LAMBTON COLLEGE

A Project on

[Hadoop Ecosystem]

121 Brunel Rd, Mississauga

ON L4Z 3E9

A Group project for analyzing stack overflow data in Hive on Top of Hadoop

Big Data Analytics DSMM

Under the supervision

of

Professor Teresa Zhu

Submitted BY:Group E
Aadarsha Chapagain (C0825975)
Onyinye Mbanefo (C0831578)
Roshan Acharya (C0831342)
Anjana Kuriakose (C0829580)

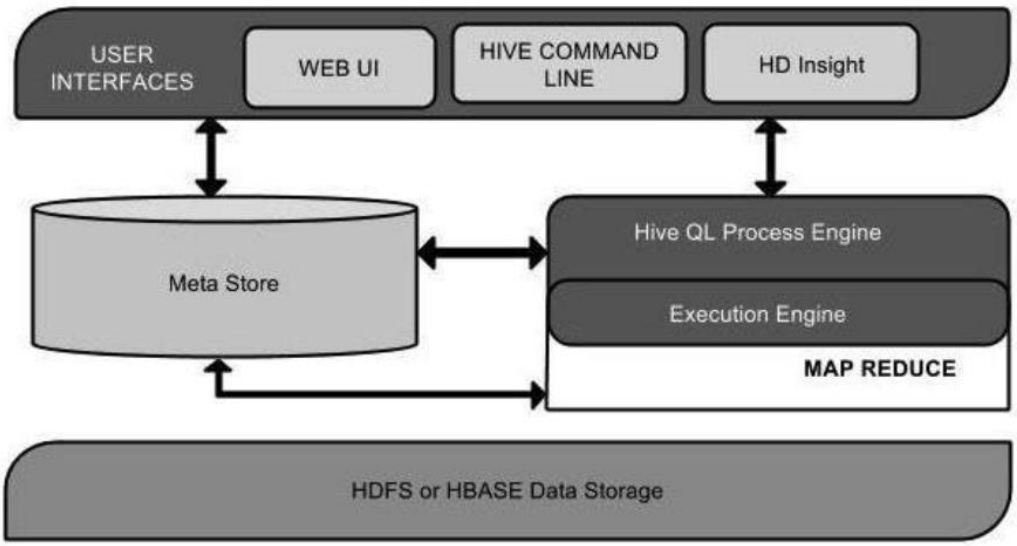
Submitted To:
Lambton College
Professor Teresa Zhu

Submission Date:
27th September 2022

Step1:

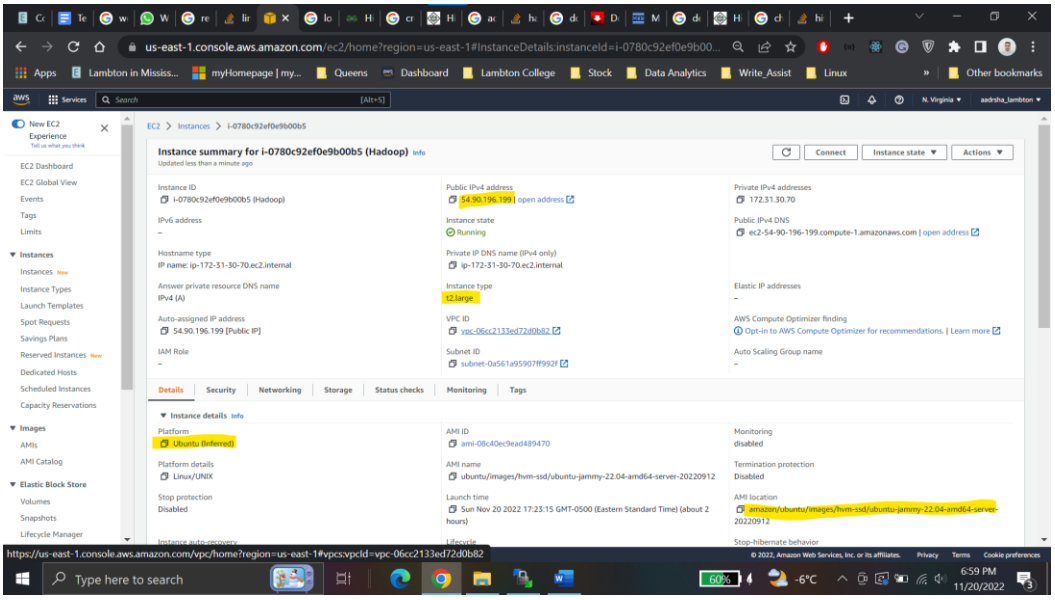
1. Download the zip file from the following location & move it to your Hadoop environment. Place this data under /LDZ/data/ in Hadoop

