

# **Hadoop Ecosystem**

2022F-T1 BDM 1003

Group D  
Anargha Manoj  
Harcharan Singh  
Jovi Fez Bartolata  
Sharon Victor

# Table of Contents

|             |  |    |
|-------------|--|----|
| <b>I.</b>   | <b>Objectives.....</b>   | 4  |
| <b>II.</b>  | <b>Version Requirements .....</b>  | 4  |
| <b>III.</b> | <b>Launch Hadoop.....</b>  | 5  |
| •           | Start instance and connect using PuTTY.....  | 5  |
| •           | Update core-site.xml .....   | 5  |
| •           | Update yarn-site.xml.....  | 6  |
| •           | Update mapred-site.xml .....   | 6  |
| •           | Start DFS, Start YARN, Run JPS .....   | 7  |
| •           | Access using browser.....  | 7  |
| <b>IV.</b>  | <b>Download Dependencies .....</b>   | 8  |
| •           | Install 7zip.....  | 8  |
| •           | Download hivexmlserde-1.0.5.3.jar .....  | 9  |
| <b>V.</b>   | <b>Initial Set-up.....</b>   | 10 |
| •           | Create /LDZ and /LDZ/data directories .....  | 10 |
| •           | Launch Hive, set properties, and add jar file .....  | 11 |
| •           | Create /DWZ database .....   | 11 |
| •           | Create external table POSTS_EXT with SerDe .....   | 12 |
| •           | Create external table POSTS_EXT_DYN with partitioning by Creation Date and Post Type ..... | 14 |
| <b>VI.</b>  | <b>Performing Daily Update .....</b>   | 15 |
| •           | Step 1: <i>get_raw_file.sh</i> .....   | 16 |
| ○           | Usage and Result of <i>get_raw_file.sh</i> .....   | 16 |
| •           | Step 2: <i>update_table.sh</i> .....   | 18 |
| ○           | Script: <i>load_data.hql</i> .....   | 18 |
| ○           | Usage and Result of <i>update_table.sh</i> .....   | 20 |
| <b>VII.</b> | <b>Performing Queries .....</b>  | 22 |
| •           | Top 10 Most Answered Stack Overflow Questions.....   | 22 |
| ○           | Script: <i>query_top10_answered.sh</i> .....   | 22 |
| ○           | Usage and Result of <i>query_top10_answered.sh</i> .....                                   | 23 |
| •           | Percentage of Stack Overflow Questions that Went Unanswered .....                          | 25 |
| ○           | Script: <i>query_uanswered.sh</i> .....  | 25 |
| ○           | Usage and Result of <i>query_uanswered.sh</i> .....  | 26 |

|  |           |
|--|-----------|
| <b>VIII. Appendix: Scripts and Commands.....</b> | <b>27</b> |
| • create_tables.hql.....                         | 27        |
| • get_raw_file.sh.....                           | 29        |
| • update_table.sh.....                           | 29        |
| • load_data.hql.....                             | 29        |
| • query_top10_answered.sh.....                   | 30        |
| • query_uanswered.sh.....                        | 31        |

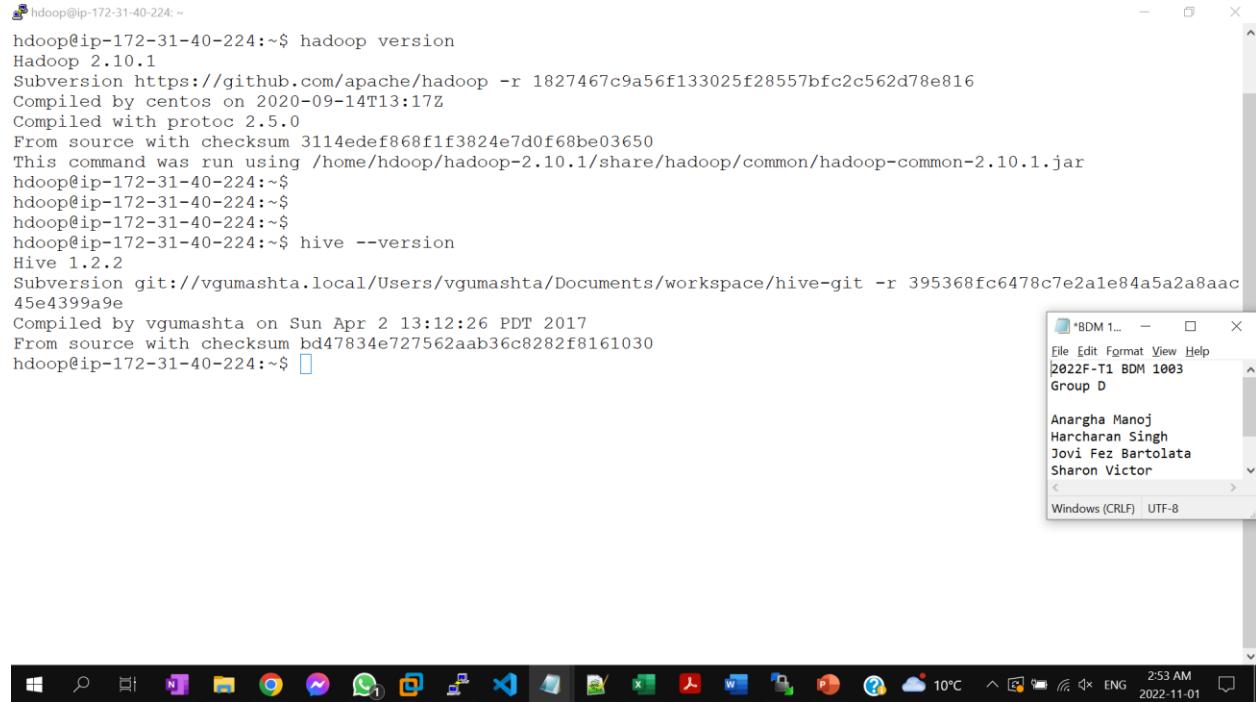
## I. Objectives

This project aims to:

- Build a data warehouse for Stack Overflow (stackoverflow.com) data in Hadoop ecosystem
- Perform queries using the archived data in Hive
- Automate the processes for daily data warehouse update and query runs

## II. Version Requirements

This project used Hadoop 2.10.1 and Hive 1.2.2. The XML SerDe file (hivexmlserde-1.0.5.3.jar) may not be compatible with the newer versions of Hive.

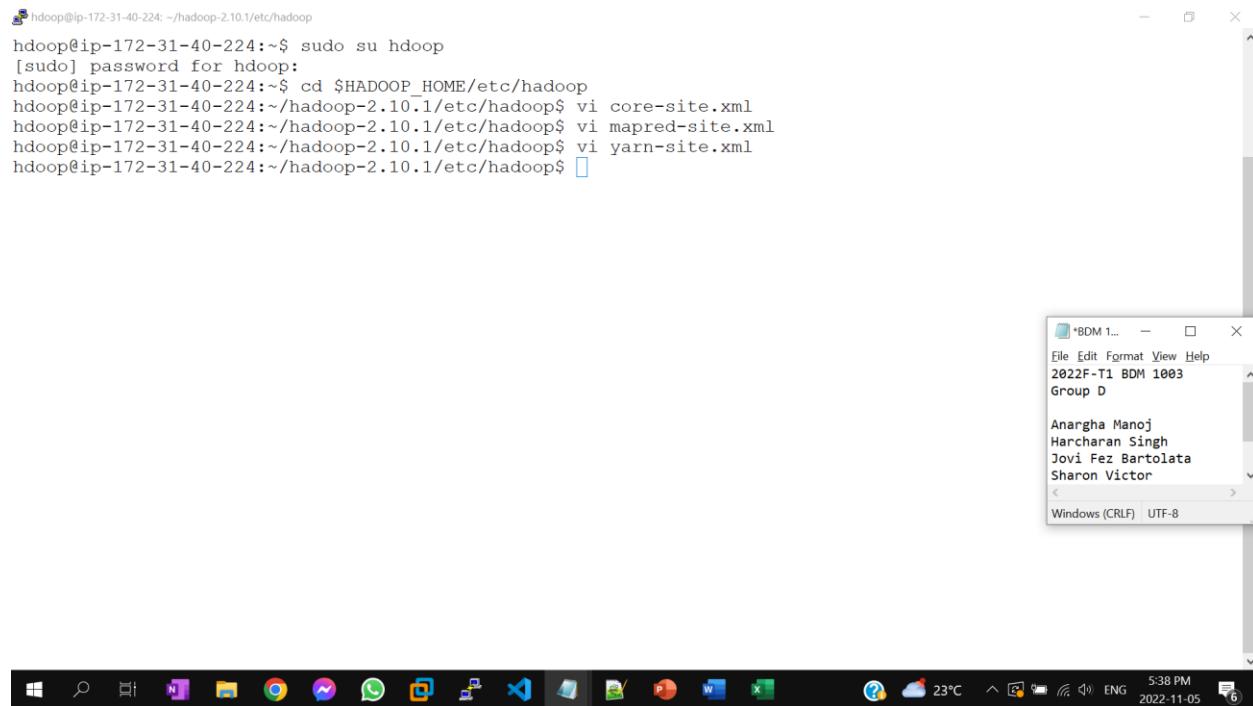


hadoop@ip-172-31-40-224: ~  
hadoop@ip-172-31-40-224:~\$ hadoop version  
Hadoop 2.10.1  
Subversion https://github.com/apache/hadoop -r 1827467c9a56f133025f28557bfc2c562d78e816  
Compiled by centos on 2020-09-14T13:17Z  
Compiled with protoc 2.5.0  
From source with checksum 3114edef868f1f3824e7d0f68be03650  
This command was run using /home/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-common-2.10.1.jar  
hadoop@ip-172-31-40-224:~\$  
hadoop@ip-172-31-40-224:~\$  
hadoop@ip-172-31-40-224:~\$  
hadoop@ip-172-31-40-224:~\$ hive --version  
Hive 1.2.2  
Subversion git://vgumashta.local/Users/vgumashta/Documents/workspace/hive-git -r 395368fc6478c7e2a1e84a5a2a8aac  
45e4399a9e  
Compiled by vgumashta on Sun Apr 2 13:12:26 PDT 2017  
From source with checksum bd47834e727562aab36c8282f8161030  
hadoop@ip-172-31-40-224:~\$

+BDM 1... — □ ×  
File Edit Format View Help  
2022F-T1 BDM 1003  
Group D  
Anargha Manoj  
Harcharan Singh  
Jovi Fez Bartolata  
Sharon Victor  
Windows (CRLF) UTF-8

### III. Launch Hadoop

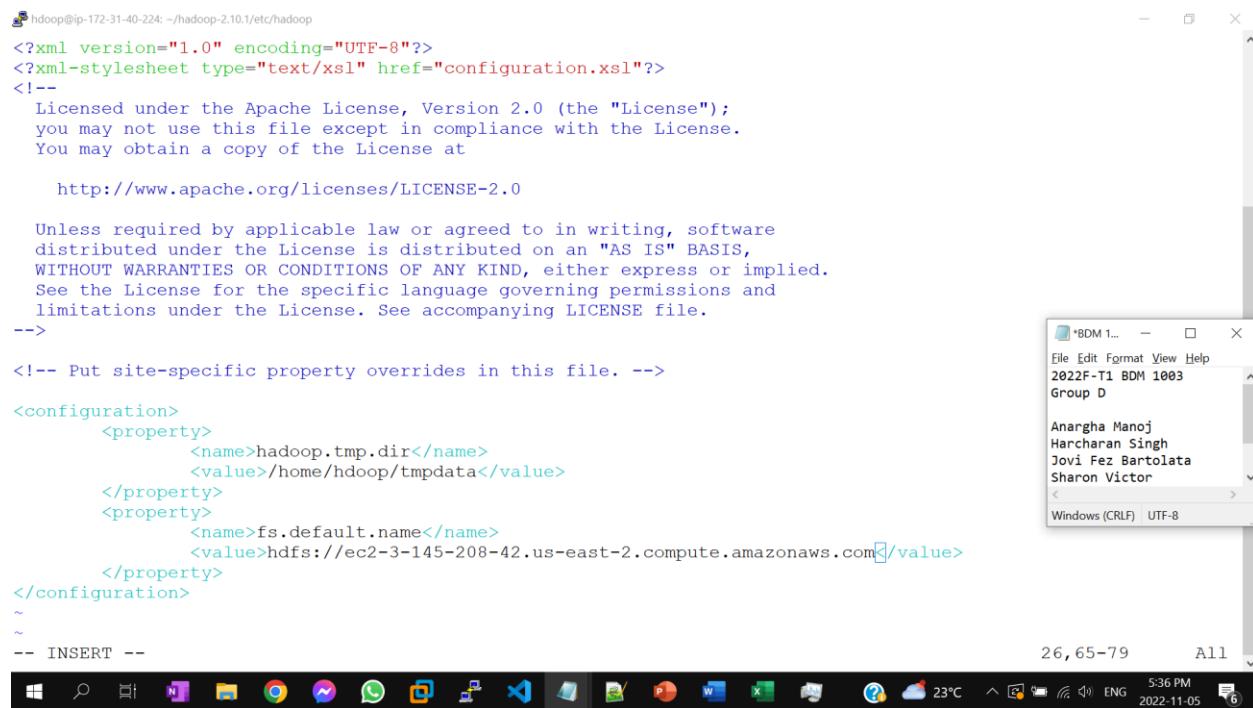
- Start instance and connect using PuTTY.



