# HDFS AND MAP REDUCE - HADOOP ADMINISTRATION

# <sup>1</sup>P.Srilakshmi, <sup>2</sup>N.Siva Rama Krishna Prasad, <sup>3</sup>Dr.M.I.Thariq Hussan

<sup>1</sup>Associate Professor, <sup>2</sup>Assistant Professor, <sup>3</sup>Professor and Head Department of Information Technology, Guru Nanak Institutions Technical Campus, Hyderabad, India

*Abstract:* HDFS administrator is one who administers and manages Hadoop clusters and all other resources in the entire Hadoop ecosystem. A Hadoop admin's job is not visible to all other groups or end users. The role of a Hadoop admin is mainly associated with tasks that involve installing and monitoring Hadoop cluster, HDFS file structure, locations, and the updated files. Map Reduce administration includes monitoring the list of applications, configuration of nodes, application status, etc. HDFS (Hadoop Distributed File System) contains the user directories, input files, and output files. The Map Reduce commands such as put and get used for storing and retrieving. After starting the Hadoop framework (daemons) by passing the command "startall.sh" on "/\$HADOOP\_HOME/sbin", pass the following URL to the browser "http://localhost:50070". It describes the file structure of HDFS and data node information in a cluster. A map reduce application is a collection of jobs such as map job, combiner, partitioner and reduce job. It is mandatory to monitor and maintain the following configuration of data node where the application is suitable. Each application consists of the number of data nodes and resources. To monitor all these things, it is imperative that we should have a user interface. After starting the Hadoop framework by passing the command "start-Il.sh" on "/\$HADOOP\_HOME/sbin", pass the following URL to the browser "http://localhost:8080". It describes the details of the particular application and its local host address of the running node.

Index Terms - HDFS Monitoring, Map Reduce Monitoring

#### I. INTRODUCTION

Hadoop administration includes both HDFS and map reduce administration. HDFS administration includes monitoring the HDFS file structure, locations, and the updated files. Map reduce administration includes monitoring the list of applications, configuration of nodes, application status, etc.

#### **1.1 HDFS Monitoring**

HDFS (Hadoop Distributed File System) contains the user directories, input files, and output files. Use the Map Reduce commands, **put** and **get**, for storing and retrieving. After starting the Hadoop framework (daemons) by passing the command "startall.sh" on "/\$HADOOP\_HOME/sbin", pass the following URL to the browser "http://localhost:50070". You should see the following screen on your browser.

The following screenshot shows how to browse the HDFS. It shows the files in the "/user/hadoop" directory.

				Hards Tests	(#):41	E. Hotosja
tites ()) - arbeit.	+ Xinan	nes faren 🗕 🛄	With Light	tian + Nam	itiniat. +/m	
Contract litera	and the later			~ a 🖪	time	9 0 0 4
-adoop	-	0.00	and the second	10 Mar 1		
Permission.	Deter	(irong),	N/W	Replication	Block Van	tes
Pendistan meneri	tractor Totality	lineşi 1.deraya	5494 1.1	feption t	Hock Nati	New Services
Hernisian mereri mar	tranie Nation Nation	Grad Nargoa Nargoa	5494 1.1 1.1	Angliadan 1	Hole Yay 17 17	New Services 10
Permission (Two et al. (Two et al. (Two et al.	Datest Nation Nation Nation	Urași Nargoja Nargoja Nargoja	509 17 18	Angliadion 11 12 1	Hock Yao 17 19 19	New Serves No

The following screenshot shows the data node information in a cluster with its configurations and capacities.

