Big Data in Large Scale Intelligent Smart City Installations

Sylva Girtelschmid
Institute of Telecooperation
Johannes Kepler University
Linz, Austria
sylva.girtelschmid@jku.at

Matthias Steinbauer
Institute of Telecooperation
Johannes Kepler University
Linz, Austria
matthias.steinbauer@jku.at

Vikash Kumar
The Telecommunications
Research Center Vienna
Vienna, Austria
kumar@ftw.at

Anna Fensel
The Telecommunications
Research Center Vienna
Vienna, Austria
fensel@ftw.at

Gabriele Kotsis
Institute of Telecooperation
Johannes Kepler University
Linz, Austria
gabriele.kotsis@jku.at

ABSTRACT

This paper highlights how the domain of Smart Cities is often modeled by ontologies to create applications and services that are highly flexible, (re)configurable, and inter-operable. However, ontology repositories and their accompanying reasoning and rule languages face the disadvantage of bad runtime behavior, especially if the models grow large in size. We propose an architecture that uses tools and methods from the domain of Big Data processing in conjunction with an ontology repository and a rule engine to overcome potential performance bottlenecks that will occur in this scenario.

Categories and Subject Descriptors

C.2.4 [**Distributed Systems**]: Distributed Applications; H.2.4 [**Systems**]: Rule-based databases

General Terms

Design, Experimentation, Performance

Keywords

Energy Efficiency, Smart City, Ontology, Semantic Modeling, Real-time Streaming, Big Data

1. INTRODUCTION

As a response to the urgent call to lower emissions of carbon dioxide, the Smart Grid concept, closely tied to the concept of Smart Building/Home and eventually Smart City, all aimed at improving energy efficiency, has emerged.

A wide array of commercial product solutions, supporting the adoption of Smart Grid technology among energy providers [10] as well as Smart Building technology among building constructors and managers are already available on the market [11]. Although these products are a sound starting point to achieve energy savings, the vision of a Smart City reaches much farther, including broader resource integration, increased ease of deployment and operation, automation, policy sharing, etc. Moreover, the commercially available Smart Building Analytics Technologies are known to be cumbersome to setup and reconfigure, making them less likely to get adopted [15]. The main priority in further development[1] is to improve the user experience. A wide area of research has formed around the Smart Grid/Building/Home concept, for readability purposes, further generalized and referred to as Smart City. The application of semantics is one such initiative.

Making use of semantic modeling to describe resources in various domains has been explored in numerous projects focusing among other topics also on Smart Cities [7, 16, 5]. The widespread adoption of semantic technologies in the Smart City domain stems from the fact that it enables flexibility in system configuration and adaptation. Additionally, it can provide intelligence via reasoning over the system. This concept allows for interoperability of diverse distributed system components, such as sensor devices, smart meters, smart plugs, etc.. It can also provide the basis for higher-level interoperability such as rule translation and policy sharing [5]. As a result, an ontology-based semantic model has become prevalent in the area of context aware and Smart City technologies [1, 12, 14]. Nonetheless, semantic processing is not capable of managing such vast amounts of data, as are expected in the Smart City scenario, in realtime [1, 2, 3, 8]. Furthermore, in [10], the call for application of Big Data technology in Smart Grid analytics is emphasized and, as suggested in [11], cloud delivery of energy management solutions is largely becoming an accepted approach.

We plan to contribute to this research by proposing a novel system architecture for Smart City applications which employs ontology reasoning and distributed stream processing framework on the cloud. With our approach, the decision making process is fully automatic and self-contained and, at the same time, the system remains robust and time efficient even in a large scale domain.

The remainder of this paper is organized as follows: in section 2, we summarize current research related to the topics of ontology-based Smart City and processing of Big Data.

iiWAS'13 , December 02-04 2013, Vienna, Austria
Copyright 2013 ACM 978-1-4503-2113-6/13/12 ...$15.00.

[1] IMS Research - `http://www.imsresearch.com/`

Section 3 describes our proposed architecture and gives requirements for a real-world implementation. The prototype we have built to prove the concept is shown in section 4 and the future work and concluding remarks can be found in sections 5 and 6, respectively.