hadoop@ip-172-31-40-224: ~/hadoop-2.10.1/etc/hadoop  
[sudo] password for hadoop:  
hadoop@ip-172-31-40-224:~\$ cd \$HADOOP\_HOME/etc/hadoop  
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/etc/hadoop\$ vi core-site.xml  
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/etc/hadoop\$ vi mapred-site.xml  
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/etc/hadoop\$ vi yarn-site.xml  
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/etc/hadoop\$

\*BDM 1... File Edit Format View Help 2022F-T1 BDM 1003 Group D Anargha Manoj Harcharan Singh Jovi Fez Bartolata Sharon Victor < > Windows (CRLF) UTF-8

- Update core-site.xml  
fs.default.name = hdfs://<new IP address>



```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
 Licensed under the Apache License, Version 2.0 (the "License");
 You may not use this file except in compliance with the License.
 You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>hadoop.tmp.dir</name>
        <value>/home/hadoop/tmpdata</value>
    </property>
    <property>
        <name>fs.default.name</name>
        <value>hdfs://ec2-3-145-208-42.us-east-2.compute.amazonaws.com</value>
    </property>
</configuration>
~
~
-- INSERT --
```

26,65-79 All

- Update yarn-site.xml  
yarn.resourcemanager.hostname = <new IP address>

```
hadoop@ip-172-31-40-224: ~/hadoop-2.10.1/etc/hadoop
You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.

-->
<configuration>

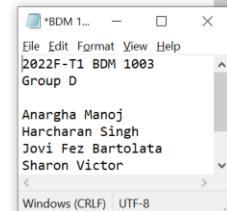
<!-- Site specific YARN configuration properties -->

<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
    <name>yarn.resourcemanager.hostname</name>
    <value>ec2-3-145-208-42.us-east-2.compute.amazonaws.com</value>
</property>
<property>
    <name>yarn.acl.enable</name>
    <value>0</value>
</property>
-- INSERT --
29,57-64        40%  
Windows (CRLF)  UTF-8  
File Edit Format View Help  
2022F-T1 BDM 1003  
Group D  
Anargha Manoj  
Harcharan Singh  
Jovi Fez Bartolata  
Sharon Victor  
< >  
*BDM 1...
```

- Update mapred-site.xml
    - mapreduce.jobtracker.address = <new IP address>
    - mapreduce.map.memory.mb = 4096
    - mapreduce.reduce.memory.mb = 4096

- Start DFS, Start YARN, Run JPS

```
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$ cd $HADOOP_HOME/sbin
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$ ./start-dfs.sh
Starting namenodes on [ec2-3-145-208-42.us-east-2.compute.amazonaws.com]
ec2-3-145-208-42.us-east-2.compute.amazonaws.com: starting namenode, logging to /home/hadoop/hadoop-2.10.1/logs/
hadoop-hadoop-namenode-ip-172-31-40-224.out
localhost: starting datanode, logging to /home/hadoop/hadoop-2.10.1/logs/hadoop-hadoop-datanode-ip-172-31-40-224.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/hadoop/hadoop-2.10.1/logs/hadoop-hadoop-secondarynamenode-ip-172-31-40-224.out
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$ ./start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/hadoop/hadoop-2.10.1/logs/yarn-hadoop-resourcemanager-ip-172-31-40-224.out
localhost: starting nodemanager, logging to /home/hadoop/hadoop-2.10.1/logs/yarn-hadoop-nodemanager-ip-172-31-40-224.out
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$ hdfs dfs -ls /
4913 SecondaryNameNode
4466 NameNode
5255 NodeManager
5355 Jps
4636 DataNode
5084 ResourceManager
hadoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$ jps
```



- Access using browser

The screenshot shows the Hadoop cluster management interface. The main content area displays the following metrics:

- Cluster Metrics:**

|  | Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Used Resources       |
|--|----------------|--------------|--------------|----------------|--------------------|----------------------|
|  | 0              | 0            | 0            | 0              | 0                  | <memory:0, vCores:0> |
- Cluster Nodes Metrics:**

|  | Active Nodes | Decommissioning Nodes | Decommissioned Nodes | Lost Nodes | Unheal |
|--|--------------|-----------------------|----------------------|------------|--------|
|  | 1            | 0                     | 0                    | 0          | 0      |
- Scheduler Metrics:**

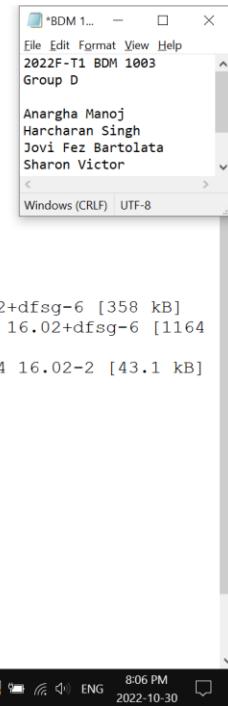
|  | Scheduler Type     | Scheduling Resource Type  | Minimum Alloc           |
|--|--------------------|---|-------------------------|
|  | Capacity Scheduler | [<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>] | <memory:1024, vCores:0> |

The sidebar on the left includes links for Cluster, About, Nodes, Node Labels, Applications, Scheduler, and Tools. The bottom of the screen shows a Windows taskbar with various icons and system status.

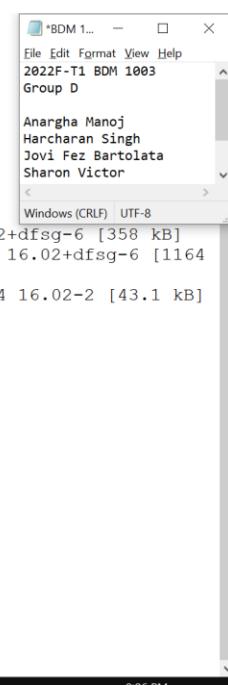
## IV. Download Dependencies

- Install 7zip

```
hadoop@ip-172-31-40-224:~$ sudo apt install p7zip-full p7zip-rar
[sudo] password for hadoop:
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  p7zip
The following NEW packages will be installed:
  p7zip p7zip-full p7zip-rar
0 upgraded, 3 newly installed, 0 to remove and 7 not upgraded.
Need to get 1565 kB of archives.
After this operation, 5868 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/universe amd64 p7zip amd64 16.02+dfsg-6 [358 kB]
Get:2 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/universe amd64 p7zip-full amd64 16.02+dfsg-6 [1164 kB]
Get:3 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/multiverse amd64 p7zip-rar amd64 16.02-2 [43.1 kB]
Fetched 1565 kB in 3s (470 kB/s)
Selecting previously unselected package p7zip.
(Reading database ... 130272 files and directories currently installed.)
Preparing to unpack .../p7zip_16.02+dfsg-6_amd64.deb ...
Unpacking p7zip (16.02+dfsg-6) ...
Selecting previously unselected package p7zip-full.
Preparing to unpack .../p7zip-full_16.02+dfsg-6_amd64.deb ...
Unpacking p7zip-full (16.02+dfsg-6) ...
Selecting previously unselected package p7zip-rar.
Preparing to unpack .../p7zip-rar_16.02-2_amd64.deb ...
Unpacking p7zip-rar (16.02-2) ...
Setting up p7zip (16.02+dfsg-6) ...
Setting up p7zip-full (16.02+dfsg-6) ...
Setting up p7zip-rar (16.02-2) ...
```



```
hadoop@ip-172-31-40-224:~$ 
Reading state information... Done
The following additional packages will be installed:
  p7zip
The following NEW packages will be installed:
  p7zip p7zip-full p7zip-rar
0 upgraded, 3 newly installed, 0 to remove and 7 not upgraded.
Need to get 1565 kB of archives.
After this operation, 5868 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/universe amd64 p7zip amd64 16.02+dfsg-6 [358 kB]
Get:2 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/universe amd64 p7zip-full amd64 16.02+dfsg-6 [1164 kB]
Get:3 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/multiverse amd64 p7zip-rar amd64 16.02-2 [43.1 kB]
Fetched 1565 kB in 3s (470 kB/s)
Selecting previously unselected package p7zip.
(Reading database ... 130272 files and directories currently installed.)
Preparing to unpack .../p7zip_16.02+dfsg-6_amd64.deb ...
Unpacking p7zip (16.02+dfsg-6) ...
Selecting previously unselected package p7zip-full.
Preparing to unpack .../p7zip-full_16.02+dfsg-6_amd64.deb ...
Unpacking p7zip-full (16.02+dfsg-6) ...
Selecting previously unselected package p7zip-rar.
Preparing to unpack .../p7zip-rar_16.02-2_amd64.deb ...
Unpacking p7zip-rar (16.02-2) ...
Setting up p7zip (16.02+dfsg-6) ...
Setting up p7zip-full (16.02+dfsg-6) ...
Setting up p7zip-rar (16.02-2) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
hadoop@ip-172-31-40-224:~$ 
```



- Download hivexmlserde-1.0.5.3.jar

Save the jar file in the /lib folder

```
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib$ wget http://search.maven.org/remotecontent/+/spss/hive/serde2/xml/hivexmlserde/1.0.5.3/hivexmlserde-1.0.5.3.jar
--2022-10-31 00:29:33-- http://search.maven.org/remotecontent?filepath=com/ibm/spss/hive/serde2/xml/hivexmlserde/1.0.5.3/hivexmlserde-1.0.5.3.jar
Resolving search.maven.org (search.maven.org)... 3.214.38.196, 34.203.165.39
Connecting to search.maven.org (search.maven.org) |3.214.38.196|:80... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://repo1.maven.org/maven2/com/ibm/spss/hive/serde2/xml/hivexmlserde/1.0.5.3/hivexmlserde-1.0.5.3.jar [following]
--2022-10-31 00:29:33-- https://repo1.maven.org/maven2/com/ibm/spss/hive/serde2/xml/hivexmlserde/1.0.5.3/hivexmlserde-1.0.5.3.jar
Resolving repo1.maven.org (repo1.maven.org)... 146.75.32.209
Connecting to repo1.maven.org (repo1.maven.org) |146.75.32.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 50656 (49K) [application/java-archive]
Saving to: 'remotecontent?filepath=com%2Fibm%2Fspss%2Fhive%2Fserde2%2FXML%2Fhivexmlserde%2F1.0.5.3%2Fhivexmlserde-1.0.5.3.jar'

remotecontent?filepath=com% 100%[=====] 49.47K --.-KB/s in 0.01s

2022-10-31 00:29:33 (4.59 MB/s) - 'remotecontent?filepath=com%2Fibm%2Fspss%2Fhive%2Fserde2%2FXML%2Fhivexmlserde%2F1.0.5.3%2Fhivexmlserde-1.0.5.3.jar' saved [50656/50656]

hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib$ mv remotecontent?filepath=com%2Fibm%2Fspss%2Fhive%2Fserde2%2FXML%2Fhivexmlserde%2F1.0.5.3%2Fhivexmlserde-1.0.5.3.jar hivexmlserde-1.0.5.3.jar
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib$ 
```



