



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.



Engenharia de Dados com Hadoop e Spark



Modelo de Dados do Apache HBase



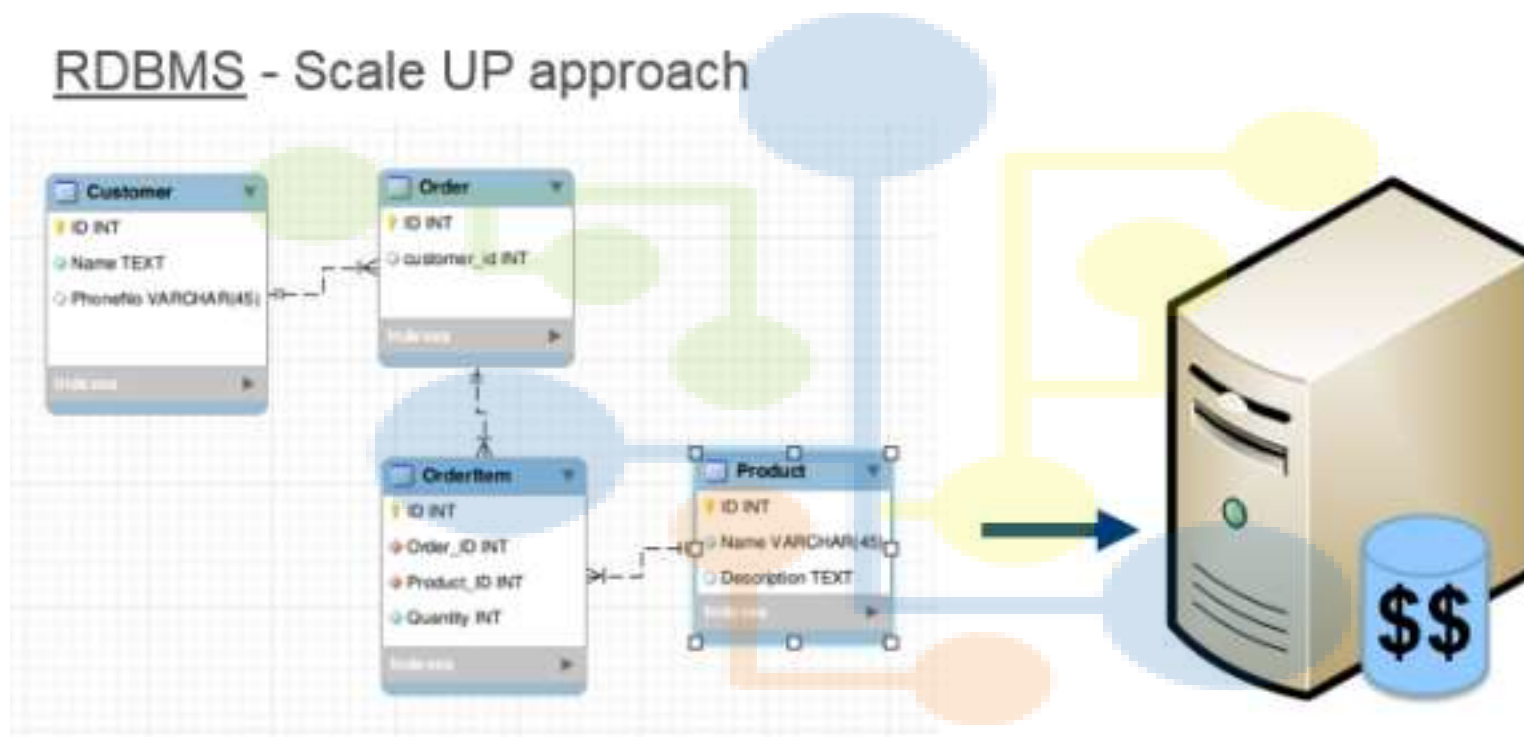


Modelo de Dados do Apache HBase





Modelo de Dados do Apache HBase



Escalabilidade Vertical



Modelo de Dados do Apache HBase

Tabelas divididas em
partições e distribuídas
no cluster

Key	colB	colC
val	val	val
xxx	val	val
Key	colB	colC
val	val	val
xxx	val	val
Key	colB	colC
val	val	val
xxx	val	val

