Real Time IoT Stream Processing and Large-scale Data Analytics for Smart City Applications

Ralf Tönjes*, Muhammad Intizar Ali†, Payam Barnaghi‡, Sorin Ganea¶, Frieder Ganz‡, Manfred Haushwirth†, Brigitte Kjærgaard‡‡, Daniel Kümper*, Alessandra Mileo†, Septimiu Nechifor††, Amit Sheth‖, Vlasios Tsiatsis** and Lasse Vestergaard§

*University of Applied Sciences Osnabrück, Germany, †Insight @ National University of Ireland, Galway, ‡‡ City of Aarhus, Denmark, ‡ University of Surrey, Guildford, UK, †† Siemens Romania, §Alexandra Institute, Aarhus, Denmark, ¶Brasov Metropolitan Agency, Romania, ‖Wright State University, Ohio, **Ericsson Research, Sweden

***Abstract*—Smart cities are evolving into a larger interconnected ecosystems and many applications and services in these ecosystems are going on-line. A key challenge in smart city applications is aggregation and processing of streaming information from various domains. In fact, large amounts of valuable data and sensor information remain still unused or are limited to specific application domains due to large number of specific technologies and format, and efficient adaptive stream processing for high-level events is still a hard task. We present a smart city framework for processing large-scale IoT data streams by enriching data streams with semantic annotations, enabling adaptive processing, aggregation and federation of data. We also address the challenges of smart adaptation, user centric decision support and reliable information processing for smart city applications.**

I. Introduction

An increasing number of cities have started to use Information and Communication Technology (ICT) enabled services to address sustainability as well as to improve the operational efficiency of the city services and infrastructure, providing citizens with better experience. As a result, cities are evolving into larger ecosystems that were previously more or less unconnected and are now going more and more on-line. The city council is the pivotal facilitator in making this on-line ecosystem of ecosystems become a reality. The achievement of sustainability for smart city frameworks is, however, challenged in several ways.

A first key challenge in developing smart city frameworks is the integration across different application domains, as well as the engagement of different city departments, city-contracted entrepreneurs and individual enterprises providing services. Large amounts of valuable data and sensory information remain unused or are limited to specific application domains due to the large number of specific technologies and formats (e.g. traffic information, parking spaces, bus timetables, waiting times at events, event calendars, environment sensors for pollution or weather warnings, GIS databases etc.). An aggregation of information from various sources is typically done manually and is often outdated or is provided static sources of data. This hinders the ability to continuously link, interpret and share dynamic knowledge across city stakeholders and citizens.

This work is supported by the EU FP7 CityPulse Project under grant No.603095. http://www.ict-citypulse.eu

At the same time, a rapid growth in Wireless Sensor and Actuator Networks (WSANs) and emerging technologies and solutions in the Internet of Things (IoT) domain is giving access to sensory information provided by citizens (participatory sensing) through smart devices (i.e. smart phones) or embedded sensors. In this work we describe a framework to enable innovative smart city applications by adopting an integrated approach to the Internet of Things and the Internet of People. We discuss how the creation and provision of reliable real-time smart city applications can be facilitated by bringing together the two disciplines of knowledge-based computing and reliability testing.

We describe the challenges, issues and solutions to collect physical world data and integrate it in to cyber and social systems. Smart city data can be seen as big data; however it is not only large in volume, it is also multi-modal, varies in quality, format, representation form and levels of dynamicity. The main difference between the sensory data streams compared to the conventional streams and data models will be rapid changes in data and dynamicity of the environment, resource constraints in sensing platforms and distribution and heterogeneity of data that make the processing and event detection a challenging task. The sensory streams in a smart city environment can emerge from various sources, with different qualities, modalities and trust and reputation that need to be identified and associated to these sources. Efficient stream reasoning mechanisms are required to interpret the meaning of events in a context-aware fashion, and share such meaning across applications. Figure 1 shows a view of the different data sources and key areas in the proposed framework.

In Section II, we identify basic requirements for the smart city applications and review the state of the art. The building blocks of the proposed smart city framework is detailed in Section III, while its functional components are discussed in Section IV. We advocate the practicality of our framework by describing real word usecase scenarios designed by our city partners in Section V. Finally, we conclude our work in Section VI.

