Ambiente Livre
Tecnologia

Hadoop

Treinamento Hadoop – Big Data Open Source - Fundamental.

Instrutor: Marcio Junior Vieira.
marcio@ambientelivre.com.br

O que é Hive

- SQL-like query language and metastore
- Um armazém de dados (datawarehouse) distribuídos. Gerencia os dados armazenados no HDFS e fornece uma linguagem de consulta baseada em SQL para consultar os dados.
- Criado pelo Facebook e hoje mantido pela comunidade Apache.

Características

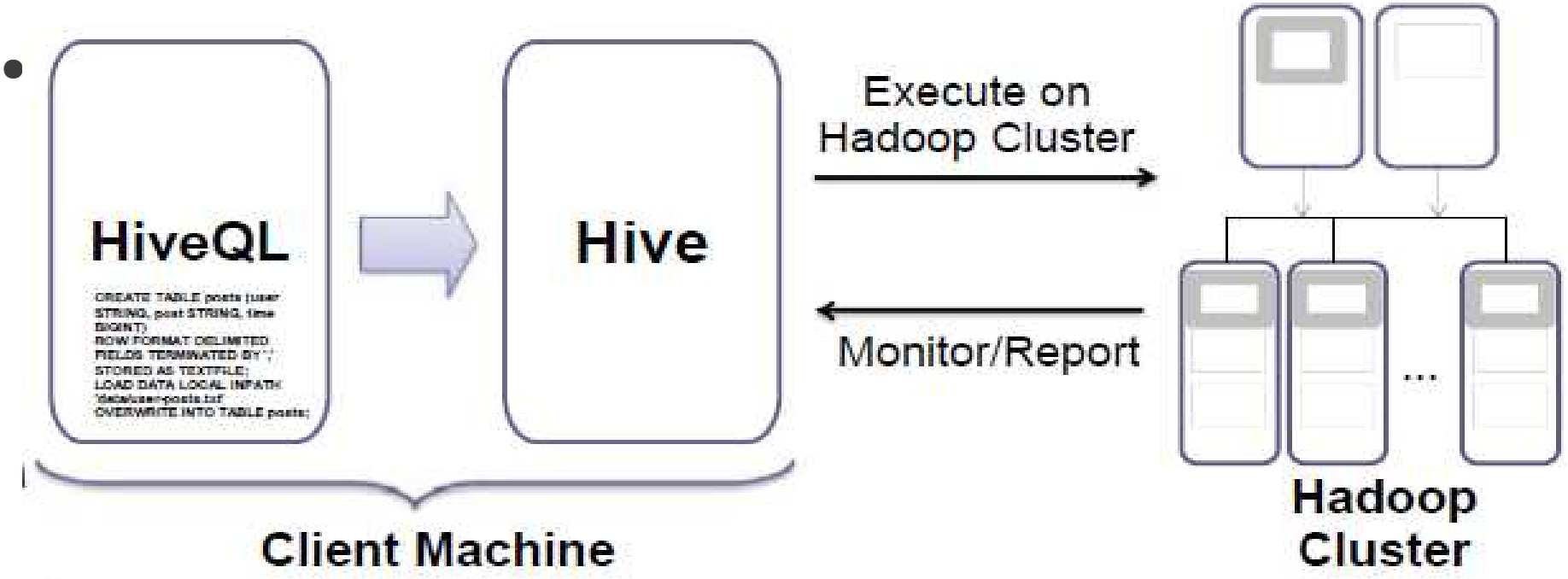
- Habilidade para trazer estrutura para vários formatos de dados
- Interface simples para consultas ad hoc, analisando e sintetizando grandes quantidades de dados
- O acesso aos arquivos em vários armazenamentos de dados, como HDFS e HBase

