



**A Training Report**  
**On**  
**“DataStack – Platform As A Service”**  
Submitted to the Rajasthan Technical University, Kota  
In partial fulfillment of the requirement for the degree of  
**MASTER OF COMPUTER APPLICATIONS**

Submitted by:-  
**Nimisha Vilayatarani**  
Roll no.: 18CPGXX264

<b>Name of Internal Guide</b> Mrs. Mamta Sharma	<b>Name of Training in-charge</b> Ms. Palak Jain
--	---

**S. S. Jain Subodh P.G. College**  
Affiliated to  
**Rajasthan Technical University, Kota**

**MCA – Batch (2018-2020)**

April, 2020

## **Acknowledgment**

I express my deep sense of gratitude to my respected and learned guide,

**Mrs. Mamta Sharma** for her valuable help and guidance. I am thankful for the encouragement she has given me in the completion of my project.

I am also grateful to our respected **Prof. K.B. Sharma**, Director and **Dr. Leena Bhatia Mam**, Head of Department for allowing me to utilize all the necessary resources of the institution to complete my project. I am also thankful to all the other faculty and staff members of our department for their kind co-operation and help.

Lastly, I would like to express my deep appreciation towards my classmates and indebtedness to my parents for providing me the moral support and encouragement.

Date :

Nimisha vilayatrani

Class - MCA - VI

Roll No. 18CPGXX264

## **Table of Content**

<b>S.No.</b>	<b>Title</b>	<b>Page No</b>
1.	Introduction to the Project	1 - 2
2.	Project Objectives	3
3.	Technologies Used	4 - 6
4.	Requirement Analysis ( <i>Hardware and Software Specifications</i> )	7
5.	System Design	8 - 12
6.	Testing ( <i>Testing methods</i> )	13 – 15
7.	Implementation & Maintenance	16
8.	Screenshots of Project	17 - 21
9.	Future Scope of the Project	22
10.	Bibliography ( <i>Websites, Books and Journals</i> )	23 - 24

### **• Introduction to Project**

Datastack is a Platform as a Service that allows users to work on Big data technologies without the complexity of building and maintaining the infrastructure and environment, saving 97% time of user.

Datastack is a web portal that is integrated with the cloud platform and providing Platform As A Service to the users.

This platform allows users to work on big data technologies such as Hadoop, spark etc. This platform eliminates the user efforts to install the entire environment that is quiet a lot time consuming task.

User can use various technologies such as Linux, python, Big data hadoop, spark on an easy to use portal built on top of AWS cloud, without

actually installing and configuring them.

Services provided by DataStack are :-

- **Dedicated Redhat Linux Operating System –**

- Datastack provides a dedicated redhat 7 linux operating system to each user on the browser. Users need not to get the system and install the linux on it to start working.
- This dedicated Operating system is running on the cloud server, and to access it and start working on it, user is given a terminal on the browser with the required permissions to access and work with their dedicated operating system.

- **Single click Hadoop and Yarn cluster setup –**

- Setting up the Hadoop cluster is quite a complex and time consuming task if done manually, and also more prone to errors. Simplifying this step, we have provided the user with a single click option that will automatically setup the entire Hadoop cluster in the background within few seconds.
- Installing the entire Hadoop cluster and yarn will take upto 40 minutes if done manually, and that's a lot time consuming and requires chronological steps to be done.
- Datastack made it extremely easy by providing hassle free single click solution to all those steps. Just click the button and within few seconds your environment is up and running.

- **Python and Jupyter Notebook setup.**

- Users can also work on python language. It is automatically installed in the users dedicated system. So user can create python programs and run them easily on the portal along with the Jupyter Notebook that provide easy to use interface to work with python.
- With python user can work on their big data task, compile and run their python programs.

- **Objectives of the Project**

- Datastack is a platform as a service which aims to simplify the complex task for the user on an easy to go web portal.



- Providing a platform for user to work on high end technologies such as Big Data Hadoop, Spark, Python.
- Eliminating all the manual efforts of installing and configuring the setup to work on such technologies which is quite a complex task and requires a lot of time.
- Bringing the users on the cloud platform from their local machine, which is more reliable, versatile and secure.



