Engenharia de Dados com Hadoop e Spark

Data Science Academy

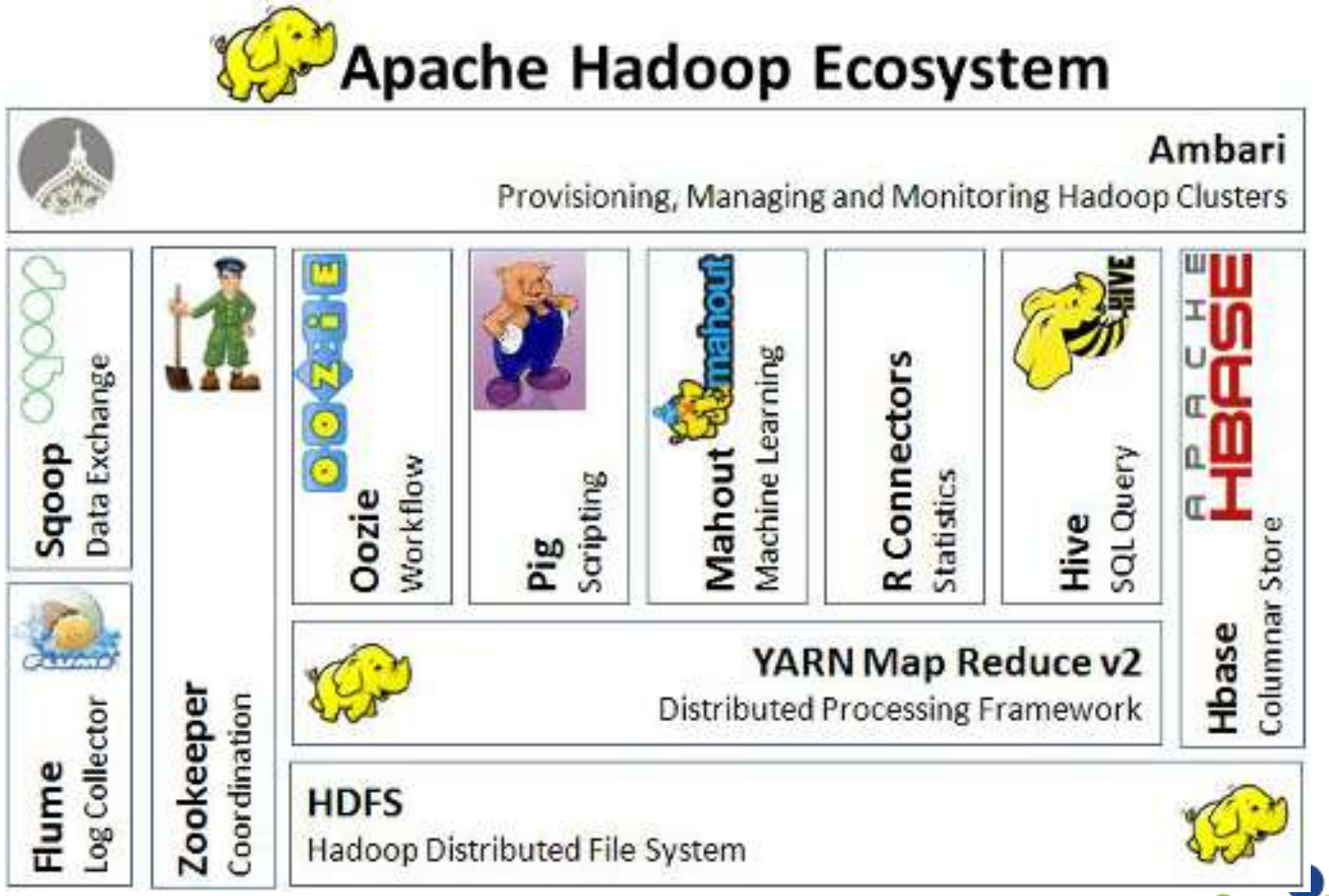
Bem-vindo

Data Science Academy

Data Science Academy

Tudo bem até aqui?

Tudo bem até aqui?



Data Science Academy

Tudo bem até aqui?

Data Science Academy

Tudo bem até aqui?

Data Science Academy

Armazenamento de Dados com Hbase e Hive

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Data Science Academy

Armazenamento de Dados – Hbase e Hive

O Hbase é para Big Data Problems

Data Science Academy

Armazenamento de Dados – Hbase e Hive

O Hbase é Bancos de Dados NoSQL, distribuído e escalável que funciona sobre o Apache Hadoop

”sharding”

Data Science Academy

Armazenamento de Dados – Hbase e Hive

O Hbase pode armazenar grandes quantidades de dados com bilhões de registros e milhões de colunas

Data Science Academy

Armazenamento de Dados – Hbase e Hive

O que é preciso para trabalhar com Hbase?

- Conhecimento sobre Hadoop
- Conhecimento sobre Sistema Operacional Linux
- Conceitos de Programação

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Como o Hbase vem sendo utilizado?

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Como o Hbase vem sendo utilizado?

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Como o Hbase vem sendo utilizado?

Data Science Academy

Armazenamento de Dados – Hbase e Hive

O Hive é um Framework para soluções de Data Warehousing com Hadoop

Armazenamento de Dados – Hbase e Hive

Pode ser integrado a ferramentas de Business Intelligence

Armazenamento de Dados – Hbase e Hive

HiveQL → Jobs MapReduce

Armazenamento de Dados – Hbase e Hive

Infraestrutura de Data Warehouse

Pode executar Jobs MapReduce

Ferramentas para ETL

Utiliza Linguagem HQL

Data Science Academy

Armazenamento de Dados – Hbase e Hive

O que estudaremos neste capítulo?

- Arquitetura do Hbase e do Hive
- Manipulação de dados com Hbase e Hive
- Criação de tabelas, importação de dados, consultas
- Vamos utilizar o HiveQL
- Consultas com o Pig
- Análise de Dados com Linguagem R

Data Science Academy

O que é Apache Hbase?

Armazenamento de Dados – Hbase e Hive

Data Science Academy

Armazenamento de Dados – Hbase e Hive

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

Data Science Academy

Armazenamento de Dados – Hbase e Hive

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

Data Science Academy

Armazenamento de Dados – Hbase e Hive

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

Armazenamento de Dados – Hbase e Hive

Hbase x HDFS

Data Science Academy

Armazenamento de Dados – Hbase e Hive

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.

Armazenamento de Dados – Hbase e Hive

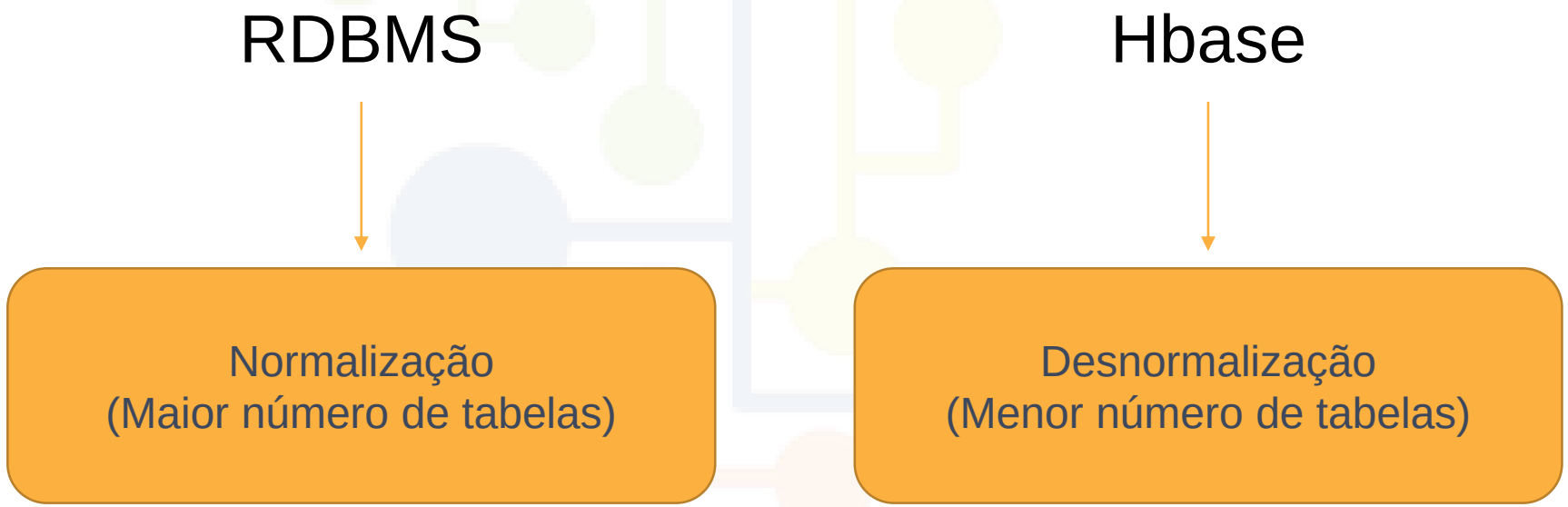
Hbase x RDBMS

Data Science Academy

Armazenamento de Dados – Hbase e Hive

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.

Armazenamento de Dados – Hbase e Hive



Armazenamento de Dados – Hbase e Hive

Quando Utilizar Hbase?

- Dados não 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

Data Science Academy

Armazenamento de Dados – Hbase e Hive

- Poucas linhas devem ser armazenadas
- Não for necessário realizar consultas cruzadas (SQL Joins)
- Cluster com poucas máquinas

Quando **Não** Utilizar Hbase?

