WEATHER FORECASTING USING HADOOP AND R

1. AMRUTH SKANDA 2.SHASHI KIRAN S 3.RAMESH MUDALAGI 4.SAMEER

DEPT. OF COMPUTER SCIENCE AND ENGINEERING, RNSIT

ABSTRACT

Weather sensors are generating data at the rate that exceeds our ability to effectively process, manage and analytics it. It is difficult to process weather big data using data management tools or by using traditional data processing technique. Hadoop, an open source software implementation of the map reduce is used to big data analytics and R is open source software for statistical analysis and graphical computing of big data. Big Data Analytics is the process of discovering meaningful patterns in large data, so that information retrieved can be transformed into usable knowledge. Knowledge of weather data in a region is essential for business, society, agriculture and energy. This paper proposes new approach for weather big data analytics using Hadoop and R. By using this technique's we can acquire weather data and we can find the hidden patterns inside the large dataset so as to transfer the retrieved information into usable knowledge for classification and prediction of climate condition.

Keywords: Big Data, HDFS, Hadoop, Map Reduce and R.

1. INTRODUCTION

Big Data true to its name deals with large volumes of data characterized by volume, variety and velocity. Many domains such as weather sensors, bioinformatics, health care, socio-networks and wireless sensor networks communities are seeking exponential growth in data. Big Data Analytics is process extracting useful information from large data and commonly available data analytics tools are unable to scale data with increase in size. Hadoop is new emerging software for large scale data analytics [1] and R for statistical analysis of big data [4]. Steganography become more important as more people join the cyberspace revolution. Steganography is the art of concealing information in ways that prevents the detection of hidden messages. Steganography includes an array of secret communication methods that hide the message from being seen or discovered.

1.1 Big Data

Big Data a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. Big Data sizes are in terms of terabytes, petabytes, hexabytes and zettabytes. A zettabytes is 1021 bytes or equivalently one thousand hexabytes or one million petabytes or one million terabytes [6]. Big data are generating through many sources as business process, transactions, socio networking sites, web servers, etc.

1.2 Hadoop

Hadoop is open-source software for distributed storage and computational capabilities. It is a framework that allows for the distributed processing of large structured, semi structured or unstructured data across a cluster of computers. It is a distributed master-slave architecture that consists of the Hadoop Distributed File System [HDFS] for storage and map reduce algorithm for computational capabilities. It is easily accessible and scalable on large cluster commodity machines or on cloud computing services such as Amazon's Elastic Compute Cloud.

1.3 Map Reduce

Map Reduce is a batch based distributed data processing and computing framework model. Its greatest advantage is the parallel scaling of raw data over multiple computing nodes. Map Reduce consists of two data processing and computing primitives are called mappers and reducers. It is parallel scaling the application to run over hundreds or thousands of computers in a cluster.Fig.1 shows the map reduce processing.

1.3.1 Mapper Step

Mapper Step is filter and transforms the input

Map (k1, v1) → list (K2, v2)

1.3.2 Reducer Step

Reducer Step is aggregate over the mapper output.
Reduce (K2, list (v2)) → list (v3)

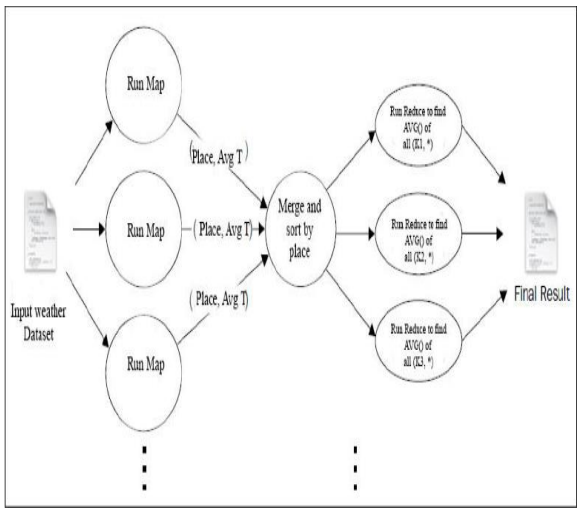


Figure: Proposed MapReduce Framework

1.4 R

R is an environment for statistical analysis and graphical Computing [4]. It's open source software.

