Curso 3: Administração de Dados Complexos em Larga Escala

-- Apache Hive, mais detalhes --

Prof. Jose Fernando Rodrigues Junior

Objetivo: apresentar mais detalhes sobre a tecnologia Apache Hive

Quando usar MapReduce?

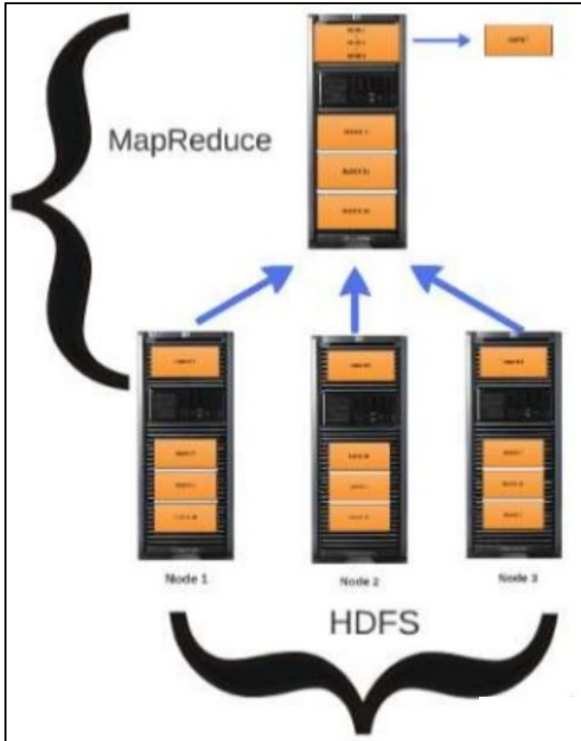
- MapReduce é adequado a problemas definidos como ***embarrassingly parallel*** (ou *perfectly parallel*);
- Problemas simples, **mas grandes**, que podem ser resolvidos mais rapidamente;
- O problema do **caixeiro viajante (grafo)**, por exemplo, não pode ser tratado com MapReduce;
- **Agregação e junção** de dados podem muito bem serem tratados com MapReduce.

Quando usar MapReduce?

Exercício *Hands on* – primeiro programa mapreduce (não distribuído) em python:

Write your first MapReduce program in 20 minutes
http://michaelnielsen.org/blog/write-your-first-mapreduce-program-in-20-minutes/

HDFS + MapReduce = Hadoop



HDFS - Reliable Shared Storage

+

MapReduce - Distributed Computation

=

hadoop

Um arcabouço distribuído para processamento e armazenamento de grandes bases de dados sobre *clusters* (conglomerados) de computadores convencionais (*commodity*).

HDFS + MapReduce = Hadoop

-O Hadoop é uma **instância** do modelo MapReduce;

-Como ele se **baseia no HDFS**, trata-se de uma instância **distribuída;**

-Como visto, o Hadoop acrescenta uma nova etapa, intrínseca ao processamento distribuído, o ***shuffle***, responsável pelo agrupamento do processamento distribuído no cluster;

-O HDFS permite a **distribuição** do processamento dentro do *cluster* de **maneira abstrata** (ou transparente);

-O Hadoop favorece o ***data locality***, isto é, o processamento deve ocorrer, sempre que possível, nos próprios nós de armazenamento de dados.

