GURU NANAK DEV INSTITUTE OF TECHNOLOGY

CLOUD COMPUTING

(PRACTICAL)

5th SEMESTER

SUBMITTEDBY:
DEEPAK BHARDWAJ

ENROLLMENT:
00726518118

(SOFTWARE DEVELOPMENT)

SUBMITTED TO:
ESHA SAXENA

INDEX

S.NO	TITLE	DATE	TEACHER SIGN
1.	To study about cloud types		
2.	To study about cloud models.		
3	Case study of IAAS.	11 sep.	
4.	Case study of PAAS (Facebook).	18 sep.	
5.	Installation and configuration of Hadoop.	25 sep.	
6.	Create an application for word count using hadoop.map/reduce.	09.oct.	
7	Databases in cloud computing	22 oct.	

EXPERIMENT-1

AIM - STUDY ABOUT TYPES OF CLOUD.

THEORY-

Cloud computing is an Internet-based computing in which shared the pool of resources are available over a broad network access, these resources can e provisioned or released with minimum management efforts and service provider interaction.

There are four types of cloud:

1. Public cloud
2. Private cloud
3. Hybrid cloud
4. Community cloud

Public cloud:

Public cloud are managed by third parties which provide cloud services over the internet to public, these services are available as pay-as-you-go billing mode.

They offer solutions for minimizing IT infrastructure costs and act as a good option for handling peak loads on the local infrastructure. They are a goto option for small enterprises, which are able to start their businesses without large upfront investments by completely relying on public infrastructure for their IT needs.

A fundamental characteristic of public clouds is multitenancy. A public cloud is meant to serve multiple users, not a single customer. A user requires a virtual computing environment that is separated, and most likely isolated, from other users.

Unlike private clouds, public clouds can save companies from the expensive costs of having to purchase, manage and maintain on-premises hardware and application infrastructure - the cloud service provider is held responsible for all management and maintenance of the system. Public

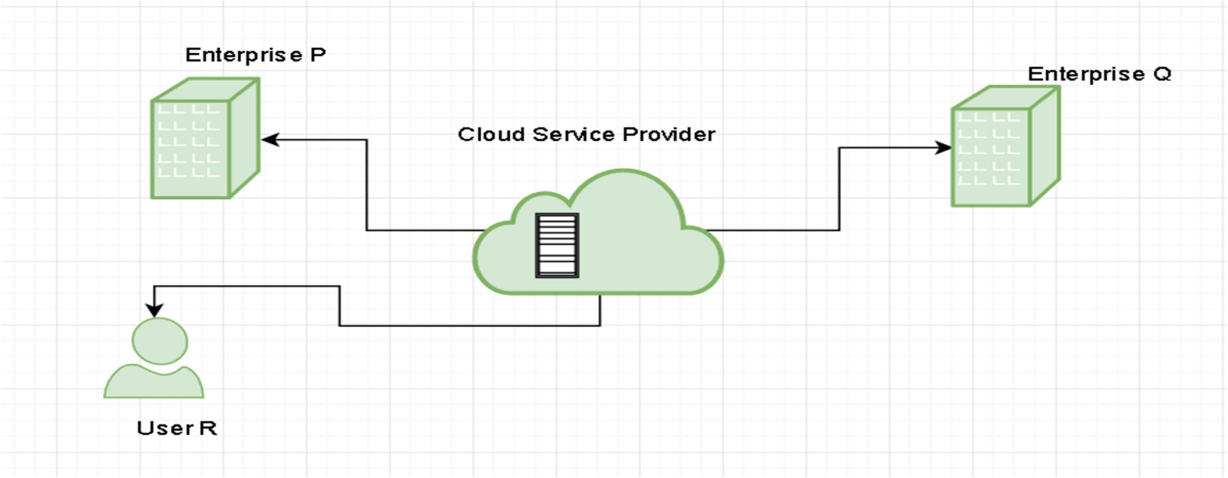
clouds can also be deployed faster than on-premises infrastructures and with an almost infinitely scalable platform. Every employee of a company can use the same application from any office or branch using their device of choice as long as they can access the Internet. While security concerns have been raised over public cloud environments, when implemented correctly, the public cloud can be as secure as the most effectively managed private cloud implementation if the provider uses proper security methods, such as intrusion detection and prevention systems (IDPS).

Advantages of Public Cloud

- **Scalability**
- **Cost**
- **Management**

Disadvantages

- **Security**
- **Flexibility:**
- **Compliance**



PRIVATE CLOUD

Private cloud is a type of cloud computing that delivers similar advantages to public cloud, including scalability and self-service, but through a proprietary architecture. Unlike public clouds, which deliver services to multiple organizations, a private cloud is dedicated to the needs and goals of a single organization.

How private clouds work

A private cloud is a *single-tenant environment*, meaning the organization using it (the tenant) does not share resources with other users. Those resources can be hosted and managed in a variety of ways. The private cloud might be based on resources and infrastructure already present in an organization's on-premises data center or on new, separate infrastructure, which is provided by a third-party organization. In some cases, the single-tenant environment is enabled solely using virtualization software. In any case, the private cloud and its resources are dedicated to a single user or tenant.

The private cloud is one of three general models for cloud deployment in an organization: public, private and hybrid (there is also multi-cloud, which is any combination of the three). All three models share common basic elements of cloud infrastructure. For example, all clouds need an operating system to function. However, the various types of software -- including virtualization and container software -- stacked on top of the operating system is what determines how the cloud will function, and distinguishes the three main models.

Advantages of a private cloud

- Enhanced security and privacy
- Improved reliability
- Improved performance:
- Increased flexibility
- Total control

Disadvantages of private cloud

- Cost
- Under-utilisation
- Platform scaling

HYBRID CLOUD

A hybrid cloud solution is one that combines many of the advantages of both public and private cloud systems, allowing information to be shared between on-premise systems and those maintained in the cloud. By acting as orchestration between all platforms, the hybrid cloud increases organizational flexibility. Remote workers, for example, would still be able to access

information located on local servers through the cloud. As companies grow in size and scope, this improved flexibility often becomes more and more essential.

This isn't just a small trend, either. The hybrid cloud market is projected to grow to $97.6 billion by 2023.1 Demand for hybrid cloud computing has never been higher, and luckily, companies are rising to the occasion by providing top-notch service.

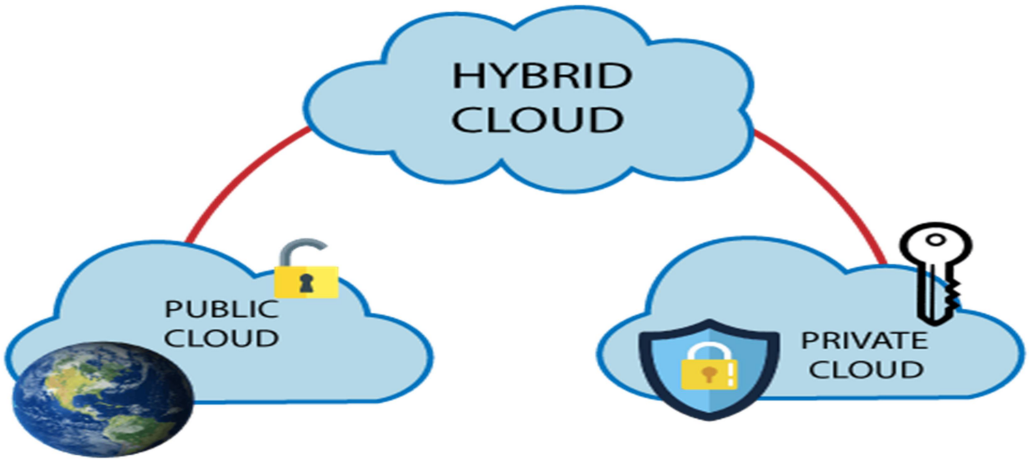
Hybrid Cloud Benefits

- Cost savings
- Unique balance of control, performance and scalability
- Speed of deployment
- Speed of deployment

Disadvantages

Overly complex security

Visibility issues



EXPERIMENT -2

AIM- To Study about cloud model.

