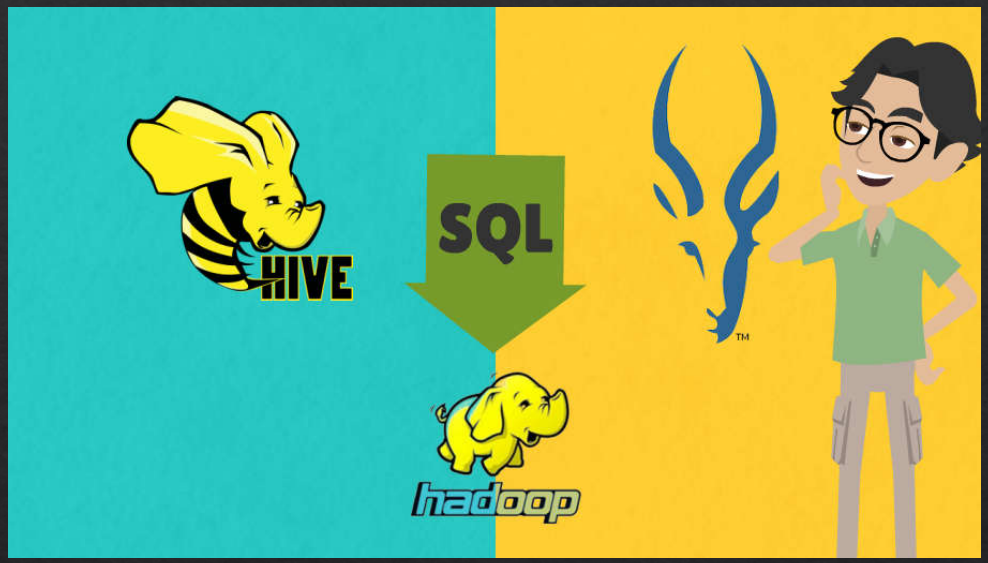


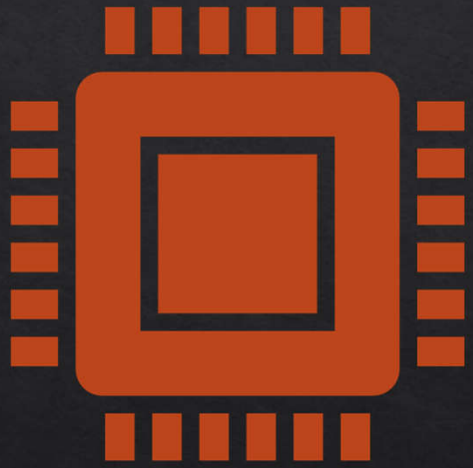
Big Data com Hive e Impala

Introdução

PROF. FERNANDO AMARAL
www.datascientist.com.br

Contexto

- Hadoop
  - Sistema em processamento de dados distribuído
  - Para processar grandes volumes de dados em batch
  - Dois elementos principais: Map Reduce e HDFS
    - Map Reduce: Processamento
    - HDFS: Sistemas de Arquivos Distribuídos



Analogia

- Imagine que você precisa construir uma casa
  - Um pedreiro comum levará 1 ano para construir uma casa
  - Você precisa construir a casa em menos tempo
  - Você contrata então um super pedreiro
  - Este super pedreiro levará 11 meses para construir a casa
  - Mas você precisa da casa em menos tempo ainda!
  - E não existem pedreiros mais rápidos!

HIVE

PROF. FERNANDO AMARAL
www.datascientist.com.br

Solução

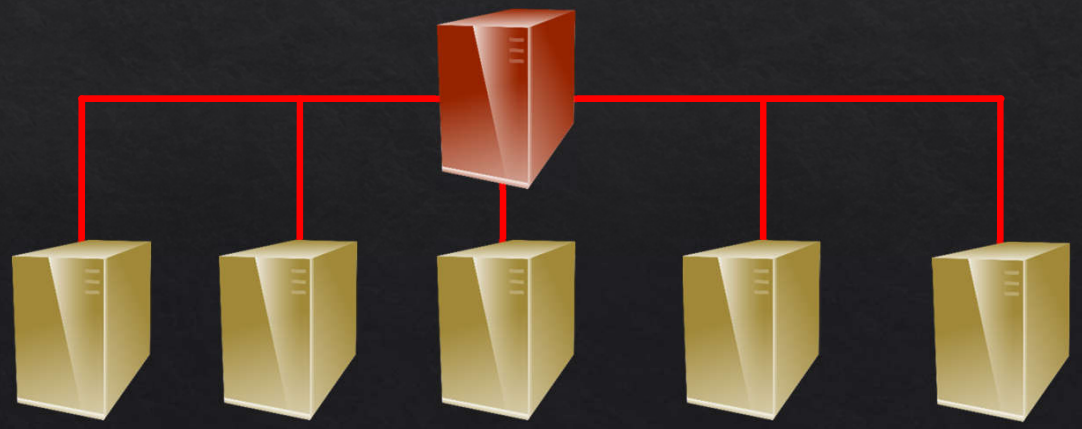
- Contratar 5 pedreiros
- Eles vão trabalhar com um objetivo comum, dividindo tarefas
- Trabalhando em “paralelo”
- Coordenados por um engenheiro
- Construção da casa levará 3 meses

