

# Project Outcomes:

## 1.TaskOutcome:

➔ start mysql in cloudera

```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 17
Server version: 5.1.66 Source distribution
```

Copyright (c) 2000, 2012, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

```
mysql> create database project ;
mysql>
```

➔create a database (database name is project)

```
mysql> create database project ;
Query OK, 1 row affected (0.15 sec)
```

```
mysql> show databases ;
```

Database
information_schema
cm
firehose
hue
metastore
mysql
nav
navms
oozie
project
retail_db
rman
sentry

```
13 rows in set (0.04 sec)
```

➔**create a first table** (table name is click\_data)

```
-----
mysql> create table clickdata(userid int,timestamp datetime,page char) ;
Query OK, 0 rows affected (0.12 sec)
```

```
mysql> describe clickdata ;
```

Field	Type	Null	Key	Default	Extra
userid	int(11)	YES		NULL	
timestamp	datetime	YES		NULL	
page	char(1)	YES		NULL	

```
3 rows in set (0.12 sec)
```

➔put data in first table using load command(click\_data table)

```
mysql> load data infile '/home/cloudera/project/clickdata' into table clickdata
-> fields terminated by ','
-> lines terminated by '\n' ;
Query OK, 13 rows affected, 13 warnings (0.01 sec)
Records: 13 Deleted: 0 Skipped: 0 Warnings: 13
```

➔show the data first table using select statement

```
mysql> select *from clickdata ;
```

userid	timestamp	page
1	2023-01-01 10:00:00	homepage
1	2023-01-01 10:01:00	product_page
2	2023-01-01 10:02:00	homepage
2	2023-01-01 10:03:00	cart_page
3	2023-01-01 10:05:00	homepage
3	2023-01-01 10:06:00	product_page
3	2023-01-01 10:07:00	cart_page
4	2023-01-01 10:09:00	homepage
4	2023-01-01 10:10:00	product_page
4	2023-01-01 10:11:00	cart_page
4	2023-01-01 10:12:00	checkout_page
5	2023-01-01 10:15:00	home_page
5	2023-01-01 10:16:00	product_page

```
13 rows in set (0.00 sec)
```

→again doing same step in next table for customer\_data and purchase\_data:

**For customer\_data table:**

```
mysql> create table customer_data(userid int,Name varchar(30),Email varchar(50)) ;
Query OK, 0 rows affected (0.06 sec)
```

```
mysql> load data infile '/home/cloudera/project/customerdata' into table customer_data
Query OK, 5 rows affected (0.01 sec)
Records: 5 Deleted: 0 Skipped: 0 Warnings: 0
```

```
mysql> select *from customer_data
-> ;
```

userid	Name	Email
1	john Doe	john.doe@example.com
2	Jane Smith	Jane.smith@example.com
3	Robert Johnson	robert.johnson@example.com
4	Lisa Brown	lisa.brown@example.com
5	Mischael Wilson	michael.wilson@example.com

5 rows in set (0.00 sec)

### **For purchase\_data table:**

```
mysql> create table purchase_data(userid int,timestamp datetime,amount int) ;  
Query OK, 0 rows affected (0.07 sec)
```

```
mysql> load data infile '/home/cloudera/project/purchasedata' into table purchase_data  
Query OK, 5 rows affected (0.01 sec)  
Records: 5 Deleted: 0 Skipped: 0 Warnings: 0
```

```
mysql> select * from purchase_data ;  
+-----+-----+-----+  
| userid | timestamp          | amount |  
+-----+-----+-----+  
|      1 | 2023-01-01 10:05:00 |     100 |  
|      2 | 2023-01-01 10:08:00 |     150 |  
|      3 | 2023-01-01 10:09:00 |     200 |  
|      4 | 2023-01-01 10:13:00 |     120 |  
|      5 | 2023-01-01 10:17:00 |      80 |  
+-----+-----+-----+  
5 rows in set (0.00 sec)
```

---

## 2.TaskOutcomes:

→import data from mysql to hive using sqoop command:

→first create a database in hive and use this database and create a table:

```
hive> create database export_db ;
OK
Time taken: 2.516 seconds
hive> show databases ;
OK
default
export_db
```

```
hive> use export_db ;
OK
Time taken: 0.496 seconds
hive> create table click_data(userid int,timestamp datetime,page varchar(30));
FAILED: SemanticException [Error 10099]: DATETIME type isn't supported yet. Please use DATE or TIMESTAMP instead
hive> create table click_data(userid int,timestamp Timestamp,page varchar(30));
OK
Time taken: 11.275 seconds
hive> show tables ;
OK
click_data
Time taken: 0.621 seconds, Fetched: 1 row(s)
```

