Engenharia de Dados com Hadoop e Spark

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

Bem-vindo(a)

Armazenamento de Dados com HBase e Hive

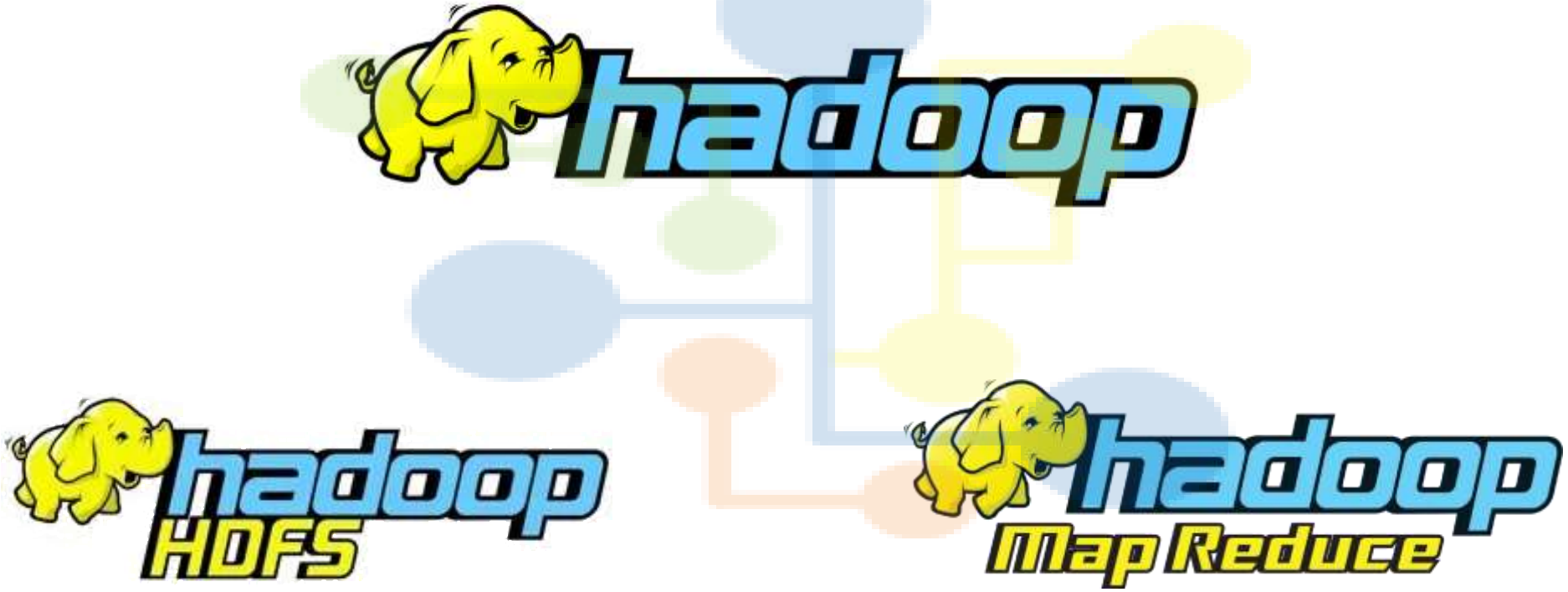
Armazenamento de Dados com HBase e Hive

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.

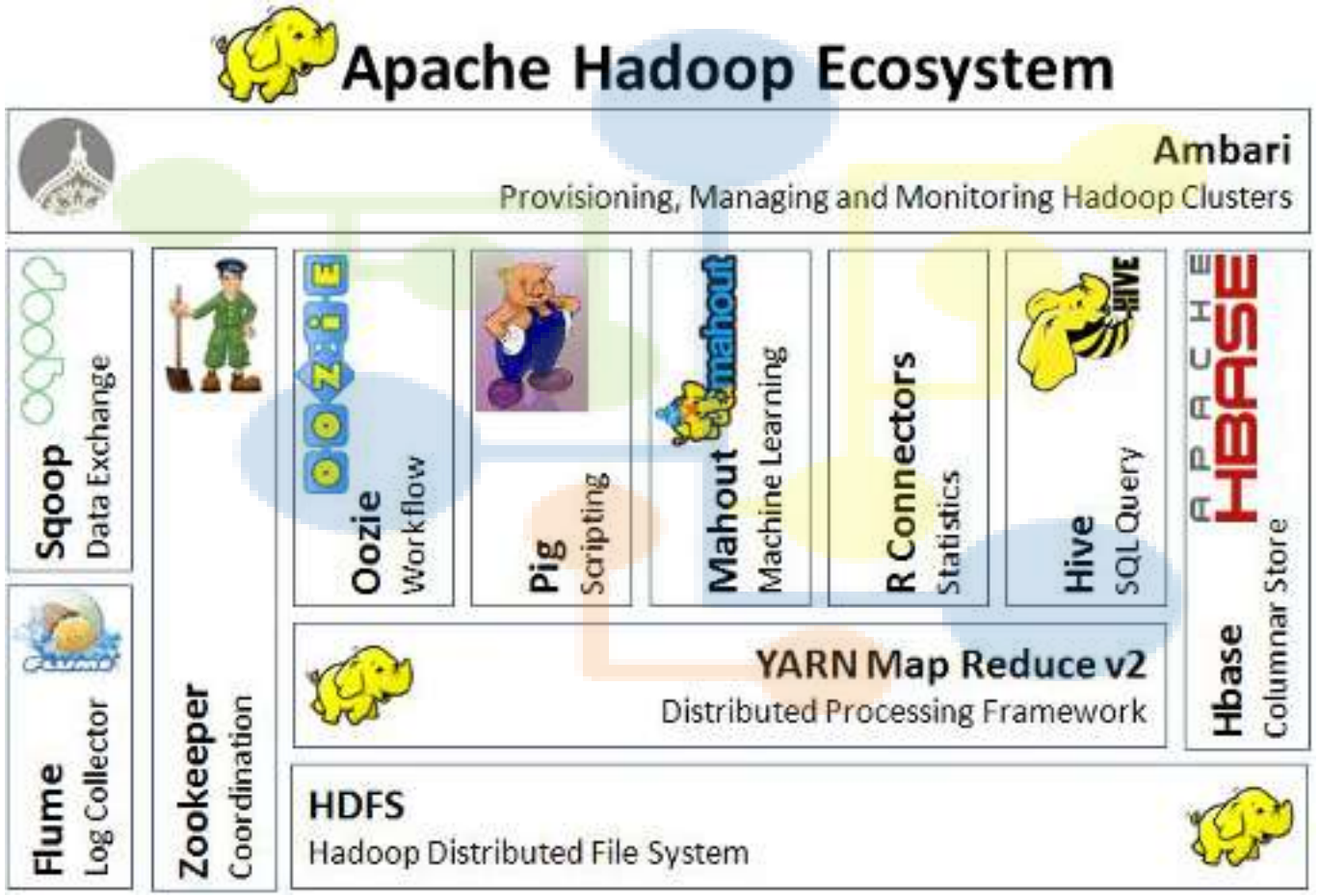
Conhecimento de Sistema Operacional Linux é com certeza um diferencial e esperamos que você esteja aproveitando esta oportunidade para aprender um pouco mais sobre Linux também.

