GALGOTIA UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Weather Forecasting Using Hadoop

A SynopsisSubmitted
in Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Technology
in
Computer Science and Engineering

by
Akash Ranjan

Under the supervision of
Guide name - Mr. Mukesh Kumar Jha

Year - 2019 Semester - VIII Session - 2016-2020

October, 2019

Declaration

We hereby declare that the project work presented in this report entitled " **Weather Forcasting Using Hadoop ",** in partial fulfillment of the requirement for the award of the degree of BachelorofTechnologyinComputerScience&Engineering,submittedtoGalgotias University Greater Noida, is based on my own work carried out at Department of Computer Science & Engineering, Galgotia University ,Greater Noida. The work contained in the report is original and project work reported in this report has not been submitted by me/us for award of any other degree.

Signature:
NAME – Akash Ranjan
Roll No- 1613101081
Place – Greater Noida

Certificate

This is to certify that the Projec treport entitled " Weather Forcasting Using Hadoop " done by **Akash Ranjan(1613101081)** is an original work carried out by them in Department of Computer Science & Engineering, Galgotias University,GreaterNoidaundermyguidance.Thematterembodiedinthisprojectworkhas notbeensubmittedearlierfortheawardofanydegreeordiplomatothebestofmyknowledge and belief.

GUIDE NAME: Mr Mukesh Kumar Jha

Acknowledgement

The merciful guidance bestowed to us by the almighty made us stick out this project to a successful end. We humbly pray with sincere heart for his guidance to continue forever.

We pay thanks to our project guide Mr. Mukesh Kumar jha who has given guidance and light to us during this project. His/her versatile knowledge has cased us in the critical times during the span of this project.

We pay special thanks to our Panel In charge Mr. Gautaam Kumar who has been always present as a support and help us in all possible way during this project.

We want to thanks our friends who have always encouraged us during this project.

At the last but not least thanks to all the faculty of CSE department who provided valuable suggestions during the period of project.

Abstract

Hadoop an apache product which is an open-source, Java based programming framework is used to support large data sets in a distributed environment. Hadoop has maximum advantage over scalable and fault-tolerant distributed processing technologies. Also,(HDFS)Hadoop Distributed File System is highly fault tolerant and used for applications that have large data sets. Hence HDFS file system using name node, data node and task tracker will perform distribution of job in Hadoop environment .Since Hadoop has overwhelming advantage in optimizing big data , we prefer to use Hadoop to analyze large datasets of weather processing.

TABLE OF CONTENT

References

LIST OF FIGURES

Page No.

Chapter 1

Introduction

1.1 Problem Definition Hadoop is the popular open source implementation of MapReduce, a powerful tool designed for deep analysis and transformation of very large data sets. Hadoop enables you to explore complex data, using custom analyses tailored to your information and questions. Hadoop is the system that allows unstructured data to be distributed across hundreds or thousands of machines forming shared nothing clusters, and the execution of Map/Reduce routines to run on the data in that cluster. Hadoop has its own filesystem which replicates data to multiple nodes to ensure if one node holding data goes down, there are at least 2 other nodes from which to retrieve that piece of information. This protects the data availability from node failure, something which is critical when there are many nodes in a cluster (aka RAID at a server level).

Hadoop has its origins in Apache Nutch, an open source web searchengine, itself a part of the Lucene project. Building a web search engine from scratch was an ambitious goal, for not only is the software required to crawl and index websites complex to write, but it is also a

challenge to run without a dedicated operations team, since there are so many moving parts. It's expensive too: Mike Cafarella and Doug Cutting estimated a system supporting a 1-billion-page index would cost around half a million dollars in hardware, with a monthly running cost of $30,000

Introduction of Hadoop

In a Hadoop cluster, data is distributed to all the nodes of the cluster as it is being loaded in. The Hadoop Distributed File System (HDFS) will split large data files into chunks which are managed by different nodes in the cluster. In addition to this each chunk is replicated across several machines, so that a single machine failure does not result in any data being unavailable. An active monitoring system then re-replicates the data in response to system failures which can result in partial storage. Even though the file chunks are replicated and distributed across several machines, they form a single namespace, so their contents are universally accessible.

