CS 6V81: Introduction to Cloud Data Security

Project Description - Query Evaluation using Hadoop

Due: December 2nd, 2011

The main aim of this project is to extend the basic MapReduce discussions for relational algebra operators that were given in class into a simple, demonstrable real-world application that is capable of evaluating simple queries using a series of MapReduce tasks. Additional goals of this project are to give the students hands-on experience with setting up a Hadoop cluster as well as programming in the MapReduce environment.

Project Team Size – Three- four people

Project Components

The main components of the project are as follows:

1. Setup a Hadoop cluster in one of the three pre-defined modes: Standalone, Pseudo-Distributed or Fully-Distributed. Note that you can choose any mode of operation provided that you are able to demonstrate your application in the selected environment.
2. Implement MapReduce programs for the three basic relational algebra operators: Selection (Equality conditions only), Projection and Join (Two-way equality joins only).
3. Build a simple query execution engine that takes as input simple Selection-Projection-Join (SPJ) queries and executes these queries as a sequence of MapReduce tasks. For example, given a query,

   Join(Selection(Condition, R1), R2, R1.A = R2.B),

   a possible execution plan may be to execute the Selection operator as the first MapReduce task followed by the Join operator as the second MapReduce task. Therefore, your execution engine must be able to construct a sequence of MapReduce tasks that are required to answer a given query. Note that you may use the form specified above as the query language that is used to submit queries to your execution engine.
4. Build a simple user interface for users to be able to submit their queries and view the results of query execution. There is no need to build a sophisticated graphical user interface, a simple command line tool is also sufficient.

Project Deliverables

You will need to submit the following documents as a part of the final project submission:

1. Project Report – A short project report (no more than three pages, one inch margins, 11pt fonts) that details the group's information, problem definition, proposed solutions for the implementation of the relational algebra operators as well as the query evaluation engine.

2. Project Code – The entire code that you have written to implement the components specified above. Please document the code using standard Javadoc comments that are available with any development environment.
3. Project Presentation – You will also have a short project (10 minutes per group) presentation in class on 12/02/2011. Your slides should include an overview of the problem and your propose implementation strategy to solve the problem.

Resources

The following resources are a good place to start learning about Hadoop:

- Documentation on setting up a Hadoop cluster

  Single-Node Setup: http://hadoop.apache.org/common/docs/current/single_node_setup.html, http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

  Cluster Setup: http://hadoop.apache.org/common/docs/current/cluster_setup.html, http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/

- Documentation on programming in MapReduce

  http://hadoop.apache.org/common/docs/current/mapred_tutorial.html
  http://developer.yahoo.com/hadoop/tutorial/module4.html

  Note that the code given in both the links is based on older versions of Hadoop. The signatures for both, the map and reduce functions, as well as some of the objects (e.g. an object called Context is used instead of OutputCollector) have changed in the newer versions of Hadoop.

- A comprehensive description of Relational Algebra operators

  http://www.pathfinder-xquery.org/files/teaching/ss09/db1/db1-03.pdf

  Note that you can use the first part of the presentation to understand the basics of various relational algebra operators as well as how they are combined together using.