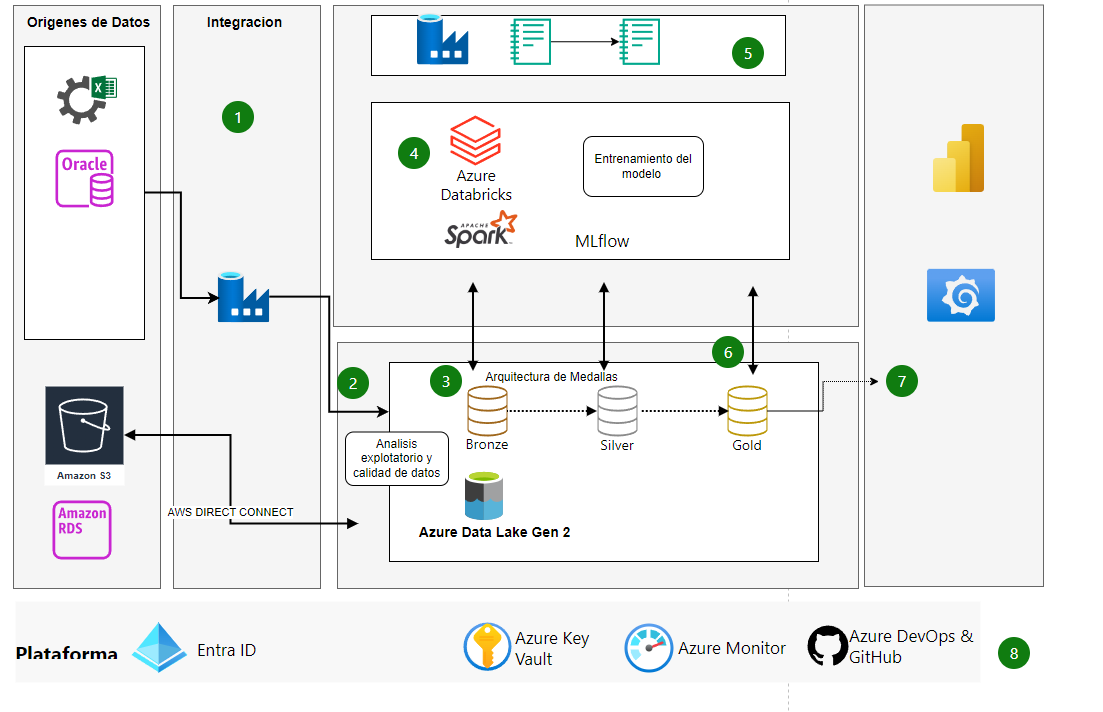
**Planos Técnicos**



***Paso 1:*** Integración de Orígenes y Estrategia de Almacenamiento Multicloud, En esta fase inicial del proyecto, se establece la base de la arquitectura mediante la implementación de mecanismos de integración robustos y flexibles que permitan capturar, transformar y mover datos desde múltiples fuentes hacia un entorno analítico centralizado. Este proceso se realiza utilizando Azure Data Factory (ADF) como componente principal de orquestación de flujos de datos, junto con una estrategia de almacenamiento multicloud que garantiza interoperabilidad, escalabilidad y continuidad operativa.

***Paso 2****:* Análisis Exploratorio, Procesamiento de Calidad y Arquitectura de Medallas,En esta segunda fase del proyecto, se realiza el análisis exploratorio de datos (EDA) como etapa clave para evaluar la calidad, comprender el comportamiento de las variables y preparar los datos para su posterior análisis avanzado y explotación empresarial. Esta fase sigue una metodología estructurada basada en la arquitectura de medallas (bronce, plata, oro), con el respaldo de un lago de datos centralizado como repositorio confiable y escalable.

Análisis Exploratorio y Procesos de Calidad

El análisis exploratorio se ejecuta con herramientas como Azure Synapse, Databricks y notebooks de Python o Spark, e incluye:

Perfilamiento de datos: detección de nulos, duplicados, outliers, sesgos y formatos inconsistentes.

