On the Continuous and Reactive Analysis of a Variety of Spatio-Temporal Data

Marco Balduini[1][0000−0002−2397−2166]

DEIB - Politecnico di Milano, Milan, Italy
marco.balduini@polimi.it

Abstract. Reactive decision making on heterogeneous streaming data is gaining importance in a wide range of situations, e.g., in the electricity management domain for reacting to anomaly consumption or in oil and gas extraction sites to detect dangerous situations. Modern cities represent a relevant scenario for reactive decision making because of the vast number of stakeholders willing to benefit from the growing availability of streams of data from various sources. In the state-of-the-art, this problem is addressed through ad-hoc solutions that fit only a specific scenario. In this paper, I report on the models and technical implementations, which I propose to enable reactive analysis of a variety of spatio-temporal data, and on their evaluation in real-world scenarios to prove their adequacy.

Keywords: Heterogeneous Spatio-Temporal Streaming Data, Continuous and Reactive Analysis, Urban Data, Streaming Data Fusion, Stream Processing, Stream Reasoning

1 Relevancy

In an increasing number of situations, a decision must be reactive[1] and must be based on a variety of streaming data. In the electricity management domain, a reactive anomaly detection system for the consumption data is useful to avoid network problems. In the oil and gas extraction sites, the analyses of sensors' readings from the wells are vital for reactive detection of dangerous situations.

The urban environment is particularly relevant when talking about reactive decision. In modern cities, a dense network of interactions between people and the environment produces a great amount of spatio-temporal fast evolving data [1] and a multitude of stakeholders are interested in reactive decisions. Tourists would value information about the current top rated and less crowded attractions around the city. Commuters would like to know the busiest roads to choose the fastest way home. Public safety agencies would like to learn about over-crowded area during a public event.

In the mid 2000s', the growing use of location-based social networks via mobile devices, improved the ability to capture the people's interests, habits, and

[1] Deciding an action in response to a stimulus before new incoming information makes the planned action useless.

preferences in a privacy-preserving manner and enabled innovative scenarios. It became possible to create an accurate and up-to-date representation of reality (a.k.a. Digital footprint or Digital reflection or Digital twin) exploiting either social media or mobile phones data, i.e. Call Data Records (CDR). For instance, analyzing social media Cho et. al. [2] were able to identify mobility patterns, while I built a location-based recommendation engine for restaurant in Korea [3]. Parallel works exploited CDR to create models to estimate the density of crowds and vehicles [4–6].

However, better decisions can result from the analyses of multiple data sources simultaneously. The growing availability of new urban data sources (e.g. IoT, WI-FI logs) stimulated the research of a conceptual model to manage data variety in a comprehensive way. The current interest is for solutions that fuse streaming heterogeneous data to enable reactive decisions.

2 Problem Statement

Before starting my PhD, I investigated for three years the modeling and the analysis of streaming data from social media [3, 7–9]. I approached the problem with Stream Reasoning [10], RDF Stream Processing (RSP) and state-of-the-art techniques based on named entity recognition and linking, and machine learning for recommendation.

Reflecting on the obtained results, I identified two main findings: (i) when dealing with data stream, a continuous ingestion mechanism avoids data losses, but continuous analysis is not always needed; an analysis can be reactive even if postponed. (ii) Ontologies are an adequate knowledge representation technique for modeling data characterized by high variety. In the previous works I counted on two assumptions: (A) adequate ontologies to model a domain are available, or they can be obtained with minimal effort by extending existing ones. Indeed, SMA[7], an ontology ables to represent location-based social media data, was defined starting from SIOC[2] by adding only few axioms. (B) Data streams can be RDF-ized at a negligible cost. Indeed, social media APIs return statuses in JSON that can be easily transformed in JSON-LD[3] exploiting standard formats, such as Activity Stream[4].

In my PhD, aiming to continuously and reactively analyze a variety of spatio-temporal data, I reflected on the finding and on the assumptions of my previous work. Social media data is semi-structured: only time and space information is presented in a structured way; the content is unstructured, e.g. free texts or images. On the contrary, IoT data, WI-FI logs, CDRs are structured. While the integration of semi-structured data is generally based on the content analysis (e.g. named entity recognition and linking), the integration of structured data requires other methods, e.g., Ontology Based Data Integration (OBDI) [11].

[2] http://sioc-project.org
[3] https://json-ld.org
[4] http://activitystrea.ms

In approaching my PhD keeping working on Stream Reasoning, I needed to check if the assumptions of my previous work still hold. Assumption A does not hold in this extended scenario, so a first problem emerges:

Rp.1 Defining a conceptual model to represent a variety of streaming data.

Moreover, Assumption B holds only to a limited extent, i.e. for social media data. Therefore, I need to face two problems :

Rp.2 Defining a streaming computational model to enable analysis on a variety of data.
Rp.3 Defining appropriate technical instantiations of the computational model in Rp.2.

