Apache Big Data Stack

Tatyana Matthews[1], Scott McCaulay[2]
Indiana University[2]
Elizabeth City State University[1]

Abstract

As the amount of data generated around the world continues to accelerate by the second, the more the term **Big Data** finds its way into scientific conversation. Because of this tremendous surge, it has become imperative that such mass data use “computing power and space” for it to be processed, analyzed, and serve other purposes [1]. Hence, in order to meet head-on the enormous challenges rendered by Big Data, open source software from the Apache Foundation is evaluated as a “Big Data Stack” to support scientific computing. The approach to handling the complications surrounding Big Data involve installing and testing as many open-source software packages from the Apache Big Data Stack as possible on **FutureGrid** machines and later making those packages accessible utilizing **Chef**. The packages will be built into projects and from that point Chef will be used to transform the infrastructure of each project’s code, making it agile and accessible through a network of servers [2]. Essentially, this research will demonstrate how the Apache Big Data Stack can be used and applied to solve complex problems regarding Big Data.

Introduction

As each day passes users of the community around the world generate petabytes and even zettabytes of data. What is produced here is Big Data, data that cannot be seized, managed, organized, or processed through use of traditional tools in a fair amount of time [1]. The source of this data is essentially users of the community and with so much of it spawning at an enormous rate, it has now more than ever become imperative to “tame” today’s Big Data obstacles [3]. In order to combat those obstacles, it is necessary to use computing power and storage in order to capture, manage, and organize the immense data so that it can be processed, analyzed, and serve additional purposes [1]. Accordingly, supercomputers could solve this problem because they do provide storage and power; however, the amount it takes to build such a machine is costly. Hence, this expensive aspect has led to the search of other alternatives.

The Apache Big Data Stack, as shown in Figure 1, is a big solution and is composed of software that is capable of providing the storage and power that Big Data needs, not to mention that it is open-source . Accordingly, the Apache Software Foundation (ASF) plays a vital role with this data through use of the Apache Big Data Stack because it provides a “flexible and agile environment” for open-source projects to flourish [4]. Essentially, such projects flourish because the very community that produces Big Data is the same community that enables software packages within the Apache Big Data Stack to “mature fast and evolve quickly” because all code is exposed to the community; thus, making the packages robust and prosper [4].

In brief with this in mind, this research will reveal how Big Data complications facing the science world today can be solved through use and application of the Apache Big Data Stack.

Requirements

In order to prepare for the fall 2014 course I590-Investigating Big Data Open Source Software and Projects, this research will require exploration of a series of open-source software packages within the Apache Big Data Stack. As these packages are explored a variety of deliverables will be produced, including demonstrations, tutorials, hands-on examples, Chef recipes, and an abstract understanding of each package explored (this is shown in Figure 2). Focus will also be on troubleshooting the installation process utilizing FutureGrid resources. Essentially, fulfilling these requirements will prepare materials for course I590 which will encompass studying the Data Stack in order to further study Big Data [5].

