



Engenharia de Dados com Hadoop e Spark



Bem-vindo(a)





Armazenamento de Dados com HBase e Hive

A faint, stylized diagram in the background illustrates a data storage architecture. It features several circular nodes connected by lines. The nodes are colored in shades of blue, green, yellow, and orange. The connections form a complex network, with some nodes acting as central hubs and others as peripheral components, representing the relationships between different data storage and processing elements in a system like HBase and Hive.

Armazenamento de Dados com HBase e Hive



Data Science Academy rodrigo.cabreu@hotmail.com 5e207d48e32fc335fa60447d

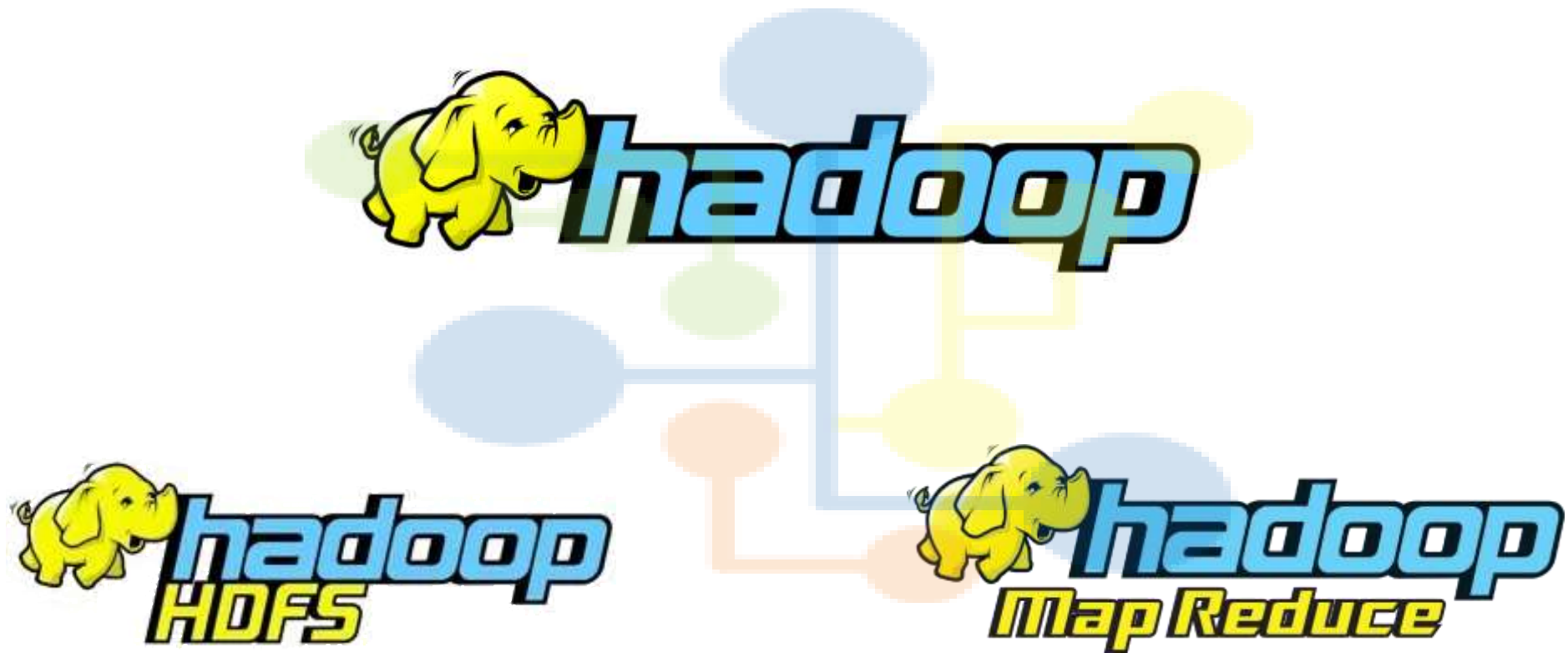
Como está sendo sua experiência com o Hadoop? Tudo bem até aqui? Como você deve ter percebido, trabalhar com o Hadoop não é uma tarefa das mais fáceis....requer prática e muito trabalho manual.

Conhecimento de Sistema Operacional Linux é com certeza um diferencial e esperamos que você esteja aproveitando esta oportunidade para aprender um pouco mais sobre Linux também.