2. RELATED WORK

This work relates to other fields of research mainly in two areas (1) an ontology-based reasoning in the Smart City domain and (2) the processing of Big Data. In this section we briefly introduce the concepts behind Smart City and cover related research. At the end, the novelty of our architecture approach is explained.

2.1 Smart City

In the conventional power distribution systems, the customer (e.g. household) is viewed as a passive consumer of energy. With the Smart City paradigm, however, comes a different scenario, in which power plants need to interconnect with customers (acting as prosumers) and with other distributed renewable energy sources. The challenge in coordinating this communication and fostering energy savings and reuse calls for ICT[2] technologies to be applied.

Much research, concerning interactions of Smart Houses and Smart Grids, has already been conducted and many related projects have emerged. The report in [4] presents a summary of projects from around the world. In many cases, the work is also carried out by the industry. These initiatives promise to shed more light on the challenges of integrating heterogeneous participants and the ability to handle the large amounts of data needed to be reasoned over. However, the progress is rather slow, largely due to privacy and security issues and the power grid's high reliability needs. Given the fact that at this point there is little available to capitalize on in terms of the actual home-grid data interchange, we would like to contribute to the advance in the research by proposing a solution that we claim is general enough to support the energy coordination in a Smart City. More specifically, we claim our solution to be: (1) adaptable to a wider range of concepts from additional domains (such as infrastructure, weather, etc. [16]) and (2) applicable to the coordination of the new power grids (often micro grids) and their distributed nature (so far in the research most successfully implemented with the multi-agent system (MAS) technology [7]). This leads us to two major requirements for our design: (1) follow the direction of previous research and base our design on semantics in order to allow for seamless communication between diverse systems and (2) satisfy a sufficient realtime response of such a complex system by applying Big Data technology. The next two paragraphs address the research related to these two requirements.

2.2 Ontology

Within the realm of computer science, ontologies are formal representations of concepts within a specific domain. Such models describe the relationships between pairs of these concepts, providing a common vocabulary for the given domain, and consequently, eliciting knowledge sharing. Ontologies can also include reasoning rules allowing processing of knowledge and deriving new information via inference [13].

The use of semantics in the Smart City problem domain has been researched in numerous works. For example, in [9], the work focuses on semantic-based architecture communication by applying standards of ICT. The author identified standards and specifications necessary for seamless data exchange among Smart Grid devices and systems and proposed a solution which relies on annotating semantic meta data to services in order to allow server and clients to share information.

Penya *et al.* [7] apply semantic tools in their distributed architecture approach within the ENERGOS project, and propose that a unique global ontology, based on the standards from CIM[3], should be sufficient to represent the Smart Grid domain as a whole.

In [16], the authors integrate a Smart Grid information model and apply it to a demand response (DR) optimization application using complex event processing (CEP). A number of ontologies (electrical equipment, organization, infrastructure, weather, and spacial and temporal ontologies) are integrated to represent a complete Smart Grid and the relationships among them have also been defined.

In our previous work ([5]), on which we base our current research, a semantically enabled Smart Building system was implemented, combining home automation techniques and data from smart meters, smart plugs, and sensors. Complex rules and policies were created to monitor and administer the centrally stored data that was updated in near realtime. The system was deployed in two real-life buildings (a school and a factory floor) and data were collected over a period of several months resulting in almost 10 million triples.

In summary, the use of ontologies has proven to be inevitable for Smart Building/Grid applications. However, some authors consider the use of ontologies in context aware environments to be very sensitive to the size of the dataset and unreasonable to use for time-critical applications [8]. Similarly, in [3], the authors have reported a limitation in semantic processing, where the reasoner and rule engine would not be able to perform in realtime when applied to the full semantics model. We too are skeptical about the performance in scenarios of Smart City where the context data are of large size, constantly changing, and often incomplete. We turn to Big Data to remedy these issues.