- **Technologies Used**

- Ansible Automation –

- Ansible is an open source automation platform. It is very, [very simple to setup](#) and yet powerful. Ansible can help you with configuration management, application deployment, task automation.
- It can also do IT orchestration, where you have to run tasks in sequence and create a chain of events which must happen on several different servers or devices.
- Unlike Puppet or Chef it doesn't use an agent on the remote host. Instead Ansible uses SSH which is assumed to be installed on all the systems you want to manage. Also it's written in Python which needs to be installed on the remote host.
- Ansible is available for free and runs on Linux, Mac or BSD. Aside from the free offering, Ansible also has an enterprise product called [Ansible](http://www.ansible.com/tower) [HYPERLINK "http://www.ansible.com/tower"](http://www.ansible.com/tower) [Tower](http://www.ansible.com/tower).
- **AWS Cloud –**
  - Amazon web service is a platform that offers flexible, reliable, scalable, easy-to-use and cost-effective cloud computing solutions.
  - AWS is a comprehensive, easy to use computing platform offered Amazon. The platform is developed with a combination of infrastructure as a service (IaaS), platform as a service (PaaS) and packaged software as a service (SaaS) offerings.
  - EC2(Elastic Compute Cloud) - EC2 is a virtual machine in the cloud on which you have OS level control. You can run this cloud server whenever you want.
- **CGI Programming –**
  - Common Gateway Interface, a specification for transferring information between a World Wide Web server and a CGI

program. A CGI program is any program designed to accept and return data that conforms to the CGI specification.

- CGI programs are the most common way for Web servers to interact dynamically with users. Many HTML pages that contain forms, for example, use a CGI program to process the form's data once it's submitted.

- **Python –**

- Python is an [interpreted](#), [high-level](#), [general-purpose programming language](#). Python was created by [Guido van Rossum](#) and first released in 1991.
- Python's design philosophy emphasizes [code readability](#) with its notable use of [significant whitespace](#). Its [language constructs](#) and [object-oriented](#) approach aim to help programmers write clear, logical code for small and large-scale projects.
- Python is [dynamically typed](#) and [garbage-collected](#). It supports multiple [programming paradigms](#), including [structured](#) (particularly, [procedural](#)), object-oriented, and [functional programming](#).
- Python is often described as a "batteries included" language due to its comprehensive [standard library](#).

- **Linux**

- Linux is a family of [open source Unix-like operating systems](#) based on the [Linux kernel](#), an [operating system kernel](#) first released on September 17, 1991, by [Linus Torvalds](#). Linux is typically [packaged](#) in a [Linux distribution](#).

- **PHP**

- PHP is a server scripting language, and a powerful tool for making dynamic and interactive Web pages. PHP is a widely-used, free, and efficient alternative to competitors such as Microsoft's ASP. PHP 7 is the latest stable release.
- PHP is a popular general-purpose scripting language that is especially suited to web development. It was originally created by Rasmus Lerdorf in 1994.
- PHP runs on various platforms (Windows, Linux, Unix, Mac OS X, etc)

- **MYSQL**

- MySQL is the most popular Open Source Relational SQL Database Management System. MySQL is one of the best RDBMS being used for developing various web-based software applications.
- MySQL is a very powerful program in its own right. It handles a large subset of the functionality of the most expensive and powerful database packages.
- MySQL uses a standard form of the well-known SQL data language.

- **Requirement analysis**

## **Hardware Specification –**

- RHEL 7 (Red Hat Enterprise Linux).



- 8GB RAM
- 500 GB Hard Disk

### **Software Specification -**

- AWS Cloud
- EC2 – Elastic Compute Cloud (basic)
- Big Data Hadoop V 2.0 or greater
- Apache Spark
- Python V 3.0 or greater
- PHP 5
- MYSQL
- Linux 7
- Ansible

### **• SYSTEM DESIGN**

Design is the first step in the development phase for any techniques and principles for the purpose of defining a device, a process or system in sufficient detail to permit its physical realization.

Once the software requirements have been analyzed and specified the software design involves three technical activities - design, coding, implementation and testing that are required to build and verify the software.

