Survey Using Big Data Tools for Doing Rainfall Prediction

A. Saranya, R. Anandan

***Abstract---** Big data collect large volume of data and gives the analysis of the data obtained using the Hadoop Framework in a fast and more efficient manner which is very accurate and makes our work very easy. Rainfall data is collected from the data set obtained and using Sqoop tool we get the data from MySql to HDFS architecture. The HDFS is mainly the database of Hadoop architecture which stores the data and distributes it to the various tools of Hadoop. It performs the accumulation of data in a way which makes decision analysis for the final output easy to obtain. Hence in our project we are focusing on getting the data from the data set and storing it in HDFS to get the analysis by Hadoop framework using Hive, Map Reduce and Pig to get the output of result which consists of all the analysis of rainfall in a city for all the years and give a clear perspective of the state of rainfall in the city at any moment of the year and also the analysis is shown through bar graph and pie chart which makes the understanding of analysis a little easier and it brings a little more significance to our work.*

***Keywords---** Hadoop, MySql, HDFS, Hive, MapReduce, Pig, Sqoop.*

I. INTRODUCTION

Big Data is a concept which defines large amount of data which can be structured or unstructured. It is widely used as the amounts of records in a database are very high which can only be handled by a machine and not by human. It provides the platform which makes the handling of such large data much easier, the processing done using big data is much faster and easier. Plus there is no limit in the size of data which can be there for processing. Big Data works on the concept of three V's Velocity, Volume, Variety which clearly highlight the importance of Big Data.

Hadoop is a framework which provides a platform to work on Big Data. The Big Data tools Hive, Pig and Map-Reduce all are used to get the best analysis of the data no matter how large the size of the data is. The framework allows us to get of the data analysis by all of the three tools. The sqqop interacts with the database MySql and it fetches the data to the HDFS from where on all the data gets distributed to all the tools.

The tools get the data from the HDFS and the data is broken into packets or splitted so a separate analysis can be done of the Dataset. Then all the data is integrated and a combined analysis of the data is done. All the three tools use there analysis mechanisms to get the best possible output and the mechanism used to operate the data are either by query or programming methodology. The final output is very helpful with all the analysis. It is efficient and reliable which handles the data even if there is power loss and data consistency is maintained.

II. SURVEY DETAILS

In paper[1] Rainfall information is gathered to anticipate the tempest notices from the hydrological information. This is considered as research thought as it expends immense number of records from the circulated framework. The paper depicts a answer for dealing with the data in light of spatial transient qualities utilizing Map Reduce Framework. The workload is characterized utilizing Support Vector Machine (SVM). It utilizes include determination and decrease calculation related with the dataset.

In paper[2] This manuscript was motivated by Congress where amid a talk of tempest water BMP plan, two inquiries were postured: 1. What precipitation force ought to be utilized for such plan? 2. To what extent (amid a tempest) should a BMP play out its water quality capacity? A more thorough inquiry is "All things considered, how hard would it be able to rain?" This paper tends to this last inquiry and sets up a system to address the other two inquiries. In this composition the experimental hyetograph that relates profundity and length, and along these lines whether a tempest is front stacked, back stacked.

In paper[3] Rainstorm Prediction using Support Vector Machine in Hadoop Cluste Precipitation information is gathered to foresee the tempest notices from the hydrological information. This is considered as an examination thought as it devours colossal number of records from the appropriated framework. This paper depicts a novel answer for deal with the information in view of spatial transient attributes utilizing a Map Reduce Framework. The workload is grouped utilizing bolster vector machine (SVM).

In paper[4] Precipitation profundities for different spans and frequencies, alluded to as profundity length recurrence (DDF) in this report, have numerous employments. A typical utilization of DDF is for the plan of structures that control and course restricted spillover, for example, parking areas, storm channels, and ducts. Another utilization of DDF is to drive waterway stream models that join precipitation attributes. Exact DDF gauges are vital for practical basic plans at stream intersections and for creating solid surge forecast models.

