

Group A

Lab 2c

Hive part 2 and file format.

Participants:

1. Aadarsha chapagain
- 2.Jyoti shukla
- 3.Rishi Phaneendra Varma
- 4.Priti Bhale
- 5.Sreya Treesa Johnny
- 6.Piyush Bhatia

Slide 2:

```
create table employee_data(id int, name string,
compensation Map<string, int>
row format delimited
fields terminated by ','
collection items terminated by '$'
map keys terminated by ':'
Stored as textfile;
```

The screenshot shows a Windows desktop environment. In the foreground, a terminal window titled 'root@sandbox-hdp:~' is open, displaying the creation of a Hive table named 'employee_data'. The command includes specifying the table's schema (id int, name string, compensation Map<string, int>), defining the row format as delimited by commas, and setting collection items to be terminated by a dollar sign (\$). It also specifies map keys to be terminated by a colon (:) and stores the data as a textfile. The terminal output shows various log messages from the Hive server and the completion of the table creation command.

In the background, a Notepad window titled 'BDM_Group_A' is visible, listing the participants for Group A. The participants listed are: 1. Aadarsha chapagain, 2.Jyoti shukla, 3.Rishi Phaneendra Varma, 4.Priti Bhale, 5.Sreya Treesa Johnny, and 6.Piyush Bhatia. The Notepad window has a 'Lenovo' logo in the bottom right corner.

```
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/3.0.1.0-187/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.0.1.0-187/hadoop/lib/slf4j-log4j12-1.7.25.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.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=hive;serviceDiscoveryMode=zooKeeper;user=hive;zookeeperNamespace=hiveserver2
21/09/22 13:22:52 [main]: INFO jdbc.HiveConnection: Connected to sandbox-hdp.hortonworks.com:10000
[Connected to: Apache Hive (version 3.1.0-3.0.1.0-187)
Driver: Hive JDBC (version 3.1.0-3.0.1.0-187)
Transaction isolation: TRANSACTION_READ_COMMITTED
Beeline version 3.1.0-3.0.1.0-187 by Apache Hive
\o: jdbc:hive2://sandbox-hdp.hortonworks.com:2] create table employee_data(id int, name string,
compensation Map<string, int>
row format delimited
fields terminated by ','
collection items terminated by '$'
map keys terminated by ':'
Stored as textfile;
INFO : Compiling command(queryId=hive_20210922132255_57337da0-3fa8-4803-ae9c-1a2eed62ba82): create table employee_data(id int, name string,
compensation Map<string, int>
row format delimited
fields terminated by ','
collection items terminated by '$'
map keys terminated by ':'
Stored as textfile;
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20210922132255_57337da0-3fa8-4803-ae9c-1a2eed62ba82); Time taken: 0.057 seconds
INFO : Executing command(queryId=hive_20210922132255_57337da0-3fa8-4803-ae9c-1a2eed62ba82): create table employee_data(id int, name string,
compensation Map<string, int>
row format delimited
fields terminated by ','
collection items terminated by '$'
map keys terminated by ':'
Stored as textfile;
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20210922132255_57337da0-3fa8-4803-ae9c-1a2eed62ba82); Time taken: 0.198 seconds
INFO : OK
No rows affected (0.434 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

Slide 7:

Select name, compensation['salary'] + compensation['comm'] + compensation['bonus']
as totalcompensation from employee_data where name='Benjamin';

```
dfs@sandbox-hdp:~
```

No rows affected (0.466 seconds)

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> show * from employee_data;
```

Error: Error while compiling statement: FAILED: ParseException line 1:5 cannot recognize input near 'show' '*' 'from' in ddl statement (state=42000,code=40000)

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select * from employee_data;
```

INFO : Compiling command(queryId=hive_20210922134101_06fd90d9-abef3-46a9-a9e7-638c76fa3e70): select * from employee_data

INFO : Semantic Analysis Completed (retryal = false)

INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:employee_data.name, type:string, comment:null), FieldSchema(name:employee_data.id, type:int, comment:null), FieldSchema(name:employee_data.compensation, type:map<string,int>, comment:null)], properties:null)

INFO : Completed compiling command(queryId=hive_20210922134101_06fd90d9-abef3-46a9-a9e7-638c76fa3e70); Time taken: 0.271 seconds

INFO : Executing command(queryId=hive_20210922134101_06fd90d9-abef3-46a9-a9e7-638c76fa3e70): select * from employee_data

INFO : Completed executing command(queryId=hive_20210922134101_06fd90d9-abef3-46a9-a9e7-638c76fa3e70); Time taken: 0.004 seconds

INFO : OK

