



INSTITUTO FEDERAL
Sudeste de Minas Gerais



Processamento Analítico de Dados e Hive

Curso: Tecnologia em Gestão da Tecnologia da Informação

Prof.: Jean Henrique de Sousa Câmara

Contato: jean.camara@ifsudestemg.edu.br

Big Data

“*Business Intelligence*”

2

- Grosseiramente falando, trata-se da exploração pelas organizações dos dados que resultam do seu próprio negócio para gerar conhecimento em benefício da própria organização
- Aplicações
 - Análise de mercado
 - Planejamento estratégico
 - Tomada de decisão
- Ferramentas
 - Relatórios, dashboards, mineração de dados, etc...

Dados Transacionais vs. Analíticos

3

- OLTP
 - Online Transaction Processing
- OLAP
 - Online Analytical Processing

Dados Transacionais vs. Analíticos

4

	OLTP	OLAP
Aplicações Típicas	E-commerce, operações bancárias, reservas aéreas, matrículas	Planejamento estratégico, Geração de modelos
Cenário	Atendimento a usuários	Retarguarda
Requisitos	Tempo real; baixa latência, alta concorrência	Cargas em batch; menos concorrência
Tarefas	Conjunto relativamente pequeno operações transacionais	Consultas analíticas complexas, muitas vezes ad-hoc
Padrão de Acesso	Leituras, escritas e atualizações randômicas; Volumes de dados relativamente pequenos	Table scans; Grandes volumes de dados

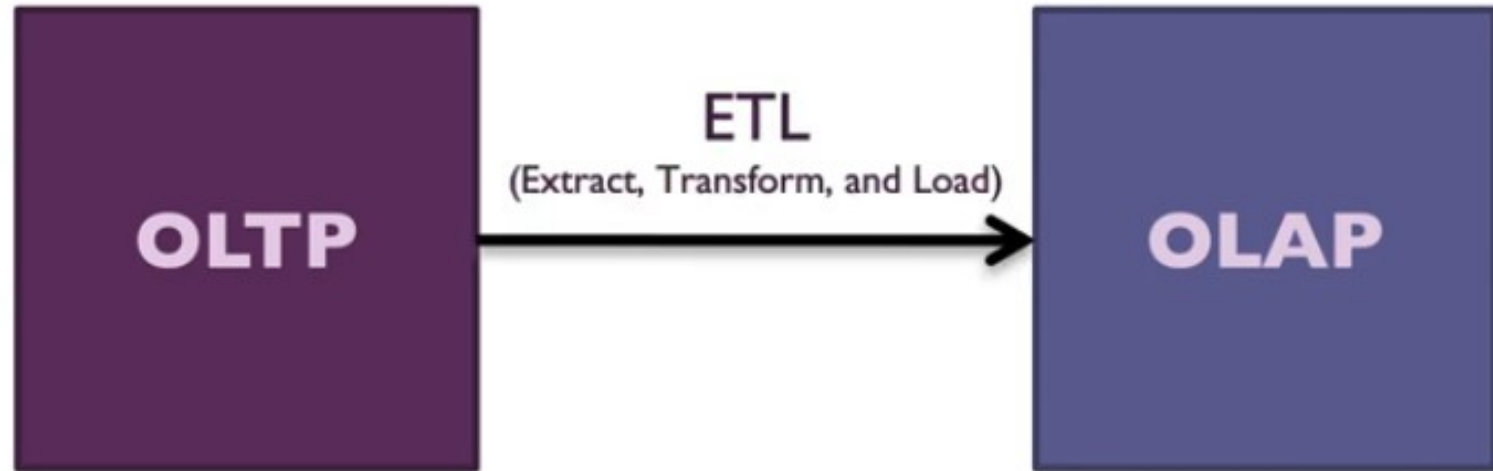
Dados Transacionais vs. Analíticos

5

- Problemas na coexistência de OLTP e OLAP
 - Gerência de memória
 - Padrões de acesso conflitantes
 - Latência variável
 - Aumento no nível de concorrência no sistema
- Solução: BDs separados
 - BD OLTP para atender grandes volumes de transações de usuários
 - BD OLAP: Data Warehouse

Arquitetura OLTP/OLAP

6



- ETL
 - Extração dos registros do BD OLTP
 - Transformação
 - Limpeza, verificação de consistência, agregação, etc
 - Carga no BD OLAP

Exemplos - OLAP

7

➤ BD

- Vendas(cod, data, loja, preco)
- Carros(placa, modelo, cor)
- Lojas(nome, cidade, estado, telefone)

➤ Consultas

- `SELECT SUM(preco) FROM Vendas;`
- `SELECT loja, AVG(preco) FROM Vendas GROUP BY loja;`
- `SELECT estado, AVG(preco) FROM Vendas JOIN Lojas
ON loja = nome WHERE data > '2020-09-11' GROUP BY
estado`

Gargalo ETL

8

The Facebook logo, consisting of the word "facebook" in a white, lowercase, sans-serif font, followed by a registered trademark symbol (®). The logo is centered on a dark blue rectangular background.

Jeff Hammerbacher, Information Platforms and the Rise of the Data Scientist.
In, *Beautiful Data*, O'Reilly, 2009.

“On the first day of logging the Facebook clickstream, more than 400 gigabytes of data was collected. The load, index, and aggregation processes for this data set really taxed the Oracle data warehouse. Even after significant tuning, we were unable to aggregate a day of clickstream data in less than 24 hours.”

Gargalo ETL

9

- ETL é tipicamente realizado a noite
 - O que acontece se leva mais de 24 horas para processar 24 horas de dados?
- Uma solução: Plataformas de Big Data (MR/Hadoop, Spark, etc...)
 - Ingestão de dados é limitada pela velocidade do sistema de arquivos
 - Escala com mais nodos
 - Massivamente paralelo
 - Permite o uso de qualquer ferramenta de processamento
 - Mais barato que SGBDs paralelos
 - ETL exige processamento em batch de qualquer forma