## ➔import mysql-table to hive using sqoop commands:



```
cloudera@quickstart: ~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ sqoop import --connect jdbc:mysql://localhost/project --username=root --password=cloudera --table=clickdata --hive-home=/user/hive/warehouse --hive-import --hive-overwrite --hive-table=export_db.click_data ;
```

```
23/07/19 02:01:48 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/84c217ad949df59c8c5d9426f78eb0cd/clickdata.jar
23/07/19 02:01:48 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/07/19 02:01:48 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/07/19 02:01:48 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/07/19 02:01:48 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/07/19 02:01:48 INFO mapreduce.ImportJobBase: Beginning import of clickdata
23/07/19 02:01:48 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
23/07/19 02:01:52 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/07/19 02:02:05 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/07/19 02:02:06 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/07/19 02:02:38 INFO db.DBInputFormat: Using read committed transaction isolation
23/07/19 02:02:39 INFO mapreduce.JobSubmitter: number of splits:1
23/07/19 02:02:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1689744761876_0001
23/07/19 02:02:54 INFO impl.YarnClientImpl: Submitted application application_1689744761876_0001
23/07/19 02:02:56 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1689744761876_0001/
23/07/19 02:02:56 INFO mapreduce.Job: Running job: job_1689744761876_0001
23/07/19 02:05:25 INFO mapreduce.Job: Job job_1689744761876_0001 running in uber mode : false
23/07/19 02:05:25 INFO mapreduce.Job: map 0% reduce 0%
23/07/19 02:07:52 INFO mapreduce.Job: map 100% reduce 0%
23/07/19 02:08:04 INFO mapreduce.Job: Job job_1689744761876_0001 completed successfully
23/07/19 02:08:06 INFO mapreduce.Job: Counters: 30
```

### File System Counters

```
FILE: Number of bytes read=0
FILE: Number of bytes written=134953
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
```

```

bytes written=454
23/07/19 02:08:06 INFO mapreduce.ImportJobBase: Transferred 454 bytes in 361.1227 seconds (1.2572 bytes/sec)
23/07/19 02:08:06 INFO mapreduce.ImportJobBase: Retrieved 13 records.
23/07/19 02:08:06 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `clickdata` AS t LIMIT 1
23/07/19 02:08:07 WARN hive.TableDefWriter: Column timestamp had to be cast to a less precise type in Hive
23/07/19 02:08:07 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/jars/hive-common-1.1.0-cdh5.4.2.jar!/hive-log4j.properties
OK
Time taken: 8.363 seconds
Loading data to table export_db.click_data
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/export_db.db/click_data/part-m-00000'
Table export_db.click_data stats: [numFiles=1, numRows=0, totalSize=454, rawDataSize=0]
OK
Time taken: 11.23 seconds
[cloudera@quickstart Desktop]$ █

```

➔ Go to hive check import status:

```

hive> select * from click_data ;
OK
1      2023-01-01 10:00:00      homepage
1      2023-01-01 10:01:00      product_page
2      2023-01-01 10:02:00      homepage
2      2023-01-01 10:03:00      cart_page
3      2023-01-01 10:05:00      homepage
3      2023-01-01 10:06:00      product_page
3      2023-01-01 10:07:00      cart_page
4      2023-01-01 10:09:00      homepage
4      2023-01-01 10:10:00      product_page
4      2023-01-01 10:11:00      cart_page
4      2023-01-01 10:12:00      checkout_page
5      2023-01-01 10:15:00      home_page
5      2023-01-01 10:16:00      product_page
Time taken: 6.069 seconds, Fetched: 13 row(s)
hive> █

```

Success full import all rows are come in hive table this is first table import (clickdata) more two table are import mysql to hive



## →import customer\_data table into hive:

TABLEREAD: 2019-07-23 02:43:40

```
[cloudera@quickstart Desktop]$ sqoop import --connect jdbc:mysql://localhost/project --username=root --password=cloudera --table=customer_data --hive-home=/user/hive/warehouse --hive-import --hive-overwrite --hive-table=export db.customer_data -m 1 ;
```

Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.

Please set \$ACCUMULO\_HOME to the root of your Accumulo installation.

