2018 ASEE ANNUAL CONFERENCE & EXPOSITION
JUNE 24–27, 2018
SALT PALACE CONVENTION CENTER
SALT LAKE CITY, UT

Paper ID #23997

Implementation of Big Data Lab for Broadband Wireless Networks Intelligent Traffic Management System: Evaluation and Challenges

Dr. Tamer Omar, California State Polytechnic University, Pomona

Tamer Omar is an Assistant professor with the Electrical and Computer Engineering Department at California State Polytechnic University. Dr. Omar earned his Ph.D. from the Electrical Engineering department at Iowa State University, USA and his MBA with emphasis on MIS from the Arab Academy for Science and Technology, Egypt and his B.S. degree in Electrical Engineering from Ain Shams University, Egypt. Dr. Omar research interests include wireless networks architecture, resources allocation in wireless networks, heterogeneous networks, self-organized networks, big data implementation and analysis, RDBMS and decision support systems. Dr. Omar has 8 years of experience in academia and more than 10 years of industrial experience in different ICT positions.

Implementation of Big Data Lab for Broadband Wireless Networks Intelligent Traffic Management System: Evaluation and Challenges

ABSTRACT

Mobile data traffic is growing in an unprecedented rate. Mobile service providers and other businesses relying on mobile traffic require talented calibers to hire with proper skills to operate and manage their networks. Broadband wireless networks and big data systems are two important technologies that current STEM students need to learn, comprehend and master to satisfy the market needs. Design and implementation of an academic big-data system and broadband wireless testbed for instruction and research purposes is a difficult task. In this work, challenges facing the design and implementation of a mobile networks and big-data lab are evaluated. This work aims at providing a comprehensive reporting about an experience gained from designing and implementing an academic lab of big-data system used for broadband wireless networks traffic analysis and management. Challenges facing the project team during the implementation are discussed and probable solutions are described. Lessons learned from different project milestones are detailed to highlight the advantages and disadvantages of different project paths adopted by the project team. Finally, recommendations to other teams willing to create similar labs are presented.

1 INTRODUCTION

With advances in mobile standards, wireless technologies are getting more reliable. More users are now dependent on wireless technologies for their business and recreational activities than ever before. This tremendously increase the amount of data transfer over mobile network. Asp per Cisco® global mobile data traffic forecast, data traffic grew 63 percent in 2016, 18-fold over the past 5 years (Cisco®, 2017). These statistics indicate the importance of mobile network as the future carrier for human communications.

The future 5G mobile networks depends on small-cells technology to increase the area spectral efficiency of the new network. This shift in technology implementation will have a great effect on the way these networks are managed. According to Bell labs the new mobile networks will rely on big-data systems to manage the network resources (Weldon, 2015). With this massive expected increase in the number of small-cells that will operate the mobile network, big-data can be used to monitor the operations of mobile networks. Using the available tools in big data systems can help collect, organize, and analyze traffic in the network.

Previously the author proposed the design, implementation and utilization aspects of an educational big-data system that imitate Mobile Service Providers (MSPs) systems to delve deep into their data stores. The purpose of the educational system was to introduce students to different concepts of big-data systems, mobile networks, systems integration, and mobile networks traffic analysis, and data management (Tamer Omar, 2016). The paper introduced the big-data system architecture together with the analytics framework designed to support the mobile network.

The proposed project to implement the lab resources was planned over three years with an estimate budget of hundred thousand dollars. During this period the project team leader was planning to implement a production mobile testbed from small-cells network, a big-data system using one of the open sources Hadoop distributions, and an anlytics system to handle the traffic anlysis of the mobile network. During the first two years of the project life cycle, the project was intrrupted with different unexpected challenges. The result of them was the change in the project scope, implementation plan and time duration.

The purpose of this paper is to highlight the project challenges and report the lesson learned from the actual implementation of the different systems. The study will also recommend applicable solutions for implementing both an acedmic big-data systems and an academic mobile networks in an educational enviroment. The paper will discuss the human resources and equipment required for sucessful implementation of the lab. Satkeholders opinions will be presented and scope defintion will be described to gain agreement from all stakeholders.

2 LAB SCOPE and PRIMARY DESIGN

The primary scope of this educational lab includes three phases for design, implementation and running an academic broadband (4G) network supported by a bid-data system and analytics frame work. The infrastructure used in developing the lab environment was proposed to support hands on to student for studying introductory concepts of broadband wireless networking and big-data systems. Also, analytics tools was integrated as a lab component to introduce students to powerful tools that can help in analyzing data within the big data repository.

The primary design includes a proposal for implementing the following systems; A production broadband wireless 4G network relying on small-cells. A big-data system with an open source distribution based on hadoop distributed file system as a data plate-form. Finally, a data Analytics frame-work from multiple tools integrated in the big-data system distribution in additon to educatinal licensed software anlytic tools.