The design activities are of main importance in this phase, because in this activity, decisions ultimately affecting the success of the software implementation and its ease of maintenance are made. These decisions have the final bearing upon reliability and maintainability of the system. Design is the only way to accurately translate the customer's requirements into finished software or a system.

Design is the place where quality is fostered in development. Software design is a process through which requirements are translated into a representation of software. Software design is conducted in two steps. Preliminary design is concerned with the transformation of requirements into data.

### **UML Diagrams:**

Actor: A coherent set of roles that users of use cases play when interacting with the use cases.

Use case: A description of sequence of actions, including variants, that a system performs that yields an observable result of value of an actor.

UML stands for Unified Modeling Language. UML is a language for specifying, visualizing and documenting the system. This is the step while developing any product after analysis. The goal from this is to produce a model of the entities involved in the project which later need to be built. The representation of the entities that are to be used in the product being developed need to be designed.

There are various kinds of methods in software design:

They are as follows:

- Use case Diagram
- Sequence Diagram
- Collaboration Diagram
- Activity Diagram
- State chat Diagram

## **USECASE DIAGRAMS:**

Use case diagrams model behavior within a system and helps the developers understand of what the user require. The stick man represents what's called an actor. Use case diagram can be useful for getting an overall view of the system and clarifying who can do and more importantly what they can't do.

- The purpose is to show the interactions between the use case and actor.
- To represent the system requirements from user's perspective.
- An actor could be the end-user of the system or an external system.

## **DATA FLOW DIAGRAMS**

The DFD takes an input-process-output view of a system i.e. data objects flow into the software, are transformed by processing elements, and resultant data objects flow out of the software.

Data objects represented by labeled arrows and transformation are represented by circles also called as bubbles. DFD is presented in a hierarchical fashion i.e. the first data flow model represents the system as a whole. Subsequent DFD refine the context diagram (level 0 DFD), providing increasing details with each subsequent level.

The DFD enables the software engineer to develop models of the information domain & functional domain at the same time. As the DFD is refined into greater levels of details, the analyst perform an implicit functional decomposition of the system. At the same time, the DFD

refinement results in a corresponding refinement of the data as it moves through the process that embody the applications.

A context-level DFD for the system the primary external entities produce information for use by the system and consume information generated by the system. The labeled arrow represents data objects or object hierarchy.

### **RULES FOR DFD:**

- Fix the scope of the system by means of context diagrams.
- Organize the DFD so that the main sequence of the actions
- Reads left to right and top to bottom.
- Identify all inputs and outputs.
- Identify and label each process internal to the system with Rounded circles.
- A process is required for all the data transformation and Transfers. Therefore, never connect a data store to a data Source or the destinations or another data store with just a Data flow arrow.
- Do not indicate hardware and ignore control information.
- Make sure the names of the processes accurately convey everything the process is done.
- There must not be unnamed process.
- Indicate external sources and destinations of the data, with Squares.
- Number each occurrence of repeated external entities.
- Identify all data flows for each process step, except simple Record retrievals.
- Label data flow on each arrow.
- Use details flow on each arrow.
- Use the details flow arrow to indicate data movements.

- **TESTING**

Testing is a process of executing a program with the intent of finding an error. Testing is a crucial element of software quality assurance and presents ultimate review of specification, design and coding.

System Testing is an important phase. Testing represents an interesting anomaly for the software. Thus a series of testing are performed for the proposed system before the system is ready for user acceptance testing.

A good test case is one that has a high probability of finding an as undiscovered error. A successful test is one that uncovers an as undiscovered error.

**Testing Objectives:**

- Testing is a process of executing a program to find the error .
- A good test case is one that has a probability of finding an as yet undiscovered error
- A successful test is one that uncovers an undiscovered error

## **Testing Principles**

- All tests should be traceable to end user requirements
- Tests should be planned long before testing begins
- Testing should begin on a small scale and progress towards testing in large
- Exhaustive testing is not possible
- To be most effective testing should be conducted by a independent third party

The primary objective for test case design is to derive a set of tests that has the highest livelihood for uncovering defects in software. To accomplish this objective two different categories of test case design techniques are used.

They are -

- White box testing.
- Black box testing.

