CLOUD STORAGE USING HADOOP AND PLAY

Devateja G[1], Kashyap P V B[2], Suraj C[3], Harshavardhan C[4], Impana Appaji[5]

[1234]Computer Science & Engineering, Academy for Technical and Management Excellence college of Engineering, Mysore, Karnataka, India, devateja94@gmail.com , kashyappvb007@gmail.com , surajdhapte999@gmail.com, harshavardhanc95@gmail.com .

[5]Assistant Professor, Computer Science & Engineering, Academy for Technical and Management Excellence college of Engineering, Mysore, Karnataka, India, impana.appaji@gmail.com.

Abstract: In this project, we are concentrating on cloud storage. It is a service model in which data is maintained, managed and backed up remotely and made available to users over a network (typically the Internet).These cloud storage providers are responsible for keeping the data available and accessible, and the physical environment protected and running. In recent years, the information that are retrieved from large datasets known as Big Data. Its difficult to transfer larger files, For these reasons, we need to manipulate (e.g. edit, split, create) big data files to make them easier to move and work with them and even split big data files to make them more manageable. For this we use Apache Hadoop frameworks. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Which is based on distributed computing having HDFS file system. This file system is written in Java and designed for portability across various hardware and software platforms. Hadoop is very much suitable for storing high volume of data and it also provide the high speed access to the data of the application which we want to use. But hadoop is not really a database : It stores data and you can pull data out of it, but there are no queries involved - SQL or otherwise. Hadoop is more of a data warehousing system - so it needs a system like Map Reduce to actually process the data.

KEYWORDS: *Big Data, Apache Hadoop, HDFS, Map Reduce, Distributed storage.*

I. Introduction

Cloud storage is a model of data storage in which the digital data is stored in logical pools, the physical storage spans multiple servers (and often locations), and the physical environment is typically owned and managed by a hosting company. These cloud storage providers are responsible for keeping the data available and accessible, and the physical environment protected and running. People and organizations buy or lease storage capacity from the providers to store user, organization, or application data.

Cloud storage services may be accessed through a co-located cloud computer service, a web service application programming interface (API) or by applications that utilize the API, such as cloud desktop storage, a cloud storage gateway or Web-based content management systems.

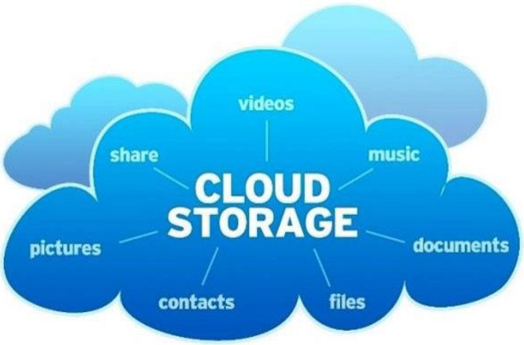


Figure 1.1 – Cloud

There are three main cloud storage models :

1. Public cloud storage services, such as Amazon’s Simple Storage Service (S3), provide a multi-tenant storage environment thats most suitable for unstructured data.

2. Private cloud storage services provide a dedicated environment protected behind an organizations firewall. Private clouds are appropriate for users who need customization and more control over their data.

3. Hybrid cloud storage is a combination of the other two models that includes at least one private cloud and one public cloud infrastructure. An organization might, for example, store actively used and structured data in a private cloud and unstructured and archival data in a public cloud.

1.2 What is Hadoop ?

“Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.”

The Hadoop Distributed File System is the block storage layer that Hadoop uses to keep its files. HDFS was de-signed to hold very large datasets reliably using data replication (Shvachko et al., 2010). This allows HDFS to stream large amounts of data to user applications in a reasonable time. Its architecture is composed of two main entities : NameNode and DataNodes A particular study that contributes in different ways to the framework was developed specifically to improve Hadoops performance.

