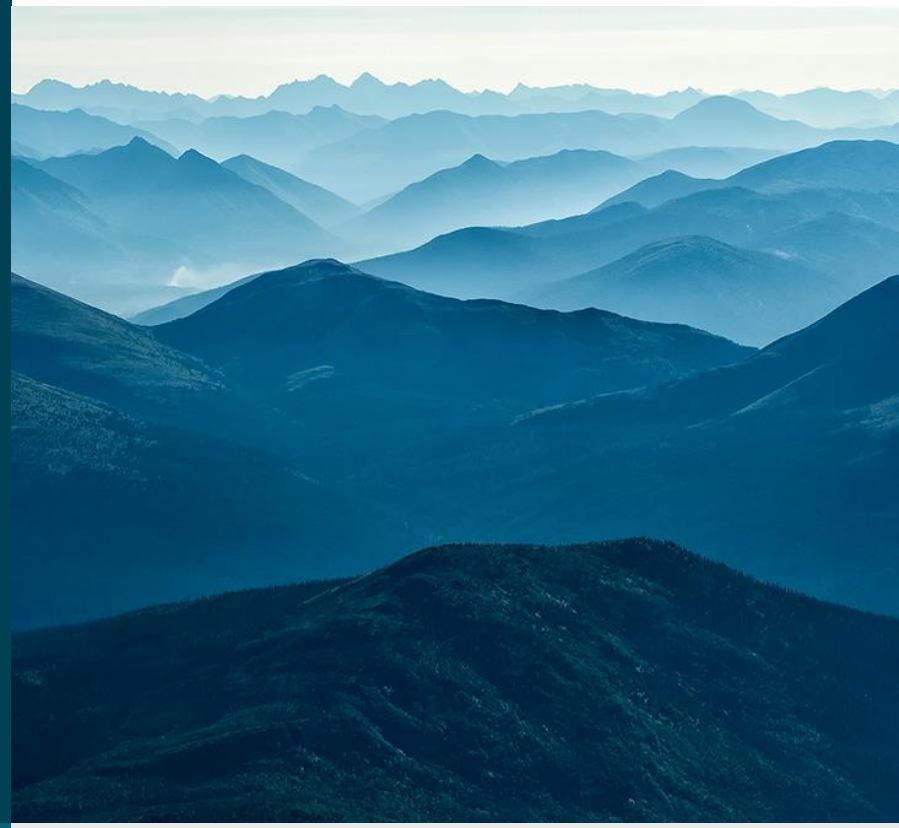
Engenharia de Dados Big Data

Aula 2 - Data Lake e Ambiente de Desenvolvimento



18/10/2021

Índice

O que é Data Lake?

Um data lake é um repositório central que contém grandes quantidades de dados em seu formato bruto.
Os dados são armazenados com tags de metadados e um identificador único, o que torna mais fácil localizar e recuperar dados entre as regiões e melhorar o desempenho, os data lakes permitem que muitos aplicativos consultem esses dados.