900	* *
9.00	* *
9.00	4.8
4 taket	. Westing
	1.4.4
and the second state	
Survey of the American	
	i fake Numero

## **1.2 Map Reduce Job Monitoring**

A Map Reduce application is a collection of jobs (Map job, Combiner, Practitioner, and Reduce job). It is mandatory to monitor and maintain the following configuration of data node where the application is suitable. Each application consists of the number of data nodes and resources. To monitor all these things, it is imperative that we should have a user interface. After starting the Hadoop framework by passing the command "start-all.sh" on "/\$HADOOP\_HOME/sbin", pass the following URL to the browser "http://localhost:8080".

ala silena -	Planta	100	With Bridge												
					Had	De Poorfee						10		2.0	
1000 A	-	* %. Sta	Andre UA		3 Word Lie	section	Antonia	ing 2. 4	1040	in. 30	14.99	11.0	1.9	÷	
timeria -				11111	ing along		+0.00	1.11	9	÷	1 +		ŧ .:	=	
Currer.		-								. Spanne	-	-	-		
Please						\$PM071	Define								
Cardina -					Aparter	Attern Typered	INVESTIGATE							1	
15 m and					Anetr.	attent Tager	A REAL PROPERTY.								
20010	101					Testification	Bucchelens.								
<ul> <li></li></ul>						Matterit	9.748 GEL9.9	11012							
20122						aliticated states of	hinters.								
forement.					- 10	ugrovice									
Seattle .										-	-	144	-1	1	
						Folul Horse	ann Preempt	sit com	- 110	100.00					6.
				Data I	harmore of	AM Contain	tern Presengt	ed a							
			Manufact	Nonline of Section	COLUMN CONTRACTOR	plied from	annext Athen	NUL COLOR	wy t, all	111.11	- 11				
							Alleria	-	-	_					
					-	Tipare nes		approx.							
			and the second												
B . B . I . B		- 67		-						0.0	6116	211		627	÷
Application	III. 8	ны 🛃	wfax Web	\$treast.					43	8.8	MoniO	911	4	112	in
Application	ш.В	HE E	vfax Web	\$roour	5				13	88	More D	9.13	1	1.2	ios
8 Applicatio	ant B	ITH E	wfax Web	broom:	Nedes of t	he cluster -	Macilla Firefe	,	• 3	88	Mue O	913	1	1.2	in:
Application	anı, R	nei 🛃	artax Web	Browner	Nodes of t	be duster •	Mazila Finfe	¢	• 5	8.95	More D	9.12	1	1 <sub>12</sub>	in:
P Applicate	nı A	ITE .	ontox Web	browner - Towner	Nodes of 1	be cluster -	Mazila Firefe	<	*3	8.8	More C	9.12		l ha	100
P Applicatio M Intera (I	ana ; M	1781 🛃	n Mapilansi Web	browner e Veteral	Nedes of t * [ Wo	the charter + rd Caurt Exa	Habila Firefe	k Katelludorg	*3 2.4	R B	Mon Q	9.12	e e	y y	ioo :
Mittan (	9-12	ин <b>(</b>	MapRedu	brooser e Uzral	Nedes of t * 0 Wa	the cluster + rd Court Exa	Mazila Finfe n. X Nip	k Nate Hudoup	* ; 2_1	tion o	Moe D	912		X A	ion e
Mittan ( Constant)	9 - 121 1 - 121	ин <b>н</b>	Mapiladu Mapiladu	brooser o Tataral	Nodes of t * 1 Wo	the cluster + rd Court Exa	Hadla Fiefe n. X Na Y C	k Katerludorp Ør Gorge	4) ( 2. X	a an	Moe Q of the D	911 ciuri 6	4	× #	ing
Mittan Mittan ( Calanati	979 - 122 19-122	ne 🥌	n Mapilita Mapilita	troon a Ucral	Nodes of t * 0 Wo	the charter - rd Caurt Exa	Mazila Firefa n. X No X C	ene Hudoop Orr Googe	4) ( 2. X	R B toda	d the	911 clari	4	x #	100 ÷
Mitter (	47 - 122 19 - 122 19 - 12	ин <b>(*</b>	Nigelinda Migelinda	troour a Uzral	Nodes of t * [] Wa	the cluster + rd Court Ex	Mazila Firefa n. X Na Y C	k Kote Hudsog Of Y Giospa	4) ( 2. X	a B tota	d the	912	2 C	x #	ini 🔹
