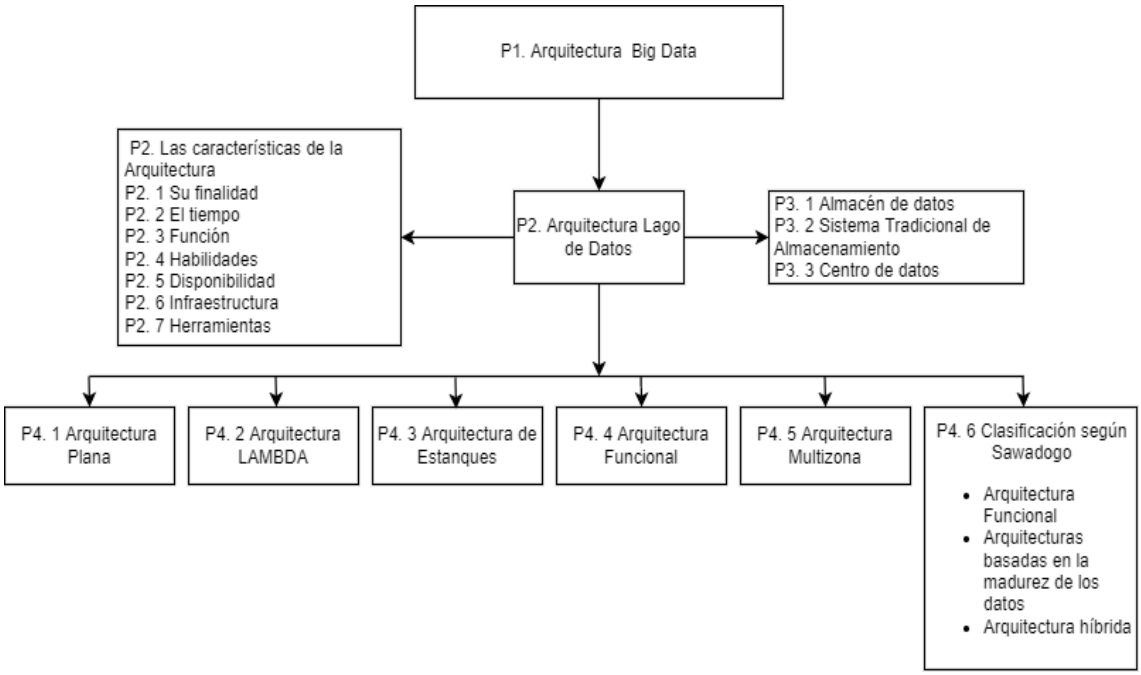
El mentefacto fue propuesto por (de Zubiría Ragó & de Zubiría Samper, 2019), lo cual permite ver la dimensión del problema de acuerdo a los grupos de análisis: isoordinados, supraordinados, excluidos, infraordinados.

Los isoordinados son las características, atributos y funciones que muestran las esencialidades que lo define al objeto principal, los supraordinados son de donde provienen el objeto principal se incluyen dentro del grupo de origen, los excluidos son aquellos que no pertenece al grupo del concepto principal aunque comparten ciertas similitudes, los infraordinados son parte del concepto principal que especifica las clases y los subtipos del grupo.

Mediante esta conceptualización se realiza el mentefacto conceptual que permite investigar los grupos de pensamiento del caso del estudio “Propuesta de una Arquitectura Lago de Datos Académicos”.

**Figura 1**

*Mentefacto, caso de estudio Arquitectura Lago de Datos Académicos*




La elaboración del Mentefacto da originalidad al método y facilita la elaboración del tesauro para búsquedas y criterios de inclusión y exclusión (de Zubiría Ragó & de Zubiría Samper, 2019).

Para dar enfoque al grupo de pensamiento en el presente Trabajo de Titulación propuesto se basa es las siguientes preguntas propuestas por (de Zubiría Ragó & de Zubiría Samper, 2019) estas son: ¿Qué la caracteriza, en esencia?, ¿En qué grupo de cosas lo incluyen?, ¿Cuáles son sus diferencias con objetos similares? y, ¿hay subtipos?.

A partir de este andamiaje de preguntas el caso de estudio, responde a lo siguiente.

¿Qué la caracteriza, en esencia? .- Las características de una ALD aplicado a las líneas de producto de software enfoque tomado en (Marianne Huchard et al., 2020).

- ✓ Adquisición: obtener, duplicar y vincular
- ✓ Categorización: listado, escribir, referencia, localización, personalización, producción, indexación, evaluación de calidad, rastreo.
- ✓ Explotación: Preparar, enriquecer, agregar, etiquetar, reconocer patrones, clasificar, limpiar, conciliar, correlacionar, transformar, explorar / producir, navegar, describir,

computar patrones o reglas estadísticas, informar, segmentar, predecir, prescribir, inferir, consultar, liberación, gestión.

- ✓ Gestión del ciclo de vida: ejecutar, borrar por motivos técnicos, agregar, resumir, guardar, archivar.
- ✓ Aseguramiento: proteger, cifrar, gestionar la confidencialidad, anonimizar, auditar, cumplir.
- ✓ Almacenamiento: almacenamiento físico, almacenamiento virtual.

Conceptos de un LD, a continuación se describe los conceptos que presentan diferentes autores con respecto a un Lago de Datos.

**Tabla 1**

*Conceptos de un Data Lake*

| Autor | Definición |
|---|---|
| (Megdiche et al., 2020) | ✓ Un Lago de datos es una solución de análisis de Big Data que ingiere datos brutos estructurados heterogéneamente de varias fuentes.<br>✓ Almacena estos datos brutos en su formato nativo.<br>✓ Permite procesar datos de acuerdo con diferentes requisitos y<br>✓ Brinda acceso a los datos a diferentes usuarios.<br>✓ Gobierna los datos para garantizar la calidad de los datos, la seguridad de los datos y el ciclo de vida de los datos. |
| (P Sawadogo & Darmont, 2021) | " Un Lago de Datos es un sistema de análisis y almacenamiento escalable para datos de cualquier tipo, que se conservan en su formato nativo y se utilizan principalmente por especialistas en datos (estadísticos, científicos de datos o analistas) para la extracción de conocimientos" |
| (Couto et al., 2019) | "El Lago de Datos es un sistema de repositorio central para el almacenamiento, procesamiento y análisis de datos sin procesar, en el que los datos se mantienen en su formato original y se procesan para ser consultados solo cuando sea necesario". |

¿En qué grupo de cosas lo incluyen?.-Los sistemas de almacenamiento de datos en la actualidad han evolucionado por las características especiales que presentan los datos como es su volumen, variedad, velocidad, veracidad, viabilidad, visualización, valor, estos

datos ya no pueden ser tratados con sistemas tradicionales lo cual se requiere de nuevas tecnologías que nos trae la era del Big Data, a partir de las cuales se crean nuevas prácticas para dar mayor flexibilidad a la variedad de datos, según (Dixon, 2010) lo define como “Data Lake” termino traducido para este trabajo como Lago de Datos.

Con los Lago de Datos, las Instituciones Educativas pueden finalmente liberar el potencial de Big Data (Krishnan, 2020). Considerar que las empresas e instituciones pueden utilizar los Lagos de Datos para obtener una mejor visibilidad de los datos, eliminar los silos de datos y capturar vistas de 360 grados de los clientes(Krishnan, 2020).

¿Cuáles son sus diferencias con objetos similares?.- A continuación se describe en la Tabla 2 las diferencias entre objetos similares.

**Tabla 2**

*Diferencias entre objetos similares*

| | **Database** | **Data Mart(Top-down)** | **Data Warehouse** | **Data Lake** |
|---|---|---|---|---|
| **Source** | Single | Single | Múltiple | Múltiple |
| **Structure** | Structured | Structures | Structures | Raw |
| **Purpose** | Determined | Determined | Determined | Undetermined |
| **Storage** | Centralized | Decentralized | Centralized | Centralized |
| **Data Format** | Detailed | Summarized | Detailed | All |
| **Flexibility** | Low | Medium | Medium | High |
| **Primary Use** | Transactional | Reporting | Analytics & Reporting | Analytics |
| **Cost** | Low | Medium | Medium | High |
| **Data Volume** | Low | low | Medium | High |
| **Development** | Top-down | Bottom-up | Top-down | All |
| **Design Time** | Medium | Medium | Hugh | Low |
| **Volatility** | Medium | Low | None | Nome |
| **Data Operations** | CRUD | CR | CRU | CR |
| **Subject Area** | Single | Single | Múltiple | Múltiple |
| **Design Schema** | Relational | Multi-dimensional | Relational | No schema |

*Nota:* Adaptado de (Goli, 2020)

¿Subtipos? .-En la actualidad existen diferentes arquitecturas de Lago de Datos, según varios autores no existe ninguna, arquitectura clara y definida tampoco un concepto concreto. Arquitectura LAMBDA, Arquitectura de Estanques, Arquitecturas Multizonas, Arquitecturas Hibridas, Arquitecturas funcionales, Arquitectura de datos Maduros.

### *1.1.2 Un modelo de script de búsqueda*

Los datos dentro del conjunto del concepto, disponibles en exclusión y supra ordinación así como las clases y subclases fueron utilizados para los criterios de búsqueda. Quedando únicamente una frase compuesta como script de búsqueda. En la Tabla 3 se muestra un script.

**Tabla 3**

*Script de búsqueda ADL*

| Base de datos | Sintaxis | Resultado de búsqueda |
|---|---|---|
| SCOPUS | ALL ( "Lago de DatosArchitecture" ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) OR LIMIT-TO ( LANGUAGE , "Spanish" ) OR LIMIT-TO ( LANGUAGE , "French" ) ) | 50 |

### *1.1.3 Preguntas de investigación*

Se sintetiza una pregunta de investigación que engloba la idea central para lo cual se encuentra enfocado el trabajo de investigación, que demande de una innovación para dar una respuesta acorde que nos permita cumplir con los requerimientos al manejar los datos académicos caso de estudio UTPL. En base a estos requerimientos se plantea la siguiente pregunta:

**¿Qué tipos de arquitecturas de Lago de Datos se aplican en la actualidad?**

Dentro de la Revisión Sistemática de Literatura, se encuentra varios autores que proponen diferentes arquitecturas.

A continuación en la Tabla 4, se describe las Arquitecturas propuestas por los autores de literatura científica.

**Tabla 4**

*Fuente evolución de la arquitectura funcional del lago de datos*

| Arquitectura | Autores | Característica |
|---|---|---|
| De una Zona | Nixon, J. (2010)<br>Fan, H. (2015) | Almacenamientos de datos sin procesar |
| Estanques | Inmon, B. (2016) | Datos brutos, Análoga, Aplicación, Textual, Archivo |
| Multi-Zonas | Nadipalli, R. (2017)<br>Amazon Web Services | Ingestión, Almacenamiento, Procesando, Gobernanza y Seguridad |
| Munti-Zonas | Memon, P. (2017) | Almacén de datos sin procesar, Almacenes de datos procesados, Catalogación de datos, Seguridad y Gobernanza. |
| Multi-Zonas | LaPlante, A. (2014)<br>Zaloni's, DL. | Zona de carga, Datos Brutos, Datos Refinados, Datos Ajustados, Descubrimiento de caja, Gobernanza. |
| Basada en la madurez | ZiKopoulos y col. (2015)<br>LaPlante y Sharme (2016)<br>Tharrington (2017) | Las funciones básicas del lago de datos suelen incluir:<br>1. Una función de ingestión de datos para conectarse con fuentes de datos;<br>2. Una función de almacenamiento de datos para conservar los datos brutos y refinados;<br>3. Una función de procesamiento de datos;<br>4. Una función de acceso a datos para permitir consultas de datos sin procesar y refinados. |
| Híbridas | Inmon (2016),<br>Ravat y Zhao(2019) | Son arquitecturas de lago de datos donde los componentes identificados dependen tanto de las funciones del lago de datos como del refinamiento de datos. |
| Funcional Arquitectura | Jhon Misra(2017)<br>Quix y Hai(2018)<br>Mehmood et al.(2019) | Está constituido por la mayoría de las arquitecturas de la zona. |

*Nota:* Adaptado de (Megdiche et al., 2020), (Pegdwendé Sawadogo & Darmont, 2021)

(Anne Laurent et al., 2020) manifiesta lo siguiente sobre urbanización de los Lagos de Datos.

1) Arquitectura empresarial
2) Arquitectura funcional
3) La arquitectura de la aplicación
4) Arquitectura Técnica

La propuesta estándar para una arquitectura de un lago de datos según (Chihoub et al., 2020)

1) La capa de ingestión
2) La capa de almacenamiento
3) La capa de transformación
4) La capa de iteración

Al innovar con una arquitectura de Lago de Datos(Pegdwendé Sawadogo & Darmont, 2021) hace referencia a una propuesta de topología de tipos de ALD, mediante sus componentes. Arquitectura funcional, Arquitectura basada en la madurez de datos, Arquitectura híbrida.

