Engenharia de Dados

10/02/2021

Programa de Aceleração Banco Inter

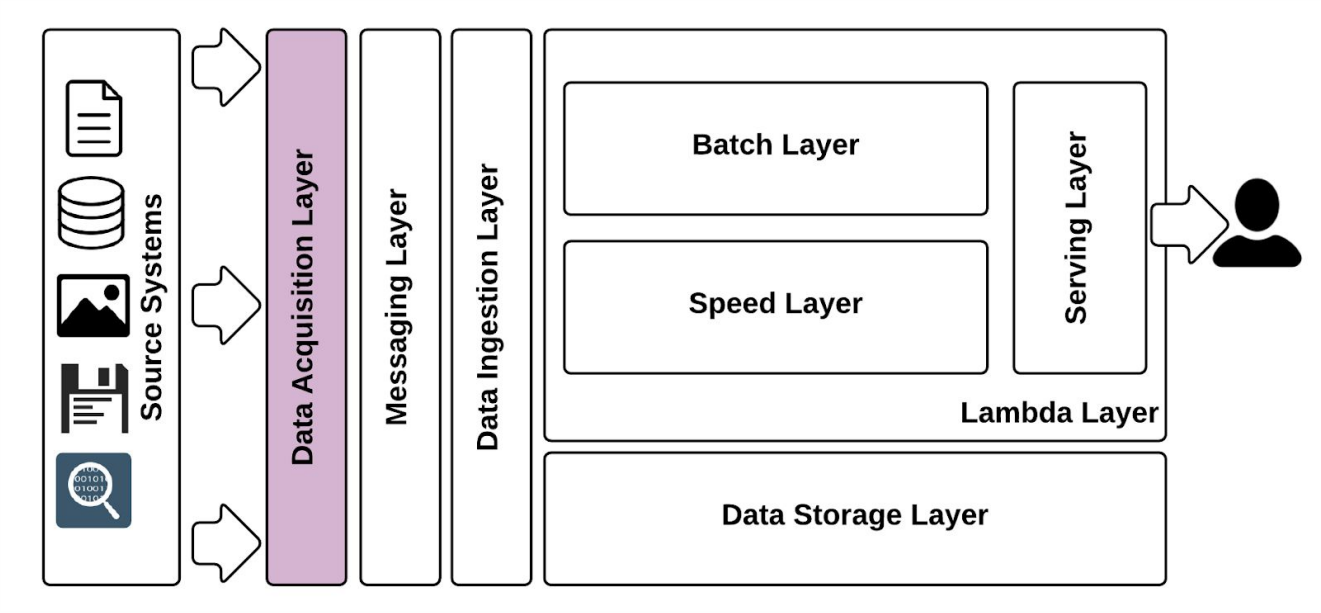
Processamento de Dados em Big Data

datasprints

Agenda

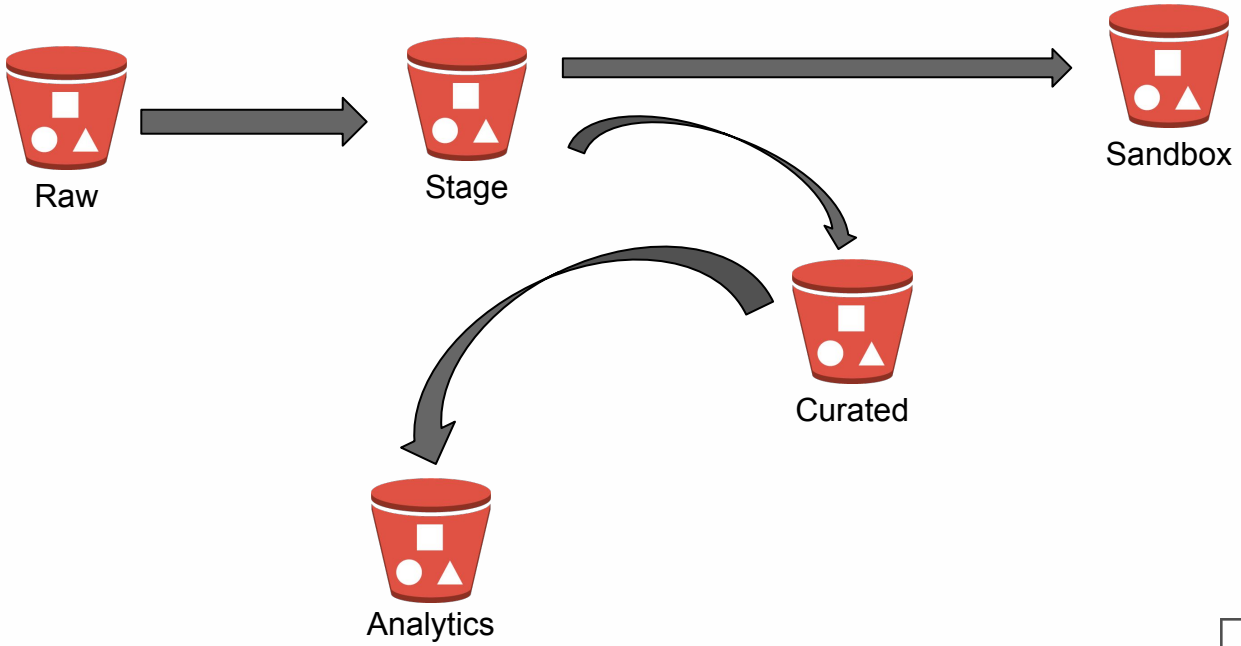
1. Camadas de um Data Lake;
2. Organização de uma Data Lake;
3. Estratégias para processamento de dados;
4. Ferramentas para processar dados;
5. Processamento de dados na AWS;
6. Data Build Tool (DBT).