2.1 Original Lab Architecture

The lab architechture shown in Figure 1 represent the original design of the lab. The three main componants of the lab integrated throught the core switch located in campus network.

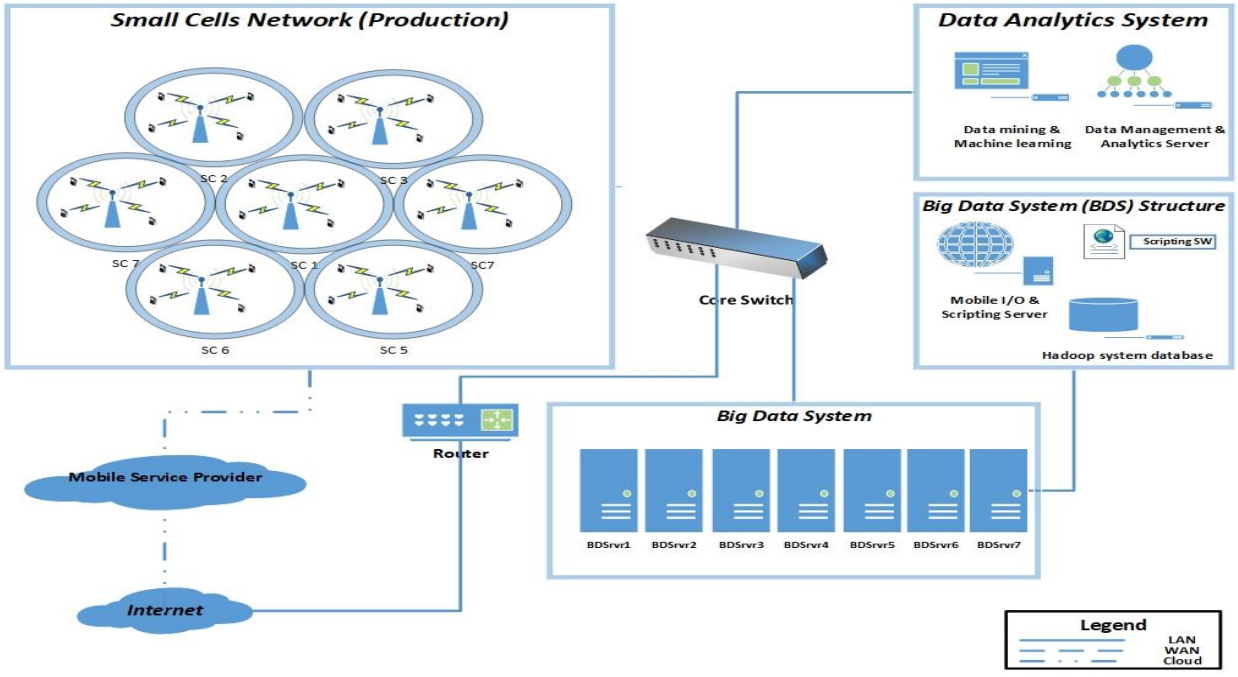

Figure 1: Lab Architecture (Design)

2.2 Wireless Network Design

The original design intended to use Mobile Service Provider (MSP) to support the 4G broadband wireless network. The planned test-bed was designed to relay on the MSP core network to provide a production network for educational purposes. Only data traffic was planned to be

utilized on the network, voice services is prohibited. Users of the network will create the required traffic for educational purposes only and the produced traffic is supposed to be under user license agreement to avoid data privacy limitations. The small-cells network resources were supposed to be under the control of the lab management team. Only meta-data used for communication signaling and traffic control was planned to be collected in the big-data system for traffic network and traffic management teaching and research purposes.

2.3 Big-Data System Design

The big data system requires a computational infrastructure with scalable resources. Since this environment need to be isolated to the lab environment only, a virtualization system relying on VMware® was planned to be used to provide the required virtualization for the HDFS cluster data nodes, name nodes and other required services.

The original big-data system intended to collect data from the wireless network was planned to use Hortonworks® distribution (HORTONW0RKS®, 2017). Hortonworks® is an open source HDFS distribution that provide a data plate form include, but not limited to, an Ambary server, Map-Reduce, Hive, SQL and No SQL, and spark. The plate-form uses the ambary server to populate and manage the data nodes, name nodes, and job tracker required to implement the big-data system.

2.4 Analytics framework

Big data systems are mostly populated by unstructured or semi-structured data. In case of this traffic management system, the data was expected to be totally unstructured in the form of traffic captures using traffic probes or sniffers such as Wireshark®. However, this type of collected data in the HDFS is unrecognized by the staging tools. Signaling and traffic management data need to be parsed using parsing programs to transform the traffic into a format recognized by the staging tools.

