Engenharia de Dados

05/02/2021

Programa de Aceleração Banco Inter

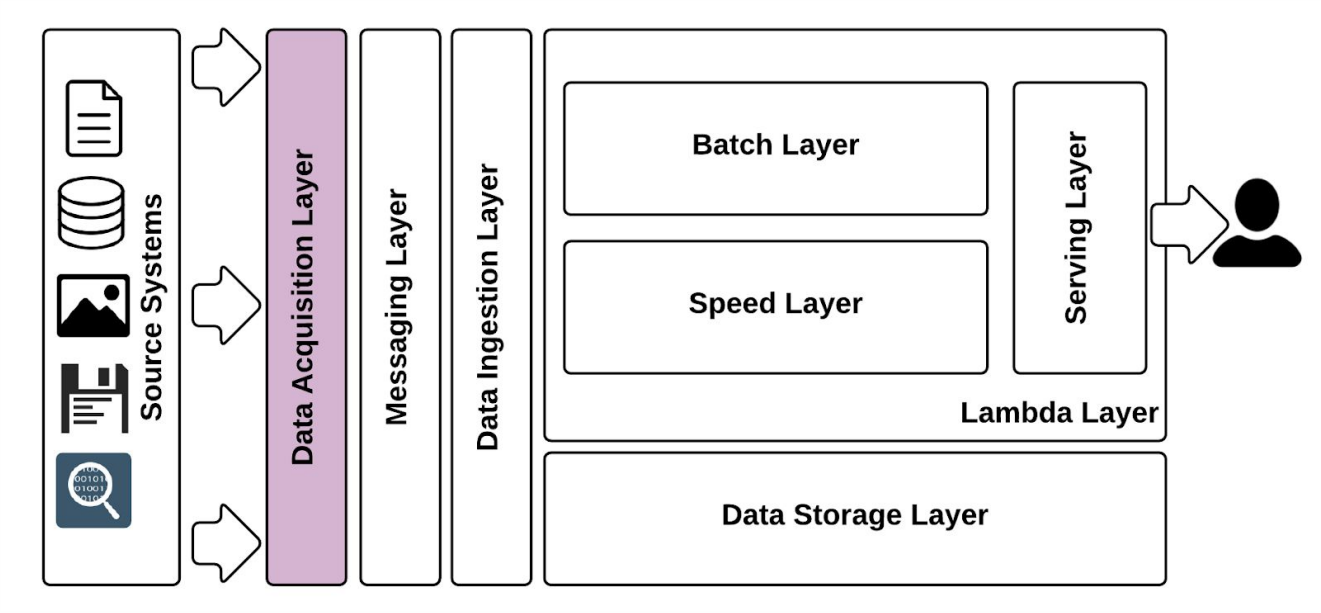
Armazenamento de Dados em Big Data

datasprints