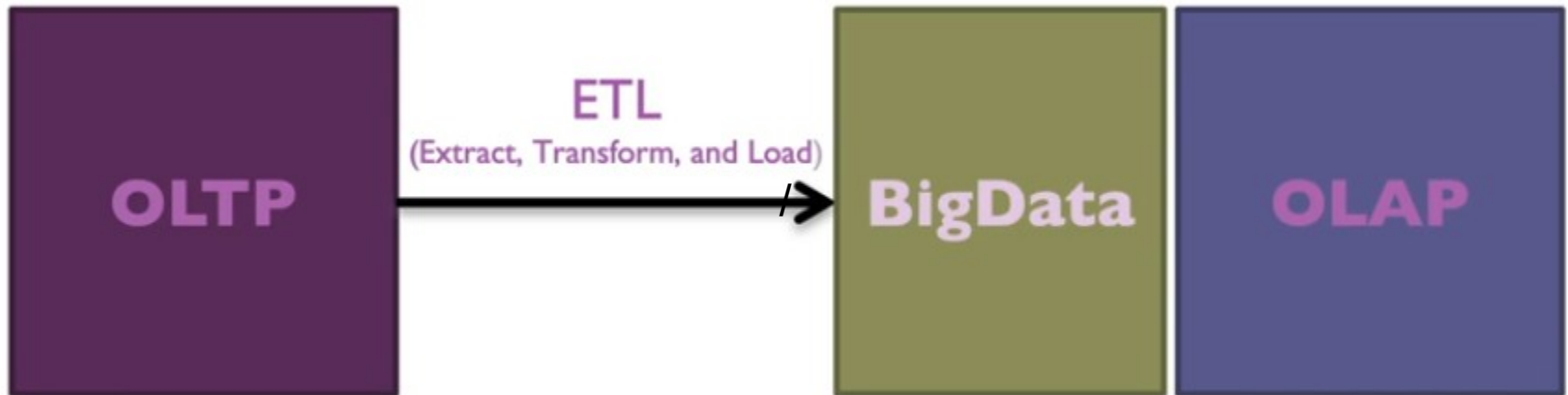
O que mudou?

10

- Queda no preço dos discos: É mais barato armazenar tudo do que tentar descobrir o que descartar
- Tipos de dados coletados: Variam de dados que obviamente são importantes a dados cujo valor é menos aparente
- Surgimento das mídias sociais e conteúdo criado por usuário: grande aumento no volume de dados
- Crescente maturidade das técnicas de mineração de dados: Demonstra o valor da análise de dados

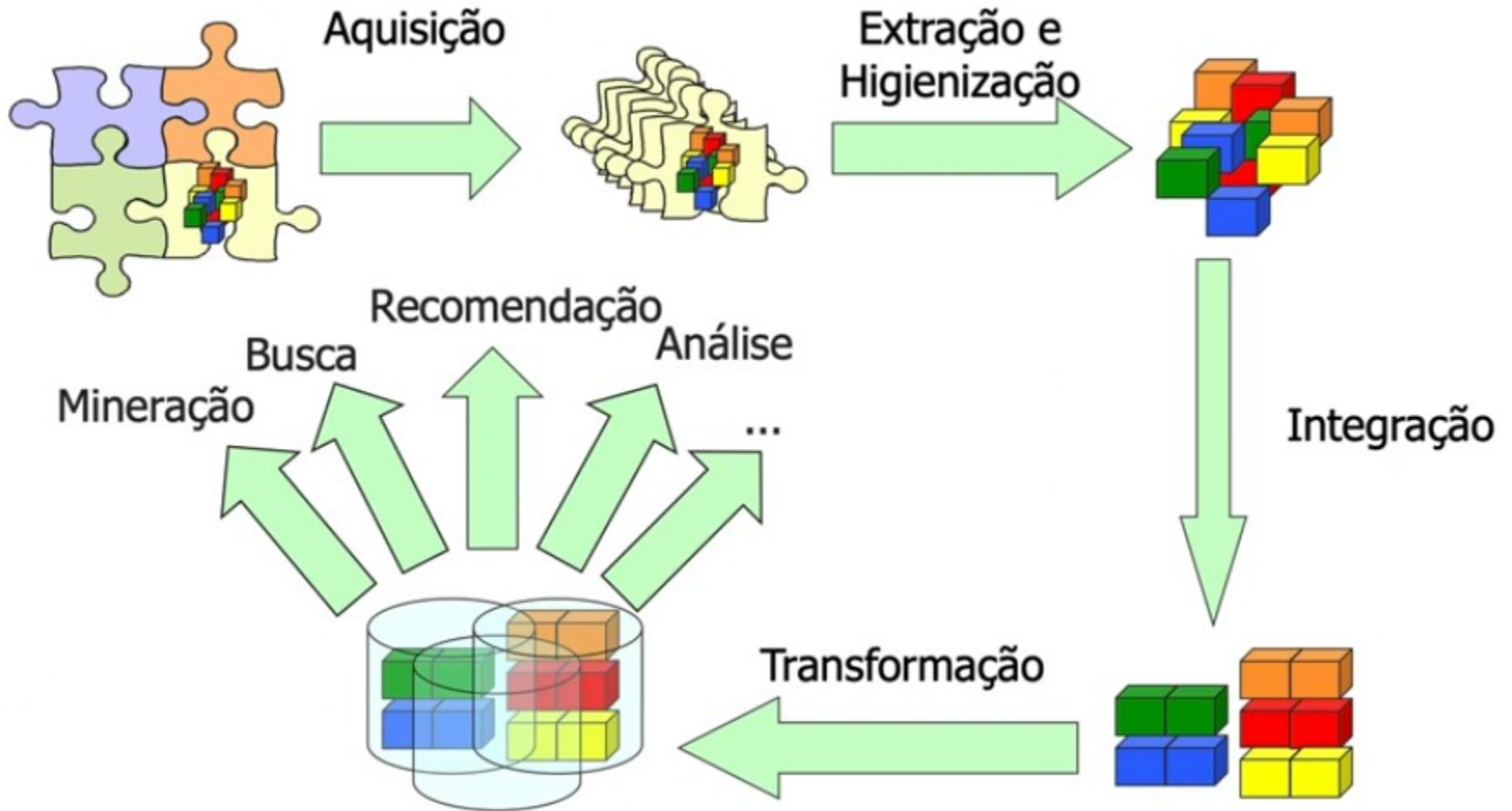
Arquitetura OLTP/OLAP/Big Data

11



O Pipeline do BigData

12



Aquisição

13

- Coleta dos dados em suas fontes originais: páginas web, redes sociais, sensores...
- Requisito: filtrar fontes relevantes para a aplicação reduzindo o ruído
 - Reduzir o volume em ordem de magnitude
 - Facilitar o trabalho dos outros serviços
- Exemplo: “Web Crawling as a service”

Extração e Higienização

14

- Extrair dados de interesse em meio a outros dados não relevantes
 - Extrair entidades nomeadas em notícias
 - Posts de redes sociais
 - Extrair componentes específicos de sinais de sensores
- Bases de conhecimento, ontologias...

