

→ Put 'custs.txt', 'custs_add' and 'txns1.txt' files to ftp

→ Customer Table

→ C2. Create customer table

```
hive (surya_training)> CREATE TABLE customer(custno INT, firstname
STRING, lastname STRING, age INT, profession STRING) ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
Time taken: 0.175 seconds, Fetched: 2 row(s)
hive (surya_training)> CREATE TABLE customer(custno INT, firstname STRING, lastname STRING, age INT, profession STRING)
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.398 seconds
hive (surya_training)>
```

```
hive (surya_training)> show tables ;
```

```
hive (surya_training)> show tables ;
OK
customer
nyse
stkv01
Time taken: 0.197 seconds, Fetched: 3 row(s)
hive (surya_training)>
```

→ D1. Load the data into the table (from local file system)

```
hive (surya_training)> LOAD DATA LOCAL INPATH 'custs.txt' OVERWRITE
INTO TABLE customer;
```

```
hive (surya_training)> LOAD DATA LOCAL INPATH 'custs.txt' OVERWRITE INTO TABLE customer;
Loading data to table surya_training.customer
OK
Time taken: 1.138 seconds
hive (surya_training)>
```

```
hive (surya_training)> SELECT COUNT(*) FROM customer ;
```

```
hive (surya_training)> SELECT COUNT(*) FROM customer ;
Query ID = bigdata1ab456422_20230529091615_3c7baf5a-6920-4c6b-ad33-42c515699c40
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
23/05/29 09:16:15 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/29 09:16:16 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1684866872278_3461, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_3461/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1684866872278_3461
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-29 09:16:26,749 Stage-1 map = 0%, reduce = 0%
2023-05-29 09:16:35,047 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.03 sec
2023-05-29 09:16:42,336 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.36 sec
MapReduce Total cumulative CPU time: 6 seconds 360 msec
Ended Job = job_1684866872278_3461
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.36 sec HDFS Read: 400195 HDFS Write: 104 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 360 msec
OK
9090
Time taken: 28.442 seconds, Fetched: 1 row(s)
hive (surya_training)>
```

→ D1. Load the data into the table (from local file system)

```
hive (surya_training)> LOAD DATA LOCAL INPATH 'custs_add' INTO TABLE
customer;
```

```
hive (surya_training)> LOAD DATA LOCAL INPATH 'custs_add' INTO TABLE customer;
Loading data to table surya_training.customer
OK
Time taken: 0.684 seconds
hive (surya_training)>
```

```
hive (surya_training)> SELECT COUNT(*) FROM customer ;
```

```
hive (surya_training)> SELECT COUNT(*) FROM customer ;
Query ID = bigdatalab456422_20230529091941_08db940c-7987-48ec-bee8-da40b6b991d5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
23/05/29 09:19:42 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/29 09:19:42 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1684866872278_3465, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_3465/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1684866872278_3465
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-29 09:19:52,444 Stage-1 map = 0%, reduce = 0%
2023-05-29 09:19:59,661 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.3 sec
2023-05-29 09:20:08,870 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.26 sec
MapReduce Total cumulative CPU time: 6 seconds 260 msec
Ended Job = job_1684866872278_3465
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.26 sec HDFS Read: 400407 HDFS Write: 105 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 260 msec
OK
10000
Time taken: 29.031 seconds, Fetched: 1 row(s)
hive (surya_training)>
```

→ Transaction Table

→ C1. Create transaction table

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -mkdir sales
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put txns1.txt sales
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -ls sales
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -ls sales
Found 1 items
-rw-r--r-- 3 bigdatalab456422 bigdatalab456422 4418144 2023-05-29 09:22 sales/txns1.txt
[bigdatalab456422@ip-10-1-1-204 ~]$
```

[Home](#) / [user](#) / [bigdatalab456422](#) / [sales](#) [Trash](#)

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	sales		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 29, 2023 02:22 AM
<input type="checkbox"/>	.		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 29, 2023 02:22 AM
<input type="checkbox"/>	txns1.txt	4.2 MB	bigdatalab456422	bigdatalab456422	-rw-r--r--	May 29, 2023 02:22 AM

Show 45 of 1 items Page 1 of 1

-- creating and adding data to sales table in hadoop fs

