Hadoop Ecosystem and Hive

2022F-T1 BDM 1003

Table of Contents

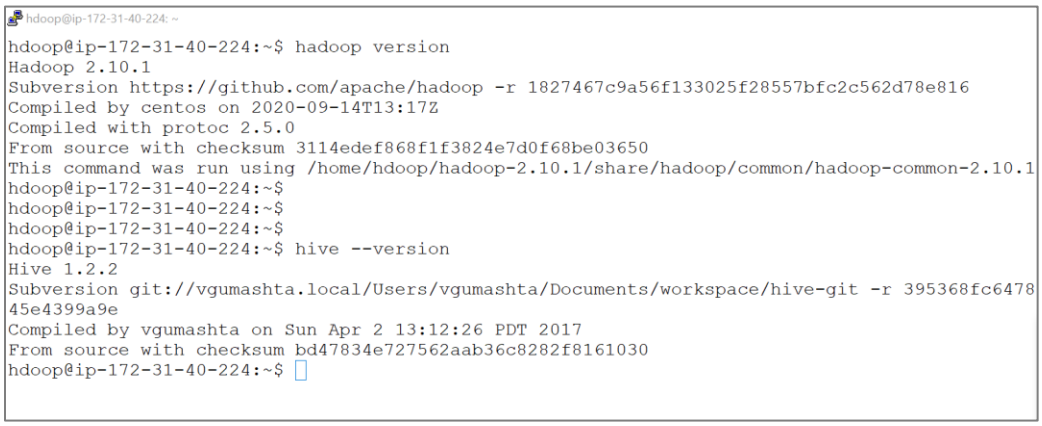
I. Objectives

This project aims to:

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

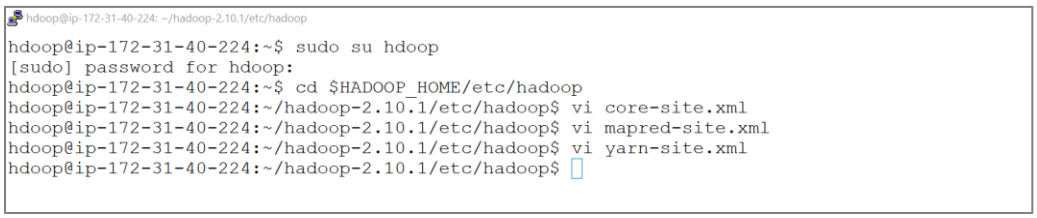
II. Version Requirements

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.



III. Launch Hadoop

- Start instance and connect using PuTTY.



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

```
hdoop@ip-172-31-40-224: ~/hadoop-2.10.1/etc/hadoop
<?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/hdoop/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 --
```

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

```
hdoop@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 --
```

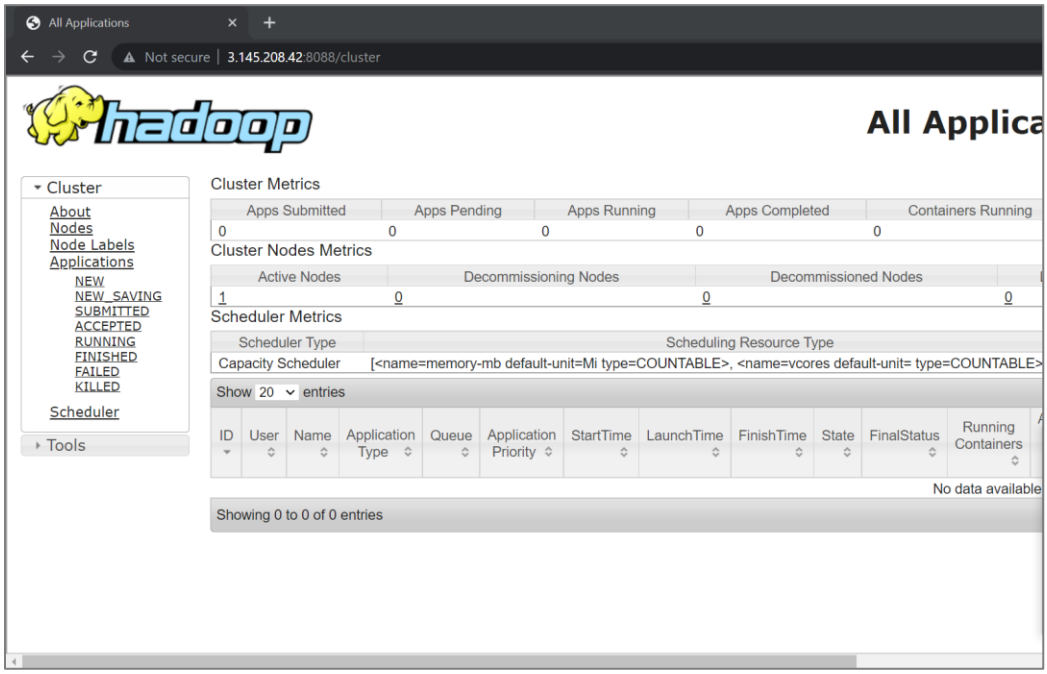
- Update mapred-site.xml
  mapreduce.jobtracker.address = <new IP address>
  mapreduce.map.memory.mb = 4096
  mapreduce.reduce.memory.mb = 4096

```
hdoop@ip-172-31-40-224: ~/hadoop-2.10.1/etc/hadoop
<configuration>
        <property>
                <name>mapreduce.jobtracker.address</name>
                <value>ec2-3-145-208-42.us-east-2.compute.amazonaws.com</value>
        </property>
        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>


        <property>
            <name>mapreduce.map.memory.mb</name>
            <value>4096</value>
        </property>
        <property>
            <name>mapreduce.reduce.memory.mb</name>
            <value>4096</value>
        </property>
</configuration>
~
~
~
~
~
~
~
~
~
~
-- INSERT --
```