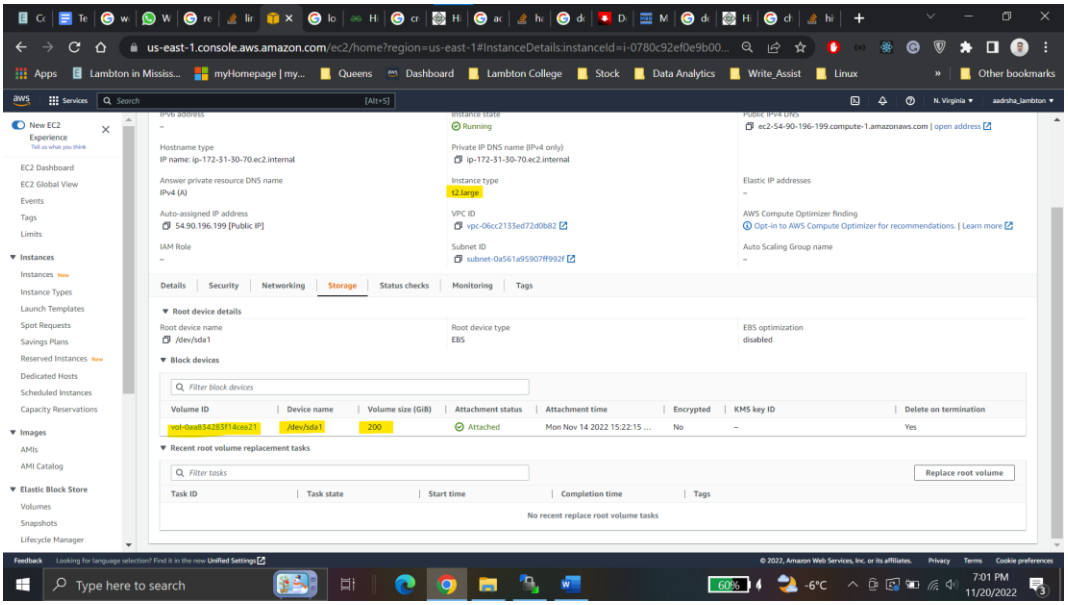
Architecture for Hive



Hadoop on Cloud Environment

Ubuntu 18.04 was installed on AWS ec2(**t2.large**) with 8gb of memory and 300gb of storage. Hadoop 2.6, Hive and mysql was installed on the machine.





Single Node on Hadoop was set up using following commands on AWS ec2.

1. create a new 'hadoop' user in ubuntu.

2. create '/home/hadoop/work' and '/home/hadoop/work/hadoopdata' folders

```
mkdir /home/hadoop/work

mkdir /home/hadoop/work/hadoopdata
```

3. Download 'hadoop-2.6.0.tar.gz' file from Apache mirrors https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/, copy 'hadoop-2.6.0.tar.gz' file into this '/home/hadoop/work' directory and extract the tar file into same directory.

tar -xvzf hadoop-2.6.0.tar.gz

4. Open the '~/.bashrc' file on all the machines and add the following lines at the end and save:

command: gedit ~/.bashrc

```
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64

export HADOOP_HOME=/home/hadoop/work/hadoop-2.6.0

export HADOOP_COMMON_HOME=$HADOOP_HOME

export HADOOP_HDFS_HOME=$HADOOP_HOME

export HADOOP_MAPRED_HOME=$HADOOP_HOME

export YARN_HOME=$HADOOP_HOME
```

```
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop

export YARN_CONF_DIR=$HADOOP_HOME/etc/hadoop

export
PATH=$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH
```

5. Enter the below commands on terminal:

```
ssh localhost

ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

6. Update the '/home/hadoop/work/hadoop-2.6.0/etc' folder files 'hadoop-env.sh','core-site.xml','hdfs-site.xml','mapred-site.xml', 'yarn-env.sh', 'yarn-site.xml','masters' and 'slaves' files as per the below configurations

hadoop-env.sh

```
=============

# The java implementation to use.

export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64

export JAVA_HOME=${JAVA_HOME}
```

core-site.xml

```
=============

<configuration>

        <property>

                <name>fs.defaultFS</name>

                <value>hdfs://localhost:8020</value>

        </property>

</configuration>
```

hdfs-site.xml

```
=============

<configuration>
```

```
        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>file:/home/hadoop/work/hadoopdata/dfs/name</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>file:/home/hadoop/work/hadoopdata/dfs/data</value>
        </property>
        <property>
                <name>dfs.namenode.checkpoint.dir</name>
                <value>file:/home/hadoop/work/hadoopdata/dfs/namesecondary</value>
        </property>
</configuration>
```

mapred-env.sh

==========

```
# export JAVA_HOME=/home/y/libexec/jdk1.8.0/
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
```

mapred-site.xml

==========

```
<configuration>
        <property>
                <name>mapreduce.framework.name</name>
```

```
                <value>yarn</value>
        </property>
</configuration>
```

yarn-env.sh

==========

```
# export JAVA_HOME=/home/y/libexec/jdk1.8.0/
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
```

yarn-site.xml

==========