Last, but not least, to verify and validate the solutions proposed to solve the problems above, I need to:

Rp.4 Assess, in real world scenarios, the feasibility and the effectiveness of the instantiations developed addressing Rp.3 using the models developed in solving Rp.1 and Rp.2.

3 Related Work

Concerning Rp.1, visual analytics is a common approach to support reactive decision making, but there was a gap between low-level time-varying geo-located data and the high-level needs of visual analytics. Vocabularies to publish the low-level data exist, e.g., geosparql vocabulary[5], event ontology[6] or time ontology[7], but the high-level part, to enable visual analytics, was missing. Social Pixel [12] represents a first attempt to create abstractions to visually represent spatio-temporal phenomena analysing social media data.

The transient nature of streaming information often requires to treat it differently from persistent data. Data streams are often consumed on the fly by continuous queries. Such a paradigmatic change was investigated by the Database community [13, 14] and, more recently, by the Semantic Web community [10] and by the Distributed System community [15]. The processing model of RDF stream processors (RSP) [16] was inspired by the work done in the Database community, in particular by the CQL stream processing model [17]. With regards to Rp.2, at the time I started my PhD, the Semantic Web stack was already extended with stream computing concepts. RDF streams, continuous extensions to SPARQL, as well as continuous reasoning concepts existed. Several RSP Engines also existed [16]. At that time I was maintaining the C-SPARQL Engine [18] and I designed, developed and evaluated SLD [3], a system that exploits RDF stream

[5] http://www.opengeospatial.org/standards/geosparql
[6] http://motools.sourceforge.net/event/event.html
[7] https://www.w3.org/TR/owl-time/

processing and OBDI to enable the layout of complex query networks that continuously analyze social media. But, as I already mentioned in Section 2, I based my works on Assumptions A and B, that don't hold in all the scenarios.

With regards to Rp.3 and to Rp.4, I assessed the work done in benchmarking [19]. In particular, in recent years, the benchmarking of single-threaded implementations against distributed systems has drawn attention. McSherry et. al. in COST [20] showed that a distributed solution, to be effective, must outperform a single-threaded one. Inspired by this work, I decided to solve Rp.3 both with single-threaded and a distributed approach and to evaluate Rp.4 using the cost-effectiveness metric.

4 Research Question

I developed my research question with the Macro, Mezzo and Micro method [21]. The three different levels aim at probing the validity (Rp.4) of the conceptual model (Rp.1), of the computational model for streaming heterogeneous data (Rp.2) and of its technical instantiations (Rp.3).

At Macro level I focused on relevancy and formulated the question: *Is it possible to support reactive decisions by managing data characterized by velocity and variety without forgetting volume?*

At Mezzo level, I focused the attention on a question for which I could find a viable solution. I concentrate my effort on spatio-temporal streaming data, I focused on the findings of my previous work and I characterized the way to support reactive decisions, i.e. visually make sense of data. So, the Mezzo level question is: *Is it possible to visually make sense of a variety of spatio-temporal streaming data by enabling continuous ingestion and reactive analysis?*

Finally, at Micro level, I formalized a question that can be evaluated. I concentrate my effort on the streaming urban data and I specify a way to exploit the visual analytics instrument to support reactive decision making, i.e. find emerging patterns and data dynamics. As a result, my research question is: *Is it possible to continuously ingest and reactively analyses a variety of streaming urban data in order to visualize emerging patterns and their dynamics?*

In answering to the Micro level question, I'm directly contributing to answer the Mezzo level question, and, indirectly, to cast some light on the Macro level question.

5 Approach and Evaluation Plan

Inspired by OBDI methods, I approached the research problems in a modular way by relaxing, in parallel, the two original assumptions presented in Section 2. This modularity reflects the research problems structures and allows me performing a continuous evaluation.

On the one hand, relaxing Assumption A, I approached the creation of a conceptual model in the form of an ontology by following the Methontology [22]

methodology, and I planned to evaluate the result using Tom Gruber's principles [23].

On the other hand, relaxing Assumption B, I planned the development of a computational model to enable continuous ingestion, wrangling and reactive analysis of heterogeneous data streams. I planned to implement such a computational model using different technologies, i.e. single-threaded and distributed, in order to prove its adequacy in different work conditions. To finalize the work I planned the evaluation of the cited implementations against already existing system (SLD) and one against the other. In particular, inspired by COST [20], I decided to evaluate the cost-effectiveness of the single-threaded system against the distributed one.

The modular approach, during the development and the evaluations phases, allowed me planning an overall evaluation. I planned to put at work a complete system, composed by an implementation of the computational model that exploits the conceptual model, in different scenario and to evaluate it: (i) in terms of guessability [24] of data visualization by the users, and (ii) in terms of performances using well-known indicators, i.e. throughput and cost-effectiveness.

