→ Bucketing / Clustering

```
Bucketing (clustering)


A,20
B,10
C,5
A,30
B,20
C,40

2 clusters or buckets

Hash Partitioning

hashcode(A) = 200
hashcode(B) = 205
hashcode(C) = 210

hashcode % number of buckets = bucket no

200 % 2 = 0
205 % 2 = 1
210 % 2 = 0
```

```
select a.custno, b.firstname,b.lastname, b.age, b.profession, round(sum(a.amount),2) as amt from txnrecords a join customer b

H1. Create partitioned table
---------------------------------------------
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;


H2. Create partitioned table (with multiple buckets)
-----------------------------------------------
create table txnrecsByCat2(txnno INT, txndate STRING, custno INT, amount DOUBLE,
product STRING, city STRING, state STRING, spendby STRING)
partitioned by (category STRING, state)
//clustered by (state) into 10 buckets
row format delimited
fields terminated by ','
stored as textfile;

H3. Create partitioned table (single bucket) on a derived column
-----------------------------------------------------------------
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 ','
```

a. Max 20,000 partitions are allowed
b. To avoid creating lot of partitions, we create clusters/buckets
c. hash partitioning: decides which record will go into which cluster

→ H2. Create partitioned table (with multiple buckets & partitioning)

```
hive (surya_training)> CREATE TABLE txnrecsByCat2(txnno INT, txndate STRING, custno INT, amount DOUBLE, product STRING, city STRING, state STRING, spendby STRING) PARTITIONED BY (category STRING) CLUSTERED BY (state) INTO 10 buckets ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
hive (surya_training)> CREATE TABLE txnrecsByCat2(txnno INT, txndate STRING, custno INT, amount DOUBLE,
                     >
                     > product STRING, city STRING, state STRING, spendby STRING)
                     >
                     > PARTITIONED BY (category STRING)
                     >
                     > CLUSTERED BY (state) INTO 10 buckets
                     >
                     > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.437 seconds
hive (surya_training)>
```

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

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

→ I2. Load data into partition table (with multiple buckets & partitioning)

```
hive (surya_training)> INSERT OVERWRITE TABLE txnrecsByCat2 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 txnrecsByCat2 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_20230530092516_5c09c625-7ad3-4699-9250-1b6751c81898
Total jobs = 2
Launching Job 1 out of 2
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>
23/05/30 09:25:18 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/30 09:25:18 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_3996, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_3996/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1684866872278_3996
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-05-30 09:27:02,024 Stage-1 map = 0%,  reduce = 0%
2023-05-30 09:27:47,005 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.32 sec
2023-05-30 09:28:13,003 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.6 sec
MapReduce Total cumulative CPU time: 10 seconds 850 msec
Ended Job = job_1684866872278_3996
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 10
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>
23/05/30 09:28:45 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/30 09:28:45 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_4018, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_4018/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1684866872278_4018
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 10
2023-05-30 09:30:42,711 Stage-2 map = 0%,  reduce = 0%
2023-05-30 09:31:52,883 Stage-2 map = 100%,  reduce = 10%, Cumulative CPU 9.69 sec
2023-05-30 09:32:53,034 Stage-2 map = 100%,  reduce = 10%, Cumulative CPU 10.62 sec
2023-05-30 09:32:55,121 Stage-2 map = 100%,  reduce = 30%, Cumulative CPU 21.38 sec
2023-05-30 09:33:55,266 Stage-2 map = 100%,  reduce = 30%, Cumulative CPU 23.27 sec
2023-05-30 09:34:03,604 Stage-2 map = 100%,  reduce = 40%, Cumulative CPU 28.16 sec
2023-05-30 09:34:05,667 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 33.5 sec
2023-05-30 09:35:06,423 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 35.94 sec
2023-05-30 09:35:21,908 Stage-2 map = 100%,  reduce = 60%, Cumulative CPU 41.05 sec
2023-05-30 09:35:22,962 Stage-2 map = 100%,  reduce = 70%, Cumulative CPU 46.61 sec
2023-05-30 09:36:23,097 Stage-2 map = 100%,  reduce = 70%, Cumulative CPU 48.84 sec
2023-05-30 09:36:28,206 Stage-2 map = 100%,  reduce = 80%, Cumulative CPU 54.42 sec
2023-05-30 09:36:35,419 Stage-2 map = 100%,  reduce = 90%, Cumulative CPU 59.17 sec
2023-05-30 09:37:35,844 Stage-2 map = 100%,  reduce = 90%, Cumulative CPU 61.44 sec
2023-05-30 09:37:42,003 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 66.68 sec
2023-05-30 09:38:42,141 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 67.64 sec
MapReduce Total cumulative CPU time: 1 minutes 7 seconds 640 msec
Ended Job = job_1684866872278_4018
Loading data to table surya_training.txnrecsbycat2 partition (category=null)


         Time taken to load dynamic partitions: 0.713 seconds
         Time taken for adding to write entity : 0.004 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.85 sec   HDFS Read: 4427194 HDFS Write: 4868792 HDFS EC Read: 0 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 10   Cumulative CPU: 67.64 sec   HDFS Read: 4920524 HDFS Write: 3510614 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 18 seconds 490 msec
OK
Time taken: 817.442 seconds
hive (surya_training)>
```