2.3 Streaming Big Data

According to [10], the amount of sensors in Smart Grids, combined with those in Smart Homes/Buildings, will vastly increase the data influx in the near future. Moreover, the coordination of energy in Smart Cities is highly time critical and dependent on reliable data [14]. This leads us to a situation in which the stream processing tools from Big Data technology are needed to ensure efficient processing of the generated data [6].

Realtime streaming platforms are tools in Big Data that are able to handle large volumes of data, arriving to the system at high velocities, by using compute clusters to balance the workload. Those systems inherit some properties of MPI[4] clusters but add scalability to the feature set. They are able to rebalance the workload if too many messages need to be processed in a certain compute node. There are three systems that can be considered mature enough for produc-

[2] Information and Communications Technology

[3] Common Information Model

[4] Message Passing Interface

tive environments: Project Storm[5] (developed at Twitter), S4[6] (developed at Yahoo), and Project Spark[7]. All these systems have in common the ability to reliably process events or messages on a distributed compute clusters. For all messages entering the system, they guarantee processing even if some of the compute nodes in the cluster fail. Finally, all systems can be dynamically reconfigured, making it possible to adjust the size of the cluster during runtime.

Our work stands apart from the presented related work in the combination of ontology-based reasoning with Big Data streaming methodology. We approach the integration of these two technologies by off-loading the basic data processing tasks (data cleansing, broken sensor detection, normalizing, threshold alerts, etc.) to the compute cluster and work with a reduced dataset (in terms of volume and throughput) in the ontology repository. As a result, we read from and, most importantly, write to the ontology only when necessary.

3. PROPOSED ARCHITECTURE

Based on what we found in literature [10, 11, 2, 8] and our initial proof-of-concept performance testing on ontology repositories, we have concluded that moving Smart City applications (which heavily make use of ontologies and rule-based reasoning) to the cloud and integrating it with a real-time computation platform is desirable. On this account, we suggest an architecture, featuring components for realtime processing and reasoning.

In the following, we discuss the proposed architecture as shown in figure 1. The main components: streaming platform, ontology repository, rule engine, and possible client applications are displayed. The figure also shows the general flow of information in the system.

Sensor data originating from any sensor found in a Smart City (e.g. smart meters, smart sensors, etc.) are sent to a realtime streaming platform. This platform is assembled into a cluster from many individual compute nodes. Due to the inherent feature of stream processing engines[8], each of these nodes is handling streams of sensor data from arbitrarily many sensors, depending on the amount of data the individual sensors are generating. The streaming platform component is responsible for detecting any considerable changes in sensor data readings or failures in sensors, and for accumulating sensor readings where applicable. Further, data cleansing processes (handling of outliers, temporary sensor outages, normalization, calibration, etc.) are applied within the compute cluster to free underlying components from these tasks. Each node in the cluster can directly access the ontology repository to be able to make changes to it.

The ontology repository records the most recent readings and accumulated information such as the average temperature of the last hour or the average power consumption of the last week, etc. Any changes in the ontology are triggering the rule engine to re-evaluate rules and take actions, if applicable.

Within the rule engine, a set of rules are stored and executed on a timely basis upon user request and whenever the compute cluster is signalling a change event. The rule engine has several means for output of rule execution results. 1) The engine can raise alerts. These are stored in the ontology repository for later querying[9] and can be sent to clients that have registered to receive alerts of a certain category. Further, alerts are used to adapt the preprocessing that is performed in the compute cluster (update sensor readings more frequently, stop reading certain sensors, etc.). 2) The outcome of rule execution can also just be an adaption of the preprocessing process (dependency between sensors has changed, sensor needs to be read less frequently, error correction needs to be adapted, etc.) in order to adapt to new situations. 3) Last but not least, the rule execution results can just be sent back to clients, especially if rule execution was initiated by a client application.