1. Background National weather forecasting is well resourced and the subject of much discussion. However it is only able to give indicative forecasts at broad scales and with a relative lack of detail. To improve decision making farmers need to know what the weather will be doing on their properties in 2, 12, 22 or 32 hours time. Access to detailed ‘customised’ weather forecasting was expected to enhance cropping farmers’ ability to optimise use of irrigation, agrochemicals and fertilisers and plan other farm operations. Small weather stations have become increasingly affordable and the ability to transmit, process and share data is improving greatly. A network of stations and access to appropriate crop and disease models was anticipated to help farmers make the best practical and timely decisions on spraying, irrigating, and harvesting crops. Farmers would not all need their own weather stations, as they would be able to obtain property relevant predictions of the time of arrival of weather events or occurrence of infection periods via email, fax or phone messages or internet options. The project investigated the opportunity to co-ordinate a regional weather station network, use computer modelling to predict disease risk, irrigation need,

and crop yields. This would reduce unnecessary expenditure on hardware, reduce weather related risks, increase profitability and save farmers and the environment from wasteful application of chemicals and/or water. We anticipated the outcomes from this project would be able to flow on to
other cropping sectors, and to be extended geographically as well.

2. **Project activities** This project involved co-ordination of three small groups of cropping farmers in Canterbury. Those involved had a commitment to developing much better site specific weather forecasts and taking advantage of related technologies to improve production efficiency and safeguard the environment. An initial small workshop outlined the available technology options and explained how these are being applied in other sectors and regions in NZ. Information and technology providers HortPlus (Trevor Atkins) outlined the existing uses of new weather and related prediction technologies for fruit industries, and potential for arable cropping in Canterbury. A number of growers and industry people described the roles they saw for

such information. Three existing weather stations operated by the Foundation for Arable Research were used to demonstrate the potential of already available forecast opportunities. The three FAR stations have potential to be included in a regional network providing data for on-farm decisions. Detailed site specific forecasts were generated by the Met Service and made available to focus group members by HortPlus MetWatch. An attempt was made to record the locations of other weather stations that were potentially able to be incorporated in a regional network. Crop & Food Research and McCain Foods are known to have additional stations. A number of sites are operated by NIWA and through the National Rural Fire network, however these may not be in suitable locations or collecting sufficient data. Surveys of participants were conducted at the beginning and end of the project to determine the uses to which weather and climate information is currently being used, and to document the benefits that accrue to those acting on the information provided. This also helped identify gaps and prioritise future development goals. Information about weather stations and interpretation was prepared in fact sheet format. The material is predominantly derived from an earlier MAF SFF project, Research to Practice, use and interpretation of weather

information (Manktelow and Porteous). This material is intended to increase the extent of awareness of the potential for these and related technologies. We anticipate again interacting with the wider community through a series of workshops that present the results of the project, and generate ongoing adoption of suitable opportunities. This has not been completed as FAR are reviewing their portfolio of weather and crop model research and services for levy payers.

3. **Weather station locations** 3.1. Weather information requirements The detailed forecasts used in this project are quite site specific. Therefore valued weather recording sites are located in or very near cropping districts. In addition, specific information is required to generate MetWatch forecasts. 1. Minimum sensoring required for MetWatch Air temperature Rainfall Leaf wetness 2. Station Outputs Output needs to include year, plus an output array for daily totals (usually 9am but midnight otherwise) for: Rainfall total Average, min, max air temperature (plus Windrun if wind speed is recorded, and total radiation if recorded)

3.2. Weather stations used Three weather stations were

ultimately used as trial sites for this project – rather than one originally planned. This was assisted by Met Service making free forecasts available for this project. The contribution is gratefully acknowledged. The sites were: 1. Methven at “The Glebe” 2. Chertesy at the “FAR Arable Site” 3. St Andrews at “Copperfields” All three stations are operated by the Foundation for Arable Research.