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

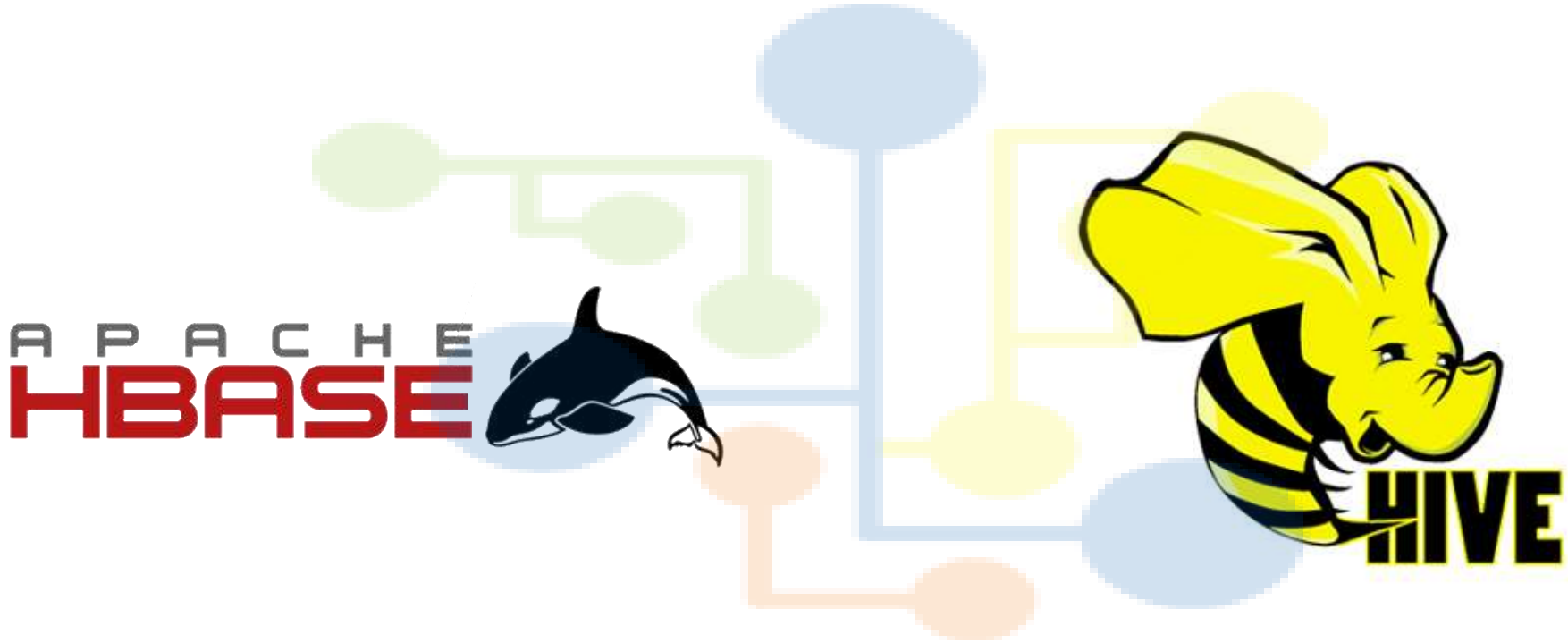
Armazenamento de Dados com HBase e Hive



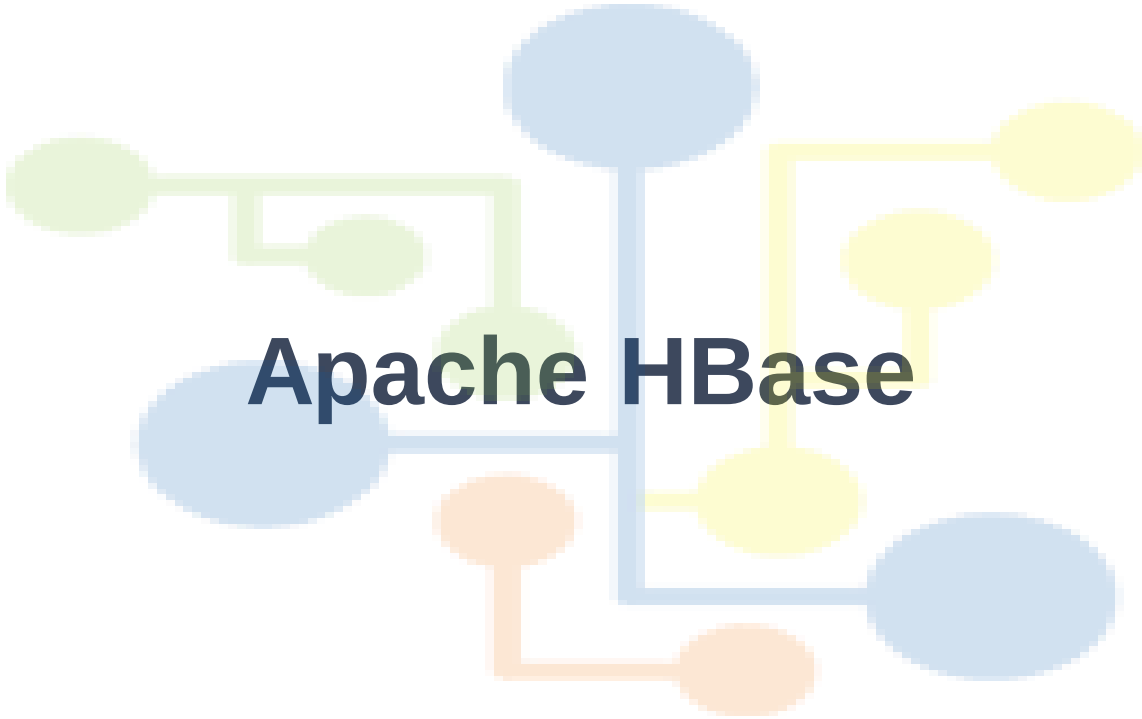
Armazenamento de Dados com HBase e Hive



Armazenamento de Dados com HBase e Hive



Armazenamento de Dados com HBase e Hive



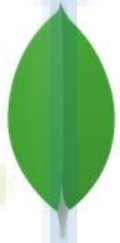
Apache HBase

Apache HBase

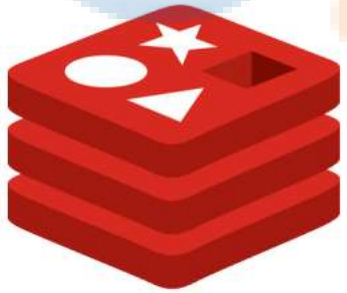
Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

