Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Master Thesis

ENERGY STATUS DATA
COLLECT • ANALYSE • UTILISE

Impact of Aggregation Methods on High Volume Energy Data Streams

The energy consumption of industry amounts to 24% of the overall consumption in the European Union [1]. From a business perspective, energy consumption also is an important factor to the overall production cost. Companies can directly benefit from energy-efficient production processes. On the other hand, the shift to renewable energies requires companies to adapt to a less flexible energy supply. It requires to not only understand the own production process but also the behavior of the electrical grid.

In parallel with these new challenges, the technology to measure energy consumption is advancing. smart meters allow to measure different physical quantities, such as voltage, frequency and harmonic distortion. They give an indication of machine behavior and the quality of the electrical grid. With sample rates up to multiple measurements per second, these devices produce huge amounts of data. This challenges the data processing infrastructure to scale up to hundreds of thousands of events per second.

In this thesis, the research centers around smart meter data from a production site. The focus is to quantify the impact of aggregation methods on time-series clustering.

For data mining applications, aggregation often is a trade-off between storage and processing capacities versus more accurate results. In this thesis, the focus is on methods on clustering energy data time series. The following questions are of particular interest:

- Are the clustering results different for different levels of aggregation? What are possible interpretations of these differences?
- What is a good way to quantify the influence of the aggregation method and of the aggregation level on the clustering result?
- Are there unusual sequences in the data which do not fit into the clustering? Does their occurrence depend on the aggregation level?
- Which information about energy consumption and the electrical grid can be extracted from fine-granular data, as compared to aggregated data?

This results in the following tasks:

- Design and implementation of an evaluation framework for different time series clustering algorithms.
- Implementation of metrics to quantify the impact of aggregation on result quality.
- Experimental evaluation on real world smart meter data.

Our technology stack builds upon modern data processing frameworks such as Apache Cassandra and Apache Spark. Experimental evaluations can be run on a cluster with 512 GB RAM and 48 Cores.

In this thesis, you are working on latest research questions and acquire practical knowledge on large-scale data analytics. You train highly demanded skills in development and evaluation of data-mining algorithms. Knowledge from a lecture such as “Big Data Analytics” is beneficial. Elementary statistical knowledge, programming skills and the ability to accomplish conceptual work are desired.

Contact: Holger Trittenbach holger.trittenbach@kit.edu Building: 50.34 Room 338

[1] O. G. O'Driscoll E, *J. Clean. Prod.,* 2013.