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 across all Wikimedia sites (ie Wikipedia, wikibooks, Wikimedia, etc.) with respect to hour, day, month and year
2. Create multi-node cluster using OpenStack (scalable and configurable)
3. Deploy mongodb to all nodes in cluster (positioned for future sharding across cluster)
4. Extract, transfer, and load datasets to one of the MongoDB instances
5. Create single node Hadoop cluster
6. Clustering of dataset using Map Reduce java program
7. Execute on a virtual cluster
8. Output on command line
9. Single CM command for installation

Technologies Used

- Shell Script
- Java
- Hadoop
- Python
- Cloudmesh cm command
- MongoDB
- Ansible

Github Repository Link

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

Pre-requisites Required for Setup

- Must have a provisioned VM with cloudmesh installed
- Must have ansible installed
- Must have git installed
- Following commands should be used for setting up your environment
    a. module load openstack
    b. source ENV/bin/activate (make sure you are using a virtual env)
    c. eval $(ssh-agent -s)
    d. ssh-add ~/.ssh/id_rsa
    e. /cloudmesh*/python setup.py install
    f. nano ~/.cloudmesh/cmd3.yaml and add - cloudmesh_wikicount.plugins
    g. India cloud activated

User Manual

- git clone https://github.com/futuresystems/465-project-datawarehousemining.git
- Go to 465-project-datawarehousemining folder, execute the following command: **cm wikicount install**

This command internally executing below scripts

- ✓ Loading Data into MongoDB using import_wiki_pagecounts_May2014_1.sh
- ✓ Hadoop Deployment on single node using Hadoop_Deployment_Automation.sh "**instance name"**
- ✓ Map reduce java program- WikiDataAnalysis.java
- ✓ Execute java file using Wiki_Data_Analysis_Automation.sh

  (Prerequisite: Availability of mongodb = wikimedia_project with collection = pagecounts_small_May14)

Output Snapshots

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

List of Domains and Total count of access

{ "_id" : "AR" , "value" : { "Domain_Access_Count" : 48.0}}

{ "_id" : "CA" , "value" : { "Domain_Access_Count" : 1985.0}}

{ "_id" : "De" , "value" : { "Domain_Access_Count" : 179.0}}

✓ **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)

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_View_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" : 42.0}}

{ "_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" : 1141.0}}

{ "_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}}
{ "_id" : "Liste_antiker_BrÃ¼ckenbauten" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_antiker_Theaterbauten" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_australischer_Erfinder_und_Entdecker" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_bedeutender_TÃ¤nzer" , "value" : { "Page_View_Count" : 2.0}}
{ "_id" : "Liste_der_BaudenkmÃ¤ler_in_Itzgrund" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_in_Berlin-Borsigwalde" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_und_PlÃ¤tze_in_Berlin-Baumschulenweg" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_und_PlÃ¤tze_in_Berlin-Blankenfelde" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_und_PlÃ¤tze_in_Berlin-Buch" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_und_PlÃ¤tze_in_Berlin-Gesundbrunnen" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_und_PlÃ¤tze_in_Berlin-Johannisthal" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_und_PlÃ¤tze_in_Berlin-Kladow" , "value" : { "Page_View_Count" : 1.0}}
{ "_id" : "Liste_der_StraÃ�en_und_PlÃ¤tze_in_Berlin-Lichtenberg" , "value" : { "Page_View_Count" : 1.0}}

{ "_id" : "Liste_des_UNESCO-Welterbes_(Afrika)" , "value" : { "Page_View_Count" : 1.0}}

{ "_id" : "Liste_des_gefÃ¤hrdeten_Welterbes" , "value" : { "Page_View_Count" : 1.0}}

{ "_id" : "Liste_deutscher_Bezeichnungen_russischer_Orte" , "value" : { "Page_View_Count" : 1.0}}

{ "_id" : "Liste_deutschsprachiger_TheaterstÃ¼cke" , "value" : { "Page_View_Count" : 2.0}}