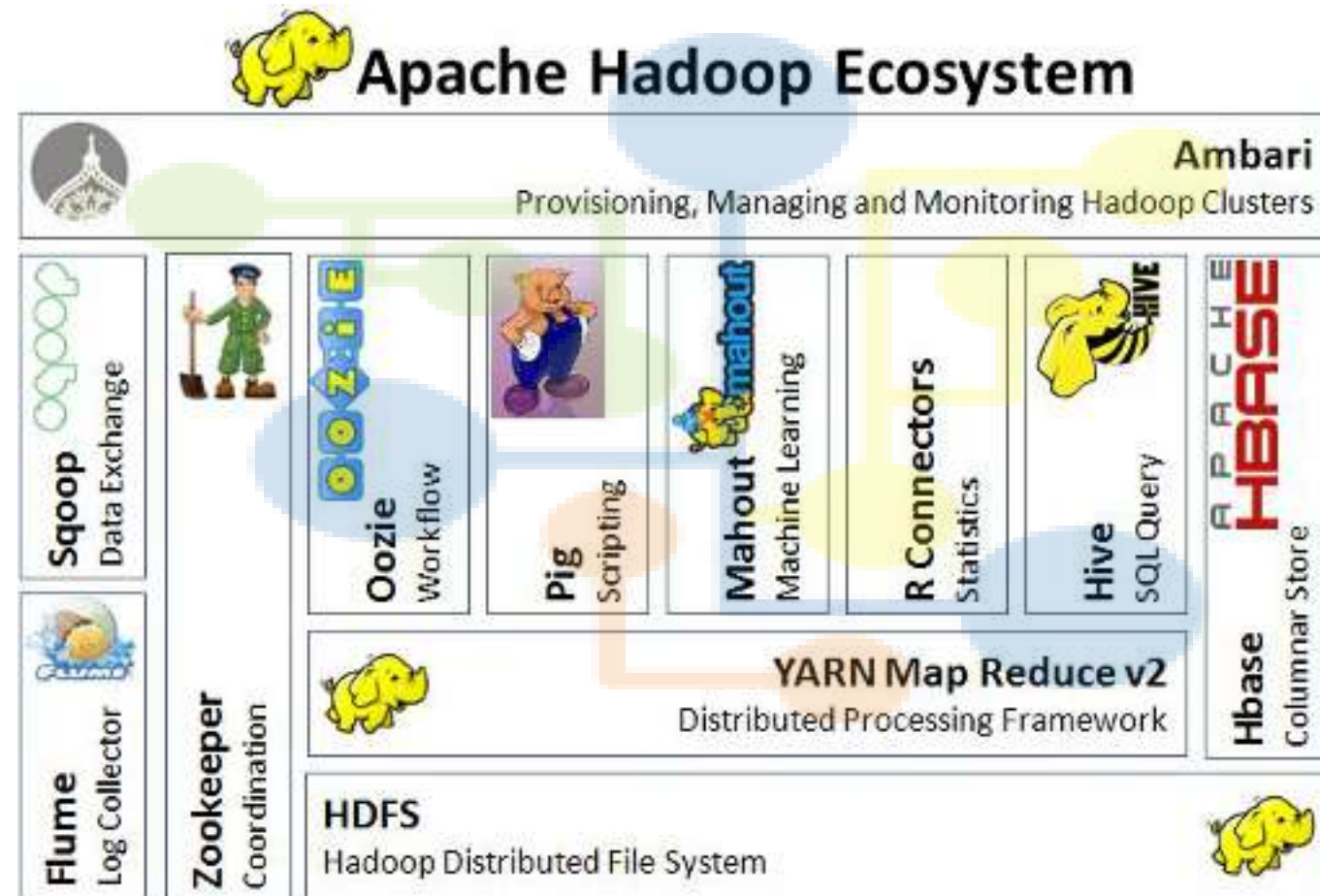
Se tiver qualquer dúvida, utilize o fórum do curso ou os canais de suporte.

