Business Analytics Using The Knowledge Graph Built Over The Russian Legal Entities Registry

Eugene Hlyzov[1], Sergey Isaev[1], Yury Emelyanov[2], Dmitry Pavlov[2], Olga Belyaeva[2], Dmitry Mouromtsev[3], Olga Parkhimovich[3], and Maxim Kolchin[3]

[1] DataFabric Ltd., St.Petersburg, Russia
{eugene.hlyzov, isaev}@datafabric.cc
[2] Vismart Ltd., St.Petersburg, Russia
{yury.emelyanov, dmitry.pavlov,olga.belyaeva}@vismart.biz
[3] ITMO University, St.Petersburg, Russia
mouromtsev@mail.ifmo.ru, olya.parkhimovich@gmail.com, kolchinmax@niuitmo.ru

Basic facts
URL of the live web application: http://tree.datafabric.cc[1] (launched on 17.07.2017).
Product type: subscription-based commercial web application.
Application domains: business intelligence, data analytics.
Semantic technologies (ST) employed: RDF, OWL, knowledge graphs, SPARQL.
Volume: 2.8 bil. triples, 10 mil. companies, 12 mil. individual entrepreneurs, 27 mil. persons, 333 Gb of raw unstructured data.

1 Motivation And The Product

Throughout the last 3-4 years Russian authorities have made great volumes of data collected and maintained by government services open and accessible to the people. The knowledge concealed in these data is of a great business value and open a new prominent opportunities for services that offer convenient and efficient ways of finding, exploring and discovering the hidden relations and structures that the data contain. In our case, the published open data required a tremendous effort in merging various bits, checking and distilling it before it could be used in an application. We demonstrate how the linked data and visual analytics are able to satisfy the information needs for official registry data exploration and save the user's time in the process of business investigation. Finally, from the market perspective, despite the existence of well-established tools for business intelligence over the open government data, we observe an unsatisfied demand for the affordable, easy-to-use and lightweight services targeted to individuals and small and medium businesses.

Our web service allows users a) to search the registry for the entity of interest b) create a graph and explore visually its connections to other companies, persons, etc. c) produce a user-generated report about entity's background. The service requires subscription fee to be paid up front. To access the full functionality of the service for reviewing purposes please use the following credentials: login: iswc2017review@example.com, password: demo.

[1] English version (for demo purposes only): http://tree-i8n.apps.datafabric.cc

2 ST In The Product

The added value of ST

Expected business value from utilizing ST: a) simpler and cheaper integration of additional sources of data; b) cheaper custom development; c) richer user experience - discovery-provoking visual exploration of company context.

The role of ST in the architecture of the application: The application is comprised of 4 major parts: data extract-transform-load (ETL) pipeline hosted on google cloud, RDF-triple store deployed on open source version of Blazegraph[2], server-side logic and request handling service running on Kotlin and Node.js and frontend user interface logic and rendering implemented using ReactJS and open source Ontodia library[3]. On all stages of data transformation from storing the data to visualizing it we employ ST: data is stored in RDF graph with OWL ontology applied to it in ETL process, data is queried with SPARQL, data is visualized in the form of a graph by conversion of RDF into nodes and edges.

Advantages of using ST in our application

- We apply ST standards to explicate the knowledge concealed in the raw data by transforming the semi-structured XML files into cumulative RDF graph with well-defined data schema expressed in a form of OWL ontology .
- On the basis of ST we align the internal graph-based data structures with their graph-based diagrammatic visualization in Ontodia library[1] thus reducing the amount of data transformation steps and simplifying our code base.
- With the help of ST we extend the graph with additional data sources, which carry their own schemas, by investing minimum effort into it.
- By employing the underlying ontology we are empowered by an effective and automated data publishing mechanism assuring convenient data consumption, which contributes greatly to visibility of our product. [4]

Challenges that arise from locking on ST

- To introduce the regular software developers to the tools of semantic web by selecting the most mature technologies and utilities.
- To solve the mismatch between the ontologies that come from knowledge engineers and requirements of our use case regarding querying the data.
- To adjust the queries and system architecture to achieve the expected commercial product stability and performance on complex data calls.

References

1. Mouromtsev, D., Pavlov, D., Emelyanov, Y., Morozov, A., Razdyakonov, D., Galkin, M.: The simple web-based tool for visualization and sharing of semantic data and ontologies. In: Int-l Semantic Web Conf. (Posters & Demos) (2015)

[2] https://github.com/blazegraph/database
[3] https://github.com/ontodia-org/ontodia
[4] https://github.com/DataFabricRus/ontology-fts