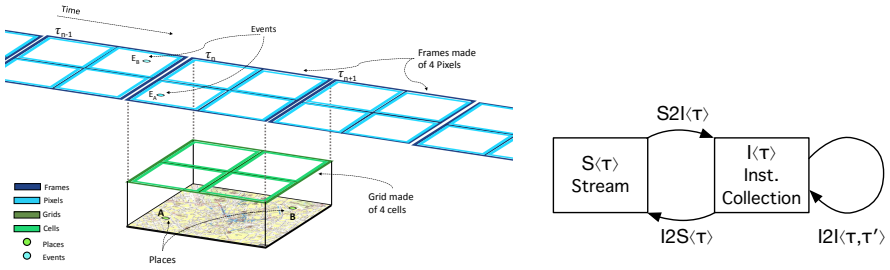
6 Hypotheses

In order to answer my research questions, I formulated a set of hypotheses that I used to operationalize my work, w.r.t. the four problems in Section 2.

- Hp.1 A conceptual model containing concepts from the image processing domain can represent spatio-temporal data in an extendable and coherent way with a minimal encoding bias and a minimal ontological commitment.
- Hp.2 A streaming computational model that defers as long as possible the data transformation is less complex, in terms of time and space, than a computational model that cast data into RDF at ingestion time.
- Hp.3 A single-threaded implementation of the streaming computational model from Hp.2 that uses the conceptual model from Hp.1 can be more cost-effective than a distributed implementation of the same model while guaranteeing the reactiveness of the system
- Hp.4 An implementation from Hp.3 can create a bridge between data analytics and data visualization that enhances the comprehension of a variety of spatio-temporal data and, at the same time, is reactive.

7 Results

To validate Hp.1, I created the FraPPE ontology. Figure 1(a) offers a graphical overview of the FraPPE concepts. The abstractions in the FraPPE ontology [25] exploit classical image processing concepts (i.e. Pixel and Frame) as well as common sense concepts (i.e. Place and Event). The intuition behind the FraPPE



(a) (b)

Fig. 1. (a) presents a high-level view of FraPPE including Frames, Pixels, Places and Events, (b) presents an overview of the operators inspired by CQL

data model is the discretization of space and time in atomic units. The representation of the geographical space is mediated by a *Grid* of *Cell*s that contain *Place*s. *Frame*, *Pixel*s and *Event*s are the time-varying representation (i.e. taken every given interval of time) of, respectively, Grid, Cells and Places. FraPPE was developed using Methontology [22], and complies with the Tom Gruber's principles [23], i.e clarity, coherence, minimal encoding bias, minimal ontological commitment, extendibility.

In parallel, I investigated a streaming computational model to enable access and analysis of a variety of streaming data. The main idea behind this part of the work is to combine my previous findings (see Section 2) with the intuition that, often, data transformation can be deferred (as stated in Hypothesis Hp.2). For example, if we need to filter a stream of JSON items in a first stage of a long query network, the execution of a path query with JSONiq[8], before transforming the data in RDF, is for sure faster than transforming the data in RDF and then executing a graph pattern matching.

Figure 1(b) shows the three proposed classes of operators inspired by CQL [17]. $\mathcal{T}$ denotes a generic type to-be-specified-later, $S\langle\mathcal{T}\rangle$ is a generic data stream and $I\langle\mathcal{T}\rangle$ a collection of instantaneous generic data items (e.g., a table, a document, or a graph, which are normally manipulated by relational, document-based or graph-based databases). Those operators allow moving from generic data streams to instantaneous generic collection an vice versa.

As a first implementation of the computational model, I developed Natron: a direct improvement of SLD that maintains the single threaded nature of the original platform. I empirically evaluated the performance of Natron against SLD and validated Hypothesis Hp.2 by proving that a deferred data transformation, namely Lazy Transformation principle, can improve the performance of a stream processing framework [26]. Inspired by the momentum of the distributed technologies and by the work presented in [20], I also implemented a horizontally scalable version of the computational model based on Spark. Both

[8] http://jsoniq.org

implementations operate on data in its original format as long as they can, and they transform it only if it is really needed. I evaluated the cost-effectiveness of the distributed implementation against Natron and I demonstrated that the single-threaded implementation can outperform the distributed one [27]. This result validated Hp.3 from an empirical perspective.

In order to validate Hp.4, Natron and FraPPE were then put at work and evaluated in real-world scenario [28–30]. Those works demonstrate the validity of the whole infrastructure in various scenarios facing heterogeneous streaming data. The guessability and the reactiveness of the visual analytics instruments enabled by the system were evaluated by tens of real-world users via questionnaires and interviews.

