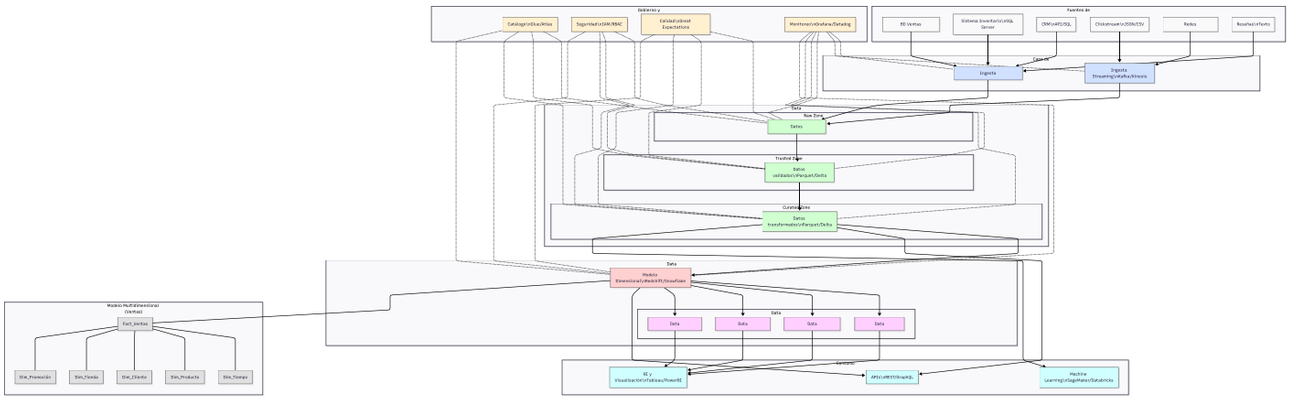
# Proyecto Integrador: Arquitectura de Datos para E-commerce y Retail Digital

## Resumen Ejecutivo

Este documento presenta una propuesta integral de arquitectura de datos para una empresa de e-commerce y retail digital que ha experimentado un crecimiento acelerado. La arquitectura propuesta aborda los desafíos de integrar múltiples fuentes de datos heterogéneas, implementar una estrategia de almacenamiento escalable, garantizar la calidad de los datos y proporcionar un modelo analítico multidimensional para la toma de decisiones.

La solución diseñada se alinea con las mejores prácticas de la industria y contempla aspectos clave como la gobernanza, la escalabilidad y la seguridad de los datos, mientras proporciona las bases para casos de uso analíticos avanzados que generarán valor para el negocio.




## 1. Arquitectura de Datos

### 1.1. Fuentes de Datos Identificadas

Se han identificado seis fuentes de datos principales que alimentarán la arquitectura:

**Fuentes Estructuradas:**

1. **Base de datos transaccional de ventas online** (SQL)
2. **Sistema de inventario** (SQL Server)
3. **CRM** (API/SQL)

**Fuentes No Estructuradas:**

1. **Registros de navegación web (clickstream)** (JSON/CSV)
2. **Redes sociales** (APIs)
3. **Comentarios y reseñas de productos** (Texto)

### 1.2. Arquitectura por Capas

La arquitectura propuesta se organiza en cinco capas funcionales:

1. **Capa de Ingesta**: Captura datos de fuentes diversas mediante procesos batch y en tiempo real.
2. **Capa de Almacenamiento**: Organizada en zonas Raw, Trusted y Curated, además de Data Warehouse y Data Marts.
3. **Capa de Procesamiento**: Transformación, limpieza y enriquecimiento de datos mediante procesos batch y streaming.
4. **Capa de Gobierno y Calidad**: Aseguramiento de políticas, metadatos, seguridad y calidad de datos.
5. **Capa de Consumo**: Acceso a los datos para análisis, visualización y APIs.

### 1.3. Principios Arquitectónicos

La arquitectura se fundamenta en principios clave:

- **Gobierno**: Propiedad clara, catalogación, linaje, políticas de retención y cumplimiento normativo.
- **Escalabilidad**: Arquitectura distribuida, separación computación/almacenamiento, procesamiento elástico.
- **Flexibilidad**: Diseño modular, tecnologías cloud-native, schema-on-read, enfoque políglota.

# 2. Estrategia de Almacenamiento y Gestión

## 2.1. Zonas de Almacenamiento

**Data Lake**

- **Zona Raw (Bronze)**: Almacenamiento de datos en su formato original sin transformaciones.
- **Zona Trusted (Silver)**: Datos validados, limpios y transformados con esquemas estandarizados.
- **Zona Curated (Gold)**: Datos agregados y optimizados para casos de uso específicos.

**Data Warehouse**

Almacenamiento estructurado para análisis avanzado con modelo relacional o multidimensional optimizado para consultas analíticas.

**Data Marts**

Subconjuntos departamentales enfocados en Marketing, Ventas, Logística y Finanzas.

## 2.2. Tecnologías Recomendadas

Se propone un enfoque cloud-native con opción de implementación en AWS:

| Componente | Tecnología AWS | Alternativa Multi-cloud |
|---|---|---|
| Data Lake Raw | S3 | MinIO / HDFS |
| Data Lake Trusted | S3 + AWS Glue | Databricks Delta Lake |
| Data Lake Curated | S3 + Athena | Apache Iceberg + Trino |
| Data Warehouse | Redshift | Snowflake |
| Data Marts | Redshift Spectrum + QuickSight | Looker, Tableau |
| Procesamiento | EMR, Lambda, Kinesis | Spark, Flink |
| Gobernanza | AWS Glue, Lake Formation | Collibra, Apache Atlas |

## 2.3. Prácticas de Gobernanza

- **Metadatos y Catalogación**: Catálogo centralizado, business glossary, etiquetado.
- **Linaje de Datos**: Trazabilidad end-to-end, análisis de impacto.
- **Seguridad y Acceso**: RBAC/ABAC, encriptación, tokenización, enmascaramiento.
- **Gestión del Ciclo de Vida**: Políticas de retención, archivado, purga.

