International workshop on Data Mining of Iot Systems (DaMIS)

A service oriented framework for analysing social network activities

Angelo Chianese[a], Francesco Piccialli[a,*]

[a] *University of Naples "Federico II, Department of Electrical Engineering and Information Technologies, Italy*

Abstract

Analysing and monitoring Social Networking activities raise multiple challenges for the evolution of Service Oriented Systems Engineering. This is particularly evident for event detection in social networks and, more in general, for large-scale Social Analytics, which require continuous processing of data. In this paper we present a service oriented framework exploring effective ways to leverage the opportunities coming from innovations and evolutions in computational power, storage, and infrastructures, with particular focus on modern architectures including in-memory database technology, in-database computation, massive parallel processing, Open Data Services, and scalability with multi-node clusters in Cloud. A prototype of this system was experimented in the contest of a specific kind of social event, an art exhibition of sculptures, where the system collected and analyzed in real-time the tweets issued in an entire region, including exhibition sites, and continuously updated analytical dashboards placed in one of the exhibition rooms.

1. Introduction

Social Networking activities generate huge volumes of data, whose analysis is becoming more and more important for both profit and non-profit organizations; at the same time, social networks are used in proactive manners by organizations and even individuals for multiple objectives, including internal and external communication, coordination, information, recommendation, advertising, promotion campaigns, and organization of social events [12]. In particular, the Twitter microblog features provide great opportunities and pose specific challenges for large-scale analytics, which stem from multiple aspects: real-time nature of messages, often based on volatile emotional reactions; fast viral propagation over the network; high peaks of interest concentrated in short timeframes; short lifetime of many hashtags; short and noisy text including non textual contents images, emoticons, other symbols, and so on. In this context, the need of advanced services requires evolutions in multiple aspects of Service Oriented Systems Engineering. One of the most important aspects is how to effectively leverage the opportunities coming from innovations and evolutions in computational power, storage, and infrastructures; our research is focused on modern architectures including in-memory database technology, in-database computation, MPP(Massive Parallel Processing), Open Data Services, scalability with multi-node clusters in Cloud infrastructures, with particular respect to the SAP HANA Cloud Plat-

* Corresponding author. Tel.: +39 0812532035.
 E-mail address: francesco.piccialli@unina.it

form(HCP)[1]. In the last five years, the research in Service Oriented Computing has evolved in multiple directions and areas. In[3] a flexible service oriented architecture is proposed for planning and decision support in environmental crisis management, which includes real time middleware that is responsible for selecting the most appropriate method of the available geospatial content and service, in order to satisfy the QoS requirements of users and/or application. A conceptual model that supports evaluation of theory and methods for the management of service-oriented decision support systems is proposed in[4], which distinguishes Data-as-a-service (DaaS), Information-as-a-service (information on demand), Models-as-services (MaaS), Analytics-as-a-service (AaaS); this study emphasizes the importance of social media analytics, Continuous Analytics as a service, and the need of a shared infrastructure, with local processing power to perform real-time data analysis, in-memory grid computations and parallel transaction and event processing. In this perspective, authors in[11] studies the users behaviors with regard to communication, movement, and consumption based on large user data sets. In the area of Social Networking Analytics, it is well recognized the outstanding importance of real-time social networking services, with emphasis on those based on microblogs like twitter, which have been used in multiple domains including event detection in emergency and disaster relief[8,9,10], healthcare, cultural heritage, tourism, marketing, social network building and enhancement.

The main scenario of this research is supporting the objectives of DATABENC, a High Tech Consortium for Cultural Assets in Italy, involving multiple Universities, cultural organizations and industries located in the Campania region. In particular, the CHIS project (Cultural Heritage Information System) includes the design, prototyping and experimentation of Service Oriented Computing solutions, which comprise the combination of heterogeneous information sources, knowledge bases, Big Data processing, and analytical services[14,15,16]. The rest of this paper is organized as follows: Section 3 summarizes the main features of the HCP and its position with respect to Microservices[2]; Section 3 presents the design principles and implementation strategies of our Service System, Section 4 reports on some Experiments performed in the context of the CHIS project, and finally Section 5 provides some conclusions.

