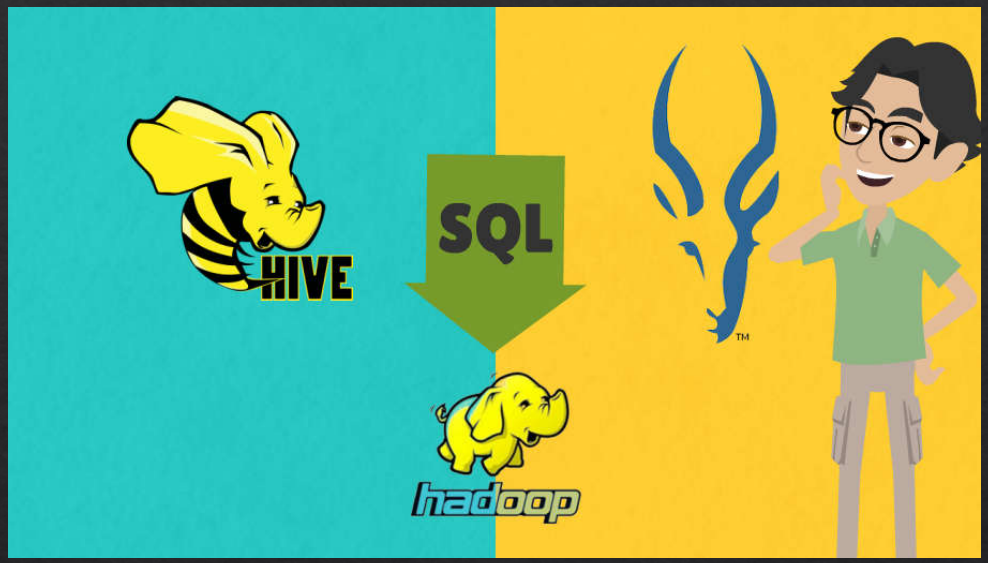

Big Data com Hive e Impala

Introdução

PROF. FERNANDO AMARAL
www.datascientist.com.br

Introdução

- Hive por padrão usa MapReduce
  - Processamento em batch
  - Alta latência

Introdução

- Não usa MapReduce
  - MPP (Massive Parallel Processing)
- Suporte a consultas ad hoc
- Acessa Hbase
- Desenvolvido em C++ (maioria Hadoop Java)
- Compatível com maioria dos formatos de dados do Hadoop
- Suporte a HDFS

Porém

- Não suporta Update
- Não tem tolerância a falhas
- Não suporta tipos complexos, com exeção de array a partir da versão CDH 5.5

Compatibilidade com Hive

- Compatível com HiveQL
- Acessa Hive Metastore (Obrigatório)
  - Acesso a bancos de dados e tabelas
- Mesmo Drive ODBC
- Mesmo ambiente de consulta (Hue)
- Importante: Alguns tipos não são suportados!

Diferenças com Hive

- Importantes
  - Tipo date não é suportado
  - Diversas funções de agregação
  - Alguns tipos de cast()

https://www.cloudera.com/documentation/enterprise/latest/topics/impala_langref_unsupported.html

Tipos de Dados

- Inteiro: INT, BIGINT, INT, SMALLINT, TINYINT
- BOOLEAN
- Texto: CHAR, STRING, VARCHAR
- DECIMAL, DOUBLE, FLOAT, REAL
- TIMESTAMP (Inclui dia, mês e ano)
- Complexos:
  - Array
  - Map
  - Struct