Meeting Log (three pages)

Nov 11, 2016

Intros of first time attendees to the call/working group. Only new person is Steve Lyman.

Old Business: the questionnaire/survey. I want to bring the survey up on screen during the meeting, and get commits from people to fill it out! Shawn showed the survey on the screen. Six respondents so far.

New Business

1. on the concept of having a reference architecture (this advice came from group members, not the moderator):
   1. it can generate practical advice; what are the first steps to do in Hadoop; [within an organization]
   2. a ref architecture helps people narrow down choices (cannot define a single reference architecture).
   3. each domain will be different.
   4. So there is consensus that having a reference architecture or multiple ones per use case is a good idea
   5. "it does help to see how some people have achieved their use case".
   6. Problem: "people think they are talking about the same thing, but they are not" lots of ways to do the same thing.
   7. "would be helpful to have some suggestions for each use case, as well as things people have tried and haven't worked, or limitations"....
   8. might end up being some consolidation of use cases tied to an architecture.
2. On concept of progress within workgroup setting (this advice came from group members, not the moderator):
   1. Don’t forget they are informal,
   2. good to have deadlines, key is getting someone to volunteer, can't coerce people to a deadline,
   3. we should be reporting back to the broader community. Also need to report back to main group 'this is the progress of the Hadoop WG so far'.
3. Possibility of seeing ‘demo’s’ in the work group con call
   1. would like to see what others are doing, will ping Shawn to see if he can line it up.
   2. Do a show and tell.
   3. 'what challenges have they had so far, what have they done so far.'
4. useful follow on survey:
   1. get benchmark of background of the different members, years of experience with big data
   2. what have folks done so far,
   3. survey (big data projects of interest, e.g. Spark, Hadoop, etc).
5. should this group be changed to "Alternative Platforms" and include Google BigQuery?
   1. Google BigQuery could be considered SQL-on-\_\_\_\_ ...
   2. maybe goes with Redshift/DWH or another version of SQL-on-big data.
   3. We want to focus on things like Impala.
   4. “we can't focus on too many things or we will achieve nothing”
   5. 'There is a lot in Hadoop ecosystem by itself'.
   6. Shouldn't change the name of the wg.
   7. "If we can hear about cost impact, when it is good to use it or not".
   8. At this time, no one believed we should change focus or widen focus
6. Call for people to mention what they use for Hadoop if they have already started, and on what people would be willing to make open source or not and why and why not
   1. Quintiles uses Parquet
   2. Lee uses Parquet and Spark
   3. Nothing so secret sauce about putting OMOP v4 or v5 into Impala or some other big data database. Any secret sauce for these are probably outside CDM or something (that is one thought).
   4. A lot of people make extensions to the CDM, sometimes those ought to be private. Maybe someone who has done it has more than Tom made.
   5. TOM HAS IT TIED TO STRINGS! This is a roadmap issue with the current committed Impala CDM build
7. Version thoughts on CDM
   1. when they go to v6 will be an interesting time.
   2. Finishing version thought: there is a script lets you migrate from 4 to 4.5 to 5.
8. what can we be doing in parallel over next few week
   1. have Tom on one of these calls to describe his code
   2. Shawn tabulates survey results and publishes them

Tom's roadmap or next step list seems good, one benefit might be on his list he could run Achilles web against impala backend: this list is:

1. CDM schema for Impala

2. SqlRender support for Impala

3. DatabaseConnection support for Impala

4. Achilles support for Impala

5. Achilles HEEL support for Impala

6. Run Achilles web against an Impala backend

Tom: “1 is done (it was committed by Patrick Ryan). I've got patches for 2 and 3, and have made some headway on 4. I sent a message to the list about 4: <http://forums.ohdsi.org/t/first-achilles-queries-working-with-hadoop/1876>. It would be very useful to have some people trying out what I've done on 4, reviewing it, and getting it completed. Once that is done then others could do 5 and 6.”