Data Warehouse & Data Mining

a Big Data S/W Project

Sakshi Gupta
Sakshi.Gupta1@ge.com
github: saksgupt
portalname: saksgupt

Michael Eddie
Michael.eddie@ge.com
github: michaeleddie789
portalname: meddie

Description

Our project aim is to process information for enhanced insight and decision-making. We are using big data tools that will turn data into useful information. Below procedure is followed for project implementation.

1. Take raw data of page view statistics from wikimedia
   https://dumps.wikimedia.org/other/pagecounts-raw/
   Dataset is related to page view statistics with respect to hour, day, month and year
2. Upload datasets to the MongoDB database
3. Create single node Hadoop cluster
4. Clustering of dataset using Map Reduce java program
5. Execute on a virtual cluster
6. Output on command line

Technologies Used

- Shell Script
- Java
- Hadoop
- MongoDB
- Ansible

Github Repository Link

https://github.com/futuresystems/465-project-datawarehousemining

Installation Instructions

1) Instance Creation

Prerequisite:

- module load openstack
- source ~/.cloudmesh/clouds/india/juno/openrc.sh
- source ENV/bin/activate

2) Loading Data into MongoDB

3) Hadoop Deployment on single node

Copy Hadoop_Deployment_Automation.sh from github to cd /home/ubuntu

Commands:

```
Sudo su –

cd /home/Ubuntu

sudo bash Hadoop_Deployment_Automation.sh instance name
```

(example : sudo bash Hadoop_Deployment_Automation.sh **saksgupt-001)**

5) Map reduce java program

Copy WikiDataAnalysis.java from github to cd /home/ubuntu

7) Run the script to execute java file

Copy Wiki_Data_Analysis_Automation.sh from github to cd /home/ubuntu

Prerequisite: Availability of mongodb = **wikimedia_project** with collection = **pagecounts_small_May14**

If not available, then execute:

{

```
wget https://dumps.wikimedia.org/other/pagecounts-raw/2014/2014-05/pagecounts-20140501-000000.gz
gunzip pagecounts-20140501-000000.gz

echo "adding year month day data to each line”

 sed "s/^/2014 5 1 0 /" < pagecounts-20140501-000000 > pagecounts-20140501-000000-prefixed

echo "replacing all spaces with commas"

 tr ' ' ',' < pagecounts-20140501-000000-prefixed >pagecounts-20140501-000000.csv

echo "convert to UTF-8"

iconv -f ISO-8859-1 -t UTF-8 pagecounts-20140501-000000.csv 20140501-000000-UT8.csv

echo "importing data into mongodb"

 mongoimport --db wikimedia_project --collection pagecounts_small_May14 --type csv --fieldFile
pagecount_headers.txt --file 20140501-000000-UT8.csv
```

Please note pagecount_headers.txt is available in github

}

Then execute the script

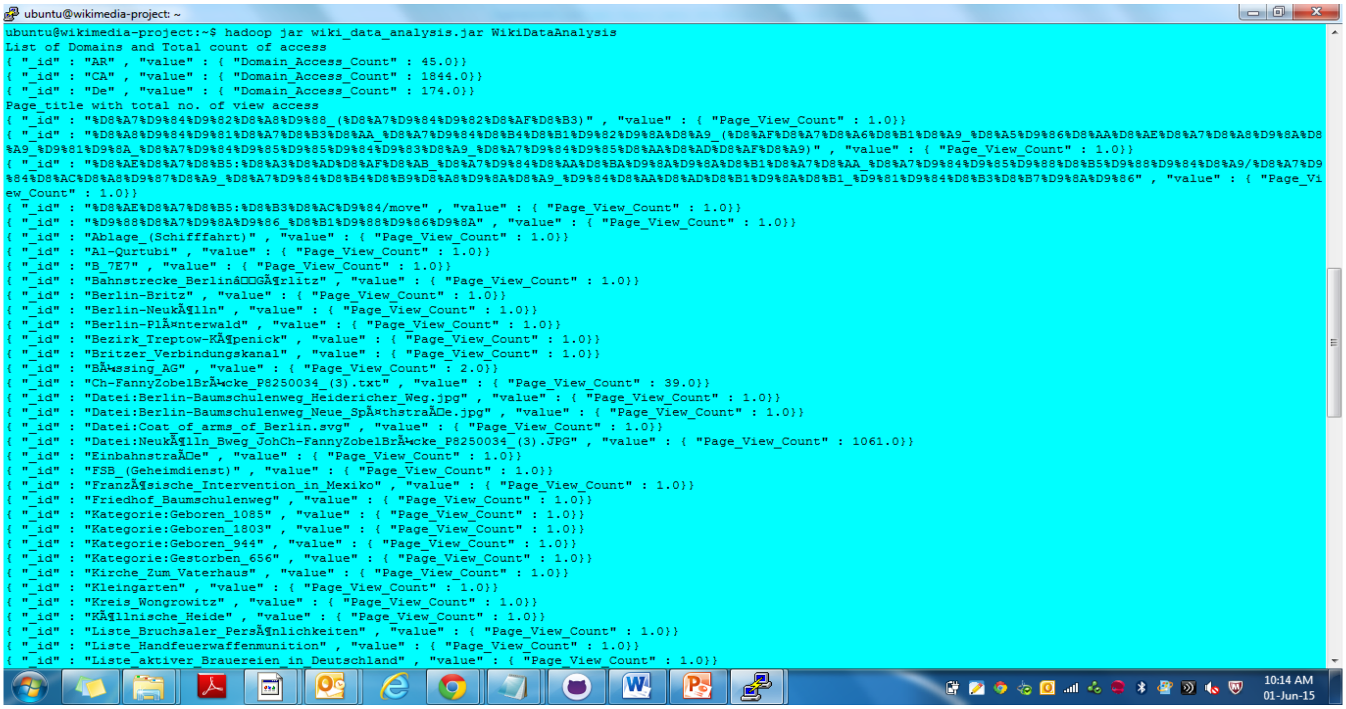
Sudo bash Wiki_Data_Analysis_Automation.sh