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Mas o que é um banco de dados não relacional (NoSQL)?

Data Science Academy

Bancos de Dados NoSQL

Data Science Academy

NO SQL

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Bancos de Dados NoSQL, são bancos de dados não-relacionais, que foram projetados para atender os requerimentos de Big Data

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Características de Bancos de Dados NoSQL:

- Ideal para soluções analíticas
- Modelos de dados flexíveis
- Escalabilidade
- Representação de dados sem esquemas
- Velocidade

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Bancos de Dados NoSQL

x

Bancos de Dados Relacionais

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Bancos de Dados NoSQL	Bancos de Dados Relacionais
Orientados a coluna ou a chaves	Orientado a linha
Tabelas populadas de forma esparsa	Tabelas populadas de forma densa
Permite armazenar dados estruturados, semi-estruturados e não-estruturados	Dados estruturados
Escalabilidade horizontal	Escalabilidade é mais complexa

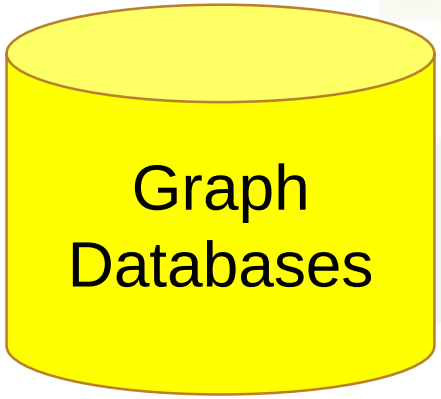
Armazenamento de Dados – Hbase e Hive

Bancos de Dados NoSQL oferecem 4 categorias principais de bancos de dados:

- Graph databases
- Document databases
- Key-values stores
- Column family stores