Integração

15

- Organização dos dados extraídos de acordo com a semântica própria da aplicação
 - Deduplicação
 - Pareamento (linkage)
 - Fusão
 - Classificações...

Transformação

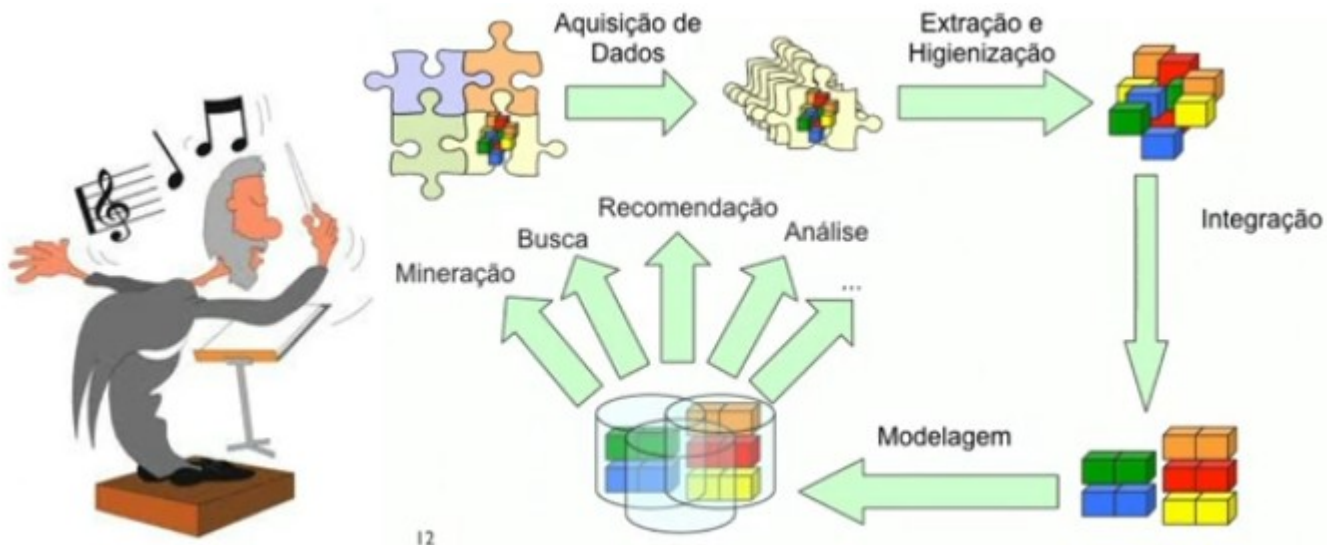
16

- Disponibilizar “visões” dos dados que atendem a necessidades específicas da análise

Orquestração

17

- Deve ser possível orquestrar os serviços usando workflows flexíveis, pois
 - Nem todos os serviços são necessários
 - A ordem pode não ser exatamente a mesma
 - Podem haver loops, condicionantes...



Vantagens do OLAP com Big Data

18

- Redução de tempo, custos, falhas, insegurança...
- Compartilhamento de dados nas várias fases do pipeline
- Melhorias dos serviços de forma transparente para as aplicações
 - Ex.: modelos probabilísticos podem melhorar continuamente quando expostos a mais instâncias de treino
- Serviços integrados reduzem o tempo com fluxo de dados massivos
 - Vários destes serviços existem de forma isolada