HDFS + MapReduce = Hadoop

-O H

-Com ia
distr

-Co ao
proce o
proce

-O H do
clust

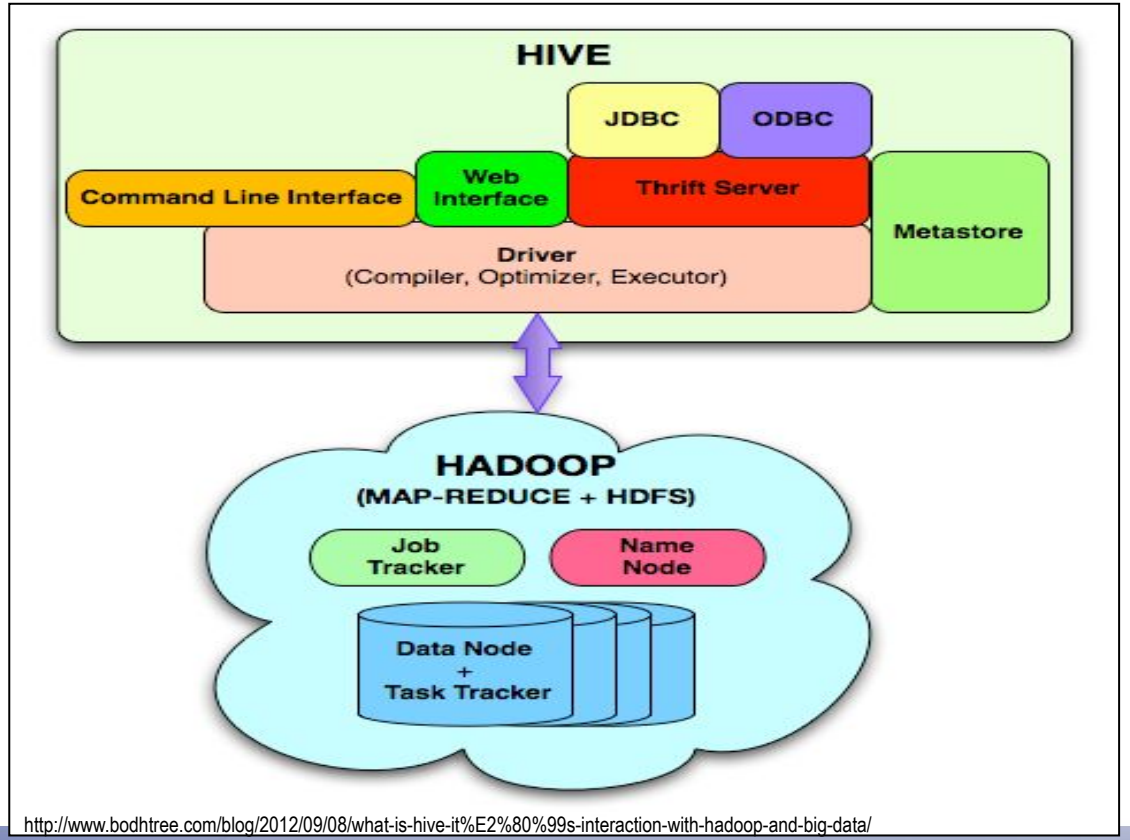
⇒ Para saber mais: Apache Spark, Kubernetes, serverless applications, etc

-O Hadoop favorece o ***data locality***, isto é, o processamento deve ocorrer, sempre que possível, nos próprios nós de armazenamento de dados.

Hive - Big Data Warehouse

- **Originalmente** funcionava com uma **API Java** para executar processamento map-reduce;
- Posteriormente, ganhou uma camada que o tornou orientado a **tabelas e a SQL** (HiveQL, na verdade);
- **WORM**: **w**rite-**o**nce, **r**ead **m**any times; em contraste a SGBDs, os quais são read and write many times ⇒ **Hive não é uma base operacional**;
- **Várias distribuições**: Apache, Cloudera Hortonworks, MapR, MS Azure HDInsight(cloud), entre outras - os vendors tornam o uso mais simples via interface, distribuição, documentação, suporte, etc; ⇒ Comparativo
- Permite o armazenamento, a consulta, e a análise de grandes bases de dados armazenadas em **HDFS**; bases na escala de **Peta** bytes são suportadas.

Visão geral da arquitetura Apache Hive



http://www.bodhtree.com/blog/2012/09/08/what-is-hive-it%E2%80%99s-interaction-with-hadoop-and-big-data/

Hive - Big Data Warehouse

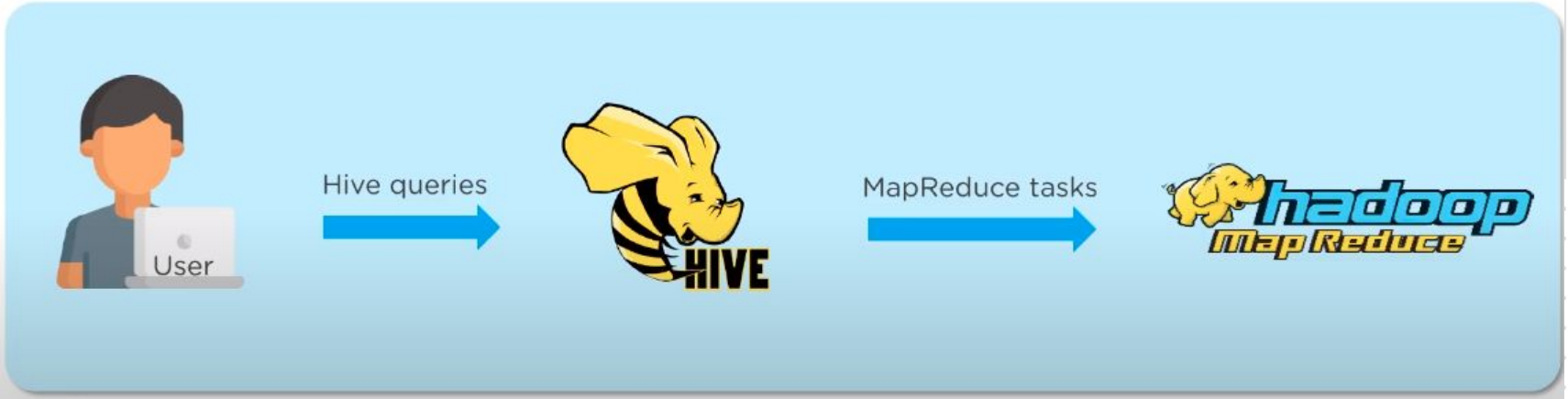
Onde seus problemas se encaixam?

