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

Master Thesis

Optimal Window Sizes for Correlation Analysis on Energy Time Series

The energy transition ("Energiewende") poses numerous challenges to energy systems. Since the production of renewable energy cannot really follow the demand, processes in industry require more flexibility, to shift consumption. At the same time, high resolution smart meters which measure the energy consumption at a very fine resolution are becoming ubiquitous, allowing for data driven analyses. Modern smart meters measure hundreds of physical quantities such as voltage, frequency and many levels of harmonic distortions on a time scale of seconds. With one smart meter per machine this leads to many attributes with unknown relevance.

However, because data collection and analysis at this scale induces higher costs, its benefits need to be clear. As a first step, one would like to know which attributes are independent and which ones are correlated. Then one can quantify the value of capturing a certain attribute or not. However, the dependencies between attributes may change over time. As a result, the correlation analysis needs to be time-sensitive. Sliding windows are a common approach to detect correlated attributes in certain time windows and to monitor changes in that correlation.

In this thesis, the research centers around smart meter data from a production site. The focus is on the impact of aggregation on correlation analysis.

Using sliding windows in time series analysis requires a careful selection of the window size. The main focus of this work is leveraging knowledge about the used correlation measures for automated selection of window sizes. The following questions are of particular interest:

- How can we identify good window sizes, and how can we do so efficiently?
- How does aggregation impact the window sizes? Are the sizes consistent regarding their represented time frame?
- How to apply this detection to data streams? Which obstacles or restrictions exist for online algorithms?

This results in the following tasks:

- Design quality metrics that quantify important aspects of a window size for specific measures of correlation.
- Design and implement an efficient approach to identify the best window sizes for a given data set.
- Experimental evaluation on high resolution 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 not a prerequisite. However, elementary statistical knowledge, programming skills and the ability to accomplish conceptual work are desired.

Contact: Michael Vollmer michael.vollmer@kit.edu Building: 50.34 Room 340