The analytic framework plan depends on data organizing, cleansing, and analyzing tools. Hive, SQL and spark are examples of these tools used for staging the data and transforming it into structured data suitable for data analysis and modeling the algorithms required for traffic optimization.

The next phase after extracting, transforming, and loading data is to use analytics tools such as R® and SAS®, the analytic tools chosen for this lab, to analyze and model the data. Both programs need to be integrated with the Hadoop distribution and use the HDFS repository as the data source

3 IMPLEMENTATION CHALLENGES

The lab implementation encountered different challenges that slow the implementation progress in addition to have an impact on the lab design. These challenges are related to either security, interference with current systems, or human resources concerns. The university research office was willing to fund the lab; however, the information technology and computing services office have multiple concerns to support the lab implementation.

3.1 Wireless Network

One of the challenges in implementing the lab environment is related to the wireless network. The problem presents in the existence of a Distributed Antenna System (DAS) in the football stadium in addition to a plan for improving cellular coverage on campus by using DAS technology to improve cellular coverage. The responsibility of the DAS management rests with the ITCS department and is supported from an external vendor. ITCS claims that introducing the small cells lab will result in interference to the existing DAS infrastructure and will complicate trouble-shooting in the event of an outage and accordingly will place additional burden upon ITCS.

The ITCS department decides that the test bed project will provide uncertainty for ITCS to ensure the reliability of the present cellular infrastructure. They require that the full control of the cellular must be managed by ITCS staff. They considered the lab as a test system that they cannot imbedded within or linked to any production-based systems. The ITCS department evaluates that introducing a small cell technology will require some of their human resources who are already overburdened with projects and that the lab test-bed will add another layer of troubleshooting for that is not fully owned and managed by central IT.

The ITCS department according to their final evaluation decide that the lab test-bed aims at facilitating an academic initiative and should not be considered a production-based environment nor interfere with such an environment that serves the entire campus and its constituents. They recommend that the lab may be best served by a joint venture or contract directly with a cellular provider that may offer an avenue to provide coverage in an area that is not impacted by the university DAS.

According to design concepts of cellular networks both DAS and small-cell systems can coexist. Each system can use different frequency band from a different MSPs. This design will mitigate any form of interference between the running production systems serving the university constituents and the lab test-bed. Another solution is using carful coverage design to provide specific areas for lab small-cells coverage, this design will allow both systems to coexist and share the same frequency bands. The main challenge in implementing the small-cells wireless network is the resistance of ITCS to allow the implementation of production system on campus, but not for valid technical reasons. In addition to the reluctance of ITCS to support such system due to their staff overload.

One of the major challenges that face the test-bed after ITCS rejection to allow the implementation of a production system is to provide the wireless network core system functionalities. Generally, the core network in broad-band wireless networks is complicated and require multiple components to be emulated. The updated lab architecture in the upcoming section shows the design considerations to allow a fully functional wireless network core.

3.2 Big-Data System

The challenges of big-data system implementation can be summarized into two main categories. The first category is considered with campus network accessibility and security credentials. System virtualization is one safe method that can be used to implement a big data system in lab-

environment. However, in order to implement the big-data system, a subnet need to be dedicated to the system data nodes in addition to a gateway that can allow the system updates from the internet. Internet accessibility for systems outside ITCS control is considered a threat on the campus network particularly if the administration privileges of the system is granted to the end user. Access control using subnetting is a proper solution to isolate the lab environment from other production systems running on the network.

The second category is considered with the high complexity of the open source big-data distributions (e.g. Hortonworks®) structure. Building big-data system require different skills to build the system (Linux, MS windows server, and application server). These skills are required to build the applications needed to run the big data system. Usually this skill set is not available to a single technologist and is distributed of different administrators in industry. These challenges make individual efforts to implement the system without external consultancy or an implementation vendor a cumbersome task. Figure 2 shows the Ambari application server interface used for managing the big data cluster of 7 data nodes. The management console is used to access all the tools available with this Hortonworks® Hadoop distribution (e.g. MapReduce, hive, etc.).

