Governança de Dados em Data Lakes

Checklist e Boas Práticas

> A diferença entre um Data Lake e um "Data Swamp" é a Governança

Abertura: Data Lake vs Data Swamp

Governança de Dados é o conjunto de regras, processos e responsabilidades que garantem que os dados são confiáveis, seguros e estão em conformidade legal. Sem governança, um Data Lake se torna um "Data Swamp" (pântano) - caótico, inseguro e inútil.

Os 3 Pilares da Governança em Data Lakes

Segurança e Conformidade

Proteger contra acesso não autorizado e garantir cumprimento de leis como LGPD, HIPAA, GDPR. Criptografia, auditoria e controle de acesso.

Qualidade de Dados

Garantir que dados são precisos, completos e consistentes. Validação em cada etapa do pipeline, testes de volume e monitoramento contínuo.

Documentação e Catálogo

Saber onde os dados estão, o que significam e seu histórico de processamento (linhagem). Data Catalog e metadados documentados.

Por que Governança é Crítica em Data Lakes? A ausência de um esquema rígido (Schema-on-Read) torna a qualidade e a segurança mais difíceis de controlar. Sem governança, dados crescem sem controle, segurança fica comprometida e conformidade legal é violada. Governança transforma um pântano em um ativo estratégico.

Os 3 Pilares da Governança em Data Lakes

Segurança e Conformidade

DEFINIÇÃO

Proteger dados contra acesso não autorizado e garantir conformidade legal

CARACTERÍSTICAS

- Criptografia em repouso e trânsito
- Controle de acesso por Roles
- Auditoria e logs de acesso
- Anonimização de dados sensíveis

TECNOLOGIAS

- AWS KMS, Azure Key Vault
- IAM, CloudTrail
- LGPD, HIPAA, GDPR

EXEMPLO

Dados de pacientes: CPF mascarado, acesso restrito a médicos, logs de quem acessou

Qualidade de Dados

DEFINIÇÃO

Garantir que dados são precisos, completos e consistentes em cada etapa

CARACTERÍSTICAS

- Validação de volume
- Verificação de consistência
- Formato otimizado (Parquet)
- Particionamento correto

TECNOLOGIAS

- Spark, SQL validações
- Delta Lake, Parquet
- Data Quality tools

EXEMPLO

Vendas: validar se volume não cai > 10%, nenhuma chave nula, formato Parquet

Documentação e Catálogo

DEFINIÇÃO

Garantir que usuários saibam onde os dados estão e o que significam

CARACTERÍSTICAS

- Catálogo de datasets
- Documentação de colunas
- Linhagem de dados
- Mapa de responsabilidades

TECNOLOGIAS

- AWS Glue Catalog
- Azure Data Catalog
- Apache Atlas

EXEMPLO

Cada dataset tem: descrição, owner, SLA, histórico de transformações

Desafios Únicos do Data Lake

1 Schema-on-Read vs Schema-on-Write

DILEMA

Data Lake: Schema-on-Read (flexível, mas difícil de governar). Data Warehouse: Schema-on-Write (rígido, mas fácil de governar).

SOLUÇÃO

Usar Data Catalog para impor schema. Validar schema em cada etapa do pipeline (Bronze → Prata → Ouro).

2 Qualidade vs Flexibilidade

DILEMA

Quanto mais flexível o Data Lake, menos controle sobre qualidade. Dados entram sem validação rigorosa.

SOLUÇÃO

Validação em cada camada. Bronze: aceita tudo. Prata: valida. Ouro: garante qualidade máxima.

3 Escalabilidade vs Governança

DILEMA

Dados crescem rapidamente (TB/dia). Governança fica para trás. Novos datasets chegam sem documentação.

SOLUÇÃO

Automatizar governança: Data Catalog automático, Lineage automático, Validação automática com DAGs.

4 Privacidade vs Usabilidade

DILEMA

Anonimizar dados reduz utilidade para análises. Manter dados brutos viola privacidade (LGPD).

SOLUÇÃO

Diferentes níveis de acesso: Bronze (bruto, acesso restrito), Prata (anonimizado), Ouro (agregado).

Checklist: Segurança e Conformidade

1

☐ Acesso Mínimo (Princípio do Menor Privilégio)

As permissões são concedidas por Papéis/Roles e não por usuários individuais. Cada usuário acessa apenas o que precisa.