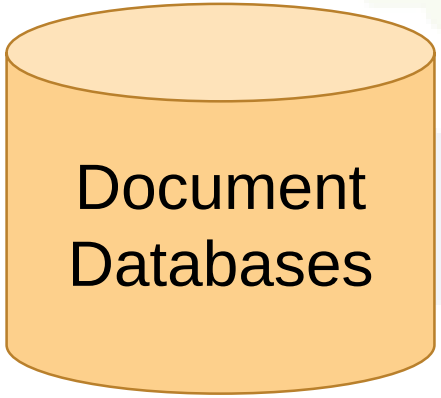
Data Science Academy

Armazenamento de Dados – Hbase e Hive



Graph	Neo4J
	FlockDB
	GraphDB
	ArangoDB

Armazenamento de Dados – Hbase e Hive

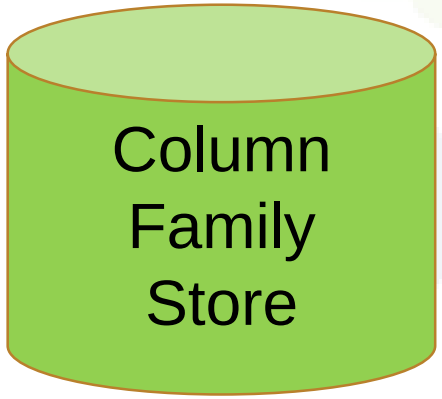


Document	MongoDB
	CouchDB
	RavenDB
	Terrastore

Armazenamento de Dados – Hbase e Hive

Key-value	Oracle NoSQL DB
	MemcacheDB
	Redis
	Voldemort

Armazenamento de Dados – Hbase e Hive



Column	**HBase**
	Cassandra*
	Hypertable
	Accumulo

* Cassandra é híbrido, Column e Key-value

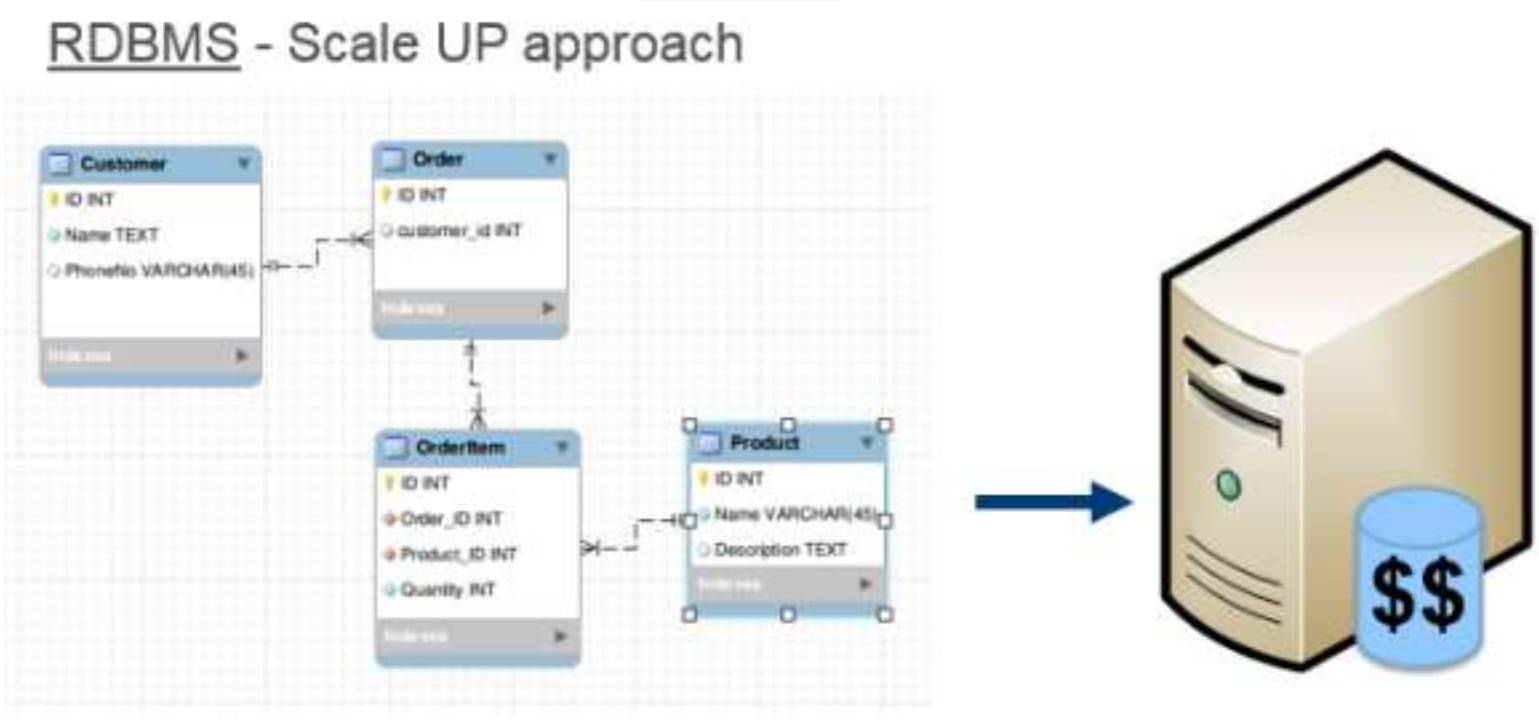
Armazenamento de Dados – Hbase e Hive

Data Science Academy

Modelo de Dados do Hbase

Data Science Academy

Armazenamento de Dados – Hbase e Hive



Escalabilidade Vertical

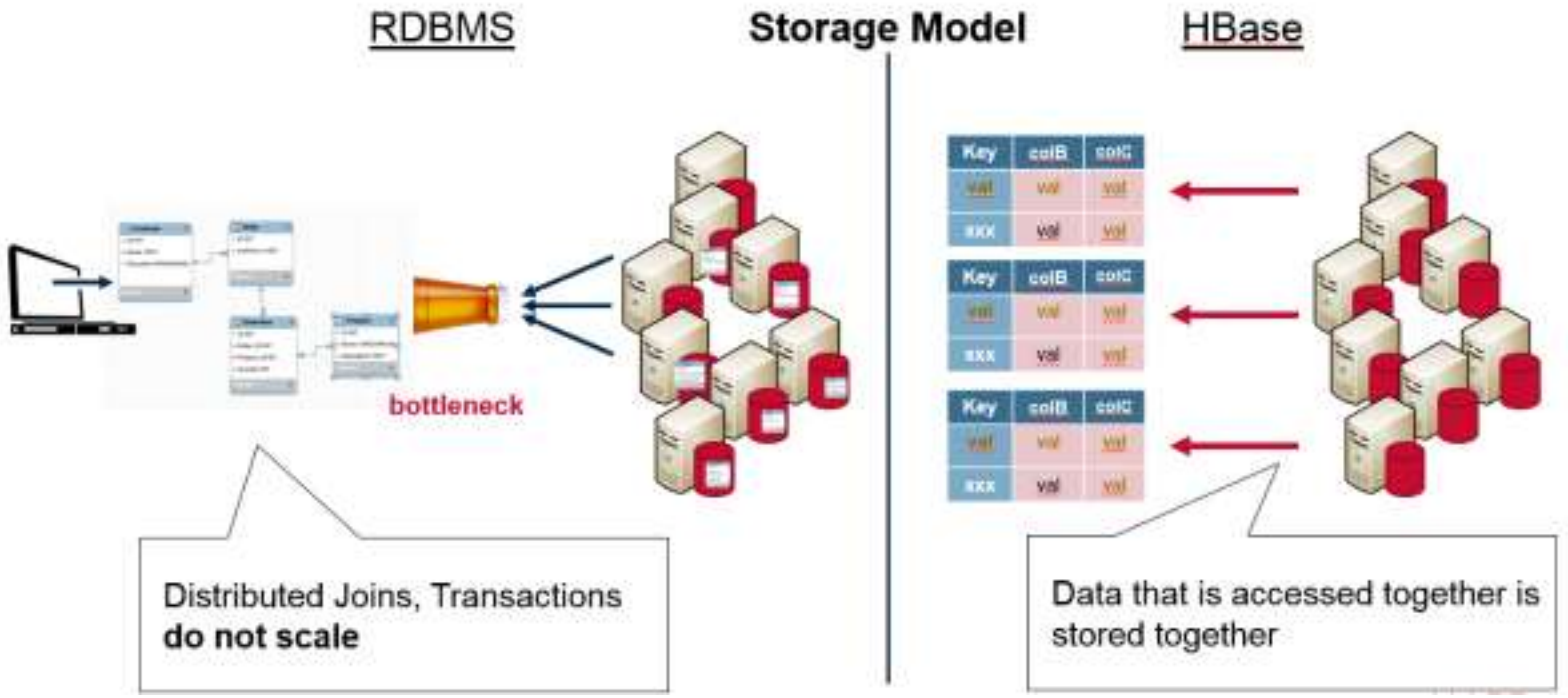
Data Science Academy

Armazenamento de Dados – Hbase e Hive

Tabelas divididas em partições e distribuídas no cluster

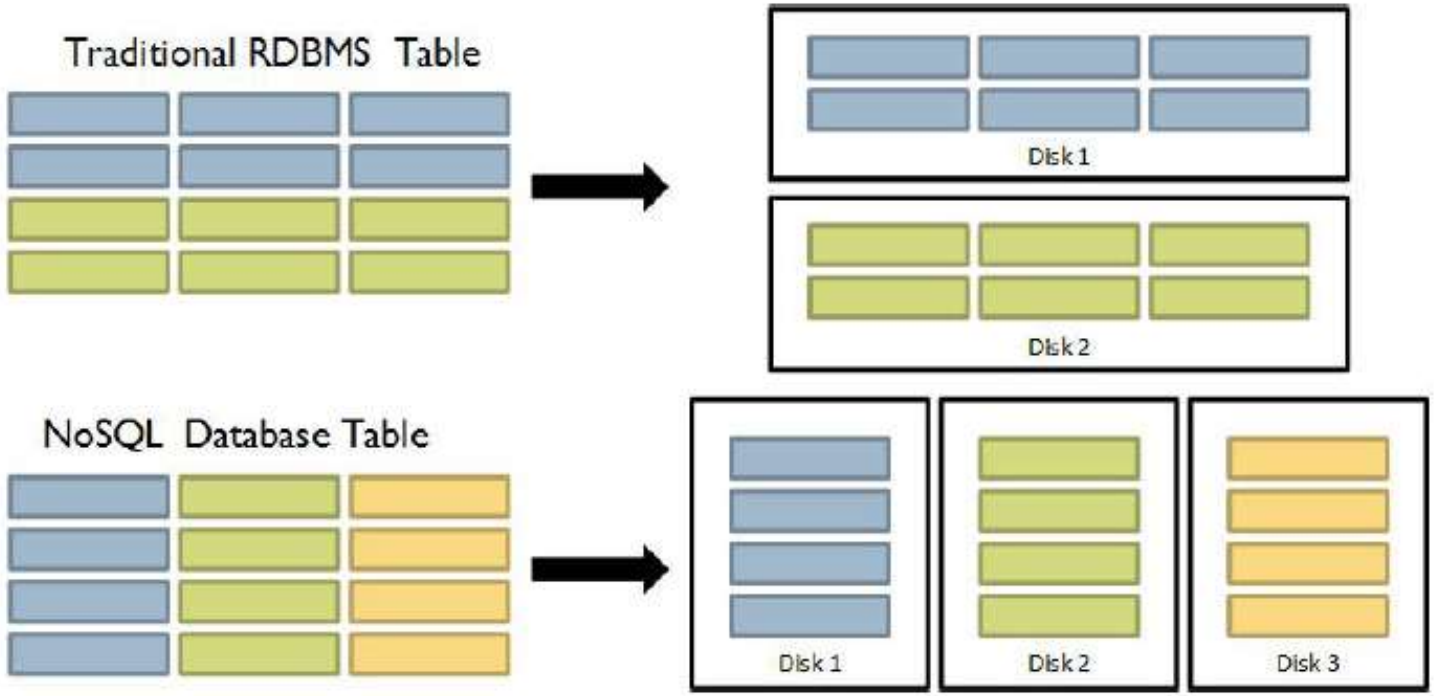
Escalabilidade Horizontal

Armazenamento de Dados – Hbase e Hive



Data Science Academy

Armazenamento de Dados – Hbase e Hive

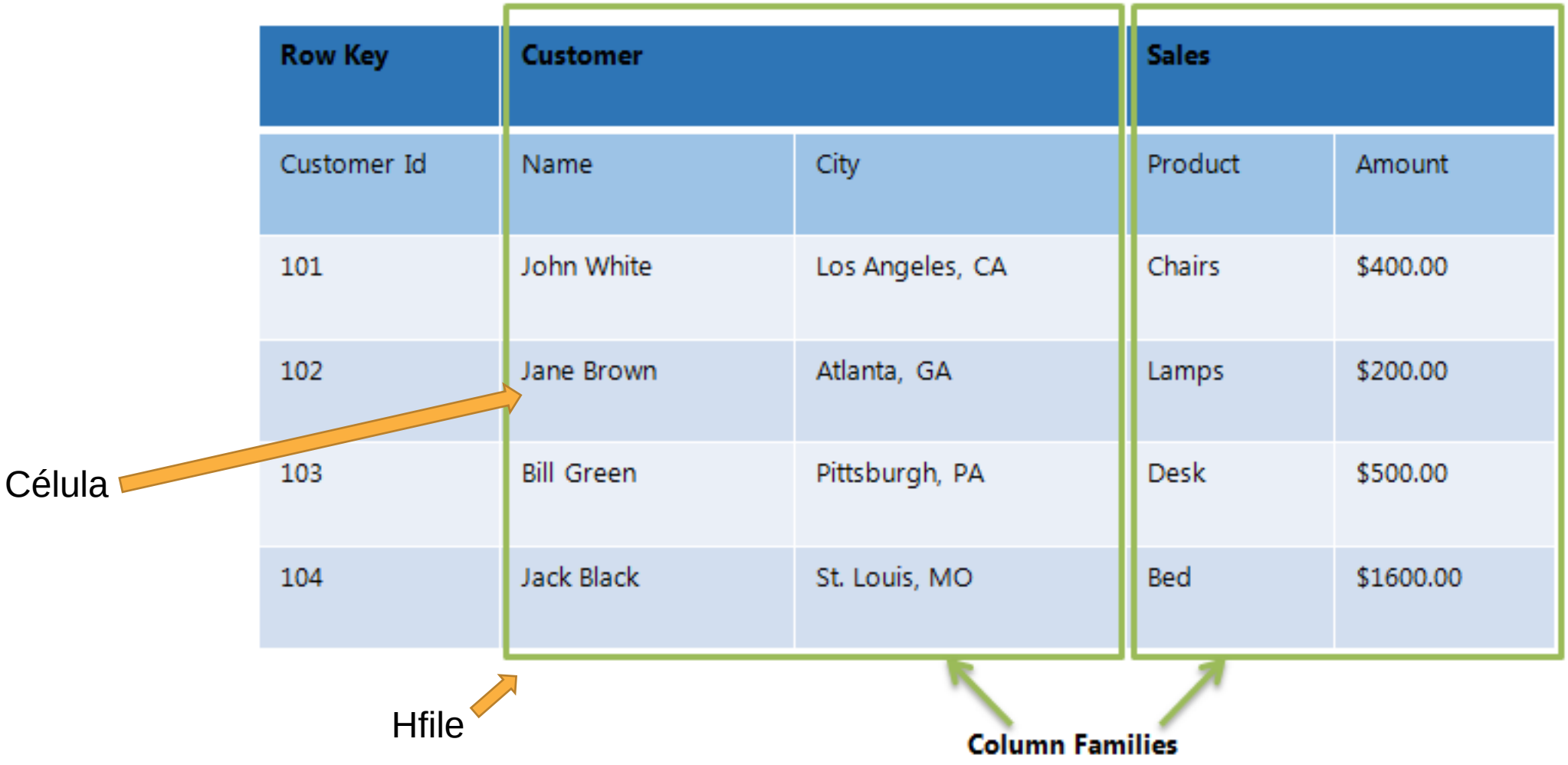


Armazenamento de Dados – Hbase e Hive

Row Key	Customer		Sales	
Customer Id	Name	City	Product	Amount
101	John White	Los Angeles, CA	Chairs	$400.00
102	Jane Brown	Atlanta, GA	Lamps	$200.00
103	Bill Green	Pittsburgh, PA	Desk	$500.00
104	Jack Black	St. Louis, MO	Bed	$1600.00

