



Name: Subrato Tapaswi	Class/Roll No.: D16AD/60	Grade:
------------------------------	---------------------------------	---------------

Title of Experiment: Use Sqoop to load data from RDBMS (weblog/ transactions data) and analyze it using HIVE/PIG.

Objective of Experiment:

The objective of this project is to use Sqoop, Hive, and Pig to efficiently extract, transform, and analyze data from a relational database management system (RDBMS), specifically weblog or transactional data

Outcome of Experiment:

Thus, we use Sqoop to load data from RDBMS (MySQL) and analyzed it using HIVE/PIG

Problem Statement:

The challenge is to efficiently extract, transform, and analyze large volumes of weblog or transactional data from a relational database using Sqoop, Hive, and Pig within a scalable and performance-optimized Hadoop ecosystem, ensuring data quality and delivering valuable insights for informed decision-making

Description / Theory:

Hadoop Eco-System:

The Hadoop ecosystem is a collection of open-source software tools and frameworks designed to process, store, and analyze large volumes of data in a distributed computing environment. Here's a brief overview of some key components within the Hadoop ecosystem:

- | | |
|---|--------------|
| 1. HDFS (Hadoop Distributed File System) | 6. Pig |
| 2. MapReduce | 7. HBase |
| 3. YARN (Yet Another Resource Negotiator) | 8. ZooKeeper |
| 4. Apache Spark | 9. Sqoop |
| 5. Hive | 10. Flume |



Vivekanand Education Society's Institute of Technology

Approved by AICTE & Affiliated to University of Mumbai

Artificial Intelligence and Data Science Department

Big Data Analytics/Odd Sem 2023-23/Experiment 3

Output:

```
[cloudera@quickstart ~]$ mysql -u root -pcloudera
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 32
Server version: 5.1.73 Source distribution
```

Copyright (c) 2000, 2013, 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 sales1;
ERROR 1007 (HY000): Can't create database 'sales1'; database exists
mysql> use sales1;
Database changed
mysql> CREATE TABLE sales(month number VARCHAR(5) not null primary key, facecream VARCHAR(20), facewash VARCHAR(20), toothpaste VARCHAR(20), bathings SOAP VARCHAR(20), shampoo VARCHAR(20), moisturizer VARCHAR(20), total_units VARCHAR(20), total_profit VARCHAR(20));
Query OK, 0 rows affected (0.92 sec)
```

```
mysql> LOAD Data Local Infile '/home/cloudera/Desktop/sales.csv' into table sales
Fields Terminated By ',' Lines Terminated By '\n';
Query OK, 13 rows affected, 1 warning (0.22 sec)
Records: 13 Deleted: 0 Skipped: 0 Warnings: 1
```

```
mysql> SELECT * FROM sales limit 5;
+-----+-----+-----+-----+-----+-----+-----+
| month_number | facecream | facewash | toothpaste | bathings SOAP | shampoo | moisturizer | total_units | total_profit |
+-----+-----+-----+-----+-----+-----+-----+
| 1           | 2500      | 1500     | 5200       | 9200          | 1200    | 150         | 21100       | 211000       |
| 10          | 1990      | 1890     | 8300       | 10300         | 2300    | 189         | 26670       | 266700       |
| 11          | 2340      | 2100     | 7300       | 13300         | 2400    | 210         | 41280       | 412800       |
| 12          | 2900      | 1760     | 7400       | 14400         | 1800    | 176         | 30020       | 300200       |
| 2           | 2630      | 1200     | 5100       | 6100          | 2100    | 120         | 18330       | 183300       |
+-----+-----+-----+-----+-----+-----+-----+
```

