3.1 Publishable summary

Introduction and Context of the DOPA Project

Today's SMEs often have to compete with global players when engaging in Big Data Analytic challenges. Those SMEs cannot afford huge investments in hardware, software, or in personnel to realize use case specific operations in the area of Information Extraction, Data Cleansing, and Data Mining. One barrier for SMEs is the need to resort to a wide range of domain specific datasets. Therefore, one goal of the DOPA project is to set up a network of different data pools with an ability of easy extensions. The need for a data supply chain environment with an underlying system for large-scale, data intensive experiments that also considers the diversity of the European region, is more than ever immensely significant for the competitiveness of European SMEs.

The DOPA project (http://www.dopa-project.eu/) will enable the access for external users, especially for SMEs, to a flexible and user-friendly data supply chain language, a wide range of complex operators for Big Data challenges and will assure an integration of different data pools. Only the linking of those data sources, which offer structured, semi-structured or unstructured data, enables the full coverage of the project's information acquisition.

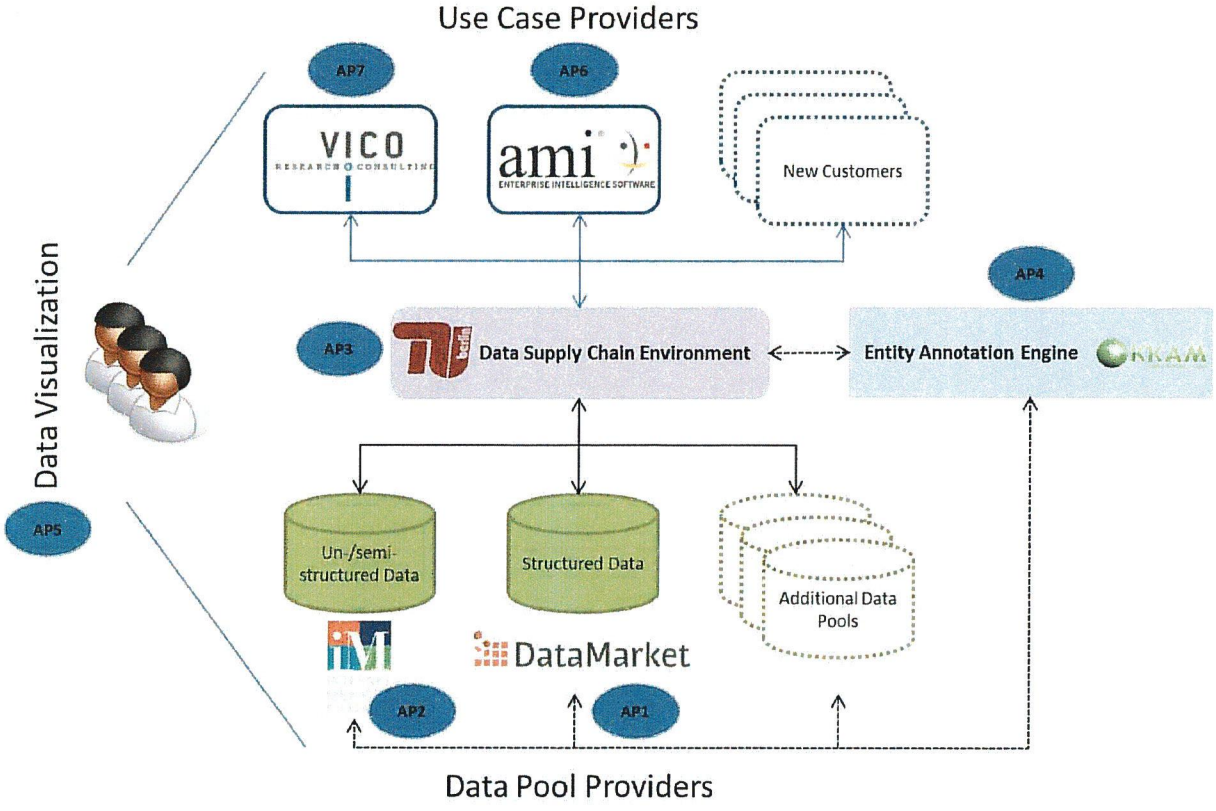

Figure 1 Overall Architecture

Objectives

DOPA – "Data Supply Chains for Pools, Services and Analytics in Economics and Finance" is an EU-funded, collaborative project carried out in the Seventh Framework Programme (FP7). The consortium consists of two data pool providers (DataMarket and Internet Memory Research), one partner responsible for the Data Linkage (OKKAM), two use case partners (AMI and VICO), and the Technical University Berlin (TUB) as consortium leader.

Main goal of the project is to enable the access to several data pools and complex Big Data Analytics for European SMEs and to encourage and support them to integrate user-defined operators in the existing system. Moreover, the DOPA project covers a wide range of semi-structured web data or structured fact data by our two data pool providers and is designed to integrate additional data pools. Therefore, a Data Linkage Service combines the input of participating data pool providers based on similar entities. The implementation of the Data Supply Chain Language enables a flexible and easy usage of the set of operators.

The intensive market analysis by our DOPA use case providers showcased a highly importance for market intelligence applications for investors, as well as for market agencies and advertisers. The project also focuses on creating risk management scenarios.

The specific objectives to be achieved when addressing the described challenges are:

- The preparation of a system to write user-specific queries due to the development of a Data Supply Chain environment on top of a massively-parallel execution system
- Capability of efficient processing of Big Data sets after the evaluation of methods for building up, scheduling, and executing Data Supply Chain operators
- Integration and connection of data pools via a Data Linkage Service
- Availability of use case specific operators with focus on scenarios in economics and finance
- Evaluation and potential extension of current data anonymization methods
- Exploitation and dissemination of project results to gain interest in the SME community

Summary of Results within first year

One of the first steps towards the specification of DOPA relevant use cases contained the analysis of the given data sources, its data structures and metadata. During the first part of period one, our data pool providers implemented different crawling strategies and a method to link different dimension values for fact data, followed by the evaluation of datasets. The analysis of the selected data sets was carried out in cooperation with the entire consortium. The extracted and offered data sets fulfil the needs for our use cases in market intelligence applications and in credit risk management scenarios. The credit risk management scenario is strongly supported by a German bank. It has given customer data for making first tests, which look promising. The evaluation of current Named Entity recognition tools led to the usage of the Stanford library, which assures best results for the data linkage process. The specification of the data supply chain language on top of an integrated massively parallel system has been finished and first use cases showcased the usability and flexibility of the declarative query language and the impact of the underlying massive-parallel execution platform. The data supply chain language is designed to enable a gradual extension of the operator sets by external users. A first draft of our interface and data visualization supports the interaction with potential customers. Feedback by partners' customers and during demo exhibitions at conferences, where first use cases could be presented, showed high interest in the solutions for Big Data Analytics in economics and finance contexts.

The project is completed by dissemination and exploitation plans to secure the progress of the project and the interactivity with potential customers. Our project website (http://www.dopa-project.eu/) informs potential customers about the progress of the project and demonstrates at the top of the homepage the current amount of datasets and active sources. Visitors of the website are kept up to date about DOPA related publications, press releases, as well as project meetings of the DOPA team.