School of Information Technology and Engineering

B.Tech Information Technology

BIG DATA ANALYTICS USING HADOOP

T Prashanth Reddy

(12BIT0077)

Roop sai Krishna

(12BIT0186)

N.Nagamani

(12BIT0258)

Guide	**Panel member**
Name: Prof.Hari Ram Vishvakarma	**Name: Prof. Menaka S**
Signature:	**Signature:**
Date:	**Date:**

MINI PROJECT FIRST REVIEW REPORT

1. Introduction

1.1 Background

Big Data" refers to datasets whose sizes are beyond the ability of typical software tools to capture, store, manage, and analyze them." The term Big Data Analytics implies methodologies and tools for processing and analyzing the data to produce useful results that cannot be inferred or calculated using other methods in an efficient manner. There are many tools and algorithms available to carry out big data analysis. These methodologies and tools are useful to extract valuable information from large amount of data.

1.2 Problem Statement

Every organization or a company collects large amounts of data from various resources. This data collected is used for organizational support and decision making. But this type of complex data is very difficult to process using traditional data processing applications. It needs advanced methods to process this type of data.

1.3 Importance

Big Data Analytics is very useful for large organizations with tons and tons of data being collected daily. These organizations can implement various big data methodologies and tools to process the enormous amount of data. It can be used by any organization to get valuable information from the data collected which is used for support and decision making.

2. Overview and Planning

2.1 Proposed System Overview

The proposed system consists of a system which acts as a Master and number of other systems which act as Slaves. The Master system controls all the slave systems and it assigns work to the slave systems. The data collected is stored in the slave systems and various methodologies are implemented by the Master system on the slave systems. The slave systems process the data and produce the results which are stored in the Master system. This results are displayed by the Master system which are useful for decision making.

2.2 Challenges

This system has many challenges which include analysis, capture, storage and visualization of data.

2.3 Assumptions

The data you collect comes entirely from the past. We can analyze what happened in the past and try to draw trends between actions and decision points and their consequences, based on the data, and we might use that to guess that under similar circumstances, if a similar decision were made, similar outcomes would occur as a result.

2.4 Architecture Specifications

Hadoop Distributed File System Hadoop is a master/slave based architecture. It primarily consists of a single NameNode, Resource Manager or Yarn, DataNodes and TaskTrackers. In general, the NameNode and Yarn run on the master node where as DataNode and TaskTracker runs on all slave nodes. We also have the provision on creating a separate NameNode and Yarn in case the cluster is huge and load on NameNode is high. NameNode is a master server that manages all the file directories.
NameNode stores metadata related to each file, the metadata mainly consists of file name, location, number of replications etc. NameNode also manages clients and applications access to these files. The DataNode are created on all the slaves usually one per node in the cluster.
DataNodes are responsible for managing storage on the node to which it is attached to. Hadoop Distributed File System exposes a file system namespace and allows user to store their files. Internally, each of these files is divided into several equi-sized blocks, replicated for fixed number of times and then are evenly distributed over the cluster. These blocks are stored in DataNodes with their location stored as a metadata on NameNode. The NameNode executes file system namespace operations like opening files, closing files etc. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode. The system design is unique as it prevents any data from being stored on NameNode.
HDFS Architecture Hadoop's Distributed File System has the traditional hierarchical file organization which is quite similar to most of the existing file systems. A user or application can very well create or delete directories and store files inside it, move file from one director to other or rename it. NameNode maintains a file system namespace, any changes to this namespace is recorded

by NameNode. We can specify number of replicas of a file to be maintained in HDFS, this helps in high fault tolerance and reliability.