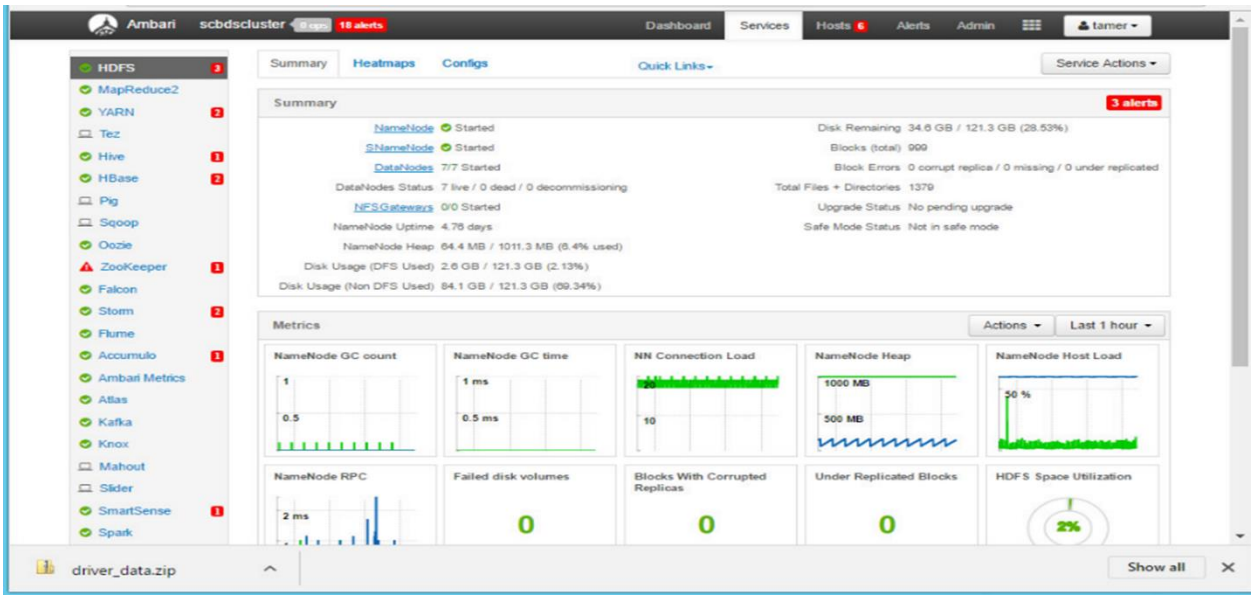

Figure 2: Ambari Application Server

There are two approaches that can be used to implement big-data systems. The first approach is to implement your own internal HDFS and Apache application server. The second approach is to relay on big-data cloud service from one of the providers available in the market. The advantage of the first approach is that it's cost effective, however as mentioned before the disadvantage is that it is human resource incentive and require high technical skills to allow the implementation of in-house system that maintained internally. The advantage of the second approach is that the cloud service provider provides the required support to maintain the big-data system in addition to provide new services as they evolve. However, the disadvantage of this approach is its high cost and the obligation to renew the system cloud resources on fixed (e.g. annual) basis.

3.3 Analytics Tools

The last component of the integrated lab environment is the analytics component. Regarding this component the challenge mainly found in the integration of the big-data repository and the analytics tool. Two main analytic tools were tested within this lab environment, the first tool is R® (The R project for statistical computing) and the second tool is the SAS® software. R® seems to be easier to integrate with the big-data framework than SAS®. However, both tools allow the possibility to integrate with the big-data system after identifying the big-data repository as one of the data sources used by the analytics software.

The second challenge retain on the problem occur during data exportation. Some forms of the raw data collected are unrecognized file extensions by the analytics system (e.g. .pcap files). This problem requires the creation of file parsers capable of reading raw data and converting them into file formats readable by the analytics tools. Most analytic tools are capable of reading and writing comma separated values (.csv) file format. Proper file parsers written in C++ or Java language can be used to parse traffic files collected from the network using different network probes and saved in specific file formats into .csv files stored in the big data system and recognized by the analytic tools.

3.4 Systems Integration

One more challenge involves lab systems integration that occur during data collection is the distributed-centralized nature of the system. This contradictory nature between the centralized cluster implementation of big-data clusters versus the distributed nature of small-cells that spread over a large geographical area enforce multiple constrains for the system implementation. The first constrain involve the requirement of a scripting system that need to be used to transfer data from cell sites to the centralized big-data cluster. This can efficiently be performed using data storage cloud services as staging phase for the big-data cluster (i.e. data uploaded first to the cloud storage before being populated into the cluster).

Another constrain involve the management of the replication factor in the big-data cluster and its effect on the amount of network traffic traversing the network to serve the big data system. This problem is a technical problem related to the distributed-centralized nature of the system. Currently the default replication factor for big-data systems implementations is 3 with a centralized cluster location. The distributed nature of broadband wireless network traffic management system requires a geographically distributed cluster data nodes with zero replications. This contradiction between the current best practice for the big-data system implementation and the small-cells distributed geographical nature need to be studied for an applicable solution.

4 ALTERNATIVE SCOPE & DESIGN

In order to address the challenges for the primary design, an updated lab architecture is used to overcome the obstacles in implementing both the broadband wireless network and the big-data system. The following section introduce the new design and highlight the changes in the lab scope to deliver an operating lab environment.