```
+-----+-----+-----+
```

employee_data.id	employee_data.name	employee_data.compensation
1	Pedram	{"salary":5000,"comm":1000,"bonus":5000}
2	Benjamin	{"salary":1000,"comm":200,"bonus":1000}

```
+-----+-----+-----+
```

2 rows selected (0.337 seconds)

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Select name, compensation['salary'] + compensation['comm'] + compensation['bonus'] as totalcompensation from employee_data where name='Benjamin';
```

INFO : Compiling command(queryId=hive_20210922134407_3abea879-f073-4881-baa6-fc1f21b98c5a): Select name, compensation['salary'] + compensation['comm'] + compensation['bonus'] as totalcompensation from employee_data where name='Benjamin'

INFO : Semantic Analysis Completed (retryal = false)

INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:name, type:string, comment:null), FieldSchema(name:totalcompensation, type:int, comment:null)], properties:null)

INFO : Completed compiling command(queryId=hive_20210922134407_3abea879-f073-4881-baa6-fc1f21b98c5a); Time taken: 0.158 seconds

INFO : Executing command(queryId=hive_20210922134407_3abea879-f073-4881-baa6-fc1f21b98c5a): Select name, compensation['salary'] + compensation['comm'] + compensation['bonus'] as totalcompensation from employee_data where name='Benjamin'

INFO : Completed executing command(queryId=hive_20210922134407_3abea879-f073-4881-baa6-fc1f21b98c5a); Time taken: 0.004 seconds

INFO : OK

```
+-----+-----+
```

name	totalcompensation
Benjamin	2200

```
+-----+-----+
```

1 row selected (0.218 seconds)

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```



Slide 8:

Create view sales_by_US_view as select *

From salesjan2009_ext where country='United States';

```
[hdfs@sandbox-hdp ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/3.0.1.0-187/hive/lib/log4j-slf4j-log4j2-2.10.0.jar!/_org.slf4j.impl.StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.0.1.0-187/hadoop/lib/slf4j-log4j12-1.7.25.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.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=hive;serviceDiscoveryMode=zooKeeper;user=hive;zooKeeperNamespace=hiveserver2
21/09/22 14:12:38 [main]: INFO jdbc.HiveConnection: Connected to sandbox-hdp.hortonworks.com:10000
Connected to: Apache Hive (Version 3.1.0.3.0.1.0-187)
Driver: Hive JDBC (version 3.1.0.3.0.1.0-187)
Transaction Isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.0.3.0.1.0-187 by Apache Hive
:0> jdbc:hive2://sandbox-hdp.hortonworks.com:> Create view sales_by_US_view as select *
INFO : Compiling command(queryId=hive_20210922141304_5f5edcbc-26b5-4e7e-a966-5e1afb2674ac): Create view sales_by_US_view
as select *
From salesjan2009_ext where country='United States'
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:productid, type:string, comment:null), FieldSchema(n
ame:salesamount, type:int, comment:null), FieldSchema(name:paymenttype, type:string, comment:null), FieldSchema(name:cust
omername, type:string, comment:null), FieldSchema(name:city, type:string, comment:null), FieldSchema(name:region, type:st
ring, comment:null), FieldSchema(name:country, type:string, comment:null)], properties:null)
INFO : completed compiling command(queryId=hive_20210922141304_5f5edcbc-26b5-4e7e-a966-5e1afb2674ac); Time taken: 0.557
seconds
INFO : Executing command(queryId=hive_20210922141304_5f5edcbc-26b5-4e7e-a966-5e1afb2674ac): Create view sales_by_US_view
as select *
From salesjan2009_ext where country='United States'
INFO : Starting task [Stage-1:DDL] in serial mode
INFO : completed executing command(queryId=hive_20210922141304_5f5edcbc-26b5-4e7e-a966-5e1afb2674ac); Time taken: 0.131
seconds
INFO : OK
No rows affected (0.9 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> ■
```

Slide 9:

Select productid, salesamount, country from sales_by_US_view.