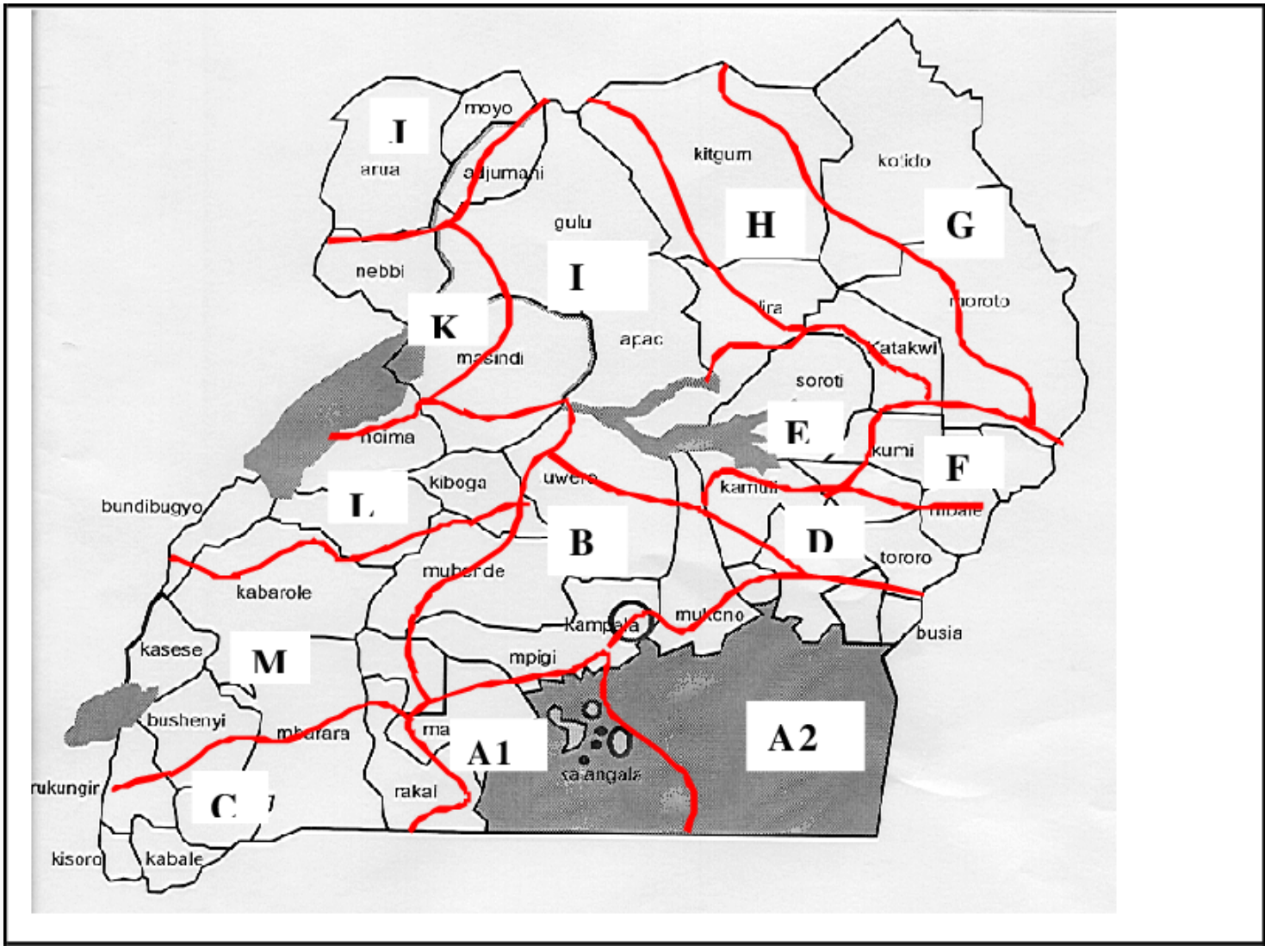
Project weather forecasts A number of problems were encountered during the establishment phase of this project. Early forecasts were notably incorrect – to the extent farmers rang to query the forecasts they received. This problem was resolved quickly, but stayed in the minds of focus group farmers. The cause was an error in co-ordinate conversion. This error mislocated stations (Methven was apparently on Mt Hutt). All stations were some 10km out, and in areas with variable topographical (Methven and St Andrews) this was particularly significant. On several occasions problems remotely accessing the three weather stations appeared to have been resolved but then resurfaced. General station maintenance was completed and stations were reprogrammed to allow 24/7 remote login. Aerials were checked and changed but did not seem to cause the faults. It is possible a download software version conflict may have been the root cause.

