User-Centered Event Data Modelling and Analytics

Stefano Valtolina, Marco Mesiti and
Luca Ferrari

Department of Computer Science
Università degli Studi di Milano, Italy

{valtolin, mesiti,
ferrari}@di.unimi.it

Koji Zettsu and
Minh S. Dao

Universal Communication Research Institute, NICT Kyoto, Japan

{zettsu,dao.minhson}
@nict.go.jp

Abstract. Conventional data analytics platforms are not adequate to be applied in the management of emergency situations. The 3V the usually characterize big data (volume, variety, velocity) along with the issue of integrating information coming from heterogeneous networks require the development of new systems. In this paper we provide the design of a data analytics platform that we are developing around the concept of event, that is simple or complex data stream gathered from physical and social sensors that are encapsulated with contextual information (space, time, thematics).

Keywords: Event ETL, Event Datawarehouse, Event OLAP, Service-Controlled Networking

1 Introduction

Nowadays we are witnesses of the proliferations of different sensor devices able to produce heterogeneous types of data (textual, visual, audio, and other rich multimedia formats) that can be profitable used for detecting, handling and advising people of the verification of emergency events such as disasters due to natural phenomena (like flooding, storming, extreme temperatures etc.). Beside the physical sensors, able to detect data about physical phenomena (like temperature, humidity, wind, rain, pressure, level of see water), there is a proliferation of social sensors able to collect data from people (like twitter data, traffic information, train or flight schedule) [1]. These events are characterized both from the temporal, spatial and thematic dimensions that can be exploited from one side for the identification of the useful information needed to face a given emergency event and from another side for the analysis and forecast of useful activities to carry out for alerting people and rescue victims from the places of the disaster. Conventional data analytics platforms cannot be exploited profitably for handling this kind of data and new advanced architectures should be developed for several reasons. First, the sensors (both physical and social) are located in different networks and made available by different institutes, agencies and NPOs. In this context, network configuration, sensor detection and discovery are difficult issues to be solved. Moreover, sensors and the data they produce should be handled in real time in order to be properly

Proc. of Third International Workshop on Cultures of Participation in the Digital Age - CoPDA 2015
Madrid (Spain), May 26th, 2015 (published at http://ceur-ws.org).

elaborated during the emergency event. Therefore, scalable and efficient solutions should be devised that can be applied on-line. ETL (extract, transform and load) solutions, usually applied off-line, need to be revised and applied on-line for feeding the DW (data warehouse) with fresh and timely data. Finally, an user-friendly environment has been conceived for properly helping the user in the different phases of the analysis processes (sensor discovery, feeding, and knowledge inference).

In this paper we propose a novel OLAP Analytics Platform at the level of event for analyzing multi-dimensional relations between multiple data streams based on event attributes (spatial, temporal and thematic attributes). The output would be spatial-temporal-thematic aggregates of multiple data streams (e.g., physical and social sensing data streams) representing complex events (e.g., social response to natural disaster) organized at different spatio-temporal granularities..

In Section 2 the new architecture will be illustrated and compared with respect to conventional Data Analytic Platform. Section 3 describes the event data model tailored for handling multidimensional complex events according to a multigranular spatio-temporal-thematic (STT) data model. Section 4 starts with discussing the description of the services that should be developed for the Event ETL, for then describing the user interface developed for creating a workflow based on the combinations of different data streams. Finally, we provide a description of the current efforts and detail for developing the Event OLAP interface, which aim is to support users in analyzing data coming from different sensors and detect possible correlation and significant events.

2 From Conventional to Event based Analytic Platform

In conventional OLAP system there is a distinction between ETL operations and OLAP operations. The former are used for feeding the DW, whereas the latter are used to query the DW once the cube has been defined. Moreover, feeding the DW is considered a pre-processing phase that is taken off-line with the purpose to solve many issues (extracting, cleaning and riconciliate heterogeneous data and denormalize data) in order to make somehow simpler the execution of the OLAP operations. Stream DW have been also proposed for handling information produced in streams and approaches based on sliding windows have been introduced to manage continuous data [2].

Recently the term “active” or “real-time” DW [3] was coined to capture the need for a DW containing data as fresh as possible. In this context the periodic population of a DW is considered outdated, and new OLAP and DSS services are required in order to work on-line, therefore they should handle updates in order to change the data representation and the machine learning models developed on top of them. In Real-time DW, data are loaded in real time from the OLTP system into data warehouse, providing a convenient way for user to real-timely read the data information and make tactical decisions. In this context, the ETL and OLAP operations must be tightly connected in order to make the process possible and their implementations need to be quite efficient and scalable. In our work, we wish to develop real-time DW that is developed around the concept of event as described before.