Mittal (Mittal) (Mittal)	4-12 1-12	ин <b>(1</b> 1. х 1	Niełeby Chi	troour	Nodes of t * 0 Wa	the cluster - rd Caurt Ex Nodes	Peoble Finite n × No + C	r Litte Huding Of Kanpo	4) (	R B tota	Non Q si the th	913 6	2 C	x ft	100 e coi
Mittal (Mittal) (Mittal)	9-32 9-32 120	ин <b>(4</b> 1. х.) ()	Mişileti Mişileti Gili	troour a land	Nodes of t * 0 Wa	the cluster - rd Caurt Ex Nodes	Musila Finfe n × No + C of the cluster	r Litte Huding Of Y Goop	* 3 2*	R B Notes	Hoe Q ci the C	911 0	2 C	x # -	100 e e e e
Mitta ( t diaar Mitta (	41-521 100-11 100-11	ues 🕌 5 - X - 1 8	Mişileti Mişileti	troour a Ucral	Nodes of 1 * [] Wa	he cluster - rd Caurt Eco Nodes	Hacilla Finde n. X X Io x C of the cluster	x Kote Hudong Qir Licopo	* 3 2. *	R B total	Hoe Q el the the	911 6	2 C	x # -	100 + 100 + 100
Mitta ( Mitta ( diaan Chief	9-30 9-30 500	une 🕌 e. x 1	MapRedu MapRedu	troour a Ucral	Nodes of 1 * [ ] Wa	he cluster - rd Court Eco Nodes	Muzilla Finifie n. X X p v C of the cluster	e ene Hudong Qir Gonga	* 3 2. *	R B Note	Hote Q cl the C	911 6	4	x #	ion in the second se
Mitta ( Mitta ( diaan	9-50 9-50 100	ues 🕌	Mapino Mapino Chi	i loral	Nodes of t	the cluster + rd Caurt Ex Nodes	Hustla Finite • Vip • C of the cluster	erinter Oriop	* 3 2 *		Hoe Q	911 6 -	4	x #	100 ···································
Mintea I e Sincer Contest to E	19-12 19-12 19-12	ина 🥌 р. * 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	MigArdy	towar	Nodes of t	the cluster + rd Court Ext Nodes	Musila Finfe a. × No v C of the cluster	eriop	4) ( 2. 4)		Hoe Q	911 6	4	x #	100 ···································
Mittal Mittal Saar	19-12 19-12 1001	unes 🥌	MigRady MigRady	a local	Nodes of t * 0 Wo	the cluster - rd Court Ex Nodes	Musila Finite a × Vio + C of the cluster	erioto Orioto	* ( 2_X)		Hoe Q	911 6	4	x #	100
Mittal Mittal Car	19-12 19-12 19-12	nes 🛃	Mapiledu Mapiledu Citel	a land	Nodes of t * 0 Wa	the charter - rd Court Ext Notices	Macilla Firefa a. × Vip v 0 of the cluster	er fictor	* 1 2		Hoe Q	912	4	x #	100 ·····
			Mig-Arch	ineer s land	Nodes of 1 * 0 Wo	the charter + rd Court Exa Nodes	Hazila Finite n × Vip + C of the cluster	er iop	* 1 2		More Q of the D	912 0	4	x ft	100 ···································
Mittal			Mig-Retu	Insur 1 Cord 1 Cord	Nodes of 1 * 0 Wo	the cluster - rd Court Ex Nodes	Hacilla Finde • • • • • • • • • • • • • • • • • • •	er here Or i op	* 1 2_X			912	4	x ft	
			MigRedu MigRedu China and Anno Anno Anno Anno Anno Anno Anno Anno	traver a laral a laral	Nodes of 1 * 0 Wo	he cluster - rd Caurt Ex Nodes	Mucilia Finde n. X Xop v C of the cluster i i i i i		* 1 2_X		d the	912	4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	× #	
Mittal				traine s larai s larai s larai s larai	Nodes of t * 1 We we we we true and and and and and and and and	the cluster - rd Court Ex Notices	Hazila Finte a. X No + C of the cluster	ene Hatop Or Goope	* 1 2		Hori C d the D	912	· · · ·	x # -	
Mittal I			Han Web	traver a land a	Nodes of 1 * 1 wo	he cluster - rd Court Ex Nodes	Hacilla Finde x X Vip y C of the cluster c C C C C C C C C C C C C C C C C C C C	ete hating Or Scope	• 3 2 ×		Hori C of the d	911	4 D 1	x #	

You should see the following screen on your browser. In the above screenshot, the hand pointer is on the application ID.

