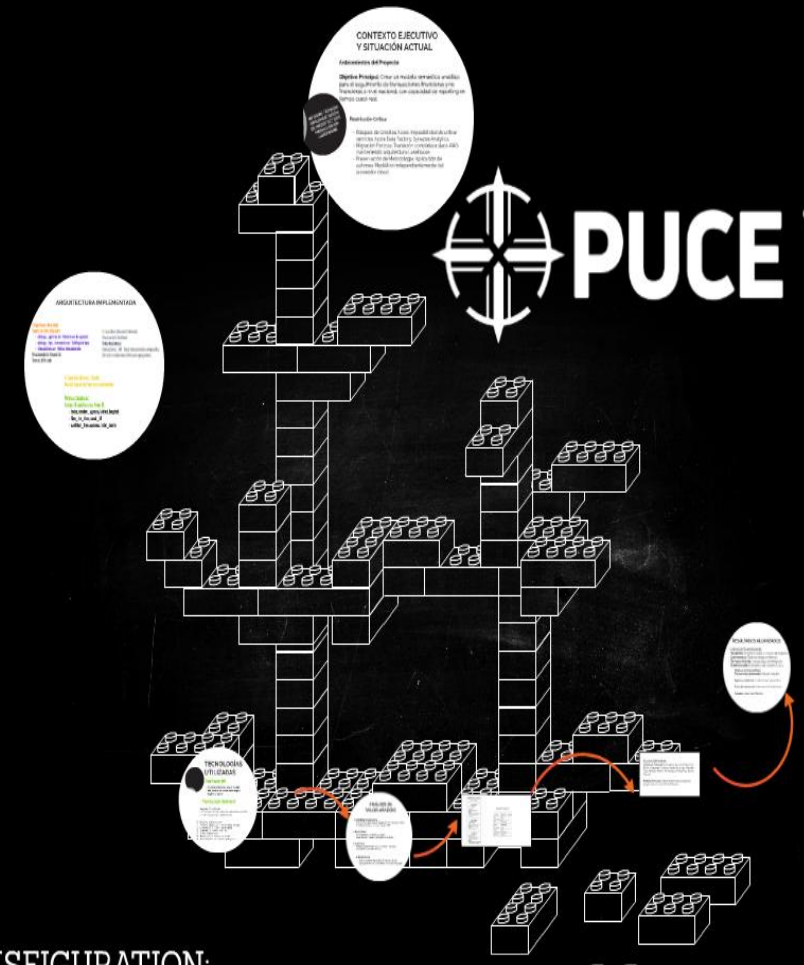


PUCE TEC

DATA TRANSFIGURATION:
INFORME TÉCNICO: IMPLEMENTACIÓN DE ARQUITECTURA MEDALLION EN LAKEHOUSE

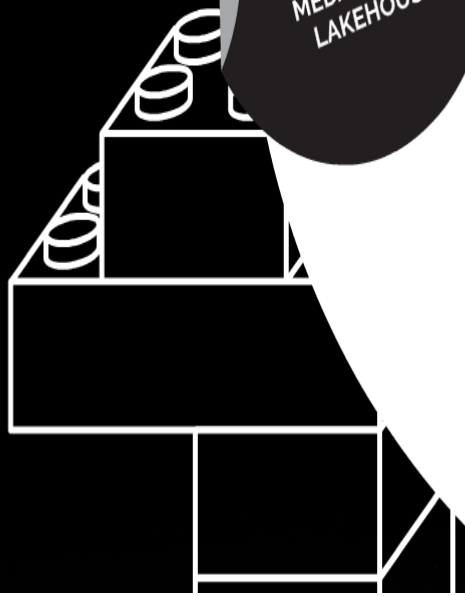
Carlos Pazmiño Zambrano
Desarrollo de Software

CONTEXTO EJECUTIVO Y SITUACIÓN ACTUAL

Antecedentes del Proyecto

Objetivo Principal: Crear un modelo semántico analítico para el seguimiento de transacciones financieras y no financieras a nivel nacional, con capacidad de reporting en tiempo cuasi-real.

INFORME TÉCNICO: IMPLEMENTACIÓN DE ARQUITECTURA MEDALLION EN LAKEHOUSE



Restricción Crítica:

- Bloqueo de Créditos Azure: Imposibilidad de utilizar servicios Azure Data Factory, Synapse Analytics.
- Migración Forzosa: Transición completa a stack AWS manteniendo arquitectura Lakehouse.
- Preservación de Metodología: Aplicación de patrones Medallion independientemente del proveedor cloud.

