PlanetData: A European Network of Excellence on Large-Scale Data Management

Irini Fundulaki

Institute of Computer Science-FORTH
fundul@ics.forth.gr

Extended Abstract

PlanetData[1] is a Network of Excellence funded by the European Commission's Seventh Framework Programme FP7/2007-2013. The aim of PlanetData is to push forward the state-of-the-art in large-scale data management and its application to the creation of *useful*, *open data* sets. This is motivated by *i)* the increasing reliance of business on large public data *ii)* the uptake of open data principles in many vertical sectors *iii)* the need of research communities to make sense out of large amounts of scientific data, to *describe* and *expose* this data in ways that encourage and enable collaboration between interested parties. The PlanetData project is built around three objectives that together ensure the creation of a durable community made up of academic and industrial partners:

RESEARCH: To bring together approaches to large-scale data management from different disciplines in order to create *holistic solutions* to the challenges faced when dealing with big data.

DATA PROVISIONING AND MANAGEMENT: To apply the conducted research to real-world data sets, to derive best practices and guidelines for organisations that own or provide large amounts of data online. In this context, PlanetData undertakes the following specific tasks:

- the development of software to support large-scale data provisioning, made available via the PlanetData Lab, supporting relational, graph, and stream processing, for researchers to test and validate their techniques and the creation of definitive vocabularies for the description of datasets.
- the creation of a catalogue of datasets in vertical domains chosen for their high adoption potential and data management needs.
- the publication of guidelines and best practices for provisioning, such that available datasets can be more readily consumed by end-users and efficiently assembled into innovative products and services.

IMPACT: to collaborate with interested parties worldwide through the PlanetData Programs. This task can be achieved by:

- providing a medium through which the research results and empirical findings of the PlanetData network can be used to improve the education level related to large-scale data management in both academia and industry.

[1] http://www.planet-data.eu/

– bringing together researchers from disparate disciplines in order to form an integrated community that can support organisations in publishing their data in a way that is purposeful, thus addressing key challenges of large-scale data management.
– encouraging industrial uptake through standardisation, and strategic dissemination and networking events.

PlanetData has three *R&D Showcases* to present in the EU networking session:
• The *HealthCare Use Case Demo*[2] showcases the use of the *access control* technology developed in PlanetData to support access to sensitive medical data. In particular, access control enforcement techniques are used to provide selective exposure of information to various users/roles on patients' *Personal Health Records* (PHR), taking into account the *purpose* for which access is requested and the *consent(s)* that the patient has signed regarding said data. The prototype developed in PlanetData supports the basic functionalities of such a system, namely *i)* the ability of a patient to grant or deny access to specific parts of his PHR *(informed consent)* for specific purposes and entities *ii)* the ability of a patient to update his/her PHR information and *iii)* the capacity of external entities to query PHR information and retrieve only the part of the data they are entitled to (based on the patients' consent and the declared purpose).
• The motivation behind *Event Registry*[3] was to build a system helping publishers to *(a)* search across the media space, and *(b)* align their own information in the space of other published materials (annotation of news with pointers to events) thereby creating a *real-time global media observatory.* It consists of a series of software components, from data sourcing to visualisation and interoperable interfaces. It also includes a cross-lingual component connecting textual information that spans over 100 languages. It is developed as a prototype to support a standardisation working group at the IPTC level, a publishers' standardisation organisation whose aim is to release recommendations to collect, annotate and interoperate information on global events and story-lines across languages, domains and granularities.
• Last but not least, the *Smart Cities* PlanetData use case that exhibits a number of applications to demonstrate the use of different forms of Linked Data (*static* and *streaming*) as an *enabler of a homogeneous data layer* over the different data sources that contain relevant information. The static data sources are obtained from open data portals including information about *territory*, *statistical information* or the *census of business units* in the city, transformed using ETL processes that eventually generate RDF data. Dynamic data sources (e.g., transport, credit card transactions) cannot be transformed through ETL processes since the delay in the generation and storage of such data would incur too much overhead. These sources are exposed as virtual streaming RDF sources with the use of morph-streams technology developed in PlanetData which interfaces streaming REST APIs, WS services and CEP-based data producers, by means of declarative R2RML mappings.

[2] http://139.91.183.31:8084/pd-demo/login.jsp
[3] http://eventregistry.org/