### **White-box testing:**

White box testing focus on the program control structure. Test cases are derived to ensure that all statements in the program have been executed at least once during testing and that all logical conditions have been executed.

### **Block-box testing:**

Black box testing is designed to validate functional requirements without regard to the internal workings of a program. Black box testing mainly

focuses on the information domain of the software, deriving test cases by partitioning input and output in a manner that provides thorough test coverage. Incorrect and missing functions, interface errors, errors in data structures, error in functional logic are the errors falling in this category.

### **Unit testing:**

Unit testing is essential for the verification of the code produced during the coding phase and hence the goal is to test the internal logic of the modules. Using the detailed design description as a guide, important paths are tested to uncover errors within the boundary of the modules. These tests were carried out during the programming stage itself. All units of Vienna SQL were successfully tested.

### **Integration testing :**

Integration testing focuses on unit tested modules and build the program structure that is dictated by the design phase.

### **System testing:**

System testing tests the integration of each module in the system. It also tests to find discrepancies between the system and its original objective, current specification and system documentation. The primary concern is the compatibility of individual modules. Entire system is working properly or not will be tested here, and specified path ODBC connection will correct or not, and giving output or not are tested here these verifications and validations are done by giving input values to the system and by comparing with expected output.

### **Acceptance Testing:**

This testing is done to verify the readiness of the system for the implementation. Acceptance testing begins when the system is complete.



Its purpose is to provide the end user with the confidence that the system is ready for use. It involves planning and execution of functional tests, performance tests and stress tests in order to demonstrate that the implemented system satisfies its requirements.

- **Implementation**

Implementation is the stage where the theoretical design is turned into a working system. The most crucial stage in achieving a new successful system and in giving confidence on the new system for the users that it will work efficiently and effectively.

The system can be implemented only after thorough testing is done and if it is found to work according to the specification.


It involves careful planning, investigation of the current system and its constraints on implementation, design of methods to achieve the change over and an evaluation of change over methods a part from planning. Two major tasks of preparing the implementation are education and training of the users and testing of the system.

The more complex the system being implemented, the more involved will be the systems analysis and design effort required just for implementation.

The implementation phase comprises of several activities and we had tried for different users . The required hardware and software acquisition

is carried out. The system may require some software to be developed. For this, programs are written and tested. The user then changes over to his new fully tested system and the old system is discontinued.

- **Screenshots**

A woman in a light blue polo shirt is shown in profile, holding a tablet. A semi-transparent purple overlay is on the left side of the image. On the right, a white 'Sign Up' form is displayed. The form includes fields for Full Name, Email, Username, Password, and Repeat Password, each with a green checkmark indicating successful input. There is also a checkbox for 'I agree to the Terms of User' and a purple 'Sign Up' button. A 'Sign in' link with a right arrow is located at the bottom right of the form.

### Sign Up

Full Name  
Harsh Gupta ✓

Email  
info.harshgupta10@gmail.com ✓

Username  
harshhg ✓

Password  
\*\*\*\*\* ✓

Repeat Password  
\*\*\*\*\* ✓

☒ I agree to the Terms of User

Sign Up Sign in →

Please wait while we setup your environment..

You will automatically be redirected within 60 seconds.. Do not refresh the page..



