My Hadoop Workouts - Hive

# ****Environment****

## Hive home directory

[hduser@Inceptez ~]$ hdfs dfs -ls -R /user/hive

drwxrwxr-x - hduser supergroup 0 2016-07-02 10:47 /user/hive/warehouse

drwxr-xr-x - hduser supergroup 0 2016-07-02 10:52 /user/hive/warehouse/retail.db

drwxr-xr-x - hduser supergroup 0 2016-07-02 10:52 /user/hive/warehouse/retail.db/txnrecords

## Metastore

🡺 Start Mysql thru root access using the command “**service mysqld start**”)

[hduser@Inceptez install]$ hive --service metastore

Starting Hive Metastore Server

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/local/hive/lib/hive-jdbc-0.14.0-standalone.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

mysql> select \* from metastore.TBLS;

+--------+-------------+-------+------------------+--------+-----------+-------+----------+---------------+--------------------+--------------------+----------------+

| TBL\_ID | CREATE\_TIME | DB\_ID | LAST\_ACCESS\_TIME | OWNER | RETENTION | SD\_ID | TBL\_NAME | TBL\_TYPE | VIEW\_EXPANDED\_TEXT | VIEW\_ORIGINAL\_TEXT | LINK\_TARGET\_ID |

+--------+-------------+-------+------------------+--------+-----------+-------+----------+---------------+--------------------+--------------------+----------------+

| 1 | 1467522307 | 1 | 0 | hduser | 0 | 1 | hivetest | MANAGED\_TABLE | NULL | NULL | NULL |

| 6 | 1467735040 | 1 | 0 | hduser | 0 | 6 | array | MANAGED\_TABLE | NULL | NULL | NULL |

+--------+-------------+-------+------------------+--------+-----------+-------+----------+---------------+--------------------+--------------------+----------------+

2 rows in set (0.00 sec)

mysql> select \* from metastore.COLUMNS\_V2;

+-------+---------+-------------+---------------+-------------+

| CD\_ID | COMMENT | COLUMN\_NAME | TYPE\_NAME | INTEGER\_IDX |

+-------+---------+-------------+---------------+-------------+

| 1 | NULL | department | string | 4 |

| 1 | NULL | designation | string | 2 |

| 1 | NULL | employeeid | int | 0 |

| 1 | NULL | firstname | string | 1 |

| 1 | NULL | salary | int | 3 |

| 6 | NULL | city | string | 4 |

| 6 | NULL | desig | array<string> | 3 |

| 6 | NULL | id | int | 0 |

| 6 | NULL | name | string | 1 |

| 6 | NULL | sal | bigint | 2 |

+-------+---------+-------------+---------------+-------------+

10 rows in set (0.00 sec)

# ****DDL Statements****

## Create Database

hive> create database retail;

hive> show databases;

default

retail

hive> use retail;

hive> set hive.cli.print.current.db=true;

hive (retail)> set hive.cli.print.current.db=false;

hive> set hive.cli.print.current.db=true;

## Create Table

hive> create table dummy1 (id int, value varchar(10));

hive> show create table dummy1;

CREATE TABLE `dummy1`(

`id` int,

`value` varchar(10))

ROW FORMAT SERDE 🡸 Row Format Serde (default delimiter) is applied when not explicitly specified

'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dummy1'

TBLPROPERTIES (

'transient\_lastDdlTime'='1470231774')

Time taken: 0.377 seconds, Fetched: 13 row(s)

hive> insert into table dummy1 values (1,'test1');

hive> select \* from dummy1;

1 test1

Time taken: 0.286 seconds, Fetched: 1 row(s)

[hduser@Inceptez ~]$ hdfs dfs -cat hdfs://localhost:54310/user/hive/warehouse/dummy1/000000\_0

1test1

hive> create table dummy2 (id int, value varchar(10)) row format delimited fields terminated by ',';

hive> desc dummy2;

id int

value varchar(10)

Time taken: 0.169 seconds, Fetched: 2 row(s)

hive> show create table dummy2;

CREATE TABLE `dummy2`(

`id` int,

`value` varchar(10))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dummy2'

TBLPROPERTIES (

'transient\_lastDdlTime'='1470232230')

Time taken: 0.151 seconds, Fetched: 13 row(s)

[hduser@Inceptez ~]$ hdfs dfs -cat hdfs://localhost:54310/user/hive/warehouse/dummy2/000000\_0

1,test1

### Internal Table

hive (retail)> 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 ','

> lines terminated by '\n'

> stored as textfile;

hive (retail)> show tables in retail;

txnrecords

hive (retail)> show create table txnrecords;

OK

createtab\_stmt

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 ','

LINES TERMINATED BY '\n'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/retail.db/txnrecords'

TBLPROPERTIES (

'COLUMN\_STATS\_ACCURATE'='true',

'numFiles'='1',

'numRows'='0',

'rawDataSize'='0',

'totalSize'='8472073',

'transient\_lastDdlTime'='1467439247')

Time taken: 0.267 seconds, Fetched: 26 row(s)

### External Table

hive (retail)> create external table externaltxnrecords(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/hduser/hiveexternaldata'; 🡸 This folder need not be available @ the time of table creation.

OK

Time taken: 0.194 seconds

hive> create external table players (pl\_id int, pl\_name string, pl\_team string, pl\_salary string, pl\_country string);

OK

CREATE EXTERNAL TABLE `players`(

`pl\_id` int,

`pl\_name` string,

`pl\_team` string,

`pl\_salary` string,

`pl\_country` string)

ROW FORMAT SERDE 🡸 Row Format Serde (default delimiter) is applied when not explicitly specified

'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/players' 🡸 default hive location used when not explicitly specified for an EXTERNAL table

TBLPROPERTIES (

'transient\_lastDdlTime'='1467477662')

Time taken: 0.431 seconds, Fetched: 16 row(s)

#### ****Alter External Table location****

hive> alter table players set location '/user/hduser/hiveexternaldata/players';

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. {0} is not absolute or has no scheme information. Please specify a complete absolute uri with scheme information. /user/hduser/hiveexternaldata/players

hive> alter table players set location 'hdfs://localhost:54310/user/hduser/hiveexternaldata/players';

OK

#### ****Put data into table location****

[hduser@Inceptez data]$ hdfs dfs -put ~/hive\_data\_01/players.txt /user/hduser/hiveexternaldata/players

hive> select \* from players;

OK

10 Sachin India 2500000 India

17 Federer Switzerland 6000000 Switzerland

10 Messi Barcelona 3000000 Argentina

7 Ronaldo Real Madrid 3250000 Portugal

Time taken: 0.742 seconds, Fetched: 4 row(s)

#### ****Remove data from table location****

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -rm /user/hduser/hiveexternaldata/players

16/07/02 23:24:04 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.

Deleted /user/hduser/hiveexternaldata/players

hive> select \* from players;

OK

Time taken: 0.133 seconds

#### ****Add more data into table location****

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -mkdir /user/hduser/hiveexternaldata/players

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -put ~/hive\_data\_01/players\* /user/hduser/hiveexternaldata/players

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls -R /user/hduser/hiveexternaldata/players

-rw-r--r-- 1 hduser supergroup 149 2016-07-02 23:27 /user/hduser/hiveexternaldata/players/players.txt

-rw-r--r-- 1 hduser supergroup 72 2016-07-02 23:27 /user/hduser/hiveexternaldata/players/players1.txt

hive> select \* from players;

OK

10 Sachin India 2500000 India

17 Federer Switzerland 6000000 Switzerland

10 Messi Barcelona 3000000 Argentina

7 Ronaldo Real Madrid 3250000 Portugal

7 Dhoni India 2750000 India

10 Rooney Manchester United 2750000 England

Time taken: 0.059 seconds, Fetched: 6 row(s)

[hduser@Inceptez hive\_data\_01]$ echo 10,'Deepak','Manchester United',1010000,'India' >> players2.txt

[hduser@Inceptez hive\_data\_01]$ echo dummy record >> players2.txt

[hduser@Inceptez hive\_data\_01]$ cat players2.txt

10,Deepak,Manchester United,1010000,India

dummy record

hive> select \* from players;

OK

10 Sachin India 2500000 India

17 Federer Switzerland 6000000 Switzerland

10 Messi Barcelona 3000000 Argentina

7 Ronaldo Real Madrid 3250000 Portugal

7 Dhoni India 2750000 India

10 Rooney Manchester United 2750000 England

10 Deepak Manchester United 1010000 India

NULL NULL NULL NULL NULL 🡸 Incorrect data i.e. not having proper datatype as per table definition.

Time taken: 0.068 seconds, Fetched: 8 row(s)

### Bucketed Table

hive> create table HiveTest

> (EmployeeID Int,FirstName String,Designation String,

> Salary Int,Department String)

> clustered by (department) into 3 buckets

> stored as orc TBLPROPERTIES ('transactional'='true');

### Partitioned Table

hive> create table dkmdb02.ManUtdPlayers1 (player varchar(30), position varchar(20), dob date)

> row format delimited fields terminated by ','

> partitioned by (country varchar(25));;

FAILED: ParseException line 3:0 cannot recognize input near 'partitioned' 'by' '(' in serde properties specification

hive> create table dkmdb02.ManUtdPlayers1 (player varchar(30), position varchar(20), dob date)

> partitioned by (country varchar(25))

> row format delimited fields terminated by ',';

OK

Time taken: 3.811 seconds

hive> show create table dkmdb02.manutdplayers1;

OK

CREATE TABLE `dkmdb02.manutdplayers1`(

`player` varchar(30),

`position` varchar(20),

`dob` date)

PARTITIONED BY (

`country` varchar(25))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/manutdplayers1'

TBLPROPERTIES (

'transient\_lastDdlTime'='1472115978')

Time taken: 0.462 seconds, Fetched: 16 row(s)

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:36 /user/hive/warehouse/dkmdb02.db/manutdplayers1

#### ****Partitioning on multiple columns****

hive> create external table dkmdb02.ManUtdPlayers2 (player varchar(30), dob date) partitioned by (country varchar(25), position varchar(20))

> row format delimited fields terminated by ','

> location '/user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2';

OK

Time taken: 0.335 seconds

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hduser/hiveexternaldata/dkmdb02.db/

drwxr-xr-x - hduser supergroup 0 2016-08-25 01:14 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers

-rw-r--r-- 1 hduser supergroup 23373 2016-08-25 00:58 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers/ManUtdPlayers1.csv

drwxr-xr-x - hduser supergroup 0 2016-08-25 20:58 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2

### Table with Array datatype

hive (default)> create table array(id int,name string,sal bigint,desig array<string>,city string)

row format delimited

fields terminated by ','

collection items terminated by '$';

hive (default)> !cat /home/hduser/hive\_data\_01/array\_ip.data;

rvindh,40000,SWE$Analyst$SAnalyst$PL,hyd

2,Bala,30000,SWE$Analyst$SAnalyst$PM,hyd

3,Chandra,50000,SWE$Analyst$PL$PM,Che

4,Gokul,30000,SWE$Analyst$SAnalyst$,hyd

hive (default)> load data inpath '/user/hduser/data\_01/array\_ip.data' into table array;

hive (default)> select \* from array;

array.id array.name array.sal array.desig array.city

NULL 40000 NULL ["hyd"] NULL

2 Bala 30000 ["SWE","Analyst","SAnalyst","PM"] hyd

3 Chandra 50000 ["SWE","Analyst","PL","PM"] Che

4 Gokul 30000 ["SWE","Analyst","SAnalyst",""] hyd

Time taken: 0.139 seconds, Fetched: 4 row(s)

hive> select desig[0],desig[1],desig[2],desig[3],desig[4] from array;

OK

\_c0 \_c1 \_c2 \_c3 \_c4

SWE Analyst SAnalyst PL NULL

SWE Analyst SAnalyst PM NULL

SWE Analyst PL PM NULL

SWE Analyst SAnalyst NULL

Time taken: 0.163 seconds, Fetched: 4 row(s)

### Table with Struct datatype

hive> create table Struct(id int,name string,sal bigint,addr struct<city:string,state:string,pin:bigint>)

> row format delimited

> fields terminated by ','

> collection items terminated by '$';

hive> load data local inpath '/home/hduser/hive\_data\_01/struct\_ip.data' overwrite into table struct;

hive> select \* from struct;

1 Arvindh 40000 {"city":"Hyderabad","state":"AP","pin":800042}

2 Bala 30000 {"city":"Chennai","state":"TamilNadu","pin":600042}

hive> select addr.city, addr.pin from struct;

Hyderabad 800042

Chennai 600042

### Table with Map datatype

hive> create table map1 (id int,name string,sal bigint,Mark map<string,int>,city string)

> row format delimited

> fields terminated by ','

> collection items terminated by '$'

> map keys terminated by '#';

hive> load data local inpath '/home/hduser/hive\_data\_01/map\_ip.data' overwrite into table map1;

hive> select \* from map1;

1 Arvindh 40000 {"mat":100,"Eng":99} hyd

2 Bala 30000 {"Sci":100,"Eng":99} chn

hive> select mark from map1;

{"mat":100,"Eng":99}

{"Sci":100,"Eng":99}

hive> select mark('mat') from map1;

FAILED: SemanticException [Error 10011]: Line 1:7 Invalid function 'mark' 🡸 Case sensitive

hive> select Mark('mat') from map1;

FAILED: SemanticException [Error 10011]: Line 1:7 Invalid function 'Mark' 🡸 Square braces required

hive> select Mark['mat'] from map1;

100

NULL

hive> select Mark['eng'] from map1;