2. The Cloud Platform

2.1. Basic Technologies

The basic SAP HANA technology is a computer appliance (discrete hardware device with integrated software /firmware), which is specifically designed to implement an in-memory transactional-analytical Database Management System, i.e. conceived to support both transactional workloads and analytic workloads. Combining OLAP and OTLP into one database, this system creates a unified view on data from transaction, analysis, decision, and planning systems, and allows for in-database real-time analytics and data mining. Processing is performed with a massively parallel (distributed) data management system that runs fully in main memory, exploiting multicore units as well as multi-node clusters. The HANA in-memory database server supports both columnar storage and row storage. The first advantage of columnar storage is the ability to facilitate analytic queries, which typically incorporate the use of grouping, ranking, sorting, and aggregating operations; it provides a better mechanism for querying large quantities of specific attributes because data is quickly pinpointed in each column, without the need to scan every row in the table. In addition, a columnar table effectively compresses the data stored in each column of the table. Because there are multiple CPU cores available, the nature of the column store structure makes parallelization possible: since data is already vertically partitioned into columns, multiple columns can be searched and aggregated at the same time by different cores. However, there are scenarios where row storage and SQL row engine are more suitable: for example, the processing of transactions where the entire row of the record is needed to satisfy the application, writing data one record at a time is required while aggregations arent required, and the table will be used for returning one record quite often. The SAP HANA appliance manages periodic snapshots of memory to disk automatically. It incorporates a fast storage layer to manage database logging; typically, the logging layer is managed on Solid State Disk (SSD) arrays or NAND flash cards. The SAP HANA DB manages data in a multicore architecture for data distribution across all cores to maximize RAM locality using scale-out (horizontally) and scale-up (vertically) functionality; in the scale-out scenario, the SAP HANA database scales beyond a single server by allowing multiple servers in one cluster. The SAP Hana Cloud Platform can combine the power of many appliance clusters, each one equipped with 2TB (or even more TB) of RAM, in order to offer cloud-equivalent Petabytes of RAM. This platform supports high-performance processing of heterogeneous types of data in multiple functional areas, in particular:

- in-Database Geospatial processing is supported with a spatially enabled database and libraries, which uses the in-memory technology to store, process, manipulate, share and retrieve spatial data directly in the database; this includes the processing of spatial vector data with spatial analytic functions and attributes (measurements of distance, surface, area, perimeter, volume; relationships such as intersects, contains; transform operators, etc.); this system implements the ISO/IEC 13249-3 standard and Open Geospatial Consortium (SQL/MM standard) and is enforced by multiple GIS partnerships (ESRI, Pitney Bowes, SAPNS2, HERE, Fichtner, Critigen et al.)
- in-Database Text Processing is performed with text analysis, fulltext indexing, fulltext search via SQL, search models, search rule sets, text mining and sentiment analysis (aka voice of the customer); this system supports more than 30 languages providing functions to classify entities (people, companies, things, etc.), identify domain facts (sentiments, topics, request). It is possible to combine the columnar store with another store and content management system(CMIS) for unstructured data sources, thus enabling the full processing of structured and unstructured data on the same platform.
- in-Database Data Mining/Predictive Analytics are supported by the Predictive Analysis Library: this provides a very large set of algorithms for predictive analytics across multiple data types and sources.

2.2. Code-to-Data

The concepts of in-database processing, and in-DB analytics in particular, stem from a general shift from traditional Data to Code programming models to the innovative Code to Data model:

- in traditional models, most of the data-intensive and control-flow code is encapsulated outside of the database layer, i.e in application tiers or others; this induces massive data copies which create bottlenecks;
- in the code-to-data model, a great part of code lies and works in the context of the database layer; this eliminates most of the data transfer between database and application tiers, which is reduced to a minimum result set. In addition, executing calculations in the database layer allows to get the maximum benefit from SAP HANA high-performance oriented features such as fast column operations, query optimization, and parallel execution.