```
<configuration>
        <property>
                <name>yarn.resourcemanager.hostname</name>
                <value>localhost</value>
        </property>
        <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>
</configuration>
```

slaves

======

localhost

7. Format the 'namenode' from current machine using this command:

hadoop namenode -format or hdfs namenode -format

8. Start the hadoop by using this command on current machine:

start-dfs.sh

start-all.sh (depricated)

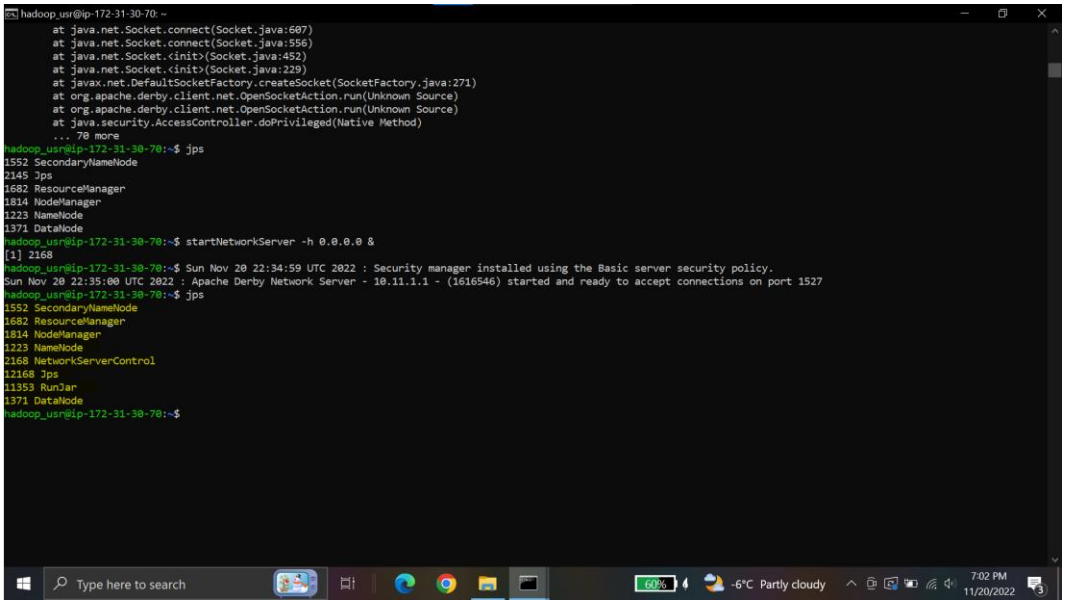
9. Stop the hadoop by using this command on current machine:

start-yarn.sh

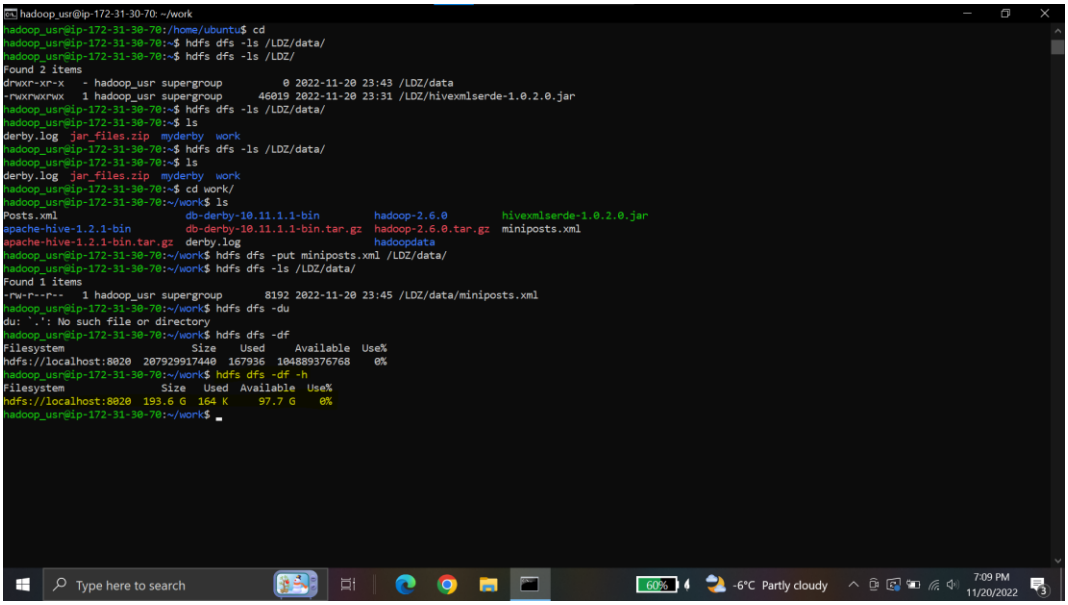
stop-all.sh (depricated)

10. jps

With JPS it can been seen that namenode and resource manager are started.

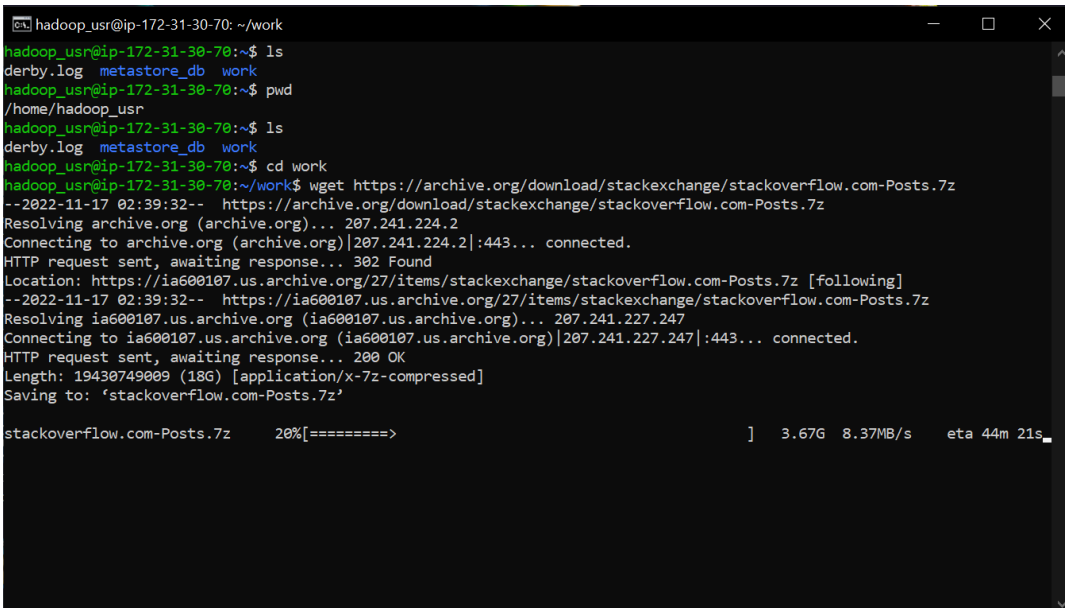


To check hadoop storage hdfs dfs -df -h command can be used



Once hadoop is installed create directory named LDZ and /LDZ/data in Hadoop and get the data in local system, unzip it and transfer it to hadoop .

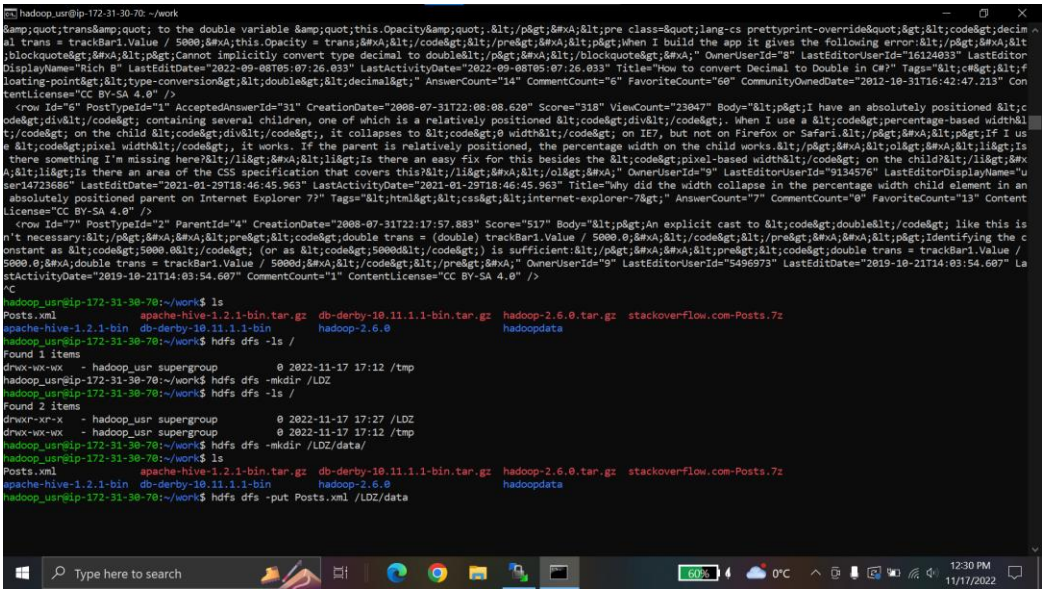
Download the zip file from the following location & move it to your Hadoop environment. Place this data under /LDZ/data/ in Hadoop



The zip file was downloaded into the ubuntu machine using **wget** command

Transfer file to HDFS

File transferred to HDFS into the location /DWZ/data using put command in HDFS



Step 2:

3. Build a data warehouse (location:/DWZ) with data partitioned based on Creation Date and then Post Type.

Install Hive in hadoo using following commands

