RESEARCH HUB – International Multidisciplinary Research Journal (RHIMRJ)
Research Paper
Available online at: www.rhimrj.com

NOAA's National Climatic Data Center (NCDC) Weather Data Analysis on Apache Hadoop Yarn Single Cluster Environment

Dr. Dhaval S. Vyas[1st]
Dean- M.Phil Programme
C.U. Shah University
Wadhwan City, Gujarat, (India)

Mr. Bhavin J. Mathiya[2nd]
Research Scholar,
C.U. Shah University
Wadhwan City, Gujarat, (India)

Dr. Vinodkumar L. Desai[3rd]
Department of Computer Science
Government Science College
Chikhli, Navsari, Gujarat, (India)

Abstract: ***Apache Hadoop performance analysis and tuning done through parameter configuration using Real time application (NOAA's National Climatic Data Center (NCDC) Weather Data Analysis. Apache Hadoop currently use in various kind of real word application like weather data analysis, Social Networking sites analysis, Medical Data analysis, Sensor Data analysis. In this NOAA's National Climatic Data Center (NCDC) Weather Data Analysis is carried out using different Apache Hadoop customize parameter configuration for performance tuning to find out Hot and Cold days based on temperature recorded. NOAA's National Climatic Data Center (NCDC) is responsible for storing, observing, retrieving and provide public access to weather data. User can download this weather data using FTP functionality.***

Keywords: ***Apache Hadoop Yarn, HDFS, MapReduce, TeraGen, TeraSort, Tera Validate, TestDFSIO(Read), TestDFSIO(Write) WordCount.***

I. INTRODUCTION

Apache Hadoop Yarn is a open source framework developed by Apache Software Foundation. It is use for handling Big Data. It provides storage as well as processing functionality. HDFS (Hadoop Distributed File System) is used for storing data as a block of file locally as well as distributable. MapReduce is programming model based on Key Value Pair. MapReduce provide Mapper and Reducer for writing programming logic.

Apache Hadoop Yarn provides solution for various kind of real time application like Sensor Data Analysis, Social Networking Data Analysis, and Scientific Data Analysis. In this research paper NOAA's National Climatic Data Center (NCDC) Weather Data Analysis on Apache Hadoop Yarn Single Cluster Environment. The rest of paper is organized as follows. Section I Introduction, Section II Related Work, Section III Apache Hadoop Yarn Architecture, Section IV Environmental Setup, Section V Result and Discussion, Section VI conclusion and future work.

II. RELATED WORK

Yamazaki et al. [1] results show that the proposal method has achieved a reduction of execution times of jobs by about 11.1% in Multi-Hadoop environment as compared to original Hadoop. Tan et al. [2] present some of the challenges and issues that are to be considered in performance tuning when running applications in Hadoop. Yang et al. [3] propose a statistic analysis approach to identify the relationships among workload characteristics, Hadoop configurations and workload performance. Lin et al. [4] propose a node performance measurement method on Hadoop. Lin et al. [4] describe in detail how to measure the performance value of each node in heterogeneous Hadoop cluster and evaluate measurement results by running MapReduce programs. Zhuoyao Zhang et al [5] offer a MapReduce performance model that estimates the program completion time for processing a new dataset.

III. APACHE HADOOP YARN ARCHITECTURE

Apache Hadoop Yarn is a framework developed by Apache Software Foundation for handling Big Data [6].

  ISSN: 2349-7637 (Online)

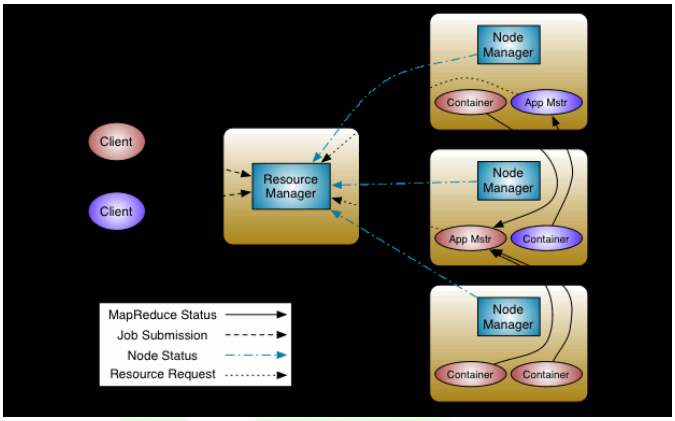

Fig. 1 Hadoop Yarn Architecture [6]

IV. ENVIRONMENT SETUP

These experiments carried out on single cluster node and Ubntu 12.04.4 LTS Operating System and Java version 1.7, Openssh Server Apache Hadoop 2.4.1 Softwares are installed. All research experiments carried out using NOAA's National Climatic Data Center (NCDC) Weather Data Analysis on Apache Hadoop Yarn Single Cluster Environment [7].

