Hadoop WG Meeting Log 12-9-16

sdolley@cloudera.com

* Roll call
* Shawn Dolley
* Dan Flavin
* Tom White
* Kirill Eitvid (Odysseus)
* Melissa (NantHealth)
* Cornelius (Nanthealth)
* Greg Klebanov (Odysseus)
* Lee Evans
* Hello from new members
* Odysseus, global firm with HQ in Boston, focus on observational research, experience in CDM conversions. Greg is CTO. Kirill is in R&D Practice. Have Big Data skills. AWS, Hadoop. Spark for conversions.
* Call out for topics that need some time today
* Nothing added at meeting start
* Impala CDM 5 code walk-through by Tom White
  1. Tom showed his new code, not yet committed, that enables Achilles to run on Impala (on CDM running on Impala)
  2. In this, he has code in 3 areas: DatabaseConnector project, SQLRender project, and Achilles project
  3. For DatabaseConnector, Tom wrote a new connector for Impala (yay!)
  4. Only high priority topic for future discussion re the DatabaseConnector work is that the Impala connector code has a lot of jars with it (inherited from other OHDSI databaseconnector code?) that have nothing to do with Impala (Tom can explain this better when we can have that dialogue)
  5. On SQLRender, it looked to me like Tom made hundreds of mappings from SQLServer SQL syntax to their equivalent in Impala (yay! \* 2)
  6. High priority topic here is a) testing, how can we do unit testing, who can test it, do we need 100% test coverage?
  7. On Achilles, biggest change is CASTING. And b) CASTING. SQL Server is known to be lenient however Impala needs explicit datatype or other instructions with CAST everywhere. When we/if we merge Tom’s code, can all database SQL have specific CASTS? Other approaches?
  8. Odysseus asked Tom if some parts of the CohortMethod should work using the SQL Render code, Tom says yes it should
  9. Odysseus runs Spark on YARN (and Parquet) and was wondering if Parquet would work OK if used in the context of Tom’s code to date, answer was yes
  10. Will it work on Hive came up, I think the answer is current code leverages Hive metastore, but if you don’t want to use Impala and you want to use exclusively Hive, that code is not yet written
  11. SparkSQL should work
  12. Need to look at SQL questions from Dan (on Merge, on Oracle, etc) (CASTING and other), will use another meeting time for this
* Impala CDM 5 code walk-through by Tom White
  1. Straightforward, no questions
* ~~Reference Architecture mini walk-through by Derek Kane (posted on wiki and you may have received via email)~~ Derek not here, maybe Tom can do this a bit and full one as follow on
* Review of survey results by Shawn Dolley (results are posted on Wiki and forum btw and you may also have received via email): completed, shawn will send out again maybe with some more readable insights
* New business/other topics: None
* Scheduling suggestions for future demos from Hadoop-based OMOP CDM-based commercial solutions (from Outcomes Insights, your organization if you have a demo, Knowledgent, others): Group decided best approach was make the demos a bit shorter and run them during the Hadoop WG call
* Next steps:
  1. Desires for more stuff: Odysseus will be thinking on this, what they want next; the group had no immediate comments on things missing they needed (vs. what track we’re currently running down…)

The Hadoop Prescription