Column Families

Armazenamento de Dados – Hbase e Hive



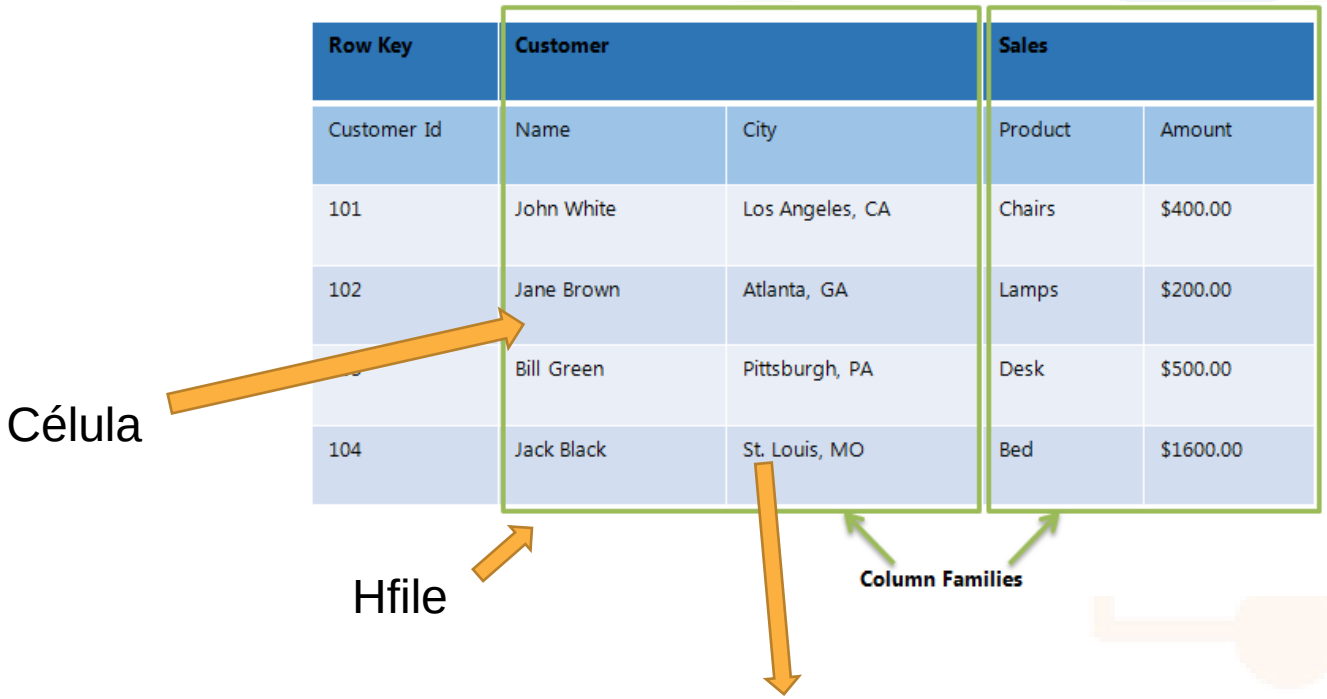
Row Key	Customer		Sales	
Customer Id	Name	City	Product	Amount
101	John White	Los Angeles, CA	Chairs	$400.00
102	Jane Brown	Atlanta, GA	Lamps	$200.00
103	Bill Green	Pittsburgh, PA	Desk	$500.00
104	Jack Black	St. Louis, MO	Bed	$1600.00

Armazenamento de Dados – Hbase e Hive

Timestamp é uma sequência de caracteres que identifica quando um evento ocorre e normalmente com dados de hora no nível de fração de segundos

Data Science Academy

Armazenamento de Dados – Hbase e Hive



Identificação da célula = rowkey + column family + column key + timestamp

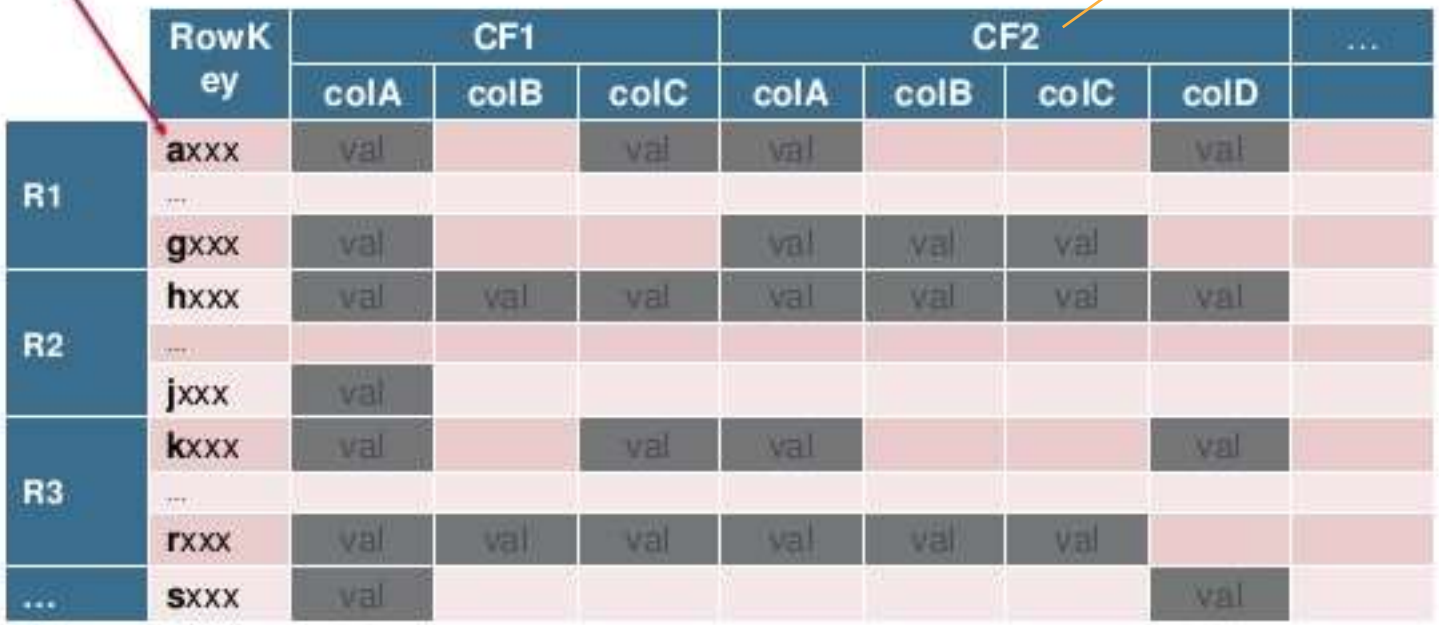
Data Science Academy

Armazenamento de Dados – Hbase e Hive

CF = Column Family

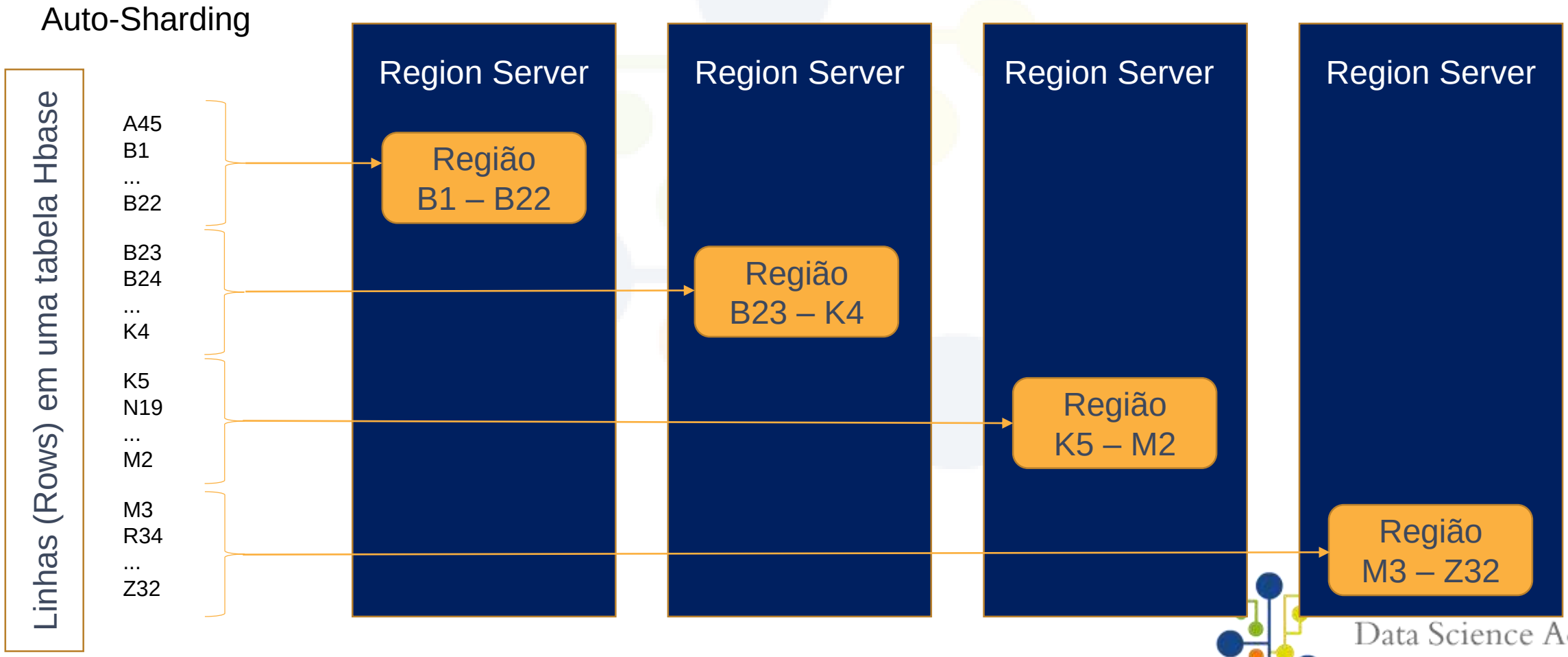
- Row Keys: Identificam unicamente um registro

