COMPUTING DATA QUALITY INDICATORS ON BIG DATA STREAMS USING A CEP

Wenlu Yang, Alzennyr Gomes da Silva, Marie-Luce Picard

COMPUTING DATA QUALITY INDICATORS ON BIG DATA STREAMS USING A CEP

Wenlu Yang *

Sorbonne Universités
UPMC Univ Paris 06, CNRS, LIP6 UMR 7606
4 place Jussieu 75005 Paris, France

Alzennyr Da Silva, Marie-Luce Picard

EDF Lab Clamart
1 avenue du Général de Gaulle
92140 Clamart, France

ABSTRACT

Big Data is often referred to as the 3Vs: Volume, Velocity and Variety. A 4th V (validity) was introduced to address the quality dimension. Poor data quality can be costly, lead to breaks in processes and invalidate the company's efforts on regulatory compliance. In order to process data streams in real time, a new technology called CEP (complex event processing) was developed. In France, the current deployment of smart meters will generate massive electricity consumption data. In this work, we developed a diagnostic approach to compute generic quality indicators of smart meter data streams on the fly. This solution is based on Tibco StreamBase CEP. Visualization tools were also developed in order to give a better understanding of the inter-relation between quality issues and geographical/temporal dimensions. According to the application purpose, two visualization methods can be loaded: (1) StreamBase LiveView is used to visualize quality indicators in real time; and (2) a Web application provides a posteriori and geographical analysis of the quality indicators which are plotted on a map within a color scale (lighter colors indicate good quality and darker colors indicate poor quality). In future works, new quality indicators could be added to the solution which can be applied in an operational context in order to monitor data quality from smart meters.

Index Terms— Data quality, Big Data, data stream, CEP, smart grids.

1. INTRODUCTION

Traditional electricity grids lack the capability of providing electricity consumption data in real time. Smart grids [1, 2] remedy these problems by taking advantage of advances in terminal devices (smart meters) and information and communication technologies (ICT). It can gather and react to information about suppliers and consumers in an automated fashion to improve the efficiency, reliability, economics, and sustainability of the production and distribution of electricity. To achieve this goal, many countries invest in the development of smart grids. In Martinique, ERDF [1] has started to install smart meters as part of a pilot program [3]. These smart meters can measure different types of data, such as load curves, indexes, electricity status such as outage and voltage spike. With the capability of transmitting measurements within configurable time intervals (as frequently as 10 min), the volume of data can increase rapidly. At the same time, these data streams can arrive continuously in multiple, rapid, time-varying, possibly unpredictable and unbounded ways, which makes data mining on big and multi-source data streams a great challenge.

Given the intrinsic characteristics of data streams, new technology is needed to handle it. In order to deal with this challenge, a new solution called CEP (Complex Event Processing) was proposed. CEP is an event processing technology that combines data from multiple data streams to infer events or patterns that suggest more complicated circumstances. The goal of complex event processing is to identify meaningful events and respond to them as quickly as possible. A number of CEPs have emerged in recent years: StreamBase[4], EsperTech [5], StreamInsight [6], STORM [7], etc.

To prepare for the smart grid, utilities must address the data quality challenge. One of the major demands for utilities is to quickly locate the region suffering from poor data quality problems and to get to know when these problems occur.

The main purpose of the current work is to create a framework to compute generic quality indicators of smart meter data streams on the fly by taking advantage of CEP technology and to give intuitive geographical and time scale views of the quality issues. This article is organized as follows. In section 2 we present the related work on smart grids. We describe the proposed framework in section 3 by explaining its architecture and characteristics. Section 4 describes the application of our framework to data streams produced by smart meters installed for an experimentation in Martinique. Section 5 presents the conclusion and future work.

*This work was produced during the author's Master internship.
978-1-4673-8457-5/15/$31.00

[1] Électricité Réseau Distribution France, `http://www.erdf.fr/`

2. RELATED WORK

In this section we present the related work on smart grids. In [8], the author focuses on the way smart meters collect data and the quality of this data. The author proposes quality indicators such as value of duplicates, zeros and spikes. The approach proposed helps to get better understanding of residential smart meter data. However, the work does not address the problem of creating real-time computing and visualization of the quality indicators.

In [9], the authors deal with the problem of data quality in a geographical information system (GIS) which highlights the importance and benefit of including geographical information into the data quality analysis. The authors developed an in-house algorithm by using voltage profile correlation analysis to connect errors in the GIS representation of the distribution network topology.

In [10], the authors developed a complex event processing framework for smart grid management. The framework integrates fundamental principles from logic-based programming and multi-agent systems in order to create a CEP for intelligent reliable management of the power system. However, the work did not address data quality issues.