- Multiple partitions are created for category
- Each partition has 10 buckets of states in it

/user/hive/warehouse/surya_training.db/txnrecsbycat2

Home / user / hive / warehouse / surya_training.db / **txnrecsbycat2** Trash

Name	Size	User	Group	Permissions	Date
↑		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:48 AM
.		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:38 AM
category=Air Sports		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
category=Combat Sports		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:38 AM
category=Dancing		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:38 AM
category=Exercise & Fitness		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
category=Games		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
category=Gymnastics		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
category=Indoor Games		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
category=Jumping		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
category=Outdoor Play Equipment		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
category=Outdoor Recreation		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:38 AM

/user/hive/warehouse/surya_training.db/txnrecsbycat2/category=Air%20Sports

Home / user / hive / warehouse / surya_training.db / txnrecsbycat2 / **category=Air Sports** Trash

Name	Size	User	Group	Permissions	Date
↑		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:38 AM
.		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:37 AM
000000_0	5.3 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:31 AM
000001_0	5.9 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:32 AM
000002_0	5.5 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:32 AM
000003_0	1.5 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:34 AM
000004_0	7.2 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:34 AM
000005_0	5.7 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:35 AM
000006_0	14.1 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:35 AM
000007_0	10.9 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:36 AM
000008_0	1.1 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:36 AM
000009_1	7.0 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:37 AM

→ modd H2. Create partitioned table (only buckets)

```
hive (surya_training)> CREATE TABLE txn_bucket(txnno INT, txndate STRING, custno INT, amount DOUBLE, category STRING, product STRING, city STRING, state STRING, spendby STRING) CLUSTERED BY (state) INTO 10 buckets ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
hive (surya_training)> CREATE TABLE txn_bucket(txnno INT, txndate STRING, custno INT, amount DOUBLE,
                     >
                     > category STRING, product STRING, city STRING, state STRING, spendby STRING)
                     >
                     > CLUSTERED BY (state) INTO 10 buckets
                     >
                     > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.116 seconds
hive (surya_training)>
```

→ modd I2. Load data into partition table (only buckets)

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

```
hive (surya_training)> INSERT OVERWRITE TABLE txn_bucket
                     >
                     > SELECT * FROM txnrecords;
Query ID = bigdatalab456422_20230530095035_e735f98c-174d-41ee-8193-767036d18e15
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 10
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>
23/05/30 09:50:37 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/30 09:50:37 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_4071, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_4071/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1684866872278_4071
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 10
2023-05-30 09:52:43,812 Stage-1 map = 0%,  reduce = 0%
2023-05-30 09:53:43,943 Stage-1 map = 0%,  reduce = 0%
2023-05-30 09:53:54,258 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.84 sec
2023-05-30 09:54:44,308 Stage-1 map = 100%,  reduce = 13%, Cumulative CPU 9.44 sec
2023-05-30 09:54:45,331 Stage-1 map = 100%,  reduce = 20%, Cumulative CPU 11.27 sec
2023-05-30 09:54:51,526 Stage-1 map = 100%,  reduce = 23%, Cumulative CPU 12.85 sec
2023-05-30 09:54:52,556 Stage-1 map = 100%,  reduce = 27%, Cumulative CPU 14.82 sec
2023-05-30 09:54:55,629 Stage-1 map = 100%,  reduce = 30%, Cumulative CPU 17.57 sec
2023-05-30 09:55:06,950 Stage-1 map = 100%,  reduce = 37%, Cumulative CPU 19.6 sec
2023-05-30 09:55:14,176 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 21.52 sec
2023-05-30 09:55:16,247 Stage-1 map = 100%,  reduce = 47%, Cumulative CPU 23.56 sec
2023-05-30 09:55:18,316 Stage-1 map = 100%,  reduce = 53%, Cumulative CPU 25.91 sec
2023-05-30 09:55:26,535 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 30.76 sec
2023-05-30 09:55:27,562 Stage-1 map = 100%,  reduce = 73%, Cumulative CPU 34.44 sec
2023-05-30 09:55:32,697 Stage-1 map = 100%,  reduce = 77%, Cumulative CPU 36.15 sec
2023-05-30 09:55:35,801 Stage-1 map = 100%,  reduce = 80%, Cumulative CPU 38.96 sec
2023-05-30 09:56:13,203 Stage-1 map = 100%,  reduce = 87%, Cumulative CPU 40.6 sec
2023-05-30 09:56:14,246 Stage-1 map = 100%,  reduce = 93%, Cumulative CPU 42.17 sec
2023-05-30 09:56:24,610 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 46.52 sec

MapReduce Total cumulative CPU time: 46 seconds 520 msec
Ended Job = job_1684866872278_4071
Loading data to table surya_training.txn_bucket
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 10   Cumulative CPU: 46.52 sec   HDFS Read: 4480604 HDFS Write: 4226867 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 46 seconds 520 msec
OK
Time taken: 354.446 seconds
hive (surya_training)>
```