Hive

Hive

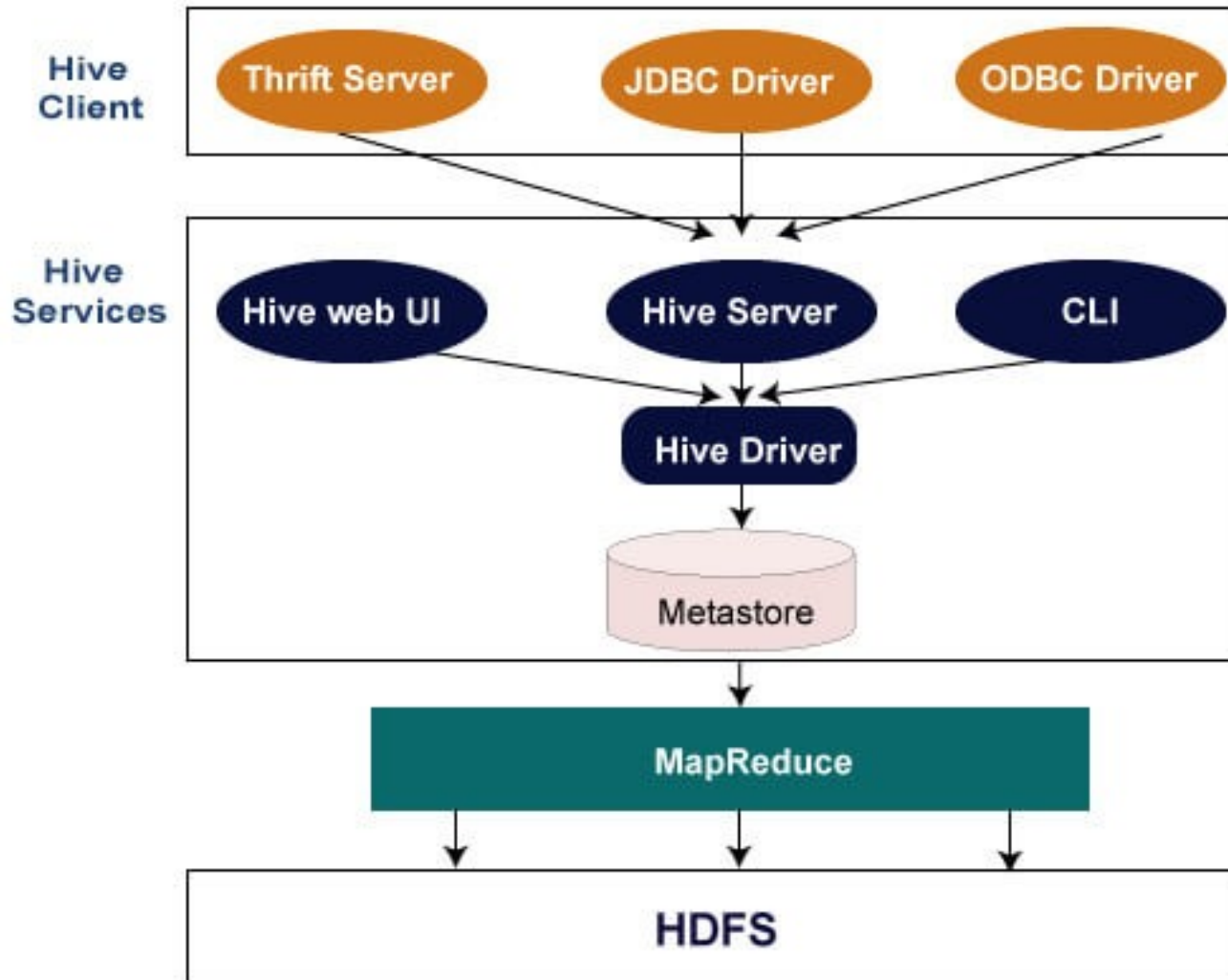
20

- Armazém de dados distribuído sobre Map Reduce/HDFS
- Características de banco de dados relacionais: bancos de dados, tabelas, views, funções e SQL
- Não implementa funções não analíticas (ex.: PK, FK)
- Open Source
- Mantido pela Fundação Apache
- Analítico, não transacional



Hive - Arquitetura

21



Hive - Tabelas

22

➤ **Externas**

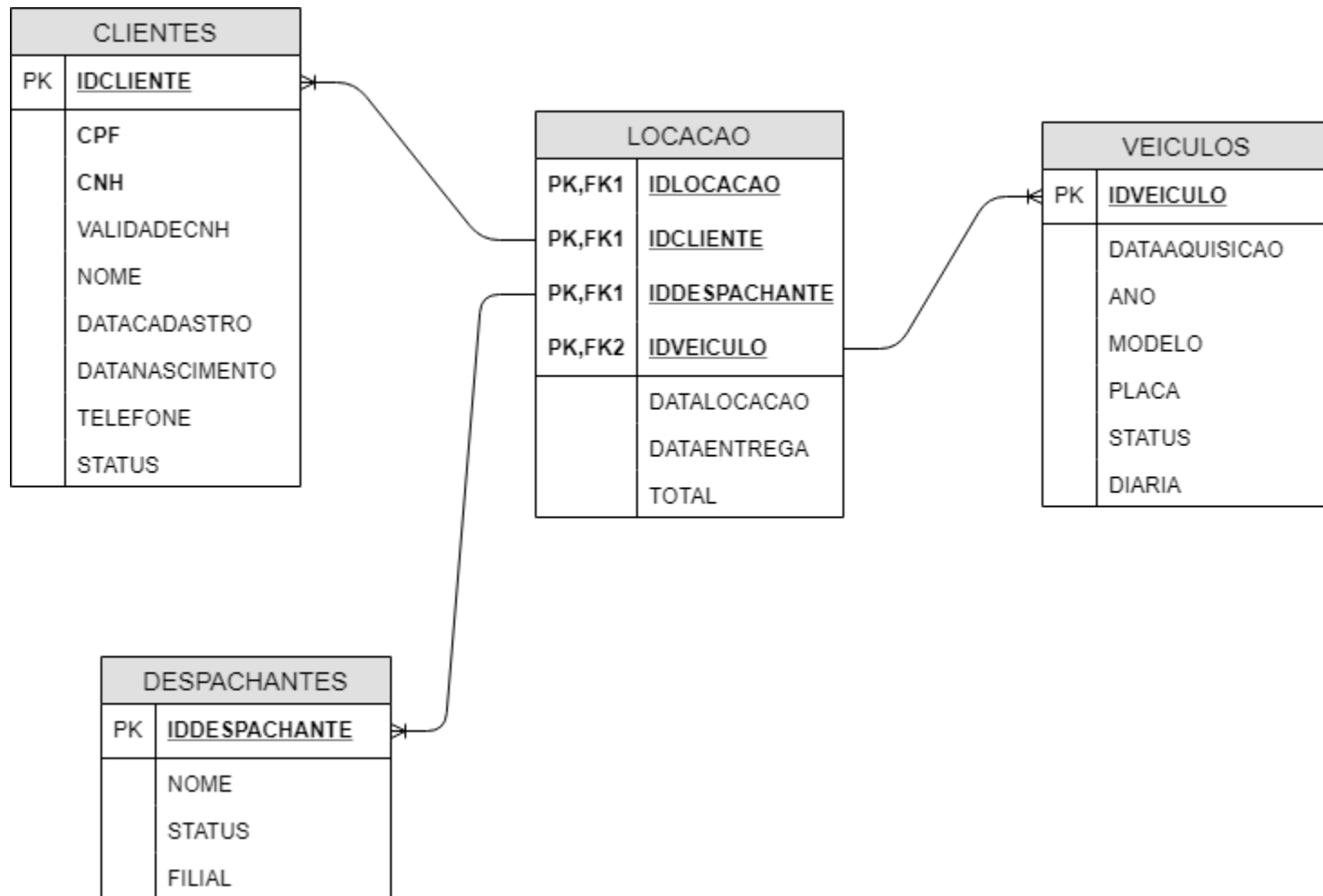
- Hive não controla o arquivo
- Arquivo pode ser utilizado por outro recurso
- Tipo recomendado
- Se excluir a tabela, apenas metadados são excluídos (arquivo permanece)

➤ **Internas (Gerenciadas)**

- Hive controla
- Acesso exclusivo pelo Hive
- Excluir tabela exclui arquivo físico