Slide 10:

Select paymenttype, sum(salesamount) from sales_by_US_view group by paymenttype

```
hdfs@sandbox-hdp:~$ jdbc:hive2://sandbox-hdp.hortonworks.com:2> Select paymenttype, sum(salesamount) from sales_by_US_view group by payment type
Error: Error while compiling statement: FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'paymenttype' (state=42000,code=10025)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> select paymenttype, sum(salesamount) from sales_by_US_view group by paymenttype;
INFO : Compiling command(queryId=hive_20210922142123_37b63433-218f-4b74-bf47-c368cb831388): Select paymenttype, sum(sale amount) from sales_by_US_view group by paymenttype
INFO : Semantic Analysis Completed (rettrial = false)
INFO : Returning Hive schema: Schema(fieldschemas:[Fieldschema(name:paymenttype, type:string, comment:null), Fieldschema(name:c1, type:bigint, comment:null)], properties:{null})
INFO : Completed compiling command(queryId=hive_20210922142123_37b63433-218f-4b74-bf47-c368cb831388); Time taken: 0.191 seconds
INFO : Executing command(queryId=hive_20210922142123_37b63433-218f-4b74-bf47-c368cb831388): Select paymenttype, sum(sale amount) from sales_by_US_view group by paymenttype
INFO : Query ID = hive_20210922142123_37b63433-218f-4b74-bf47-c368cb831388
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20210922142123_37b63433-218f-4b74-bf47-c368cb831388
INFO : Tez session hasn't been created yet. Opening session
INFO : Dag name: Select paymenttype, sum(salesa...paymenttype (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1632227054967_0034)
BDM_Gro...
File Edit Format View Help
Group A
Participants
1. Aadarsha chapagain
2. Jyoti shukla
3. Rishi Phaneendra Varma
4. Priti Bhale
5. Sreya Treesa Johny
6. Piyush Bhatia
Lenovo
BDN
VERTICES      ROWS      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... Container SUCCEEDED   1    1    0    0    0    0
Reducer 2 ..... Container SUCCEEDED   2    2    0    0    0    0
----- VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 27.01 s
INFO : Status: DAG finished successfully in 26.83 seconds
INFO :
INFO : Query Execution Summary
INFO :
INFO : OPERATION DURATION
INFO :
INFO : Compile Query 0.19s
INFO : Prepare Plan 6.00s
INFO : Get Query Coordinator (AM) 0.00s
INFO : Submit Plan 0.34s
INFO : Start DAG 1.29s
INFO : Run DAG 26.83s
INFO :
```

Slide 11:

Create a script to create a table and change permission and run via hive -f

A screenshot of a Windows desktop environment. In the foreground, a terminal window titled 'hdfs@sandbox-hdp:' shows the execution of a shell script to create a table named 'salesjan2009_by_script'. The command 'cat creat_table_script' displays the table definition, which includes columns for productid, salesamount, paymenttype, customername, city, region, and country. The command 'chmod 777 creat_table_script' changes the file permissions. The command 'hive -f creat_table_script' runs the script. The terminal also lists several log files and a Java file named 'QueryResult.java'. In the background, a Notepad window titled 'BDM_Gro...' is open, showing a list of participants for 'Group A': 1. Aadarsha chapagain, 2. Jyoti shukla, 3. Rishi Phaneendra Varma, 4. Priti Bhole, 5. Sreya Treesa Johny, and 6. Piyush Bhatia. The desktop taskbar at the bottom shows various icons and the date/time as 8:52 AM, 10/9/2021.

```
[hdfs@sandbox-hdp ~]$ vi creat_table_script
[hdfs@sandbox-hdp ~]$ cat creat_table_script
use test_db
create table salesjan2009_by_script(
productid string, salesamount int,
paymenttype string, customername string, city string, region string, country string)
Row format delimited
Fields terminated by ',' stored as textfile;
Load data inpath '/user/aadarsha/salesJan2009.csv' into table SalesJan2009_by_script;
[hdfs@sandbox-hdp ~]$ chmod 777 creat_table_script
[hdfs@sandbox-hdp ~]$ ls -l
total 804
-rwxrwxrwx 1 hdfs hadoop 85 Sep 22 13:35 compensation.csv
-rwxrwxrwx 1 hdfs hadoop 323 Sep 22 14:33 creat_table_script
-rw-r--r-- 1 root root 299158 Oct 8 2021 data_set_trimmed.csv
-rw-r--r-- 1 hdfs hadoop 14301 Sep 21 18:02 pig_1632247038985.log
-rw-r--r-- 1 hdfs hadoop 5228 Sep 21 18:05 pig_1632247442848.log
-rw-r--r-- 1 root root 39867 Sep 21 22:04 QueryResult.java
-rw-r--r-- 1 root root 65835 Sep 21 17:54 SalesJan2009.csv
-rw-r--r-- 1 root root 299158 Oct 8 2021 sample_data.csv
-rw-r--r-- 1 hdfs hadoop 39594 Sep 21 21:45 TBLS.java
-rw-r--r-- 1 hdfs hadoop 30555 Sep 22 21:41 TXNS.java
[hdfs@sandbox-hdp ~]$ hdfs dfs -put test_table_script /user/aadarsha
[hdfs@sandbox-hdp ~]$ hive -f creat_table_script
```

