



Formação Engenheiro de Dados

Hive

Hadoop

- ❖ Arquitetura fantástica para Big Data!
- ❖ Porem extremamente complexo e de baixo nível!
- ❖ Requer escrita e compilação de programa em Java

Hive



DataWarehouse com Hadoop



Suporta grande volume de dados: distribuído



Gera tarefas map-reduce para maioria das consultas



HiveSQL: linguagem de consulta semelhante a SQL



Hive DDL (Hive Data Definition Language)



Suporta consultas interativas



Pode ser operado por:

Linha de comando
Web (Ambar ou Hue)

Hive



Dados estruturados



Requer schema



Como um
datawarehouse, feito
para inserção e leitura,
não atualização

Dividido em
clusters -
atualização em
vários nós

Metastore

- ❖ Core do Hive
- ❖ Mantem metadados
- ❖ Padrão armazenado em banco de dados derby

- ❖ Técnicas de otimização
 - ❖ Desnormalização
 - ❖ Bucketing
 - ❖ Partition

Arquitetura

- ❖ Banco de Dados
- ❖ Tabelas
- ❖ Partição: equivalente a coluna, podendo dividir dados horizontalmente
- ❖ Bucket: partição vertical dos dados baseados em hashs
- ❖ Fisicamente no HDFS:
 - ❖ /user/hive/warehouse/

Modos



Local: Um nó Hadoop,
pequeno volume de dados



MapReduce: Vários nós, maior
volume de dados

Shell

- ◊ beeline
- ◊ hive (descontinuado)

Outras Interfaces

The screenshot shows the Hue - Editor interface for Cloudera. The top navigation bar includes links for Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. The main area has tabs for Query, Hive, and a search bar. On the left, there's a sidebar with icons for Tables, Edits, and a Recent section. The central pane displays a Hive query:

```
dc.cliente,
dc.estado,
dc.sexo,
fv.quantidade,
fv.valorunitario,
fv.valortotal,
fv.desconto,
dp.produto,
dt.data,
dt.dia,
dt.mes,
dt.ano,
dt.trimestre,
dv.nome
from dimensaocliente dc
join fatovendas fv on fv.chavecliente = dc.chavecliente
join dimensao produto dp on dp.chaveproduto = fv.chaveproduto
join dimensao tempo dt on dt.chavetempo = fv.chavetempo
join dimensao vendedor dv on dv.chavevendedor = fv.chavevendor
```

Below the query, there are tabs for Query History, Saved Queries, and Results (16). The results table has columns: dc.cliente, dc.estado, dc.sexo, fv.quantidade, fv.valorunitario, fv.valortotal, fv.desconto, and dp.produto. The results are as follows:

	dc.cliente	dc.estado	dc.sexo	fv.quantidade	fv.valorunitario	fv.valortotal	fv.desconto	dp.produto
1	Adelino Gago	RJ	M	1	7820.85009765625	7820.85009765625	0	Bicicleta Aro 29 Mountain Bike Endorphine 6.3 - 24 Ma
2	Alberto Cezimbra	AM	M	1	97.75	97.75	0.98000001907348633	Bicicleta Aro 29 Mountain Bike Endorphine 6.3 - 24 Ma
3	Alberto Monsanto	RN	M	1	135	135	1.3500000238418579	Bicicleta Aro 29 Mountain Bike Endorphine 6.3 - 24 Ma
4	Adelino Gago	RJ	M	1	7820.85009765625	7820.85009765625	0	Bicicleta Gometws Endorphine 7.3 - Shimano Aluminí
5	Adriana Guedelha	RO	F	2	2955	5910	59.099998474121094	Bicicleta Gometws Endorphine 7.3 - Shimano Aluminí
6	Alberto Cezimbra	AM	M	1	97.75	97.75	0.98000001907348633	Bicicleta Gometws Endorphine 7.3 - Shimano Aluminí
7	Adelino Gago	RJ	M	1	7820.85009765625	7820.85009765625	78.209999084472656	Bicicleta Gometws Endorphine 6.1 Shimano Aluminí

Tipos de dados

Tipos

TINYINT, SMALLINT, INT, BIGINT	Inteiro	1,2,4 e 8 bytes
FLOAT, DOUBLE, DECIMAL	Ponto Flutuante	4, 8 DECIMAL; Customizado
CHAR	Texto	255
VARCHAR	Texto	1 até 65355
STRING	Texto	Customizado
Array	Vetor	
Map		
Date	Data	YYYY-MM-DD
Timestamp	Data e hora	formato Unix padrão

Tipos de Tabela

- ◊ Internal
 - ◊ Fortemente acoplada
 - ◊ Define-se o schema e se carregam os dados (data on schema)
 - ◊ Armazenado em /user/hive/warehouse
 - ◊ Excluir a tabela, exclui dados e schema
 - ◊ Utilizado, para dados locais
 - ◊ Move os dados para o warehouse do hive
- ◊ External
 - ◊ Fracamente acoplada
 - ◊ Schema on data