Santa Clara University

Department of Computer Engineering

Advanced Database Systems (COEN 380)

Group Project-3

Project Overview:

The goal of this project is two folds: (1) Become familiar with executing SQL on Hadoop using Hive[1]. (2) Good understanding and compare the query optimization material in the class against the query plan generated by a mature database such as Oracle.

Overview:

In this project, start with identifying the set of SQL statements that you will be using throughout the project. These SQL statements should include: Selection, Predicates, Aggregates Natural JOIN, Theta JOIN, and Multiple flavors of Sub-queries. Groups should clear early the SQL statements with the professor.

Once the SQL statements are finalized, proceed with two experiments:

1. Execute the above SQL statements, whenever possible over Hadoop/Hive. Identify any SQL statement that is not supported on Hive; remember Hive is a subset of SQL 1992 only.

2. Use the Oracle Explain Plan tool from Oracle to view the Oracle generated query plan to each one of your SQL statements and try to map the query plan in what was covered in the query optimization in the class.

Project Demo:

Each group will give demo to the professor on the date shown on the syllabus. Please submit PPT before the demo including:

1. **Hadoop/Hive Section**: please write all agreed on SQL statements and identify which ones you could not run on Hive?

1. **Hive:** https://cwiki.apache.org/confluence/display/Hive/Tutorial

2. **Oracle Query Optimizer Section**: please write all agreed on SQL statements and on each page include the SQL statement + the query plan as shown by the Oracle Explain Plan tool.