Hive – Locadora de Veículos Autônomos

23



Hive - Clientes

24

clientes.idcliente	clientes.cnh	clientes.cpf	clientes.validadecnh	clientes.nome	clientes.datacadastro	clientes.datanascimento	clientes.telefone	clientes.status
1	16528447080	44971318961	2021-09-30	Alvito Espargosa	2019-01-25	1973-03-15	51390834589	Ativo
2	83327244065	44976832780	2028-02-24	Blasco Canto	2019-01-15	1969-10-31	51922375520	Ativo
3	54513828080	27672609233	2024-12-11	Carmina Verissimo	2019-04-09	1967-11-03	51738662107	Ativo
4	54075418316	75523770439	2025-08-02	Cleusa Lamego	2019-05-26	1961-10-26	51824283267	Ativo
5	41148381120	55079718160	2020-08-21	Clóvis Carrasco	2019-01-28	1974-11-02	51056509510	Ativo
6	85588747194	30471250272	2020-06-05	Cátia Fróis	2019-05-23	1965-07-08	51901121962	Ativo
7	10003401804	45948223160	2027-05-27	Emilio Faro	2019-04-17	1966-09-23	51608924744	Ativo
8	41555470319	15942100562	2026-05-16	Hipólito Granja	2019-05-13	1974-09-21	51020081027	Ativo
9	86405032817	43316295637	2027-11-25	Iara Cardoso	2019-04-04	1981-06-18	51815092282	Ativo
10	15516586747	76742255766	2021-12-25	Isaura Farias	2019-05-04	1971-04-22	51081565935	Ativo
11	44818765309	63408388553	2027-05-03	Iuri Alancastre	2019-02-18	1970-07-15	51505707400	Ativo
12	11257675109	41824436778	2020-05-22	Laura Marcondes	2019-06-30	1985-08-29	51532760959	Ativo
13	88357618405	49548848905	2021-03-25	Micael Mangueira	2019-05-24	1994-01-08	51050644140	Ativo
14	58831402110	84776516129	2026-03-08	Márcio Taveira	2019-02-13	1963-06-07	51545396437	Ativo
15	86405301141	89650561775	2027-05-06	Noêmia Tupinambá	2019-06-28	1987-03-12	51073079084	Ativo
16	70686300220	95360639928	2028-07-19	Paula Padilha	2019-04-09	1996-05-16	51468982948	Ativo
17	70805566009	61460867546	2022-05-31	Rebeca Torcuato	2019-02-09	1994-12-06	51398031738	Ativo
18	84611132765	85761272593	2023-03-24	Rosana Betancour	2019-05-10	1976-06-08	51298196500	Ativo
19	72887644190	40239379806	2022-09-06	Sabino Abrantes	2019-05-27	1982-02-16	51805190959	Ativo
20	73201278572	56448355307	2025-02-02	Severino Leiria	2019-02-15	1965-08-06	51837010991	Ativo
21	35215454663	33281183954	2027-07-26	Tobias Garcés	2019-06-15	1962-08-20	51038698605	Ativo
22	77853823525	55351556215	2028-08-14	Vanderlei Açores	2019-01-18	1998-05-06	51069055089	Inativo
23	05013312205	32675365584	2021-12-15	Vicente Rosa	2019-05-14	1983-11-09	51782871990	Ativo
24	57460105538	54721641167	2028-02-03	Xisto Mendoza	2019-06-06	1969-07-27	51862632282	Ativo
25	02251571280	83497776322	2025-02-08	Zeferino Matoso	2019-05-01	1973-03-23	51230738729	Ativo

Hive - Veículos

25

veiculos.idveiculo	veiculos.dataaquisicao	veiculos.ano	veiculos.modelo	veiculos.placa	veiculos.status	veiculos.diaria
1	2019-01-25	2019	Tesla Model S P90D	LWJ2929	Disponível	1800.0
2	2019-01-04	2019	Volvo XC90 T8 Hybrid	LWJ2930	Disponível	1600.0
3	2019-01-10	2019	BMW 750i xDrive	LWJ2931	Disponível	1450.0
4	2019-01-28	2019	Mercedes-Benz S65 AMG Coupe	LWJ2932	Disponível	1950.0
5	2019-01-12	2019	Infiniti Q50S 3.7 Sedan	LWJ2933	Disponível	2100.0
6	2019-01-30	2019	Volvo XC90 T8 Hybrid	LWJ2934	Disponível	1600.0
7	2019-01-06	2019	BMW 750i xDrive	LWJ2935	Disponível	1450.0
8	2019-01-24	2019	Infiniti Q50S 3.7 Sedan	LWJ2936	Disponível	2100.0
9	2019-01-27	2019	BMW 750i xDrive	LWJ2937	Disponível	1450.0
10	2019-01-19	2019	Tesla Model S P90D	LWJ2938	Disponível	1800.0
11	2019-01-27	2019	Infiniti Q50S 3.7 Sedan	LWJ2939	Disponível	2100.0
12	2019-01-27	2019	BMW 750i xDrive	LWJ2940	Disponível	1450.0
13	2019-01-21	2019	Mercedes-Benz S65 AMG Coupe	LWJ2941	Disponível	1950.0
14	2019-01-10	2019	Infiniti Q50S 3.7 Sedan	LWJ2942	Disponível	2100.0
15	2019-01-07	2019	Volvo XC90 T8 Hybrid	LWJ2943	Disponível	1600.0
16	2019-01-25	2019	Volvo XC90 T8 Hybrid	LWJ2944	Disponível	1600.0
17	2019-01-22	2019	Volvo XC90 T8 Hybrid	LWJ2945	Disponível	1600.0
18	2019-01-24	2019	BMW 750i xDrive	LWJ2946	Disponível	1450.0
19	2019-01-16	2019	Mercedes-Benz S65 AMG Coupe	LWJ2947	Disponível	1950.0
20	2019-01-30	2019	Volvo XC90 T8 Hybrid	LWJ2948	Disponível	1600.0
21	2019-01-04	2019	Tesla Model S P90D	LWJ2949	Disponível	1800.0
22	2019-01-23	2019	Infiniti Q50S 3.7 Sedan	LWJ2950	Disponível	2100.0
23	2019-01-13	2019	BMW 750i xDrive	LWJ2951	Disponível	1450.0
24	2019-01-05	2019	Volvo XC90 T8 Hybrid	LWJ2952	Disponível	1600.0
25	2019-01-06	2019	Mercedes-Benz S65 AMG Coupe	LWJ2953	Indisponível	1950.0
26	2019-01-05	2019	BMW 750i xDrive	LWJ2954	Disponível	1450.0
27	2019-01-30	2019	Tesla Model S P90D	LWJ2955	Disponível	1800.0
28	2019-01-28	2019	BMW 750i xDrive	LWJ2956	Disponível	1450.0
29	2019-01-15	2019	Infiniti Q50S 3.7 Sedan	LWJ2957	Disponível	2100.0
30	2019-01-12	2019	BMW 750i xDrive	LWJ2958	Disponível	1450.0

Hive - Despachantes

26

despachantes.iddespachante	despachantes.nome	despachantes.status	despachantes.filial
1	Carmina Pestana	Ativo	Santa Maria
2	Deolinda Vilela	Ativo	Novo Hamburgo
3	Emídio Dornelles	Ativo	Porto Alegre
4	Felisbela Dornelles	Ativo	Porto Alegre
5	Graça Ornellas	Ativo	Porto Alegre
6	Matilde Rebouças	Ativo	Porto Alegre
7	Noêmia Orriça	Ativo	Santa Maria
8	Roque Vásquez	Ativo	Porto Alegre
9	Uriel Queiroz	Ativo	Porto Alegre
10	Viviana Sequeira	Ativo	Porto Alegre