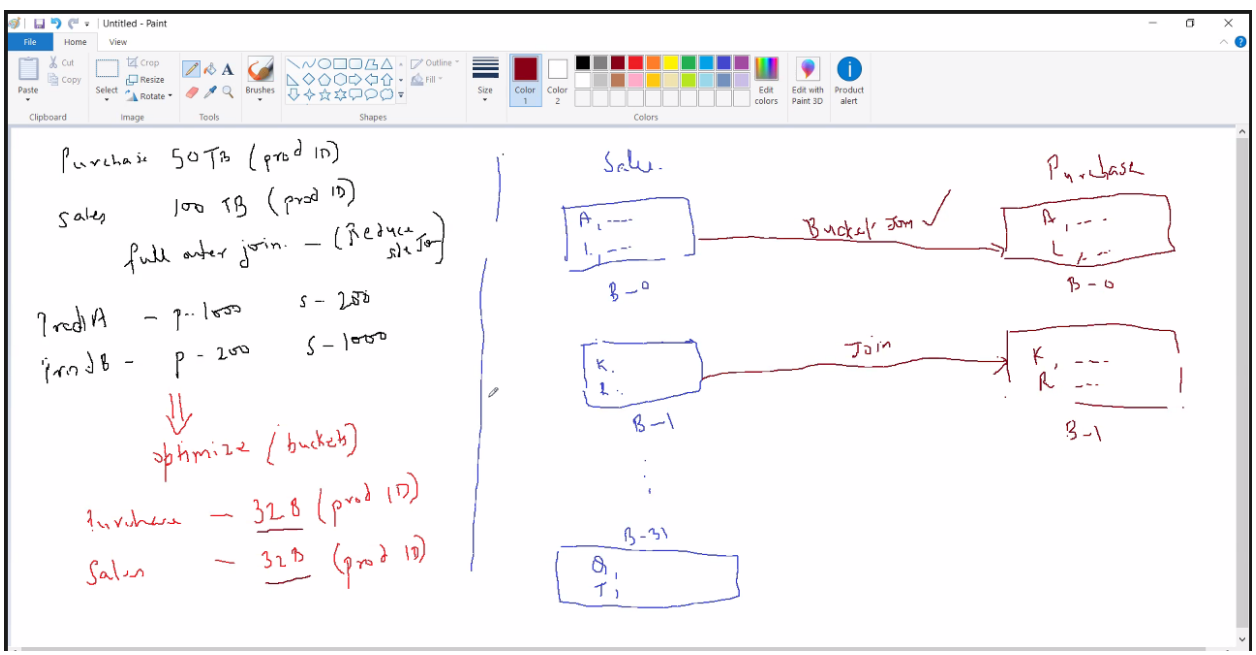
- Only 10 buckets are created for states

/user/hive/warehouse/surya_training.db/txn_bucket

Home / user / hive / warehouse / surya_training.db / **txn_bucket** Trash

Name	Size	User	Group	Permissions	Date
↑		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:48 AM
.		bigdatalab456422	hive	drwxrwxrwx	May 30, 2023 02:56 AM
000000_0	319.5 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:54 AM
000001_0	333.8 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:54 AM
000002_0	373.5 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:54 AM
000003_0	111.4 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:55 AM
000004_0	376.9 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:56 AM
000005_0	373.0 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:55 AM
000006_0	959.2 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:56 AM
000007_0	769.1 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:55 AM
000008_0	75.9 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:55 AM
000009_0	434.6 KB	bigdatalab456422	hive	-rwxrwxrwx	May 30, 2023 02:55 AM