```
hive (surya_training)> CREATE TABLE txnrecords(txnno INT, txndate
STRING, custno INT, amount DOUBLE, category STRING, product STRING,
city STRING, state STRING, spendby STRING) ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' STORED AS textfile LOCATION
'/user/bigdatalab456422/sales';
```

```
hive (surya_training)> CREATE TABLE txnrecords(txnno INT, txndate STRING, custno INT, amount DOUBLE,
> category STRING, product STRING, city STRING, state STRING, spendby STRING)
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS textfile
> LOCATION '/user/bigdatalab456422/sales';
OK
Time taken: 0.072 seconds
```






-- textfile data is directly accessible in table, but it is not mapped to the /user/hive.warehouse dir but it is in hadoop fs directory sales

```
hive (surya_training)> SELECT * FROM txnrecords LIMIT 10;
```

```
hive (surya_training)> SELECT * FROM txnrecords LIMIT 10;
OK
0      06-26-2011      4007024 40.33 Exercise & Fitness Cardio Machine Accessories Clarksville Tennessee credit
1      06-26-2011      4006742 198.44 Exercise & Fitness Weightlifting Gloves Long Beach California credit
2      06-01-2011      4009775 5.58 Exercise & Fitness Weightlifting Machine Accessories Anaheim California credit
3      06-05-2011      4002199 198.19 Gymnastics Gymnastics Rings Milwaukee Wisconsin credit
4      12-17-2011      4002613 98.81 Team Sports Field Hockey Nashville Tennessee credit
5      02-14-2011      4007591 193.63 Outdoor Recreation Camping & Backpacking & Hiking Chicago Illinois credit
6      10-28-2011      4002190 27.89 Puzzles Jigsaw Puzzles Charleston South Carolina credit
7      07-14-2011      4002964 96.01 Outdoor Play Equipment Sandboxes Columbus Ohio credit
8      01-17-2011      4007361 10.44 Winter Sports Snowmobiling Des Moines Iowa credit
9      05-17-2011      4004798 152.46 Jumping Bungee Jumping St. Petersburg Florida credit
Time taken: 0.103 seconds, Fetched: 10 row(s)
```




This table is not in /user/hive/warehouse, but in \$HADOOP_PATH/sales

[Home](#) / [user](#) / [hive](#) / [warehouse](#) / [surya_training.db](#) [Trash](#)

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑		hive	hive	drwxrwxrwx	May 28, 2023 10:54 PM
<input type="checkbox"/>	 .		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 02:11 AM
<input type="checkbox"/>	 customer		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 02:18 AM
<input type="checkbox"/>	 nyse		bigdatalab456422	hive	drwxrwxrwx	May 23, 2023 11:29 PM
<input type="checkbox"/>	 stkvol		bigdatalab456422	hive	drwxrwxrwx	May 24, 2023 12:30 AM

Show of 3 items Page of 1 [⏪](#) [⏩](#) [⏴](#) [⏵](#)

[Home](#) / [user](#) / [bigdatalab456422](#) / [sales](#) [Trash](#)

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 29, 2023 02:29 AM
<input type="checkbox"/>	 .		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 29, 2023 02:34 AM
<input type="checkbox"/>	 txns1.txt	4.2 MB	bigdatalab456422	bigdatalab456422	-rw-r--r--	May 29, 2023 02:34 AM

Show of 1 items Page of 1 [⏪](#) [⏩](#) [⏴](#) [⏵](#)

→ Hive optimization

1TB x time

10TB 10x

- Creating ORC and Parquet files
- Columnar tables
- Inbuilt indexes

ORC (Optimized Row Columnar) : improves performance when hive R/W and processes data

Advantages :

- Compressed data, reduces data storage by over 90%
- Inbuilt indexes
- Columnar tables
- Hive data is processed faster as compared processing of text files
- Efficient execution
- Data is encrypted, so data is more secure

g. ORC supports ACID properties, but Parquet does not support ACID properties

→ create table using ORC file

→ C1. Create transaction table

```
hive (surya_training)> CREATE TABLE txn_orc(txnno INT, txndate
STRING, custno INT, amount DOUBLE, category STRING, product STRING,
city STRING, state STRING, spendby STRING) STORED AS ORC;
```

