Arquitetura e Estrutura de Data Lakes em Nuvem

Onde as grandes empresas guardam seus dados brutos em terabytes

> **Onde Google, Amazon e Netflix armazenam seus dados brutos para Machine Learning e análises exploratórias?**

Abertura: O que é um Data Lake?

Onde as grandes empresas guardam todos os seus dados brutos, não estruturados e em terabytes?

A resposta é o Data Lake!

Diferente do **Data Warehouse** (otimizado para relatórios), o **Data Lake** armazena dados em seu formato original, sendo a base para **Big Data, Machine Learning e análises exploratórias**. É onde os dados brutos vivem antes de serem transformados em insights.

Data Lake

Armazena dados brutos de múltiplas fontes. Flexível, escalável, pronto para exploração.

Data Warehouse

Armazena dados limpos e estruturados. Otimizado para relatórios e BI gerenciais.

Arquitetura em Camadas

Bronze (bruta) → Prata (refinada) → Ouro (curada). Transformação progressiva de dados.

Data Lake vs Data Warehouse

Aspecto	Data Lake	Data Warehouse
Estrutura	**Flexível** (Schema-on-Read)	**Rígida** (Schema-on-Write)
Dados	Brutos, não estruturados	Limpos, estruturados
Formato	Múltiplos (JSON, CSV, logs, vídeos)	Tabelas normalizadas
Uso Principal	Big Data, ML, exploração	Relatórios, BI, análises
Custo	**Baixo** (armazenamento barato)	**Alto** (processamento intensivo)
Escalabilidade	**Horizontal** (fácil)	**Vertical** (limitada)

Arquitetura em Camadas: Bronze, Prata, Ouro

Bronze
RAW / BRUTA

DADOS

Formato original, sem alterações. Brutos e não estruturados.

CARACTERÍSTICAS

- Pode conter nulos
- Pode ter duplicatas
- Múltiplos formatos
- Retenção: 2-3 anos

ACESSO

Apenas Engenharia de Dados e sistemas de ingestão.

Exemplo:
Logs brutos de navegação, CSVs de transações, JSONs de APIs

Prata
REFINED / REFINADA

DADOS

Limpos, validados e padronizados. Prontos para Data Science.

CARACTERÍSTICAS

- Nulos tratados
- Duplicatas removidas
- Formato Parquet
- Retenção: 1 ano

ACESSO

Engenharia, Data Scientists, Analistas.

Exemplo:
Tabelas Parquet de clientes, transações, produtos limpos

Ouro
CURATED / CURADA

DADOS

Agregados, sumarizados e prontos para relatórios e BI.

CARACTERÍSTICAS

- Dados agregados
- Pronto para dashboards
- Formato Parquet
- Retenção: 3-5 anos

ACESSO

BI, Gerentes, Executivos.

Exemplo:
Vendas diárias, clientes segmentados, KPIs agregados

Estratégia de Particionamento para Performance

Particionamento é dividir os dados em **diretórios (pastas) menores** com base em uma coluna de alta cardinalidade. Exemplo: em vez de um único arquivo com 1TB de dados, você cria pastas por ano/mês/dia. Isso melhora drasticamente a performance e reduz custos.

Exemplo 1: Varejo (Vendas)

```
s3://datalake/prata/vendas/ ├── ano=2025/mes=01/ | ├── vendas_2025-01-15.parquet | └── vendas_2025-01-16.parquet ├── ano=2025/mes=02/ | ├── vendas_2025-02-01.parquet | └── ... └── ano=2024/mes=12/ └── vendas_2024-12-31.parquet
```

Exemplo 2: Saúde (Prontuários)

```
s3://datalake/prata/prontuarios/ ├── ano=2025/mes=01/dia=15/ | ├── prontuarios_2025-01-15.parquet ├── ano=2025/mes=01/dia=16/ | ├── prontuarios_2025-01-16.parquet └── ano=2024/mes=12/dia=31/ └── prontuarios_2024-12-31.parquet
```

Usar Colunas Frequentes em WHERE

Particione por colunas que aparecem frequentemente em cláusulas WHERE (ex: data, região, país).

Evitar Particionamento Excessivo

Particionar por hora em dados históricos cria milhões de pastas. Use ano/mês/dia no máximo.

Tamanho Ideal de Arquivo

Cada arquivo particionado deve ter 128MB-1GB. Muito pequeno = overhead, muito grande = lento.

Formato Parquet

Use Parquet (comprimido, colunar) em vez de CSV. Reduz tamanho em 10x e melhora queries.

Impacto na Performance e Custo

Políticas de Segurança e Acesso

Princípio de Acesso Mínimo: Cada usuário/sistema deve ter acesso **apenas aos dados necessários** para sua função. Bronze é restrita (apenas Engenharia), Prata é moderada (Engenharia + Data Science), Ouro é aberta (todos os analistas).