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

Apache HBase



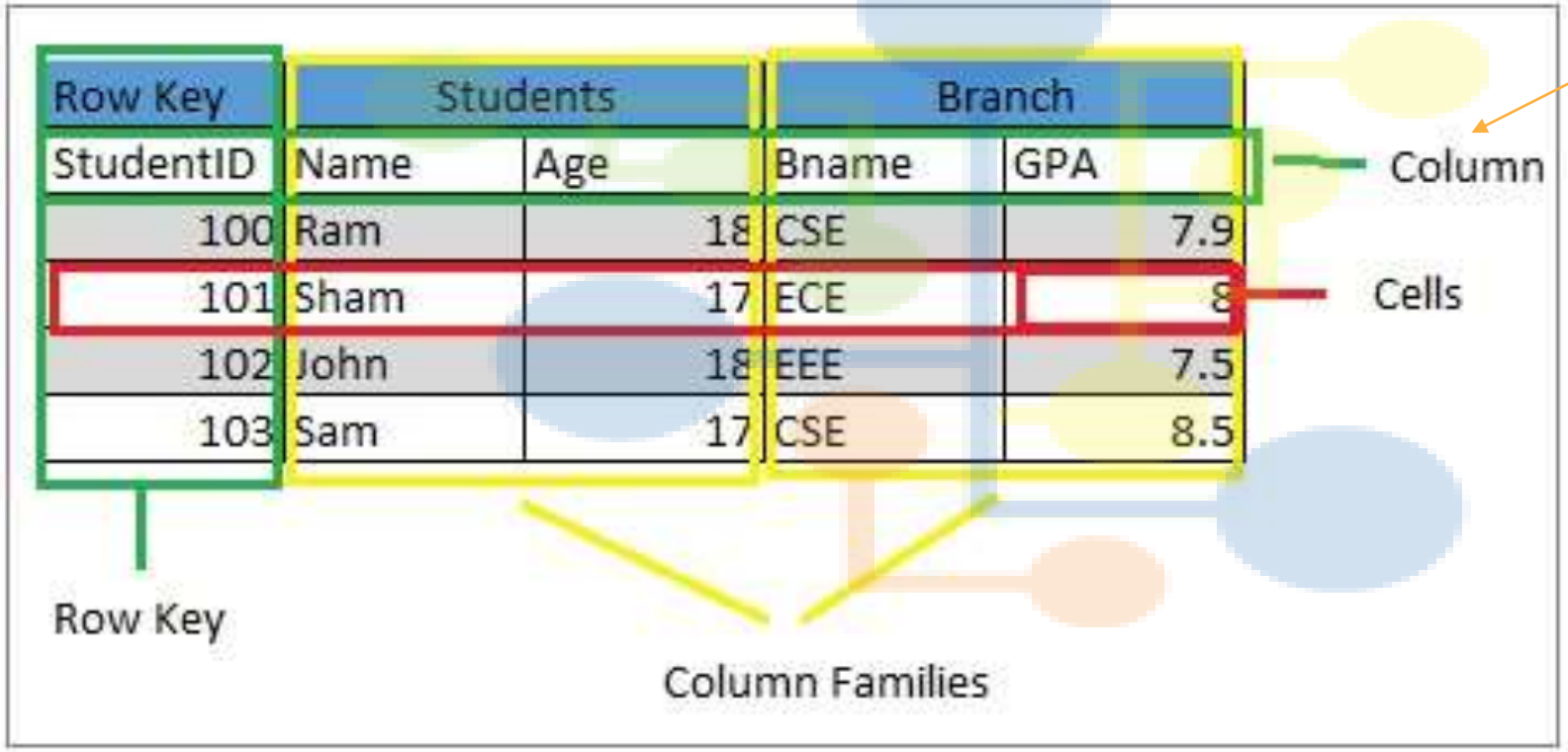