```
hive (surya_training)> CREATE TABLE txn_orc(txnno INT, txndate STRING, custno INT, amount DOUBLE,
>
> category STRING, product STRING, city STRING, state STRING, spendby STRING)
> STORED AS ORC;
```

```
OK
Time taken: 0.092 seconds
hive (surya_training)>
```

```
hive (surya_training)> INSERT OVERWRITE TABLE txn_orc SELECT * FROM
txnrecords;
```

```
hive (surya_training)> INSERT OVERWRITE TABLE txn_orc SELECT * FROM txnrecords;
Query ID = bigdatalab456422_20230529102351_b59557b8-e280-45ec-9e95-b9d1f99e8fa3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/05/29 10:23:52 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/29 10:23:52 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1684866872278_3530, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:8086/proxy/application_1684866872278_3530/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1684866872278_3530
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-05-29 10:24:03,529 Stage-1 map = 0%, reduce = 0%
2023-05-29 10:24:13,811 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.01 sec
MapReduce Total cumulative CPU time: 6 seconds 10 msec
Ended Job = job_1684866872278_3530
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://nameservice1/user/hive/warehouse/surya_training.db/txn_orc/.hive-staging_hive_2023-05-29_10-23-51_606_9153754882217344466-1/-ext-10000
Loading data to table surya_training.txn_orc
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 6.01 sec HDFS Read: 4423450 HDFS Write: 506063 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 10 msec
OK
Time taken: 24.043 seconds
hive (surya_training)>
```

```
hive (surya_training)> SELECT * FROM txn_orc LIMIT 10;
```

```
hive (surya_training)> SELECT * FROM txn_orc LIMIT 10;
OK
0      06-26-2011      4007024 40.33  Exercise & Fitness  Cardio Machine Accessories  Clarksville  Tennessee  credit
1      05-26-2011      4006742 198.44 Exercise & Fitness  Weightlifting Gloves  Long Beach  California  credit
2      06-01-2011      4009775 5.58  Exercise & Fitness  Weightlifting Machine Accessories  Anaheim  California  credit
3      06-05-2011      4002199 198.19 Gymnastics  Gymnastics Rings  Milwaukee  Wisconsin  credit
4      12-17-2011      4002613 98.81 Team Sports  Field Hockey  Nashville  Tennessee  credit
5      02-14-2011      4007591 193.63 Outdoor Recreation  Camping & Backpacking & Hiking  Chicago  Illinois  credit
6      10-28-2011      4002190 27.89 Puzzles  Jigsaw Puzzles  Charleston  South Carolina  credit
7      07-14-2011      4002964 96.91 Outdoor Play Equipment  Sandboxes  Columbus  Ohio  credit
8      01-17-2011      4007361 10.44 Winter Sports  Snowmobiling  Des Moines  Iowa  credit
9      05-17-2011      4004798 152.46 Jumping Bungee Jumping  St. Petersburg  Florida  credit
Time taken: 0.15 seconds, Fetched: 10 row(s)
hive (surya_training)>
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -cat
/user/hive/warehouse/surya_training.db/txn_orc/* ;
```



```
hive (surya_training)> CREATE TABLE txn_parquet(txnno INT, txndate STRING, custno INT, amount DOUBLE,
>
> category STRING, product STRING, city STRING, state STRING, spendby STRING)
>
> STORED AS PARQUET;
OK
Time taken: 0.095 seconds
hive (surya_training)>
```

```
hive (surya_training)> INSERT OVERWRITE TABLE txn_parquet SELECT *
FROM txnrecords;
```

```
hive (surya_training)> SELECT * FROM txn_parquet LIMIT 10;
OK
Time taken: 0.079 seconds
hive (surya_training)> INSERT OVERWRITE TABLE txn_parquet SELECT * FROM txnrecords;
Query ID = bigdatalab456422_20230529173433_36a43519-1d81-43fa-9ad9-3391d469204f
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
23/05/29 17:34:33 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/29 17:34:33 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1684866872278_3739, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_3739/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1684866872278_3739
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-05-29 17:34:43,257 Stage-1 map = 0%, reduce = 0%
2023-05-29 17:34:52,666 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.42 sec
MapReduce Total cumulative CPU time: 5 seconds 420 msec
Ended Job = job_1684866872278_3739
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://nameservice1/user/hive/warehouse/surya_training.db/txn_parquet/.hive-staging_hive_2023-05-29_17-34-33_259_598362340879233563-1/-ext-1000
0
Loading data to table surya_training.txn_parquet
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 5.42 sec HDFS Read: 4423501 HDFS Write: 787857 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 420 msec
OK
Time taken: 21.123 seconds
hive (surya_training)>
```