V. RESULTS AND OBSERVATIONS

Table 5.1 and Fig.5.1 depicts NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days Execution time (Sec) for the different data sizes. The shorter Execution time (Sec) indicates better performance and respectively the Execution time (Sec) indicate worse performance.

Table 5.1 NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days

Data Size	Execution time (Seconds)
20 Lacks (500MB)	57
40 Lacks (1000MB)	108

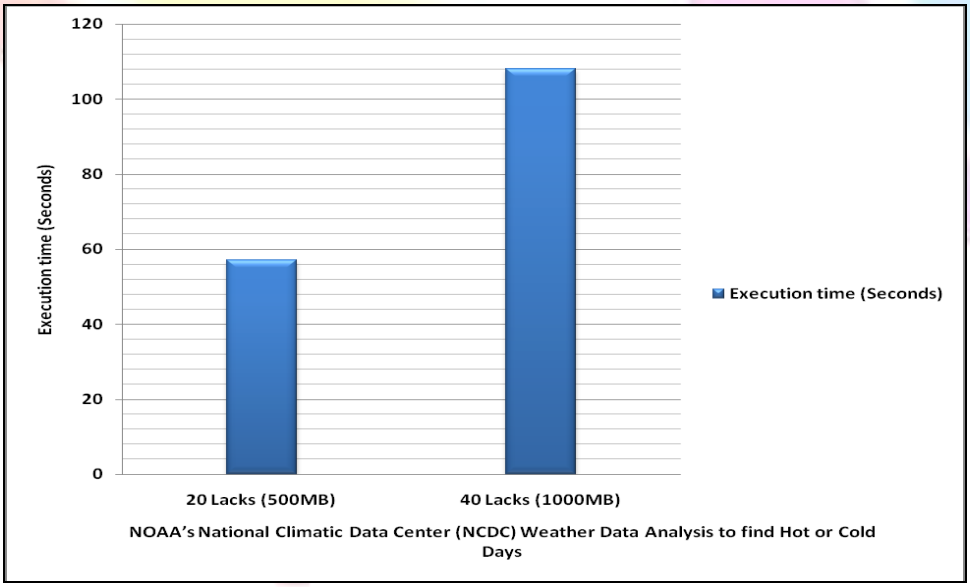

Fig 5.1 Execution time (Seconds) of NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days with different Data Size

Table 5.2 and Fig. 5.2 depict NOAA's National Climatic Data Center (NCDC) Weather Data Analysis Execution time (Sec) for the different data sizes as well as different Hadoop's Customized Compression Parameter Configuration Settings like mapreduce.output.fileoutputformat.compress, mapreduce.output. fileoutputformat compress.codec. The shorter Execution time (Sec) indicates better performance and respectively the Execution time (Sec) indicate worse performance.

ISSN: 2349-7637 (Online)

Table 5.2 NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days Using Different mapreduce.output.fileoutputformat.compress Configuration Settings (mapred-default.xml)

	Execution time (Seconds)				
Data Size	**mapreduce.output.fileoutputformat.compress = FALSE (Default)**	**mapreduce.output.fileoutputformat.compress=TRUE, mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.DefaultCodec**	**mapreduce.output.fileoutputformat.compress=TRUE, mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.BZip2Codec**	**mapreduce.output.fileoutputformat.compress=TRUE, mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.GzipCodec**	**mapreduce.output.fileoutputformat.compress=TRUE, mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.Lz4Codec**
20 Lacks (500MB)	57	63	58	56	54
40 Lacks (1000MB)	108	97	110	95	92

