Towards Quality Aware Sensor Data Stream Processing in a Smart City Living Lab

Aboubakr Benabbas
University of Bamberg
aboubakr.benabbas@uni-bamberg.de

Simon Steuer
University of Bamberg
simon.steuer@uni-bamberg.de

Daniela Nicklas
University of Bamberg
daniela.nicklas@uni-bamberg.de

ABSTRACT

Smart cities have gained a lot of attention in recent years. Under this vision, people develop and plan how life can be improved in cities that constantly grow, which makes city management harder using only conventional methods. Hence many cities rely on live data gathered by multi-sensory constellations.

Living Lab Ecosystems aim at offering a common platform for research, private and public institutions for data collection and processing for research and development purposes.

One of the big challenges is the quality of data generated by the different sensors, based on which important decisions are made. In this paper, we propose a Quality Aware Sensor Data Stream Processing as part of a Living Lab infrastructure that continuously monitors data collection and enriches the data with relevant quality information.

Keywords

Data Stream Processing, Data Quality, Sensor Networks, Complex Event Processing

1. INTRODUCTION

Cities are getting bigger and bigger. Statistics show that cities are becoming larger than urban areas. This poses a lot of challenges to city management, who has to deal with all aspects of daily life and handle the problems as they come. However, not only mega cities have to face those problems. The philosophy behind smart cities aims at involving inhabitants, city authorities and companies in the management of the different utilities. IBM [7] defined a smarter city as a city that optimally exploits the data from the available data networks for a better management and control over the different processes. Realisation of smart cities underlies different applications and communication channels that enable the integration of the aforementioned stakeholders into the whole system. In addition to mega cities, even smaller communities can benefit from such approaches to manage their resources in a sustainable way; they need to cope with an aging population that needs access to mobility or health care. Initiatives targeting these challenges extend the term "Smart City" to be a "Smart Region".

The use of a Living Lab as a platform for testing new technology, implementing new research ideas in the fields of smart region and smart city has a big potential. It enables a productive and transparent cooperation between research institutions (e.g. universities), public institutions (e.g. city management), companies and citizens. It facilitates data collection from different data sources and sensors in the activity region of the Living Lab and the implementation of new solutions for city problems.

The University of Bamberg will operate a Living Lab that includes different components and interfaces to handle incoming data streams from a multi-sensory installation. The Living Lab can process data for the clients and give back results as well as acquiring sensor data from external sources. The data gained can be in turn published for the general public.

To achieve our goals, a lot of requirements have to be met, and controlling and managing data quality is one of those requirements. Data quality is very crucial for decision making, since unknown data quality will yield decisions of unknown quality. Our contribution is a proposal for a Quality Aware Sensor Data Stream Processing in a Living Lab platform that receives data and continuously enriches it with the important quality information.

The use of sensors as the main source of data brings us to the ever increasing use of multi-sensor applications that perform its computations based on the combination of data from different sources. These applications have to be provided with enough data quality to make sure the outcome meets the application-specific requirements. Where large areas are equipped with many cost effective sensors means that a high spatial redundancy is available and can be exploited to make up for the lack of precision and faulty behaviour of the sensors. Besides we can have sensors that offer content-redundancy like providing related observations about the same feature of interest. This can also be used to determine the quality of one measurement using the other observation as a determining factor.

The paper is structured as follows: in Section 2 we introduce a main use case for a data quality component in the Living Lab. In Section 3 we discuss the related work in the

29^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 30.05.2017 - 02.06.2017, Blankenburg/Harz, Germany.

area of data quality and derive the quality dimensions of this quality component. In Section 4 we present the general architecture of the system, the underlying Sensor Model and the query plan that generates that quality enriched data. In the last section we discuss challenges of data quality in sensor data and future work.

2. USE CASE: PEOPLE COUNTING

The use cases for a Living Lab infrastructure are manifold. They range from environmental monitoring (e.g., the water level of a river) over multi-modal mobility management to smart energy solutions. In this paper we present a use case of public safety in street festivals. The festival organizers get permission from the city management to organize a street festival and pledge to comply with the safety guidelines defined by the public order office.