```
hive (surya_training)> SELECT * FROM txn_parquet LIMIT 10;
```

```
hive (surya_training)> SELECT * FROM txn_parquet LIMIT 10;
OK
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/hive-exec-2.1.1-cdh6.2.1.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/parquet-hadoop-bundle-1.9.0-cdh6.2.1.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/parquet-format-2.3.1-cdh6.2.1.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/jars/parquet-pig-bundle-1.9.0-cdh6.2.1.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [shaded.parquet.org.slf4j.helpers.NOPLoggerFactory]
0 06-26-2011 4007024 40.33 Exercise & Fitness Cardio Machine Accessories Clarksville Tennessee credit
1 05-26-2011 4006742 198.44 Exercise & Fitness Weightlifting Gloves Long Beach California credit
2 06-01-2011 4009775 5.58 Exercise & Fitness Weightlifting Machine Accessories Anaheim California credit
3 06-05-2011 4002199 198.19 Gymnastics Gymnastics Rings Milwaukee Wisconsin credit
4 12-17-2011 4002613 98.81 Team Sports Field Hockey Nashville Tennessee credit
5 02-14-2011 4007591 193.63 Outdoor Recreation Camping & Backpacking & Hiking Chicago Illinois credit
6 10-28-2011 4002190 27.89 Puzzles Jigsaw Puzzles Charleston South Carolina credit
7 07-14-2011 4002964 96.01 Outdoor Play Equipment Sandboxes Columbus Ohio credit
8 01-17-2011 4007361 10.44 Winter Sports Snowmobiling Des Moines Iowa credit
9 05-17-2011 4004798 152.46 Jumping Bungee Jumping St. Petersburg Florida credit
Time taken: 0.09 seconds, Fetched: 10 row(s)
hive (surya_training)>
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -cat
/user/hive/warehouse/surya_training.db/txn_parquet/* ;
```


→ Partition Table

→ creating partition table

```
Untitled - Notepad
File Edit Format View Help
creating a partition table

00000000,06-26-2011,4007024,040.33,Exercise & Fitness,Cardio Machine Accessories,Clarksville,Tennessee,credit
00000001,05-26-2011,4006742,198.44,Exercise & Fitness,Weightlifting Gloves,Long Beach,California,credit
00000002,06-01-2011,4009775,005.58,Exercise & Fitness,Weightlifting Machine Accessories,Anaheim,California,credit
00000003,06-05-2011,4002199,198.19,Gymnastics,Gymnastics Rings,Milwaukee,Wisconsin,credit
00000004,12-17-2011,4002613,098.81,Team Sports,Field Hockey,Nashville ,Tennessee,credit
00000005,02-14-2011,4007591,193.63,Outdoor Recreation,Camping & Backpacking & Hiking,Chicago,Illinois,credit
00000006,10-28-2011,4002190,027.89,Puzzles,Jigsaw Puzzles,Charleston,South Carolina,credit
00000007,07-14-2011,4002964,096.01,Outdoor Play Equipment,Sandboxes,Columbus,Ohio,credit

data = 1 TB
blocks = 8192

select sum(amount) from txnrecords where category = 'Exercise & Fitness';
8000 mappers, containers, RAM

actual data = 10%

select sum(amount) from txnrecords where category = 'Team Sports';
8000 mappers

actual data = 5%

select sum(amount) from txnrecords where category = 'Gym';
```

Handwritten notes:
800 mappers.
8000 blocks - specific
4000 blocks - 4000 mappers
24000
162
1500