ARQUITECTURA IMPLEMENTADA

1. Capa Bronze (Raw Data)

Fuentes de Datos Originales:

- catalogo_agencias.csv Dimensiones de agencias
- catalogo_tipo_transaccion.csv Catálogo de tipos
- transacciones.csv Hechos transaccionales

Almacenamiento: Amazon S3

Formato: CSV crudo

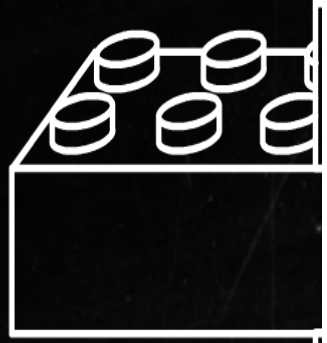
2. Capa Silver (Cleaned & Enriched)

Procesamiento Realizado:

Tablas Resultantes:

transacciones_full Datos transaccionales enriquecidos

Estructura normalizada y lista para agregaciones



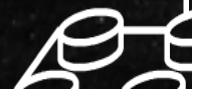
3. Capa Gold (Business Ready)

Modelo Semántico Final: transaccionalidad

Métricas Calculadas:

Campos Disponibles para Power BI:

- fecha, nombre_agencia, latitud, longitud
- tipo_trx_desc, canal_id
- cantidad_transacciones, total_monto



TECNOLOGÍAS UTILIZADAS

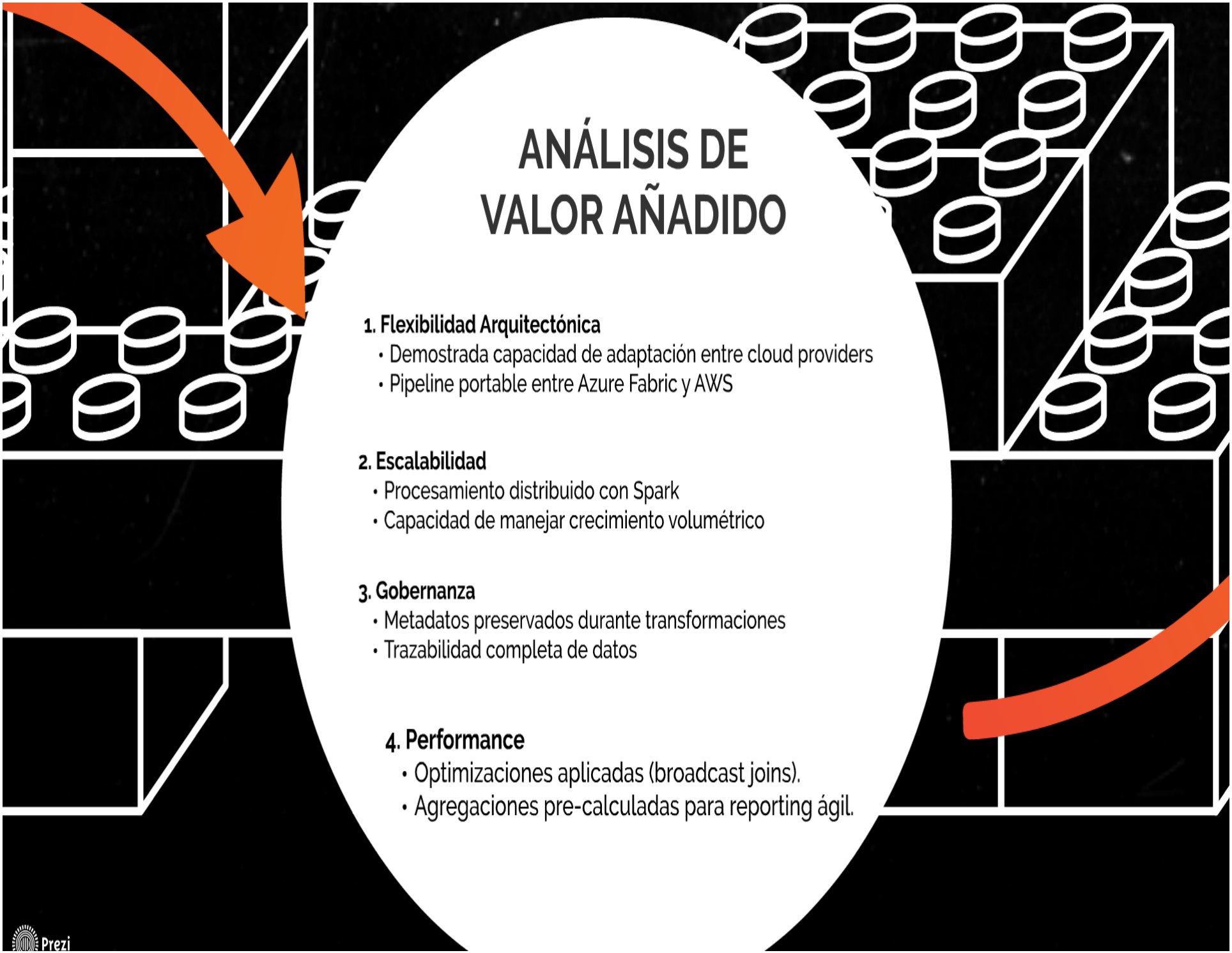
Cloud Provider: AWS

- **S3: Almacenamiento object storage**
- **IAM: Gestión de credenciales seguro**
- **Región: us-east-2**