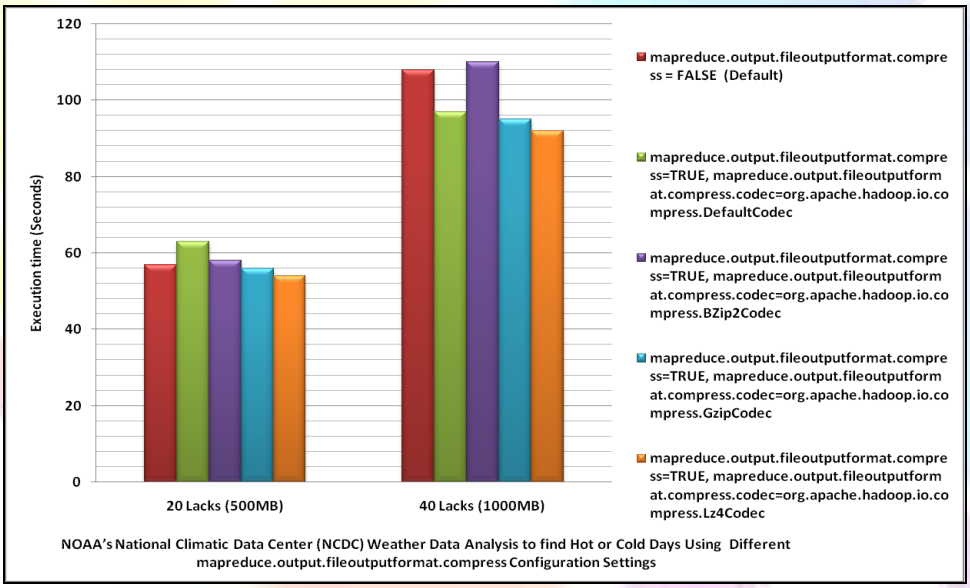

Fig 5.2 Execution time (Seconds) of NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days with different Data Size and Different mapreduce.output.fileoutputformat.compress Configuration Settings

Table 5.3 depicts and Fig. 5.3 the NOAA's National Climatic Data Center (NCDC) Weather Data Analysis Execution time (Sec) for the different data sizes as well as different Hadoop's Customized Compression Parameter Configuration Settings like mapreduce.map.output.compress. mapreduce.map.output.compress.codec. The shorter Execution time (Sec) indicates better performance and respectively the Execution time (Sec) indicate worse performance.

Table 5.3 NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days Using Different mapreduce.map.output.compress Configuration Settings

	Execution time (Seconds)				
Data Size	**mapreduce.map.output.compress = FALSE (Default)**	**mapreduce.map.output.compress =TRUE, mapreduce.map.output.compress.codec**	**mapreduce.map.output.compress =TRUE, mapreduce.map.output.compress.codec**	**mapreduce.map.output.compress =TRUE, mapreduce.map.output.compress.codec**	**mapreduce.map.output.compress =TRUE, mapreduce.map.output.compress.codec**

20 Lacks (500MB)	57	55	60	53	54
40 Lacks (1000MB)	108	90	110	92	91

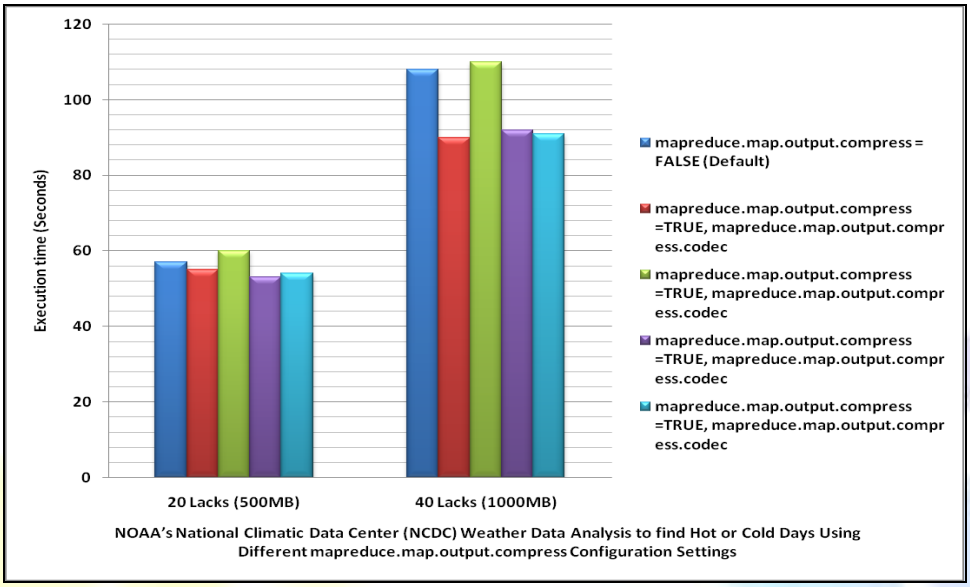

Fig 5.3 Execution time (Seconds) of NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days with different Data Size and Different mapreduce.map.output.compress Configuration Settings

Table 5.4 and Fig. 5.4 depicts the NOAA's National Climatic Data Center (NCDC) Weather Data Analysis Execution time (Sec) for the different data sizes as well as different Hadoop's Customized dfs.blocksize Parameter Configuration Settings like 128 MB (Default), 64 MB, 256 MB, 512MB. The shorter Execution time (Sec) indicates better performance and respectively the Execution time (Sec) indicate worse performance.