II. Requirements Analysis and State of the Art

In this section, we map the basic requirements for smart city applications into four broad categories and compare them with the state of the art.

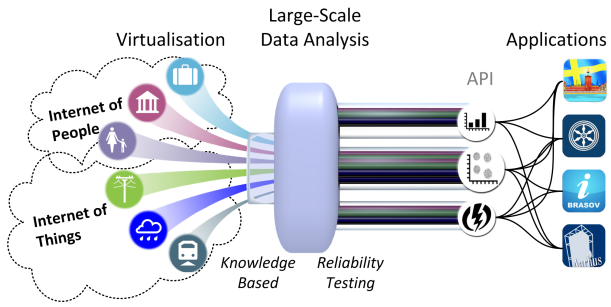

Fig. 1. Integrated approach for IoT and social media stream based smart city applications

1) Federation of heterogeneous data streams using semantic description and annotation: Data Federation combines heterogeneous sets of data to provide a unified view. In the context of smart city data the dynamicity and heterogeneity of various IoT streams and their effective utilisation by integrating them with the existing data sources is a key challenge. Smart city frameworks should provide mechanisms to (i) *seamlessly integrate real world data streams*, (ii) *automated search, discovery and federation of data streams*, and (iii) *adaptive techniques to handle fail-overs at run-time*.

A number of modelling methods for formally representing sensors and sensor networks using semantic Web technologies have been proposed[16], [5]. Semantic Sensor Networks Incubator Group (SSN-XG) has developed a semantic description framework for Semantic Sensor Networks (SSN) based on the concepts of systems, processes, and observations [1]. The common limitation of the existing methods of semantic annotations is that properties related to the dynamicity of the physical world to a large extent are not modelled. The existing works focus on IoT resources and data but less on IoT streams and their features. Also the adoptive techniques to handle the failover of the IoT streams require further investigation.

2) Large scale IoT processing and data analytics: The smart city framework should provide an open platform for discovery, integration, federation and processing of large-scale IoT streams from heterogeneous sensory resources. The large volume and heterogeneity of data makes the modelling of the real world streams different from conventional models. Querying and accessing the data in many cases will require real-time (or near-real-time) discovery and access to the streams (and their data) and the ability to mash-up different kinds of streaming data from various sources. Smart city applications not only require to efficiently process large scale IoT streams but also need efficient methods to perform data anlytics in dynamic environment by aggregating, summarizing and abstracting sensor data on demand.

Data analytics is an essential research topic in the IoT domain that has attracted many different research areas such as statistics, machine-learning and data mining [11]. Existing approaches introduce general analysis techniques without having a motivation for smart city applications. WEKA [13] is a toolbox for data mining tasks that can analyse static data, but misses features for stream handling. MOA [6], an advancement of WEKA for streaming data is able to handle streaming data, including data from social media. Though, MOA follows a centralized approach and lacks scalability for large-scale applications. The SAMOA [9] project aims in merging streaming data analytic techniques from MOA with distributed processing engines such as Apache Storm and Apache S4. These current analytic frameworks have to be evaluated for applicability in the smart city environment and the impact on privacy has to be taken into account.

3) Real-time IoT information extraction, event detection and stream reasoning: Smart city applications should be able to process event streams in a real time, extract relevant information and identify values that do not follow the general trends. Beyond the identification of relevant events, extraction of high level knowledge form heterogeneous, multi-modal data streams is an important component of IoT. Existing stream reasoning techniques use background knowledge and streaming queries to reason over data streams. There are some existing systems that facilitate streaming queries based on semantic models [4], [3], [8]. Current techniques of stream reasoning do not cater to the needs of IoT due to the lack of proper treatment of uncertainty (e.g. possible reasons of traffic jam vs. most probable reason of traffic jam) in the IoT environment.

Event detection is an interdisciplinary area of research and borrows research insights from novelty detection and anomaly detection. Event detection in social streams is carried out in [17]. Existing event detection and processing techniques are typically limited to a single one modality (e.g., processing only sensor data). However, there is a lack of an event detection approach that would take into account multiple data streams considering their varying nature in terms of trustworthiness, reliability, and other aspects.

