15-415 Databases
Project #3 (Cassandra)

Assigned: Tuesday, November 15th, 2011
Due: Tuesday, December 6th, 2011

Overview

This assignment asks you to revise your project #1 to make use of *Apache Cassandra* or *HBase.* In particular, the idea is to explore the difference in the way big data style distributed databases represent data in sparse maps-of-maps, as compared to the table-like relations of traditional relational databases. The focus of the assignment is not on the actual distributed storage or the details of any particular distributed database API.

The short version of the assignment is that you'll replace the SQL tying your Lab #1 application to the RDBMS with calls to Cassandra's or HBase's API. This will require rethinking your schemas and data organization. You should be able to support the same feature space – but some operations will become significantly more expensive, while others will become cheaper. In your report, you should document how the paradigm shift affects your projects performance in computation and space.

Teams

You may work with any subset of your Project #1 team, including by yourself. Other formulations are possible – please see a member of the course staff.

APIs

You may use any available API for Cassandra or HBase.

HBase Cluster Size/Cassandra Ring Size

The focus of this assignment is on the model, not the distributed database, itself. Please start by using a single node, either on your laptop or a cluster system. After Thanksgiving, we'll make a large cluster available so you can see it work in the wild.

The Report

The report should contain a few sections (a) an *Application Description*, which reminds us about the purpose of your project #1, (b) an *Explanation of Changes to the Logical Design*, which

describes how *and why* things have changed from the perspective of the data's organization, (c) an *Explanations of Limitations or Enhanced Functionality*, which documents any user visible changes since project #1, (c) *Dependencies,* that describe the precise versions of the operating system and platform, Cassandra or Hadoop/HBase, high level client libraries, and any other software we need to build your application, (d) *Instructions for Building and Testing*, that let us know how to build and test your project.

Deliverables

Please submit, into the usual AFS space, (a) your report, (b) your source code, and (c) a sample of your data sufficient for testing.