The public order office and the organizers want to continuously monitor the situation on the streets where the festival is taking place. An estimation of people's count is a valuable information for all parties in this use case. The stakeholders in this situation are the city management represented by the public order office, the festival organizers, and the different businesses in the festival area.

The public order office needs an overview on the situation on the streets. The office makes assumptions about the capacity of the streets and the flow of people. It has also guidelines for safety measures. The University of Bamberg operates the Living Lab, which receives data from different sensors. In order to estimate the number of people on a certain area, people counting cameras and mobile devices scanners (FT) are deployed. FT stands for *FlowTrack* that is a commercial product whose function is to scan a certain area for mobile devices [6].

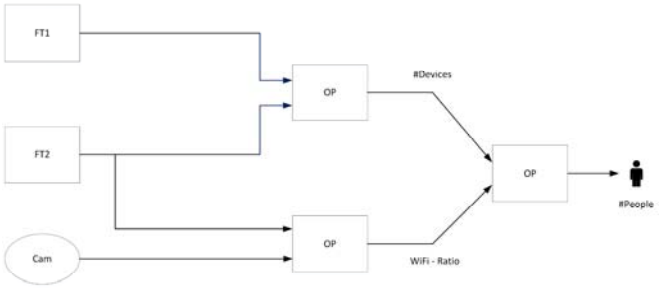

Figure 1: people counting process

The cameras report people entering a defined area and mobile devices' scanners (FTs) report the mobile devices in approximately the same region. Mobile devices' scanners (FT in the figures) are devices that scan a certain area and record the MAC addresses of mobile devices in the area. Data from both sensors will be sent to the Living Lab for processing by the stream processing service.

Figure 1 shows the processing that needs to be performed in the Living Lab. A stream processing component computes the results using several operators (OPs). The operators either process the data directly as it comes from the sensors or process data already processed by other operators. The

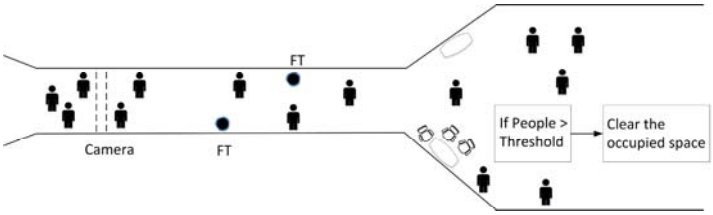

Figure 2: People counting in street festivals

FTs data is fed to the OPs and the result is the number of distinct devices. Another operator is fed data from FT and camera and computes the WiFi-Ratio. WiFi-Ratio is the relationship between the number of people counted by the camera and the number of mobile devices seen by the FTs. The expected result at the end of the stream processing chain is an estimation of people gathered in the area covered by the sensors (cameras and scanners). The numbers of street occupancy in the different locations of the festival are monitored continuously by the public order office, and measures are taken if people at some place exceed the tolerated number. In such situations a good estimation of people is important, however data coming from sensors is not always good enough.

Exact information provided by the processing of sensors' data is crucial for this scenario because it enables the authorities to take decisions to avoid any unwanted outcomes. Figure 2 depicts one standard situation in a street festival. If the number of people exceeds a certain threshold on a small business street, shops and locals occupying parts of the streets are immediately notified to clear the occupied space from any objects(tables and chairs) to make more room and improve the fluidity of movement.

From the aforementioned use case we can derive two main requirements:

- Live data quality assessment: we need to have a real time assessment of the data quality, to make sure that estimations and computations using the data will yield good results.

- Multi-Sensor data quality: since the outcome of any process depends on the combination of different data sources, we would like to ascertain the effects of data quality of one source on the others.

3. RELATED WORK

Many authors have discussed the issue of data quality. In this section, we focus on the most relevant one to our use case. Sheikh et al. proposed a middleware that decouples applications from the data producing sensors [13]. The middleware aims to make up for the quality limitations of the context sensors. This research restricts context information on data about humans, i.e., the focus on human users and not entities in general, whereas we do not want to limit our approach to data generated by humans.

Schiffers [14] defines quality of context as "any information that describes the quality of information that is used as context information".