It describes the following information:

- The application name
- Type of that application
- Current status, Final status

The following screenshot shows the details of a particular application.

The following screenshot describes the currently running nodes information. Here, the screenshot contains only one node. A hand pointer shows the local host address of the running node.

#### **II. HADOOP ADMINISTRATOR**

In the Hadoop world, a system administrator is called a Hadoop Administrator. Hadoop admin roles and responsibilities include setting up Hadoop clusters. Other duties involve backup, recovery and maintenance. Hadoop administration requires good knowledge of hardware systems and excellent understanding of Hadoop architecture.

#### 2.1 Roles and Responsibilities

Responsible for implementing and support of the Enterprise Hadoop environment involves designing, capacity arrangement, cluster set up, performance fine-tuning, monitoring, structure planning, scaling and administration.

The administrator consultant works closely with infrastructure, network, database, business intelligence and application teams to ensure that business applications are highly available and performing within agreed on service levels.

- 1. Need to implement concepts of Hadoop eco system such as YARN, Map Reduce, HDFS, HBase, Zookeeper, Pig and Hive.
- 2. In charge of installing, administering, and supporting Windows and Linux operating systems in an enterprise environment.
- 3. Accountable for storage, performance tuning and volume management of Hadoop clusters and Map Reduce routines.
- 4. In command of setup, configuration and security for Hadoop clusters using Kerberos.
- 5. Monitor Hadoop cluster connectivity and performance.
  - Manage and analyze Hadoop log files.
  - File system management and monitoring.
  - Develop and document best practices
  - HDFS support and maintenance.
  - Setting up new Hadoop users.
  - Responsible for the new and existing administration of Hadoop infrastructure.
  - Include DBA Responsibilities like data modeling, design and implementation, software installation and configuration, database backup and recovery, database connectivity and security.

Hadoop admin roles and responsibilities on popular job portals like Dice.com, Glassdoor.com, and Monster.com are as follows:

• Hadoop administration includes ongoing administration of the Hadoop infrastructure.

- Keeps track of Hadoop Cluster connectivity and security
- Capacity planning and screening of Hadoop cluster job performances.
- HDFS maintenance and support
- Setting up new Hadoop users.

# 2.2 Hadoop Administration Skills

- Strong scripting skills in Linux environment
- Hands on experience in Oozie, HCatalog, Hive
- Knowledge of HBase for efficient Hadoop administration

# **III. HOW TO SECURE HADOOP**

Current versions of Hadoop are more secure than earlier versions. However, by default they are not secure out-ofthe-box, therefore you need to get your hands on the configuration files and make sure that security relevant options are actually enabled.

Also, it is important to keep in mind that the only way of making your cluster secure is to protect it at different layers, from the lower (that is OS-level security) up to applicationlevel and network-level security.



Fig1: Architecture of Layers of Hadoop Security

Here is a minimum set of options we strongly recommend to enable in order to secure Hadoop clusters:

- 1. Enable HDFS extended ACLs by adding the following properties to *hdfs-site.xml* dfs.namenode.acls.enabled true
- 2. Enable Hadoop security module and strong authentication (Kerberos) by adding the following properties to *core-site.xml* hadoop.security.authentication kerberos Hadoop.security.authorization true
- Secure HDFS by adding the following properties to hdfs-site.xml (in particular, enable HTTPS and advanced authorization) dfs.block.access.token.enable true dfs.namenode.keytab.file /etc/hadoop/conf/hdfs.keytab dfs.namenode.kerberos.principal hdfs/ HOST@YOUR-REALM.COM

