

**Data Analytics Using AWS In Retail Industry**

**TY B.Tech. Mini Project Report**

**SUBMITTED BY**

Siddhant Jain- SCETYC22

Harshal Shinde- SCETYC06

Devendra Dhale- SCETYC08

**GUIDED BY**

**Prof. Shubhangi Kale**

**SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY**

**MIT ACADEMY OF ENGINEERING, ALANDI(D), PUNE-412105**

**MAHARASHTRA (INDIA)**

**MAY, 2020**



Data Analytics Using AWS In Retail Industry

**TY B.Tech. Mini Project Report**

*submitted in partial fulfillment of the*

*requirements for the award of the degree*

*of*

**Bachelor of Technology**

*in*

**INFORMATION TECHNOLOGY**

**BY**

**SIDDHANT JAIN, HARSHAL SHINDE, DEVENDRA DHALE**

**SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY**

**MIT ACADEMY OF ENGINEERING, ALANDI(D), PUNE-412105**

**MAHARASHTRA (INDIA)**

**MAY, 2020**



**CERTIFICATE**

It is hereby certified that the work which is being presented in the TY B.Tech. Mini Project Report entitled **“*Data Analytics Using AWS In Retail Industry*”,** in partial fulfillment of the requirements for the award of the **Bachelor of Technology in Information Technology** and submitted to the **School of Computer Engineering and Technology, Alandi(D), Pune, Affiliated to Savitribai Phule Pune University (SPPU), Pune** is an authentic record of work carried out during an Academic Year 2019-2020, under the supervision of **Name of Advisor(s),** **School of Electrical Engineering.**

**Siddhant Jain PRN No.** **120170445 Exam Seat No. T198029**

**Harshal Shinde PRN No.** **120170164 Exam Seat No. T198010**

**Devendra Dahale PRN No.** **120170193 Exam Seat No. T198012**

**Date:07/05/2020**

|  |  |
| --- | --- |
| *Signature of Project Advisor* | *Signature of Dean* |
| **Project Adviser** | **Dean** |
| School of Electrical Engineering, | School of Electrical Engineering, |
| MIT Academy of Engineering, Alandi(D), Pune | MIT Academy of Engineering, Alandi(D), Pune |

**(STAMP/SEAL)**

|  |  |
| --- | --- |
| *Signature of Internal examiner/s* | *Signature of External examiner/s* |
| *Name………………………………* | *Name………………………………* |
| *Affiliation…………………………* | *Affiliation…………………………* |

**ACKNOWLEDGEMENT**

*Thank all them who have helped in your project….*

*We would like to express our gratitude to the School of Computer Engineering and Technology of MIT Academy of Engineering, who encouraged us to do this minor project in Data Analytics using AWS in Retail Industry*

*We wish to express our profound thanks to our project guides, Mrs. Shubhangi Kale for their full cooperation, guidance and advices for the betterment of our project.*

*We are grateful that such interesting courses are included in the curriculum and for the proper support and guidance.*

*We would be failing in our duty if we do not thank all the other staff and faculty members for* *their experienced advice and evergreen co-operation*

1.Siddhant Jain sign

2.Harshal Shinde sign

3.Devendra Dahale sign

**ABSTRACT**

With the increasing competitiveness within the retail sector, it is extremely important that service processes are optimized in order to satisfy the customer expectations. So the purpose of the project is to analyze data and making changes adaptable to earn customer favorability is very important to generate profits.

Retail data analytics deal with identifying potential customers based on their past purchases, finding the most appropriate way to handle them via targeted marketing strategies and then deciding what the next offering should be.

The main functions of Data Analytics are: Price Optimization, Less Expensive Business Development Future, Performance Prediction and Demand Prediction

Retail data analytics enable retailers to gather information about customers and support them with increasing customer outreach as well as sales. The above functions help in utilizing the insights gained from data analytics for better risk management and improved performance.

|  |  |  |
| --- | --- | --- |
| **LIST OF FIGURES** | | |
| **Fig. No.** | **Fig. Name** | **Page No.** |
| **Fig 3.1** | **Block Diagram** | **5** |
| **Fig 3.2** | **Work Flow Diagram** | **6** |
| **Fig 4.1** | **Starting AWS ec2 instances** | **8** |
| **Fig 4.2** | **Launching Rstudio on ec2 with port no. 8787** | **9** |
| **Fig 4.3** | **Rstudio launched on AWS** | **9** |
| **Fig 4.4** | **Dashboard of shiny app consisting of data visualization** | **10** |
| **Fig 4.5** | **Sales Revenue according to country** | **10** |
| **Fig 4.6** | **Frequent Combination of product** | **11** |
| **Fig 4.7** | **Network of combination done using D3** | **11** |
| **Fig 4.8** | **Option in product combination using confidence** | **12** |
| **Fig 4.9** | **Recommendation to particular customer by id** | **12** |
| **Fig 4.10** | **Options in recommendation** | **13** |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **LIST OF TABLES** | | |
| **Table. No.** | **Table. Name** | **Page No.** |
| Table 2.1 | Literature Survey | 10 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| Acknowledgements | | | i |
| Abstract | | | ii |
| List of Figures | | | iii |
| List of Tables | | | iv |
|  | Introduction | | 1 |
|  | 1.1 | Motivation for the project |  |
|  | 1.2 | Problem Statement |  |
|  | 1.3 | Objectives and Scope |  |
|  | 1.4 | Organization of the report |  |
|  | Literature Survey | | 4 |
|  | System Design | | 7 |
|  | 3.1 | Block diagram/ Proposed System setup |  |
|  | 3.2 | Circuit Diagram (If Applicable) and explaination |  |
|  | 3.3 | Mechanical Drawing (If Applicable) |  |
|  | 3.4 | Use Case Diagram (If Applicable) |  |
|  | 3.5 | Sequence Diagram (If Applicable) |  |
|  | 3.6 | Activity Diagram (If Applicable) |  |
|  | 3.7 | Related mathematical modelling |  |
|  | 3.8 | Hardware and Software Requirements |  |
|  | Implementation and Results | | 22 |
|  | 4.1 Algorithm and flowcharts (If applicable) | |  |
|  | 4.2 Results | |  |
|  | 4.3 Discussion | |  |
|  | Conclusion and Future scope | | 25 |
| References | | | 26 |
| Appendix    (Datasheets/ Psuedo Code/Report Etc…(If applicable)) | | |  |

1. **INTRODUCTION**

As we become a more digital society, the amount of data being created and collected is growing and accelerating significantly. Analysis of this ever-growing data becomes a challenge with traditional analytical tools. We require innovation to bridge the gap between data being generated and data that can be analyzed effectively.