First of all, we need to point out the adoption of a different data model. Indeed, conventional DWs adopt the relational model for the representation of the information on the storage. This is not adequate in the Event warehouse because of the heterogeneity of the data formats used and also the fact that information is stored at different granularities (ranging from simple raw values of temperature to structured and complex data like the objects extracted from a picture, or the mood inferred from tweet data with their level of trustiness). Second, conventional systems process stored data, whereas in the Event warehouse streams of data should be processed and collected and analyses through temporal windows. These streams can be stored in the DW or only aggregated information are stored. Orthogonally, contextual information can be associated with the stream by means of machine learning algorithms and exploited for extracting knowledge. For what concern the implementation of the architecture and of the required operations, we are considering the use of cloud-based architectures because we need to guarantee high throughput and a low end-to-end latency from the system, despite possible fluctuations in the workload. We are currently evaluating several architectures like Apache S4 [4], D-Streams [5], Storm [6] and StreamCloud [7]. Apache S4 [4] and Storm [6] allow the design of parallel applications for dealing with stream of data.

3 STT Event Data Model

In order to represent events, we consider the set of basic and structured (like list and records) types whose values are represented thorough a JSON-like notation. These types allow to represent different aggregation of values without imposing the restrictions of the relational model and so be able to handle a great variety of datatypes. Moreover, we consider spatial and temporal types at different granularities. Temporal granularities include *seconds*, *minutes*, *days* with the usual meaning adopted in the Gregorian calendar, whereas, *meters*, *kilometers*, feet, *yards*, *provinces* and *countries* are examples of spatial granularities. Different granularities provide different partitions of their domains because of the diverse relationships that can exist among granularities, depending on the inclusion and the overlapping of granules [8].

For example, temporal granularity *seconds* is finer than *minutes*, and granularity *months* is finer than *years*. Likewise, spatial granularity *provinces* is finer than *regions*. Thematics are also considered for associating to a given value a semantic annotation. For example, the annotation `HR,MR,LR` can be associated to the real number representing the degree (high, medium, low) of rain that is precipitated in a given area. Several thematics can be identified depending on the context where the datum is acquired or processed. Thematics can indeed be inferred by machine learning algorithms. Relying on the concepts of temporal and spatial granularities, we present the concept of *event*, that is a value associated with a spatial object at a given time according to given thematics.   Therefore, an *event* is a value represented at a given spatio-temporal granularity for which thematic information is added. Relying on the concept of events, we can characterize an event stream that a source can produce. A source can be sensor (either physical or virtual) or a service (for example by aggregating sensors/services streams).

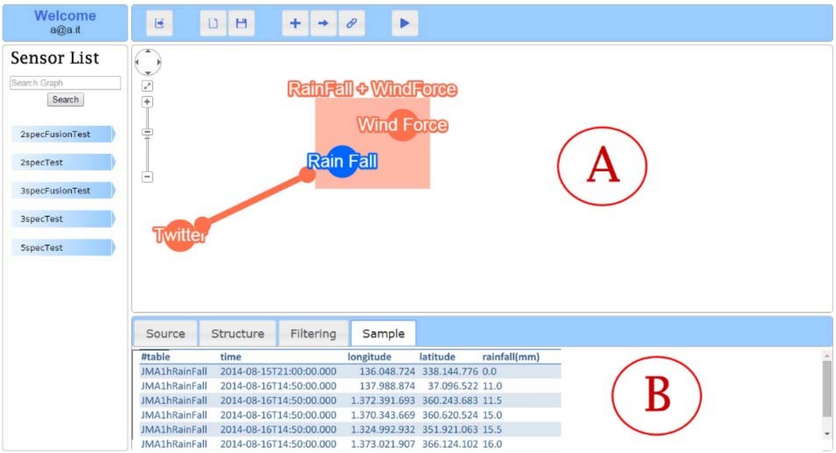
4 Event Management

At the bottom of our architecture, a middleware provides description of the available sensors and offers several services including service discovery, service monitoring, execution control, and service message exchanges. These services are exploited for executing `Event ETL` operators over a programmable network (like SDN) efficiently and effectively (e.g., filter by in-network data processing). Currently we developed three kinds of operators that are relevant in our context: conversion, merging and connecting operators. Conversion operators are used for changing the spatio-temporal-thematics granularity of an event stream. Merging operators are used for merging events at the same spatio-temporal granularity in a single data structure. Connecting operators are used to link a data source to another by applying a SQL joint-like operator. Following, for reasons of space the paper does not describe in detail the implementation of these operators focusing only on the graphical visual interface using which the user can identify the sensors and apply the operators for converting, merging and combining their data stream. Finally, we provide the description of the Event OLAP system used for analyzing the data and for extracting knowledge.

