

Weather Data Analysis

VISHWANATH KODRE¹, SABYASACHI ROY CHOUDHURY¹, AND ABHIJIT THAKRE¹

¹School of Informatics and Computing, Bloomington, IN 47408, U.S.A.

¹Corresponding authors: sabyasachi087@gmail.com, vkodre@gmail.com, athakre@gmail.com

project-000, April 30, 2017

The project aims to analyze any relationship between change in climate, geo- magnetic field and natural disasters with focusing on use of Hadoop Framework for data analysis and Ansible for automating deployment and monitoring.

© 2017 <https://creativecommons.org/licenses/>. The authors verify that the text is not plagiarized.

Keywords: Cloud, I524

<https://github.com/cloudmesh/classes/blob/master/docs/source/format/report/report.pdf>

1. INTRODUCTION

The study of environmental science and climatic changes around has been done for decades, the study has always been predictive based on the past experiences and forecasting of the weather conditions around us. With use of modern days technologies it determining the climatic changes and with analysis done around it has helped human being to prepare and face the natural calamities. Though with current equipment weather department has strengthen their arms but has not been able to be full proof and many time its not been able to predict/ forecast the climatic changes effectively. The study of the whether data and geo graphical changes is ongoing evolving process. Thus more and more researcher needs modern days tools and technologies to leverage it and forecast more accurately.

1.1. Objective

The goal of this is to study the weather data and analyze the relationship between the geo graphical changes such change in geo magnetic field and/or natural disaster. With use of Hadoop for distributed data analysis aims to finds any pattern that might exists between these parameters. The course of the analysis will also provides visualization of these parameters in order to identify any pattern in a more intuitive way. By leveraging the power ansible for application deployment over cluster and monitoring the application performance to determine scalability and throughput. The conclusion will be determine by establishing any existing pattern, analysis done over it and by visualizing it.

2. DATA SOURCES

Weather data has been recorded since 19th century. This data can be used to estimate climate changes and forecasting. The same data can be can be used to find any existing pattern with natural disasters. Following sources has been compiled for weather, natural disaster and geo magnetic fields.

- Weather-Data[1]
- Natural Disaster[2]
- Geo Magnetic Field[3]

3. HIGH LEVEL DESIGN

The design of the application is thought of leveraging power of analysis with deployment on the cluster environment where application units for execution, database for persistence and visualization.

```
\begin{figure}[htbp]
\centering
\fbbox{\includegraphics[width=\linewidth]{images/weather_analysis}}
\caption{Architecture}
\label{Reference:false-color}
\end{figure}
```

The project is divided into following steps:

```
\begin{itemize}
\item Data cleaning and persistence - The raw data cannot be used
\item Core Analysis Program - Core analysis program will be run
set on a given location and duration and compute relationship
MapReduce implementation and is the heart of the application.
Hadoop framework. Hadoop will execute the program in a distributed
\item Deployment and Monitoring - The application needs multiple
i) Deployment and configuration of Hadoop on the multiple nodes
ii) Starting Hadoop servers, inserting/reading data.
iii) Execution of the commands to run the analysis using Hadoop
response to HDFS or some output file.
iv) This output can be then passes to the visualization step
\item Visualization - Finally once the programs completes execution
kit and the output file, graphs and patterns depicting the relationship
\item BenchMarking - The application can be benchmarked for throughput
performance for strong scaling. The report will be represented
```

\end{itemize}

Ansible is open source automation tool. It can be used for deployment. It also serves for monitoring of the state of the application. The script deploys Java, Hadoop, Hbase on the independent cluster.

\subsection{Inventory Configuration}

THIS SECTION IS NATURALLY NOT ACCEPTABLE AS IT MUST BE PROPER REFERENCES.

WE CAN NOT ACCEPT SUCH A SECTION AS IT DOES NOT COMPILE EVEN IN LATEX. IT CONTAINS THE LIST OF HOSTNAME OR NODES THAT CAN BE USED FOR THE CURRENT PROJECT.

\section{Data Curation}

Getting data ready is the very first and basic step for analysis. We have chose NCDC as our source of data. NCDC exposes data in different formats like weatherCluster, etc.

\begin{itemize}

\item I) Datasets : This groups data into monthly daily , yearly , etc. We will be considering only few.

\item II) Data Categories: This groups data into data categories like temperature, pressure, etc. We will consider only few.

\item III) Data Types: This group Data Categories into further smaller sub types. (URL : <https://www.ncdc.noaa.gov/cdo-web/api/v2/locations>)

\item IV) Location Categories: This groups data in terms of location categories like country, state etc. (URL : <https://www.ncdc.noaa.gov/cdo-web/api/v2/locations>)

\item V) Location: Groups data in terms of country, state etc. (URL : <https://www.ncdc.noaa.gov/cdo-web/api/v2/locations>)

\item VI) Station : This collects stations details based on stationId, start_date and end_date. (URL : <https://www.ncdc.noaa.gov/cdo-web/api/v2/stations>)

\item VII) Data : This collects actual weather data for the registrationId, start_date and end_date. (URL : <https://www.ncdc.noaa.gov/cdo-web/api/v2/data>)

\end{itemize}

\subsection{Collecting Data}

NCDC uses token based authentication for security and with each token, more than 10,000 hits are allowed per day. Well, we can use the token to access the data.

\subsection{Technology Stack}

We have considered python , Apache thrift and Apache Hbase for data access and storage. We will be covering a short introduction to each of them.

\subsubsection{Install Apache Thrift \cite{thrift-python-install}}

\begin{itemize}

\item 1) Download Thrift from "<http://redrockdigimark.com/apache-thrift/0.10.0/thrift-0.10.0.tar.gz>".

\item 2) Extract the file and run ./configure.

\item 3) Execute sudo make install or simply make to generate binaries. If using make, the thrift binaries has to put in the path.

\item 4) After thrift is available in path , it can be tested by running "thrift --version".

\item 5) Then the python module has to be created to be used within python. To do this download "<https://github.com/apache/thrift/blob/master/doc/tutorial/python.md>".

\item 6) Then execute "thrift --gen py <path/to/Hbase.thrift>" to generate "gen-py" folder.

\item 7) Add this folder to python path by export PYTHONPATH=<path/to>/gen-py/:\$PYTHONPATH.

\end{itemize}

For details you can check [\href{https://acadgild.com/blog/connecting-hbase-with-python-application-using-thrift-server/}](https://acadgild.com/blog/connecting-hbase-with-python-application-using-thrift-server/)

Now thrift is ready to use. Execute "/hbase thrift start" to start the thrift server. Hbase has to be started separately.

\subsubsection{Install Happybase \cite{happybase-install}}

Happybase is a wrapper program written on thrift to facilitate data access layer in python in a more readable and swift way.

Once done test the python library by running the following test code

\begin{lstlisting}[label=python,caption=Connect-Hbase]

```
import happybase as hbase
```

```
connection = hbase.Connection('localhost')
```

```
print connection.tables()
```

\end{lstlisting}

Once happybase, thrift and hbase is functional, we are ready to download our weather data. Weather data download is divided into two parts.

\begin{itemize}

\item Download Weather Stations - To consume rest services we have used python's inbuilt request response module. Let us see how to use it.

\end{itemize}

\section{Deployment Using Ansible}