The Trento Big Data Platform for Public Administration and Large Companies: Use cases and Opportunities

Ivan Bedini
Trento RISE
i.bedini@trentorise.eu

Benedikt Elser
Trento RISE
b.elser@trentorise.eu

Yannis Velegrakis
University of Trento
velgias@disi.unitn.eu

1. INTRODUCTION

Data analysis is used to drive almost every aspect of our modern society, including mobile services, retail manufacturing, financial services, life sciences, and physical sciences. Novel approaches are required to extract value from such data without omitting the opportunities enabling service innovation design. TrentoRISE, in collaboration with the dbTrento group [8] is developing an Information Infrastructures that will be used to promote services, research and development in the Trentino area and used to offer a better living to the citizens. As a starting point, a new generation platform has been designed to be the central place in the Trentino territory for the collection, cleaning, integration and analysis of BigData.

2. BIGDATA PLATFORM

Making the data public has no value if none can actually access, study and use that data. The BigData platform at Trento RISE aims at creating a flexible and powerful infrastructure that allows accessing an heterogeneous collection of large data sources to unveil new information of interest. It is an open source minded platform that has been architecturally built to be easily replicated and personalized following the "as a service" paradigm.

The platform, depicted in Figure 1, provides all necessary tools to manage the overall procedure processing the data, from their raw collection, until producing useful information to help in decision making. The first step captures related data from different sources through an in-house built *topical crawler*. The crawlers can be adapted to collect specific data for different use cases. These captured data are then filtered and stored into a Data Mart, where the heterogeneous data are integrated in a uniform and extensible knowledge base, and cleaned up for making them available to the BigData frameworks layer. The Presentation and Analysis layers access the generated output from the Big Data framework allowing further knowledge inference and presentations. This

Proceedings of the VLDB Endowment, Vol. 6, No. 11

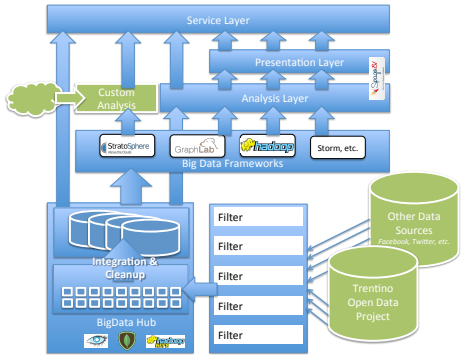

Figure 1: BigData Service Platform Architecture

is done through Business Intelligence tools, like SpagoBI [1], or other ad-hoc implementations, to provide a correlated view and the desired analytics. The final analysis is provided as a service. This service can be a specific analysis or even a generic platform for big data that a user can access for his/her own development need.

3. THE PLATFORM FEATURES

The Big Data Platform characterized by a number of features that are a result of a number of analysis studies [3] we have performed and distinguish it from other similar platforms.

[Opinions and sentimental analysis] Such algorithms depend on the quality of the provided dictionary. In our platform, we are able to create specific dictionaries for a provided topic. Using our topic driven crawler, that is able to identify content, that is related to any given topic. By turning that data into a dictionary, we can improve the outcome of our sentiment analysis algorithms.

[Profiling] It has been said many times that our actions define who we are. The same applies to social media. What we say and what we do is what defines our actual personality. One of the directions of the BigData group is to analyze the social media and extract characteristics of the persons that participating in them. This diminishes knowledge discover, yet, it is fundamentally different. In knowledge discovery from web or other contents, the information to be discovered is explicitly stated in the content of the pages, thus advanced analysis techniques can offer a large portion of the contained knowledge. However, the social data contain the actual reasons and knowledge hidden in the words and the

[1] http://www.spagoworld.org

actions described. For instance, an aggressive person will never declare it explicitly, but such a characteristic will be inferred implicitly from the tone, the words and the actions. To successfully generate profiles we are based on logical data analyser that allows us to find for every entity in the datasets its position in a multidimensional space, where the dimensions of the space are properties or opinions on various issues, and the position of the entity representation on that dimension is the polarity, opinion or strength of knowledge on the respective topic. Having generated the profiles, we are in position to discover not only individuals with certain characteristics but also groups of people that have certain properties collectively.