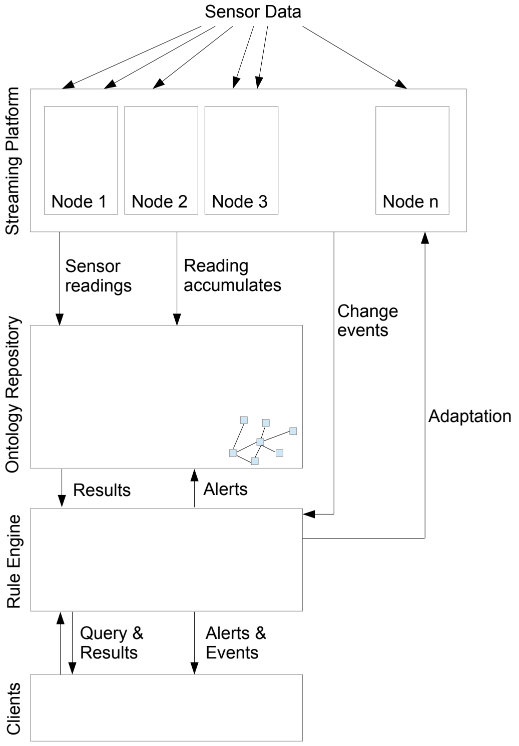

Figure 1: High-level overview of the Smart City data processing architecture

Finally, client applications are able to register to receive results of rules stored in the rule engine (alerts, events, etc.), send queries to the rule engine for one time execution, and store new rules in the rule engine for regular or event driven execution.

4. PROTOTYPE

Although S4 would provide the advantage of integrating

[5] `http://storm-project.net`

[6] `http://incubator.apache.org/s4/`

[7] `http://spark-project.org`

[8] Working on streams of messages/events, they are able to balance these streams between compute nodes in the cluster.

[9] Especially useful when clients are polling the repository for alerts, which is the case in our earlier work. These legacy clients therefore remain compatible.

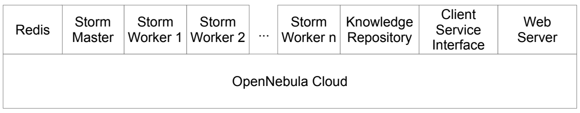

Figure 2: Stack of the components as setup on top of OpenNebula

with Spring configuration[10] and Spark would already provide tight integration with other Big Data tools such as Hadoop and Hive, we have chosen to use Storm as the basis for the prototype implementation. The deciding factor for our choice was the availability of a simple administrative interface for monitoring the cluster and the possibility to specify upper bounds for parallelism. This, in our case, can be used to vary the cluster size for running experiments.

As already explained to some extend in the previous section, Storm is a distributed and fault-tolerant realtime computational platform. It knows two basic processing primitives: Spouts and Bolts. Spouts are used to stream data to the system. They connect to the data sources, in our case to Redis[11]-backed sorted sets of sensor readings and pass this data encapsulated in so called Tuples to the processing units (Bolts). These Bolts are able to consume and emit streams of Tuples. They build the main building blocks of a Storm system. It is in these Bolts where our preprocessing takes place. Spouts and Bolts are tied together in a so called Topology that describes the flow of messages within Storm. It allows to specify groupings, and the number of concurrent instances for a certain Bolt or Spout can be bounded in the Topology.

The ontology repository, describing a representational subgroup of facets typical for a Smart City model, is mostly reused from the previous work of [5], where ontologies of Smart Buildings and the connector software talking to the hardware (sensors, smart meters, smart plugs, etc.) were already put to test. However, for our prove of concept in this work, we implement our ontology as a set of simple Java classes as is required by our rule engine. The repository does not store extensive data (sensor readings, timings, etc.) as was the case in [5]. Instead, only the most recent readings and accumulated information are being recorded, effectively reducing the work complexity done by the reasoning engine.

As our reasoning engine in the prototype, we are using the Drools Fusion module of Drools from JBoss[12]. Drools is an object-oriented business rule management system (BRMS) with a forward chaining inference based rules engine, i.e. a production rule system. Drools uses an enhanced implementation of the Rete algorithm and is very flexible when it comes to adapting to any problem domain. Drools Fusion is the Drools module for enabling complex event processing capabilities, offering features such as temporal reasoning and reasoning over an absence of events[13]. We have chosen this rule engine thanks to it being rated as one of the fastest and most flexible open source rule engine available as of now, which allowed us to rapidly build our test scenarios.