id 1-1000

id 1000-2000

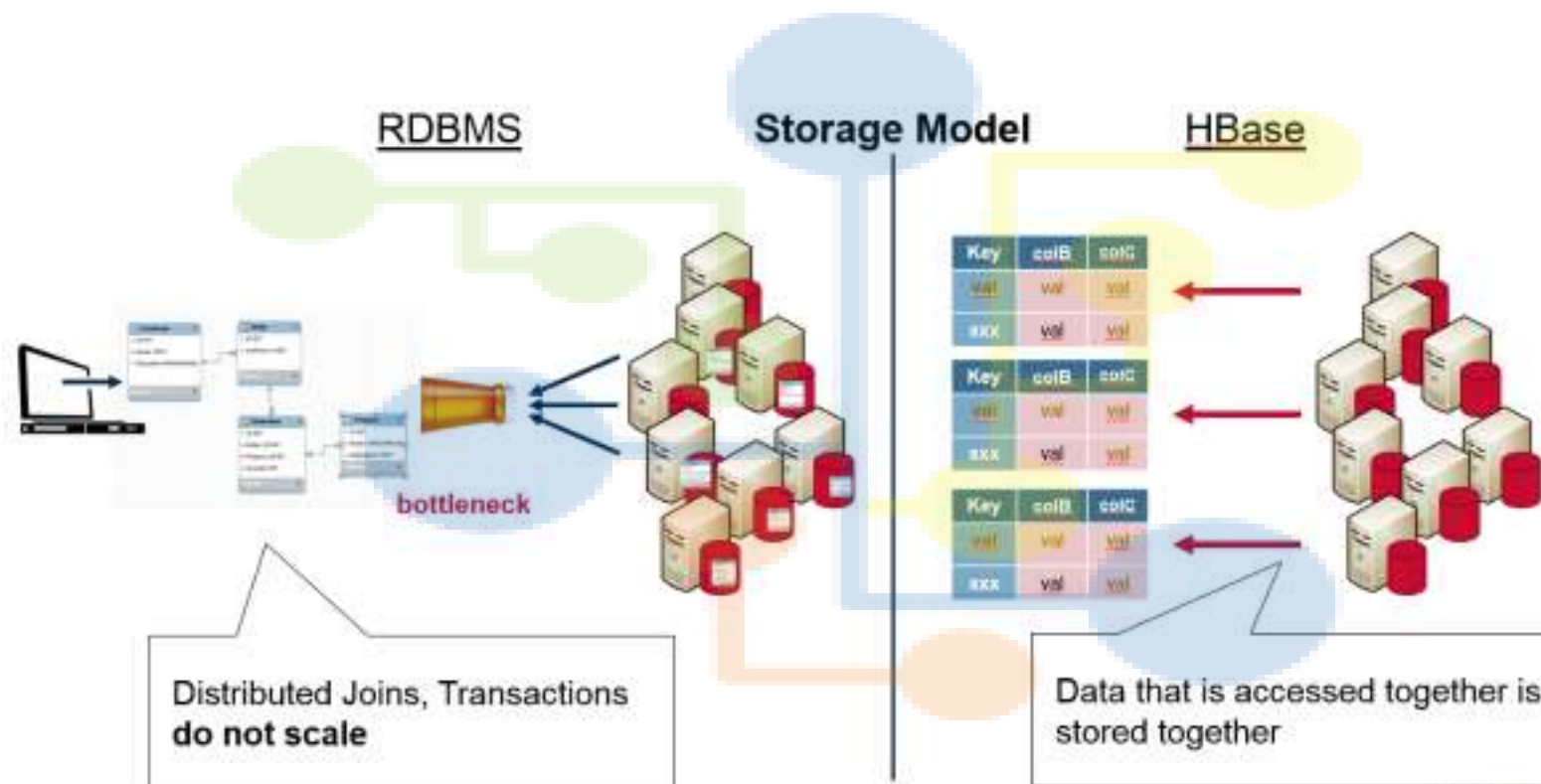
id 2000-3000



Escalabilidade Horizontal

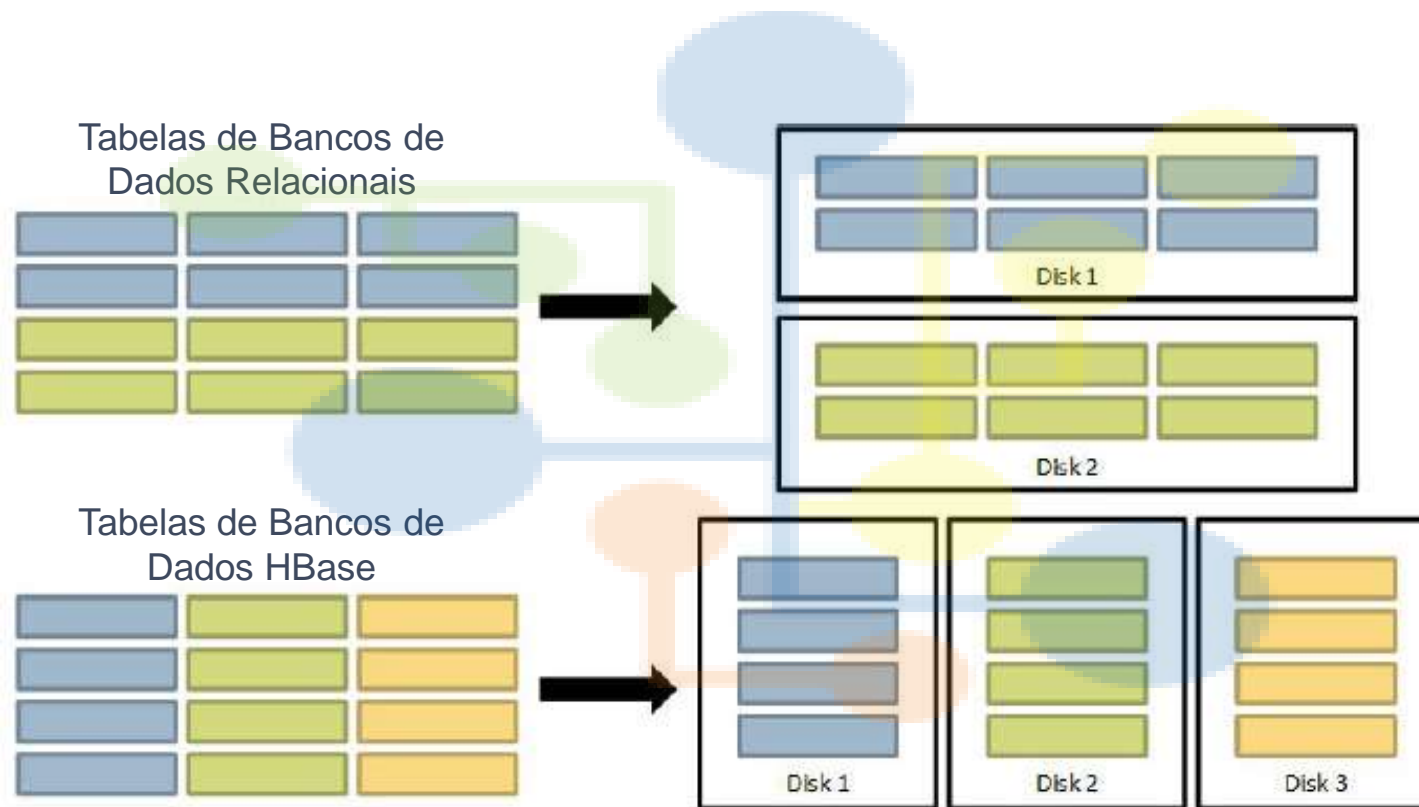


Modelo de Dados do Apache HBase





Modelo de Dados do Apache HBase





Modelo de Dados do Apache HBase

Célula

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

Hfile

Column Families



Modelo de Dados do Apache HBase

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



Modelo de Dados do Apache HBase

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
	Bill Green	Pittsburgh, PA	Desk	\$500.00
104	Jack Black	St. Louis, MO	Bed	\$1600.00

Célula

Hfile

Column Families

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



Modelo de Dados do Apache HBase

- Row Keys: Identificam unicamente um registro

CF = Column Family

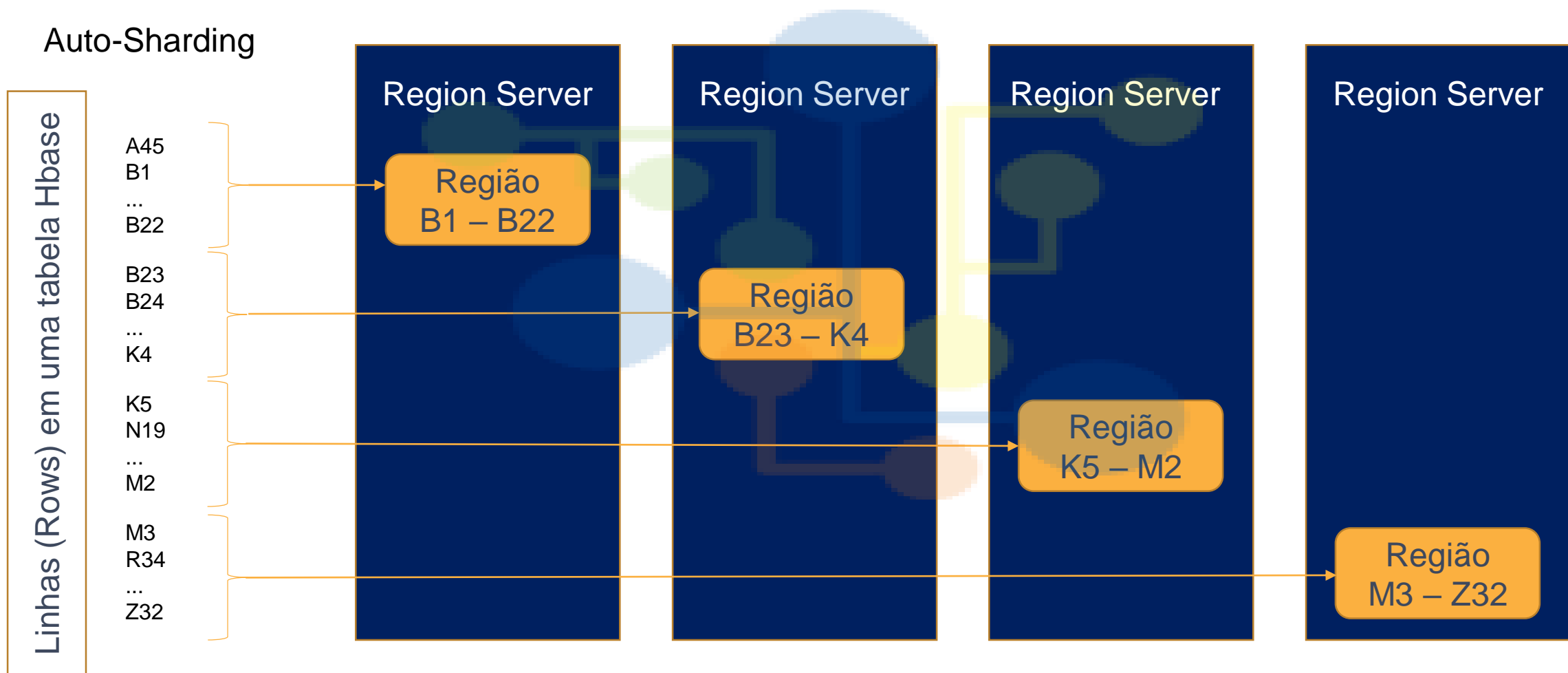
Regiões

	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	



Modelo de Dados do Apache HBase

Auto-Sharding





Orientação a Linha x Orientação a Coluna





Orientação a Linha x Orientação a Coluna

Bancos de Dados relacionais são orientados a linha:

Tabela Clientes

ID Cliente	Nome	Cidade
1	Bob	SP
2	Zico	RJ

Tabela Vendas

ID Cliente	ID Produto	Valor
1	1002	500
2	1008	700

1, Bob, SP, 1002, 500
2, Zico, RJ, 1008, 700



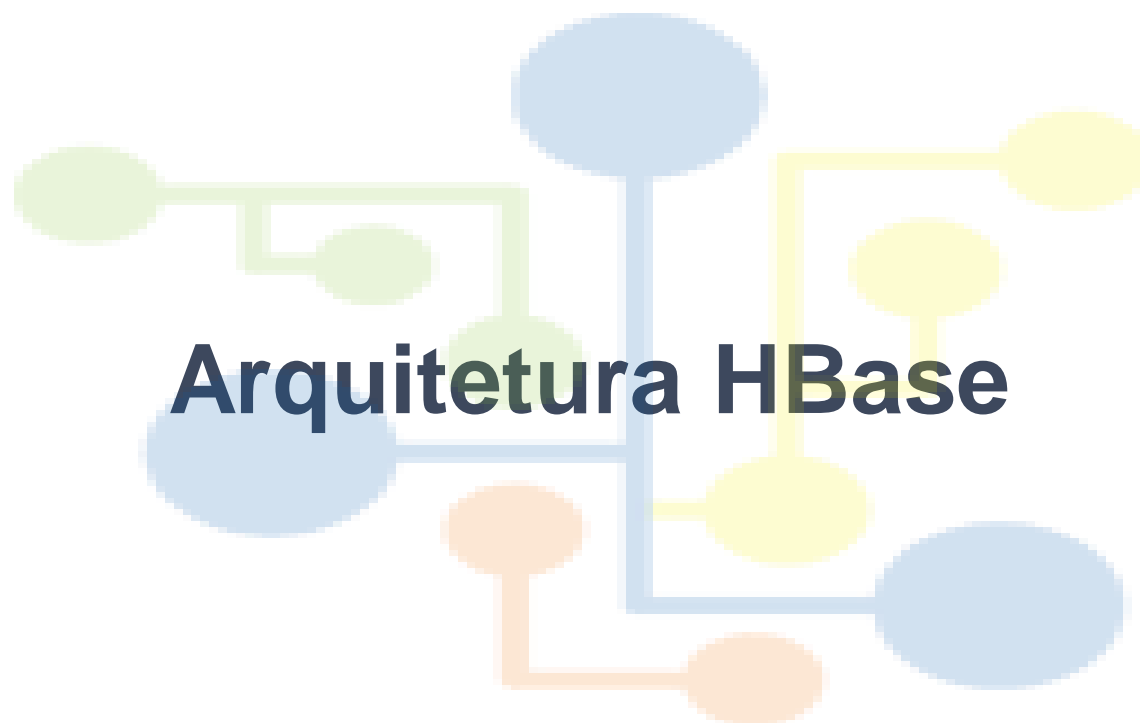
Orientação a Linha x Orientação a Coluna

O HBase é orientado a coluna:

Rowkey	Clientes		Vendas	
ID Cliente	Nome Cliente	Cidade	ID Produto	Valor
1	Bob	SP	1002	500
2	Zico	RJ	1008	700