4) Reliable information processing, QoI, testing and monitoring: Data quality issues and provenance play an important role in smart city scenarios. Smart city frameworks should provide methods and techniques (i) *to evaluate accuracy, trustworthiness, and provenence of IoT streams*, (ii) *to resolve conflicts in case of contradictory information*, and (iii) *continous monitoring and testing to dynamically update QoI and trustworthiness.*

There is a wide spread research work in QoI and data provenance in the service area, however there has been very little research on combined QoI, trustworthiness and provenance in IoT and smart city environments. The problem of QoI has been discussed in homogeneous WSAN[7] and identifies issues closely related to QoI, such as accuracy, timeliness, reliability/confidence, completeness, relevancy and usability. Smart city environments open new challenges due to distributed ownership, high dynamicity and heterogeneity of the systems involved.

Testing and monitoring in the research area of smart cities utilises various testing methods like test first [2], model based testing or fuzz testing [12]. An adapted test data generation [10] is used to reach high efficiency. The extensive description of provided services can be used to derive adapted test cases[15]. Test of resource constrained devices always has to consider energy costs of test runs if devices e.g. run on batteries [14]. Current testing approaches in smart city environments are not able to rely on collected test data and require energy consuming procedures to ensure application

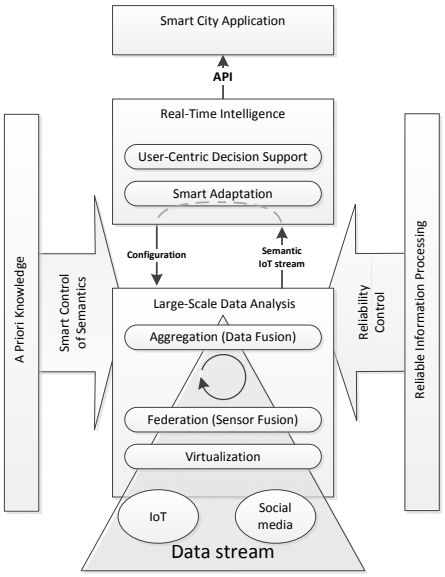

Fig. 2. Framework for data stream based smart city applications

availability. Beyond distinct information if applications function is working or services QoS parameters are satisfying there is no output information about the information quality sensors deliver compared to their neighbours.

III. Smart City Framework

The CityPulse framework addresses the complex task of stream processing by providing large-scale data analysis and real-time intelligence functionalities (Fig. 2). In this way, the control intelligence is separated from powerful parameterized big data processing algorithms. In addition to that, reliability measures are used to assess data quality and optimize the processing pipeline accordingly. Details of each component of the smart city framework are provided in the remainder of this section.

a) Virtualisation and semantic annotation: A major hindrance of the uptake of Smart City applications is the heterogeneity of the data coming from the urban fabric, human sensors, social web data and city information systems. As a result the applications are usually data specific, often assuming implicit knowledge about the data and limiting the possibility of reusing the data for other applications. To overcome these silo architectures virtualisation is employed to provide a common abstract interface to the higher layers that deal with information processing and knowledge extraction. Moreover, this layer is the key enabler to ensure easy discovery and automated processing of data: This layer assigns a semantic annotation to the data streams, thus making their characteristics and capabilities available in a machine processable way.

b) Federation: heterogeneous data streams are made accessible to data consumers as one integrated data stream by using on-demand data integration. Hence this component combines heterogeneous sets of data streams to create one unified view of all the data. The IoT environment and the physical world are subject to constant changes and the devices that provide the underlying layer for IoT are often resource constrained, pervasive, unreliable and heterogeneous in nature. To overcome this problem in an automated way this layer employs a knowledge-based approach using semantics and linked data to represent data stream relationships and to support mash-ups. Thus adaptive methods allow for real-time federation of multi-modal data streams.