Initial user survey A postal survey of the focus group members receiving

daily weather forecasts was undertaken near the start of the project. The survey form is presented in Appendix 7.6.1. Unfortunately only a small number of users responded to the survey, even after repeated approaches.

. Results of Initial Survey Details are presented in Appendix 7.6.2. Weather forecasting was generally regarded ‘very important’ for farm management. Half of the group checked forecasts daily, one quarter twice daily, and the others several times. Information was obtained from TV, radio and internet, and newspapers. Forecast information was used for work planning and for protection from adverse weather such as frost. Spring and summer were predominantly named as the critical periods. Half of the respondents check weather forecasts several times a day at critical times. Responses to ‘how often access was wanted’ were generally the same as ‘actual use’. Wind speed and direction, rain information, maximum and minimum air temperatures and humidity were important to most of the group, while solar radiation was of interest to only 3 respondents. Most farmers measured their rainfall, max and min temperatures and humidity. Less than half measured wind speed or direction and solar radiation. Rainfall was recorded by all but one respondent. Soil temperature and disease pressure were the most highly sought related data (6/7), followed closely by soil moisture, leaf wetness and pest pressure. Ground air temperature was only wanted by three of the farmers. Few were measuring all that they are seeking and even less were recording it.

Map of Arable Relevant Weather Stations .



Weather Data Analysis Using Hadoop: Applications and Challenges .

Abstract. Weather data is very crucial in every aspect of human daily life. It plays an important role in many sectors such as agriculture, tourism, government planning, industry and so on. Weather has a variety of parameters like temperature, pressure, humidity and wind speed. The meteorological department deployed sensors for each weather parameter at different geographical locations to collect data. This data is stored mostly in the unstructured format. Thus, a big amount of data has been collected and archived. Therefore, storage and processing of this big data for accurate weather prediction is a huge challenge. Hadoop an apache product it used to support big data sets in a distributed environment. Hadoop has greatest advantages over scalable and fault-tolerant distributed processing technologies. This paper explains a system that uses the historical weather data of a region and apply the MapReduce and Hadoop techniques to analysis these historical data.

1. Introduction Big Data has become one of the buzzwords in IT, during the last couple of years. Originally it was created by companies which had to manage fast increase rates of data such as web data, data resulting from scientific or business simulations or other data sources. Some of those business companies' models are basically based on indexing and using

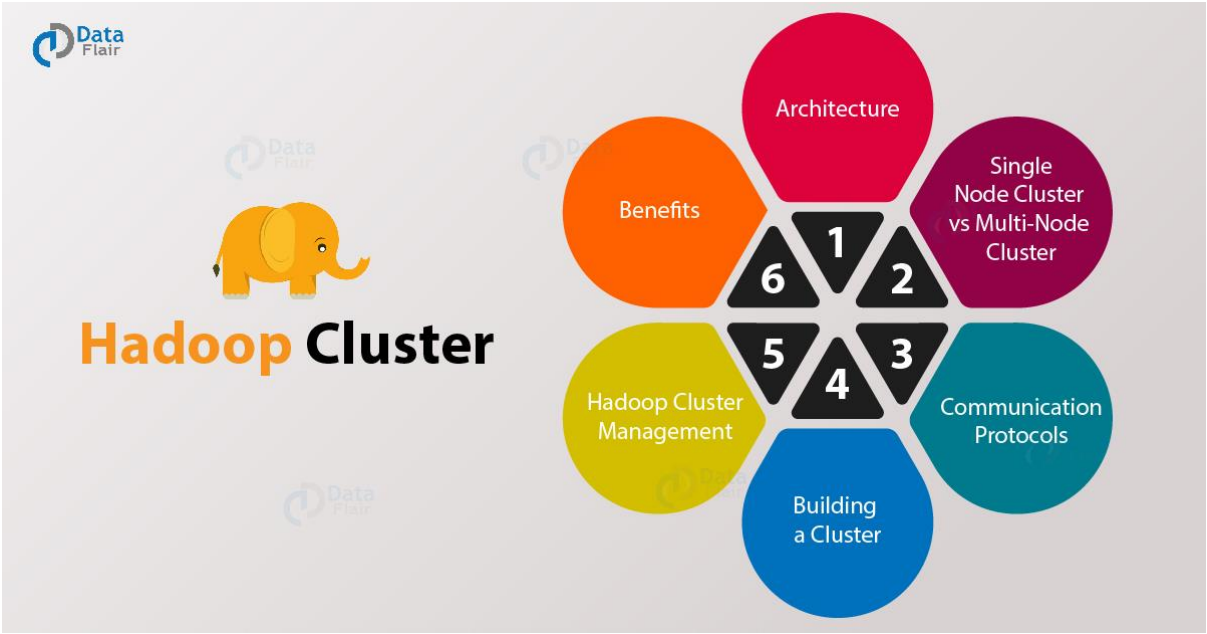
this large amount of data. The challenges to handle the fast growing of data amount on the web e.g. lead Google to develop the Google File System [1] and MapReduce [2]. Furthermore, most of the cities have become smart. Thus, many sensors devices utilized in smart city can be used to measure weather parameters [3]. Which led the weather department collect and analysis huge amount of data like temperature [4]. These different sensors value such as temperature, humidity to predict the rain fall etc. Consequently, when the number of sensors/devices increases, the data becomes high volume and velocity [5]. Therefore, there is an essential of a scalable analytics tool to process massive amount of data. However, the conventional method of process the data is very slow. Compare to process the sensor data with Hadoop framework which remove the scalability bottleneck. Hadoop is a framework used for handling huge amount of data. Mainly the processing engine is MapReduce, which is currently one