Hive - Locação

27

locacao.idlocacao	locacao.idcliente	locacao.iddespachante	locacao.idveiculo	locacao.datalocacao	locacao.dataentrega	locacao.total
1	3	5	22	2019-02-20	2019-02-20	1996.0
2	5	5	7	2019-06-30	2019-07-03	1843.0
3	17	5	3	2019-03-09	2019-03-11	2016.0
4	24	3	6	2019-04-19	2019-04-22	1857.0
5	8	10	3	2019-02-24	2019-02-26	2049.0
6	14	5	23	2019-05-15	2019-05-19	1986.0
7	13	2	6	2019-05-01	2019-05-04	1989.0
8	22	5	29	2019-03-29	2019-04-03	1874.0
9	17	7	13	2019-05-10	2019-05-13	2027.0
10	1	2	21	2019-03-17	2019-03-19	2010.0
11	11	5	19	2019-05-12	2019-05-17	2006.0
12	19	10	2	2019-02-26	2019-02-26	1915.0
13	20	6	9	2019-06-12	2019-06-17	2089.0
14	24	5	3	2019-06-04	2019-06-06	1942.0
15	10	10	1	2019-02-22	2019-02-23	2071.0
16	17	4	13	2019-05-29	2019-05-29	2035.0
17	4	1	11	2019-04-09	2019-04-13	1810.0
18	9	8	11	2019-04-21	2019-04-23	2017.0
19	3	7	29	2019-05-11	2019-05-15	2095.0
20	1	10	2	2019-06-17	2019-06-17	2020.0
21	13	6	1	2019-04-12	2019-04-15	2077.0
22	16	6	8	2019-06-02	2019-06-03	1867.0
23	19	9	17	2019-05-12	2019-05-17	1862.0
24	13	6	20	2019-04-06	2019-04-07	1855.0
25	21	4	21	2019-04-26	2019-04-28	2033.0
26	25	10	8	2019-03-16	2019-03-21	2034.0
27	8	2	3	2019-05-06	2019-05-09	1865.0
28	15	3	1	2019-06-05	2019-06-07	1997.0
29	22	7	8	2019-05-07	2019-05-12	2032.0
30	13	3	7	2019-05-06	2019-05-06	2011.0

Hive – Tipos de Dados

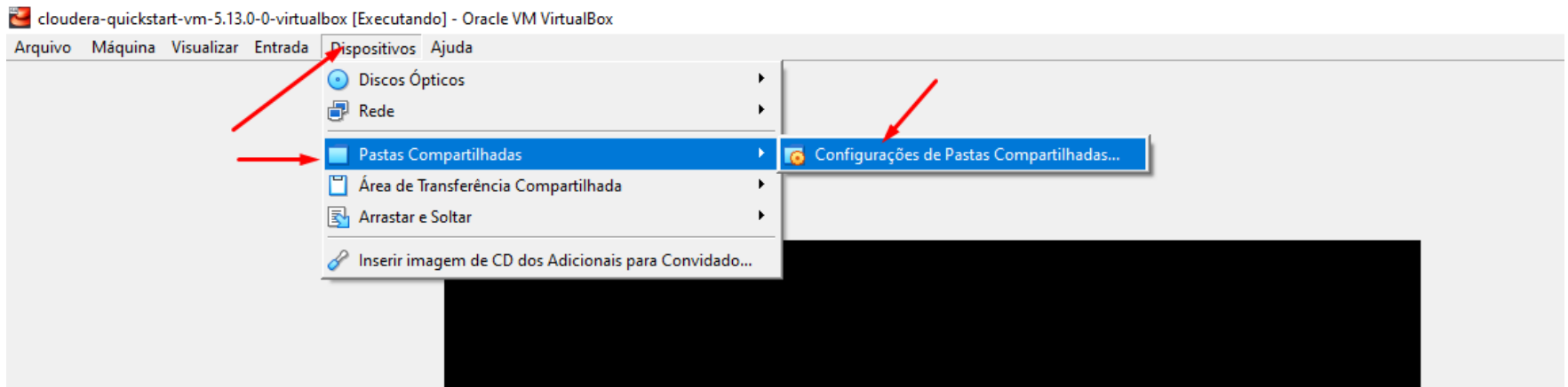
28

Tipo		
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	<chave, valor>	
Date	Data	YYYY-MM-DD
Timestamp	Data e Hora	Formato Unix padrão