- Start DFS, Start YARN, Run JPS

```
hdoop@ip-172-31-40-224: ~/hadoop-2.10.1/sbin
hdoop@ip-172-31-40-224:~$ cd $HADOOP_HOME/sbin
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$
hdoop@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/hdoop/h
hadoop-hdoop-namenode-ip-172-31-40-224.out
localhost: starting datanode, logging to /home/hdoop/hadoop-2.10.1/logs/hadoop-hdoop-datanode
out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/hdoop/hadoop-2.10.1/logs/hadoop-hdoop-s
p-172-31-40-224.out
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$ ./start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/hdoop/hadoop-2.10.1/logs/yarn-hdoop-resourcemanage
.out
localhost: starting nodemanager, logging to /home/hdoop/hadoop-2.10.1/logs/yarn-hdoop-nodeman
224.out
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$ jps
4913 SecondaryNameNode
4466 NameNode
5255 NodeManager
5355 Jps
4636 DataNode
5084 ResourceManager
hdoop@ip-172-31-40-224:~/hadoop-2.10.1/sbin$
```

- Access using browser



IV. Download Dependencies

- Install 7zip

```
hdoop@ip-172-31-40-224: ~
hdoop@ip-172-31-40-224:~$ sudo apt install p7zip-full p7zip-rar
[sudo] password for hdoop:
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+
Get:2 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/universe amd64 p7zip-full amd64 1
kB]
Get:3 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/multiverse amd64 p7zip-rar amd64
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) ...
```

```
hdoop@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+
Get:2 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/universe amd64 p7zip-full amd64 1
kB]
Get:3 http://us-east-2.ec2.archive.ubuntu.com/ubuntu bionic/multiverse amd64 p7zip-rar amd64
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) ...
hdoop@ip-172-31-40-224:~$
```