Varejo: Permissões por Camada

Perfil	Bronze	Prata	Ouro
Engenharia	✓ RW	✓ RW	✓ R
Data Science	✗	✓ R	✓ R
BI/Analytics	✗	✗	✓ R
Executivos	✗	✗	✓ R

Saúde: Permissões por Camada

Perfil	Bronze	Prata	Ouro
Engenharia	✓ RW	✓ RW	✓ R
Médicos	✗	✓ R*	✓ R*
Pesquisadores	✗	✓ R**	✓ R
Administrativo	✗	✗	✓ R

Conformidade LGPD/HIPAA

LGPD (Brasil): Dados pessoais devem ser acessados apenas por quem precisa. Auditoria obrigatória. **HIPAA (EUA):** Dados de saúde exigem criptografia e acesso controlado. Saúde: Médicos acessam apenas prontuários de seus pacientes (Prata anonimizada com filtros por paciente). Pesquisadores acessam dados agregados (sem PII).

Atividade 1: Planejamento Estrutural de Data Lake

Instruções: Divida a turma em grupos de 3-4 alunos. Cada grupo recebe um cenário de negócio e deve criar um esquema lógico de Data Lake, definindo camadas, particionamento e políticas de acesso.

Grupo A: Varejo

E-commerce

Contexto: Uma grande empresa de e-commerce precisa armazenar dados de transações, estoques, logs de navegação e dados de clientes em um Data Lake para análises exploratórias, Machine Learning e relatórios gerenciais.

Tipos de Dados:

- Logs de navegação (alta frequência)
- Transações de vendas (estruturadas)
- Dados de clientes (estruturados)
- Dados de produtos (estruturados)
- Dados de estoques (estruturados)

Sua Tarefa:

→ Defina as camadas (Bronze, Prata, Ouro) e quais dados vão para cada

→ Escolha a estratégia de particionamento para cada camada

→ Defina políticas de acesso (Engenharia, Data Science, BI)

→ Desenhe o roadmap: como um dado flui de Bronze → Ouro

Grupo B: Saúde

Clínica/Hospital

Contexto: Uma rede de clínicas precisa armazenar dados de prontuários eletrônicos, resultados de exames, dados de sensores médicos e registros de pacientes em um Data Lake para análises clínicas, pesquisa e relatórios administrativos.

Tipos de Dados:

- Prontuários eletrônicos (XML, estruturados)
- Resultados de exames (múltiplos formatos)
- Dados de sensores (tempo real, alta frequência)
- Registros de pacientes (estruturados)
- Dados de consultas (estruturados)

Sua Tarefa:

→ Defina as camadas e quais dados vão para cada (considere PII/LGPD)

→ Escolha a estratégia de particionamento (sensores: hora? dia?)

→ Defina políticas de acesso com conformidade LGPD/HIPAA

→ Desenhe o roadmap: como um dado clínico é transformado

Atividade 2: Demonstração de Organização e Catalogação

Instruções: O instrutor deve projetar a tela (ou usar um simulador de console) para demonstrar a criação da estrutura de Data Lake. Mostrar: (1) Criação de diretórios, (2) Particionamento visual, (3) Catalogação conceitual.

Demonstração 1: Estrutura de Diretórios

```
├─ s3://datalake/
├─   bronze/
├─     logs/
├─       navegacao.log
├─   prata/
├─     clientes/
├─       clientes.parquet
├─   ouro/
├─     relatorios/
```

Resultado: Estrutura lógica clara com separação de camadas (Bronze, Prata, Ouro).

Demonstração 2: Particionamento Visual

Upload de Dados Particionados:

`s3://datalake/prata/vendas/`
`ano=2025` / `mes=01` / `dia=15` /
`vendas_2025-01-15.parquet`

Benefício: Queries filtram por data sem ler todos os dados. Performance ↑ Custo ↓

Exemplo de Query:
SELECT * FROM vendas WHERE ano=2025 AND mes=01
← Lê apenas 1 pasta!

Demonstração 3: Catalogação (Inferência de Metadados)

Arquivo Parquet (Prata):

- id_cliente: integer
- nome: string
- email: string
- data_cadastro: date

Como Funciona:

1. **Data Catalog** (AWS Glue, Azure Data Catalog) lê o arquivo Parquet
2. **Infere automaticamente** os tipos de dados
3. **Cria metadados** (schema)
4. **Ferramentas SQL** (Athena) usam esses metadados para queries

Ferramentas de Nuvem: Storage e Catalogação

Storage de Objetos

AWS S3, Azure Data Lake Storage Gen2 e Google Cloud Storage são a base física do Data Lake. Armazenam arquivos em formato original (JSON, CSV, Parquet, logs, vídeos) de forma escalável e barata.

Catalogação (Data Catalog)

O Data Lake armazena arquivos, mas não sabe o que há dentro deles. O Data Catalog (AWS Glue, Azure Data Catalog) lê os arquivos e cria o schema (metadados), permitindo que ferramentas SQL consultem os dados.