1. Download 'apache-hive-1.2.1-bin' file from apache mirrors https://archive.apache.org/dist/hive/hive-1.2.1/, copy 'apache-hive-1.2.1-bin' file into this'/home/hadoop/work' directory and extract the tar files into same directory.

Using terminal:tar -xvzf apache-hive-1.2.1-bin

2. Open the '~/.bashrc' file using command: gedit ~/.bashrc in the terminal add below lines at the end of the document.

export HIVE_HOME=/home/hadoop/work/apache-hive-1.2.1-bin

export PATH=$HIVE_HOME/bin:$PATH

3. Copy mysql-connector-java-5.1.38 to /home/hadoop/work/apache-hive-1.2.1-bin/lib

Copy /home/hadoop/work/db-derby-10.11.1.1-bin/lib/derbyclient.jar to /home/hadoop/work/apache-hive-1.2.1-bin/lib

4. The following configuration files are required for hive to be run in different modes.

1.Hive-site.xml (Main configuration file)

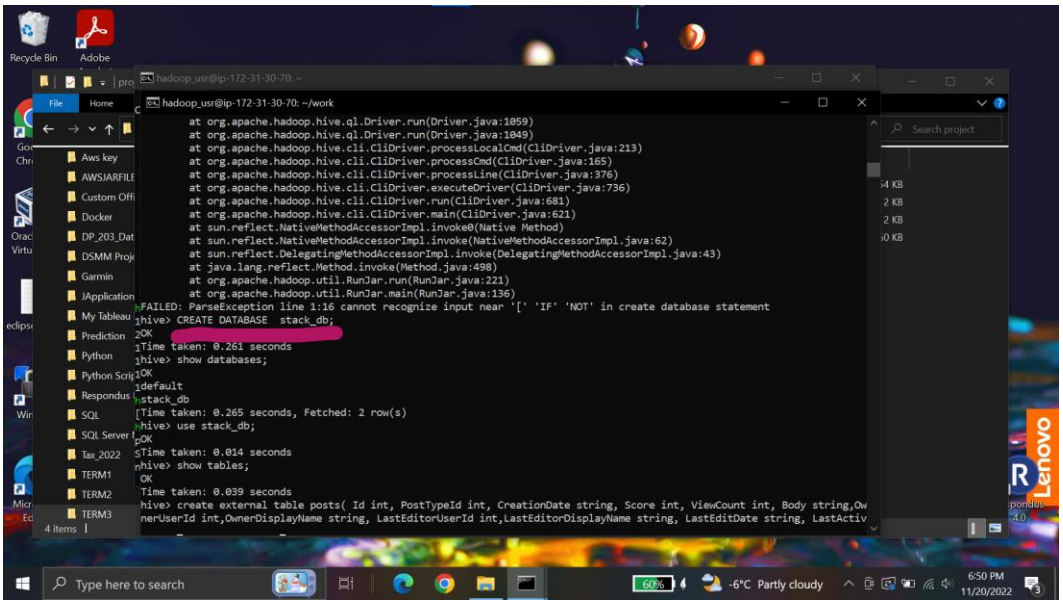
2.hive-site.xml_local (For this mode, copy paste this script into Main configuration file)

3.Hive-site.xml_derby (For this mode, copy paste this script into Main configuration file)

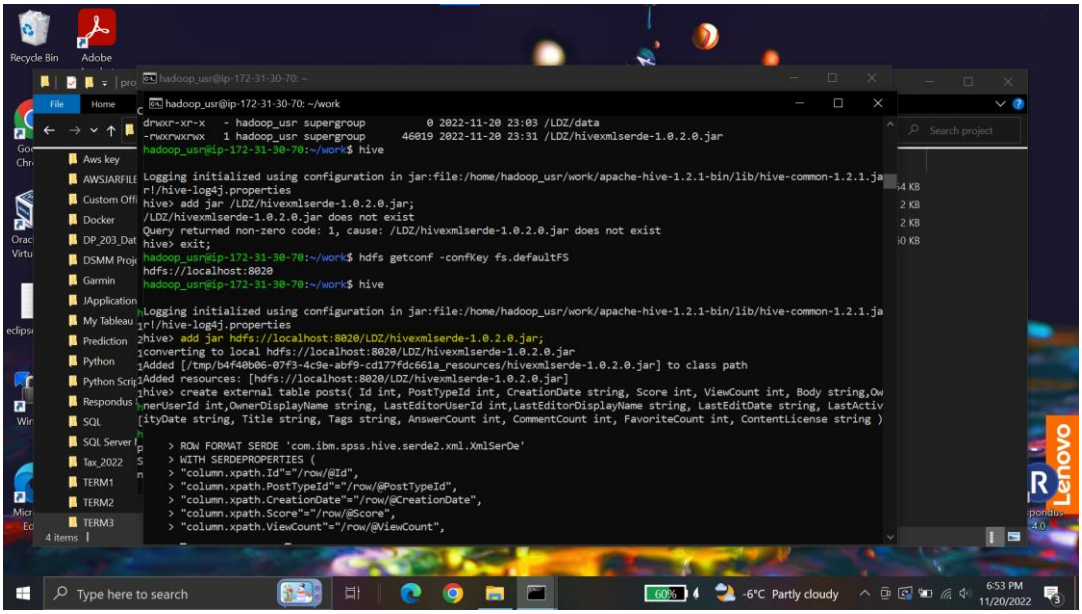
4.hive-site.xml_mysql (For this mode, copy paste this script into Main configuration file)

5. Open hive-site.xml from /home/hadoop/work/apache-hive-1.2.1-bin/conf. Edit configuration with below properties.

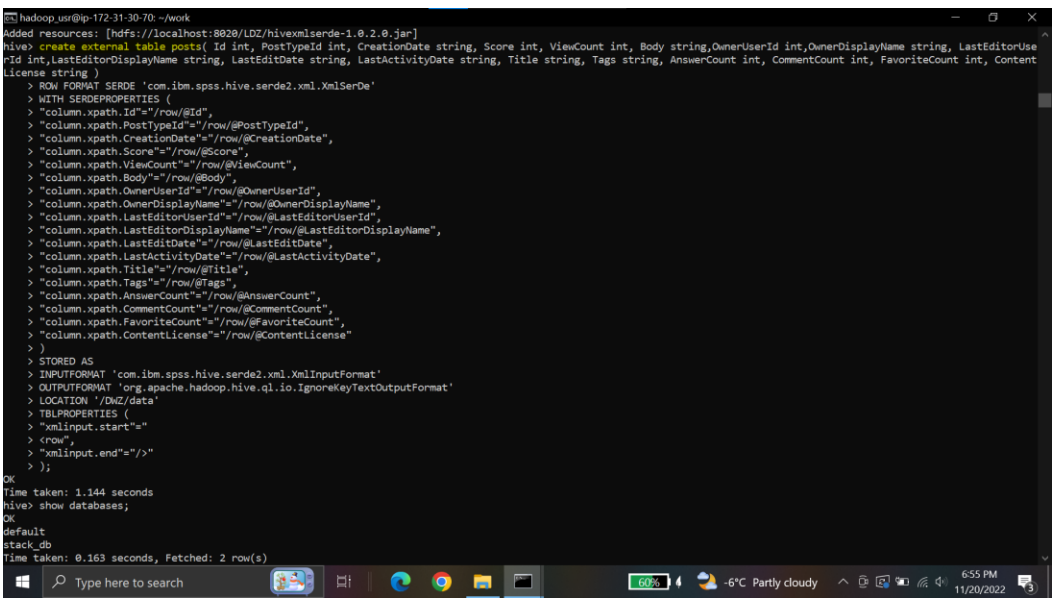
Once Hive is installed, add xml create database and table in Hadoop



We need serializer before creating table in hive, so download the serializer put it in hadoop and add jar file to hive providing the location



Create External table in hive