c) Aggregation: While the above federation provides a kind of sensor fusion, the aggregation aims at data fusion, thus reducing the enormous amount of data by aggregation techniques, e.g. clustering, summarisation, filtering and pattern recognition. The data is time and location dependent, is often transient and often needs to be processed in correlation with other data to create higher-level abstractions. Hence aggregation and federation are applied consecutively and iteratively to extract the relevant data and to detect events. During run-time the tool chain is dynamically adapted and con-figured in a goal-oriented way by the following smart adaptation layer.

d) Smart Adaptation: This layer provides methods for higher-level information processing to interpret the semantic data in the current context, transforming the lower-level dynamic information (e.g., changes in sensor readings) to higher-level contextual abstractions. Data interpretation is challenged by the real-time demands of the targeted smart city applications. In most of the existing solutions matchmaking between the requirements expressed by applications and available data is carried out at design-time, neglecting the fact that the properties of underlying services depend on dynamic phenomena (e.g. sensor readings - network availability, weather conditions, and temperature). To overcome this limitation, the smart adaptation component consumes the semantic data streams from the lower layers and monitors the need for adaptation in the configuration of the data processing pipeline.

e) User centric decision support: aims at utilizing contextual information, usage patterns and preferences to provide optimal configurations of smart city applications. Users (citizens, enterprises or city councils) can specify their requirements and preferences explicitly or have them implicitly derived from the application's usage patterns. The social and context analysis enables matchmaking and discovery mechanisms to match the data according to the users preferences and context-dependent attributes (location, time, etc.).

f) Reliable Information Processing: The IoT systems often operate in dynamic environments that are subject to change and prone to errors. Developing reliable data processing and information extraction methods requires accuracy and trust issues to be taken into consideration when dealing with IoT data. This is further emphasised when data from citizens (e.g. smart phone sensors or social media) is used. Therefore accuracy and trust are considered by dedicated measures and methods in data acquisition, federation and aggregation. To cope with malfunctioning or even disappearing sensors, noisy and conflicting data, techniques for monitoring (run-time) and testing (design-time) are integrated in the framework, thereby ensuring reliable information processing.

g) Smart City Applications: Many of today's cities are faced with challenges resulting from changes and new demands that a rapidly growing digital economy imposes on current applications and information systems. In order to foster rapid prototyping the framework provides application programming

interfaces (API) to smart city application programmers, enabling access to and management of the building blocks of our smart city framework described in this section.

IV. Functional Components

The three main functional components of the proposed smart city framework are described in this section.

A. Large-Scale IoT Stream Processing

Data virtualisation is the process of offering data consumers a common data access interface that hides the technical aspects of data streams, such as location, storage structure, access language, and streaming technology. Virtualisation should include IoT sensing objects, either alone or in conjunction with associated external data stream processing, in a way that enables the generation of "smart" sensing feeds that can autonomously satisfy requests at different level of sensing detail. The virtualised objects also need to be enriched with metadata as well as with processing capabilities, in order to achieve self-awareness, self-manageability, self-(re-)configurability. This component provides an integrated platform based on linked open data and Web standards to virtualize resources and provide uniform access and integration support irrespective of the information source. Using semantic annotations for all data and data streams, data becomes machine understandable and processable. This meta-knowledge is employed to automate the discovery and processing control for data analytics, facilitating event detection, aggregation and reasoning.

B. Reliable Information Processing

In order to achieve reliable data and resilient applications at design as well as at run-time, the smart city framework utilises the following two components.

1) QoI Datastore and Reputation Systems: This component adapts and develops methods for automated rating of information against accuracy, trustworthiness and QoI using techniques such as possibility theory, fuzzy sets, Bayesian networks and evidence theory. Analysis algorithms not only require initial training phase but also require to be adjusted in the later phases by the continuous analysis of data streams. Active reputation system to identify false positives and false negatives can evaluate the technical reliability and the provenance based trustworthiness of the data streams. This components is utilized during IoT stream processing to provide an adaptive control-loop for the continuous assessment of reliability of the data streams and extracted information. Smart city applications can also demand their requirements in terms of accuracy, trustworthiness, QoI and resource constraints which can be satisfied by this component.