Regiões



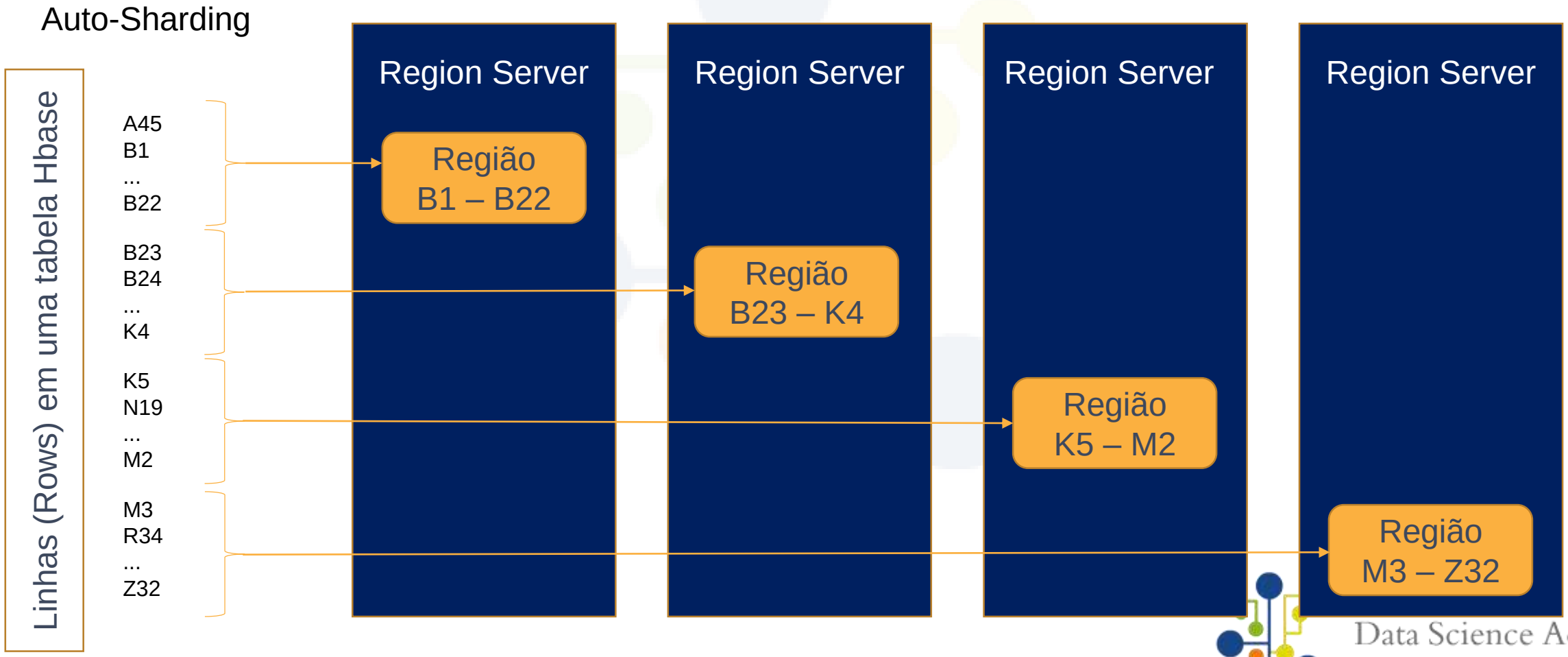
	RowKey	CF1			CF2				...
		colA	colB	colC	colA	colB	colC	colD	
R1	axxx	val		val	val			val	
	...								
	gxxx	val			val	val	val		
R2	hxxx	val	val	val	val	val	val	val	
	...								
	jxxx	val							
R3	kxxx	val		val	val			val	
	...								
	rxxx	val	val	val	val	val	val		
...	sxxx	val						val	