create external table posts(Id int, PostTypeId int, CreationDate string, Score int, ViewCount int, Body string,OwnerUserId int,OwnerDisplayName string, LastEditorUserId int,LastEditorDisplayName string, LastEditDate string, LastActivityDate string, Title string, Tags string, AnswerCount int, CommentCount int, FavoriteCount int, ContentLicense string)

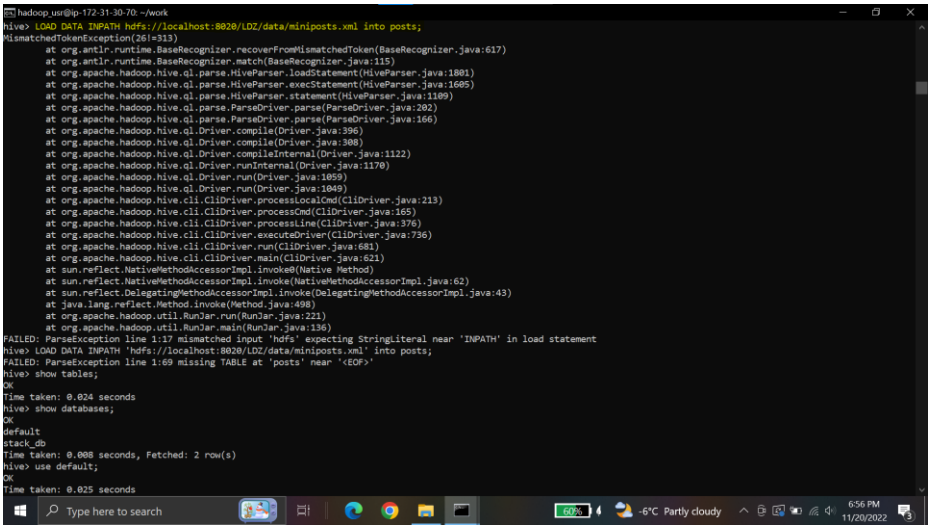
ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'

WITH SERDEPROPERTIES (

"column.xpath.Id"="/row/@Id",

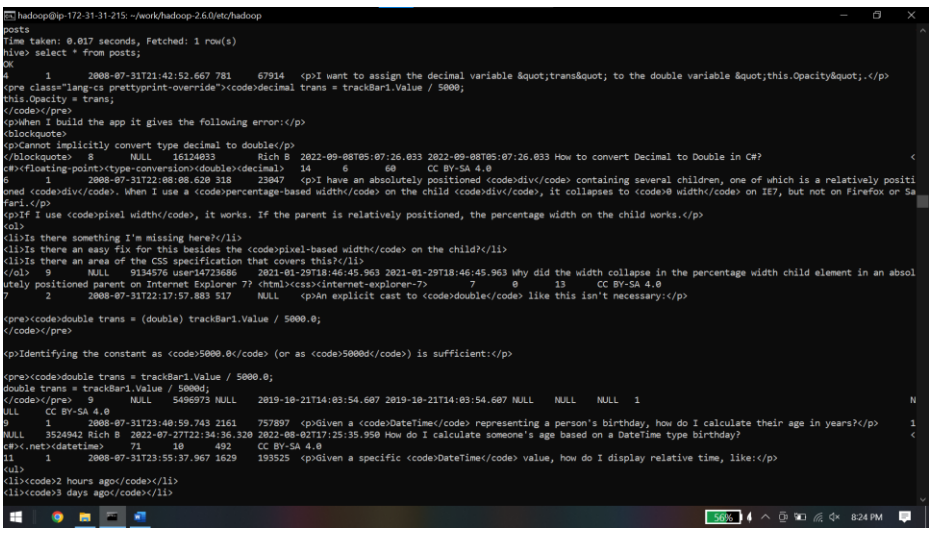
```
"column.xpath.PostTypeId"="/row/@PostTypeId",
"column.xpath.CreationDate"="/row/@CreationDate",
"column.xpath.Score"="/row/@Score",
"column.xpath.ViewCount"="/row/@ViewCount",
"column.xpath.Body"="/row/@Body",
"column.xpath.OwnerUserId"="/row/@OwnerUserId",
"column.xpath.OwnerDisplayName"="/row/@OwnerDisplayName",
"column.xpath.LastEditorUserId"="/row/@LastEditorUserId",
"column.xpath.LastEditorDisplayName"="/row/@LastEditorDisplayName",
"column.xpath.LastEditDate"="/row/@LastEditDate",
"column.xpath.LastActivityDate"="/row/@LastActivityDate",
"column.xpath.Title"="/row/@Title",
"column.xpath.Tags"="/row/@Tags",
"column.xpath.AnswerCount"="/row/@AnswerCount",
"column.xpath.CommentCount"="/row/@CommentCount",
"column.xpath.FavoriteCount"="/row/@FavoriteCount",
"column.xpath.ContentLicense"="/row/@ContentLicense"
)
STORED AS
INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'
OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'
LOCATION '/DWZ/data'
TBLPROPERTIES (
"xmlinput.start"="
<row",
"xmlinput.end"="/>"
);
```

Load data into posts table