In paper[5] In the course of recent years, the creators and numerous others at Google have executed several unique reason calculations that procedure a lot of crude information, for example, crept records, web ask for logs,

Manuscript received February 01, 2019

A. Saranya, Research Scholar, Department of Computer Science & Engineering, VELS Institute of Science and Technology.

Assistant Professor (O.G), Department of Software Engineering, SRM Institute of Science & Technology, (e-mail: aksaranya@gmail.com)

Dr.R. Anandan, Professor, Department of Computer Science & Engineering, VELS Institute of Science and Technology. (e-mail: anandan.se@velsuniv.ac.in)

and so on., to figure different sorts of determined information, for example, upset lists, different portrayals of the chart structure of web reports, synopses of the quantity of pages slithered per have, the arrangement of most incessant questions in a given day, and so on. Most such calculations are reasonably clear.

In paper[6] Most precipitation information is put away in positions that are difficult to break down and mine. In these configurations, the measure of information is colossal. In this paper, they propose procedures to compress the crude precipitation information into a model that encourages storm investigation and mining, and diminishes the information estimate. The outcome is to change over crude precipitation information into significant tempest driven information, which is then put away in a social database for simple investigation and mining.

III. PROPOSED METHODOLOGY

In this project we are reading rainfall facts by way of using Hadoop device at the side of a few Hadoop ecosystems like HDFS, map reduce, sqoop, hive and pig. Maximum rainfall data is saved in formats which is not easy to understand and refine. The data volume is very high in the codes provided, by using the usage of these tools we are able to technique no hassle of information, no information misplaced problem, we can get excessive throughput, maintenance value additionally very less and it's miles an open source software, it is compatible on all of the systems given that it is java based. the output obtained by using processing huge chunks of information thru Hadoop structure is very useful for analyzing and predicting purposes.

Data Source

Dataset is obtained from the net which acts as the primary source of data. The data input is the input given by the client to the administrator which refines the data from the dataset using the Hadoop architecture.

IV. SYSTEM ARCHITECTURE

When the rainfall data is provided initially stored in data base by using sqoop the stored in database is fetched and it is given to hadoop distributed file system. The data will be provided to different bigdata tools like hive, pig and mapreduce.

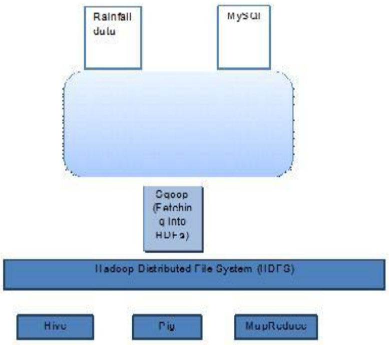

Rainfall data
MySQL
Sqoop (Fetching into HDFs)
Hadoop Distributed File System (HDFS)
Hive
Pig
MapReduce

Fig. 1: Architecture diagram

V. EXPERIMENTAL SETUP

Initially the data source which I took for prediction is sent to sqoop for importing into HDFS. Then the data is being accessed and by using appropriate data analytic tools analysis is carried out and the output is generated.

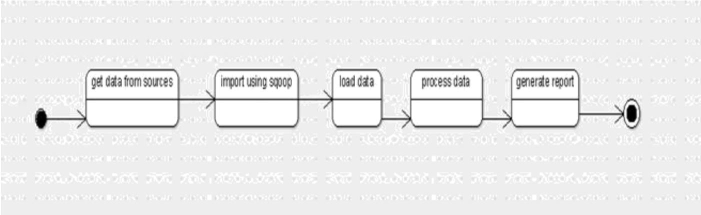
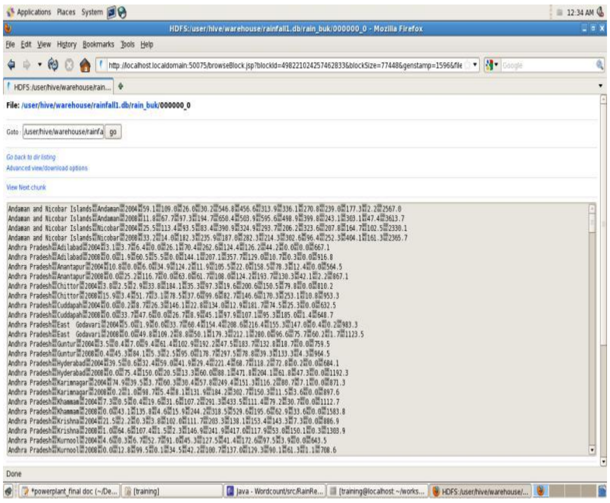