```
[cloudera@quickstart ~]$ sqoop list-tables --connect jdbc:mysql://localhost/sales --username root --password "cloudera"
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
```



Vivekanand Education Society's Institute of Technology

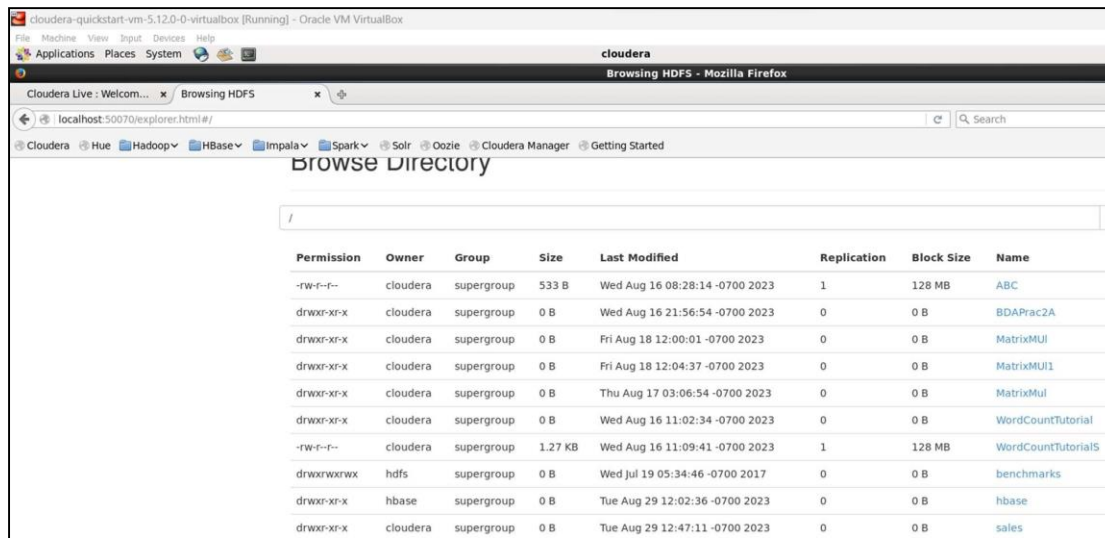
Approved by AICTE & Affiliated to University of Mumbai

Artificial Intelligence and Data Science Department Big Data Analytics/Odd Sem 2023-23/Experiment 3

```
23/10/11 18:48:32 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
23/10/11 18:48:32 WARN tool.BaseSqoopTool: Setting your password on the command-
line is insecure. Consider using -P instead.
23/10/11 18:48:36 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
sales
```

Importing tables from RDMS to HDFS using Sqoop:

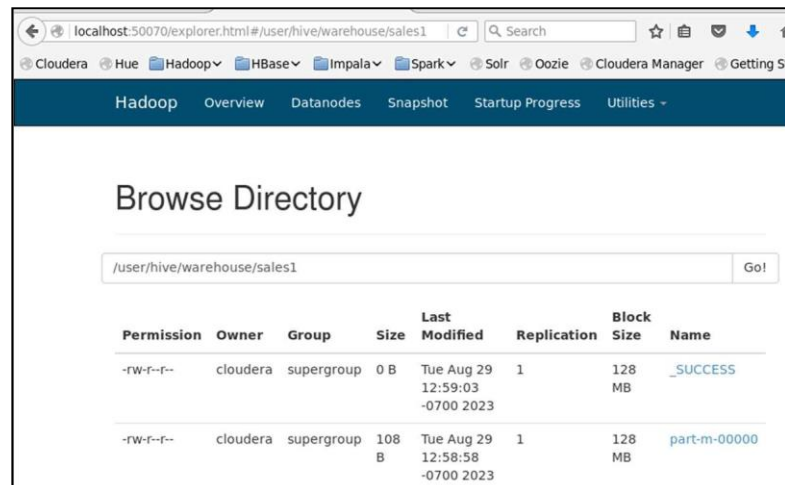
```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/sales --u
sername=root --password="cloudera" --table=sales --target-dir=/sales/sales --incr
emental append --check-column order_id --fields-terminated-by='\t';
```



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	supergroup	533 B	Wed Aug 16 08:28:14 -0700 2023	1	128 MB	ABC
drwxr-xr-x	cloudera	supergroup	0 B	Wed Aug 16 21:56:54 -0700 2023	0	0 B	BDAPrac2A
drwxr-xr-x	cloudera	supergroup	0 B	Fri Aug 18 12:00:01 -0700 2023	0	0 B	MatrixMUJ
drwxr-xr-x	cloudera	supergroup	0 B	Fri Aug 18 12:04:37 -0700 2023	0	0 B	MatrixMUJ1
drwxr-xr-x	cloudera	supergroup	0 B	Thu Aug 17 03:06:54 -0700 2023	0	0 B	MatrixMul
drwxr-xr-x	cloudera	supergroup	0 B	Wed Aug 16 11:02:34 -0700 2023	0	0 B	WordCountTutorial
-rw-r--r--	cloudera	supergroup	1.27 KB	Wed Aug 16 11:09:41 -0700 2023	1	128 MB	WordCountTutorialS
drwxrwxrwx	hdfs	supergroup	0 B	Wed Jul 19 05:34:46 -0700 2017	0	0 B	benchmarks
drwxr-xr-x	hbase	supergroup	0 B	Tue Aug 29 12:02:36 -0700 2023	0	0 B	hbase
drwxr-xr-x	cloudera	supergroup	0 B	Tue Aug 29 12:47:11 -0700 2023	0	0 B	sales

Importing Table From HDFS to HIVE:

```
[cloudera@quickstart ~]$ sqoop import-all-tables --connect jdbc:mysql://localhost/sales --username
root --password "cloudera" --warehouse-dir /user/hive/warehouse
```



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	supergroup	0 B	Tue Aug 29 12:59:03 -0700 2023	1	128 MB	_SUCCESS
-rw-r--r--	cloudera	supergroup	108 B	Tue Aug 29 12:58:58 -0700 2023	1	128 MB	part-m-00000



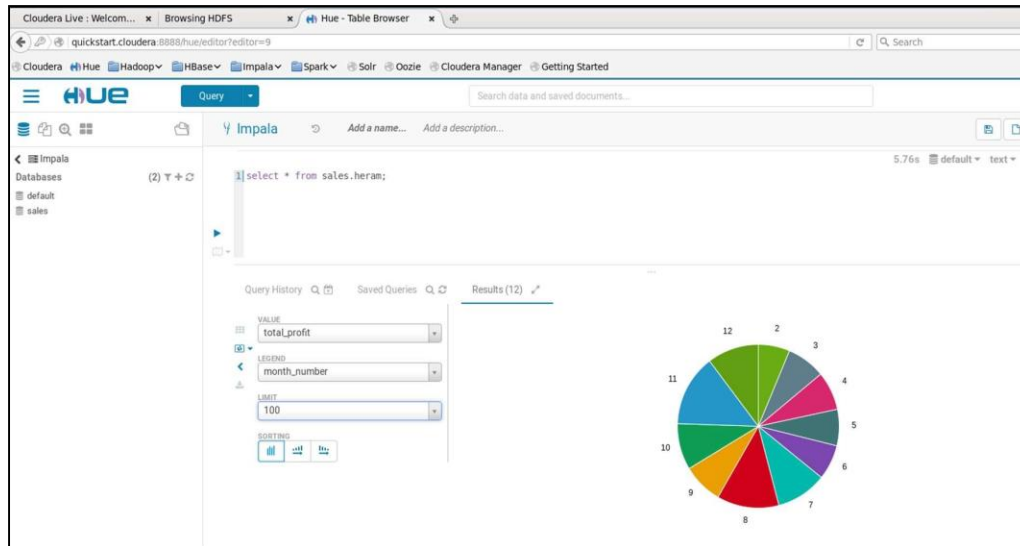
Vivekanand Education Society's Institute of Technology

Approved by AICTE & Affiliated to University of Mumbai

Artificial Intelligence and Data Science Department

Big Data Analytics/Odd Sem 2023-23/Experiment 3

Going to Hue Editor, Importing table, Writing Query And Doing Visualization.



Running Some Queries:

The screenshot shows the Hue Editor interface. The query editor contains the following SQL query:

```
1 SELECT *
2 FROM sales.hera
3 WHERE total_profit >= 300200;
```

The query results are displayed as a table with 9 columns: month_number, facecream, facewash, toothpaste, bathingssoap, shampoo, moisturizer, total_units, and total_profit. The table is titled "Results (3)".

month_number	facecream	facewash	toothpaste	bathingssoap	shampoo	moisturizer	total_units	total_profit
1	8	3700	1400	5860	9960	2860	1400	36140
2	11	2340	2100	7300	13300	2400	2100	41280
3	12	2900	1760	7400	14400	1800	1760	30020

Result and Discussion:

We started by creating a database and table in MySQL on Cloudera, imported the data into HDFS using Sqoop, and then used Hive for structured table creation. We analyzed the data using SQL queries and visualizations in Hue. This experiment highlighted Sqoop's data ingestion, Hive's analytical capabilities, and the power of Hadoop for large-scale data tasks, emphasizing the importance of data integration and analysis tools in the data-driven landscape.