BIG DATA
DEVELOPMENT

Project 1.2

ACADGILD

Project 2.1- State-Wise Development Analysis In India

Table of Contents

1. Executive Summary

1.1 Project Overview

To develop the System to analyze the log data (In XML format) of government progress of various development activities.

1.2 Purpose and Scope of this Specification

The purpose of this project is to capture the data for analyzing the progress of various activities.

In scope

The following requirement will be addressed in phase 1 of Project:

- Developing system to handle the incoming log feed and store the information in Hadoop Cluster (Flume)
- Analyze the data and understand the progress
- Store the results in Hbase/RDBMS

Out of scope
We can use this data and visualization and get more insights

2. Product/Service Description

2.1 Assumptions

Log will be generated in XML format and stored in a server

2.2 Constraints

Describe any item that will constrain the design options, including

- This system may not be used for searching for now. But it will be used for analysis and saving the relevant information as of now
- System will be using Hbase as a database

3. Requirements

- The FLUME job which will format the data and place the data to HDFS
- Pig/MapReduce job for parsing the XML data.
- Create Pig scripts/MapReduce jobs to analyze the data
- Create the Sqoop job to store the data in database

Priority Definitions

The following definitions are intended as a guideline to prioritize requirements.

- Priority 1 – Create FLUME job for fetching log files from spool directory the data
- Priority 2 – MapReduce/pig job to preprocess

Download the dataset using the below link:

Link:

https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing

Refer the below steps to understand the actual steps to create the above project.

Step 1:

Copy dataset from local file system to HDFS using flume.

Note: use the conf file by downloading from below link.

Click here to download

Command:

```
flume-agent agent –n agent1 –c conf –f <path to filecopy.conf>
```

Step 2:

Input file is in the XML format use Map reduce or pig to parse the data and get the results for the below problem statements.

4. Problem statement

1. Find out the districts who achieved 100 percent objective in BPL cards

Export the results to mysql using sqoop

2. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.

Export the results to MySQL using Sqoop.