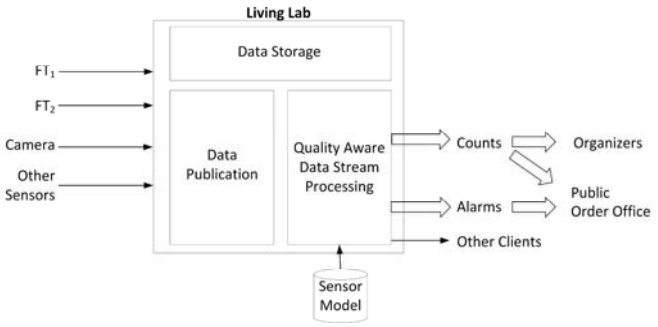

Figure 3: The Living Lab (LL)

Batini et al. offer an interesting view on the data quality dimensions and their respective definitions in [3].

The work of Batini defines clearly dimensions of quality like accuracy and completeness, and provides clear definitions of some of those dimensions. The list of dimensions presented in the work cover all the data dimensions needed.

From the related work described above, we can see that most work in the domain of data quality focuses on describing the quality with the use of different parameters and the value range for each parameter. The research in this area also does not examine the impact of data quality on the processing results and focus on data quality without considering methods to compute it online.

Since sensor data is known for being error prone and its exposure to many disturbances, it has to be collected from multiple sources and integrated together to make up for the lost quality. In the context of data stream processing was the focus initially on Quality of Service. QoS dimensions in data stream processing are classified by Schmidt [11] into time-based dimensions like throughput and latency and content-based dimensions such as sampling rate data mining quality.

Further work on QoS in data stream processing aims at producing Data Stream Management Systems that are quality-aware. In [12] Schmidt et al. develop a deterministic data stream processing system called QStream that offers QoS parameters to users to choose from.

In [1] Abadi et al. propose a dynamic optimization model at the operator level to optimize different QoS metrics across a combined server and sensor network. In [8] Klein and Lehner present a flexible model for the propagation and processing of data quality in a stream processing network for sensor data in a smart environment, where the approach relies on adapting operators.

Geisler et al. [5] proposed an ontology based framework for data quality in data streams. The Data Quality framework is an ontology that manages all Data Quality related metadata. The related work misses the importance of analysing historical data to get insights on the influence of external factors on the accuracy of the data.

From the above we also point out to the importance of our work in the area of data quality in a Living Lab, where much of the processing relies on a combination of data from different sources (sensors). Live data quality assessment enables the Living Lab stakeholders to get insights on the data used for their decision making processes and for the operators of the Living Lab to monitor the infrastructure. Statistical analysis of the historical data can produce knowledge about the influence of external factors on the sensor measurements.

From the related work in the area of data quality in [3] and the work achieved by [9], we can derive the relevant data quality dimensions for our use case. We define each property and give its formula (analog to [3]).

Accuracy we define accuracy as the veracity of values delivered by the different sensors. As data from sensors is error-prone, we want to check continuously the values delivered by the deployed sensors and compute the accuracy of the measurement based on defined conditions of the sensor model.

Completeness is given by the breadth, width and scope of data for the given task. Completeness answers this question: how sufficient is the information provided by the data? Completeness can be described by completeness of schema, completeness of columns, and completeness of population. In our use case we can compute the completeness of the populations of the flowtracks based on the number of people seen in the area.

Since our focus is on live processing of data streams, we have to take a look at the existing work in the area of data quality in stream processing. In our previous work, we implemented quality-aware processing of sensor data in a Data Stream Management System (DSMS) in an automated manner [9]. In our approach the processing results are enriched with additional quality information. The method relies on using an existing ontology to describe sensors and their capabilities and qualities in a sensor model. The information provided by the sensor model is used to compute the data quality continuously. Whereas the focus in the previous work was on determining the quality dimensions of sensor data based on observing sensors and probabilistic models, we try here to use statistics about historical data to find out the factors that influence data quality (like accuracy) and put it into the sensor model.