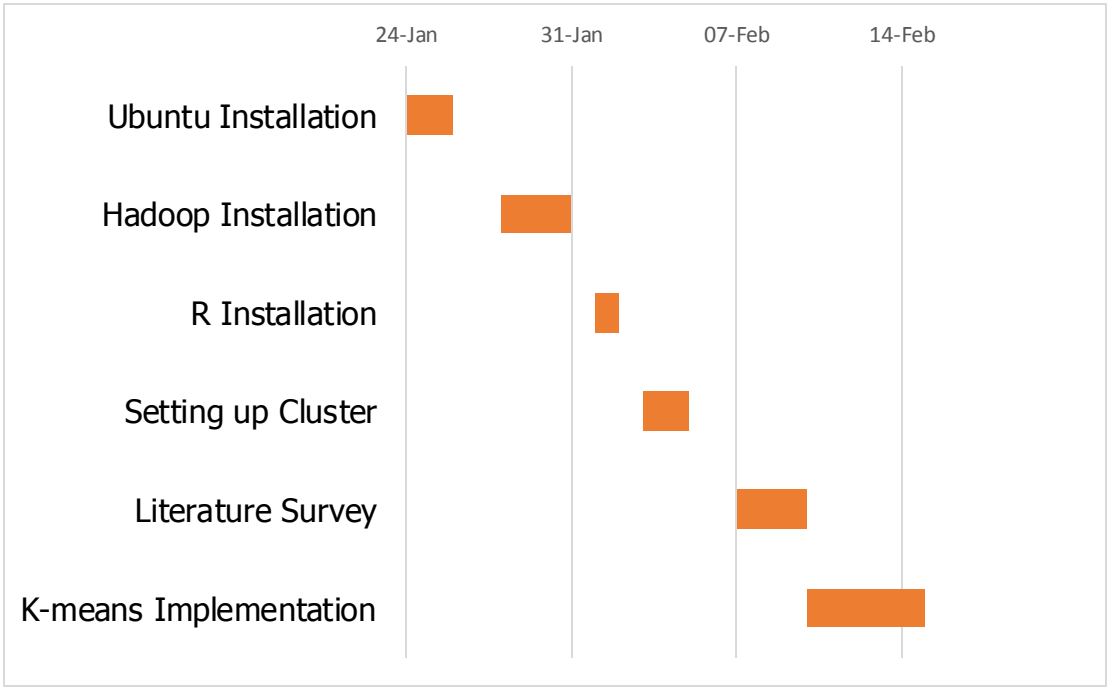
2.5 Hardware Requirements (Optimum requirements)

A cluster of systems with single Master and many slave systems.
Systems must be connected through LAN
2.13 GHz
100 GB of Disk Drive
4 GB