1.4.1 Advantages of R.

- R is free! Other statistical software platforms Cost thousands of dollars.
- R providing all manner of data analytic techniques.
- R is a powerful platform for interactive data analytics and exploration.
- R contains advanced statistical routines packages, they not yet available in other statistical software.
- R runs on a wide variety of platforms like UNIX,Windows, and Mac OS X.

1.5 Weather Data.

Weather sensors are located across the globe and collecting weather data every hour [3]. Weather data provided by different sensors are unclean. It contains information about weather station identifier, observation date, observation time, latitude, longitude, wind direction, air temperature, dew point temperature, atmospheric pressure, etc.

The rest of this paper is organized as follows. We discuss System Architecture in Section 2. We provide back ground on our Experimental Setup in Section 3. We present our results and discussion in Section 4. Conclusion and future work are presented in Section 5.

2. SYSTEM ARCHITECTURE

Fig.2 shows the HDFS architecture. HDFS is a distributed File System for storing and processing very large files that are hundreds of megebytes, gigabytes or terabytes in size and designed to running on clusters of commodity hardware [2].

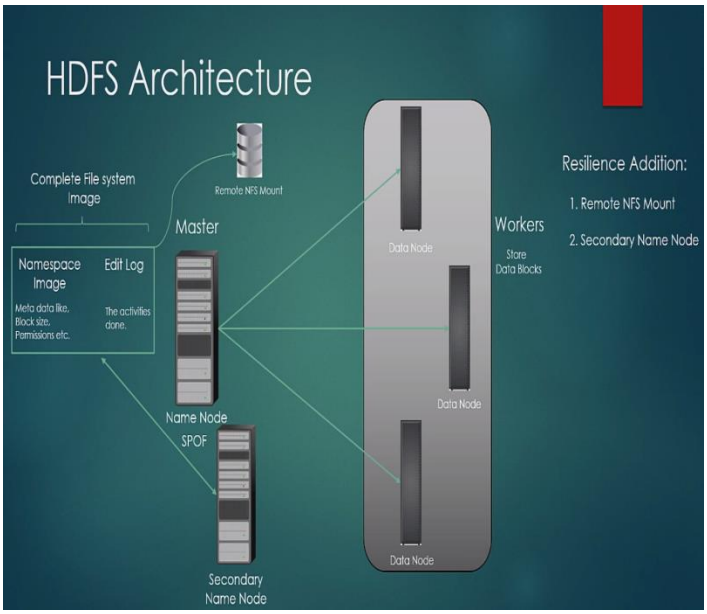


Figure: HDFS Architecture

2.1 Hadoop Distributed File System Architecture.

HDFS architecture consists of the following.

2.1.1 Name Node
2.1.2 Data Nodes
2.1.3. Job Tracker.
2.1.4 Task Tracker
2.1.5 Secondary Name Node

2.1.1 Name Node

Hadoop employs Master/Slave architecture for both distributed storage and distributed computation. The Name Node is the master of HDFS that manages the file system Namespace and assigns input/output tasks to slave data node. It breaks the large file into small file blocks, small file blocks are distributed to different data node in a cluster.

2.1.2 Data Nodes

Data Nodes are the work horses of the HDFS. Slave machine in HDFS cluster will host a data node to perform store and retrieve file blocks. Data nodes executing the job assign by the name node and writing result back into name node.

2.1.3. Job Tracker.

Job Tracker runs on a master node of cluster. It tracking the task assign to different data node. If a task fails, the job tracker will automatically re-launch the task to other data node.

2.1.4 Task Tracker

Task Tracker runs on a slave node of cluster and only one task tracker per slave node. Task tracker manages the execution of individual tasks on each slave node.

2.1.5 Secondary Name Node

It communicates with the name node to take snapshots of the HDFS metadata at periodic intervals and there is only one secondary name node on each cluster.

2.1.6 HDFS Advantages

- HDFS is highly fault-tolerant by detecting faults and automatic recovery.
- . HDFS data access through Map Reduce.
- Applications that run on HDFS have large data sets in terms of terabytes to petabytes in size.
- Simple and Robust Processing model.
- Scalability to process large amount of data into hundreds of nodes in a single cluster.