Size	Classification	Tools
Lines *Sample Data*	Analysis and Visualization	Whiteboard, `bash`,...
KBs - low MBs *Prototype Data*	Analysis and Visualization	Matlab, Octave, R, Processing, `bash`,...
MBs - low GBs *Online Data*	Storage	MySQL (DBs),...
	Analysis	NumPy, SciPy, Weka, BLAS/ LAPACK,...
	Visualization	Flare, AmCharts, Raphael, Protovis,...
GBs - TBs - PBs *Big Data*	Storage	HDFS, HBase, Cassandra,...
	Analysis	Hive, Mahout, Hama, Giraph,..

https://www.slideshare.net/Cataldo/apache-mahout-tutorial-recommendation-20132014?next_slideshow=1

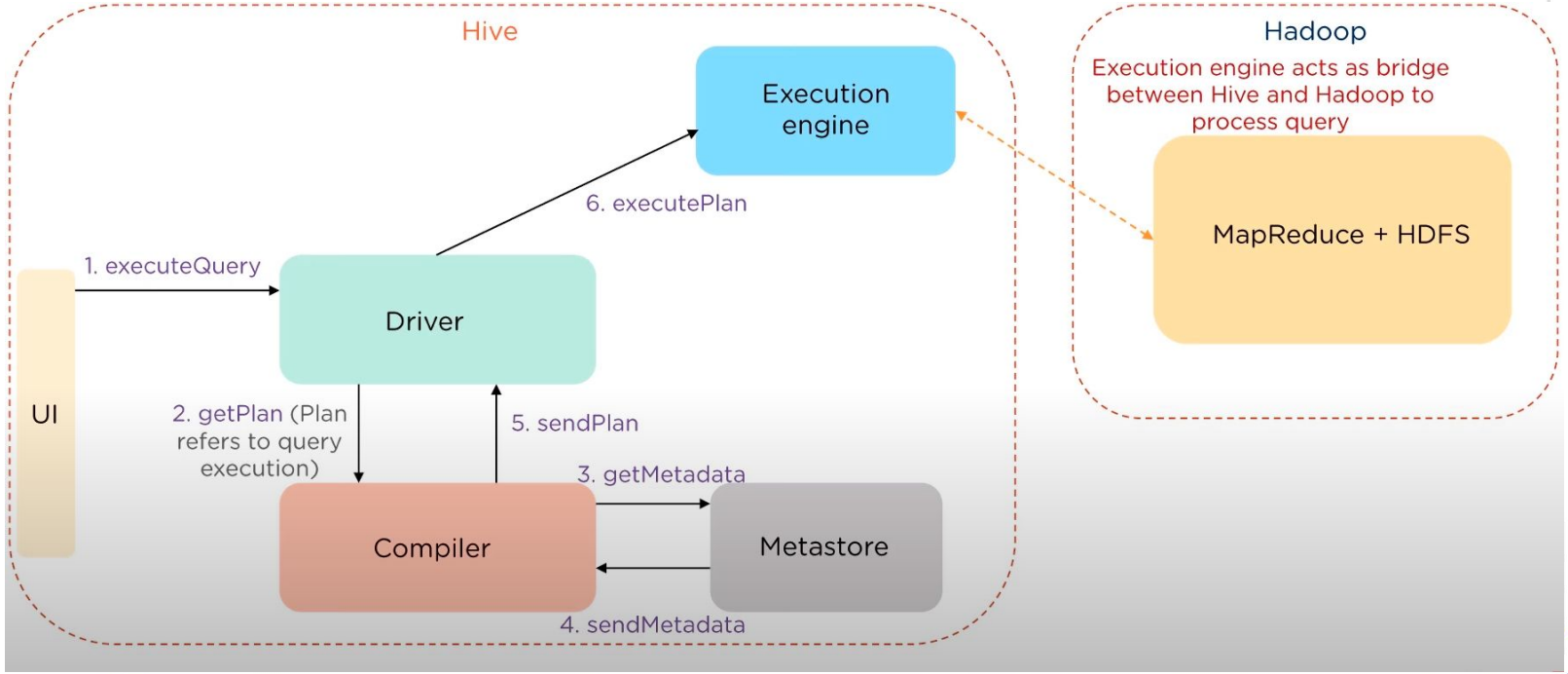
Hive - Big Data Warehouse

- Hive: processamento DW sobre Apache Hadoop



Hive - Big Data Warehouse

- Hive: processamento DW sobre Apache Hadoop



Hive - Big Data Warehouse

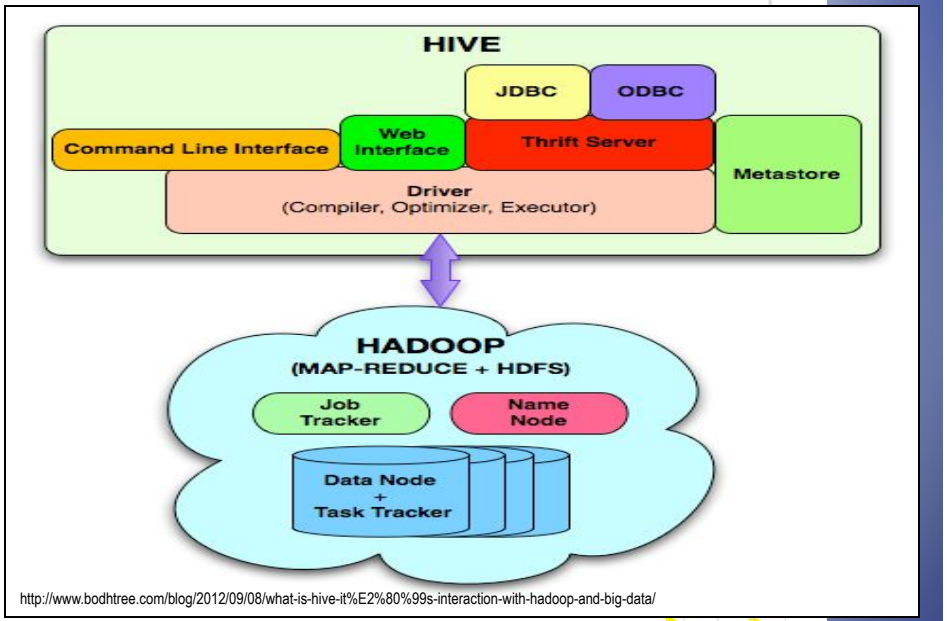
- Em resumo, o Hive é usado da mesma maneira que um **terminal SQL**;
- Diferente de um SGBD, o Hive executa as consultas **SQL de modo distribuído** sobre o Apache Hadoop;