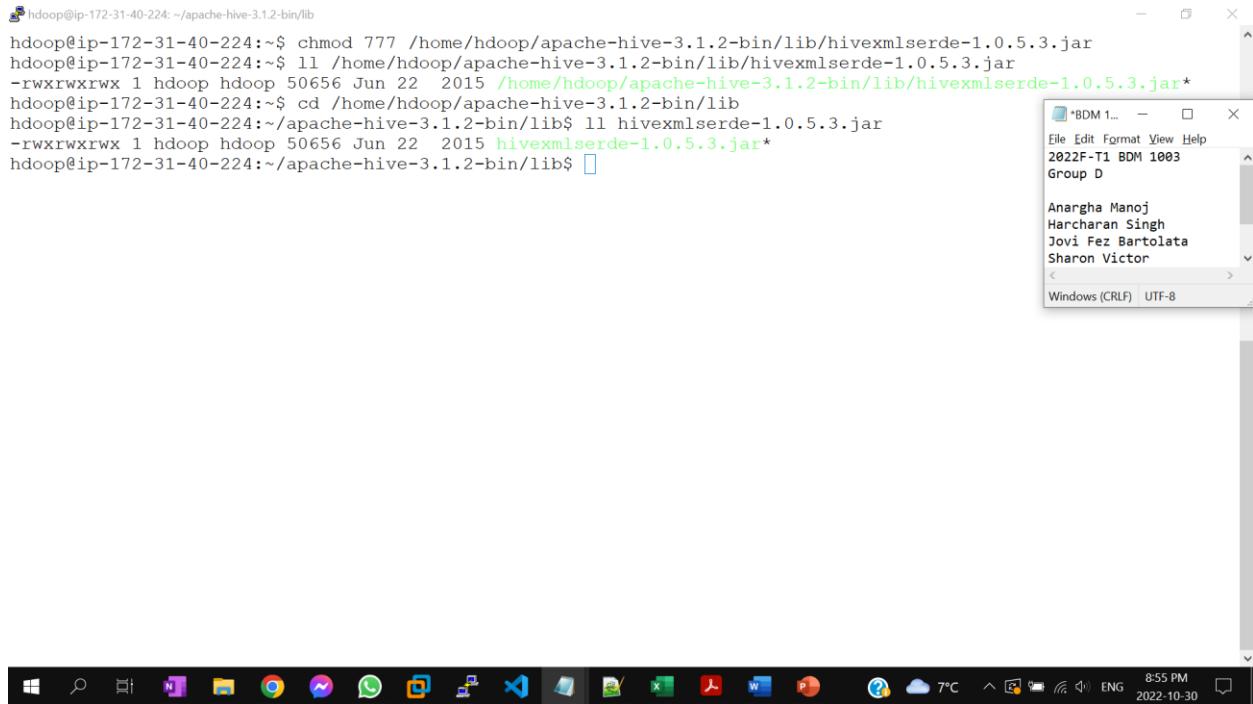
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib

```
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib$ ls
HikariCP-2.6.1.jar
ST4-4.0.4.jar
accumulo-core-1.7.3.jar
accumulo-fate-1.7.3.jar
accumulo-start-1.7.3.jar
accumulo-trace-1.7.3.jar
aircompressor-0.10.jar
ant-1.9.1.jar
ant-launcher-1.9.1.jar
antlr-runtime-3.5.2.jar
antlr4-runtime-4.5.jar
aopalliance-repackaged-2.5.0-b32.jar
apache-curator-2.12.0.pom
apache-jsp-9.3.20.v20170531.jar
apache-jstl-9.3.20.v20170531.jar
arrow-format-0.8.0.jar
arrow-memory-0.8.0.jar
arrow-vector-0.8.0.jar
asm-5.0.1.jar
asm-commons-5.0.1.jar
asm-tree-5.0.1.jar
audience-annotations-0.5.0.jar
avatica-1.11.0.jar
avro-1.7.7.jar
bonecp-0.8.0.RELEASE.jar
calcite-core-1.16.0.jar
calcite-druid-1.16.0.jar
calcite-linq4j-1.16.0.jar
commons-cli-1.2.jar
commons-codec-1.7.jar
hive-streaming-3.1.2.jar
hive-testutils-3.1.2.jar
hive-upgrade-acid-3.1.2.jar
hive-vector-code-gen-3.1.2.jar
hivexmlserde-1.0.5.3.jar
hk2-api-2.5.0-b32.jar
hk2-locator-2.5.0-b32.jar
hk2-utils-2.5.0-b32.jar
hpc-0.7.2.jar
htrace-core-3.2.0-incubating.jar
httpclient-4.5.2.jar
httpcore-4.4.4.jar
ivy-2.4.0.jar
jackson-annotations-2.9.5.jar
jackson-core-2.9.5.jar
jackson-core-asl-1.9.13.jar
jackson-databind-2.9.5.jar
jackson-dataformat-smile-2.9.5.jar
jackson-mapper-asl-1.9.13.jar
jamon-runtime-2.3.1.jar
janino-2.7.6.jar
javassist-3.20.0-GA.jar
javax.annotation-api-1.2.jar
javax.inject-2.5.0-b32.jar
javax.jdo-3.2.0-m3.jar
javax.servlet-api-3.1.0.jar
javax.servlet.jsp-2.3.2.jar
javax.servlet.jsp-api-2.3.1.jar
javax.ws.rs-api-2.0.1.jar
javolution-5.5.1.jar
```



HP Support Assistant  
Recommended action needed.

## Change permission of jar file



```
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib
```

```
hadoop@ip-172-31-40-224:~$ chmod 777 /home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar
hadoop@ip-172-31-40-224:~$ ll /home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar
-rwxrwxrwx 1 hadoop hadoop 50656 Jun 22 2015 /home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar*
hadoop@ip-172-31-40-224:~$ cd /home/hadoop/apache-hive-3.1.2-bin/lib
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib$ ll hivexmlserde-1.0.5.3.jar
-rwxrwxrwx 1 hadoop hadoop 50656 Jun 22 2015 hivexmlserde-1.0.5.3.jar*
```

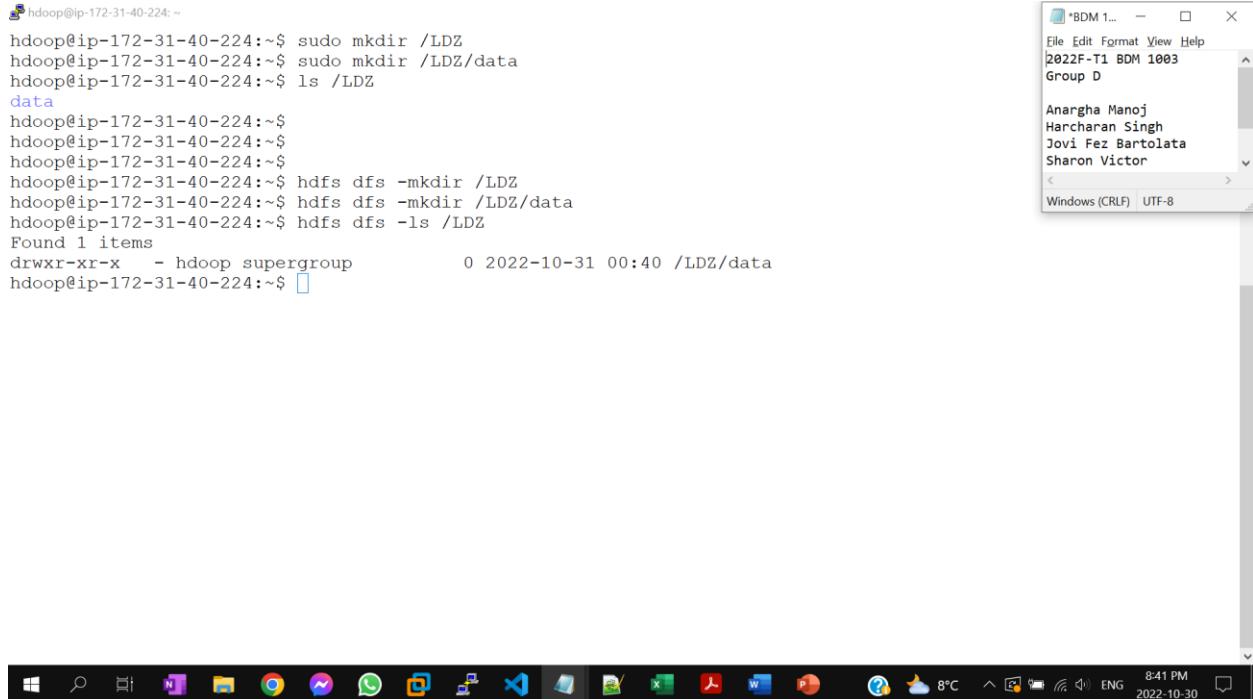
```
hadoop@ip-172-31-40-224:~/apache-hive-3.1.2-bin/lib$ 
```

```
*BDM 1... — X
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF) | UTF-8
```

Windows taskbar at the bottom: 8:55 PM, ENG, 2022-10-30.

## V. Initial Set-up

- Create /LDZ and /LDZ/data directories



```
hadoop@ip-172-31-40-224:~
```

```
hadoop@ip-172-31-40-224:~$ sudo mkdir /LDZ
hadoop@ip-172-31-40-224:~$ sudo mkdir /LDZ/data
hadoop@ip-172-31-40-224:~$ ls /LDZ
data
hadoop@ip-172-31-40-224:~$
hadoop@ip-172-31-40-224:~$ hdfs dfs -mkdir /LDZ
hadoop@ip-172-31-40-224:~$ hdfs dfs -mkdir /LDZ/data
hadoop@ip-172-31-40-224:~$ hdfs dfs -ls /LDZ
Found 1 items
drwxr-xr-x - hadoop supergroup          0 2022-10-31 00:40 /LDZ/data
hadoop@ip-172-31-40-224:~$ 
```

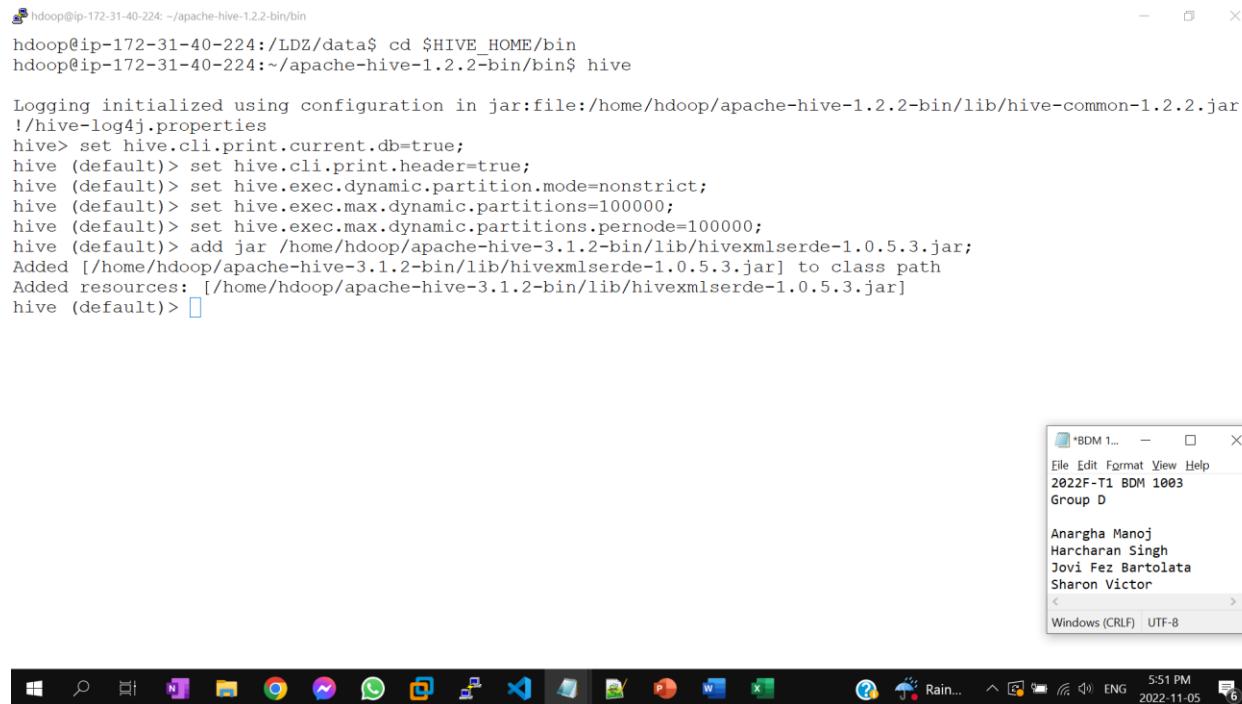
```
*BDM 1... — X
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF) | UTF-8
```

Windows taskbar at the bottom: 8:41 PM, ENG, 2022-10-30.