Big data tools and technologies offer opportunities and challenges in being able to analyze data efficiently to better understand customer preferences, gain a competitive advantage in the marketplace, and grow your business. Data management architectures have evolved from the traditional data warehousing model to more complex architectures that address more requirements, such as real-time and batch processing; structured and unstructured data; high-velocity transactions; and so on.

Amazon Web Services (AWS) provides a broad platform of managed services to help you build, secure, and seamlessly scale end-to-end big data applications quickly and with ease. Whether your applications require real-time streaming or batch data processing,

AWS provides the infrastructure and tools to tackle your next big data project. No hardware to procure, no infrastructure to maintain and scale—only what you need to collect, store, process, and analyze big data. AWS has an ecosystem of analytical solutions specifically designed to handle this growing amount of data and provide insight into your business.

The AWS Advantage in Big Data Analytics Analyzing large data sets requires significant compute capacity that can vary in size based on the amount of input data and the type of analysis. This characteristic of big data workloads is ideally suited to the pay-as-you-go cloud computing model, where applications can easily scale up and down based on demand. As requirements change, you can easily resize your environment (horizontally or vertically) on AWS to meet your needs, without having to wait for additional hardware or being required to over invest to provision enough capacity. For mission-critical applications on a more traditional infrastructure, system designers have no choice but to over-provision, because a surge in additional data due to an increase in business need must be something the system can Amazon Web Services – Big Data Analytics Options on AWS handle. By contrast, on AWS you can provision more capacity and compute in a matter of minutes, meaning that your big data applications grow and shrink as demand dictates, and your system runs as close to optimal efficiency as possible.

In addition, you get flexible computing on a global infrastructure with access to the many different geographic regions that AWS offers, along with the ability to use other scalable services that augment to build sophisticated big data applications. These other services include Amazon Simple Storage Service (Amazon S3) to store data and AWS Glue to orchestrate jobs to move and transform that data easily. AWS IoT, which lets connected devices interact with cloud applications and other connected devices. As the amount of data being generated continues to grow, AWS has many options to get that data to the cloud, including secure devices like AWS Snowball to accelerate petabyte-scale data transfers, delivery streams with Amazon Kinesis Data Firehose to load streaming data continuously, migrating databases using AWS Database Migration Service, and scalable private connections through AWS Direct Connect. AWS recently added AWS Snowball Edge, which is a 100 TB data transfer device with on-board storage and compute capabilities.

The following services for collecting, processing, storing, and analyzing big data are described in order:

• Amazon Kinesis

AWS Lambda

• Amazon Elastic MapReduce

• Amazon Glue

• Amazon Machine Learning

• Amazon DynamoDB

• Amazon Redshif

t • Amazon Athena

• Amazon Elasticsearch Service

• Amazon QuickSight

In addition to these services, Amazon EC2 instances are available for selfmanaged big data applications.

Data movement is also big part of data analytics so for that specifically For Real-time data movement we have [Amazon Kinesis, Data Firehouse](https://aws.amazon.com/kinesis/data-firehose/?c=a&sec=srv), [Amazon Kinesis Video Streams](https://aws.amazon.com/kinesis/video-streams/?c=a&sec=srv), [Amazon Kinesis Data Streams](https://aws.amazon.com/kinesis/data-streams/?c=a&sec=srv), [Amazon Kinesis Data Analytics](https://aws.amazon.com/kinesis/data-analytics/?c=a&sec=srv)

Now for storing the processed data we have Data Lake and also categorizing that data For Object storage we have [Amazon S3](https://aws.amazon.com/s3/?c=a&sec=srv) and Formation. Now we see backup is a essential part of data analytics so For Backup and archive we have [Amazon S3 Glacier](https://aws.amazon.com/glacier/?c=a&sec=srv) and [AWS Backup](https://aws.amazon.com/backup/?c=a&sec=srv).

AWS also serve us with some inbuilt algorithm such as [AWS Deep Learning AMIs](https://aws.amazon.com/machine-learning/amis/?c=a&sec=srv) Platform services

* 1. **Motivations**
     + With the growing e commerce market. It is necessary to understand what a customer want and what he is willing to pay for that product.
     + At different regions and area what are the different needs of the product which we need to understand
     + What could be the right price of the product that should attract more customers.
     + Any retail industry should be able to make right decision in future by analysis their past data.
  2. **Problem Statement**

With the help of AWS implement data analysis for retail industry.

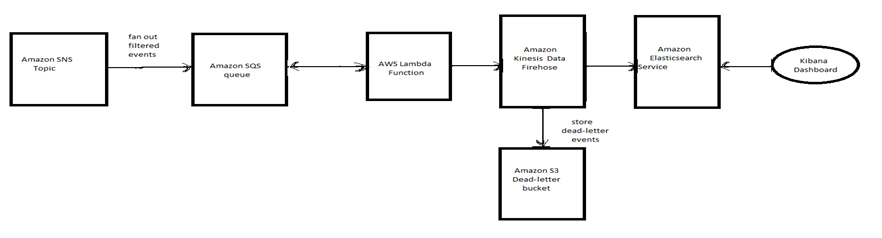
* 1. **Objectives and Scope**
* To determine when prices are to be dropped -Price Optimization
* Future performance prediction
* To accommodate small-scale retailers
* To analyze what needs to be promoted by marketing departments and what is not –Forecasting trends
* Which customers would likely desire a certain product- Customer Identification

1. **LITERATURE / Requirement Gathering SURVEY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SR. NO.** | **TITLE OF PAPER** | **AUTHORS** | **PUBLISHER** | **KEY FINDINGS OF THE AUTHORS** | **REMARKS** |
| 1 | Cloud Based Big Data Analytics A Review | Amitkumar Manekar and G. Pradeepini | IEEE | There are many advantages that cloud offers over on-premise data analytics implementation. | Cloud is able to deploy direct interconnections between data and analytics and can reduce latency to a great extent. |
| 2 | Google Cloud and Analysis of Realtime Process Data | R. Langmann | IEEE | A need for more efficient data processing techniques is important to make sure that real-time visualisation is done. | Hosting analytics on public cloud platforms poses a major problem of data security and privacy that is a vast topic to research upon. |
| 3 | Data Analytics using Cloud Computing | Prakhar Maheshwariand Alankar Singha | IEEE | Organisations are able to consolidate data from all sources in cloud without which collecting data from all internal applications and social networks | The techniques used for analytics vary according to the requirements of the organization. |
| 4 | Predictive Analytics for Banking User Data using AWS | Ranjith Ramesh | IEEE | Amazon Machine Learning is a service that makes it was for all skill levels to use machine learning technology along with AWS | The binary classification model is sued to mainly predict binary outputs Most of the binary classification are a yes or no questions |