IMPLEMENTAÇÃO
IAM com Roles, não usuários individuais

BENEFÍCIO
Reduz risco de acesso indevido

2

☐ Segregação de Camadas

A camada Bronze está isolada da camada Ouro em redes separadas. Bronze: dados brutos (engenheiros). Prata: dados limpos (cientistas). Ouro: dados finais (analistas).

IMPLEMENTAÇÃO
VPCs, Security Groups, Network Policies

BENEFÍCIO
Protege dados sensíveis em camadas

3

☐ Criptografia de Dados

Todos os buckets de armazenamento (S3, ADLS) possuem criptografia ativada. Dados em repouso (storage) e em trânsito (HTTPS/TLS) devem ser criptografados.

IMPLEMENTAÇÃO
AWS KMS, Azure Storage Encryption, TLS

BENEFÍCIO
Protege contra roubo de dados

4

☐ LGPD/GDPR/Anonimização

Dados sensíveis (CPF, Nomes, Emails) são mascarados ou anonimizados antes de serem movidos para a camada Prata. Conformidade com LGPD, GDPR, HIPAA.

IMPLEMENTAÇÃO
Hashing, Masking, Tokenization em Spark

BENEFÍCIO
Conformidade legal, privacidade garantida

5

☐ Auditoria e Logs

Existe um log de quem acessou e modificou cada arquivo. Registrar: Usuário, Timestamp, Ação (read/write/delete), Resultado. Armazenar em local seguro e imutável.

IMPLEMENTAÇÃO
CloudTrail (AWS), Activity Log (Azure)

BENEFÍCIO
Rastreabilidade, investigação de incidentes

Atividade 1: Desenvolvimento do Checklist de Segurança

Cenário: Um hospital precisa proteger dados de pacientes (CPF, Nome, Histórico Médico) em um Data Lake. Vocês devem definir as regras de segurança para as 3 camadas (Bronze, Prata, Ouro) e criar um checklist de conformidade LGPD.

Passos da Atividade

1 Definir Papéis e Permissões

Listar 3 papéis (Engenheiro, Cientista, Médico). Para cada papel, definir: quais camadas pode acessar, quais operações (read/write/delete), restrições específicas.

2 Definir Transformações de Segurança

Para dados sensíveis (CPF, Nome): definir técnica de proteção (Masking, Hashing, Tokenização). Exemplo: CPF → SHA256(CPF), Nome → "Paciente_XXX".

3 Definir Criptografia e Auditoria

Especificar: criptografia em repouso (KMS), criptografia em trânsito (HTTPS), logs de auditoria (quem acessou, quando, o quê).

4 Montar o Checklist de Segurança

Consolidar todos os itens em um checklist: Acesso Mínimo, Segregação, Criptografia, LGPD/Anonimização, Auditoria. Cada item deve ter: descrição, implementação, responsável.

Exemplo: Segurança nas 3 Camadas

Bronze (Brutos)

Acesso: Engenheiros
Dados: CPF, Nome, Histórico completo
Criptografia: KMS ativada

Prata (Limpos)

Acesso: Cientistas
Dados: CPF mascarado, Nome anonimizado
Criptografia: KMS ativada

Ouro (Finais)

Acesso: Médicos, Analistas
Dados: Apenas agregações
Criptografia: KMS ativada

Checklist de Segurança (5 Itens Essenciais)

- ☐ Acesso Mínimo: Papéis/Roles definidos?
- ☐ Criptografia: KMS ativada em todos os buckets?
- ☐ Auditoria: Logs de acesso registrados?
- ☐ Segregação: Camadas isoladas em redes?
- ☐ LGPD: Dados sensíveis mascarados/anonimizados?
- ☐ Conformidade: Documentação de políticas pronta?

Checklist: Qualidade e Documentação

Validação de Volume

O QUE É?

Verificar se o número de linhas ingestadas está dentro do esperado

IMPLEMENTAÇÃO

```
if (linhas_recebidas / linhas_esperadas) < 0.9:
  raise Exception("Queda de volume!")
```

BENEFÍCIO

Detecta problemas na ingestão rapidamente

Verificação de Consistência

O QUE É?

Verificar que campos obrigatórios (PKs) não têm valores nulos

IMPLEMENTAÇÃO

```
SELECT COUNT(*) FROM dados
WHERE customer_id IS NULL
if count > 0: raise Exception()
```

BENEFÍCIO

Garante integridade dos dados

Padrão de Formato

O QUE É?