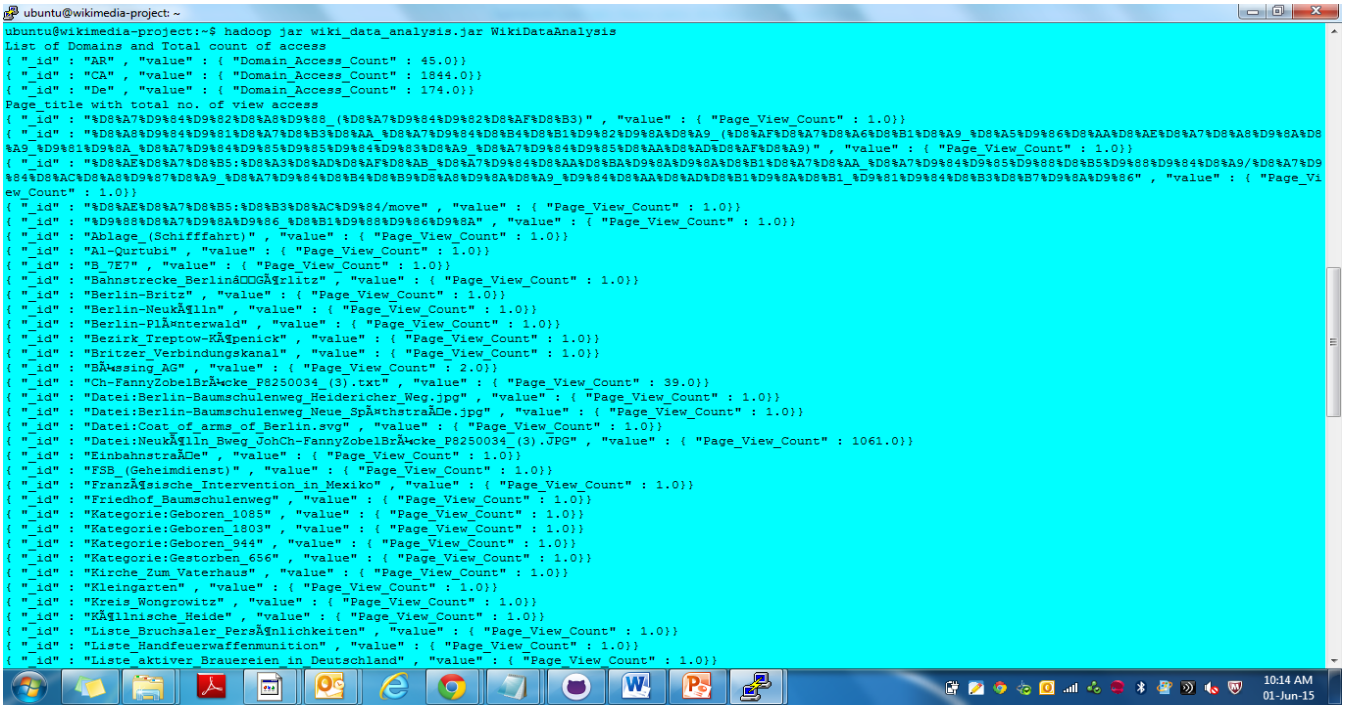
{ "_id" : "Liste_dÃ¤nischer_Schachspieler" , "value" : { "Page_View_Count" : 1.0}}

{ "_id" : "Liste_emigrierter_deutschsprachiger_Sozialwissenschaftler" , "value" : { "Page_View_Count" : 1.0}}

{ "_id" : "Liste_groÃ�er_Siedlungen_in_Russland" , "value" : { "Page_View_Count" : 1.0}}

{ "_id" : "Liste_italienischer_HÃ¶rfunksender" , "value" : { "Page_View_Count" : 1.0}}

PS: Snapshot is not visible in IE browser



Lessons Learned

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.