mongoDB



redis

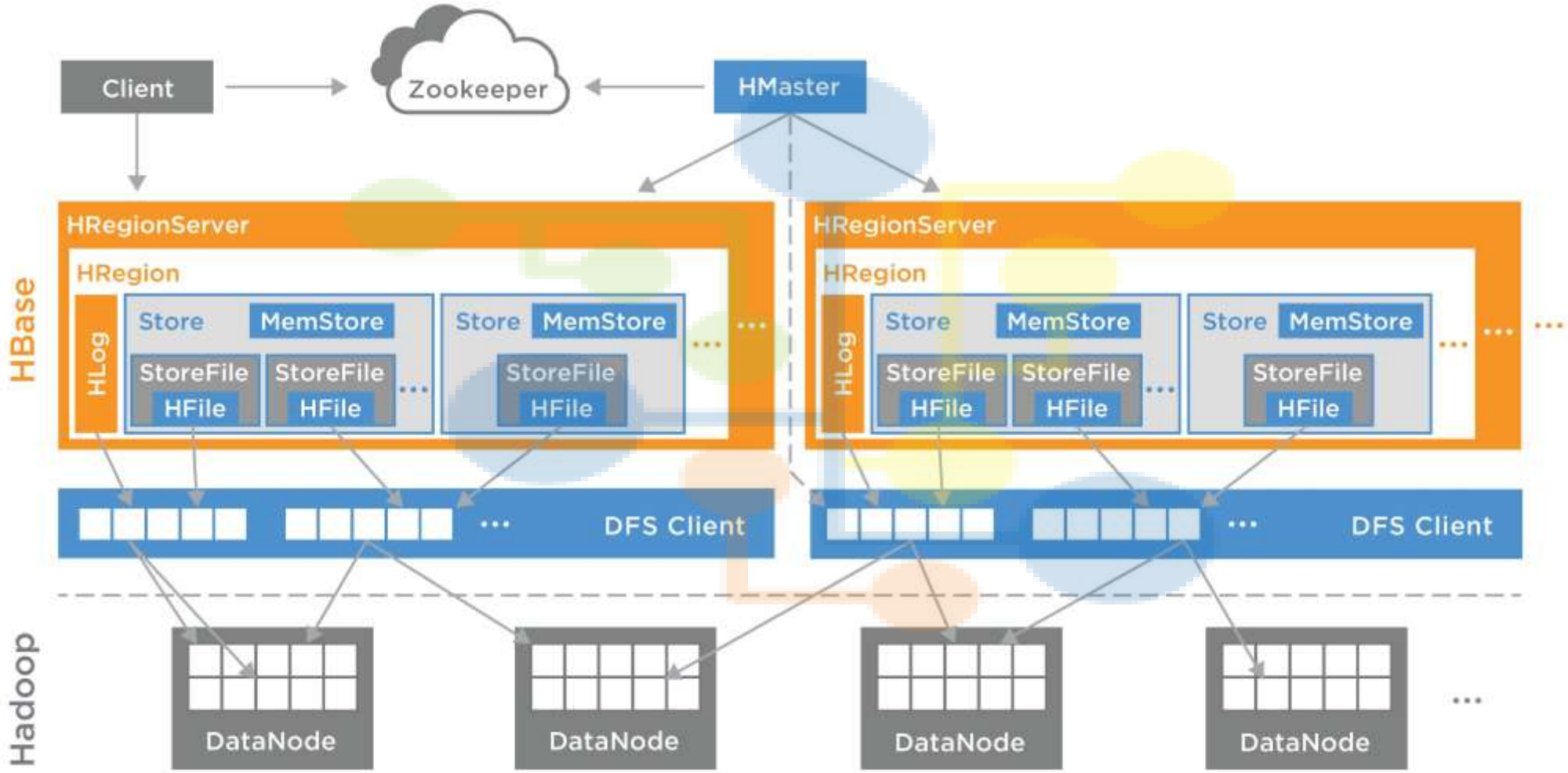
Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