23/07/19 02:43:40 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.2

23/07/19 02:43:03 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera.com

23/07/19 02:45:03 INFO mapreduce.Job: Running job: job\_1689744761876\_0002

23/07/19 02:46:49 INFO mapreduce.Job: Job job\_1689744761876\_0002 running in uber mode : false

23/07/19 02:46:49 INFO mapreduce.Job: map 0% reduce 0%

23/07/19 02:48:54 INFO mapreduce.Job: map 100% reduce 0%

23/07/19 02:49:04 INFO mapreduce.Job: Job job\_1689744761876\_0002 completed successfully

23/07/19 02:49:06 INFO mapreduce.Job: Counters: 30

### File System Counters

FILE: Number of bytes read=0

FILE: Number of bytes written=134972

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=87

HDFS: Number of bytes written=193

HDFS: Number of read operations=4

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

### Job Counters

Launched map tasks=1

```
23/07/19 02:49:06 INFO mapreduce.ImportJobBase: Transferred 193 bytes in 278.8163 seconds (0.6922 bytes/sec)
23/07/19 02:49:06 INFO mapreduce.ImportJobBase: Retrieved 5 records.
23/07/19 02:49:07 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `customer_data` AS t LIMIT 1
23/07/19 02:49:07 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/jars/hive-common-1.1.0-cdh5.4.2.jar!/hive-log4j.properties
OK
Time taken: 9.699 seconds
Loading data to table export_db.customer_data
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/export_db.db/customer_data/part-m-00000'
Table export_db.customer_data stats: [numFiles=1, numRows=0, totalSize=193, rawDataSize=0]
OK
Time taken: 9.629 seconds
[cloudera@quickstart Desktop]$ █
```

## See the results:

```
hive> select *from customer_data ;
OK
1      john Doe      john.doe@example.com
2      Jane Smith    Jane.smith@example.com
3      Robert Johnson robert.johnson@example.com
4      Lisa Brown     lisa.brown@example.com
5      Mischael Wilson michael.wilson@example.com
Time taken: 0.587 seconds, Fetched: 5 row(s)
hive> █
```

## →import purchase\_data table into hive:

```
[cloudera@quickstart Desktop]$ sqoop import --connect jdbc:mysql://localhost/project --username=root --password=cloudera
rehouse --hive-import --hive-overwrite --hive-table=export db.purchase_data -m 1 ;
```

```
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
```

```
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
```

```
Bytes Written=139
```

```
23/07/19 03:22:28 INFO mapreduce.ImportJobBase: Transferred 139 bytes in 282.1412 seconds (0.4927 bytes/sec)
```

```
23/07/19 03:22:28 INFO mapreduce.ImportJobBase: Retrieved 5 records.
```

```
23/07/19 03:22:29 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `purchase_data` AS t LIMIT 1
```

```
23/07/19 03:22:29 WARN hive.TableDefWriter: Column timestamp had to be cast to a less precise type in Hive
```

```
23/07/19 03:22:29 INFO hive.HiveImport: Loading uploaded data into Hive
```

```
Logging initialized using configuration in jar:file:/usr/jars/hive-common-1.1.0-cdh5.4.2.jar!/hive-log4j.properties
```

```
OK
```

```
Time taken: 8.462 seconds
```

```
Loading data to table export db.purchase_data
```

```
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/export_db.db/purchase_data/part-m-000000'
```

---

```
hive> select *from purchase_data ;
OK
1      2023-01-01 10:05:00      100
2      2023-01-01 10:08:00      150
3      2023-01-01 10:09:00      200
4      2023-01-01 10:13:00      120
5      2023-01-01 10:17:00       80
Time taken: 3.056 seconds, Fetched: 5 row(s)
hive> █
```

---

Last table come in hive (purchase\_data table).

### **3.TaskOutcomes:**

#### **1.Data Cleaning:**

➔Distinct element:

```

hive> select distinct userid from click_data ;
Query ID = cloudera_20230719044848_79bd735f-bfce-4238-82fd-
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input
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_1689744761876_0005, Tracking URL = http:
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_16
Hadoop job information for Stage-1: number of mappers: 1; n
2023-07-19 04:51:11,970 Stage-1 map = 0%, reduce = 0%
2023-07-19 04:52:12,765 Stage-1 map = 0%, reduce = 0%
2023-07-19 04:53:27,529 Stage-1 map = 0%, reduce = 0%
2023-07-19 04:54:35,095 Stage-1 map = 0%, reduce = 0%
2023-07-19 04:54:55,802 Stage-1 map = 100%, reduce = 0%, C
2023-07-19 04:55:56,786 Stage-1 map = 100%, reduce = 0%, C
2023-07-19 04:56:20,078 Stage-1 map = 100%, reduce = 67%,
2023-07-19 04:56:42,067 Stage-1 map = 100%, reduce = 100%,
MapReduce Total cumulative CPU time: 19 seconds 40 msec
Ended Job = job_1689744761876_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 19.22 se
Total MapReduce CPU Time Spent: 19 seconds 220 msec
OK
1
2
3
4
5
Time taken: 509.729 seconds, Fetched: 5 row(s)
hive> █