Table 5.4 NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days Using Different dfs.blocksize Configuration Settings

	Execution time (Seconds)				
Data Size	**128 MB (Default)**	**64 MB**	**256 MB**	**512MB**	**1024MB**
20 Lacks (500MB)	57	55	54	59	60
40 Lacks (1000MB)	108	102	110	111	113

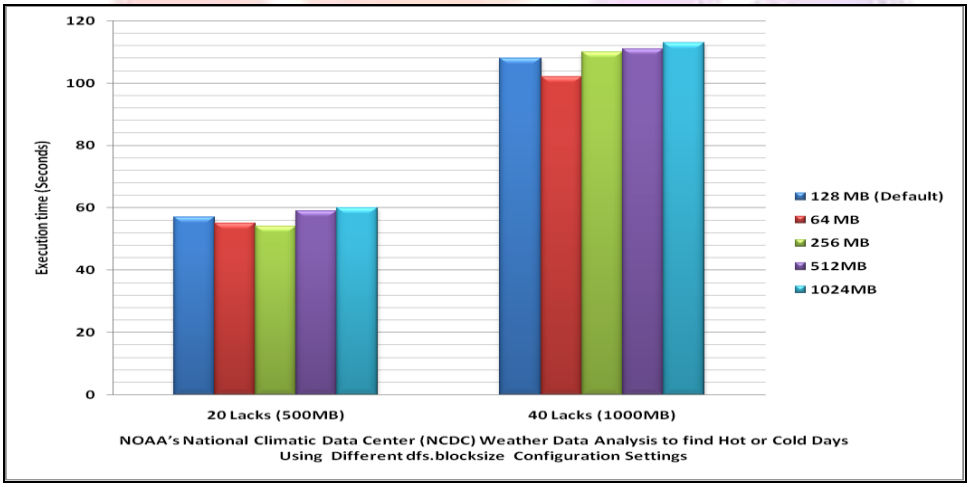

Fig 5.4 Execution time (Seconds) of NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days with different Data Size and Different mapreduce.map.output.compress Configuration Settings

Table 5.5 depicts the NOAA's National Climatic Data Center (NCDC) Weather Data Analysis Execution time (Sec) for the different data sizes as well as different Hadoop's Customized preduce.task.io.sort.factor Parameter Configuration Settings like 10 (Default), 30, 50 ,70, 100. The shorter Execution time (Sec) indicates better performance and respectively the Execution time (Sec) indicate worse performance. S

Table 5.5 and Fig.5.5 NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days Using Different mapreduce.task.io.sort.factor Configuration Settings (Impetus Technologies).

Data Size	**Execution time (Seconds)**				
	mapreduce.task.io.sort.factor = 10 (Default)	**mapreduce.task.io.sort.factor = 30**	**mapreduce.task.io.sort.factor = 50**	**mapreduce.task.io.sort.factor = 80**	**mapreduce.task.io.sort.factor = 100**
20 Lacks (500MB)	57	61	60	55	53
40 Lacks (1000MB)	108	95	100	104	106

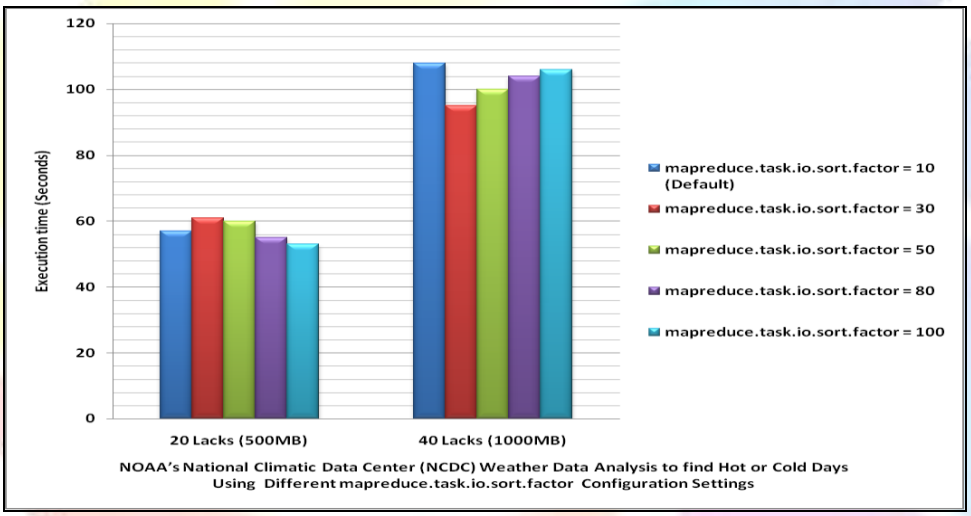

Fig 5.5 Execution time (Seconds) of NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days with different Data Size and Different mapreduce.task.io.sort.factor Configuration Settings