2) Testing and Monitoring Environment: During design-time the testing and monitoring environment initially tests implemented applications for their functionality and evaluates their performance using benchmarking datasets. While during run-time, a continuous monitoring and analysis of the data sources and real-time data streams can proactively maintain conflict resolution and fault recovery. Monitored information is represented in uniform metrics allowing domain overlapping comparison of the real time information.

C. Real-Time IoT Intelligence

Users of smart city applications should be provided with solutions that take into account their individual needs and usage patterns to provide better decision support. Individual factors and ever-evolving urban fabric reflected in dynamicity of data streams, data mash-ups and services should provide user-centric, dynamicity-aware and adaptive smart city applications. We refer to these capabilities as *Real-Time IoT Intelligence* which provides mechanisms to enable the user-centric and adaptive decision support. This component comprises of two main functionalities:

1) Adaptivity: Real-time adaptivity will be based on the identification of unexpected events. Aggregated high-level events can be unexpected either because they do not fit the technical configuration requirements guiding the federation process, or because their relevance to a particular context is different than expected. The former will determine the need for a change in the configuration settings for federation and aggregation of events, while the latter will require to adapt the decision support process to either extend or reduce the context of reference. Detection of such unexpected events will be based on continuous query processing over semantic streams, which will trigger Stream Reasoning to identify the criticality and determine which adaptation is needed. Appropriate result will be fed back to the decision support component and suggestions will be made for the dynamic reconfiguration of stream discovery and aggregation mechanisms.

2) User-centric stream reasoning: This functional component aims at considering a set of user-centric factors (explicit and implicit requirements, preferences, usage profile patterns) to provide optimal configuration of smart city applications. Our approach is based on modeling user requirements and preferences using Linked Data and open vocabularies in order to provide a lightweight, interoperable and well-established foundation for decision making support and match-making of city services. Users can thus specify their requirements and preferences explicitly. In addition, requirements and preferences can be implicitly derived from their individual application usage patterns, and taken into account to provide suggestions to users based on their previous behaviour. Data mining techniques also play a role in building user preference models resulting in an individualised experience.

V. Usecases

Scenarios from our city partners Brasov (Romania) and Aarhus (Denmark) show how the smart city framework can create synergies within the respective municipalities.

A. Traffic management as a support element for sustainable development of the urban fabric

Brasov is known as being the green capital of Romania, having at the same time a long and successful history of large-scale industry and transportation development. The city of Brasov uses video surveillance through various networks of video cameras and Geographical Information System (GIS) infrastructure. These networks allow a basic level of traffic surveillance for identifying any traffic jams that occur within the city boundaries. The medium and long-term objective is to develop such video surveillance networks as well as various

complementary networks in order to achieve an efficient mobility management and public safety. Another objective is to raise the level of use of ICT in locating public transportation, e-ticketing for public transport users, information of passengers both in public areas as well as in the public transport vehicles.

The smart city framework will integrate knowledge about GPS location and video surveillance to provide a real time analysis of the traffic conditions and use it to organize and optimize traffic flows in the Brasov City area, and relate this knowledge with the public transport e-ticketing system, management of traffic lights, quality of air in the most congested areas of the city, etc. to provide users a personalized support to decisions related to transportation.

B. Open innovation platform

In recent years the city of Aarhus (second largest city in Denmark) is moving towards being a smart city through open innovation, where co-creation and citizen empowerment is in focus. One outcome of this effort is the open data platform[1]. The idea of the platform is to give anyone (organization and people from different sectors) free access to play with public city data, fostering innovation and creation of new services for citizens.

Despite the fact that the platform is still in an early stage, a lot of activities are already happening. Citizens and local University have been using the data for developing targeted applications, and the municipality has had competitions where they collaborated with companies and private users. Driven by emerging needs for real-time applications, the platform is beginning to focus on dynamic data, and there are already a few real-time datasets on the platform including public trash bin weighings and flow of library rentals. Real-time analytics on such data, ability to aggregate and process them to obtain higher-level insights and dynamic monitoring of certain trends are extremely valuable for Aarhus, since it allows to analyze how data is used, by whom and for what. This information is relevant for further developing the platform and the ecosystem, and it can be used by decision-makers, to map out innovation in the city.

