

CS 410/510 Data Management in the Cloud

Spring 2018 Quarter

Project Part 2: System Profile

Due: Thursday, 10 May 2017 at the beginning of class

You should do this assignment in project groups. You should only talk to your group, the TA and the instructor about this assignment. You may also post questions and comments to the course discussion list on Piazza.

Please turn in your completed assignment on paper. Please put names of all group members (last, first) on separate lines at the top of the first page.

Task 2: System Profile (Written Report – Project Teams)

In this part, project groups should choose from a set of existing cloud-scale data management systems and compile a "system profile" that characterizes that system. This project task is designed to foster an understanding of the common features and the differences between cloud-scale data management systems. The set of systems groups can choose from is the following. If you wish to use a system not on this list, please ask Prof. Maier. Systems that are the primary subject of a lecture (Amazon DynamoDB, MongoDB, BigTable, Neo4j, etc.) cannot be used for the project. Note that the choice of system is for the rest of the project.

- Riak (<http://wiki.basho.com/>)
- Project Voldemort (<http://project-voldemort.com/>)
- CouchDB (<http://couchdb.apache.org/>)
- SimpleDB (<http://www.amazon.com/simplydb/>)
- HBase (<http://hbase.apache.org/>)
- Cassandra (<http://cassandra.apache.org/>)
- OrientDB (<http://www.orienttechnologies.com/>)
- Accumulo (<https://accumulo.apache.org/>)
- Redis (<https://redis.io/>)

Note: Please add your system to your team listing on Piazza once you decide. While teams don't have to pick different systems, it would be good to have some variety.

For the chosen system, the profile will include the details listed below. Each project group should prepare is a ~3-page short paper. Lengths given for each section are estimates. Grammar, clarity and formatting will figure into the scoring of this part.

- 1 Data Model (1/2 page)
 - a What data model described in class (Lect 2 - Data Models) is most like your system's data model?
 - b List the objects in the data model (e.g., collection, column, field.)
 - c Give examples of the data manipulation API, including functions to create, retrieve, update and delete data items.
- 2 Query Support (1/2 page)
 - a What is the query language for your system, if any?
 - b Can your system support range queries? If so, give an example.
 - c Can your system support navigation (graph) queries? If so, give an example.
 - d Are queries automatically optimized?
- 3 Indexes (1/3 page)
 - a Does your system support indices?
 - b Are primary and secondary indices supported?
 - c Are there any limitations on what can be indexed?
 - d Are indexes managed manually or automatically?
- 4 Storage (1/3 page)
 - a List storage options supported (disk, RAM, SSD, database, Cloud)
- 5 Transactions and Concurrency Control (1/2 page)
 - a What consistency options does the system support? (If only one, describe it.)
 - b What type of consistency guarantees does each option provide?
 - c What technology is used to implement consistency? (e.g., locks, R/W quorum, MVCC)
- 6 Scalability and Replication (1/2 page)

- a Does the system support replication?
- b If yes, is the replication synchronous or asynchronous?
- c Does the system support sharding?
- d Is the sharding automatic?
- e Can the system dynamically reshard (say when a node is added)?

7 Platform/Deployment (1/3 page)

- a How is the system usually deployed? What cloud infrastructures are supported?
- b What deployment scenarios are supported, i.e. embedded, client/server, multi-core CPU, cloud, etc.?
- c What language bindings and communication protocols are supported?