- Launch Hive, set properties, and add jar file

```
hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:/LDZ/data$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ hive

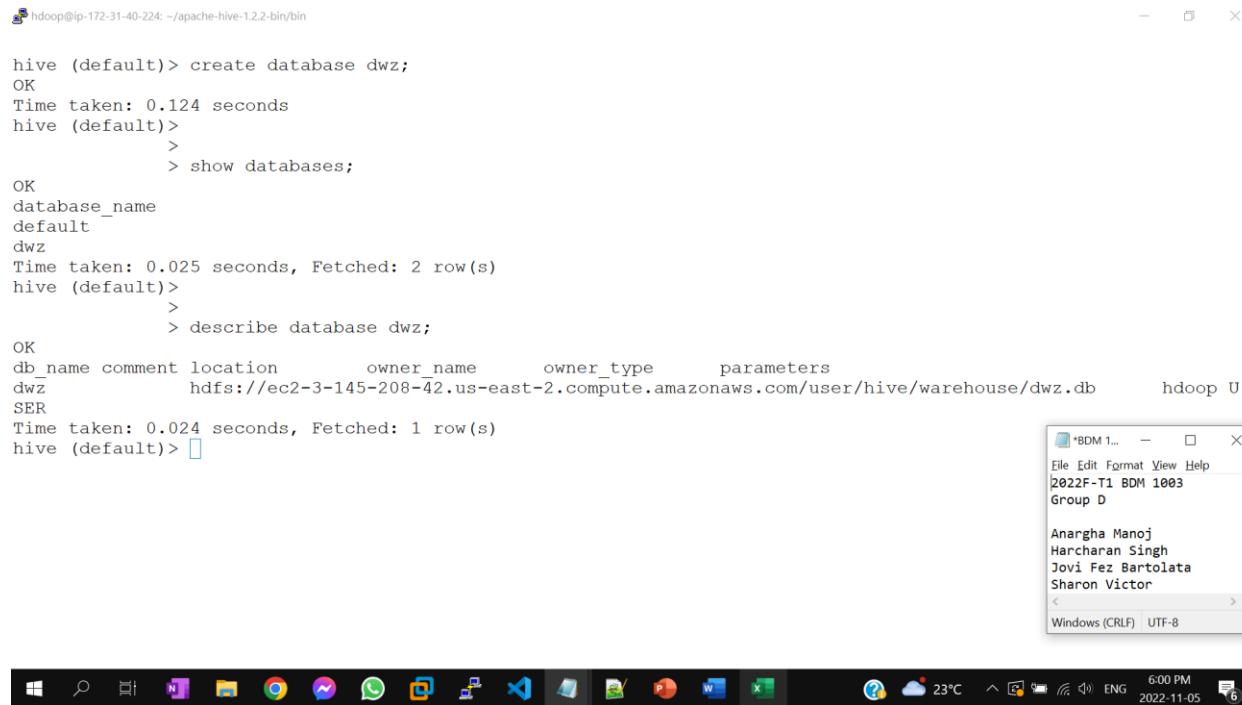
Logging initialized using configuration in jar:file:/home/hadoop/apache-hive-1.2.2-bin/lib/hive-common-1.2.2.jar
!/hive-log4j.properties
hive> set hive.cli.print.current.db=true;
hive (default)> set hive.cli.print.header=true;
hive (default)> set hive.exec.dynamic.partition.mode=nonstrict;
hive (default)> set hive.exec.max.dynamic.partitions=100000;
hive (default)> set hive.exec.max.dynamic.partitions.pernode=100000;
hive (default)> add jar /home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar;
Added [/home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar] to class path
Added resources: [/home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar]
hive (default)>
```



The screenshot shows a Windows desktop environment. At the top is a taskbar with various icons. In the center is a terminal window with a black background and white text, displaying the command-line session above. To the right of the terminal is a smaller window titled 'BDM 1...' containing a list of names: Anargha Manoj, Harcharan Singh, Jovi Fez Bartolata, and Sharon Victor. The desktop background is light blue.

- Create /DWZ database

```
hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hive (default)> create database dwz;
OK
Time taken: 0.124 seconds
hive (default)>
> show databases;
OK
database_name
default
dwz
Time taken: 0.025 seconds, Fetched: 2 row(s)
hive (default)>
>
> describe database dwz;
OK
db_name comment location          owner_name      owner_type      parameters
dwz           hdfs://ec2-3-145-208-42.us-east-2.compute.amazonaws.com/user/hive/warehouse/dwz.db      hdoop U
SER
Time taken: 0.024 seconds, Fetched: 1 row(s)
hive (default)>
```



The screenshot shows a Windows desktop environment, identical to the one above. It features a taskbar with various icons, a terminal window with a black background and white text showing the creation of a database and its description, and a Notepad window titled 'BDM 1...' listing the same four names: Anargha Manoj, Harcharan Singh, Jovi Fez Bartolata, and Sharon Victor. The desktop background is light blue.

- Create external table POSTS\_EXT with SerDe

For the commands, please refer to Appendix: create\_tables.hql

```
hdooop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hive (default)> use dwz;
OK
Time taken: 0.026 seconds
hive (dwz)> create external table posts_ext (
    > CreationDateTime string,
    > LastEditorUserId string,
    > LastEditorDisplayName string,
    > LastEditDateTime string,
    > LastActivityDateTime string,
    > CommunityOwnedDateTime string,
    > ContentLicense string,
    > Id string,
    > PostTypeId string,
    > AcceptedAnswerId string,
    > Score string,
    > ViewCount string,
    > Title string,
    > Body string,
    > OwnerUserId string,
    > Tags string,
    > AnswerCount string,
    > CommentCount string,
    > FavoriteCount string,
    > ParentId string
    > )
    > row format serde 'com.ibm.spss.hive.serde2.xml.XmlSerDe' with serdeproperties (
    > "column.xpath.CreationDateTime"="/row/@CreationDate",
    > "column.xpath.LastEditorUserId"="/row/@LastEditorUserId",
    > "column.xpath.LastEditorDisplayName"="/row/@LastEditorDisplayName",
    > "column.xpath.LastEditDateTime"="/row/@LastEditDate",
    > "column.xpath.LastActivityDateTime"="/row/@LastActivityDate",
    > )
```

| *BDM 1...                  |  |
|----------------------------|--|
| File Edit Format View Help |  |
| 2022F-T1 BDM 1003          |  |
| Group D                    |  |
| Anargha Manoj              |  |
| Harcharan Singh            |  |
| Jovi Fez Bartolata         |  |
| Sharon Victor              |  |
| < >                        |  |
| Windows (CRLF) UTF-8       |  |

```
Windows taskbar showing various icons and system status.
```

```
hdooop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
    > "column.xpath.CommunityOwnedDateTime"="/row/@CommunityOwnedDate",
    > "column.xpath.ContentLicense"="/row/@ContentLicense",
    > "column.xpath.Id"="/row/@Id",
    > "column.xpath.PostTypeId"="/row/@PostTypeId",
    > "column.xpath.AcceptedAnswerId"="/row/@AcceptedAnswerId",
    > "column.xpath.Score"="/row/@Score",
    > "column.xpath.ViewCount"="/row/@ViewCount",
    > "column.xpath.Title"="/row",
    > "column.xpath.Body"="/row",
    > "column.xpath.OwnerUserId"="/row/@OwnerUserId",
    > "column.xpath.Tags"="/row/@Tags",
    > "column.xpath.AnswerCount"="/row/@AnswerCount",
    > "column.xpath.CommentCount"="/row/@CommentCount",
    > "column.xpath.FavoriteCount"="/row/@FavoriteCount",
    > "column.xpath.ParentId"="/row/@ParentId"
    > )
    > stored as
    > inputformat 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'
    > outputformat 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'
    > location '/LDZ/data'
    > tblproperties (
    > "xmlinput.start"="<row ",
    > "xmlinput.end"=" />"
    > );
```

| *BDM 1...                  |  |
|----------------------------|--|
| File Edit Format View Help |  |
| 2022F-T1 BDM 1003          |  |
| Group D                    |  |
| Anargha Manoj              |  |
| Harcharan Singh            |  |
| Jovi Fez Bartolata         |  |
| Sharon Victor              |  |
| < >                        |  |
| Windows (CRLF) UTF-8       |  |

```
Windows taskbar showing various icons and system status.
```

```

hdoo@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
> describe formatted posts_ext;
OK
col_name      data_type      comment
# col_name      data_type      comment

creationdatetime    string      from deserializer
lasteditoruserid   string      from deserializer
lasteditordisplayname string      from deserializer
lasteditdatetime   string      from deserializer
lastactivitydatetime string      from deserializer
communityowneddatetime string      from deserializer
contentlicense    string      from deserializer
id               string      from deserializer
posttypeid       string      from deserializer
acceptedanswerid string      from deserializer
score            string      from deserializer
viewcount         string      from deserializer
title            string      from deserializer
body             string      from deserializer
owneruserid      string      from deserializer
tags              string      from deserializer
answercount       string      from deserializer
commentcount     string      from deserializer
favoritecount    string      from deserializer
parentid          string      from deserializer

# Detailed Table Information
Database:           dwz
Owner:              hdoo
CreateTime:         Sat Nov 05 22:01:31 UTC 2022
LastAccessTime:     UNKNOWN

Windows (CRLF)  UTF-8

```

```

+BDM 1... - X
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF)  UTF-8

```

```

hdoo@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
Protect Mode:        None
Retention:           0
Location:            hdfs://ec2-3-145-208-42.us-east-2.compute.amazonaws.com/LDZ/data
Table Type:          EXTERNAL_TABLE
Table Parameters:
  COLUMN_STATS_ACCURATE  false
  EXTERNAL                TRUE
  numFiles                 0
  numRows                  -1
  rawDataSize              -1
  totalSize                 0
  transient_lastDdlTime   1667685691
  xmlinput.end               />
  xmlinput.start              <row

# Storage Information
SerDe Library:      com.ibm.spss.hive.serde2.xml.XmlSerDe
InputFormat:         com.ibm.spss.hive.serde2.xml.XmlInputFormat
OutputFormat:        org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:          No
Num Buckets:        -1
Bucket Columns:     []
Sort Columns:        []
Storage Desc Params:
  column.xpath.AcceptedAnswerId  /row/@AcceptedAnswerId
  column.xpath.AnswerCount        /row/@AnswerCount
  column.xpath.Body               /row
  column.xpath.CommentCount      /row/@CommentCount
  column.xpath.CommunityOwnedDateTime /row/@CommunityOwnedDate
  column.xpath.ContentLicense    /row/@ContentLicense
  column.xpath.CreationDateTime  /row/@CreationDate

Windows (CRLF)  UTF-8

```

```

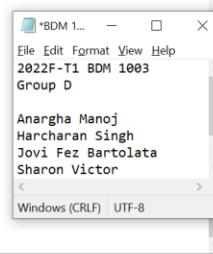
+BDM 1... - X
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF)  UTF-8

```

```

hdooop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
SerDe Library: com.ibm.spss.hive.serde2.xml.XmlSerDe
InputFormat: com.ibm.spss.hive.serde2.xml.XmlInputFormat
OutputFormat: org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
Bucket Columns: []
Sort Columns: []
Storage Desc Params:
  column.xpath.AcceptedAnswerId /row/@AcceptedAnswerId
  column.xpath.AnswerCount /row/@AnswerCount
  column.xpath.Body /row
  column.xpath.CommentCount /row/@CommentCount
  column.xpath.CommunityOwnedDateTime /row/@CommunityOwnedDate
  column.xpath.ContentLicense /row/@ContentLicense
  column.xpath.CreationDateTime /row/@CreationDate
  column.xpath.FavoriteCount /row/@FavoriteCount
  column.xpath.Id /row@Id
  column.xpath.LastActivityDateTime /row/@LastActivityDate
  column.xpath.LastEditDateTime /row/@LastEditDate
  column.xpath.LastEditorDisplayName /row/@LastEditorDisplayName
  column.xpath.LastEditorUserId /row/@LastEditorUserId
  column.xpath.OwnerUserId /row/@OwnerUserId
  column.xpath.ParentId /row/@ParentId
  column.xpath.PostTypeId /row/@PostTypeId
  column.xpath.Score /row/@Score
  column.xpath.Tags /row/@Tags
  column.xpath.Title /row
  column.xpath.ViewCount /row/@ViewCount
  serialization.format 1
Time taken: 1.115 seconds, Fetched: 73 row(s)
hive (dwz)> 