```
hadoop_usr@ip-172-31-30-70: ~/work
hive> LOAD DATA INPATH hdfs://localhost:8020/LDZ/data/miniposts.xml into posts;
MismatchedTokenException(26!=313)
        at org.antlr.runtime.BaseRecognizer.recoverFromMismatchedToken(BaseRecognizer.java:617)
        at org.antlr.runtime.BaseRecognizer.match(BaseRecognizer.java:115)
        at org.apache.hadoop.hive.ql.parse.HiveParser.loadStatement(HiveParser.java:1801)
        at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:1605)
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1109)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:202)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:166)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:396)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:308)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1122)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1170)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1059)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1049)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:213)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:165)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:376)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:736)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:683)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:621)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:17 mismatched input 'hdfs' expecting StringLiteral near 'INPATH' in load statement
hive> LOAD DATA INPATH 'hdfs://localhost:8020/LDZ/data/miniposts.xml' into posts;
FAILED: ParseException line 1:69 missing TABLE at 'posts' near '<EOF>'
hive> show tables;
OK
Time taken: 0.024 seconds
hive> show databases;
OK
default
stack_db
Time taken: 0.008 seconds, Fetched: 2 row(s)
hive> use default;
OK
Time taken: 0.025 seconds
```

Show few data in hive in table posts



