Designing and Building a Hybrid Data Cloud*

Juan F. Sequeda, Dave Griffith, and Bryon Jacob

data.world, Austin, Texas, USA
juan,dave.griffith,bryon@data.world

A data cloud is where your data is available to your people and your machines - it's where your data assets are leveraged. data.world launched a data cloud in 2016 as a collaborative and open web platform where anybody can sign up to work with data [1, 2]. data.world's data cloud platform is build using knowledge graph technology and semantic web standards including RDF and SPARQL but also standards such as CSVW [5, 6] among others. All data loaded into data.world is translated into RDF. The platform prioritizes query response time over update flexibility. Once data is bulk ingested, the result is an immutable RDF dataset in the RDF HDT (Header-Dictionary-Triples) file format [3]. This architecture is optimized for exploratory queries and allows to treat datasets as independent graphs which can then be loaded together as named graphs for optimized joins. Knowledge Graphs form a backbone to a data cloud because they focus on the relationships between your data assets - between the databases you own, and with publicly and commercially available databases.

data.world is now taking their data cloud and applying it to enterprise needs becoming the first cloud native data catalog powered by Knowledge Graph technology. One of the obstacles is that enterprise customers do not want to move their enterprise data into the cloud for security and privacy reasons. Thus, naturally a data cloud will evolve to be hybrid: combining data uploaded to a public cloud, such as data.world, and elements of on-prem systems such as data warehouses and relational databases behind an enterprise customers firewall.

Knowledge Graph Virtualization technology is key to address this obstacle. The goal is to virtualize enterprise data as if it were a knowledge graph and have the ability to query it in real time in SPARQL, without the need to move the data to a centralized storage. This is done by mapping the source database to a target ontology. An example of this technology is Capsenta's Ultrawrap [7] and Gra.fo (https://gra.fo/).

Recently, data.world and Capsenta have merged in order to combine a data cloud with Knowledge Graph Virtualization technology in order to enable a hybrid data cloud. This integration of technologies will help bring organizations into the data value chain by leveraging their on-prem data with data in the cloud.

In this talk, we will discuss:

- our architecture to combine Ultrawrap with the data.world data cloud platform.
- engineering obstacles that we have encountered and the solutions that we have developed.
- use cases that leverage on-prem data with cloud data in various domains.
- the methodology required to operationalize ontologies and mappings [4] in enterprise settings.

– open challenges which we believe there is still interesting science to be done.

References

1. Bryon Jacob, Dave Griffith, Triet Le: data.world: A Platform for Global-Scale Semantic Publishing. ISWC Industry Track 2017.
2. Bryon Jacob, Jonathan Ortiz: Data.world: A Platform for Global-Scale Semantic Publishing. ISWC Industry Track 2017.
3. Javier D. Fernandez, Miguel A. Martnez-Prieto, Claudio Gutierrez, Axel Polleres, Mario Arias: Binary RDF representation for publication and exchange (HDT). Journal of Web Semantics 19: 22-41 (2013)
4. Juan F. Sequeda, Willard J. Briggs, Daniel P. Miranker, and Wayne P. Heideman. A Pay-as-you-go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases. ISWC 2019.
5. Jeni Tennison (Ed). CSV on the Web: A Primer. W3C Working Group Note 25 February 2016
6. Jeremy Tandy, Ivan Herman, Gregg Kellog (Eds). Generating RDF from Tabular Data on the Web. W3C Recommendation 17 December 2015
7. Juan F. Sequeda, Daniel P. Miranker: Ultrawrap: SPARQL execution on relational data. Journal of Web Semantics 22: 19-39 (2013)