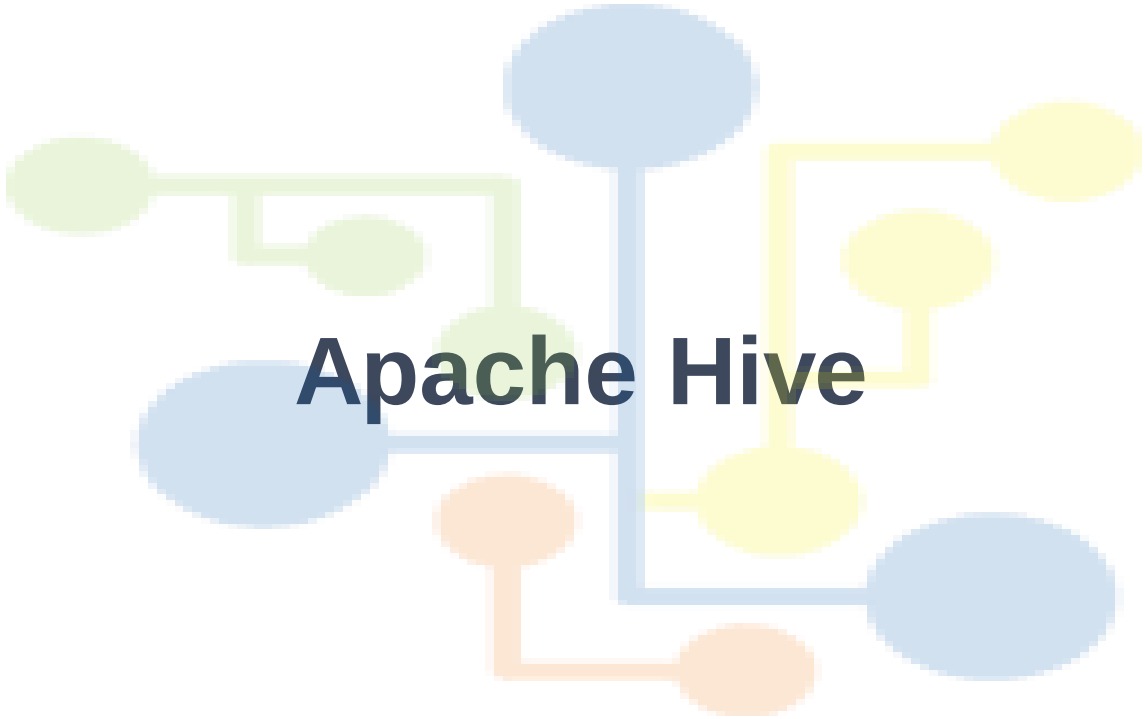
Apache HBase



Representa os dados de uma coluna!