```
hadoop@ip-172-31-31-215: ~/work/hadoop-2.6.0/etc/hadoop
posts
Time taken: 0.017 seconds, Fetched: 1 row(s)
hive> select * from posts;
OK
4       1       2008-07-31T21:42:52.667 781     67914   <p>I want to assign the decimal variable &amp;quot;trans&amp;quot; to the double variable &amp;quot;this.Opacity&amp;quot;.</p>
<pre class="lang-cs prettyprint-override"><code>decimal trans = trackBar1.Value / 5000;
this.Opacity = trans;
</code></pre>
<p>When I build the app it gives the following error:</p>
<blockquote>
<p>Cannot implicitly convert type decimal to double</p>
</blockquote>   8       NULL    16124033        Rich B  2022-09-08T05:07:26.033 2022-09-08T05:07:26.033 How to convert Decimal to Double in C#?                                <
c#><floating-point><type-conversion><double><decimal>   14      6       60      CC BY-SA 4.0
6       1       2008-07-31T22:08:08.620 318     23047   <p>I have an absolutely positioned <code>div</code> containing several children, one of which is a relatively positi
oned <code>div</code>. When I use a <code>percentage-based width</code> on the child <code>div</code>, it collapses to <code>0 width</code> on IE7, but not on Firefox or Sa
fari.</p>
<p>If I use <code>pixel width</code>, it works. If the parent is relatively positioned, the percentage width on the child works.</p>
<ol>
<li>Is there something I'm missing here?</li>
<li>Is there an easy fix for this besides the <code>pixel-based width</code> on the child?</li>
<li>Is there an area of the CSS specification that covers this?</li>
</ol>   9       NULL    9134576 user14723686    2021-01-29T18:46:45.963 2021-01-29T18:46:45.963 Why did the width collapse in the percentage width child element in an absol
utely positioned parent on Internet Explorer 7? <html><css><internet-explorer-7>        7       0       13      CC BY-SA 4.0
7       2       2008-07-31T22:17:57.883 517     NULL    <p>An explicit cast to <code>double</code> like this isn't necessary:</p>

