## Private Cloud Setup with User Storage Management

**Course Title: Distributed System**

**Submitted To: Dr. Ayaz ul Hassan Khan**

**Submitted By:**

**Syed Taqi Ali (IS-009)**

**Bilal Ahmed Khan (IS-022)**

**Muhammad Anus (IS-023)**

## Private Cloud Setup with User Storage Management

Introduction

With the rapid growth of the internal volume of enterprise data, it causes many problems in terms of storage, such as the lack of adequate hardware storage devices, high maintenance costs, and incomplete storage environment and so on. Cloud Computing technology can provide customers with the corresponding demand services that can effectively address these issues.

Cloud Computing provides us by which we can access the applications as utilities, over the Internet. It allows us to create, configure, and customize applications online. With Cloud Computing users can access database resources via the internet from anywhere for as long as they need without worrying about any maintenance or management of actual resources.

However there are security and privacy concerns to be addressed before customers readily outsource their sensitive data. To be effective, a cloud storage service should provide, at a minimum, features and security guarantees like confidentiality, integrity, availability, reliability, efficient retrieval and data sharing Most cloud consumers want their cloud data and its usage to remain private, so along with encryption of their data objects, they also want their access patterns to be oblivious to the cloud provider.

Purpose

We will design and configure an algorithm to cloud user storage management on HADOOP (VMWARE) at operating system of Linux. Manage user directory by using HDFS (Hadoop Distributed File System) as well create remote user interface. User also managed or checked its storage capacity either available or used.

In this Report we also highlight the feature (Scalability, storage management, Network management, user I.D, multiple node,) of different platform like (Amazon, Google, iCloud, Apache Hadoop)

Cloud Storage Features

* **Scalability :** When they store data on-premises, organizations have to forecast their needs far into the future and purchase excess capacity in order to have enough space as their data continues to grow. But with cloud storage, more resources are automatically available as the organization needs them, with no need for over provisioning.
* **Security and File Encryption :** Security is a top concern when you take your business to the cloud. Your company's private data [could be compromised](https://www.businessnewsdaily.com/5215-dangers-cloud-computing.html) if the service you use lacks the right features. Make sure your service encrypts your data at least once before it ever leaves your servers. The best services also allow you to pick your own encryption key so that not even that company's employees can view your files. You should also have the ability to protect your files with a password, particularly when accessing them from a mobile device, as well as the capability to set different levels of access for different employees. Top-notch security features will let you relax when sending private information about your clients to the cloud.
* **Cloud Storage Management Complexity :**

The idea behind cloud storage management complexity is to use computer data storage software for policy-based provisioning and management of data storage independent of the underlying hardware.

Traditional data storage cannot overcome today’s challenges of scale, integration, and flexibility. If your solution for managing data growth is simply to buy more storage capacity, sooner or later you’ll be facing dramatically increased costs for both storage and management. Manually managing across heterogeneous storage systems, silos and clouds is not only error-prone but also leads to administrative overhead.

Software-defined storage addresses these challenges by separating the software that provides the intelligence for storage from the traditional hardware platform. The results include easier storage management, lower storage costs, and anywhere-anytime access to support cloud storage.

**DESIGN AND IMPLEMENTATION OF PRIVATE CLOUD STORAGE PLATFORM**

There are two ways to install Hadoop.

**Single node** and **Multi node**.

**Single Node** cluster means only one Data Node running and setting up all the Name Node, Data Node, Resource Manager and Node Manager on a single machine. This is used for studying and testing purposes. For example, let us consider a sample data set inside a healthcare industry. So, for testing whether the Oozie (Oozie is a server based Workflow Engine specialized in running workflow jobs with actions that run Hadoop Map/Reduce and Pig jobs. Oozie is a Java Web-Application that runs in a Java servlet-container) jobs have scheduled all the processes like collecting, aggregating, storing and processing the data in a proper sequence, we use single node cluster. It can easily and efficiently test the sequential workflow in a smaller environment as compared to large environments which contains terabytes of data distributed across hundreds of machines

.

**Multi node** cluster there are more than one Data Node running and each Data Node is running on different machines. The multi node cluster is practically used in organizations for analyzing Big Data. Considering the above example, in real time when we deal with peta bytes of data, it needs to be distributed across hundreds of machines to be processed. Thus, here we use multi node cluster.

**Prerequisites**

* **VIRTUAL BOX**: it is used for installing the operating system on it.
* **OPERATING SYSTEM**: we can install Hadoop on Linux based operating systems. Ubuntu and CentOS are very commonly used. In this project, we are using Linux.
* **JAVA**: we need to install the Java 8 package on your system.
* **HADOOP**: we need to install Hadoop package.

