HADOOP FRAMEWORK & CLOUD COMPUTING
A HANDS-ON WORKSHOP

- Type of submission: Proposal of a Workshop

- Author(s): Thuan L Nguyen
    - Affiliation(s): The University of North Texas
    - Complete Address:
        - 1155 Union Circle – Denton TX 76203
    - Telephone Number(s): 972 333 3234
    - Email Address: Thuan.Nguyen@utexas.edu
    - Name of the Dean(s) of the Affiliation School(s):
        - Victor Prybutok

- Track(Topic): Big Data

HADOOP FRAMEWORK AND CLOUD COMPUTING: A HANDS-ON WORKSHOP

ABSTRACT

The emergent big data technology can help established firms drastically transformed themselves, bring forth a whole new industry, and enable companies of any size to innovate, gain competitive advantage, and enhance business performance However, employing the emergent technology successfully is not easy. Participants of this hands-on workshop will learn how to set up a Hadoop system in the public cloud provided by Amazon Web Services and making it ready for big data analysis. The system consists of Hadoop Distributed File System, MapReduce, Apache Yarn, Apache Hive, and the database PostgreSQL. The experience of setting up a Hadoop system in the cloud offers an appreciation of a critical part of the process of teaching, learning, and analyzing big data.

Keywords: BIG DATA, BIG DATA ANALYTICS, BIG DATA AND CLOUD, HADOOP, APACHE HADOOP FRAMEWORK, HADOOP SYSTEMS IN CLOUD

INTRODUCTION

As a nascent field, big data or data science evolve with a warp speed, which has definitely caught the attention of scholars and practitioners in various industries (George & Lavie, 2016; McAfee & Brynjolfsson, 2012). The speed of evolving is intriguing enough to make academic researchers and business leaders wonder what emerging opportunities that big data and data science can offer (George & Lavie, 2016). Big data analytics has come out as a new important field of study for both researchers and practitioners, demonstrating the significant demand for solutions to business problems in a data-driven knowledge-based economy (Chen, Chiang, & Storey, 2012).

As a result, big data analytics along with business intelligence has emerged to be more and more essential to academic researchers, industrial practitioners, and business entrepreneurs (Chen, Chiang, & Storey, 2012). Along the path to success, the implementation of big data technology faces numerous obstacles (George & Lavie, 2016; Stourm & Ebbes, 2017). Many firms have invested or had a plan to invest in big data shortly (Gartner, 2015). However, about half of these organizations were not sure about what would be the results of their investment (Gartner, 2015).

A plausible explanation for the above problem is that big data technologies require skill sets that may be new to many information technology (IT) departments of firms in various business sectors, and professionals with these skills are hard to find (McAfee & Brynjolfsson, 2012). According to the McKinsey Global Institute (MGI) report, by 2018, the United States will have to tackle a serious shortage of professionals with critical skills of big data analytics as well as face a severe lack of managers with crucial knowledge and skills of making data-driven decisions. As a result, teaching and learning big data technology have become an urgent need.

BIG DATA AND CLOUD COMPUTING: A FORMIDABLE COMBINATION

According to Ferkoun (2014), the democratization of information technology has a significant impact on both the cloud computing and big data. Ferkoun also pointed out that cloud computing and big data have proved to be an ideal combination although both continue to be in constant evolution. Also, Trovati, Hill, Anjum, Zhu, & Liu (2016) identified technologies that facilitate the development of new business models employing big data approaches. The authors also examined the applications and implementations with which big data analytics is performed in cloud architectures. In the same line of thinking, In summary, big data analytics and the cloud computing have been proving to be a formidable combination that fosters faster advances in both fields. Cloud technology makes big data more powerful and more acceptable while big data helps turn cloud computing into an indispensable technology employed by any enterprise.

