

## 15-415 Databases Project #4 (Cassandra)

**Assigned:** Thursday, November 16, 2010  
**Due:** Thursday, December 2, 2010

### Overview

This assignment asks you to revise your project #1 to make use of *Apache Cassandra*. In particular, the idea is to explore the difference in the way *Google BigTable*-like 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 current application to the RDBMS with calls to Cassandra's API, *Thrift*, or one of the higher-level APIs. You'll likely find that some of the features of your RDBMS application, as it was previously designed, are not a good fit for Cassandra's model. Where this is the case, you should make straight-forward changes to adapt your solution. But, if the changes are deeply rooted, it is not necessary to implement these features. Instead, in your report, you should document and explain your adaptations – and the adaptations you did not make. In those cases where you did not implement the changes, you should not simply dismiss doing so as too much work – instead you should explain, specifically, why Cassandra's model was distant from the RDBMs model in a way that impacted your application.

### 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. May we suggest *Pelops*?

### Cassandra Version

May we suggest using 0.70.x rather than 0.68.x? We know it is technically the development, rather than stable, version. But, in our experience, it is more stable and, at present, better supported by the high-level clients.

## 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 Adaptations and Limited Functionality*, which describes how *and why* things have changed from the perspective of the data's organization and the user visible functionality since project #1, (c) *Dependencies*, that describe the precise versions of the operating system and platform, Cassandra, 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.

## Perhaps Useful Resources

- Cassandra High-level Clients: <http://wiki.apache.org/cassandra/ClientOptions>
- Getting Started with Cassandra: <http://wiki.apache.org/cassandra/GettingStarted>
- Cassandra By Example: <http://www.rackspacecloud.com/blog/2010/05/12/cassandra-by-example/>

Remember: We're Here to Help! Email: [staff-415@cs.cmu.edu](mailto:staff-415@cs.cmu.edu)