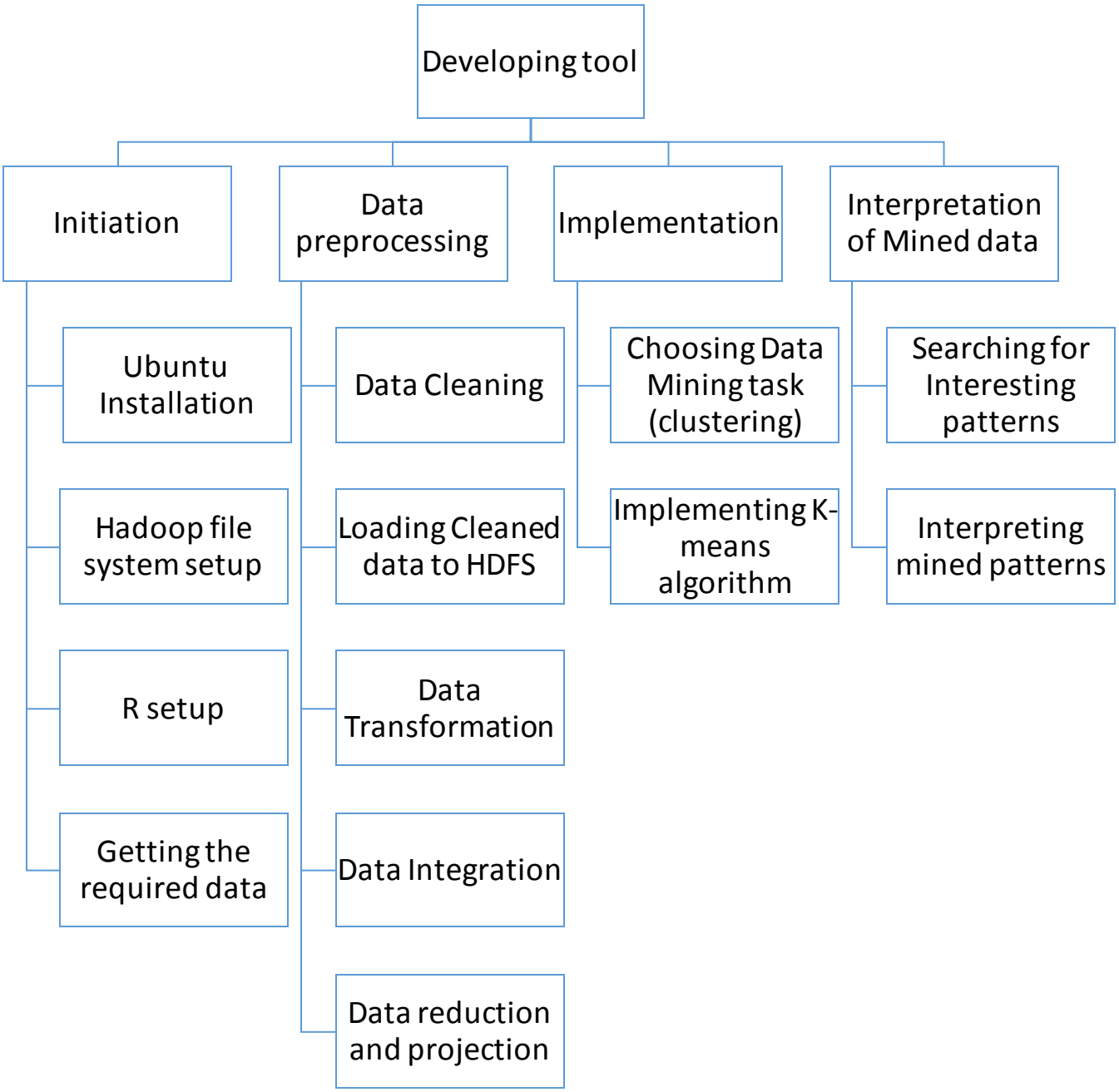
2.6 Software Requirements

Ubuntu Operating system (Linux-based)
Apache Hadoop
R studio
Java JDK 6 or higher

2.7 Project Schedule (Gantt chart)



2.8 Work Breakdown Structure



3. Literature Survey and Review

3.1 Literature Survey

The process of the research into complex data basically concerned with the revealing of hidden patterns. **Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review"** describe the big data content, its scope, methods, samples, advantages and challenges of Data. The critical issue about the Big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. Life sciences etc .By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets. The overall Evaluation describe that the data is increasing and becoming complex. The challenge is not only to collect and manage the data also how to extract the useful information from that collected data. According to the Intel IT Center, there are many challenges related to Big Data which are data growth, data infrastructure, data variety, data visualization, data velocity.

Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk big data analytics with Apache Hadoop" Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model . In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns.

Aditya B. Patel, Manashvi Birla, Ushma Nair (*6-8 Dec. 2012)* "Addressing Big Data Problem Using Hadoop and Map Reduce" reports the experimental work on the Big data problems. It describe the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets.

Kanungo, Tapas ; Mount, D.M. ; Netanyahu, N.S. ; Piatko, C.D. ; Silverman, R. ; Wu, A.Y. ,2002. "An efficient k-means clustering algorithm: analysis and implementation".
This study reveals an efficient implementation of Lloyd's k-means clustering algorithm, called the filtering algorithm. The algorithm is easy to implement and only requires that a k d-tree be built once for the given data points. Efficiency is

achieved because the data points do not vary throughout the computation and, hence, this data structure does not need to be recomputed at each stage.

Zakrzewska, D.; Murlewski, J.Intelligent Systems Design and Applications, 2005. "Clustering Algorithms for Bank Customer Segmentation"