```
Untitled - Notepad
File Edit Format View Help
00000004,12-17-2011,4002613,098.81,Team Sports,Field Hockey,Nashville ,Tennessee,credit
00000005,02-14-2011,4007591,193.63,Outdoor Recreation,Camping & Backpacking & Hiking,Chicago,Illinois,credit
00000006,10-28-2011,4002190,027.89,Puzzles,Jigsaw Puzzles,Charleston,South Carolina,credit
00000007,07-14-2011,4002964,096.01,Outdoor Play Equipment,Sandboxes,Columbus,Ohio,credit

data = 1 TB
blocks = 8192

select sum(amount) from txnrecords where category = 'Exercise & Fitness';
8000 mappers, containers, RAM

actual data = 10%

select sum(amount) from txnrecords where category = 'Team Sports';
8000 mappers

actual data = 5%

select sum(amount) from txnrecords where category = 'Gym';

define a new partitioned table txnrecsbycat - partition col - category

select sum(amount) from txnrecsbycat where category = 'Exercise & Fitness';
800 blocks , 800 mappers
```

Handwritten notes:
2.1. ✓
1001
100 TB → 200 GB.
800 blocks - 800 mappers
800 blocks - 800 mappers

→ H1. Create partitioned table

```
hive (surya_training)> CREATE TABLE txnrecsByCat(txnno INT, txndate
STRING, custno INT, amount DOUBLE, product STRING, city STRING, state
STRING, spendby STRING) PARTITIONED BY (category STRING) ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
hive (surya_training)> CREATE TABLE txnrecsByCat(txnno INT, txndate STRING, custno INT, amount DOUBLE,
> product STRING, city STRING, state STRING, spendby STRING)
> PARTITIONED BY (category STRING)
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.131 seconds
hive (surya_training)>
```

```
hive (surya_training)> DESC txnrecsByCat;
```

```
hive (surya_training)> DESC txnrecsByCat;
OK
txnno                int
txndate              string
custno              int
amount              double
product             string
city                string
state               string
spendby             string
category            string

# Partition Information
# col_name            data_type            comment
category             string
Time taken: 0.113 seconds, Fetched: 14 row(s)
hive (surya_training)>
```

```
hive (surya_training)> set
hive.exec.dynamic.partition.mode=nonstrict;
hive (surya_training)> set hive.exec.dynamic.partition=true;
```

→ I1. Load data into partition table (single bucket)


















```
hive (surya_training)> INSERT OVERWRITE TABLE txnrecsByCat
PARTITION(category) SELECT txn.txnno, txn.txndate,txn.custno,
txn.amount,txn.product,txn.city,txn.state, txn.spendby, txn.category
FROM txnrecords txn DISTRIBUTE BY category;
```

```
hive (surya_training)> INSERT OVERWRITE TABLE txnrecsByCat PARTITION(category)
>
> SELECT txn.txnno, txn.txndate,txn.custno, txn.amount,txn.product,txn.city,txn.state, txn.spendby, txn.category FROM txnrecords txn DISTRIBUTE BY
category;
Query ID = bigdatalab456422_20230529114258_e47ccb0e-cc5e-4634-a79b-8b6851f076bc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
23/05/29 11:42:59 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/29 11:42:59 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1684866872278_3571, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_3571/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1684866872278_3571
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-29 11:43:20,267 Stage-1 map = 0%, reduce = 0%
2023-05-29 11:43:31,564 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.19 sec
2023-05-29 11:43:40,803 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 9.57 sec
MapReduce Total cumulative CPU time: 9 seconds 570 msec
Ended Job = job_1684866872278_3571
Loading data to table surya_training.txnrecsbycat partition (category=null)

Time taken to load dynamic partitions: 0.41 seconds
Time taken for adding to write entity : 0.004 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 9.57 sec HDFS Read: 4428810 HDFS Write: 3500532 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 570 msec
OK
Time taken: 44.713 seconds
hive (surya_training)>
```





/user/hive/warehouse/surya_training.db/txnrecsbycat

Home / user / hive / warehouse / surya_training.db / txnrecsbycat Trash

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:39 AM
<input type="checkbox"/>	 .		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Air Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Combat Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Dancing		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Exercise & Fitness		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Games		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Gymnastics		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Indoor Games		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Jumping		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Outdoor Play Equipment		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Outdoor Recreation		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Puzzles		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Racquet Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Team Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 category=Water Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM

/user/hive/warehouse/surya_training.db/txnrecsbycat/category=Air%20Sports

Home / user / hive / warehouse / surya_training.db / txnrecsbycat / category=Air Sports Trash

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 .		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 04:43 AM
<input type="checkbox"/>	 000000_0	64.1 KB	bigdatalab456422	hive	-rwxrwxrwx	May 29, 2023 04:43 AM

Show 45 of 1 items Page 1 of 1

→ H3. Create partitioned table (single bucket) on a derived column