**Table.2.1**

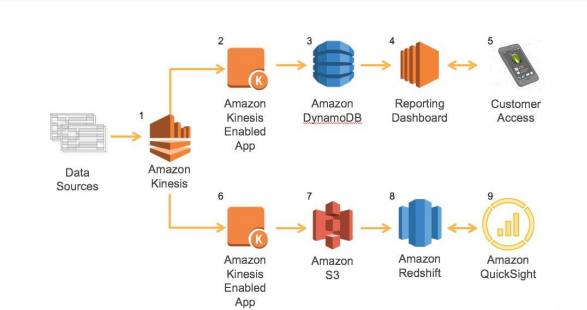
1. **SYSTEM DESIGN**



**FIG 3.1 BLOCK DIGRAM**

**WORKING:-**

* Connect to raw data from any sources
* Ingest the data into a storage solution with any form at any speed
* Store the variety of data (structured, semi structured or unstructured) in a scalable, durable and secure solution
* Process this data to transform it into usable data sets ready to be consumed and analyzed (aggregated, joined…). These data sets needs to be stored in a different location than raw data
* Visualize the final data with business intelligence tools and dashboards to seek answers and insights. Machine learning is used in this step to apply algorithms to go beyond usual analytics.

****

**Fig 3.2-Work Flowagram**

* The process begins with each A/C unit providing a constant data stream to Amazon Kinesis Data Streams. This provides an elastic and durable interface the units can talk to that can be scaled seamlessly as more and more A/C units are sold and brought online.
* Using the Amazon Kinesis Data Streams-provided tools such as the Kinesis Client Library or SDK, a simple application is built on Amazon EC2 to read data as it comes into Amazon Kinesis Data Streams, analyze it, and determine if the data warrants an update to the real-time dashboard. It looks for changes in system operation, temperature fluctuations, and any errors that the units encounter.
* This data flow needs to occur in near real time so that customers and maintenance teams can be alerted as quickly as possible if there is an issue with the unit. The data in the dashboard does have some aggregated trend information, but it is mainly the current state as well as any system errors. So, the data needed to populate the dashboard is relatively small. Additionally, there will be lots of potential access to this data from the following sources: o Customers checking on their system via a mobile device or browser o Maintenance teams checking the status of its fleet o Data and intelligence algorithms and analytics in the reporting platform spot trends that can be then sent out as alerts, such as if A/C fan has been running unusually long with the building temperature not going down. DynamoDB was chosen to store this near real-time data set because it is both highly available and scalable; throughput to this data can be easily scaled up or down to meet the needs of its consumers as the platform is adopted and usage grows.
* The reporting dashboard is a custom web application that is built on top of this data set and run on Amazon EC2. It provides content based on the system status and trends as well as alerting customers and maintenance crews of any issues that may come up with the unit.
* The customer accesses the data from a mobile device or a web browser to get the current status of the system and visualize historical trends.

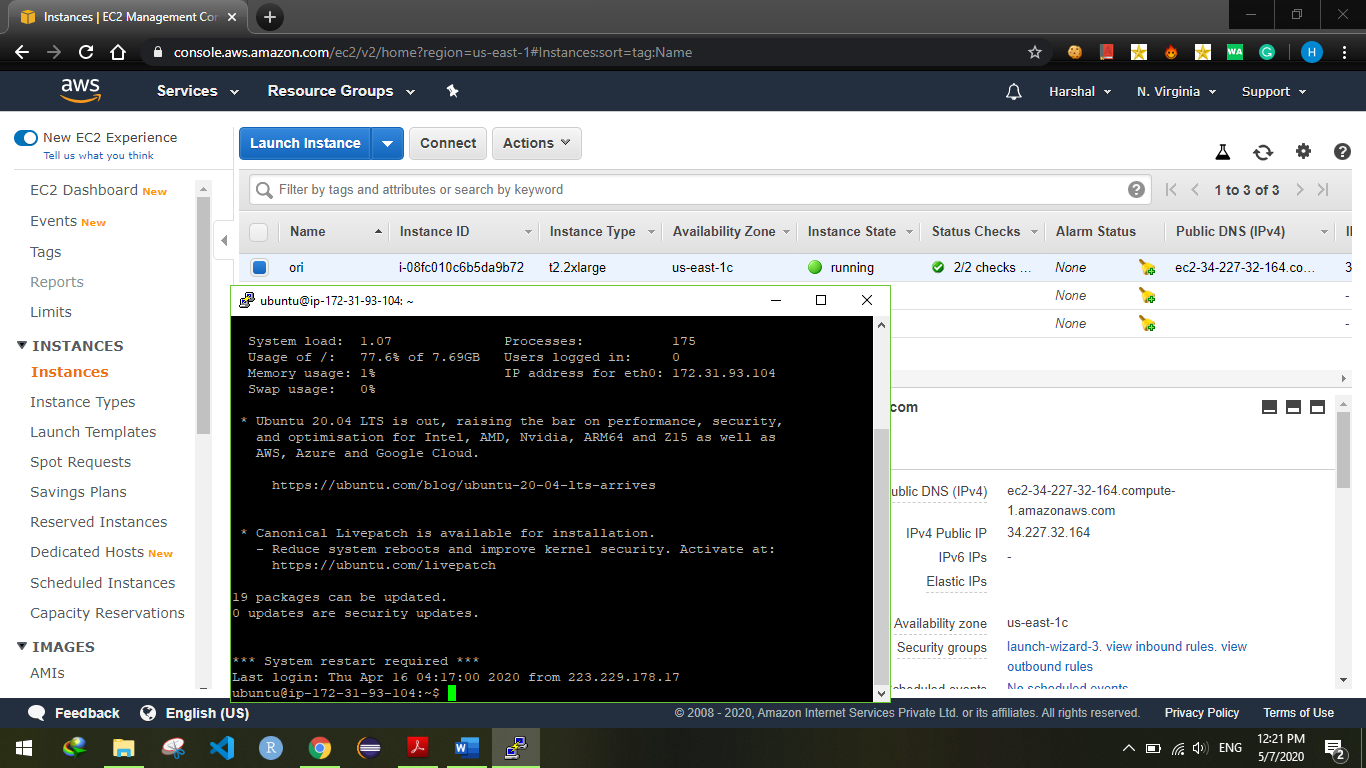
**CONTRIBUTORS:-**

The following individuals and organizations contributed to this document: • Erik Swensson, Manager, Solutions Architecture, Amazon Web Services • Erick Dame, Solutions Architect, Amazon Web Services • Shree Kenghe, Solutions Architect, Amazon Web Services