→ Bucket Join



a. Joins of two buckets from two tables where join is based on same key/hash
b. Results into faster join / processing , as only buckets are read for join, instead of entire table while joining

→ Types of Tables

```
A,30
B,20
C,40

2 clusters or buckets

Hash Partitioning

hashcode(A) = 200
hashcode(B) = 205
hashcode(C) = 210

hashcode % number of buckets = bucket no

200 % 2 = 0
205 % 2 = 1
210 % 2 = 0


Types of tables

1) managed table
2) external


User define function
```

a. <u>Managed tables</u>
   i. created by default
   ii. If you drop a managed table, you also lose/delete the data as well as schema from the meta-store.
   iii. And if someone is using the data from that managed table and you delete the table, data becomes inaccessible to user
   iv. Managed by hive

   v. `hive (surya_training)> DESC FORMATTED txnrecords;`

```
hive (surya_training)> DESC FORMATTED txnrecords;
OK
# col_name              data_type               comment

txnno                   int
txndate                 string
custno                  int
amount                  double
category                string
product                 string
city                    string
state                   string
spendby                 string

# Detailed Table Information
Database:               surya_training
OwnerType:              USER
Owner:                  bigdatalab456422
CreateTime:             Mon May 29 09:29:46 UTC 2023
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://nameservice1/user/bigdatalab456422/sales
Table Type:             MANAGED_TABLE
Table Parameters:
        transient_lastDdlTime   1685352586

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             ,
        serialization.format    ,
Time taken: 0.088 seconds, Fetched: 35 row(s)
hive (surya_training)>
```

hive (surya_training)> SHOW TABLES ;

```
hive (surya_training)> SHOW TABLES ;
OK
customer
nyse
stkvol
txn_bucket
txn_orc
txn_parquet
txnrecords
txnrecsbycat
txnrecsbycat2
txnrecsbycat3
txnrecsbycat4
Time taken: 0.035 seconds, Fetched: 11 row(s)
hive (surya_training)>
```

Table exists in location

Home / user / bigdatalab456422 / **sales** Trash

Name	Size	User	Group	Permissions	Date
↑		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 29, 2023 02:29 AM
.		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 29, 2023 02:34 AM
txns1.txt	4.2 MB	bigdatalab456422	bigdatalab456422	-rw-r--r--	May 29, 2023 02:34 AM

Show 45 of 1 items Page 1 of 1

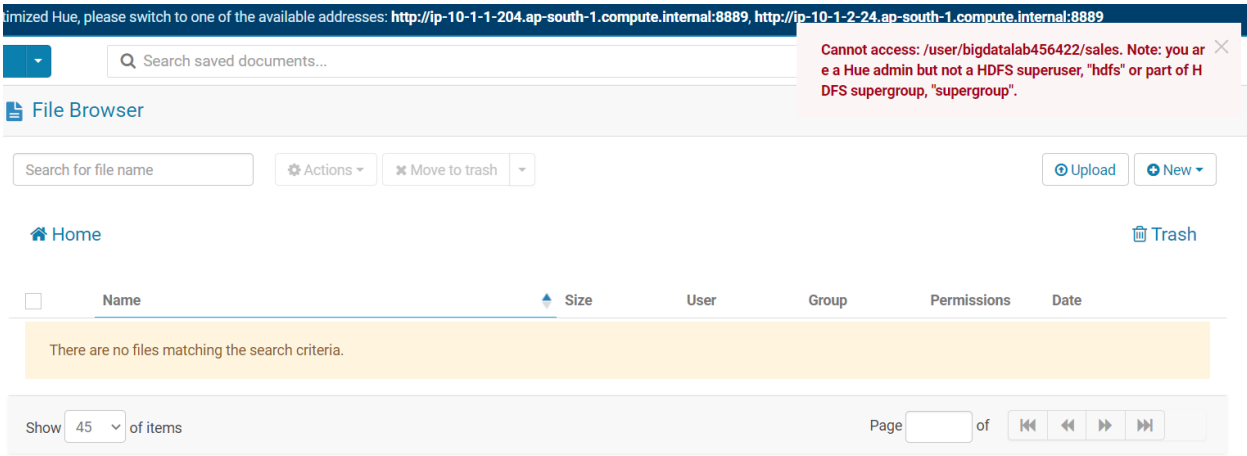
hive (surya_training)> DROP TABLE txnrecords ;

```
hive (surya_training)> DROP TABLE txnrecords ;
OK
Time taken: 0.165 seconds
hive (surya_training)>
```

hive (surya_training)> SHOW TABLES ;

```
hive (surya_training)> SHOW TABLES ;
OK
customer
nyse
stkvol
txn_bucket
txn_orc
txn_parquet
txnrecsbycat
txnrecsbycat2
txnrecsbycat3
txnrecsbycat4
Time taken: 0.035 seconds, Fetched: 10 row(s)
hive (surya_training)>
```

