Towards User-friendly and High-performance Analytics with Big Data Historian

Martin Possolt[1], Václav Jirkovský[1], Marek Obitko[2]

[1]Czech Institute of Informatics, Robotics and Cybernetics
Czech Technical University in Prague
Zikova 4, Prague, Czech Republic
[2]Rockwell Automation Research and Development Center
Argentinská 1610/4, 170 00 Prague, Czech Republic

{martin.possolt,vaclav.jirkovsky}@cvut.cz
mobitko@ra.rockwell.com

Abstract. We are witnessing the trend of increasing data production in various domains including industrial automation. This trend requires means for data capturing, storing, and analyzing. Furthermore, a versatile data model is needed to enable easy knowledge representation as well as change management. In this paper, we utilize Semantic Big Data Historian, which can cope with previously mentioned requirements, for a demonstration of promising analytic approach combining Big Data methods and a user-friendly modular platform. The approach is demonstrated on data from a hydroelectric power station. The station has been dealing with the interesting problem of prediction when to momentarily stop their turbine to increase generated power after the restart. In this contribution, we discuss several approaches how to process and analyze data from power station sensors for achieving the best results.

Key words: Multilayer perceptron, Big Data, Ontology, Hydroelectric power station.

1 Introduction

Nowadays, we are witnessing the trend of increasing data production in various domains including industrial automation problem of data processing in industrial automation domain. This trend requires means for data capturing, storing, and analyzing. Many companies are facing the problem of processing this huge amount of data and a company capability to process data in the shortest time, to provide faultless processing, and to derive new previously unknown knowledge represents a competitive advantage. These data are produced by sensors and machines from a shop floor as well as by high-level systems, e.g. MES[1] and ERP[2].

[1] Manufacturing Execution Systems
[2] Enterprise Resource Planning

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 27-30.

The essential requirement is the proper understanding of given data models (sensors, machines, etc.) together with an understanding and a utilization of knowledge coming from various surrounding systems across a factory or external data sources. This requirement may be expressed as semantic integration problem [1], and the suitable solution is the employment of Semantic Web technologies and model description in ontologies [2].

We proposed and developed Semantic Big Data Historian (SBDH) [3] to cope with previously mentioned requirements. SBDH was designed to overcome common deficiencies of a legacy historian software which is usually optimized to allow fast and compressed storage of data. However, a legacy historian software does not pay much attention to analytics nor to heterogeneous data models integration. Thus, SBDH stores data according to a global data model in the form of ontology knowledge base and particular data are represented as RDF triples. This approach facilitates proper understanding of given data, and subsequently, it enables better data integration as well as data querying. On the other hand, a read/write time could be a bottleneck in the case of semantic data description and therefore SBDH stores RDF triples corresponding to time series in the form of Hybrid SBDH Model [4]. The historian architecture is divided into four layers – data acquisition (collects data from sensors and other related systems), transformation layer (transforms data to the unified semantic form according to the ontology), data storage layer (reads and writes data from storage – implemented by means of Apache Spark[3] and Apache Cassandra[4]), and analytic layer (represented by Apache Spark and KNIME[5]).

KNIME Analytics Platform is an open source data analysis toolbox. It offers various components for machine learning and data mining from data preprocessing, through modeling and data analysis to visualization. One of its biggest advantages is a graphical user interface which allows building a workflow very easily by connecting nodes, each with single operation task. There is no need to have any programming background. An example of a workflow will be illustrated later.

2 How to Analyze Big Data from Industrial Automation

The fast, high-quality, and versatile approach to coping with analytical tasks over production data may mean a significant advantage for many manufacturing companies. Nowadays, the Big Data paradigm and Cloud technologies become widely accepted by many companies and we can utilize such technologies for facilitating our task. Moreover, a processing of data stored in the form of ontologies (i.e. RDF triples) brings bigger demands for data handling. We have tried to overcome this issue by storing RDF triples in specialized structures (e.g. Hybrid SBDH Model [4]) but still RDF data handling is a very demanding task. Thus, we separated the processing of an analytical task into two layers – processing of all data with the help of Apache Spark;

[3] http://spark.apache.org
[4] http://cassandra.apache.org
[5] https://www.knime.org

and a utilization of KNIME for enabling the user-friendly analytic interface, where a user may conduct very complex task without any complex programming skills.

