Lección 2: Enfoques para Almacenamiento y Gestión de Datos

1. Análisis del Esquema Arquitectónico

Partiendo del esquema arquitectónico diseñado en la Lección 1, es fundamental definir las estrategias de almacenamiento y gobernanza que permitirán materializar esta visión. La arquitectura basada en capas (ingesta, almacenamiento, procesamiento, gobierno y consumo) requiere un enfoque detallado para cada una de las zonas de almacenamiento.

2. Zonas de Almacenamiento

2.1. Data Lake

Zona Raw (Bronze)

- **Propósito**: Almacenamiento de datos crudos en su formato original, sin transformaciones.
- **Características**:
  - Preserva la integridad y autenticidad de los datos originales
  - Mantiene el historial completo para auditoría y linaje
  - Almacena datos estructurados, semi-estructurados y no estructurados
- **Datos almacenados**:
  - Dumps de bases de datos transaccionales (ventas, inventario)
  - Logs de clickstream sin procesar
  - Datos de redes sociales en formato JSON
  - Datos de sensores y dispositivos IoT
  - Comentarios y reseñas en texto plano
- **Tecnologías recomendadas**:
  - **Almacenamiento**: Amazon S3, Azure Data Lake Storage Gen2, Google Cloud Storage
  - **Formato de archivos**: Parquet, ORC, JSON, CSV, Avro
  - **Gestión**: AWS Lake Formation, Azure Data Factory, Delta Lake

Zona Trusted (Silver)

- **Propósito**: Almacenamiento de datos validados, limpios y transformados.
- **Características**:
  - Datos con calidad validada
  - Esquemas estandarizados y normalizados
  - Datos enriquecidos con información adicional
- **Datos almacenados**:
  - Tablas de transacciones con errores corregidos
  - Datos de clickstream procesados y estructurados
  - Sentimiento extraído de comentarios y redes sociales
  - Datos de inventario normalizados
- **Tecnologías recomendadas**:
  - **Almacenamiento**: Mismo que Raw, pero con estructura mejorada
  - **Procesamiento**: Apache Spark, Databricks
  - **Calidad**: Great Expectations, Deequ
  - **Formato**: Principalmente Parquet, Delta

Zona Curated (Gold)

- **Propósito**: Datos agregados, transformados y optimizados para análisis específicos.
- **Características**:
  - Orientada a casos de uso de negocio
  - Agregaciones predefinidas

    - Optimizada para rendimiento de consultas
- **Datos almacenados**:
    - KPIs de negocio agregados
    - Dashboards precomputados
    - Vistas materializadas para análisis frecuentes
    - Datasets de machine learning
- **Tecnologías recomendadas**:
    - **Almacenamiento**: Mismo que Trusted
    - **Optimización**: Delta Lake, Iceberg
    - **Consulta**: Presto, Athena, Spark SQL

2.2. Data Warehouse

- **Propósito**: Almacenamiento optimizado para análisis estructurado y reportería.
- **Características**:
    - Modelo relacional o multidimensional
    - Optimizado para consultas analíticas complejas
    - Alto rendimiento para reporting
- **Datos almacenados**:
    - Modelo dimensional de ventas
    - Análisis de inventario y cadena de suministro
    - Datos de clientes unificados
    - Métricas de rendimiento del negocio
- **Tecnologías recomendadas**:
    - **Plataformas**: Snowflake, Amazon Redshift, Google BigQuery, Azure Synapse
    - **Modelado**: dbt, Dataform

2.3. Data Marts

- **Propósito**: Subconjuntos del Data Warehouse orientados a departamentos específicos.
- **Características**:
    - Centrados en necesidades departamentales
    - Facilitan el self-service BI
    - Simplifican el acceso a datos relevantes
- **Data Marts específicos**:
    - **Marketing**: Comportamiento de clientes, efectividad de campañas
    - **Ventas**: Análisis de ventas por región, producto y tiempo
    - **Logística**: Gestión de inventario, tiempos de entrega
    - **Finanzas**: Ingresos, costos, márgenes
- **Tecnologías recomendadas**:
    - Mismas que Data Warehouse o plataformas específicas como Microsoft Analysis Services

3. Relación entre Zonas de Almacenamiento

El flujo de datos entre las diferentes zonas sigue un patrón de maduración progresiva:

1. **Raw** → **Trusted**: Proceso de validación, limpieza y estandarización.
2. **Trusted** → **Curated**: Transformaciones orientadas a casos de uso específicos.
3. **Curated** → **Data Warehouse**: Carga de datos limpios y transformados al modelo dimensional.
4. **Data Warehouse** → **Data Marts**: Distribución de datos a departamentos específicos.

Este flujo unidireccional garantiza la integridad y calidad de los datos a medida que avanzan por la arquitectura.