4. APPROACH

Within the data-management of the LL, we implement quality-aware processing by enriching the data stream queries with partial plans to monitor the data quality. The implementation for this enrichment comes from a sensor data model. In this chapter, we first introduce the general architecture of this approach. Then we give an example for the sensor data model before we show the enrichment of the query plan. We will take the camera as a first step, for which we want to compute the accuracy. The completeness will be addressed in a future work.

4.1 Architecture

As we see in the Figure 3, sensors generate data and send it to the Living Lab, where data stream processing components process the data. Organizers and public order office

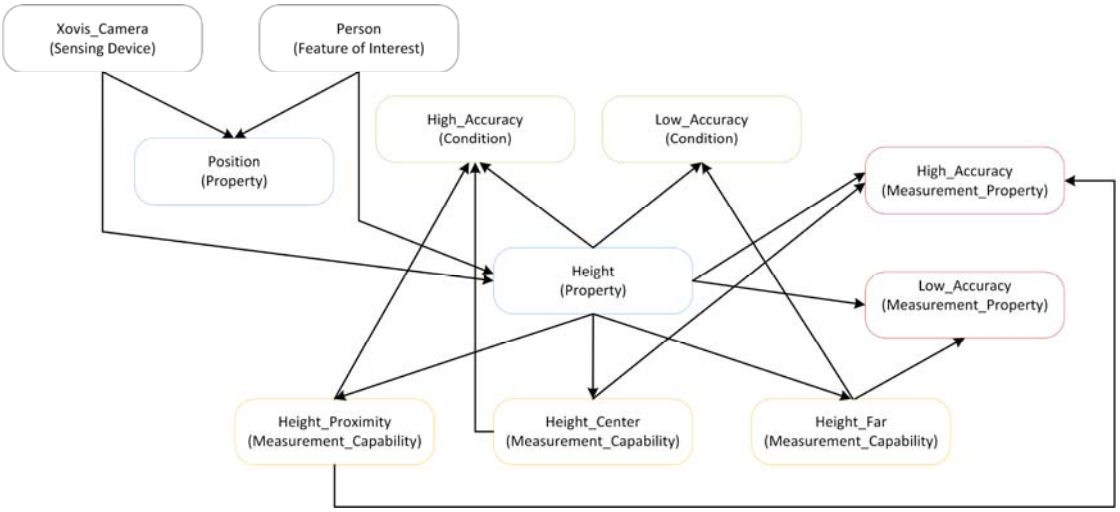

Figure 4: The SSN Model of the camera

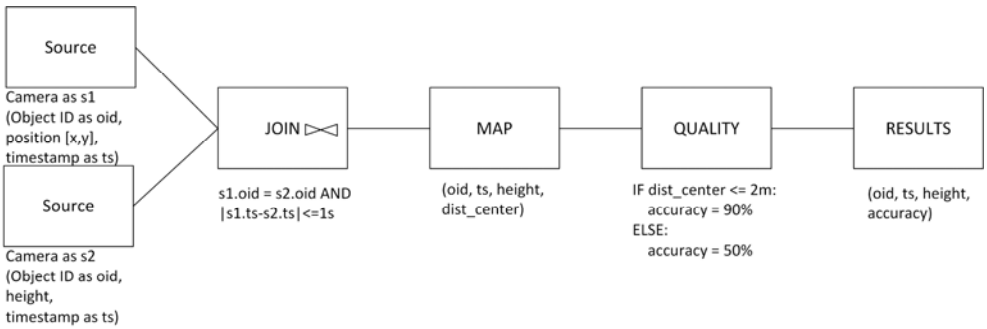

Figure 5: The Quality Query Plan

receive the counts as a result of the processing.

The Data Stream Management System system receives the data and the sensor model information to enrich the query plan with quality-aware operators. The operators can then enrich data with quality properties. The enrichment and query transformation is depicted in Figure 6

4.2 The Sensor Model

To describe the sensors, their capabilities, deployments and their combinations we use Semantic Sensor Network (SSN) ontology [4]. This ontology can specify the survival ranges of sensors and the sensors performance within those ranges. The ontology offers also the possibility to describe the field of deployment of sensors, where the duration and purpose of deployment is indicated.