- Download hivexmlserde-1.0.5.3.jar

  Save the jar file in the /lib folder

```
hdoop@ip-172-31-40-224: ~/apache-hive-3.1.2-bin/lib
hdoop@ip-172-31-40-224:~$ cd /home/hdoop/apache-hive-3.1.2-bin/lib
hdoop@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/serd
de/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/hiv
.jar [following]
--2022-10-31 00:29:33--  https://repo1.maven.org/maven2/com/ibm/spss/hive/serde2/xml/hivexmlse
mlserde-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.(
de-1.0.5.3.jar'

remotecontent?filepath=com% 100%[=========================================>]  49.47K  --.-KB/

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

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

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

Change permission of jar file

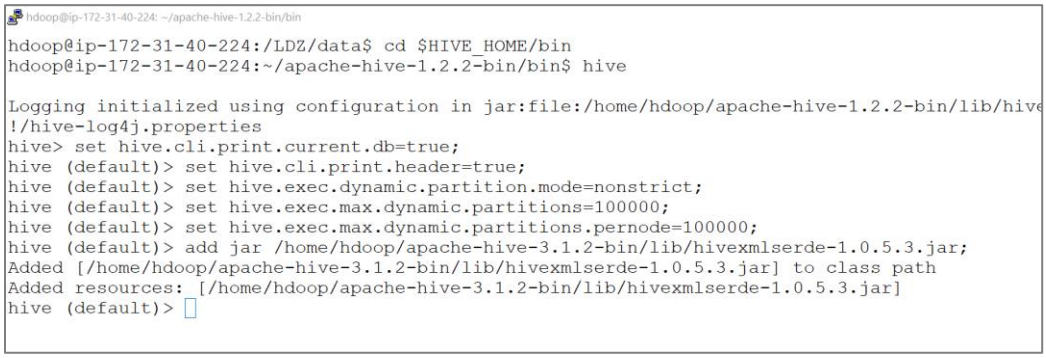
```
hdoop@ip-172-31-40-224: ~/apache-hive-3.1.2-bin/lib
hdoop@ip-172-31-40-224:~$ chmod 777 /home/hdoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3
hdoop@ip-172-31-40-224:~$ ll /home/hdoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar
-rwxrwxrwx 1 hdoop hdoop 50656 Jun 22  2015 /home/hdoop/apache-hive-3.1.2-bin/lib/hivexmlserde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

V. Initial Set-up

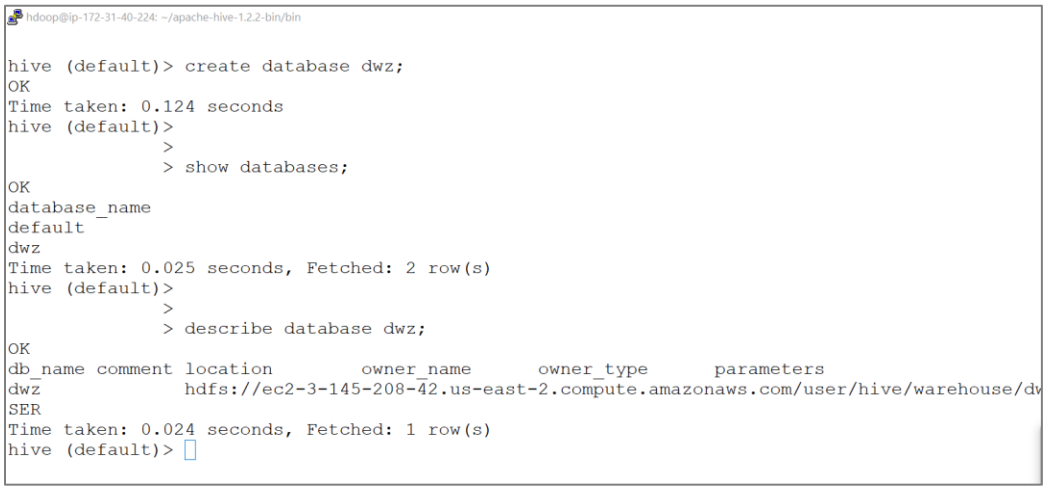
- Create /LDZ and /LDZ/data directories

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

- Launch Hive, set properties, and add jar file



- Create /DWZ database



- Create external table POSTS_EXT with SerDe
  For the commands, please refer to Appendix: create_tables.hql

```
hdoop@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",
```

```
hdoop@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"=" />"
          > );
OK
Time taken: 1.249 seconds
hive (dwz)>
          >
          > describe formatted posts_ext;
OK
col_name        data_type       comment
```

```
hdoop@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:                  hdoop
CreateTime:             Sat Nov 05 22:01:31 UTC 2022
LastAccessTime:         UNKNOWN
```

```
hdoop@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
```

```
hdoop@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.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
        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)>
```

- Create external table POSTS_EXT_DYN with partitioning by Creation Date and Post Type
  For the commands, please refer to Appendix: create_tables.hql

```
hdoop@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)>
          >
          >
```

```
hdoop@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
```

```
hdoop@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:                  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/war
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: 0.244 seconds, Fetched: 52 row(s)
hive (dwz)>
```

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*.

```
hdoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ pwd
/home/hdoop/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi get_raw_file.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 get_raw_file.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll get_raw_file.sh
-rwxrwxrwx 1 hdoop hdoop 280 Nov  5 22:30 get_raw_file.sh*
hdoop@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 $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$
```