```
hive (surya_training)> CREATE TABLE txnrecsByCat3(txnno INT, txndate
STRING, custno INT, amount DOUBLE, product STRING, city STRING, state
STRING) PARTITIONED BY (category STRING,spendby STRING) ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
hive (surya_training)> CREATE TABLE txnrecsByCat3(txnno INT, txndate STRING, custno INT, amount DOUBLE,
> product STRING, city STRING, state STRING)
> PARTITIONED BY (category STRING,spendby STRING)
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.081 seconds
hive (surya_training)>
```

```
hive (surya_training)> DESC txnrecsByCat3;
```

```
hive (surya_training)> DESC txnrecsByCat3;
OK
txnno          int
txndate        string
custno         int
amount         double
product        string
city           string
state          string
category       string
spendby        string

# Partition Information
# col_name      data_type      comment
category        string
spendby         string
Time taken: 0.108 seconds, Fetched: 15 row(s)
hive (surya_training)>
```


















→ I3. Load data into partition table (single bucket)

```
hive (surya_training)> FROM txnrecords txn INSERT OVERWRITE TABLE
txnrecsByCat3 PARTITION(category,spendby) SELECT txn.txnno,
txn.txndate,txn.custno, txn.amount,txn.product,txn.city,txn.state,
txn.category, txn.spending DISTRIBUTE BY category,spending;
```






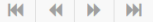
```
hive (surya_training)> FROM txnrecords txn INSERT OVERWRITE TABLE txnrecsByCat3 PARTITION(category,spendby)
>
> SELECT txn.txnno, txn.txndate,txn.custno, txn.amount,txn.product,txn.city,txn.state, txn.category, txn.spendby DISTRIBUTE BY category,spendby;
Query ID = bigdatalab456422_20230529120218_ae66b8ab-30b2-4cbd-9240-c67fe1080a73
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
23/05/29 12:02:19 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/29 12:02:19 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1684866872278_3613, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_3613/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1684866872278_3613
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-29 12:02:30,497 Stage-1 map = 0%, reduce = 0%
2023-05-29 12:02:38,848 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.31 sec
2023-05-29 12:02:48,084 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.32 sec
MapReduce Total cumulative CPU time: 10 seconds 320 msec
Ended Job = job_1684866872278_3613
Loading data to table surya_training.txnrecsbycat3 partition (category=null, spendby=null)

chmod: File hdf://nameservice1/user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Jumping/spendby=cash does not exist.
chmod: File hdf://nameservice1/user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Racquet Sports/spendby=credit does not exist.
chmod: File hdf://nameservice1/user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Winter Sports/spendby=cash does not exist.
chmod: changing permissions of 'hdf://nameservice1/user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Outdoor Play Equipment/spendby=cash': Directory/File do
es not exist /user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Outdoor Play Equipment/spendby=cash
chmod: changing permissions of 'hdf://nameservice1/user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Exercise & Fitness/spendby=credit': Directory/File does
not exist /user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Exercise & Fitness/spendby=credit
Time taken to load dynamic partitions: 0.852 seconds
Time taken for adding to write entity : 0.004 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 10.32 sec HDFS Read: 4428817 HDFS Write: 3165791 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 320 msec
OK
Time taken: 32.689 seconds
hive (surya_training)>
```

/user/hive/warehouse/surya_training.db/txnrecsbycat3

Home	/ user / hive / warehouse / surya_training.db / txnrecsbycat3					Trash
<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:00 AM
<input type="checkbox"/>	 .		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Air Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Combat Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Dancing		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Exercise & Fitness		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Games		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Gymnastics		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Indoor Games		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Jumping		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Outdoor Play Equipment		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Outdoor Recreation		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Puzzles		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Racquet Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Team Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 category=Water Sports		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM

/user/hive/warehouse/surya_training.db/txnrecsbycat3/category=Air%20Sports

Home	/ user / hive / warehouse / surya_training.db / txnrecsbycat3 / category=Air Sports					Trash
<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 .		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 spendby=cash		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
<input type="checkbox"/>	 spendby=credit		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:02 AM
Show 45 of 2 items						
			Page	1	of 1	

→ H3. Create partitioned table (single bucket) on a derived column