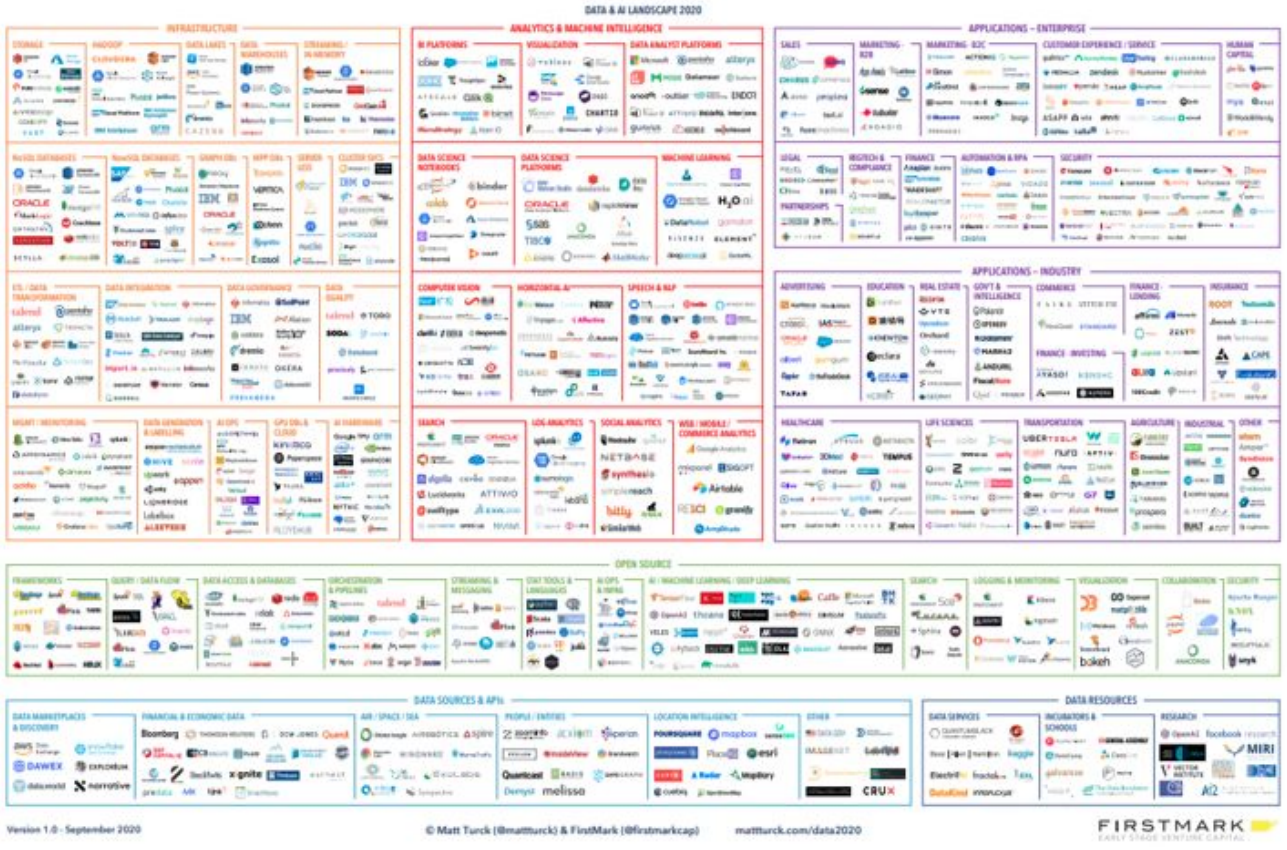
Agenda

1. Camadas de um Data Lake;
2. Ecossistema Hadoop;
3. Banco de SQL e NoSQL;
4. Armazenamento de dados na Nuvem;
5. Armazenamento de dados na AWS.