Según(C Giebler et al., 2020) propone una arquitectura de modelos de zonas mediante la evaluación de diferentes ALD, en cada una de las zonas.

1) Zona de aterrizaje: Caso de uso independiente
2) Zona de datos crudos: Caso de uso independiente
3) Zona de armonización: Caso de uso independiente
4) Zona de datos destilados: Caso de uso dependiente
5) Zona de datos exploratoria: Caso de uso dependiente
6) Zonas de entrega: Caso de uso dependiente

Características de las arquitecturas una zona según (Dixon, 2010) hacen referencia (Megdiche et al., 2020)(Pegdwendé Sawadogo & Darmont, 2021) (Sharma, 2018)(Ravat & Zhao, 2019)

- ✓ La primera visión de la arquitectura DL, es una arquitectura plana con una zona única que almacena todos los datos sin procesar en su formato nativo
- ✓ Esta arquitectura, estrechamente ligada al entorno HADOOP
- ✓ Permite cargar datos heterogéneos y voluminosos a bajo costo

Características de las arquitecturas cinco estanques de datos según (Inmon, 2016) hacen referencia (Megdiche et al., 2020) (Pegdwendé Sawadogo & Darmont, 2021) (Ravat &

Zhao, 2019)(C Giebler et al., 2020): El Lago de Datos se divide en varias secciones, denominadas estanques de datos.

- ✓ Estanque de datos brutos que almacena los datos recién ingeridos y los datos que no encajan en otros estanques.
- ✓ Estanques de datos analógicos
- ✓ Estanques de datos de aplicación
- ✓ Estanques de datos textuales, almacena datos clasificados del estanque de datos sin procesar por sus características.
- ✓ Estanque de datos de archivo almacena los datos que ya no se utilizan.

**Estanque de datos brutos:** Es en realidad, es una zona de tránsito, ya que los datos se acondicionan y transfieren a otro estanque de datos, es decir, el estanque de datos analógico, de aplicación o textual. Este estanque de datos brutos se diferencia de los otros estanques ya que no está asociado con ningún sistema de metadatos.

**Estanques de datos analógicos:** Es un lugar donde, naturalmente, se almacenan datos analógicos. El proceso de preparación de los datos analógicos consiste principalmente en la reajuste de datos, con respecto al volumen de los datos volumen de datos que sea viable, ligero, especifico, y reestructurar los datos del estanque.

**Estanques de datos de aplicación:** Este estanque se llena con información que proviene de la ejecución de una o más aplicaciones. Los datos de esta aplicación son probablemente los más limpios del DL, porque han sido generados por una aplicación. Todos los datos del estanque de aplicaciones estarán estructurados de manera uniforme y contienen valores que son relevantes para la ejecución de alguna actividad comercial. Pero existe la posibilidad de que los datos en este estanque provengan de diferentes aplicaciones. Este origen de datos de múltiples aplicaciones es lo que le puede dar al analista un momento difícil.

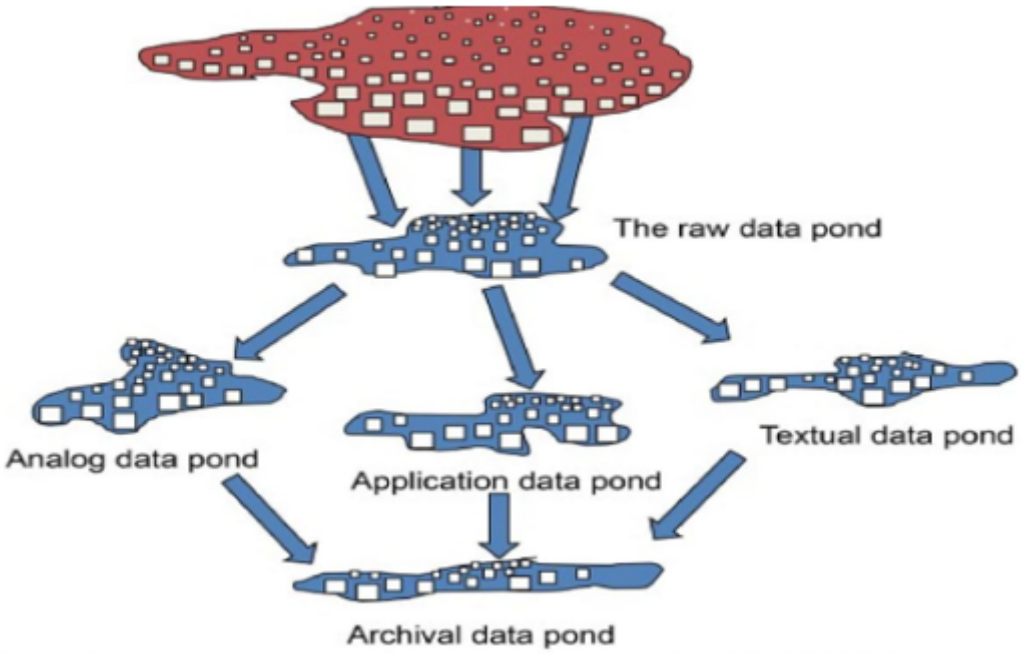
**Estanques de datos textuales:** Es donde se colocan los datos textuales no estructurados. El texto aquí puede provenir de cualquier lugar.

**Estanque de datos de archivo:** El propósito de este estanque de datos de archivo es guardar los datos que no se utilizan activamente, pero que aún pueden ser necesarios en

el futuro. Los datos archivados pueden provenir de los estanques de datos analógicos, de aplicación y textuales.

**Figura 2**

*Arquitectura Lago de Datospropuesta por Inmon*




*Nota:* Adaptado de (Inmon, 2016)

Características de las arquitecturas multi zona (Nadipalli, 2017a)hacen referencia (Ravat & Zhao, 2019) (Megdiche et al., 2020): Arquitectura DL de Amazon Web Services (AWS) con cuatro zonas.

- ✓ Ingestión, almacenamiento, procesamiento y gobernanza y seguridad.
- ✓ Los datos sin procesar se cargan en la zona de ingestión.
- ✓ Los datos brutos ingeridos se almacenan en la zona de almacenamiento
- ✓ Cuando se necesitan datos, se procesan en la zona de procesamiento
- ✓ El objetivo de Govern & secure zone es controlar la seguridad de los datos, la calidad de los datos, la gestión de metadatos y el ciclo de vida de los datos.

Características de las arquitecturas multi-zona de Lago de Datos (Sharma, 2018), hacen referencia (Megdiche et al., 2020) (Pegdwendé Sawadogo & Darmont, 2021) (C Giebler et al., 2020)

La principal ventaja de esta arquitectura es que los datos pueden ser ingeridos al lago de datos desde cualquier lugar, como procesamiento de transacciones en línea(OLTP), sistemas de almacenamiento de datos operativos(ODS), servicios en la nube. Estos sistemas fuente incluyen muchos formatos diferentes, como datos de archivos, datos de bases de datos, ETL, datos de transmisión e incluso datos que ingresan a través de API.

Las principales zonas de esta arquitectura son: Zona de carga transitoria, Datos brutos o sin procesar, Datos refinados o de confianza, Descubrimiento de pruebas, Datos Ajustados o zona de consumo, Gobernanza.

- ✓ Separa las zonas de procesamiento y almacenamiento en una zona de datos refinada, una zona de datos de confianza y una zona de pruebas de detección.
- ✓ La zona refinada permite integrar y estructurar datos.
- ✓ La zona de datos de confianza almacena todos los datos limpios.

**Zona de carga transitoria:** Donde se hacen comprobaciones básicas de la calidad de los datos mediante MapReduce o Spark en el clúster de Hadoop.

**Zona de datos sin procesar o datos brutos:** Maneja datos en formato casi sin procesar provenientes de la zona transitoria, se cargan Hadoop, y los datos personales se pueden redactar para que se pueda acceder a ellos sin revelar información de identificación personal, información de la industria de tarjetas de pago (PCI), u otros tipos de datos significativamente confidenciales.

**Zona de datos de confianza:** Este repositorio de confianza contiene tanto datos maestros y datos de referencia. Realiza métodos de estándar de limpieza y validación.

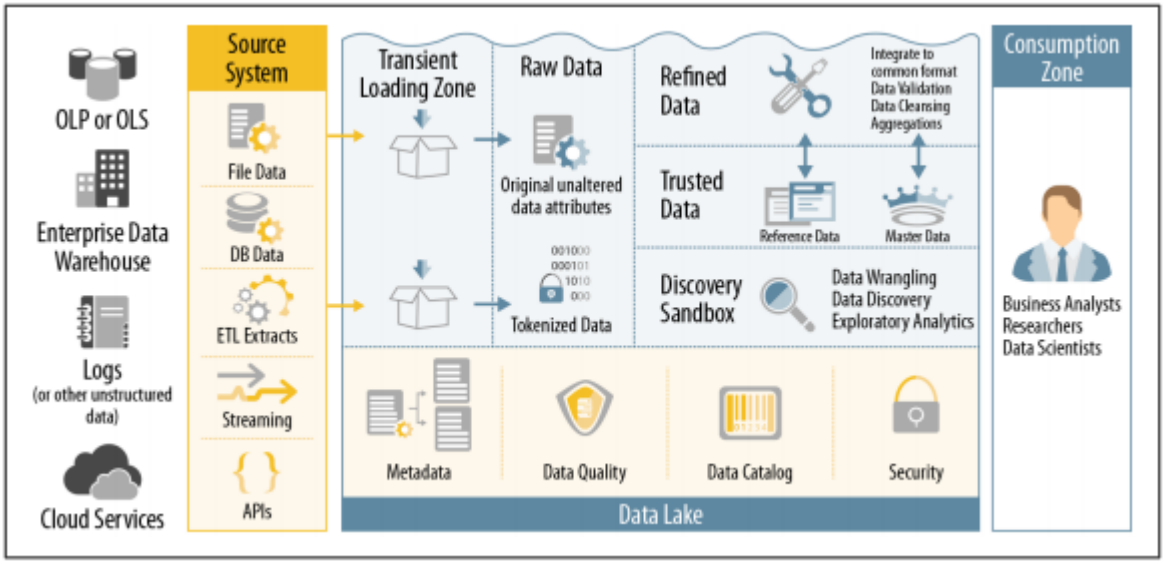
**Zona de descubrimiento de pruebas:** Desde el área de confianza, los datos se mueven a la zona de pruebas de descubrimiento donde pueden estar a los que acceden los científicos de datos a través de operaciones de búsqueda de datos o negociación de datos.

**Zona de consumo:** Existe para que los analistas de negocios, los investigadores y los científicos de datos se sumerjan en el lago de datos para ejecutar informes, hacer análisis del tipo "qué pasaría si" y, de otro modo, consumir los datos para generar conocimientos comerciales para una toma de decisiones informada.

**El zona de gobierno:** Finalmente permite administrar, monitorear y gobernar metadatos, datos calidad, catálogo de datos y seguridad. Aunque las empresas pueden variar en la forma en que estructuran la plataforma de integración, en general, la gobernanza debe ser parte de la solución.

**Figura 3**

*Arquitectura de Lago de Datos propuesta por Zaloni*




*Nota:* Adaptado de (Sharma, 2018)

Esta arquitectura se basa en las siguientes funciones principales:

**Ingestión de datos:** El uso de la ingestión administrada con un lago de datos abre enormes posibilidades. Puede ingerir rápida y fácilmente datos no estructurados y ponerlos a disposición para su análisis sin necesidad de transformarlos de ninguna manera. Con la ingestión administrada, ingresa todos los datos en una tabla gigante organizada con etiquetas de metadatos. En otras palabras, puede especificar quién tiene acceso a los datos en cada celda y bajo qué circunstancias, desde el comienzo de la ingestión.

- ✓ Escalable, extensible para capturar la transmisión de datos por lotes.
- ✓ Proporcionar capacidad para la lógica empresarial, los filtros, la validación, la calidad de los datos, el enrutamiento, etc. Requisitos comerciales

➔ Tecnologías:

- ❖ Apache Flume
- ❖ Apache Kafka
- ❖ Apache sqoop
- ❖ NFS Gateway

➔ Gobierno de datos: Un proceso de ingesta administrado hace cumplir las reglas de gobierno que se aplican a todos los datos que pueden ingresar al LD.

- ❖ Cifrado: Los datos deben protegerse mediante encriptación, si su visibilidad es un problema, deben estar encriptados, antes de entrar al lago de datos.
- ❖ Procedencia y linaje: Es esencial evitar que los datos ingresen al LD, si se desconoce su procedencia.
- ❖ Captura de metadatos: El proceso de ingestión administrado permite establecer reglas de gobierno que capturan los metadatos.
- ❖ Limpieza de datos: Establecer estándares de limpieza de datos que se aplican a medida que se ingieren los datos para garantizar que solo los datos limpios ingresen a los LD.