```
hive (surya_training)> CREATE TABLE txnrecsByCat4(txnno INT, txndate
STRING, custno INT, amount DOUBLE, category String, product STRING,
city STRING, state STRING, spendby STRING) PARTITIONED BY (month
STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS
TEXTFILE;
```

```
hive (surya_training)> CREATE TABLE txnrecsByCat4(txnno INT, txndate STRING, custno INT, amount DOUBLE,
> category String, product STRING, city STRING, state STRING, spendby STRING)
> PARTITIONED BY (month STRING)
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.083 seconds
hive (surya_training)>
```

```
hive (surya_training)> DESC txnrecsByCat4;
```

```
hive (surya_training)> DESC txnrecsByCat4;
OK
txnno          int
txndate        string
custno         int
amount         double
category       string
product        string
city           string
state          string
spendby        string
month          string

# Partition Information
# col_name      data_type      comment
month          string
Time taken: 0.102 seconds, Fetched: 15 row(s)
hive (surya_training)>
```

→ I3. Load data into partition table (single bucket)















```
hive (surya_training)> FROM txnrecords txn INSERT OVERWRITE TABLE
txnrecsByCat4 PARTITION(month) SELECT txn.txnno,
txn.txndate,txn.custno,
txn.amount,txn.category,txn.product,txn.city,txn.state,
txn.spendby,substring(txn.txndate,1,2) DISTRIBUTE BY
substring(txn.txndate,1,2);
```

```
hive (surya_training)> FROM txnrecords txn INSERT OVERWRITE TABLE txnrecsByCat4 PARTITION(month)
>
> SELECT txn.txnno, txn.txndate,txn.custno, txn.amount,txn.category,txn.product,txn.city,txn.state, txn.spendby,substring(txn.txndate,1,2) DISTRIBU
TE BY substring(txn.txndate,1,2);
Query ID = bigdatalab456422_20230529120834_9198f7d3-58e8-4b16-9558-c7f41086f459
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
23/05/29 12:08:34 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/29 12:08:34 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1684866872278_3639, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_3639/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1684866872278_3639
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-29 12:08:48,735 Stage-1 map = 0%, reduce = 0%
2023-05-29 12:08:58,131 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.06 sec
2023-05-29 12:09:06,365 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.2 sec
MapReduce Total cumulative CPU time: 11 seconds 200 msec
Ended Job = job_1684866872278_3639
Loading data to table surya_training.txnrecsbycat4 partition (month=null)

Time taken to load dynamic partitions: 0.299 seconds
Time taken for adding to write entity : 0.001 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 11.2 sec HDFS Read: 4429675 HDFS Write: 4226779 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 200 msec
OK
Time taken: 36.021 seconds
hive (surya_training)>
```

/user/hive/warehouse/surya_training.db/txnrecsbycat4




[Home](#) / [user](#) / [hive](#) / [warehouse](#) / [surya_training.db](#) / [txnrecsbycat4](#) [Trash](#)

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:08 AM
<input type="checkbox"/>	 .		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=01		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=02		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=03		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=04		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=05		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=06		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=07		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=08		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=09		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=10		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=11		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 month=12		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM

Show of 12 items Page of 1 [⏪](#) [⏴](#) [⏵](#) [⏩](#)

/user/hive/warehouse/surya_training.db/txnrecsbycat4/month=01

[Home](#) / [user](#) / [hive](#) / [warehouse](#) / [surya_training.db](#) / [txnrecsbycat4](#) / [month=01](#) [Trash](#)

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 .		bigdatalab456422	hive	drwxrwxrwx	May 29, 2023 05:09 AM
<input type="checkbox"/>	 000000_0	352.6 KB	bigdatalab456422	hive	-rwxrwxrwx	May 29, 2023 05:09 AM

Show of 1 items Page of 1 [⏪](#) [⏴](#) [⏵](#) [⏩](#)