NULL

NULL

hive> select Mark['Eng'] from map1;

99

99

### Table with XML Serde

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -put serde1.xml /user/hduser/data\_01

[hduser@Inceptez hive\_data\_01]$ hadoop fs -ls -R /user/hduser/data\_01

-rw-r--r-- 1 hduser supergroup 657 2016-07-12 20:14 /user/hduser/data\_01/serde1.xml

[hduser@Inceptez hive\_data\_01]$ hadoop fs -cat /user/hduser/data\_01/serde1.xml

<records>

<record>

<empid>82338</empid>

<empname>Ashfaq</empname>

<income>200000</income>

<age>29</age>

</record>

<record>

<empid>83993</empid>

<empname>Vivek</empname>

<income>300000</income>

<age>33</age>

</record>

<record>

<empid>83994</empid>

<empname>Vishwa</empname>

<income>200000</income>

<age>33</age>

</record>

<record>

<empid>83994</empid>

<empname>Arun</empname>

<income>200000</income>

<age>30</age>

</record>

<record1>

<empid>83994</empid>

<empname>Arun</empname>

<income>200000</income>

<age>30</age>

</record1>

</records>

<recs>

<record2>

<empid>83996</empid>

<empname>Arun</empname>

<income>200000</income>

<age>30</age>

</record2>

</recs>

hive (default)> create external table xml\_Emp1(empid string, empname string,income BIGINT,age int)

> row format serde 'com.ibm.spss.hive.serde2.xml.XmlSerDe'

> WITH SERDEPROPERTIES (

> "column.xpath.empid"="/record/empid/text()",

> "column.xpath.empname"="/record/empname/text()",

> "column.xpath.income"="/record/income/text()",

> "column.xpath.age"="/record/age/text()"

> )

> STORED AS

> INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'

> OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'

> location '/user/hduser/data\_01'

> TBLPROPERTIES ("xmlinput.start"="<record>","xmlinput.end"="</record>");

OK

Time taken: 5.477 seconds

hive (default)> show create table xml\_emp1;

OK

createtab\_stmt

CREATE EXTERNAL TABLE `xml\_emp1`(

`empid` string COMMENT 'from deserializer',

`empname` string COMMENT 'from deserializer',

`income` bigint COMMENT 'from deserializer',

`age` int COMMENT 'from deserializer')

ROW FORMAT DELIMITED

STORED AS INPUTFORMAT

'com.ibm.spss.hive.serde2.xml.XmlInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hduser/data\_01'

TBLPROPERTIES (

'COLUMN\_STATS\_ACCURATE'='false',

'numFiles'='0',

'numRows'='-1',

'rawDataSize'='-1',

'totalSize'='0',

'transient\_lastDdlTime'='1468334761',

'xmlinput.end'='</record>',

'xmlinput.start'='<record>')

Time taken: 0.49 seconds, Fetched: 21 row(s)

hive (default)> select \* from xml\_Emp1;

OK

xml\_emp1.empid xml\_emp1.empname xml\_emp1.income xml\_emp1.age

82338 Ashfaq 200000 29

83993 Vivek 300000 33

83994 Vishwa 200000 33

83994 Arun 200000 30

Time taken: 1.416 seconds, Fetched: 4 row(s)

### Observations

* External table took default hive location when not explicitly specified
* Hive Insert capability from 0.14 version only.

hive> insert into table employee values(1,'Deepak',82000.00,1);

FAILED: ParseException line 1:27 cannot recognize input near 'values' '(' '1' in select clause

[root@sandbox ~]# locate hive-cli

/usr/lib/hive/lib/hive-cli-0.12.0.2.0.6.0-76.jar

/usr/lib/oozie/share/lib/hive/hive-cli-0.12.0.2.0.6.0-76.jar

/var/lib/oozie/oozie-server/webapps/oozie/WEB-INF/lib/hive-cli-0.12.0.2.0.6.0-76.jar

## Storage Formats

### Delimited

hive (dkmdb02)> create table test1 (id int, col1 varchar(6), col2 varchar(6)) row format delimited fields terminated by ',';

hive (dkmdb02)> show create table test1;

createtab\_stmt

CREATE TABLE `test1`(

`id` int,

`col1` varchar(6),

`col2` varchar(6))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test1'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474473425')

Time taken: 0.101 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test1;

hive (dkmdb02)> select \* from test1;

test1.id test1.col1 test1.col2

1 R1C1 R1C2

2 R2C1 R2C2

3 R3C1 R3C2

Time taken: 0.502 seconds, Fetched: 3 row(s)

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls /user/hive/warehouse/dkmdb02.db/test1

Found 1 items

-rw-r--r-- 1 hduser supergroup 39 2016-09-21 21:29 /user/hive/warehouse/dkmdb02.db/test1/test1.txt

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/test1/test1.txt

01,R1C1,R1C2

02,R2C1,R2C2

03,R3C1,R3C2

### Row Format Serde (Default)

hive (dkmdb02)> create table test2 (id int, col1 varchar(6), col2 varchar(6));

hive (dkmdb02)> show create table test2;

createtab\_stmt

CREATE TABLE `test2`(

`id` int,

`col1` varchar(6),

`col2` varchar(6))

ROW FORMAT SERDE

'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test2'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474473818')

Time taken: 0.234 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test2;

>>> Will not load properly as row format is serde

hive (dkmdb02)> select \* from test2;

test2.id test2.col1 test2.col2

NULL NULL NULL

NULL NULL NULL

NULL NULL NULL

Time taken: 0.115 seconds, Fetched: 3 row(s)

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls /user/hive/warehouse/dkmdb02.db/test2

Found 1 items

-rw-r--r-- 1 hduser supergroup 39 2016-09-21 21:34 /user/hive/warehouse/dkmdb02.db/test2/test1.txt

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/test2/test1.txt

01,R1C1,R1C2

02,R2C1,R2C2

03,R3C1,R3C2

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -rm /user/hive/warehouse/dkmdb02.db/test2/test1.txt

Deleted /user/hive/warehouse/dkmdb02.db/test2/test1.txt

hive (dkmdb02)> select \* from test2;

test2.id test2.col1 test2.col2

hive (dkmdb02)> insert into table test2 select \* from test1;

hive (dkmdb02)> select \* from test2;

test2.id test2.col1 test2.col2

1 R1C1 R1C2

2 R2C1 R2C2

3 R3C1 R3C2

Time taken: 0.216 seconds, Fetched: 3 row(s)

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls /user/hive/warehouse/dkmdb02.db/test2

Found 1 items

-rw-r--r-- 1 hduser supergroup 36 2016-09-21 21:39 /user/hive/warehouse/dkmdb02.db/test2/000000\_0

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/test2/000000\_0

1R1C1R1C2

2R2C1R2C2

3R3C1R3C2

### Textfile

hive (dkmdb02)> create table test3 (id int, col1 varchar(6), col2 varchar(6)) row format delimited fields terminated by ',' stored as textfile;

hive (dkmdb02)> show create table test3;

createtab\_stmt

CREATE TABLE `test3`(

`id` int,

`col1` varchar(6),

`col2` varchar(6))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test3'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474474398')

Time taken: 0.107 seconds, Fetched: 14 row(s)

>>> show create table indicates it's same as test1.

### Sequence file

hive (dkmdb02)> create table test4 (id int, col1 varchar(6), col2 varchar(6)) row format delimited fields terminated by ',' stored as sequencefile;

hive (dkmdb02)> show create table test4;

createtab\_stmt

CREATE TABLE `test4`(

`id` int,

`col1` varchar(6),

`col2` varchar(6))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.SequenceFileInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveSequenceFileOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test4'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474475384')

Time taken: 0.1 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test4;

Loading data to table dkmdb02.test4

Failed with exception Wrong file format. Please check the file's format.

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.MoveTask

hive (dkmdb02)> insert into table test4 select \* from test2;

hive (dkmdb02)> select \* from test4;

test4.id test4.col1 test4.col2

1 R1C1 R1C2

2 R2C1 R2C2

3 R3C1 R3C2

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls /user/hive/warehouse/dkmdb02.db/test4

Found 1 items

-rw-r--r-- 1 hduser supergroup 159 2016-09-21 22:03 /user/hive/warehouse/dkmdb02.db/test4/000000\_0

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/test4/000000\_0

SEQ"org.apache.hadoop.io.BytesWritableorg.apache.hadoop.io.Textiݗ▒▒g▒▒)▒▒zˋ7

1,R1C1,R1C2

2,R2C1,R2C2

3,R3C1,R3C2[hduser@Inceptez hive\_data\_01]$

### RC file

hive (dkmdb02)> create table test5 (id int, col1 varchar(6), col2 varchar(6)) row format delimited fields terminated by ',' stored as rcfile;

hive (dkmdb02)> show create table test5;

createtab\_stmt

CREATE TABLE `test5`(

`id` int,

`col1` varchar(6),

`col2` varchar(6))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.RCFileInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.RCFileOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test5'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474475945')

Time taken: 0.1 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test5;

Loading data to table dkmdb02.test5

Failed with exception Wrong file format. Please check the file's format.

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.MoveTask

hive (dkmdb02)> insert into table test5 select \* from test4;

hive (dkmdb02)> select \* from test5;

test5.id test5.col1 test5.col2

1 R1C1 R1C2

2 R2C1 R2C2

3 R3C1 R3C2

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls /user/hive/warehouse/dkmdb02.db/test5

Found 1 items

-rw-r--r-- 1 hduser supergroup 111 2016-09-21 22:11 /user/hive/warehouse/dkmdb02.db/test5/000000\_0

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/test5/000000\_0

RCFhive.io.rcfile.column.number3K!▒Ɇ"w\▒▒▒▒i4+▒

▒

▒R1C1R2C1R3C1R1C2R2C2R3C2[hduser@Inceptez hive\_data\_01]$

>>> RCFILE stores columns of a table in form of record in a columnar manner. It first partitions rows horizontally into row splits and then it vertically partitions each row split in a columnar way. RCFILE first stores the metadata of a row split, as the key part of a record, and all the data of a row split as the value part. This means that RCFILE encourages column oriented storage rather than row oriented storage.

### ORC file

hive (dkmdb02)> create table test6 (id int, col1 varchar(6), col2 varchar(6)) row format delimited fields terminated by ',' stored as orcfile;

hive (dkmdb02)> show create table test6;

createtab\_stmt

CREATE TABLE `test6`(

`id` int,

`col1` varchar(6),

`col2` varchar(6))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test6'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474476407')

Time taken: 0.094 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test6;

hive (dkmdb02)> select \* from test6;

OK

test6.id test6.col1 test6.col2

Failed with exception java.io.IOException:java.io.IOException: Malformed ORC file hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test6/test1.txt. Invalid postscript.

>>> Will not load properly as storage format is ORC

hive (dkmdb02)> insert into table test6 select \* from test5;

hive (dkmdb02)> select \* from test6;

test6.id test6.col1 test6.col2

1 R1C1 R1C2

2 R2C1 R2C2

3 R3C1 R3C2

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls /user/hive/warehouse/dkmdb02.db/test6

Found 1 items

-rw-r--r-- 1 hduser supergroup 375 2016-09-21 22:21 /user/hive/warehouse/dkmdb02.db/test6/000000\_0

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/test6/000000\_0

ORC '

8▒▒be!!f%>.▒ CgC!▒ cgC 8▒▒be!!f%>.▒ Cg#!▒ cg# F$`R1C1R2C1R3C1R1C2R2C2R3C2j▒b▒``▒▒▒ьb`▒IBL3▒iF▒8;▒f▒▒▒L@▒▒g▒▒%X▒▒8▒▒ ▒`▒▒b▒`▒▒▒`b▒``▒▒▒▒`V▒▒b2t6b 2v6▒▒@3▒IH▒▒`8▒(▒▒▒,+▒▒`▒▒$▒▒#▒▒▒▒,▒▒▒▒c▒

!▒▒▒

▒▒

4`▒▒X▒

▒q0 ▒I▒X U▒q▒:

;H ▒▒Ō$$&e▒▒"

(30▒▒ORCPuTTYPuTTY[hduser@Inceptez hive\_data\_01]$ PuTTYPuTTY

### Parquet

hive (dkmdb02)> create table test6 (id int, col1 varchar(6), col2 varchar(6)) row format serde 'parquet.hive.serde.ParquetHiveSerDe';

FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. Cannot validate serde: parquet.hive.serde.ParquetHiveSerDe

hive (dkmdb02)> create table test6 (id int, col1 varchar(6), col2 varchar(6)) row format serde 'parquet.hive.serde.ParquetHiveSerDe' stored as INPUTFORMAT 'parquet.hive.DeprecatedParquetInputFormat' OUTPUTFORMAT 'parquet.hive.DeprecatedParquetOutputFormat';

FAILED: SemanticException Cannot find class 'parquet.hive.DeprecatedParquetInputFormat'

[hduser@Inceptez hive\_data\_01]$ locate parquet

...

