Project Abstracts for Spring 2017
Cloud Computing (CSC 547)

Projects 1 through 4 (rev3.0 – 02/09/17)
Configure a cloud for Big Data Analytics

Instructor: John Streck (jpstreck@ncsu.edu)
Advising Mentors:
Dr. Michael Kowolenko (mdkowole@ncsu.edu)
Professor James Hall (jlhall4@ncsu.edu)
Shireesh Bhat (sbhat@ncsu.edu)

Project overview: The Laboratory for Analytical Sciences, High Performance Cloud Research Group is focused on the intersection of "big data" infrastructure, and the development of novel, computationally intensive algorithms to solve specific questions & problems using a cloud platform. The analytical cloud will have a mixed set of services that will offer both the attributes of a Platform as a Service (PaaS) cloud and a Software as a Service (SaaS) cloud offering.

Goal: This project is focused on developing an <u>open source</u> high performance Big Data (Analytics) Cloud infrastructure such that a statistician & data science person, from a student to a seasoned researcher, can use this platform as an IDE (integrated development environment) of sorts to develop and test new algorithms for the use in advanced analytical studies. The cloud should be architected such that the user can be proficient in a minimum learning cycle. The cloud should also be architected such that the design is portable to any number of cloud providers and within a chosen cloud is easily scalable to fit the data and problem presented.

Use Cases: The use cases for applying the cloud infrastructure could include Insurance services, Financial Services, Character recognition, etc. One of the potential ways for validating the cloud infrastructure model integrated with machine learning libraries (Project 2) would be to construct a model based on the first half of the data set and comparing the second half of the data set with the predicted values from the model and comparing the deviation from the actual values for the predicted values and using machine learning to refine the regression model. The use case and the dataset should be negotiated with the instructor and the mentor.

Layer Model: Figure 1 helps in visualizing the building blocks which make up the project. All the building blocks may or may not be needed for constructing the cloud infrastructure but the figure gives an idea on where the blocks fit in, in the Open Stack layer model of services.
(reference: https://dague.net/2014/08/26/openstack-as-layers/).

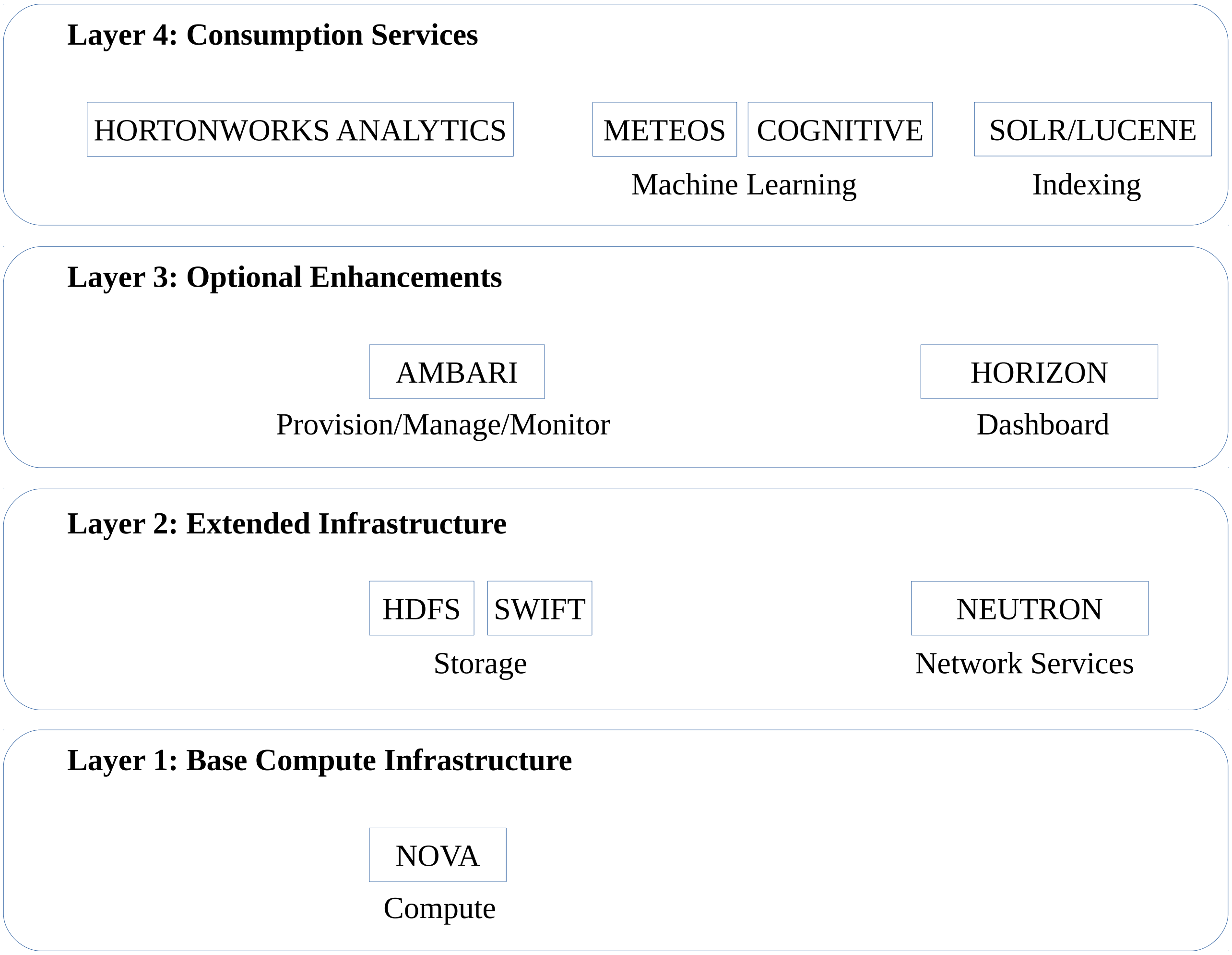

Figure 1

Project 1 (P1)

Initial direction is to construct a analytics cloud using the Hortonworks analysis package (will be provided by the instructor) integrated with Solr/Lucene. (The data set can be selected from the existing sources for big data sets). The cloud should allow relatively easy user provisioning of new nodes.