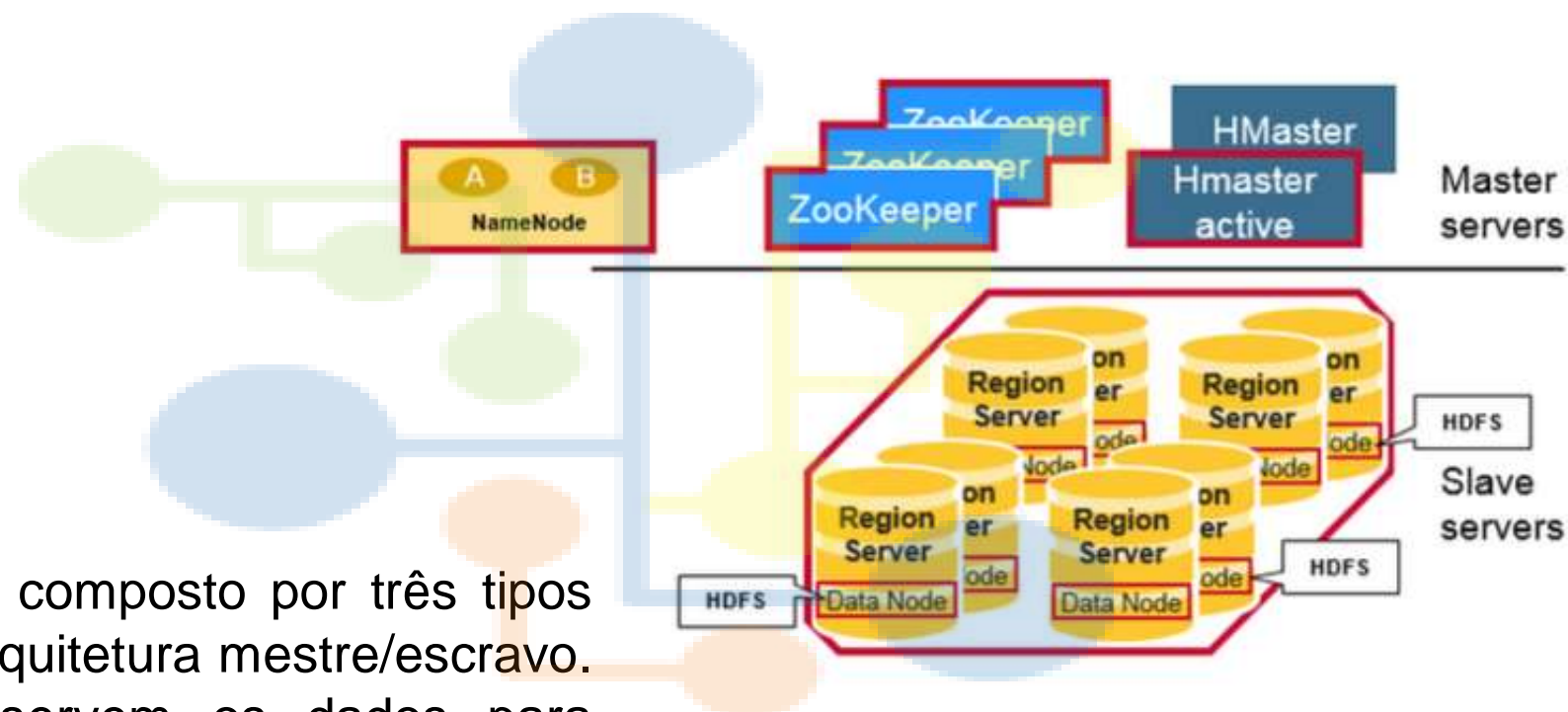


Arquitetura HBase





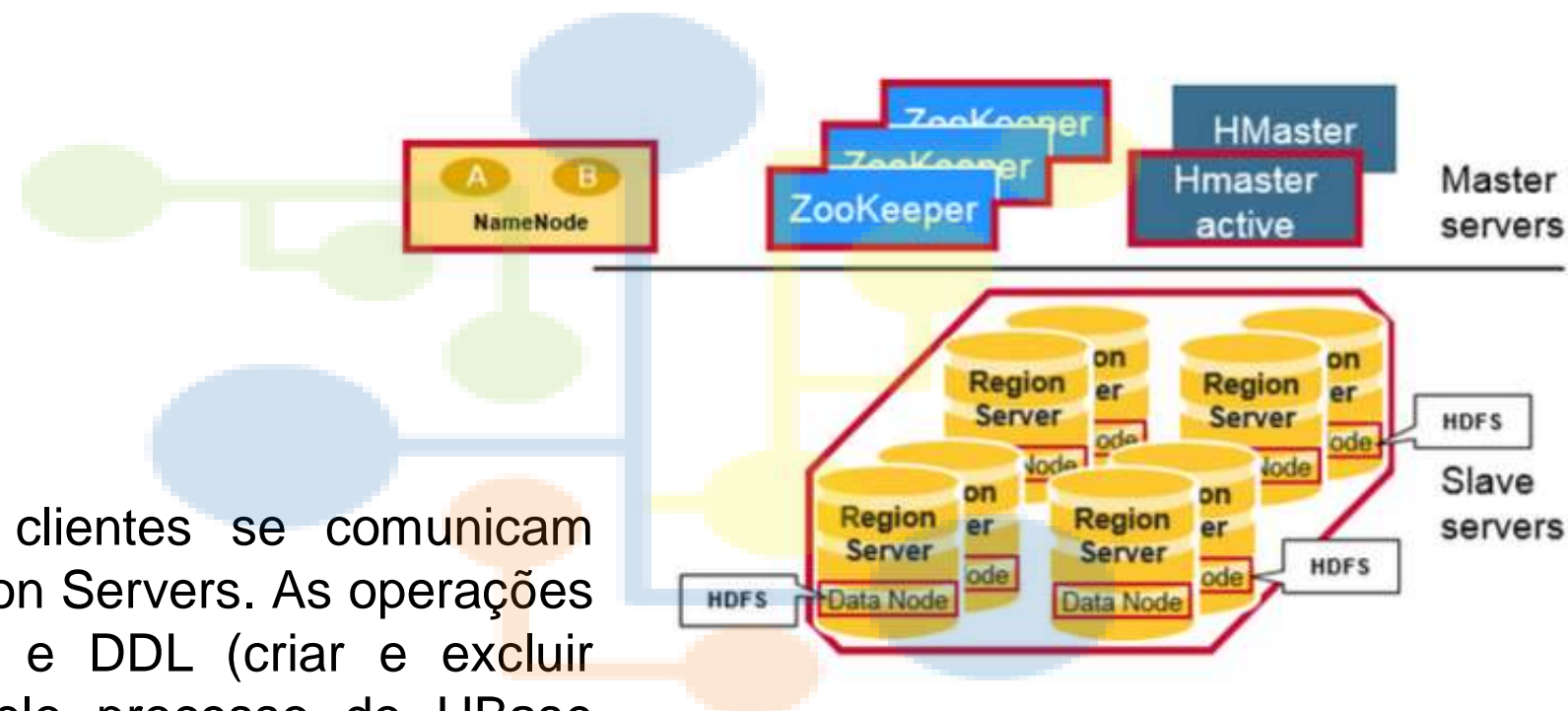
Arquitetura HBase



Fisicamente, o HBase é composto por três tipos de servidores em uma arquitetura mestre/escravo. Servidores de Região servem os dados para leituras e gravações.



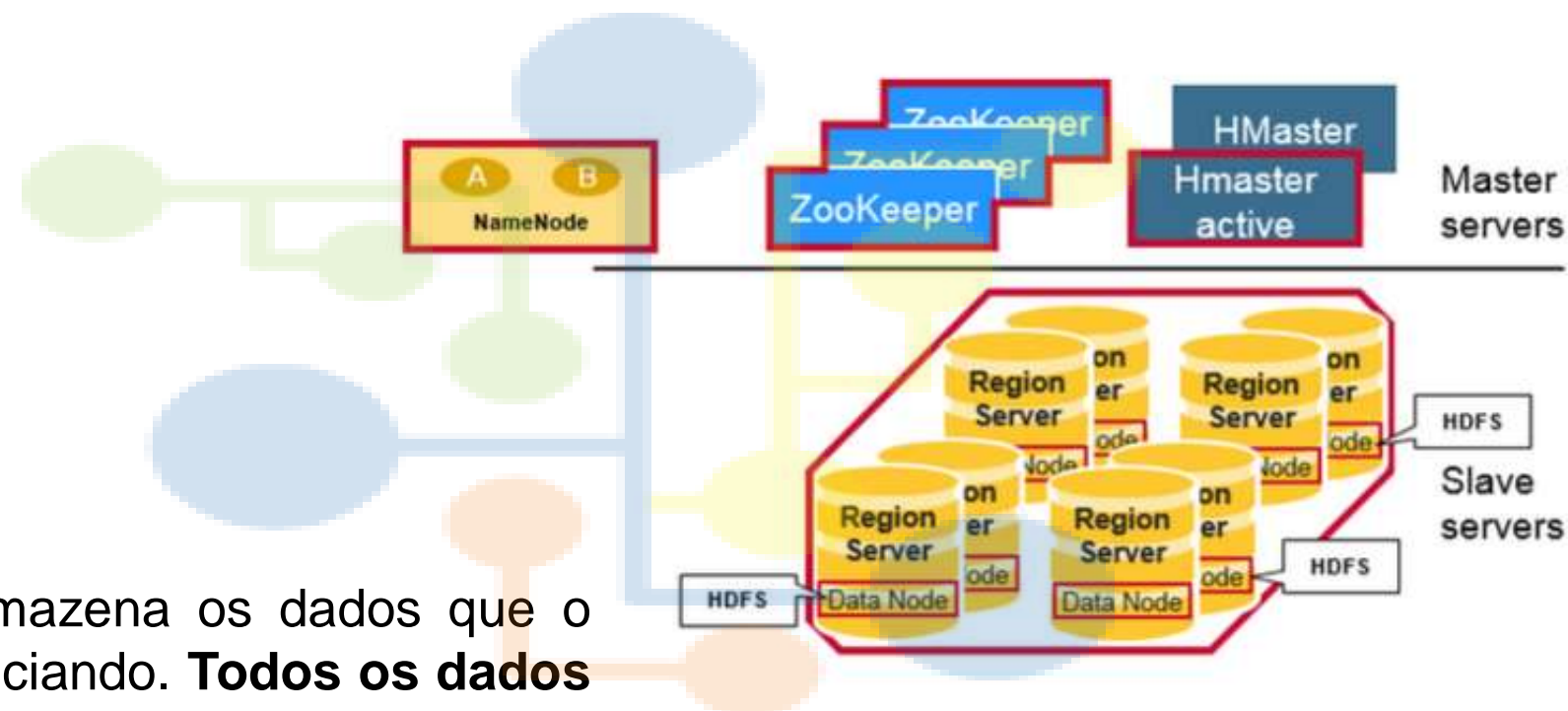
Arquitetura HBase



Ao acessar dados, os clientes se comunicam diretamente com os Region Servers. As operações de atribuição de região e DDL (criar e excluir tabelas) são tratadas pelo processo do HBase Master. O Zookeeper mantém o estado de cluster ativo.