Camadas de um Data Lake



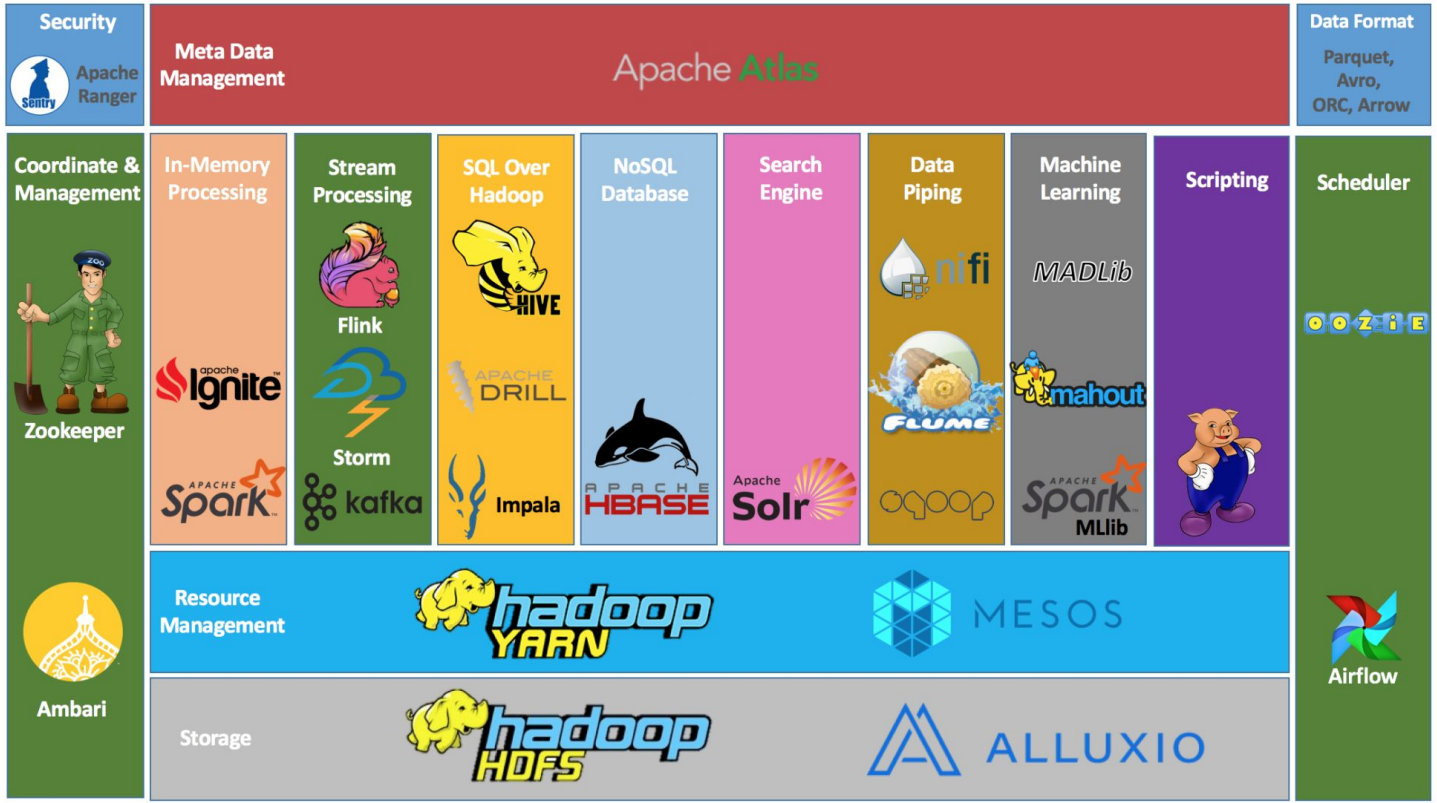
Landscape Big Data



https://mattturck.com/data2020/

datasprints

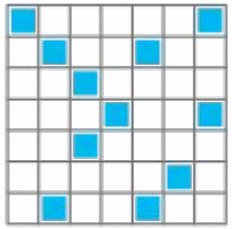
Ecossistema Hadoop



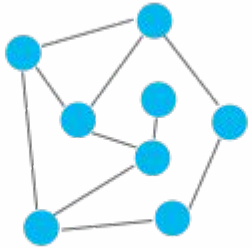
Banco de Dados SQL

- Bancos SQL:
  - Postgres;
  - MySQL;
  - SQLServer;
  - Redshift;