Fig. 2: Information moved from first stage to report stage

VI. Experimental Results & Analysis

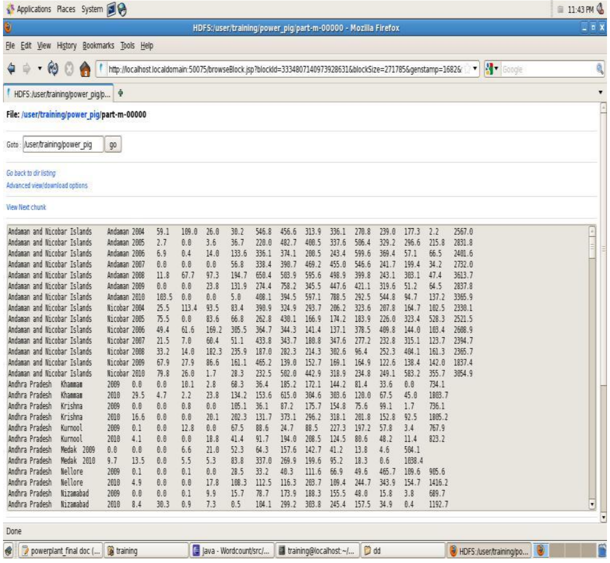
The data set consist of monthly, weekly, hourly based rain information of particular state. In that state the data is divided in to region wise and for all the region the information is provided in above informed manner. First fig show the unstructured data then second fig shows converted unstructured to structured data


Applications Places System
12:34 AM
HDFS:/user/hive/warehouse/rainfall1.db/rain_buk/000000_0 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
HDFS:/user/hive/warehouse/rain...
Goto : /user/hive/warehouse/rainfa
go
Go back to dir listing
Advanced view/download options
View Next chunk
Done
*powerplant_final doc (~/De...
[training]
Java - Wordcount/src/RainRe...
[training@localhost ~/works...
HDFS:/user/hive/warehouse/...