To enable the quality-aware processing we model the conditions under which a sensor provides which quality. In addition we model if there are sensors that could monitor these conditions. If such sensors exist, we can use them to assess the quality of the main sensor. One example in [9] consists of the combination of a temperature sensor as a main sensor and a position sensor as monitor. The monitor is deployed nearby the main sensor and gives the distance to the nearest object above the temperature sensor. If the distance sensor gives a value of less than 1 meter then the accuracy of the temperature sensor is reduced, i.e, there is an object that covers the temperature sensor and the values produced can no longer be trusted.

For the cameras that we use in our installation, the main factor that influences the quality of the measurement is the distance to the center. As first experiments show, the error in height measurement increases with the distance (see Figure 7). This effect can be described with SSN. Figure 4 shows the SSN model of the camera sensor, which reports the position and the height of every person walking through its deployment area. The height property has a measurement property, which is its accuracy. We can have either high accuracy and low accuracy and both depend on the condition of the distance to the center area of the camera. The height measurement has a low accuracy and a high accuracy and each one depends on the condition set by the position of the tracked person. High accuracy can be achieved when the person is in the center of the camera area or in the proximity area (up to 2 meters). A low accuracy is reported if the person is tracked far from the center area of the camera.

4.3 The Data Stream Management Query Plan

To implement the set of quality operators we follow the ideas of Kuka and Nicklas [10], where the SSN ontology is used to describe the sensors and the quality properties of their observations. We use also a customizable Data Stream

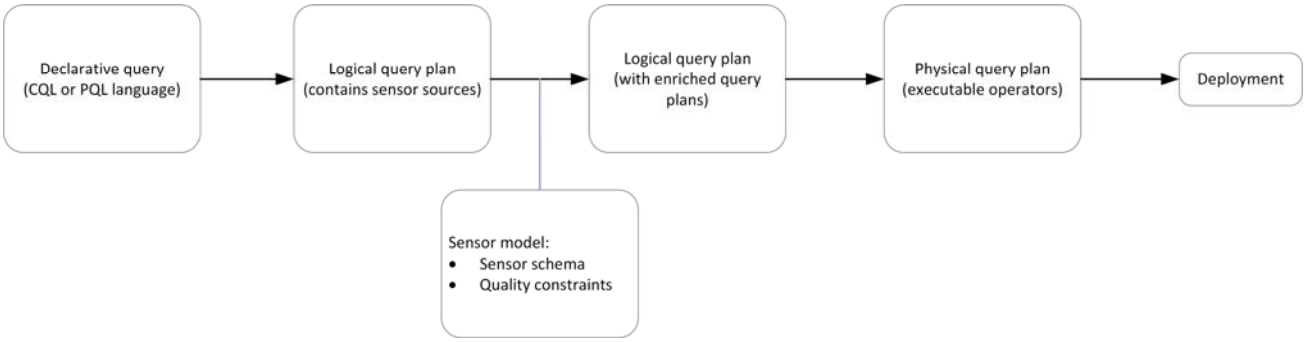

Figure 6: Query enrichment and transformation

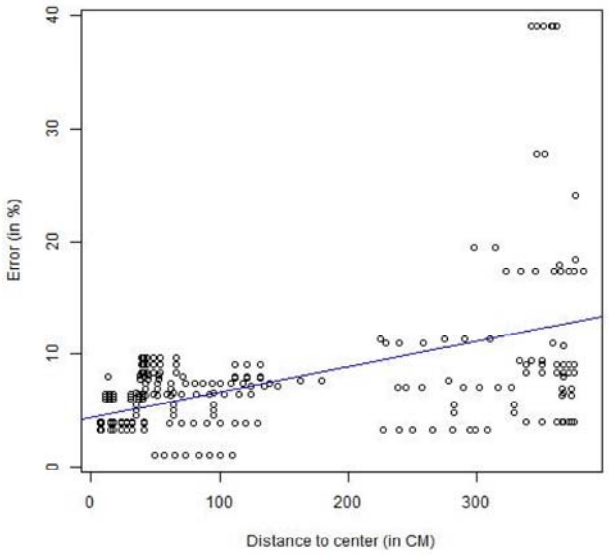

Figure 7: The correlation between the accuracy of height measurements and the distance to the camera: within a distance less than 2 meters the error of height measurements is within 10 percent

Management System called Odysseus [2].

