



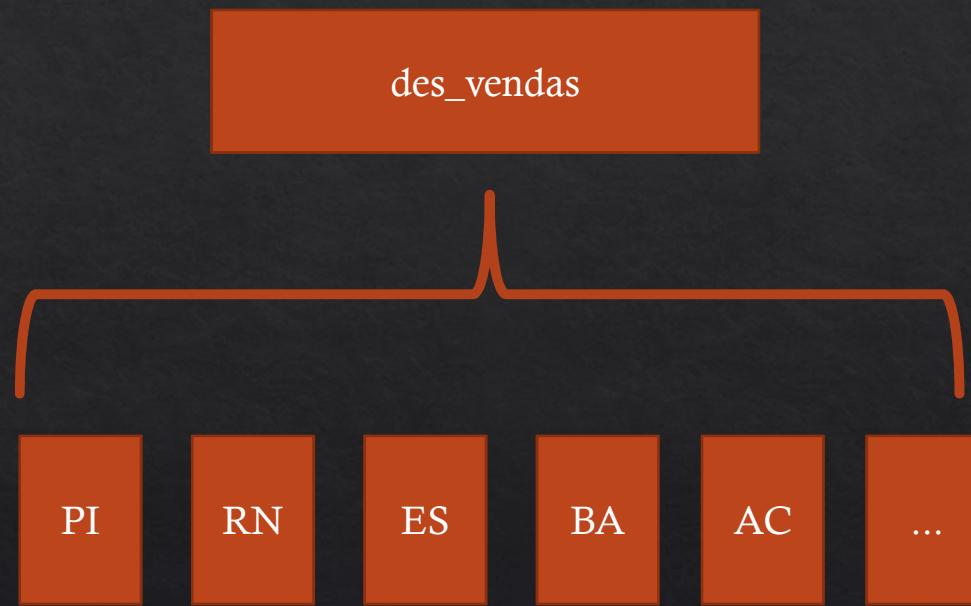
# Formação Engenheiro de Dados

Hive: Partition e Bucketing

# Partitions

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

# Partitions



# Partitions

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

```
create table des_vendas_part(quantidade int, valortotal float )
PARTITIONED BY (estado char(2));
```

```
INSERT OVERWRITE TABLE des_vendas_part PARTITION(estado)
SELECT quantidade, valortotal, estado from des_vendas;
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/hive/warehouse/ed.db/dev.vendas_part/
Found 1 items
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=AC
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=AL
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=AM
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=AP
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=BA
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=CE
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=DF
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=ES
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=GO
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=MA
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=MG
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=MS
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=RS
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=TO
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=MT
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=PA
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=PB
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=PE
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=PI
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=RN
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=RO
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=RS
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=SE
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=SC
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=SE
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:43 /user/hive/warehouse/ed.db/dev.vendas_part/estado=SP
drwxrwxrwx - cloudera supergroup 0 2019-07-02 08:42 /user/hive/warehouse/ed.db/dev.vendas_part/estado=TO
```

Loading partition {estad0=PI}  
 Loading partition {estad0=RN}  
 Loading partition {estad0=ES}  
 Loading partition {estad0=BA}  
 Loading partition {estad0=RR}  
 Loading partition {estad0=RJ}  
 Loading partition {estad0=SC}  
 Loading partition {estad0=DF}  
 Loading partition {estad0=SP}  
 Loading partition {estad0=MG}  
 Loading partition {estad0=PE}  
 Loading partition {estad0=AM}  
 Loading partition {estad0=T0}  
 Loading partition {estad0=CE}  
 Loading partition {estad0=PB}  
 Loading partition {estad0=MA}  
 Loading partition {estad0=MT}  
 Loading partition {estad0=RS}  
 Loading partition {estad0=AL}  
 Loading partition {estad0=GO}  
 Loading partition {estad0=MS}  
 Loading partition {estad0=AC}  
 Loading partition {estad0=RO}  
 Loading partition {estad0=AP}  
 Loading partition {estad0=SE}  
 Loading partition {estad0=PR}  
 Loading partition {estad0=PA}

# Partitions

# 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 des_vendas_buck(quantidade int, valortotal float,  
estado char(2) ) clustered by(estado)  
into 4 buckets;
```

```
INSERT OVERWRITE TABLE des_vendas_buck  
SELECT quantidade, valortotal, estado from des_vendas;
```

## Partitions

Baixa cardinalidade

Um atributo ou mais

## Bucketing

Cardinalidade alta e variável

Apenas um atributo

# Partições Vs Bucketing