Deactivation and Decommissioning Web Log Analysis Using Big Data Technology

Santosh Joshi (Graduate Student Assistant), Dr. Himanshu Upadhyay, Dr. Leonel Lagos
Applied Research Center, Florida International University

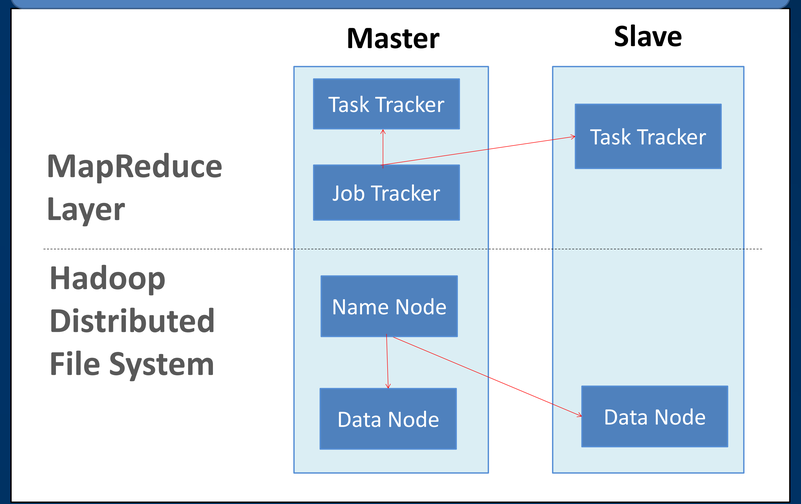
Introduction/Purpose

- The D&D Knowledge Management Information Tool (D&D KM- IT) is a web-based knowledge management information tool built for the D&D user community.
- Big data is a massive volume of structured and unstructured datasets which is so large that it's difficult to process using traditional database techniques.
- Web logs are the repository of files which are generated automatically for any operation on the website.
- Web log files generated from the D&D KM-IT will be processed using the Apache Hadoop Framework to extract meaningful data.

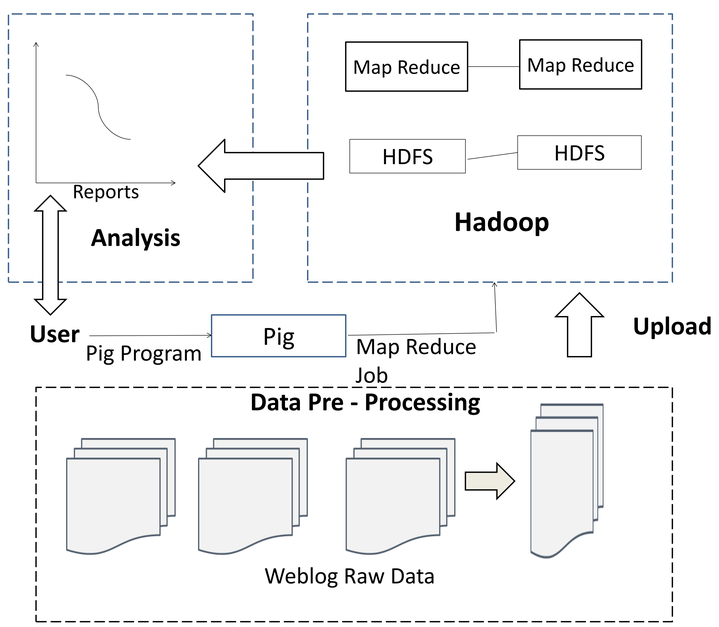
Hadoop

- Hadoop is an open-source software framework for storing and processing big data in a distributed environment on large clusters of commodity hardware.
- Hadoop framework consists of two main layers: Hadoop distributed file system (HDFS) and the Execution engine (MapReduce).
- Automatic parallelization & distribution; clean and simple programming abstraction.

Hadoop Architecture



Implementation Diagram



Implementation

- Web log files from the D&D KM-IT server are fetched using Apache Flume and loaded into the Hadoop distributed file system.
- Using the Apache Hcatalog tool, the web log data is parsed and converted into a structured table format.
- A Pig program is developed to query the tables to retrieve and store the significant information onto the HDFS.
- The extracted data is fed to business visualization tools such as Microsoft Excel for analysis and reporting.

Results

- The reports generated through the web log data processing will be analyzed to improve the usability and performance of the D&D KM-IT.

Conclusion/Future Work

- User keyword information can be explored to add new trending topics to the D&D KM-IT application.
- Statistics generated can be compared to analysis reports from existing analytical tools.