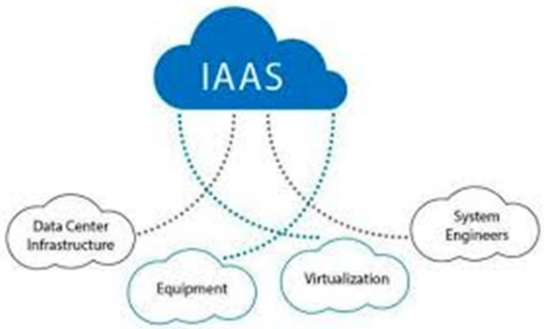
Theory-

There are the following three types of cloud service models -

1. Infrastructure as a Service (IaaS)
2. Platform as a Service (PaaS)
3. Software as a Service (SaaS)

IAAS

Infrastructure as a service (IaaS) is a form of cloud computing that provides virtualized computing resources over the internet. IaaS is one of the three main categories of cloud computing services, alongside software as a service (SaaS) and platform as a service (PaaS).



Characteristics of IaaS

There are the following characteristics of IaaS -

- Resources are available as a service
- Services are highly scalable
- Dynamic and flexible
- GUI and API-based access
- Automated administrative tasks

Example: DigitalOcean, Linode, Amazon Web Services (AWS), Microsoft Azure, Google Compute Engine (GCE), Rackspace, and Cisco Metacloud.

Advantages of IaaS cloud computing layer

There are the following advantages of IaaS computing layer -

1. Shared infrastructure

IaaS allows multiple users to share the same physical infrastructure.

2. Web access to the resources

Iaas allows IT users to access resources over the internet.

3. Pay-as-per-use model

IaaS providers provide services based on the pay-as-per-use basis. The users are required to pay for what they have used.

4. Focus on the core business

IaaS providers focus on the organization's core business rather than on IT infrastructure.

5. On-demand scalability

On-demand scalability is one of the biggest advantages of IaaS. Using IaaS, users do not worry about to upgrade software and troubleshoot the issues related to hardware components.

Disadvantages of IaaS cloud computing layer

1.Security is one of the biggest issues in IaaS. Most of the IaaS providers are not able to provide 100% security.

2. Maintenance & Upgrade

Although IaaS service providers maintain the software, but they do not upgrade the software for some organizations.

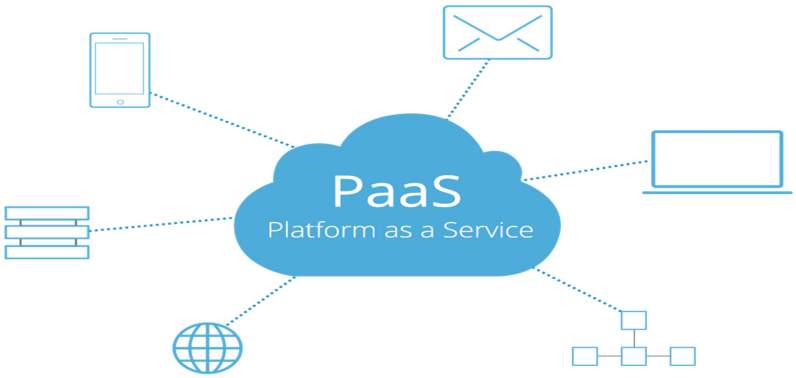
3. Interoperability issues

It is difficult to migrate VM from one IaaS provider to the other, so the customers might face problem related to vendor lock-in.

Security

Platform as a Service (PaaS)

PaaS cloud computing platform is created for the programmer to develop, test, run, and manage the applications.



Characteristics of PaaS

There are the following characteristics of PaaS -

- Accessible to various users via the same development application.
- Integrates with web services and databases.
- Builds on virtualization technology, so resources can easily be scaled up or down as per the organization's need.
- Support multiple languages and frameworks.
- Provides an ability to "**Auto-scale**".

Example: AWS Elastic Beanstalk, Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos, Magento Commerce Cloud, and OpenShift.

Software as a Service (SaaS) SaaS is also known as "**on-demand software**". It is a software in which the applications are hosted by a cloud service provider. Users can access these

applications with the help of internet connection and web browser.

Characteristics of SaaS

There are the following characteristics of SaaS -

- Managed from a central location
- Hosted on a remote server
- Accessible over the internet
- Users are not responsible for hardware and software updates. Updates are applied automatically.
- The services are purchased on the pay-as-per-use basis

Example: BigCommerce, Google Apps, Salesforce, Dropbox, ZenDesk, Cisco WebEx, ZenDesk, Slack, and GoToMeeting.

EXPERIMENT-3

AIM- Case study of iaas.(FUJITSU)

THEORY-

CHALLENGE

Following the merger of Fujitsu companies in the UK and Ireland, Fujitsu's IT department conducted a comprehensive review of its IT infrastructure looking at a number of factors including operational flexibility, the consolidation of systems from previously separate businesses, for improved reliability, cost effectiveness and a much lower environmental impact. The internal IT team knew that a hardware refresh could not be delayed any further as a significant number of servers were over eight years old. Additionally a number of the systems were located in a co-location provider's data centre, whose rates would increase at the end of the year, adding to the financial burden. The team reviewing the forward roadmap considered the possibility of using cloud computing and soon realised that it would be rapid to deploy and would provide quick wins in terms of cost and energy efficiency. The environment would have to be sufficiently secure and reliable with equivalent end user response times to achieve the business requirements. "We wanted to reduce the cost of our IT infrastructure, upgrade performance and IaaS was clearly the best way to achieve these objectives," explains Sean Barker, Head of Architecture strategy team at Fujitsu UK. "We would also provide feedback to our IaaS development team to enhance the service for Fujitsu's customers."

SOLUTION

Fujitsu IaaS is the cornerstone in Fujitsu's cloud services offering for customers, announced in November 2009 the first customer started using the service for production workloads in April 2010. Fujitsu's own use of the IaaS hosting environment started a couple of months later, providing a seamless extension to the IT infrastructure owned by the IT department.

IaaS provides the flexibility and cost effectiveness anticipated by the business and delivers the security, resilience and performance required to host live business systems. The flexibility and high utilisation it provides, lowers Fujitsu's infrastructure costs, and makes expenditure on additional data centre premises unnecessary. Using IaaS to host the business systems reduces the total amount of physical server capacity required. Fujitsu had a few separate virtual server environments and numerous dedicated legacy servers hosting single applications where the utilisation was often less than 10%.

Barker provides an illuminating analogy to describe the benefits of IaaS: "We think of it in terms of vehicles. Previously, everyone had their own car which was for most of the time was left unused in the car park. But with a car pool system we would need far fewer vehicles in total and whenever I need a car, I take it from the pool and not only that, I can choose a large or small car depending on my needs. That's how IaaS works – instead of owning IT, use what you want when you want it. That means a much improved utilisation and lower costs for those workloads where dedicated IT is not needed all the time." Fujitsu's internal IT department could see the advantage of pooling resources to achieve utilisation levels of over 70% with the associated savings in the number of servers required, plus a lower power consumption and hence a lower environmental impact. The team had the choice of designing and building a common shared pool themselves or using one that is already available. Using Fujitsu IaaS saved some development and set up costs and enabled a higher level of utilisation though a shared pool, shared with other organisations to achieve a greater efficiency than a private pool would achieve. The system images and connecting network had to be designed and built, but the infrastructure and internal data centre network is ready; built, installed in a managed data centre, and maintained by the IaaS Shared Service team. Having the infrastructure managed by the IaaS team saved the IT department from the low level infrastructure management tasks. "The beauty of our new shared infrastructure is that we no longer need to worry about provisioning hardware for specific applications or storage because we now have the option of using three IaaS options; virtual machines in shared or dedicated servers, or physical servers, all on pay-as-you-go terms" adds Barker. "Every new internal project now