- O Hive também funciona sobre a tecnologia Spark, uma evolução do Hadoop com melhor desempenho
  ⇒ documentação Spark sobre o Hive.

Hive - Big Data Warehouse

- DW e OLAP se baseiam em consultas **a dados estruturados, e operações de agregação**;
- Desta maneira, como o MapReduce suporta processamento SQL, então **é possível processar DW e OLAP sobre MapReduce**;
- Isso é possível via **Hive** e, de maneira mais eficiente, sobre arcabouços como **Apache Kylin**, Druid, Kyvos, e Apache Lens, denominadas ***Distributed Analytics Engines (DAEs).***

Hive - Big Data Warehouse

-Em suma: é possível fazer DW/OLAP sobre as tecnologias HIVE + HADOOP;



http://www.bodhtree.com/blog/2012/09/08/what-is-hive-it%E2%80%99s-interaction-with-hadoop-and-big-data/

Hive - Big Data Warehouse

-Em suma: é possível fazer DW/OLAP sobre as tecnologias HIVE + HADOOP;

Exercício *Hands on* - criação de um Data Warehouse em Hive passo a passo:

Cloudera - Getting Started with Hortonworks Data Platform Sandbox
https://www.cloudera.com/tutorials/getting-started-with-hdp-sandbox/3.html