dfs.namenode.kerberos.internal.spnego.principal HTTP/\_HOST@YOUR-REALM.COM dfs.secondary.namenode.keytab.file /etc/hadoop/conf/hdfs.keytab dfs.secondary.namenode.kerberos.principal hdfs/\_HOST@YOUR-REALM.COM dfs.secondary.namenode.kerberos.internal.spnego.prin cipal. HTTP/ HOST@YOUR-REALM.COM dfs.datanode.data.dir.perm 700 dfs.datanode.address 0.0.0.0:1004 dfs.datanode.http.address 0.0.0:1006 dfs.datanode.keytab.file /etc/hadoop/conf/hdfs.keytab dfs.datanode.kerberos.principal dfs/ HOST@YOUR-REALM.COM dfs.web.authentication.kerberos.principal HTTP/\_HOST@YOUR\_REALM dfs.http.policy HTTPS\_ONLY

- Use Web HDFS (REST API for HDFS), make sure Kerberos authentication is on by adding the following properties to *hdfs-site.xml*dfs.web.authentication.kerberos.principal HTTP/\_HOST@YOUR-REALM.COM
  dfs.web.authentication.kerberos.keytab
  /etc/hadoop/conf/HTTP.keytab
- 5. Enable transparent data encryption and configure the Key Provider (which will take of generating and providing encryption keys). You can use the *hdfs crypto* to test your configuration.
- 6. Finally, as a general security warning, make sure that firewall rules are correctly set and limit the access from the Internet only to necessary services. Indeed, by default Hadoop will expose two web interfaces, one for the Resource Manager (on port 8088) and one for Name Node (on port 50070), which are not protected by authentication, therefore they can potentially leak sensitive and critical information

-991-4.44
an a

COLUMN STATES	the second s		
All and a second	and the shares in	1	
Oversiane -	ALC: 10 10 10 10 10 10 10 10 10 10 10 10 10		
-			
	11.0000.0000	duite .	
(pages)	04110000000000000	********	
Cone (h	prosecup produced and a		
Bearing	3410312100004		
Summary			
the second second			
Wine in American	and the state of the second		
deal merry and high	and the second sec	-	
10 10 10 10 10 10	A M C M C M CONTRACTOR AND AND	0.000000110	
Tellpine Identity		- paleon.	
011mm			
au 101 (au)		114	
and insuring		18840	
and stands		- A.	
101 million (1997)			
instantion.		144	
Destantante.		- 4-	
 Children and			second in the local second

3.1 Continuous Security and Vulnerability Assessment

Monitoring a Hadoop cluster becomes a headache for system administrators and developers, therefore an automated tool can be of great help. That's why we embedded these Hadoop security analysis into our product Elastic Workload Protector in order to help you continuously monitor the security of your Hadoop cluster(s) and be notified as soon as there is a potential issue or mis-configuration.

-	Discovered Hadoop Vulnerabilities or Misconfigurations Operating System
NAME:	<b>6</b> 3
Descryther	Present first basis the Bard House Bard have been basis AND/OCH 2011/PTTT DIAMAGE IT havens a meaning and management of theme basis as using authority and management of theme Modified Process (Marcel Marcel Andrews), and and an anti-Andrews (Marcel Andrews), and an anti-Andrews Modified Process (Marcel Andrews), and and and an anti-Andrews (Marcel Andrews), and and an anti- Modified Process (Marcel Andrews), and and an anti-Andrews (Marcel Andrews), and and and and Modified Process (Marcel Andrews), and and an anti- Marcel Process (Marcel Andrews), and and and an anti- Marcel Marcel Andrews), and and an anti- Marcel Marcel Andrews (Marcel Andrews), and and and and and and and Marcel Marcel Andrews), and and an anti- Marcel
CVE:	Installer
100 e -	Ne DDs

# **IV. HADOOP CLUSTER INSTALLATION**

Installing a Hadoop cluster typically involves unpacking the software on all the machines in the cluster or installing it via a packaging system as appropriate for your operating system. It is important to divide up the hardware into functions.