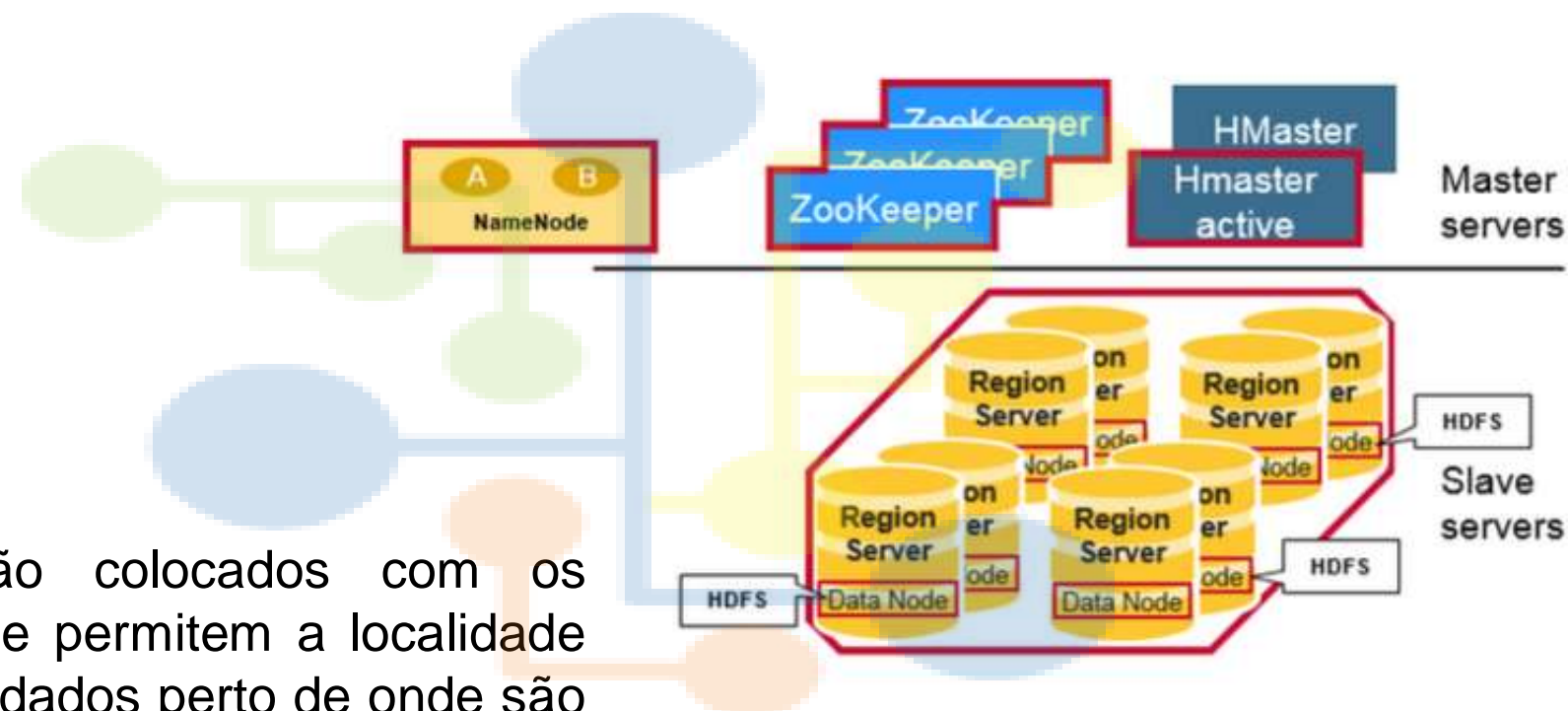
Arquitetura HBase



O Hadoop DataNode armazena os dados que o Region Server está gerenciando. **Todos os dados do HBase são armazenados em arquivos HDFS.**



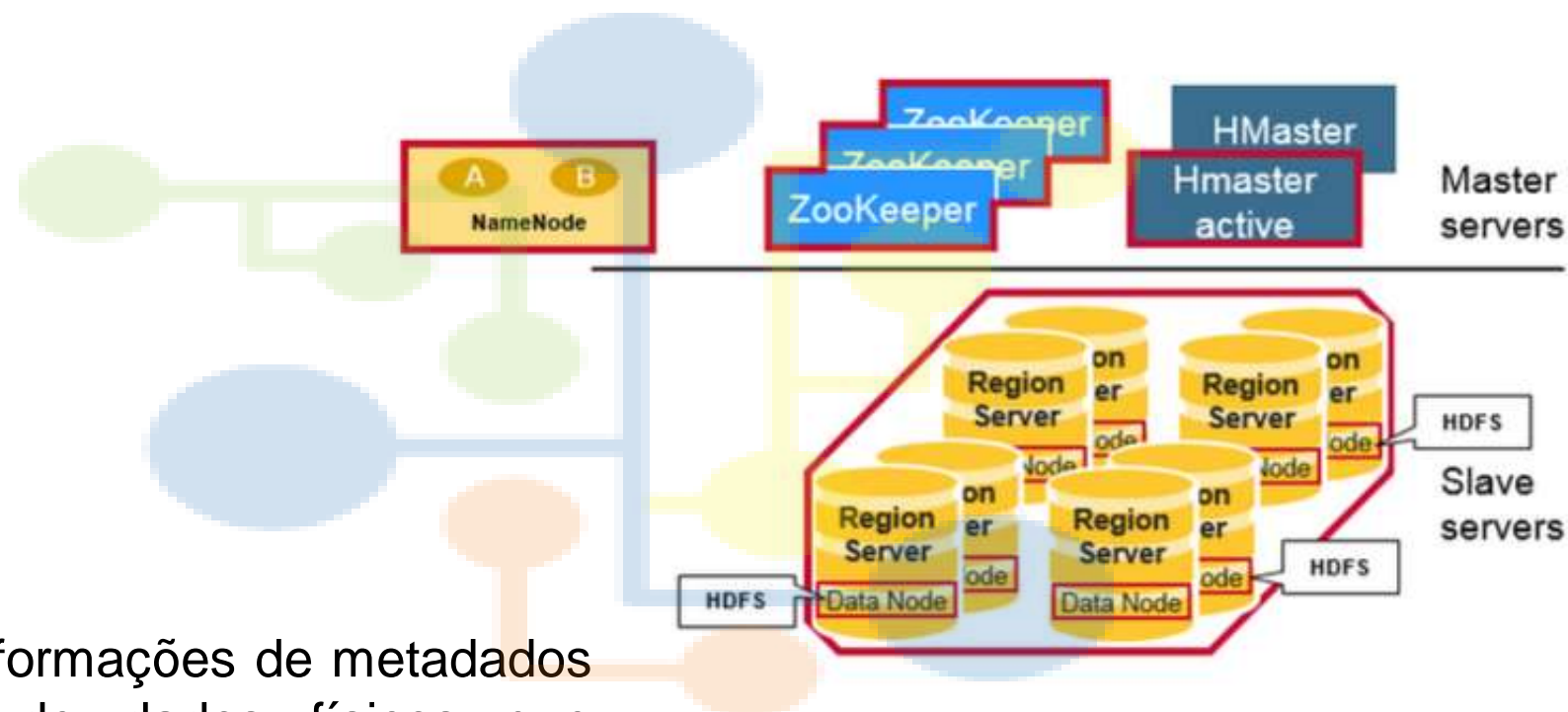
Arquitetura HBase



Os Region Servers são colocados com os DataNodes do HDFS, que permitem a localidade dos dados (colocando os dados perto de onde são necessários) para os dados servidos pelos Region Servers.



Arquitetura HBase

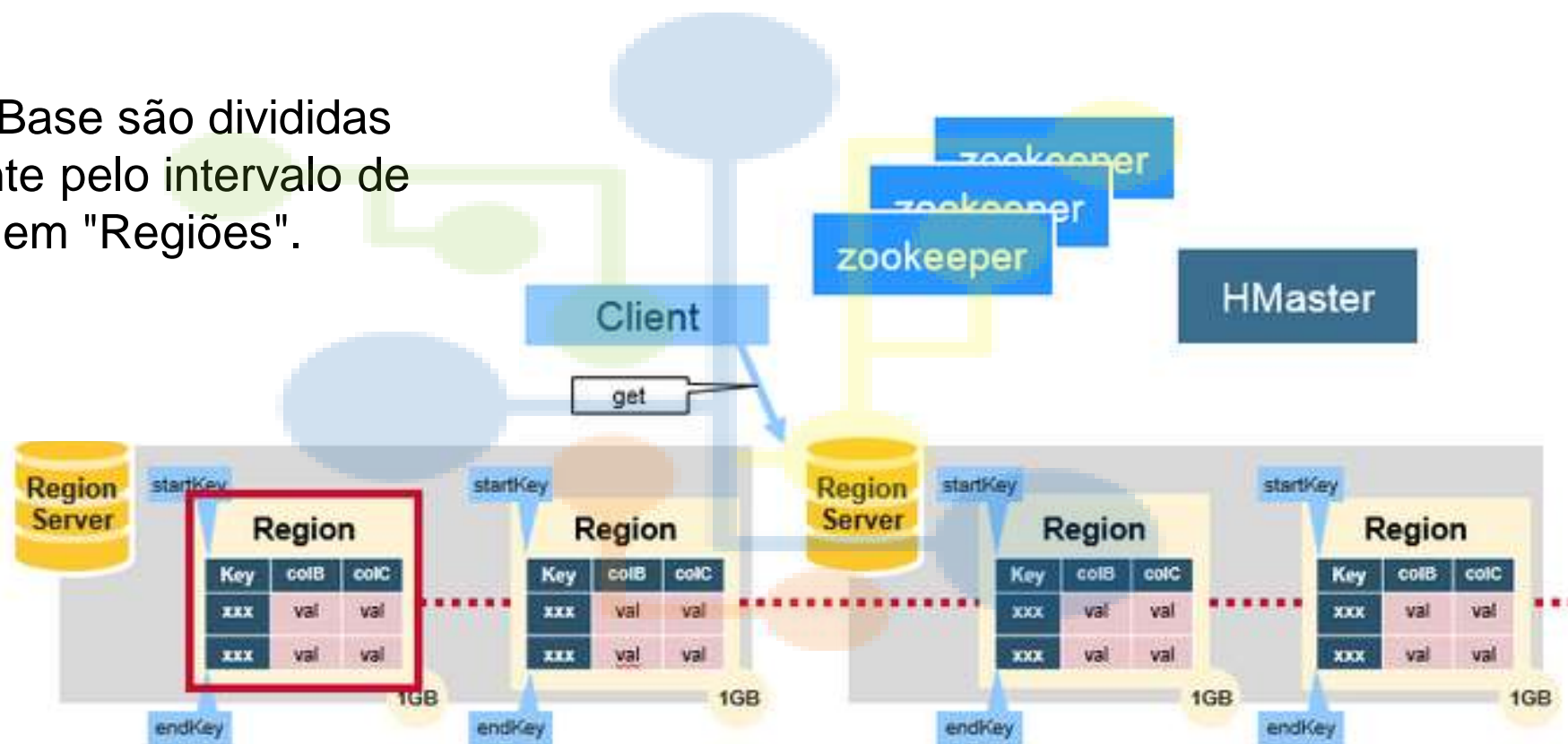


O NameNode mantém informações de metadados para todos os blocos de dados físicos que compõem os arquivos.



