Big data Visualization with Apache Zeppelin

NAVEENKUMAR RAMARAJU[1,*] AND VEERA MARNI[1,*]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors: naveenkumar2703@gmail.com, narayana1043@gmail.com

project-008, March 27, 2017

Apache Zeppelin is an open source notebook for data analytics and visualization. In this project we deploy Apache Zeppelin in cluster and visualize data stored in Spark across cluster using Apache Zeppelin interpreter that employs Python and Scala in same notebook.

Keywords: Zeppelin, Apache, Big data, Visualization

https://github.com/cloudmesh/sp17-i524/blob/master/project/S17-IR-P008/report/report.pdf

1. INTRODUCTION

Apache Zeppelin[1] is an interactive notebook that is used for data ingestion, discovery, analytics, visualization and collaboration. It has built in Spark integration and supports multiple language backends like Python, Hadoop HDFS, R etc. Multiple languages can be used within same Zeppelin script and share data between them. In this project we aim to deploy Zeppelin 0.7 along with in built Spark and backend languages R and Python across cluster using Ansible. Then install additional visualization packages provided by Apache Zeppelin Helium APIs.

We also aim to load a large data set into Spark across cluster and perform data analytics and visualization in cloud using Zeppelin. We have not decided about data set at this point.

2. EXECUTION PLAN

Deploy Spark, Zeppelin, Helium, R and Python using ansible by March 31.

Find a data set by March 31.

Data set found.

Tamilnadu Electricity Board Hourly Readings Data Set

This data can be effectively produced the result to fewer parameter of the Load profile can be reduced in the Database

Data Set Characteristics: Multivariate
Number of Instances: 45781
Attribute Characteristics: Real
Number of Attributes: 5
Associated Tasks: Classification, Regression, Clustering
Source:K.Kalyani ,kkalyanims@gmail.com,T.U.K Arts College,Karanthai,Thanjavur.

Relevant Papers: Efficient Electricity Utilization By IHBMO
Attribute Information: forkva,forkw,type,sector,service.

Benchmark the deployment times of individual and all items by April 7.

Load the data distributed across machines using Spark by April 7.

Perform Visualization on loaded data with Zeppelin using Spark, Scala and Python in same environment and validate the collaboration across cluster by April 14.

See, if configuration of Apache clusters inside Zeppelin could be done employing Ansible at deployment time by April 17.

Finish report and submit project on April 21.

3. DEPLOYMENT

TBD

4. BENCHMARKS

TBD

5. VISUALIZATION WITH ZEPPELIN

TBD

6. SUPPLEMENTAL MATERIAL

TBD

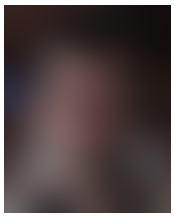
ACKNOWLEDGEMENTS

TBD

REFERENCES

[1] Apache Zeppelin, "Zeppelin 0.7.0 Documentation," Web Page, Apache Software Foundation, Mar. 2017. [Online]. Available: https://zeppelin.apache.org/docs/0.7.0/

AUTHOR BIOGRAPHIES



Naveenkumar Ramaraju yet to update his Bio

Veera Marni yet to update his bio

A. WORK BREAKDOWN

TBD

Naveenkumar Ramaraju TBD

Veera Marni TBD

B. REPORT CHECKLIST

- □ Have you written the report in word or LaTeX in the specified format?
- □ Have you included the report in github/lab?
- □ Have you specified the names and e-mails of all team members in your report. E.g. the username in Canvas?
- □ Have you included the HID of all team members?
- □ Does the report have the project number added to it?
- □ Have you included all images in native and PDF format in gitlab in the images folder?
- □ Have you added the bibliography file in bibtex format?
- □ Have you submitted an additional page that describes who did what in the project or report?
- □ Have you spellchecked the paper?
- □ Have you made sure you do not plagiarize?
- □ Have you made sure that the important directories are all lower case and have no underscore or space in it?
- □ Have you made sure that all authors have a README.rst in their HID github/lab repository?
- □ Have you made sure that there is a README.rst in the project directory and that it is properly filled out?
- □ Have you put a work breakdown in the document if you worked in a group?

C. POSSIBLE TECHNOLOGY PAPER OUTLINE

The next sections are just some suggestions, your may want to add sections and subsections as you see fit. Images and references do not count towards the 2 page length. Please use the `\section`, `\subsection`, and `\subsubsection` commands in your paper. do not introduce hardcoded numbers. Use the `\ref` and `\label` commands to refer to the sections.

Abstract Put in the abstract a summary what this paper is about

1. Introduction Introduce the technology and provide general useful information.

2. Architecture If applicable include a description about architectural details. This may include a figure. Make sure that if you copy a figure you put the **[?]** in the caption also. Otherwise it is plagiarism.

2.1. API comment on the API which could include language bindings

2.2. Shell Access If applicable comment on how the tool can be used from the command line

2.3.Graphical Interface If applicable comment on if the technology has a GUI

3. Licensing Often tools may have different versions, some fre, some for pay. Comment on this. For example while a tool may offer a comercial version this version may be too costly for others. Identify especially the difference between features for free vs commercial tools.

Sometimes you may need to introduce this also in the introduction as there may be a big difference and without the knowledge you do not provide the user an adequate introduction.

4. Ecosystem Some technologies have a large ecosystem developed around them with extensions plugins and other useful tools. Identify if they exists and comment on what they can achieve

provide potentially a mindmap or a figure illustrating how the technology fits in with other technologies if applicable.

4. Use Cases

4.1. Use Cases for Big Data Locate and describe major usecases that demonstrate the technology while focussing on big data related use cases. Make sure you do proper references with the [**?**] command. Do not put URLs in the text.

4.2. Other Use Cases Some technologies may not just be used for big data, find other makor use cases from other areas if applicable. Make sure you do proper references with the [**?**] command. Do not put URLs in the text.

5. Educational material Put information here how someone would find out more about the technology. Use important material and do not list hundrets of web pages, be selective.

6. Conclusion Put in some conclusion based on what you have researched

Acknowledgement Put in the information for this class and who may sponsor you. Examples will be given later