Typically one machine in the cluster is designated as the Name Node and another machine as the Resource Manager, exclusively. These are the masters. Other services (such as Web App Proxy Server and Map Reduce Job History server) are usually run either on dedicated hardware or on shared infrastructure, depending upon the load.

The rest of the machines in the cluster act as both Data Node and Node Manager.

# 4.1 Configuring Hadoop in Non-Secure Mode

Hadoop's Java configuration is driven by two types of important configuration files:

- Read-only default configuration -core-default.xml, hdfs-default.xml, yarn-default.xml and mapred-default.xml.
- Site-specific configuration -etc/hadoop/coresite.xml,etc/hadoop/hdfs-site.xml, etc/hadoop/yarn-site.xml and etc/hadoop/mapred-site.xml.

Additionally, you can control the Hadoop scripts found in the bin/ directory of the distribution, by setting site-specific values via the etc/hadoop/hadoop-env.sh and etc/hadoop/yarnenv.sh.

To configure the Hadoop cluster you will need to configure the environment in which the Hadoop daemons execute as well as the configuration parameters for the Hadoop daemons.

HDFS daemons are Name Node, Secondary Name Node, and Data Node. YARN daemons are Resource Manager, Node Manager, and Web App Proxy. If Map Reduce is to be used, then the Map Reduce Job History Server will also be running. For large installations, these are generally running on separate hosts.

## 4.2 Day-to-Day Hadoop Cluster Care

The lifetime of a Hadoop Administrator revolves around making, managing and watching the Hadoop Cluster. However, cluster administration isn't an identical activity practiced through and thru by directors from round the globe. The most variable during this case is that the "Distribution of Hadoop" or in easy words a 'cluster' based mostly wherever you decide on the cluster watching tools. The various distributions of Hadoop are Cloudera, Horton works, Apache and MapR. Apache distribution is in fact the Open supply Hadoop distribution.



As associate administrator, if I need to setup Hadoop clusters on the Horton works/Cloudera distribution, my job are easy as a result of all the configurations files are gift on startup. However, within the case of the open supply Apache distribution of Hadoop, we've got to manually setup all the configurations like Core-Site, HDFS-Site, YARN-Site and Map Red-Site.

Once we've got created the cluster, we've got to make sure that the Cluster is active and accessible the least bit times. For this, all the nodes within the cluster need to be setup. They're Name Nodes, Data Node, Active & amp; Standby Name Nodes, Resource Manager and also the Node Manager.

Name Nodes is that the Heart of the cluster. It consists of information that helps the cluster to acknowledge the info and coordinate all the activities. Since lots depends on the Name Nodes, we've got to make sure 100 percent dependableness and for this, we've got one thing known as the Standby Name Nodes that acts because the backup for the Active Name Nodes. Name Nodes stores the information, whereas the particular information is hold on within the information Node within the type of Blocks. The Resource Manager takes care of the cluster's CPU and memory resources the least bit times for all the roles whereas the applying Master manages the particular jobs. If all the above services are running and are active at all times, then your Hadoop Cluster is ready for use.

When setting up the Hadoop Cluster, the administrator will also need to decide the cluster size based on the amount of data that is to be stored in the HDFS. Since the replication factor of HDFS is 3, 15 TB of free space is required to store 5 TB of data in the Hadoop cluster. The replication factor is set at 3 in order to increase the Redundancy and Reliability. Cluster growth based on storage capacity is a very effective technique that is implemented in the clusters. We can add new systems to the existing cluster and thereby increase the storage space to any number of times.

Another important activity we have to perform as a Hadoop Administrator is that we have to monitor the cluster on a regular basis. While monitor the Cluster it is to ensure that it is up and running at all times and to keep track of the performance. Clusters can be monitored using the various cluster monitoring tools. We choose the appropriate cluster monitoring tools based on the distribution of Hadoop that you are using.

# 4.3 Monitoring tools for distribution of Hadoop

Open Source Hadoop/Apache Hadoop à Nagios/Ganglia/Ambari

Shell scripting/Python Scripting.