Armazenamento de Dados com HBase e Hive

Data Science Academy
Data Science Academy rodrigo.cabreu@hotmail.com 5e207d48e32fc335fa60447d



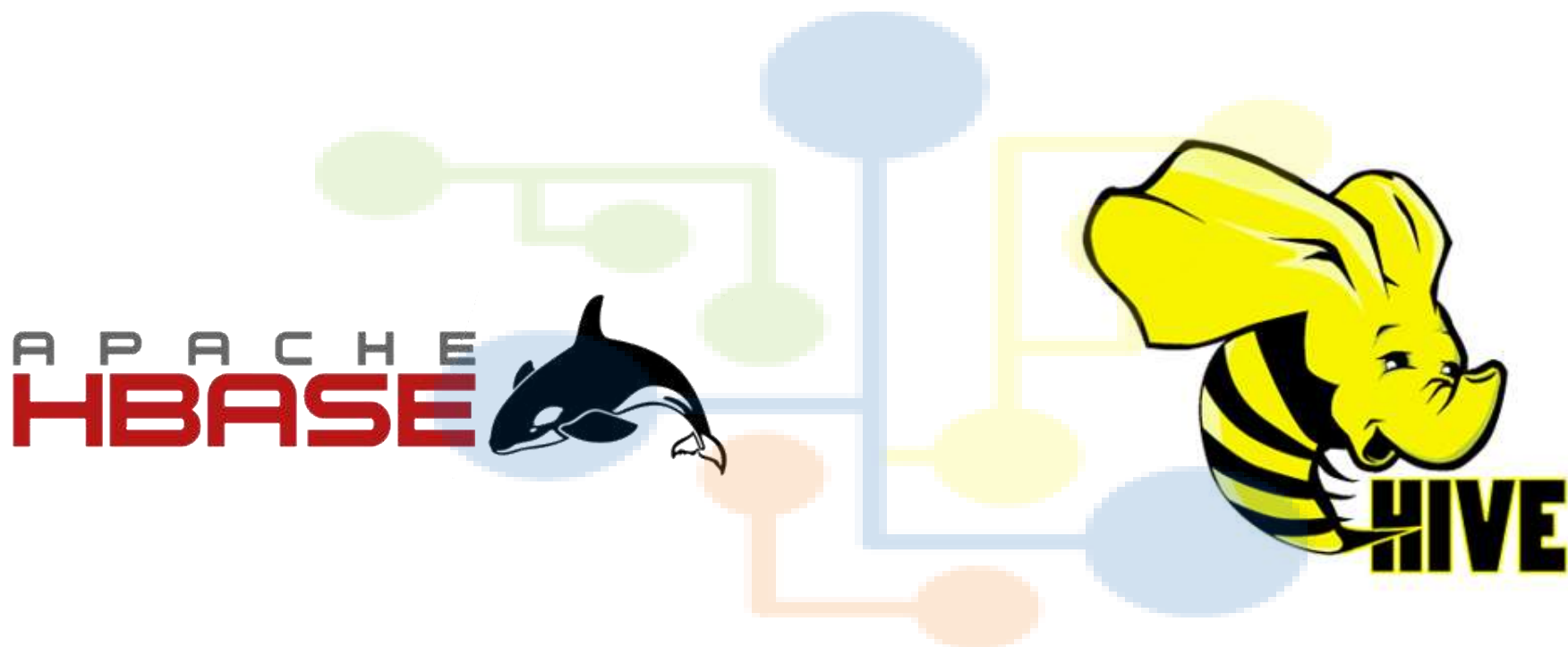
Armazenamento de Dados com HBase e Hive



Armazenamento de Dados com HBase e Hive

Data Science
Academy

Data Science Academy rodrigo.cabreu@hotmail.com 5e207d48e32fc335fa60447d



Armazenamento de Dados com HBase e Hive

Data Science
Academy

Data Science Academy rodrigo.cabreu@hotmail.com 5e207d48e32fc335fa60447d

The Cloudera logo is displayed in a dark blue, sans-serif font. It is positioned on the left side of the slide, partially overlapping a faint, stylized network diagram in the background. The diagram consists of various colored circles (blue, green, yellow, orange) connected by thin lines, representing a distributed system architecture.

cloudera

The Hortonworks logo is located on the right side of the slide. It features three green silhouettes of elephants walking to the right, positioned above the word "Hortonworks" in a bold, black, sans-serif font. The logo is set against a background of the same faint network diagram seen behind the Cloudera logo.