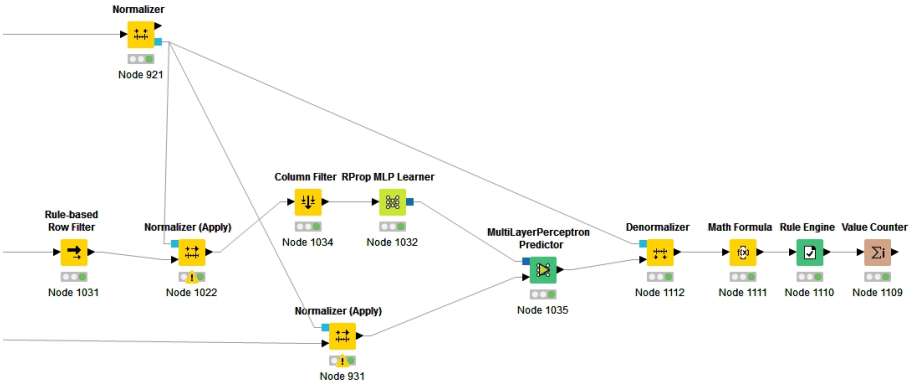
This two layers processing can be applied in different ways. First, in the case of ad-hoc task where the decomposition of the task should be near to the optimum for the best processing times, the prevalent part of processing is dedicated to the Spark layer (if it is reasonable from the nature of the task, i.e., degree of parallelism) and KNIME serves only as a presentation layer.

In the second case, SBDH contains set of pre-defined analytical methods for Apache Spark (e.g., filtering, application of various parameterized methods as computation of fast furrier transformation from this time interval and with a given time window). Then, KNIME is used for finishing preceding processing from Apache Spark and for providing visualization methods as well.

More detailed overview of the first approach to this two layer analytics is provided in [5].

3 KNIME for User Friendly Analytics

As it was mentioned, data analysis was tested on the data from the hydroelectric power station. There seems to be a problem with waste deposition on the blades of the turbine which causes a decrease of generated power. It is known from experience that if the turbine is momentarily stopped and restarted, the shock wave of water cleans the blades and increases the power to the normal level depending on other conditions. The power is dependent on water density and flow, turbine efficiency and water level difference before and after the station. Because the restart of the turbine is not very good for it, we analyze the data from station sensors in order to predict when it is desirable to stop it, i.e., when the effect of restart is greater than its abrasion.



Picture 1. Example of a workflow from KNIME

In the Picture 1 there is a screenshot of the workflow from KNIME. After some preprocessing the data from the whole year of operation are normalized and sent to MultiLayer Perceptron (MLP) Learner node where they are used as training data. The

neural network was then tested on the data from the following year (MLP Predictor node). The next nodes serve for the evaluation of the result. With this approach, we achieve over 80% accuracy of predicting of the right time to stop the turbine.

One of the reasons that the accuracy is not higher can be a small amount of training data. We had data from only a few years available (there were not many stops) and the difference between these years in the conditions can be substantial.

4 Conclusions

The requirements to improve the performance and versatility of analytics are pervading many domains including industrial automation domain. As in many cases, the trade-off between versatility (it is related also to user friendliness) and good performance has to be established.

In this paper, we shortly introduced the solution how to analyze big amount data with the help of SBDH and KNIME. KNIME does not allow to process big amounts of data, but it offers many pre-built analytic blocks for easy composing of a complex task and as we illustrated, can be used in combination with Big Data storage

Acknowledgment: This research has been supported by Rockwell Automation Laboratory for Distributed Intelligent Control (RA-DIC) and by institutional resources for research by the Czech Technical University in Prague, Czech Republic.

References

1. A. Doan a A. Y. Halevy, „Semantic integration research in the database community: A brief survey," *AI magazine*, sv. 26, č. 1, p. 83, 2005.
2. M. Obitko a V. Mařík, „Integrating transportation ontologies using se-mantic web languages," v *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*, Springer Berlin Heidelberg, 2005.
3. V. Jirkovský, M. Obitko a V. Mařík, „Understanding Data Heterogeneity in the Context of Cyber-Physical Systems Integration," *IEEE Transactions on Industrial Informatics*, pp. 660-667, 2017.
4. V. Jirkovský, M. Possolt a M. Obitko, „RDF Storage for Semantic Big Data Historian," v *Proceedings of WIKT & DaZ 2016*, Bratislava, Slovakia, 2016.
5. M. J., *Transformace uživatelských dotazů pro analýzu dat v průmyslové automatizaci*, BS thesis. České vysoké učení technické v Praze. Vypočetní a informační centrum., 2016.