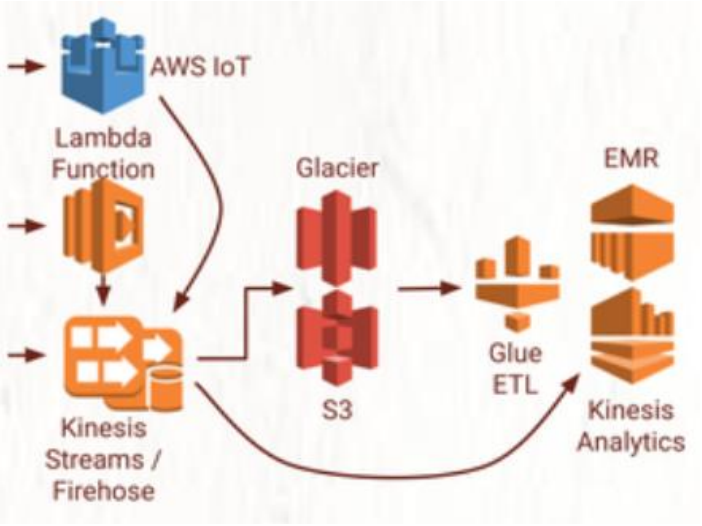
1

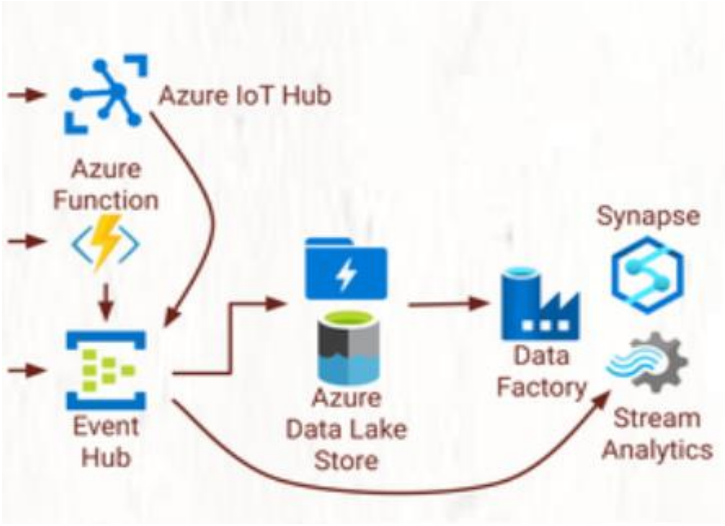
Data Lake AWS, GCP, Azure

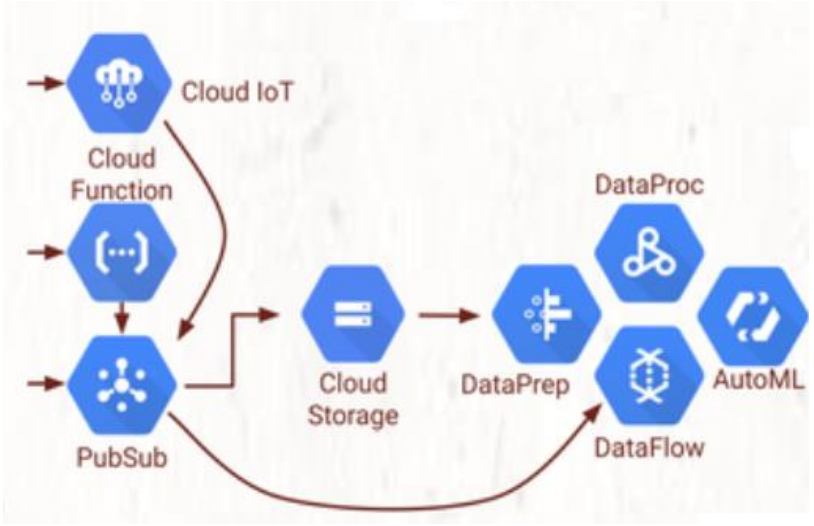
aws

Azure

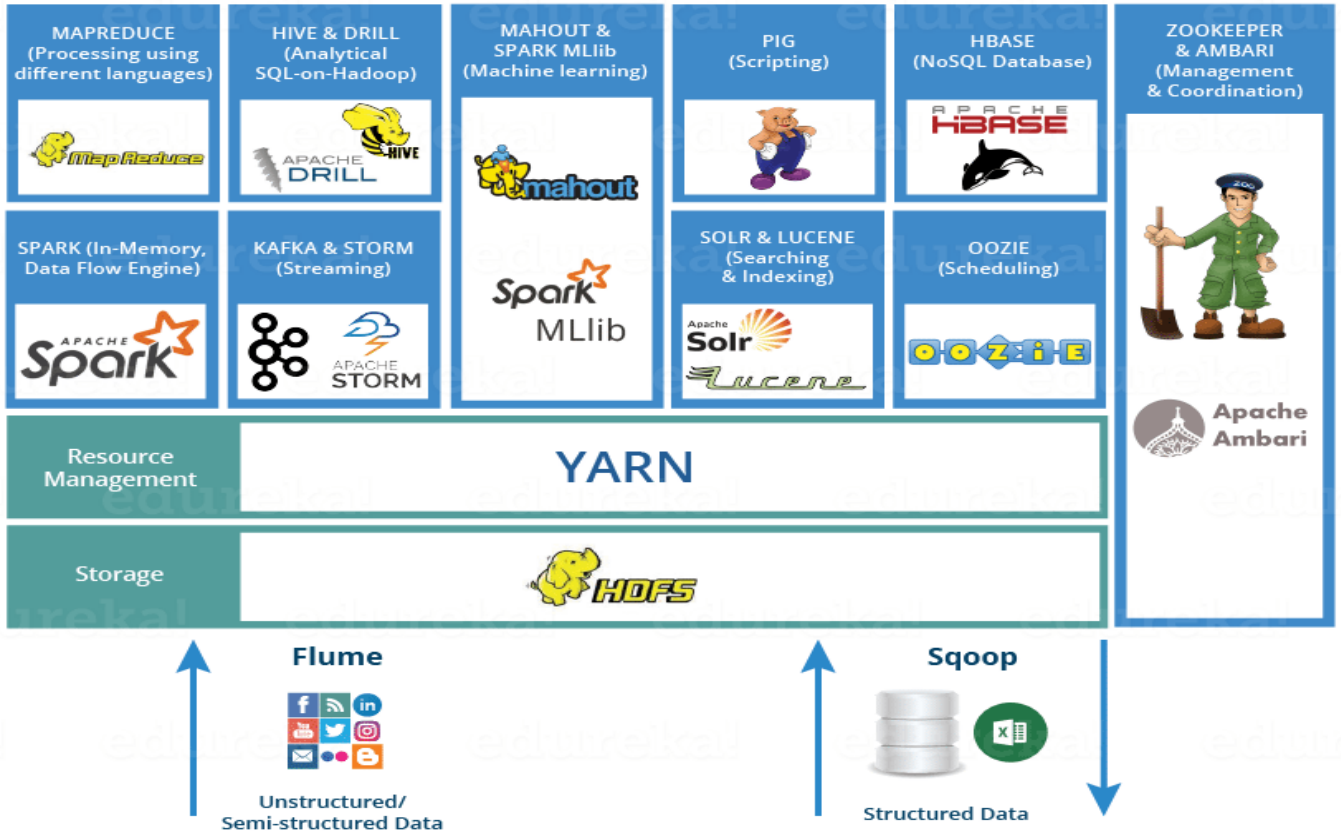
Google Cloud







Ambiente de Desenvolvimento



2

Ambiente de Desenvolvimento – Análise de Dados

Ambiente de Desenvolvimento – Hive

Apache Hive

- Criado pelo Facebook em 2007
  - Facilitar a consulta aos dados do Hadoop
  - Criação do Data warehouse no Hadoop
  - Facilitar o processamento do MapReduce

- Ferramenta para permitir fácil acesso aos dados via SQL
  - Data warehouse construído em cima do Hadoop
  - Camada de acesso a dados armazenados no HDFS
  - Não é um SGBD

- Criar tabelas no Hive (Metadados)
- Dados são armazenadas no HDFS

Ambiente de Desenvolvimento – Hive

Componentes:

HCatalog
- Camada de gerenciamento de armazenamento para o Hadoop
- Permite que usuários com diferentes ferramentas de processamento de dados leiam e gravem os dados

WebHCat
- Servidor web para se conectar com o Metastore Hive

HiveServer2 (HS2)
- Serviço que permite aos clientes executar consultas no Hive Metastore
- Todos os metadados das tabelas e partições do Hive são acessados através do Hive

Metastore
- Usado pelo Hive para armazenar metadados das tabelas do Hive armazenadas no HDFS. Metastores podem ser locais, incorporados ou remotos

Beeline
- Cliente Hive
- Faz uso de JDBC para se conectar ao HiveServer2

Ambiente de Desenvolvimento – Hive

Estrutura de Dados

Conector para vários formatos:

- Arquivos de texto com valores separados por vírgula e tabulação (CSV / TSV)
- Parquet
- ORC
- AVRO
- JSONFILE
- Outros ...

Ambiente de Desenvolvimento – Hive

Estrutura de Dados

Dados estruturados e semi-estruturados

Hierarquia dos dados

- Database
  - Table
    - Partition - Coluna de armazenamento dos dados no sistema de arquivo (diretórios)
      - Bucket - Dados são divididos em uma coluna através de Hash

Exemplo de caminho