automatically uses IaaS which is making it much easier for us to get the most out of our IT infrastructure." Moving to a shared pool enables the company to consolidate the systems previously dispersed across five data centres, down to two data centres near London. These are twinned to provide operational continuity for business critical systems. This yielded rack space in Fujitsu data centres for use by customers and assisted in the rationalisation plans to close one Fujitsu data centre and move out of another co-location data centre where the co-locator was increasing their accommodation and power rates. Overall the savings for the IT budget amount to a hosting charge cost reduction of 20% per server and additional benefits in other parts of the business. The first phase deployment includes; IT enterprise management tools, an extranet service, virtual firewalls, anti-virus management, news service, data warehouse, and application development projects. The first production workloads include systems for; business reporting, warehouse management, HR document management, sales information, CAD, BlackBerry service, and shared SQL. The second phase includes the migration of approximately 250 production workloads as well as mail and messaging services to a Fujitsu Mail-as-aService platform built on the IaaS infrastructure. The migration priority for the production workloads is not a pure IT decision and involvves consultation and scheduling with the associated business operations to arrive at a workable roadmap and timeframe. Not all environments are suitable for deployment on virtual machines in a shared IaaS pool, either because of licensing limitations by the software vendor or for technical reasons such as for high performance database servers which are currently more efficiently operated without a virtualisation layer. Such workloads will be deployed on a physical server in the IaaS pool to achieve the right balance of system performance and cost effectiveness.

IMPACT

The IaaS solution is changing the way Fujitsu works and brings with it a number of benefits. "Taking a cloud computing approach has multiple advantages. Firstly we can roll out new applications and services much more rapidly," adds Barker. "Previously, requisitioning a new server could take months, from initial paperwork to raising a purchase order to delivery, installation and configuration. That's not only a significant delay for new deployments but a lot of man hours. Now, we have access to a server in a matter of hours and an application platform can be deployed across the estate in hours not months. "We don't need to commit in advance when

planning hardware or software deployment so over-provisioning is a thing of the past." Electricity consumption will be reduced by 85%, not just from the increased utilisation, but also because IaaS servers and storage systems are energy efficient and they are installed in Fujitsu's most efficient data centres. London North Data Centre for example has a Power Usage Efficiency rating of 1.4. For the service and scope of this programme we expect the carbon footprint to reduce from 1,500 tons per year to 30-15 tons per year as each server is decommissioned. As Fujitsu continues to reap the benefits of IaaS, it is looking at other areas where cloud computing could provide savings and better performance, for example one area of Fujitsu's expertise is Salesforce.com. The internal IT department are reviewing the current Siebel CRM system which requires a major upgrade, and are instead investigating a move to Salesforce.com "Third party hosted applications can be scalable, robust and offer consistent performance," explains Steve Ranaghan, Systems Strategy Manager at Fujitsu. "The 'per person per month' pricing model means that the systems expenditure is aligned with business use and means that the systems are readily available to additional users as and when required." David Smith, CIO, concludes, "As you would expect from an IT service provider our investment focus always prioritises serving our clients over our internal IT requirements;' the shoemaker's children with no shoes' cliché resonates. It is therefore imperative that when we make an investment that we do so in a way that maximises our return, minimises the funding required and delivers a capability that enables the agility that my company demands. Our deployment of Fujitsu IaaS has delivered on all fronts and has given us a flexible computing platform that meets the company's needs today and for years to come."

EXPERIMENT-4

AIM-Case study of facebook on paas

THEORY-

Organization

Ravio, creator of the blockbuster “Angry Birds” game series, turned to Google App Engine when it came time to adapt its mobile apps for web browsers. The Finland based company needed a platform that could support explosive demand and provide robust capabilities to deliver a superior user experience. Google App Engine provides both while requiring minimal maintenance, which gives the company’s developers time to focus in improving the games.

Challenge

Ravio knew that bringing its game online presented an enormous opportunity. In early2011, a development team began planning a version of “Angry Birds” for Google Chrome. The company wanted to launch the game ar Google’s annual I/O conference that spring just a few months away.

The developers needed platform that would scale effortlessly. The mobile app had already hit more than 140 millions downloads and the team expected demand for the free online version to be overwhelming.They also wanted a low maintenance system that would make it easy to update features and bring new titles online.

Solution

The developers chose Google app engine to build the game because they knew it would allow them to work quickly and provide the scalability needed to support an enormous user base."Angry Birds Chrome " Finished on schedule, followed by other titles such as "Angry Birds Google+" and "Angry Birds Friends". Rovio also created customized versions for companies, sports teams and other partners.

"Google app engine allows us to launch games very quickly with teams of one or two developers per game,"says Stefan Hauk , Rovio's lead server developer for web games. "Because Google manages allthe servers, there is little required of us in terms of maintenance".

Hauk and his fellow developers use a number use a number of App Engine features to improve the games, including:

- **High-Replication Datastore** for scalable, long term storage of game data.
- **Memcache API** to boost performance by providing temporary, high- speed data access through a high performance memory cache.
- **Task queues** to run certain complex operations in the background, improving game responsiveness for users.
- **User API** to authenticate users with their Google usernames and password, which provides a seamless experience when accessing a game.

App Engine allows the developers to add new features easily and continuously improve the game for users. They can deploy new versions with a single command and switch back to the previous versions if needed. They can also rely on App Engine to scale automatically to support heavy demand from the moment the games launch.

“Because our web games are popular immediately, we don’t have the option of scaling them over time,” Hauk says. “Google App Engine makes the process painless, since it can instantly launch as many servers as we need and scale back down has passed its usage peak.”

Results

Milllions of gamers have flocked to Rovio’s web games since their launch. The company’s most popular offering, the Facebook game “Angry Birds Friends,” logs more than 13 millions users every month. Since the developers don’t need to install or maintain hardware, they candevote their attention to enhancing the games which have received overwhelmingly positive reviews.

“Google App Engine automates a lot of processes, which has made our jobs easier,” Hauk says.”At other companies I’ve been with there was always a need to be on call after hours to deal with server problems. This isn’t necessary here, because Google App Engine just works.”

The ability to build and deploy quickly has allowed Rovio to capitalize on expanding its audience and to act on business opportunities.

“There have been times when we’ve been asked to build a customized game in a week or two,” Hauk says, “we know that App Engine will enable us to do this and that it will scale for us no matter how many users we get.”

EXPERIMENT-5

AIM- Installation and configuration of Hadoop.

THEORY-

Install Hadoop: Setting up a Single Node Hadoop Cluster

Install Hadoop

Step 1: Click here to download the Java 8 Package. Save this file in your home directory.

Step 2: Extract the Java Tar File.

Command: tar -xvf jdk-8u101-linux-i586.tar.gz

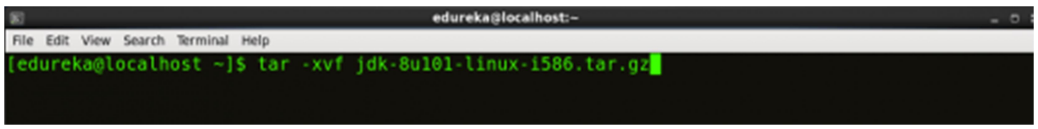


Fig: Hadoop Installation – Extracting Java Files

Step 3: Download the Hadoop 2.7.3 Package.

Command: wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz

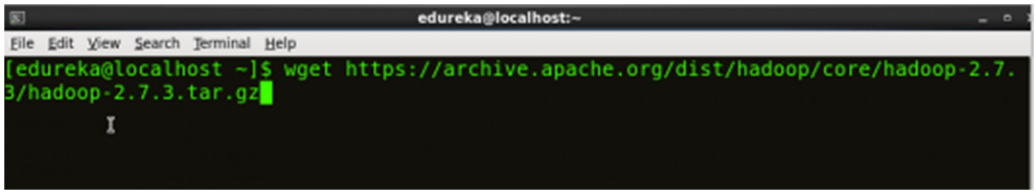


Fig: Hadoop Installation – Downloading Hadoop

Step 4: Extract the Hadoop tar File.