VI. Conclusion

The current Internet of Things platforms and Smart City applications often focus on providing connectivity and communication and also support data collection and employ data analytics. however, the main focus is usually on the collected and stored datasets with an emphasis on high performance computing and data mining solutions. While the recent efforts in this area have enabled emerging technologies and solutions to develop novel techniques for smart city application and use-cases scenarios, there is however a gap in providing efficient and scalable methods that enable real-time processing and interpretation of streaming sensory and social media data in smart city environments. In this paper we discuss principles of large-scale data analytics for real-time smart city data processing and interpretation and discuss how various sources of raw sensory data can be combined and processed to extract actionable-knowledge that can be used by citizens and/or decision support systems that are used by the city authorities. We also introduce integrating Physical, Cyber and Social data and analytical solutions for processing multi-modal smart city data. The paper describes an architecture and discusses the functional components for a smart city framework. The key requirements, technologies and solutions to enable the architecture are also discussed. We provide samples of the use-case scenarios that are being developed in the CityPulse project. The future work will focus on development of the components and evaluation of the proposed architecture and the functional components based on real-time city data.

[1] (www.odaa.dk)

References

[1] W3C Semantic Sensor Networks Incubator Group (SSN-XG). http://www.w3.org/2005/incubator/ssn/.

[2] J. Andrea. Envisioning the next-generation of functional testing tools. *Software, IEEE*, 24(3):58–66, 2007.

[3] D. Anicic, P. Fodor, S. Rudolph, and N. Stojanovic. Ep-sparql: a unified language for event processing and stream reasoning. In *Proceedings of the 20th international conference on World wide web*, pages 635–644. ACM, 2011.

[4] D. F. Barbieri, D. Braga, S. Ceri, E. Della Valle, and M. Grossniklaus. C-sparql: Sparql for continuous querying. In *Proceedings of the 18th international conference on World wide web*, pages 1061–1062. ACM, 2009.

[5] P. Barnaghi, S. Meissner, M. Presser, and K. Moessner. Sense and sens ability: Semantic data modelling for sensor networks. In *Conference Proceedings of ICT Mobile Summit 2009*, 2009.

[6] A. Bifet, G. Holmes, B. Pfahringer, J. Read, P. Kranen, H. Kremer, T. Jansen, and T. Seidl. Moa: a real-time analytics open source framework. In *Machine Learning and Knowledge Discovery in Databases*, pages 617–620. Springer, 2011.

[7] C. Bisdikian, R. Damarla, T. Pham, and V. Thomas. Quality of information in sensor networks. In *1st Annual Conference of ITA (ACITA07)*, 2007.

[8] A. Bolles, M. Grawunder, and J. Jacobi. Streaming sparql-extending sparql to process data streams. In *The Semantic Web: Research and Applications*, pages 448–462. Springer, 2008.

[9] G. De Francisci Morales. Samoa: a platform for mining big data streams. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 777–778. International World Wide Web Conferences Steering Committee, 2013.

[10] M. Fischer and R. Tonjes. Generating test data for black-box testing using genetic algorithms. In *Emerging Technologies & Factory Automation (ETFA), 2012 IEEE 17th Conference on*, pages 1–6. IEEE, 2012.

[11] F. Ganz, P. Barnaghi, and F. Carrez. Information abstraction for heterogeneous real world internet data. 2013.

[12] P. Godefroid, M. Y. Levin, D. A. Molnar, et al. Automated whitebox fuzz testing. In *NDSS*, volume 8, pages 151–166, 2008.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[14] G. Kortuem, F. Kawsar, D. Fitton, and V. Sundramoorthy. Smart objects as building blocks for the internet of things. *Internet Computing, IEEE*, 14(1):44–51, 2010.

[15] D. Kuemper, E. Reetz, and R. Tonjes. Test derivation for semantically described iot services. In *Future Network and Mobile Summit (FutureNetworkSummit), 2013*, pages 1–10. IEEE, 2013.

[16] D. J. Russomanno, C. Kothari, and O. Thomas. Sensor ontologies: from shallow to deep models. In *System Theory, 2005. SSST'05. Proceedings of the Thirty-Seventh Southeastern Symposium on*, pages 107–112. IEEE, 2005.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.