/user/hive/warehouse/banco.db/tabela/data=010119/000000_0

Ambiente de Desenvolvimento – Hive

Liguagem

Hive Query Language

HiveQL
HQL

- Instruções SQL são transformadas internamente em Jobs de MapReduce

Ambiente de Desenvolvimento – Hive

Liguagem

Listar todos os BD
- show database;

Estrutura sobre o bd
- describe database <nomeBD>;

Listar as tabelas
- show tables;

Estrutura da tabela
- describe <nomeTabela>;
- describe formatted <nomeTabela>;
- describe extended <nomeTabela>;

Ambiente de Desenvolvimento – Hive

Liguagem

Criar BD
• create database <nomeBanco>;

Local diferente do conf. Hive
• create database <nomeBanco> location “/diretorio”;

Adicionar comentário
• create database <nomeBanco> comment “descrição”;

Ex
• create database test location “/user/hive/warehouse/test” comment “banco de dados para treinamento”
• default
o /user/hive/warehouse/test.db

Ambiente de Desenvolvimento – Hive

Liguagem

Tipo
- Internas
- Externas

Partição
- Não particionada
- Particionada
  Dinâmico
  Estático

Tabela interna
- create table user(cod int, name string);
- drop table

Apaga os dados e metadados

Tabela externa
- create external table e_user(cod int, name string)

location '/user/teste/data_users';
- drop table

Usar para compartilhar os dados com outras ferramentas
Apaga apenas os metadados
Dados ficam armazenado no sistema de arquivos

Ambiente de Desenvolvimento – Hive

Views

Salvar consultas
  Tratar como tabelas
  Objetos Lógicos

- Esquema é fixo quando criado a View

- Alterar tabela não altera a view

Comando
- create view <nomeView> as select * from nome_table;

Ambiente de Desenvolvimento – Hive

Tabelas Temporárias

- Criar tabelas voláteis no HDFS
- Dados são persistidos apenas durante a sessão do Hive
- Comando

```
CREATE TEMPORARY TABLE tmp AS

SELECT c2, c3, c4

FROM mytable;
```

*Caso não seja especificada a base, a tabela é criada na base default. Ex: default.tmp


indra
At the core