



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

Name: Shreya Singh	Class/Roll No.: D16AD/55	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)
2. MapReduce
3. YARN (Yet Another Resource Negotiator)
4. Apache Spark
5. Hive
6. Pig
7. HBase
8. ZooKeeper
9. Sqoop
10. Flume



Hive:

Hive is like a translator for Hadoop. It allows you to write queries in a language similar to SQL (called HiveQL) and then translates those queries into MapReduce jobs that can be executed on a Hadoop cluster.

It's great for data analysts who are familiar with SQL because they can use Hive to query and analyze data stored in Hadoop's distributed file system (HDFS).

Pig:

Pig is a platform that simplifies the process of writing data transformations for Hadoop. Instead of writing complex Java code for MapReduce, you can use a simple scripting language called Pig Latin.

Pig is handy when you need to process and clean large amounts of data before analyzing it. It's especially useful for ETL (Extract, Transform, Load) tasks.

Sqoop:

Sqoop is a tool for efficiently transferring data between Hadoop and relational databases (like MySQL or Oracle). It helps you import data from databases into Hadoop or export data from Hadoop back to databases.

Sqoop is essential when you have data in traditional databases that you want to analyze with Hadoop. It makes the data import/export process straightforward and automated.



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 -uroot -pcloudera
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 22
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 sales;
Query OK, 1 row affected (0.00 sec)
```

```
mysql> use sales;
Database changed
```

```
mysql> LOAD DATA Local Infile '/home/cloudera/Desktop/Heramb/Heran.csv' into table sales1 Fields Terminated By ',' Lines Terminated By '\n';
Query OK, 13 rows affected, 9 warnings (0.02 sec)
Records: 13 Deleted: 0 Skipped: 0 Warnings: 0
```

```
mysql> select * from sales1;
```

month_number	facecream	facewash	toothpaste	bathingsoap	shampoo	moisturizer	total_units	total_profit
1	2500	1500	5200	9200	1200	1500	21100	211000
2	2630	1200	5100	6100	2100	1200	18330	183300
3	2140	1340	4550	9550	3550	1340	22470	224700
4	3400	1130	5870	8870	1870	1130	22270	222700
5	3600	1740	4560	7760	1560	1740	20960	209600

```
mysql> show tables
-> ;
```

Tables_in_sales
sales1



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

Importing tables from RDMS to HDFS using Sqoop:

```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/sales --username=root --password="cloudera" --table=sales1 --target-dir=/sales/sales --incremental append --check-column month_number --fields-terminated-by='\t';
```

The screenshot shows the Cloudera Manager web interface. The 'Browse Directory' view for the HDFS path '/sales/sales' is displayed. The table lists the contents of the directory, including files like 'ABC', 'BDAFile2A', 'MatrixML', 'MatrixML1', 'MatrixML', 'WordCountTutorial', 'WordCountTutorial5', 'benchmarks', 'hbase', and 'sales'.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxr-xr-x	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	BDAFile2A
drwxr-xr-x	cloudera	supergroup	0 B	Fri Aug 18 12:00:01 -0700 2023	0	0 B	MatrixML
drwxr-xr-x	cloudera	supergroup	0 B	Fri Aug 18 12:04:37 -0700 2023	0	0 B	MatrixML1
drwxr-xr-x	cloudera	supergroup	0 B	Thu Aug 17 03:06:54 -0700 2023	0	0 B	MatrixML
drwxr-xr-x	cloudera	supergroup	0 B	Wed Aug 16 11:02:34 -0700 2023	0	0 B	WordCountTutorial
-rwxr-xr-x	cloudera	supergroup	1.27 KB	Wed Aug 16 11:09:41 -0700 2023	1	128 MB	WordCountTutorial5
drwxrwxr-x	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 12:52:36 -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
```