In [11], the authors investigate the data traffic problem in smart grids. In this scenario, huge volumes of data produced by smart meters cannot be fully delivered to the data centers due to limited bandwidth. To deal with this problem, the authors developed a solution for optimizing the traffic flow which allows different levels of data quality.

Despite the current number of existing works on smart-grid design and management, few have incorporated data quality issues into the framework definition. Our work is an attempt to integrate data quality analysis into the smart grid analytic framework, by taking advantage of the emerging CEP technology and providing intuitive visualization modules.

3. DESCRIPTION OF THE PROPOSED SOLUTION

In this section we present the solution developed as a diagnostic approach to evaluate quality issues on data streams produced by smart meters.

3.1. Overview of the solution architecture

Our solution is composed of three main stages: (i) a pre-processing stage which involves ETL (Extract Transform Load) operations and the database construction, (ii) a computation stage during which data quality indicators are calculated by StreamBase, and (iii) a visualization stage which can be produced by a Web interface or by the LiveView module. Figure 1 presents the architecture of the solution.

The framework can follow two paths depending on the application purpose. For both paths, the input data streams

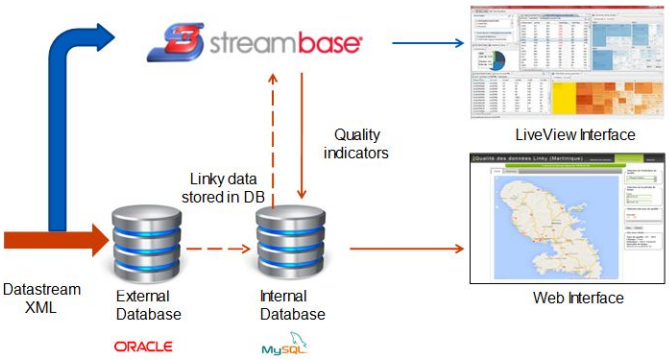

Fig. 1. Architecture of the solution

are the same. Data streams produced by the smart meters are in XML format which needs to be pre-processed into meter reading tuples. Then, the data streams are divided into two paths. In Figure 1, the upper blue arrows represent the path leading to a real-time visualization of the quality indicators by using the Liveview module, while the lower orange arrows present the path leading to a posteriori visualization and analysis of the quality indicators by using a Web interface.

A. Real-time visualization within the LiveView module

In the real-time visualization path, the readings of smart meters are extracted from the XML file to feed StreamBase which calculates the quality indicators on the fly (defined in Section 3.2). The result of the calculation is sent to the LiveView module which provides a real-time visualization of the quality indicators based on dynamic tables and dynamic graphics. Figure 2 shows a screenshot of the LiveView interface.

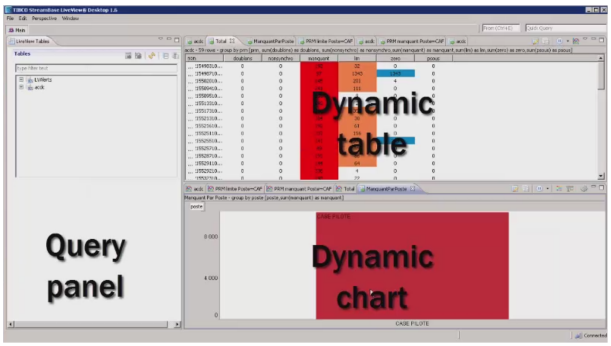

Fig. 2. Liveview interface

In this Figure, queries on the incoming data stream, results of queries in dynamic tables, as well as in dynamic graphics are easily accessible on the interface. In this method, data changes can be detected as soon as they arrive.

B. A posteriori visualization within a Web application

In this method, the extracted meter readings are permanently stored in an Oracle database which is the reference data store for different studies on electricity consumption data

in our Research Department. For our specific work on quality data, we created a dedicated MySQL database. Data of the target clients are extracted from the Oracle database and stored in the MySQL database. StreamBase reads this data, calculates its quality indicators and sends the results to the MySQL database. Thus, the MySQL database stores both the original meter readings and the quality indicators. Then, a Web application supplies an interface to interact with the data stored in the MySQL database and provides a posteriori analysis of quality indicators. Figure 3 shows a screenshot of the Web application.