Slide 26:

create table salesjan2009_clustered(

productid string,salesamount int,paymenttype string,customernam

clustered by (country) into 10 buckets

stored as ORCFILE

A screenshot of a Windows desktop environment. In the foreground, a terminal window titled 'hdfs@sandbox-hdp:' shows the creation of a clustered table named 'salesjan2009_clustered'. The command 'show tables' lists existing tables like emp, emp_data, emp_static, emp_dynamic, emp_data_dyn, emp_data_static, and salesjan2009. The command 'CREATE TABLE salesjan2009_clustered' defines the new table with columns for productid, salesamount, paymenttype, and customername, clustered by country into 10 buckets. The command 'show tables' is run again to verify the new table. In the background, a Notepad window titled 'BDM_Gro...' is open, showing the same list of participants for 'Group A'. The desktop taskbar at the bottom shows various icons and the date/time as 9:20 AM, 10/9/2021.

```
[hdfs@sandbox-hdp ~]$ jdbc:hive2://sandbox-hdp.hortonworks.com:2181;password=
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2181/default;password=
INFO : Command completed (retVal = false).
INFO : Semantic Analysis Completed (retVal = false).
INFO : Returning Hive schema: Schema[fieldschemas: [fieldschema(name:tab_name, type:string, comment:from deserializer)]].
INFO : Completed compiling command(queryId=hive_20210922150310_fc37a9c4-b4ce-4258-949e-634980a3cd93); Time taken: 0.019 seconds
INFO : Executing command(queryId=hive_20210922150310_fc37a9c4-b4ce-4258-949e-634980a3cd93); show tables
INFO : Starting task [Stage-0:DQL] in serial mode
INFO : Completed executing command(queryId=hive_20210922150310_fc37a9c4-b4ce-4258-949e-634980a3cd93); Time taken: 0.012 seconds
INFO : OK
+-----+
| tab_name |
+-----+
| emp       |
| emp_data  |
| emp_data_part |
| emp_dynamic |
| emp_data_dyn_part |
| emp_data_static |
| salesjan2009 |
| salesjan2009_ext |
+-----+
Rows displayed (0.039 seconds)

0: jdbc:hive2://sandbox-hdp.hortonworks.com:2181;password=
+hive:serviceDiscoveryMode=zookeeper;user=hive;zookeeperNamespace=hiveserver2
(hdfs@sandbox-hdp ~)$ hdfs dfs -ls /warehouse/tblspace/salesjan2009
Found 9 items
drwxrwxrwx - hive    hadoop 0 2021-09-22 13:40 /warehouse/tblspace/managed/hive/emp/_static_part
drwxrwxrwx - hive    hadoop 0 2021-09-22 13:40 /warehouse/tblspace/managed/hive/employee_data
drwxrwxrwx - hive    hadoop 0 2021-09-22 09:51 /warehouse/tblspace/managed/hive/employee_data_dyn_part
drwxrwxrwx - hive    hadoop 0 2021-09-22 09:53 /warehouse/tblspace/managed/hive/emp/emp_data_static_part
drwxrwxrwx - hive    hadoop 0 2021-09-22 09:53 /warehouse/tblspace/managed/hive/emp/emp_data_dyn_part
drwxrwxrwx - hive    hadoop 0 2021-09-22 03:36 /warehouse/tblspace/managed/hive/hive_db
drwxrwxrwx - hive    hadoop 0 2021-09-22 14:58 /warehouse/tblspace/managed/hive/salesjan2009_clustered
drwxrwxrwx - hive    hadoop 0 2021-09-22 14:58 /warehouse/tblspace/managed/hive/sys_db
drwxrwxrwx - hive    hadoop 0 2021-09-22 14:51 /warehouse/tblspace/managed/hive/test_db
[hdfs@sandbox-hdp ~]$
```