/usr/local/nifi/work/nar/extensions/nifi-flume-nar-0.6.0.nar-unpacked/META-INF/bundled-dependencies/parquet-hive-bundle-1.4.1.jar

...

hive (dkmdb02)> list jars;

hive (dkmdb02)> add jar /usr/local/nifi/work/nar/extensions/nifi-flume-nar-0.6.0.nar-unpacked/META-INF/bundled-dependencies/parquet-hive-bundle-1.4.1.jar;

Added [/usr/local/nifi/work/nar/extensions/nifi-flume-nar-0.6.0.nar-unpacked/META-INF/bundled-dependencies/parquet-hive-bundle-1.4.1.jar] to class path

Added resources: [/usr/local/nifi/work/nar/extensions/nifi-flume-nar-0.6.0.nar-unpacked/META-INF/bundled-dependencies/parquet-hive-bundle-1.4.1.jar]

hive (dkmdb02)> list jars;

/usr/local/nifi/work/nar/extensions/nifi-flume-nar-0.6.0.nar-unpacked/META-INF/bundled-dependencies/parquet-hive-bundle-1.4.1.jar

hive (dkmdb02)> create table test7 (id int, col1 varchar(6), col2 varchar(6)) row format serde 'parquet.hive.serde.ParquetHiveSerDe';

hive (dkmdb02)> show create table test7;

createtab\_stmt

CREATE TABLE `test7`(

`id` int COMMENT '',

`col1` varchar(6) COMMENT '',

`col2` varchar(6) COMMENT '')

ROW FORMAT SERDE

'parquet.hive.serde.ParquetHiveSerDe'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test7'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474478075')

Time taken: 0.211 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test7;

hive (dkmdb02)> select \* from test7;

test7.id test7.col1 test7.col2

NULL NULL NULL

NULL NULL NULL

NULL NULL NULL

>>> Will not load properly as Row format is Parquet

hive (dkmdb02)> insert into table test7 select \* from test6;

FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask

hive (dkmdb02)> create table test8 (id int, col1 varchar(6), col2 varchar(6)) row format serde 'parquet.hive.serde.ParquetHiveSerDe' stored as INPUTFORMAT 'parquet.hive.DeprecatedParquetInputFormat' OUTPUTFORMAT 'parquet.hive.DeprecatedParquetOutputFormat';

hive (dkmdb02)> show create table test8;

OK

createtab\_stmt

CREATE TABLE `test8`(

`id` int COMMENT '',

`col1` varchar(6) COMMENT '',

`col2` varchar(6) COMMENT '')

ROW FORMAT SERDE

'parquet.hive.serde.ParquetHiveSerDe'

STORED AS INPUTFORMAT

'parquet.hive.DeprecatedParquetInputFormat'

OUTPUTFORMAT

'parquet.hive.DeprecatedParquetOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test8'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474478718')

Time taken: 0.12 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test8;

hive (dkmdb02)> select \* from test8;

OK

test8.id test8.col1 test8.col2

Failed with exception java.io.IOException:java.lang.RuntimeException: hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test8/test1.txt is not a Parquet file. expected magic number at tail [80, 65, 82, 49] but found [51, 67, 50, 10]

>>> Will not load properly as storage format is Parquet

hive (dkmdb02)> insert into table test8 select \* from test6;

hive (dkmdb02)> select \* from test8;

OK

test8.id test8.col1 test8.col2

SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".

SLF4J: Defaulting to no-operation (NOP) logger implementation

SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.

Failed with exception java.io.IOException:java.lang.RuntimeException: hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test8/test1.txt is not a Parquet file. expected magic number at tail [80, 65, 82, 49] but found [51, 67, 50, 10]

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -ls /user/hive/warehouse/dkmdb02.db/test8

Found 2 items

-rw-r--r-- 1 hduser supergroup 418 2016-09-21 22:57 /user/hive/warehouse/dkmdb02.db/test8/000000\_0

-rw-r--r-- 1 hduser supergroup 39 2016-09-21 22:56 /user/hive/warehouse/dkmdb02.db/test8/test1.txt

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -rm /user/hive/warehouse/dkmdb02.db/test8/test1.txt

Deleted /user/hive/warehouse/dkmdb02.db/test8/test1.txt

hive (dkmdb02)> select \* from test8;

OK

test8.id test8.col1 test8.col2

1 R1C1 R1C2

2 R2C1 R2C2

3 R3C1 R3C2

Time taken: 0.109 seconds, Fetched: 3 row(s)

[hduser@Inceptez hive\_data\_01]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/test8/000000\_0

PAR1$$<<R3C1R1C1R1C1R2C1R3C1<<R3C2R1C2R1C2R2C2R3C2LH

hive\_schema%id

%col1%

%col2%<idff<&n

col1~~&n<R3C1R1C1&▒

col2~~&▒<R3C2R1C2▒(parquet-mr version 1.5.0▒PAR1[hduser@Inceptez hive\_data\_01]$

### Avro

hive (dkmdb02)> create table test9 (id int, col1 varchar(6), col2 varchar(6)) row format delimited fields terminated by ',' stored as avro;

hive (dkmdb02)> show create table test9 ;

OK

createtab\_stmt

CREATE TABLE `test9`(

`id` int COMMENT 'from deserializer',

`col1` varchar(6) COMMENT 'from deserializer',

`col2` varchar(6) COMMENT 'from deserializer')

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/test9'

TBLPROPERTIES (

'transient\_lastDdlTime'='1474556659')

Time taken: 0.366 seconds, Fetched: 14 row(s)

hive (dkmdb02)> load data local inpath '/home/hduser/hive\_data\_01/test1.txt' into table test9;

hive (dkmdb02)> select \* from test9;

OK

test9.id test9.col1 test9.col2

Failed with exception java.io.IOException:java.io.IOException: Not a data file.

>>> Will not load properly as storage format is Avro

### JSON Serde

hive> CREATE TABLE clickstr (a string, c string, nk int, tz string, gr string, g string, h string,l string, hh string, r string, u string, t int, hc int, cy string, al string) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe';

hive (default)> show create table clickstr;

OK

createtab\_stmt

CREATE TABLE `clickstr`(

`a` string COMMENT 'from deserializer',

`c` string COMMENT 'from deserializer',

`nk` int COMMENT 'from deserializer',

`tz` string COMMENT 'from deserializer',

`gr` string COMMENT 'from deserializer',

`g` string COMMENT 'from deserializer',

`h` string COMMENT 'from deserializer',

`l` string COMMENT 'from deserializer',

`hh` string COMMENT 'from deserializer',

`r` string COMMENT 'from deserializer',

`u` string COMMENT 'from deserializer',

`t` int COMMENT 'from deserializer',

`hc` int COMMENT 'from deserializer',

`cy` string COMMENT 'from deserializer',

`al` string COMMENT 'from deserializer')

ROW FORMAT SERDE

'com.cloudera.hive.serde.JSONSerDe'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/clickstr'

TBLPROPERTIES (

'transient\_lastDdlTime'='1469722576')

Time taken: 0.379 seconds, Fetched: 26 row(s)

LOAD DATA LOCAL INPATH '/home/hduser/mydata01/UsaGovData.txt' OVERWRITE INTO TABLE clickstr;

hive (default)> LOAD DATA LOCAL INPATH '/home/hduser/mydata01/UsaGovData.txt' OVERWRITE INTO TABLE clickstr;

Loading data to table default.clickstr

Table default.clickstr stats: [numFiles=1, numRows=0, totalSize=832181, rawDataSize=0]

OK

Time taken: 1.555 seconds

hive (default)> CREATE EXTERNAL TABLE IF NOT EXISTS clickstrpart (a string, nk int,g string, h string,l string,hh string, r string ,u string, t bigint,gr string, cy string, tz string,hc bigint,al string)

> PARTITIONED BY (c string) > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

hive (default)> show create table clickstrpart;

OK

createtab\_stmt

CREATE EXTERNAL TABLE `clickstrpart`(

`a` string,

`nk` int,

`g` string,

`h` string,

`l` string,

`hh` string,

`r` string,

`u` string,

`t` bigint,

`gr` string,

`cy` string,

`tz` string,

`hc` bigint,

`al` string)

PARTITIONED BY (

`c` string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/clickstrpart'

TBLPROPERTIES (

'transient\_lastDdlTime'='1469723018')

Time taken: 0.109 seconds, Fetched: 27 row(s)

[hduser@Inceptez ~]$ hdfs dfs -ls /user/hive/warehouse

drwxr-xr-x - hduser supergroup 0 2016-07-28 21:50 /user/hive/warehouse/clickstr

drwxr-xr-x - hduser supergroup 0 2016-07-28 21:53 /user/hive/warehouse/clickstrpart

[hduser@Inceptez ~]$ hdfs dfs -ls -R /user/hive/warehouse/click\*

-rw-r--r-- 1 hduser supergroup 832181 2016-07-28 21:50 /user/hive/warehouse/clickstr/UsaGovData.txt

hive (default)> SET hive.exec.dynamic.partition=true;

hive (default)> set hive.exec.dynamic.partition.mode=nonstrict;

hive (default)> INSERT OVERWRITE TABLE clickstrpart PARTITION(c) SELECT a,nk,g,h,l,hh,r,u,t,gr,cy,tz,hc,al,c FROM clickstr;

FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 HDFS Read: 0 HDFS Write: 0 FAIL

hive (default)> list jars;

hive (default)> add jar /usr/local/hive/lib/hive-serdes-1.0-SNAPSHOT.jar;

Added [/usr/local/hive/lib/hive-serdes-1.0-SNAPSHOT.jar] to class path

Added resources: [/usr/local/hive/lib/hive-serdes-1.0-SNAPSHOT.jar]

hive (default)> list jars;

/usr/local/hive/lib/hive-serdes-1.0-SNAPSHOT.jar

hive (default)> INSERT OVERWRITE TABLE clickstrpart PARTITION(c) SELECT a,nk,g,h,l,hh,r,u,t,gr,cy,tz,hc,al,c FROM clickstr;

[hduser@Inceptez ~]$ hdfs dfs -ls -R /user/hive/warehouse/click\*

-rw-r--r-- 1 hduser supergroup 832181 2016-07-28 21:50 /user/hive/warehouse/clickstr/UsaGovData.txt

drwxr-xr-x - hduser supergroup 0 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=AR

-rw-r--r-- 1 hduser supergroup 2596 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=AR/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=AU

-rw-r--r-- 1 hduser supergroup 7596 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=AU/000000\_0

… more partitions

drwxr-xr-x - hduser supergroup 0 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=VN

-rw-r--r-- 1 hduser supergroup 1082 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=VN/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=ZA

-rw-r--r-- 1 hduser supergroup 1006 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=ZA/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=\_\_HIVE\_DEFAULT\_PARTITION\_\_

-rw-r--r-- 1 hduser supergroup 88981 2016-07-28 22:05 /user/hive/warehouse/clickstrpart/c=\_\_HIVE\_DEFAULT\_PARTITION\_\_/000000\_0

hive (default)> SELECT c,u,count(\*) as cnt from clickstrpart GROUP BY c,u ORDER BY cnt DESC LIMIT 10;

c u cnt

\_\_HIVE\_DEFAULT\_PARTITION\_\_ NULL 120

US http://www.cpc.ncep.noaa.gov/products/predictions/long\_range/seasonal.php?lead=1 87

US http://bachmann.house.gov/press-release/bachmann-introduces-legislation-rescind-congressional-pay-raise 81

US http://www.nasa.gov/mission\_pages/station/expeditions/expedition34/newyear.html 77

\_\_HIVE\_DEFAULT\_PARTITION\_\_ http://www.nasa.gov/mission\_pages/station/expeditions/expedition34/newyear.html 53

GB http://www.nasa.gov/mission\_pages/station/expeditions/expedition34/newyear.html 52

US http://paul.house.gov/index.php?option=com\_content&view=article&id=2038:new-years-resolutions-for-congress&catid=64:2012-texas-straight-talk&Itemid=69 51

US http://www.nasa.gov/multimedia/imagegallery/image\_feature\_2419.html 31

US http://studentaid.ed.gov/fafsa/filling-out#documents 31

US http://www.speaker.gov/press-release/boehner-president-obama-should-lead-not-cast-blame 17

Time taken: 62.9 seconds, Fetched: 10 row(s)

hive (default)> SELECT u,count(\*) as cnt from clickstrpart GROUP BY u ORDER BY cnt DESC LIMIT 10;

u cnt

http://www.nasa.gov/mission\_pages/station/expeditions/expedition34/newyear.html 295

NULL 120

http://www.cpc.ncep.noaa.gov/products/predictions/long\_range/seasonal.php?lead=1 105

http://bachmann.house.gov/press-release/bachmann-introduces-legislation-rescind-congressional-pay-raise 98

http://www.nasa.gov/multimedia/imagegallery/image\_feature\_2419.html 72

http://paul.house.gov/index.php?option=com\_content&view=article&id=2038:new-years-resolutions-for-congress&catid=64:2012-texas-straight-talk&Itemid=69 59

http://studentaid.ed.gov/fafsa/filling-out#documents 37

http://earthobservatory.nasa.gov/NaturalHazards/view.php?id=80052&src=twitter-nh 22

http://www.nasa.gov/news/whats\_next\_2013.html 20

http://www.speaker.gov/press-release/boehner-president-obama-should-lead-not-cast-blame 20

Time taken: 57.932 seconds, Fetched: 10 row(s)

## Using UDFs

### UDF using Java

Write your own UDF as java class and save it as a jar

**Helloworld.class**

**package** Inceptez.Training.Hello;

**import** org.apache.hadoop.hive.ql.exec.UDF;

**import** org.apache.hadoop.io.Text;

**public** **class** helloword

**extends** UDF

{

**public** Text evaluate(Text input)

{

**return** **new** Text("Hello " + input.toString());

}

}

hive (default)> ADD JAR /home/hduser/hive/HiveUdf.jar;

Added [/home/hduser/hive/HiveUdf.jar] to class path

Added resources: [/home/hduser/hive/HiveUdf.jar]

hive (default)> create function hello as 'Inceptez.Training.Hello.helloword';