Apache HBase



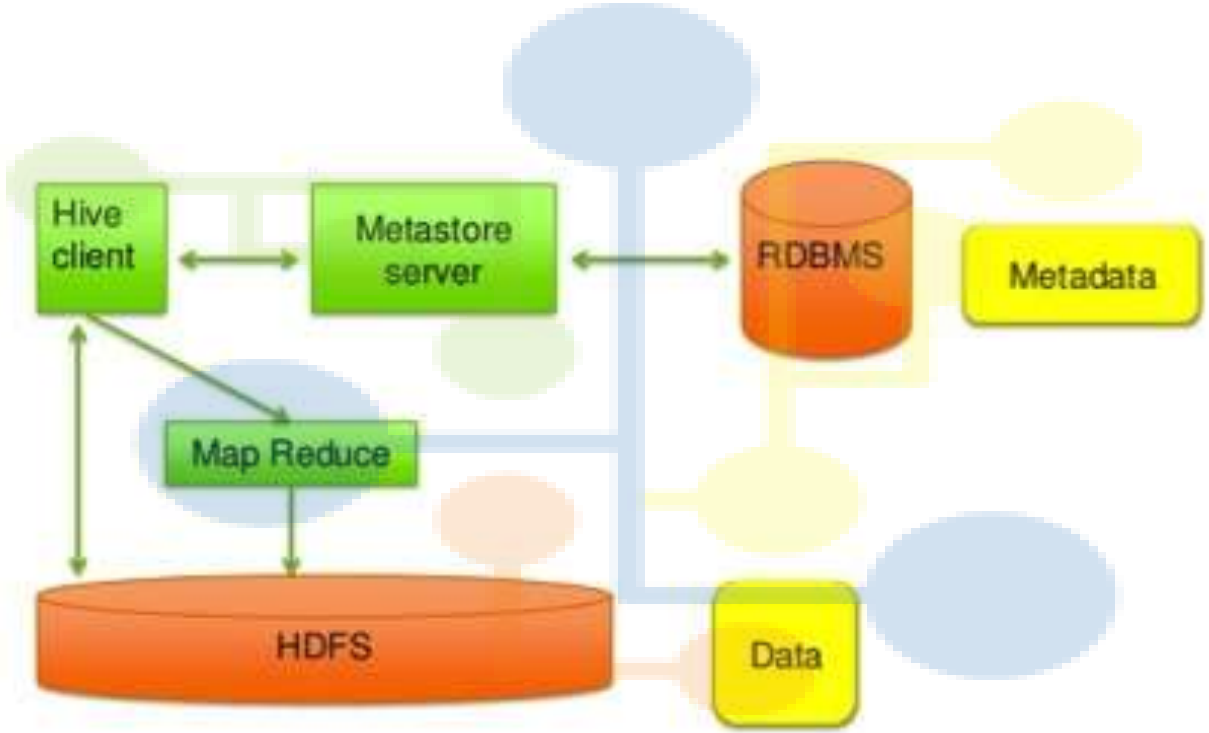


Apache Hive

Apache Hive

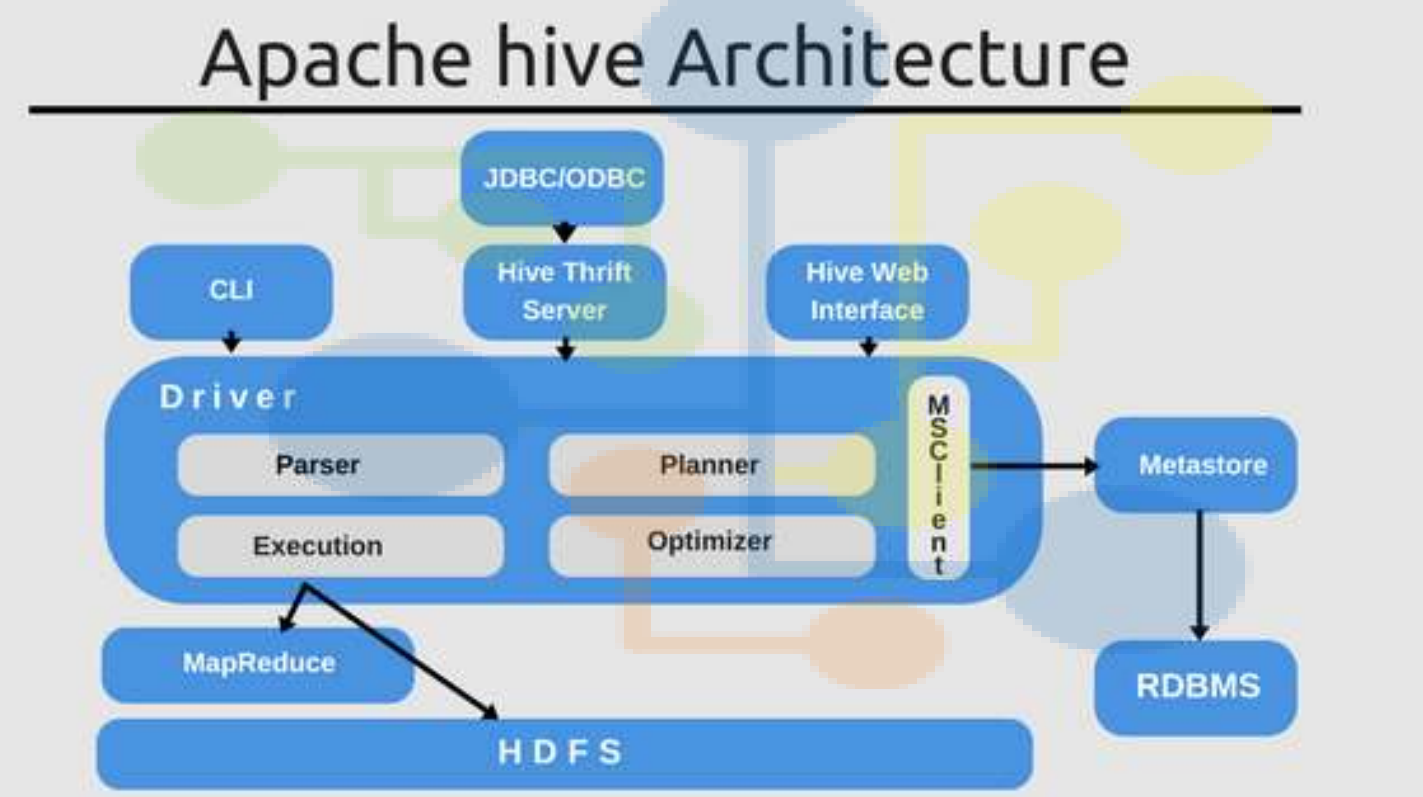
Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Apache Hive