# 3. Plan de Calidad de Datos

## 3.1. Dimensiones de Calidad

El plan contempla dimensiones clave: Exactitud, Completitud, Consistencia, Actualidad, Unicidad, Validez e Integridad.

## 3.2. Controles por Zona

**Zona Raw**:

- Verificación de completitud de archivos
- Validación de estructura básica
- Detección de anomalías en volumen

**Zona Trusted**:

- Validación de tipos de datos
- Detección de valores nulos en campos obligatorios
- Verificación de integridad referencial básica

**Zona Curated**:

- Validación de reglas de negocio complejas
- Verificación de agregaciones y cálculos
- Detección de anomalías en tendencias

**Data Warehouse**:

- Validación de integridad referencial completa
- Verificación de conformidad dimensional
- Validación de consistencia temporal

### 3.3. Proceso de Monitoreo y Remediación

- **Monitoreo Proactivo**: Dashboards de calidad, alertas automáticas, reportes periódicos.
- **Remediación**: Proceso estructurado de detección, clasificación, investigación, corrección y verificación.
- **Gestión de Incidentes**: Registro centralizado, clasificación de severidad, SLAs de resolución.

### 3.4. Integración en la Arquitectura

El plan de calidad se integra en puntos estratégicos del flujo: pre-ingesta, post-ingesta, pre-transformación, post-transformación, pre-carga y post-carga.

## 4. Modelo Multidimensional para Ventas

### 4.1. Justificación del Área de Negocio

Se seleccionó el área de Ventas por su impacto directo en ingresos, necesidades analíticas frecuentes, riqueza de dimensiones y disponibilidad de datos.

### 4.2. Tablas de Hechos

- **FACT_VENTAS**: Transacciones de venta con métricas como cantidad, monto_bruto, descuento, impuesto, monto_neto, costo y margen.
- **FACT_ACTIVIDAD_WEB**: Comportamiento en el sitio con métricas como tiempo en página, clicks y flags de actividad.

### 4.3. Dimensiones Principales

- **DIM_TIEMPO**: Jerarquía temporal completa (día → semana → mes → trimestre → año)
- **DIM_PRODUCTO**: Información de productos con jerarquía de categorización
- **DIM_CLIENTE**: Datos de clientes con segmentación
- **DIM_TIENDA**: Información de canales de venta (físicos y digitales)
- **DIM_PROMOCION**: Detalles de promociones y campañas
- **DIM_METODO_PAGO**: Formas de pago disponibles
- **Dimensiones adicionales**: Empleado, Página web, Dispositivo

### 4.4. Esquema y Justificación

Se implementó un esquema estrella para optimizar el rendimiento de consultas analíticas, minimizar joins y facilitar la comprensión del modelo. Las dimensiones se desnormalizaron para mejorar el rendimiento a costa de un mayor espacio de almacenamiento.

## 5. Implementación y Hoja de Ruta

### 5.1. Enfoque de Implementación

Se recomienda una implementación incremental:

**Fase 1 (0-3 meses)**:

- Establecer infraestructura básica (zonas del Data Lake)
- Implementar ingesta de fuentes críticas
- Definir controles de calidad básicos

**Fase 2 (3-6 meses)**:

- Implementar Data Warehouse y primer modelo dimensional
- Desarrollar procesos ETL/ELT completos
- Extender controles de calidad

**Fase 3 (6-12 meses)**:

- Implementar Data Marts departamentales
- Desarrollar capacidades avanzadas de análisis
- Automatización completa de procesos

### 5.2. Recursos Necesarios

- **Infraestructura**: Entorno cloud (preferiblemente AWS)
- **Herramientas**: Suite de ingesta, procesamiento, calidad y visualización
- **Equipo**: Data Engineers, Data Architects, Data Quality Analysts, Data Stewards

### 5.3. Métricas de Éxito

- Reducción del tiempo para integrar nuevas fuentes de datos
- Disminución de incidentes relacionados con calidad de datos
- Adopción de herramientas analíticas por usuarios de negocio
- ROI demostrado en casos de uso implementados

## 6. Conclusiones y Recomendaciones

La arquitectura propuesta establece una base sólida para la gestión y análisis de datos en la empresa de e-commerce y retail digital. Las principales ventajas de esta solución son:

1. **Escalabilidad**: Capacidad para adaptarse al crecimiento continuo del negocio.
2. **Flexibilidad**: Adaptabilidad a nuevas fuentes y requisitos analíticos.
3. **Gobierno**: Control y trazabilidad completa del ciclo de vida de los datos.
4. **Calidad**: Aseguramiento sistemático de la fiabilidad de la información.
5. **Valor analítico**: Modelo dimensional optimizado para responder preguntas clave del negocio.

Se recomienda:

- Establecer un comité de gobierno de datos para supervisar la implementación
- Priorizar casos de uso con alto impacto en ingresos o reducción de costos
- Invertir en capacitación del personal para maximizar el valor de la arquitectura
- Revisar y actualizar periódicamente la arquitectura para incorporar nuevas tecnologías y prácticas

La implementación exitosa de esta arquitectura posicionará a la empresa para aprovechar al máximo el valor de sus datos y obtener ventajas competitivas en el mercado de e-commerce y retail digital.