Dados

- A poucas décadas, quando se precisava de mais capacidade de processamento, aumentava-se o servidor (pedreiro)
- Porém existem limites físicos para esse aumento (além de outras consequências)
- A solução é colocar vários servidores trabalhando em conjunto!

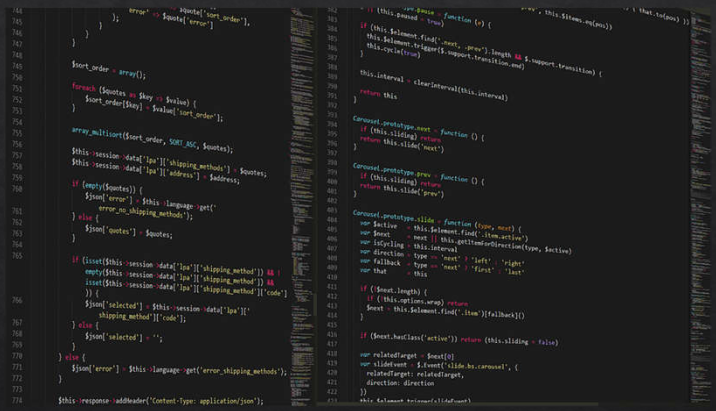
Hadoop

- Processar grandes volumes de dados
- Uma rede de computadores trabalhando em conjunto (Cluster)
- Conceito Master (Engenheiro) e Slaves (Pedreiros)



Hadoop

- Solução de Sucesso para o processamento de Grandes Volumes de Dados em Batch
- Porém...
- Hadoop foi criado em Java
- Para “programa-lo”, é preciso criar um programa em Java