**Supported Platforms**

* GNU/Linux is supported as a development and production platform. Hadoop has been demonstrated on GNU/Linux clusters with 2000 nodes.
* Windows is also a supported platform but the followings steps are for Linux only. To set up Hadoop on Windows,

**Install Hadoop**

**Step 1:** To download the Java 8 Package. Save this file in your home directory.

**Step 2:** Extract the Java Tar File.

***Untar Java - Install Hadoop - Edureka*Command*:*** **tar -xvf jdk-8u101-linux-i586.tar.gz**

Figure 1: Hadoop Installation – Extracting Java Files

**Step 3:**Download the Hadoop 2.7.3 Package.

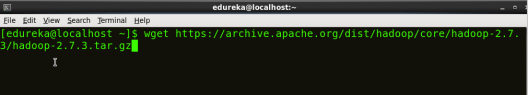
**Command:**  **wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz**

Figure 2: Hadoop Installation – Downloading Hadoop

**Step 4:**Extract the Hadoop tar File.

Extract Hadoop Package - Install Hadoop - Edureka**Command:** **tar -xvf hadoop-2.7.3.tar.gz**

Figure 3: Hadoop Installation – Extracting Hadoop Files

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

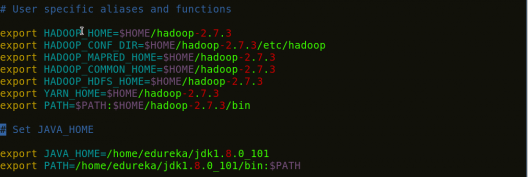
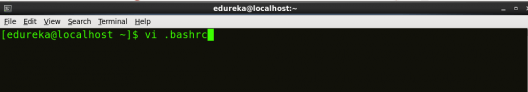
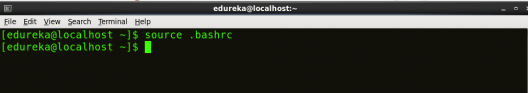
**Open. bashrc file. Now, add Hadoop and Java Path as shown below.**

Figure 4: Hadoop Installation – Setting Environment Variable

**Command:**  **vi .bashrc**

Then, save the bash file and close it.

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

**Command:**  **source .bashrc**

Figure 5: 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.

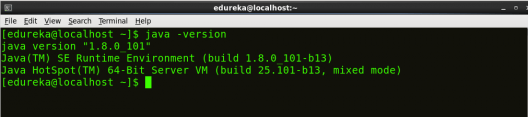
**Command:** **java -version**

Figure 6: Hadoop Installation – Checking Java Version

**Command:** **hadoop version**

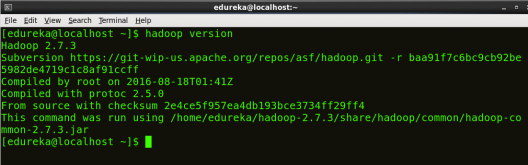


Figure 7: Hadoop Installation – Checking Hadoop Version

**Step 6:** Edit the [Hadoop Configuration files](https://www.edureka.co/blog/explaining-hadoop-configuration/" \t "_blank).

**Command:** **cd hadoop-2.7.3/etc/hadoop/**

**Command: ls**

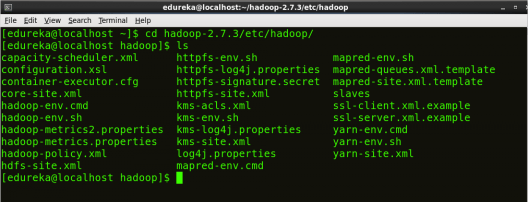
All the Hadoop configuration files are located in hadoop-2.7.3/etc/hadoop directory as you can see in the snapshot below:

Figure 8: Hadoop Installation – Hadoop Configuration Files

**Step 7:** Open ***core-site.xml*** and edit the property mentioned below inside configuration tag:

***Core-site.xml*** informs Hadoop daemon where Name Node runs in the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & Map Reduce.