4.1 Visual Event ETL

ETL operations have been proposed in different contexts depending on the kinds of data to handle (structured and semi-structured). In [2,3,9] there is a good treatment at the conceptual level for feeding a DW. Moreover, in [10] there are approaches for the semi-automatic generation of ETL operations depending on the user needs and context of use. ETL operations are usually coupled with graphical visual data-flow for helping the user in the identification of the original data sources, the application of the operations for extracting, cleaning, transforming and combining their data. Once the ETL specification is completed, some strategies are proposed for the optimization of the data-flow and for the efficient execution of the loading schedule. These approaches have been mainly developed for producing relational data to feed conventional DW system. In [11] an approach has been presented for feeding arbitrary target sources (either relational or based on a noSql system).

In our context, we designed a Web environment where the users have the possibility to drag and drop different sensor data sources and visually apply on them a set of operations. This application offers an engine and graphical environment for data transformation and mashup. As depicted in **Fig. 1** (section A), this mash-up consists of a user interface that contextually displays icons of data sources or operations in order to link, filter or merge data coming from different sensors. It is based on the idea of providing a visual workflow generator for letting the end user creating aggregation, filtering, and porting of data originated by sources. An advanced use of such visual paradigm allows the end users to have an online generation of sample data coming from the data sources dragged-and-dropped on the canvas, or as result of the operations carried out on them (see **Fig. 1** section B). This End User development strategy offers a solution able to gather information from across the net and trigger specific filters and operations. This solution enables a very simple and easy to learn solution based on the definition of sources to use for collecting data and on the possibility to apply converting, merging and connecting operations in a visual way.



#table	time	longitude	latitude	rainfall(mm)
JMA1hRainFall	2014-08-15T21:00:00.000	136.048.724	338.144.776	0.0
JMA1hRainFall	2014-08-16T14:50:00.000	137.988.874	37.096.522	11.0
JMA1hRainFall	2014-08-16T14:50:00.000	1.372.391.693	360.243.683	11.5
JMA1hRainFall	2014-08-16T14:50:00.000	1.370.343.669	360.620.524	15.0
JMA1hRainFall	2014-08-16T14:50:00.000	1.324.992.932	351.921.063	15.5
JMA1hRainFall	2014-08-16T14:50:00.000	1.373.021.907	366.124.102	16.0

Fig. 1. The image present a screenshot of the User interface of the Visual Event ETL. In section A the image illustrates the canvas used by the user to drag and drop data sources. The social sensor “Twitter” is linked to the merge of two different sensors: “RainFall” and ”WindForce”. In Section B is presented a set of sample data coming from the selected sensor: The “RainFall”. In this part of the interface the user can set up filters on the data sources or conditions on the operations

Future activates aim at developing machine-learning algorithms for exploiting the possible operations that users can apply to a set of data sources or for providing them useful predictions of what can happen by integrating the selected data.

4.2 Event OLAP

Traditional OLAP front-ends, designed primarily to support routine reporting and analysis and offering visualization merely for expressive presentation of the data [12], are not suitable in the context of Event OLAP.

Event OLAP aims at providing a much more powerful data analysis environment, accessible through an intuitive user interface for quickly analyzing STT events. Event OLAP delivers a web-based interactive environment that allows non-technical users but experts of the analysis domain, to explore data in real-time by slicing and dicing, pivoting, filtering, and summarizing them in an intuitive way. As depicted in **Fig. 2**, it is based on a 3D map-based visualization through which carried out analysis and monitoring of trajectories (that is, spatio-temporal movements) of data streams that are part of an event (e.g. a thunderstorm). The aim of this environment is to provide users with a visualization of the data streams coming from different sensors in order to capture possible visual correlations among data. The visual perception of these correlations can be used for integrating correlations that can be detected in automatic way by applying data mining techniques on the data streams. Visual display of events' trajectories and collection of movement statistics can pro-vide useful indicators about how an event stream can influence another event streams. Moreover, current efforts are addressed to endow the web application with the possible to modify the workflow designed by using the Visual Event ETL in order to visualize the related changes in real time on the on the Event OLAP system.