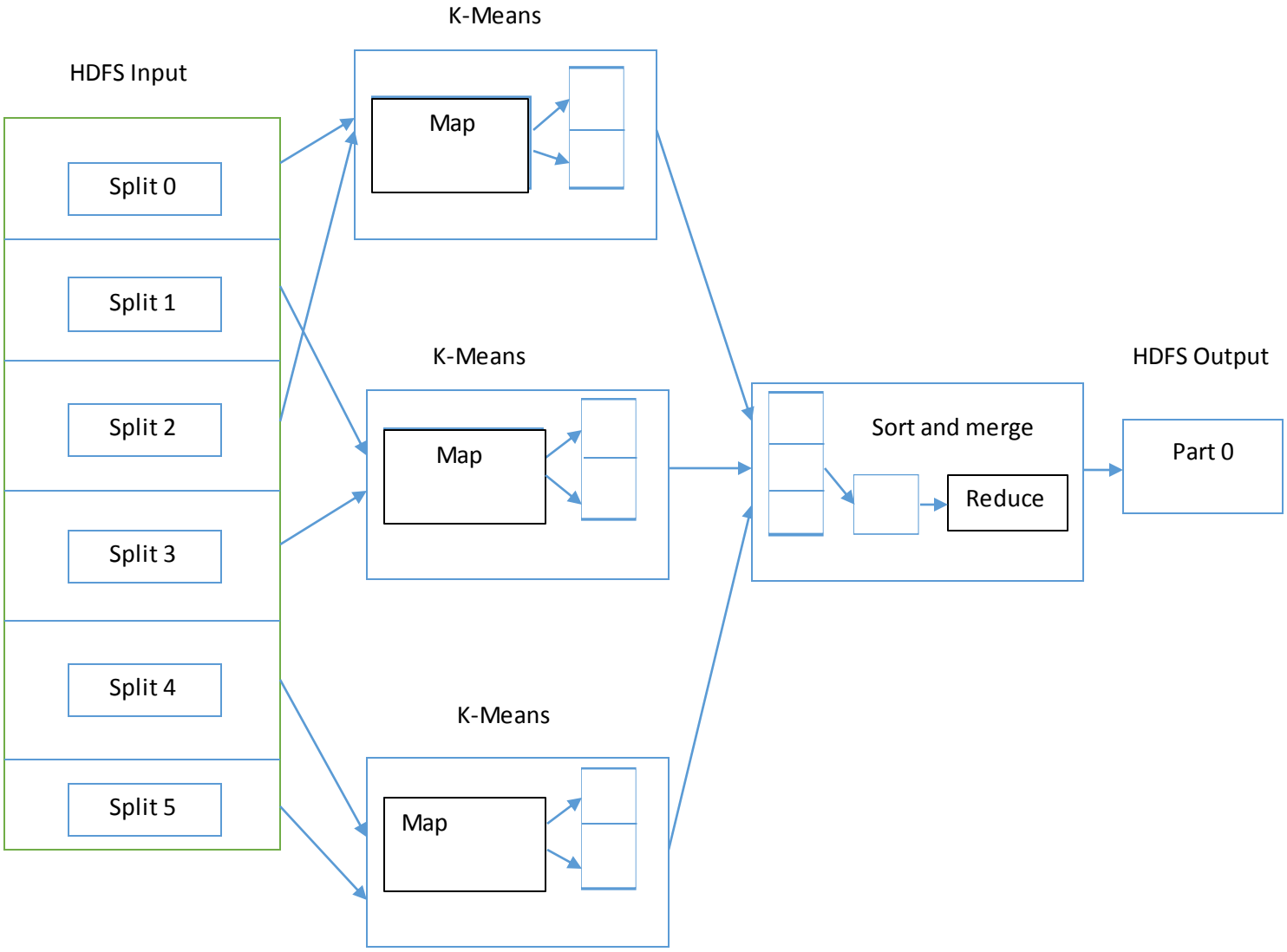
This paper considered three algorithms of cluster analysis: **k-means**, two phase clustering and DBSCAN in bank customer segmentation. The tests showed that all the algorithms have their shortcomings and advantages. **K-means algorithm is very efficient for large multidimensional datasets**, however depends strongly on the choice of input parameter k. It is not recommended in the case of data sets with noise. Two-phase clustering algorithm has a very good performance for data with noise and small amount of dimensions. In DBSCAN algorithm, wrong choice of input parameters, may resulted in a bad quality.