Criar Pasta Compartilhada VM

29

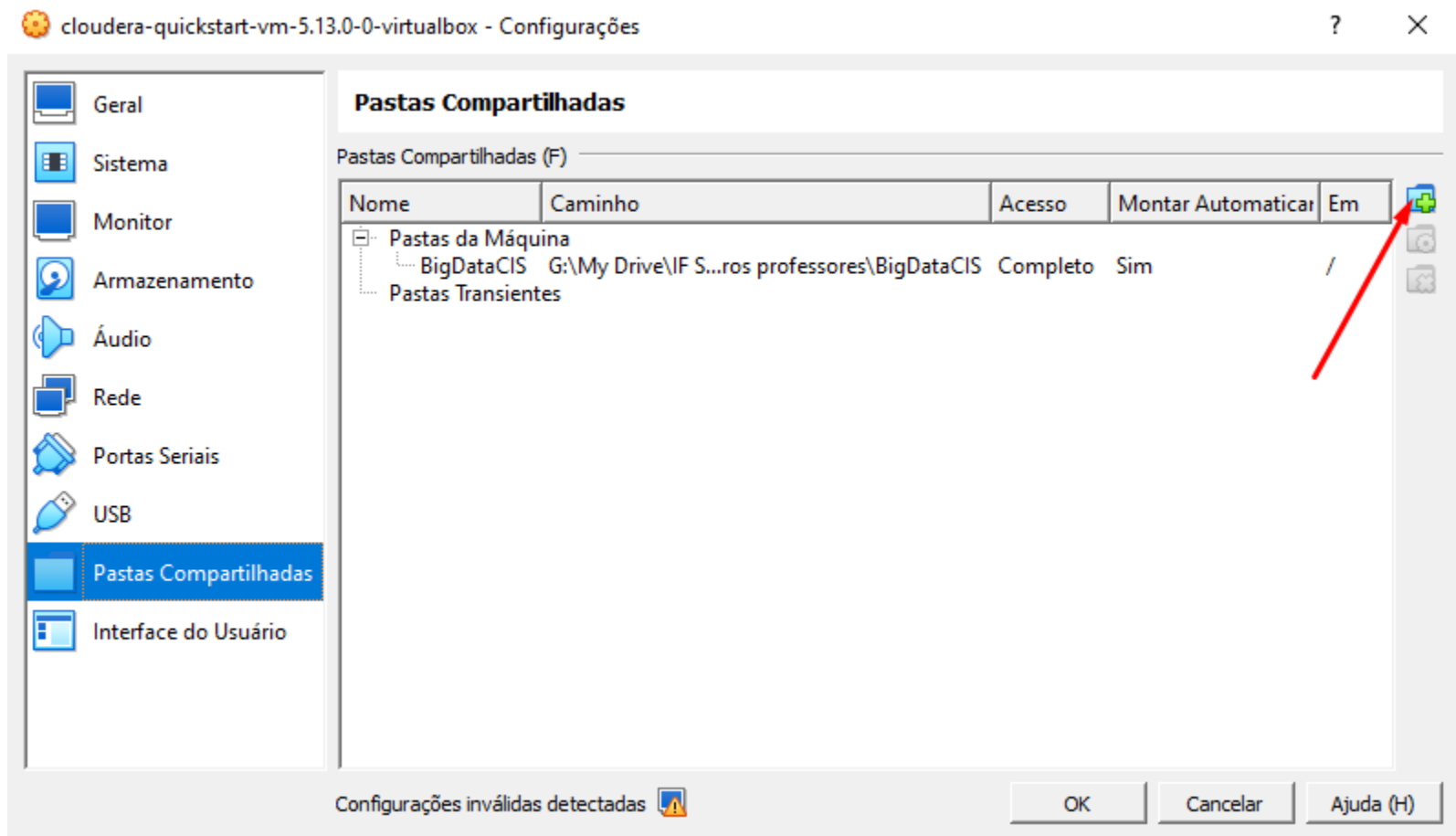
- Abra o VirtualBox
- Inicie a Máquina Virtual
- Clique em Dispositivos → Pastas Compartilhadas → Configurações de Pastas Compartilhadas...



Criar Pasta Compartilhada VM

30

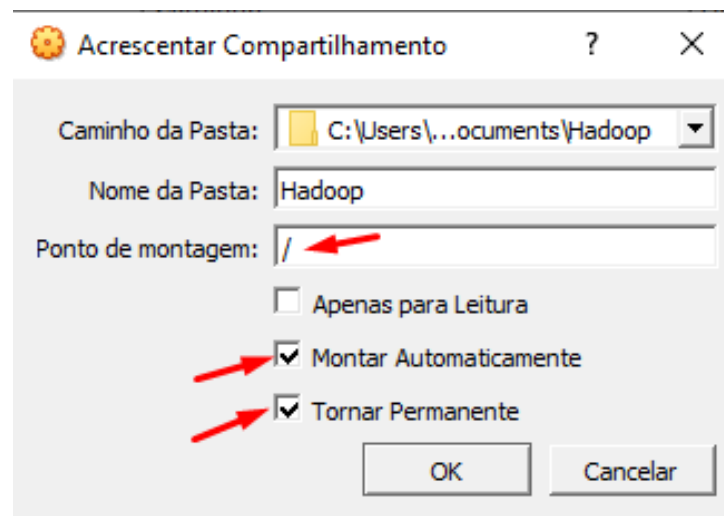
- Clique no Ícone de adicionar nova pasta



Criar Pasta Compartilhada VM

31

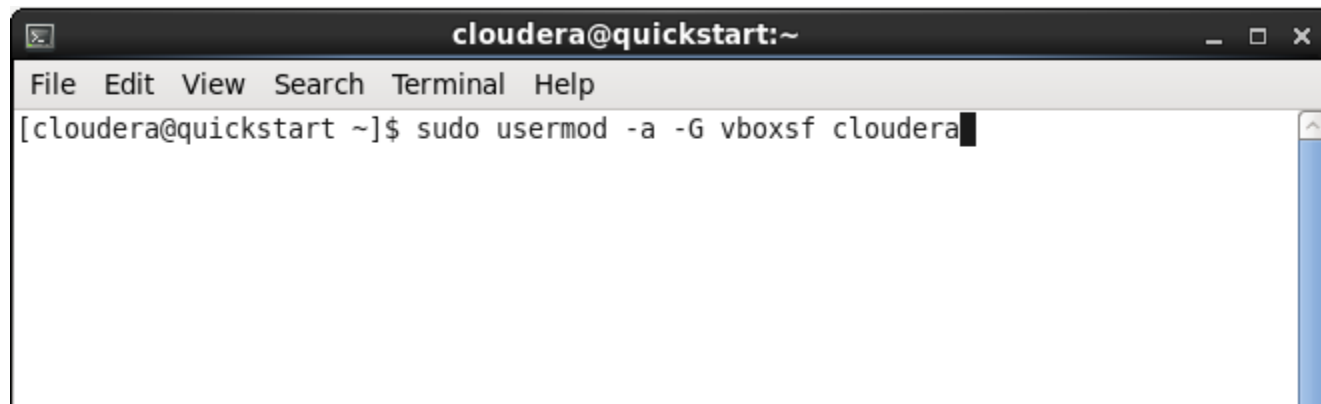
- Em Caminho da Pasta, selecione a opção “Outro”
 - Selecione no seu computador a pasta que deseja compartilhar
 - Sugestão: Crie uma pasta chamada Hadoop dentro de Documentos



Criar Pasta Compartilhada VM

32

- Clique em OK e depois em OK novamente
- Abra o terminal e execute o seguinte comando
sudo usermod -a -G vboxsf cloudera
- Reinicie a Máquina Virtual (System → Shut Down
→ Restart)

A screenshot of a terminal window titled 'cloudera@quickstart:~'. The window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The command prompt shows '[cloudera@quickstart ~]\$ sudo usermod -a -G vboxsf cloudera' with a cursor at the end of the line.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ sudo usermod -a -G vboxsf cloudera
```


Acessando o Hive

33

- Baixe os seguintes arquivos do SIGAA
 - clientes.csv
 - despachantes.csv
 - locacao.csv
 - veiculos.csv
- Coloque esses arquivos na pasta que você acabou de compartilhar (Documentos\Hadoop)

Acessando o Hive

34

- Crie uma pasta no HDFS para colocar esses arquivos