OK

Time taken: 0.575 seconds

hive (default)> select hello(empname) from xml\_Emp1;

OK

\_c0

Hello Ashfaq

Hello Vivek

Hello Vishwa

Hello Arun

Time taken: 0.403 seconds, Fetched: 4 row(s)

### UDF using Python

hive (default)> select \* from u\_data;

OK

u\_data.userid u\_data.movieid u\_data.rating u\_data.unixtime

1 101 6 1369721454

2 102 7 1369721455

3 103 8 1369721456

4 104 9 1369721457

5 105 9 1369721458

Time taken: 0.149 seconds, Fetched: 5 row(s)

/home/hduser/hive\_udfs

[hduser@Inceptez hive\_udfs]$ cat weekday\_mapper.py

import sys

import datetime

for line in sys.stdin:

line = line.strip()

userid, movieid, rating, unixtime = line.split('\t')

weekday = datetime.datetime.fromtimestamp(float(unixtime)).isoweekday()

print '\t'.join([userid, movieid, rating, str(weekday)])

hive (default)> add file /home/hduser/hive\_udfs/weekday\_mapper.py;

Added resources: [/home/hduser/hive\_udfs/weekday\_mapper.py]

hive (default)> SELECT

> TRANSFORM (userid, movieid, rating, unixtime)

> USING 'python weekday\_mapper.py'

> AS (userid, movieid, rating, weekday)

> FROM u\_data;

Query ID = hduser\_20160715214646\_0c1e8e88-21f5-4685-9622-ea0589c2a6b0

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1468591924483\_0007, Tracking URL = http://Inceptez:8088/proxy/application\_1468591924483\_0007/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1468591924483\_0007

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0

2016-07-15 21:46:38,403 Stage-1 map = 0%, reduce = 0%

2016-07-15 21:46:54,729 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.66 sec

MapReduce Total cumulative CPU time: 1 seconds 660 msec

Ended Job = job\_1468591924483\_0007

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Cumulative CPU: 1.66 sec HDFS Read: 308 HDFS Write: 50 SUCCESS

Total MapReduce CPU Time Spent: 1 seconds 660 msec

OK

userid movieid rating weekday

1 101 6 2

2 102 7 2

3 103 8 2

4 104 9 2

5 105 9 2

Time taken: 41.642 seconds, Fetched: 5 row(s)

# DML Statements

## Load data into Table

hive (retail)> LOAD DATA LOCAL INPATH '/home/hduser/hive/data/txns' OVERWRITE INTO TABLE txnrecords; 🡸 Local file

Loading data to table retail.txnrecords

Table retail.txnrecords stats: [numFiles=1, numRows=0, totalSize=8472073, rawDataSize=0]

OK

hive (dkmdb01)> load data inpath 'user/hduser/data\_01/players1v2.txt' into table players1;

FAILED: SemanticException Line 1:17 Invalid path ''user/hduser/data\_01/players1v2.txt'': No files matching path hdfs://localhost:54310/user/hduser/user/hduser/data\_01/players1v2.txt

hive (dkmdb01)> load data inpath 'data\_01/players1v2.txt' into table players1; 🡸 HDFS file

Loading data to table dkmdb01.players1

Table dkmdb01.players1 stats: [numFiles=2, totalSize=170]

OK

## Static Partitioning using Insert

hive> insert into table dkmdb02.manutdplayers1 partition(country='Germany') select player, position, dob from default.manutdplayers where country = 'Germany';;

Query ID = hduser\_20160825144646\_1473db54-34de-481d-acf1-91ceaac76282

Total jobs = 3

Launching Job 1 out of 3

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1472113754651\_0002, Tracking URL = http://Inceptez:8088/proxy/application\_1472113754651\_0002/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1472113754651\_0002

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0

2016-08-25 14:47:44,439 Stage-1 map = 0%, reduce = 0%

2016-08-25 14:48:10,918 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.62 sec

MapReduce Total cumulative CPU time: 4 seconds 620 msec

Ended Job = job\_1472113754651\_0002

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: hdfs://localhost:54310/tmp/hive/hduser/de761e10-d0f1-4bc7-9946-8cb52dc08556/hive\_2016-08-25\_14-46-52\_042\_6115899871486036047-1/-ext-10000

Loading data to table dkmdb02.manutdplayers1 partition (country=Germany)

[Error 30017]: Skipping stats aggregation by error org.apache.hadoop.hive.ql.metadata.HiveException: [Error 30001]: StatsPublisher cannot be initialized. There was a error in the initialization of StatsPublisher, and retrying might help. If you dont want the query to fail because accurate statistics could not be collected, set hive.stats.reliable=false

Partition dkmdb02.manutdplayers1{country=Germany} stats: [numFiles=1, numRows=-1, totalSize=85, rawDataSize=-1]

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Cumulative CPU: 4.62 sec HDFS Read: 23622 HDFS Write: 85 SUCCESS

Total MapReduce CPU Time Spent: 4 seconds 620 msec

OK

Time taken: 82.25 seconds

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany

-rw-r--r-- 1 hduser supergroup 85 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany/000000\_0

[hduser@Inceptez Sample\_Data]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany/000000\_0

Bastian Schweinsteiger,Midfielder,1984-08-01

Ron-Robert Zieler,Goalkeeper,1989-02-12

hive> show partitions dkmdb02.manutdplayers1;

OK

country=Germany

Time taken: 0.287 seconds, Fetched: 1 row(s)

## Static Partitioning using Alter

hive> alter table dkmdb02.manutdplayers1 add partition (country = 'USA') location '/user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA';

OK

Time taken: 0.358 seconds

hive> show partitions dkmdb02.manutdplayers1;

OK

country=Germany

country=USA

Time taken: 0.272 seconds, Fetched: 2 row(s)

hive> select \* from dkmdb02.manutdplayers1;

OK

Bastian Schweinsteiger Midfielder 1984-08-01 Germany

Ron-Robert Zieler Goalkeeper 1989-02-12 Germany

Time taken: 0.288 seconds, Fetched: 2 row(s)

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:53 /user/hive/warehouse/dkmdb02.db/manutdplayers1

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany

-rw-r--r-- 1 hduser supergroup 85 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:53 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA

[hduser@Inceptez Sample\_Data]$ hdfs dfs -put MU-USA.csv /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:53 /user/hive/warehouse/dkmdb02.db/manutdplayers1

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany

-rw-r--r-- 1 hduser supergroup 85 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:55 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA

-rw-r--r-- 1 hduser supergroup 149 2016-08-25 14:55 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA/MU-USA.csv

[hduser@Inceptez Sample\_Data]$ hdfs dfs -put -f MU-USA.csv /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA

[hduser@Inceptez Sample\_Data]$ hdfs dfs -cat /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA/MU-USA.csv

Jeff Baicher,Forward,1968-11-16

James Brown,Forward,1908-12-31

Tim Howard,Goalkeeper,1979-03-06

Jonathan Spector,Defender,1986-03-01

hive> select \* from dkmdb02.manutdplayers1;

OK

Bastian Schweinsteiger Midfielder 1984-08-01 Germany

Ron-Robert Zieler Goalkeeper 1989-02-12 Germany

Jeff Baicher Forward 1968-11-16 USA

James Brown Forward 1908-12-31 USA

Tim Howard Goalkeeper 1979-03-06 USA

Jonathan Spector Defender 1986-03-01 USA

Time taken: 0.241 seconds, Fetched: 6 row(s)

## Dynamic Partitioning using Insert

hive> insert into table dkmdb02.manutdplayers1 partition(country)

> select player, position, dob from default.manutdplayers where country in ('Argentina','France','Netherlands');

FAILED: SemanticException [Error 10044]: Line 1:18 Cannot insert into target table because column number/types are different 'country': Table insclause-0 has 4 columns, but query has 3 columns.

hive> insert into table dkmdb02.manutdplayers1 partition(country)

> select player, position, dob, country from default.manutdplayers where country in ('Argentina','France','Netherlands');

Query ID = hduser\_20160825150909\_777bcfde-1d9e-4c0b-9154-c99bf6f0771e

Total jobs = 3

Launching Job 1 out of 3

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1472113754651\_0003, Tracking URL = http://Inceptez:8088/proxy/application\_1472113754651\_0003/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1472113754651\_0003

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0

2016-08-25 15:09:51,511 Stage-1 map = 0%, reduce = 0%

2016-08-25 15:10:13,459 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.37 sec

MapReduce Total cumulative CPU time: 4 seconds 370 msec

Ended Job = job\_1472113754651\_0003

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: hdfs://localhost:54310/tmp/hive/hduser/de761e10-d0f1-4bc7-9946-8cb52dc08556/hive\_2016-08-25\_15-09-07\_441\_2348231848106717214-1/-ext-10000

Loading data to table dkmdb02.manutdplayers1 partition (country=null)

Time taken for load dynamic partitions : 824

Loading partition {country=Argentina}

Loading partition {country=Netherlands}

Loading partition {country=France}

Time taken for adding to write entity : 1

[Error 30017]: Skipping stats aggregation by error org.apache.hadoop.hive.ql.metadata.HiveException: [Error 30001]: StatsPublisher cannot be initialized. There was a error in the initialization of StatsPublisher, and retrying might help. If you dont want the query to fail because accurate statistics could not be collected, set hive.stats.reliable=false

Partition dkmdb02.manutdplayers1{country=Argentina} stats: [numFiles=1, numRows=-1, totalSize=223, rawDataSize=-1]

Partition dkmdb02.manutdplayers1{country=France} stats: [numFiles=1, numRows=-1, totalSize=421, rawDataSize=-1]

Partition dkmdb02.manutdplayers1{country=Netherlands} stats: [numFiles=1, numRows=-1, totalSize=404, rawDataSize=-1]

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Cumulative CPU: 4.37 sec HDFS Read: 23622 HDFS Write: 1048 SUCCESS

Total MapReduce CPU Time Spent: 4 seconds 370 msec

OK

Time taken: 70.006 seconds

hive> show partitions dkmdb02.manutdplayers1;

OK

country=Argentina

country=France

country=Germany

country=Netherlands

country=USA

Time taken: 0.083 seconds, Fetched: 5 row(s)

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

16/08/25 15:11:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Argentina

-rw-r--r-- 1 hduser supergroup 223 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Argentina/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=France

-rw-r--r-- 1 hduser supergroup 421 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=France/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany

-rw-r--r-- 1 hduser supergroup 85 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Netherlands

-rw-r--r-- 1 hduser supergroup 404 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Netherlands/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:00 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA

-rw-r--r-- 1 hduser supergroup 133 2016-08-25 15:00 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA/MU-USA.csv

**Note:** Both the properties needed for dynamic partition have already been enabled in the hive-site.xml

<property>

<name>hive.exec.dynamic.partition</name>

<value>**true**</value>

<description>Whether or not to allow dynamic partitions in DML/DDL.</description>

</property>

<property>

<name>hive.exec.dynamic.partition.mode</name>

<value>**nonstrict**</value>

<description>

In strict mode, the user must specify at least one static partition

in case the user accidentally overwrites all partitions.

In nonstrict mode all partitions are allowed to be dynamic.

</description>

hive> set hive.exec.dynamic.partition=false;

hive> insert into table dkmdb02.manutdplayers1 partition(country)

> select player, position, dob, country from default.manutdplayers where country = 'Spain';

FAILED: SemanticException 1:18 Dynamic partition is disabled. Either enable it by setting hive.exec.dynamic.partition=true or specify partition column values. Error encountered near token 'country'

hive> set hive.exec.dynamic.partition=true;

hive> set hive.exec.dynamic.partition.mode=strict;

hive> insert into table dkmdb02.manutdplayers1 partition(country)

> select player, position, dob, country from default.manutdplayers where country = 'Spain';

FAILED: SemanticException [Error 10096]: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive.exec.dynamic.partition.mode=nonstrict

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

hive> insert into table dkmdb02.manutdplayers1 partition(country)

> select player, position, dob, country from default.manutdplayers where country = 'Spain';

OK

Time taken: 50.206 seconds

hive> show partitions dkmdb02.manutdplayers1;

OK

country=Argentina

country=France

country=Germany

country=Netherlands

country=Spain

country=USA

Time taken: 0.161 seconds, Fetched: 6 row(s)

[hduser@Inceptez ~]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

drwxr-xr-x - hduser supergroup 0 2016-08-25 20:08 /user/hive/warehouse/dkmdb02.db/manutdplayers1

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Argentina

-rw-r--r-- 1 hduser supergroup 223 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Argentina/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=France

-rw-r--r-- 1 hduser supergroup 421 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=France/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany

-rw-r--r-- 1 hduser supergroup 85 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Netherlands

-rw-r--r-- 1 hduser supergroup 404 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Netherlands/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 20:08 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Spain

-rw-r--r-- 1 hduser supergroup 189 2016-08-25 20:08 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Spain/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:00 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA

-rw-r--r-- 1 hduser supergroup 133 2016-08-25 15:00 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA/MU-USA.csv

#### ****Partitioning on multiple columns****

hive> set hive.exec.dynamic.partition.mode=strict;

hive> insert into table dkmdb02.manutdplayers2 partition (country='England', position) select player, dob, country, position from default.ManUtdPlayers where country = 'England' and position <> 'Forward';

FAILED: SemanticException [Error 10044]: Line 1:18 Cannot insert into target table because column number/types are different 'position': Table insclause-0 has 3 columns, but query has 4 columns.