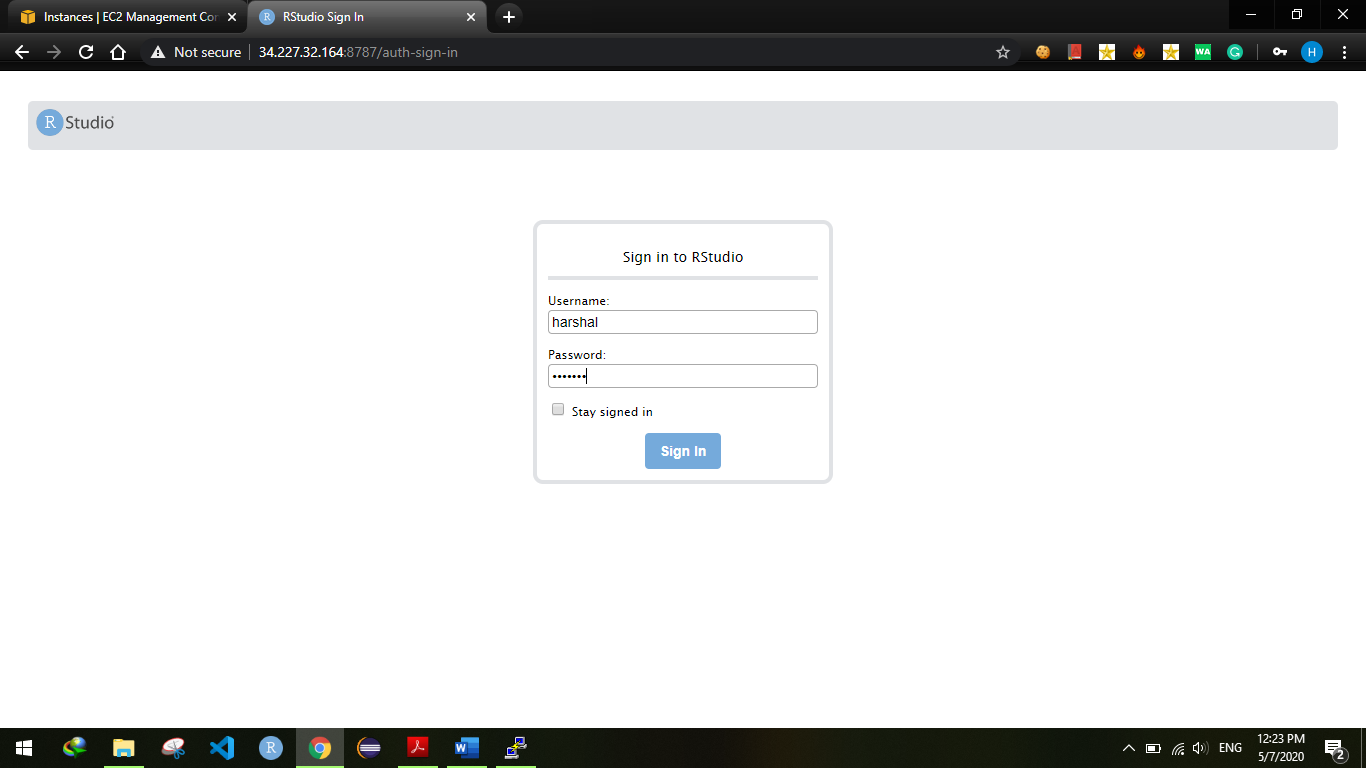
1. **IMPLEMENTATION DETAILS**

|  |
| --- |
| * We used tidyverse for data manipulation and visualization, and Spark for data processing. * R sparklyr package will enable me to do data analysis using the typical tidyverse approach, * while leveraging the power of Spark in the background. As I don’t have a real Spark cluster at my disposal, * We used a local Spark instance, which comes with the sparklyr package. * Compared to using a cluster syntax - wise, only the connection string is different. * After those basic steps, * We employ machine learning algorithms on Spark to uncover more complex customer behavior patterns, * like which products are frequently purchased together. |

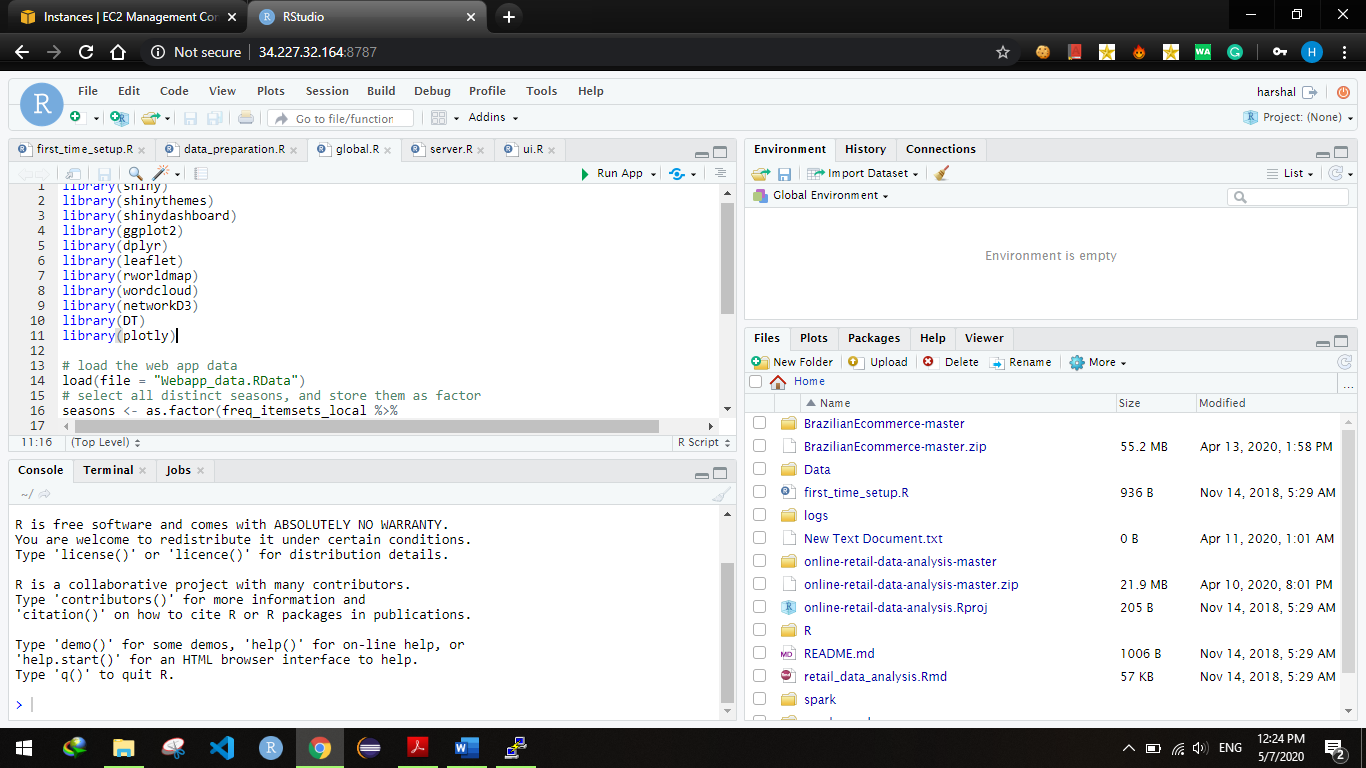
**Screenshots of implementation of Data analytics in retail Industry using AWS**