```
hadoop fs -mkdir /user/cloudera/locacao
```

- Acesse a pasta compartilhada

```
cd /media/sf_Hadoop
```

- Transfira todos os arquivos csv para o HDFS

```
hadoop fs -copyFromLocal *.csv /user/cloudera/locacao
```

Acessando o Hive

35

- Verifique se os arquivos foram copiados corretamente

```
hadoop fs -ls /user/cloudera/locacao
```

```
[cloudera@quickstart sf_Hadoop]$ hadoop fs -ls /user/cloudera/locacao
Found 4 items
-rw-r--r--  1 cloudera cloudera      2365 2022-11-05 12:45 /user/cloudera/locacao/clientes.csv
-rw-r--r--  1 cloudera cloudera       390 2022-11-05 12:45 /user/cloudera/locacao/despachantes.csv
-rw-r--r--  1 cloudera cloudera     3833 2022-11-05 12:45 /user/cloudera/locacao/locacao.csv
-rw-r--r--  1 cloudera cloudera      1939 2022-11-05 12:45 /user/cloudera/locacao/veiculos.csv
```

- Veja o conteúdo de um dos arquivos csv

```
hadoop fs -cat /user/cloudera/locacao/clientes.csv
```

Acessando o Hive

36

- Inicie o shell beeline

- Cliente Hive

beeline

```
[cloudera@quickstart sf_Hadoop]$ beeline
Beeline version 1.1.0-cdh5.13.0 by Apache Hive
beeline> █
```

- Conecte ao Hive Server

!connect jdbc:hive2://

- Vai pedir usuário e senha, basta deixar em branco e apertar enter

Acessando o Hive

37

- Crie um banco de dados teste

CREATE DATABASE MEUDB;

```
0: jdbc:hive2://> CREATE DATABASE MEUDB;  
OK  
No rows affected (2.239 seconds)
```

- Exiba todos os bancos de dados

SHOW DATABASES;

- Apague o banco de dados MEUDB

DROP DATABASE MEUDB CASCADE;

Acessando o Hive

38

- Crie um banco de dados chamado Locacao
`CREATE DATABASE LOCACAO;`
- Para selecionar e usar o banco de dados criado, rode o comando
`USE LOCACAO;`
- Nos próximos slides vamos criar uma tabela externa para cada arquivo csv

Criar tabela Clientes

39

- `CREATE EXTERNAL TABLE CLIENTES (idcliente int, cnh string, cpf string, validadecnh date, nome string, datacadastro date, datanascimento date, telefone string, status string) row format delimited fields terminated by ',' STORED AS TEXTFILE;`
- `LOAD DATA INPATH '/user/cloudera/locacao/clientes.csv' INTO TABLE CLIENTES;`
- `SELECT * FROM CLIENTES;`

Criar tabela Veiculos

40

- `CREATE EXTERNAL TABLE VEICULOS (idveiculo int, dataaquisicao date, ano int, modelo string, placa string, status string, diaria double) row format delimited fields terminated by ',' STORED AS TEXTFILE;`
- `LOAD DATA INPATH
'/user/cloudera/locacao/veiculos.csv' INTO TABLE
VEICULOS;`
- `SELECT * FROM VEICULOS;`

Criar tabela Despachantes

41

- CREATE EXTERNAL TABLE DESPACHANTES
(iddespachante int, nome string, status string, filial
string) row format delimited fields terminated by ','
STORED AS TEXTFILE;
- LOAD DATA INPATH
'/user/cloudera/locacao/despachantes.csv' INTO
TABLE DESPACHANTES;
- SELECT * FROM DESPACHANTES;

Criar tabela Locacao

42

- `CREATE EXTERNAL TABLE LOCACAO (idlocacao int, idcliente int, iddespachante int, idveiculo int, datalocacao date, dataentrega date, total double) row format delimited fields terminated by ',' STORED AS TEXTFILE;`
- `LOAD DATA INPATH '/user/cloudera/locacao/locacao.csv' INTO TABLE LOCACAO;`
- `SELECT * FROM LOCACAO;`

Metadados

43

- Exiba todas as tabelas criadas

`SHOW TABLES;`

- Exiba os metadados da tabela Clientes

`DESCRIBE CLIENTES;`

col_name	data_type	comment
idcliente	int	
cnh	string	
cpf	string	
validadecnh	date	
nome	string	
datacadastro	date	
datanascimento	date	
telefone	string	
status	string	

Metadados

44

- Outros comandos para exibir metadados

DESCRIBE FORMATTED LOCACAO;

DESCRIBE DATABASE LOCACAO;

- Consultando Hcatalog

- Saia do beeline (Ctrl+C)

mysql -u root -pcloudera

show databases;

use metastore;

show tables;

select * from DBS;

select * from TBLS where DB_ID=12;

HiveQL

HiveQL

46

- É uma linguagem de consulta para Hive para analisar e processar dados estruturados
- Hive Query Language
 - É muito semelhante ao SQL e altamente escalonável

Operadores	
Relacionais	=, <>, <, <=, >, >=
NULL	IS NULL, IS NOT NULL
Aritméticos	+, -, *, /
Lógicos	AND, OR, NOT, IN, NOT IN

Exercícios

47

- Valor: **2 pontos**
- Deverá ser entregue pelo **SIGAA**
- Pode ser feito no computador ou manuscrito. Se for manuscrito deve-se digitalizar para enviar
- O arquivo a ser enviado deve ser **PDF**

Exercícios

48

1. Selecione a data de aquisição, o modelo e a diária de todos os veículos.
2. Selecione todos os veículos cujo o status seja Disponível e a diária seja R\$ 1600 ou mais.
3. Selecione todas as locações ordenadas por data de locação.
4. Selecione o maior valor de locação.
5. Selecione todos os veículos que possuem BMW no nome.
6. Selecione a data e o nome do despachante de todas as locações.
7. Insira um despachante cujo ID é 11, nome 'José Vilela', status 'Ativo' e filial de 'Muriaé'.
8. Repita a consulta da questão 6, porém agora o despachante José Vilela também deve aparecer no resultado com o valor da data da locação NULL.
9. Selecione os modelos dos veículos e a soma total de locação por modelo de veículo.
10. Selecione os modelos de veículos, o mês e o ano de locação.

Dúvidas?

49



jean.camara@ifsudestemg.edu.br