Cloud era Hadoop à Cloud era Manager + Open Source Hadoop tools

Horton works à Apache Ambari + Open Source Hadoop tools.



Ganglia is used for monitoring Compute Grids i.e a bunch of servers working on the same task to achieve a common goal. It is like a cluster of clusters.

Ganglia is also used to monitor the various statistics of the cluster. Nagios is used for monitoring the different servers, the different services running in the servers, switches, and network bandwidth via SNMP etc.

Nagios and Ganglia are open source which is why both are slightly difficult to manage when compared to Ambari and Cloud era Manager. The former is the monitoring tool used by Horton works distribution while Cloud era uses the latter.

Apache Ambari and Cloud era Manager are more popular tools because they come along with the Hadoop Distributions providing you with around 10,000 statistics to monitor. But the drawback is that they are not open source.

# 4.4 Key skills For Hadoop admin

- Mater of UNIX commands
- Sound knowledge UNIX based File System
- Excellent hold on shell scripting
- Deep understanding and knowledge of Operating System, Scheduling, Process Management.
- Command on Hadoop cluster setup Single Node, Pseudo Distributed and fully distributed Mode.
- Knowledge of Networking.
- What does a Hadoop Admin do?

- Installation and Configuration
- Cluster Maintenance
- Resource Management
- Security Management
- Troubleshooting
- Cluster Monitoring
- Backup and Recovery Task

If working with open source Apache Distribution then hadoop admin's has to manually setup all the configurations- Core-Site, HDFS-Site, YARN-Site and Map Red-Site. However, when working with popular hadoop distribution like Horton works, Cloud era or MapR the configuration files are setup on startup and the hadoop admin need not configure them manually.

#### 4.5 Backup and recovery tasks

Many people don't consider backups since Hadoop has 3X replication by default. Also, Hadoop is often a repository for data that resides in existing data warehouses or transactional systems, so the data can be reloaded. That is not the only case anymore! Social media data, ML models, logs, third-party feeds, open APIs, IoT data, and other sources of data may not be reloadable, easily available, or in the enterprise at all. So, this is not critical single-source data that must be backed up and stored forever.

There are a lot of tools in the open-source space that allow you to handle most of your backup, recovery, replication, and disaster recovery needs. There are also some other enterprise hardware and software options.

#### **V. CONCLUSION**

Hadoop distributed file systems provides a high throughput access to data of an application and is suitable for applications that need to work with large data sets. It is designed to hold terabytes or petabytes of data and provides higher throughput access to this data.

Files containing data are stored redundantly across number of machines for higher availability and durability to failure. Moving computation is faster than the moving data. By default, the Map Reduce framework gets input data from the Hadoop Distributed File System (HDFS). The reduce phase uses results from map tasks as input to a set of parallel reduce tasks. The reduce tasks consolidate the data into final results.

By default, the Map Reduce framework stores results in HDFS.

#### REFERENCES

- [1] https://www.dezyre.com/article/what-are-the-jobresponsibilities-of-a-hadoop administrator/306
- [2] https://elasticsecurity.com/2016/08/10/big-data-securityhow-to-secure-your-hadoop-cluster/
- [3] https://hadoop.apache.org/docs/r2.7.2/hadoopprojectdist/had oophdfs/TransparentEncryption.html# Configuration
- [4] https://www.slideshare.net/amalgjose89/deployment-andmanagement-of-hadoop-clusters
- [5] https://www.dezyre.com/article/hadoop-cluster-overviewwhat-it-is-and-howtosetup-one/356
- [6] https://www.slideshare.net/amalgjose89/deployment-andmanagement-of-hadoopclusters
- [7] https://www.slideshare.net/amalgjose89/deployment-andmanagement-of-hadoopclusters

- [8] http://www.xoomtrainings.com/blog/what-day-to-dayactivities-does-a-hadoop-admin-do
- [9] https://cfs22.simplicdn.net/ice9/free\_resources\_article\_t humb/Applications\_of\_big\_data\_infog raphic.png
- [10] https://mapr.com/hadoop-security-and-big-data-governance-mapr/