Validaciones y reglas de negocio: establecimiento de controles de calidad (por ejemplo, rangos permitidos, valores requeridos, consistencia entre campos).

Trazabilidad y documentación: registro de incidencias de calidad, metadatos descriptivos y origen de cada conjunto de datos.

Beneficios del procesamiento de calidad:

* Aumenta la confianza en la analítica y modelos predictivos.
* Permite detectar errores en origen antes de que escalen al usuario final.
* Mejora la gobernanza y la transparencia en los procesos de toma de decisiones.
* Reduce retrabajos y evita errores en reportes o automatizaciones.

***Paso 3,*** Metodología de Arquitectura de Medallas

El procesamiento se organiza según la arquitectura de medallas, un enfoque que categoriza los datos según su nivel de transformación y confiabilidad:

| **Nivel** | **Descripción** |
| --- | --- |
| **Bronce** | Datos crudos tal como se ingieren, sin modificaciones. Se conservan para auditoría y trazabilidad. |
| **Plata** | Datos depurados, normalizados, con control de calidad aplicado. Son consistentes y listos para consumo analítico. |
| **Oro** | ***Paso 6,*** Datos enriquecidos, agregados o modelados para casos de uso específicos (BI, IA, reporting, etc.). |

Beneficios de esta metodología:

* Fomenta la modularidad y escalabilidad de los pipelines.
* Garantiza la reproducibilidad y auditoría en cada etapa del procesamiento.
* Permite segmentar accesos según niveles de madurez del dato.
* Optimiza el rendimiento al reutilizar transformaciones comunes.

Almacenamiento en Lago de Datos

Todos los datos procesados bajo esta metodología se almacenan en un Data Lake centralizado sobre Azure (Data Lake Storage Gen2), con arquitectura desacoplada y compatible con múltiples motores de análisis.

Beneficios del lago de datos:

* Escalabilidad horizontal para grandes volúmenes y variedad de datos.
* Separación de cómputo y almacenamiento, optimizando costos y desempeño.
* Soporte para datos estructurados, semiestructurados y no estructurados.
* Integración nativa con servicios como Power BI, Synapse, Databricks y Purview.
* Base para implementar catálogos, linaje y control de versiones.

***Paso 4:*** Procesamiento Avanzado y Modelado con Notebooks en Azure Databricks

En esta etapa, se sugiere la habilitacion un entorno de trabajo colaborativo y altamente escalable para el procesamiento avanzado de datos, la creación de modelos analíticos, y la integración de inteligencia artificial a través de **notebooks interactivos en Azure Databricks**, que combinan poder de computo, flexibilidad en el desarrollo y soporte para ciencia de datos e ingeniería avanzada.

**Notebooks para Limpieza y Preparación de Datos**

Azure Databricks nos permitirá al equipo de datos construir notebooks en Python, Scala o SQL donde se pueden:

* Ejecutar procesos complejos de **limpieza, transformación y normalización de datos**.
* Integrar validaciones de calidad avanzadas (reconciliación, consistencia temporal, outliers, etc.).
* Aplicar reglas de negocio personalizadas y crear datasets listos para modelado.
* Reutilizar notebooks como parte de pipelines automatizados con integración a Azure Data Factory.

*Beneficios:*

* Ambiente colaborativo con control de versiones y comentarios por celda.
* Ejecución distribuida sobre clústeres escalables (Spark).
* Reutilización de código como módulos y jobs programados.
* Mayor trazabilidad y gobernanza del proceso analítico.

**Construcción de Modelos Analíticos**

Los mismos notebooks se utilizan para entrenar, evaluar y desplegar modelos de machine learning:

* Uso de bibliotecas como **MLlib, scikit-learn, XGBoost, LightGBM o TensorFlow**.
* Entrenamiento de modelos predictivos (regresión, clasificación, series de tiempo).
* Hiperparámetro tuning y validación cruzada integrada.
* Registro de modelos en **MLflow**, con seguimiento de versiones, métricas y datasets usados.

