Quality-aware Sensor Data Stream Management in a Living Lab Environment

Aboubakr Benabbas, supervised by: Prof. Dr. Daniela Nicklas
Chair of Mobile Systems
University of Bamberg, Germany
aboubakr.benabbas@uni-bamberg.de
daniela.nicklas@uni-bamberg.de

***Abstract*—Sensor data is error-prone. Developers of pervasive applications must take the limitations of sensors into account when processing the data. To relieve the developers from the task of data cleaning and quality monitoring, we need a set of tools to model sensor data quality and to integrate the quality information into the stream data processing. In this dissertation, the goal is to provide a framework of tools to semi-automatically generate sensor models and stream processing queries for sensors with quality and context information for a quality-aware data stream processing.**

***Index Terms*—data stream processing, data quality, sensors, context**

I. Motivation

The emergence of pervasive applications along the need for scalable and quality-aware data management for complex sensor-based applications give data quality an important role. To offer developers the best possible input and users the best possible output and experience for such applications, sensor data quality has to be evaluated thoroughly. For different sensor platforms, a testing environment is a good place to monitor sensors and evaluate their capabilities in terms of data quality. Such environment is a Living Lab that serves as an ecosystem for collaboration between public and private institutions, to enable data collection and testing for research and development purposes. Since sensor data is known for being faulty and context-dependent, the use of raw sensor data for some application might not satisfy the quality requirements. Even when developers know the limitations of the sensors, the lack of means to describe the sensor data quality and tools to integrate the quality-aware processing makes application development more complex. With the use of multi-sensory data sources, context can be measured and used for data quality estimation. The existence of some expression means like ontologies can be used as leverage to facilitate the semi-automatic integration of data quality-aware processing into applications.

II. Problem Statement

The problem of data quality assessment in many sensor-based applications can be traced back to:

- Lack of information about the context related data quality or not being used in the data processing. In many use cases, data comes from multiple sensors, with one type indicating how another type performs under certain conditions. The non-existence of such context information does not enable the application developers to benefit from such inter-sensor dependency.
- Lack of a framework to describe the sensor quality constraints and integrate them into the data processing, so that developers can be unburdened from the task of dealing with the sensor data quality.
- Lack of tools to allow the developer to quickly integrate data quality assessment and cleansing into their application without the need to develop tailored solutions for different types of sensors.

III. Related Work

Quality-aware Data Stream Management for pervasive applications has been the subject of many research works. The work of Batini [1] defines clear dimensions of quality like accuracy and completeness. Some approaches mainly consider Quality of Service (QoS) [2] [3] [4] [5]. This is a dimension that does not cover the quality of the sensor data itself and is thus orthogonal to this approach.

In [6] [7], quality-aware sensor data filtering and compression of data are proposed to preserve the accuracy of data. These contributions focus either on QoS or the impact of missing data on the results. They do not try to compute quality dimensions such as accuracy of the sensor data and its influence on the processing results.

Considering data quality as context-dependent, Kuka [8] implemented quality-aware processing of sensor data in a Data Stream Management System (DSMS). In this approach the processing results are enriched with additional accuracy value. Kuka's work determines the accuracy of sensors based on probabilistic models, the proposed work in this dissertation uses historical data to find accuracy factors and put it into the sensor model. Geisler et al. [9] proposed an ontology based management for data quality in data streams. The work does not provide any methods to compute the quality metrics.In [10], data quality is measured by latency and delay.

In data stream processing, we see that most approaches regard quality in terms quality of service, description of dimensions or the response time without much emphasis on the input data quality. Furthermore, work in the area of sensor data quality does not consider the relationship between

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

different sensors. Also, the lack of a clear method to describe data quality constraints complicates the task of integrating quality-aware processing into applications. From the above discussion, the importance of our work in the area of data quality is highlighted, where much of the processing relies on a combination of data from different sources (sensors).