4.1 Updated Lab Architecture

Figure 3 shows the updated lab architecture. The new architecture replaces the small-cells production system by two lab components. The first component is a simulator to the access network in addition to the second component which is an emulator the core network. After evaluation of the big-data implementation using open source distribution, the cloud services approach was adopted to provide the required infrastructure to the big-data system. Using the cloud storage services, the two systems are integrated by collecting the data from the simulator into the cloud storage. Finally, data analytics using R® was achievable by integrating R® services with the cloud services

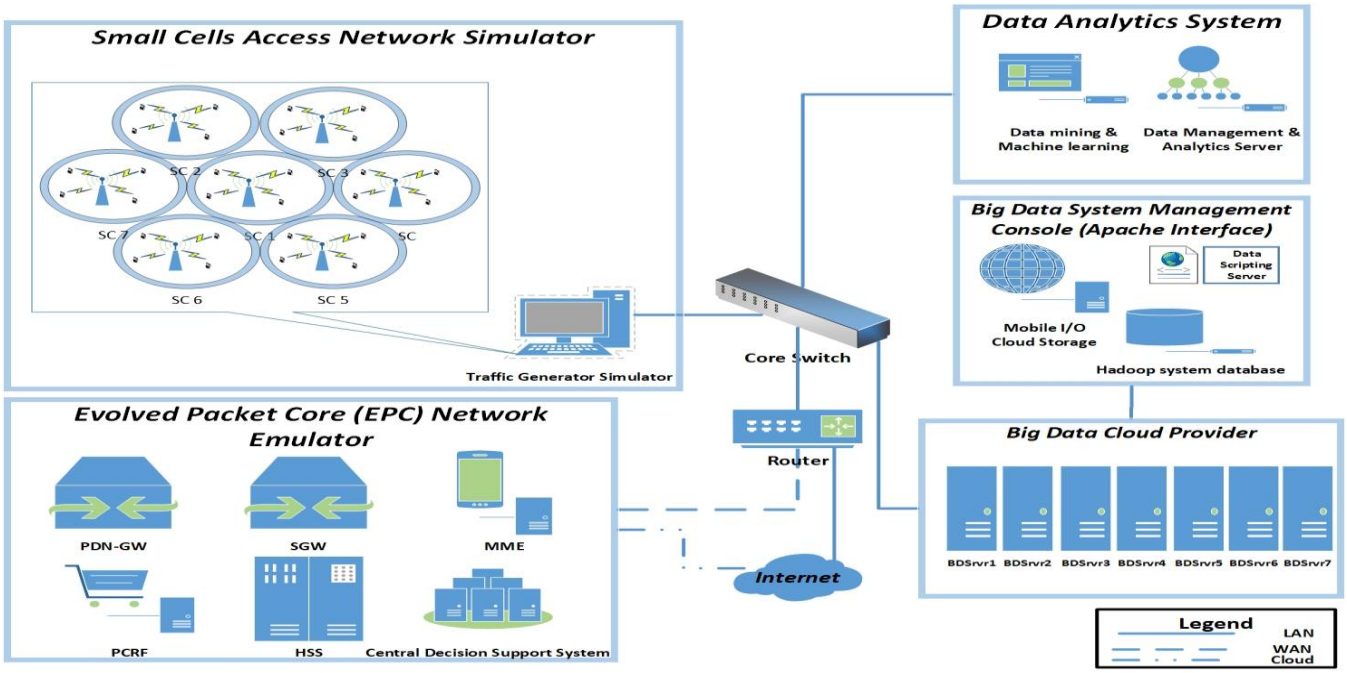

Figure 3: Lab Architecture (Updated Design)

4.2 Wireless Network

The new network design updated to address the challenges appears due to the proposed implementation of a production broadband wireless network environment in the original design. An OPNET® network simulator is used to replace the access network functions in the original design (Riverbed®, 2017). OPNET® is chosen because it includes two modules that can simulate the wireless access networks based on LTE technology. The two modules used by OPNET® to provide the access network functions are the Modeler Wireless Suite and the LTE Specialized Model. One of the disadvantages of OPNET® is its inability to model the core network. However, OPNET provide the capability to model the radio characteristics of the wireless network and to adjust a large set of wireless network parameters.

The other component used to complete the simulation of wireless network is the core network emulator. Polaris Networks® provides a suite of LTE network equipment emulators that allows end-to-end simulation of an LTE network in the lab (Polaris Networks®, 2017). The suite can emulate the core network functions and work with either Polaris Networks® access network

emulator component or receive the access network information from other external systems. The disadvantage of the access network emulator offered by Polaris Networks® is its inability to model the LTE radio characteristics of the wireless network.

Although each of the two network components, the access network and core networks, is fully functional separately and is considered necessary to provide students with basic concepts and hands-on about broadband wireless networks. The integration of the two components is overwhelming and not guaranteed. Up to the moment, the author couldn't identify a complete integrated broadband wireless network simulator or emulator that can provide an integrated fully functional system.