**Almacenamiento y retención de datos:** Los proveedores de servicios en la nube como AWS ofrecen una variedad de opciones de almacenamiento a diferentes precios, según sus requisitos de accesibilidad.

- ❖ Según los requisitos, los datos se colocan en hadoop HDFS, Hivem Hbase, Elastic Search o en memoria.
- ❖ Gestión de metadatos
- ❖ Se proporciona retención de datos basada en políticas.

➔ Tecnologías:

- ❖ HDFS
- ❖ Hive Tables

- Hbase/MapR DB
- Elastic Search

**Procesamiento de datos:** El procesamiento es la etapa en la que los usuarios comerciales o los científicos de datos pueden transformar los datos a un formato estandarizado. Los usuarios comerciales pueden realizar diferentes estandarizaciones y transformaciones dependiendo de sus necesidades únicas. Dependiendo de su necesidad puede utilizarse herramientas para la transmisión en lotes o en tiempo real.

Para casos de uso por lotes, las organizaciones generalmente usan Pig, Hive, Spark y MapReduce. Para tratamiento de transmisión de la información en tiempo real, se encuentran disponibles las siguientes herramientas como Spark-Streaming, Kafka, Flume y Storm.

- El procesamiento se proporciona en tiempo real y por lotes
- Proporciona flujos de trabajo para el procesamiento de trabajo
- Proporciona manejo de datos en llegada tardía

➔ Tecnologías:

- Map Reduce
- Hive
- Spark
- Storm
- Drill

**Visualización y Acceso a los datos, APIs:** En esta etapa es donde se consumen los datos del lago de datos.

Los usuarios empresariales también pueden utilizar paneles, ya sea creados a medida para satisfacer sus necesidades, o Microsoft SQL Server Reporting Services (SSRS), Oracle Business Intelligence, Enterprise Edition(OBIEE) o IBMCognos.

- Aplicaciones que proporcionan información empresarial valiosa

- ❖ Los datos estarán disponibles para los consumidores mediante API, MQ Feed y acceso a BD

➔ Tecnologías:

- ❖ Qlik/Tableau/Spotfire
- ❖ REST APIs
- ❖ Apache Kafka
- ❖ JDBC

Características de la Arquitectura multi zona funcional genérica propuesta por (Ravat & Zhao, 2019) quienes hacen referencia (Pegdwendé Sawadogo & Darmont, 2021) (Megdiche et al., 2020) (Cravero et al., 2020)(C Giebler et al., 2020): Contiene cuatro zonas: ingesta, proceso, acceso, gobernanza.

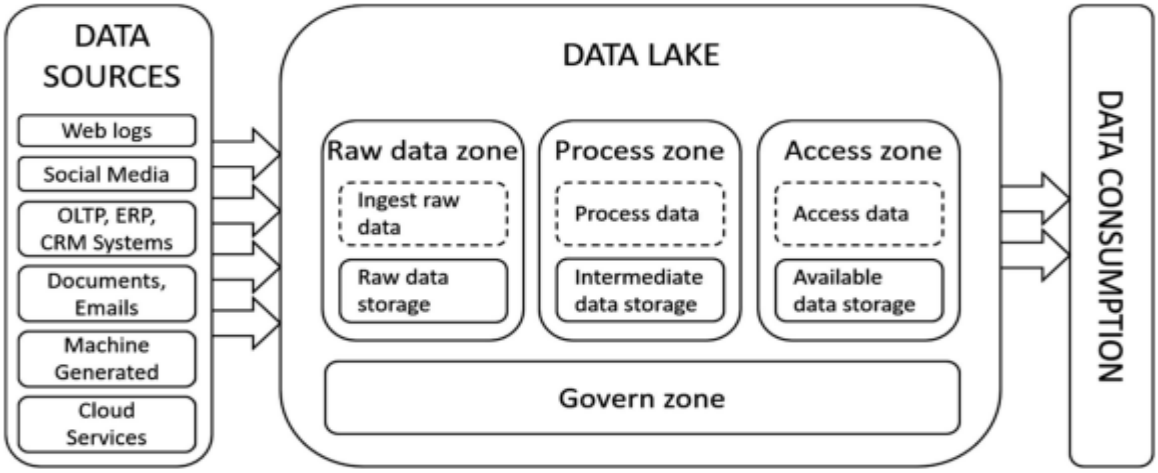
- ❖ Cada zona, excepto la zona de gobierno, tiene un área de tratamiento y un área de almacenamiento de datos que almacena el resultado de los procesos.
- ❖ Zona de datos brutos: todos los tipos de datos se ingieren sin procesar y se almacenan en su formato nativo. La ingestión puede ser por lotes, en tiempo real o híbrida. Esta zona permite a los usuarios encontrar la versión original de los datos para su análisis para facilitar los tratamientos posteriores. El formato de datos sin procesar almacenados puede ser diferente del formato de origen. Por lo que es obligatorio configurar un sistema de gestión de metadatos para LD.
- ❖ Zona de proceso: en esta zona, los usuarios pueden transformar los datos según sus necesidades y almacenar todos los datos intermedios. El procesamiento de datos incluye procesamiento por lotes y / o en tiempo real. Esta zona permite a los usuarios procesar datos (selección, proyección, unión, agregación, etc.) para su análisis de datos.
- ❖ Zona de acceso: la zona de acceso almacena todos los datos disponibles para el análisis de datos y proporciona el acceso a los datos. Esta zona permite el

consumo de datos de autoservicio para diferentes analíticas (informes, análisis estadístico, análisis de inteligencia empresarial, algoritmos de aprendizaje automático).

- Zona de gobernanza: La gobernanza de datos se aplica en todas las demás zonas. Está a cargo de asegurar la seguridad de los datos, la calidad de los datos, el ciclo de vida de los datos, el acceso a los datos y la gestión de los metadatos.

**Figura 4**

*Arquitectura Funcional Lago de Datos*




*Nota:* Adaptado de (Ravat & Zhao, 2019)

Propuesta de tipología de arquitectura según (Pegdwendé Sawadogo & Darmont, 2021) quien hace referencia (Cravero et al., 2020): Tres tipos de Arquitecturas:

Características:

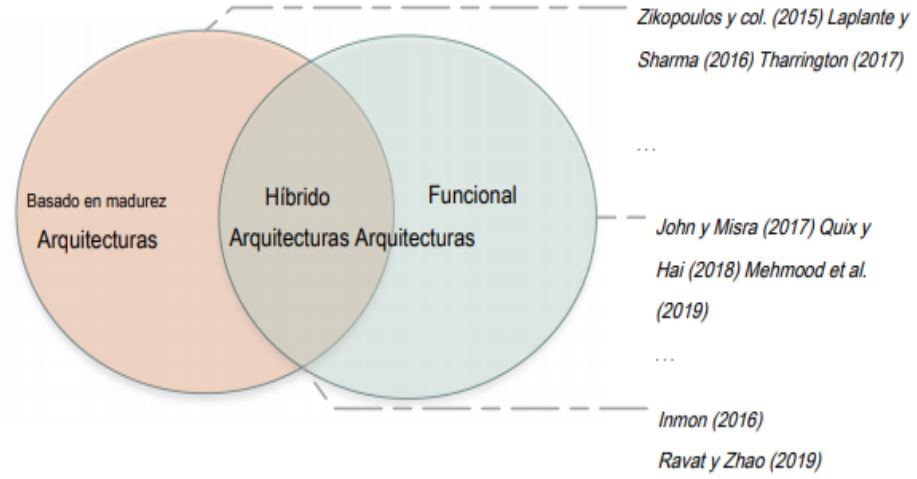
**Las arquitecturas funcionales:** Siguen algunas funciones básicas para definir los componentes de un lago.

**Arquitecturas basadas en madurez de datos:** Son arquitecturas de LD donde los componentes se definen con respecto al nivel de refinamiento de los datos. En otras palabras, está constituido por la mayoría de las arquitecturas de la zona.

**Arquitecturas híbridas**: Arquitecturas de LD, donde los componentes identificados dependen tanto de las funciones del lago de datos como del refinamiento de datos.

**Figura 5**

*Propuesta de tipología de Arquitectura*




*Nota: Adaptado de* (Pegdwendé Sawadogo & Darmont, 2021)

Propuesta de un metamodelo referencial para la gestión de lagos de datos de nivel empresarial(C Giebler et al., 2020)

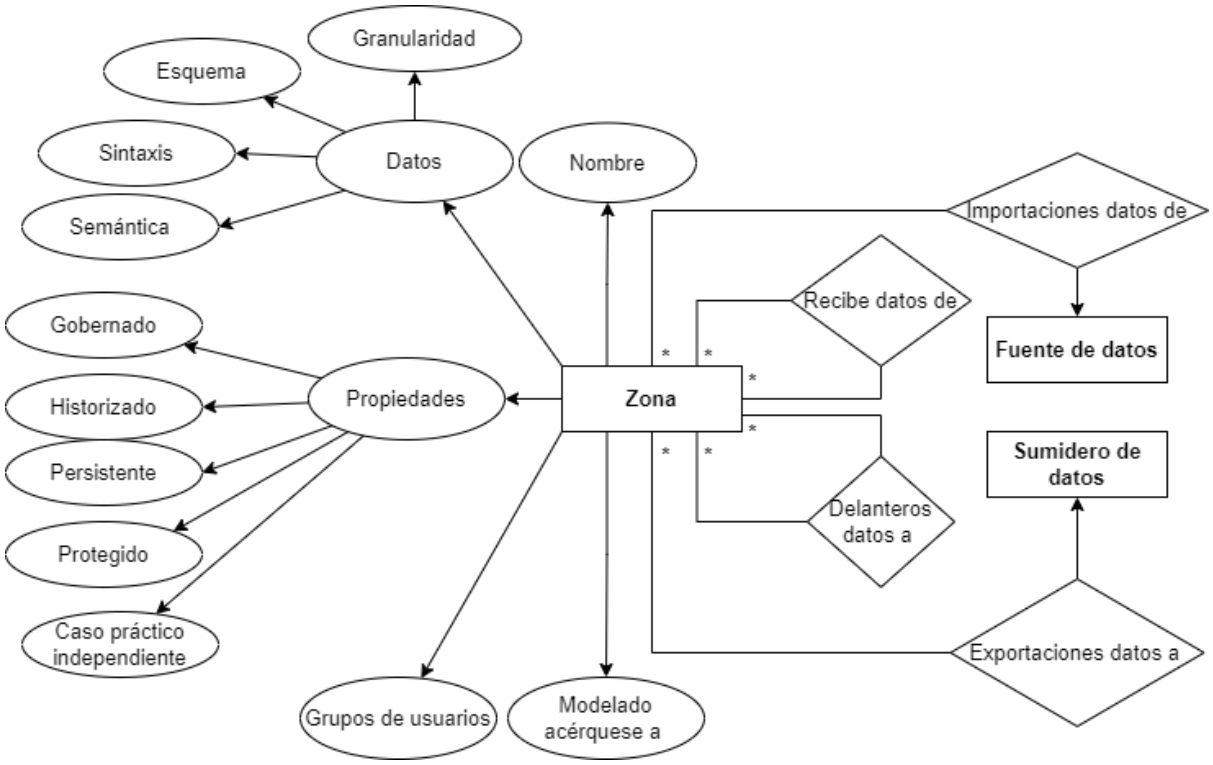
Según lo citado, enfatiza que no hay una evaluación sistemática de los conceptos sobre modelos de una ALD, por lo cual se enfoca en evaluar cada uno de los modelos proponiendo un metamodelo de zona, mediante la aplicación de requisitos y casos de uso típicos de las empresas: *Requisitos*: Pre-procesado, Limpiado, Integrado, Gobernado, Informes y OLAP, Análisis avanzados, Volver a escribir.

*Casos de uso*: Finanzas, Calidad de Gestión, Fabricación, Los Servicios del Cliente final.

A partir de aquí, cada zona puede describirse de acuerdo a un conjunto de atributos, en un diagrama E-R de un metamodelo.