[**Goal Oriented Recommendations**] For every action or event that is currently observed in life there is some goal that is leading to the execution of that action. By analyzing the social media one has the opportunity to investigate what are the goals that are driving individuals and organizations into performing a series of actions, events, or statements and evaluate their effectiveness into achieving the desired goal. A large line of research is devoting into studying the above issues, especially in the context of social media. It is highly intriguing to manage to figure out why a person decides to perform an action or a sequence of actions in the first place and whether the resulting reactions meet his or her expectations. In addition to this, user reactions can be used as a form or evaluation for complex multidimensional tasks, that can then be translated into trends of improvements of the different dimensions. The above findings can be used to bring recommendation systems into a new level that can predict user preferences for items never seen before even if past evaluations are not encouraging their recommendation.

[**Flexible Integration**] Integration has traditionally been performed by experts that accurately decide how data structures in different systems relate to each other or model the same real world objects. Such decisions are supported entity identification tools that compute a likelihood score that such a case is true. If this score is high, then the structures are merged. Unfortunately, for Big Data this is not a feasible solution since the level that the likelihood is considered high is not clear, plus, different situations may require different likelihood thresholds. The Big Data group platform is able to provide flexible on-the-fly integration [5] that depending on the items of interests, decides what needs to be integrated and what now. This mode is more suitable for Big Data since no a-priori decisions need to be made, yet, the level of complexity is highly increasing, which makes the task particularly challenging.

[**Event Detection**] A timely notification of things, that happen for a specific topic is a desirable feature. One possible source for inferring these events is Twitter, an increasingly popular real time communication platform. By analyzing and clustering the flow and contents of Twitter messages, we can deduce hot topics from that data, rank and classify them. The spacial and temporal pattern of messages allow us to fine-tune that model. A correlation with other traditional data sources, such as news sites and blogs allows us to even lower the false positive rate. In this way we can develop a system, that informs about events, as they appear.

Other features, not mentioned here include query relaxation for improved user experience [6], data evolution [2], keyword search [1], semantic tecnologies [4], entity-based retrieval [7], e.a.

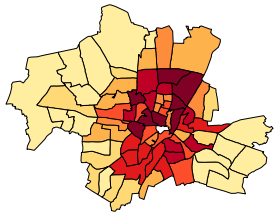

Figure 2: Distribution of available Rent-a-Cars across Munich.

4. USE CASES

[**Public administration**] The public sector needs to reduce IT costs while increasing services to the citizens, their efficiency, velocity and completeness. One of the most efficient way to obtain such results is to invest in new technologies models, such as cloud computing, platform for Intelligent Information Management and sharing the available data.

[**Transportation**] Traffic networks span an enormous graph across the globe. However operators of these networks often do not have the tools at hand, to effectively mine their backend data, such as the complete historical data set, of customers using their service. As an example, in Figure 2 we analyzed a dataset collected over two years from a car sharing service in Munich. It represents the mean availability of cars in regions of Munich at noon. This analysis can be easily used to improve customer satisfaction, by identifying bottlenecks.

[**Customer analysis**]. Customer loyalty comparison shopping is making retail competitive. The capacity of creating targeted discount programs, e-mail or social media correspondence seduce shoppers. Brand loyalty is created by attention to detail and targeted customer service. A complete customer view, based on access data, social media analysis enables this personalizion.

References

[1] S. Bergamaschi, F. Guerra, M. Interlandi, R. Trillo-Lado, and Y. Velegrakis. Quest: A keyword search system for relational data based on semantic and machine learning techniques. *PVLDB*, 6, 2013.

[2] S. Bykau, F. Rizzolo, and Y. Velegrakis. A query answering system for data with evolution relationships. In *SIGMOD Conference*, pages 989–992, 2013.

[3] B. Elser and A. Montresor. An evaluation study of bigdata frameworks for graph processing. In *IIEEE BigData 2013)*, 2013.

[4] O. Hassanzadeh, A. Kementsietsidis, and Y. Velegrakis. Data management issues on the semantic web. In *ICDE*, pages 1204–1206, 2012.

[5] E. Ioannou, N. Rassadko, and Y. Velegrakis. On generating benchmark data for entity matching. *J. Data Semantics*, 2(1):37–56, 2013.

[6] D. Motin, A. Marascu, S. B. Roy, G. Das, T. Palpanas, and Y. Velegrakis. A probabilistic optimization framework for the empty-answer problem. *PVLDB*, 6, 2013.

[7] D. Mottin, T. Palpanas, and Y. Velegrakis. Entity ranking using click-log information. *Intelligent Data Analysis*, 17(5), 2013.

[8] T. Palpanas and Y. Velegrakis. dbtrento: the data and information management group at the university of trento. *SIGMOD Record*, 41(3):28–33, 2012.