PhD Lecture

In partial fulfillment of the terms for obtaining the PhD degree, Søren Kejser Jensen will give a lecture on the following subject:

Model-Based Time Series Management at Scale

on Monday 4th of November 2019, 12:30, in room 0.2.13 at Selma Lagerløfs Vej 300

Abstract:
Sensor networks are increasingly deployed to monitor and simplify management of physical entities such as planes and wind turbines. As the produced amount of sensor data increases, so do the requirements for methods and systems that can store and analyze the vast quantities of sensor data being collected. However, the systems used in industry are in general not designed to manage sensor data at large scale. This forces practitioners to only store simple aggregates, for example averages over a 10-minute window, instead of the raw time series. As a remedy, this thesis proposes a model-based Time Series Management System (TSMS) named ModelarDB that supports ingestion, storage, and multi-dimensional analysis of time series at scale. A model in this context is any representation from which a time series can be reconstructed within a user-defined error bound (possibly zero). As part of ModelarDB, this thesis proposes methods for model-based management of time series.

Firstly, the thesis provides a comprehensive literature survey of TSMSs proposed by academia or developed through industrial research. A set of classification criteria are proposed that illustrate the commonalities and differences of each system. The surveyed systems are grouped based on these criteria and the user-facing functionality they provide to give a simple to consume overview of the TSMSs. The survey should significantly reduce the work required for researchers entering this topic of research and allow practitioners to easily compare the benefits and drawbacks of each TSMS. Secondly, the thesis addresses the challenge of using models for storage and analysis of individual time series. A model-agnostic ingestion method is proposed that split each time series into dynamically sized sub-sequences and represents each using a model from a set of model types. These are either included with the system or user-defined using a proposed API. Methods for model-based query processing are proposed to efficiently execute aggregate queries directly on models instead of on reconstructed data points. In addition, methods for predicate push-down and for improving the performance of projections using static code-generation are proposed. A general schema is proposed for storing time series as models, with specific optimizations proposed for using the schema with a distributed key-value store. The proposed methods for model-based management of time series allow storage and analysis of time series at larger scale than other solutions by significantly decreasing the amount of storage required. Thirdly, the thesis addresses the challenge of using models for storage and analysis of correlated time series. As a result, the thesis proposes extensions to ModelarDB’s ingestion method, the included model types, and its model-based query processing methods. In addition, a method for grouping correlated time series is proposed. The proposed methods for management of correlated time series further increases the scale at which time series can be managed. Finally, the thesis provides a demonstration of the current version of ModelarDB.

Members of the assessment committee are Associate Professor Simonas Šaltenis, Aalborg University, Professor Klemens Böhm, Karlsruhe Institute of Technology, and Associate Professor Yongluan Zhou, Copenhagen University. Professor Torben Bach Pedersen and Associate Professor Christian Thomsen are Søren Kejser Jensen’s supervisors. Moderator Associate Professor Chenjuan Guo.

All interested parties are welcome. After the defense the department will be hosting a small reception in cluster 3.