**Figura 6**

*Modelo entidad relación*



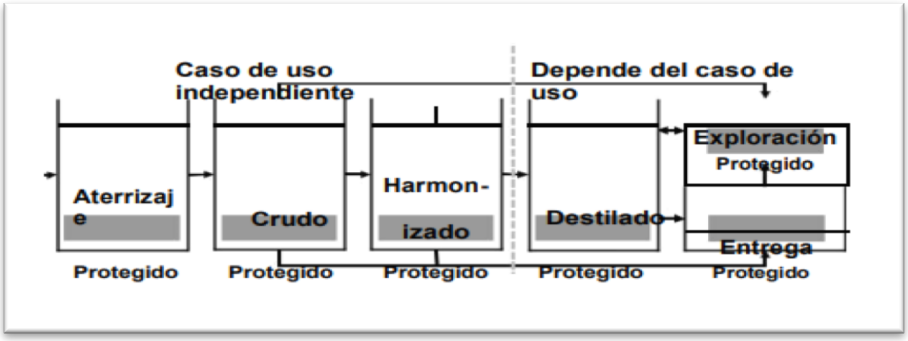

*Nota:* Adaptado de (C Giebler et al., 2020)

Atributos a evaluar: Granularidad(En bruto-Agregado), Esquema(Cualquiera-Consolidado), Sintaxis(Sin cambios-Consolidado), Semántica(Sin cambios-Procesado), Propiedades, Grupos de usuarios, Enfoque de modelado.

Zonas: Aterrizaje, Crudo, Armonizado, Destilado, Exploración, Entrega.

**Figura 7**

*El modelo de referencia zonal, comprende seis zonas*




*Nota:* Adaptado de (C Giebler et al., 2020)

**Zona de Aterrizaje:** cumple con funciones específicas en caso de que los datos requieran una alta tasa de ingesta, y no existe la implementación técnica en la Zona Cruda.

Las características: granularidad bruta, su esquema es el mismo de origen, la sintaxis puede cambiar pueden ser anonimizados para cumplir con la norma legal, la semántica de los datos se mantiene, las propiedades gobernada, no historiada, no persistente, en cuanto a los grupos de usuarios, no tiene usuarios finales, solo los sistemas y procesos, pueden introducir datos en ella, finalmente no se define ningún modelado.

**Zona Cruda:** Datos en formato de origen, esta zona difiera de la Zona de Aterrizaje, con respecto a las propiedades, almacena los datos de forma persistente y pueden ser manipulados y eliminados además historiados, los grupos de usuarios en esta zona son los científicos de datos, considerando que el uso de los datos de la parte protegida está muy restringido.

**Zona Armonizada:** Dentro de los atributos el esquema y la sintaxis de los datos cambian en comparación con los datos de origen, los datos de diferentes orígenes se integran en un esquema consolidado, y la sintaxis de los datos se consolidan en esta zona, los usuarios y propiedades no cambian, su modelo estandarizado se enfoca en la técnica de Data Vault (Corinna Giebler et al., 2019)

***Zona de Destilación:*** Datos disponibles para su uso, la granularidad de los datos se modifica, la semántica cambia, el esquema hace referencia al caso de uso, en lo que refiere a las propiedades a partir de la Zona Destilada se va a depender de los casos de uso que soporte. En cuanto a los usuarios los grupos de usuarios de la zona armonizada también forman parte de esta Zona, utiliza también modelado de datos con Data Vault (Corinna Giebler et al., 2019)

***Zona de Explotación:*** La granularidad, el esquema, la sintaxis y la semántica se pueden alterar de cualquier forma en que sea obligatoria para la exploración.

Esta zona no se encuentra gobernada, los científicos de datos pasarían a ser el grupo de usuarios, en esta zona se pueden utilizar datos que no son sensibles, así como la utilización de los datos también depende del caso de uso, aquí no necesariamente tienen que

estar historiados, los análisis como resultados pueden ser enviados a la Zona Destilada antes de ser eliminados de la Zona Exploratoria, esta zona es la menos inflexible, sin embargo, los análisis de los datos sensibles se deben de realizar en la parte protegida de la Zona.

***Zona de Entrega:*** En esta zona los datos se gobiernan y almacenan de forma persistente, en este caso los usuarios son variados, como por ejemplo, científicos de datos , expertos, empresariales así también como los procesos y sistemas, leen los datos de la zona de entrega, su modelado depende del caso de uso.

### *1.1.4 Modelo protocolo de revisión*

Para el desarrollo de este protocolo se considera criterios de inclusión y exclusión, para el protocolo de búsqueda.

Se considera las clases excluyentes y las clases altas del mentefacto conceptual. En nuestro caso de estudio se considera la base de datos SCOPUS, sin exclusión de años.

SCOPUS es una base de datos de referencias bibliográficas y citas de la empresa Elsevier, de literatura y contenido web de calidad, con herramientas para el seguimiento análisis y visualización de la investigación.

### *1.1.5 Modelo de (RSL)*

En la Tabla 5 se muestra lo que se obtuvo del siguiente script de búsqueda: (title(“Data Lake*”) AND TITLE-ADS-KEY(“Architecture*”))

**Tabla 5**

*Modelo de RSL*

| Base de datos | Total encontrados | Aceptados | Incluidos |
|---|---|---|---|
| Scopus | 54 | 43 | 7 |

### *1.1.6 Listado de revistas*

Mediante el protocolo de búsqueda y la aplicación de selección de literatura. Se ha encontrado los siguientes resultados:

**Tabla 6**

*Número de revistas*

| Nombre de la Base de Datos | Número de artículos |
|---|---|
| Scopus | 14 |
| Total: | 14 |

**Tabla 7**

*Revistas*

| Ord | Nombre de Revista | Nro. papers | JCR | | SJR | | h5 Google | Value |
|---|---|---|---|---|---|---|---|---|
| | | | IF | Cuartil | IF | Cuartil | | |
| 1 | Journal of Intelligent Information Systems | 198 | | | 0,424 | Q2 | 23 | 965,448 |
| 2 | IEEE Access | 41669 | | | 0,587 | Q1 | 119 | 727676,1642 |
| 3 | Future Internet | 690 | | | 0,434 | Q2 | 27 | 4042,71 |
| 4 | International Journal of Advanced Computer Science and Applications | 2911 | | | 0,193 | Q3 | 35 | 14747,8537 |
| 5 | Security and Communication Networks | 1060 | | | 0,446 | Q2 | 40 | 9455,2 |
| 6 | Information Systems Frontiers | 321 | | | 1,086 | Q1 | 46 | 4008,969 |
| 7 | ACM Transactions on Information Systems | 170 | | | 0,672 | Q1 | 30 | 856,8 |

$$Value = (\#\ papers * 25\%).(JCR\ IF).(SJR\ IF).(h5\ index)$$

### *1.1.7 Modelo de resultado de búsquedas*

Palabras del diccionario de sinónimos para búsqueda de criterios, Estructura semántica para buscar trabajos específicos, script de búsqueda BD, selección de artículos, organización de resultados.

RQ1: Hace referencia a pregunta #1

A1: Hace referencia a las variables

**Tabla 8**

*Respuesta a pregunta RQ1*

| **RQ1** | **¿Qué tipos de arquitecturas de lago de datos se aplican en la actualidad?** | | ***f*** |
|---|---|---|---|
| | A1(Una Zona) | (Dixon, 2010)(Ravat & Zhao, 2019)(Megdiche et al., 2020) | 1 |
| | A2(Cinco Estanques) | (Inmon, 2016)(Ravat & Zhao, 2019)(Megdiche et al., 2020)(C Giebler et al., 2020) | 1 |
| | A3(Multi Zonas) | (Nadipalli, 2017b)(Ravat & Zhao, 2019)(Megdiche et al., 2020)(C Giebler et al., 2020)[ 14] | 2 |
| | A4(Basados en Madurez) | (Pegdwendé Sawadogo & Darmont, 2021)(Sharma, 2018)(Couto et al., 2019) | 1 |
| | A5(Híbridos) | (Inmon, 2016)(Ravat & Zhao, 2019)(P Sawadogo & Darmont, 2021) | 2 |
| | A6(Funcional) | (Ravat & Zhao, 2019)(Pegdwendé Sawadogo & Darmont, 2021) (Sharma, 2018) | 1 |
| | A7(Urbanización) | (A. Laurent et al., 2020) | 1 |
| | A8(Cinco Capas) | (Chihoub et al., 2020)(A. Laurent et al., 2020) | 1 |
| | A9(Modelo) | (C Giebler et al., 2020) | |

### *1.1.8 Resultados*

**Tabla 9**

*Visión general ALD*

| Diseño de la estructura | Actividades | Datos | Tecnología | Usuarios |
|---|---|---|---|---|
| Arquitectura Plana, propuesta por (Dixon, 2010) | | | | |
| Zona única | Recolección y almacenamiento de datos sin procesar | Registros web, dispositivos sensores, datos operativos(ODS), sistemas de procesamiento de transacciones en línea(OLTP) | HADOOP | No refleja |
| Arquitectura Estanques, propuesta por (Inmon, 2016) | | | | |
| Estanque de datos transitorio, sin procesar, bruto, crudos | Ingerir, almacenamiento, transferir a otro estanque | | | |
| estanque de datos analógicos | Almacenamiento, gran velocidad, procesamiento | datos IoT | | |
| Estanques de datos de aplicación | Almacenamiento, ingerir, integrar, transformar, preparar. | estructurados de aplicaciones (DBMS) | | |
| Estanque de datos textuales | Almacenamiento, gestión, análisis | no estructurados | | |
| Estanque de datos de archivo | Archivar datos inútiles | | | |

| Multizona AWS(Nadipalli, 2017b) | | | | |
|---|---|---|---|---|
| Ingestión | Carga de datos | datos sin procesar | | |
| Almacenamiento | Ingerir datos | datos brutos | | |
| Procesamiento | Procesamiento de datos a petición | datos procesados | | |
| Gobernanza<br>Seguridad | control, seguridad, calidad, gestión de metadatos, ciclo de vida | | | |
| **Multizonas (Sharma, 2018)** | | | | |
| Zona de carga transitoria | Ingerir calidad de datos, comprobaciones básicas de calidad | Datos de transmisión en tiempo real y por lotes | por lotes: Pig, Apache Hive, Tormenta apache, Taladro Apache, Apache Spark y MapReduce. Transmisión; Spark-Streaming, Kafka, Flume y Storm | procesos |
| Datos crudos, sin procesar | carga, estandarizar y depurar, redactar datos confidenciales, transformación de datos | identificación personal, información de salud personal, información de tarjetas, datos vulnerables. | | Científicos de datos, analistas de negocio |
| Zona de datos | Procesar y almacenar, integrar, | | HDFS, Hive | |

| | | | | |
|---|---|---|---|---|
| Refinados | estructurar, métodos estándar de limpieza y validación | | TABLES, Hbase/MapR, Escalabilidad, Flexibilidad, DB,Elastic Search | |
| Zona de confianza | Procesar y almacenar todos los datos depurados | Datos maestros: limpiar y validar(Nombre, dirección)(fechas de nacimiento, número de cédula); y datos de referencia: única fuente de verdad | Map Reduce, Hive, Spark, Storm, Drill | |
| Zona descubrimiento, Pruebas | análisis exploratorio, Descubrir y analiza | | | Usuarios y los científicos |
| Zona de consumo | Ejecutar informes, hacer análisis, consumir datos, generar conocimiento comerciales, para tomar decisiones informadas. | | Tableau, Qlik, Microsoft SQL Server Reporting Services(SSRS), Oracle Business Intelligence, Enterprise Edition(OBIEE) o IBMCognos, Spotfire, reset Apis, apache kafka, JDBC, App | Analistas de negocio, investigadores, científicos de datos |
| Gobernanza | administra, monitorea, integra, | | Ambari, cloudera | |

| | | | | |
|---|---|---|---|---|
| | gobierna metadatos, calidad, catálogo de datos, seguridad | | Manager, Cloudera Navigator, MapR MCS, AWS(almacenamiento en la Nube S3) | |
| Multizona Genérica (Ravat & Zhao, 2019) | | | | |
| Zona de datos sin procesar | Ingerir datos sin procesar, almacenar | Por lotes y en tiempo real o híbrida | | Sistemas |
| Zona de proceso | transformación a petición, almacenar datos intermedios | datos procesados por lotes, y en tiempo real | | Usuarios: selección, proyección, unión, agregación |
| Zona de acceso | Almacena de datos disponibles para el análisis, acceso a datos, consumo de datos (Informes, análisis estadístico, análisis de inteligencia empresarial, algoritmos de aprendizaje automático) | | | |
| Zona de Gobernanza | Aplica a todas las demás zonas: seguridad, calidad, ciclo de vida, acceso y gestión de los metadatos. | | | |
| Modelo de referencia de Zona (Corinna Giebler et al., 2020) | | | | |
| Zona de aterrizaje | -Ingerir datos y luego enviar a la zona de datos crudos por lotes | Datos a grandes velocidades y volumen, crudos y granularidad bruta, datos por lotes ERP, datos de flujo | Arquitectura lambda, Kafka, HDFS, bases de | Sistemas, proceso |