### Sign In

Username

harshhg

✓

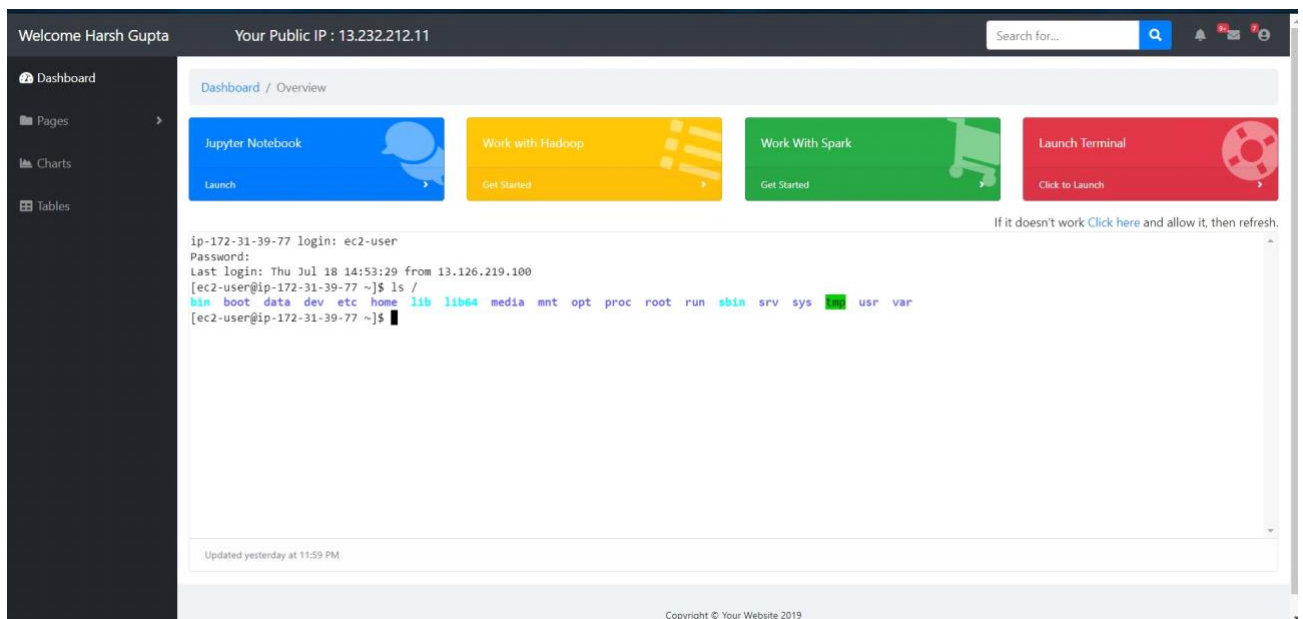
Password

\*\*\*\*\*

☐ I agree to the Terms of User

Sign in

Sign up →



**Please wait while we setup your HADOOP and YARN CLUSTER..**

You will automatically be redirected within 30 seconds.. Do not refresh the page..



Your Public IP : 13.232.212.11

Search for...

Dashboard / Overview

Jupyter Notebook

Launch

Work with Hadoop

Get Started

Work With Spark

Get Started

Launch Terminal

Click to Launch

Login: ec2-user | Password:12345

ec2-13-232-212-11 login: ec2-user

Password:

Last login: Thu Jul 18 14:58:22 from 13.126.219.100

[ec2-user@ec2-13-232-212-11 ~]\$ jps

4449 Jps

4082 NameNode

3796 NodeManager

3660 ResourceManager

4284 DataNode

[ec2-user@ec2-13-232-212-11 ~]\$

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview

'ec2-13-232-212-11.ap-south-1.compute.amazonaws.com:10000' (active)

Started:	Thu Jul 18 14:58:24 UTC 2019
Version:	2.7.3, rbaa91f7c5bc9cb52be5982de4719c1c8af91ccf
Compiled:	2016-08-18T01:41Z by root from branch-2.7.3
Cluster ID:	CID-20692e1c-2883-4442-a9f1-5fdb1092b5b2

Dashboard / Overview

Jupyter Notebook

Launch

Work with Hadoop

Get Started

Work With Spark

Get Started

Login: ec2-user | Password:12345

ec2-13-232-212-11 login: ec2-user

Password:

Last login: Thu Jul 18 14:58:22 from 13.126.219.100

[ec2-user@ec2-13-232-212-11 ~]\$ jps

4449 Jps

4082 NameNode

3796 NodeManager

3660 ResourceManager

4284 DataNode

[ec2-user@ec2-13-232-212-11 ~]\$

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview

'ec2-13-232-212-11.ap-south-1.compute.amazonaws.com:10000' (

Started:	Thu Jul 18 14:58:24 UTC 2019
----------	------------------------------

< Data Stack Online

I want to know about datastack.

8:31 PM ✓

Data Stack

Hii harsh, Datastack is a platform where you can use Big data components without worrying about platform. We provide you hadoop cluster, linux terminal, jupyter notebook, spark environment so that you can use them hassle free.

8:31 PM

I got a problem with HDFS cluster.

