data.world: A Platform for Global-Scale Semantic Publishing

Bryon Jacob[1][0000-0003-0470-9300] ,Dave Griffith[1][0000-0001-9700-0012], Triet Le[1][0000-0001-5619-5802]

[1] data.world, 7000 North Mopac Expressway #425, Austin, TX 78731 USA
bryon@data.world dave.griffith@data.world triet.le@data.world

1 Introduction

data.world (https://data.world/) is a collaborative web platform with a user base consisting primarily of users who are not Semantic Web experts, and datasets that are not initially semantically annotated or linked. By using web standards for the automated translation of those tabular data formats into RDF, data.world leverages the iterative data work done by the users of the platform to build a connected network of linked datasets. data.world is an open platform where anyone can sign up for a free account to work with open data - it was launched in July of 2016, and as of a year later is in active use by a community of tens of thousands of users and organizations.

An oft-cited fact is that finding, understanding, and preparing data for use can take eighty percent of the time spent on an analysis project. These projects usually involve multiple data sources in a variety of formats. Semantic Web offers a powerful set of tools (universal structure for data, federated query) to deal with this diversity. Metadata can be iteratively layered into datasets, by different actors at different times.

data.world focuses on collaborations where each actor is empowered to participate in the iterative development of the data resource, by helping to clean, annotate, and contextualize the data. Data can be worked on in the open, or in access-controlled datasets and projects. Structured data is converted into RDF and can be queried via SPARQL, but the original data is retrievable as well, so that users can continue working in familiar modes while enjoying the benefits of semantic web technology.

2 Managing diverse data and user base

The majority of structured data in the world is tabular in nature. CSV and other text variants, spreadsheets, relational database tables, and many "long tail" data formats are all representations of tabular data. CSVW provides a model for modeling tables within an RDF graph structure. CSVW tables use RDF schema and types, can be mixed and queried together with graph structures defined directly in RDF, and can be serialized for transmission or storage in any textual or binary form that RDF can take.

Across projects and domains, and often within a single project, there are a diverse set of actors. There are knowledge engineers who create ontologies and knowledge bases; data scientists and statisticians who produce models and visualizations; analysts and scientists who use spreadsheets or visual analytics platforms; and end-users and stakeholders consuming the conclusions of the work. data.world emphasizes collaboration between these personas.

3 Architecture for Scalable, Heterogeneous Data Publication

data.world prioritizes query responsiveness over update flexibility. Updates are handled as bulk ingest, the output of the ingest is an immutable RDF dataset in the HDT (Header-Dictionary-Triples) file format. This HDT architecture is optimized for the queries that characterize exploratory usage - and allows us to treat datasets as independent graphs, but loadable together as named graphs for optimized joins.

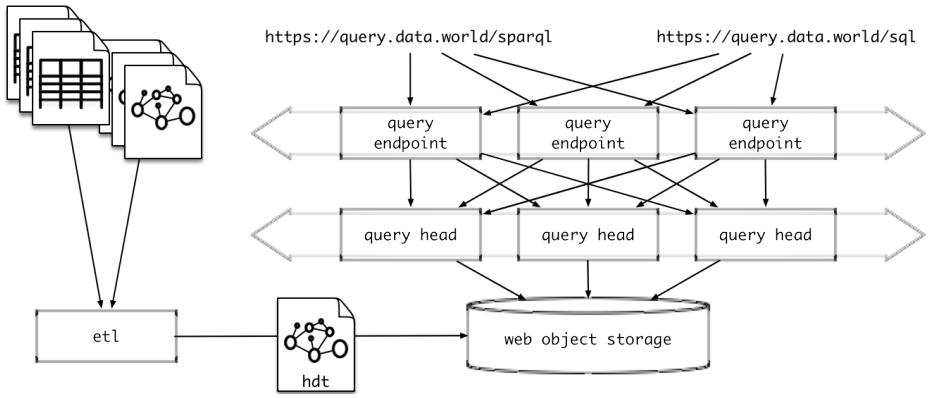

Fig. 1. data.world high-level ETL and query architecture overview – data files are ingested through the etl process, then derivative data rendered as HDT is persisted. Query heads load and cache HDT to execute SPARQL queries on demand, and query endpoints expose SPARQL endpoints and SQL query endpoints to the web – the endpoints parse and rewrite SQL into SPARQL, and parse SPARQL queries to rewrite and route them. Each layer is scalable, with performance proportional to the size of the query set, not to the overall collection.

4 Future Work and Conclusions

Dataset versioning and provenance is an area of active research and development for data.world – our presentation will cover our current work there. Our HDT-based architecture works well for exploratory queries, but it is suboptimal for large analytical (non-selective) queries. We will talk about the work we are doing to leverage a hybrid query architecture to support both simultaneously.

Our hypothesis is that to the surface area of the web of Linked Data, we need to nurture of the network of people who are working with data. A component of every data project is collecting, cleaning and preparing the data – and we can use Semantic Web technology to both facilitate that work and leverage that work to grow the web of linked data. We have seen promising indicators that support this, and as the community grows we hope to present a quantitative assessment of that growth. One group that leverages data.world is Data For Democracy (http://datafordemocracy.org/) - hundreds of data scientists, analysts, and programmers using data.world to build data dictionaries, capture cleaned data alongside raw data, and highlight the relationships between data. This work is annotating data and enriching metadata, which is turning raw data in CSVs and spreadsheets into meaningful sources of linked data.