Arquitetura HBase

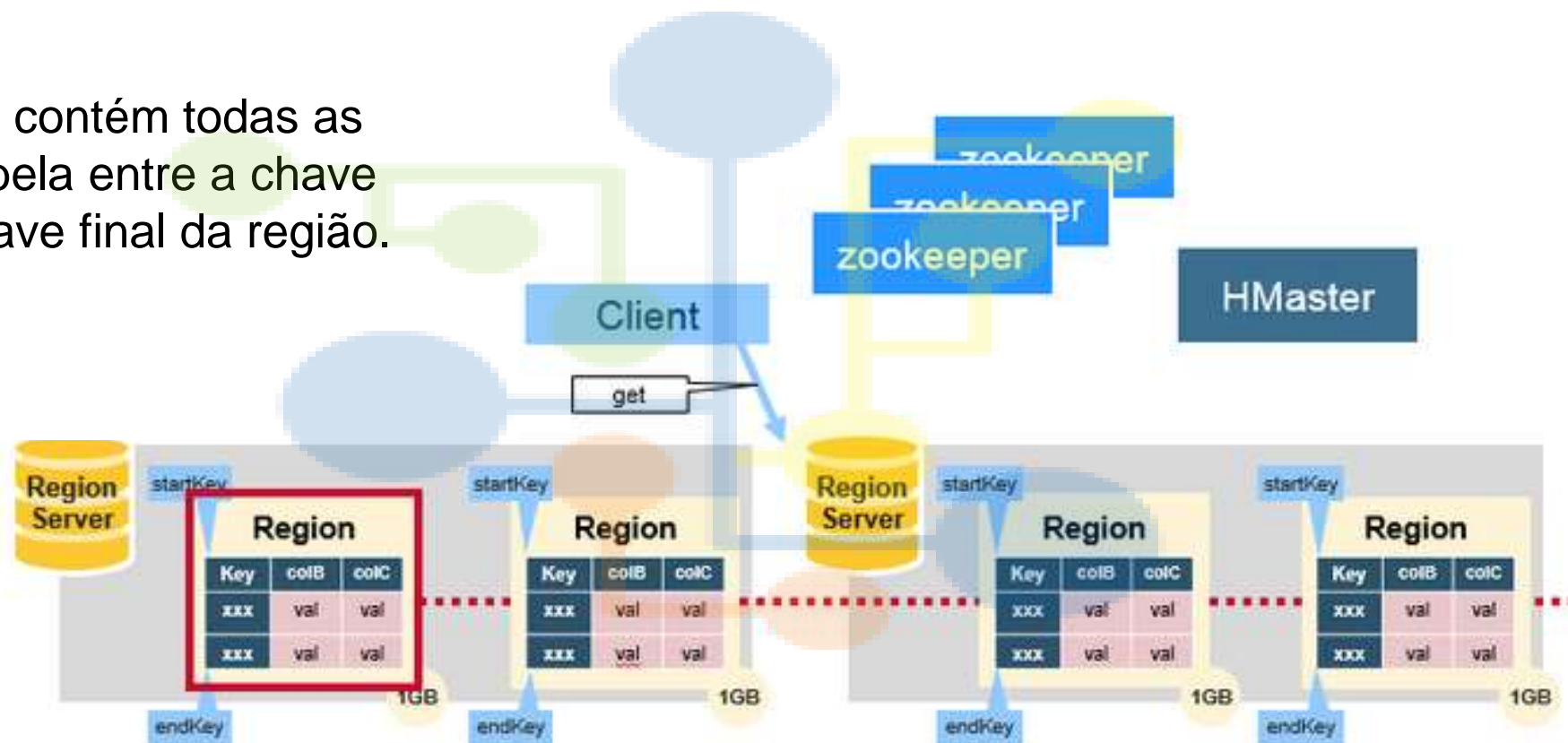
As tabelas HBase são divididas horizontalmente pelo intervalo de rowkeys em "Regiões".





Arquitetura HBase

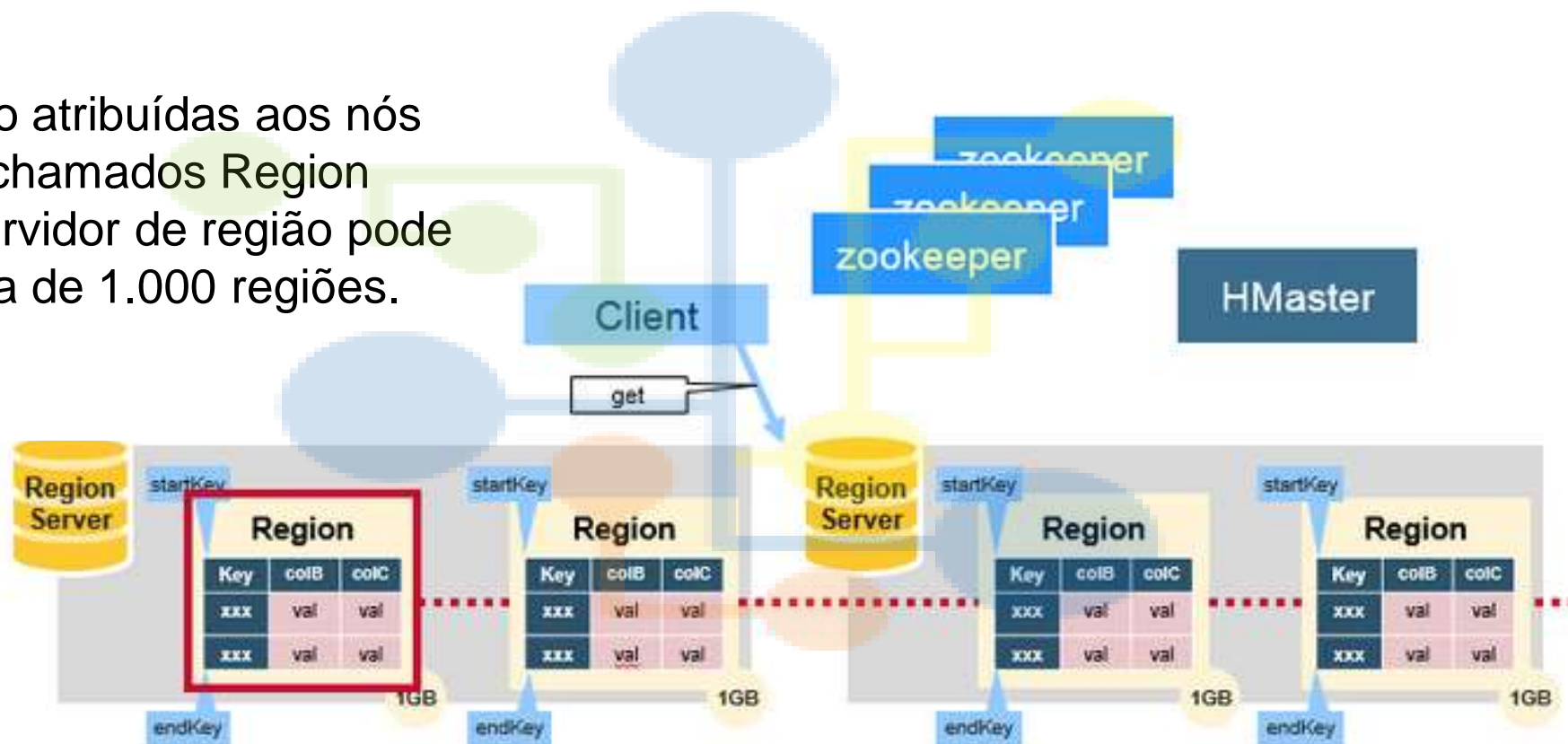
Uma região contém todas as linhas da tabela entre a chave inicial e a chave final da região.