1.2.1 Hadoop Ecosystem

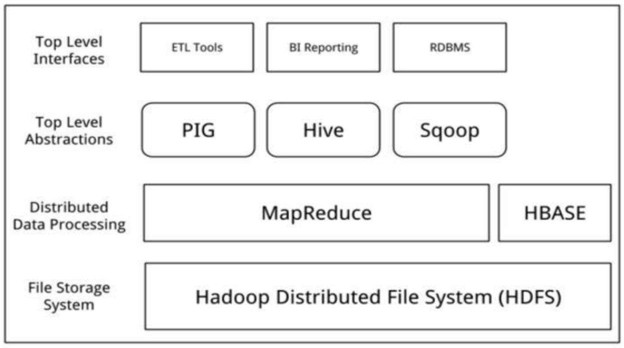


Figure 1.2 – Hadoop Ecosystem

Hadoop is based on distributed computing having HDFS file system (Hadoop Distributed File System).HDFS stores file system metadata and application data separately. HDFS stores metadata on a dedicated server, called the Name Node. Name Node is a type of master node, which is having the information or we can say that meta data about the all data node there is address(use to talk), free space, data they store, active data node , passive data node, task tracker, job tracker and many other configuration such as replication of data. Application data are stored on other servers called Data Node. Data Node is a type of slave node in the hadoop, which is used to save the data and there is task tracker in data node which is use to track on the ongoing job on the data node and the jobs which coming from name node.

1.2.2 HDFS Architecture

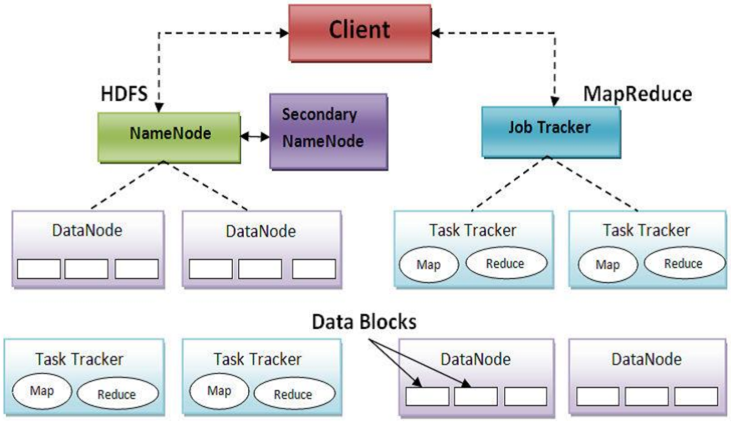


Figure 1.3 – HDFS Architecture

Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly faulttolerant and designed using low-cost hardware.

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

DataNode : It stores data in the hadoop file system. A functional file system has more than one DataNode, with the data replicated across them.

NameNode : It keeps the directory of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these file (Meta data of files).

Secondary Namenode : Its purpose is to have a checkpoint in HDFS. It is just a helper node for NameNode.

1.2.3 Play Frameworks

Figure 1.4 – Play Frameworks

Play Framework makes it easy to build web applications with Java & Scala. Play is based on a lightweight, stateless, web-friendly architecture. Built on Akka, Play provides predictable and minimal resource consumption (CPU, memory, threads) for highly-scalable applications.

Play is an open source web application framework, written in Scala and Java, which follows the modelviewcontroller (mvc) architectural pattern. It aims to optimize developer productivity by using convention over configuration, hot code reloading and display of errors in the browser.

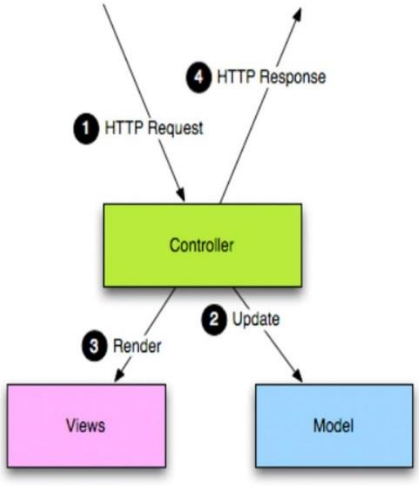


Figure 1.5 – Play MVC

II. Literature Survey

While doing survey we found that the term Cloud refers to a Network or Internet. In other words, we can say that Cloud is something, which is present at remote location. Cloud can provide services over network. Cloud storage is an effective means to solve the storage and management challenges of the huge growing data. In order to settle the problems existing in the application of cloud storage for remote sensing data, a distributed storage module.[??]