4.3 Big-Data System

The original design as shown in Figure 2 adopt the Hortonwork® Hadoop distribution. Hortonworks® provides an open source HDFS distribution supported by several tools for data extraction, transformation and loading (ETL). Hortonworks® offers a sandbox and free tutorials to introduce new implementers to the system. Hortonworks data platforms (HDP®) is an open source Apache Hadoop distribution based on centralized architecture (YARN). An installation guide is supported to help with the Apache Ambari installation, creating the data repository, and cluster preparation and deployment. Using VMware vSphere® the required virtual infrastructure for the data nodes and Ambari server is deployed on an ubuntu and Microsoft ® Windows Server respectively. The virtual machines are connected through an isolated subnet, private pool of IP's address to allow internet and remote access, to provide the required connectivity between the cluster node, broadband LTE access network simulator and core network emulator, and the analytics framework. The implementation process as stated earlier is a complicated process and require multiple skills (e.g. Linux, application server, networking and data repository management) to deploy the system. Also, continuous updates and upgrades are required to keep the system up, running and secured against external and internal cybersecurity threats. The main advantage of this solution is its cost effectiveness. However, the unprecedented load for system implementation, operation and improvement is considered impractical unless the lab is supported by qualified personal from the institution IT team.

The alternative to avoid the burden of the system management roles is to use cloud services supported by a cloud service provider. The new design replaces the open source distribution by Amazon Web Services (AWS) storage services, data transfer, and support. The service includes a running cluster on demand for approximately 8 hours/day or 20% of the month. The service covers the following AWS; Elastic Compute Cloud (EC2) to provide a fair compute capacity to operate the big-data cluster, Elastic Map Reduce (EMR) that run the managed Hadoop framework, Simple Storage Service (S3) to provide a cloud storage required to export the data from the broadband LTE access network simulator and core network emulator, data transfer out service for moving data across AWS services to and from EC2, and 24/7 business support (Amazon®, 2017). As shown in Figure 4 the EMR service, in the form of the cluster generation, can be initiated and destroyed according to availability requirements using step by step configuration or saved templates to avoid extra resources utilization during times where the service is not required (no student labs or research activities). The services have multi-user

capability and are controlled by an operator that can provide the basic accessibility and resources assignment tasks.