| | | | | |
|---|---|---|---|---|
| | -Transformaciones básicas, anonimizados, con igual semántica.<br>Gobernado, no especializado, sin modelado | procedentes del campo | datos relacionales, Spark streaming3 | |
| Zona cruda | Transformaciones básicas, sin esquema, sintaxis básica, con igual semántica, gobernado, historiado, persistente, cumplimiento de la normativa , e integridad de los datos, almacenamiento, anonimizar | datos granulados, datos historiados, Datos por lotes y de flujo, datos en streaming, datos anonimizados | HDFS, bases de datos relacionales ERP, almacenamiento de datos, archivos JSON, | Científicos de datos: copiar datos a la zona exploratoria, Sistemas procesos |
| Zona armonizada | Actuar en función de la demanda, copia de los datos de la zona bruta, datos maestros accesibles, gestión de datos maestros, limpieza de datos, esquema y sintaxis diferente, integración basada en enlaces, sintaxis consolidada, fusión de datos, modelo estandarizado, data vault(modelar datos de la empresa). | datos armonizados y consolidados, datos modelados, datos heterogéneos procedentes de diversas fuentes | | Científicos de datos |
| Zona de destilación | datos disponibles para el uso, esquema consolidado, sintaxis | dato agregado granularidad, dato semántica, gobernado, historiado, | SQL | Científicos de datos, sistemas, procesos. |

| | | | | |
|---|---|---|---|---|
| | consolidado, semántica compleja, cálculo de los KPI. proceso de datos para un determinado grupo de casos de uso, Data Vault(modelado estandarizado). | protegido caso de uso dependiente, persistente, datos en lote, datos en flujo, datos modelados. | | |
| Zona de explotación | almacenar transformaciones de resultados, extraer datos de cualquier otra zona, cualquier enfoque de modelado, descubrir KPIs, | datos procesados | Hadoop, Python 3 o diversas herramientas de visualización. | científicos de datos |
| Zona de entrega | dependiendo del caso de uso, usos y aplicaciones específicas, elaboración de informes y el OLAP, funcionalidad similar a la de los data marts, datos operativos data warehousing, enviar datos a sumideros, modelado dimensional OLAP, o tablas planas, esquema de estrella, gobierno y almacén de forma persistente, preparación de datos informes y OLAP. | datos de apoyo a usuarios con pocos conocimientos sobre análisis de datos | Bases de datos relacionales, herramientas de análisis | Científicos de datos, expertos en el dominio, usuarios empresariales |

### *1.1.9 Conclusiones*

- ✓ Al realizar el análisis sistemático de literatura encontré que en (Ravat & Zhao, 2019), hace un análisis de las diferentes ALD y lo denomina evolución de zona única a múltiples-zonas, proponiendo una arquitectura genérica funcional y un concepto de LD.
- ✓ La clasificación de arquitecturas por componentes propuesta en (Pegdwendé Sawadogo & Darmont, 2021), hace referencia a tres tipos de Arquitecturas: Funcional, Arquitectura de la madurez de datos, Arquitectura híbrida, hace referencia a un estudio sintetizado de las propuestas de diferentes autores. Y una propuesta de gestión de metadatos.
- ✓ En (C Giebler et al., 2020), hace una evaluación de las ALD existentes, en cada una de las zonas y propone un modelo de referencia. Así como su estudio es evaluado para los atributos de la zona, en casos de uso como requisitos para diferentes casos de estudio.

# Capítulo dos

# Problema

## 2.1 Contexto

La conectividad global en la industria 4.0, y las mega tendencias traen consigo realidades paralelas, a nivel industrial, comercial, financiero, educativo, organizacional, político, civil. Estos sectores generan grandes cantidades de datos, de diversos tipos.

En el sector educativo las mega tendencias y las tecnologías persistentes traen propuestas de valor bastante diferente a lo que se ha utilizado en el pasado en su procesamiento de datos para sacar valor.

En nuestro caso, la institución educativa UTPL, contiene 23 carreras en su Modalidad Presencial, 17 en su Modalidad Abierta y a Distancia y 5 posgrados vigentes, son más de 65.000 profesionales que se han graduado en esta institución y actualmente cuenta con una planta docente de 1070 profesores.

El desafío de Big Data como medio de almacenamiento masivo de datos educativos es esencial en la dinámica de innovación tecnológica institucional. Los datos educativos que se producen son masivos y complejos, con una gestión adecuada se podrá extraer información valiosa según (Munshi & Alhindi, 2021). Las instituciones educativas recopilan grandes cantidades de datos a nivel organizacional y técnico.

Todo esto provoca grandes desafíos tecnológicos en la forma de organizar, procesar y analizar la capacidad de almacenamiento de los datos académicos. Esto requiere tener modelos con tecnologías adecuadas para aprovechar diferentes técnicas y métodos para analizar estos datos y mejorar de forma continua.

Beyer de Garther(Krishnan, 2020), y otros expertos relacionan que los grandes volúmenes de datos, o Big Data, requieren grandes cambios en el servidor, la infraestructura de almacenamiento y la arquitectura de administración de la información en la mayoría de las empresas, la arquitectura de Big Data se da en base a la fuente de datos.

En las instituciones educativas existen diferentes fuentes de datos, voluminosos, variados que son todo un desafío gobernarlos, según (Munshi & Alhindi, 2021) para sacarle provecho a estos datos se utilizan técnicas de inteligencia artificial y minería de datos. Esto significa que los enormes conjuntos de datos no se pueden administrar con sistemas de administración de bases de datos tradicionales.

Las tecnologías de Big Data son una muy buena opción para llevar a cabo una gestión de datos académicos adecuada, propuesta como conceptos emergentes Lago de Datos, dicho concepto se propone para abordar los problemas de Big Data, especialmente aquellos inducidos por la velocidad, volumen y variedad de datos, fue introducido por primera vez por (Dixon, 2010).

Este trabajo de investigación tiene como enfoque fundamental, hacer una propuesta de un modelo de Arquitectura Lago de Datos Académicos, para la gestión acorde a los requisitos en sus características así como encajar el ciclo de vida de los datos en los componentes de la arquitectura seleccionada.

Las arquitecturas de un Lago de Datos que se considera de acuerdo al análisis sistemático de literatura son las propuestas por los siguientes autores: Arquitectura plana (Dixon, 2010), arquitectura de Estanques (Inmon, 2016), arquitectura multi-zonas (Megdiche et al., 2020), (Pegdwendé Sawadogo & Darmont, 2021), (A Laurent et al., 2020), (Chihoub et al., 2020), (Nadipalli, 2017b), (Sharma, 2018), (Ravat & Zhao, 2019), (C Giebler et al., 2020), (Couto et al., 2019).

Estos autores hacen referencia al concepto de LD, características LD, propuestas de Arquitecturas de LD, características sobre los componentes como requisito de la LDA, Tecnologías LD, aplicabilidad LD, Funciones LD.

Entre otras situaciones muy importantes que se dan en el manejo de aplicabilidad de su Arquitectura, el manejo y la gestión de los metadatos según (Pegdwendé Sawadogo & Darmont, 2021)

## 2.2 Alcance

El presente trabajo tiene como finalidad dar un alcance específico basado en un contexto institucional académico, la Universidad Técnica Particular de Loja, en su estructura organizacional cuenta con el gobierno de la UTPL está integrado por la máxima autoridad de cogobierno, que es el Consejo Superior; la primera autoridad ejecutiva, que es el Rector; los vicerrectores; las autoridades ejecutivas, académicas, administrativas, de gestión y de apoyo; y los órganos colegiados que no constituyen cogobierno, de acuerdo con el siguiente orden jerárquico: a) Junta Ejecutiva Universitaria, b)Junta Académica, c)Junta de Facultad, d)Consejo de Departamento, e)Otros.

Las fuentes de datos provienen de todo el conjunto organizacional internas y externas conectadas con las áreas de dominio académico, ***Facultades***, ***Departamentos, Dominios Académicos.***

Se propone dar una solución acorde al manejo de datos estructurados, no estructurados y semiestructurados, con arquitecturas Big Data que permita ser un componente para la evaluación de mecanismos y mejora de PEDI.

Centrándonos únicamente en una propuesta que mejore el acceso, proceso, presentación, y como gobernanza en la gestión de metadatos, acceso de usuarios, privacidad de los datos. Mediante el diseño de un prototipo “Arquitectura de un Lago de Datos Académicos”.

## 2.3 Problema

La gran cantidad de datos que en la actualidad generan las instituciones educativas provenientes de diferentes fuentes, provoca que los sistemas de datos tradicionales no sean los más idóneos para la gestión de estos datos, impulsando así un desarrollo de nuevas tecnologías que permitan flexibilidad en su volumen, velocidad, variedad, y veracidad en el manejo de los datos mediante tecnologías de Big Data.

Los datos institucionales de la unidad de gestión académica de la UTPL, en la actualidad plantean la problemática de llevar a cabo una gestión de los datos mediante

Arquitecturas de Big Data, como un componente adicional para su mecanismo de evaluación y mejorar su PEDI, ya que en la actualidad se maneja con un sistema informático.

## 2.4 Objetivos

### *2.4.1 Objetivo General*

Proponer una Arquitectura Lago de Datos Académico, caso de estudio UTPL.

### *2.4.2 Objetivos Específicos*

- Analizar las propuestas sobre Arquitectura Lago de Datos, para identificar tipologías y conocer su estado actual, aplicabilidad, componentes y factores en general sobre las soluciones de los Lagos de Datos planteadas a nivel académico e industrial.
- Evaluar y seleccionar un prototipo de Arquitectura Lago de Datos Académicos UTPL, en base a las tipologías y factores de un Lago de Datos.
- Diseñar un prototipo que permita gestionar el manejo de datos Heterogéneos Académicos UTPL.

### *2.4.3 Entregables*

- ✓ Informe Trabajo Fin de Máster

# Capítulo tres

# Propuesta

En la propuesta se describen las siguientes fases:

a. Fase de Inicio: Se analizo los diferentes prototipos de una Arquitectura Lago de Datos, para identificar los diferentes factores que lo componen.

b. Fase de Evaluación: Se evaluó una Arquitecturas Lago de Datos en base a los casos de estudio en un caso de uso, se consideró además el mejor concepto emergente que se adaptó a la propuesta, así como las características a nivel estándar.

c. Fase de Selección de un prototipo: Se seleccionó el prototipo que cumplió con las características más adecuadas para un ALDA, adaptado a el caso de estudio UTPL

## 3.1 Fase de inicio

Esta fase comprendió la discusión del análisis de los RSL, encontrados como trabajos relacionados existentes en la literatura Científica.

### *3.1.1 Visión general sobre ADL más reconocidas.*

En la Tabla 9, se observa que existen diferentes ALD como son: Arquitectura de zona única, Arquitectura por estanques, y Arquitectura de Zonas, las variantes de estas arquitecturas se resumen como; características, actividades, tipos de datos, tecnologías, usuarios, que forman parte de cada una de las diferentes zonas de las ALD propuestas hasta la actualidad. Esto nos permitió ver con claridad que no existe en detalle la forma de implementar y ejecutar cada una de las zonas.

La arquitectura plana propuesta por (Dixon, 2010), es la que da el inicio a que se desprendan los demás estudios en el año 2010, según (A Laurent et al., 2020) fue introducido por primera vez el término LD, mediante lo cual solo se pretendía que fueran enormes conjuntos de datos estructurados y no estructurados y que los usuarios pudieran hacer actividades de acceder para trabajar en minería, muestreo o análisis de hecho la única tecnología que manifiesta o propone es Hadoop.

En (C Giebler et al., 2020) se dice que la arquitectura de estanques no cumple con el concepto de LD, por contener cinco estanques disjuntos, esto hace que a medida que se vayan procesando los datos vayan perdiendo información original, lo cual hace que quede fuera del análisis de evaluación de los modelos.

A partir de la cual podríamos decir que el siguiente conjunto son las arquitecturas que se han basado en zonas múltiples.

Según (Ravat & Zhao, 2019) plantea la evolución de diferentes arquitecturas de LD, de una zona única a una zona múltiple, también hace hincapié a que no existe una arquitectura de LD reconocida.

En (C Giebler et al., 2020) se confirma que los modelos de un de una ALD, son vagos y sin valoraciones, lo cual no deja claro cuál de los modelos de zona son aplicables en la implementación práctica de un ALD en las empresas que para nuestro caso sería para datos Académicos.

En (Chihoub et al., 2020) define las características más importantes de los lagos de datos, esta visión da inicio para describir una arquitectura que se basa en una arquitectura genérica, y se muestra como una propuesta estándar, la misma que contiene las siguientes capas:

1. La capa de ingestión;
2. La capa de almacenamiento
3. La capa de transformación
4. La capa de interacción

Se ha descartado las arquitecturas propuestas por (Dixon, 2010), (Inmon, 2016) (Nadipalli, 2017b)[14] (Sharma, 2018) ya que son consideradas en el análisis de evolución de las zonas en (Megdiche et al., 2020), (Ravat & Zhao, 2019); se descartó también la arquitectura propuesta por (Pegdwendé Sawadogo & Darmont, 2021) ya que esta agrupa a las arquitecturas según los componentes.