8:37 PM ✓

Data Stack

I am sorry for your inconvenience. Please try refreshing the page, and still if it doesn't work, you can write your queries to our developer - info.harshgupta10@gmail.com.

8:37 PM

Type your message...

If unable to view the content, please refresh the page.

```

Login: ec2-user | Password:12345
ec2-13-232-212-11 login: ec2-user
Password:
Last login: Thu Jul 18 14:58:22 from 13.126.219.100
[ec2-user@ec2-13-232-212-11 ~]$ jps
4449 Jps
4082 NameNode
3796 NodeManager
3660 ResourceManager
4284 DataNode
[ec2-user@ec2-13-232-212-11 ~]$ hdfs dfs -mkdir /demo
[ec2-user@ec2-13-232-212-11 ~]$

```

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

## Browse Directory

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	ec2-user	supergroup	0 B	7/18/2019, 8:40:49 PM	0	0 B	<a href="#">demo</a>

Hadoop, 2016.

Copyright © Your Website 2019

```

mysql> use project;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> desc register;
+-----+-----+-----+-----+-----+-----+
| Field      | Type          | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| name       | varchar(20)   | YES  |     | NULL    |       |
| email      | varchar(30)   | YES  |     | NULL    |       |
| username   | varchar(20)   | YES  |     | NULL    |       |
| password   | varchar(20)   | YES  |     | NULL    |       |
| instance_id | varchar(30)   | YES  |     | NULL    |       |
| region     | varchar(20)   | YES  |     | NULL    |       |
| public_ip  | varchar(20)   | YES  |     | NULL    |       |
| public_dns | varchar(100)  | YES  |     | NULL    |       |
| private_ip | varchar(20)   | YES  |     | NULL    |       |
| hadoop     | char(10)      | YES  |     | NULL    |       |
+-----+-----+-----+-----+-----+-----+
10 rows in set (0.03 sec)

```

- **Future Scope of the Project**

The package was designed in such a way that future modifications can be done easily. The following conclusions can be deduced from the development of the project.

- Automation of the entire system improves the efficiency
- It provides a friendly graphical user interface which proves to be better when compared to the existing system.
- It gives appropriate access to the authorized users depending on their permissions.
- It effectively overcomes the delay in communications.
- Updating of information becomes so easier.
- System security, data security and reliability are the striking features.
- The System has adequate scope for modification in future if it is necessary.

- **BIBLIOGRAPHY**

## **Big data Hadoop –**

### **Text Books:**

- Seema Acharya ,Subhashini Chellappan ,“Big Data and Analytics (WIND)”, Wiley, ISBN: 8126554789, 2015.
- Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, “Professional Hadoop Solutions”, Wiley, ISBN: 9788126551071, 2015.
- Chris Eaton, Dirk deroos et al. , “Understanding Big data ”, McGraw Hill, 2012.
- Alberto Cordoba, “Understanding the Predictive Analytics Lifecycle”, Wiley, 2014.

### **References:**

- Tom White, “HADOOP: The definitive Guide” , O Reilly 2012. 6 IT2015
- VigneshPrajapati, “Big Data Analytics with R and Haoop”, Packet Publishing 2013.
- Tom Plunkett, Brian Macdonald et al, “Oracle Big Data Handbook”, Oracle Press, 2014.
- Jay Liebowitz, “Big Data and Business analytics”,CRC press, 2013.

## **Cloud Computing -**

### **Text Books:**

- RajkumarBuyya, J.Broberg, A. Goscinski, “Cloud Computing Principles and Paradigms”, Wiley, 2011.



- A. Srinivasan, J. Surish “ Cloud Computing A Practical Approach for Learning and implementation””, Pearson, 2014.
- Ron Schmelzer et al. “XML and Web Services”, Pearson Education, 2002.
- Thomas Erl, “Service Oriented Architecture: Concepts, Technology, and Design”, Pearson Education, 2005.

## **References:**

- Barrie Sosinsky, “ Cloud Computing Bible”, Wiley., 2010
- Tim Mather, “Cloud Security and Privacy”, O'REILLY., 2009
- <https://docs.aws.amazon.com/>
- Frank P.Coyle, “XML, Web Services and the Data Revolution”, Pearson Education, 2002