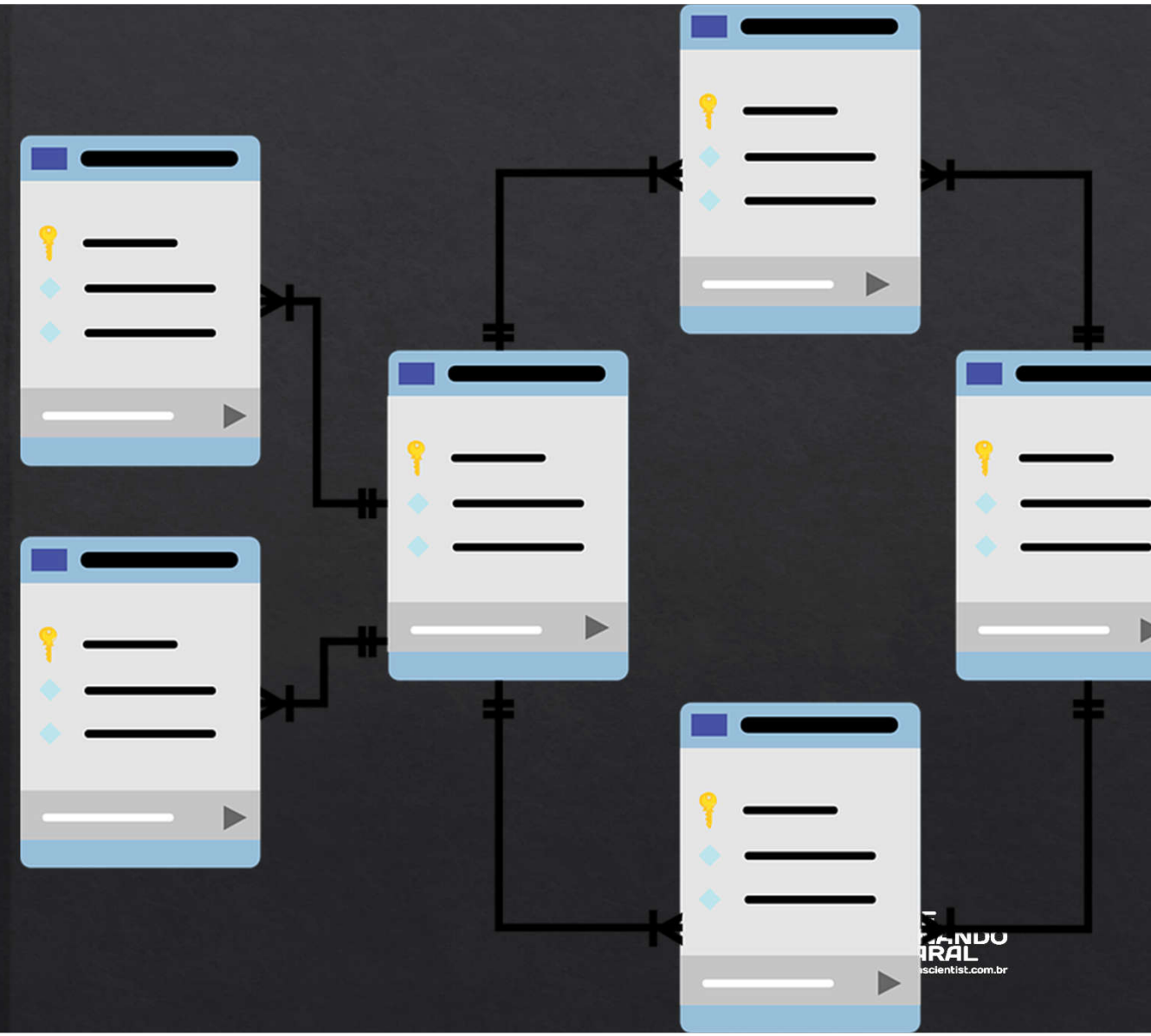
Imperativo VS Declarativo

Programação imperativa: especificação das etapas para atingir o objetivo. Como fazer. Ex: Java.

Programação declarativa: especificação dos objetivos. O que fazer. Ex: SQL

Hadoop

- O modelo de programação imperativo se apresentou um obstáculo, pois eram necessários especialistas em Java, escrita de programação, compilação, depuração testes etc.
- Facebook então resolveu criar o Hive:
  - Baseado na arquitetura Map Reduce / HDFS
  - Porém estruturados de forma semelhante a um banco de dados relacional
  - Utiliza linguagem de programação SQL: popular e fácil



Hive

- Armazém de dados distribuído sobre Map Reduce /HDFS
- Características de banco de dados relacionais: bancos de dados / tabelas / views / funções / SQL
- Não implementa funções não analíticas (ex: PK, FK)
- Open Source
- Mantido pela Fundação Apache
- Analítico, não transacional

351 systems in ranking, August 2019

Rank			DBMS	Database Model	Score		
Aug 2019	Jul 2019	Aug 2018			Aug 2019	Jul 2019	Aug 2018
1.	1.	1.	Oracle	Relational, Multi-model	1339.48	+18.22	+27.45
2.	2.	2.	MySQL	Relational, Multi-model	1253.68	+24.16	+46.87
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	1093.18	+2.35	+20.53
4.	4.	4.	PostgreSQL	Relational, Multi-model	481.33	-1.94	+63.83
5.	5.	5.	MongoDB	Document	404.57	-5.36	+53.59
6.	6.	6.	IBM Db2	Relational, Multi-model	172.95	-1.19	-8.89
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model	149.08	+0.27	+10.97
8.	8.	↓ 7.	Redis	Key-value, Multi-model	144.08	-0.18	+5.51
9.	9.	9.	Microsoft Access	Relational	135.33	-1.98	+6.24
10.	10.	10.	Cassandra	Wide column	125.21	-1.80	+5.63
11.	11.	11.	SQLite	Relational	122.72	-1.91	+8.99
12.	12.	↑ 13.	Splunk	Search engine	85.88	+0.39	+15.39
13.	13.	↑ 14.	MariaDB	Relational, Multi-model	84.95	+0.52	+16.66
14.	14.	↑ 18.	Hive	Relational	81.80	+0.93	+23.86
15.	15.	↓ 12.	Teradata	Relational, Multi-model	76.64	-1.18	-0.77
16.	16.	↓ 15.	Solr	Search engine	59.12	-0.52	-2.78
17.	17.	↑ 19.	FileMaker	Relational	58.02	+0.12	+1.96
18.	↑ 20.	↑ 21.	Amazon DynamoDB	Multi-model	56.57	+0.15	+4.91
19.	↓ 18.	↓ 17.	HBase	Wide column	56.54	-1.00	-2.27
20.	↓ 19.	↓ 16.	SAP Adaptive Server	Relational	55.86	-0.79	-4.57

Crescimento

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

Apache Impala

- Alternativa da Cloudera para "SQL em Hadoop"
- Objetivo é ser mais rápido que o Hive padrão (com engine MR)
- Impala não é baseado em MapReduce
- Tem menor latência e melhor desempenho em consultas médias
- SQL Baseado no Hive