Table 5.6 and Fig.5.6 depicts the NOAA's National Climatic Data Center (NCDC) Weather Data Analysis Execution time (Sec) for the different data sizes as well as different Hadoop's Customized mapreduce.reduce.speculative Parameter Configuration Settings like TRUE (Default), FALSE. The shorter Execution time (Sec) indicates better performance and respectively the Execution time (Sec) indicate worse performance.

Table 5.6 NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days Using Different mapreduce.reduce.speculative Configuration Settings

Data Size	**Execution time (Seconds)**	
	TRUE(Default)	**FALSE**
20 Lacks (500MB)	57	60
40 Lacks (1000MB)	108	99

ISSN: 2349-7637 (Online)

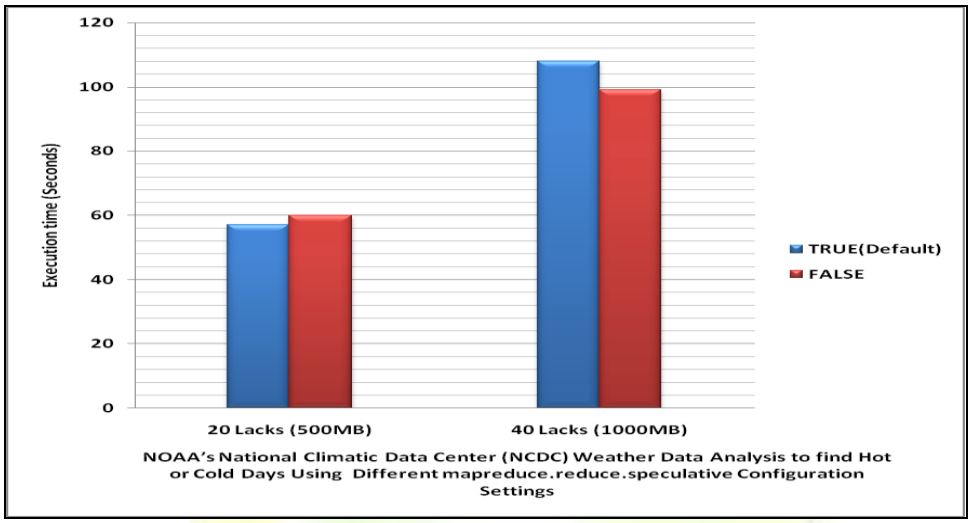

Fig 5.6 Execution time (Seconds) of NOAA's National Climatic Data Center (NCDC) Weather Data Analysis to find Hot or Cold Days with different Data Size and Different mapreduce.reduce.speculative Configuration Settings

VI. CONCLUSION AND FUTURE WORK

In this research NOAA's National Climatic Data Center (NCDC) Weather Data Analysis done on Apache Hadoop Yarn Single Cluster Environment. In future we will do experiments on other real time applications data like social networking, scientific data analysis etc.

REFERENCES

1. Yamazaki, Kazuki, et al. "Implementation and Evaluation of the JobTracker Initiative Task Scheduling on Hadoop." Computing and Networking (CANDAR), 2013 First International Symposium on. IEEE, 2013.
2. Tan, Yu Shyang, and Bu-sung Lee. "MapReduce Framework: Some Challenges." Annual International Conference on CCV. 2010.
3. Yang, Hailong, et al. "Statistics-based Workload Modeling for MapReduce." Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International. IEEE, 2012.
4. Lin, Wenhui, and Jun Liu. "Performance Analysis of MapReduce Program in Heterogeneous Cloud Computing." Journal of Networks 8.8 (2013): 1734-1741.
5. Zhuoyao Zhang; Cherkasova, L.; Boon Thau Loo, "Optimizing cost and performance trade-offs for MapReduce job processing in the cloud," Network Operations and Management Symposium (NOMS), 2014 IEEE , vol., no., pp.1,8, 5-9 May 2014
6. http://hadoop.apache.org/
7. Abhishek "Hadoop Project on NCDC (National Climate Data Center – NOAA) Dataset "9 August 9 2015 web. 26 Arpil 2016 < https://www.eduonix.com/blog/bigdata-and-hadoop/hadoop-project-on-ncdc-national-climate-data-center-noaa-dataset/>.

  ISSN: 2349-7637 (Online)