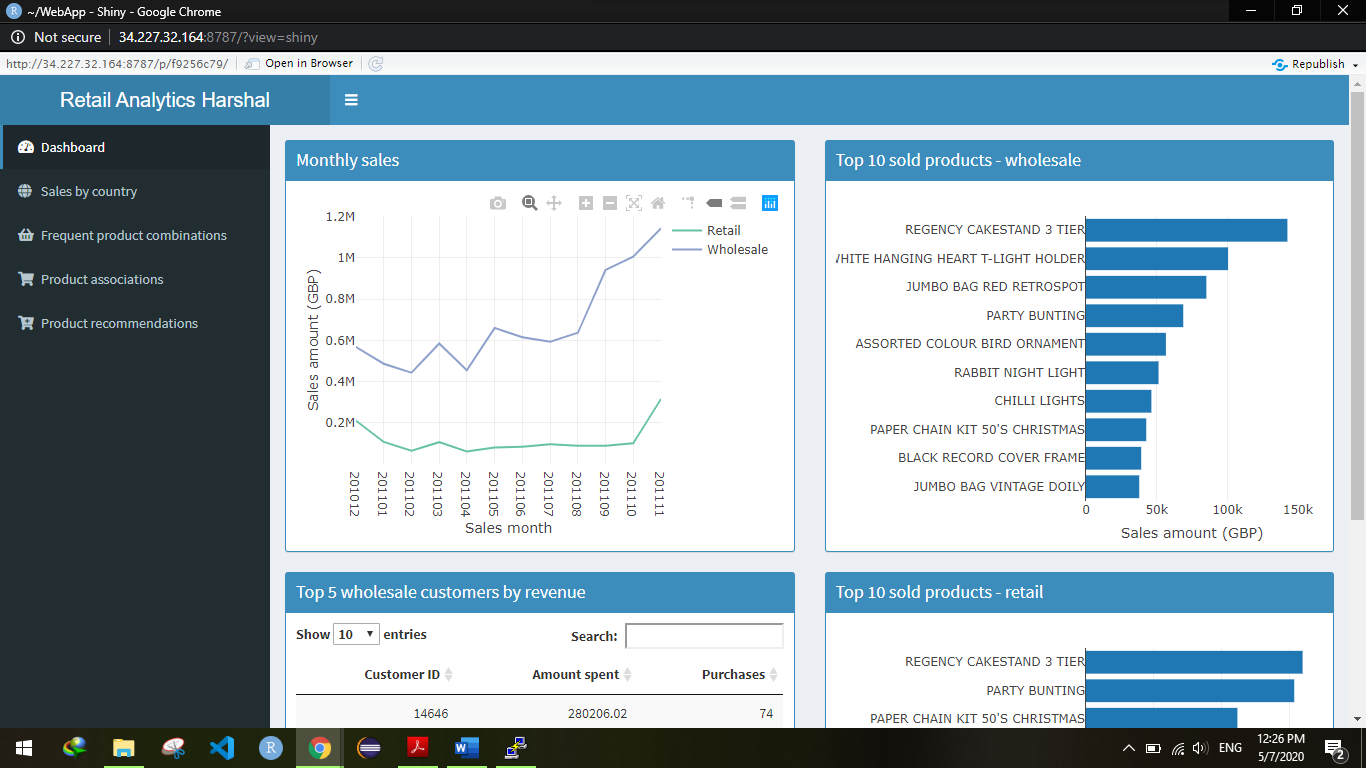
**Fig 4.1 Starting AWS ec2 Instance**



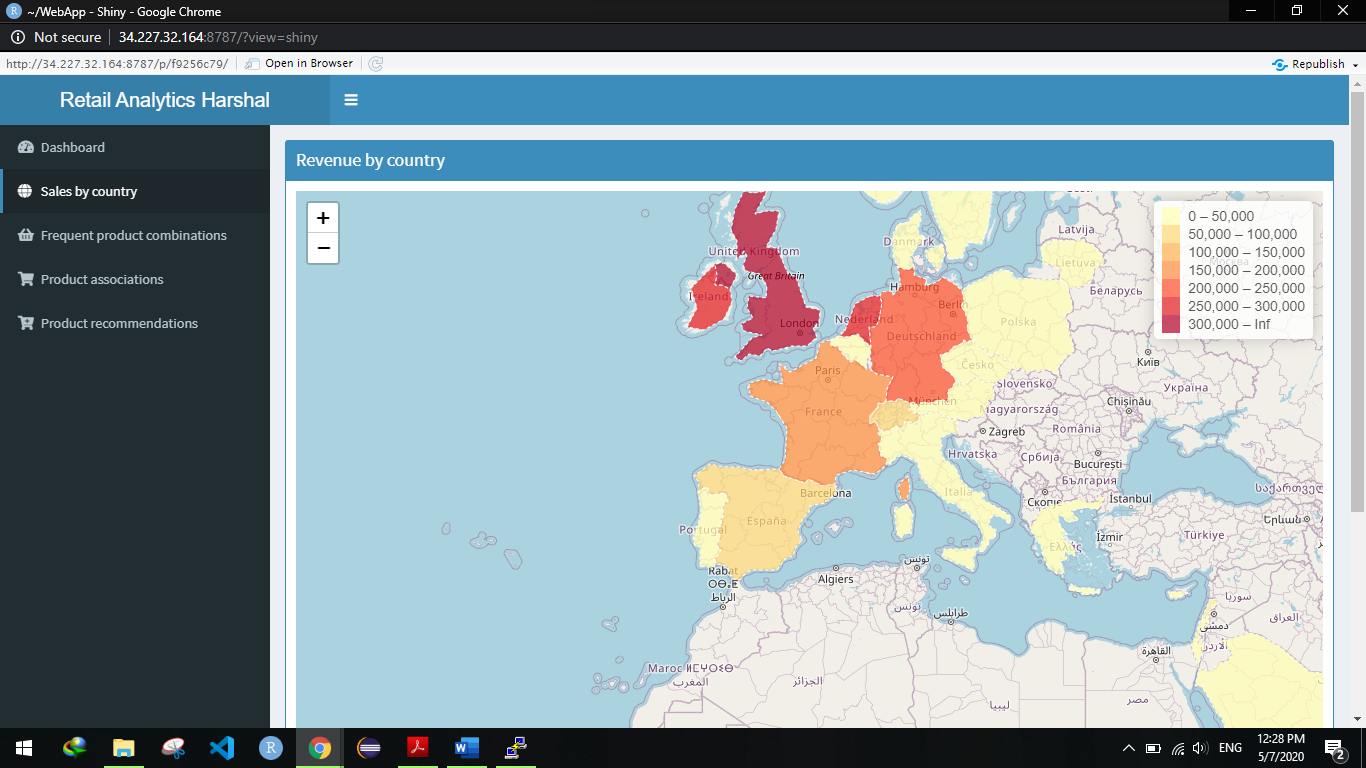
**Fig 4.2 Launching Rstudio on ec2 with port no. 8787**