Armazenamento de Dados – Hbase e Hive



Data Science Academy

Armazenamento de Dados – Hbase e Hive

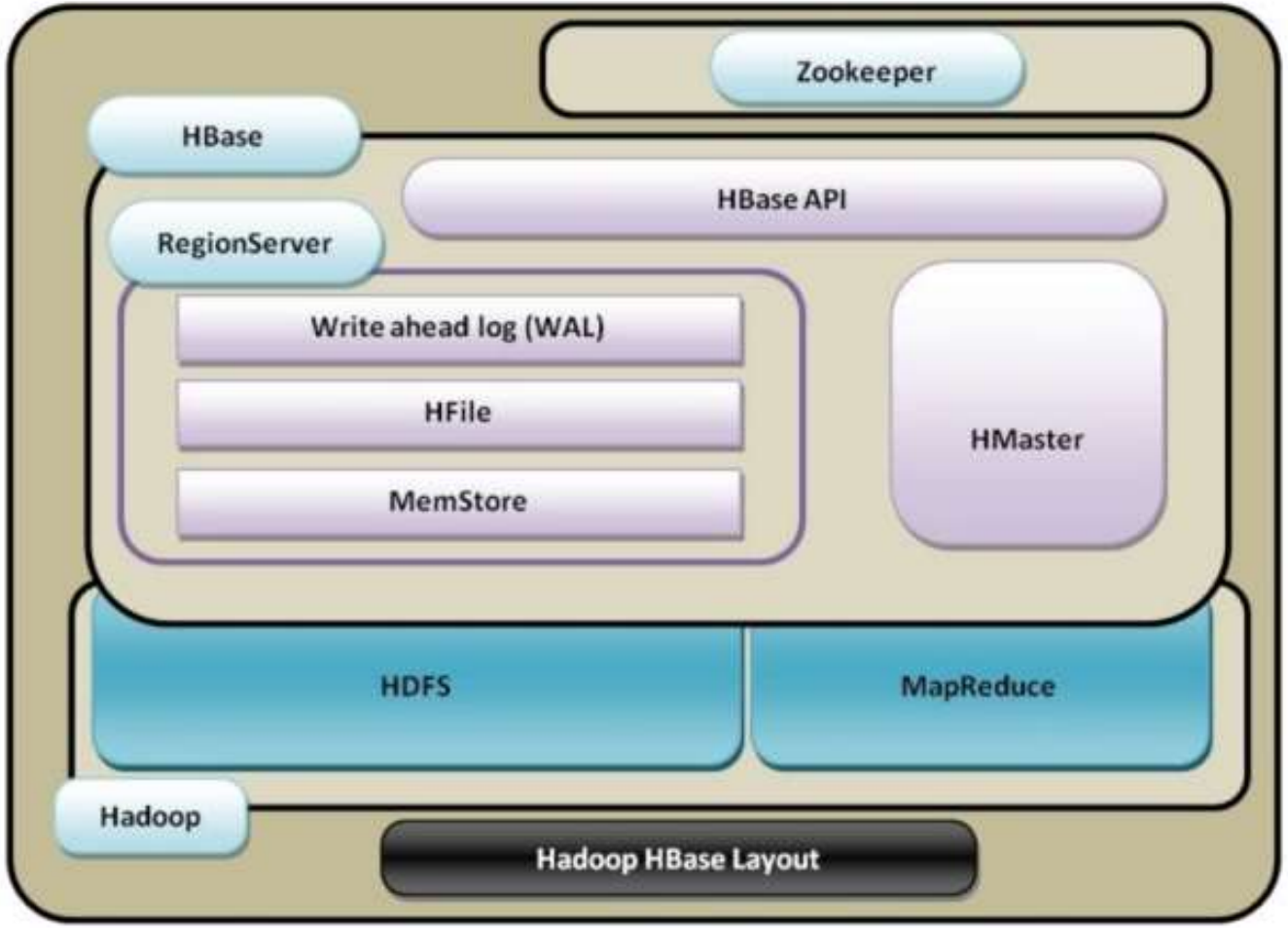


Data Science Academy

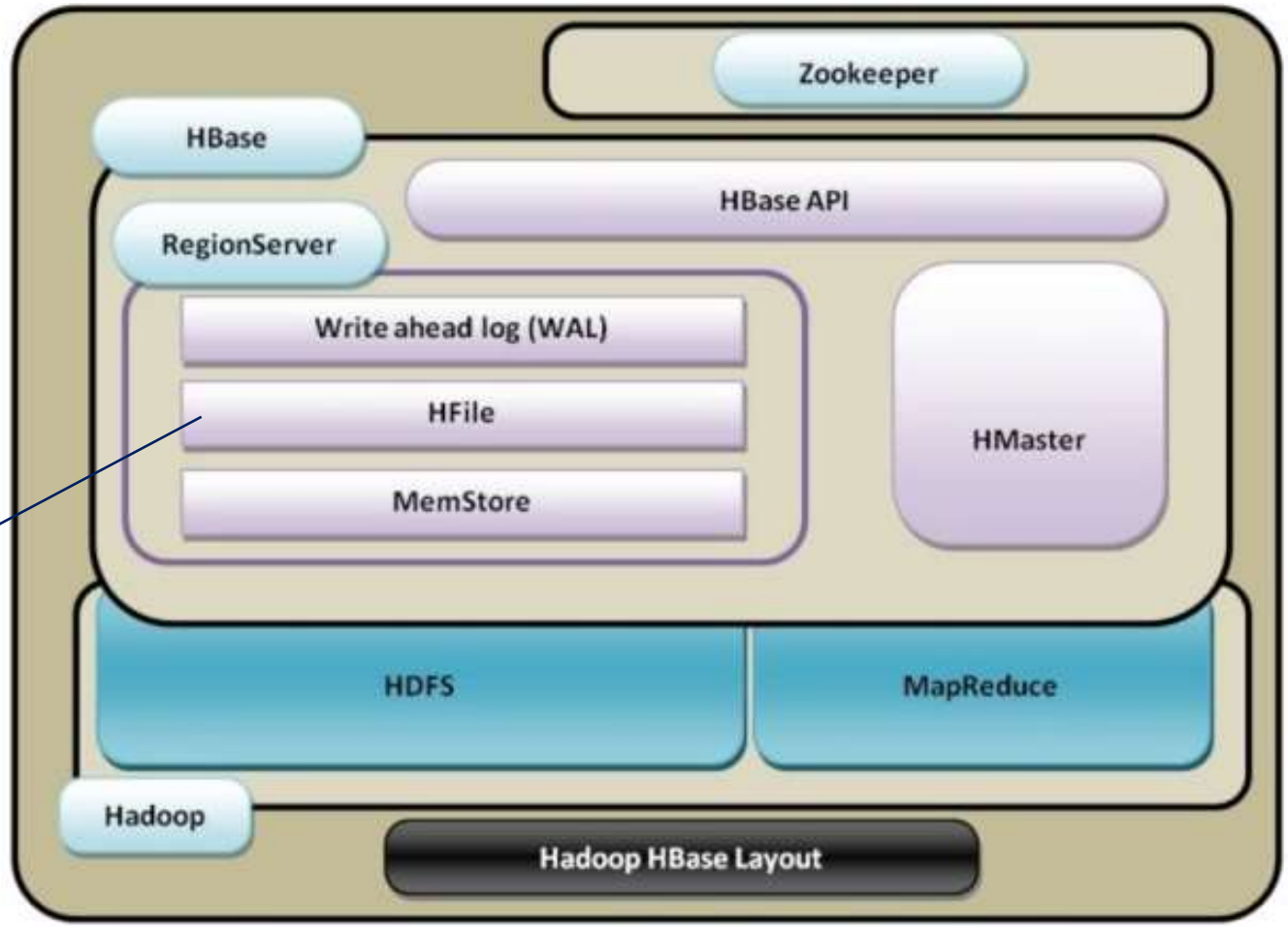
Arquitetura Hbase

Data Science Academy

Armazenamento de Dados – Hbase e Hive

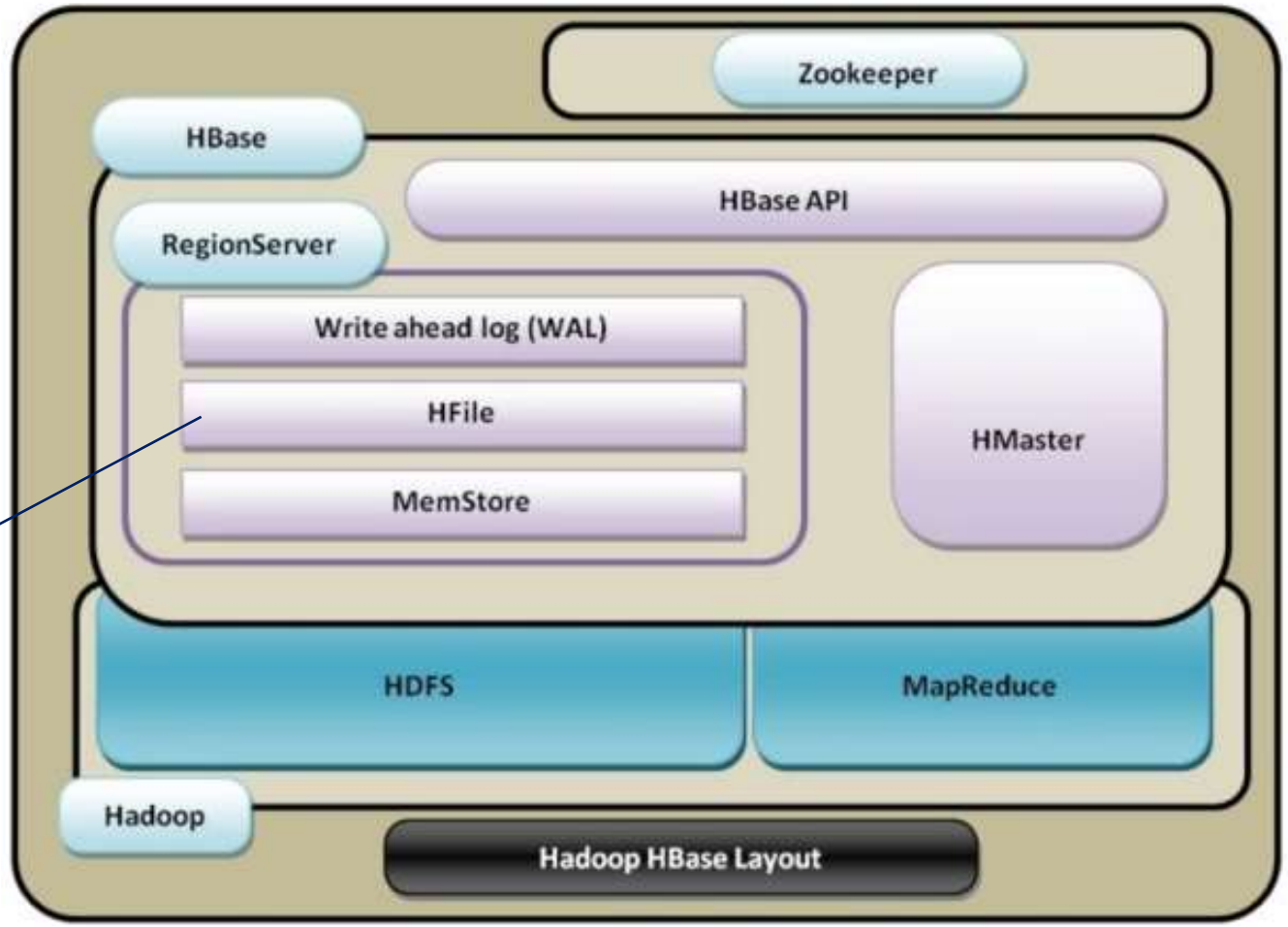


Armazenamento de Dados – Hbase e Hive



Armazena os dados em stores

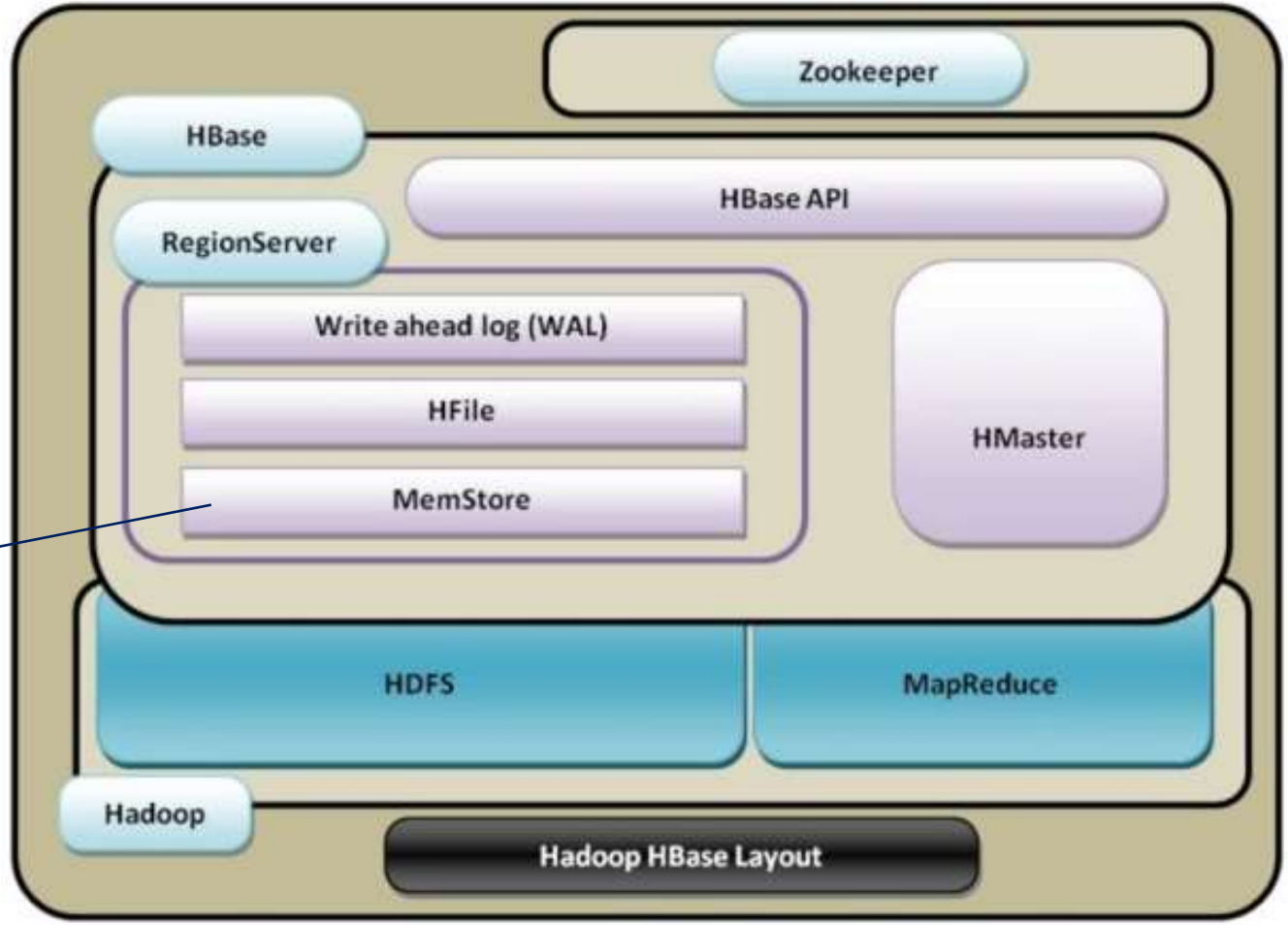
Armazenamento de Dados – Hbase e Hive



Armazena os dados em stores

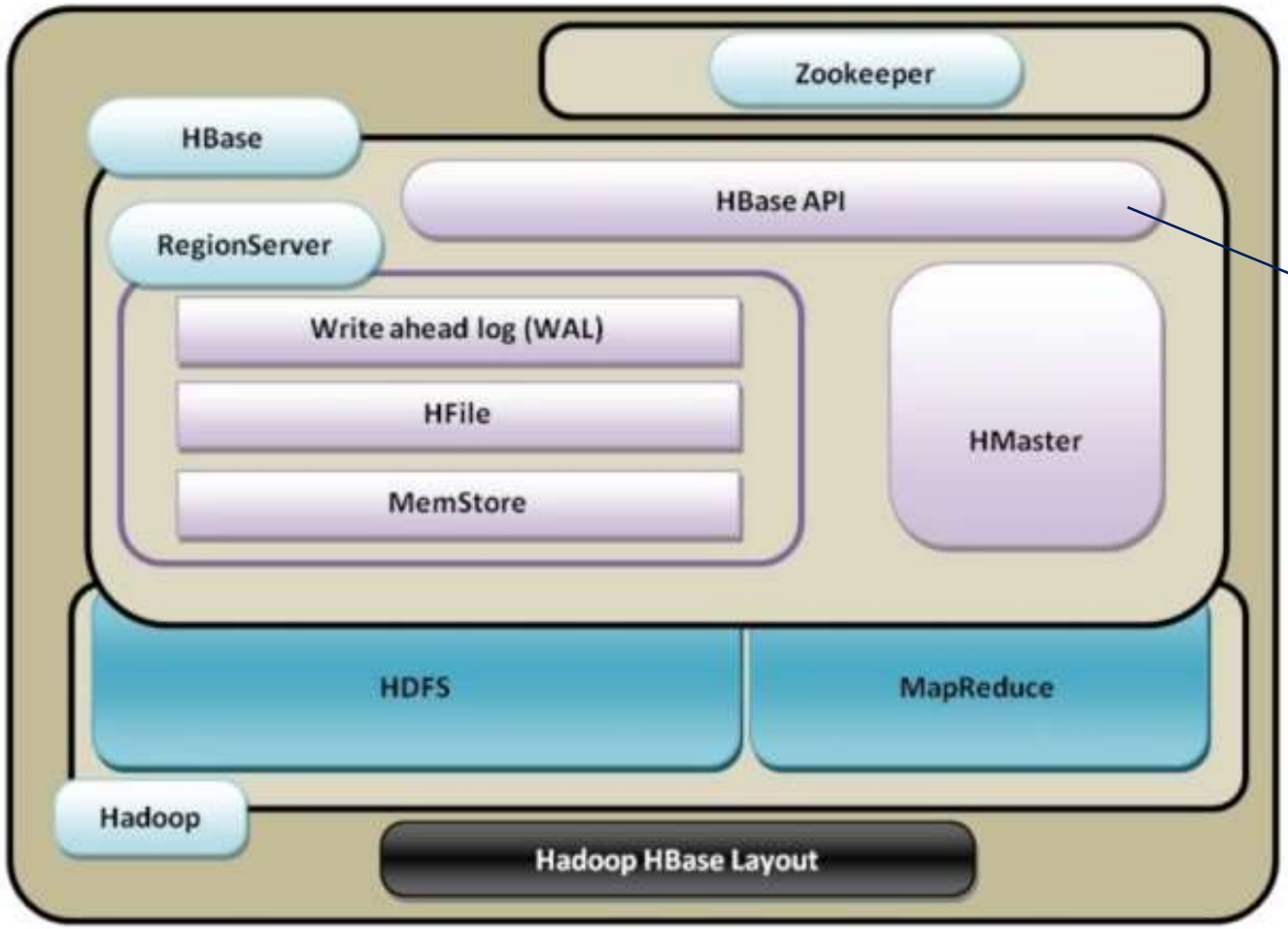
Write Ahead Log

Armazenamento de Dados – Hbase e Hive



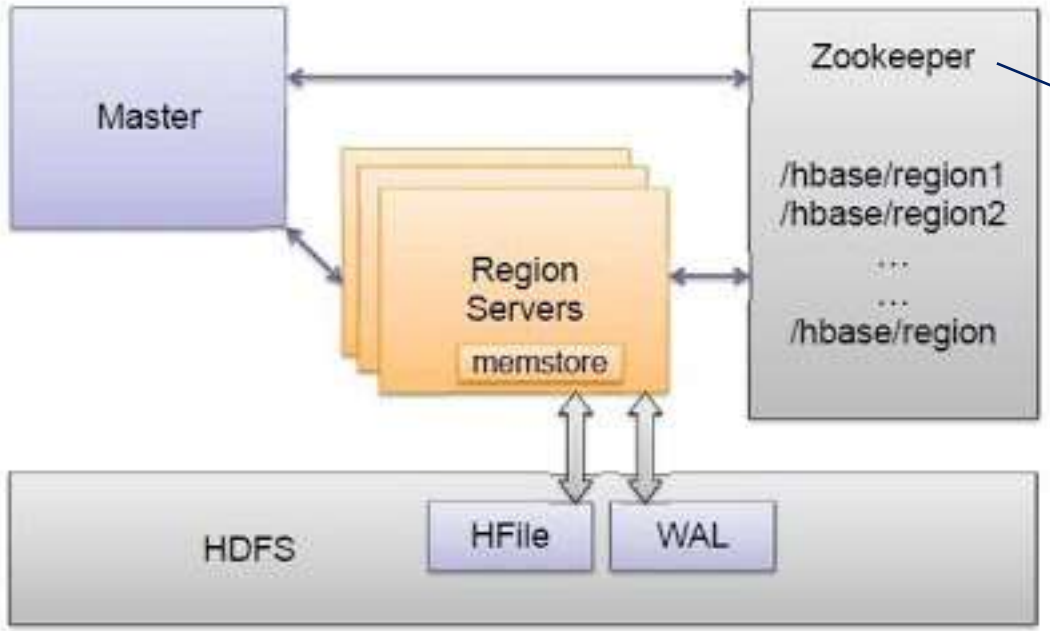
Armazena as stores

Armazenamento de Dados – Hbase e Hive



Acesso aos dados

Armazenamento de Dados – Hbase e Hive



Gestão dos
Region Servers

Data Science Academy

Operações CRUD

Data Science Academy

Armazenamento de Dados – Hbase e Hive

CRUD

Create, Read, Update e Delete

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Hbase API Java

- Hbase foi escrito em Java
- Suporte para operações CRUD
- Tudo que pode ser feito via linha de comando, pode ser feito com API Java
- API Java oferece o método mais rápido de acesso ao Hbase

Armazenamento de Dados – Hbase e Hive

Data Science Academy

Introdução ao Apache Hive

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Armazenamento de Dados – Hbase e Hive