Hortonworks



Apache HBase

A faint, stylized diagram in the background representing a distributed system architecture. It features several nodes (circles) in blue, green, yellow, and orange, connected by lines. A central blue node is connected to other nodes, including a green node on the left, a yellow node on the right, and an orange node at the bottom. The text 'Apache HBase' is overlaid on this diagram.

Apache HBase



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d



Apache HBase



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

APACHE
HBASE

mongoDB



redis

Apache HBase



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

The diagram illustrates the structure of an Apache HBase table. It features a table with a 'Row Key' column and two column families: 'Students' and 'Branch'. The 'Students' family contains 'Name' and 'Age' columns, while the 'Branch' family contains 'Bname' and 'GPA' columns. Annotations include: a green box around the 'Row Key' column labeled 'Row Key'; yellow boxes around the 'Students' and 'Branch' column families labeled 'Column Families'; a green box around the 'GPA' column labeled 'Column'; a red box around the row for 'Sham' labeled 'Cells'; and an orange arrow pointing to the 'GPA' column with the text 'Representa os dados de uma coluna!'.

Row Key	Students		Branch	
StudentID	Name	Age	Bname	GPA
100	Ram	18	CSE	7.9
101	Sham	17	ECE	8
102	John	18	EEE	7.5
103	Sam	17	CSE	8.5

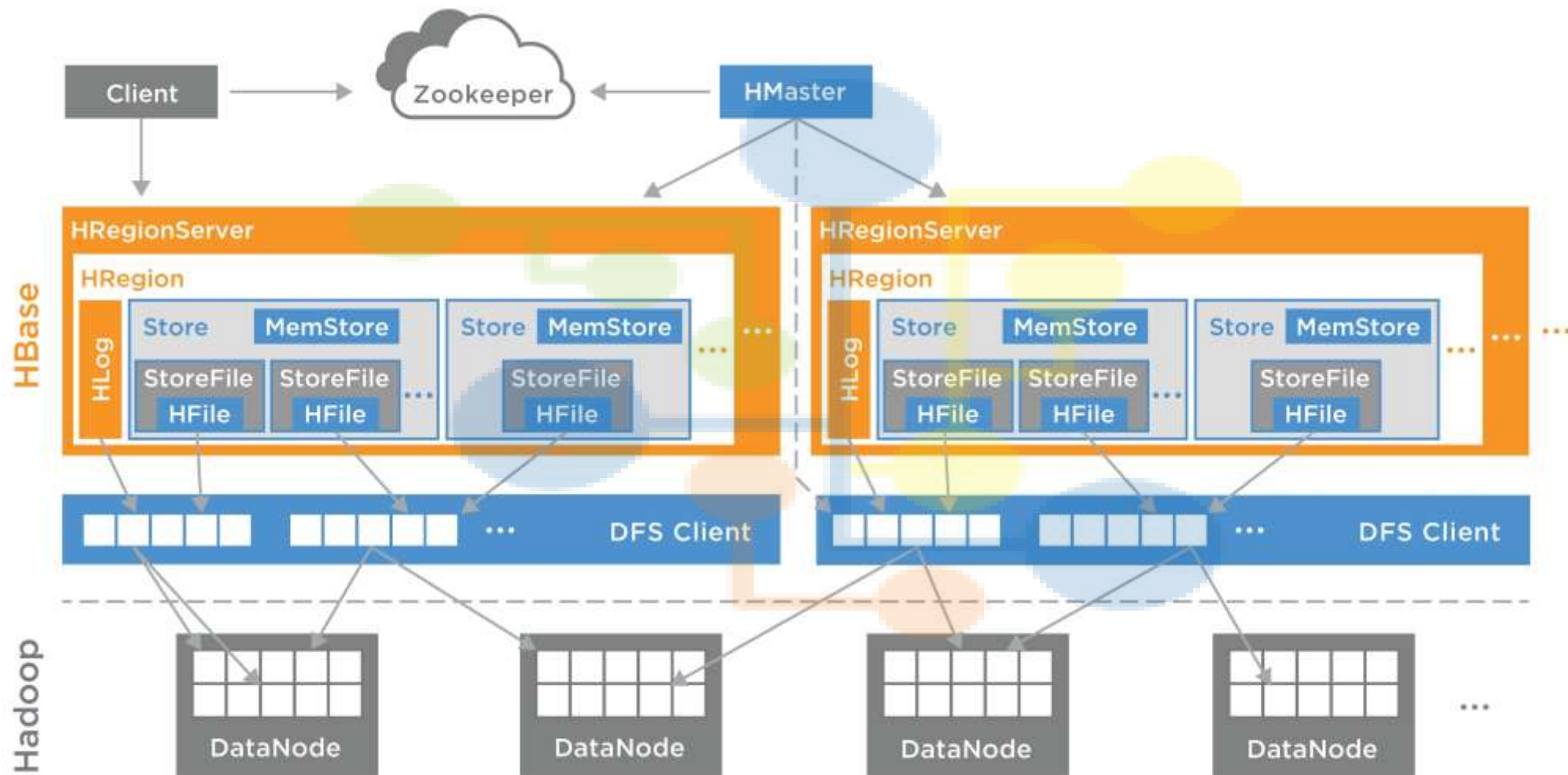
Representa os
dados de uma
coluna!