Slide 27:

insert into salesjan2009_clustered

select * from salesjan2009_ext;

hdfs@sandbox-hdp:~

```
D: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
D: jdbc:hive2://sandbox-hdp.hortonworks.com:2> insert into salesjan2009_clustered
D: jdbc:hive2://sandbox-hdp.hortonworks.com:2>   select * from salesjan2009_ext;
INFO : Compiling command(queryId=hive_20210922150742_b76034c3-1716-4dde-9930-36f86bd4a9e1): insert into salesjan2009_clustered
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:[Fieldschema(name:salesjan2009_ext.productid, type:string, comment:nul
D: Fieldschema(name:salesjan2009_ext.salesamount, type:int, comment:null), Fieldschema(name:salesjan2009_ext.paymenttype
D: type:string, comment:null), Fieldschema(name:salesjan2009_ext.customername, type:string, comment:null), Fieldschema(nam
D: Fieldschema(name:salesjan2009_ext.city, type:string, comment:null), Fieldschema(name:salesjan2009_ext.state, type:string, comment:null)
D: Fieldschema(name:salesjan2009_ext.country, type:string, comment:null), properties:null])
INFO : Completed compiling command(queryId=hive_20210922150742_b76034c3-1716-4dde-9930-36f86bd4a9e1); Time taken: 0.272
seconds
D: jdbc:hive2://sandbox-hdp.hortonworks.com:2> insert into salesjan2009_clustered
INFO : Executing command(queryId=hive_20210922150742_b76034c3-1716-4dde-9930-36f86bd4a9e1): insert into salesjan2009_clustered
INFO : Query ID = Hive_20210922150742_b76034c3-1716-4dde-9930-36f86bd4a9e1
INFO : Total jobs = 1
INFO :Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20210922150742_b76034c3-1716-4dde-9930-36f86bd4a9e1
INFO : Tez session hasn't been created yet. Opening session
INFO : Dag name: insert into salesjan2009..._salesjan2009_ext (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_163227054967_0035)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	10	10	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0

VERTICES: 0/03 [=====>] 100% ELAPSED TIME: 9.52 s

Participants

- 1. Aadarsha chapagain
- 2.Jyoti shukla
- 3.Rishi Phaneendra Varma
- 4.Priti Bhale
- 5.Sreya Treesa Johnny
- 6.Piyush Bhatia

Slide 28:

```
hdfs@sandbox-hdp:~ - Shell In / x +
```

Re ← → C ① localhost:4200

Apps My blog Translate Travel to canada Lambton College Share Data Analytics Dilema Gmail YouTube Maps Descriptive, Predict...