File: /user/hive/warehouse/rainfall1.db/rain_buk/000000_0

```
Andaman and Nicobar Islands␁Andaman␁2004␁59.1␁109.0␁26.0␁30.2␁546.8␁456.6␁313.9␁336.1␁270.8␁239.0␁177.3␁2.2␁2567.0
Andaman and Nicobar Islands␁Andaman␁2008␁11.8␁67.7␁97.3␁194.7␁650.4␁503.9␁595.6␁498.9␁399.8␁243.1␁303.1␁47.4␁3613.7
Andaman and Nicobar Islands␁Nicobar␁2004␁25.5␁113.4␁93.5␁83.4␁390.9␁324.9␁293.7␁206.2␁323.6␁207.8␁164.7␁102.5␁2330.1
Andaman and Nicobar Islands␁Nicobar␁2008␁33.2␁14.0␁182.3␁235.9␁187.0␁282.3␁214.3␁302.6␁96.4␁252.3␁404.1␁161.3␁2365.7
Andhra Pradesh␁Adilabad␁2004␁3.1␁3.7␁6.4␁0.0␁26.1␁70.4␁262.6␁124.4␁126.2␁44.2␁0.0␁0.0␁667.1
Andhra Pradesh␁Adilabad␁2008␁0.0␁1.9␁60.5␁5.5␁0.0␁144.1␁207.1␁357.7␁129.0␁10.7␁0.3␁0.0␁916.8
Andhra Pradesh␁Anantapur␁2004␁10.8␁0.0␁6.0␁34.9␁124.2␁11.9␁105.5␁22.0␁158.5␁78.3␁12.4␁0.0␁564.5
Andhra Pradesh␁Anantapur␁2008␁0.0␁25.2␁116.7␁0.0␁63.0␁61.7␁108.0␁124.2␁193.7␁130.3␁42.1␁2.2␁867.1
Andhra Pradesh␁Chittor␁2004␁3.8␁2.5␁2.9␁33.8␁184.1␁35.3␁97.3␁19.6␁200.6␁150.5␁79.8␁0.0␁810.2
Andhra Pradesh␁Chittor␁2008␁15.9␁3.4␁51.7␁3.1␁78.5␁37.6␁99.6␁82.7␁146.6␁170.3␁253.1␁10.8␁953.3
Andhra Pradesh␁Cuddapah␁2004␁0.0␁0.2␁8.7␁26.3␁146.1␁22.8␁134.0␁12.9␁181.7␁74.5␁25.3␁0.0␁632.5
Andhra Pradesh␁Cuddapah␁2008␁0.0␁33.7␁47.6␁0.0␁26.7␁8.9␁45.1␁97.9␁107.1␁95.3␁185.0␁1.4␁648.7
Andhra Pradesh␁East  Godavari␁2004␁5.0␁1.9␁0.0␁33.7␁60.4␁154.4␁208.6␁216.4␁155.3␁147.0␁0.4␁0.2␁983.3
Andhra Pradesh␁East  Godavari␁2008␁0.0␁49.8␁109.2␁8.8␁50.1␁179.3␁212.1␁280.0␁96.6␁75.7␁60.2␁1.7␁1123.5
Andhra Pradesh␁Guntur␁2004␁3.5␁0.4␁7.0␁9.4␁61.4␁102.9␁192.2␁47.5␁183.7␁132.8␁18.7␁0.0␁759.5
Andhra Pradesh␁Guntur␁2008␁0.4␁45.3␁84.1␁5.3␁2.5␁95.0␁178.7␁297.5␁78.8␁39.3␁133.3␁4.3␁964.5
Andhra Pradesh␁Hyderabad␁2004␁39.5␁0.6␁32.4␁59.0␁41.9␁29.4␁221.4␁68.7␁118.2␁72.8␁0.2␁0.0␁684.1
Andhra Pradesh␁Hyderabad␁2008␁0.0␁75.4␁150.0␁20.5␁13.3␁60.0␁88.1␁471.8␁204.1␁61.8␁47.3␁0.0␁1192.3
Andhra Pradesh␁Karimnagar␁2004␁74.9␁39.5␁3.7␁60.3␁30.4␁57.8␁249.4␁151.3␁116.2␁80.7␁7.1␁0.0␁871.3
Andhra Pradesh␁Karimnagar␁2008␁0.2␁1.0␁98.7␁5.4␁8.1␁131.9␁184.2␁302.7␁150.3␁11.5␁3.6␁0.0␁897.6
Andhra Pradesh␁Khammam␁2004␁7.3␁0.5␁0.4␁19.6␁31.6␁107.2␁291.3␁433.5␁111.4␁79.2␁30.7␁0.0␁1112.7
Andhra Pradesh␁Khammam␁2008␁0.0␁43.1␁135.8␁4.6␁15.9␁244.2␁318.5␁529.6␁195.6␁62.9␁33.6␁0.0␁1583.8
Andhra Pradesh␁Krishna␁2004␁21.5␁2.2␁0.3␁3.8␁102.0␁111.7␁203.3␁138.1␁153.4␁143.3␁7.3␁0.0␁886.9
Andhra Pradesh␁Krishna␁2008␁1.0␁64.6␁107.4␁1.5␁2.3␁146.9␁241.9␁417.0␁117.9␁53.0␁150.1␁0.3␁1303.9
Andhra Pradesh␁Kurnool␁2004␁4.6␁0.3␁6.7␁52.7␁91.0␁45.3␁127.5␁41.4␁172.6␁97.5␁3.9␁0.0␁643.5
Andhra Pradesh␁Kurnool␁2008␁0.0␁12.8␁99.5␁0.1␁34.5␁42.2␁100.7␁137.0␁129.3␁90.1␁61.3␁1.1␁708.6
```