**Command:** **vi core-site.xml**

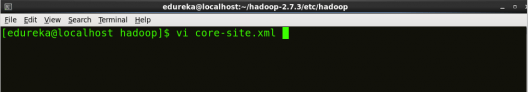
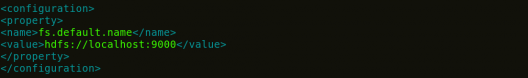
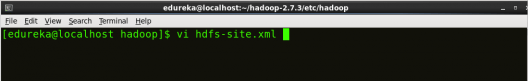


Figure 9: Hadoop Installation – Configuring core-site.xml

**Step 8:** Edit ***hdfs-site.xml***and edit the property mentioned below inside configuration tag:

***Hdfs-site.xml*** contains configuration settings of HDFS daemons (i.e. Name Node, Data Node, Secondary Name Node). It also includes the replication factor and block size of HDFS.

**Command: vi hdfs-site.xml**

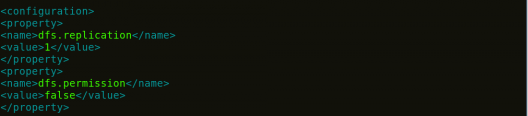


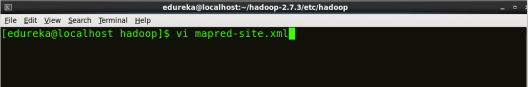
Figure 10: Hadoop Installation – Configuring hdfs-site.xml

**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.**

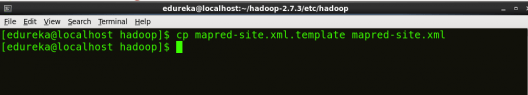
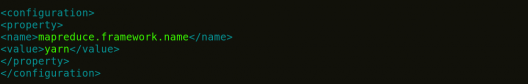
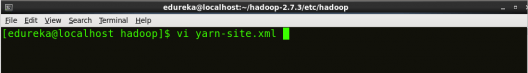


Figure 11: Hadoop Installation – Configuring mapred-site.xml

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

***Yarn-site.xml*** contains configuration settings of Resource Manager and Node Manager like application memory management size, the operation needed on program & algorithm, etc.

**Command: vi yarn-site.xml**



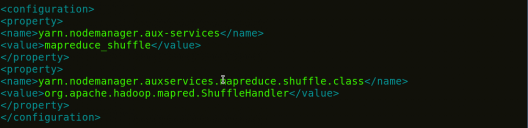


Figure 12: Hadoop Installation – Configuring yarn-site.xml

**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**

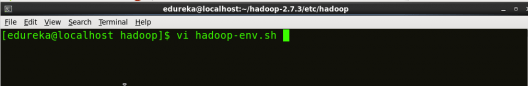
Property of hadoop-env - Install Hadoop - Edureka

Figure 13: 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**

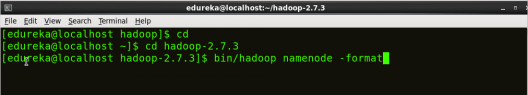


Figure 14: Hadoop Installation – Formatting NameNode

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 file system. 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***

### Hadoop NameNode UI - Install Hadoop - EdurekaStep 15: Now open the Mozilla browser and go to localhost:50070/dfshealth.html to check the NameNode interface.

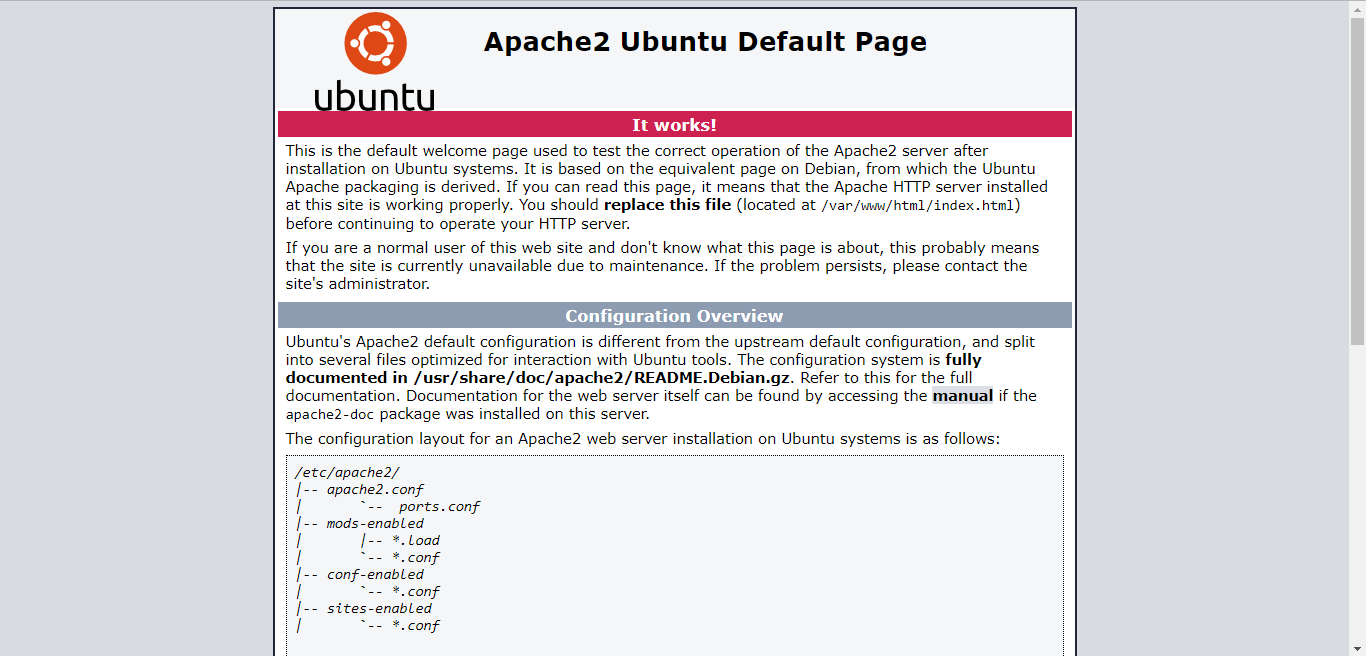
Figure 15: Hadoop Installation – Starting WebUI