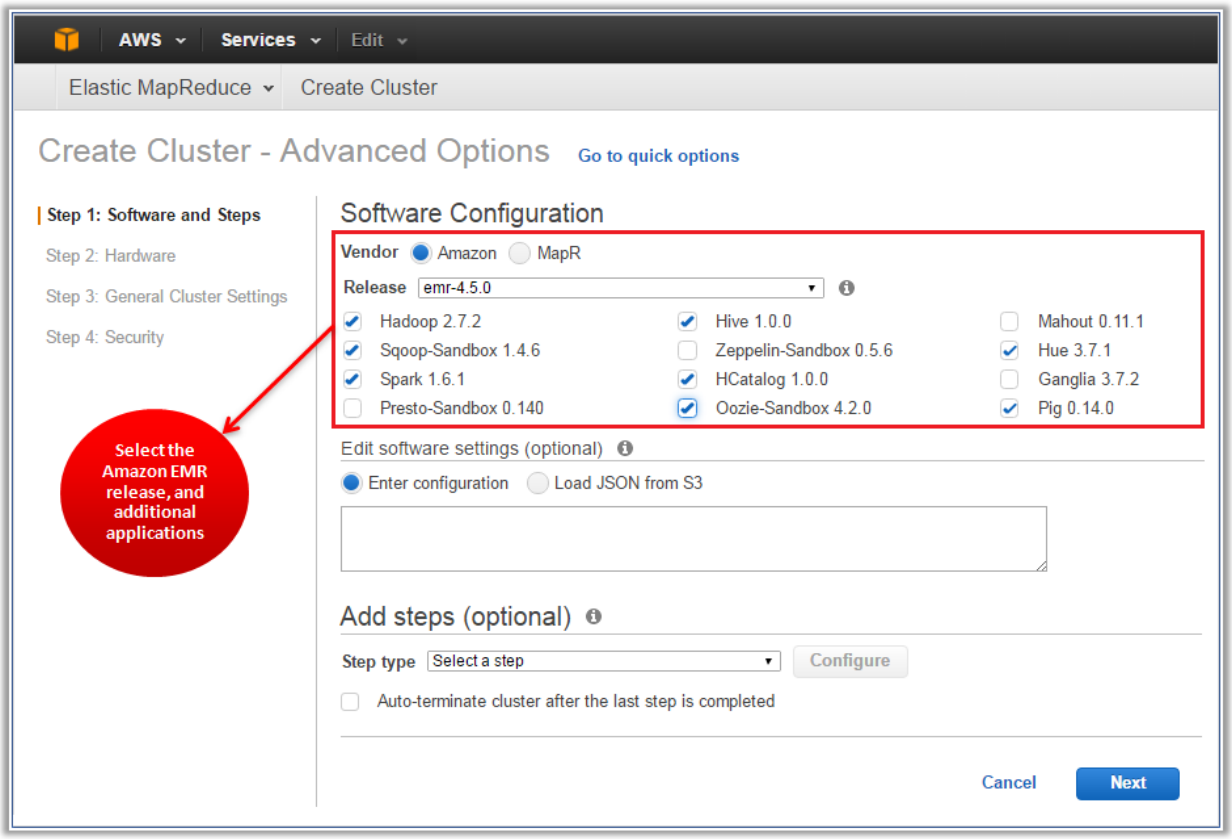

Figure 4: Creating the Amazon EMR Cluster (Oracle, 2017)

4.4 Analytics Tools & System Integration

In order to perform the required analytics, the AWS need to integrate with the analytic tools. RStudio is an open source statistics software and RStudio server allow access to R® via web browser. RStudio Integrated Development Environment (IDE) is used to integrate the RStudio server with the AWS. In addition to R® other analysis tools such as hive and hue can be used to perform preliminary analysis task like the one presented in Figure 5.

To achieve integration of the three main components of the lab environment NETLAB® is adopted to provide the virtual lab environment required to build a network instance for each student (NETLAB®, 2017). The student can schedule time slots to work on their network instances and this decrease the number of required concurrent licenses used to operate the OPNET® simulator and the Polaris core network emulator. The AWS service can be accessed through the web-based interface to load and manipulate data as required by students after creating an account for each student to access the EC2, EMR and S3 services. NETLAB® provides the advantage of creating, saving and deleting network instances and the virtual computing infra-structure required to operate the network. Before the end of the scheduling period, Students can save their network configurations. NETLAB® destroy the instances once the

students scheduling period is over. Students can recall saved configurations on their profiles during future configurations to continue working on their lab activities. NETLAB® provides a proper sharing environment that allows maximum utilization of lab resources all time with secured remote access capabilities.

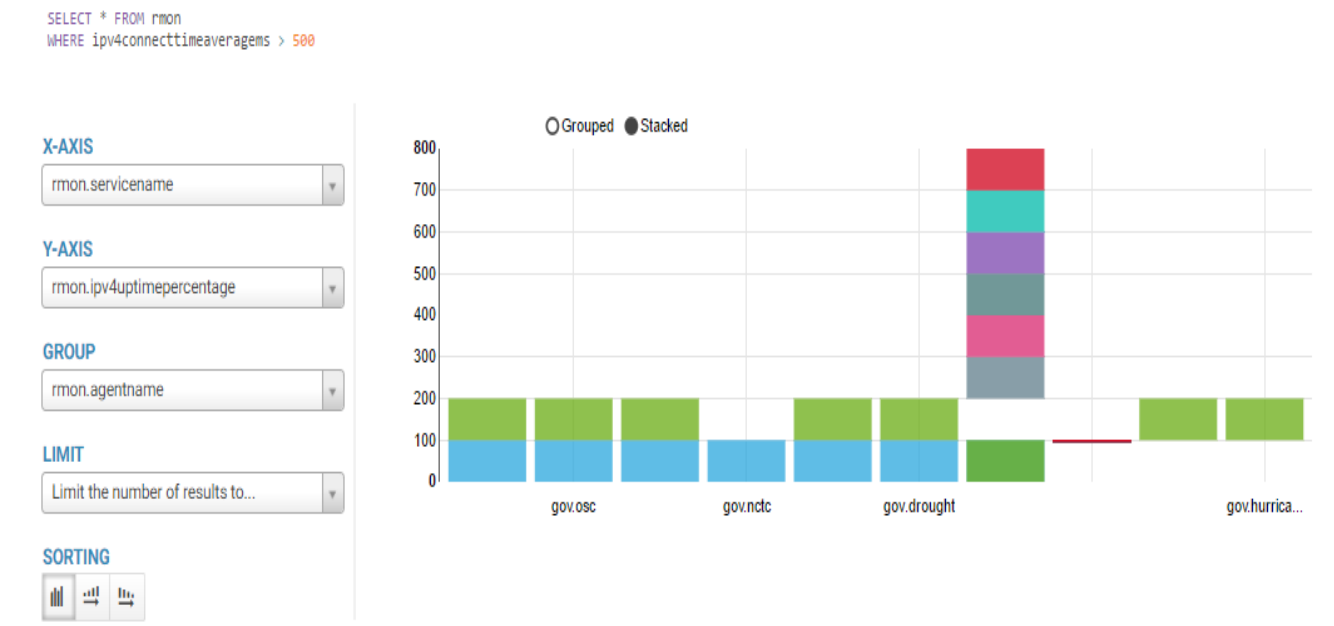

Figure 5: Graphical representation of data using Hive

5 STUDENTS EXPERIENCE

As of now and due to the multiple implementation challenges, the lab production environment is not completed. The lab testing environment for each individual system is finished and partial integration of systems are working according to the updated design. The integration of the whole lab into NETLAB® isn't tested. However graduate and undergraduate students were part of the design, implementation and testing activities of the lab components.