Características

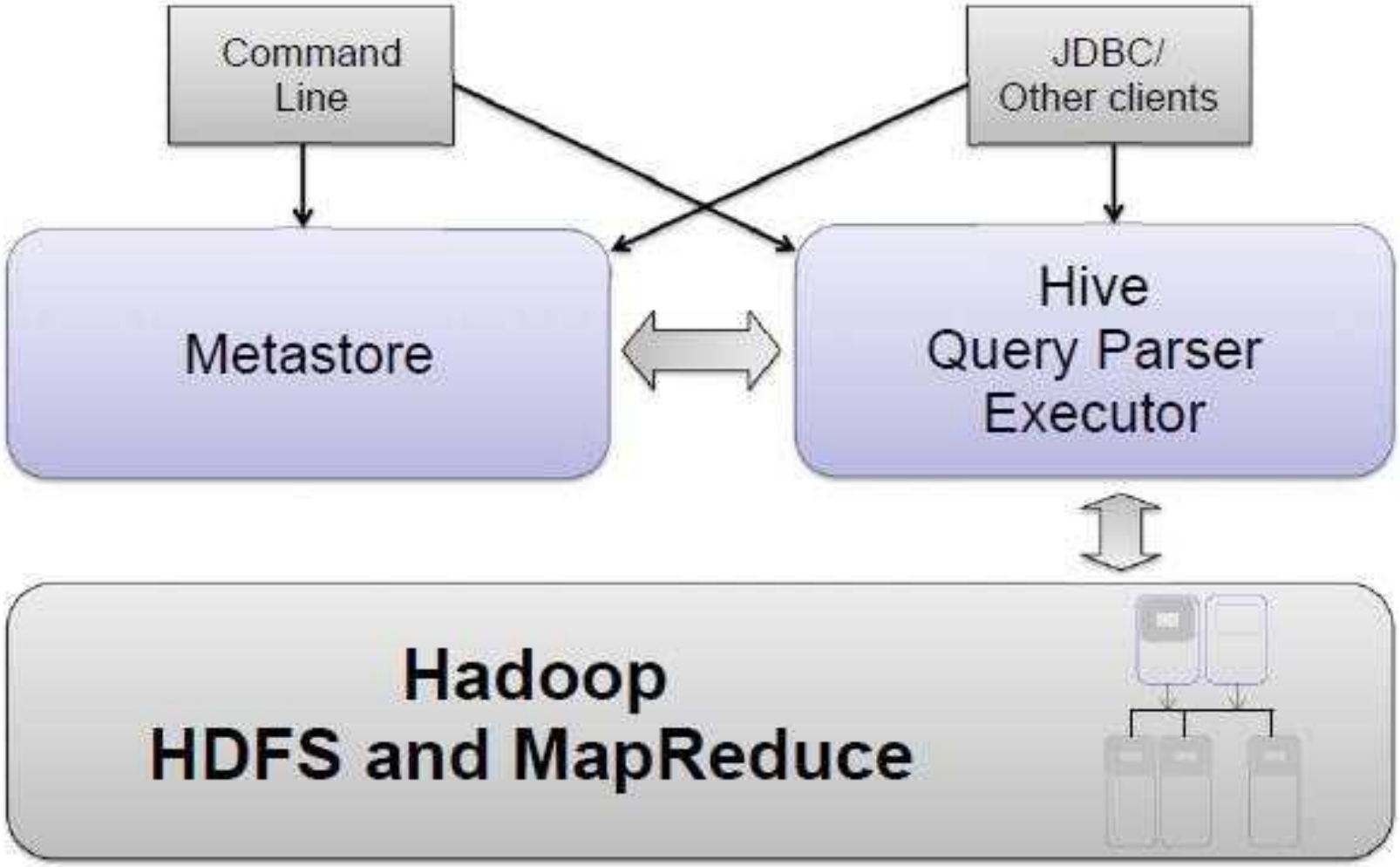
- NÃO fornece baixa latência ou consultas em tempo real
- Mesmo consultando pequenas quantidades de dados pode demorar alguns minutos
- Projetado para escalabilidade e facilidade de uso em vez de respostas de baixa latência
-