Arquitetura HBase

As regiões são atribuídas aos nós do cluster, chamados Region Servers. Um servidor de região pode atender cerca de 1.000 regiões.





Introdução ao Apache Hive

Apache Hive



Data Science
Academy

Data Science Academy marcelo_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d



O Hive é um framework para soluções de Data Warehousing, que executa no ambiente do Hadoop, construído inicialmente pelo time de desenvolvimento do Facebook em 2007.

Outro ponto que levou ao desenvolvimento do Hive foi o baixo desempenho das soluções de mercado para realizar operações de Full Scan em grandes volumes de dados.





Hive Query Language

SQL → Jobs MapReduce

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



Usamos o Apache Hive quando precisarmos realizar consultas ou manipulações em grandes conjuntos de dados, tais como seleção de registros ou colunas, agregação, sumarização, contagem de elementos, filtros ou atualizações em massa.

Essas tarefas não precisam ser feitas em tempo real e o que queremos é obter insights a partir de grandes conjuntos de dados, Big Data.

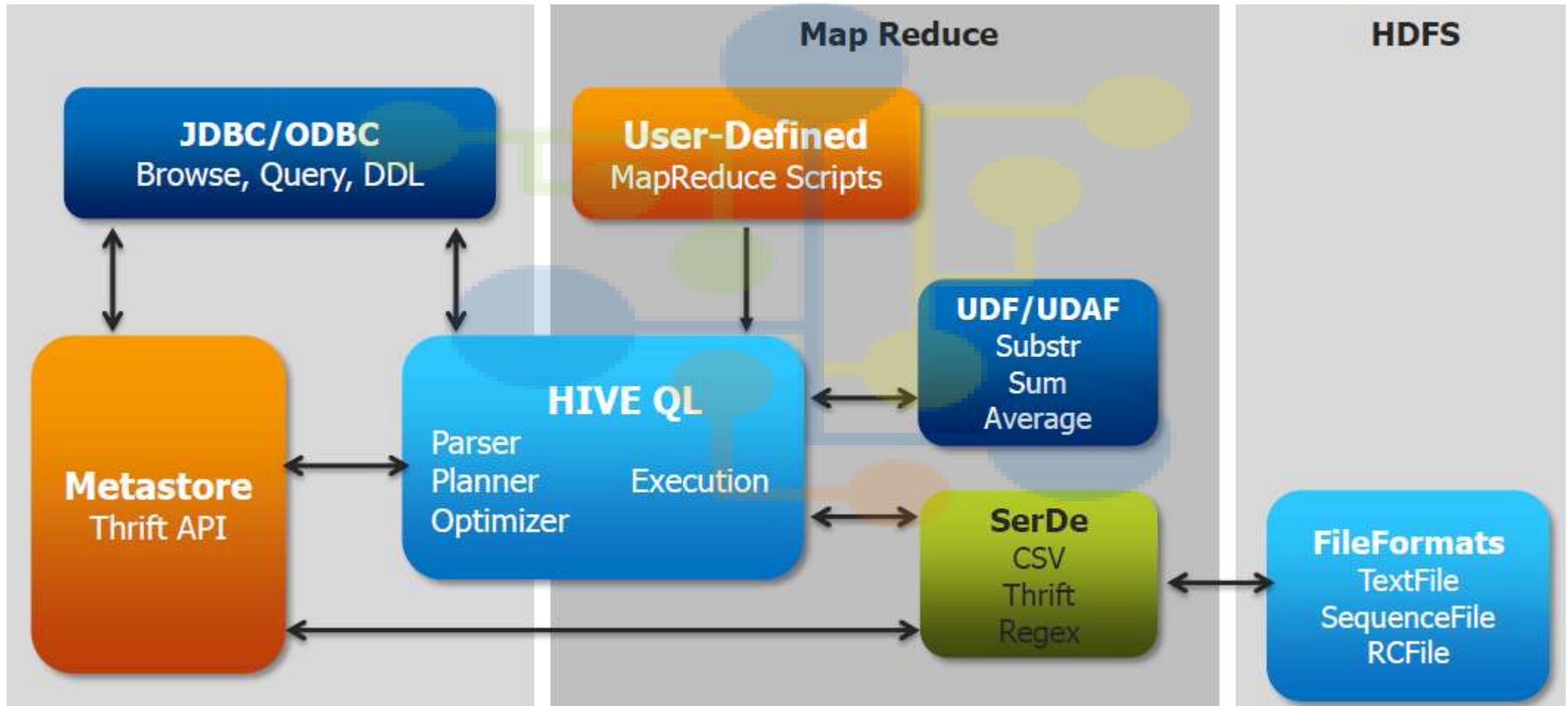


Apache Hive



Data Science
Academy

Data Science Academy marcelo_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

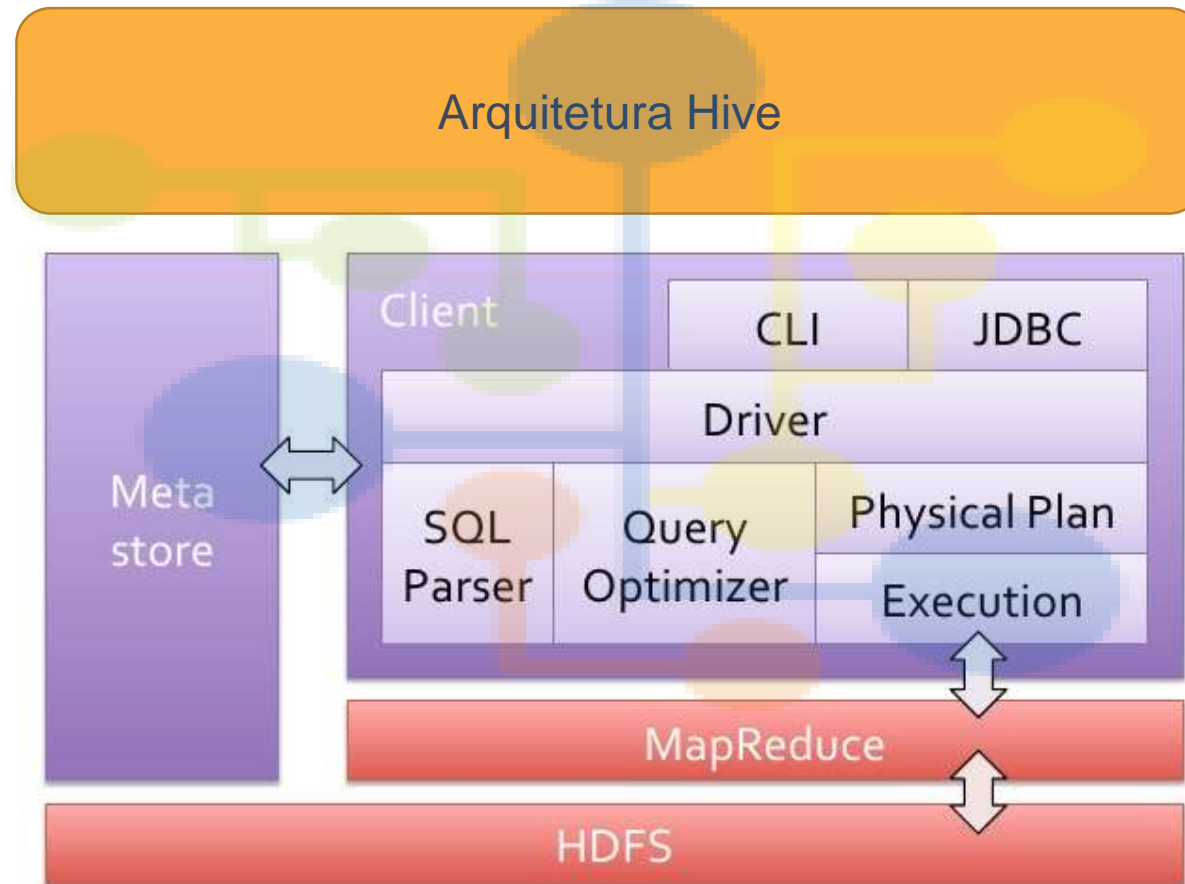


Apache Hive



Data Science
Academy

Data Science Academy marcelo_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