*Beneficios:*

* Plataforma unificada para ingeniería y ciencia de datos.
* Tiempos de entrenamiento reducidos gracias a procesamiento distribuido.
* Integración directa con Azure Machine Learning para despliegue en producción.

***Paso 5*, Orquestación de Notebooks en Azure Databricks**

Mediante el uso de pipelines en ADF, se pueden ejecutar notebooks alojados en Azure Databricks como parte de un flujo ETL/ML. Cada notebook representa una etapa clave del ciclo de vida de los datos:

1. **Carga e Ingesta de Datos**
   * Notebooks diseñados para leer datos desde múltiples fuentes (bases SQL, S3, APIs, archivos planos).
   * Aplicación de estructuras temporales o staging en el lago de datos (nivel bronce).
2. **Procesamiento y Calidad**
   * Notebooks que aplican transformaciones, reglas de negocio y validaciones.
   * Generación de datasets consistentes, depurados y documentados (nivel plata).
3. **Modelado y Predicción**
   * Entrenamiento, evaluación y despliegue de modelos predictivos o de LLMs.
   * Exportación de resultados a dashboards o sistemas operacionales (nivel oro)

**Beneficios Técnicos**

| **Categoría** | **Beneficio** |
| --- | --- |
| **Escalabilidad** | Servicios como Azure Data Lake, Synapse y Databricks escalan bajo demanda para soportar grandes volúmenes de datos. |
| **Interoperabilidad** | Azure Data factory facilita la integración con múltiples fuentes de datos (on-premise, cloud, IoT, SaaS, etc.). |
| **Procesamiento en Tiempo Real** | Uso de Azure Stream Analytics, Event Hubs y Kafka para ingerir y procesar datos en tiempo real. |
| **Alta Disponibilidad y Resiliencia** | Servicios distribuidos con replicación geográfica y recuperación ante fallos. |
| **Seguridad Avanzada** | Azure incluye controles como Azure Defender, encriptación, RBAC apara control de accesos y políticas de cumplimiento ( ISO, etc.). |
| **Automatización y DevOps** | Integración con Azure DevOps y CI/CD para desplegar pipelines de datos automatizados asi como . |
| **Gobierno y Catalogación** | Con Azure Purview se facilita la gestión de metadatos, linaje de datos y cumplimiento normativo. |
| **Analítica Avanzada e IA** | Azure Fabric o Azure Databricks, Machine Learning y Cognitive Services permiten analítica predictiva y prescriptiva. |

**Beneficios de Negocio**

| **Categoría** | **Beneficio** |
| --- | --- |
| **Toma de Decisiones Basada en Datos** | La analítica avanzada permite decisiones más rápidas y precisas con información confiable. |
| **Reducción de Costos Operativos** | La arquitectura PaaS y serverless permite pagar solo por lo que se usa, reduciendo gastos de infraestructura. |
| **Time-to-Market Más Rápido** | Las soluciones preconfiguradas de Azure aceleran el despliegue de productos y servicios basados en datos. |
| **Innovación Continua** | Posibilita pruebas ágiles de nuevas hipótesis, modelos de negocio o soluciones analíticas. |
| **Cumplimiento Normativo** | Facilita la trazabilidad, control de acceso y auditoría exigidos por entes reguladores. |
| **Ventaja Competitiva** | Permite construir modelos de inteligencia artificial, customer 360, predicción de demanda, etc. |
| **Visión 360 del Cliente y Operaciones** | Integración de datos de diferentes áreas para una visión unificada que mejora la experiencia del cliente y la eficiencia operativa. |
| **Resiliencia Digital** | Capacidades nativas de respaldo, recuperación, monitoreo y continuidad operativa para afrontar interrupciones. |