All of our components are running inside of virtual machine containers on top of an OpenNebula[14] cloud installation set up at one of our cloud computing labs. This cloud is featuring 24 CPU cores 72 GB of RAM which allows us to deploy test installations and to conduct experiments with varying cluster sizes. All virtual machines are connected to a 1GBit switch in bridged networking mode which sets the upper bound of our internal bandwidth limit. A diagram illustrating the setup is shown in figure 2.

Being fault-tolerant and scalable, Storm guarantees that each emitted Tuple is processed by the correct order of Bolts and the cluster adapts to the addition and removal of compute nodes dynamically. Thanks to these properties, it is possible to seamlessly grow the system further (from just handling data for a number of buildings up to the size of whole cities) without drastically affecting the performance of the system.

4.1 Evaluation

As a smaller scale real-life experiment, we have streamed an inflated version of previously collected data ([5]) at high velocity to our cluster, in order to simulate the activities in a virtual Smart City.

The simulation to test our prototype is driven by a configurable Smart City model created anew for each test run. This simple model takes care of generating energy consuming entities (for the prototype testing so far, it generates buildings with sensors and appliances in their rooms). This model can be readily extended to include additional sensor data or other entities that play a role in a Smart City, once their data are available to us. From our previous work, we have temperature, humidity, and lighting data for a number of rooms as well as information from which presence can be deduced. The sensor data arriving at the Spouts are either directly forwarded (going through a no-action Bolt) to update the ontology and reevaluate rules or they are preprocessed in a number of Bolts. In the later case, the ontology is updated only selectively, causing the rules not to be evaluated as often. Every ontology update triggers the rule engine, which is making simple decisions for every building. For example, it evaluates a given rule to be true for a given building if temperature reaches a certain threshold. It also checks for presence (whenever presence is needed in the decision making process) and then triggers actions accordingly (such as turning on/off a given heating appliance or raising an alert).

We have been able to observe that for scenarios in which a small number of buildings are handled the rule engine performs satisfyingly even without the application of any prior filtering but has significant improvement when preprocessing is employed. As we have moved to larger setups, the performance was gradually deteriorating when no filtering was used whereas with the streaming platform filtering it was still satisfying.

Additionally, one has to take into account the fact that in a real-life application the rules are likely to be much more complex as well as more numerous making the case for the use of preprocessing within a streaming platform even more sound. We also believe that our simulation proves the point even though the data used are not directly related to the data that would need to be processed in a real-life Smart City system. We claim this relevant since in both cases the streaming data are essentially timestamp-value Tuples.

[10] http://www.springsource.org

[11] http://redis.io

[12] www.jboss.org/drools/

[13] http://www.jboss.org/drools/drools-fusion.html

[14] http://www.opennebula.org

5. FUTURE WORK

The prototype shows that our approach is both feasible for the problem scope and scalable to larger problem sizes. Still, there are issues that need to be addressed to make this architecture applicable in real-world scenarios.

One such matter is the lack of any security restrictions in both our architecture and implementation. This was omitted on purpose due to time constraints. However, since high standards in security and privacy are by all means indispensable [14], we plan to integrate security and privacy mechanisms in updated versions of this architecture.

Our project is still in an early stage and thus, we are lacking performance evaluation in large scale setups. The problem we face is that there are no complete datasets publicly available in the Smart City category that possess Big Data properties. As a future improvement to our system, we plan to extend our city model to simulate more complex reasoning scenarios by including additional datasets to reason over, such as weather and traffic data.

In continuation, we are also planning to devote substantial attention to improving the streaming platform component where more advanced filtering, failure detection, and data cleansing needs to take place.

6. CONCLUSION

In this paper, we have proposed and presented an architecture for efficient processing of sensor data from Smart City installations. We concluded that in such larger scale scenarios, the influx of data, needed to be processed in order to optimize energy usage, requires smart reasoning mechanisms, such as ontologies, to live up to its full potential. Since the current ontology-based knowledge databases would cause performance bottlenecks in large scale installations, we have addressed this matter by combining ontology-based reasoning with Big Data processing. Namely, our architecture uses Big Data streaming clusters for basic processing needs, while the time consuming ontology driven reasoning is applied only when necessary.

