Documento Técnico: Arquitectura Multicloud para Optimización de Estrategias de Precios en una Tienda de Ropa en Línea

1. Objetivo del Proyecto

Diseñar una arquitectura escalable, segura y eficiente que permita a una tienda de ropa en línea optimizar sus estrategias de precios y categorización de productos mediante análisis avanzado y machine learning. La solución integrará datos desde AWS hacia Azure, aplicando una arquitectura medallion y modelos predictivos, con visualización de resultados a través de Power BI.

2. Pilares de una Arquitectura de Datos Moderna

Diseñar una arquitectura de datos moderna y robusta va más allá de seleccionar tecnologías. Implica adoptar un conjunto de **pilares fundamentales** que garanticen que la solución no solo funcione hoy, sino que sea escalable, segura, eficiente y alineada con los objetivos del negocio en el largo plazo. Estos pilares sirven como una guía estructural para tomar decisiones informadas en términos de diseño, implementación y operación. Incorporarlos desde las etapas tempranas permite crear plataformas que soporten de manera sostenible la innovación, la analítica avanzada y el cumplimiento normativo, al tiempo que optimizan recursos y maximizan el valor del dato.

A continuación se detallan los pilares clave considerados en el diseño propuesto para esta solución multicloud con enfoque analítico:

Data and AI Governance

La gobernanza de datos e inteligencia artificial garantiza que la información utilizada sea confiable, segura y cumpla con políticas corporativas y regulatorias. Esto incluye la trazabilidad de los datos, la calidad, el linaje y la gestión del ciclo de vida de los modelos de machine learning. Una gobernanza sólida asegura que los modelos predictivos utilizados para definir precios se construyen sobre bases confiables y auditables.

Interoperability and Usability

La arquitectura propuesta está diseñada para ser interoperable entre plataformas, específicamente entre AWS y Azure. Esto permite aprovechar servicios especializados de cada nube mientras se mantienen estándares comunes de acceso, almacenamiento y procesamiento. La usabilidad se ve reforzada mediante interfaces familiares como Power BI y SQL, que facilitan el consumo de datos por usuarios de negocio.

Operational Excellence

Este pilar garantiza que todos los procesos y flujos de datos se ejecuten de manera confiable en producción. Incluye prácticas como automatización de pipelines, manejo de errores, logging centralizado y monitoreo activo. Al establecer una arquitectura medallion (bronze, silver, gold), se facilita la trazabilidad y el control de calidad en cada fase del procesamiento.

Security, Privacy, and Compliance

Proteger la aplicación, los datos y los modelos ante amenazas es esencial. La solución contempla controles RBAC (Role-Based Access Control), encriptación en tránsito y en reposo, autenticación mediante Azure Active Directory. Esto proporciona una base segura para compartir información entre entornos multicloud y usuarios internos.

Reliability

La capacidad del sistema para recuperarse ante fallos y seguir operando es crucial. El uso de almacenamiento en S3 junto con procesamiento en Databricks (tolerante a fallos y con autoescalado) garantiza una alta disponibilidad. Además, el monitoreo continuo permite detectar y resolver problemas de manera proactiva.

Performance Efficiency

La arquitectura está diseñada para adaptarse dinámicamente a cambios en la carga de trabajo. El uso de Databricks permite escalar horizontalmente los clústeres según necesidad, optimizar consultas con Delta Lake y aplicar caching donde sea necesario. Esto asegura tiempos de respuesta adecuados, incluso ante grandes volúmenes de datos.

Cost Optimization

Finalmente, la eficiencia en costos se logra mediante el uso de almacenamiento económico en S3, instancias spot donde sea posible, y políticas de autoapagado para clústeres en Databricks. La separación de procesamiento y almacenamiento también permite escalar cada componente de manera independiente, evitando sobrecostos innecesarios.

3. Componentes Principales de la Arquitectura

a. Origen de Datos

- Sistema transaccional de la tienda (ERP, eCommerce, CRM).
- Exportación diaria o en tiempo real hacia AWS S3.

b. Almacenamiento Inicial

- **Amazon S3 (Data Lake Raw)**
    - Almacenamiento económico y escalable.
    - Repositorio inicial de archivos CSV/JSON/parquet.

c. Procesamiento y ETL

- **Azure Databricks**
    - Lectura directa desde S3 usando acceso multicloud (token o IAM Role + External Location).
    - Arquitectura medallion con capas:
        - **Bronze**: Datos crudos desde S3.
        - **Silver**: Datos limpios y normalizados.
        - **Gold**: Datos agregados listos para analítica y machine learning.

d. Machine Learning

- **MLflow en Databricks**
    - Entrenamiento de modelos de regresión para predicción de precios.
    - Seguimiento, versionado y despliegue de modelos.
    - Posible uso de AutoML.

e. Visualización

- **Power BI (Azure)**
    - Conexión directa a Databricks SQL Warehouse.
    - Dashboards para el equipo comercial con predicciones, precios sugeridos, rendimiento de productos y análisis de categorías.