Usar formato otimizado (Parquet/Delta Lake) em camadas Prata e Ouro

IMPLEMENTAÇÃO

```
df.write.format("parquet")
  .mode("overwrite")
  .save(path)
```

BENEFÍCIO

Melhor performance, compressão e schema enforcement

Particionamento Correto

O QUE É?

Particionar dados por coluna de query frequente (data, região)

IMPLEMENTAÇÃO

```
PARTITION BY year, month, day
Exemplo: /gold/sales/
year=2025/month=01/day=15/
```

BENEFÍCIO

Queries mais rápidas, menos dados lidos

Catalogação no Data Catalog

O QUE É?

Registrar todos os datasets com metadados (nome, owner, frequência)

IMPLEMENTAÇÃO

```
Dataset: gold_sales
Owner: analytics@company.com
Frequência: Diária (2am)
```

BENEFÍCIO

Descoberta de dados, documentação centralizada

Linhagem de Dados

O QUE É?

Documentar histórico de transformações (Bronze → Prata → Ouro)

IMPLEMENTAÇÃO

```
bronze_raw → prata_cleaned
→ gold_aggregated
(Usar Apache Atlas ou Glue)
```

BENEFÍCIO

Entender origem dos dados, facilita debug

Atividade 2: Desenvolvimento do Checklist de Qualidade

Cenário: Plataforma de e-commerce com dados de pedidos, clientes e produtos. Seu grupo deve criar um checklist de qualidade que garanta: volume consistente de pedidos, nenhuma chave nula, formato otimizado, particionamento correto, catalogação completa e linhagem documentada.

Passos da Atividade

1. **Definir Validações de Volume**
   Estabelecer: volume esperado diário (ex: 10.000 pedidos), mínimo aceitável (90%), máximo aceitável (110%). Documentar o que fazer se sair desse intervalo.

2. **Definir Verificações de Consistência**
   Listar campos obrigatórios (order_id, customer_id, order_date, amount). Para cada um, descrever: tipo, se pode ser nulo, validação esperada.

3. **Definir Padrões de Formato e Particionamento**
   Especificar: formato final (Parquet/Delta Lake), compressão, particionamento (por data, região). Exemplo: /gold/orders/year=2025/month=01/day=15/

4. **Definir Catalogação e Linhagem**
   Documentar: nome do dataset, descrição, owner, frequência, SLA. Desenhar a linhagem: bronze_raw → silver_cleaned → gold_aggregated

EXEMPLO: CHECKLIST DE QUALIDADE PARA E-COMMERCE

Volume Esperado
10.000 pedidos/dia (min: 9.000, max: 11.000)

Campos Obrigatórios
order_id, customer_id, order_date, amount (NOT NULL)

Formato Final
Parquet, particionado por year/month/day

SLA
Atualizar até 3am, 99.9% disponibilidade

CHECKLIST: ITENS A INCLUIR NO SEU CHECKLIST

Exemplo Integrado: Banco de Dados

Cenário: Um banco precisa gerenciar dados de clientes (CPF, Nome, Saldo) e transações (ID, Valor, Data) em um Data Lake. Como aplicar os 3 pilares (Segurança, Qualidade, Documentação) de forma integrada?

Segurança e Conformidade

ACESSO POR PAPÉIS

Engenheiros: Bronze (dados brutos). Cientistas: Prata (dados limpos). Médicos/Analistas: Ouro (agregados).

ANONIMIZAÇÃO LGPD

CPF → SHA256(CPF), Nome → "Cliente_XXX" antes de Prata

CRIPTOGRAFIA

KMS em repouso, HTTPS em trânsito

```
Bronze: CPF, Nome, Saldo (bruto)
Prata: CPF_hash, Nome_anon, Saldo
Ouro: Apenas agregações
```

Qualidade de Dados

VALIDAÇÃO DE VOLUME

Esperado: 100.000 clientes/dia. Alerta se cair < 90.000

CONSISTÊNCIA

customer_id, transaction_id: NOT NULL. Nenhuma chave pode ser nula

FORMATO E PARTICIONAMENTO

Parquet, particionado por year/month/day

```
Ouro: /transactions/
year=2025/month=01/day=15/
data.parquet
```

Documentação e Catálogo

CATALOGAÇÃO

Dataset: gold_customer_360. Owner: analytics@bank.com. Frequência: Diária (1am)

DOCUMENTAÇÃO DE COLUNAS

customer_id: ID único. transaction_date: Data da transação. amount: Valor em R$

LINHAGEM

bronze_raw → silver_cleaned → gold_aggregated

```
SLA: Atualizar até 2am
Backup: joao@bank.com
Última atualização: 2025-01-15 01:45
```

Integração dos 3 Pilares: Segurança define quem acessa e como proteger. Qualidade garante que dados são confiáveis. Documentação permite que todos entendam e usem os dados corretamente. Juntos, transformam um Data Lake em um ativo estratégico confiável.