Processing Engine: Apache Spark

- Lenguaje: Python/PySpark
- Optimizaciones: Broadcast joins para dimensiones pequeñas
- Transformaciones: SQL + DataFrame API

```
# Pipeline implementado
1. Lectura desde S3 → DataFrames Pandas
2. Conversión a Spark DataFrames
3. Limpieza y normalización
4. Joins semánticos
5. Agregaciones business-ready
6. Persistencia en tablas queryable
```


ANÁLISIS DE
VALOR AÑADIDO
1. Flexibilidad Arquitectónica
• Demostrada capacidad de adaptación entre cloud providers
• Pipeline portable entre Azure Fabric y AWS
2. Escalabilidad
• Procesamiento distribuido con Spark
• Capacidad de manejar crecimiento volumétrico
3. Gobernanza
• Metadatos preservados durante transformaciones
• Trazabilidad completa de datos
4. Performance
• Optimizaciones aplicadas (broadcast joins).
• Agregaciones pre-calculadas para reporting ágil.

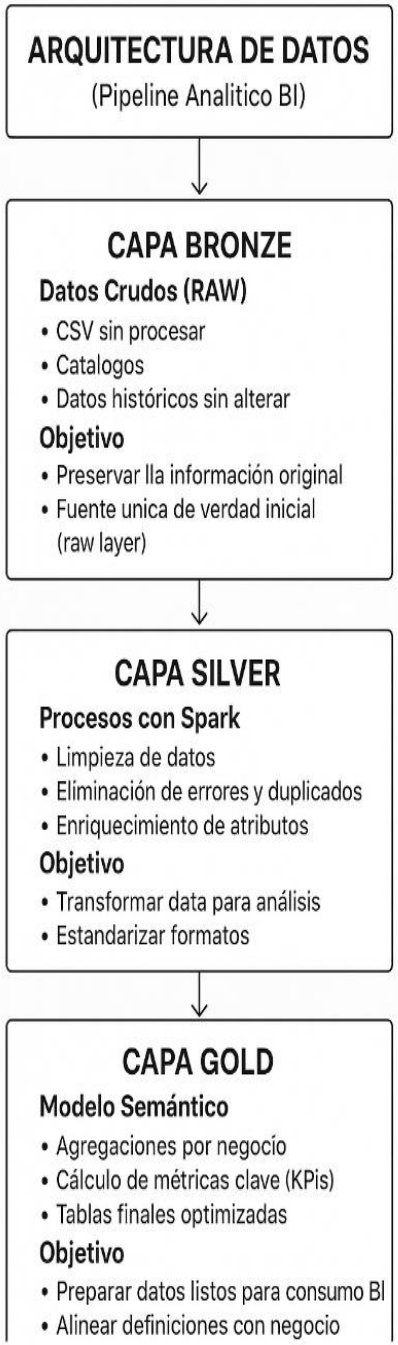


tabla de transacción

col_name	data_type	comment
fecha	date	
nombre_agencia	string	
latitud	decimal(9,6)	
longitud	decimal(9,6)	
tipo_trx_desc	string	
canal_id	bigint	
cantidad_transa-	decimal(15,2)	
total_monto	decimal	

Alcance y Limitaciones

Cobertura Geográfica: 12 agencias a nivel nacional (Quito, Guayaquil, Cuenca, Santo Domingo, Machala, Loja, Ambato, Manta, Esmeraldas, Riobamba, Ibarra, Tulcán)

Período Temporal: Datos desde marzo 2025 con proyección de crecimiento histórico

RESULTADOS ALCANZADOS

- **Calidad de Datos Lograda:**
- **Integridad:** 0 valores nulos en claves principales
- **Consistencia:** Tipos de datos uniformes
- **Enriquecimiento:** Geolocalización integrada
- **Normalización:** Estructura dimensional clara

Métricas del Dataset Final:
Transacciones procesadas: Dataset completo

Agencias cubiertas: 12 ubicaciones geográficas

Tipos de transacción: Financiera/No Financiera

Canales: 4 tipos identificados