```



File Edit Format View Help  
2022F-T1 BDM 1003  
Group D  
Anargha Manoj  
Harcharan Singh  
Jovi Fez Bartolata  
Sharon Victor  
Windows (CRLF) UTF-8

- Create external table POSTS\_EXT\_DYN with partitioning by Creation Date and Post Type  
For the commands, please refer to Appendix: create\_tables.hql

```

hdooop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hive (dwz)> create external table posts_ext_dyn (
  > CreationDateTime timestamp,
  > LastEditorUserId string,
  > LastEditorDisplayName string,
  > LastEditDateTime timestamp,
  > LastActivityDateTime timestamp,
  > CommunityOwnedDateTime timestamp,
  > ContentLicense string,
  > Id string,
  > PostTypeId string,
  > AcceptedAnswerId string,
  > Score int,
  > ViewCount bigint,
  > Title string,
  > Body string,
  > OwnerUserId string,
  > Tags string,
  > AnswerCount bigint,
  > CommentCount bigint,
  > FavoriteCount bigint,
  > ParentId string
  > )
  > partitioned by (CreationDate date, PostType string)
  > row format delimited
  > stored as textfile
  > ;
OK
Time taken: 0.396 seconds
hive (dwz)> 

```



File Edit Format View Help  
2022F-T1 BDM 1003  
Group D  
Anargha Manoj  
Harcharan Singh  
Jovi Fez Bartolata  
Sharon Victor  
Windows (CRLF) UTF-8

```

hdoo@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
> describe formatted posts_ext_dyn;
OK
col_name      data_type      comment
# col_name      data_type      comment

creationdatetime      timestamp
lasteditoruserid      string
lasteditordisplayname      string
lasteditdatetime      timestamp
lastactivitydatetime      timestamp
communityowneddatetime      timestamp
contentlicense      string
id      string
posttypeid      string
acceptedanswerid      string
score      int
viewcount      bigint
title      string
body      string
owneruserid      string
tags      string
answercount      bigint
commentcount      bigint
favoritecount      bigint
parentid      string

# Partition Information
# col_name      data_type      comment

creationdate      date
posttype      string

Windows (CRLF) UTF-8
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF) UTF-8
6:05 PM 23°C ENG 2022-11-05 6

hdoo@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
# col_name      data_type      comment

creationdate      date
posttype      string

# Detailed Table Information
Database:      dwz
Owner:      hdoo
CreateTime:      Sat Nov 05 22:04:27 UTC 2022
LastAccessTime:      UNKNOWN
Protect Mode:      None
Retention:      0
Location:      hdfs://ec2-3-145-208-42.us-east-2.compute.amazonaws.com/user/hive/warehouse/dwz.db/post
s_ext_dyn
Table Type:      EXTERNAL_TABLE
Table Parameters:
    EXTERNAL      TRUE
    transient_lastDdlTime      1667685867

# Storage Information
SerDe Library:      org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:      org.apache.hadoop.mapred.TextInputFormat
OutputFormat:      org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
Compressed:      No
Num Buckets:      -1
Bucket Columns:      []
Sort Columns:      []
Storage Desc Params:
    serialization.format      1
Time taken: 0.244 seconds, Fetched: 52 row(s)
hive (dwz)> 
Windows (CRLF) UTF-8
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF) UTF-8
6:05 PM 23°C ENG 2022-11-05 6

```

## VI. Performing Daily Update

The daily update is divided into two parts: (1) downloading raw file and (2) loading data into table. This is to ensure that the download is successful before performing updates to the table.

Note: This automation assumes that the IP address is static.

- Step 1: `get_raw_file.sh`

Create the shell script `get_raw_file.sh`, save it in the `/bin` folder, then change file permission to make it executable. When executed, this will download the input file and unzip it. Ensure that there is no error before running `update_table.sh`.

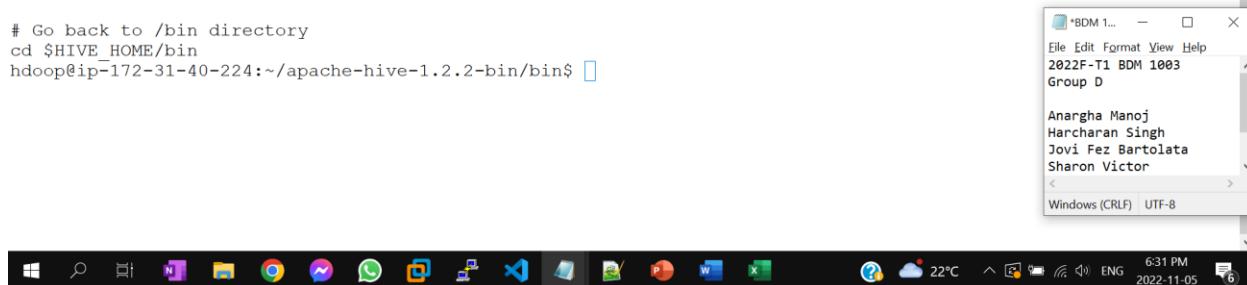
```
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ pwd
/home/hadoop/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi get_raw_file.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 get_raw_file.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll get_raw_file.sh
-rwxrwxrwx 1 hadoop hadoop 280 Nov 5 22:30 get_raw_file.sh*
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cat get_raw_file.sh
# Go to cd /LDZ/data
cd /LDZ/data

# Download input file
sudo wget https://archive.org/download/stackexchange/stackoverflow.com-Posts.7z

# Unzip input file
sudo 7z x stackoverflow.com-Posts.7z

# Display information
ll -h /LDZ/data

# Go back to /bin directory
cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$
```



- Usage and Result of `get_raw_file.sh`
  1. Go to `$HIVE_HOME/bin` or `/home/hadoop/apache-hive-1.2.2-bin/bin`
  2. Run `./get_raw_file.sh`

Note: This step downloads and unzips the complete raw file (92 GB).

```
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ./get_raw_file.sh
--2022-11-05 22:31:25-- https://archive.org/download/stackexchange/stackoverflow.com-Posts.7z
Resolving archive.org (archive.org)... 207.241.224.2
Connecting to archive.org (archive.org) |207.241.224.2|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://ia800107.us.archive.org/27/items/stackexchange/stackoverflow.com-Posts.7z [
--2022-11-05 22:31:26-- https://ia800107.us.archive.org/27/items/stackexchange/stackoverflow.com-Posts.7z
Resolving ia800107.us.archive.org (ia800107.us.archive.org)... 207.241.232.17
Connecting to ia800107.us.archive.org (ia800107.us.archive.org) |207.241.232.17|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 19430749009 (18G) [application/x-7z-compressed]
Saving to: 'stackoverflow.com-Posts.7z'

stackoverflow.com-Posts.7z 100%[=====] 18.10G 2.21MB/s in 99m 47s

2022-11-06 00:11:13 (3.10 MB/s) - 'stackoverflow.com-Posts.7z' saved [19430749009/19430749009]
```

[sudo] password for hadoop:

```
7-Zip [64] 16.02 : Copyright (c) 1999-2016 Igor Pavlov : 2016-05-21
p7zip Version 16.02 (locale=C.UTF-8,Utf16=on,HugeFiles=on,64 bits,4 CPUs Intel(R) Xeon(R) Platinum 8259CL CPU @
2.50GHz (50657),ASM,AES-NI)
```

Scanning the drive for archives:
1 file, 19430749009 bytes (19 GiB)

Extracting archive: stackoverflow.com-Posts.7z

```
-- 
Path = stackoverflow.com-Posts.7z
Type = 7z
```



```
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
```

Scanning the drive for archives:
1 file, 19430749009 bytes (19 GiB)

Extracting archive: stackoverflow.com-Posts.7z

```
-- 
Path = stackoverflow.com-Posts.7z
Type = 7z
Physical Size = 19430749009
Headers Size = 117
Method = BZip2
Solid = -
Blocks = 1
```

Everything is Ok

```
Size: 98441935589
Compressed: 19430749009
./get_raw_file.sh: line 11: ll: command not found
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll /LDZ/data
total 115110072
drwxr-xr-x 2 root root 4096 Nov 6 00:11 .
drwxr-xr-x 3 root root 4096 Nov 5 21:48 ..
-rw-r--r-- 1 root root 98441935589 Sep 30 04:54 Posts.xml
-rw-r--r-- 1 root root 19430749009 Oct 9 17:02 stackoverflow.com-Posts.7z
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll -h /LDZ/data
total 110G
drwxr-xr-x 2 root root 4.0K Nov 6 00:11 .
drwxr-xr-x 3 root root 4.0K Nov 5 21:48 ..
-rw-r--r-- 1 root root 92G Sep 30 04:54 Posts.xml
-rw-r--r-- 1 root root 19G Oct 9 17:02 stackoverflow.com-Posts.7z
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ 
```



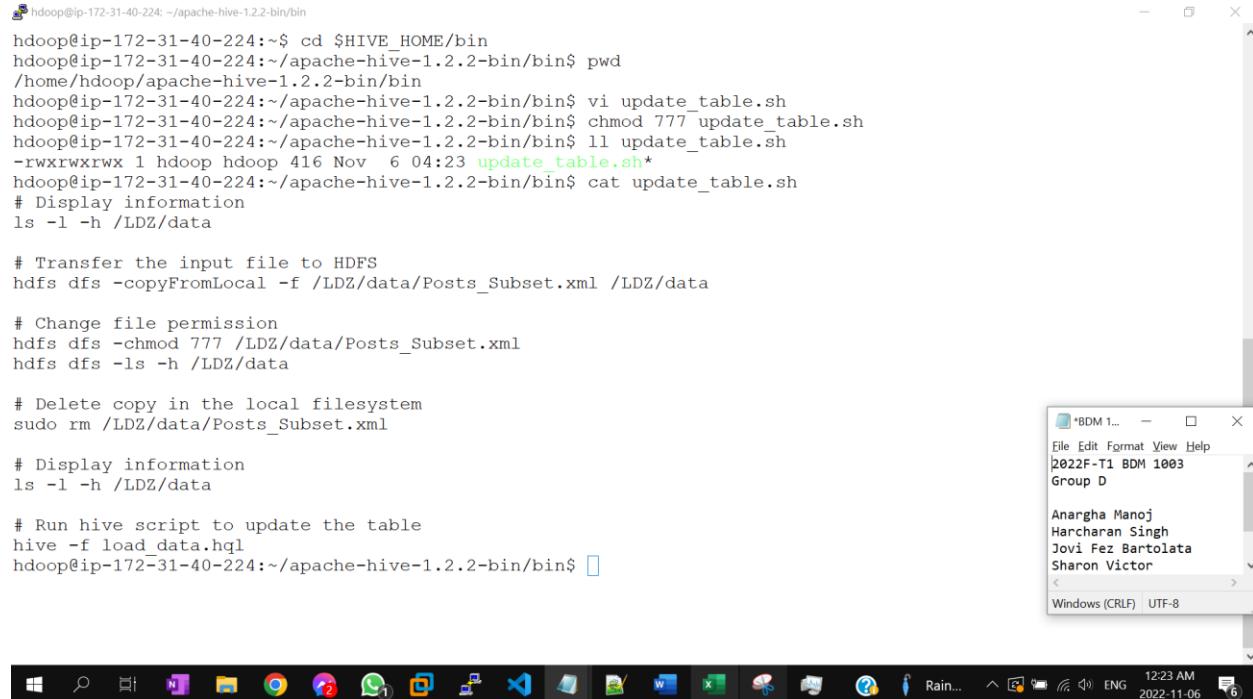
```
*BDM 1...
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF) UTF-8
```

```
*BDM 1...
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF) UTF-8
```



- Step 2: update\_table.sh

Create the shell script *update\_table.sh*, save it in the /bin folder, then change file permission to make it executable. When executed, this will upload the file into HDFS, and will run the Hive script *load\_data.hql* to load the data into the table.



hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin

```

hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ pwd
/home/hadoop/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi update_table.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 update_table.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll update_table.sh
-rwxrwxrwx 1 hadoop hadoop 416 Nov  6 04:23 update_table.sh*
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cat update_table.sh
# Display information
ls -l -h /LDZ/data

# Transfer the input file to HDFS
hdfs dfs -copyFromLocal -f /LDZ/data/Posts_Subset.xml /LDZ/data

# Change file permission
hdfs dfs -chmod 777 /LDZ/data/Posts_Subset.xml
hdfs dfs -ls -h /LDZ/data

# Delete copy in the local filesystem
sudo rm /LDZ/data/Posts_Subset.xml

# Display information
ls -l -h /LDZ/data

# Run hive script to update the table
hive -f load_data.hql
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ 
```

\*BDM 1... File Edit Format View Help  
2022F-T1 BDM 1003 Group D  
Anargha Manoj  
Harcharan Singh  
Jovi Fez Bartolata  
Sharon Victor  
< >  
Windows (CRLF) UTF-8

12:23 AM 2022-11-06 6

- Script: load\_data.hql

Create a Hive script *load\_data.hql*, save it in the /bin folder, then change file permission to make it executable. This is the script to load the data into the table and is executed in the *update\_table* step.

```
hadoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ls
beeline  get_raw_file.sh  hiveserver2  query_top10_answered.sh  update_table.sh
derby.log  hive          metastore_db  query_uanswered.sh
ext        hive-config.sh metatool      schematool
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ rm update_table.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ls /LDZ/data
Posts.xml  stackoverflow.com=Posts.7z
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ sudo cp /LDZ_old/data/Posts_Subset.xml /LDZ/data
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cd
hadoop@ip-172-31-40-224:~$ clear
hadoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ pwd
/home/hadoop/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi load_data.hql
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 load_data.hql
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll load_data.hql
-rwxrwxrwx 1 hadoop hadoop 1482 Nov  6 04:21 load_data.hql*
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cat load_data.hql
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

add jar /home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar;

use dwz;

insert overwrite table posts_ext_dyn partition(CreationDate, PostType)
```

```
hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin$ insert overwrite table posts_ext_dyn partition(CreationDate, PostType)
select
cast(regexp_replace(CreationDateTime, 'T',' ') as timestamp) as CreationDateTime,
LastEditorUserId,
LastEditorDisplayName,
cast(regexp_replace(LastEditDateTime, 'T',' ') as timestamp) as LastEditDateTime,
cast(regexp_replace>LastActivityDateTime, 'T',' ') as timestamp) as LastActivityDateTime,
cast(regexp_replace(CommunityOwnedDateTime, 'T',' ') as timestamp) as CommunityOwnedDateTime,
ContentLicense,
Id,
PostTypeId,
AcceptedAnswerId,
cast(Score as int) as Score,
cast(ViewCount as bigint) as ViewCount,
case when PostTypeId='1' then substr>Title,instr>Title="')+8,instr(substr>Title,instr>Title="')+8,'")-1) else NULL end as Title,
substr(Body,instr(Body,' Body="')+7,instr(substr(Body,instr(Body,' Body="')+7,'')-1) as Body,
OwnerId,
Tags,
cast(AnswerCount as bigint) as AnswerCount,
cast(CommentCount as bigint) as CommentCount,
cast(FavoriteCount as bigint) as FavoriteCount,
ParentId,
to_date(CreationDateTime) as CreationDate,
case when PostTypeId='1' then 'Question' when PostTypeId='2' then 'Answer' end as PostType
from posts_ext
;

describe formatted posts_ext_dyn;
hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin$
```

- Usage and Result of *update\_table.sh*

1. Go to \$HIVE\_HOME/bin or /home/hadoop/apache-hive-1.2.2-bin/bin
2. Run ./update\_table.sh

Note: The input file used in this step is a subset only. It covers Creation Dates from 2008-07-31 to 2009-08-31, and file size is 1.1 GB.

```
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ./update_table.sh
total 111G
-rw-r--r-- 1 root root 92G Sep 30 04:54 Posts.xml
-rw-r--r-- 1 root root 1.1G Nov  6 05:48 Posts_Subset.xml
-rw-r--r-- 1 root root 19G Nov  6 03:55 stackoverflow.com-Posts.7z
Found 1 items
-rwxrwxrwx 1 hdoop supergroup      1.1 G 2022-11-06 05:48 /LDZ/data/Posts_Subset.xml
total 110G
-rw-r--r-- 1 root root 92G Sep 30 04:54 Posts.xml
-rw-r--r-- 1 root root 19G Nov  6 03:55 stackoverflow.com-Posts.7z

Logging initialized using configuration in jar:file:/home/hadoop/apache-hive-1.2.2-bin/lib/hive-common-1.2.2.jar!
!hive-log4j.properties
Added [/home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar] to class path
Added resources: [/home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar]
OK
Time taken: 0.912 seconds
Query ID = hdoop_20221106054903_4a359b10-3d13-4382-a4ca-047ce26ac219
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_1667684710775_0007, Tracking URL = http://ec2-3-145-208-42.us-east-2.compute.amazonaws.com:8088/proxy/application_1667684710775_0007/
Kill Command = /home/hadoop/hadoop-2.10.1/bin/hadoop job -kill job_1667684710775_0007
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 0
2022-11-06 05:49:17,520 Stage-1 map = 0%,  reduce = 0%
2022-11-06 05:50:18,384 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 80.87 sec

```



```
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
Loading partition {creationdate=2008-11-18, posttype=Answer}
Loading partition {creationdate=2009-04-07, posttype=Answer}
Loading partition {creationdate=2009-07-23, posttype=Question}
Loading partition {creationdate=2008-09-28, posttype=Answer}
Loading partition {creationdate=2009-06-20, posttype=Question}
Loading partition {creationdate=2008-12-13, posttype=Question}
Loading partition {creationdate=2009-07-23, posttype=Answer}
Loading partition {creationdate=2009-01-21, posttype=Question}
Loading partition {creationdate=2008-11-07, posttype=Answer}
Loading partition {creationdate=2009-04-18, posttype=Answer}
Loading partition {creationdate=2009-03-27, posttype=Question}
Loading partition {creationdate=2009-02-24, posttype=Question}
Loading partition {creationdate=2008-08-02, posttype=Answer}
    Time taken for adding to write entity : 129
Partition dwz.posts_ext_dyn{creationdate=2008-07-31, posttype=Answer} stats: [numFiles=1, numF
2206, rawDataSize=2204]
Partition dwz.posts_ext_dyn{creationdate=2008-07-31, posttype=Question} stats: [numFiles=1, numRow
e=2856, rawDataSize=2852]
Partition dwz.posts_ext_dyn{creationdate=2008-08-01, posttype=Answer} stats: [numFiles=1, numRows=89, totalSiz
e=77991, rawDataSize=77902]
Partition dwz.posts_ext_dyn{creationdate=2008-08-01, posttype=Question} stats: [numFiles=1, numRows=48, totalsi
ze=62587, rawDataSize=62539]
Partition dwz.posts_ext_dyn{creationdate=2008-08-02, posttype=Answer} stats: [numFiles=1, numRows=68, totalSize
=73205, rawDataSize=73137]
Partition dwz.posts_ext_dyn{creationdate=2008-08-02, posttype=Question} stats: [numFiles=1, numRows=25, totalsi
ze=27410, rawDataSize=27385]
Partition dwz.posts_ext_dyn{creationdate=2008-08-03, posttype=Answer} stats: [numFiles=1, numRows=97, totalSize
=104781, rawDataSize=104684]
Partition dwz.posts_ext_dyn{creationdate=2008-08-03, posttype=Question} stats: [numFiles=1, numRows=33, totalSi
ze=31681, rawDataSize=31648]
Partition dwz.posts_ext_dyn{creationdate=2008-08-04, posttype=Answer} stats: [numFiles=1, numRows=238, totalsiz
e=23800, rawDataSize=23795]
```



```

hdoo@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5 Cumulative CPU: 3641.38 sec HDFS Read: 1172994484 HDFS Write: 1010570135 SUCCESS
Stage-Stage-5: Map: 8 Cumulative CPU: 38.5 sec HDFS Read: 14351683 HDFS Write: 14290331 SUCCESS
Total MapReduce CPU Time Spent: 0 days 1 hours 1 minutes 19 seconds 880 msec
OK
creationdatetime      lasteditoruserid      lasteditordisplayname      lasteditdatetime      lastactivitydat
etime    communityowneddatetime   contentlicense    id    posttypeid    acceptedanswerid    score    viewcou
nt    title    body    owneruserid    tags    answercount    commentcount    favoritecount    parentid    c
reationdate    posttype
Time taken: 3367.152 seconds
OK
col_name      data_type      comment
# col_name      data_type      comment

creationdatetime      timestamp
lasteditoruserid      string
lasteditordisplayname      string
lasteditdatetime      timestamp
lastactivitydatetime      timestamp
communityowneddatetime      timestamp
contentlicense      string
id      string
posttypeid      string
acceptedanswerid      string
score      int
viewcount      bigint
title      string
body      string
owneruserid      string
tags      string
answercount      bigint

Windows (CRLF) UTF-8
18°C 1:23 AM ENG 2022-11-06 6

*+BDM 1... - X
File Edit Format View Help
2022F-T1 BDM 1003
Group D
Anargha Manoj
Harcharan Singh
Jovi Fez Bartolata
Sharon Victor
< >
Windows (CRLF) UTF-8
18°C 1:23 AM ENG 2022-11-06 6

hdoo@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
commentcount      bigint
favoritecount      bigint
parentid      string

# Partition Information
# col_name      data_type      comment

creationdate      date
posttype      string

# Detailed Table Information
Database:      dwz
Owner:      hdoop
CreateTime:      Sat Nov 05 22:04:27 UTC 2022
LastAccessTime:      UNKNOWN
Protect Mode:      None
Retention:      0
Location:      hdfs://ec2-3-145-208-42.us-east-2.compute.amazonaws.com/user/hive/warehouse/dwz.db/post
s_ext_dyn
Table Type:      EXTERNAL_TABLE
Table Parameters:
  EXTERNAL      TRUE
  transient_lastDdlTime      1667685867

# 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:      []

Windows (CRLF) UTF-8
18°C 1:23 AM ENG 2022-11-06 6

```

```

hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
# col_name          data_type            comment
creationdate        date
posttype           string

# Detailed Table Information
Database:          dwz
Owner:             hadoop
CreateTime:        Sat Nov  5 22:04:27 UTC 2022
LastAccessTime:    UNKNOWN
Protect Mode:      None
Retention:         0
Location:          hdfs://ec2-3-145-208-42.us-east-2.compute.amazonaws.com/user/hive/warehouse/dwz.db/post
s_ext_dyn
Table Type:        EXTERNAL_TABLE
Table Parameters:
  EXTERNAL           TRUE
  transient_lastDdlTime 1667685867

# 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:
  serialization.format 1
Time taken: 0.076 seconds, Fetched: 52 row(s)
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ 

```

The screenshot shows a Windows desktop environment. At the top is a taskbar with various icons. Below it is a terminal window showing the command-line output of a Hive query. To the right of the terminal is a Notepad window titled "BDM 1..." containing a list of names: Anargha Manoj, Harcharan Singh, Jovi Fez Bartolata, and Sharon Victor. The Notepad window also shows file menu options like File, Edit, Format, View, Help, and encoding settings (Windows (CRLF) | UTF-8). The system tray at the bottom right shows the date (2022-11-06), time (1:23 AM), battery level (18°C), and network status.