In SAP HANA the basic support for data-intensive in-database processing is the SQLScript language, an Extension of ANSI standard SQL, which includes table-driven system for the configuration of specific engines and analysis types, user defined functions and stored procedures, as well as built-in libraries such as Application Function Library(AFL), Business Function Library(BFL), and Predictive Analysis Library(PAL). The in-DB strategy for data-intensive processing is combined with the push control flow down strategy:

- UI Rendering with presentation logic is completely in the Client (web browser, mobile apps);
- Control-flow / procedural logic is server-side
- all code artifacts are stored in the SAP HANA Repository.

In the context of SAP HANA Extended Application Services(XS), the main Control Flow Technologies include Server-Side JavaScript (XSJS), OData (XSODATA), and XMLA. The XSJS interacts directly with the SAP HANA XS runtime environment. Server Side JavaScript Libraries allow to perform repetitive tasks, the reuse of program elements as well as the reuse of 3rd party/open source code. XSJS code can be imported into other XSJS programs and be called from other XSJS programs.

2.3. Open Data and User Interfaces

The OData (Open Data Protocol) defines an abstract data model and a REST (Representational State Transfer) protocol that let any client access information exposed by any data source. Data is transferred over HTTP using either the JSON (JavaScript) format or Atom (XML). The OData also provides a uniform way to represent metadata about the data, allowing computers to know more about the type system, relationships and structure of the data. Even a simple web browser can view the data exposed through an OData service. The OData protocol is just like HTTP and can be used to perform CRUD operations (Create, Read, Update, Delete); constraints can be specified for exposing read-only endpoints. OData provides an entire query language directly in the URL; this means that by changing

the URL, the data returned from an OData feed (Data Source) also changes. Being able to control what data are returned from the consumer side means that the OData feed has complete control over what parts of the content to use. An OData service exposes data stored in database tables or views as OData collections for analysis and display by client applications; the client can be a web browser, or HTML5 application or any other application which supports OData. The OData Client Libraries are pre-built libraries to request OData and display results. The User Interface model assumes a client via HTTP with application services in the Cloud; the SAP standard UI technologies, Fiori and UI5, follow this mode. In particular, SAPUI5 includes an enterprise-ready HTML5 rendering library, CSS3, Client-side JavaScript, which targets multiple devices: smartphone, tablet, and desktop. SAPUI5 supports the Model View Controller pattern(MVC), and especially in a SOA context the typical way to populate the Model is based on OData services.

2.4. HCP and Microservices

The SAP HANA Cloud Platform(HCP) offers a Platform as a Service (PaaS), which enables Software as a Service(SaaS) and Leverages an Infrastructure as a Service(IaaS). Many of its features appear suited to support modern architectural principles and styles, such as microservices[2]. First, microservices prefer letting each service manage its own database, and one option for supporting this principle is using multi-tenant database containers. Multi-tenancy is natively supported in the HANA in-memory DBMS: all tenant databases in the same system share the same system resources (memory and CPU Cores). However, each tenant database is fully isolated with its own database users, catalog, repository, persistence (data files and log files) and database services. Another major benefit of multitenant database containers feature is that it simplifies development and deployment of secure, multi-tenant cloud-based applications. In a scale out scenario, a tenant database can span across multiple SAP HANA nodes; this means, the size of the tenant database is not limited by the size of a single SAP HANA node. One of the main characteristics of a microservice architecture is also the provisioning of a RESTful API, and SAP HANA OData Services provide a fully fledged, flexible system in this respect[13].