hive> insert into table dkmdb02.manutdplayers2 partition (country='England', position) select player, dob, position from default.ManUtdPlayers where country = 'England' and position <> 'Forward';

Query ID = hduser\_20160825211111\_c2f8c3bd-9e4b-4b2b-a3e2-cb8d578e0885

Total jobs = 3

Launching Job 1 out of 3

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job\_1472132069327\_0002, Tracking URL = http://Inceptez:8088/proxy/application\_1472132069327\_0002/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1472132069327\_0002

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0

2016-08-25 21:12:03,446 Stage-1 map = 0%, reduce = 0%

2016-08-25 21:12:30,274 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.54 sec

MapReduce Total cumulative CPU time: 4 seconds 540 msec

Ended Job = job\_1472132069327\_0002

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: hdfs://localhost:54310/tmp/hive/hduser/70b9991c-b3c8-409d-8a72-b6883f3234aa/hive\_2016-08-25\_21-11-15\_564\_1176943574379215478-1/-ext-10000

Loading data to table dkmdb02.manutdplayers2 partition (country=England, position=null)

Time taken for load dynamic partitions : 863

Loading partition {country=England, position=Midfielder}

Loading partition {country=England, position=Defender}

Loading partition {country=England, position=Goalkeeper}

Time taken for adding to write entity : 5

[Error 30017]: Skipping stats aggregation by error org.apache.hadoop.hive.ql.metadata.HiveException: [Error 30001]: StatsPublisher cannot be initialized. There was a error in the initialization of StatsPublisher, and retrying might help. If you dont want the query to fail because accurate statistics could not be collected, set hive.stats.reliable=false

Partition dkmdb02.manutdplayers2{country=England, position=Defender} stats: [numFiles=1, numRows=-1, totalSize=1827, rawDataSize=-1]

Partition dkmdb02.manutdplayers2{country=England, position=Goalkeeper} stats: [numFiles=1, numRows=-1, totalSize=918, rawDataSize=-1]

Partition dkmdb02.manutdplayers2{country=England, position=Midfielder} stats: [numFiles=1, numRows=-1, totalSize=1871, rawDataSize=-1]

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Cumulative CPU: 4.54 sec HDFS Read: 23622 HDFS Write: 4616 SUCCESS

Total MapReduce CPU Time Spent: 4 seconds 540 msec

OK

Time taken: 78.842 seconds

hive> show partitions dkmdb02.ManUtdPlayers2; OK

country=England/position=Defender

country=England/position=Forward

country=England/position=Goalkeeper

country=England/position=Midfielder

Time taken: 0.269 seconds, Fetched: 4 row(s)

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2

drwxr-xr-x - hduser supergroup 0 2016-08-25 21:12 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England

drwxr-xr-x - hduser supergroup 0 2016-08-25 21:12 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Defender

-rw-r--r-- 1 hduser supergroup 1827 2016-08-25 21:12 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Defender/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 21:04 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Forward

-rw-r--r-- 1 hduser supergroup 2214 2016-08-25 21:04 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Forward/MU-Eng-Fwd.csv

drwxr-xr-x - hduser supergroup 0 2016-08-25 21:12 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Goalkeeper

-rw-r--r-- 1 hduser supergroup 918 2016-08-25 21:12 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Goalkeeper/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 21:12 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Midfielder

-rw-r--r-- 1 hduser supergroup 1871 2016-08-25 21:12 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Midfielder/000000\_0

## Static Partitioning using Load

hive> load data local inpath '/home/hduser/Desktop/Work/Sample\_Data/MU-Eng.csv' into table dkmdb02.manutdplayers1 partition (country = 'England');

Loading data to table dkmdb02.manutdplayers1 partition (country=England)

Partition dkmdb02.manutdplayers1{country=England} stats: [numFiles=1, numRows=0, totalSize=9479, rawDataSize=0]

OK

Time taken: 0.494 seconds

hive> show partitions dkmdb02.manutdplayers1;

OK

country=Argentina

country=England

country=France

country=Germany

country=Netherlands

country=Spain

country=USA

Time taken: 0.169 seconds, Fetched: 7 row(s)

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

16/08/25 20:27:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

drwxr-xr-x - hduser supergroup 0 2016-08-25 20:26 /user/hive/warehouse/dkmdb02.db/manutdplayers1

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Argentina

-rw-r--r-- 1 hduser supergroup 223 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Argentina/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 20:26 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=England

-rw-r--r-- 1 hduser supergroup 9479 2016-08-25 20:26 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=England/MU-Eng.csv

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=France

-rw-r--r-- 1 hduser supergroup 421 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=France/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany

-rw-r--r-- 1 hduser supergroup 85 2016-08-25 14:48 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Germany/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Netherlands

-rw-r--r-- 1 hduser supergroup 404 2016-08-25 15:10 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Netherlands/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 20:08 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Spain

-rw-r--r-- 1 hduser supergroup 189 2016-08-25 20:08 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=Spain/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-25 15:00 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA

-rw-r--r-- 1 hduser supergroup 133 2016-08-25 15:00 /user/hive/warehouse/dkmdb02.db/manutdplayers1/country=USA/MU-USA.csv

hive> select \* from dkmdb02.manutdplayers1 where country = 'England' limit 10;

OK

Peter Abbott Forward 1953-10-01 England

Ted Adams Goalkeeper 1906-11-30 England

Alphonso Ainsworth Forward 1913-07-31 England

Nicky Ajose Forward 1991-10-07 England

George Albinson Midfielder 1897-02-14 England

Reg Allen Goalkeeper 1919-05-03 England

Arthur Allman Defender 1890-12-24 England

Alfred Ambler Midfielder 1879-07-01 England

Ben Amos Goalkeeper 1990-04-10 England

John Anderson Midfielder 1921-10-11 England

Time taken: 0.612 seconds, Fetched: 10 row(s)

#### ****Partitioning on multiple columns****

hive> load data local inpath '/home/hduser/Desktop/Work/Sample\_Data/MU-Eng-Fwd.csv' into table dkmdb02.ManUtdPlayers2 partition (country = 'England', position = 'Forward');

Loading data to table dkmdb02.manutdplayers2 partition (country=England, position=Forward)

Partition dkmdb02.manutdplayers2{country=England, position=Forward} stats: [numFiles=1, numRows=0, totalSize=2214, rawDataSize=0]

OK

Time taken: 0.996 seconds

hive> show partitions dkmdb02.ManUtdPlayers2;

OK

country=England/position=Forward

Time taken: 0.185 seconds, Fetched: 1 row(s)

[hduser@Inceptez Sample\_Data]$ hdfs dfs -ls -R /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2

drwxr-xr-x - hduser supergroup 0 2016-08-25 21:04 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England

drwxr-xr-x - hduser supergroup 0 2016-08-25 21:04 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Forward

-rw-r--r-- 1 hduser supergroup 2214 2016-08-25 21:04 /user/hduser/hiveexternaldata/dkmdb02.db/ManUtdPlayers2/country=England/position=Forward/MU-Eng-Fwd.csv

## Dynamic Partitioning using Load (WON'T WORK)

hive> load data local inpath '/home/hduser/Desktop/Work/Sample\_Data/MU-Bra-Por.csv' into table dkmdb02.manutdplayers1 partition (country);

FAILED: SemanticException org.apache.hadoop.hive.ql.metadata.HiveException: MetaException(message:Invalid partition key & values; keys [country, ], values [])

**The quick context is that:**

1.Load data simply copies data, it doesn't read it so it cannot figure out what to partition.

2.Would suggest that you load data into an intermediate table first (or using an external table pointing to all the files) and then letting partition dynamic insert to kick in to load it into a partitioned table.

# Hive Queries

## Simple Select (Count, Sum, Group By etc)

hive (retail)> select \* from txnrecords where txnno = 00095903;

OK

95903 02-07-2016 4005514 52.82 Jumping Pogo Sticks Scottsdale Arizona credit

Time taken: 0.527 seconds, Fetched: 1 row(s)

hive (retail)> set hive.cli.print.header=true;

hive (retail)> select \* from txnrecords where txnno = 00095903;

OK

txnrecords.txnno txnrecords.txndate txnrecords.custno txnrecords.amount txnrecords.category txnrecords.product txnrecords.city txnrecords.state txnrecords.spendby

95903 02-07-2016 4005514 52.82 Jumping Pogo Sticks Scottsdale Arizona credit

hive (retail)> select count(\*) from txnrecords;

Query ID = hduser\_20160702112020\_2c296705-e1cf-480d-ba5b-d06c11b1ebc4

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>

Starting Job = job\_1467433224759\_0001, Tracking URL = http://Inceptez:8088/proxy/application\_1467433224759\_0001/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1467433224759\_0001

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2016-07-02 11:20:50,645 Stage-1 map = 0%, reduce = 0%

2016-07-02 11:21:02,929 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.69 sec

2016-07-02 11:21:13,867 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.18 sec

MapReduce Total cumulative CPU time: 3 seconds 180 msec

Ended Job = job\_1467433224759\_0001

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.18 sec HDFS Read: 8472295 HDFS Write: 6 SUCCESS

Total MapReduce CPU Time Spent: 3 seconds 180 msec

OK

95904

Time taken: 49.966 seconds, Fetched: 1 row(s)

hive (retail)> select count(1) from txnrecords;

\_c0

95904

hive (default)> select sum(amount) from cust\_trxn group by age;

## Explain

hive (retail)> explain select \* from txnrecords limit 10;

OK

Explain

STAGE DEPENDENCIES:

Stage-0 is a root stage

STAGE PLANS:

Stage: Stage-0

Fetch Operator

limit: 10

Processor Tree:

TableScan

alias: txnrecords

Statistics: Num rows: 13753 Data size: 8472073 Basic stats: COMPLETE Column stats: NONE

Select Operator

expressions: txnno (type: int), txndate (type: string), custno (type: int), amount (type: double), category (type: string), product (type: string), city (type: string), state (type: string), spendby (type: string)

outputColumnNames: \_col0, \_col1, \_col2, \_col3, \_col4, \_col5, \_col6, \_col7, \_col8

Statistics: Num rows: 13753 Data size: 8472073 Basic stats: COMPLETE Column stats: NONE

Limit

Number of rows: 10

Statistics: Num rows: 10 Data size: 6160 Basic stats: COMPLETE Column stats: NONE

ListSink

Time taken: 0.446 seconds, Fetched: 20 row(s)

## Insert

hive (dkmdb01)> insert into players1 values (10,'Sachin','India',43);

FAILED: ParseException line 1:12 missing TABLE at 'players1' near '<EOF>'

hive (dkmdb01)> insert into table players1 values (10,'Sachin','India',43);

hive (default)> insert into table cust\_trxn

> select a.custno,a.firstname,a.age,a.profession,b.amount,b.product

> from customer a join txnrecords b on a.custno = b.custno;

hive (dkmdb01)> insert overwrite table players2 partition(pl\_country='India') select pl\_id,pl\_name,pl\_age from players1 where pl\_country = 'India';

## Copy data from Hive table into a file

hive (dkmdb01)> insert overwrite local directory '/home/hduser/hive2local\_data' row format delimited select \* from employee1;

[hduser@Inceptez hive2local\_data]$ cd /home/hduser/hive2local\_data

[hduser@Inceptez hive2local\_data]$ ls

000000\_0

[hduser@Inceptez hive2local\_data]$ cat 000000\_0

1Deepak750002016-06-13 11:36:27

2Sachin15000002016-06-13 11:37:51

3Messi25000002016-06-13 11:38:04

4Federer20000002016-06-13 12:18:19

5Ronaldo27500002016-06-18 02:58:31

6Rooney20000002016-07-09 10:41:36

hive (dkmdb01)> insert overwrite local directory '/home/hduser/hive2local\_data/employee1.data1' row format delimited select \* from employee1;

[hduser@Inceptez hive2local\_data]$ ls

000000\_0 employee1.data1

[hduser@Inceptez hive2local\_data]$ cd employee1.data1/

[hduser@Inceptez employee1.data1]$ ls

000000\_0

hive (dkmdb01)> insert overwrite local directory '/home/hduser/hive2local\_data' row format delimited fields terminated by '|' select \* from employee1;

[hduser@Inceptez hive2local\_data]$ ls -l

total 4

-rw-r--r-- 1 hduser hduser 221 Jul 11 22:27 000000\_0

[hduser@Inceptez hive2local\_data]$ date

Mon Jul 11 22:28:33 IST 2016

[hduser@Inceptez hive2local\_data]$ cat 000000\_0

1|Deepak|75000|2016-06-13 11:36:27

2|Sachin|1500000|2016-06-13 11:37:51

3|Messi|2500000|2016-06-13 11:38:04

4|Federer|2000000|2016-06-13 12:18:19

5|Ronaldo|2750000|2016-06-18 02:58:31

6|Rooney|2000000|2016-07-09 10:41:36

## Delete

hive (dkmdb01)> delete from players1;

FAILED: SemanticException [Error 10297]: Attempt to do update or delete on table dkmdb01.players1 that does not use an AcidOutputFormat or is not bucketed

hive (dkmdb01)> delete from testdml3 where id = 6;

Query ID = hduser\_20160712224141\_7e283be5-5566-42a1-82db-353079296179

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks determined at compile time: 3

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>

Starting Job = job\_1468334110122\_0009, Tracking URL = http://Inceptez:8088/proxy/application\_1468334110122\_0009/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1468334110122\_0009

Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3

2016-07-12 22:41:22,506 Stage-1 map = 0%, reduce = 0%

2016-07-12 22:42:17,283 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 5.37 sec

2016-07-12 22:42:26,151 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.42 sec

