CSC 560	Modern DBMS Architectures	Alexander Dekhtyar

Assignment 6: CouchBase vs. MySQL Efficiency study

Due date: Tuesday, November 20, in-class

Assignment

This is a team assignment. Each team will conduct one study and submit one set of deliverables. However, it is a requirement of this assignment that each team member has developed and executed an experiment on the MySQL and CouchBase clusters.

Aims. Using two different clusters, build an application to measure the performance of a read/write workload between a MySQL based cluster and a Couchbase Server 2.0 based cluster. MySQL Cluster will be configured with 4 nodes, one as a the master, and three as read-only using replication. Couchbase Cluster will be configured with 4 nodes. Information about configuration of both systems is provided below.

Process. Each team will receive access to six EC2 instances on the Amazon Web Services cloud. Four instances will form a MySQL or CouchBase cluster; two remaining instances will be used to run client processes.

Team will build and deploy a *workload harness*[1] using the Java Client Library for CouchBase or JDBC driver for MySQL, and will configure the MySQL and CouchBase clusters. The workload harnesses will collect the appropriate performance measures and will output them for the team to analyze.

Experiment design and metrics collected. Each experiment designed by a team shall consist of the following components:

[1] An client application/collection of applications that dispense(s) the workload to the MySQL/CouchBase cluster and collects measurements.

1. **Data.** What data will be stored in the CouchBase/MySQL data stores.
2. **Workload.** What type of activity with the data will be performed.
3. **Contol variables.** What parameters of the workloads will be controlled by the experiment design.
4. **Dependent variables/Measures.** What measures (metrics) of system performance will be collected.

Data. Each team will select its own data loads for the CouchBase and MySQL storage. A single experiment shall use the same data in both CouchBase and MySQL (except to accommodate the data model needs). You can use the same data in multiple experiments, or choose new data loads for each experiment.

Workloads. There are three major categories of simple workloads that need to be tested:

- **Read-heavy workloads.** These workloads are dominated by simple read operations. For the purpose of this assignment, a read-heavy workload is a workload that is 80% or more read operations.
- **Write-heavy workloads.** These workloads are dominated by simple write operations. For the purpose of this assignment, a read-heavy workload is a workload that is 80% or more write operations.
- **Mixed workloads.** The workloads with a more even balance of read and write operations than an 80-20 split are considered **mixed workloads**.

Each team will design a number of workloads for testing the performance of the MySQL and CouchBase clusters. The following conditions shall be satisfied:

1. The total number of different workloads tested in this assignment shall be equal to the number of students on the team.
2. At least one workload of each type: **read-heavy**, **write-heavy** and **mixed** shall be implemented and tested. For **mixed workloads**, try implementing at least one workload that is close to a 50-50 balance.

Control variables. Performance experiments are commonly about selecting one or more parameters to be controlled when building the workload, deciding which dependent variables will be measured and, based on the outcomes of the performance experiment, building a graph of the dependent variable values changing with the changes in the values of control variables.

Each team will develop its own set of control and dependent variables. Note, that the control variables should correspond to parameters of the workload that your workload harness can easily alter on different runs of the experiment. For example, a popular control variable in experiments related to distributed and parallel system performance is throughput - the number of atomic tasks/operations per unit of time, sent to the system under the test.

Metrics/Measurements. We are interested in analyzing the performance of CouchBase vs. MySQL for the purpose of discovering the following:

1. Write I/O performance and limitations
2. Read I/O performance and limitations
3. Time-delay between writes and reads of same document
4. Time-delay between writes and index updates of same document
5. Any potential problems with each cluster design

Each team shall deteremine the specific ways to measure the performance factors describe above. Time-delay, also known as *latency* between different actions of the system can be measured in terms of response from client to server and back.

Setup Information

Each team gets access to six Amazon Web Services EC2 instances to be used for the assignment. The IP addresses of the machines will be handed out separately.

Each team will configure the MySQL four-node cluster (the default distributed configuration for MySQL) with one master and three slave nodes.

Each team will configure a four-node CouchBase cluster.

Instructions for MySQL and CouchBase cluster configurations are provided.

Deliverables

There are two deliverables for this project: code and written report.

Code. Deliver the entire code base for your assignment. Please make sure different components of the code are commented, so that a person familiar with the assignment (and having read your report) could figure out which code performs which functions. Include a `README` file detailing the organization of the code base and description of the data used for the testing.

Submit the code to the TRAC repository on the course wiki by the assignment deadline date.

Report. Write a report detailing the experiments you have conducted and their results. The report shall contain the following:

- Introduction. The introduction can be short. Outline your approach to the performance testing, briefly describe the data you are testing on and the variables you are controlling and measuring.
- Experiment design. The experiment design section should start with a *holistic overview* of all the experiments designed and implemented by each team. This part should integrate the input of all team members into a cohesive description. After the holistic overview, include the formal description of each experimental design. These can be results of individual contributions of team members (or contributions of subsets of team members).
- Implementation. Describe briefly the implementation of the experiments, mention any technical challenges encountered and how they were met. This section may be short, if there is nothing for you to report except the basic information about the cluster setup. However, remember that your document must be readable in isolation - i.e., must not rely on the reader having familiarity with the specifics of the assignment. This means that the text about the four-node cluster setup needs to be included, even though it matches the requirements of the assignmnet.
- Evaluation. Describe the results of each experiment. Build and include appropriate graphs. Since there is no space limitation - make sure to include reasonably large versions of graphs to make results visible.
- Analysis. Can be part of the the Evaluation section. A holistic overview of the results of your testing, and conclusions drawn from the observations made. This section, again, needs to integrate the results obtained from all individual contributions of team members.
- Conclusions. A short conclusions section (if separate from Analysis) to finalize your conclusions and address any limitations of your testing.
- Bibliography and Appendices. As no related work is required for this assignment, bibliography can be kept to a minimum (if you are referencing any work/documents in the write-up, please provide appropriate citations). Appendices can be used to do a data dump - if you have raw performance results that are too big to fit into graphs, or if you've constructed more graphs than is feasible to discuss in the report, dump all information into an appendix of a group of appendices.

- **Formatting.** LaTeX is preferred, but I will accept Word or Google-docs documents as long as they are formatted as academic papers with title, list of authors, abstract, etc. We might release your write-ups as technical reports, please be aware of that.

Submission. Each team submits one copy of the writeup. Email the PDF of the writeup to me on the due day. Eventually, each team will submit a revised (to take any comments into account) version of the writeup on the course wiki, but for the November 20 submission, we will go with an email to dekhtyar@calpoly.edu.

Good Luck!