```
7 rows selected (0.183 seconds)
D: jdbc:hive2://sandbox-hdp.hortonworks.com:2> describe formatted salesjan2009_clustered;
INFO : Compiling command(queryId=hive_20210922152913_4effe546f-36c2-47f1-84dc-2f5842862d7b): describe formatted salesjan2009_clustered
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:[Fieldschema(name:col_name, type:string, comment:from deserializer), Fieldschema(name:productid, type:string, comment:from deserializer), Fieldschema(name:salesamount, type:int, comment:from deserializer), Fieldschema(name:paymenttype, type:string, comment:from deserializer), Fieldschema(name:customername, type:string, comment:from deserializer), Fieldschema(name:city, type:string, comment:from deserializer), Fieldschema(name:state, type:string, comment:from deserializer), Fieldschema(name:region, type:string, comment:from deserializer), Fieldschema(name:country, type:string, comment:from deserializer), Fieldschema(name:latitude, type:double, comment:from deserializer), Fieldschema(name:longitude, type:double, comment:from deserializer), Fieldschema(name:owner, type:string, comment:from deserializer), Fieldschema(name:create_time, type:timestamp, comment:from deserializer), Fieldschema(name:last_update_time, type:timestamp, comment:from deserializer), Fieldschema(name:location, type:string, comment:from deserializer)])
INFO : Completed compiling command(queryId=hive_20210922152913_4effe546f-36c2-47f1-84dc-2f5842862d7b); Time taken: 0.166 seconds
INFO : Starting task [Stage-0:FORMAT] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20210922152913_4effe546f-36c2-47f1-84dc-2f5842862d7b; describe formatted salesjan2009_clustered
INFO : Starting task [Stage-0:OBJS] in serial mode
INFO : Completed executing command(queryId=hive_20210922152913_4effe546f-36c2-47f1-84dc-2f5842862d7b); Time taken: 0.025 seconds
INFO : Ok
```

col_name	data_type	comment
productid	string	
salesamount	int	
paymenttype	string	
customername	string	
city	string	
region	string	
country	string	
latitude	double	
longitude	double	
owner	string	
create_time	hive_timestamp	
last_update_time	hive_timestamp	
location	string	

Table Type: MANAGED_TABLE

Table Parameters:

- COLUMN_STATS_ACCURATE: true
- transactional: true
- transactional_properties: bucketing_version=1, numBuckets=10, numRows=1996, sizeInBytes=1000000, totalSize=32725
- transient: true
- transient_lastDdlTime: 16322323278

Storage Information

Storage Library: org.apache.hadoop.hive.oai.oio.orc.OrcSerde

InputFormat: org.apache.hadoop.hive.oai.oio.orc.OrcInputFormat

OutputFormat: org.apache.hadoop.hive.oai.oio.orc.OrcOutputFormat

Compressed: false

Num Buckets: 10

Bucketed By: country

Sort Columns:

Storage Desc Params:

Serialization Format: 1

39 rows selected (0.318 seconds)
D: jdbc:hive2://sandbox-hdp.hortonworks.com:2>

Participants

- 1. Aadarsha chapagain
- 2.Jyoti shukla
- 3.Rishi Phaneendra Varma
- 4.Priti Bhale
- 5.Sreya Treesa Johnny
- 6.Piyush Bhatia

Slide 29:

10 buckets are created and orc file is not in human readable form.

Slide 32:

```
select country, count(*) as count from SalesJan2009_ext
```

group by (country)

order by count;

Here we can see that the data is skewed and do not have normal distribution.

The screenshot shows a Windows desktop environment. On the left, a Jupyter Notebook interface is open, displaying a list of countries with their corresponding population counts. The code cell at the bottom of the notebook is as follows:

```
df = pd.read_csv('countries.csv')
df.head()
df.info()
```

On the right, a file explorer window titled "BDM_Gro..." is open, showing a list of participants: Aaderesha chapeagain, Jyoti shukla, Rishabh Phaneendra Varma, Priti Bhale, Sreya Treesa Johny, and Piyush Bhatia.

Slide 33: create table salesjan2009_skewed(

productid string,salesamount int,paymenttype string,customername string,city string,region string,country string)

skewed by (country) on ('United States', 'United Kingdom') stored as directories

row format delimited

fields terminated by ',';

We need to specify by which column and value for skewed table;

```
hdfs@sandbox-hdp:~ - Shell In x +  
localhost:4200  
Apps My blog Translate Travel to canada Lambton College Share Data Analytics Dilema Gmail YouTube Maps Descriptive, Predict...  
| South Korea | 2 |  
| Ukraine | 2 |  
| Malta | 4 |  
| Poland | 4 |  
| Philippines | 4 |  
| Finland | 4 |  
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> create table salesjan2009_skewed(  
string)  
O:  
V: > skewed by (country) on ('United States', 'United Kingdom') stored as directories  
> row format delimited  
> fields terminated by ','  
INFO : Compiling command(queryId=hive_20210922155227_b5f729c1-c73e-4653-b103-bfe7a0875777): create table salesjan2009_skewed(  
productid string,salesamount int,paymenttype string,customername string,city string,region string,country string)  
skewed by (country) on ('United States', 'United Kingdom') stored as directories  
row format delimited  
fields terminated by ','  
fields terminated by ''  
INFO : Semantic Analysis Completed (retrial = false)  
INFO : Retaining Hive schema: Schema(fieldschema=null, properties=null)  
INFO : Compiling and running command(queryId=hive_20210922155227_b5f729c1-c73e-4653-b103-bfe7a0875777) Time taken: 0.042 seconds  
INFO : Compiling and running command(queryId=hive_20210922155227_b5f729c1-c73e-4653-b103-bfe7a0875777): create table salesjan2009_skewed(  
productid string,salesamount int,paymenttype string,customername string,city string,region string,country string)  
skewed by (country) on ('United States', 'United Kingdom') stored as directories  
row format delimited  
fields terminated by ''  
INFO : Starting task [stage-0@DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20210922155227_b5f729c1-c73e-4653-b103-bfe7a0875777) Time taken: 0.15 seconds  
INFO : OK  
No rows affected (0.241 seconds)  
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>   
  