Amazon Web Services

STORAGE

S3 (Simple Storage Service)

Armazenamento de objetos escalável. Suporta particionamento, versionamento e controle de acesso granular.

CATALOGAÇÃO

AWS Glue

Data Catalog que infere schemas. Integra com Athena para queries SQL diretas no S3.

Athena

Query SQL sem servidor. Lê dados do S3 usando metadados do Glue.

Vantagem:

Ecossistema completo e maduro. Melhor para escala massiva.

Microsoft Azure

STORAGE

Data Lake Storage Gen2

Armazenamento otimizado para Big Data. Suporta sistema de arquivos hierárquico (ADLS Gen2).

CATALOGAÇÃO

Azure Data Catalog

Gerencia metadados e descobre dados. Integra com Synapse Analytics para queries.

Synapse Analytics

Query SQL e Spark. Processa dados em ADLS Gen2 com performance alta.

Vantagem:

Integração com ecossistema Microsoft. Melhor para empresas Windows.

Google Cloud

STORAGE

Cloud Storage

Armazenamento de objetos escalável. Suporta classes de armazenamento (Standard, Nearline, Coldline).

CATALOGAÇÃO

Data Catalog

Gerencia metadados e descobre dados. Integra com BigQuery para queries.

BigQuery

Data Warehouse serverless. Queries SQL ultra-rápidas em dados do Cloud Storage.

Vantagem:

BigQuery é o melhor para análises rápidas. Ideal para Data Science.

Debate e Consolidação

Apresentação de Grupos (15-20 min): Cada grupo apresenta seu esquema de Data Lake (5 min cada). Mostrem: camadas, particionamento, políticas de acesso e roadmap de dados. Após cada apresentação, debate com a turma.

Grupo A: Varejo
Apresentação: 5 minutos

Grupo B: Saúde
Apresentação: 5 minutos

Questões Críticas para Debate:

O particionamento por hora é excessivo para a camada Ouro? Por quê?

Dica: Pense em quantas pastas seriam criadas em 1 ano (365 × 24 = 8.760 pastas).

A segurança da camada Bronze está muito rígida? Poderia ser menos restritiva?

Dica: Considere: dados brutos têm qualidade? Quem realmente precisa acessar?

Como você trataria dados sensíveis (PII) na camada Prata?

Dica: Anonimização, hasheamento, ou simplesmente não copiar para Prata?

Qual é o trade-off entre retenção de dados e custo de armazenamento?

Dica: 5 anos de dados = custo alto. 3 meses = perda de histórico para análises.

Alinhamento Coletivo: Lições Aprendidas

- → Não existe um "tamanho único" para todos os Data Lakes. Cada negócio tem necessidades diferentes.
- → Particionamento bem feito é crítico: melhora performance em 10-100x e reduz custo em 20x.
- → Segurança não é opcional: LGPD, HIPAA e conformidade exigem acesso controlado por camada.
- → O roadmap de dados (Bronze → Prata → Ouro) é a chave para transformar dados brutos em insights.

Conclusão: Impacto da Arquitetura de Data Lake

Arquitetura em Camadas

Bronze (bruta) → Prata (refinada) → Ouro (curada). Transformação progressiva de dados brutos em insights.

Particionamento Estratégico

Dividir dados por colunas frequentes (data, região). Melhora performance 10-100x e reduz custos.

Acesso Mínimo

Cada usuário acessa apenas dados necessários. Bronze restrita, Prata moderada, Ouro aberta.

Catalogação e Metadados

Data Catalog infere automaticamente tipos de dados. Ferramentas SQL usam esses metadados para queries.

Performance

Particionamento bem feito: Queries 10-100x mais rápidas. Sem particionamento: lê 1TB inteiro. Com particionamento: lê apenas 50GB.

Custo

Redução de 80-90%: Menos dados lidos = menos processamento. Formato Parquet comprime 10x vs CSV. Retenção estratégica economiza armazenamento.

Segurança

Conformidade LGPD/HIPAA: Acesso controlado por camada. Auditoria de quem acessa o quê. Dados sensíveis isolados em Bronze.

Checklist de Validação: Seu Data Lake está bem planejado?

- ☑ Defini as 3 camadas (Bronze, Prata, Ouro)?
- ☑ Defini políticas de acesso por perfil?
- ☑ Documentei o roadmap (Bronze → Ouro)?
- ☑ Estimei volume e frequência de dados?
- ☑ Escolhi colunas de particionamento?
- ☑ Planejei retenção de dados?
- ☑ Considerei conformidade (LGPD/HIPAA)?
- ☑ Escolhi ferramentas (S3, Glue, Athena)?

Reflexão Final

A **estrutura de um Data Lake é uma decisão de engenharia** que impacta diretamente a **performance, custo e segurança** de toda a estratégia de dados da empresa. Um Data Lake bem arquitetado permite que dados brutos se transformem em insights valiosos. Um Data Lake mal planejado vira um "Data Swamp" (pântano de dados) — caótico e inutilizável. A escolha é sua!