Table is deleted now along with its data, as it shows error while accessing its file



b. <u>External Tables</u>

i. To avoid losing data from the table while only deleting the schema, you use external table

ii. Only structure is deleted from meta-store, but data is kept intact

iii. Data is independent of the schema

iv. `hive (surya_training)> CREATE EXTERNAL 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 EXTERNAL 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: 0.076 seconds
hive (surya_training)>
```

Shows newly created external table

`hive (surya_training)> show tables ;`

```
hive (surya_training)> show tables ;
OK
customer
nyse
stkvol
txn_bucket
txn_orc
txn_parquet
txnrecords
txnrecsbycat
txnrecsbycat2
txnrecsbycat3
txnrecsbycat4
Time taken: 0.036 seconds, Fetched: 11 row(s)
hive (surya_training)>
```

/user/bigdatalab456422/sales

Home / user / bigdatalab456422 / **sales** Trash

Name	Size	User	Group	Permissions	Date
↑		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 30, 2023 03:34 AM
.		bigdatalab456422	bigdatalab456422	drwxr-xr-x	May 30, 2023 03:34 AM

Show 45 of 0 items Page 1 of 1

→ User Defined Function (UDF)

a. Need to create custom defined function using java
b. Create a function UDF to convert UDF to date, time

```
hive (surya_training)> CREATE TABLE testing(id string,unixtime
string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

```
hive (surya_training)> CREATE TABLE testing(id string,unixtime string)
                     >
                     > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.091 seconds
hive (surya_training)>
```

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

```
hive (surya_training)> DESC testing ;
OK
id                      string
unixtime                string
Time taken: 0.061 seconds, Fetched: 2 row(s)
hive (surya_training)>
```

```
hive (surya_training)> LOAD DATA LOCAL INPATH 'counter.txt'
INTO TABLE testing;
```

```
hive (surya_training)> LOAD DATA LOCAL INPATH 'counter.txt' INTO TABLE testing;
Loading data to table surya_training.testing
OK
Time taken: 0.712 seconds
hive (surya_training)>
```

```
SELECT * FROM testing;
```

```
hive (surya_training)> SELECT * FROM testing;
OK
one     1470000000000
two     1389523259550
three   1389523259550
four    1389523259550
five    1479589200000
Time taken: 0.073 seconds, Fetched: 5 row(s)
hive (surya_training)>
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ jar tvf udfhive.jar
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ jar tvf udfhive.jar
    25 Tue May 30 16:59:30 UTC 2023 META-INF/MANIFEST.MF
  1097 Tue May 30 16:57:42 UTC 2023 hive/UnixtimeToDate.class
   756 Tue May 30 16:57:42 UTC 2023 .classpath
   380 Tue May 30 16:50:58 UTC 2023 .project
[bigdatalab456422@ip-10-1-1-204 ~]$
```

```
hive (surya_training)> add jar udfhive.jar;
```

```
hive (surya_training)> add jar udfhive.jar;
Added [udfhive.jar] to class path
Added resources: [udfhive.jar]
hive (surya_training)>
```

```
hive (surya_training)> list jars ;
```

```
hive (surya_training)> list jars ;
udfhive.jar
hive (surya_training)>
```

