



# Big Data com Hive e Impala

Partition e Bucketing

# Nesta seção



Partition e Bucketing



Tabelas Temporarias e Views



Arquivos ORC

# Partitions

- ❖ Divide as tabelas baseado em partições lógicas e físicas
- ❖ Um partição lógica é uma pasta no HDFS
  - ❖ Por exemplo, dividir locações por veículos
- ❖ Objetivo: Otimização de consulta, já que a partição fica fisicamente separada
  - ❖ Se fizermos uma consulta em vendas onde where veiculo = 'Tesla Model S', a consulta vai executar de forma muito mais otimizada!
- ❖ Não há benefícios em simples consultas (ex, select \* from fatovendas)

# Partitions

locacao

BMW 750i xDrive

Infiniti Q50S 3.7  
Sedan

Mercedes-Benz  
S65 AMG Coupe

Tesla Model S  
P90D

Volvo XC90 T8  
Hybrid

# Partitions

```
set hive.exec.dynamic.partition.mode=nonstrict;

create table locacaoanalitico(cliente string, despachante string,
datalocacao date, total double) PARTITIONED BY(veiculo string);

INSERT OVERWRITE TABLE locacaoanalitico PARTITION(veiculo)
select cli.nome, des.nome, loc.datalocacao, loc.total, veic.modelo
from locacao loc
join despachantes des on (loc.iddespachante = des.iddespachante)
join clientes cli on (loc.idcliente = cli.idcliente)
join veiculos veic on (loc.idveiculo = veic.idveiculo);
```

```

Loading data to table locacao.locacaoanalitico partition (veiculo=null)
Time taken for load dynamic partitions : 557
Loading partition {veiculo=Tesla Model S P90D}
Loading partition {veiculo=BMW 750i xDrive}
Loading partition {veiculo=Mercedes-Benz S65 AMG Coupe}
Loading partition {veiculo=Volvo XC90 T8 Hybrid}
Loading partition {veiculo=Infiniti Q50S 3.7 Sedan}

```

locacaoanalitico.cliente	locacaoanalitico.despachante	locacaoanalitico.datelocacao	locacaoanalitico.total	locacaoanalitico.veiculo
Clóvis Carrasco	Graca Ornelas	2019-06-30	1843.0	BMW 750i xDrive
Rebeca Torcato	Graca Ornelas	2019-03-09	2016.0	BMW 750i xDrive
Thiago Oliveira	Viviana Sequira	2019-04-24	2029.0	BMW 750i xDrive
Márcio Taverid	Graca Ornelas	2019-05-15	1989.0	BMW 750i xDrive
Severino Leiria	Matilde Rebouças	2019-06-12	2089.0	BMW 750i xDrive
Edson Góes	Deolinda Vilela	2019-06-04	1932.0	BMW 750i xDrive
Hipólito Granja	Deolinda Vilela	2019-05-06	1865.0	BMW 750i xDrive
Miguel Manquera	Emídio Dornelles	2019-05-06	2011.0	BMW 750i xDrive
Edson Góes	Deolinda Vilela	2019-05-29	2029.0	BMW 750i xDrive
Clóvis Carrasco	Carminda Pestana	2019-06-01	1989.0	BMW 750i xDrive
Thiago Oliveira	Rosane Queiroz	2019-05-31	1770.0	BMW 750i xDrive
Emílio Faro	Viviana Sequira	2019-05-02	1988.0	BMW 750i xDrive
Tobias Garcés	Carminda Pestana	2019-06-15	1888.0	BMW 750i xDrive
Edson Góes	Emídio Dornelles	2019-05-18	1993.0	BMW 750i xDrive
Edson Góes	Graca Ornelas	2019-04-21	1898.0	BMW 750i xDrive
Vanderlei Acores	Rogue Vásquez	2019-05-22	1836.0	BMW 750i xDrive
Blasco Canto	Nodenia Gracia	2019-06-22	2078.0	BMW 750i xDrive
Edson Góes	Deolinda Vilela	2019-05-11	2089.0	BMW 750i xDrive
Cátia Fróis	Carminda Pestana	2019-06-08	1853.0	BMW 750i xDrive
Emílio Faro	Rogue Vásquez	2019-05-13	1843.0	BMW 750i xDrive
Edson Góes	Carminda Pestana	2019-05-26	1881.0	BMW 750i xDrive
Paula Padilha	Carminda Pestana	2019-06-15	1812.0	BMW 750i xDrive

# Partitions

# Fisicamente

```
[cloudera@quickstart locacao2]$ hdfs dfs -ls /user/hive/warehouse/locacao.db/locacaoanalitico
Found 5 items
drwxrwxrwx  - cloudera supergroup      0 2019-08-07 05:53 /user/hive/warehouse/locacao.db/locacaoanalitico/veiculo=BMW 750i xDrive
drwxrwxrwx  - cloudera supergroup      0 2019-08-07 05:53 /user/hive/warehouse/locacao.db/locacaoanalitico/veiculo=Infiniti Q50S 3.7 Sedan
drwxrwxrwx  - cloudera supergroup      0 2019-08-07 05:53 /user/hive/warehouse/locacao.db/locacaoanalitico/veiculo=Mercedes-Benz S65 AMG Coupe
drwxrwxrwx  - cloudera supergroup      0 2019-08-07 05:53 /user/hive/warehouse/locacao.db/locacaoanalitico/veiculo=Tesla Model S P90D
drwxrwxrwx  - cloudera supergroup      0 2019-08-07 05:53 /user/hive/warehouse/locacao.db/locacaoanalitico/veiculo=Volvo XC90 T8 Hybrid
```

# Bucketing

- ❖ Partições: variam conforme os dados.
  - ❖ Potencial problema: milhares ou milhões de partições!
- ❖ Bucketing: número Fixo de partições, não muda conforme os dados
  - ❖ Divide fisicamente de forma balanceada em partições
  - ❖ Se existirem 300 partições e 50 dados diferentes, 250 ficarão vazios!
  - ❖ Ótimo para tabelas que operam em Joins

# Bucketing

```
create table locacaoanalitico2 (cliente string, despachante  
string, datalocacao date, total double, veiculo string) clustered  
by (veiculo) into 4 buckets;
```

```
INSERT OVERWRITE TABLE locacaoanalitico2  
select cli.nome, des.nome, loc.datalocacao, loc.total, veic.modelo  
from locacao loc  
join despachantes des on (loc.iddespachante = des.iddespachante)  
join clientes cli on (loc.idcliente = cli.idcliente)  
join veiculos veic on (loc.idveiculo = veic.idveiculo);
```

## Partitions

Baixa cardinalidade

Um atributo ou mais

## Bucketing

Cardinalidade alta e variável

Apenas um atributo

# Partições Vs Bucketing