P3 will be done on RHEL 7 which has the Open Stack Platform 10 inbuilt. The initial design can start on standard VCL (with the correct hardware chosen) for both groups and later can be demonstrated on the AWS cloud system.

Functional Requirements:

1) The cloud platform should utilize the compute infrastructure of at least two compute nodes.
2) The cloud platform will be injected with a database for analyzing the data.
3) The cloud platform will take as configuration parameters: the number of records/documents to analyze, the time to index, additional compute nodes to add.
4) The cloud platform should allow for provisioning of new nodes when the analysis is in progress and demonstrate the effective utilization of all the existing and new compute nodes.

5) Auto scaling by adding new compute nodes is a mandatory requirement. Auto scaling by removing existing compute nodes would require more effort and will be an optional requirement.

Non-Functional Requirements:
1) Analyze the impact on the type (sql vs nosql) database selected.
2) Analyze the performance impact on the number of compute nodes used.
3) How does the threshold time, which is the configuration parameter for getting the output alter the results.
4) What are the interesting pieces of information which were obtained or deduced using the analysis tool.

Project 2 (P2)

Initial direction is to construct a analytics cloud using the Hortonworks analysis package (will be provided by the instructor) integrated with Solr/Lucene. (The data set can be selected from the existing sources for big data sets). The cloud should allow relatively easy user provisioning of new nodes.
P2 will be done on RHEL 7 which has the Open Stack Platform 10 inbuilt. The initial design can start on standard VCL (with the correct hardware chosen) for both groups and later can be demonstrated on the AWS cloud system.

Functional Requirements:
1) The cloud platform should utilize the compute infrastructure of at least two compute nodes.
2) The cloud platform will be injected with a database for analyzing the data.
3) The cloud platform will take as configuration parameters: the number of records/documents to analyze, the time to index.
4) The cloud platform should integrate with machine learning modules or libraries (python).
5) Validate the machine learning model.

Non-Functional Requirements:
1) Analyze the impact on the type (sql vs nosql) database selected.
2) Analyze the performance impact on the number of compute nodes used.
3) How does the threshold time, which is the configuration parameter for getting the output alter the results.
4) What are the interesting pieces of information which were obtained or deduced using the analysis tool.

Project 3 (P3)

Initial direction is to construct a analytics cloud using the Hortonworks analysis package (will be provided by the instructor) integrated with Solr/Lucene. (The data set can be selected from the existing sources for big data sets). The cloud should allow relatively easy user provisioning of new nodes.
P3 will be done on RHEL 7 which has the Open Stack Platform 10 inbuilt. The initial design can start on standard VCL (with the correct hardware chosen) for both groups and later can be demonstrated on the AWS cloud system.

Functional Requirements:
1) The hadoop platform (analytics package) should be made to run over a virtual machine running CentOS. As a next step, the hadoop platform can be made to run on AWS while the rest of the services continue to be run locally.
2) The cloud platform should utilize the compute infrastructure of at least two compute nodes.
3) The cloud platform will be injected with a database for analyzing the data.
4) The cloud platform will take as configuration parameters: the number of records/documents to analyze, the time to index.

Non-Functional Requirements:
1) Analyze the impact on the type (sql vs nosql) database selected.
2) Analyze the performance impact on the number of compute nodes used.
3) How does the threshold time, which is the configuration parameter for getting the output alter the results.
4) What are the interesting pieces of information which were obtained or deduced using the analysis tool.

Project 4 (P4)

Initial direction is to construct a analytics cloud using the Hortonworks analysis package (will be provided by the instructor) integrated with Solr/Lucene. (The data set can be selected from the existing sources for big data sets). The cloud should allow relatively easy user provisioning of new nodes.
P4 will be done on RHEL 7 which has the Open Stack Platform 10 inbuilt. The initial design can start on standard VCL (with the correct hardware chosen) for both groups and later can be demonstrated on the AWS cloud system.

Functional Requirements:
1) The hadoop platform (analytics package) should be made to run over a virtual machine running Ubuntu. As a next step, the hadoop platform can be made to run on AWS while the rest of the services continue to be run locally.
2) The cloud platform should utilize the compute infrastructure of at least two compute nodes.
3) The cloud platform will be injected with a database for analyzing the data.
4) The cloud platform will take as configuration parameters: the number of records/documents to analyze and the time to index.

Non-Functional Requirements:
1) Analyze the impact on the type (sql vs nosql) database selected.
2) Analyze the performance impact on the number of compute nodes used.
3) How does the threshold time, which is the configuration parameter for getting the output alter the results.
4) What are the interesting pieces of information which were obtained or deduced using the analysis tool.