2016-07-12 22:43:00,633 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 8.45 sec

2016-07-12 22:43:01,763 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.32 sec

MapReduce Total cumulative CPU time: 10 seconds 320 msec

Ended Job = job\_1468334110122\_0009

Loading data to table dkmdb01.testdml3

[Error 30017]: Skipping stats aggregation by error org.apache.hadoop.hive.ql.metadata.HiveException: [Error 30001]: StatsPublisher cannot be initialized. There was a error in the initialization of StatsPublisher, and retrying might help. If you dont want the query to fail because accurate statistics could not be collected, set hive.stats.reliable=false

Table dkmdb01.testdml3 stats: [numFiles=3, totalSize=1833]

MapReduce Jobs Launched:

Stage-Stage-1: Map: 3 Reduce: 3 Cumulative CPU: 10.32 sec HDFS Read: 2416 HDFS Write: 508 SUCCESS

Total MapReduce CPU Time Spent: 10 seconds 320 msec

OK

row\_\_id

Time taken: 117.973 seconds

hive (dkmdb01)> select \* from testdml3;

OK

testdml3.id\_val testdml3.id

Deepak 1

Sachin 2

Messi 3

Federer 4

Ronaldo 5

Time taken: 0.397 seconds, Fetched: 5 row(s)

## Update

hive (default)> create table dkmdb01.testdml1 (id int, id\_val varchar(10)) row format delimited fields terminated by ',';

OK

Time taken: 0.154 seconds

hive (default)> show create table dkmdb01.testdml1;

OK

createtab\_stmt

CREATE TABLE `dkmdb01.testdml1`(

`id` int,

`id\_val` varchar(10))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/testdml1'

TBLPROPERTIES (

'transient\_lastDdlTime'='1468339581')

Time taken: 0.101 seconds, Fetched: 13 row(s)

hive (dkmdb01)> select \* from testdml1;

OK

testdml1.id testdml1.id\_val

1 Deepak

2 Sachin

3 Messi

4 Federer

5 Ronaldo

6 Rooney

Time taken: 0.083 seconds, Fetched: 6 row(s)

hive (dkmdb01)> update testdml1 set id\_val = 'W Rooney' where id = 6;

FAILED: SemanticException [Error 10297]: Attempt to do update or delete on table dkmdb01.testdml1 that does not use an AcidOutputFormat or is not bucketed

hive (dkmdb01)> alter table testdml1 set tblproperties ('transactional'='true');

OK

Time taken: 0.273 seconds

hive (dkmdb01)> show create table testdml1;

OK

createtab\_stmt

CREATE TABLE `testdml1`(

`id` int,

`id\_val` varchar(10))

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/testdml1'

TBLPROPERTIES (

'COLUMN\_STATS\_ACCURATE'='false',

'last\_modified\_by'='hduser',

'last\_modified\_time'='1468340261',

'numFiles'='1',

'numRows'='-1',

'rawDataSize'='-1',

'totalSize'='55',

'transactional'='true',

'transient\_lastDdlTime'='1468340261')

Time taken: 0.246 seconds, Fetched: 21 row(s)

hive (dkmdb01)> update testdml1 set id\_val = 'W Rooney' where id = 6;

FAILED: SemanticException [Error 10297]: Attempt to do update or delete on table testdml1 that does not use an AcidOutputFormat or is not bucketed

hive (dkmdb01)> alter table testdml1 clustered by (id) into 3 buckets;

OK

Time taken: 0.157 seconds

hive (dkmdb01)> show create table testdml1;

OK

createtab\_stmt

CREATE TABLE `testdml1`(

`id` int,

`id\_val` varchar(10))

CLUSTERED BY (

id)

INTO 3 BUCKETS

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/testdml1'

TBLPROPERTIES (

'COLUMN\_STATS\_ACCURATE'='false',

'last\_modified\_by'='hduser',

'last\_modified\_time'='1468340949',

'numFiles'='1',

'numRows'='-1',

'rawDataSize'='-1',

'totalSize'='55',

'transactional'='true',

'transient\_lastDdlTime'='1468340949')

Time taken: 0.162 seconds, Fetched: 24 row(s)

hive (dkmdb01)> update testdml1 set id\_val = 'W Rooney' where id = 6;

Diagnostic Messages for this Task:

Error: java.lang.RuntimeException: org.apache.hadoop.hive.ql.metadata.HiveException: Hive Runtime Error while processing row (tag=0) {"key":{"reducesinkkey0":null},"value":{"\_col1":"W Rooney"}}

at org.apache.hadoop.hive.ql.exec.mr.ExecReducer.reduce(ExecReducer.java:265)

at org.apache.hadoop.mapred.ReduceTask.runOldReducer(ReduceTask.java:444)

at org.apache.hadoop.mapred.ReduceTask.run(ReduceTask.java:392)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:163)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:415)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1628)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:158)

Caused by: org.apache.hadoop.hive.ql.metadata.HiveException: Hive Runtime Error while processing row (tag=0) {"key":{"reducesinkkey0":null},"value":{"\_col1":"W Rooney"}}

at org.apache.hadoop.hive.ql.exec.mr.ExecReducer.reduce(ExecReducer.java:253)

... 7 more

Caused by: org.apache.hadoop.hive.ql.metadata.HiveException: org.apache.hadoop.hive.serde2.SerDeException: Error: expecting 2 but asking for field 2

data=[Ljava.lang.Object;@6a1ad1a1

tableType=struct<id:int,id\_val:varchar(10)>

dataType=struct<\_col0:struct<transactionid:bigint,bucketid:int,rowid:bigint>,\_col1:int,\_col2:varchar(10)>

at org.apache.hadoop.hive.ql.exec.FileSinkOperator.processOp(FileSinkOperator.java:725)

at org.apache.hadoop.hive.ql.exec.Operator.forward(Operator.java:815)

at org.apache.hadoop.hive.ql.exec.SelectOperator.processOp(SelectOperator.java:84)

at org.apache.hadoop.hive.ql.exec.mr.ExecReducer.reduce(ExecReducer.java:244)

... 7 more

Caused by: org.apache.hadoop.hive.serde2.SerDeException: Error: expecting 2 but asking for field 2

data=[Ljava.lang.Object;@6a1ad1a1

tableType=struct<id:int,id\_val:varchar(10)>

dataType=struct<\_col0:struct<transactionid:bigint,bucketid:int,rowid:bigint>,\_col1:int,\_col2:varchar(10)>

at org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe.doSerialize(LazySimpleSerDe.java:415)

at org.apache.hadoop.hive.serde2.AbstractEncodingAwareSerDe.serialize(AbstractEncodingAwareSerDe.java:50)

at org.apache.hadoop.hive.ql.exec.FileSinkOperator.processOp(FileSinkOperator.java:660)

... 10 more

FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 3 Cumulative CPU: 5.18 sec HDFS Read: 280 HDFS Write: 0 FAIL

Total MapReduce CPU Time Spent: 5 seconds 180 msec

hive (dkmdb01)> select \* from testdml1;

OK

testdml1.id testdml1.id\_val

1 Deepak

2 Sachin

3 Messi

4 Federer

5 Ronaldo

6 Rooney

Time taken: 0.163 seconds, Fetched: 6 row(s)

hive (dkmdb01)> show create table testdml3;

OK

createtab\_stmt

CREATE TABLE `testdml3`(

`id\_val` varchar(10),

`id` int)

CLUSTERED BY (

id)

INTO 3 BUCKETS

ROW FORMAT SERDE

'org.apache.hadoop.hive.ql.io.orc.OrcSerde'

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/testdml3'

TBLPROPERTIES (

'transactional'='true',

'transient\_lastDdlTime'='1468342454')

Time taken: 0.149 seconds, Fetched: 17 row(s)

hive (dkmdb01)> update testdml3 set id\_val = 'W Rooney' where id = 6;

Query ID = hduser\_20160712223333\_f9d4cd8c-a5d9-4114-a736-03d9429b60af

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks determined at compile time: 3

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>

Starting Job = job\_1468334110122\_0008, Tracking URL = http://Inceptez:8088/proxy/application\_1468334110122\_0008/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1468334110122\_0008

Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3

2016-07-12 22:34:22,119 Stage-1 map = 0%, reduce = 0%

2016-07-12 22:35:27,931 Stage-1 map = 0%, reduce = 0%

2016-07-12 22:36:24,517 Stage-1 map = 22%, reduce = 0%, Cumulative CPU 5.1 sec

2016-07-12 22:36:35,131 Stage-1 map = 44%, reduce = 0%, Cumulative CPU 6.91 sec

2016-07-12 22:36:55,579 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 7.28 sec

2016-07-12 22:36:59,265 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.94 sec

2016-07-12 22:38:00,378 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 10.09 sec

2016-07-12 22:38:06,312 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 11.52 sec

2016-07-12 22:38:14,483 Stage-1 map = 100%, reduce = 78%, Cumulative CPU 12.89 sec

2016-07-12 22:38:15,574 Stage-1 map = 100%, reduce = 89%, Cumulative CPU 14.26 sec

2016-07-12 22:38:16,707 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 16.5 sec

MapReduce Total cumulative CPU time: 16 seconds 500 msec

Ended Job = job\_1468334110122\_0008

Loading data to table dkmdb01.testdml3

[Error 30017]: Skipping stats aggregation by error org.apache.hadoop.hive.ql.metadata.HiveException: [Error 30001]: StatsPublisher cannot be initialized. There was a error in the initialization of StatsPublisher, and retrying might help. If you dont want the query to fail because accurate statistics could not be collected, set hive.stats.reliable=false

Table dkmdb01.testdml3 stats: [numFiles=2, totalSize=1329]

MapReduce Jobs Launched:

Stage-Stage-1: Map: 3 Reduce: 3 Cumulative CPU: 16.5 sec HDFS Read: 1592 HDFS Write: 654 SUCCESS

Total MapReduce CPU Time Spent: 16 seconds 500 msec

OK

\_col0 \_col1 \_col2

Time taken: 267.895 seconds

hive (dkmdb01)> select \* from testdml3;

OK

testdml3.id\_val testdml3.id

Deepak 1

Sachin 2

Messi 3

Federer 4

Ronaldo 5

W Rooney 6

Time taken: 0.173 seconds, Fetched: 6 row(s)

## Show

**hive> show databases;**

default

retail

**hive (retail)> show tables in retail;**

txnrecords

**hive (retail)> show create table txnrecords;**

OK

createtab\_stmt

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 ','

LINES TERMINATED BY '\n'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/retail.db/txnrecords'

TBLPROPERTIES (

'COLUMN\_STATS\_ACCURATE'='true',

'numFiles'='1',

'numRows'='0',

'rawDataSize'='0',

'totalSize'='8472073',

'transient\_lastDdlTime'='1467439247')

Time taken: 0.267 seconds, Fetched: 26 row(s)

## Describe

**hive> desc HiveTest;**

OK

employeeid int

firstname string

designation string

salary int

department string

Time taken: 0.151 seconds, Fetched: 5 row(s)

**hive> desc formatted players;**

OK

# col\_name data\_type comment

pl\_id int

pl\_name string

pl\_team string

pl\_salary string

pl\_country string

# Detailed Table Information

Database: default

Owner: hduser

CreateTime: Sat Jul 02 22:19:28 IST 2016

LastAccessTime: UNKNOWN

Protect Mode: None

Retention: 0

Location: hdfs://localhost:54310/user/hduser/hiveexternaldata/players

Table Type: EXTERNAL\_TABLE

Table Parameters:

COLUMN\_STATS\_ACCURATE false

EXTERNAL TRUE

last\_modified\_by hduser

last\_modified\_time 1467479060

numFiles 0

numRows -1

rawDataSize -1

totalSize 0

transient\_lastDdlTime 1467479060

# 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 ,

line.delim \n

serialization.format ,

Time taken: 0.349 seconds, Fetched: 40 row(s)

hive> desc formatted dkmdb02.manutdplayers1 partition country=England;

FAILED: ParseException line 1:48 missing ( at 'country' near '=' in constant

hive> desc formatted dkmdb02.manutdplayers1 partition (country=England);

FAILED: ParseException line 1:57 cannot recognize input near 'England' ')' '<EOF>' in constant

hive> desc formatted dkmdb02.manutdplayers1 partition (country='England');

OK

# col\_name data\_type comment

player varchar(30)

position varchar(20)

dob date

# Partition Information

# col\_name data\_type comment

country varchar(25)

# Detailed Partition Information

Partition Value: [England]

Database: dkmdb02

Table: manutdplayers1

CreateTime: Thu Aug 25 20:26:51 IST 2016

LastAccessTime: UNKNOWN

Protect Mode: None

Location: hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/manutdplayers1/country=England

Partition Parameters:

COLUMN\_STATS\_ACCURATE true

numFiles 1

numRows 0

rawDataSize 0

totalSize 9479

transient\_lastDdlTime 1472137012

# 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: 0.629 seconds, Fetched: 38 row(s)

## Create Like & Timestamp

hive (dkmdb01)> create table emp1\_like like employee1;

OK

Time taken: 0.793 seconds

hive (dkmdb01)> show create table employee1;

OK

createtab\_stmt

CREATE EXTERNAL TABLE `employee1`(

`emp\_id` int,

`emp\_name` varchar(25),

`emp\_sal` decimal(11,2),

`row\_ts` timestamp)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hduser/sqphv01/employee1'

TBLPROPERTIES (

'COLUMN\_STATS\_ACCURATE'='true',

'numFiles'='0',

'numRows'='-1',

'rawDataSize'='-1',

'totalSize'='0',

'transient\_lastDdlTime'='1468422624')