## VII. Performing Queries

- Top 10 Most Answered Stack Overflow Questions
    - Script: *query\_top10\_answered.sh*
- Create the shell script *query\_top10\_answered.sh*, save it in the /bin folder, then change file permission to make it executable.

```

hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi query_top10_answered.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 query_top10_answered.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll query_top10_answered.sh*
-rwxrwxrwx 1 hadoop hadoop 789 Nov  1 06:42 query_top10_answered.sh*
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cat query_top10_answered.sh
echo ""
echo "---- Top 10 Most Answered Stack Overflow Questions ----"
echo "Input CreationDate in the format: yyyy-mm-dd"
read CreationDate

hive -e "
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

select CreationDate, Rank, AnswerCount, Title
from
(
  select a.CreationDate, rank() over (partition by a.CreationDate order by a.AnswerCount desc) as Rank, a.AnswerCount, a.Title
  from
  (
    select CreationDate, Title, sum(AnswerCount) as AnswerCount
    from dwz.posts_ext_dyn
    where PostType='Question'
    group by CreationDate, Title
  ) as a
) as b

```

The screenshot shows a Windows desktop environment. At the top is a taskbar with various icons. Below it is a terminal window showing the contents of the *query\_top10\_answered.sh* script. To the right of the terminal is a Notepad window titled "BDM 1..." containing the same list of names: Anargha Manoj, Harcharan Singh, Jovi Fez Bartolata, and Sharon Victor. The Notepad window also shows file menu options like File, Edit, Format, View, Help, and encoding settings (Windows (CRLF) | UTF-8). The system tray at the bottom right shows the date (2022-11-01), time (2:43 AM), battery level (10°C), and network status.

```

hdoo@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
echo ""
echo "---- Top 10 Most Answered Stack Overflow Questions ----"
echo "Input CreationDate in the format: yyyy-mm-dd"
read CreationDate

hive -e "
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

select CreationDate, Rank, AnswerCount, Title
from
(
  select a.CreationDate, rank() over (partition by a.CreationDate order by a.AnswerCount desc) as Rank, a.AnswerCount, a.Title
  from
  (
    select CreationDate, Title, sum(AnswerCount) as AnswerCount
    from dwz.posts_ext_dyn
    where PostType='Question'
    group by CreationDate, Title
  ) as a
) as b
where b.Rank <= 10 and b.CreationDate = '$CreationDate'
;
"

"query_top10_answered.sh" 27L, 789C

```

27,1 All 244 AM 10°C ENG 2022-11-01

- Usage and Result of *query\_top10\_answered.sh*

1. Go to \$HIVE\_HOME/bin or /home/hdoop/apache-hive-1.2.2-bin/bin
2. Run ./query\_top10\_answered.sh
3. Input Creation Date (yyyy-mm-dd)
4. Press Enter

```

hdoo@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ./query_top10_answered.sh

---- Top 10 Most Answered Stack Overflow Questions ----
Input CreationDate in the format: yyyy-mm-dd
2009-07-22

Logging initialized using configuration in jar:file:/home/hdoop/apache-hive-1.2.2-bin/lib/hive-common-1.2.2.jar!
!/hive-log4j.properties
Query ID = hdoop_20221106053820_34fce988-e57d-4c19-8eb0-4f27d3f1b837
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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_1667684710775_0004, Tracking URL = http://ec2-3-145-208-42.us-east-2.compute.amazonaws.com:8080/proxy/application_1667684710775_0004/
Kill Command = /home/hdoop/hadoop-2.10.1/bin/hadoop job -kill job_1667684710775_0004
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2022-11-06 05:38:41,056 Stage-1 map = 0%, reduce = 0%
2022-11-06 05:38:55,917 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 17.22 sec
2022-11-06 05:39:04,404 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 24.6 sec
2022-11-06 05:39:10,869 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 31.0 sec
2022-11-06 05:39:18,406 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 38.48 sec
MapReduce Total cumulative CPU time: 38 seconds 480 msec
Ended Job = job_1667684710775_0004
Launching Job 2 out of 2

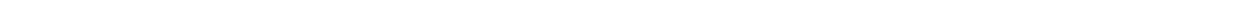
```

141 AM 17°C ENG 2022-11-06

```

hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
Number of reduce tasks not specified. Estimated from input data size: 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_1667684710775_0005, Tracking URL = http://ec2-3-145-208-42.us-east-2.compute.amazonaws.com:8080/proxy/application_1667684710775_0005/
Kill Command = /home/hadoop/hadoop-2.10.1/bin/hadoop job -kill job_1667684710775_0005
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-11-06 05:39:33,637 Stage-2 map = 0%, reduce = 0%
2022-11-06 05:39:43,081 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 8.63 sec
2022-11-06 05:39:52,506 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 15.33 sec
MapReduce Total cumulative CPU time: 15 seconds 330 msec
Ended Job = job_1667684710775_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 38.48 sec HDFS Read: 322296965 HDFS Write: 0 SUCCESS
S
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 15.33 sec HDFS Read: 18700194 HDFS Write: 701 SUCCESS
Total MapReduce CPU Time Spent: 53 seconds 810 msec
OK
creationdate      rank      answercount      title
2009-07-22        1          31      Is there a pretty print for PHP?
2009-07-22        2          26      How can I add a key/value pair to a JavaScript object?
2009-07-22        3          24      How to fix &quot;Incorrect string value&quot; errors?
2009-07-22        4          23      How to determine if a list of polygon points are in clockwise order?
2009-07-22        5          22      Why do this() and super() have to be the first statement in a constructor?
2009-07-22        6          21      Calculate difference in keys contained in two Python dictionaries
2009-07-22        7          18      VBA editor auto-deletes spaces at the ends of lines
2009-07-22        7          18      What security issues should I look out for in PHP?

```



```

hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1667684710775_0005, Tracking URL = http://ec2-3-145-208-42.us-east-2.compute.amazonaws.com:8080/proxy/application_1667684710775_0005/
Kill Command = /home/hadoop/hadoop-2.10.1/bin/hadoop job -kill job_1667684710775_0005
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-11-06 05:39:33,637 Stage-2 map = 0%, reduce = 0%
2022-11-06 05:39:43,081 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 8.63 sec
2022-11-06 05:39:52,506 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 15.33 sec
MapReduce Total cumulative CPU time: 15 seconds 330 msec
Ended Job = job_1667684710775_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 38.48 sec HDFS Read: 322296965 HDFS Write: 18691865 SUCCESS
S
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 15.33 sec HDFS Read: 18700194 HDFS Write: 701 SUCCESS
Total MapReduce CPU Time Spent: 53 seconds 810 msec
OK
creationdate      rank      answercount      title
2009-07-22        1          31      Is there a pretty print for PHP?
2009-07-22        2          26      How can I add a key/value pair to a JavaScript object?
2009-07-22        3          24      How to fix &quot;Incorrect string value&quot; errors?
2009-07-22        4          23      How to determine if a list of polygon points are in clockwise order?
2009-07-22        5          22      Why do this() and super() have to be the first statement in a constructor?
2009-07-22        6          21      Calculate difference in keys contained in two Python dictionaries
2009-07-22        7          18      VBA editor auto-deletes spaces at the ends of lines
2009-07-22        7          18      What security issues should I look out for in PHP?
2009-07-22        7          18      Using the distinct function in SQL
2009-07-22       10         17      Detecting design mode from a Control's constructor
Time taken: 92.706 seconds, Fetched: 10 rows(s)
hadoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin$ 

```

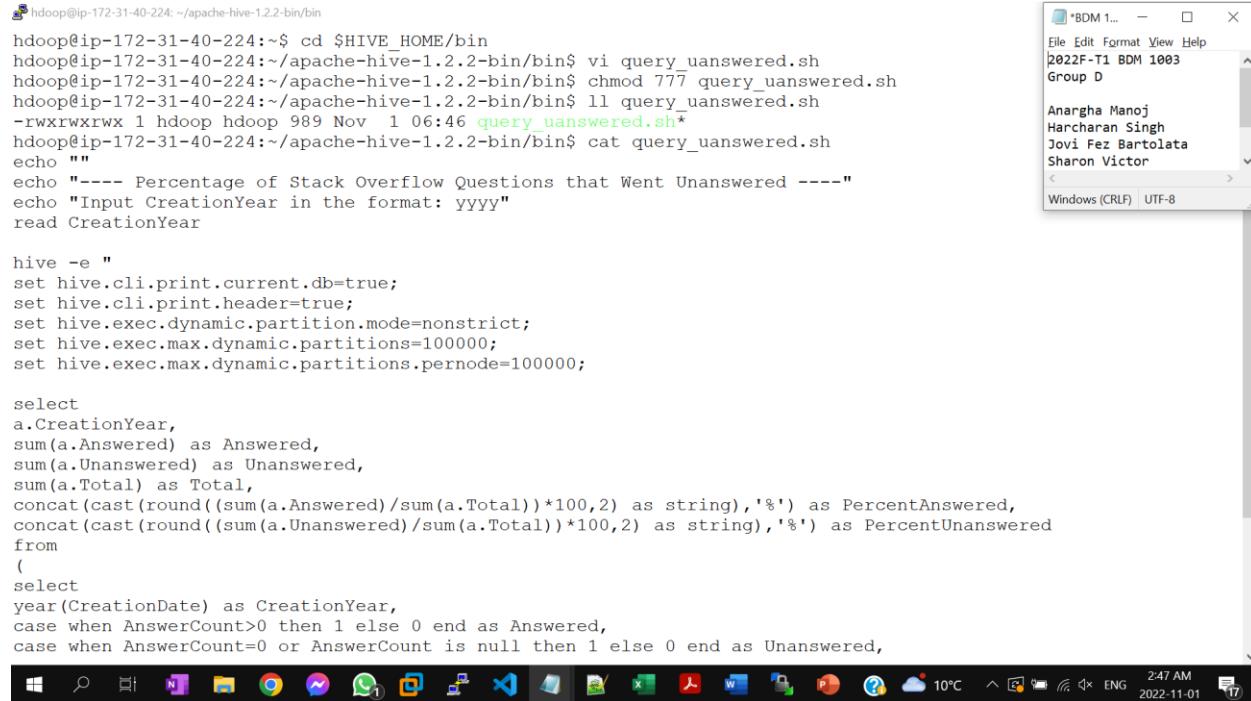
Answer: Among the questions created on July 22, 2009, “Is there a pretty print for PHP?” has the highest number of answers (31). The rest of the top 10 most answered questions are also listed in the result.

- Percentage of Stack Overflow Questions that Went Unanswered
  - Script: `query_uanswered.sh`  
Create the shell script `query_uanswered.sh`, save it in the `/bin` folder, then change file permission to make it executable.

```
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi query_uanswered.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 query_uanswered.sh
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll query_uanswered.sh
-rwxrwxrwx 1 hadoop hadoop 989 Nov  1 06:46 query_uanswered.sh*
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cat query_uanswered.sh
echo ""
echo "---- Percentage of Stack Overflow Questions that Went Unanswered ----"
echo "Input CreationYear in the format: yyyy"
read CreationYear

hive -e "
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

select
a.CreationYear,
sum(a.Answered) as Answered,
sum(a.Unanswered) as Unanswered,
sum(a.Total) as Total,
concat(cast(round((sum(a.Answered)/sum(a.Total))*100,2) as string),'%') as PercentAnswered,
concat(cast(round((sum(a.Unanswered)/sum(a.Total))*100,2) as string),'%') as PercentUnanswered
from
(
select
year(CreationDate) as CreationYear,
case when AnswerCount>0 then 1 else 0 end as Answered,
case when AnswerCount=0 or AnswerCount is null then 1 else 0 end as Unanswered,
```