Applied on the sensor defined above, we can use the query plan in Figure 5 to compute the quality of the measurements made by the cameras.

The first node in the query plan gets the data from the camera, and brings it in a format readable by the following operators of the system. The *Join* combines the data from both streams about the same person. The *Map* computes the distance to the center of the camera for each tracked person. The *Quality* operator measures the quality based on the interval function defined by the statistical analysis made on historical data in a learning phase. With this plan, the DSMS component produces quality enriched data. The Living Lab can send the results to its clients or the clients can request them via the web services provided by the Living Lab.

5. CONCLUSION AND FUTURE WORK

This paper presents an approach to assess the quality of sensor data in a Living Lab and to enrich live data with quality meta data. The Living Lab has many use cases that make the idea of offering quality-enriched data very appealing. Our experience with sensor data shows that continuously monitoring the quality of data from sensors has a huge impact on the processing results. The city of Bamberg is home to street festivals that take place on a regular basis. We intend to use those events to test the infrastructure and the quality components to see how well they can perform on a real life situation. Other plans include the use of different sensor types in other use cases to make the quality aware processing cover as many sensor types as possible. We want also to compute the completeness of the data for the flowtracks from our use case.

6. REFERENCES

[1] D. J. Abadi, Y. Ahmad, M. Balazinska, U. Cetintemel, M. Cherniack, J.-H. Hwang, W. Lindner, A. Maskey, A. Rasin, E. Ryvkina, and others. The Design of the Borealis Stream Processing Engine. In *CIDR*, volume 5, pages 277–289, 2005.

[2] H.-J. Appelrath, D. Geesen, M. Grawunder, T. Michelsen, and D. Nicklas. Odysseus: A highly customizable framework for creating efficient event stream management systems. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, DEBS '12, pages 367–368, New York, NY, USA, 2012. ACM.

[3] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] M. Compton, P. Barnaghi, L. Bermudez, R. Garcia-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. L. Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor. The SSN ontology of the W3C semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:25 – 32, 2012.

[5] S. Geisler, S. Weber, and C. Quix. An ontology-based data quality framework for data stream applications.

In *16th International Conference on Information Quality*, pages 145–159, 2011.

[6] B. C. N. GmbH. FlowTrack. www.bluecellnetworks.com/flowtrack/. Accessed: 2016-03-23.

[7] M. Kehoe, M. Cosgrove, S. Gennaro, C. Harrison, W. Harthoorn, J. Hogan, J. Meegan, P. Nesbitt, and C. Peters. Smarter cities series: a foundation for understanding IBM smarter cities. *Redguides for Business Leaders, IBM*, 2011.

[8] A. Klein and W. Lehner. Representing data quality in sensor data streaming environments. *J. Data and Information Quality*, 1(2):10:1–10:28, Sept. 2009.

[9] C. Kuka. *Qualitaetissensitive Datenstromverarbeitung zur Erstellung von dynamischen Kontextmodellen*. PhD thesis, University of Oldenburg, 2014.

[10] C. Kuka and D. Nicklas. Enriching sensor data processing with quality semantics. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 437–442, March 2014.

[11] S. Schmidt. *Quality-of-service-aware data stream processing*. PhD thesis, 2006.

[12] S. Schmidt, H. Berthold, and W. Lehner. Qstream: Deterministic querying of data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 1365–1368. VLDB Endowment, 2004.

[13] K. Sheikh, M. Wegdam, and M. Van Sinderen. Middleware support for quality of context in pervasive context-aware systems. In *Pervasive Computing and Communications Workshops, 2007. PerCom Workshops' 07. Fifth Annual IEEE International Conference on*, pages 461–466. IEEE, 2007.

[14] Thomas and M. Schiffers. Quality of context: What it is and why we need it. In *In Proceedings of the 10th Workshop of the OpenView University Association: OVUA 03*, 2003.