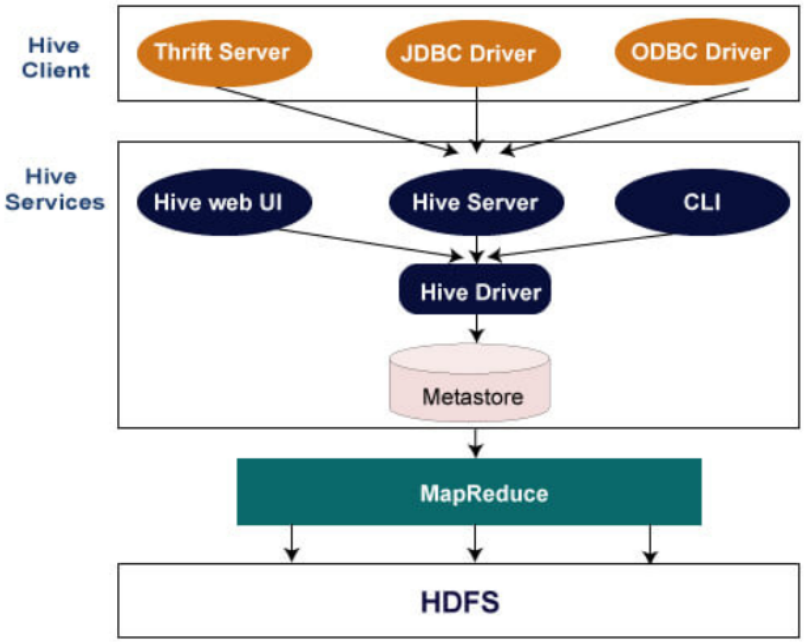
```
hive (surya_training)> CREATE TEMPORARY FUNCTION userdate AS
'hive.UnixtimeToDate';
```

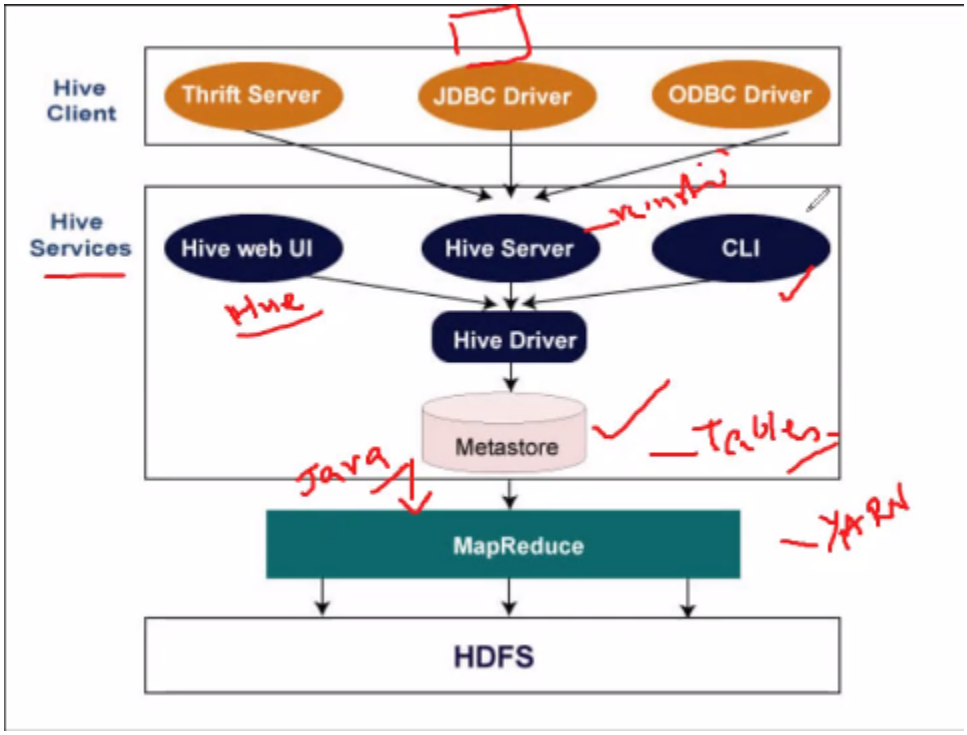
```
hive (surya_training)> CREATE TEMPORARY FUNCTION userdate AS 'hive.UnixtimeToDate';
OK
Time taken: 0.018 seconds
hive (surya_training)>
```