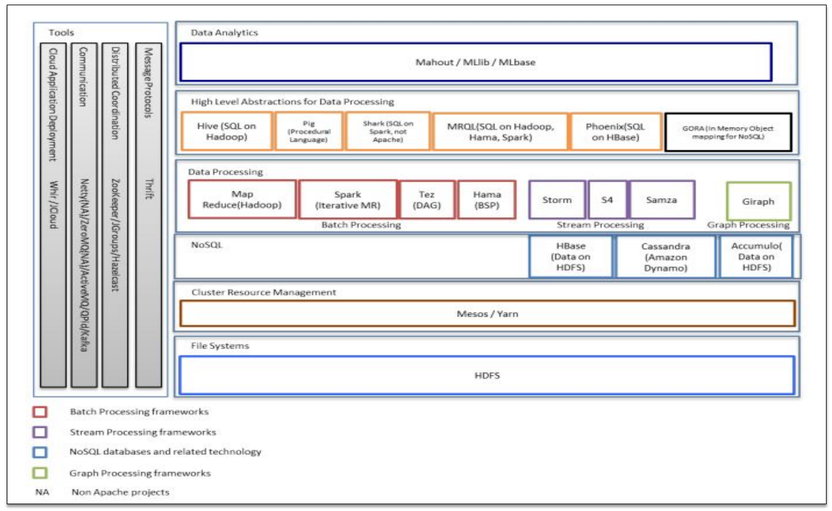

Fig 1: Apache Big Data Stack

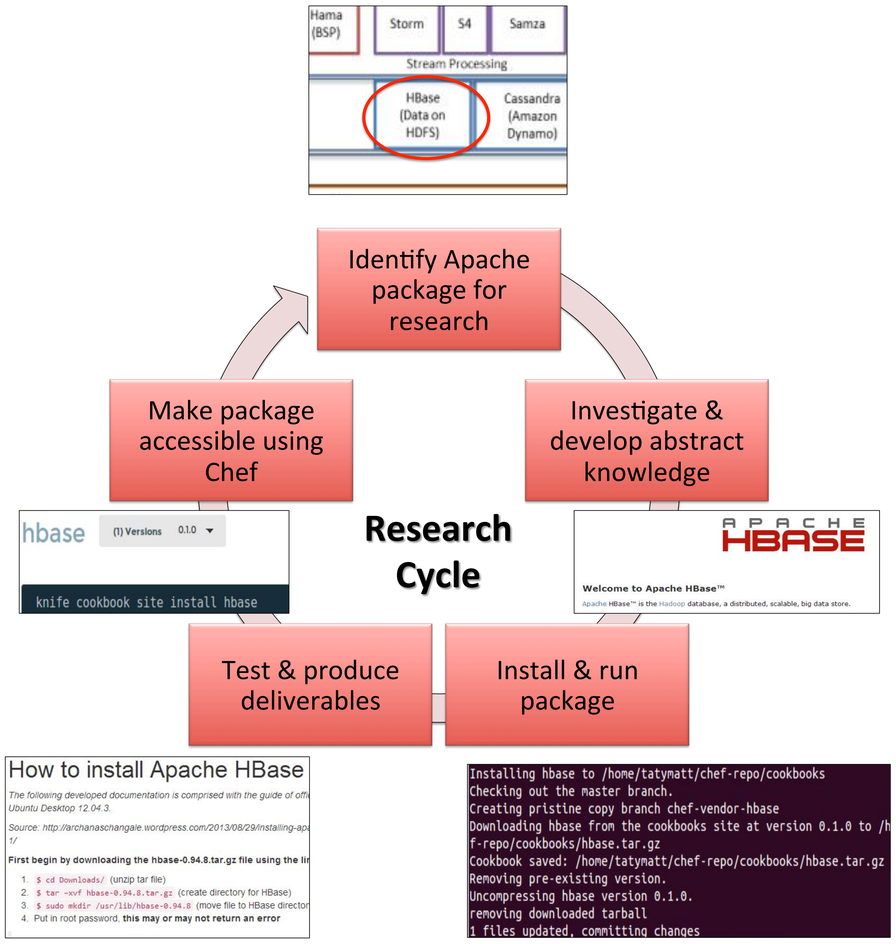

Fig 2: Research Cycle

Conclusion

Research was successful being that the Apache Big Data Stack was explored by studying its open-source software packages, developing abstract knowledge, creating tutorials, documentation, and demonstrations in preparation for course I590. Chef Enterprise cookbooks were also found, which when applied will make the packages used accessible to users of the community. In addition, research demonstrated how the Apache Big Data Stack could be used and applied to capture, store, manage, and organize Big Data so that it could be utilized for analysis, processing, and serve additional purposes [1].

Further Work

- Run HBase and other Apache software, in addition, investigate more extensively the packages within the Apache Big Data Stack
- Generate documentation, descriptions, tests, tutorials, Chef recipes, and demonstrations for each package

Once accomplished, students engaging in Dr. Geoffrey Fox’s fall course, I590-Big Data Open Source Software & Projects, will be able to directly interact with the deliverables produced and investigate the Apache Big Data Stack for themselves.

```
24  next unless app['hbase_master_role'] ||
25  (node.run_list.roles & (app['hbase_mast
26    Chef::Log.debug("hbase_role: #{hbase_
27    top = app['hbase']['top']
28    hadoop = app[:hadoop]
29    hadoop_dir = hadoop[:home]
30    hadoop_user_home = hadoop[:user_home]
31    hbase = app[:hbase]
32    hbase_dir = hbase[:home]
33    zookeeper = app[:zookeeper]
```

Fig 3: Apache HBase Chef recipe

Acknowledgments

I would like to extend a hand of appreciation to Scott McCaulay for his guidance, contributions, and help with completing this research. Also, I would like to thank Dr. Geoffrey Fox for introducing this research and laying down the groundwork for profound investigation. Furthermore, I would like to thank Dr. Lamara Warren for providing me with this opportunity to engage in a stimulating research experience as well as develop as a researcher.

The FutureGrid project is funded by the National Science Foundation (NSF) and is led by Indiana University with University of Chicago, University of Florida, San Diego Supercomputing Center, Texas Advanced Computing Center, University of Virginia, University of Tennessee, University of Southern California, Dresden, Purdue University, and Grid 5000 as partner sites. This material is based upon work supported in part by the National Science Foundation under Grant No. 0910812.

References

[1] S. Kamburugamuve, “Survey of Apache Big Data Stack,” Ph.D. Qualifying Exam, Dept. Inf. Comput., Indiana Univ., Bloomington, IN, 2013.
[2] (2014). *Get Chef* [Online]. Available: http://www.getchef.com/chef/
[3] V. Borkar, M..J. Carey, C. Li, “Inside Big Data management: ogres, onions, or parfaits?” in *Proceedings of the 15th International Conference on Extending Database Technology*, New York, NY, 2012, pp. 3-14.
[4] G. Fox. (2014, May 30). *Multi-faceted Classification of Big Data Uses and Proposed Architecture Integrating High Performance Computing and the Apache Stack* [Online]. Available: http://www.slideshare.net/Foxsden/multifaceted-classification-of-big-data-uses-and-proposed-architecture-integrating-high-performance-computing-and-the-apache-stack#btnNext
[5] G. Fox, “Syllabus for I590 Big Data Open Source Software and Projects Fall 2014,” unpublished.

Primary Contact

Dr. Geoffrey Fox, Indiana University, gcf@indiana.edu