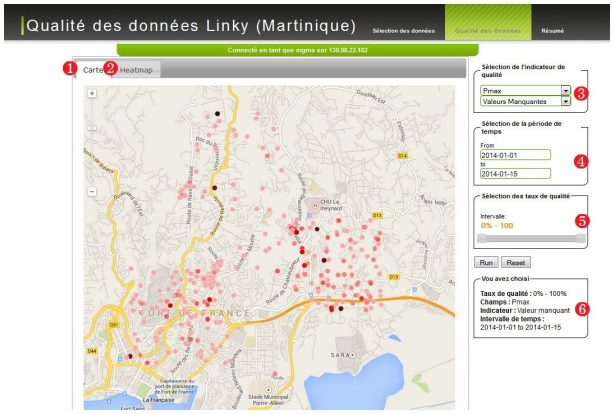

Fig. 3. Web application (Geographical view)

The Web application provides the possibility of visualizing the quality indicators on arbitrary time periods and both geographical and temporal dimensions within a color scale (lighter colors indicate good quality and darker colors indicate poor quality). It can give clues about the underlying cause of data quality issues.

3.2. Data quality indicators

Smart meters measure various types of data: instant power, indexes, electricity status such as outage and voltage spike, etc. Each data type has particular characteristics and arrives at a specific polling rate. For example, instant power readings are measured every 30 minutes and are more frequent than index readings which are measured only once (or twice, depending on the contract subscription option) a day. Table 1 shows an example of the data types produced by smart meters and the associated quality indicators we define in our framework.

Our goal is to define indicators reflecting data quality issues such as missing readings, duplicated readings, unsynchronized readings, warning-level readings, abnormal readings, zero readings, etc. As the quality issues can be interpreted in different ways and the readings from smart meters come at different rates, finding indicators which are able to represent this problem and give intuitive insight of the severity of the quality issues over different time periods is not a trivial question.

Data type	**Polling rate**	**Quality indicators**
Index or accumulated energy consumption (kWh)	1 or 2 reading(s) per day	*- missing values* *- repeated values* *- unsynchronized values* *- decreasing values* *- erroneous values (<0 or >threshold)*
Instant power(kW)	every 30 minutes	*- missing values* *- repeated values* *- unsynchronized values* *- null values* *- values superior to the contract power*
Maximum power within a day (kVA)	1 per day	*- missing values* *- repeated values* *- unsynchronized values* *- null values* *- values superior to the contract power*
Instant voltage(V)	every 10 minutes	*- repeated values* *- unsynchronized values* *- erroneours values ($> threshold_1$ or $< threshold_2$)*
Outage	Occasional (an alert is recorded when an outage occurs).	*- number of occurences* *- average duration*

Table 1. Data types and the associated quality indicators

In this work, we propose a uniform method to calculate different quality indicators. This method simplifies the workflow of processing different data types by providing a unified representation of quality issues. Moreover, the method is able to aggregate the quality indicators on different time scales.

The quality indicators are calculated using a sliding window model. Only the k elements within the sliding window are used to calculate the quality indicators.

Data streams produced by smart meters have a timestamp and a measurement. The timestamp represents the date and time when the measurement was stamped, and the measurement can be a vector of different values. Let $1, 2, 3, ...$ denote the timestamp and $\vec{M}$ denote the measurement vector. The data elements of the incoming data stream at timestamp

i can be represented by $(i, \vec{M}_i)$. We defined n quality indicators represented by a vector $\vec{D}_i$. The n quality indicators of one tuple at timestamp i can be represented as $\vec{D}_i = [D_{(1)}, D_{(2)}, ..., D(n)]$. Each indicator $D_{(j)}$ is calculated by a function $f_j(\vec{M})$ as follows.

$$f_j(\vec{M}) = \begin{cases} 1 & \text{if constraint } A \text{ is satisfied} \\ 0 & \text{if constraint } A \text{ is not satisfied} \end{cases}$$

Constraint *A* is a rule involving the attributes of the measurement. For example, to analyze the problem of unsynchronized measurements, we can take two attributes from $\vec{M}$. Let $rtime$ be the real timestamp and $stime$ be the expected timestamp. The constraint *A* could be: $rtime \neq stime$. If the constraint is satisfied, the quality indicator is set to 1, otherwise it is set to 0. As a result, the quality indicators $\vec{D}$ at timestamp i can be re presented by:

$$\vec{D}_i = [D_{(1)}, D_{(2)}, ..., D(n)] = [f_1(\vec{M}_i), f_2(\vec{M}_i), ..., f_n(\vec{M}_i)]$$

And the output of quality indicators at timestamp i can be presented by $(i, \vec{D}_i)$. This way, instead of treating each data type as isolated, the task consists in finding the corresponding constraint *A* which represents the quality indicator. Another advantage of this method is that the quality indicators can be easily aggregated within a given time period. Figure 4 illustrates the process of calculating quality indicators in a specific time period. Let k be the number of elements within the sliding window of size one day. Then, the quality indicator of day m is calculated as:

$$\vec{Dm} = \sum_{i=1}^{k} \vec{D}_i$$

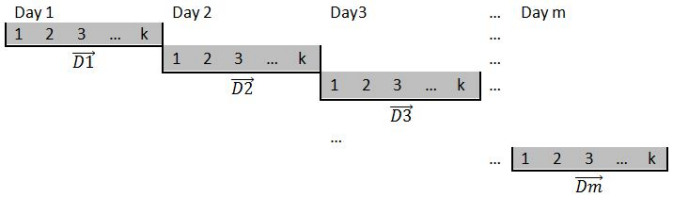

Fig. 4. Computation of the quality indicators within a sliding window

In our framework, the quality indicators are computed by StreamBase.