Fig. 3: Unstructured Data


Applications Places System
11:43 PM
HDFS:/user/training/power_pig/part-m-00000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
HDFS:/user/training/power_pig/p...
Goto : /user/training/power_pig
go
Go back to dir listing
Advanced view/download options
View Next chunk
Done
powerplant_final doc (...
training
Java - Wordcount/src/...
training@localhost:~/...
dd
HDFS:/user/training/po...

File: /user/training/power_pig/part-m-00000

```
Andaman and Nicobar Islands	Andaman 2004	59.1	109.0	26.0	30.2	546.8	456.6	313.9	336.1	270.8	239.0	177.3	2.2	2567.0
Andaman and Nicobar Islands	Andaman 2005	2.7	0.0	3.6	36.7	220.0	482.7	400.5	337.6	506.4	329.2	296.6	215.8	2831.8
Andaman and Nicobar Islands	Andaman 2006	6.9	0.4	14.0	133.6	336.1	374.1	200.5	243.4	599.6	369.4	57.1	66.5	2401.6
Andaman and Nicobar Islands	Andaman 2007	0.0	0.0	0.0	56.8	338.4	390.7	469.2	455.0	546.6	241.7	199.4	34.2	2732.0
Andaman and Nicobar Islands	Andaman 2008	11.8	67.7	97.3	194.7	650.4	503.9	595.6	498.9	399.8	243.1	303.1	47.4	3613.7
Andaman and Nicobar Islands	Andaman 2009	0.0	0.0	23.8	131.9	274.4	758.2	345.5	447.6	421.1	319.6	51.2	64.5	2837.8
Andaman and Nicobar Islands	Andaman 2010	103.5	0.0	0.0	5.0	408.1	394.5	597.1	788.5	292.5	544.8	94.7	137.2	3365.9
Andaman and Nicobar Islands	Nicobar 2004	25.5	113.4	93.5	83.4	390.9	324.9	293.7	206.2	323.6	207.8	164.7	102.5	2330.1
Andaman and Nicobar Islands	Nicobar 2005	75.5	0.0	83.6	66.8	262.8	430.1	166.9	174.2	183.9	226.0	323.4	528.3	2521.5
Andaman and Nicobar Islands	Nicobar 2006	49.4	61.6	169.2	305.5	364.7	344.3	141.4	137.1	378.5	409.8	144.0	103.4	2608.9
Andaman and Nicobar Islands	Nicobar 2007	21.5	7.0	60.4	51.1	433.8	343.7	180.8	347.6	277.2	232.8	315.1	123.7	2394.7
Andaman and Nicobar Islands	Nicobar 2008	33.2	14.0	182.3	235.9	187.0	282.3	214.3	302.6	96.4	252.3	404.1	161.3	2365.7
Andaman and Nicobar Islands	Nicobar 2009	67.9	27.9	86.6	161.1	465.2	139.0	152.7	169.1	164.9	122.6	138.4	142.0	1837.4
Andaman and Nicobar Islands	Nicobar 2010	79.8	26.0	1.7	28.3	232.5	502.0	442.9	318.9	234.8	249.1	583.2	355.7	3054.9
Andhra Pradesh	Khammam	2009	0.0	0.0	10.1	2.8	68.3	36.4	185.2	172.1	144.2	81.4	33.6	0.0	734.1
Andhra Pradesh	Khammam	2010	29.5	4.7	2.2	23.8	134.2	153.6	615.0	304.6	303.6	120.0	67.5	45.0	1803.7
Andhra Pradesh	Krishna	2009	0.0	0.0	0.8	0.0	105.1	36.1	87.2	175.7	154.8	75.6	99.1	1.7	736.1
Andhra Pradesh	Krishna	2010	16.6	0.0	0.0	20.1	202.3	131.7	373.1	296.2	318.1	201.8	152.8	92.5	1805.2
Andhra Pradesh	Kurnool	2009	0.1	0.0	12.8	0.0	67.5	88.6	24.7	88.5	227.3	197.2	57.8	3.4	767.9
Andhra Pradesh	Kurnool	2010	4.1	0.0	0.0	18.8	41.4	91.7	194.0	208.5	124.5	80.6	48.2	11.4	823.2
Andhra Pradesh	Medak	2009	0.0	0.0	0.0	6.6	21.0	52.3	64.3	157.6	142.7	41.2	13.8	4.6	504.1
Andhra Pradesh	Medak	2010	9.7	13.5	0.0	5.5	5.3	83.8	337.0	269.9	199.6	95.2	18.3	0.6	1038.4
Andhra Pradesh	Nellore	2009	0.1	0.0	0.1	0.0	28.5	33.2	40.3	111.6	66.9	49.6	465.7	109.6	905.6
Andhra Pradesh	Nellore	2010	4.9	0.0	0.0	17.8	108.3	112.5	116.3	203.7	109.4	244.7	343.9	154.7	1416.2
Andhra Pradesh	Nizamabad	2009	0.0	0.0	0.1	9.9	15.7	78.7	173.9	188.3	155.5	48.0	15.8	3.8	689.7
Andhra Pradesh	Nizamabad	2010	8.4	30.3	0.9	7.3	0.5	104.1	299.2	303.8	245.4	157.5	34.9	0.4	1192.7
```

