A Review: Hadoop Map Reduce In Big Data Analytics
B.Prasanna¹ G.Alamelu² K.Krishnaveni³
¹,²,³Assistant Professor E.M.G Yadava Women’s College, Madurai, Tamilnadu, India
naguprasi6@gmail.com

Abstract- The term ‘Big Data’ describes new techniques and technologies to capture, store, distribute, manage and analyze larger-sized datasets with high-speed and different structures. Big data can be structured, unstructured or semi-structured, resulting in failure of expected data management methods. Data is created from various different sources and can appear in the system at various charges. In order to process these large amounts of data in a low-cost and efficient way, parallelism is used. Big Data is a data whose scale. Diversity and density need new design, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop is the foundation platform for structuring Big Data, and solves the problem of creation it useful for analytics purposes. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of product serves. It is designed to level up from a single server to thousands of machines, with a very high degree of fault tolerance.

Keywords: Big Data, Hadoop, Map Reduce.

I. INTRODUCTION

Big Data:

Big Data is a term that describes large volumes of high speed, complex and data that need advanced techniques and technologies to enable tasks like capture, storage, distribution, management, and analysis of the information. It is a computing infrastructure that can take in, authenticate and analyze high volume of data, and analyzing divers’ data (structured/unstructured) from multiple sources. [8]

Characteristics of Big Data:

Big Data is really critical to handle as it is rising as one of the fastest technologies in current period. The importance of big data is in logical use which can help in generating informative decision to provide better and quick service. The big data has three characteristics, known as data volume, velocity and variety, it is also known as 3Vs [9]

Map Reduce Model

[11] Map-Reduce was created by one of the largest search engine companies – Google for application development on data-centers with thousands of computing nodes. Nowadays, it has been used across a broad range of domains within Google, including [10]

- Large – scale machine learning problems
- Clustering problems for the Google News and Froogle products
- Extracting data to generate reports of popular queries (e.g. Google Zeitgeist and Google Trends)
- Extracting properties of web pages for new experiments and products (e.g. extraction of physical locations from a large amount of Web Pages for localized search)
- Processing of satellite imagery data
- Language model dealing out for statistical machine Translation
- Large-scale graph computations

The programming model can be summarized as follows [10] Floods are common natural disasters in the world. Each year they cause considerable damage to peoples’s lives properties. The system neither directly display the areas affected by flooding, nor show the difference between two flood events. Based on the water levels, it is hard for users to directly determine which houses, roads, and structures will be affected by the predicted flooding. To deal with this problem, it is necessary to visualize the output from hydrological modeling in a Geographic Information System (GIS) [1] Geographical Information System toll is additionally used to provide analyzing the data that helps us to find the areas where floods are likely to occur. [2] The sudden floods that occur rapidly which cause huge destruction to the surroundings. The risk map is generated for the rainfall based on certain parameters. Digital Elevation Model(DEM) is applied to Geographical Information System(GIS) to describe about the rainfall and the various parameters are calculated which helps to get rid of the floods that are about to occur in that particular region. [4] The wet spa model which is the mixture of altitude, soil and the data about the use of land that predicts floods with the parameters of the rainfall. [5] Heavy rainfall that causes floods in rivers and the nearby areas that is in charge of regular day life. An approach is used to automatically view the floods in the particular area by using cyber surveillance systems and image process methods to receive the immediate flooding results [6].

Hadoop

The GMR (Google Map Reduce) was invented by Google back in their earlier days so they could usefully index all the rich textural and structural information they were collecting, and then present meaningful and actionable results to users. Map Reduce (you map the operation out to all of those servers and then you reduce the results back into a single result set), is a software paradigm for processing a large data set in a distributed parallel way. Since Google’s Map Reduce and Google file system (GFS) are proprietary, an open source Google’s Map Reduce platform by using thousands of cluster nodes[7].
**Hadoop cluster architecture:**

A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a Job Tracker, Task Tracker, Name Node and Data Node. A slave or worker node acts as both a Data Node and Task Tracker, though it is possible to have data-only worker nodes and compute-only worker nodes. These are normally used only in nonstandard applications. Hadoop requires Java Runtime Environment (JRE). The standard start-up and shutdown scripts require Secure Shell to be between nodes in the cluster. In a larger cluster, the HDFS is managed through a dedicated Name Node server to host the file system index, and a secondary Name Node that can generate snapshots of the name node’s memory structures, thus preventing file-system corruption and reducing loss of data[4]

II. PROPOSED WORK

In the existing work there is no automatic system to predict the floods in advance and no proper alert system is maintained. In the proposed work, the data must be collected and analyzed for particular area and various attributes such as the rainfall data that has occurred previous years, surface of the soil etc are taken for analysis purpose. Here analysis of risk is based on these attributes to check whether the flood may cross the limit or not. It has been a challenge in predicting the floods that cause damage i.e., mostly in the jam-packed areas. In this paper, we analyze the rainfall datasets based upon the threshold value. The force of flood is exact and alert message is sent to nearby locations which help us to get pre warning of flood molest and is a full preset system.

Architecture:

The user application is created by which the user is allowed to approach the data from the server to begin with, the user must be created and only then they are allowed to access the network. Once the user account is created they are allowed to enter their account to access the application. Depending on the appeal of the user, the server will give the response to the user. All the user detail are stored in server database. The User Interface Frame (UIF) is designed to communicate with the server through the network coding. The server will test the entire user information in their own database and validate them when they are mandatory. Also the Server will store the complete information of the users in their database and establish the connection to work together with the Users.

The server will update the each User’s activities in its database. The server will validate each user before they access the Application which helps to avoid the Unauthorized User from accessing the Application. As the air level rises it becomes cool and will not be able to have the moisture content as it could when it was warmer. Eventually the air that rises to a certain point where it is completely saturated it cannot hold any more water. Based on the user query it goes and fetch with database and bring information about the user query which is useful for detection and analyzing the rainfall. Once the flood is detected it gives an alert to the user so that they can easily rescue from the rain flood detection.

Data and Methodology

The rainfall dataset is collected for the particular area where the analysis process must be performed to figure out whether there are possibilities of flood occurrence in that region. Analysis of risk is based on various attributes to determine whether the floods may cross the limit or not. Map Reduce technique is used to present the exact reduced data from the huge quantity of input datasets. This algorithm divides the input datasets into separate pieces of data which are further processed simultaneously to produce the desired reduced data.
III. CONCLUSION

The main focus of this paper is to show up some Map Reduce implementation worked well to complete a specific purpose and compared with previously available frameworks. A notable presentation development over the existing system seems after comparison.

Floods which are one of the irremediable calamities destroy the duration of life for many people who are situated around. The certain measures for safety must be considered for floods that mostly occur in the highly jam-packed areas. The Map Reduce techniques are well suited for springy the exact reduced data from the vast quantity of rainfall datasets which are given as input.

The aim of this paper was giving an overview of the Map Reduce concepts in big data analytics. Map Reduce was developed by Google to handle big data analysis which is unstructured data such as web related data. But, it also can be used for structured data. We have discussed a number of Map Reduce models still researchers can develop a more efficient Map Reduce with improved functionalities.

REFERENCE

[1]. EARLY WARNING AND MAPPING FOR FLOOD DISASTERS
[6]. Shi-Wei Lo. “Cyber surveillance for flood disasters” open access sensors 2015, 15,2369-2387.