Se ha considerado la ALD y su concepto propuesto por (Ravat & Zhao, 2019), el modelo de referencia citado en (C Giebler et al., 2020) que contiene casos de uso, en sus

diversos requisitos, el modelo de las características de los datos y atributos de zona, en (Chihoub et al., 2020), se considera las capas estándares de que debe contener una ALD.

En (Hamadou et al., 2020) se revisó los requisitos funcionales y no funcionales que debe contener una ALD, las características propuestas por (M Huchard et al., 2020) de una ALD aplicado a las líneas de producto de software enfoque, propuestos.

### 3.2 Fase de evaluación

En esta fase se realizó la evaluación del prototipo en base a lo expuesto anteriormente. En el análisis de la literatura se ha encontrado diversos marcos ALD, pero se ha considerado aquellos modelos que se encuentran con más relevancia, ajustados a las características, definición y caso de estudio, que están considerados en la literatura científica, la ALD propuesta en (Ravat & Zhao, 2019), y el modelos de zona(C Giebler et al., 2020).

Porque utilizan la mayoría de las características, y son versiones mejoradas de los modelos anteriores.

Me he basado en un enfoque(M Huchard et al., 2020), sobre ingeniería de líneas de producto de software, aplicado al modelado de características ALD y modelo entidad relación de atributos de zona representado en (C Giebler et al., 2020)

Para realizar esta evaluación de la ALD, se considera el concepto tomado en (Ravat & Zhao, 2019), las características de la ALD(Chihoub et al., 2020), por ser un estándar basado en una arquitectura genérica, relacionando la entrada, el proceso, la salida y gobernanza de un LD, para el caso de estudio datos académicos UTPL.

En ALD (Ravat & Zhao, 2019), se enfoca en presentar una arquitectura genérica, a través de un caso de uso para el cuidado de la salud, mediante la aplicación de la gestión de metadatos y la mayoría de modelos buscan en sus propuestas proponer la gobernanza de metadatos como algo esencial, para que no se genere los Pantanos de Datos.

Su modelo presenta la Zona de datos crudos, Zona de datos procesados, Zona de acceso, Zona de gobernanza.

La Zona de Gobernanza se aplica en todas las demás Zonas y está a cargo de asegurar la seguridad de los datos, la calidad de los datos, el ciclo de vida de los datos, el acceso a los datos y la gestión de los metadatos.

### 3.3 Fase de Selección de un prototipo Arquitectura Data Lake

En (C Giebler et al., 2020) su planteamiento hace un metamodelo de instancias posibles de acuerdo a cada zona, donde demuestra su interacción con las demás zonas así como sus características y atributos, en su esquema de marco de ALD, no se muestra con claridad el manejo de administración de metadatos aunque aplica la técnica de Data Vautl como modelado de datos(Corinna Giebler et al., 2019) en las zonas Armonizada y Destilada dependen del caso de estudio, lo cual es muy bueno, pero dificulta no tener desarrollada la implementación de cada zona.

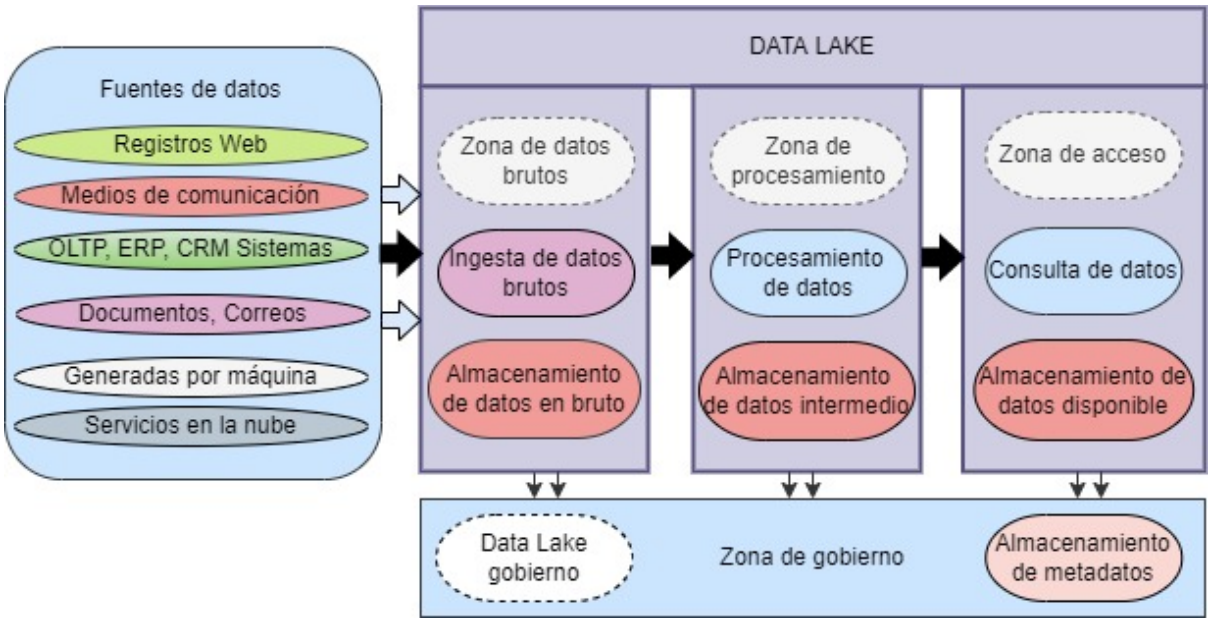
Según el análisis realizado por (Pegdwendé Sawadogo & Darmont, 2021) las Zonas corresponden al diseño de la parte arquitectónica de una ALD, estas zonas se gobiernan de acuerdo a la administración de los metadatos.

De acuerdo a estos mismo autores el enfoque de la propuesta de(Ravat & Zhao, 2019), pertenece a las arquitecturas de tipo Híbridas, las mismas que combinan las arquitecturas basadas en la madurez de los datos y las arquitecturas funcionales que tienen claras las funciones a implementar en una ALD.

En nuestra propuesta nos interesa el enfoque de ALD realizado en (Ravat & Zhao, 2019), por ajustarse a la definición de conceptos emergentes, características estándares, administración de los metadatos, gobernanza en todas las zonas y un caso de estudio que se alinea a la gestión de datos académicos.

**Figura 8**

*Arquitectura Lago de Datos*




*Nota:* Adaptado de (Ravat & Zhao, 2019)

Se ha considerado el metamodelo de zonas que describe una zona como un diagrama ER. de zonas propuesto en (C Giebler et al., 2020).

La Administración de metadatos para un ALD propuesta en (Pegdwendé Sawadogo & Darmont, 2021) presenta una identificación de dos tipologías principales de metadatos dedicados a lagos de datos: Metadatos funcionales y estructurales.

Por ser datos de orden académicos se considera la categoría que representa a los metadatos funcionales por encontrarse dentro del grupo de metadatos comerciales, metadatos operativos, metadatos técnicos.

A pesar que existen diferentes tecnologías que pueden ser utilizadas, muchos autores citan la tecnología Hadoop para la implementación de LD, pero no necesariamente es la única tecnología, a continuación de acuerdo a las funciones se especifican algunas herramientas.

La propuesta de las diferentes herramientas han sido tomadas de (Couto et al., 2019), como las más citadas en los artículos científicos.

**Ingestión de datos:** Tecnologías de Fundación Apache, también pueden servir para agregar, convertir y limpiar datos antes de la ingestión. Flink y Samza, Flume, acciones de adquisición y recolección Kafka la más citada, sqoop. Protocolos para la transferencia de datos comunes (wget, rsync, FTP, HTTP, etc.).

**Almacenamiento:** Algunas tecnologías de herramientas tiene como tarea integrar, normalizar datos, Hadoop, Apache Casandra y MongoDB, son las más citadas por los autores.

**Procesamiento:** Analizar, procesar y transformar los datos para poder sacar información valiosa de ellos. Apache Spark con mayor acogida, también se suma a la lista Apache Hadoop.

**Presentación:** Tecnologías Microsoft Power BI y Tableau como los más citados.

# Capítulo cuatro

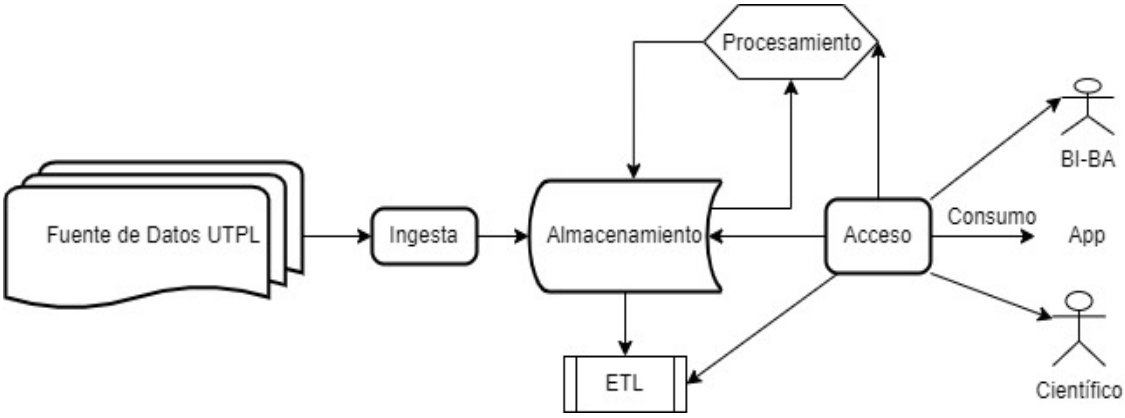
# Desarrollo

El presente capítulo abarca la descripción del marco del diseño arquitectónico que corresponde a las funciones, requisitos, procesos del ALDA, atributos de cada zona, así como el conocimiento de las tecnologías a utilizar en la aplicación del caso de estudio.

A partir de haber identificado en la fase anterior una arquitectura genérica (Ravat & Zhao, 2019) se establece las funciones establecidas en el proceso detalladas en la Figura 9 el diseño, funcionamiento y composición de cada una de las zonas, se detalla más adelante en el apartado de propuesta de la ALDA y hace referencia a las actividades, datos, tecnologías, usuarios, considerados en la Tabla 9 y sus posibles relaciones.

**Figura 9**

*Flujo de datos Funciones ADLA*




Las principales funciones del Caso de Estudio Datos Académicos UTPL, se describen a continuación.

## 4.1 Fuente de datos UTPL

El origen de los datos de gestión académica son de procedencia interna y externa, con estructura primarias y secundarias.

- **Datos estructurados:** Bases de datos relacionales, sensores, formularios, registros de red, Excel, etc.

- **Datos semiestructurados:** XML, archivos comprimidos, protocolos, almacenamiento en NoSql, JSON
- **Datos no estructurados**: Word, pdf, imagen, videos, páginas web, social media, etc.

## 4.2 Ingesta

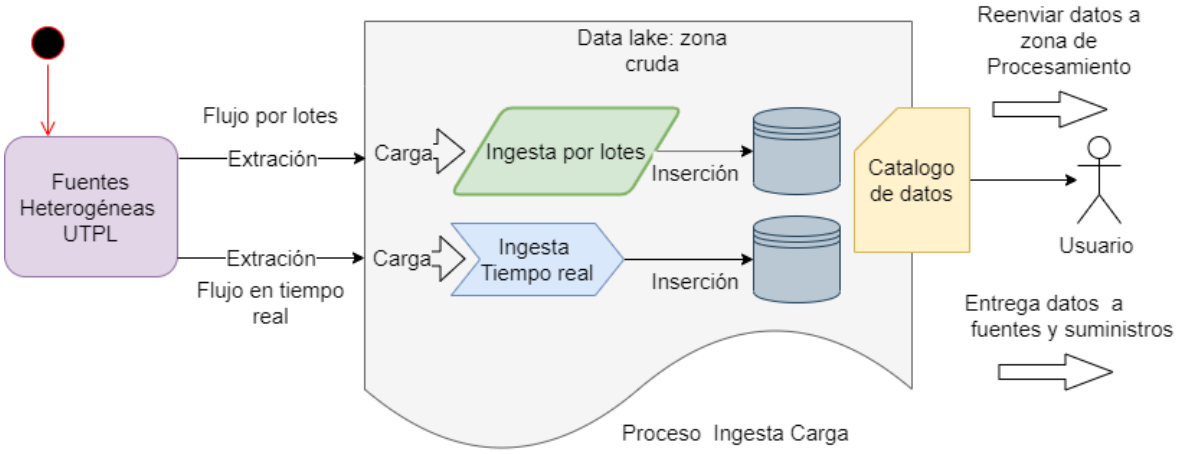
La inserción de los diversos tipos de datos es una función que se realiza mediante flujo por lotes o en tiempo real, se consigue para extraer desde diferentes dominios de fuentes de datos y se efectúa mediante ingesta gestionada. La información debe permitir proporcionar algunos metadatos, linaje, limitar las restricciones del acceso.