Cloud storage with unparalleled scalability and high throughput can meet the scalability requirements of data growing and reduce investment for new equipment. However, the distributed file systems in cloud storage platforms are based on stream access.[??]

As there is lot of data lies in the industry but there is nothing before Big data comes into picture. When we talk about Big Data, the first name comes in mind is ”HADOOP” a well know product in the market of big data. Hadoop is based on HDFS, which is hadoop distributed file system. In which data is equally (ideally) distributed on each node in the hadoop system. When we (client) want to fetch or add or modify or delete some data from hadoop, then hadoop system collect the data from each node of our interest and do the meaningful actions of our interest.[??]

2.1.4 Cloud Computing Technologies

1. Microsoft Cloud Technologies

Microsoft is a foremost provider of cloud technologies and applications with results that matches with all type of business needs. It provides all type of services whether it is PaaS, IaaS or SaaS. The Infrastructure-as-a-Service Microsoft provides the windows server and system canter. And in case of Platform-as-a-Service it provide Windows Azure, with this you can easily build, host and scale applications in Microsoft Datacenter without up-front expenses just pay for what you use. Other PaaS services are SQLSERVER and VISUAL STUDIO. On the other way office365, share-point servers, dynamic CRM and exchange server are the Software-as-a Services provided by Microsoft. With this we can say that Microsoft Cloud services are the complete package for your business.

2. **Oracle Cloud Technologies** Oracle also provides the complete enterprise read public cloud solution including IaaS, PaaS and SaaS. With this you only need to concentrate on your business

out worrying about IT management .oracle offers the following services Database, it is available Database-as-a-Service along with accessing the Database in the Cloud directly through standard network connections, or as a Platform as a Service, with a complete development and deployment environment. You can avail its services as a single schema based service, or a virtual machine with a fully configured, running Oracle Database instance. To use oracle cloud database you just need to create an account with a valid email id and login with the provided credentials, you can enjoy this service for free for 30 days trial after that you can choose their given plans as per your need.

3. Google Cloud Technologies

Google cloud also provides the services such as Software-as-a Service, Platform-as-a-Service and Infrastructure-as-a-Service. Google cloud enables developers to build, test and deploy applications on Googles highly scalable and secure infrastructure. As we know that Google has already provided infrastructure that allows Google to return billions of search results in milliseconds, provide storage for about 425 million Gmail users and serve 6 billion hours of YouTube video per month. Google has the ability to build, organize and operate a huge network of servers and fiber-optic cables. All this in aggregate makes Google the King of all cloud. Cloud SQL-cloud SQL provides you the fully managed, relational My-SQL database to store and manage data. Google deal with the replication, patch management and database management to ensure availability and performance. My-SQL database deployed in the cloud without any difficulty.

III. Requirement Specification

3.1 Software Requirements

•OS - Ubuntu 14.04 LTS

•JDK 1.8

•Apache Hadoop-2.6.0

•Play Frameworks-2.6.2

•Eclipse IDE

3.2 Hardware Requirements

RAM 4GB

•Hard Disk 500GB

•LAN

3.3 Functional Requirements

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality. In this system following are the functional requirements. Signup Authentication Login Validation Forgot Password Upload Download Sharing One to one Sharing One to Many Sharing Trash

3.4 Non-Functional Requirements

Non functional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviors. They may relate to emergent system properties such as reliability, response time and store occupancy. Non functional requirements arise through the user needs, because of budget constraints, organizational policies, the need for interoperability with other software and hardware systems or because of external factors such as

Availability Whenever user wants to access his/her uploaded files the user is able to access ired data.