5.1 Benefits

Although the students population worked on the system cannot represent a sufficient population to evaluate the lab outcomes. Generally, students worked on the lab activities shows a high interest in learning networking and data sciences concepts. Undergraduate students are more convenient with learning the basics of wireless networking, data collection and network monitoring, with little interest in the analytics activities.

Graduate students shows more interest in analytics, data organization and manipulation. Researching the big-data applications in the networking domain seems a highly attractive topic. Different aspects related to the lab capabilities were studied by masters' students for developing their thesis. Group of the students were interested in designing and applying wireless network architectures to test traffic performance and enhance network resources utilization using self-

healing. Other group of the student were interested in using generated data by the network to perform analysis and provide insights about the traffic in different network locations.

According to the educational plans and in addition to research activities, the lab was prepared to teach two curriculums. The first curriculum teaches an introduction to wireless mobile systems. Examples of topics included are mobile radio propagation, cellular concepts, channel allocation, existing wireless systems and network protocols. The second curriculum teaches data sciences and big-data analytics. Examples of topics included are big-data overview, data analytics life cycle, and basic data analytics methods using R®.

5.2 Drawbacks

In addition to lab challenges, the testing phase shows that some of the students, especially undergraduates, struggles with understanding the system components. One of the reasons is the lack of basic knowledge about either wireless networks or big-data. Students that did not receive any prerequisite courses in either data sciences or networking find the system architecture overwhelming. The students require at least on introductory networking course and introductory course data management, in addition to basic programming skills.

For course offering, the lab system architecture need to be introduced to students. The students will be required to rely on network traffic as source of data in the data sciences course and rely on the HDFS as an analytic frame work to analyze network performance in case of networking course. This interdependency between the two courses introduce a difficulty to students that are not well introduced to the basics of the other course. Some introductory materials are required for students in such cases to provide a successful overall learning experience to the students.

6 CONCLUSION

Broad-band wireless networking and big-data are two important educational subjects. Industry are looking for talented students to fill enormous amount positions in these two areas. Currently there is a lack in qualified candidates to pursue the responsibilities required by these jobs. This study presents a lab that will introduce undergraduate and graduate students' to conceptual and advanced topics in both wireless networking and big-data areas.

The paper represent the original design of the lab and the challenges appear during the implementation process. The lab consist of three individual systems that present the components of the lab environment. The three components are integrated to provide the infrastructure required to deploy the lab environment. Typical difficulties escalated during systems implementation are discussed in details. Possible solutions to address the elevated problems and implement the original are proposed. However, in case of difficulty to adopt the original design as in the currents case study. An updated design is suggested to overcome the complications escalated on the original design. The new design and its implementation is discussed in details. Partial integration of the system is presented due to the lack of full integration results to date.

The benefits and drawbacks to students are highlighted to show the potential of the lab environment to student success. Finally a summary for course curriculums are emphasized to show the capabilities of the lab environment to introduce a successful learning experience.

References

Amazon®. (2017). *Amazon EMR Product Details*. Retrieved from https://aws.amazon.com/emr/details/

Cisco®. (2017). *Cisco Visual Networking Index: Forecast and Methodology, 2016–2021*. Retrieved from https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

HORTONW0RKS®. (2017). *HORTONW0RKS DATA PLATFORM (HDP)*. Retrieved from https://hortonworks.com/products/data-platforms/hdp/

NETLAB®. (2017). *NETLAB+: Where practice leads to success*. Retrieved from https://www.netdevgroup.com/products/

Oracle. (2017). Retrieved from http://www.ateam-oracle.com/preparing-amazon-elastic-mapreduce-emr-for-oracle-data-integrator-odi/

Polaris Networks®. (2017). *Complete Packet Core Emulation for testing Base Stations and UEs*. Retrieved from http://www.polarisnetworks.net/

Riverbed®. (2017). *OPNET Technologies*. Retrieved from https://www.riverbed.com/products/steelcentral/opnet.html

Tamer Omar, S. H. (2016). Broadband Wireless Networking in the Era of Big Data. *ASEE Annual Conference & Exposition.*

Weldon, M. K. (2015). *The Future X Network: A Bell Labs Perspective.* CRC Press.