Las fuentes de datos de la UTPL son heterogéneas, una vez que se ha localizado el origen de los datos internos y eternos, se procede a su posterior carga, e ingestión, el proceso se muestra en la Figura 10

El estado de los datos debe ser en estado crudo y se ingesta en la primera Zona del DL, denominada Zona de Datos Crudos.

**Figura 10**

*Proceso Ingesta DLA*




### *4.2.1 Flujo por lotes*

Datos Académicos, ERP, CSV, CRM, XML, csv, tecnologías a utilizar, U-Sql Pig, Hive, Spark y MapReduce.

### *4.2.2 Flujo de transmisión en tiempo real*

Social media UTPL, tecnología a utilizar, Spark-Streaming, Kafka, Flume, Storm, Sqoop.

### *4.2.3 Proceso Extracción, Carga, Transformación(ECT)*

El proceso a seguir es de Extracción, Carga, transformación, donde los datos, se cargan en su formato original, con una administración gestionada.

### *4.2.4 Inserción*

Almacenamiento de datos con modelo de esquema en lectura, en este caso de estudio se propone el almacenamiento en tres zonas, zona de datos crudos o sin procesar, en zona de datos procesados, zona de datos de acceso, según (Pegdwendé Sawadogo & Darmont, 2021) el almacenamiento puede ser utilizando, sistema de bases de datos relacionales DBMS relacionales, una segunda el almacenamiento distribuido. Siendo las más citadas las siguientes tecnologías HDFS, tablas Apache hive, HBase, MapR, Apache Casandra, MongoDB, MySQL, PstgreSQL, Oracle.

### *4.2.5 Proceso*

Transformación de datos a formas estandarizadas, al igual su procesamiento se puede dar por lotes; el procesamiento por lotes se puede llevar a cabo mediante procesos Extracción, Transformación, Carga(ETC), como tecnología más utilizada citadas Apache Spark.

### *4.2.6 Catálogo de datos:*

Según(Sharma, 2018) es la forma de llevar una ingesta gestionada, que permiten la adquisición, almacenamiento, provisión y análisis de metadatos técnicos, funcionales y operativos, en (Groger & Hoos, 2019)revela tres tipos de herramientas para la gestión de metadatos en el lago de datos que representan el mercado: directorios de datos integrados en el sistema, catálogos de datos y plataformas de gestión de lago de datos, herramientas para catálogo de datos, Apache Atlas.

## 4.3 Almacenamiento

Se puede revisar los accesibles servicios en la nube como son Azure, Google Cloud, Amazon AWS, ya que permite una gran oportunidad de opciones para los diversos tipos de almacenamiento, conociendo que su precio depende del consumo, en sus requisitos de accesibilidad. Según los requisitos, los datos se colocan en hadoop hdfs, hivem hbase, elastic search o en memoria, gestión de metadatos también se facilita suspensión de datos basada en políticas que tiene como actividades e integrar, normalizar datos, hadoop, apache casandra y mongodb, son las más citadas por los autores.

## 4.4 Requerimientos y procesos ADLA

La arquitectura se basa en los requerimientos de la unidad académica, así como los requerimientos que se pueden manejar en un LD, considerando así los requisitos funcionales, no funcionales y procesos de orquestación, que se suscitan de los requisitos como capas.

### *4.4.1 Requisitos funcionales*

Las principales acciones que correspondan a la obtención de datos como entradas, flujos de datos, salidas, identificar los metadatos dentro de todos los datos cargados en las diferentes zonas, la gobernanza de los metadatos, la confidencialidad de datos sensibles, gestión de acceso a usuarios autorizados a cierta información (Hamadou et al., 2020), depende de las políticas establecidas en cada dominio de estudio presentes en cada caso de uso de la UTPL.

### *4.4.2 Requisitos no funcionales:*

Las características internas y esenciales en la Arquitectura, implica que su diseño e implementación sea mediante la ingestión gestionada de datos en tiempo real y por lotes, utilizando infraestructura de instalaciones en la nube, con herramientas de código abierto, que aseguren una escalabilidad, seguridad, disponibilidad, flexible para integrar fácilmente herramientas(Hamadou et al., 2020), a medida de las necesidades de ALDA.

### *4.4.3 Procesos de flujo de Zonas.*

Según (C Giebler et al., 2020) cada zona puede describirse mediante un conjunto limitado de atributos, los mismos que se describen en la Tabla 9 y se aplicara su relación para el funcionamiento de cada zona, aplicado al caso de estudio gestión de datos académicos UTPL.

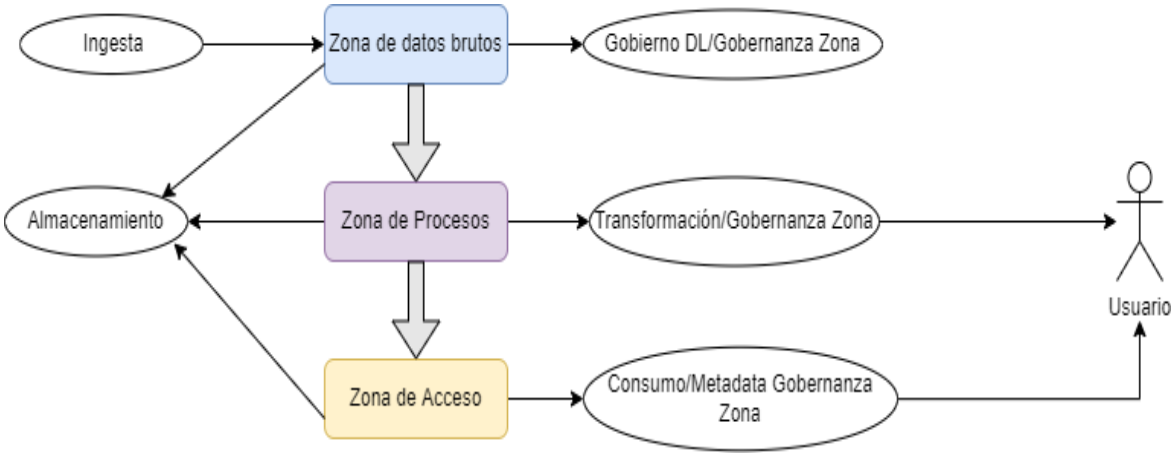
### *4.4.4 Orquestación de Zonas y Capas ADLA*

En esta propuesta he clasificado las zonas de acuerdo a lo que se ha encontrado en la literatura general para hacer referencia a la ALDA, aplicada en el caso de estudio Gestión de datos Académicos de la UTPL.

**Tres principales zonas más reconocidas**: Zona de datos brutos, zona de procesamiento, zona de acceso, estas zonas son centrales y hasta cierto punto se combinan en estados híbridos, como técnicas, funcionales, y de referencia pero son las zonas más genéricas como se muestra en la Figura 11.

**Figura 11**

*Zonas principales DLA*




En (Hamadou et al., 2020), se presentan cinco capas separadas, fuentes de datos, recopilación, almacenamiento, exploración, consumidores y cuatro capas más transversales, Gestión de Acceso, Gobernanza de metadatos, privacidad y anonimización(GDPR)(Voigt & Bussche, 2017) y Administración de recursos.

Capas externas: Fuentes de datos, Consumidores de datos.

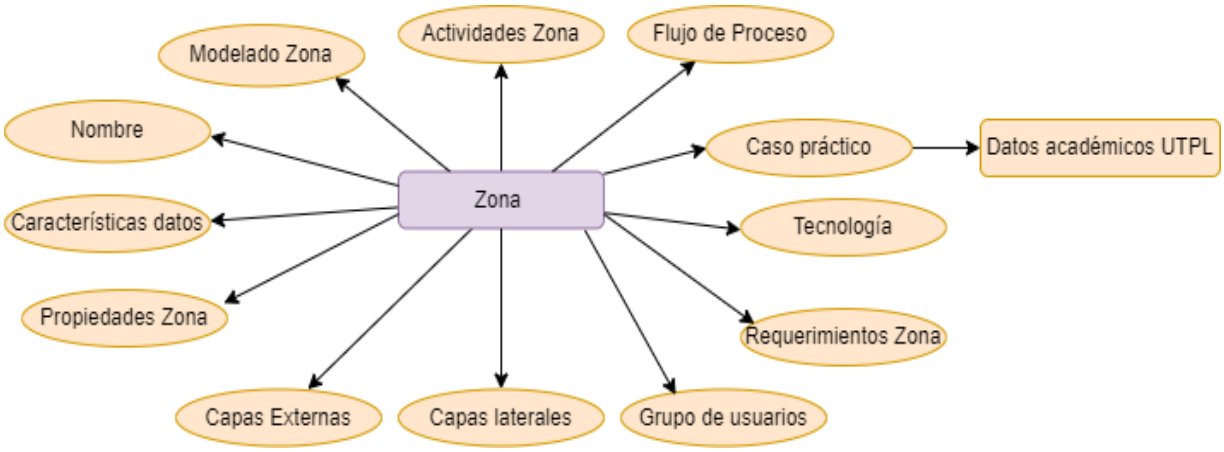
Capas transversales: Gestión de Acceso, Gobernanza de metadatos, Privacidad y anonimización(GDPR)(Voigt & Bussche, 2017), Administración de recursos.

### *4.4.5 Atributos de zonas*

En (C Giebler et al., 2020) se conceptualiza la zona mediante un conjunto limitado de atributos (por ejemplo, el grado de procesamiento de los datos contenidos) y sus interacciones con otras zonas y el mundo exterior (por ejemplo, de dónde importa los datos), comprendiendo que un Lago de Datos contiene también la Gobernanza, Confidencialidad, gestión de acceso, administración de recursos en cada zona, y trayendo a consideración también lo detallado en Tabla 9, se ha construido el modelo de interacción de cada zona el mismo que se muestra en la siguiente Figura 12.

**Figura 12**

*Iteración de cada Zona*




## 4.5 Planteamiento de ALDA

La función más importante de un lago de datos es el análisis predictivo y avanzado, el presente LD, se organiza en función del caso de estudio de Datos Académicos UTPL, un LD, su característica principal es conservar sin exclusión alguna todos los datos de origen, el prototipo se muestra en la Figura 13.

Se compone de tres zonas internas las mismas que son Zona de datos Crudos, Zona de Procesos, Zona de Acceso a los Datos(exploración), el almacenamiento de datos se considera en cada una de estas zonas y se lo representa como una función del LD, dos capas entrada y salida, Fuente de datos, Consumo de datos, una capa determinadas como función de ingesta, y las capas laterales que son consideradas para las capas de funciones así como las capas de zonas internas, Gestión de Acceso, Gobernanza de Metadatos, Privacidad y Anonimato(GDPR)(Voigt & Bussche, 2017)

### *4.5.1 Zona de datos Crudos*

Esta zona vendría a ser donde se cargan todos los datos sin ser procesados, se encarga principalmente de la recopilación de los datos extraídos de las fuentes de datos UTPL. Posteriormente serán almacenados, gestión de gobierno del LD, los datos almacenados pueden contener diferente formato que el de origen.