BDM_Gro... File Edit Format View Help Group A Participants 1. Aadarsha chapagain 2. Jyoti shukla 3. Rishi Phaneendra Varma 4. Priti Bhale 5. Sreya Treesa Johny 6. Piyush Bhatia  
Windows (CRLF) UTF-8
```

Slide 34: Skewed values are stored as sub directories

Slide 35:

```
insert into salesjan2009_skewed  
select * from salesjan2009_ext;
```

```
hdfs@sandbox-hdp:~ - Shell In / x
localhost:4200
Apps My blog Translate Travel to canada Lambton College Share Data Analytics Dilema Gmail YouTube Maps Descriptive, Predict...
INFO : Executing command[queryId:hive_20210922160205_69b06d82-6175-454e-b64b-e0c385868435]: insert into salesjan2009_skewed
select * from salesjan2009_ext
INFO : Query ID = hive_20210922160205_69b06d82-6175-454e-b64b-e0c385868435
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20210922160205_69b06d82-6175-454e-b64b-e0c385868435
INFO : Session is already open
INFO : Dag name: insert into salesjan2009...salesjan2009_ext (Stage-1)
INFO : Tez session was closed. Reopening...
INFO : Session re-established.
INFO : Session re-established.
INFO : Status: Running (Executing on YARN cluster with App id application_1632227054967_0037)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
----- BD VERTICES: 0/02 [=====>] 100% ELAPSED TIME: 10.93 s
----- INFO : Status: DAG finished successfully in 9.94 seconds
INFO :
INFO : Query Execution Summary
INFO :
INFO : OPERATION DURATION
INFO : -----
INFO : Compile Query 0.33s
INFO : Prepare Plan 0.17s
INFO : Get Query Coordinator (AM) 0.00s
INFO : Submit Plan 8.95s
INFO : Start DAG 3.38s
INFO : Run DAG 9.94s
INFO :
INFO :
INFO : Task Execution Summary
INFO : -----
INFO : VERTICES DURATION(ms) CPU_TIME(ms) GC_TIME(ms) INPUT_RECORDS OUTPUT_RECORDS
INFO : -----
INFO : Map 1 5117.00 6,250 199 1,996 1
INFO : Reducer 2 1679.00 1,140 20 1 0
INFO : -----
BDM_Gro... File Edit Format View Help Group A Participants 1. Aadarscha chapagain 2.Jyoti shukla 3.Rishi Phaneendra Varma 4.Priti Bhole 5.Sreya Treesa Johnny 6.Piyush Bhatia Windows (CRLF) UTF-8
```

Slide 89: create table salesjan2009_parquet1(

```
productid string,salesamount int,paymenttype string,customername string,city string,region  
string,country string)
```

stored as PARQUET

```
TBLPROPERTIES ("parquet.compress"="snappy");
```

Storing file as Parquet format. It is columnar file format and data are partitioned by columns.

The screenshot shows a Windows desktop environment with several windows open:

- Terminal Window:** Shows the command-line interface for HDFS and Hive. The user has run the command `desc formatted salesjan2009_parquet1` and is viewing the detailed table information, including columns, data types, and TBLPROPERTIES.
- File Explorer Window:** Shows the file system structure under `hdfs://sandbox-hdp.hortonworks.com:8020`, specifically the `salesjan2009_parquet1` directory.
- Notepad Window:** Titled "BDM.Gro...", it contains a list of participants: 1. Aadarsha chapagain, 2. Jyoti shukla, 3. Rishi Phaneendra Varma, 4. Priti Bhale, 5. Sreya Treesa Johny, 6. Piyush Bhatia.