Fig. 4: Structured data

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

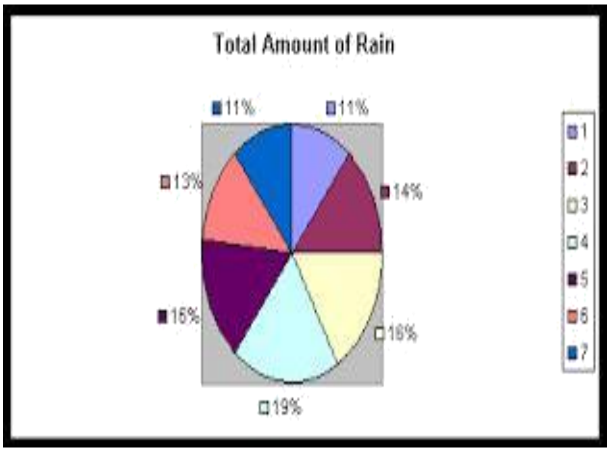


Fig. 5: Results (Theoretical)

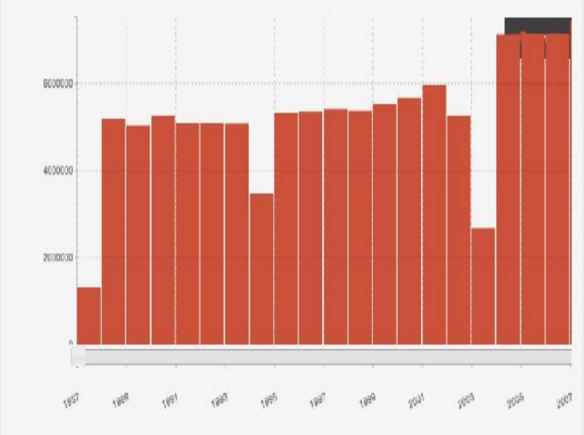


Fig. 6: Results (Graphs & Pie-Chart)

The results are obtained after processing the dataset through sqoop and getting it into the HDFS which allows us to analyze the data through giving query in hive, commands in pig, and programming in Map Reduce which allows us to get the analysis of the rainfall data which is very easily understandable and analysis makes the prediction part very much easier.

The results obtained are very much accurate as the analysis are obtained using the Hadoop tools which are very efficient even for a large amount of data.