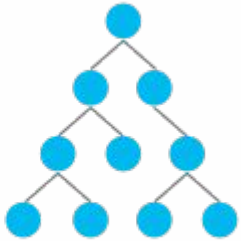
Banco de Dados NoSQL



Column-Family



Graph



Document

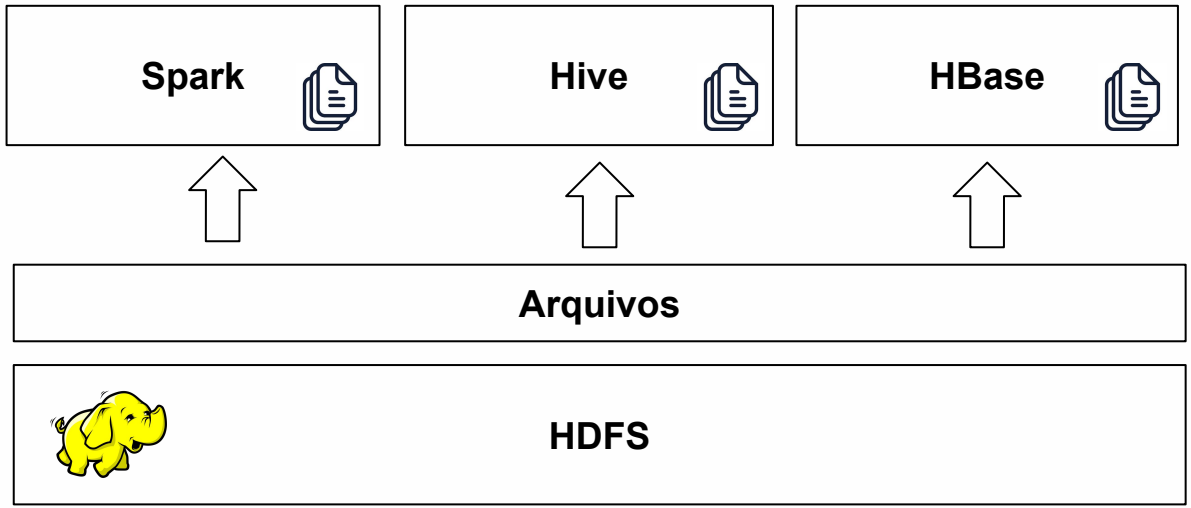
Key-Value

Banco de Dados NoSQL

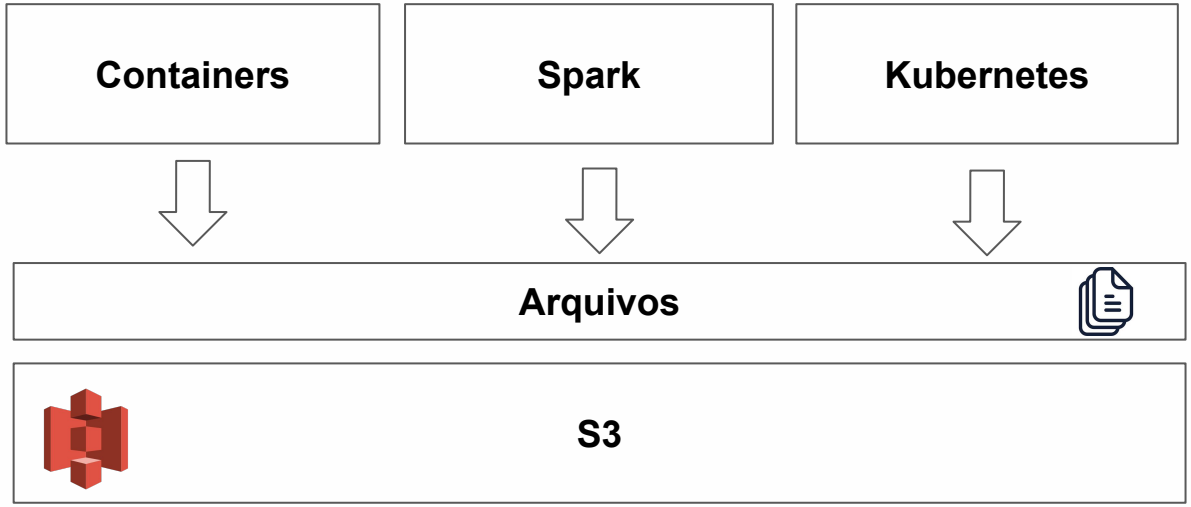
- Família de colunas:
  - HBase;
  - Cassandra;
- Chave e valor:
  - Redis;
  - EhCache.
- Orientado a documentos:
  - Mongo;
  - DynamoDB.
- Grafo:
  - Neo4j;
  - Cosmos DB.

https://db-engines.com/en/ranking

Processamento e Armazenamento



Processamento e Armazenamento



Armazenamento na AWS

- Principais:
  - S3;
  - RDS;
  - DynamoDB;
  - Elastic Cache Service;
  - Elastic Search Service;
  - Amazon Glacier.

Atividades propostas

1. Executar e avaliar código do Airflow no repositório;
2. Executar e modificar código do Lambda com DynamoDB do repositório.

Obrigado!

datasprints