4. CASE STUDY

In this section, we describe the experiments realized on electricity consumption data produced by smart meters installed in Martinique [3]. The data concerns 1020 individual clients and 15 substations in 4 different districts. Five types of data (instant power, maximum power, index, instant voltage, outage) were recorded from July 2013 to May 2014.

A. Real-time visualization

We simulate the arrival of meter readings from data stored in the Oracle database. After configuring the connection between the Oracle database, the StreamBase CEP and the Liveview module, we can dynamically visualize in LiveView the incoming quality indicators calculated by StreamBase. We can then set different rules to highlight tuples satisfying specific conditions. We can also execute online queries on dynamic tables and visualize the results on the fly. For example, if we are interested in detecting missing values, we can set a rule to change the color of a table cell if a condition is satisfied (e.g. number of missing values superior to a given threshold). We can also calculate aggregated values (e.g. number of missing values per district). The result can be shown in both dynamic tables and dynamic graphs. This interface provides a powerful dashboard which gives a general overview of data quality for monitoring purposes.

B. Web application visualization

After logging into the Web application, we can specify different parameters (districts, electricity contract options, the time period, the quality indicator we want to visualize, etc.) in order to extract data from the MySQL database. Then, a query is executed in the background in order to launch the data. We can choose to visualize the results on map, on a time table or in a summary view. For example, in the geographical view, we can get a better understanding of which regions suffer from bad quality problems.

5. CONCLUSION

In this work, we developed a diagnostic approach to compute generic quality indicators of data streams on the fly. We proposed a uniform method to calculate different quality indicators based on different data types. The implementation is based on Tibco StreamBase CEP. Visualization tools were also developed in order to give a better understanding of the inter-relation between quality issues and geographical/temporal dimensions.

Although the current number of clients equipped with smart meters in our experiment is relatively small ($\sim$ 1020 clients), the solution proves to be efficient to represent the quality issues. This framework is appropriated for processing and aggregating quality indicators on different data types and time scales. In the near future, we intend to apply this solution to a larger number of massive data streams provided by the upcoming generalized deployment of smart meters in France.

Future work may also include: adding other dimensions (weather, aging of electrical equipment, number of interventions in the electrical network, etc.) to the quality analysis and developing a data mining module which can give insight into poor data quality causes.

6. REFERENCES

[1] BB Alagoz, A Kaygusuz, and A Karabiber, "A user-mode distributed energy management architecture for smart grid applications," *Energy*, vol. 44, no. 1, pp. 167–177, 2012.

[2] Seth Blumsack and Alisha Fernandez, "Ready or not, here comes the smart grid!," *Energy*, vol. 37, no. 1, pp. 61–68, 2012.

[3] Joseph Maire, Laure Chossegros, and Gilles Rondy, "Edf evaluates smart metering in martinique: Expectations & local review," in *23rd International Conference on Electricity Distribution (CIRED)*, Lyon, France, 2015.

[4] "Tibco streambase," `http://www.StreamBase.com`.

[5] "Espertech," `http://www.espertech.com`.

[6] Mohamed Ali, Badrish Chandramouli, Jonathan Goldstein, and Roman Schindlauer, "The extensibility framework in microsoft streaminsight," in *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, Washington, DC, USA, 2011, ICDE '11, pp. 1242–1253, IEEE Computer Society.

[7] "Apache storm," `https://storm.incubator.apache.org/`.

[8] Juan Shishido, "Smart meter data quality insights," in *ACEEE Summer Study on Energy Efficiency in Buildings*, 2012.

[9] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *Smart Grid, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[10] Srivathsan Srinivasagopalan, Supratik Mukhopadhyay, and Ramesh Bharadwaj, "A complex-event-processing framework for smart-grid management," in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2012 IEEE International Multi-Disciplinary Conference on*. IEEE, 2012, pp. 272–278.

[11] Miriam Allalouf, Gidon Gershinsky, Liane Lewin-Eytan, and Joseph Naor, "Smart grid network optimization: Data-quality-aware volume reduction," 2014.