Camadas de um Data Lake



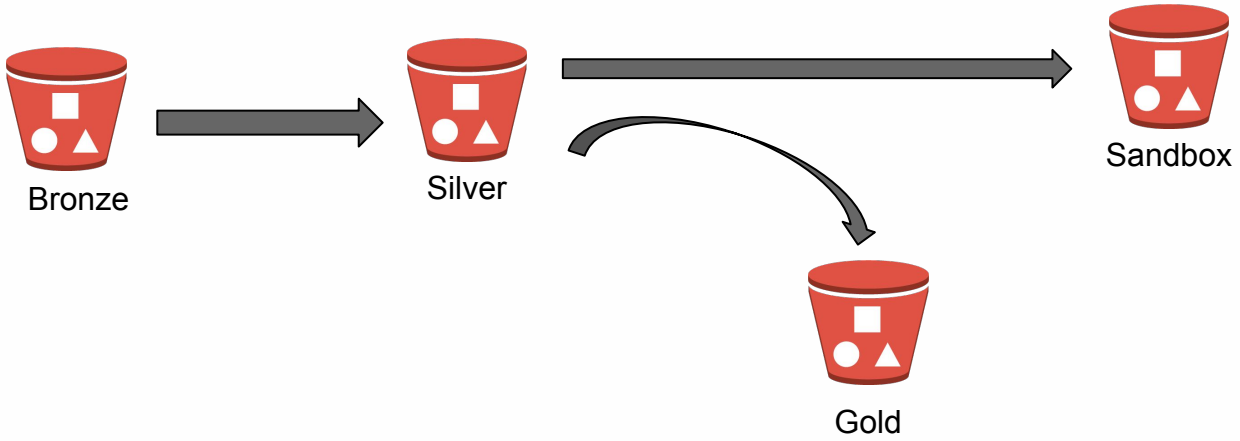
Organização de um Data Lake

- Possível modelo de Armazenamento 1:

