**Hadoop - Zeppelin Proof of Concept**

**Project Proposal**

**Stephanie Akpakoun**

**Nick Warnke**

**Motivation**

The goal of this project is to provide a proof of concept for Zeppelin, which is a tool used on top of Hadoop to visualize data using web-based notebooks. Some of the features Zeppelin offers are data ingestion, data discovery, data analytics, data visualization, and collaboration. At Union Pacific, it will be primarily used to examine raw data and make it consumable for non-IT users. The type of information that we will be analyzing is train events and movements. Union Pacific currently uses 64 core servers to store this data. Millions of rows of train event data are stored every day in Teradata servers and pulled into Oracle tables for use. A group within Union Pacific built the system currently in place on a mainframe that extracts data from the Teradata tables and inserts them into work tables via XCOM jobs. This data is then being pulled, analyzed and displayed in various applications built by different groups based on their specific needs. The current system has two main drawbacks.

First, many groups have built and maintained several applications that provide very similar functionalities. This leads to users having to access many different systems to get answers to different questions. These applications consume financial and human resources and can all be consolidated into one system that handles all user requests.

Second, the large volume of data has grown beyond the capacity of Oracle. The options were to either invest in bigger and more expensive servers or migrate the data to Hadoop which is an affordable and practical solution. Hadoop specializes in Big Data storage and is optimized to process large data sets. Therefore, the choice was made to move to Hadoop which provides a way to distribute computation across thousands of machines instead of using more powerful machines in a singleton manner.

The goal of the new system is to consolidate all of the current applications that have been built on top of the original system into one dashboard view that will be customizable by the user and not require UI development. The users of this system are departments such as Finance, Marketing and Sales, as well anyone with an interest in car events, costs, and waybills. It will help provide business insights for executives to help them make critical decisions. In this proof of concept, we will be analyzing Operational Statistics (OPStats) application data. This will lead us to offer a recommendation on whether or not Zeppelin is a viable tool to use with Hadoop as a self-service data analysis tool. We will be working with Chloe Ng Ping at Union Pacific.

**Functionalities**

The proof of concept process will lead us to put in place a system to manage the data.

* Create a data environment that will analyze and visualize the complex and large data set to create meaningful relations.
* Replicate currently generated reports.
* Analyze the data and compare it with the frequently asked questions in order to determine the data relations that will answer these questions.
* Design a customer self-service environment.

The system should be prepared to handle a wide range of scenarios. It should not require the user to have an extensive knowledge of data analytics. The reports should provide information in a way that accurately answers user queries with minimal assistance from IT.

It will remove the need for a developer to make UI changes.

* + Design tables from the pool of Hadoop data that will be used in Zeppelin to answer specific user questions.
  + Create a data subset (pre-processed data set).
  + Create queries to generate Zeppelin graphical reports.
  + Export the reports to spreadsheets.
* Replicate the data from Teradata to Hadoop.

Some of the data for OpStats has already been migrated into control tables in Hadoop, but a lot of it still needs to be imported into Hadoop using Sqoop..

**Implementation Strategy**

The project will be primarily using Zeppelin, which is a web-based notebook for interactive data analysis. Each notebook within Zeppelin allows the user to create a report. It is our goal to create several notebooks within Zeppelin to accomplish data analysis for a project known as Operational Statistics (OpStats). Primarily, the language that we will be writing the code in will be Hive Query Language. The Hive QL code will be written in the notebooks which will also serve as the VCS for the project.

**Technical Challenges**

Our technical challenges come with the complexity of the data to be analyzed. We will need to sort through large tables of data and sift out junk data. This will entail correctly analyzing and representing the data from a large data set to make it useful to users. Since we have no experience in Hadoop or big data, this could be challenging.

**Team**

Our team is comprised of two Union Pacific interns, both of which have basic understanding of Oracle SQL. We are both new to Hadoop and big data analysis.

**Stephanie Akpakoun:** Java, Spring MVC, AngularJS, Oracle SQL, Javascript, HTML, Java Servlet Pages, Hibernate, JDBC.

**Nick Warnke:** Spring MVC, AngularJS, Java, Oracle SQL, Web development using HTML, Java Servlet Pages, and JavaScript, Hibernate, JDBC.