4. Tecnologías y Servicios Recomendados por Zona

Enfoque Cloud-Native (AWS)

Zona	Servicio	Justificación

Zona	Servicio	Justificación
Raw	S3	Escalabilidad ilimitada, bajo costo, diferentes clases de almacenamiento
Trusted	S3 + AWS Glue	Catálogo de datos integrado, transformaciones ETL serverless
Curated	S3 + Athena	Consultas SQL sobre datos optimizados en S3
Data Warehouse	Redshift	Alto rendimiento para consultas analíticas, integración nativa con S3
Data Marts	Redshift Spectrum + QuickSight	Análisis federado y visualizaciones para departamentos

Enfoque Híbrido/Multi-Cloud

Zona	Servicio	Justificación
Raw	MinIO / HDFS	Solución on-premise compatible con S3 API
Trusted	Databricks Delta Lake	Transacciones ACID, esquema evolutivo, compatible multi-cloud
Curated	Apache Iceberg + Trino	Formato de tabla abierto, consultas federadas
Data Warehouse	Snowflake	Multi-cloud, separación computación/almacenamiento
Data Marts	Looker, Tableau	Herramientas BI multi-fuente

5. Prácticas de Gobernanza y Gestión de Datos

5.1. Metadatos y Catalogación

- **Catálogo de datos**: Implementación de un catálogo centralizado que documenta todas las fuentes, tablas, campos y sus definiciones.
- **Business glossary**: Definiciones de negocio estandarizadas para términos clave.
- **Etiquetado**: Sistema de etiquetas para clasificación de datos (sensibilidad, departamento, caso de uso).
- **Herramientas recomendadas**: AWS Glue Catalog, Collibra, Alation, Apache Atlas.

5.2. Linaje de Datos

- **Trazabilidad end-to-end**: Registro del origen, transformaciones y uso de los datos.
- **Impacto y dependencias**: Análisis de impacto para cambios en fuentes o transformaciones.
- **Auditoría**: Registro de accesos y modificaciones para cumplimiento normativo.
- **Herramientas recomendadas**: Collibra, Informatica Enterprise Data Catalog, IBM InfoSphere.

5.3. Seguridad y Acceso

- **RBAC/ABAC**: Control de acceso basado en roles y atributos.
- **Encriptación**: Encriptación en reposo y en tránsito para todas las zonas.
- **Tokenización**: Para datos sensibles como información financiera.
- **Enmascaramiento**: Para datos PII en entornos no productivos.
- **Herramientas recomendadas**: AWS IAM, HashiCorp Vault, CyberArk.

5.4. Gestión del Ciclo de Vida

- **Políticas de retención**: Definición de períodos de retención por tipo y sensibilidad de datos.
- **Archivado**: Estrategias para mover datos históricos a almacenamiento frío.
- **Purga**: Procesos automatizados para eliminar datos obsoletos o que exceden períodos de retención.
- **Herramientas recomendadas**: AWS S3 Lifecycle Policies, Azure Lifecycle Management.

5.5. Calidad de Datos

- **Perfiles de datos**: Análisis estadístico de distribuciones y anomalías.

- **Validación**: Reglas de validación implementadas en el proceso de ingesta.
- **Monitoreo**: Alertas automatizadas para detección de problemas de calidad.
- **Herramientas recomendadas**: Great Expectations, Talend Data Quality, AWS Deequ.

6. Justificación Técnica del Enfoque de Almacenamiento

6.1. ¿Por qué un enfoque híbrido Data Lake + Data Warehouse?

1. **Flexibilidad + Rendimiento**: El Data Lake proporciona la flexibilidad necesaria para almacenar datos heterogéneos, mientras que el Data Warehouse ofrece rendimiento optimizado para consultas analíticas.
2. **Cobertura completa de casos de uso**: Desde análisis exploratorio ad-hoc hasta dashboards predefinidos de alto rendimiento.
3. **Evolución gradual**: Permite comenzar con casos de uso prioritarios en el Data Warehouse mientras se construye el Data Lake para casos futuros.
4. **Optimización de costos**: Almacenamiento de bajo costo para datos raw, inversión en computación solo para datos de alto valor.

6.2. Beneficios de la arquitectura por zonas

1. **Preservación de datos originales**: La zona Raw mantiene la versión original de todos los datos para referencia y auditoría.
2. **Mejora progresiva de calidad**: Cada zona añade un nivel adicional de validación y transformación.
3. **Aislamiento de problemas**: Los problemas en una zona no afectan necesariamente a las demás.
4. **Optimización específica**: Cada zona puede optimizarse para su propósito particular (ingestión rápida, consultas eficientes, etc.).

7. Consideraciones de Implementación

7.1. Enfoque de Migración

Se recomienda un enfoque incremental:

1. Establecer la infraestructura básica (zonas del Data Lake, Data Warehouse)
2. Migrar primero las fuentes críticas (ventas, inventario, clientes)
3. Implementar progresivamente casos de uso analíticos
4. Expandir a fuentes no estructuradas y análisis avanzados

7.2. Monitoreo de Infraestructura

- Implementación de observabilidad end-to-end
- Dashboards de utilización y rendimiento
- Alertas para anomalías y problemas de capacidad
- Herramientas: Grafana, Prometheus, CloudWatch