8 Reflections

During my PhD, I collected positive evidences that a system such as Natron (based on a streaming computational model and on the Lazy Transformation principle), and a conceptual model such as FraPPE (containing concepts inspired by image processing) represents an adequate solution to enable visual analytics of heterogeneous streaming urban data in a reactive way. Unfortunately, so far, the evaluation was conducted only exploiting the multiple implementations of the two proposed models. This approach poses limits to the positive evaluation of Hypothesis Hp.2. I now need to perform a formal evaluation of Hypothesis Hp.2. In the remaining part of my PhD, I intend to define a formal algebra for the computational model in order to estimate the time and space complexity of the operators and to define cost models that can be exploited to automatically optimize query networks designed by users with a limited know-how on the internals of my implementations.

Acknowledgments. I worked under the supervision of Prof. E. Della Valle.

References

1. Rob Kitchin. The real-time city? big data and smart urbanism. *GeoJournal*, 79(1):1–14, 2014.
2. Eunjoon Cho et al. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090. ACM, 2011.
3. Marco Balduini et al. Social listening of city scale events using the streaming linked data framework. In *ISWC (2)*, volume 8219 of *LNCS*, pages 1–16. Springer, 2013.
4. Nathan Eagle et al. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.
5. Richard A. Becker et al. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.
6. F. Calabrese et al. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Trans. Intelligent Transportation Systems*, 12(1):141–151, 2011.
7. Marco Balduini et al. BOTTARI: an augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams. *J. Web Sem.*, 16:33–41, 2012.

8. Marco Balduini et al. Reality mining on micropost streams - deductive and inductive reasoning for personalized and location-based recommendations. *Semantic Web*, 5(5):341–356, 2014.
9. Marco Balduini et al. Recommending venues using continuous predictive social media analytics. *IEEE Internet Computing*, 18(5):28–35, 2014.
10. Emanuele Della Valle et al. It's a streaming world! reasoning upon rapidly changing information. *IEEE Intelligent Systems*, 24(6):83–89, 2009.
11. Maurizio Lenzerini. Data integration: A theoretical perspective. In *PODS*, pages 233–246. ACM, 2002.
12. Vivek K. Singh et al. Social pixels: genesis and evaluation. In *ACM Multimedia*, pages 481–490. ACM, 2010.
13. Brian Babcock et al. Models and issues in data stream systems. In *PODS*, pages 1–16. ACM, 2002.
14. Minos N. Garofalakis et al., editors. *Data Stream Management - Processing High-Speed Data Streams*. Data-Centric Systems and Applications. Springer, 2016.
15. Matei Zaharia et al. Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters. 2012.
16. Daniele Dell'Aglio et al. Stream reasoning: A survey and outlook. *Data Science*, (Preprint):1–25, 2017.
17. Arvind Arasu et al. The CQL continuous query language: semantic foundations and query execution. *VLDB J.*, 15(2):121–142, 2006.
18. Davide Francesco Barbieri et al. C-SPARQL: a continuous query language for RDF data streams. *Int. J. Semantic Computing*, 4(1):3–25, 2010.
19. Arvind Arasu et al. Linear road: A stream data management benchmark. In *VLDB*, pages 480–491. Morgan Kaufmann, 2004.
20. F. McSherry et al. Scalability! but at what cost? In *HotOS*. USENIX Association, 2015.
21. Jeffrey R. Lacasse et al. Making assessment decisions: Macro, mezzo, and micro perspectives. In *Critical Thinking in Clinical Assessment and Diagnosis*, pages 69–84. Springer, 2015.
22. M. Fernández-López et al. Methontology: from ontological art towards ontological engineering. 1997.
23. Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.
24. Jackie Moyes et al. Icon design and its effect on guessability, learnability, and experienced user performance. *People and computers*, (8):49–60, 1993.
25. Marco Balduini et al. Frappe: A vocabulary to represent heterogeneous spatio-temporal data to support visual analytics. In *ISWC (2)*, volume 9367 of *LNCS*, pages 321–328. Springer, 2015.
26. Marco Balduini et al. SLD revolution: A cheaper, faster yet more accurate streaming linked data framework. In *ESWC (Satellite Events)*, volume 10577 of *LNCS*, pages 263–279. Springer, 2017.
27. Marco Balduini et al. Cost-aware streaming data analysis: Distributed vs single-thread. (in press).
28. Emanuele Della Valle et al. Listening to and visualising the pulse of our cities using social media and call data records. In *BIS (Workshops)*, volume 228 of *LNBIP*, pages 3–14. Springer, 2015.
29. Marco Balduini et al. Citysensing: Fusing city data for visual storytelling. *IEEE MultiMedia*, 22(3):44–53, 2015.
30. Marco Balduini et al. Models and practices in urban data science at scale. (in press).