2. Related Works [7] describe the analysis of huge amounts of climatic data by using MapReduce with Hadoop. Huge amounts of climatic data

collected, stored and processed for accurate prediction of weather. Climatic data collected by using different types of sensors to store the following parameters temperature, humidity etc. weather datasets collected from National Climatic Data Center (NCDC). Daily Global Weather Measurements 1929-2009 (NCDC, GSOD) dataset is one of the biggest datasets available for weather forecast. Its total size is around 20 GB. Results show that temperature analyzed effectively by Using MapReduce with Hadoop. [6] gives a detailed description of build a platform that is extremely flexible and scalable to be able to analyse petabytes of data across an extremely wide increasing wealth of weather variables. Data processed by Apache Hadoop and Apache Spark. Experiments performed to select the best tools among Hadoop using Pig and Hive Queries. [8] explains the meteorological data storage as well as analysis platform based on Hadoop framework with the help of online logistic regression algorithm for prediction. This platform is based on distributed file system HDFS which

includes distributed database HBase, data warehouse management and useful query processing tool Hive, data migration tool Sqoop. The best data mining prediction algorithm regression also integrated into the system. This architecture has an ability of mass storage of meteorological data, efficient query, and analysis, climate change prediction.

3. . Experimental Setup and Results The experiments were carried out in a physical cluster environment, the researcher used three computers. Hadoop cluster on Linux Ubuntu 14.04 where one computer ran a NameNode and ResourceManager and the remaining ran Datanode and DataManager. Each of the computers has the following configuration: Core i7 processor, 4 GB main memory, and 1 TB disk space as shown in figure 1. The researcher used a Hadoop-2.7.1. The max replication factor “dfs.replication.max” is used to set the replication limit of blocks.



4. The proposed System use dataset of NCDC contains the following parameters: station number, station name, date, country, Precipitation, Temperature, and Wind and so on as shown in figure 2. The data files are stored in HDFS as shown in figure 3. Then, weather files are split and goes to different mappers. The output of each mapper is a set of pairs (key, value) where key is consists of station name, date and value is contains the parameters: Precipitation, Temperature, and Wind. Then the output of mappers is merged and sort by key. Finally, all results sent to the reducers.

5. Conclusion In case of using traditional systems, to process millions of sensors data it is time consuming. Today IoT and the meteorological department uses various of sensors devices to collect data (e.g.

temperature, humidity etc). Hadoop/MapReduce is a framework for processing huge amount in distributable way across large number of computers cluster. Using this framework, the sensors data can be analyzed efficiently. The major benefit of Hadoop framework speeds up the processing of huge data. Where the volume of data is increasing every day.