  - Usage and Result of *get_raw_file.sh*
    1. Go to $HIVE_HOME/bin or /home/hdoop/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).

```
hdoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ./get_raw_file.sh
--2022-11-05 22:31:25--  https://archive.org/download/stackexchange/stackoverflow.com-Posts.7
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
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... connec
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

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

[sudo] password for hdoop:

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) Pla
 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
```

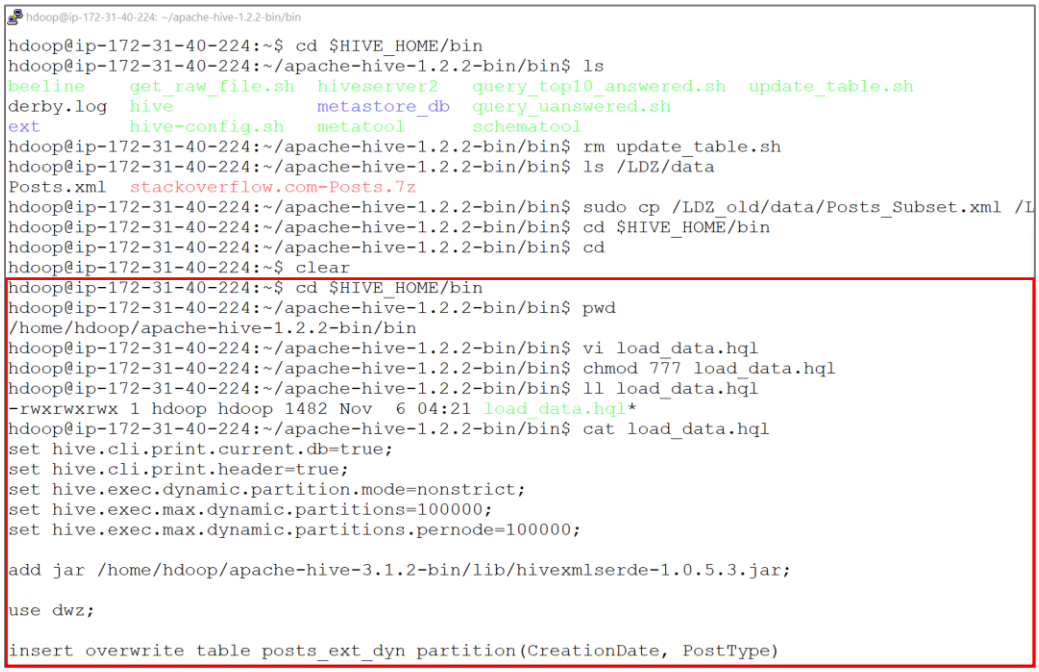
```
hdoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
Scanning the drive for archives:
1 file, 19430749009 bytes (19 GiB)

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

- Step 2: update_table.sh

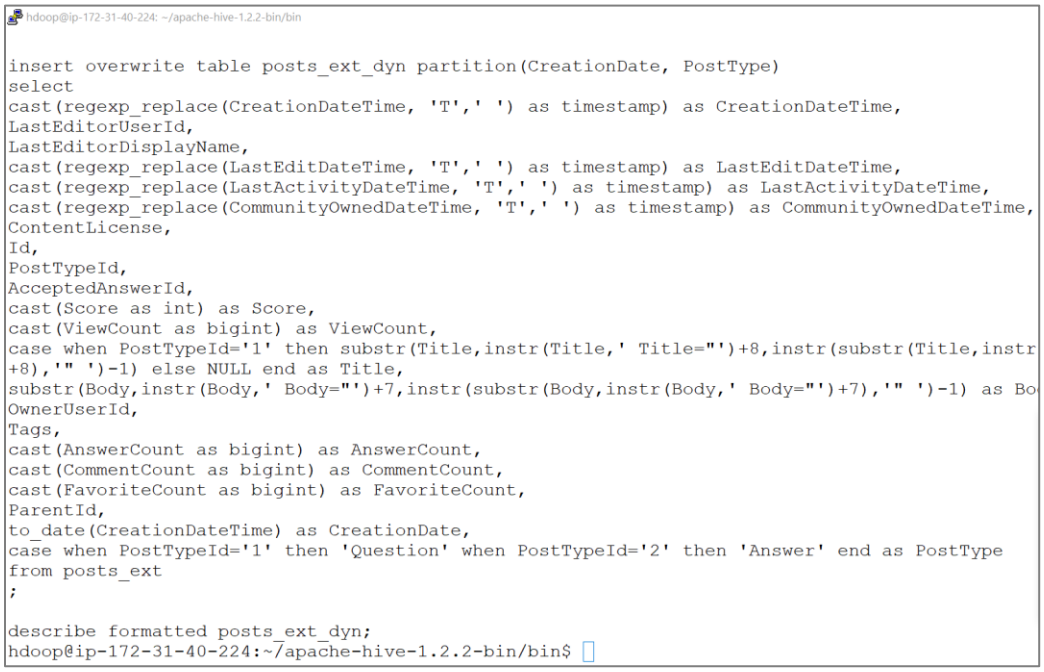
  Create the shell script *update_table.sh*, save it in the /bin folder, then change file permission to make it executable. When executed, this will upload the file into HDFS, and will run the Hive script *load_data.hql* to load the data into the table.