Command: tar -xvf hadoop-2.7.3.tar.gz

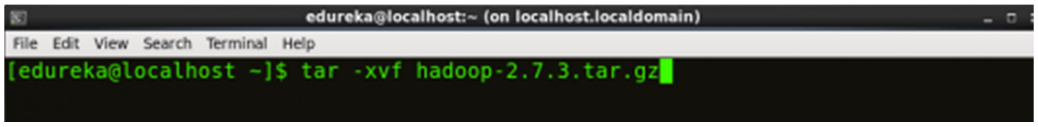
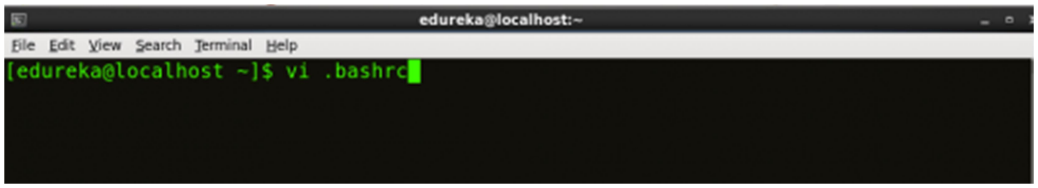


Fig: Hadoop Installation – Extracting Hadoop Files

Step 5: Add the Hadoop and Java paths in the bash file (.bashrc).

Open. bashrc file. Now, add Hadoop and Java Path as shown below.

Command: vi .bashrc



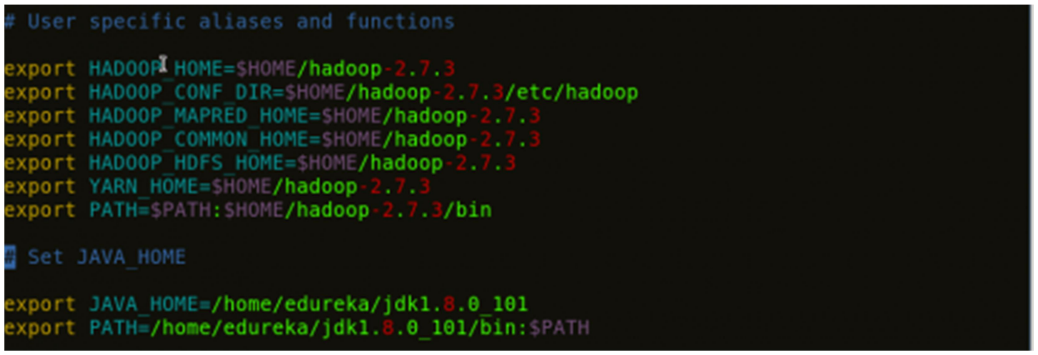


Fig: Hadoop Installation – Setting Environment Variable

Then, save the bash file and close it.

For applying all these changes to the current Terminal, execute the source command.

Command: source .bashrc

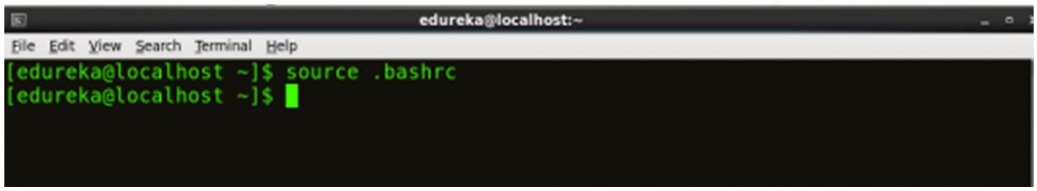


Fig: Hadoop Installation – Refreshing environment variables

To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

Command: java -version

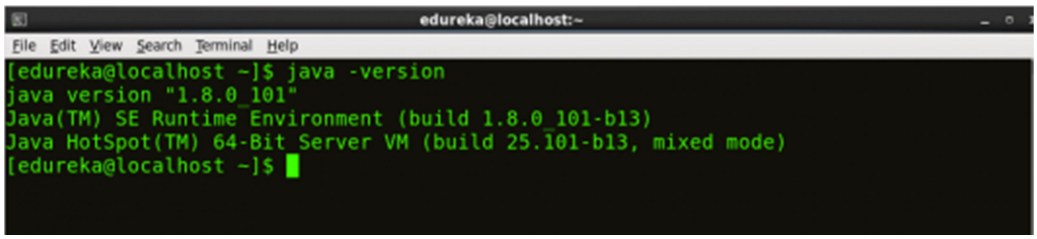
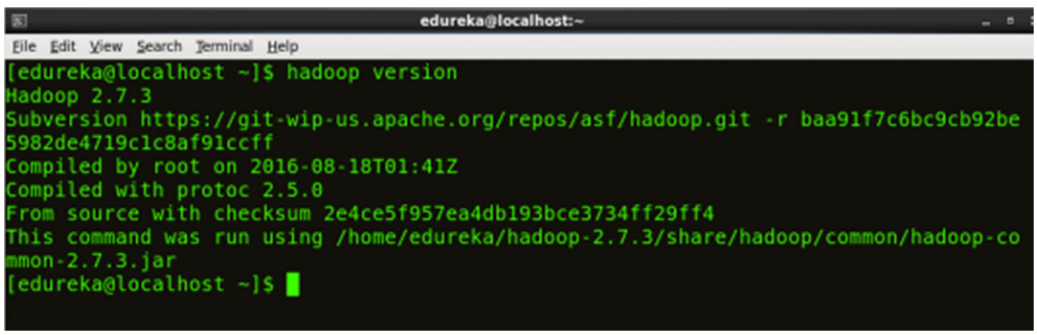


Fig: Hadoop Installation – Checking Java Version

Command: hadoop version



To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

Command: java -version

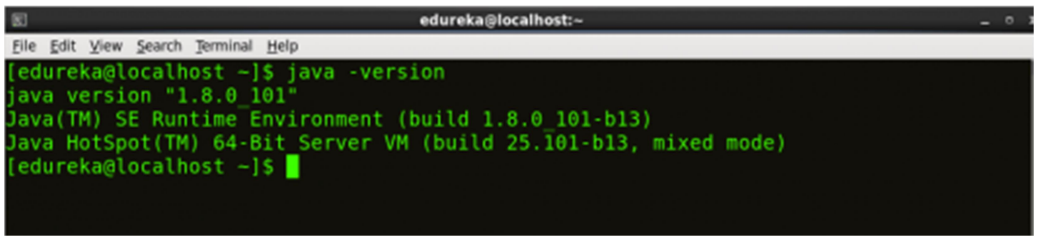
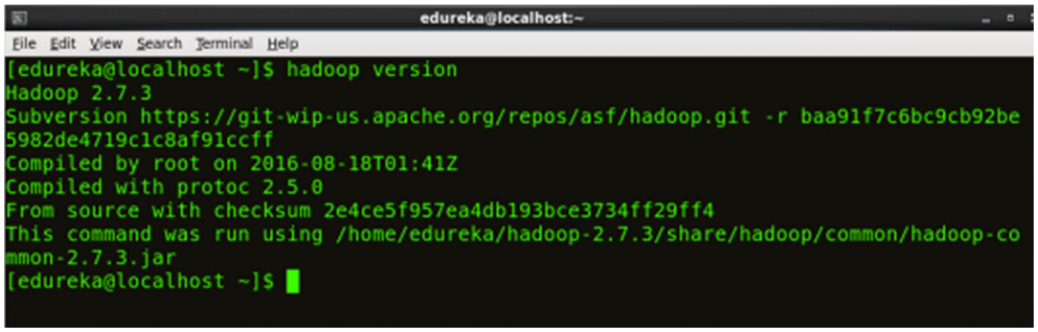