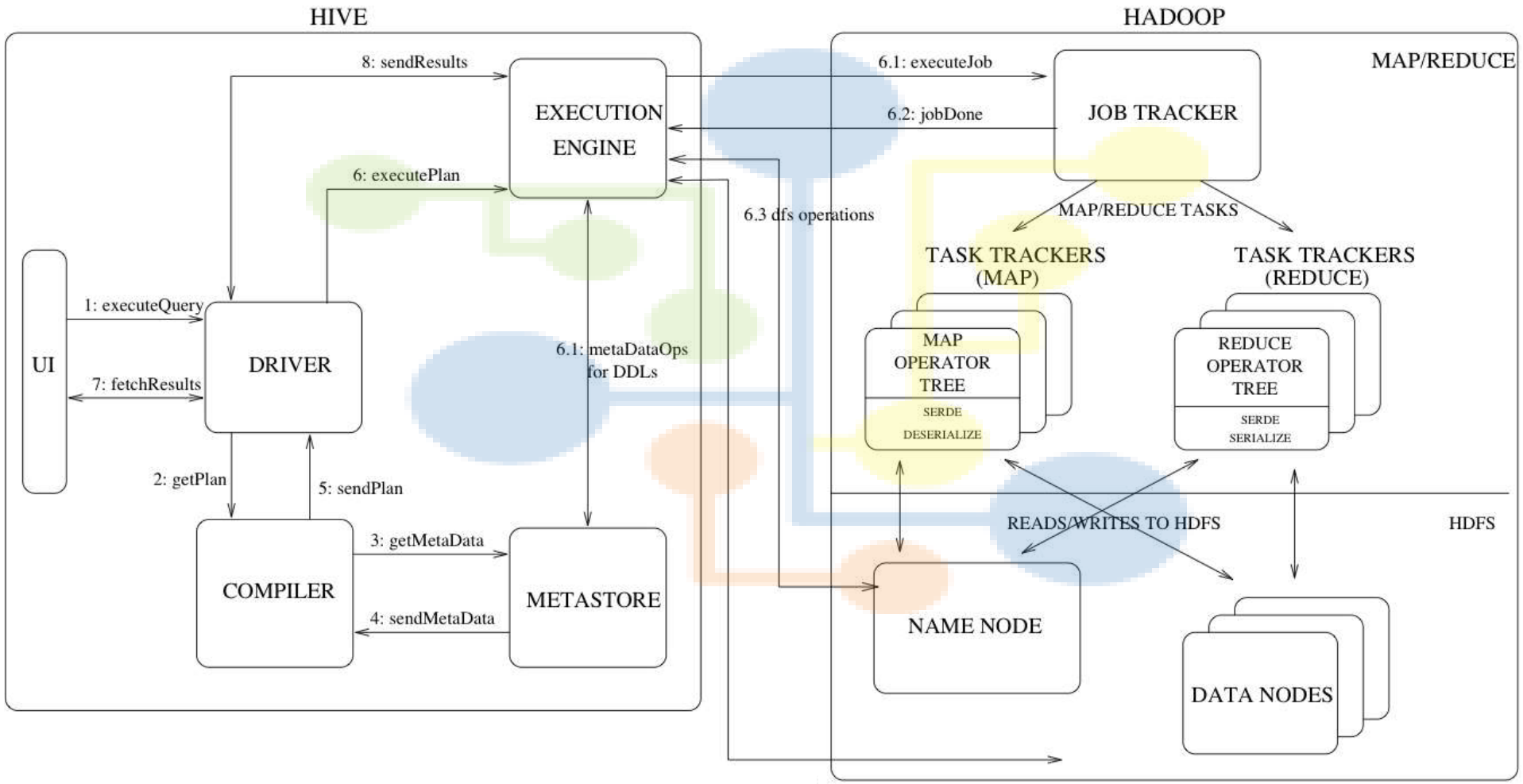


Apache Hive

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



Apache Hive


HIVE
HADOOP
MAP/REDUCE
8: sendResults
6.1: executeJob
6.2: jobDone
EXECUTION ENGINE
JOB TRACKER
6: executePlan
6.3 dfs operations
MAP/REDUCE TASKS
TASK TRACKERS (MAP)
TASK TRACKERS (REDUCE)
1: executeQuery
UI
DRIVER
7: fetchResults
6.1: metaDataOps for DDLs
MAP OPERATOR TREE
SERDE DESERIALIZE
REDUCE OPERATOR TREE
SERDE SERIALIZE
2: getPlan
5: sendPlan
READS/WRITES TO HDFS
HDFS
3: getMetaData
COMPILER
METASTORE
4: sendMetaData
NAME NODE
DATA NODES

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

Conhecendo o Apache HBase

Conhecendo o Apache HBase

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

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

Conhecendo o Apache HBase

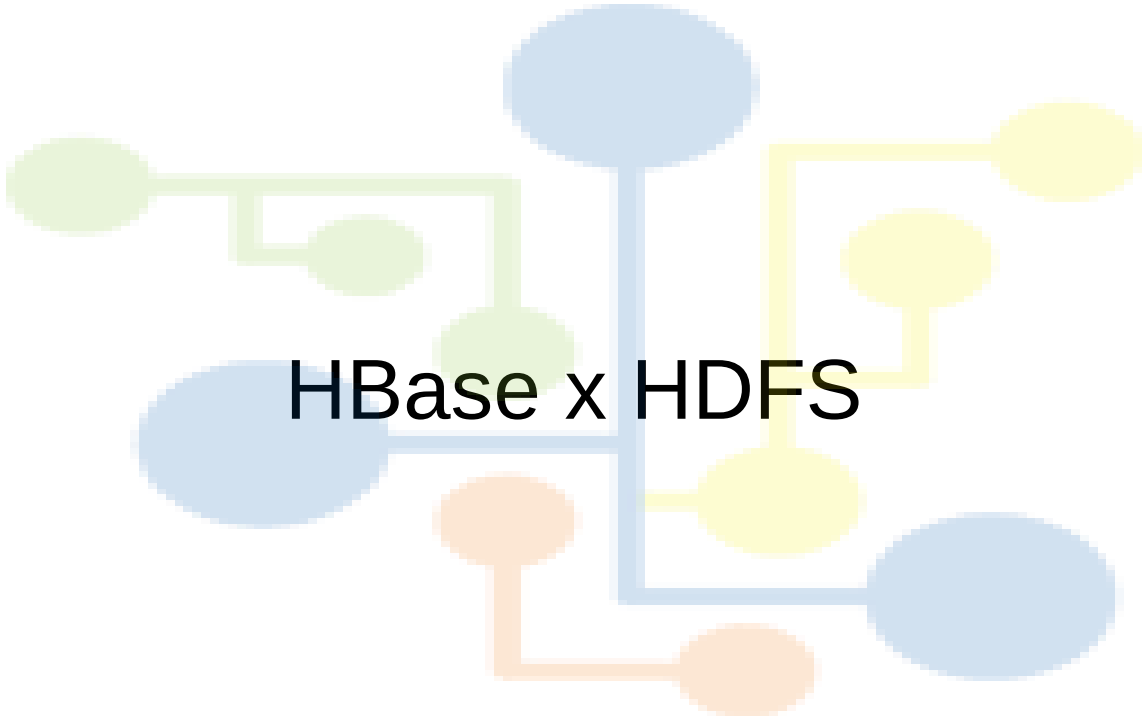
Hbase é um banco de dados distribuído, open-source, não-relacional, inspirado no Google Big Table.