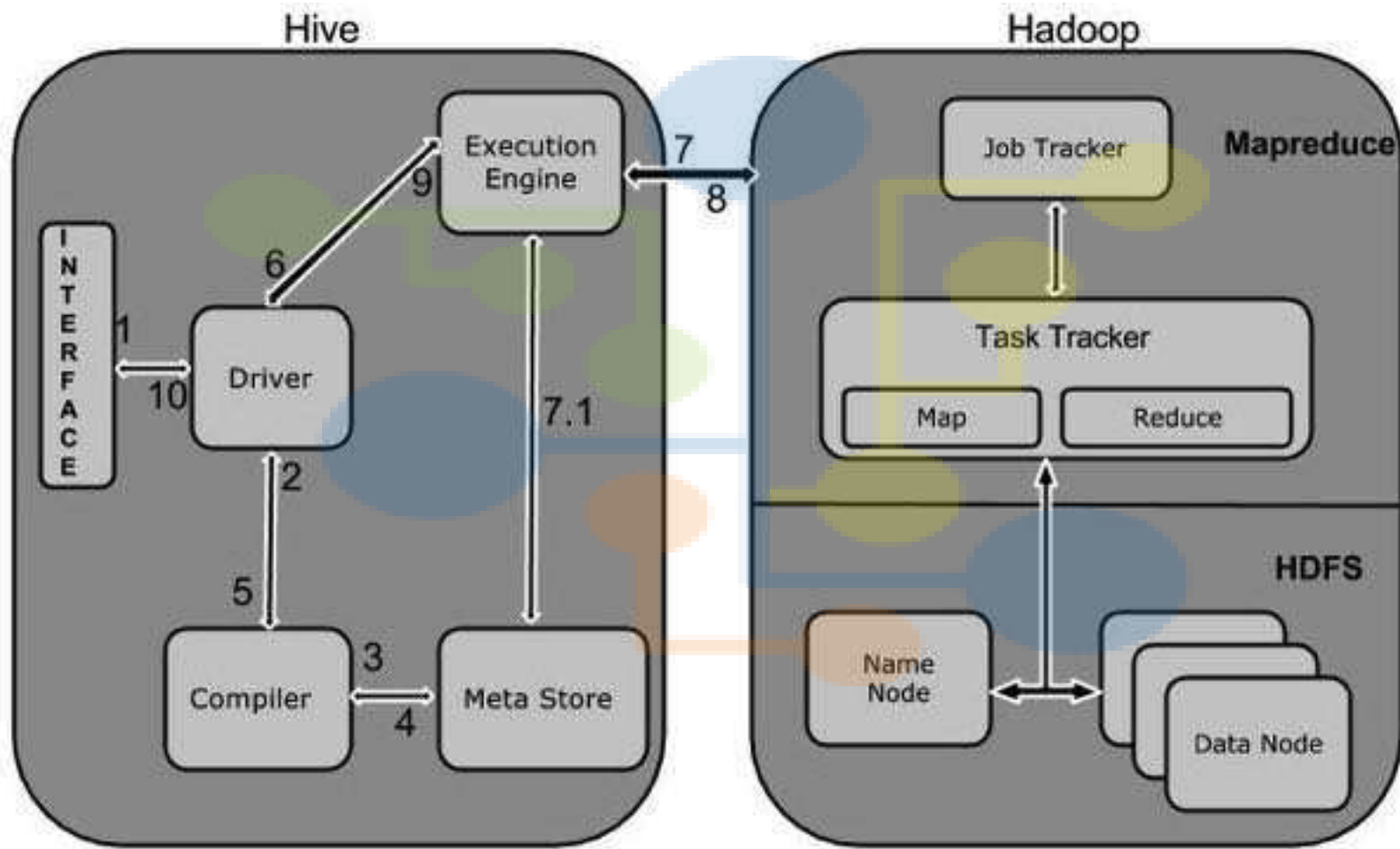


Apache Hive



Data Science
Academy

Data Science Academy marcelo_eidi12@hotmail.com 5d5c42d55e4cde68f38b457d

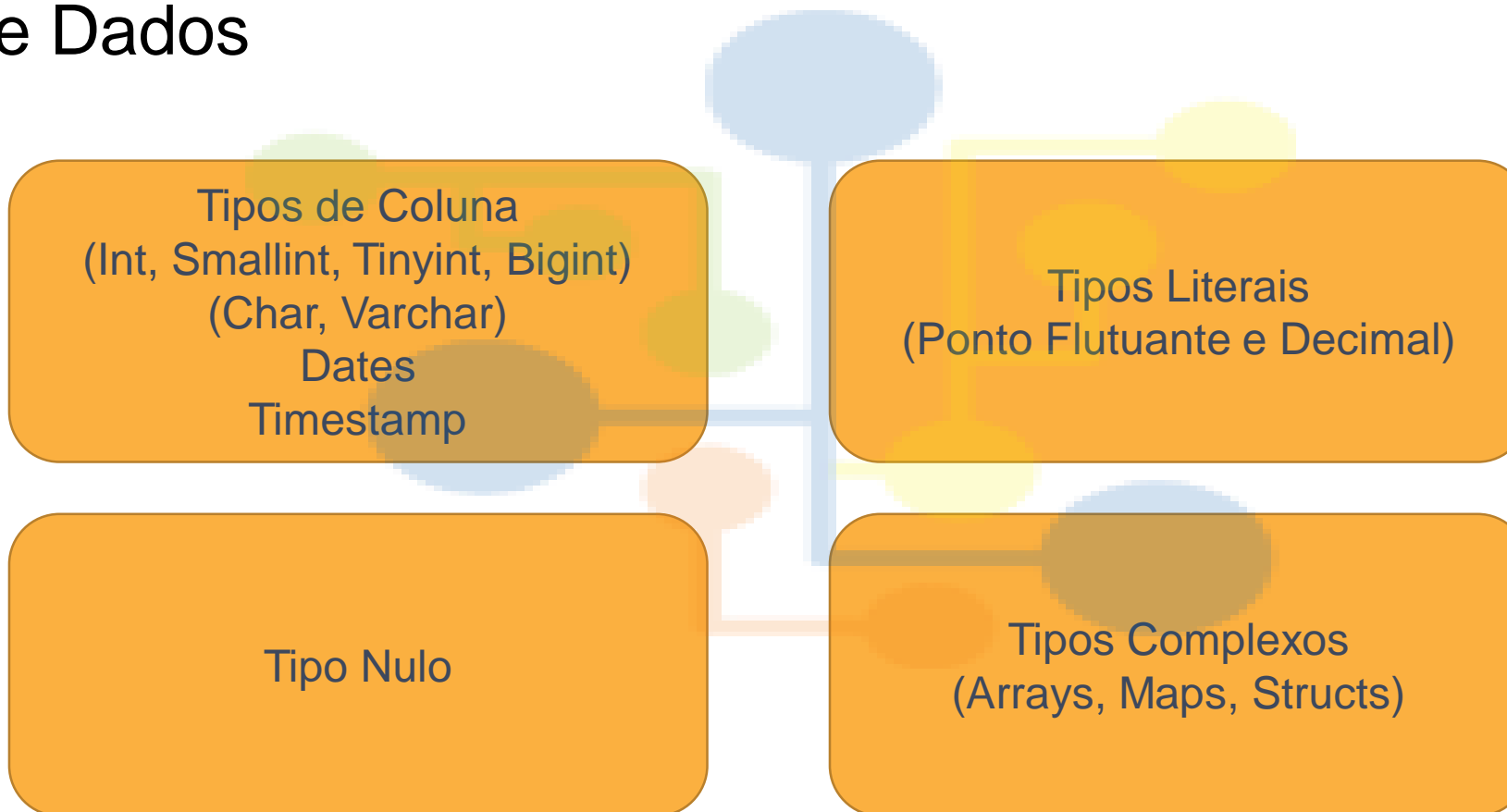




Tipos de Dados

A faint, stylized diagram in the background of the slide. It consists of several circular nodes connected by lines. The nodes are colored in light blue, light green, light yellow, and light orange. The connections form a complex, branching structure that resembles a network or a data flow diagram.

Tipos de Dados





Obrigado