7. ACKNOWLEDGMENTS

This work is supported by the Austrian government and the City of Vienna within the competence center program COMET. It is part of the joint strategic project “Analysis on semantically structured data from smart buildings and smart grids” which is carried out by the Telecommunications Research Center Vienna (FTW) and the Institute of Telecooperation at the Johannes Kepler University Linz.

8. REFERENCES

[1] A. Crapo, R. Piasecki, and X. Wang. The smart grid as a semantically enabled internet of things. In *Implementing Interoperability Advancing Smart Grid Standards, Architecture and Community: Information Interoperability Track*, Grid-Interop Forum 2011, 2011.

[2] M. de Mues, A. Alvarez, A. Espinoza, and J. Garbajosa. Towards a distributed intelligent ict architecture for the smart grid. In *Industrial Informatics (INDIN), 2011 9th IEEE International Conference on*, pages 745–749, 2011.

[3] A. Espinoza, M. Ortega, C. Fernandez, J. Garbajosa, and A. Alvarez. Software-intensive systems interoperability in smart grids: A semantic approach. In *Industrial Informatics (INDIN), 2011 9th IEEE International Conference on*, pages 739–744, 2011.

[4] V. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G. Hancke. Smart grid and smart homes: Key players and pilot projects. *Industrial Electronics Magazine, IEEE*, 6(4):18–34, 2012.

[5] V. Kumar, A. Fensel, G. Lazendic, and U. Lehner. Semantic policy-based data management for energy efficient smart buildings. In P. Herrero, H. Panetto, R. Meersman, and T. Dillon, editors, *On the Move to Meaningful Internet Systems: OTM 2012 Workshops*, volume 7567 of *Lecture Notes in Computer Science*, pages 272–281. Springer Berlin Heidelberg, 2012.

[6] C. Lynch. Big data: How do your data grow? *Nature*, 455(7209):28–29, 2008.

[7] A. Pena and Y. Penya. Distributed semantic repositories in smart grids. In *Industrial Informatics (INDIN), 2011 9th IEEE International Conference on*, pages 721–726, 2011.

[8] W. Qin, Y. Shi, and Y. Suo. Ontology-based context-aware middleware for smart spaces. *Tsinghua Science & Technology*, 12(6):707 – 713, 2007.

[9] S. Rohjans. A standard-compliant ict-architecture for semantic data service integration in smart grids. In *Innovative Smart Grid Technologies (ISGT), 2013 IEEE PES*, pages 1–6, 2013.

[10] C. L. Stimmel and B. Gohn. Smart grid data analytics: Smart meter, grid operations, asset management, and renewable energy integration data analytics: Global market analysis and forecasts. *Pike Research: A Part of Navigant*, Research Report(Executive Summary):1–16, 3Q 2012.

[11] C. Talon, S. Jaffe, and R. Nicholson. Idc marketscape: Worldwide smart building energy analytics 2011. *IDC Enrgy Insights: Distributed Energy Strategies*, Vendor Assessment(E1230178):1–19, September 2011.

[12] S. Tomic, A. Fensel, M. Schwanzer, M. K. Veljovic, and M. Stefanovic. Semantics for energy efficiency in smart home environment. In *Applied Semantic Web Technologies*. CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Fl 33487-2742, 2012.

[13] M. Uschold and M. Gruninger. Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, 33(4):58–64, Dec. 2004.

[14] A. Wagner, S. Speiser, and A. Harth. Semantic web technologies for a smart energy grid: Requirements and challenges. In *ISWC 2010*. Springer, November 2010.

[15] D.-Y. Yu, E. Ferranti, and H. Hadeli. An intelligent building that listens to your needs. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 58–63, New York, NY, USA, 2013. ACM.

[16] Q. Zhou, S. Natarajan, Y. Simmhan, and V. Prasanna. Semantic information modeling for emerging applications in smart grid. In *Information Technology: New Generations (ITNG), 2012 Ninth International Conference on*, pages 775–782, 2012.