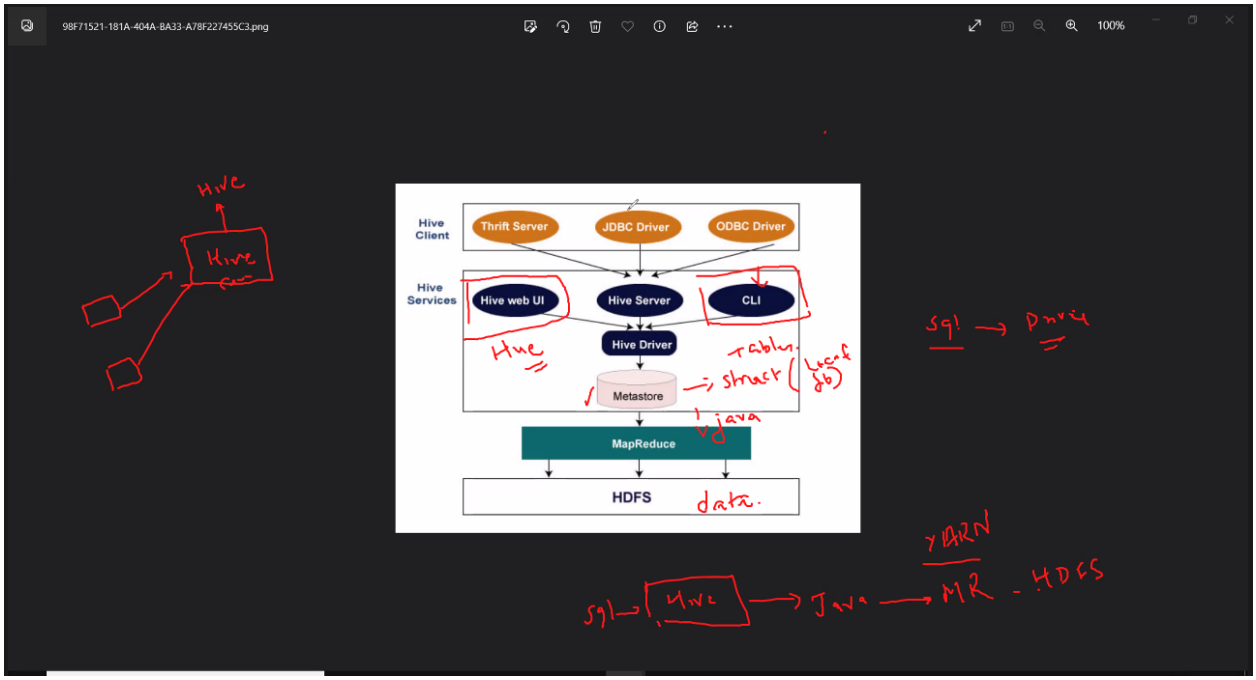
```
hive (surya_training)> SELECT id, userdate(unixtime) FROM
testing;
```

```
hive (surya_training)> SELECT id, userdate(unixtime) FROM testing;
Query ID = bigdatalab456422_2023053011350 8_ccb44fc8-6cdf-43b2-9b72-9ea7f93b189a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/05/30 11:35:09 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/30 11:35:09 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_4107, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1684866872278_4107/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1684866872278_4107
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-05-30 11:35:27,918 Stage-1 map = 0%,  reduce = 0%
2023-05-30 11:35:38,133 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.74 sec
MapReduce Total cumulative CPU time: 3 seconds 740 msec
Ended Job = job_1684866872278_4107
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.74 sec   HDFS Read: 4297 HDFS Write: 255 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 740 msec
OK
one     7/31/16 9:20 PM
two     1/12/14 10:40 AM
three   1/12/14 10:40 AM
four    1/12/14 10:40 AM
five    11/19/16 9:00 PM
Time taken: 31.419 seconds, Fetched: 5 row(s)
hive (surya_training)>
```