IV. Contributions and Approach

The goal of the doctoral research project is to enable scalable and quality-aware data management for complex sensor-based applications with the Living Lab as the main environment. I want to semi-automatically generate sensor models to describe the sensing capabilities of the sensor devices and integrate it into the data stream management system. Next, I want to create tools that take those system models and generate queries for different systems. My work intends to investigate existing methods to compute observable and relevant dimensions of the sensor data quality: namely accuracy and completeness. The choice of these dimensions comes from the inherent characteristics of sensors; either being absent and unreachable, which affects the data completeness, or delivering inaccurate observations of the monitored feature of interest when running outside the specified functional environment. I take the Living Lab as my development environment, where both indoor and outdoor sensors are deployed. In the end of the research project I will have:

- Tools to generate sensor quality models and quality-aware queries for sensors with/without knowledge of the application requirements
- Generate dynamically quality dimension values for the incoming data on the fly
- Integrate the above into the Living Lab platform

The different components of the system are shown in Fig.1, where the *Sensor Quality Management* manages the sensor models and generates the quality-enriched queries. The Data Stream Management System component receives the data from the sensors and runs queries on them to compute the quality values for the desired dimensions. The Applications can take the data with computed quality values and make decisions based on the application and quality requirements.

The research project can be divided into three work packages. The first package provides the *Sensor Quality Management* to generate the sensor quality models and their quality-enriched queries. This part defines the templates for the sensor models and the algebra to generate models and queries from available sensor information. In the second package, we implement methods to compute accuracy and completeness in the sensor data streams. This part examines methods to estimate the values of the aforementioned quality dimensions and how they are expressed in the models and the queries on the DSMS level. The third package gives applications as use cases for experiments and evaluation for each of the above work packages.

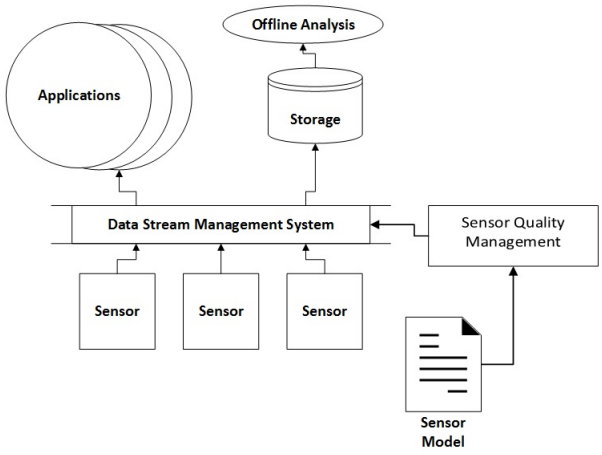

Fig. 1. System Overview

V. Conclusion

The aim of the thesis is to provide tools to enrich sensor data with quality measurements by creating sensor models and quality-aware queries based on sensor specifications. The planned work intends to provide a framework with tools for generating quality-aware sensor models, queries for data stream processing and using quality estimation methods to compute accuracy and completeness of data.

Acknowledgment

I would like to thank my supervisor Prof. Dr. Daniela Nicklas for her continuous support and advice. Many thanks go to Hannes Hornig and Tim Rütermann, with whom I had the pleasure of collaborating.

References

[1] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
[2] S. Schmidt, "Quality-of-service-aware data stream processing," Ph.D. dissertation, University of Dreseden, 2006.
[3] S. Schmidt *et al.*, "Qstream: Deterministic querying of data streams," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 2004.
[4] Abadi *et al.*, "The Design of the Borealis Stream Processing Engine." in *CIDR*, 2005.
[5] A. Klein and W. Lehner, "Representing data quality in sensor data streaming environments," *J. Data and Information Quality*, vol. 1, no. 2, 2009.
[6] Mohomed *et al.*, "HARMONI: Context-aware Filtering of Sensor Data for Continuous Remote Health Monitoring," in *Sixth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Hong Kong, China, 2008.
[7] Cappiello *et al.*, "Quality- and energy-aware data compression by aggregation in WSN data streams," in *IEEE International Conference on Pervasive Computing and Communications*, Galveston, TX, USA, 2009.
[8] C. Kuka, "Qualitaetissensitive Datenstromverarbeitung zur Erstellung von dynamischen Kontextmodellen," Ph.D. dissertation, University of Oldenburg, 2014.
[9] S. Geisler *et al.*, "Ontology-based data quality management for data streams," *Journal of Data and Information Quality*, 2016.
[10] D. Yates *et al.*, "Data quality and query cost in pervasive sensing systems," *Pervasive and Mobile Computing*, 2008.