Architectural Considerations to Enable Large Scale Analytics from the OHDSI Web Platform

Frank J. DeFalco[1,2], Martijn J. Schuemie[1,2]

[1]Janssen Research & Development, LLC, Raritan, NJ

[2]OHDSI collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY

ABSTRACT

- While tremendous progress has been made implementing statistical analysis methods by the OHDSI community the level of technical skills required for their execution has limited their accessibility.
- Applications such as ATLAS provide a friendly user interface to leverage parts of the OHDSI technology stack but the methods that require long execution time and large scale computing power have remained outside of their scope due to architectural limitations.
- New considerations in the architecture now remove these limitations and will enable new capabilities accessible to all users from within the OHDSI web platform.

BACKGROUND

- The OHDSI architecture has existing capabilities for long running processes through the Spring Batch Job Manager implementation hosted within the WebAPI.
- The limitation of this implementation was that any job would execute upon the infrastructure hosting the web platform.
- When considering the execution of long running analyses such as those enabled by the various R based methods compute capacity and scalability quickly becomes a rate-limiting factor.
- In particular packages such as CohortMethod[1] and PatientLevelPrediction[2] require more compute capacity than what is provided by the typical web server used to host the WebAPI.

DISCUSSION

- The introduction of two new components, the R Service Bus[3] and RPooli[4], provides the ability to execute R methods across a distributed infrastructure and provide the required scale and compute capacity.
- An example use case would be for Population Level Estimation:
  - Define treatment, comparator and outcome cohorts
  - Define study specifications
  - Submit specifications to the OHDSI Service Wrapper in order for execution
  - Leverage the CohortMethod package to generate study results.
- The new components are accessed as new tasks executed through the existing Job Manager framework (figure 1).
- The primary component contacted by the Job Manager is the R Service Bus which in turn communicates with RPooli to launch R nodes for the large scale analytics processing.
- Once the R Service Bus initializes a node within RPooli then analyses such as a comparative cohort analysis can occur in an independent R process.
- Access to the underlying methods library packages will be managed by a wrapper package that will encapsulate approved study implementations leveraging best practice guidelines for each package.
- This type of analysis would access the necessary data and leverage the existing R packages to generate output that can then be saved and reviewed (figure 2).

CONCLUSIONS

- The architectural considerations introduced here will enable a new class of use cases from within the tools provided within the OHDSI web platform.
- Additional research and implementation will be required to fully assess the extent of the platform's scalability across cloud and distributed infrastructure.

CONTACT

fdefalco@its.jnj.com

fdefalco@ohdsi.org

FIGURE 1

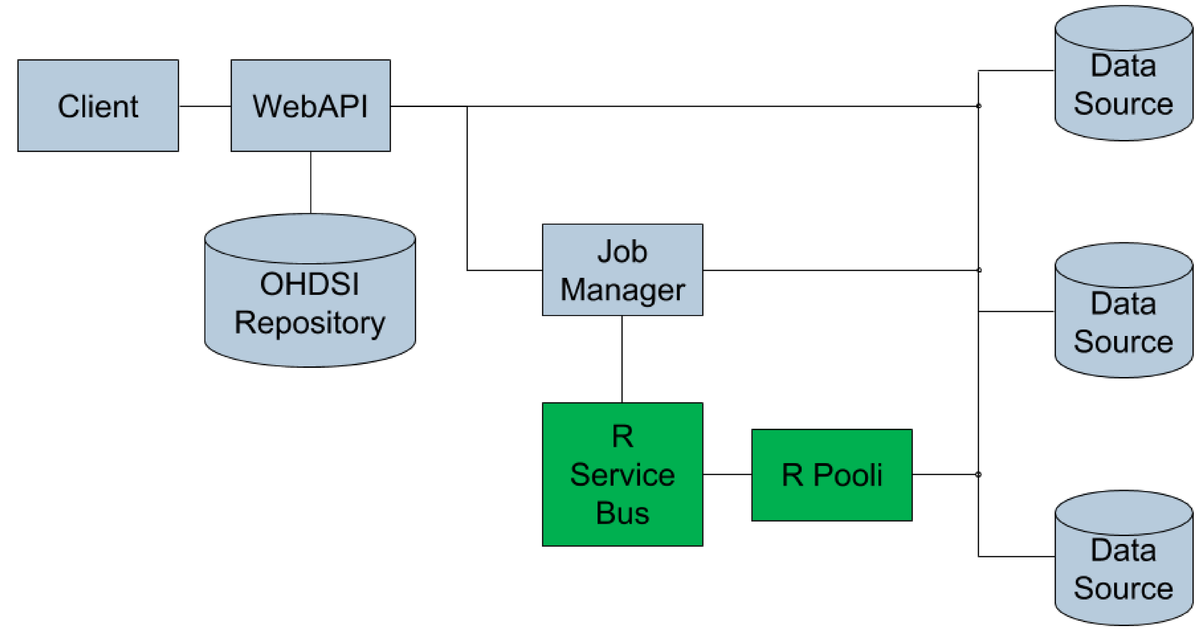
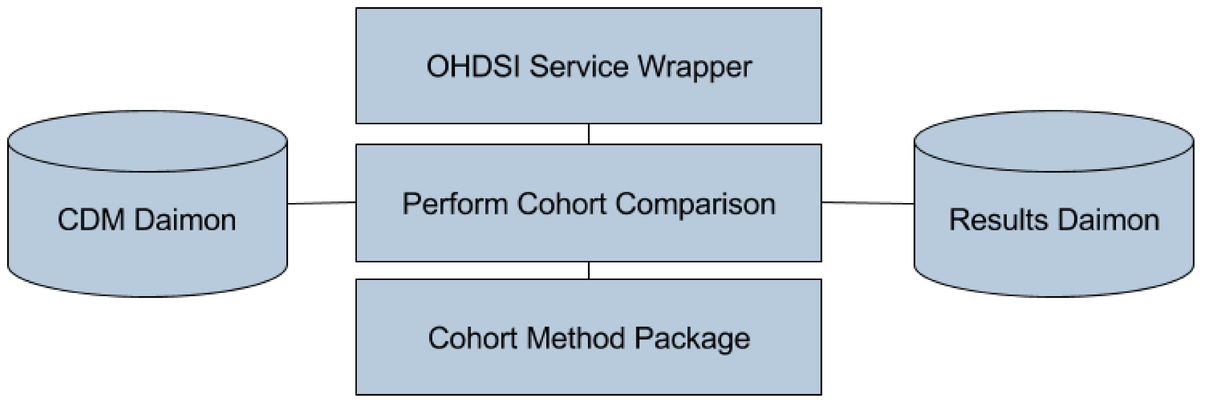


FIGURE 2



REFERENCES

Martijn J. Schuemie, Marc A. Suchard and Patrick B. Ryan (2015). CohortMethod: New-user cohort method with large scale propensity and outcome models. R package version 2.0.3. (https://github.com/OHDSI/CohortMethod)

2. Martijn J. Schuemie, Marc A. Suchard, Patrick B. Ryan, Jenna Reps and Peter Rijnbeek (2015). PatientLevelPrediction: Package for patient level prediction using data in the OMOP Common Data Model. R package version 1.1.0. (https://github.com/OHDSI/PatientLevelPrediction)

3. Open Analytics R Service Bus (2016). (https://www.openanalytics.eu/r-service-bus)

4. Open Analytics RPooli (2016). (https://github.com/openanalytics/RPooli)