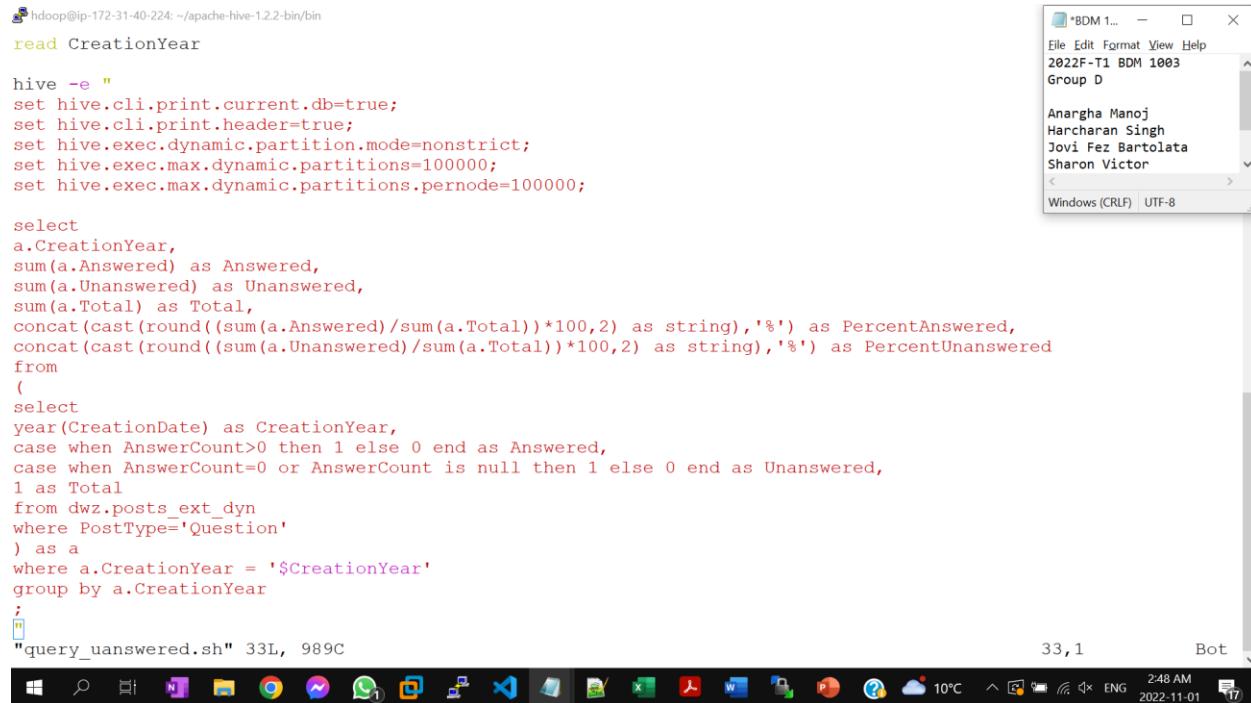


```
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
read CreationYear

hive -e "
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

select
a.CreationYear,
sum(a.Answered) as Answered,
sum(a.Unanswered) as Unanswered,
sum(a.Total) as Total,
concat(cast(round((sum(a.Answered)/sum(a.Total))*100,2) as string),'%') as PercentAnswered,
concat(cast(round((sum(a.Unanswered)/sum(a.Total))*100,2) as string),'%') as PercentUnanswered
from
(
select
year(CreationDate) as CreationYear,
case when AnswerCount>0 then 1 else 0 end as Answered,
case when AnswerCount=0 or AnswerCount is null then 1 else 0 end as Unanswered,
1 as Total
from dwz.posts_ext_dyn
where PostType='Question'
) as a
where a.CreationYear = '$CreationYear'
group by a.CreationYear
;
"

"query_uanswered.sh" 33L, 989C
```



o Usage and Result of *query\_uanswered.sh*

1. Go to \$HIVE\_HOME/bin or /home/hadoop/apache-hive-1.2.2-bin/bin
2. Run ./query\_uanswered.sh
3. Input Creation Year (yyyy)
4. Press Enter

```
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
hadoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ./query_uanswered.sh
---- Percentage of Stack Overflow Questions that Went Unanswered ----
Input CreationYear in the format: yyyy
2008

Logging initialized using configuration in jar:file:/home/hadoop/apache-hive-1.2.2-bin/lib/hive-common-1.2.2.jar!
!/hive-log4j.properties
Query ID = hadoop_20221106054415_72e00ceb-9f9c-48d3-b0af-1b5e8842785e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 2
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_1667684710775_0006, Tracking URL = http://ec2-3-145-208-42.us-east-2.compute.amazonaws.com:8088/proxy/application_1667684710775_0006/
Kill Command = /home/hadoop/hadoop-2.10.1/bin/hadoop job -kill job_1667684710775_0006
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2022-11-06 05:44:35,381 Stage-1 map = 0%, reduce = 0%
2022-11-06 05:44:48,150 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 13.68 sec
2022-11-06 05:44:57,778 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 22.19 sec
2022-11-06 05:45:04,118 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 27.9 sec
2022-11-06 05:45:11,636 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 31.45 sec
MapReduce Total cumulative CPU time: 31 seconds 450 msec
Ended Job = job_1667684710775_0006
MapReduce Jobs Launched:
```

```
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin
Logging initialized using configuration in jar:file:/home/hadoop/apache-hive-1.2.2-bin/lib/hive-common-1.2.2.jar!
!/hive-log4j.properties
Query ID = hadoop_20221106054415_72e00ceb-9f9c-48d3-b0af-1b5e8842785e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 2
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_1667684710775_0006, Tracking URL = http://ec2-3-145-208-42.us-east-2.compute.amazonaws.com:8088/proxy/application_1667684710775_0006/
Kill Command = /home/hadoop/hadoop-2.10.1/bin/hadoop job -kill job_1667684710775_0006
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2022-11-06 05:44:35,381 Stage-1 map = 0%, reduce = 0%
2022-11-06 05:44:48,150 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 13.68 sec
2022-11-06 05:44:57,778 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 22.19 sec
2022-11-06 05:45:04,118 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 27.9 sec
2022-11-06 05:45:11,636 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 31.45 sec
MapReduce Total cumulative CPU time: 31 seconds 450 msec
Ended Job = job_1667684710775_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 31.45 sec HDFS Read: 322305583 HDFS Write: 33 SUCCESS
Total MapReduce CPU Time Spent: 31 seconds 450 msec
OK
a.creationyear    answered      unanswered      total      percentanswered percentunanswered
2008      57495      74      57569      99.87%     0.13%
Time taken: 58.572 seconds, Fetched: 1 row(s)
hadoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$
```

Answer: 0.13% or 74 of the 57,569 questions created in 2008 were unanswered.

## VIII. Appendix: Scripts and Commands

- `create_tables.hql`

```
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

add jar /home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar;

create database dwz;

describe database dwz;

use dwz;

create external table posts_ext (
CreationDateTime string,
LastEditorUserId string,
LastEditorDisplayName string,
LastEditDateTime string,
LastActivityDateTime string,
CommunityOwnedDateTime string,
ContentLicense string,
Id string,
PostTypeId string,
AcceptedAnswerId string,
Score string,
ViewCount string,
Title string,
Body string,
OwnerUserId string,
Tags string,
AnswerCount string,
CommentCount string,
FavoriteCount string,
ParentId string
)
row format serde 'com.ibm.spss.hive.serde2.xml.XmlSerDe' with
serdeproperties (
"column.xpath.CreationDateTime"="/row/@CreationDate",
"column.xpath.LastEditorUserId"="/row/@LastEditorUserId",
"column.xpath.LastEditorDisplayName"="/row/@LastEditorDisplayName",
"column.xpath.LastEditDateTime"="/row/@LastEditDate",
"column.xpath.LastActivityDateTime"="/row/@LastActivityDate",
"column.xpath.CommunityOwnedDateTime"="/row/@CommunityOwnedDate",
"column.xpath.ContentLicense"="/row/@ContentLicense",
"column.xpath.Id"="/row/@Id",
"column.xpath.PostTypeId"="/row/@PostTypeId",
"column.xpath.AcceptedAnswerId"="/row/@AcceptedAnswerId",
```

```

"column.xpath.Score"="/row/@Score",
"column.xpath.ViewCount"="/row/@ViewCount",
"column.xpath.Title"="/row",
"column.xpath.Body"="/row",
"column.xpath.OwnerUserId"="/row/@OwnerUserId",
"column.xpath.Tags"="/row/@Tags",
"column.xpath.AnswerCount"="/row/@AnswerCount",
"column.xpath.CommentCount"="/row/@CommentCount",
"column.xpath.FavoriteCount"="/row/@FavoriteCount",
"column.xpath.ParentId"="/row/@ParentId"
)
stored as
inputformat 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'
outputformat 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'
location '/LDZ/data'
tblproperties (
"xmlinput.start"="<row ",
"xmlinput.end"=" />"
);

describe formatted posts_ext;

create external table posts_ext_dyn (
CreationDateTime timestamp,
LastEditorUserId string,
LastEditorDisplayName string,
LastEditDateTime timestamp,
LastActivityDateTime timestamp,
CommunityOwnedDateTime timestamp,
ContentLicense string,
Id string,
PostTypeId string,
AcceptedAnswerId string,
Score int,
ViewCount bigint,
Title string,
Body string,
OwnerUserId string,
Tags string,
AnswerCount bigint,
CommentCount bigint,
FavoriteCount bigint,
ParentId string
)
partitioned by (CreationDate date, PostType string)
row format delimited
stored as textfile
;

describe formatted posts_ext_dyn;

```

- `get_raw_file.sh`

```
# Go to cd /LDZ/data
cd /LDZ/data

# Download input file
sudo wget https://archive.org/download/stackexchange/stackoverflow.com-
Posts.7z

# Unzip input file
sudo 7z x stackoverflow.com-Posts.7z

# Display information
ls -l -h /LDZ/data

# Go back to /bin directory
cd $HIVE_HOME/bin
```

- `update_table.sh`

```
# Display information
ls -l -h /LDZ/data

# Transfer the input file to HDFS
hdfs dfs -copyFromLocal -f /LDZ/data/Posts_Subset.xml /LDZ/data

# Change file permission
hdfs dfs -chmod 777 /LDZ/data/Posts_Subset.xml
hdfs dfs -ls -h /LDZ/data

# Delete copy in the local filesystem
sudo rm /LDZ/data/Posts_Subset.xml

# Display information
ls -l -h /LDZ/data

# Run hive script to update the table
hive -f load_data.hql
```

- `load_data.hql`

```
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

add jar /home/hadoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar;

use dwz;

insert overwrite table posts_ext_dyn partition(CreationDate, PostType)
select
```

```

cast(regexp_replace(CreationDateTime, 'T',' ') as timestamp) as
CreationDateTime,
LastEditorUserId,
LastEditorDisplayName,
cast(regexp_replace(LastEditDateTime, 'T',' ') as timestamp) as
LastEditDateTime,
cast(regexp_replace(LastActivityDateTime, 'T',' ') as timestamp) as
LastActivityDateTime,
cast(regexp_replace(CommunityOwnedDateTime, 'T',' ') as timestamp) as
CommunityOwnedDateTime,
ContentLicense,
Id,
PostTypeId,
AcceptedAnswerId,
cast(Score as int) as Score,
cast(ViewCount as bigint) as ViewCount,
case when PostTypeId='1' then substr>Title,instr>Title,'
Title='')'+8,instr(substr>Title,instr>Title,' Title='')'+8),'" ')-1) else
NULL end as Title,
substr(Body,instr(Body,' Body='')'+7,instr(substr(Body,instr(Body,'
Body='')'+7),'" ')-1) as Body,
OwnerUserId,
Tags,
cast(AnswerCount as bigint) as AnswerCount,
cast(CommentCount as bigint) as CommentCount,
cast(FavoriteCount as bigint) as FavoriteCount,
ParentId,
to_date(CreationDateTime) as CreationDate,
case when PostTypeId='1' then 'Question' when PostTypeId='2' then
'Answer' end as PostType
from posts_ext
;

describe formatted posts_ext_dyn;

```

- query\_top10\_answered.sh

```

echo ""
echo "---- Top 10 Most Answered Stack Overflow Questions ----"
echo "Input CreationDate in the format: yyyy-mm-dd"
read CreationDate

hive -e "
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

select CreationDate, Rank, AnswerCount, Title
from
(

```

```

    select a.CreationDate, rank() over (partition by a.CreationDate order
    by a.AnswerCount desc) as Rank, a.AnswerCount, a.Title
    from
    (
        select CreationDate, Title, sum(AnswerCount) as AnswerCount
        from dwz.posts_ext_dyn
        where PostType='Question'
        group by CreationDate, Title
    ) as a
) as b
where b.Rank <= 10 and b.CreationDate = '$CreationDate'
;
"
```

- query\_uanswered.sh

```

echo ""
echo "---- Percentage of Stack Overflow Questions that Went Unanswered --"
echo "Input CreationYear in the format: yyyy"
read CreationYear

hive -e "
set hive.cli.print.current.db=true;
set hive.cli.print.header=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.exec.max.dynamic.partitions=100000;
set hive.exec.max.dynamic.partitions.pernode=100000;

select
a.CreationYear,
sum(a.Answered) as Answered,
sum(a.Unanswered) as Unanswered,
sum(a.Total) as Total,
concat(cast(round((sum(a.Answered)/sum(a.Total))*100,2) as string),'%')
as PercentAnswered,
concat(cast(round((sum(a.Unanswered)/sum(a.Total))*100,2) as string),'%')
as PercentUnanswered
from
(
select
year(CreationDate) as CreationYear,
case when AnswerCount>0 then 1 else 0 end as Answered,
case when AnswerCount=0 or AnswerCount is null then 1 else 0 end as
Unanswered,
1 as Total
from dwz.posts_ext_dyn
where PostType='Question'
) as a
where a.CreationYear = '$CreationYear'
group by a.CreationYear
;
"
```