Fig: Hadoop Installation – Checking Java Version

Command: hadoop version



To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

Command: java -version

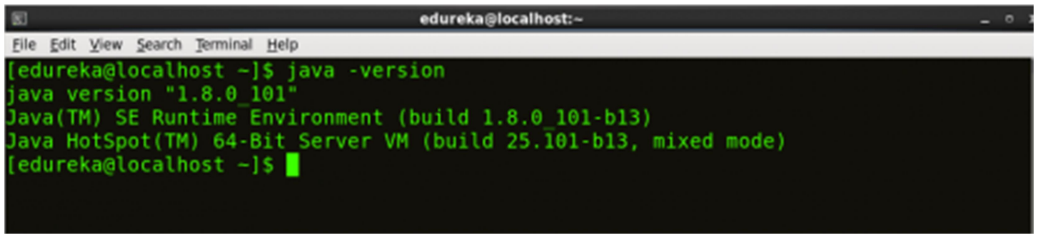


Fig: Hadoop Installation – Checking Java Version

Command: hadoop version

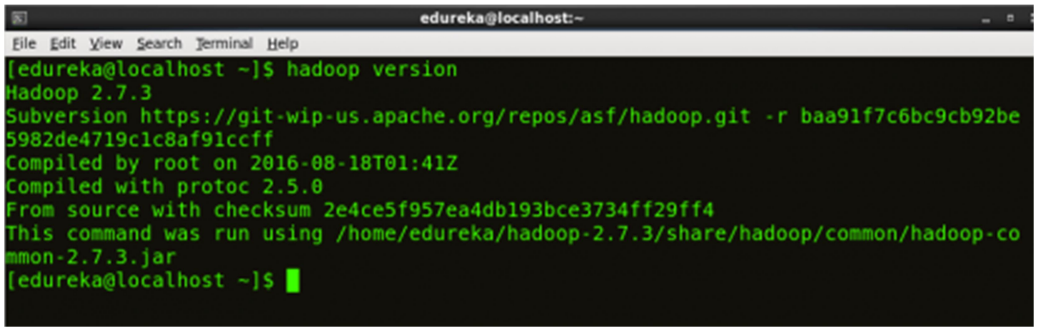


Fig: Hadoop Installation – Configuring hdfs-site.xml

```
1   <?xml version="1.0" encoding="UTF-8"?>
2   <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3   <configuration>
4   <property>
5   <name>dfs.replication</name>
6   <value>1</value>
7   </property>
8   <property>
9   <name>dfs.permission</name>
10  <value>false</value>
11  </property>
12  </configuration>
```

Step 9: Edit the mapred-site.xml file and edit the property mentioned below inside configuration tag:

mapred-site.xml contains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.

In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template.

Command: cp mapred-site.xml.template mapred-site.xml

Command: vi mapred-site.xml.

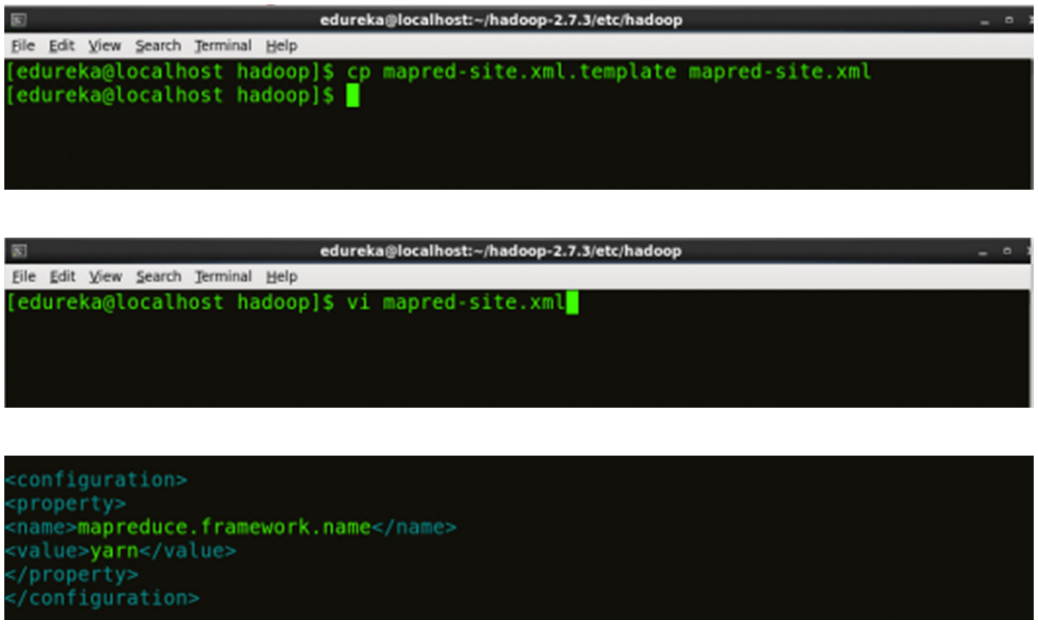


Fig: Hadoop Installation – Configuring mapred-site.xml

```
1
2   <?xml version="1.0" encoding="UTF-8"?>
3   <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
4
5   <configuration>
6   <property>
7
8   <name>mapreduce.framework.name</name>
    <value>yarn</value>
    </property>
```

</configuration>

Step 10: Edit yarn-site.xml and edit the property mentioned below inside configuration tag:

yarn-site.xml contains configuration settings of ResourceManager and NodeManager like application memory management size, the operation needed on program & algorithm, etc.

Command: vi yarn-site.xml

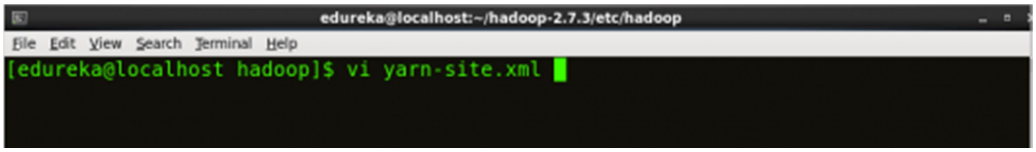


Fig: Hadoop Installation – Configuring yarn-site.xml

```
1
2  <?xml version="1.0">
3  <configuration>
4  <property>
5  <name>yarn.nodemanager.aux-services</name>
6  <value>mapreduce_shuffle</value>
7  </property>
8  <property>
9  <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
10 <value>org.apache.hadoop.mapred.ShuffleHandler</value>
11 </property>
   </configuration>
```

Step 11: Edit hadoop-env.sh and add the Java Path as mentioned below:

hadoop-env.sh contains the environment variables that are used in the script to run Hadoop like Java home path, etc.

Command: vi hadoop–env.sh

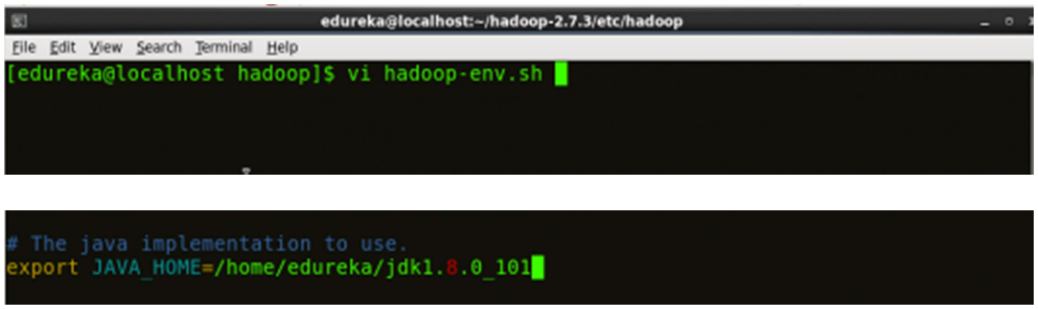


Fig: Hadoop Installation – Configuring hadoop-env.sh

Step 12: Go to Hadoop home directory and format the NameNode.