### *4.5.2 Zona de Procesos*

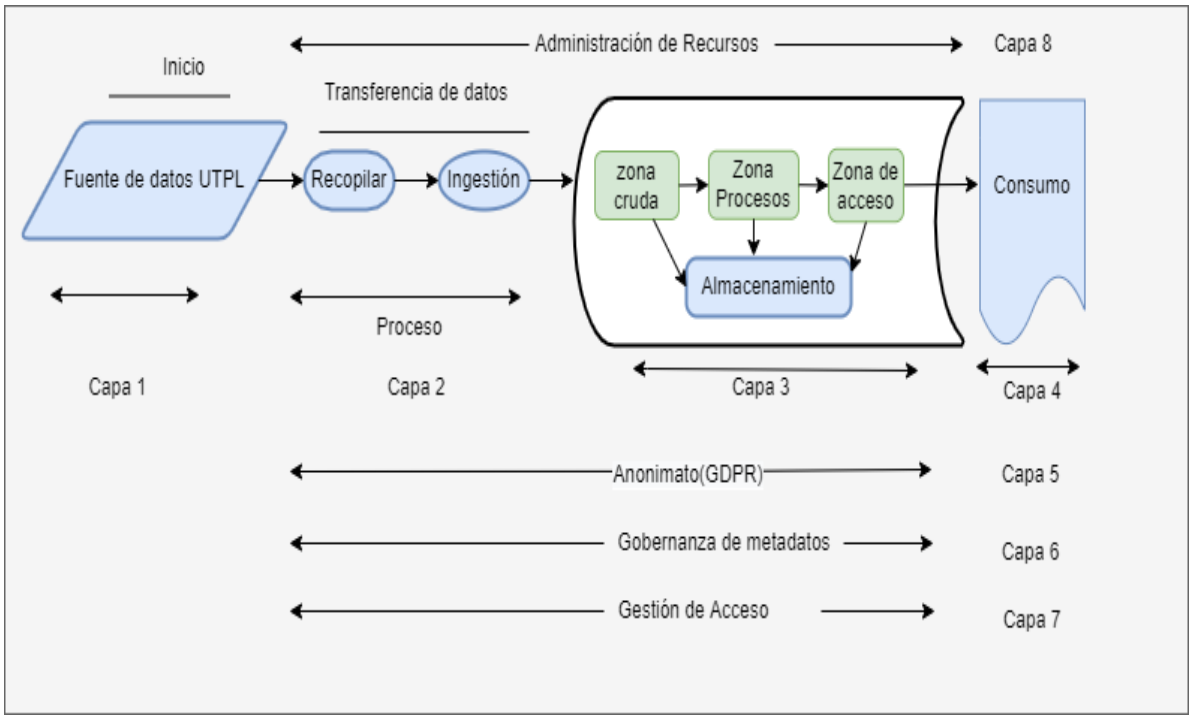
Aquí se considera acceder a los datos del caso de estudio dominio estudiantes UTPL, Los usuarios que mantienen acceso a esta zona para este caso de estudio pueden ser analista de datos, científicos de datos y pueden hacer actividades de selección, proyección, unión, agregación, se almacenan los datos.

### *4.5.3 Zona de Acceso*

Aquí se realizan las actividades de informes, análisis, la zona de acceso almacena todos los datos disponibles para el análisis de datos y proporciona el acceso a los datos. Esta zona permite el consumo de datos de autoservicio para diferentes analíticas (informes, análisis estadístico, análisis de inteligencia empresarial, algoritmos de aprendizaje automático).

**Figura 13**

*Diseño de Propuesta Arquitectura Lago de Datos Académico*




**Atributos de Zona de Datos Crudos:**

- **Actividad de Zona:** Función de Ingesta, Función Almacenamiento
- **Capa Externa:** Fuente de datos
- **Capa Lateral:** Gestión de metadatos, Confidencialidad, Gestión de Acceso, Administración de recursos.
- **Características de datos:** Granularidad crudo, Esquema cualquiera, Sintaxis Básico, Semántica sin cambios.
- **Propiedades Zona:** Gobernada, historiado, no especializada, no persistente, parte protegida, caso de uso general datos académicos, grupo de usuarios sistemas, modelado cualesquiera.
- **Usuarios:** Sistemas, procesos
- **Transferencia de Datos entre zonas:** Zona de proceso, Función Ingesta.

### *4.5.4 Gobernanza de Metadatos*

Esta capa soporta la comprensión del ciclo de vida de los datos, para percibir los procesos de los datos, en la ingestión, almacenamiento y procesamiento, acceso a procesos, se considera la tecnología como Apache Atlas.

### *4.5.5 Gestión de Acceso*

Esta capa permite administrar el acceso para orquestar varias zonas y capas que solo usuarios autorizados tengan acceso a los datos, que permita manejar la confidencialidad de los datos, es necesario herramientas con reglas claras para la autorización, Apache Ranger.

### *4.5.6 Gestión de Recursos*

Es necesario orquestar los recursos que se contienen dentro de la ALDA, esta capa permite que todas las capas se ejecuten en todo momento y se mantenga una buena administración con el uso de recursos, una tecnología más usada Apache Hadoop YARN.

## Conclusiones

En este trabajo propuse una Arquitectura Lago de Datos Académico, caso de estudio UTPL. Lo más importante del lago de datos es que permite la solución a la gestión de datos heterogéneos porque permiten la ingesta unificada de información evitando así los silos de información al mantener centralizada la información. Además permite darle un valor adicional a la analítica de datos por permitir sacar resultados homogéneos, a bajos costos. Los trabajos relacionados con la educación e industria fueron relevantes porque permitieron conocer los diferentes aplicaciones que se le puede dar a un lago de datos. Lo más arduo fue proponer los requisitos que pueda cumplirse para cada capa y zonas.

Se analizo las principales propuestas sobre Arquitectura Lago de Datos, mediante las cuales se pudo identificar y conocer su estado actual, aplicabilidad, componentes y factores en general entre las soluciones de Lagos de Datos planteadas. Lo más relevante es que permitió descubrir tipologías esenciales para el conocimiento de los lagos de datos porque mediante este análisis se permitió conocer su historia, alcance, principales característica.

Lo que mejor ayudo a comprender los lagos de datos fue sus componentes, zonas y aplicabilidad, porque existen marcos con diferentes propuestas pero mantienen los componentes de la zona, las dificultades presentes en estos marcos fue no existir arquitecturas por definición, los procesos presentes en las zonas son ligeros y poco comprensibles porque sus evaluaciones son inexistentes lo cual dificulta su aplicabilidad en un caso de estudio.

Se evaluó una Arquitectura Lago de Datos en base a los factores existentes lo que más me beneficio fueron las características estándares de un lago de datos porque contienen procesos de entrada, procesamiento y salida para la gestión de datos heterogéneos, considerando como un enfoque del modelo a seguir la arquitectura propuesta por (Ravat & Zhao, 2019) porque contiene los diferentes factores analizados. Lo que complica a esta arquitectura seleccionada es la complejidad de diferentes requisitos para el caso de estudio de la UPL, porque no define con claridad como aplicar los accesos a usuarios, administración de recursos, protección de datos.

Se diseño un prototipo para gestionar datos Heterogéneos Académicos UTPL, lo más notable en este diseño fue llevar a cabo una arquitectura compuesta por zonas, capas y atributos de zona porque es fundamental para el caso de estudio de la UTPL, lo que contribuyó a estos resultados fue el análisis de las diferentes tipologías, porque se permitió conocer la aplicabilidad en similares casos de estudio.

## Recomendaciones

Se recomienda utilizar ingesta gestionada de los datos ya que el mayor problema de los lagos de datos es el estanque de datos, citados en la mayoría de los artículos, lo cual se lo puede superar con una, gestión y gobernanza de metadatos adecuado, se sugiere además implementar esta ALDA, mediante tecnologías de almacenamiento en la nube, que permitan la creación de un catálogo de datos, a través de las herramientas mencionadas en este estudio.

Se debe considerar los trabajos que ahondan más en temas del tratamiento de la información para la administración de diferentes zonas, poder identificar con claridad los procesos a seguir en la aplicación de modelos de zonas y modelados de datos.

Plantearse como un reto a futuro la implementación de las diferentes zonas y capas que permita el manejo de la ingesta de la información gestionada para obtener un catálogo de datos, de la UTPL.

**Referencias.**

Chihoub, H., Madera, C., Quix, C., & Hai, R. (2020). Architecture of Data Lakes. In *Data Lakes* (pp. 21–39). Wiley. https://doi.org/10.1002/9781119720430.ch2

Couto, J., Borges, O., Ruiz, D. D., Marczak, S., & Prikladnicki, R. (2019). A Mapping Study about Data Lakes: An Improved Definition and Possible Architectures. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, *2019-July*, 453–458. https://doi.org/10.18293/SEKE2019-129

Cravero, A., Lefiguala, I., Tralma, R., & Gonzalez, S. (2020). *Data Lake architecture proposal for the Analysis Directorate of a Regional University*. *2020-November*. https://doi.org/10.1109/SCCC51225.2020.9281154

de Zubiría Ragó, A., & de Zubiría Samper, M. (2019). *Pedagogía conceptual: una puerta al futuro de la educación*. Ediciones de la U. https://bit.ly/37z1eyB

Dixon, J. (2010). *Pentaho, Hadoop y lagos de datos | Blog de James Dixon*. https://bit.ly/3KTnXUB

Giebler, C, Groger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2020). *A Zone Reference Model for Enterprise-Grade Data Lake Management*. 57–66. https://doi.org/10.1109/EDOC49727.2020.00017

Giebler, Corinna, Groger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2020). A Zone Reference Model for Enterprise-Grade Data Lake Management. *2020 IEEE 24th International Enterprise Distributed Object Computing Conference (EDOC)*, 57–66. https://doi.org/10.1109/EDOC49727.2020.00017

Giebler, Corinna, Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Modeling Data Lakes with Data Vault: Practical Experiences, Assessment, and Lessons Learned. In A. H. F. Laender, B. Pernici, E.-P. Lim, & J. P. M. de Oliveira (Eds.), *Conceptual Modeling* (pp. 63–77). Springer International Publishing.

Goli, A. (2020). *Comparación - Base de datos vs DataMart vs Data Warehouse vs Data*. https://bit.ly/3Lb7uvh

Groger, C., & Hoos, E. (2019). Ganzheitliches metadatenmanagement im data lake:

Anforderungen, it-werkzeuge und herausforderungen in der praxis. *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft Fur Informatik (GI)*, *P-289*, 435–452. https://doi.org/10.18420/btw2019-26

Hamadou, H. B., Bach Pedersen, T., & Thomsen, C. (2020). The Danish National Energy Data Lake: Requirements, Technical Architecture, and Tool Selection. *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, 1523–1532. https://doi.org/10.1109/BigData50022.2020.9378368

Huchard, M, Laurent, A., Libourel, T., Madera, C., & Miralles, A. (2020). Exploiting software product lines and formal concept analysis for the design of data lake architectures. In *Data Lakes* (pp. 41–56). wiley. https://doi.org/10.1002/9781119720430.ch3

Huchard, Marianne, Laurent, A., Libourel, T., Madera, C., & Miralles, A. (2020). Exploiting Software Product Lines and Formal Concept Analysis for the Design of Data Lake Architectures. In *Data Lakes* (pp. 41–56). Wiley. https://doi.org/10.1002/9781119720430.ch3

Inmon, B. (2016). *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Technics Publications.

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews, Version 1.0. *Empirical Software Engineering*, *33*(2004), 1–26.

Krishnan, K. (2020). Big Data introduction. In *Building Big Data Applications* (pp. 1–16). Elsevier. https://doi.org/10.1016/B978-0-12-815746-6.00001-6

Laurent, A., Laurent, D., & Madera, C. (2020). Data Lakes. In Anne Laurent, D. Laurent, & C. Madera (Eds.), *Data Lakes*. Wiley. https://doi.org/10.1002/9781119720430

Laurent, A, Laurent, D., & Madera, C. (2020). Introduction to data lakes: Definitions and discussions. In *Data Lakes* (pp. 1–20). wiley. https://doi.org/10.1002/9781119720430.ch1

Laurent, Anne, Laurent, D., & Madera, C. (2020). Introduction to Data Lakes: Definitions and Discussions. In *Data Lakes* (pp. 1–20). Wiley. https://doi.org/10.1002/9781119720430.ch1

Megdiche, I., Ravat, F., & Zhao, Y. (2020). A use case of data lake metadata management. In *Data Lakes* (pp. 97–122). wiley. https://doi.org/10.1002/9781119720430.ch5

Munshi, A. A., & Alhindi, A. (2021). Big Data Platform for Educational Analytics. *IEEE Access*, *9*, 52883–52890. https://doi.org/10.1109/ACCESS.2021.3070737

Nadipalli, R. (2017a). *Effective Business Intelligence with QuickSight*. Packt Publishing. https://bit.ly/3wh3dlG

Nadipalli, R. (2017b). *Effective Business Intelligence with QuickSight*. Packt Publishing.

Ravat, F., & Zhao, Y. (2019). Data Lakes: Trends and Perspectives. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11706 LNCS* (pp. 304–313). https://doi.org/10.1007/978-3-030-27615-7_23

Sawadogo, P, & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), 97–120. https://doi.org/10.1007/s10844-020-00608-7

Sawadogo, Pegdwendé, & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), 97–120. https://doi.org/10.1007/s10844-020-00608-7

Sharma, B. (2018). *Architecting Data Lakes*.

Torres, P., Gonzalez Gonzalez, C. S., Aciar, S., & Rodriguez Morales, G. (2018). Methodology for systematic literature review applied to engineering and education. *IEEE Global Engineering Education Conference, EDUCON*, *2018-April*(April), 1364–1373. https://doi.org/10.1109/EDUCON.2018.8363388

Voigt, P., & Bussche, A. von dem. (2017). *The EU general data protection regulation (GDPR) : a practical guide*. 385.