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

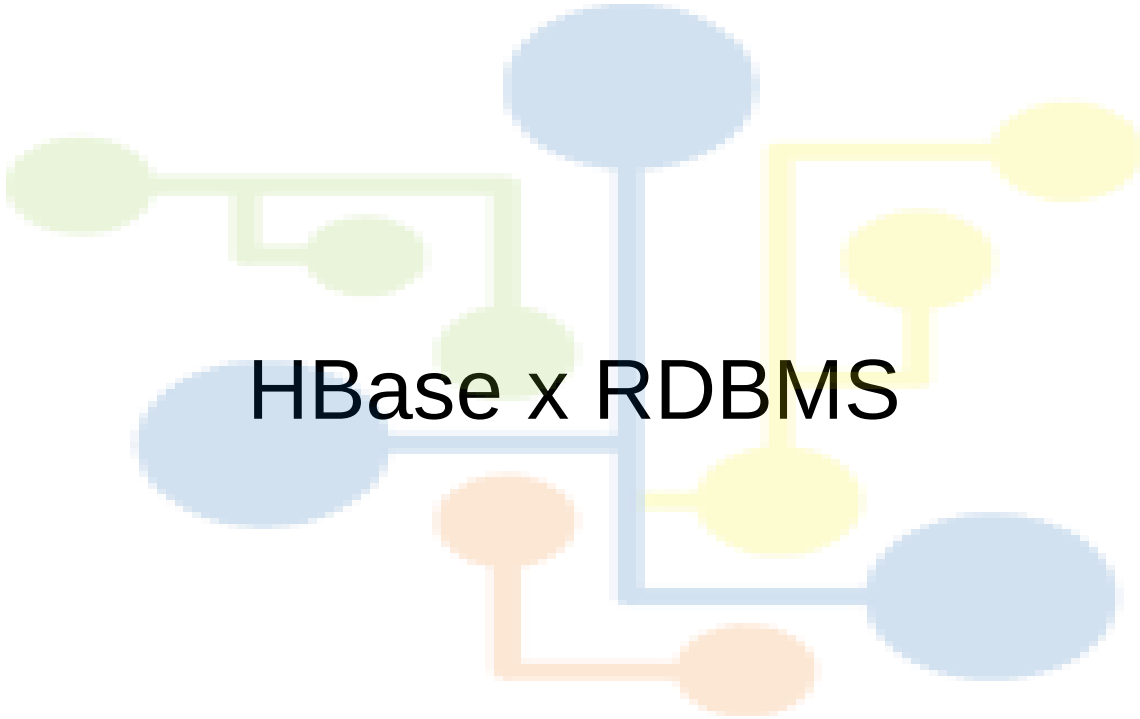


HBase x HDFS

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

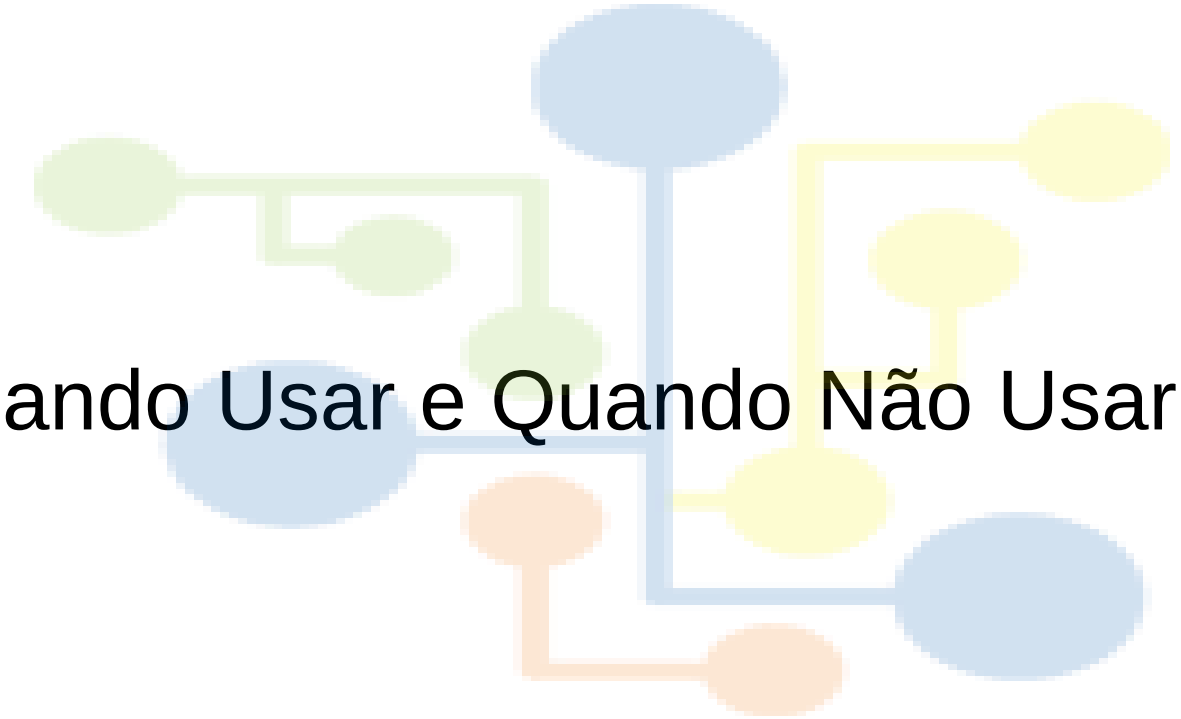


HBase x RDBMS

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?

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

Obrigado