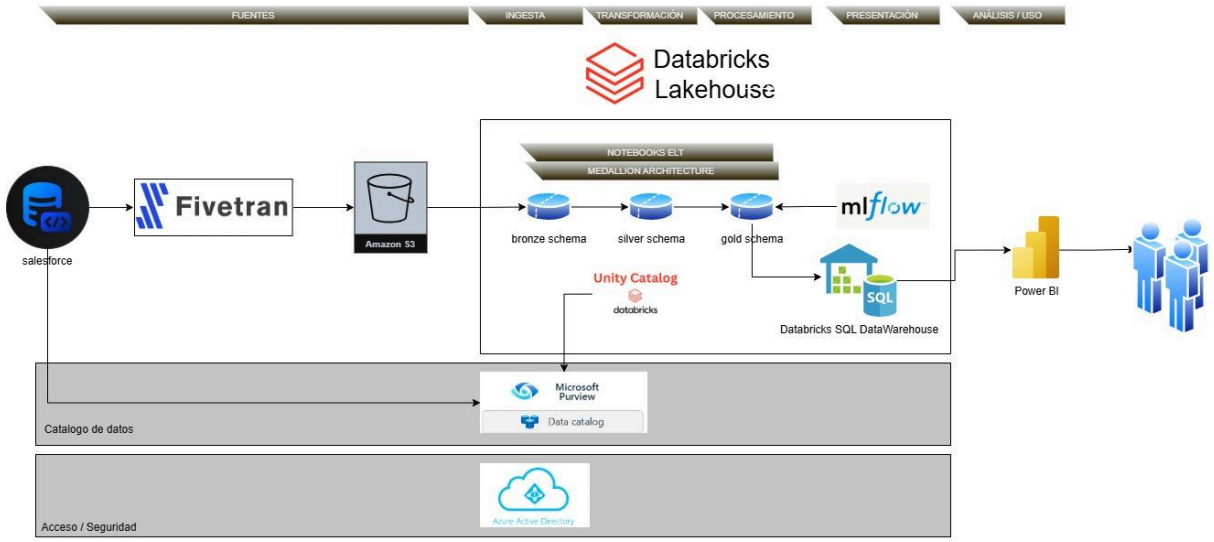
f. Gobernanza y Seguridad

- **Unity Catalog + Azure Purview**
    - Data lineage, clasificación, políticas de acceso y compliance.
- **RBAC (Role-Based Access Control) + Azure Active Directory (AAD)**
    - Control de accesos y privilegios por rol.

g. Monitoreo y Logging

- **Azure Monitor + Log Analytics**
    - Seguimiento del rendimiento de pipelines, consumo de recursos, estado de los modelos, y alertas de seguridad.

4. Diagrama Arquitectónico Conceptual



5. Justificación de Tecnologías

Tecnología	Justificación
AWS S3	Almacenamiento económico, escalable y compatible con múltiples formatos. Ideal para ingestión inicial de datos.
Azure Databricks	Potente para procesamiento distribuido, ETL, ML, con arquitectura medallion y multi cloud enablement.
MLflow	Permite gestionar el ciclo de vida de los modelos de machine learning.
Power BI	Herramienta familiar para usuarios de negocio, con visualizaciones ricas y conexión directa a Databricks**.**
Unity Catalog	Control centralizado de acceso a datos con auditoría y políticas RBAC.
Azure Monitor	Supervisión de recursos, fallas, y alertas**.**

6. Seguridad, Cumplimiento y Gobernanza

- Accesos gestionados por AAD y Unity Catalog.
- Datos en tránsito protegidos con TLS.
- Datos en reposo cifrados con claves administradas por el cliente (CMK).
- Clasificación de datos sensibles con Azure Purview.
- Logs de auditoría habilitados**.**

7. Beneficios Técnicos

- Escalabilidad multicloud: Flexibilidad para integrar nuevos orígenes o destinos en AWS y Azure.
- Eficiencia operativa: Procesos automatizados con Databricks Workflows.
- Despliegue continuo de modelos: MLflow permite actualizar modelos sin interrumpir el negocio.
- Altas prestaciones: Caching, autoescalado y ejecución optimizada en Databricks.

8. Beneficios de Negocio

- Mejora en pricing dinámico: Predicciones precisas de precios optimizan márgenes y volumen de ventas.

- Mayor visibilidad: Dashboards en Power BI ofrecen insights accionables a los equipos comerciales.

- Tiempo de respuesta más rápido: Automatización reduce la dependencia de análisis manual.

- Cumplimiento y trazabilidad: Seguridad y gobierno facilitan auditorías y evitan riesgos legales.