Apache HBase



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d





Apache Hive

A faint, stylized diagram of a network structure is visible in the background. It consists of several circular nodes connected by lines. The nodes are colored in shades of blue, green, yellow, and orange, matching the Data Science Academy logo. The lines are also colored to match the nodes they connect, creating a complex web-like pattern.

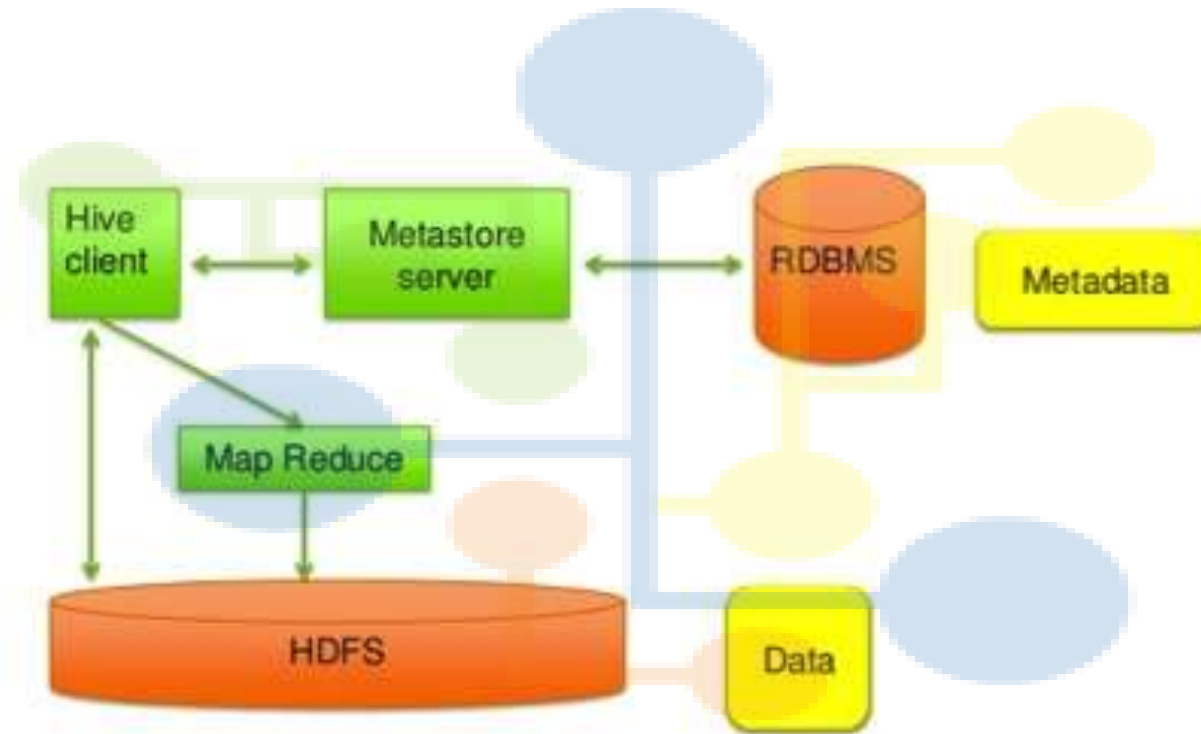


Apache Hive



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

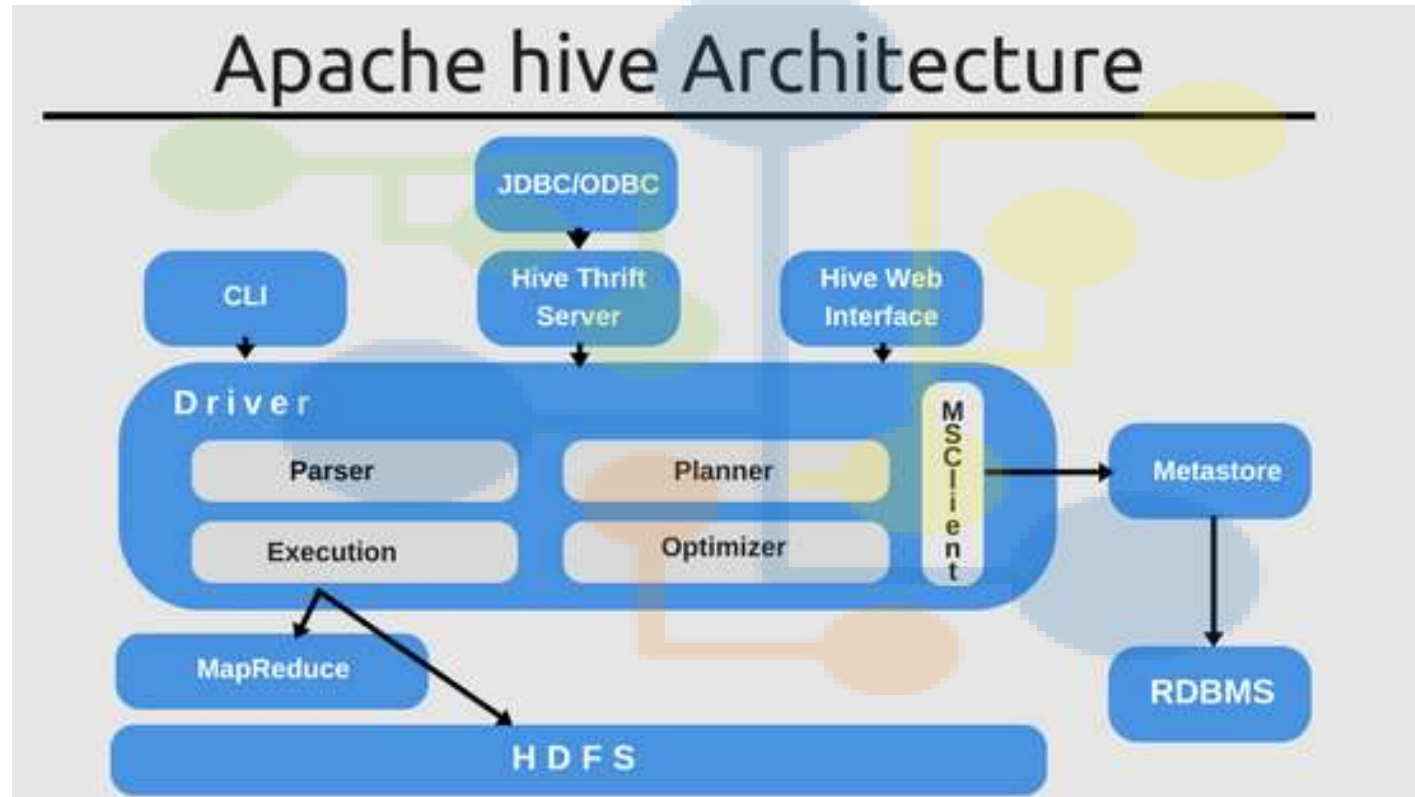


Apache Hive



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

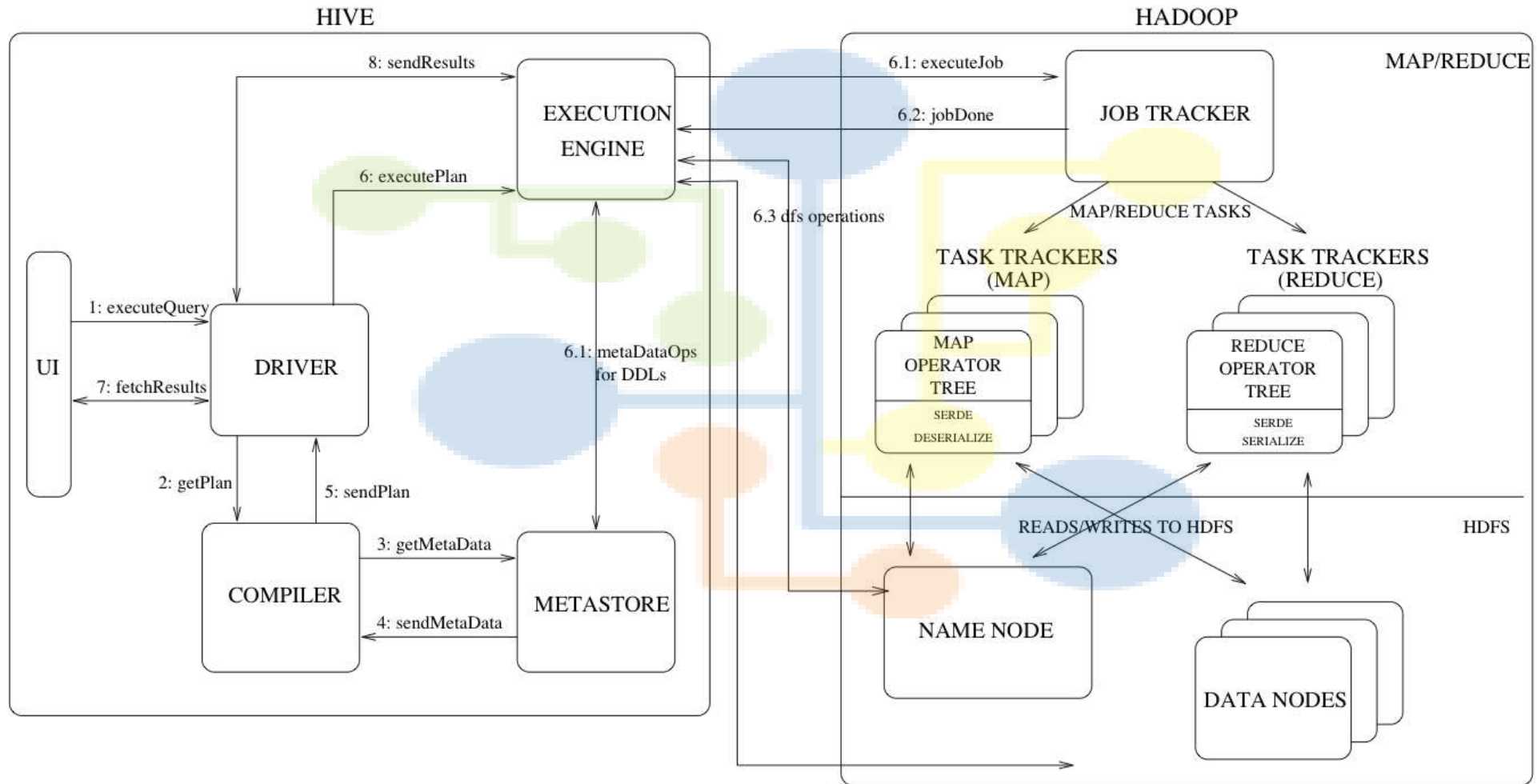


Apache Hive



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d





Conhecendo o Apache HBase

A faint, stylized diagram in the background, consisting of several interconnected nodes (circles) in blue, green, yellow, and orange, connected by lines, suggesting a network or data flow.



Conhecendo o Apache HBase



Hbase é mais um "DataStore" do que um "Database".



Conhecendo o Apache HBase



Hbase é um banco de dados distribuído, open-source, não-relacional, inspirado no Google Big Table.