Reliability Our system performs its functions well in normal cases as well as un-expected circumstances because HDFS (Hadoop Distributed File System) provides fault-tolerant future. Hence our system is able to maintain reliability even through partial failure happens.

Ease of Use The front end is designed in such a way that it provides an interface which allows the user to interact in an easy manner.

Modularity The complete product is broken up into many modules and well-defined interfaces are developed to explore the benefit of flexibility of the product.

Robustness This software is being developed in such a way that the over all performance is optimized and the user can expect the results within a limited time with utmost relevancy and correctness. Java itself possesses the feature of robustness, which implies the failure of the system is negligible.

IV. System Analysis and Design

System analysis is detailed study of various operations performed by the system and their relationships within and outside the system. Analysis begins when ther user or manager begins a study of the program using the existing system. During analysis data collected on various files, decision points and transactions helded by present system.The commonly used tools in the system are Data Flow Diagram,trainings,experiences and common sense are required for the collection of the relevant information needed to develop the system. The success of the system depends largly on how clearly the problem is defined, throughly investigated and properly carried out through the choice of the solution. A good analysis model should provide not only the mechanism of the problen understanding but also framework of the solution. Thus, it should be studied throughly by collecting data about the system.

a.Data Flow Diagram (DFD)

A data flow diagram is graphical representation of the flow of data through an information system, modeling its process aspects.Often they are a preliminary step used to create an overview of the system which can later be elaborated. DFDs can also be used for the visualization of data processing.

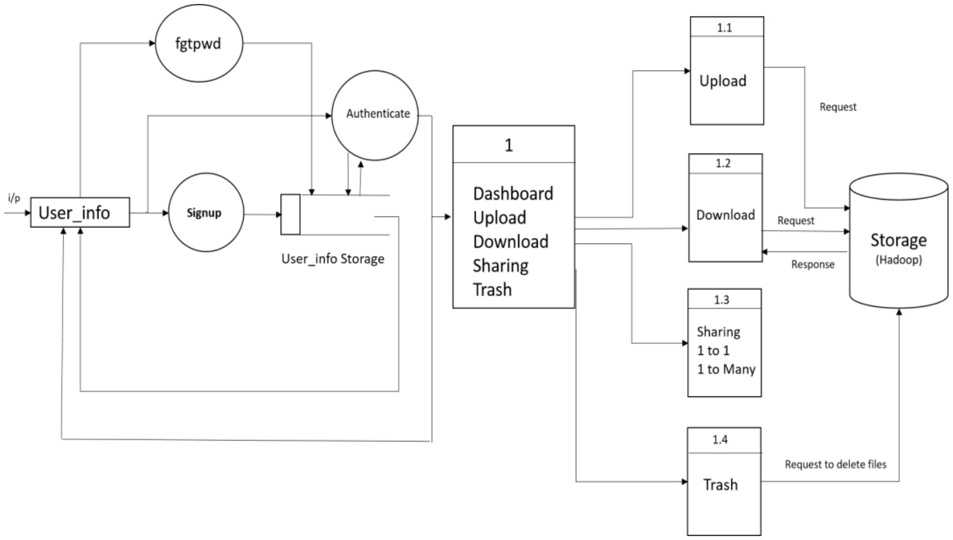


Figure 4.1 – Data Flow Diagram

b.MVC Design Method

Play Framework actually makes use of a simplified variant of the MVC design called the model-delegate. This design combines the view and the controller object into a single element that draws the component to the screen and handles GUI events known as the UI delegate. Communication between the model and the UI delegate becomes a two-way street. Each Swing component contains a model and a UI delegate. The model is responsible for maintaining information about the components state. The UI delegate is responsible for maintaining information about how to draw the component on the screen.

The design method that has been followed to design the architecture of the system is MVC design pattern. Play uses the model-view-controller (MVC) architecture as the fundamental design behind each of its components. Essentially, MVC breaks GUI component into three elements. Each of these elements plays a crucial role in how the component behaves. The MVC design pattern separates a software component into three distinct pieces : a model, a view, and a controller.