```
hdoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ pwd
/home/hdoop/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi update_table.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 update_table.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll update_table.sh
-rwxrwxrwx 1 hdoop hdoop 416 Nov  6 04:23 update_table.sh*
hdoop@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
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$
```

- 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.



```
hdoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ 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
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ rm update_table.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ls /LDZ/data
Posts.xml  stackoverflow.com-Posts.7z
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ sudo cp /LDZ_old/data/Posts_Subset.xml /L
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ cd
hdoop@ip-172-31-40-224:~$ clear
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ pwd
/home/hdoop/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi load_data.hql
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 load_data.hql
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll load_data.hql
-rwxrwxrwx 1 hdoop hdoop 1482 Nov  6 04:21 load_data.hql*
hdoop@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/hdoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar;

use dwz;

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



```
hdoop@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,' Title="')+8,instr(substr(Title,instr
+8),'" ')-1) else NULL end as Title,
substr(Body,instr(Body,' Body="')+7,instr(substr(Body,instr(Body,' Body="')+7),'" ')-1) as Bo
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;
hdoop@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/hdoop/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:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ./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/hdoop/apache-hive-1.2.2-bin/lib/hiv
!/hive-log4j.properties
Added [/home/hdoop/apache-hive-3.1.2-bin/lib/hivexmlserde-1.0.5.3.jar] to class path
Added resources: [/home/hdoop/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.compu
088/proxy/application_1667684710775_0007/
Kill Command = /home/hdoop/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, num
2206, rawDataSize=2204]
Partition dwz.posts_ext_dyn{creationdate=2008-07-31, posttype=Question} stats: [numFiles=1, n
e=2856, rawDataSize=2852]
Partition dwz.posts_ext_dyn{creationdate=2008-08-01, posttype=Answer} stats: [numFiles=1, num
=77991, rawDataSize=77902]
Partition dwz.posts_ext_dyn{creationdate=2008-08-01, posttype=Question} stats: [numFiles=1, n
ze=62587, rawDataSize=62539]
Partition dwz.posts_ext_dyn{creationdate=2008-08-02, posttype=Answer} stats: [numFiles=1, num
=73205, rawDataSize=73137]
Partition dwz.posts_ext_dyn{creationdate=2008-08-02, posttype=Question} stats: [numFiles=1, n
ze=27410, rawDataSize=27385]
Partition dwz.posts_ext_dyn{creationdate=2008-08-03, posttype=Answer} stats: [numFiles=1, num
=104781, rawDataSize=104684]
Partition dwz.posts_ext_dyn{creationdate=2008-08-03, posttype=Question} stats: [numFiles=1, n
ze=31681, rawDataSize=31648]
Partition dwz.posts_ext_dyn{creationdate=2008-08-04, posttype=Answer} stats: [numFiles=1, num
```

```
hdoop@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: 10105
Stage-Stage-5: Map: 8   Cumulative CPU: 38.5 sec   HDFS Read: 14351683 HDFS Write: 14290331 S
Total MapReduce CPU Time Spent: 0 days 1 hours 1 minutes 19 seconds 880 msec
OK
creationdatetime        lasteditoruserid        lasteditordisplayname   lasteditdatetime
etime   communityowneddatetime  contentlicense  id      posttypeid      acceptedanswerid
nt      title   body    owneruserid     tags    answercount     commentcount    favoritecount
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
```

```
hdoop@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/war
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:         []
```

```
hdoop@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:                  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/war
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)
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$
```