3. EXPERIMENTAL SETUP

For our experiment we configured and deployed Secure Shell (SSH), Hadoop, HDFS and R.

3.1 Java Installation.

Running Hadoop requires java (version 1.6 or higher). You can download the latest JDK from oracle website.

3.2Secure Shell (SSH) configuration.

Hadoop requires SSH access to communicate Master Node and Data Node. SSH utilizes standard public key cryptography to create a pair of key for user verification one public key, one private key. Every Data Node having the public key and the master communicates with the Data node by sending the private key.

3.3 Hadoop Installation.

Hadoop is open source software. Download it from Apache download mirrors and extract the contents of hadoop package to a location of your choice. Configure hadoop-env.sh ,core-site.xml and mapred-site.xml. For our experiment, we setup the hadoop multimode cluster architecture consists of the the following.

- Master: The master node of the cluster and host of the NameNode and JobTracker daemons.
- Backup: The server that hosts the Secondary Name Node daemon.
- Four DataNodes: The slave boxes of the cluster running both Data Node and Task Tracker daemons.

3.1.4 R Installation.

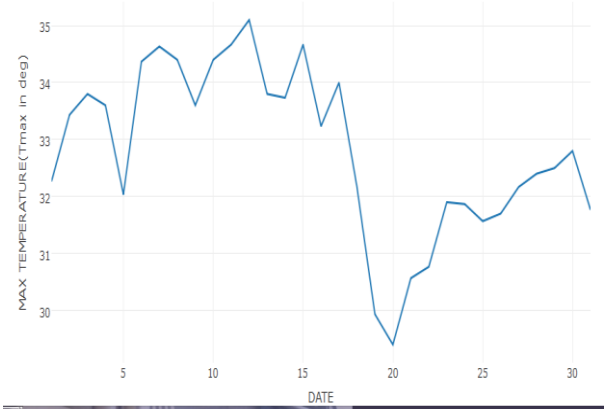
R is open source software, freely available from the Comprehensive R Archive Network. It's using for graphical analysis of our result.

4. RESULTS AND DISCUSSION

In this section, we have analyzed the 66 years weather data published by National Climatic Data Center (NCDC) [3]. Weather big data analytics is classified into following 3 ways.

4.1. Day's wise weather data analytics.

In day's wise weather data analytics we computed maximum temperature of each day for the period 1947-2012 for different locations. following fig. describes day wise maximum temperature computed of May month X- axis denotes days from 01-31 and Y-axis denotes temperature in degree centigrade. The fig 17 describes years (x-axis) versus maximum temperature (yaxis).



5. Conclusion and further work

National Climate Data Center (NCDC) weather sensors are generating weather data every hour. It's difficult to process, manage and analytics weather sensors data. The author explored the solution to weather big data using HDFS, Map Reduce, Hadoop and R. The result obtained from the various analytics is helpful for business, society, agriculture andenergy applications, etc. Future work includes enhancing adaptive and dynamic weather data analytics. Enhancedynamically to match the nature of rapidly changeable weather data and sudden events.

6. REFRENCES

[1] Apache Software Foundation. Official apache Hadoop website, http://hadoop.apache.org/, Aug, 2012.

[2] The Hadoop Architecture and Design http://hadoop.apache.org/common/docs/r0. 16.4/hdfs_design.html, Aug, 2012

[3] National Climatic Data Center (NCDC)

[4] R in Action Data analysis and graphics with R by Robert I. Kabacoff.

[5] Addressing Big Data Problem Using Hadoop and Map Reduce Aditya B. Patel, Manashvi Birla, Ushma Nair.

[6] Hadoop in Action by Chuck Lam

[7] Hung-Chih Yang, Ali Dasdan, Ruey-Lung Hsiao, and D. Stott Parker from Yahoo and UCLA, "Map-Reduce- Merge: Simplified Data Processing on Large Clusters", paper published in Proc. of ACM SIGMOD, pp. 1029– 1040, 2007.

[8] White, Tom. Hadoop The Definitive Guide 2nd Edition. United States : O'Reilly Media, Inc., 2010.

Guide: Dr.G T Raju
HOD,
Dept. of CSE,
RNSIT