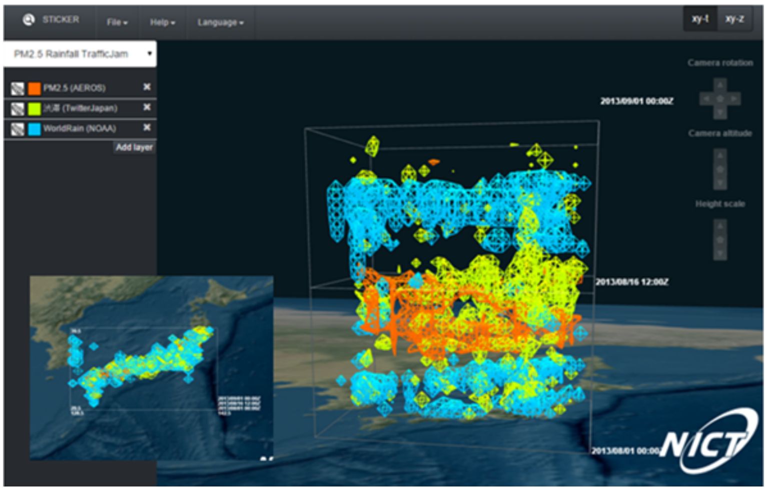


Fig. 2. The image presents a screen shot of the event OLAP web application. The image illustrates as in a specific range of time (2013, August) the concurrence of streams coming from the sensor "PM2,5" that collects data of over a specific level of fine dusts threshold and "Twitter" that collects tweet containing the keyword "Asthma" is extremely high. This event can bring the user to formulate specific hypothesis about the correlation highlighted in the visual interface.

Other activities aim at extending this web application with a location intelligence visualization strategy to identify patterns and trends by seeing and analyzing data in a map view with spatial analysis tools such as thematic maps and spatial statistics. This location intelligence service will help to find data by using spatial relationships to filter relevant data. Moreover, a temporal condition of this location intelligence service is applied for providing spatio-temporal clusters, simulation and visualization, map animation and movement tracking. To finish, in order to take into account the social aspect of the events collected by the Event OLAP, it will be endowed with a set of functionalities for creating a social network of users that will promote the creation of communities around each event. This social component exploiting ad-hoc computing techniques is able to study social network dynamics and to promote crowd-sourcing analysis for new and meaningful uses of data.

Proc. of Third International Workshop on Cultures of Participation in the Digital Age - CoPDA 2015
Madrid (Spain), May 26th, 2015 (published at http://ceur-ws.org).

5 Concluding Remarks

In this paper we have presented the architecture of an Event OLAP system specifically conceived for the management of emergency situation. We have detailed some peculiarity of the system and compare it with conventional OLAP systems and current research efforts. We are currently working on the formal specification of the needed operations at the ETL, DW and OLAP levels. Moreover, we are choosing the Cloud architecture in which the different services will be implemented and tested. For the testing we will exploit the collection of event data made available by NICT.

6 References

1. M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing Social Media Messages in Mass Emergency: A Survey eprint arXiv:1407.7071, 2014.
2. M. Gorawski and A. Gorawska. Research on the stream ETL process. In Int'l Conf. on Beyond Databases, Architectures, and Structures, 61-71, 2014.
3. H. Zhou, D. Yang, and Y. Xu. An ETL strategy for real-time data warehouse. In Practical Applications of Intelligent Systems, volume 124 of Advances in Intelligent and Soft Computing, 329-336. Springer, 2012.
4. L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: distributed stream computing platform. In IEEE Int'l Conf. on Data Mining Workshops, 170-177, 2010.
5. M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica. Discretized streams: fault-tolerant streaming computation at scale. In ACM Symposium on Operating Systems Principles, 423-438, 2013.
6. N. Marz. Storm: Distributed and fault-tolerant realtime computation, 2012.
7. V. Gulisano, R. Jiménez-Peris, M. Patino-Martinez, C. Soriente, and P. Valduriez. Streamcloud: An elastic and scalable data streaming system. IEEE Trans. Parallel Distrib. Syst., 23(12):2351-2365, 2012.
8. E. Camossi, E. Bertino, M. Mesiti, and G. Guerrini. Handling expiration of multigranular temporal objects. J. Log. Comput., 14(1):23-50, 2004.
9. P. Vassiliadis, A. Simitsis, and S. Skiadopoulos. Conceptual modeling for ETL processes. In Proc. Int'l Workshop on Data Warehousing and OLAP, 14-21, 2002.
10. V. Theodorou, A. Abellò, M. Thiele, and W. Lehner. A framework for user-centered declarative etl. In Proc. Int'l Workshop on Data Warehousing and OLAP,67-70, 2014.
11. M. Mesiti and S. Valtolina. Towards a user-friendly loading system for the analysis of big data in the internet of things. In IEEE Computer Software and Applications Conference - workshops, 2014, 312-317, 2014.
12. M. Scholl and S. Mansmann. Visual on-line analytical processing (OLAP). In Encyclopedia of Database Systems, 3388-3395. Springer US, 2009.