GORA-485 Apache Kudu datastore for Gora Reports

Summary of activities during GSoC 2019.:

Proposal: GORA-485 Apache Kudu datastore for Gora

Warm up issues (bonding period):

- https://github.com/apache/gora/pull/171
- https://github.com/apache/gora/pull/156

Design documentation:

- https://docs.google.com/document/d/1colS1ooQZIvuJcnx6DSsZIgesokK8TjwaEfgdygt4mo/edit?u

Main interactions and discussions with the community:

- https://www.mail-archive.com/dev@gora.apache.org/msg10089.html
- http://mail-archives.apache.org/mod_mbox/gora-dev/201906.mbox/%3CCAKRR2PUmKtW7vSUTXn6EO-cKkHMwrDwe0EGAk8L8ZJwisCM4vw%40mail.gmail.com%3E
- http://mail-archives.apache.org/mod_mbox/kudu-user/201907.mbox/%3CCA%2BWTGFDWgB6QbNWd0vQdTZC9t4%2Bc33wcUUNiyEHXecsBZevTYA%40mail.gmail.com%3E
- http://mail-archives.apache.org/mod_mbox/kudu-user/201907.mbox/%3CCA%2BWTGFBBdhjXHixnz2AEViEJoFr6PonHKvwsKdyvoE_%2Bz3FOFA%40mail.gmail.com%3E

Final documentation for the Kudu Backend:

- ~~GORA-625~~ - **Add documentation for the Kudu DataStore** RESOLVED

Final source Pull Request:

- https://github.com/apache/gora/pull/178

Weekly Reports:

Report #1

Period: May 27 - June 2

Activities:

- Create a new branch GORA-485
- Create basic structure for the Kudu datastore.
- Research about Kudu connection mechanisms.

Report #2

Period: June 3 - June 9

Activities:

- Implement embedded Kudu instance with KuduTestHarness (Tests).
- Implement connection to Kudu cluster with kudu-client.
- Read schema from mapping file.
- Read connection parameters.
- Implement basic schema operations (getSchemaName, deleteSchema, schemaExists, close).

Questions:

- How to represent partitioning configurations on the mapping file.
- KuduTestHarness requires the Maven plugin os-maven-plugin, which needs Maven 3.1.1+, is it a problem for Apache Gora?

Report #3

Period: June 10 - June 16

Activities:

- Add partitioning configurations in the XML mapping file.
- Create schema using kudu-client createTable method.
- Load the mapping from configuration file.
- Improve Dependency Management.

Report #4

Period: June 17 - June 23

Activities:

- Replace the class Column.DataType for Type (kudu-client) in order to avoid duplication.
- Add javadoc descriptions (main public methods).
- Enable testCreateSchema, testCreateSchema.
    - Ignore temporarily because they need query support: testAutoCreateSchema, testTruncateSchema, testDeleteSchema
- Implement method flush (add KuduSession to the datastore).
- Implement method exists (pending tests).

Questions:

- What is the Gora's policy regarding flush()?

  KuduClient has multiple flushing modes and also can set time interval for automatic flush.

  Should theses behaviors be configurable using gora.properties file? or just use the default configurations.

Report #5

Period: June 24 - June 30

Activities:

- Implement basic serialization/deserialization (using SpecificDatumReader/SpecificDatumWriter similarly to other backends).
- Implement get, put(insert and deletes in Kudu).
- Implement exists (scanner with no projected columns).
- Enable tests for put, exists .
- Fix issue with the deleteSchema method (table deleted).

Questions:

- In the Employee example there is a field called 'dateOfBirth'. I tried to map that field with the UNIXTIME_MICROS datatype of Kudu (I intuitively assumed this is a date.). However, in the java world the Employee field is a Long value and the kudu datatype is a Timestamp. So, I was wondering whether I should force the usage of the UNIXTIME_MICROS datatype for this field or just use a LONG datatype in Kudu.
- What is the Gora's policy regarding flush()?

  KuduClient has multiple flushing modes and also can set time interval for automatic flush.

  Should theses behaviors be configurable using gora.properties file? or just use the default configurations.

Report #6

Period: July 1 - July 7

Activities:

- Basic implementation of Queries (method datastore.execute(query)).
- Improve serialization/deserialization for the binary datatype.
- Enable more tests in the TestKuduStore class.

Questions:

- In the Employee example there is a field called 'dateOfBirth'. I tried to map that field with the UNIXTIME_MICROS datatype of Kudu (I intuitively assumed this is a date.). However, in the java world the Employee field is a Long value and the kudu datatype is a Timestamp. So, I was wondering whether I should force the usage of the UNIXTIME_MICROS datatype for this field or just use a LONG datatype in Kudu.
- What is the Gora's policy regarding flush()?

  KuduClient has multiple flushing modes and also can set time interval for automatic flush.

  Should theses behaviors be configurable using gora.properties file? or just use the default configurations.

Report #7

Period: July 8 - July 14

Activities:

- Improve and test queries implementation (Scanners) *
- Improve get() and exists() implementations *
- Add flush configurations.
- Replace Insert with Upsert in put().
- * More details about this in this mailing list thread.

Report #8

Period: July 15 - July 21

Activities:

- Implement delete by query using a Kudu Scanner*.
- Implement query partitions using the partitions of Kudu (range partitions only).
- Enable remaining tests.
- * More details about this in this mailing list thread.

Report #9

Period: July 22 - July 28

Activities:

- Map Reduce Tests.
    - `Enable mapreduce tests for the Kudu Datastore.`
      `It forwards some some configurations from the gora.propertiesfile to the`
      `Job configurations in order to ensure thatthe datastore instances start up successfully.`
- Replace javafx.util.Pair with java.util.AbstractMap.SimpleEntry
    - javafx is not present in OpenJDK and it could be replaced with AbstractMap.SimpleEntry

Future activities:

- Create a document with documentation for the new data store.

Report #10

Period: July 29 - August 04

Activities:

- Write a first draft of the documentation for the Kudu backend.
    - https://docs.google.com/document/d/1iHex5maLmqrWnIIVfEFm3LtARtARq7pgoYaThjGIKa0/edit?usp=sharing
- Send a PR for code review.
    - https://github.com/apache/gora/pull/178

Future activities:

- Improve documentation adding styles.
- Fix issues in the PR found by the community members.

Report #11

Period: August 05 - August 11

Activities:

- Add num-Replicas parameters within the mapping file.
    - Because this parameter could be necessary for certain use cases.
- Minor issues were addressed based on the comments of Furkan.
- Test Kudu backend with a docker image.

Future activities:

- Improve documentation adding styles.

Report #12

Period: August 12 - August 18

Activities:

- Move module versions to 1.0-SNAPSHOT.
- Submit documentation.
- Rebase Pull Request.
- Fix Carlos comments.

Report #13

Period: August 19 - August 25

Activities:

- Create a patch with documentation for the Kudu backend and post it as an Jira issue.
- Fill up the Google Summer of Code forms for the final evaluation.
- Improve the code by solving the issues pointed out by Kevin.