Time taken: 0.189 seconds, Fetched: 20 row(s)

hive (dkmdb01)> show create table emp1\_like;

OK

createtab\_stmt

CREATE TABLE `emp1\_like`(

`emp\_id` int,

`emp\_name` varchar(25),

`emp\_sal` decimal(11,2),

`row\_ts` timestamp)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/emp1\_like'

TBLPROPERTIES (

'transient\_lastDdlTime'='1468423450')

Time taken: 0.186 seconds, Fetched: 15 row(s)

hive (dkmdb01)> insert into table employee1 select 7,'Raikkonen',2100000,from\_unixtime(unix\_timestamp()) from players1;

hive (dkmdb01)> select 7,'Raikkonen',2100000,from\_unixtime(unix\_timestamp()) from players1;

## Join

hive (dkmdb01)> create table employee (emp\_id int, emp\_name varchar(10), emp\_sal decimal(9,2), emp\_org\_id int) row format delimited fields terminated by ',';

hive (dkmdb01)> create table org (org\_id int, org\_name varchar(10)) row format delimited fields terminated by ',';

[hduser@Inceptez hive\_data\_01]$ cat employee.data1

2,Farooq,70000,1

3,Kathick,96000,1

4,Hari,90000,2

5,Vadivel,100000,3

6,Mani,75000,4

7,Vignesh,75000,2

[hduser@Inceptez hive\_data\_01]$ vi org.data1

[hduser@Inceptez hive\_data\_01]$ cat org.data1

1,RBS

2,CTS

3,VDSI

4,HCL

hive> load data local inpath '/home/hduser/hive\_data\_01/employee.data1' into table dkmdb01.employee;

hive> load data local inpath '/home/hduser/hive\_data\_01/org.data1' into table dkmdb01.org;

hive> insert into table dkmdb01.org values (5,'TCS');

hive> insert into table dkmdb01.employee values (8,'XXX',10000,6);

hive> select \* from dkmdb01.org;

5 TCS

1 RBS

2 CTS

3 VDSI

4 HCL

hive> select \* from dkmdb01.org order by 1; 🡸 order by ‘1’ didn’t give correct result

4 HCL

3 VDSI

2 CTS

1 RBS

5 TCS

hive> select \* from dkmdb01.org order by org\_id asc;

1 RBS

2 CTS

3 VDSI

4 HCL

5 TCS

hive> select \* from dkmdb01.employee order by emp\_id asc;

1 Deepak 82000 1

2 Farooq 70000 1

3 Kathick 96000 1

4 Hari 90000 2

5 Vadivel 100000 3

6 Mani 75000 4

7 Vignesh 75000 2

8 XXX 10000 6

hive> set hive.groupby.orderby.position.alias=true;

hive> select \* from dkmdb01.employee order by 1;

FAILED: SemanticException [Error 10219]: Position in ORDER BY is not supported when using SELECT \*

hive> select emp\_id, emp\_name, emp\_sal, emp\_org\_id from dkmdb01.employee order by 1;

1 Deepak 82000 1

2 Farooq 70000 1

3 Kathick 96000 1

4 Hari 90000 2

5 Vadivel 100000 3

6 Mani 75000 4

7 Vignesh 75000 2

8 XXX 10000 6

hive> select \* from employee, org where emp\_org\_id = org\_id;

1 Deepak 82000 1 1 RBS

2 Farooq 70000 1 1 RBS

3 Kathick 96000 1 1 RBS

4 Hari 90000 2 2 CTS

5 Vadivel 100000 3 3 VDSI

6 Mani 75000 4 4 HCL

7 Vignesh 75000 2 2 CTS

Time taken: 64.555 seconds, Fetched: 7 row(s)

hive> select \* from employee where emp\_org\_id <> 1;

OK

8 XXX 10000 6

4 Hari 90000 2

5 Vadivel 100000 3

6 Mani 75000 4

7 Vignesh 75000 2

Time taken: 0.24 seconds, Fetched: 5 row(s)

hive> select \* from employee, org where emp\_org\_id <> org\_id; 🡸 Not-equal join Not supported

Warning: Map Join MAPJOIN[8][bigTable=employee] in task 'Stage-3:MAPRED' is a cross product

1 Deepak 82000 1 5 TCS

1 Deepak 82000 1 2 CTS

1 Deepak 82000 1 3 VDSI

1 Deepak 82000 1 4 HCL

8 XXX 10000 6 5 TCS

8 XXX 10000 6 1 RBS

8 XXX 10000 6 2 CTS

8 XXX 10000 6 3 VDSI

8 XXX 10000 6 4 HCL

2 Farooq 70000 1 5 TCS

2 Farooq 70000 1 2 CTS

2 Farooq 70000 1 3 VDSI

2 Farooq 70000 1 4 HCL

3 Kathick 96000 1 5 TCS

3 Kathick 96000 1 2 CTS

3 Kathick 96000 1 3 VDSI

3 Kathick 96000 1 4 HCL

4 Hari 90000 2 5 TCS

4 Hari 90000 2 1 RBS

4 Hari 90000 2 3 VDSI

4 Hari 90000 2 4 HCL

5 Vadivel 100000 3 5 TCS

5 Vadivel 100000 3 1 RBS

5 Vadivel 100000 3 2 CTS

5 Vadivel 100000 3 4 HCL

6 Mani 75000 4 5 TCS

6 Mani 75000 4 1 RBS

6 Mani 75000 4 2 CTS

6 Mani 75000 4 3 VDSI

7 Vignesh 75000 2 5 TCS

7 Vignesh 75000 2 1 RBS

7 Vignesh 75000 2 3 VDSI

7 Vignesh 75000 2 4 HCL

Time taken: 69.485 seconds, Fetched: 33 row(s)

hive> select \* from employee join org on emp\_org\_id <> org\_id;

FAILED: SemanticException [Error 10017]: Line 1:35 Both left and right aliases encountered in JOIN 'org\_id'

## Rank

hive> load data local inpath '/home/hduser/hive\_data\_01/salesorders\_new.data' overwrite into table dkmdb01.salesorders3;

hive> select \* from dkmdb01.salesorders3 limit 3;

OK

2015-01-06 East Jones Pencil 95 1.99 189.05

2015-01-23 Central Kivell Binder 50 19.99 999.5

2015-02-09 Central Jardine Pencil 36 4.99 179.64

Time taken: 0.097 seconds, Fetched: 3 row(s)

CREATE TABLE DKMDB01.salesorders(ord\_dt date, region varchar(8), rep varchar(15), item varchar(15), unit int, unit\_cost decimal(9,2), total decimal(10,2));

mysql> load data local infile '/home/hduser/hive\_data\_01/salesorders\_new.data' into table DKMDB01.salesorders fields terminated by ',' lines terminated by '\n';

Query OK, 43 rows affected (0.00 sec)

Records: 43 Deleted: 0 Skipped: 0 Warnings: 0

mysql> select a.\*, rank() over (partition by region, order by total desc) as rank from dkmdb01.salesorders3;

ERROR 1305 (42000): FUNCTION rank does not exist

mysql> select a.\*, row\_number() over (partition by region, order by total desc) as rank from dkmdb01.salesorders3;

ERROR 1305 (42000): FUNCTION row\_number does not exist

hive> select a.\*, rank() over (partition by region, order by total desc) as rank from dkmdb01.salesorders3 a;

FAILED: ParseException line 1:52 missing ) at 'by' near 'by'

line 1:55 missing EOF at 'total' near 'by'

hive> select a.\*, rank() over (partition by region order by total desc) as rank from dkmdb01.salesorders3 a;

2016-12-04 Central Jardine Binder 94 19.99 1879.06 1

2016-02-01 Central Smith Binder 87 15 1305 2

2015-01-23 Central Kivell Binder 50 19.99 999.5 3

2016-05-31 Central Gill Binder 80 8.99 719.2 4

2016-07-21 Central Morgan Pen Set 55 12.49 686.95 5

2016-06-17 Central Kivell Desk 5 125 625 6

2015-02-26 Central Gill Pen 27 19.99 539.73 7

2015-11-25 Central Kivell Pen Set 96 4.99 479.04 8

2015-06-25 Central Morgan Pencil 90 4.99 449.1 9

2015-05-05 Central Jardine Pencil 90 4.99 449.1 9

2016-01-15 Central Gill Binder 46 8.99 413.54 11

2015-10-05 Central Morgan Binder 28 8.99 251.72 12

2015-09-01 Central Smith Desk 2 125 250 13

2016-03-24 Central Jardine Pen Set 50 4.99 249.5 14

2015-02-09 Central Jardine Pencil 36 4.99 179.64 15

2015-04-18 Central Andrews Pencil 75 1.99 149.25 16

2016-12-21 Central Andrews Binder 28 4.99 139.72 17

2016-04-10 Central Andrews Pencil 66 1.99 131.34 18

2015-12-12 Central Smith Pencil 67 1.29 86.43 19

2016-05-14 Central Gill Pencil 53 1.29 68.37 20

2016-11-17 Central Jardine Binder 11 4.99 54.89 21

2016-10-31 Central Andrews Pencil 14 1.29 18.06 22

2016-08-07 Central Kivell Pen Set 42 23.95 15.9 23

2016-09-10 Central Gill Pencil 7 1.29 9.03 24

2015-07-29 East Parent Binder 81 19.99 1619.19 1

2015-12-29 East Parent Pen Set 74 15.99 1183.26 2

2015-10-22 East Jones Pen 64 8.99 575.36 3

2015-06-08 East Jones Binder 60 8.99 539.4 4

2016-04-27 East Howard Pen 96 4.99 479.04 5

2016-07-04 East Jones Pen Set 62 4.99 309.38 6

2015-11-08 East Parent Pen 15 19.99 299.85 7

2015-04-01 East Jones Binder 60 4.99 299.4 8

2015-09-18 East Jones Pen Set 16 15.99 255.84 9

2015-01-06 East Jones Pencil 95 1.99 189.05 10

2015-08-15 East Jones Pencil 35 4.99 174.65 11

2015-07-12 East Howard Binder 29 1.99 57.71 12

2016-02-18 East Jones Binder 4 4.99 19.96 13

2016-10-14 West Thompson Binder 57 19.99 1139.43 1

2016-08-24 West Sorvino Desk 3 275 825 2

2015-03-15 West Sorvino Pencil 56 2.99 167.44 3

2016-09-27 West Sorvino Pen 76 1.99 151.24 4

2016-03-07 West Sorvino Binder 7 19.99 139.93 5

2015-05-22 West Thompson Pencil 32 1.99 63.68 6

Time taken: 59.151 seconds, Fetched: 43 row(s)

hive> select a.\*, row\_number() over (partition by region, order by total desc) as rank from dkmdb01.salesorders3 a;

FAILED: ParseException line 1:58 missing ) at 'by' near 'by'

line 1:61 missing EOF at 'total' near 'by'

hive> create table dkmdb02.salesorders1 (ord\_dt date, rep varchar(15), item varchar(15), unit int, unit\_cost decimal(9,2), total decimal(10,2))

> partitioned by (region varchar(8))

> stored as orc;

hive> show create table dkmdb02.salesorders1;

OK

CREATE TABLE `dkmdb02.salesorders1`(

`ord\_dt` date,

`rep` varchar(15),

`item` varchar(15),

`unit` int,

`unit\_cost` decimal(9,2),

`total` decimal(10,2))

PARTITIONED BY (

`region` varchar(8))

ROW FORMAT SERDE

'org.apache.hadoop.hive.ql.io.orc.OrcSerde'

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb02.db/salesorders1'

TBLPROPERTIES (

'transient\_lastDdlTime'='1472289019')

Time taken: 0.1 seconds, Fetched: 19 row(s)

[hduser@Inceptez ~]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db

drwxr-xr-x - hduser supergroup 0 2016-08-27 14:40 /user/hive/warehouse/dkmdb02.db/salesorders1

hive> insert into table dkmdb02.salesorders1 partition by (region) select ord\_dt, rep, item, unit, unit\_cost, total, region from dkmdb01.salesorders3;

FAILED: ParseException line 1:49 extraneous input 'by' expecting ( near '<EOF>'

hive> insert into table dkmdb02.salesorders1 partition(region) select ord\_dt, rep, item, unit, unit\_cost, total, region from dkmdb01.salesorders3;

[hduser@Inceptez ~]$ hdfs dfs -ls -R /user/hive/warehouse/dkmdb02.db/salesorders1

drwxr-xr-x - hduser supergroup 0 2016-08-27 14:45 /user/hive/warehouse/dkmdb02.db/salesorders1/region=Central

-rw-r--r-- 1 hduser supergroup 1000 2016-08-27 14:45 /user/hive/warehouse/dkmdb02.db/salesorders1/region=Central/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-27 14:45 /user/hive/warehouse/dkmdb02.db/salesorders1/region=East

-rw-r--r-- 1 hduser supergroup 877 2016-08-27 14:45 /user/hive/warehouse/dkmdb02.db/salesorders1/region=East/000000\_0

drwxr-xr-x - hduser supergroup 0 2016-08-27 14:45 /user/hive/warehouse/dkmdb02.db/salesorders1/region=West

-rw-r--r-- 1 hduser supergroup 840 2016-08-27 14:45 /user/hive/warehouse/dkmdb02.db/salesorders1/region=West/000000\_0

hive> select a.\*, rank() over (partition by region order by total desc) as rank from dkmdb02.salesorders1 a where rank = 1;