VII. Performing Queries

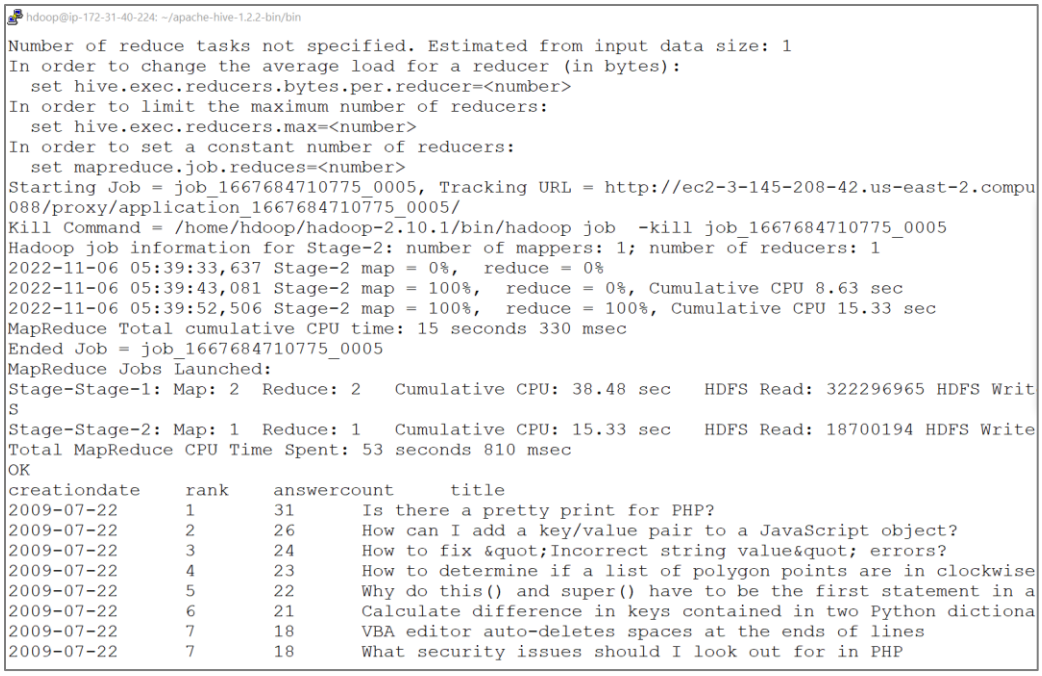
- Top 10 Most Answered Stack Overflow Questions
  - Script: query_top10_answered.sh
    Create the shell script *query_top10_answered.sh*, save it in the /bin folder, then change file permission to make it executable.

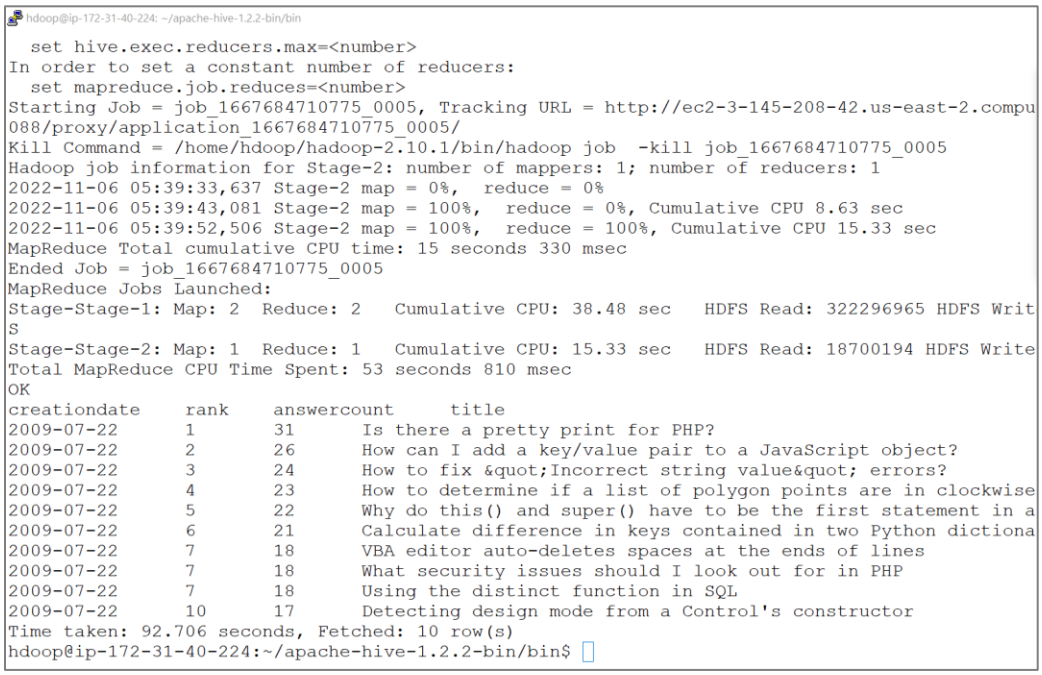
```
hdoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi query_top10_answered.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 query_top10_answered.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll query_top10_answered.sh
-rwxrwxrwx 1 hdoop hdoop 789 Nov  1 06:42 query_top10_answered.sh*
hdoop@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 des
erCount, 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
```

```
hdoop@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 des
erCount, 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
```