Hive

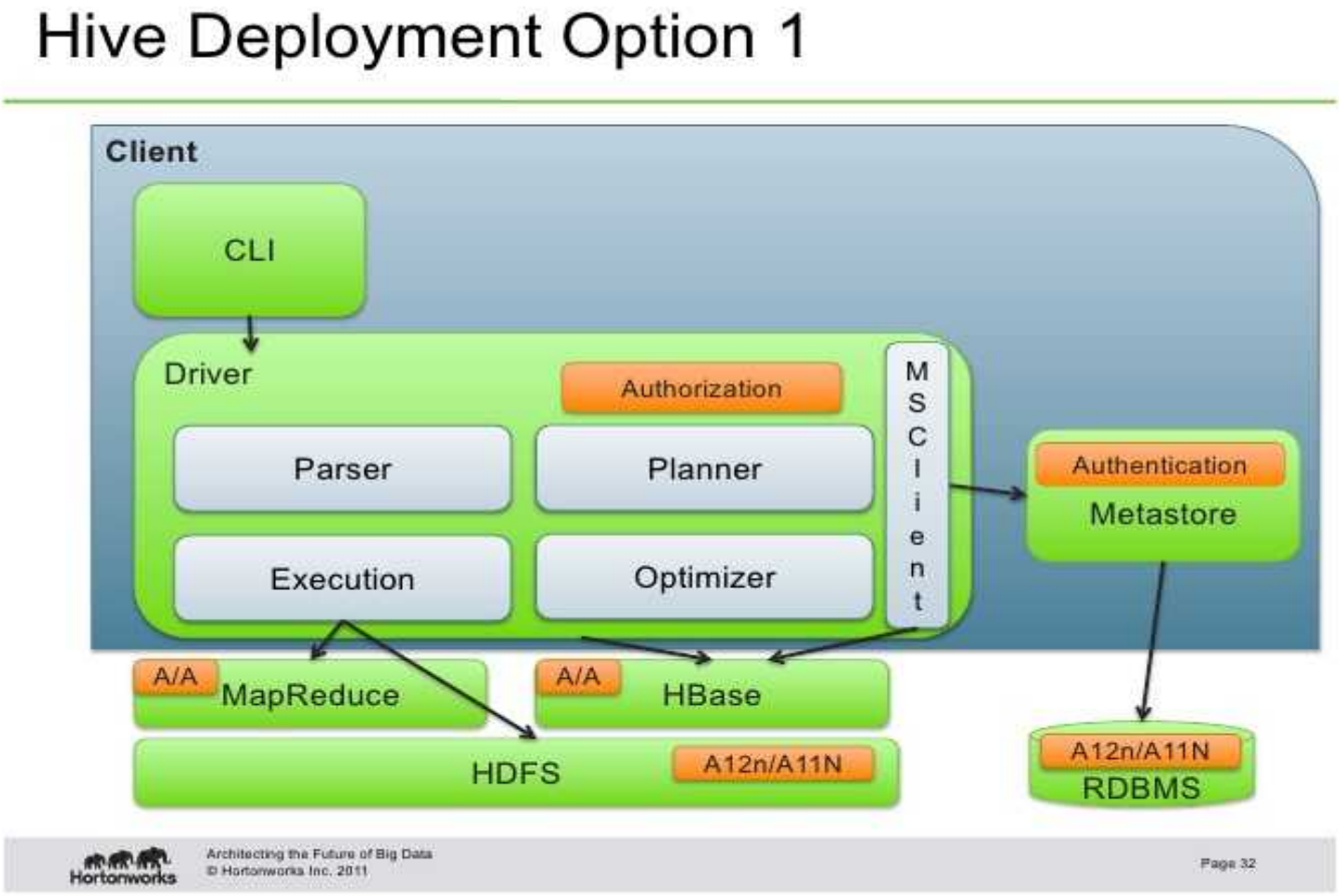
- Traduz declarações HiveQL em um conjunto de MapReduce Jobs que são executadas em um Hadoop Cluster
-



Arquitetura Hive



Hive Integrado ao HBase



Interfaces

- **Command Line Interface (CLI)**
- **Hive Web Interface**
  cwiki.apache.org/confluence/display/Hive/HiveWebInterface
- **Java Database Connectivity (JDBC)**
  cwiki.apache.org/confluence/display/Hive/HiveClient

```
hive> select count (1) from posts;
Total MapReduce jobs = 1
Launching Job 1 out of 1
...
Starting Job = job_1343957512459_0004, Tracking URL =
http://localhost:8088/proxy/application_1343957512459_0004/
Kill Command = hadoop job -Dmapred.job.tracker=localhost:10040 -kill
job_1343957512459_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2012-08-02 22:37:24,962 Stage-1 map = 0%, reduce = 0%
2012-08-02 22:37:30,497 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.87 sec
2012-08-02 22:37:31,577 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.87 sec
2012-08-02 22:37:32,664 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.64 sec
MapReduce Total cumulative CPU time: 2 seconds 640 msec
Ended Job = job_1343957512459_0004
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 1 Accumulative CPU: 2.64 sec HDFS Read: 0 HDFS Write: 0
SUCESS
Total MapReduce CPU Time Spent: 2 seconds 640 msec
OK
4
Time taken: 14.204 seconds
```