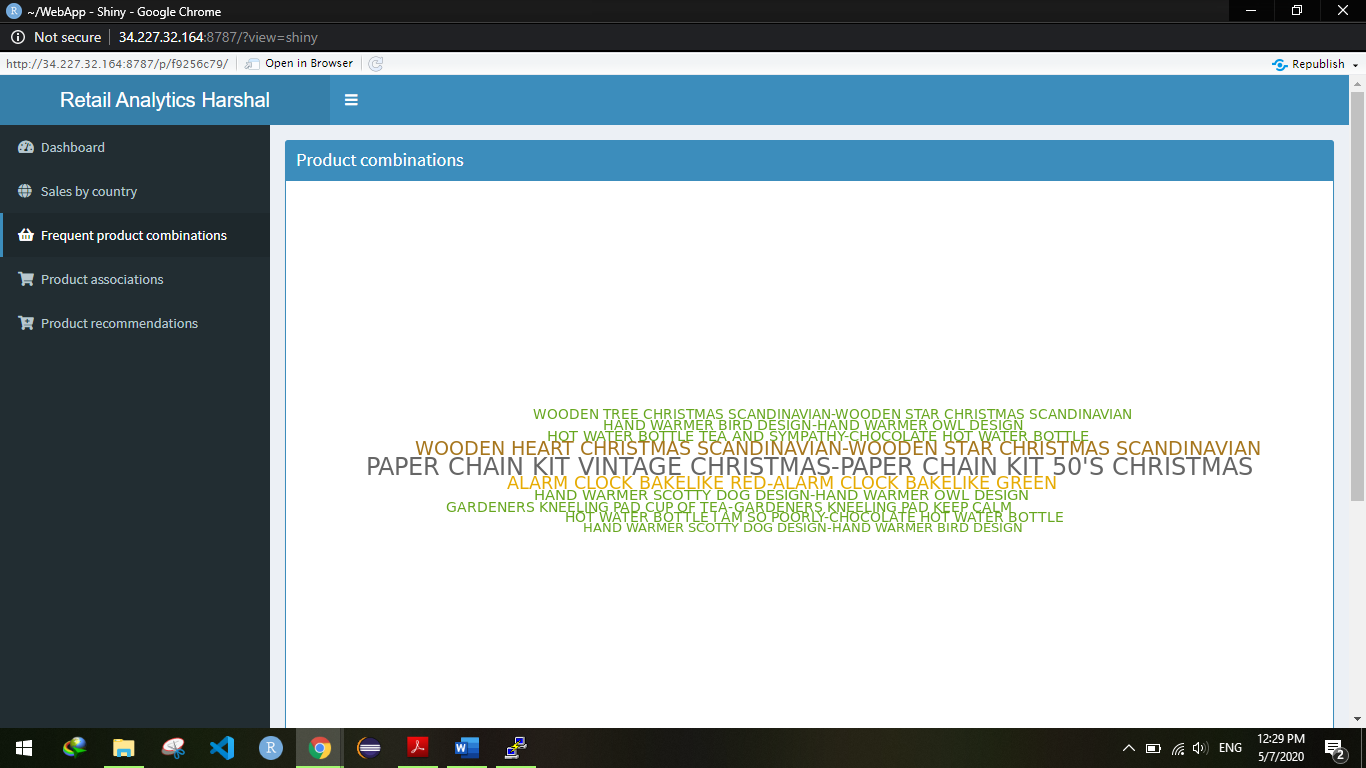
**Fig 4.3 Rstudio launched on AWS**



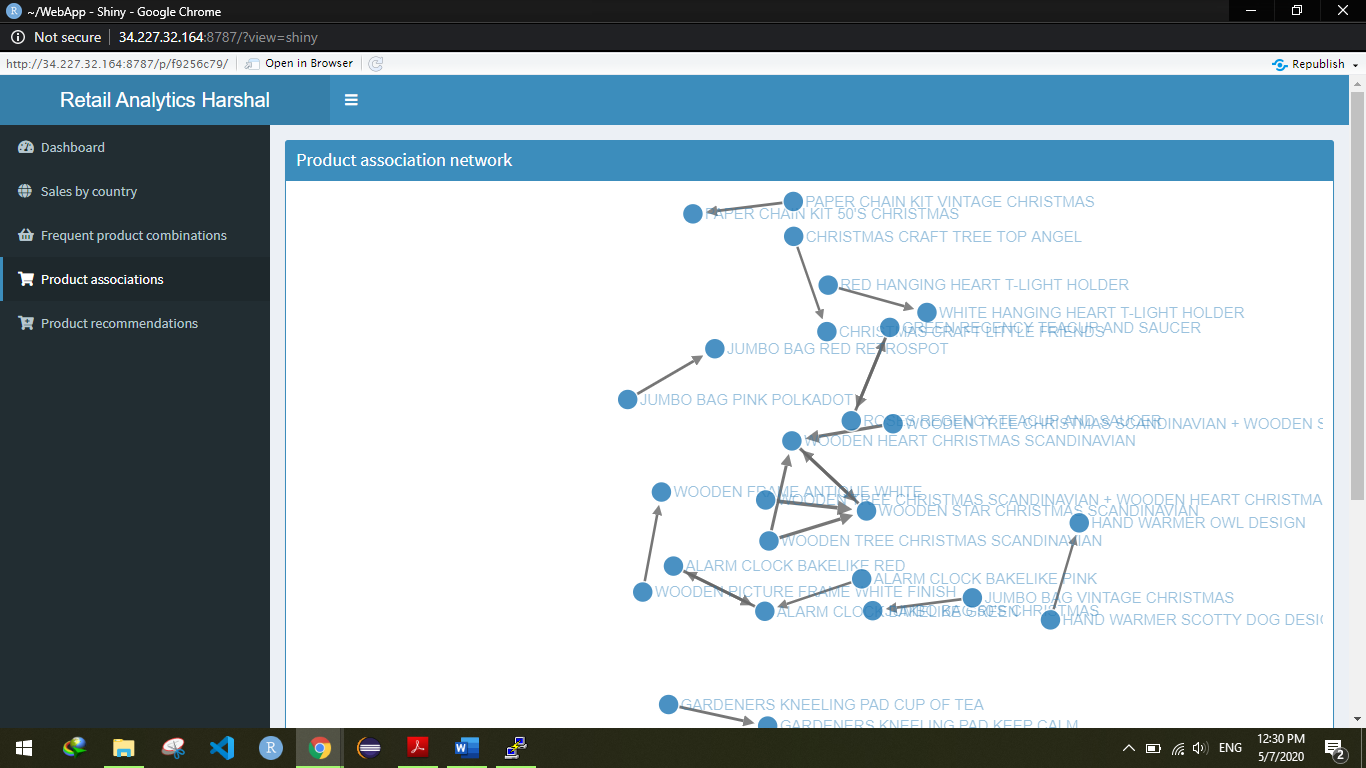
**Fig 4.4 Dashboard of shiny app