- 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

```
hdoop@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/hiv
!/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.compu
088/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
```





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.

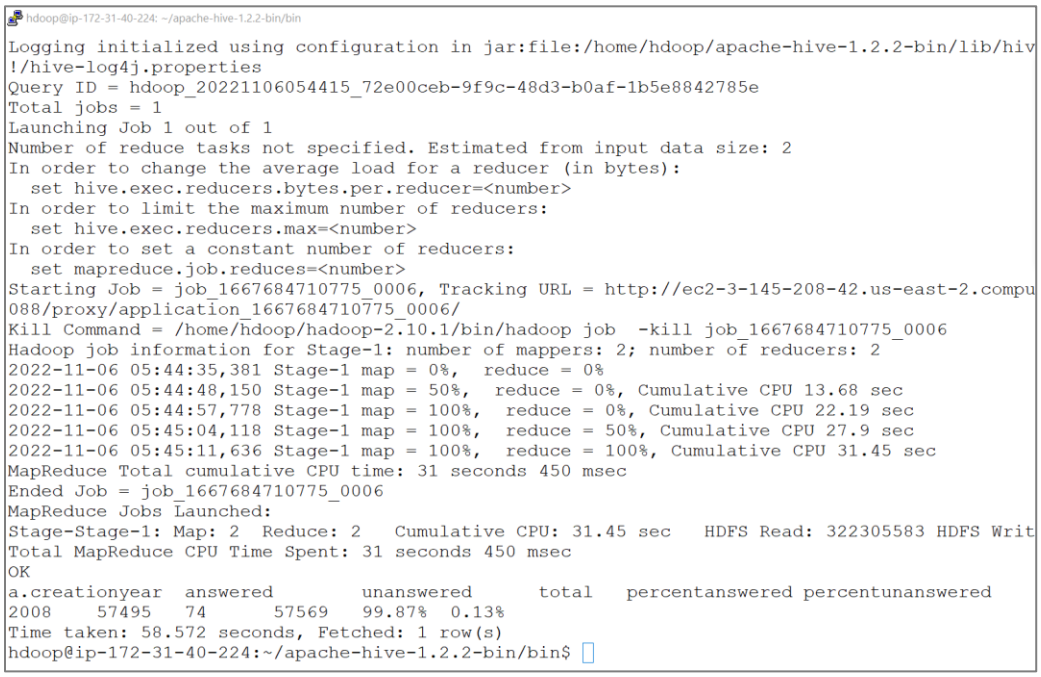
- Percentage of Stack Overflow Questions that Went Unanswered
  - Script: query_uanswered.sh
    Create the shell script *query_uanswered.sh*, save it in the /bin folder, then change file permission to make it executable.

```
hdoop@ip-172-31-40-224: ~/apache-hive-1.2.2-bin/bin
hdoop@ip-172-31-40-224:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ vi query_uanswered.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ chmod 777 query_uanswered.sh
hdoop@ip-172-31-40-224:~/apache-hive-1.2.2-bin/bin$ ll query_uanswered.sh
-rwxrwxrwx 1 hdoop hdoop 989 Nov  1 06:46 query_uanswered.sh*
hdoop@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 PercentUnanswere
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,
```

```
hdoop@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 PercentUnanswere
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
```

○ Usage and Result of *query_uanswered.sh*

1. Go to $HIVE_HOME/bin or /home/hdoop/apache-hive-1.2.2-bin/bin
2. Run ./query_uanswered.sh
3. Input Creation Year (yyyy)
4. Press Enter



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/hdoop/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/hdoop/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
;
"
```