3. The Service System

3.1. Overview

The S-InTime system (see 1) is conceived for fast integration and analysis of data coming from multiple domains, including catalogues and other semantic resources on cultural assets managed by both non-profit and profit organizations, their transactional data, as well as data gathered from Internet of Things and Social Media.
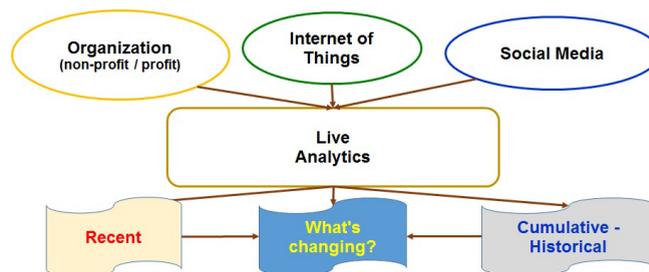

Fig. 1. System Overview.

The concept of live analytics entails not only the ability to support highly interactive analytical services with responsive user interfaces, but also real-time data processing with the capability of producing early-timely advices on new phenomena, which can be relevant for the objectives of one or more types of users. For this purpose, the system compares information obtained from cumulative-historical analysis over entire periods with information caught in recent / incoming data. In the following paragraphs, we explain how we addressed the main requirements of this kind

Table 1. Montly counts of CH-related tweets with sentiments and new entries

MONTH	CH SENTIM	CH NEW ENTRY
2015-12	54	66
2016-01	308	107
2016-02	1071	97
2016-03	885	35
2016-04	933	29
2016-05	957	30
2016-06	1007	16
2016-07	394	9

of system by exploiting the opportunities and the peculiar architectural paradigms supported by HCP, with examples of Services for Social Networking data.

4. Preliminary Experiments

We applied the architectural and programming principles of the system in the implementation of a set of Twitter Analytics services. The system extensively monitored the presence of Sentiments in CH-related tweets: Table 1 reports the monthly counts of CH-related tweets including sentiments, and the turtle Figure 2 shows the cumulated percentage over the entire period of observation, while Figure 2 is referred to the most recent day, revealing a (possibly temporary) increase of interest with respect to the average long-term trend.
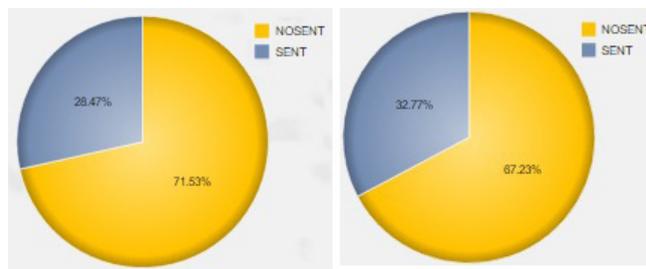

Fig. 2. On the left the presence of Sentiments in CH-related tweets-entire period, on the right the presence of Sentiments in CH-related tweets-Last Day
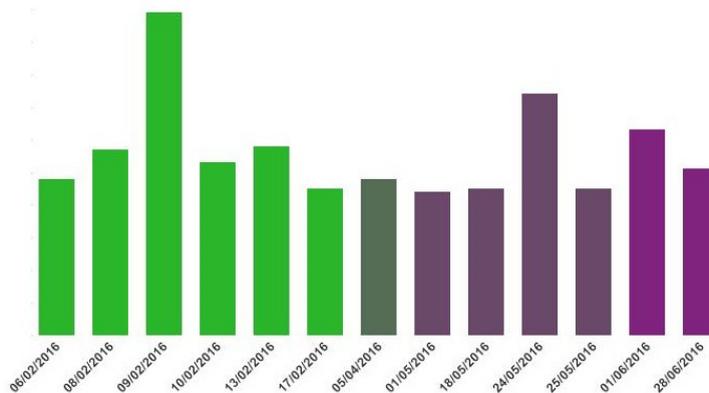

Fig. 3. Top CH-Sentiment-prone days

We also monitored the trend of top CH-Sentiment-prone days, which are dates where the tweets including sentiments on CH-related issues exceeded 45% of the average daily counts: Figure 3 reveals that sentimental peaks occur in different months, while within a month they are either concentrated in 1-2 consecutive weeks or a few almost isolated dates. More investigation is needed to measure the correlation of this phenomenon with other cultural / touristic events or communication campaigns in the same dates.