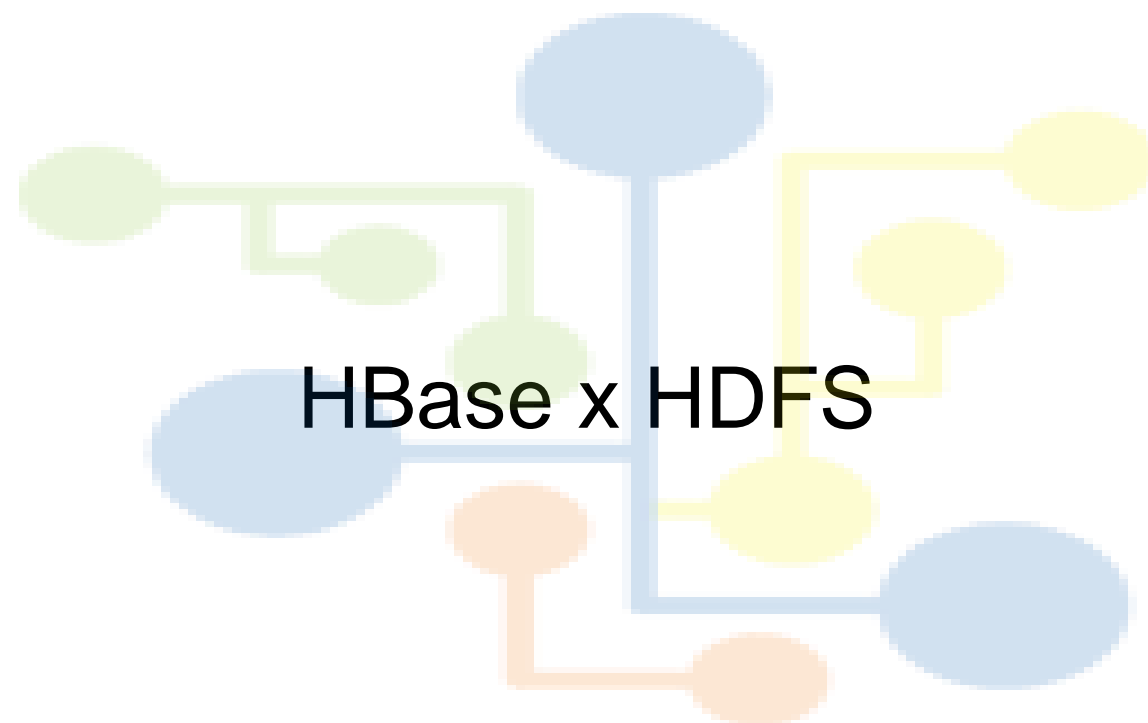
Conhecendo o Apache HBase

Principais características do HBase:

- Escalabilidade horizontal
- Processos consistentes de leitura/escrita (Hbase Read/Hbase Write)
- Particionamento automático
- Recuperação automática de falhas
- Java API para acesso aos dados



Conhecendo o Apache HBase





Conhecendo o Apache HBase

HBase	HDFS
HBase é um banco de dados NoSQL construído para trabalhar sobre o HDFS.	Sistema de arquivos distribuído para armazenamento de grandes conjuntos de dados.
Suporta consultas a grandes tabelas de dados.	Não suporta consultas a registros individuais de dados.
Baixa latência de acesso aos dados, mesmo em tabelas de bilhões de registros.	Alta latência e processamento em batch.
HBase armazena dados em formato key/value.	Armazena os dados em arquivos.



Conhecendo o Apache HBase



HBase x RDBMS



Conhecendo o Apache HBase

HBase	RDBMS
Utiliza regiões.	Utiliza tabelas.
Suporta o filesystem HDFS.	Suporta filesystems FAT , NTFS, EXT, NFS.
Conceito de Write-Ahead Logs (WAL) para armazenar alterações nos dados.	Conceito de commit logs para armazenar as alterações nos dados.
A coordenação dos processos é feita pelo Apache Zookeeper.	A coordenação dos processos é feita pelo sistema gerenciador de bancos de dados (Oracle, SQL Server, MySQL, etc...)
Linhas são identificadas unicamente pelas rowkeys .	Linhas são identificadas unicamente por chaves primárias.
Regiões podem ser particionadas.	Tabelas podem ser particionadas.
Conceito de linha, família de colunas , coluna e célula.	Conceito de linha, coluna e célula.
Suporta bilhões de registros.	Apresenta problemas de performance com bilhões de registros.



Conhecendo o Apache HBase

Afinal, Quando Usar e Quando Não Usar o HBase?



Conhecendo o Apache HBase

Quando Utilizar HBase?

Dados não-estruturados ou semi-estruturados

Alta escalabilidade

Dados versionados

Quando é necessário acesso baseado em chave

Alto volume de dados devem ser armazenados

Armazenamento de dados orientado a coluna



Conhecendo o Apache HBase

Poucas linhas devem ser armazenadas

Não for necessário realizar consultas cruzadas (SQL Joins)

Cluster com poucas máquinas

Quando **Não**
Utilizar HBase?



Obrigado

