ServiceX Working Meeting

December 13, 2019

https://indico.cern.ch/event/870876/

Notes courtesy Chris Weaver, UChicago

Ben G is working to extract a common library for the transformer interface.

Gordon asked whether the plan is for transformers to include that library, or to start from an official base image. Ben intends the first method, as he thinks the second would be difficult to make work universally.

Brian B asked about unit tests with nanoAOD. Lindsey has provided a ten event nanoAOD file which may be small enough to bundle for a test.

Ben is thinking about how to use CI to run nightly unit tests to protect against regressions, like computing a particle peak and making sure that it doesn't move.

Mark N: What are top two priorities?

Ben: Should be represented by the top items on the board. Some cleanup of the board is needed.

Rob has identified autoscaling as a high priority, so it is at the top. The next item is including Kafka in the helm chart.

There are two aspects to the C++ transformer: code generator and validator. The code generator is packages as a docker container, so it just needs to be inserted into the automatic build system for service-x containers. There are tests in the repository. Gordon is developing all of this on the master branch, there is nothing in particular in the development branch. The other part is the one which runs the generated code. It is almost working, and once it is Gordon will be ready to start integrating with the core of service-x. He expects to start working on that over the weekend or early next week. He asked whether he should work in a copy of the repository, or in some separate branch and then submit a pull request. Ben requested that he use a branch. Gordon would like the code reviewed once he has written it, and at some point he will need guidance on creating configmaps. Ben will be in Seattle on Friday.

Ben: The new transformer will speed up transforming XAOD, and Mason is working to control uproot which can be used to do KyungEon's selection.

Mason P has started looking at how to combine his work with Gordon's. He is somewhat confused about which code depends on which and how his code should connect to FuncADL. He and Gordon will discuss this early next week.

Brian B asked what the status is of Service-X for CMS, and of integration with SkyHook.

Ben: Task #48 is to make the/a DID Finder communicate with CMS Rucio. He is currently at the point where he has been asked to provide a copy of his passport to be granted access. In theory there is little coding to do, but he is unsure exactly what will happen once he gets access to try it out.

Ben has been working with Jeff L on the SkyHook integration. He sent Jeff a Parquet file, and things seem to be making sense so far.

Jeff reported that Lincoln has partitioned off three machines to run Skyhook, and after a little work everything has installed cleanly on CentOS. He had a little trouble with a three node cluster, so Lincoln has added a fourth machine to act as a controller. He expects to have it running by the end of the day. He also has an example notebook from Brian, the data retrieval portion of which he thinks he is close to duplicating. He will then set up a container running the skyhook listener on the SSL cluster. Ben has given them an example Kafka table in arrow format, which they have been able to store into skyhook. Lastly, they are writing an API to allow Ben to send a table to skyhook. He thinks that will take a couple of days.

Gordon asked about the skyhook demo which happened recently, which he unfortunately missed. Jeff will forward him the slides from the demo.

Mark N asked whether there is a diagram showing how Skyhook will fit into the overall system. Ben thinks that at the moment there is only a proof-of-concept, so this may change. Skyhook may be treated as a generic analysis system, or it might be used as a type of cache for service-x internally. Mark would like to discuss this in more detail at some point.

Carlos M mentioned that a key benefit of skyhook will be saving CPU cycles by reducing data transfer. However, the performance is not yet known; this will be measured soon.

Gordon asked whether there is a place where a CMS development instance of Service-X can be run. Ben thinks that this can just be done on the SSL, which is not ATLAS-specific. Marc W suggested that running at FNAL might also be a possibility. Ben noted that the SSL River cluster is not close to a CMS Tier 2, so that might not be ideal, and might be a reason to go to FNAL. Task #83 has been created to track this.

Ben wants more people to be familiar with the Zenhub board, and feel that they have ownership of it. He and Gordon have been using it actively. Two columns to be aware of: Ready represents

items which are well-understood, and should be worked on next, starting from whatever is at the top of the column. Backlog contains longer-term items, which should be revisited periodically, and the items people believe should be done sooner should be moved to the top.

Gordon wasn't sure whether he may have created some redundant entries, so he suggested that if this is the case Ben should close the duplicates and point him to the main one to work on.

Mark N likes the collaboration on the board, but he was concerned about whether it will become disorganized. He had not previously been aware of the significance of the vertical ordering.

Ben described that stories should describe single features which are desired, whether or not they touch multiple repositories. Epics enclose multiple stories for larger concerns, e.g. making KyungEon's analysis work. Another epic, identified by Rob, is being able to reliably deploy an entire Service-X platform from the helm chart. The third currently is getting the C++ transformer running, and perhaps supporting CMS should be a fourth.

Gordon feels a bit nervous about some of the items in the analysis epic. He is worried that some of them are out of scope for service-x, and might stretch the project too thinly.

Ben would generally like to address these questions of scope on the Friday calls, but it can also happen on another call.

Gordon noted that there are some important milestones to keep in mind, such as the 18 month review which is coming up. He thinks it is also important to make sure that CMS is well represented, as he doesn't want CMS to get the impression that this is an ATLAS-focused effort and go off to build their own, separate system. The focus on ATLAS has been largely organic, because a number of ATLAS people became excited about the project, so it's important to build interest on the CMS, since some of their expertise will be needed to get components like the CMS DID Finder working. Mark N mentioned that if Ben has problems getting CMS access he should get in touch.

Mark N thinks that the coffea integration is important, and since that is a primarily CMS project, that does tie strongly to CMS.

Gordon feels that the integration with coffea is fairly shallow, which he acknowledges as good engineering by the coffea team, but poses an issue for claiming a major connection there. He thinks that more people are needed from the CMS side who are interested in this vision, and have time to devote to it up front.

Ben noted that the existence of nanoAOD makes service-x less badly needed than it is when working with ATLAS XAOD.

Mark and Marc noted that nanoAOD use is not yet widespread in CMS, and it may not be able to support all types of analysis.

Gordon noted that while he thinks some of his code can be reused for targets like CMS miniAOD, but he really doesn't have the insight to know what would be needed to make it effective. He suggested drafting what the team thinks will be needed for this, and then getting feedback from CMS analyzers on whether it sounds correct.

Ben thought that the C++ transformer may then be the centerpiece of building interest from CMS.