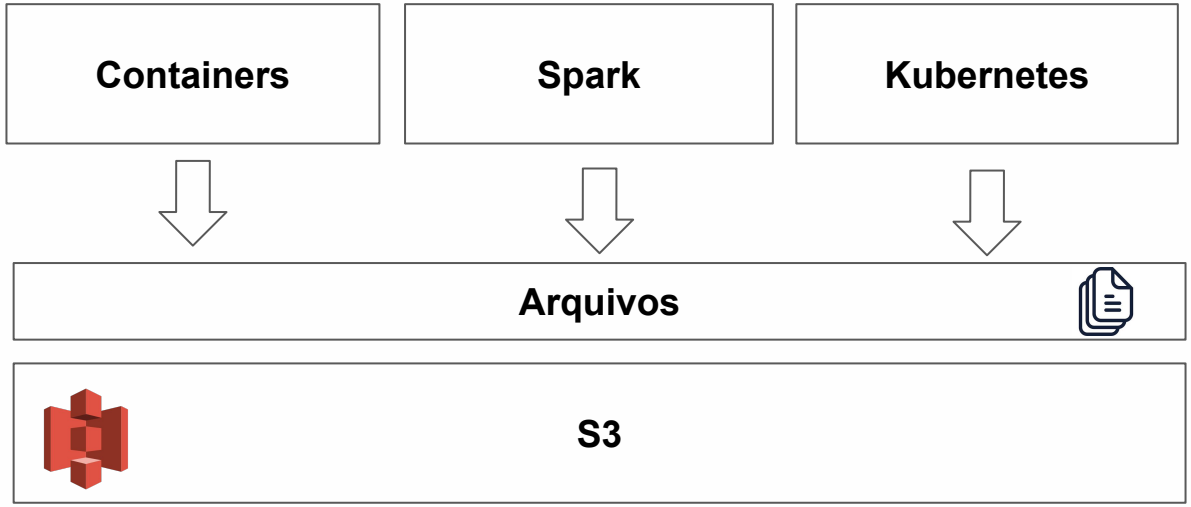


Organização de um Data Lake

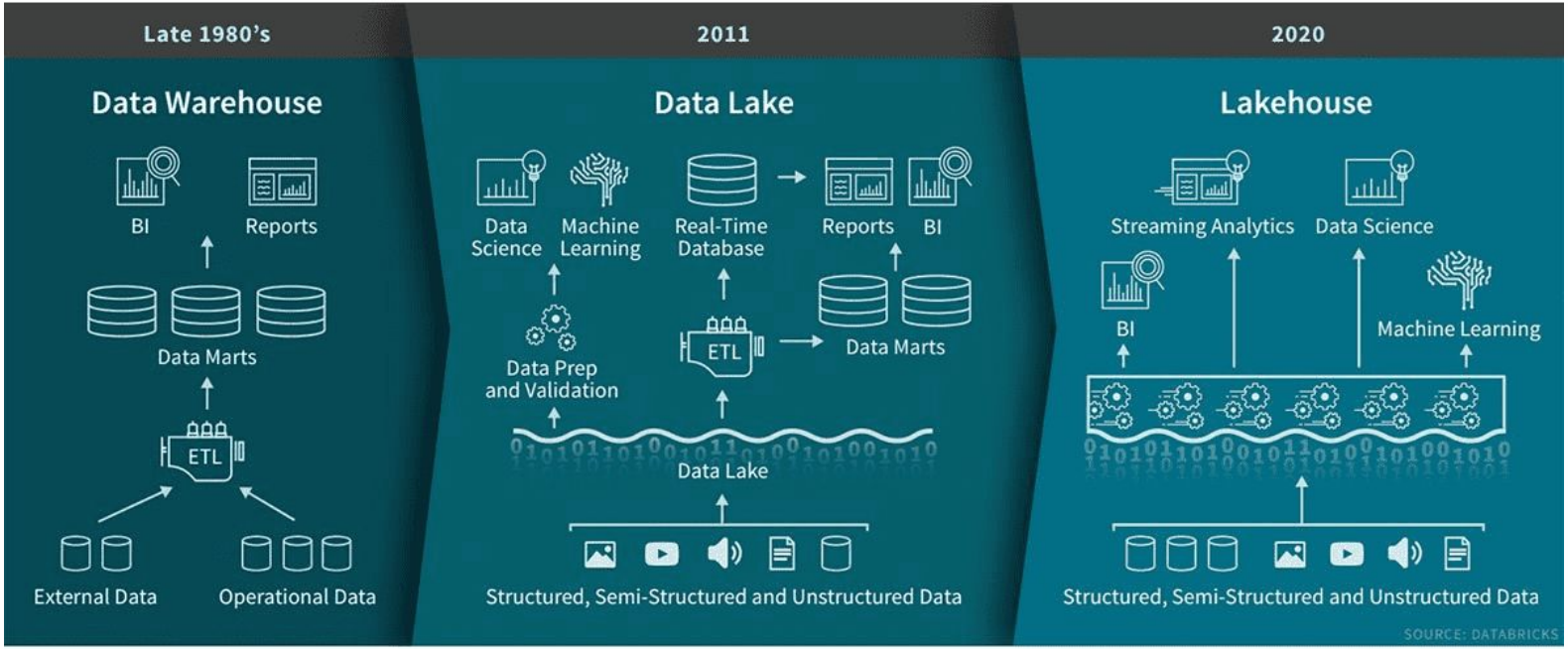
- Possível modelo de Armazenamento 2:



Processamento e Armazenamento



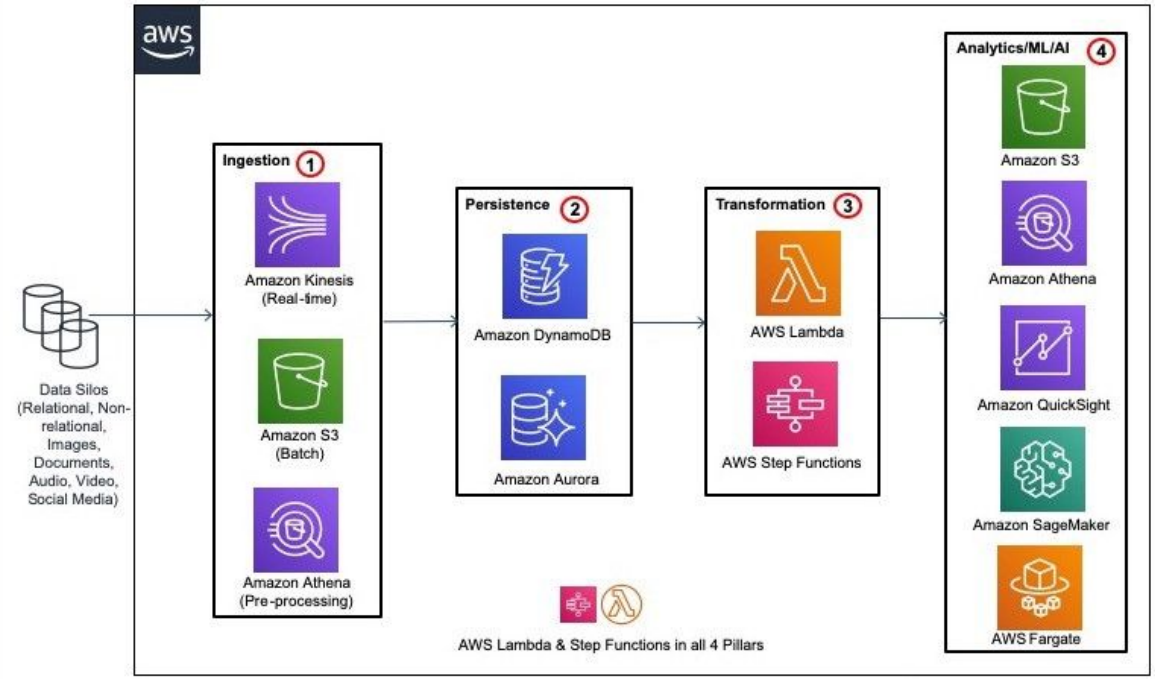
Estratégias de Processamento



https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

datasprints

Estratégias de Processamento



https://towardsdatascience.com/architecting-serverless-data-integration-hubs-on-aws-for-enterprise-data-delivery-2020-a27b42569518

Ferramentas para Processamento

- Alguns Exemplos:
  - DBT;
  - Spark;
  - Apache Hudi;
  - Glue ETL;
  - Athena;
  - Presto;
  - NiFi;
  - Dremio;
  - outros.

Processamento de dados na AWS

- Glue:
  - Grawlers;
  - ETL;
  - Catálogo de dados.
- Athena:
  - Query em arquivos não estruturados;
  - Forma distribuída;
  - Banco e tabelas;

Processamento de dados na AWS

- EMR:
  - Spark e Hadoop;
  - Processamento distribuído;
  - Diversos frameworks e serviços.
- Lambda:
  - Funções serverless;
  - Integração com diversos serviços AWS;
  - Processamento sob demanda.