**Installation of Apache in linux Mint 1.8**

If Linux is the brains of the operation, then Apache is the heart. A web server delivers web pages to the user by processing HTTP requests, and Apache is the most widely used of its kind. There are alternatives such as Nginx, but Apache is never a bad choice. It has proven stability and long-running community support.

**Command:** sudo apt-get install apache2

And you get the result of Apache in your web browser



**SURVEY OF CLOUD USER STORAGE MANAGEMENT**

In this report, we have shown the survey of cloud user storage management and compare some well-known platforms by giving rating (High, Medium and Low) w.r.t some different features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PLATFORMS** | **FEATURES** | | | |
| **SCALIBILTY** | **STAORAGE MANAGEMENT** | **PRICING** | **SECURITY** |
| **AMAZON** | **HIGH**  (Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. This means customers of all sizes and industries can use it to store and protect any amount of data for a range of use cases, such as websites, mobile applications, backup and restore, archive, enterprise applications, IOT devices, and big data analytics.) | **HIGH**  (With S3 bucket names, prefixes, object tags, and S3 Inventory, you have a range of ways to categorize and report on your data, and subsequently can configure other S3 features to take action. S3 Batch Operations makes it simple, whether you store thousands of objects or a billion, to manage your data in Amazon S3 at any scale. With S3 Batch Operations, you can copy objects between buckets, replace object tag sets, modify access controls, and restore archived objects from Amazon S3 Glacier, with a single S3 API request or a few clicks in the Amazon S3 Management Console. You can also use S3 Batch Operations to run AWS Lambda functions across your objects to execute custom business logic, such as processing data or transcoding image files. | **MEDIUM**  (Amazon Cloud Drive offers better value since you're essentially paying $1 a month per GB. In other words, you pay $20 a month for 20GB of storage, $50 a month for 50GB and so forth.) | **MEDIUM**  (Despite massive investments in public cloud security, organizations still have reservations about the security of sensitive data, systems, and services in the cloud. The security technology challenge is only exacerbated by the dramatic shortage of skilled cyber security professionals. Major points covered in the Cloud Security Report, published by Cyber Security Research |
| **GOOGLE** | **HIGH**  (Cloud Storage offers unlimited object storage and individual objects can be as large as 5TB. Objects can be overwritten no more than once per second and there is no limit to read frequency. Objects larger than 5MB should be uploaded with multipart or resemble uploading.) |  | **LOW**  (Google Drive’s new pricing structure is great for users, who benefit from more attractive prices from one of the foremost companies in data storage, especially when you compare them against paid plans from Google’s cloud competitors. It now offers the first 15 GB per month for free. For $100 a month) |  |
| **I Cloud** |  |  | **HIGH**  (Apple’s prices for extra storage aren't nearly as generous: $20 for 15GB of storage, $40 for 20GB and $100 for 50GB.) | **MEDIUM**  (One of the biggest security concerns is about software attackers. If any unauthenticated user is granted with a temporary access by introducing a malware, an attacker can attain the access to the I Cloud documents stored locally, and therefore causing it to spread stored personal information or append files to the I Cloud system which could terribly damage the Apple product which the user is using currently.) |