3.2 Literature Summary

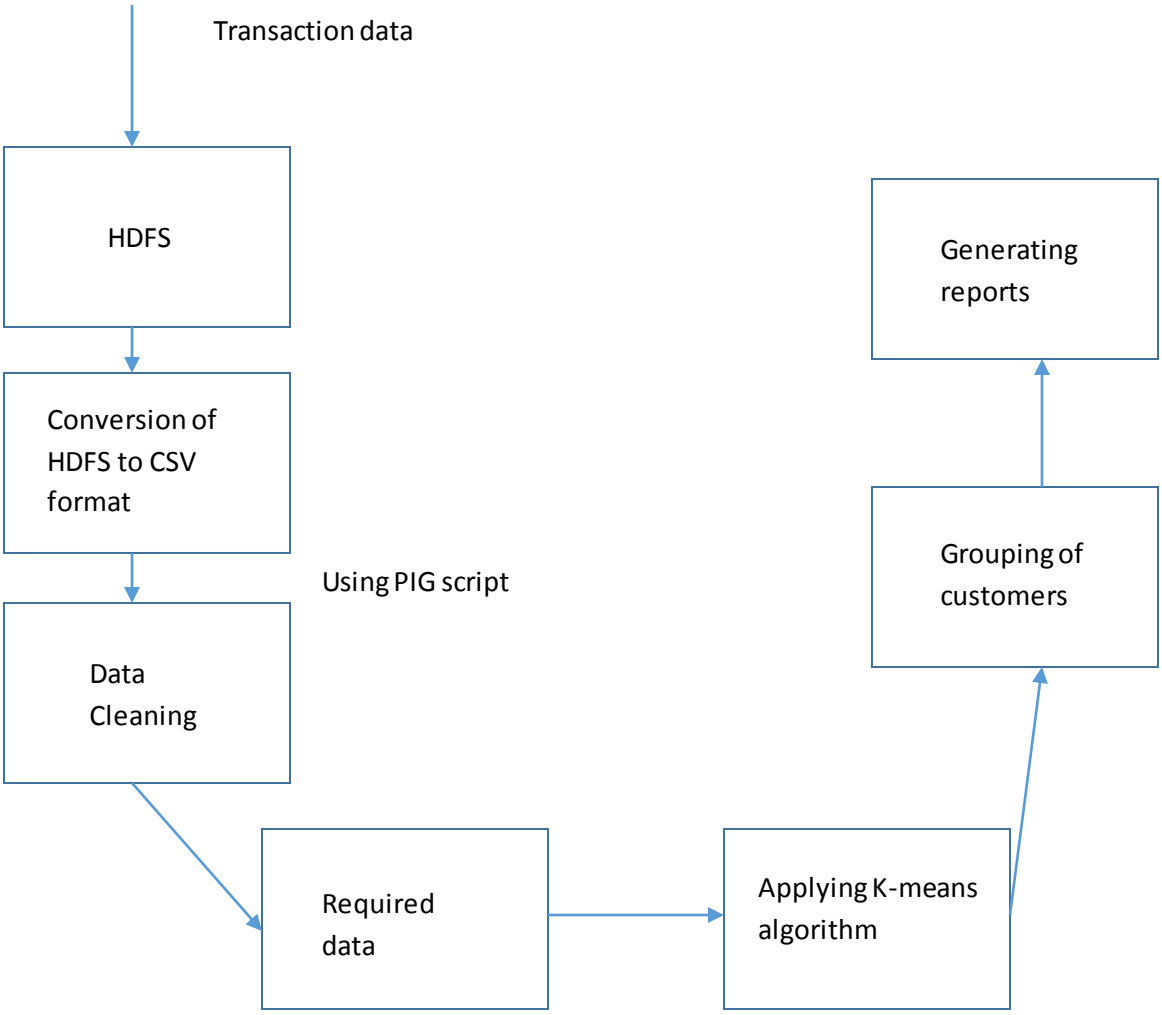
Study reveals that K-Means Algorithm always has K clusters. There is always at least one item in each cluster. The clusters are non-hierarchical and they do not overlap. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters. K-means clustering algorithm is easy to implement and apply on large datasets. It has been successfully used in various fields, including market segmentation, computer vision, geo-statistics, astronomy and agriculture. It often is used as a preprocessing step for other algorithms, for example to find a starting configuration.

4. System Design

4.1 High-Level Design



4.2 Low-Level Design



Web References:

- http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1578784&queryText%3Dk-means+algorithm+in+banking
- http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1017616&matchBoolean%3Dtrue%26queryText%3DAn+efficient+k-means+clustering+analysis+and+implementation
- http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Shared+disk+big+data+analytics+with+Apache+%09Hadoop