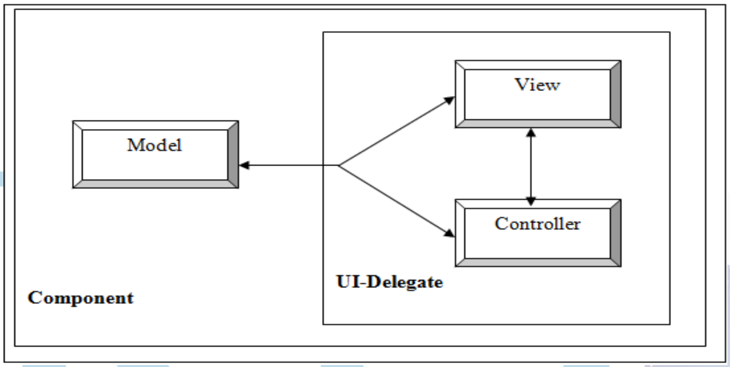


Figure 4.2 – Model-View-Controller

c. Usecase Diagram

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a use case analysis which purpose is to present the graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases) and any dependencies between those usecases

The use case diagram of the project is shown below :

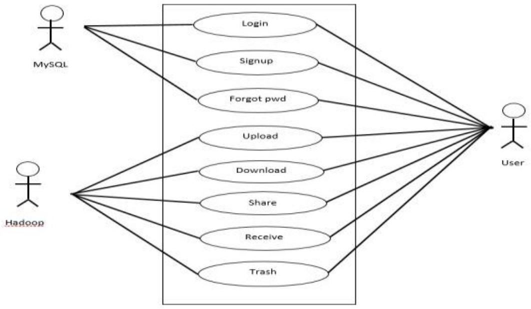
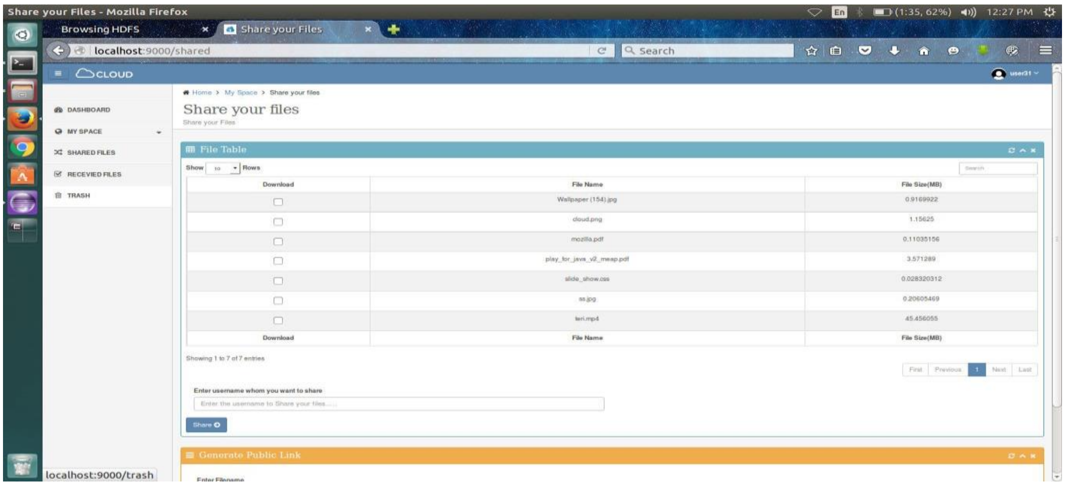


Figure 4.3 – Model-View-Controller

V. Implementation

This chapter deals with implementation details of the project. Implementation phase consists of all the processes involved in getting new software or hardware operating properly in its environment, including installation, configuration, and running, testing, and making necessary changes. The chapter mainly focuses on implementation of the main constituents of the project namely login,signup,fotgot password, upload,download,sharing and trash.