Command: cd

Command: cd hadoop-2.7.3

Command: bin/hadoop namenode -format

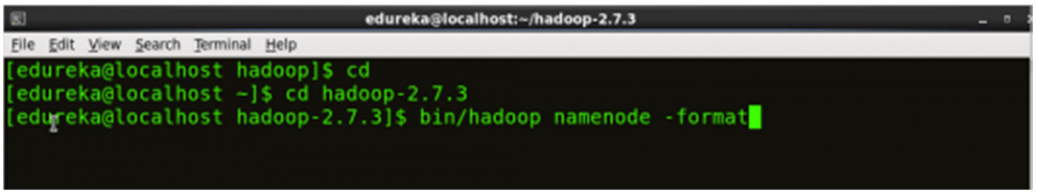


Fig: Hadoop Installation – Formatting NameNode

This formats the HDFS via NameNode. This command is only executed for the first time. Formatting the file system means initializing the directory specified by the dfs.name.dir variable.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

Step 13: Once the NameNode is formatted, go to hadoop-2.7.3/sbin directory and start all the daemons.

Command: cd hadoop-2.7.3/sbin

Either you can start all daemons with a single command or do it individually.

Command: ./start-all.sh

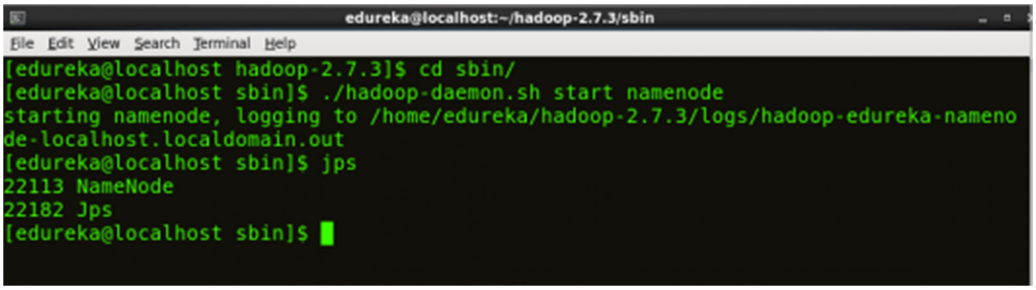
The above command is a combination of start-dfs.sh, start-yarn.sh & mr-jobhistory-daemon.sh

Or you can run all the services individually as below:

Start NameNode:

The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and tracks all the file stored across the cluster.

Command: ./hadoop-daemon.sh start namenode



Start DataNode:

On startup, a DataNode connects to the Namenode and it responds to the requests from the Namenode for different operations.

Command: ./hadoop-daemon.sh start datanode

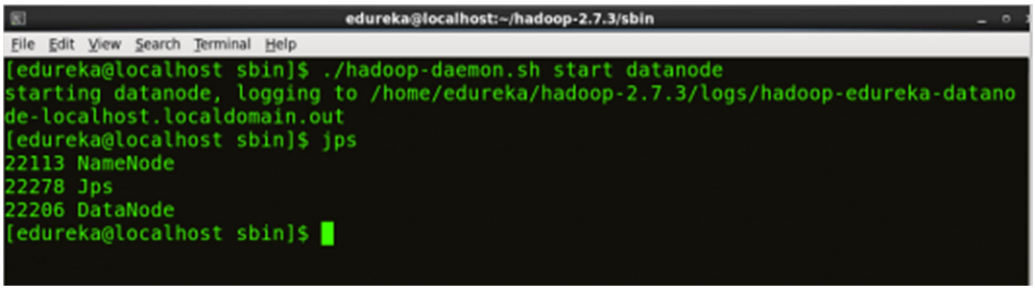


Fig: Hadoop Installation – Starting DataNode

Start ResourceManager:

ResourceManager is the master that arbitrates all the available cluster resources and thus helps in managing the distributed applications running on the YARN system. Its work is to manage each NodeManagers and the each application's ApplicationMaster.

Command: ./yarn-daemon.sh start resourcemanager

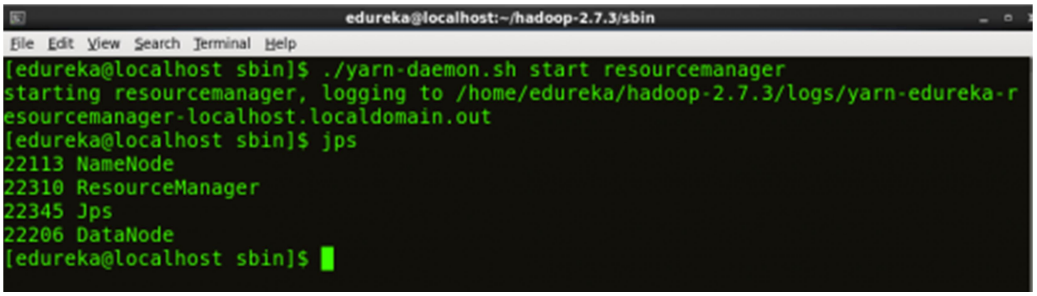
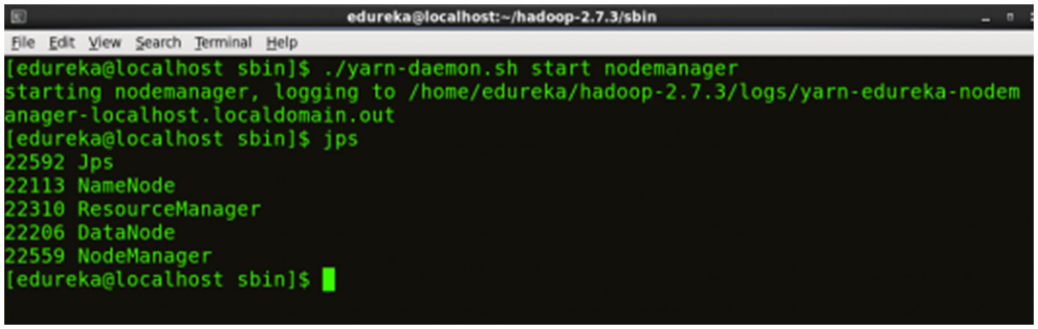


Fig: Hadoop Installation – Starting ResourceManager

Start NodeManager:

The NodeManager in each machine framework is the agent which is responsible for managing containers, monitoring their resource usage and reporting the same to the ResourceManager.

Command: ./yarn-daemon.sh start nodemanager



Start JobHistoryServer:

JobHistoryServer is responsible for servicing all job history related requests from client.

Command: ./mr-jobhistory-daemon.sh start historyserver

Step 14: To check that all the Hadoop services are up and running, run the below command.

Command: jps

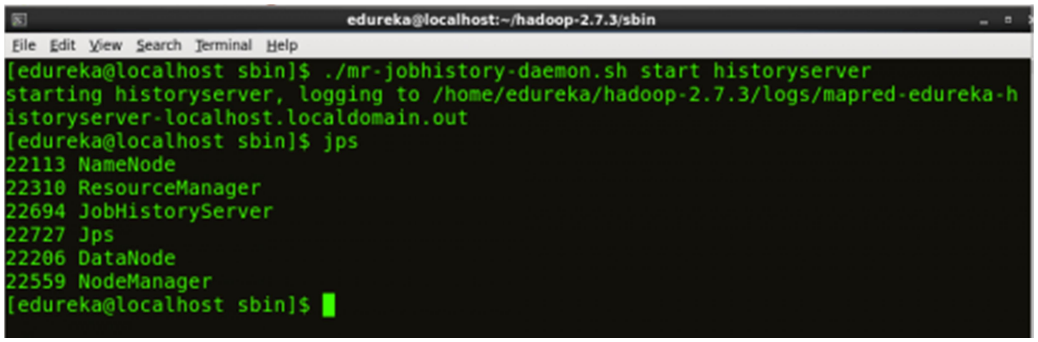


Fig: Hadoop Installation – Checking Daemons

Step 15: Now open the Mozilla browser and go to localhost:50070/dfshealth.html to check the NameNode interface.