FAILED: SemanticException [Error 10004]: Line 1:108 Invalid table alias or column reference 'rank': (possible column names are: ord\_dt, rep, item, unit, unit\_cost, total, region)

hive> select \* from (select a.\*, rank() over (partition by region order by total desc) as rank from dkmdb02.salesorders1 a ) tab where tab.rank = 1;

2016-12-04 Jardine Binder 94 19.99 1879.06 Central 1

2015-07-29 Parent Binder 81 19.99 1619.19 East 1

2016-10-14 Thompson Binder 57 19.99 1139.43 West 1

Time taken: 42.539 seconds, Fetched: 3 row(s)

hive (dkmdb01)> select \* from (select rank() over (partition by item order by total desc) as rank, region, rep, item, total from salesorders3) x where rank <= 3;

1 Central Jardine Binder 1879.06

2 East Parent Binder 1619.19

3 Central Smith Binder 1305

1 West Sorvino Desk 825

2 Central Kivell Desk 625

3 Central Smith Desk 250

1 East Jones Pen 575.36

2 Central Gill Pen 539.73

3 East Howard Pen 479.04

1 East Parent Pen Set 1183.26

2 Central Morgan Pen Set 686.95

3 Central Kivell Pen Set 479.04

1 Central Morgan Pencil 449.1

1 Central Jardine Pencil 449.1

3 East Jones Pencil 189.05

## To\_Date, Datediff, Case

hive> select to\_date("2016-07-13 20:43:13" timestamp) from dkmdb01.emloyee1 limit 1;

FAILED: ParseException line 1:37 extraneous input 'timestamp' expecting ) near '<EOF>'

hive> select to\_date("2016-07-13 20:43:13") from dkmdb01.employee1 limit 1;

OK

2016-07-13

Time taken: 2.334 seconds, Fetched: 1 row(s)

hive> select datediff((to\_date("2016-10-07 20:43:13"),(to\_date("2016-07-13 20:43:13")) from dkmdb01.employee1 limit 1;

FAILED: ParseException line 1:47 missing ) at ',' near '<EOF>'

hive> select datediff(to\_date("2016-10-07 20:43:13",to\_date("2016-07-13 20:43:13")) from dkmdb01.empl oyee1 limit 1;

FAILED: ParseException line 1:78 missing ) at 'from' near '<EOF>'

hive> select datediff("2016-10-06","2016-08-30") from dkmdb01.employee1 limit 1;

OK

37

Time taken: 0.38 seconds, Fetched: 1 row(s)

hive> select case WHEN datediff("2016-10-06","2016-08-30") < 30 then 'A'

> WHEN datediff("2016-10-06","2016-08-30") > 30 then 'B'

> ELSE 'C'

> END

> from dkmdb01.employee1 limit 1;

OK

B

Time taken: 0.612 seconds, Fetched: 1 row(s)

hive> select "CASE EXAMPLE", case WHEN datediff("2016-10-06","2016-08-30") < 30 then 'A'

> WHEN datediff("2016-10-06","2016-08-30") > 30 then 'B'

> ELSE 'C'

> END

> from dkmdb01.employee1 limit 1;

OK

CASE EXAMPLE B

Time taken: 0.215 seconds, Fetched: 1 row(s)

# Sqoop to Hive

## Create Hive table

sqoop create-hive-table --connect jdbc:mysql://localhost/DKMDB01 --username root --password root --table student --hive-table dkmdb01.student;

hive (dkmdb01)> show create table student;

OK

createtab\_stmt

CREATE TABLE `student`(

`stu\_id` int,

`stu\_name` string,

`stu\_dep` int,

`row\_ts` string)

COMMENT 'Imported by sqoop on 2016/07/13 21:09:01'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\u0001'

LINES TERMINATED BY '\n'

STORED AS INPUTFORMAT

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

'hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/student'

TBLPROPERTIES (

'transient\_lastDdlTime'='1468424352')

Time taken: 0.161 seconds, Fetched: 17 row(s)

hive (dkmdb01)> dfs -ls hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/student;

Found 1 items

-rw-r--r-- 1 hduser supergroup 185 2016-07-13 21:19 hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/student/part-m-00000

## Sqoop Job to Import Data incrementally into Hive table

sqoop job --create sqp2hv\_inc\_imp\_stud -- import --connect jdbc:mysql://localhost/DKMDB01 --direct --username root --password root --table student --m 1 --hive-import --hive-table dkmdb01.student --incremental append --check-column stu\_id --last-value 0;

sqoop job --list

Available jobs:

exp\_job1

import\_all\_DKMDB01

inc\_imp\_lm

myjob1

myjob2

sqp2hv\_inc\_imp\_stud

sqoop job --exec sqp2hv\_inc\_imp\_stud

Enter password:

hive (dkmdb01)> select \* from student;

OK

student.stu\_id student.stu\_name student.stu\_dep student.row\_ts

1 Deepak 1 2016-06-21 14:54:56

2 Farooq 1 2016-06-21 14:55:21

3 Vadivel 2 2016-06-21 14:55:43

4 Hari 3 2016-06-21 14:56:05

5 Karthick 1 2016-06-21 14:56:21

6 Mani 4 2016-06-21 15:11:08

Time taken: 0.11 seconds, Fetched: 6 row(s)

sqoop job --exec sqp2hv\_inc\_imp\_stud

Enter password:

hive (dkmdb01)> select \* from student;

OK

student.stu\_id student.stu\_name student.stu\_dep student.row\_ts

1 Deepak 1 2016-06-21 14:54:56

2 Farooq 1 2016-06-21 14:55:21

3 Vadivel 2 2016-06-21 14:55:43

4 Hari 3 2016-06-21 14:56:05

5 Karthick 1 2016-06-21 14:56:21

6 Mani 4 2016-06-21 15:11:08

7 Syed 1 2016-07-13 15:53:48

8 Rajesh 1 2016-07-13 15:54:44

9 Vignesh 3 2016-07-13 15:55:01

Time taken: 0.091 seconds, Fetched: 9 row(s)

hive (dkmdb01)> dfs -ls hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/student;

Found 2 items

-rw-r--r-- 1 hduser supergroup 185 2016-07-13 21:19 hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/student/part-m-00000

-rw-r--r-- 1 hduser supergroup 92 2016-07-13 21:26 hdfs://localhost:54310/user/hive/warehouse/dkmdb01.db/student/part-m-00000\_copy\_1

# Hive Shell

## Execute Linux commands

hive (retail)> !pwd;

/home/hduser

hive (retail)> !ls hive;

data

helloword.java

hive-site\_rms\_transction.xml

HiveUdf.jar

## Exexute HDFS commands

hive (retail)> dfs -ls -R /user/hive;

drwxrwxr-x - hduser supergroup 0 2016-07-02 10:47 /user/hive/warehouse

drwxr-xr-x - hduser supergroup 0 2016-07-02 10:52 /user/hive/warehouse/retail.db

drwxr-xr-x - hduser supergroup 0 2016-07-02 10:52 /user/hive/warehouse/retail.db/txnrecords

hive> dfs -mkdir -p /user/hduser/hiveexternaldata;

## List Files & Jars

hive (default)> list files;

/home/hduser/hive\_udfs/weekday\_mapper.py

hive (default)> list jars;

/home/hduser/hive/HiveUdf.jar

## Hive Configuration Variables commonly used

### hive.cli.print.current.db=true;

Print current database name in CLI

### hive.cli.print.header=true;

Print column header in the query result

### hive.enforce.bucketing=true;

The command set hive.enforce.bucketing = true; allows the correct number of reducers and the cluster by column to be automatically selected based on the table. Otherwise, you would need to set the number of reducers to be the same as the number of buckets as in set mapred.reduce.tasks = 256; and have a CLUSTER BY ... clause in the select. (Note: Not needed in Hive 2.x onward)

### hive.support.concurrency=true;

### hive.exec.dynamic.partition=true;

To dynamically partition based on the column value.

### hive.exec.dynamic.partition.mode=nonstrict;

### hive.txn.manager=org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;

### hive.compactor.initiator.on=true;

### hive.compactor.worker.threads=1;

### hive.groupby.orderby.position.alias=true;

Enable group by/order by using postion (i.e. number)

# Hive Script in a file

[hduser@Inceptez hive\_scripts]$ cat case1.sql

select \* , case

when age<30 then 'low'

when age>=30 and age < 50 then 'middle'

when age>=50 then 'old'

else 'others'

end

from cust\_trxn

limit 10;

[hduser@Inceptez hive\_scripts]$ hive -f case1.sql

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-0.14.0.jar!/hive-log4j.properties

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/local/hive/lib/hive-jdbc-0.14.0-standalone.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

OK

4007024 Cameron 59 Actor 40.33 Cardio Machine Accessories old

4006742 Gregory 36 Accountant 198.44 Weightlifting Gloves middle

4009775 Ruby 44 Designer 5.58 Weightlifting Machine Accessories middle

4002199 Keith 44 Police officer 198.19 Gymnastics Rings middle

4002613 Hugh 43 Engineering technician 98.81 Field Hockey middle

4007591 Jennifer 54 Electrician 193.63 Camping & Backpacking & Hiking old

4002190 Sheryl 62 Designer 27.89 Jigsaw Puzzles old

4002964 Ken 67 Recreation and fitness worker 96.01 Sandboxes old

4007361 Terri 52 Loan officer 10.44 Snowmobiling old

4004798 Geoffrey 65 Chemist 152.46 Bungee Jumping old

Time taken: 4.298 seconds, Fetched: 10 row(s)

## Inside Hive Shell

hive (default)> source /home/hduser/hive\_scripts/case1.sql

> ;

OK

cust\_trxn.custno cust\_trxn.firstname cust\_trxn.age cust\_trxn.profession cust\_trxn.amount cust\_trxn.product \_c1

4007024 Cameron 59 Actor 40.33 Cardio Machine Accessories old

4006742 Gregory 36 Accountant 198.44 Weightlifting Gloves middle

4009775 Ruby 44 Designer 5.58 Weightlifting Machine Accessories middle

4002199 Keith 44 Police officer 198.19 Gymnastics Rings middle

4002613 Hugh 43 Engineering technician 98.81 Field Hockey middle

4007591 Jennifer 54 Electrician 193.63 Camping & Backpacking & Hiking old

4002190 Sheryl 62 Designer 27.89 Jigsaw Puzzles old

4002964 Ken 67 Recreation and fitness worker 96.01 Sandboxes old

4007361 Terri 52 Loan officer 10.44 Snowmobiling old

4004798 Geoffrey 65 Chemist 152.46 Bungee Jumping old

Time taken: 0.804 seconds, Fetched: 10 row(s)

# Run Hive queries from command line

[hduser@Inceptez hive\_scripts]$ hive -e 'select max(amount) from txnrecords'

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-0.14.0.jar!/hive-log4j.properties

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/local/hive/lib/hive-jdbc-0.14.0-standalone.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

Query ID = hduser\_20160715205656\_fcefa377-c7bd-41ec-bf0c-7968f8d576e6

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>

Starting Job = job\_1468591924483\_0004, Tracking URL = http://Inceptez:8088/proxy/application\_1468591924483\_0004/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job\_1468591924483\_0004

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2016-07-15 20:57:31,228 Stage-1 map = 0%, reduce = 0%

2016-07-15 20:58:03,007 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.93 sec

2016-07-15 20:58:39,609 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.9 sec

MapReduce Total cumulative CPU time: 8 seconds 900 msec

Ended Job = job\_1468591924483\_0004

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.9 sec HDFS Read: 8472285 HDFS Write: 6 SUCCESS

Total MapReduce CPU Time Spent: 8 seconds 900 msec

OK

200.0

Time taken: 107.069 seconds, Fetched: 1 row(s)

select \* from ${tab\_name};

[root@sandbox myhqls]# hive -hiveconf tab\_name=employee -f sel\_parm\_tab.hql

FAILED: ParseException line 1:14 cannot recognize input near '$' '{' 'tab\_name' in join source

[root@sandbox myhqls]# cat sel\_parm\_tab.hql

select \* from ${hiveconf:tab\_name};

[root@sandbox myhqls]# hive -hiveconf tab\_name=employee -f sel\_parm\_tab.hql

16/08/21 01:42:20 INFO Configuration.deprecation: mapred.input.dir.recursive is deprecated. Instead, use mapreduce.input.fileinputformat.input.dir.recursive

16/08/21 01:42:20 INFO Configuration.deprecation: mapred.max.split.size is deprecated. Instead, use mapreduce.input.fileinputformat.split.maxsize

16/08/21 01:42:20 INFO Configuration.deprecation: mapred.min.split.size is deprecated. Instead, use mapreduce.input.fileinputformat.split.minsize

16/08/21 01:42:20 INFO Configuration.deprecation: mapred.min.split.size.per.rack is deprecated. Instead, use mapreduce.input.fileinputformat.split.minsize.per.rack

16/08/21 01:42:20 INFO Configuration.deprecation: mapred.min.split.size.per.node is deprecated. Instead, use mapreduce.input.fileinputformat.split.minsize.per.node

16/08/21 01:42:20 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces

16/08/21 01:42:20 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-0.12.0.2.0.6.0-76.jar!/hive-log4j.properties

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple\_bindings for an explanation.

SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

OK

1 Deepak 82000 1

2 Farooq 70000 1

3 Kathick 96000 1

4 Hari 90000 2

5 Vadivel 100000 3

6 Mani 75000 4

7 Vignesh 75000 2

Time taken: 6.537 seconds, Fetched: 7 row(s)