The result is the analysis of the rainfall in different periods in a particular state or a particular part of a district. The result can be obtained in a theoretical as well as a graphical manner which can be used to compare the data of a particular district with itself and as well as with a different state or a district.

The pie chart shows the total amount of rainfall in Tamil Nadu (example) in the previous years. The pie chart results directly show us the region with most rainfall and the year in which the rainfall was very high.

The theoretical as well as graphical and pie chart representation makes the analysis understanding very easier, which allows the prediction for future resources very easy. The results show us the complete analysis of the rainfall in a particular state in Tamil Nadu.

VII. RESULTS & ANALYSIS

The analysis of the results obtained can be done very easily because the output clearly depicts the area with high or low rainfall. The state in Tamil Nadu which experiences most of the rainfall goes up to 19% and the least rainfall in a state goes up to 11% which makes us easy to predict the future data as well as get the clear image of rainfall in a state in Tamil Nadu, whether the rainfall makes it good for development of housing facilities or not. The analysis can be done that there is too much fluctuation of rainfall in Tamil Nadu over the coming years, which is clearly depicted in the diagram. So, it's not a good time in the year to get some projects going in Tamil Nadu.

The theoretical representation shows us the complete theoretical analysis of rainfall in a State in different years over a period. Precisely it gives us the data of how the rainfall has been in major parts of a state notably Andhra Pradesh, Andaman & Nicobar Islands in a period of time. The analysis or any comparison after this can be easily done on the basis of the output. The analysis shows us that Andaman has a rainfall percentage than Andhra in the coming years and we can clearly predict that this pattern will be followed in the coming years.

VIII. CONCLUSION& FUTURE WORK

In this work, the storm class records are designed and implemented within the map reduce characteristic. To evaluate the rainfall information in Hadoop environment. Hadoop environment is hive, pig, map reduce, if you want evaluation to find some deep evaluation on rainfall, which clearly validate the efficacy of our fashions in phrases of different programs and furthermore reveal a few exhilarating rainfall water portal locating. In destiny the sparks one hundred times quicker than Hadoop.

"We are using spark as we are able to get result hundred instances faster than Hadoop. The secret is that it runs in-memory at the cluster and that it is not tied to map. This makes repeated access to the same entry a lot quicker. Spark can run as a standalone or on top of Hadoop YARN, in which it is able to examine information at once from HDFS."

REFERENCE PAPERS

[1] Rainfall analysis and rainstorm prediction using MapReduce Framework C.P Shabariram ; K.E. Kannammal ; T. Manoj praphakar Computer Communication and Informatics (ICCCI), 2016 International Conference on 30 May 2016.

[2] Rainstorm Prediction using Support Vector Machine in Hadoop Cluster C.P Shabariram data mining and knowledge engineering, vol7 2015.

[3] Depth-Duration Frequency of Precipitation for Texas, U.S. Department of the Interior, U.S. Geological Survey, Page Contact Information: Contact USGS, December 07 2016.

[4] Map Reduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawat, OSDI 2004.

[5] Extracting storm-centric characteristics from raw rainfall data for storm analysis and mining, Kulsawasd Jitkajornwanich, Ramez Elmasri, John McEnery, Chengkai Li, Univ. of Texas at Arlington, Arlington, TX, ACM Digitstal library, nov 2012.

[6] Shabariram, C.P. (2015).Rainstorm Prediction using Support Vector Machine in Hadoop Cluster. Data Mining and Knowledge Engineering, 7(9), 316-320.

[7] Jitkajornwanich, K., Elmasri, R., McEnery, J., & Li, C. (2012, November). Extracting storm-centric characteristics from raw rainfall data for storm analysis and mining. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (pp. 91-99). ACM.

[8] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[9] Cleveland, T. G., & Thompson, D. B. (2008). Rainfall Intensity in Design. In World Environmental and Water Resources Congress 2008: Ahupua'A (pp. 1-10).

[10] Asquith, W. H. (1998). Depth-duration frequency of precipitation for Texas. US Department of the Interior, US Geological Survey.