Data warehouse & Data Mining

a Big data Project

Michael Eddie

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

Sakshi Gupta

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

imagination at work

Objective

To process information for enhanced insight and decision-making. We will be using business intelligence tools that will turn data into useful information.

Below procedure will be followed for project implementation :-

- Take dataset from wikiprojects

(Source: https://dumps.wikimedia.org/other/pagecounts-raw/)

- Upload datasets to the Mongo database
- Clustering based on type of dataset using Apache Mahout (Explained in further slides)
- Execute on a virtual cluster
- Visualize with D3

Technologies

- MongoDB
- D3 visualization
- Apache Mahout
- Python

Dataset -Page view statistics for Wikimedia projects

- Considering 2014 data:

  8760 txt files (12 months* 30/31 days * 24 hours)

- Create collections with additional column for year, month & hour

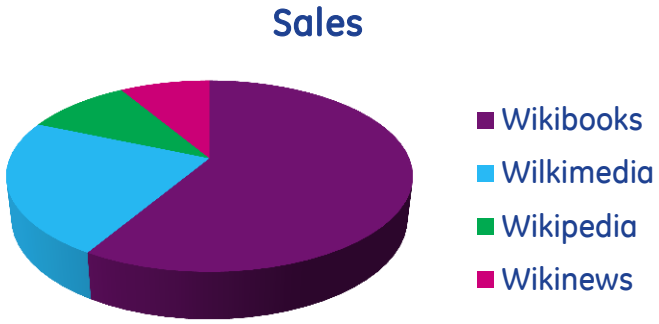
  Converting text files to csv to load the data

- Facing difficulty to load complete data set with a script in one go so Sharding MongoDB

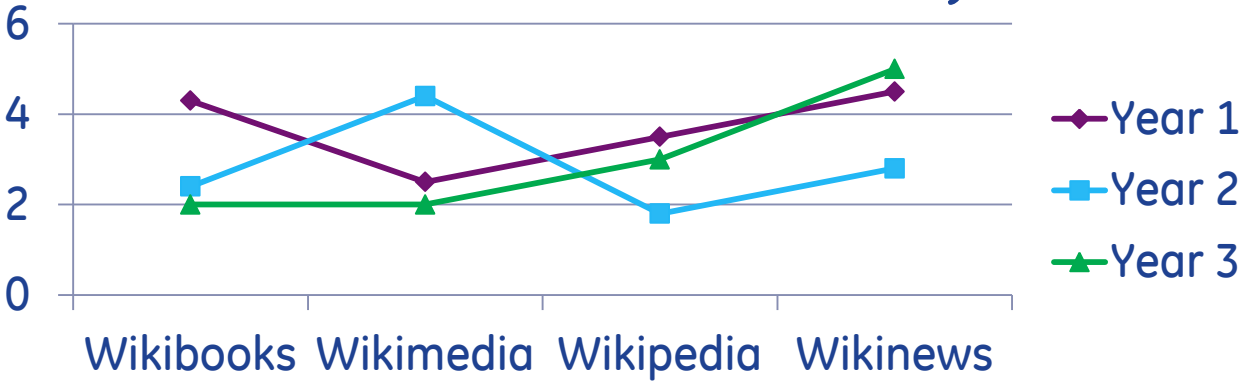
Year	Month	Hour	Domain	Page_Title	Count_Views	Total_Response_Size
2014	1	1	af	Afrikaans	1	1
2014	1	2	af	Albani%C3%AB	1	1
2014	1	3	af	Albert_Einstein	2	2
2014	2	1	aa	Record_to_Report	5	8

What all could be analyzed from Dataset?

- Contribution of 13 domains in a year



- Top 20 visited links by users in a year /across months/across years
- Trends of 13 domains across months/years



Implementation Status

- Common instance for team is created
- Common repository has been setup
- Sample data has been loaded
- Load more data into collection
- Clustering using Apache mahout
- Visualization of data into charts
- Devops- Ansible or cm


imagination at work