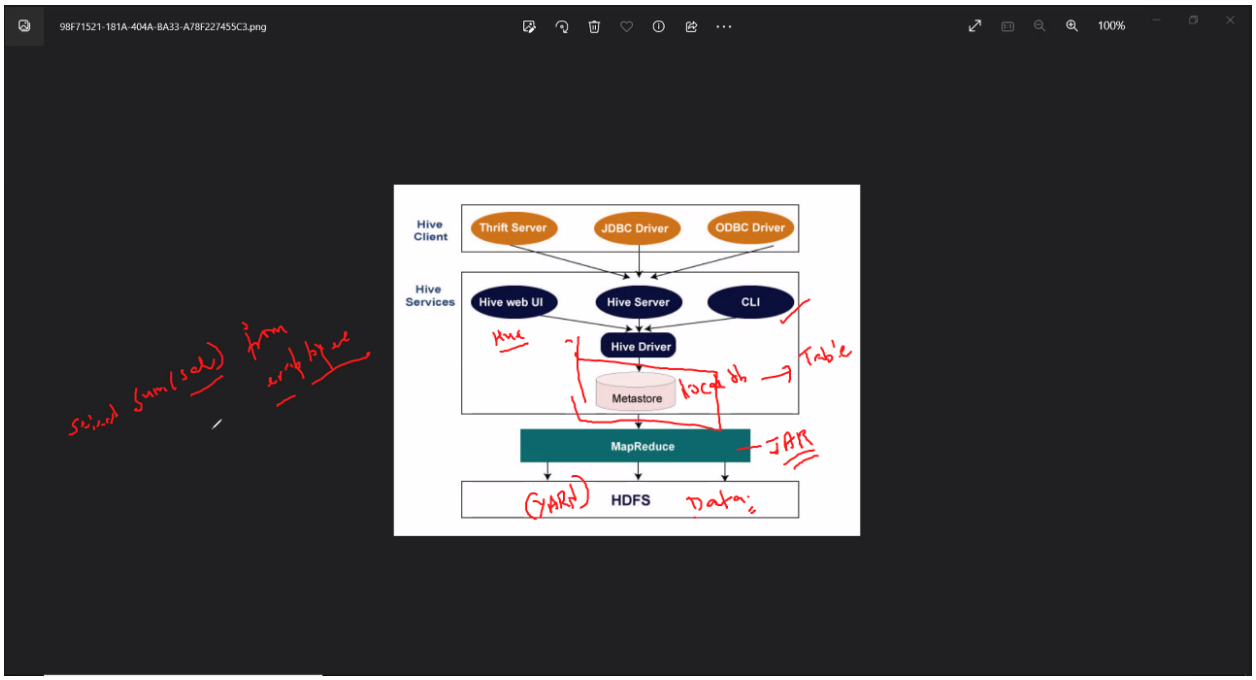
→ Hive Architecture

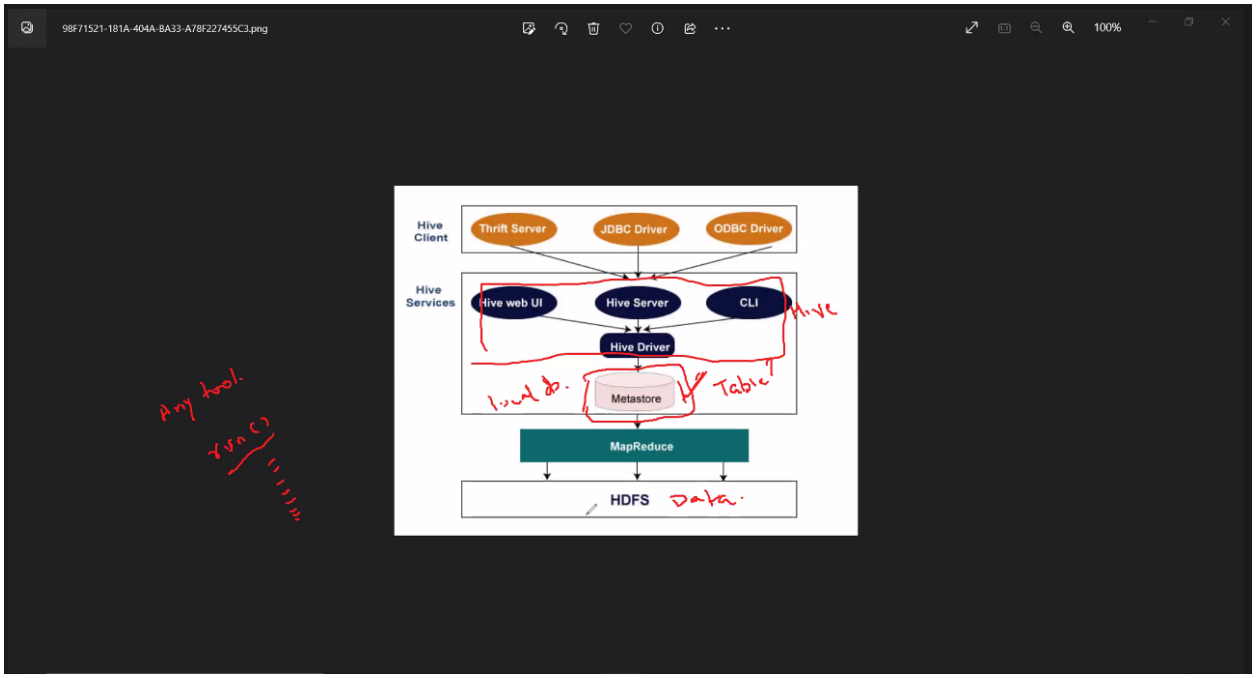

Hive Client
Thrift Server
JDBC Driver
ODBC Driver
Hive Services
Hive web UI
Hive Server
CLI
Hive Driver
Metastore
MapReduce
HDFS


Hive Client
Thrift Server
JDBC Driver
ODBC Driver
Hive Services
Hive web UI
Hive Server
CLI
Hive Driver
Metastore
MapReduce
HDFS



Metastore : local DB → stores table structure
HDFS : contains actual data





→ Indexing

a. Makes query execution faster, but hive has limited indexing capabilities
b.