<pre><code>double trans = (double) trackBar1.Value / 5000.0;
</code></pre>

<p>Identifying the constant as <code>5000.0</code> (or as <code>5000d</code>) is sufficient:</p>

<pre><code>double trans = trackBar1.Value / 5000.0;
double trans = trackBar1.Value / 5000d;
</code></pre>   9       NULL    5496973 NULL    2019-10-21T14:03:54.607 2019-10-21T14:03:54.607 NULL    NULL    NULL    1                                                N
ULL     CC BY-SA 4.0
9       1       2008-07-31T23:40:59.743 2161    757897  <p>Given a <code>DateTime</code> representing a person's birthday, how do I calculate their age in years?</p>      1
NULL    3524942 Rich B  2022-07-27T22:34:36.320 2022-08-02T17:25:35.950 How do I calculate someone's age based on a DateTime type birthday?                                <
c#><.net><datetime>     71      10      492     CC BY-SA 4.0
11      1       2008-07-31T23:55:37.967 1629    193525  <p>Given a specific <code>DateTime</code> value, how do I display relative time, like:</p>
<ul>
<li><code>2 hours ago</code></li>
<li><code>3 days ago</code></li>
```

```
hive> select id, posttypeid, creationdate from posts limit 20;
OK
4       1       2008-07-31T21:42:52.667
6       1       2008-07-31T22:08:08.620
7       2       2008-07-31T22:17:57.883
9       1       2008-07-31T23:40:59.743
11      1       2008-07-31T23:55:37.967
12      2       2008-07-31T23:56:41.303
13      1       2008-08-01T00:42:38.903
14      1       2008-08-01T00:59:11.177
4       1       2008-07-31T21:42:52.667
6       1       2008-07-31T22:08:08.620
7       2       2008-07-31T22:17:57.883
9       1       2008-07-31T23:40:59.743
11      1       2008-07-31T23:55:37.967
12      2       2008-07-31T23:56:41.303
13      1       2008-08-01T00:42:38.903
14      1       2008-08-01T00:59:11.177
16      1       2008-08-01T04:59:33.643
17      1       2008-08-01T05:09:55.993
18      2       2008-08-01T05:12:44.193
19      1       2008-08-01T05:21:22.257
Time taken: 0.087 seconds, Fetched: 20 row(s)
hive>
```

Step 3:

Querying Database:

What are the top 10 most answered questions in Stack Overflow posts for a particular creation date?

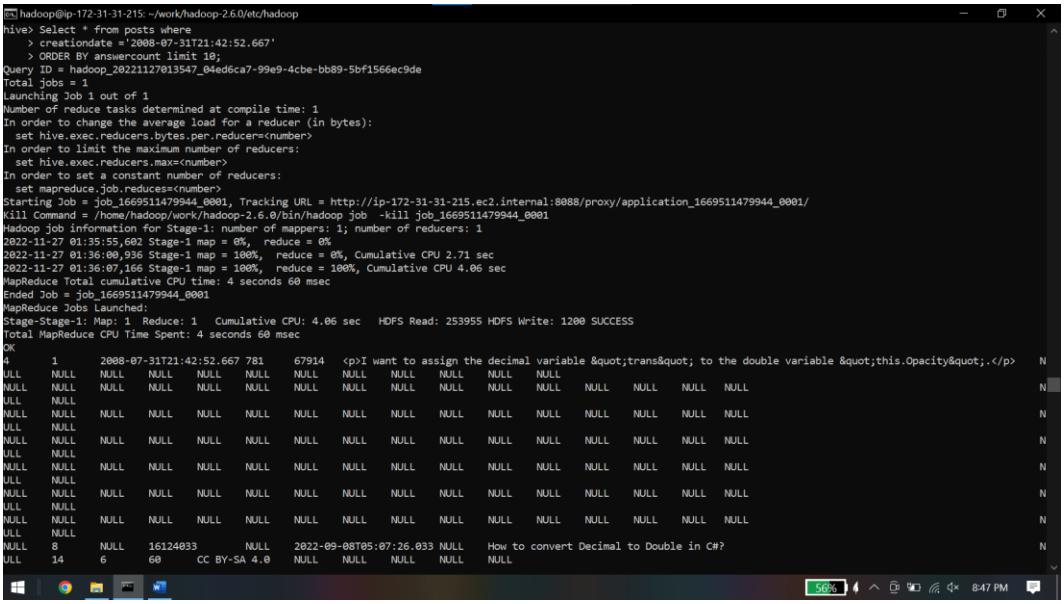
Query:

Select * from posts where

creationdate ='2008-07-31T21:42:52.667'

ORDER BY answercount limit 10;

;



```
hive> Select * from posts where
    > creationdate ='2008-07-31T21:42:52.667'
    > ORDER BY answercount limit 10;
Query ID = hadoop_20221127013547_04ed6ca7-99e9-4cbe-bb89-5bf1566ec9de
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1669511479944_0001, Tracking URL = http://ip-172-31-31-215.ec2.internal:8088/proxy/application_1669511479944_0001/
Kill Command = /home/hadoop/work/hadoop-2.6.0/bin/hadoop job  -kill job_1669511479944_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-27 01:35:55,602 Stage-1 map = 0%,  reduce = 0%
2022-11-27 01:36:00,936 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.71 sec
2022-11-27 01:36:07,166 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.06 sec
MapReduce Total cumulative CPU time: 4 seconds 60 msec
Ended Job = job_1669511479944_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.06 sec   HDFS Read: 253955 HDFS Write: 1200 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 60 msec
OK
4       1       2008-07-31T21:42:52.667 781     67914   <p>I want to assign the decimal variable &quot;trans&quot; to the double variable &quot;this.Opacity&quot;.</p>    N
ULL     NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL                            N
ULL     NULL
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL                            N
ULL     NULL
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL                            N
ULL     NULL
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL                            N
ULL     NULL
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL                            N
ULL     NULL
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL                            N
ULL     NULL
NULL    8       NULL    16124033        NULL    2022-09-08T05:07:26.033 NULL    How to convert Decimal to Double in C#?                                 N
ULL     14      6       60      CC BY-SA 4.0    NULL    NULL    NULL    NULL    NULL
```

What’s the percentage of Stack Overflow questions that went unanswered in say 2015?

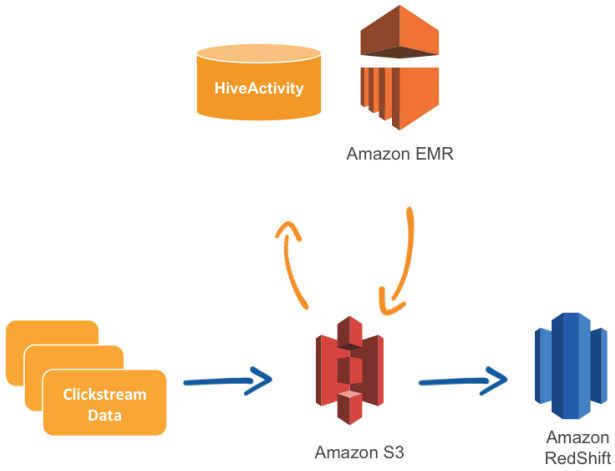
Query:

Select * from posts where creationdate like '20015 %' AND answercount=0;

Step 4

Automation

This process can be automated using pipeline. A batch of data can be uploaded daily and required transformation can be performed. After performing the transformation, the data will be loaded into the warehouse. From data stored in the warehouse’s analytics can be performed.



Conclusion

Hence the required architecture for the analysis of the stackoverflow data was built using

- Hadoop
- Hive

On Hive, mysql mode was used and the data was loaded in the external table and queries were performed to answer the given questions.