consisting of data visualization using line graph and bar charts**



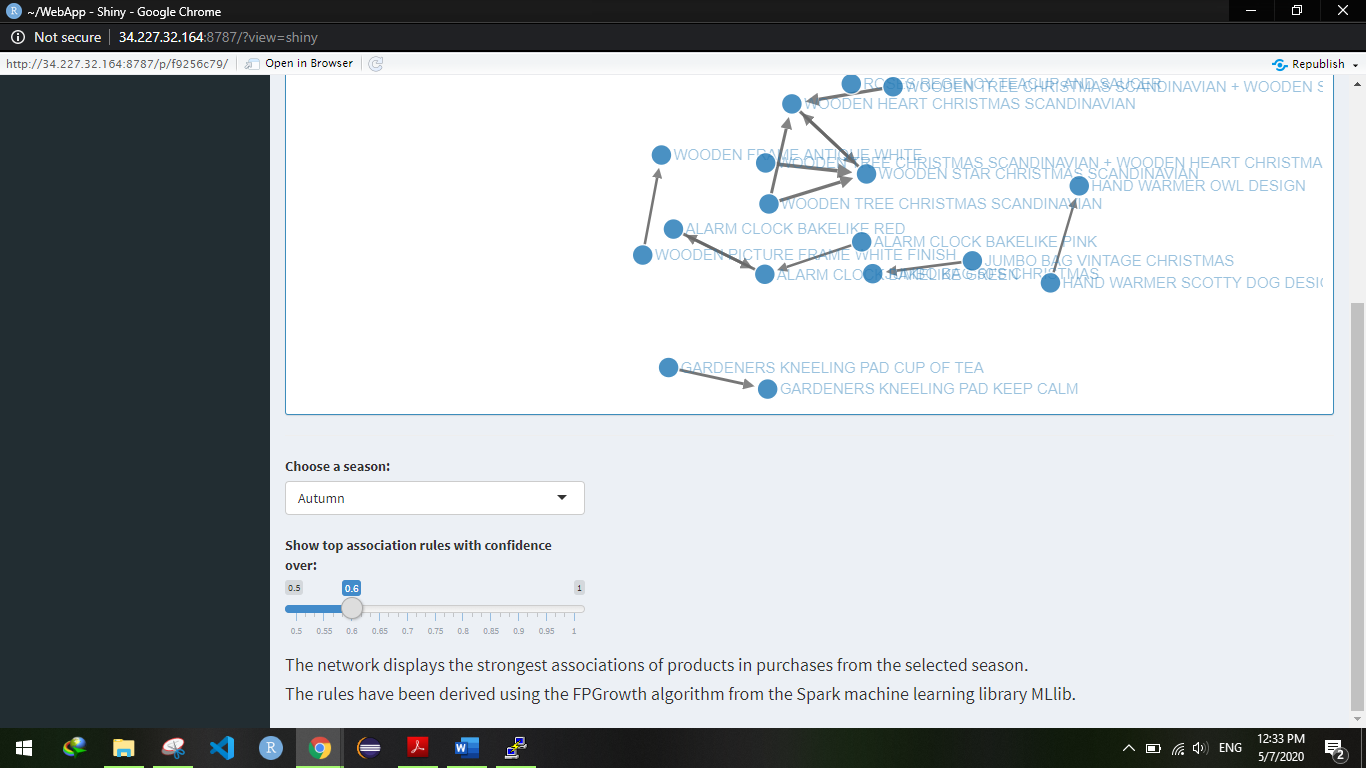
**Fig 4.5 Sales(revenue) according to country**