DESIGN OF THE WORKSHOP

The main objective of this workshop is to learn and practice the steps to set up a Hadoop system that includes Hadoop Distributed File System (HDFS), MapReduce, Apache Yarn, Apache Hive, and the database PostgreSQL. The skills can be applied to teaching, learning, and performing big data analytics including data visualization.

Although no programming knowledge is required for this hands-on workshop, the participants need to have the following fundamental computing skills:

- Read and follow written instructions
- Create, modify, and save files using a text editor
- Use, navigate and interact with files and folders using command lines

If a computer lab can be used, the lab should be available to be set up before the workshop starts. Otherwise, the workshop room should have a podium, a projector, a screen, and Internet connections via either wireline or wireless networks. If only the wireless connection is available, either a guest account with username and password or the SSID and the security key of the wireless network must be provided. Each participant needs to bring his/her laptop. The attendees also need to create an account of Amazon AWS. They may also take advantage of AWS Free Tier offered by Amazon AWS so that they can use the system free of charges while attending the workshop.

Objectives of the Workshop

In the hands-on workshop, the attendees will learn about the following:

1. How to set up an Amazon AWS instance based on an existing AWS community AMI
2. How to create an AWS Elastic Block Store (EBS) volume
3. How to attach an EBS volume to a running AWS instance
4. How to connect to an AWS instance using PuTTY
5. How to set up and configure the Hadoop system that consists of Hadoop Distributed File System (HDFS), MapReduce, Apache Hadoop Yarn, Apache Hive, and the database PostgreSQL, by running a script
6. How to create directories and files on an EXT4 (Linux) file system

7. How to create directories and files on HDFS
8. How to test the newly setup Hadoop system
9. How to start and safe-shutdown the Hadoop system

Flow of the Workshop

The participants will complete the following major steps to set up a Hadoop framework in the cloud in the workshop:

1. Getting started: Set up an Amazon Web Service (AWS) instance
2. Create and attach a new AWS Elastic Block Store (EBS) volume
3. Connect to AWS instance using PuTTY
4. Set up the Hadoop framework
5. Test the Hadoop Distributed File System (HDFS)
6. Obtain HDFS admin report
7. Safe shutdown the system
8. Start the system again and do an exercise on big data analytics

Exercise of Big Data Analytics Using the Framework in the Cloud

After successfully setting up the Hadoop framework in the cloud, the participants will work on an exercise of big data analytics with the following steps:

1. Upload a big data set from the local storage to the remote HDFS (Hadoop Distributed File System) of the Hadoop system
2. Configure Hive and Spark for the exercise
3. Perform text queries to have some insight into the data set
4. Create Hive schemas for the data in the data sets
5. Perform queries on the data set using Hive SQL-like commands
6. Perform queries on the data set using Spark-SQL commands
7. Build a supervised machine learning model (either regression or classification)
8. Run and evaluate the model

ACKNOWLEDGEMENTS

It is credited to the School of Information of the University of California at Berkeley that has created the AWS AMI “ami-be0d5fd4” and generously released it for public use as a community AMI. The author also acknowledges the school for all the documents related to setting up an instance based on its AMI.

REFERENCES

Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), pp. 1165 – 1188.

Ferkoun, M. (2014). *Cloud computing and big data: An ideal combination*. Retrieved from https://www.ibm.com/blogs/cloud-computing/2014/02/cloud-computing-and-big-data-an-ideal-combination/

Gartner (2015). Gartner survey shows more than 75 percent of companies are investing or planning to invest in big data in the next two years. Retrieved from http://www.gartner.com/newsroom/id/3130817.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90, pp. 61 – 67.

Stourm, L., and Ebbes, P. (2017). Analytics in the Era of Big Data: Opportunities and Challenges. Retrieved from http://www.hec.edu/Knowledge/Point-of-View/Analytics-in-the-Era-of-Big-Data-Opportunities-and-Challenges

Trovati, M., Hill, R., Anjum, A., Zhu, S. Y., & Liu, L. (2016). *Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications*. New York, NY: Springer.