Atividade 3: Apresentação e Alinhamento Coletivo

Objetivo: Cada grupo apresenta seus checklists de Segurança e Qualidade (5-10 min). Consolidamos um Checklist Mestre com os melhores itens. Discutimos responsabilidade ética e governança como processo contínuo.

Passos da Apresentação

1 Apresentação do Grupo (5-10 min)

Cada grupo apresenta: cenário, checklist de Segurança (5 itens), checklist de Qualidade (6 itens). Justificar por que cada item é importante.

2 Consolidação do Checklist Mestre (10 min)

Facilitar: combinar os melhores itens de cada grupo. Resultado: Checklist Mestre com 17 itens (5 Segurança + 6 Qualidade + 6 Documentação).

3 Discussão Ética (5 min)

Questões para debate: Qual é a responsabilidade do Engenheiro de Dados? Como garantir conformidade legal? O que fazer se descobrir um problema?

4 Reflexão Final (5 min)

Reforçar: Governança é um processo contínuo. O checklist deve ser atualizado quando novas leis surgem ou novas fontes de dados chegam.

CHECKLIST MESTRE: 17 ITENS DE GOVERNANÇA

- ☐ Acesso Mínimo (Roles)
- ☐ LGPD/Anonimização
- ☐ Verificação de Consistência
- ☐ Catalogação no Data Catalog
- ☐ Documentação de Colunas
- ☐ Mapa de Responsabilidades
- ☐ Segregação de Camadas
- ☐ Auditoria e Logs
- ☐ Padrão de Formato
- ☐ Linhagem de Dados
- ☐ Documentação de Transformações
- ☐ Histórico de Mudanças
- ☐ Criptografia de Dados
- ☐ Validação de Volume
- ☐ Particionamento Correto
- ☐ Documentação de Datasets
- ☐ Documentação de SLAs

QUESTÕES PARA DISCUSSÃO ÉTICA

❓ Qual é a responsabilidade do Engenheiro de Dados na governança?

❓ Como garantir conformidade legal (LGPD) sem sacrificar usabilidade dos dados?

Conclusão: Governança é um Processo Contínuo

Governança não é um projeto, é um processo contínuo de melhoria e adaptação

1. **Processo Contínuo**
   Revisar e atualizar governança regularmente. Não é "fazer uma vez e esquecer".

2. **Envolver Todas as Partes**
   Engenheiros, Cientistas, Analistas, Compliance, Segurança. Governança é responsabilidade coletiva.

3. **Automatizar ao Máximo**
   Data Catalog automático, Lineage automático, Validação automática. Reduz carga manual.

4. **Documentar Tudo**
   Datasets, Transformações, Responsabilidades, Mudanças. Documentação é viva, não estática.

5. **Monitorar Continuamente**
   Alertas para violações, Dashboards de governança, Auditorias regulares. Visibilidade total.

CHECKLIST MESTRE: 17 ITENS DE GOVERNANÇA

Segurança (5)

- ✓ Acesso Mínimo
- ✓ Segregação
- ✓ Criptografia
- ✓ LGPD/Anonimização
- ✓ Auditoria

Qualidade (6)

- ✓ Validação Volume
- ✓ Consistência
- ✓ Formato
- ✓ Particionamento
- ✓ Catalogação
- ✓ Linhagem

Documentação (6)

- ✓ Datasets
- ✓ Colunas
- ✓ Transformações
- ✓ SLAs
- ✓ Responsabilidades
- ✓ Histórico Mudanças

REFLEXÃO FINAL: RESPONSABILIDADE COMPARTILHADA

*A qualidade e a conformidade legal dos dados dependem do **rigor aplicado na construção dos pipelines** e na **definição das permissões**. Cada membro da equipe tem responsabilidade na governança. Domine esses conceitos e você será capaz de construir sistemas de dados que são **confiáveis, seguros e impactam realmente o negócio**.*