Output Snapshots

- ✓ List all the Domains along with the count of the access
  (One Domain can have multiple page titles which will have different view counts so this is the aggregation sum till Domain level)

- ✓ List all the Page titles along with view access count

  (One page_title can be accessed multiple times in different hours so total view count shows the final number)



```
ubuntu@wikimedia-project:~$ hadoop jar wiki_data_analysis.jar WikiDataAnalysis
List of Domains and Total count of access
{ "_id" : "AR" , "value" : { "Domain_Access_Count" : 45.0}}
{ "_id" : "CA" , "value" : { "Domain_Access_Count" : 1844.0}}
{ "_id" : "De" , "value" : { "Domain_Access_Count" : 174.0}}
Page_title with total no. of view access
{ "_id" : "%D8%A7%D9%84%D9%82%D8%A8%D9%88_(%D8%A7%D9%84%D9%82%D8%AF%D8%B3)" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "%D8%A8%D9%84%D9%81%D8%A7%D8%B3%D8%AA_%D8%A7%D9%84%D8%B4%D8%B1%D9%82%D9%8A%D8%A9_(%D8%AF%D8%A7%D8%A6%D8%B1%D8%A9_%D8%A5%D9%86%D8%AA%D8%AE%D8%A7%D8%A8%D9%8A%D8
%A9_%D9%81%D9%8A_%D8%A7%D9%84%D9%85%D9%85%D9%84%D9%83%D8%A9_%D8%A7%D9%84%D9%85%D8%AA%D8%AD%D8%AF%D8%A9)" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "%D8%AE%D8%A7%D8%B5:%D8%A3%D8%AD%D8%AF%D8%AB_%D8%A7%D9%84%D8%AA%D8%BA%D9%8A%D9%8A%D8%B1%D8%A7%D8%AA_%D8%A7%D9%84%D9%85%D9%88%D8%B5%D9%88%D9%84%D8%A9/%D8%A7%D9
%84%D8%AC%D8%A8%D9%87%D8%A9_%D8%A7%D9%84%D8%B4%D8%B9%D8%A8%D9%8A%D8%A9_%D9%84%D8%AA%D8%AD%D8%B1%D9%8A%D8%B1_%D9%81%D9%84%D8%B3%D8%B7%D9%8A%D9%86" , "value" : { "Page_Vi
ew_Count" : 1.0}}
{ "_id" : "%D8%AE%D8%A7%D8%B5:%D8%B3%D8%AC%D9%84/move" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "%D9%88%D8%A7%D9%8A%D9%86_%D8%B1%D9%88%D9%86%D9%8A" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Ablage_(Schifffahrt)" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Al-Qurtubi" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "B_7E7" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Bahnstrecke_Berlinâ□□GÃ¶rlitz" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Berlin-Britz" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Berlin-NeukÃ¶lln" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Berlin-PlÃ¤nterwald" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Bezirk_Treptow-KÃ¶penick" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Britzer_Verbindungskanal" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "BÃ¼ssing_AG" , "value" : { "Page_View_Count" : 2.0}}
{ "_id" : "Ch-FannyZobelBrÃ¼cke_P8250034_(3).txt" , "value" : { "Page_View_Count" : 39.0}}
{ "_id" : "Datei:Berlin-Baumschulenweg_Heidericher_Weg.jpg" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Datei:Berlin-Baumschulenweg_Neue_SpÃ¤thstraÃ□e.jpg" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Datei:Coat_of_arms_of_Berlin.svg" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Datei:NeukÃ¶lln_Bweg_JohCh-FannyZobelBrÃ¼cke_P8250034_(3).JPG" , "value" : { "Page_View_Count" : 1061.0}}
{ "_id" : "EinbahnstraÃ□e" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "FSB_(Geheimdienst)" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "FranzÃ¶sische_Intervention_in_Mexiko" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Friedhof_Baumschulenweg" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Kategorie:Geboren_1085" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Kategorie:Geboren_1803" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Kategorie:Geboren_944" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Kategorie:Gestorben_656" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Kirche_Zum_Vaterhaus" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Kleingarten" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Kreis_Wongrowitz" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "KÃ¶llnische_Heide" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_Bruchsaler_PersÃ¶nlichkeiten" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_Handfeuerwaffenmunition" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_aktiver_Brauereien_in_Deutschland" , "value" : { "Page_View_Count" : 1.0}}
```

Learnings

This project provided the end to end understanding of how to think of a problem and take it to a working solution.

Starting from the phrasing of problem statement ,understanding the dataset, loading the data into DB, interfaces among various technologies, deal with unstructured NO SQL huge data, understanding Hadoop , clustering using Map and reduce and deduce an analytical solution which might be helpful for strategic planning or Decision making.

Moreover, we learned how to ensure that your system is reproducible (Devops) in case of instance failure/deletion.

It’s a learning through complete hands-on on the machine.