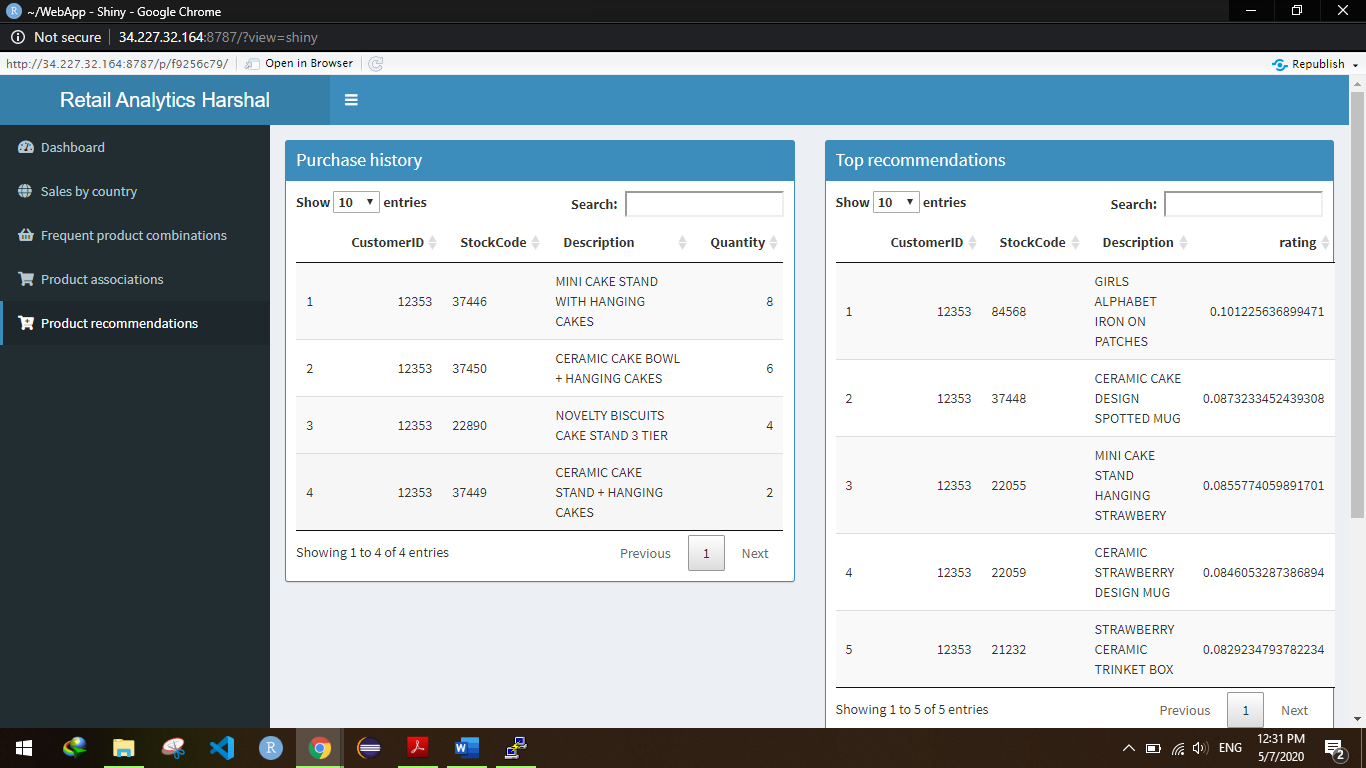
**Fig 4.6 Frequent combination of product (each line represents on shopping combination)**



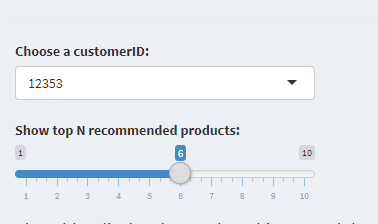
**Fig 4.7 Network of product combination done using D3 network in R**



**Fig4.8 Option in product combination to adjust according to season and form association rules using confidence**



**Fig4.9 Recommendation to particular customer by customer id**



**Fig 4.10 Options in recommendation to choose customer id and no. of products**

1. **CONCLUSION & FUTURE SCOPE**

As more and more data is generated and collected, data analysis requires scalable, flexible, and high performing tools to provide insights in a timely fashion. However, organizations are facing a growing big data ecosystem where new tools emerge and “die” very quickly. Therefore, it can be very difficult to keep pace and choose the right tools. This whitepaper offers a first step to help you solve this challenge. With a broad set of managed services to collect, process, and analyze big data, the AWS platform makes it easier to build, deploy, and scale big data applications. This allows you to focus on business problems instead of updating and managing these tools. AWS provides many solutions to address your big data analytic requirements. Most big data architecture solutions use multiple AWS tools to build a complete solution. This approach helps meet stringent business requirements in the most cost-optimized, performant, and resilient way possible. The result is a flexible, big data architecture that is able to scale along with your business.

**FUTURE SCOPE:-**

The future of AWS is bright. Infact, it's the future of modern day computing. Cloud computing, machine learning, IOT, etc are some of the domains which have a lot to offer in the near future.

With Amazon's latest addition here in India, Alexa, Uber's already operating self-driven vehicles in Florida and parts of the US, Tesla's Space-X project (rather Elon Musk's project) surely gives an illusion of what is coming in the upcoming years.

AWS made an entry way back in 2006, and today they have users exceeding 1,000,000. Here is just a instance on the growth rate of AWS S3 (Simple Storage Service), one of the most popular.

**REFERENCES**

1. Divyakant Agrawal ,Sudipto Das and Amr El Abbadi “Big Data andCloud Computing: Current State and Future Opportunities”, EDBT2011 , March 22–24, 2011, Uppsala, Sweden ACM 978-14503-0528-0/11/0003.
2. Gupta, Rajeev, Himanshu Gupta, and Mukesh Mohania, Cloud computing and big data analytics: what is new from databases perspective?.., International Conference on Big Data Analytics. Springer Berlin Heidelberg, 2012
3. Tharam Dillon et. al. “Cloud Computing: Issues and Challenges”,2010 24th IEEE International Conference on Advanced Information Networking and Applications, 2010 IEEE DOI10.1109/AINA.2010.187, pp- 27-33.
4. Fiore, Sandro, et al, Big data analytics for climate change and biodiversity in the EU Brazil CC federated cloud infrastructure., Proceedings of the 12th ACM International Conference on Computing Frontiers. ACM, 2015
5. Talia, Domenico, Toward cloud-based big-data analytics. IEEE Computer Science (2013): 98-101.

**Appendix (If applicable)**

This section can have:

1. Datasheets
2. Psuedo Code
3. Supplementary simulation
4. Theorem, proofs, Leema etc
5. Log Reports etc….
6. Any other technical matter which is not in the report but may be related to project

Any of the above can be in this report like this:

A1: Datasheets

A2: Log Reports