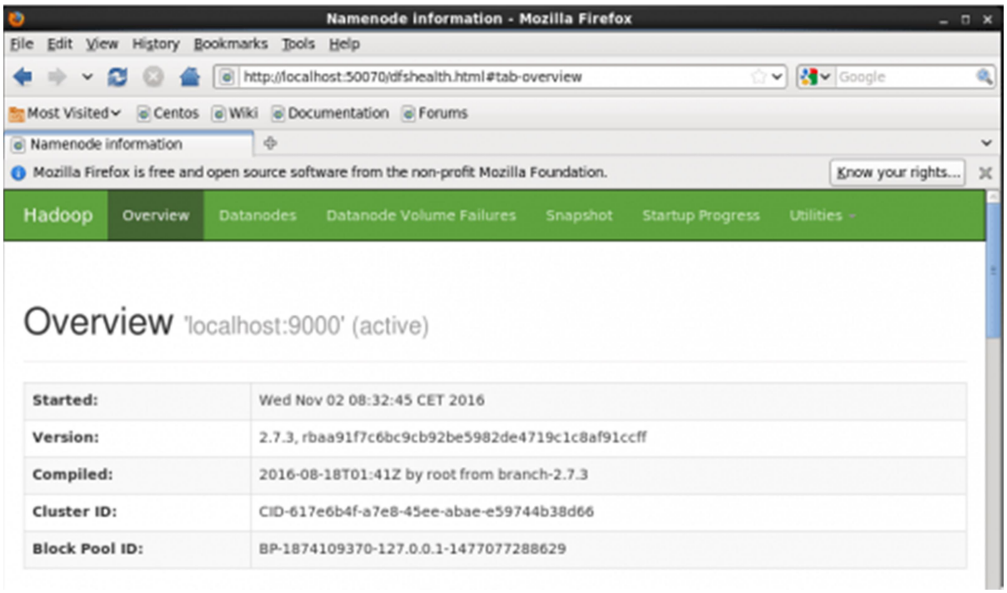


Fig: Hadoop Installation – Starting WebUI

EXPERIMENT-6

AIM- Create an application for word count using hadoop.map/reduce.

THEORY-

Hadoop MapReduce

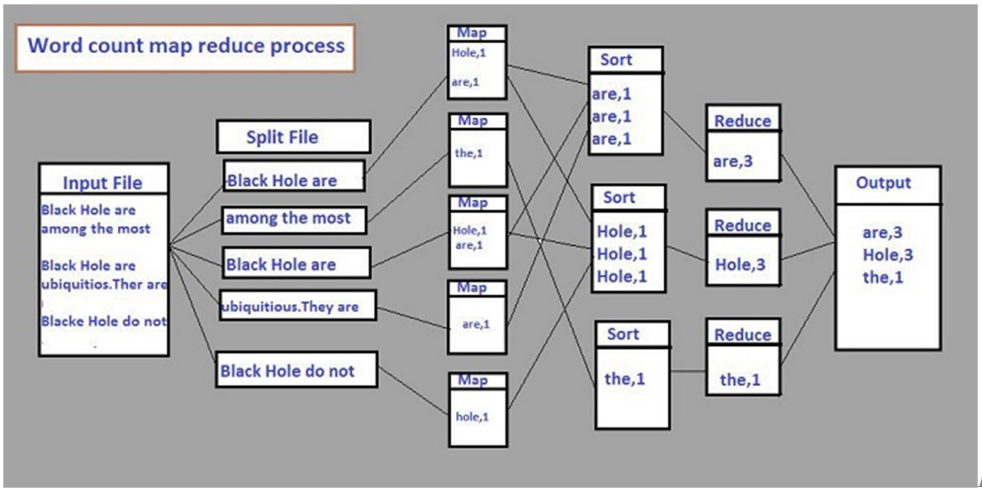
Hadoop MapReduce is a system for parallel processing which was initially adopted by Google for executing the set of functions over large data sets in batch mode which is stored in the fault-tolerant large cluster.

The input data set which can be a terabyte file broken down into chunks of 64 MB by default is the input to Mapper function. The Mapper function then filters and sort these data chunks on Hadoop cluster data nodes based on the business requir**ement.**

After the distributed computation is completed, the output of the mapper function is passed to reducer function which combines all the elements back together to provide the resulting output.

An example of Hadoop MapReduce usage is “word-count” algorithm in raw Java using classes provided by Hadoop libraries. Count how many times a given word such as “are”, “Hole”, “the” exists in a document which is the input file.

To begin, consider below figure, which breaks the word-count process into steps.



Hadoop MapReduce Word Count Process

The building blocks of Hadoop MapReduce programs are broadly classified into two phases, the map and reduce.

Both phases take input data in form of (key, value) pair and output data as (key, value) pair. The mapper program runs in parallel on the data nodes in the cluster. Once map phase is over, reducer run in parallel on data nodes.

The input file is split into 64 MB chunk and is spread over the data nodes of the cluster. The mapper program runs on each data node of the cluster and generates (K1, V1) as the key-value pair.

Sort and shuffle stage creates the iterator for each key for e.g. (are, 1,1,1) which is passed to the reduce function that sums up the values for each key to generate (K2, V2) as output. The illustration of the same is shown in above figure (word count MapReduce process).

Hadoop MapReduce Algorithm for Word Count Program

1. Take one line at a time
2. Split the line into individual word one by one (tokenize)
3. Take each word

4. Note the frequency count (tabulation) for each word
5. Report the list of words and the frequency tabulation

Hadoop MapReduce Code Structure for Word Count Program

Step 1

Import Hadoop libraries for Path, configuration, I/O, Map Reduce, and utilities.

```
import org.apache.hadoop.mapred.*;

import org.apache.hadoop.io.*;

import org.apache.hadoop.util.*;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.conf.*;
```

Step 2

The summary of the classes defined in the “word count map reduce” program is as below :

```
public class WordCount {



public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

----------

----------

--------

}
```

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

.....

.....

.....



}



public static void main(String[] args) throws Exception {

=========================

======================

================

}
```

We have created a package in the eclipse and defined a class named “WordCount”. The “WordCount” class has two nested class and one main class. “Mapper” and “Reducer” are the reserved keywords.

The source code for the same is written by Hadoop developer. We are extending the “Mapper” and “Reducer” class by the “Map” and “Reduce” respectively using inheritance.

Let us understand what is LongWritable, Text, IntWritable. For the same, we need to first understand serialization and de-serialization in java.

Object serialization is a mechanism where an object can be represented as a sequence of bytes that includes the object's data as well as information about the object's type and the types of data stored in the object.

The serialized object is written in a file and then de-serialized to recreate the object back into memory.

For example word "Hai" has a serializable value of say "0010110" and then once it is written in a file, you can de-serialized back to "Hai".

In Hadoop MapReduce framework, mapper output is feeding as reducer input. These intermediate values are always in serialized form.

Serialization and de-serialization in java are called as Writable in Hadoop MapReduce programming. Therefore, Hadoop developers have converted all the data types in serialized form. For example, Int in java is IntWritable in MapReduce framework, String in java is Text in MapReduce framework and so on.

The input and output of the mapper or reducer is in (key, value) format. For example, we have a file which contains text input and text outputs say the sample data as (1, aaa). The key is considered to be the precision of input data. The precision for (1, aaa) is defined as "01234". 0 for "1", 1 for "," and so on which makes it to "01234".

Therefore, for a text input/output file, the precision of first value is considered to be as key and the rest are values. In this case, "0" is considered as the key while as "(1, aaa)" as value.

Similarly, if you have another data in the file say (2, bbb). The precision for (1, bbb) is defined as "56789". Key here will be 5 and the value will be (1, bbb).

Now, let us try to understand the below with an example:

```
Mapper<LongWritable, Text, Text, IntWritable> {
```

Consider, we have the first line in the file as “Hi! How are you”.

The mapper input key value is (0, Hi!), (4, How), (8, are), (12, you). Therefore, the key generated by mapper class has a data type “LongWritable” i.e. the first parameter and the value generated by mapper class is “Text”.

The mapper output value would be the word and the count of the word i.e. (Hi!,1), (How,1), (are,1), (you, 1).

If the word “are” repeated twice in the sentence then the mapper output would be (are,1,1). Hence, the key of the mapper output is “Text” while as the value is “IntWritable”. This output to the mapper is getting

This output to the mapper is getting fed as the input to the reducer. Therefore, if the reducer input is (are, 1, 1) then the output of the reducer will be (are,2). Here, the reducer

output data type has the key as “Text” and value as “IntWritable”.

Step 3

Define the map class. The key and value input pair have to be serializable by the framework and hence need to implement the Writable interface.

Output pairs do not need to be of the same types as input pairs. Output pairs are collected with calls to context.

Inside the static class “map” we are declaring an object with the name “one” to store the incremental value of the given word and the particular word is stored in the variable named “word”.