Share your files

Share your Files

Download	File Name	File Size(MB)
□	Wallpaper (154).jpg	0.9169922
□	cloud.png	1.15625
□	mozilla.pdf	0.11035156
□	play_for_java_v2_meap.pdf	3.571289
□	slide_show.css	0.028320312
□	ss.jpg	0.20605469
□	teri.mp4	45.456055
Download	File Name	File Size(MB)

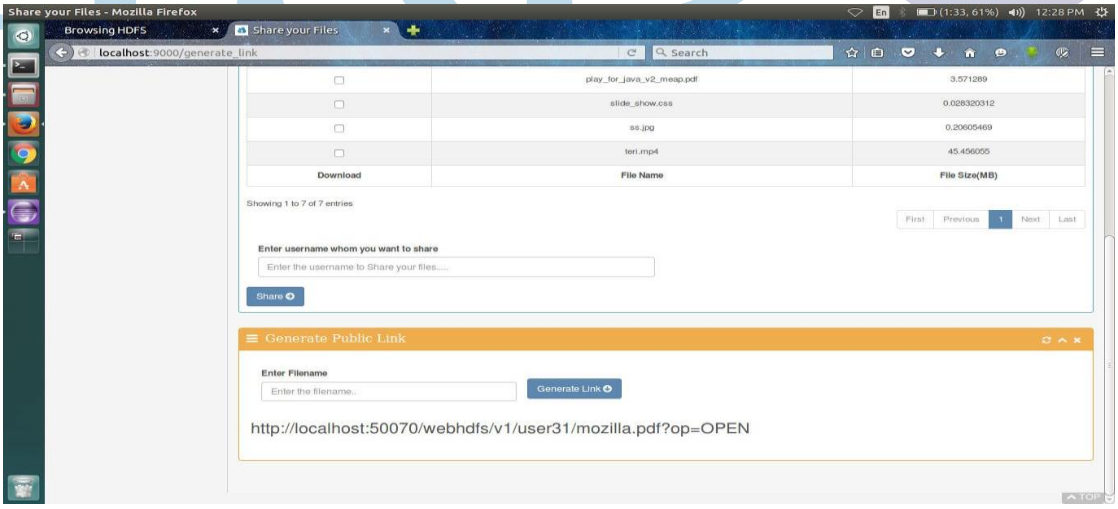
Showing 1 to 7 of 7 entries

Enter username whom you want to share

Enter the username to Share your files.....

Enter Filename

Figure 5.1 Shared Page (one to one)



□	play_for_java_v2_meap.pdf	3.571289
□	slide_show.css	0.028320312
□	ss.jpg	0.20605469
□	teri.mp4	45.456055
Download	File Name	File Size(MB)

Showing 1 to 7 of 7 entries

Enter username whom you want to share

Enter the username to Share your files.....

Enter Filename

Enter the filename..

http://localhost:50070/webhdfs/v1/user31/mozilla.pdf?op=OPEN

Figure 5.2 – Shared Page (one to many)

VI. CONCLUSION

Everyday a large amount of data is getting dumped into machines. The major challenge is not to store large data sets in our systems but to retrieve and analyze the large data in the organizations that too data present in different machines at different locations. Hadoop offers a proven solution to the modern challenges facing legacy systems. Hadoop is an open-source software platform by the Apache Foundation for building clusters of servers for use in distributed computing. Hadoop can handle large volumes of structured and unstructured data more efficiently than the traditional enterprise data warehouse.

REFERENCES

[1] Apache Hadoop: https://hadoop.apache.org/
[2] Wikipedia: https://en.wikipedia.org/wiki/Apache_Hadoop
[3] https://opensource.com/life/14/8/intro-apache-hadoop-big-data
[4] https://radar.oreilly.com/2012/02/what-is-apache-hadoop.html
[5] https: //www .nobius.org/□dbg/practical-file-systemdesign.pdf
[6] https: //www.maxi-pedia.comiwhat+is+DFS.