The screenshot shows the Cloudera Manager web interface. The 'Browse Directory' view for the HDFS path '/user/hive/warehouse/sales1' is displayed. The table lists the contents of the directory, including files like '_SUCCESS' and 'part-m-00300'.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxr-xr-x	cloudera	supergroup	0 B	Tue Aug 29 12:59:03 -0700 2023	1	128 MB	_SUCCESS
-rwxr-xr-x	cloudera	supergroup	108 B	Tue Aug 29 12:58:56 -0700 2023	1	128 MB	part-m-00300



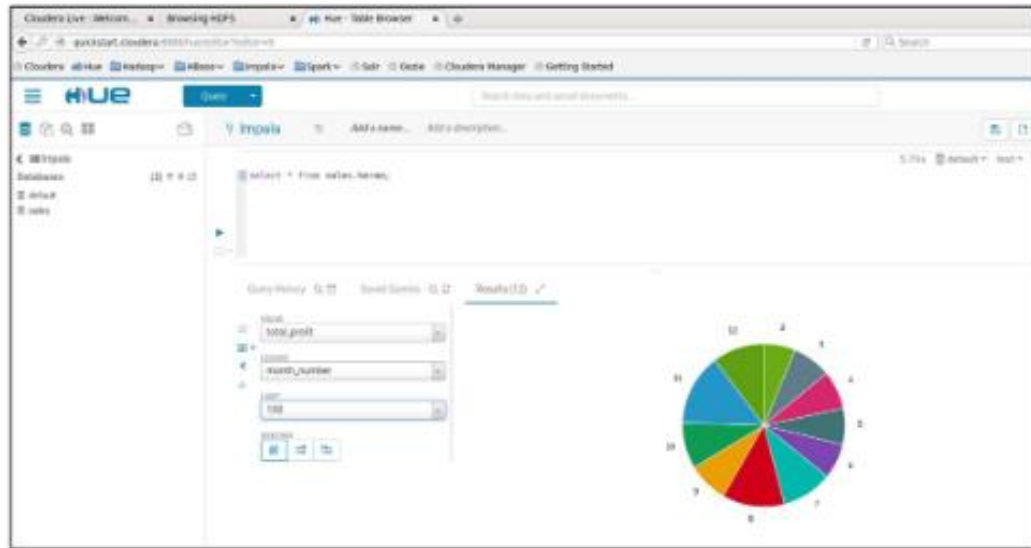
Vivekanand Education Society's Institute of Technology

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 with a query and its results. The query is: `SELECT * FROM sales.heram WHERE total_profit > 300000;`. The results are displayed in a table with 10 columns: `month_number`, `facecream`, `facewash`, `toothpaste`, `bodyingsnap`, `shampoo`, `moisturizer`, `total_price`, and `tax`.

	month_number	facecream	facewash	toothpaste	bodyingsnap	shampoo	moisturizer	total_price	tax
1	8	3700	1400	5800	9400	2800	1800	30100	261
2	11	2340	2100	7900	13300	2400	2100	41000	412
3	12	2400	1700	7400	14400	1800	1700	30200	300

The screenshot shows the Hue Editor interface with an insert query: `Insert into sales.heram values(13,121,345,56,435,43,43,568,234235);`. Below the query, a success message is displayed: `Success.`



Result and Discussion:

We began by accessing MySQL on Cloudera and creating a database and table to serve as our source of data. Next, we employed Sqoop to seamlessly import this data into the Hadoop Distributed File System (HDFS), facilitating easy access and processing. Once the data resided in HDFS, we leveraged HIVE to import the data into its structured tables, making it readily available for querying and analysis.

Finally, using the Hue editor, we crafted SQL queries and performed visualizations to gain insights from the imported data. This experiment enabled us to seamlessly bridge the gap between traditional RDBMS data and Hadoop's distributed processing capabilities, showcasing the power of Sqoop as a data ingestion tool and the analytical prowess of HIVE and PIG.

In conclusion, our experiment successfully demonstrated the efficient transfer of data from an RDBMS to HDFS using Sqoop, followed by the analysis of this data through HIVE and PIG. This process showcased the versatility of Hadoop in handling large-scale data analysis tasks and emphasized the importance of data integration and analysis tools like Sqoop, HIVE, and PIG in today's data-driven landscape.