```
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
```

```
private final static IntWritable one = new IntWritable(1);
```

```
private Text word = new Text();

public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

String line = value.toString();

StringTokenizer tokenizer = new StringTokenizer(line);

while (tokenizer.hasMoreTokens()) {

word.set(tokenizer.nextToken());

context.write(word, one);

}

}

}
```

The above piece of code takes each line as an input and stores it into the variable “line”. StringTokenizer allows an application to break a string into tokens. For example:

StringTokenizerst = new StringTokenizer(“my name is kumar”,” “);

The output of the above line will be: my

name

```
                    is

                    kumar
```

If the “tokenizer” variable has more number of tokens to count then the while loop will get open. The context will take care of executing the for loop i.e. to read line by line of the file and store the output as the particular word and their occurrences. For example: if you have “hai, hai, hai” then the context will store (hai, 1, 1, 1)

Step 4

Reduce class will accept shuffled key-value pairs as input.The code then totals the values for the key-value pairs with the same key and outputs the totaled key-value pairs; e.g. <word,3>

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

public void reduce(Text key, Iterable<IntWritable> values, Context context)

throws IOException, InterruptedException {

int sum = 0;

for (IntWritableval : values) {

sum += val.get();

}

context.write(key, new IntWritable(sum));

}

}
```

Step 5

The main method sets up the Map Reduce configuration by defining the type of input. In this case, the input is text.The code then defines the Map, Combine, and Reduce classes, as well as specifying the input/output formats.

```
public static void main(String[] args) throws Exception {



Configuration conf = new Configuration();

Job job = new Job(conf, "wordcount");//Name for the job is “wordcount”

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(IntWritable.class);

job.setMapperClass(Map.class); // Mapper Class Name

job.setReducerClass(Reduce.class); //Reducer Class Name

job.setInputFormatClass(TextInputFormat.class);

job.setOutputFormatClass(TextOutputFormat.class);

FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

job.waitForCompletion(true);



}

}
```

Step 6

The full Java code for the “word count” program is as below:

```
import java.io.IOException;

import java.util.*;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.conf.*;

import org.apache.hadoop.io.*;

import org.apache.hadoop.mapreduce.*;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;



public class WordCount {

public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {



private final static IntWritable one = new IntWritable(1);

private Text word = new Text();
```

```
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

String line = value.toString();

StringTokenizer tokenizer = new StringTokenizer(line);

while (tokenizer.hasMoreTokens()) {

word.set(tokenizer.nextToken());

context.write(word, one);

}

}

}

public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

public void reduce(Text key, Iterable<IntWritable> values, Context context)

throws IOException, InterruptedException {

int sum = 0;

for (IntWritableval : values) {

sum += val.get();

}
```

```
context.write(key, new IntWritable(sum));

}

}

public static void main(String[] args) throws Exception {



Configuration conf = new Configuration();

Job job = new Job(conf, "wordcount");

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(IntWritable.class);

job.setMapperClass(Map.class);

job.setReducerClass(Reduce.class);

job.setInputFormatClass(TextInputFormat.class);

job.setOutputFormatClass(TextOutputFormat.class);

FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

job.waitForCompletion(true);

}

}
```

EXPERIMENT-7

AIM- Databases in cloud computing.

THEORY-

Cloud computing is basically the commodification of data storage and computing time with the help of standardized technologies. Cloud databases are databases that run on cloud computing platforms such as Salesforce, GoGrid, Rackspace, and Amazon EC2. Users can independently run cloud databases on the cloud with either of the two deployment models - virtual-machine image or by purchasing access to database services that are maintained by cloud database providers. Although cloud databases provide significant benefits over traditional deployments, sometimes traditional architectures should be integrated with cloud platforms. However, cloud databases have been providing a comprehensive solution for every customer who demands custom-built, high-performance infrastructure for a relational database supported and backed by MySQL-specialized engineers. Cloud databases are best suited for customers who are focused on getting their applications developed without getting hassled with infrastructure-related issues.

1. Amazon Web Services

Amazon offers a wide array of cloud database services, which includes NoSQL as well as relational databases. Amazon RDS – Relational Database Service runs on either Oracle, SQL, or MySQL server instances whereas Amazon SimpleDB is primarily a schema-less database that is meant to handle smaller workloads. Amazon DynamoDB falls on the NoSQL databases, which is a Solid State Drive – SSD - that is capable of automatically replicating workloads across three

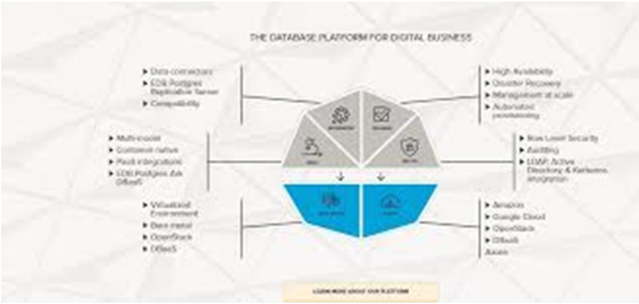
different availability zones. According to AWS CTO Werner Vogels, DynamoDB is the fastest growing database service in the history of AWS. Furthermore, Amazon offers supplementary data-management services such as Redshift – a data warehouse and Data Pipeline – a data integrating service for easier data management.

2. SAP

SAP, the giant in offering enterprise software, now offers a cloud database platform called HANA for complementing the on-premise database-related tools of an organization. One of the major database tools complemented by SAP HANA includes Sybase, and this tool is available in the AWS cloud.

3. EnterpriseDB

Although EnterpriseDB was designed to focus on open-source PostgreSQL databases, its true claim-to-fame was its capability to work on Oracle database applications. The Postgres Plus, Advanced Server of EnterpriseDB, enables businesses to use applications that are designed for Oracle on-premise databases, which run in cloud from HP and AWS. It comprises of scheduled backups as well as binary replications.



4. Garantia Data

Garantia Data has been offering gateway service for customers who prefer running Memcached (in-memory NoSQL) databases as well as open-source Redis in the public cloud of AWS. The software of Garantia enables easy configuration of open-source data

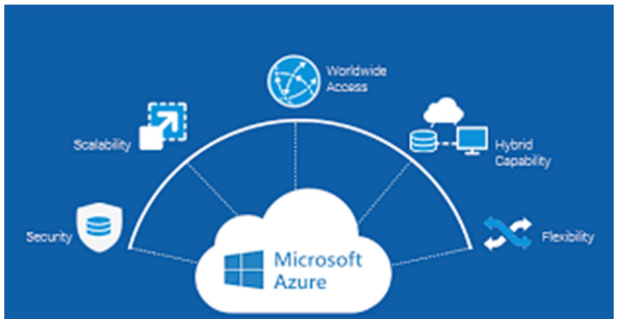
Garantia Data
The In-Memory NoSQL Cloud

5. Cloud SQL by Google

This database service comprises of two main products - Cloud SQL that describes a relational database and BigQuery analysis tool, which can run queries on vast sets of data stored in the cloud.

6. Azure by Microsoft

Azure cloud-computing platform offered by Microsoft offers a relational database that enables users to access SQL databases either on Microsoft cloud or on hosted servers on virtual machines.



7. Rackspace

Databases offered by Rackspace come in managed or hosted cloud databases. Rackspace provides high performance and incorporates a SAN storage network based on the OpenStack platform.