References

1. https://hcp.sap.com
2. M.Fowler, J.Lewis, Microservices a definition of this new architectural term, http://martinfowler.com/articles/microservices.html, 25/03/2014
3. V.Vescoukis, N.Doulamis, S. Karagiorgou, A service oriented architecture for decision support systems in environmental crisis management, Future Generation Computer Systems, Volume 28, Issue 3, March 2012, pp 593-604 doi:10.1016/j.future.2011.03.010
4. H. Demirkan, D. Delen, Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decision Support Systems, Volume 55, Issue 1, April 2013, Pages 412-421, doi:10.1016/j.dss.2012.05.048
5. Z.Sun, K.D. Strang, J.Yearwood, Analytics service oriented architecture for enterprise information systems, iiWAS '14 Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, Pages 508-516, ACM New York, NY, USA 2014, doi:10.1145/2684200.2684358
6. V.Alkkiomaki, K.Smolander, Anatomy of one service-oriented architecture implementation and reasons behind low service reuse, Service Oriented Computing and Applications June 2016, Volume 10, Issue 2, pp 207-220, DOI 10.1007/s11761-015-0181-2
7. H.Lampesberger, Technologies for Web and cloud service interaction: a survey, Service oriented computing and applications Springer, June 2016, Volume 10, Issue 2, pp 71-110, DOI 10.1007/s11761-015-0174-1
8. Chiu D, Lin D, Kafeza E, Wang M, Hu H, Hu H, Zhuang Y (2010) Alert based disaster notification and resource allocation. Inf Syst Front 12(1):29-47
9. S.Kumar, G.Barbier, M.Abbasi, H.Liu,TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief, Fifth International AAAI Conference on Weblogs and Social Media, 2011 Association for the Advancement of Artificial Intelligence (www.aaai.org).
10. T.Sakaki, M.Okazaki, Y.Matsuo, Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development, IEEE Transactions on Knowledge and Data Engineering (Volume:25 , Issue: 4), April 2013, DOI 10.1109/TKDE.2012.29
11. Zhenhua Wang, Lai Tu, Zhe Guo, Laurence T. Yang, Benxiong Huanga, Analysis of user behaviors by mining large network data sets, Future Generation Computer Systems, 37, pp. 429-437, 2014
12. Duc T. Nguyen, Jai E. Jung, Real-time event detection for online behavioral analysis of big social data Future Generation Computer Systems, http://dx.doi.org/10.1016/j.future.2016.04.012, 2016
13. SAP HANA OData Servicehttp://www.saphanacentral.com/p/sap-hana-odata-service.html
14. Chianese, A., Piccialli, F., A smart system to manage the context evolution in the Cultural Heritage domain, Computers & Electrical Engineering, 2016, doi:10.1016/j.compeleceng.2016.02.008
15. S. Cuomo, P. De Michele, F. Piccialli, A. Galletti: A Cultural Heritage case study of visitor experiences shared on a Social Network Proc. of 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC, pp. 539-544, (2015)
16. Chianese, A., Piccialli, F. and Riccio, G, Designing a smart multisensor framework based on beaglebone black board, Lecture Notes in Electrical Engineering, Vol. 330, pp-391-397, 2015
17. Matthias Steiner, Microservices on SAP HANA Cloud Platform, http://scn.sap.com/community/developer-center/cloud-platform/blog/2015/11/24/microservices-on-hcp
18. DEV300: Architecture Guidelines for Microservices Based on SAP HANA Cloud Platform http://www.slideshare.net/saphcp/dev300-architecture-guidelines-for-microservices-based-on-sap-hana-cloud-platform, SAP TechEd Nov 24, 2015