```

---

➔Data filter: use the where clause.

```
hive> select *from click_data where userid=1 ;
OK
1      2023-01-01 10:00:00      homepage
1      2023-01-01 10:01:00      product_page
Time taken: 86.939 seconds, Fetched: 2 row(s)
hive> select *from click_data where userid in(1,2,3) ;
OK
1      2023-01-01 10:00:00      homepage
1      2023-01-01 10:01:00      product_page
2      2023-01-01 10:02:00      homepage
2      2023-01-01 10:03:00      cart_page
3      2023-01-01 10:05:00      homepage
3      2023-01-01 10:06:00      product_page
3      2023-01-01 10:07:00      cart_page
Time taken: 2.064 seconds, Fetched: 7 row(s)
hive> █
```

---

**➔Data Aggregation :**

**Sum Function:**

```

hive> select sum(amount) from purchase_data ;
Query ID = cloudera_20230719050404_d9155598-7d1f-42b6-9c67-
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_1689744761876_0007, Tracking URL = http:
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_16
Hadoop job information for Stage-1: number of mappers: 1; r
2023-07-19 05:10:31,991 Stage-1 map = 0%, reduce = 0%
2023-07-19 05:11:34,034 Stage-1 map = 0%, reduce = 0%
2023-07-19 05:12:34,397 Stage-1 map = 0%, reduce = 0%
2023-07-19 05:13:30,802 Stage-1 map = 100%, reduce = 0%, (
2023-07-19 05:14:31,431 Stage-1 map = 100%, reduce = 0%, (
2023-07-19 05:15:06,229 Stage-1 map = 100%, reduce = 67%,
2023-07-19 05:15:24,269 Stage-1 map = 100%, reduce = 100%,
MapReduce Total cumulative CPU time: 12 seconds 400 msec
Ended Job = job_1689744761876_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 12.4 sec
Total MapReduce CPU Time Spent: 12 seconds 400 msec
OK
650
Time taken: 653.201 seconds, Fetched: 1 row(s)
hive> █

```

## **Min Function:**

```

hive> select min(amount) from purchase_data ;
Query ID = cloudera_20230719060000_2d9ef657-5ee2-4822-9011-bcff3e0e5079
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 17.03 sec HDFS Read: 6813 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 30 msec
OK
80
Time taken: 895.457 seconds, Fetched: 1 row(s)

```

## **Count:**

```
hive> select count(userid) from click_data ;
Query ID = cloudera_20230719050202_b2e34e99-cfec-4c0f-9e7a-8f60518d43ad
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_1689744761876_0006, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1689744761876_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1689744761876_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-07-19 05:04:22,007 Stage-1 map = 0%,   reduce = 0%
2023-07-19 05:05:23,520 Stage-1 map = 0%,   reduce = 0%
2023-07-19 05:06:24,781 Stage-1 map = 0%,   reduce = 0%
2023-07-19 05:07:25,172 Stage-1 map = 0%,   reduce = 0%

~~
2023-07-19 05:11:45,769 Stage-1 map = 100%,   reduce = 67%, Cumulative CPU 8.44 s
ec
2023-07-19 05:12:38,813 Stage-1 map = 100%,   reduce = 100%, Cumulative CPU 13.59
sec
MapReduce Total cumulative CPU time: 13 seconds 590 msec
Ended Job = job_1689744761876_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Reduce: 1   Cumulative CPU: 14.58 sec   HDFS Read: 7280 H
DFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 580 msec
OK
13
Time taken: 639.062 seconds, Fetched: 1 row(s)
```

## **Max Function:**

```
hive> select sum(amount) from purchase_data ;
Query ID = cloudera_20230719050404_d9155598-7d1f-42b6-9c67-4535b0611e64
Total jobs = 1
Stage-Stage-1: Map: 1   Reduce: 1   Cumulative CPU: 18.1 sec   HDFS Read: 6827 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 100 msec
OK
200
Time taken: 925.797 seconds, Fetched: 1 row(s)
```



### **3.Data Transformation:**

#### **Concat:**

```
hive> select concat (userid,' ',page) from click_data ;
OK
1 homepage
1 product_page
2 homepage
2 cart_page
3 homepage
3 product_page
3 cart_page
4 homepage
4 product_page
4 cart_page
4 checkout_page
5 home_page
5 product_page
Time taken: 2.71 seconds, Fetched: 13 row(s)
hive> █
```

---

#### **String Manipulation:**

```
hive> select upper(page) from click_data;
OK
HOMEPAGE
PRODUCT_PAGE
HOMEPAGE
CART_PAGE
HOMEPAGE
PRODUCT_PAGE
CART_PAGE
HOMEPAGE
PRODUCT_PAGE
CART_PAGE
CHECKOUT_PAGE
HOME_PAGE
PRODUCT_PAGE
Time taken: 1.001 seconds, Fetched: 13 row(s)
hive> █
```

---