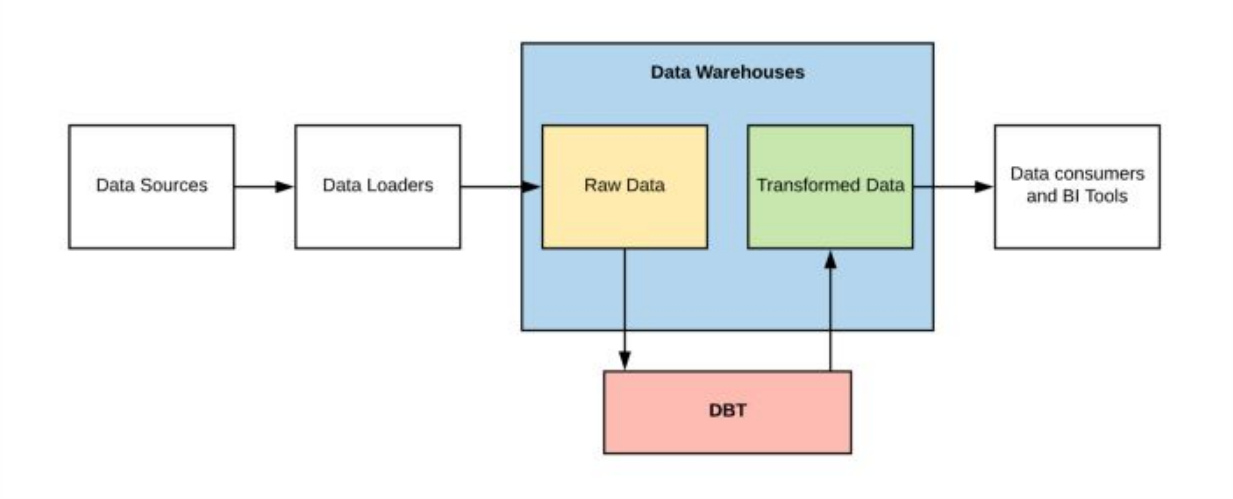
Processamento de dados na AWS

- AWS ECS:
  - Serviço de containers;
  - Execução em Fargate e EC2.
- AWS EKS
  - Kubernetes gerenciado;
  - Possibilidade de utilizar o Fargate.

DBT (Data Build Tool)

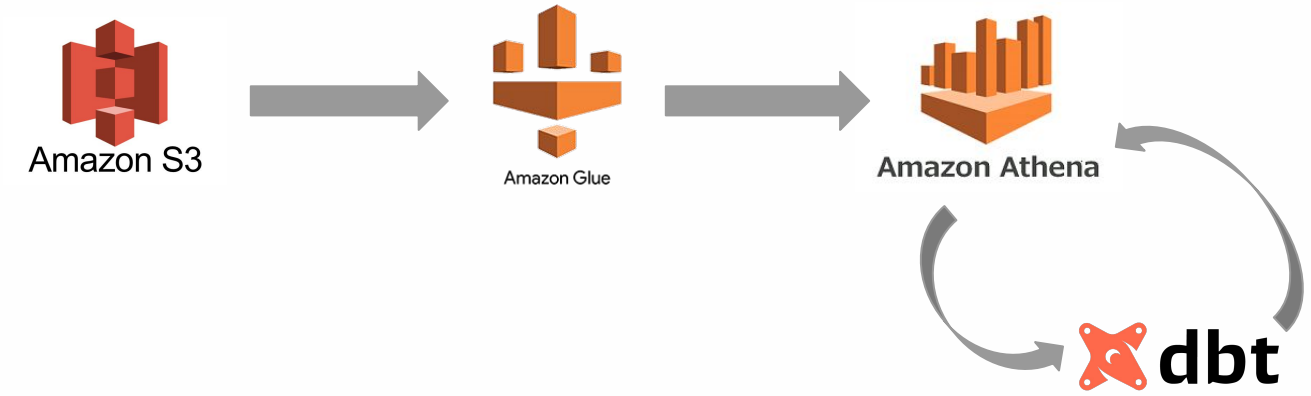
- Ferramenta moderna de ETL;
- Controle de fluxo para Stack's de DW;
- Funcionalidades:
  - CLI e UI;
  - Data Lineage;
  - Interface para Python;
  - ETL em SQL:
    - Modularização;
    - Tabelas temporárias;

DBT - Arquitetura



https://medium.com/the-telegraph-engineering/dbt-a-new-way-to-handle-data-transformation-at-the-telegraph-868ce3964eb4

DBT - Case



Atividades propostas

1. Produtizar pipeline com o DBT;
2. Hello World com Dremio;

Obrigado!

datasprints