ATLAS with a BigQuery backend running Execution Engine - a Software demo

Jose Posada[a], Priya Desai[b], Konstantin Yaroshovets[c], Gregory Klebanov[c]

[a] Center for Biomedical Informatics Research (BMIR), Stanford School of Medicine
[b] Technology and Digital Services (TDS), Stanford School of Medicine and Stanford Health Care
[c]Odysseus Data Services Inc

Background

Stanford has adopted an ecosystem view[1] of the modern clinical research tools. Built on the foundation of STRIDE, the ecosystem has since expanded to STAnford medicine Research data Repository (STARR) ecosystem. The overall design principles of the ecosystem are based on Data Commons and includes compute and storage infrastructure, data lake, data warehouses, data processing pipelines, APIs, tools, user training, and support. Our overarching goal is to streamline science for researchers.

Backbone of the STARR ecosystem is STARR-OMOP, an analytical clinical data warehouse that uses OMOP Common Data Model. One of the reasons for Stanford to choose OMOP was OHDSI in its entirety, not just the data model, we wanted the tools, the network, the community. Another critical part of our ecosystem is our data center. The compute and storage infrastructure has grown from on-premise data center to embrace cloud, not just for its larger storage and compute capacity, but also for specialized solutions. One such specialized solution is Google BigQuery, a managed distributed data warehousing solution. Stanford had previously implemented Google Cloud BigQuery for a Big Data genetics initiative[2], so it was natural to try BigQuery for STARR-OMOP. BigQuery brings two very significant features, one is the fact that it is a managed service and unlike traditional databases, it doesn't require DBA tinkering for performance. It is performance out-of-the-box. The data engineering team can focus on data standardization, completeness and quality instead of indexing, sharding, and scaling. The second big feature is the data science friendly APIs. Researchers can use their laptops or HPC environments to use their Jupyter Notebooks and never really get out of the tools they do data science with.

In a previously published manuscript, we show that ATLAS benchmarking suite using SynPUF runs 3 to 10x faster on BigQuery when compared to PostgreSQL (Manuscript[1], Supplementary Table S9.3). We also show that Achilles queries run in ATLAS using STARR-OMOP data present near real time user experience. Out of 725 total queries available in Achilles, 660 queries took less than 17 seconds, and median execution time was 3 sec (Manuscript[1], Supplementary Table S9.1).

While direct or API based SQL query using BigQuery is highly performant, the OHDSI toolkits do not directly use BigQuery. Instead, the tools use shared libraries such as DatabaseConnector, and SQLRenderer that translate the query to BigQuery SQL dialect. Optimization of the OHDSI toolkits to run on BigQuery is a journey we embarked on nearly two years ago. This journey has since led to successful deployment and utilization of ATLAS at Stanford. . We have also embraced the execution of ATLAS PLE and PLP analyses through ARACHNE Execution Engine[3]. The engine allows us to fully execute estimation and prediction studies right inside ATLAS. This presentation will demonstrate Stanford ATLAS running on top of STARR-OMOP including the ARACHNE Execution Engine.

Methods

In a close academia-industry-consortium collaboration, Stanford, Google Cloud and Odysseus, have worked with multiple OHDSI collaborators to ensure that core OHDSI R library and methods were updated to work with Google BigQuery. There were a number of fixes across multiple important shared libraries, including DatabaseConnector, and SQLRenderer. The team also identified a number of Google BigQuery limits that were causing bottlenecks for code execution, which Google removed, or significantly raised, multiple limits, including a number of concurrent inserts or table metadata updates.

To make it easy to use the ARACHNE Execution Engine on BigQuery, we have also released public docker images[4] so data scientists can have a fully operational working environment in a matter of minutes. We have embraced the ROhdsiWebApi package[5] and bulk imported more than 1000 thousands pre-defined cohorts including the ones from the OHDSI phenotype library[6]. Odysseus has since released ATLAS in Google Marketplace[7].

We have also been able to add Optum DoD OMOP dataset[8] in BigQuery for researchers who have access to the data via Stanford Center for Population Health Studies.

Results

The underlying STARR-OMOP database that Stanford ATLAS uses, specifically a de-identified version, is updated weekly and contains data from approximately 2.7 million patients. Over 75% of patients have at least one diagnosis code, over 50% have medication information, over 75% have laboratory test information, and over 90% of patients have clinical notes data available. The resulting BigQuery database has 340 million rows in observation tables, 2.5 billion rows in measurement table, 275 million rows in drug_exposure table, 130 million rows in procedure_occurance table, and 160 million rows in condition_occurance table. Since our first launch of the OMOP database in late 2019, we have brought in ~28 flowsheet elements that have added significantly to our measurements table. Specifically, we extract vitals such as blood pressure, oxygen level, heart rate, respiratory rate, Sequential Organ Failure Assessment (SOFA) scores, Glasgow Coma Scale Scores, and Deterioration Index scores. The underlying BigQuery performance has not been impacted with the significant increase in the measurements table.

Stanford launched ATLAS on its weekly refreshed OMOP data in Big Query in Q2 2020 with execution engine turned on. In the last 12 months, Stanford's ATLAS has supported ~125 users who launched ~3400 jobs including cohorts and full studies (~1,500 cohorts, 353 concepts sets, 56 Characterizations studies, 30 incidence rates studies, 14 Patient Level Prediction studies, 8 Estimation studies)

The Optum DoD OMOP dataset has data from 77 million patients. The resulting BigQuery database has 910 million rows in observation tables, 4.6 billion rows in measurement table, 3.2 billion rows in drug_exposure table, 3.5 billion rows in procedure_occurance table, and 5.5 billion rows in condition_occurance table.

Stanford has used its STARR-OMOP and ATLAS to participate in 11 published network studies[9], 10 of these are in COVID-19. Stanford is participating in 13 other studies.

Conclusion

At Stanford, we have successfully deployed Google Cloud BigQuery for ATLAS and found the execution engine to be of very high value for research.

References/Citations

1. Datta S, et al. A new paradigm for accelerating clinical data science at Stanford Medicine, arXiv:2003.10534, Mar 2020, https://arxiv.org/abs/2003.10534
2. Pan C, et al., Cloud-based interactive analytics for terabytes of genomic variants data, Bioinformatics, Volume 33, Issue 23, Pages 3709–3715, Dec 2017, https://doi.org/10.1093/bioinformatics/btx468
3. ARACHNE Execution Engine: https://github.com/OHDSI/ArachneExecutionEngine
4. Docker image with pre-installed R and OHDSI (HADES) R packages based on Ubuntu/Debian Linux: https://hub.docker.com/r/odysseusinc/r-env
5. ROhdsiWebApi package: https://github.com/OHDSI/ROhdsiWebApi
6. OHDSI phenotype library: https://data.ohdsi.org/PhenotypeLibrary/
7. ATLAS v2.7.8 with BigQuery support is available on Google Cloud Marketplace. : https://cloud.google.com/blog/topics/healthcare-life-sciences/powering-open-source-healthcare-research-on-google-cloud
8. Optum DoD dataset: https://redivis.com/StanfordPHS/datasets/1433
9. STARR-OMOP network studies: https://med.stanford.edu/starr-omop/summary.html#omop_network_studies