Hive Query Language

SQL → Jobs MapReduce

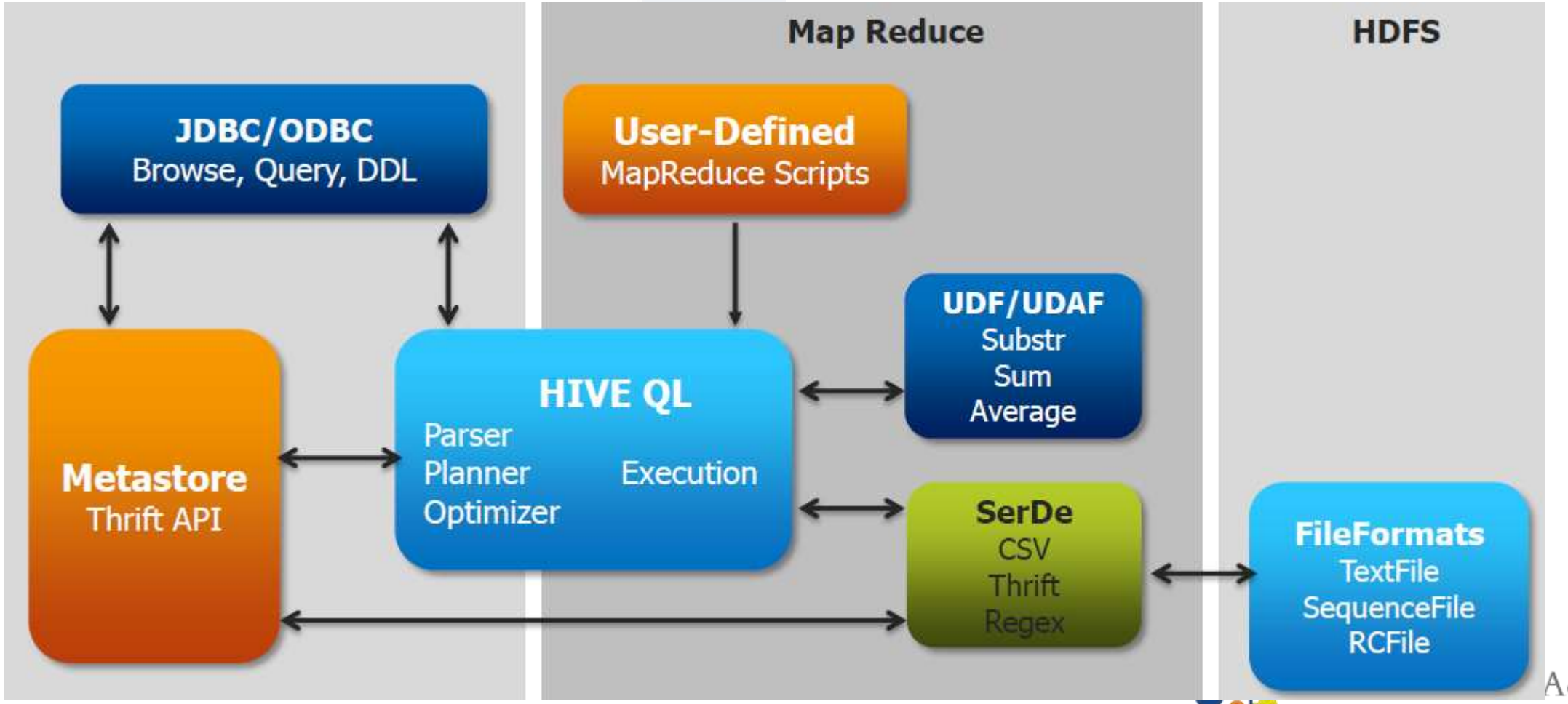
Data Science Academy

Armazenamento de Dados – Hbase e Hive

O que o Hive não é?

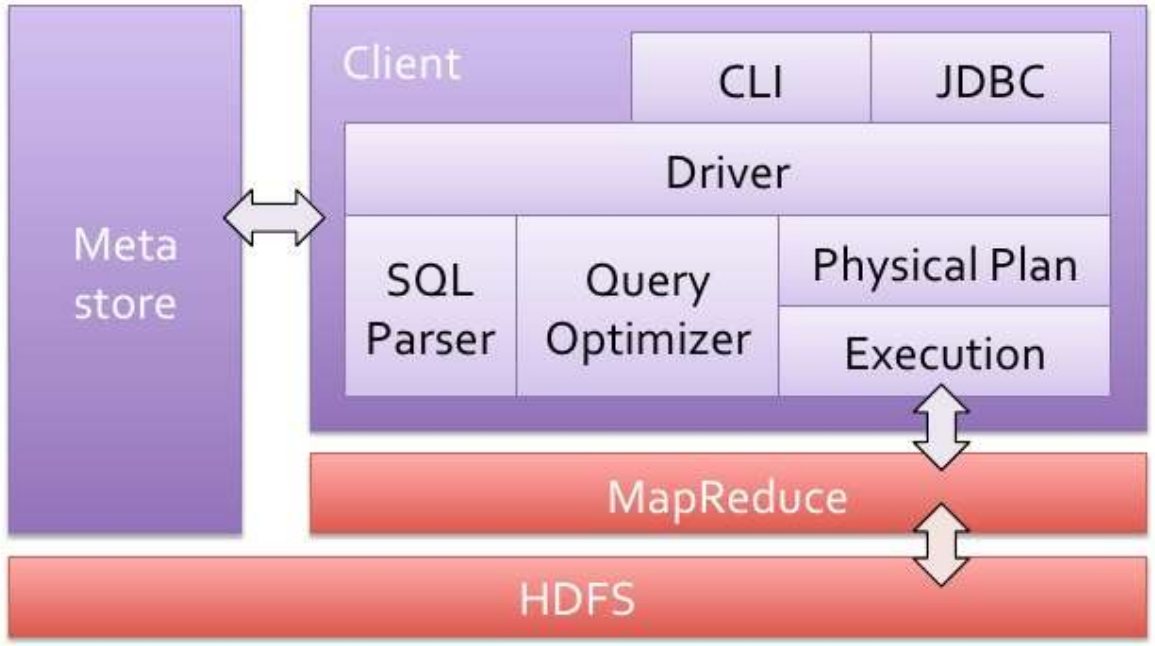
- Um banco de dados relacional
- Um projeto para Online Transaction Processing (OLTP)
- Uma solução para consultas em tempo real e atualizações em nível de linha

Armazenamento de Dados – Hbase e Hive

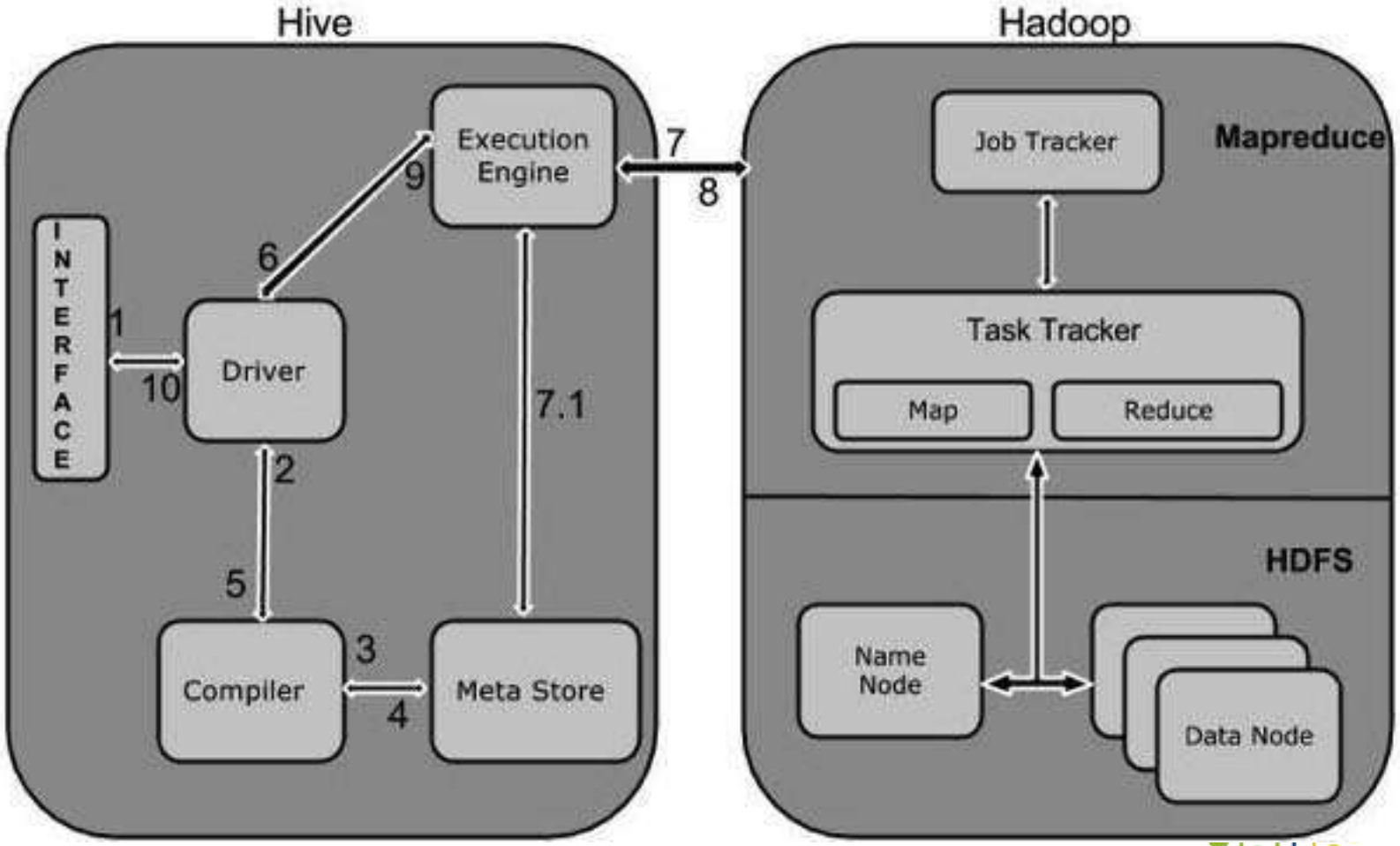


Armazenamento de Dados – Hbase e Hive

Arquitetura Hive



Armazenamento de Dados – Hbase e Hive



Data Science Academy

Armazenamento de Dados – Hbase e Hive

Tipos de Dados

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Tipos de Dados

Tipos de Coluna
(Int, Smallint, Tinyint, Bigint)
(Char, Varchar)
Dates
Timestamp

Tipos Literais
(Ponto Flutuante e Decimal)

Tipo Nulo

Tipos Complexos
(Arrays, Maps, Structs)

Data Science Academy

Armazenamento de Dados – Hbase e Hive

Funções Built-in

Data Science Academy

Obrigado

Data Science Academy

Data Science Academy