ESR 2.3: Model-based Storage for Time Series

Abduvoris Abduvakhobov
abduvorisa@cs.aau.dk

Motivation

Meetings with manufacturers, owners and energy traders:

- Modern turbines are monitored by up to 300 high-quality sensors generating up to 200 GB per day
- Simple aggregates (e.g. 10-minute averages) are stored instead of high frequency series, thereby removing useful fluctuations and outliers
- High velocity and high volume of data makes storing it in a raw format infeasible
- Compression must be either lossless or lossy depending on use cases
- Each sensor produces a data stream sampled in regular intervals e.g. in 10 Hz series or irregular intervals

State of the Art

Time Series Management Systems (TSDB)

- Store time series that consistsof a time stamp and value
- Optionally contain metadata or tags
- Process queries on time series
- Queries contain timestamp or a time range

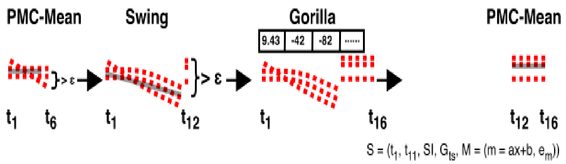
Categorization of TSDB based on storage architecture (Jensen et al. 2017):

Internal Data Stores	External Data Stores	Extension for RDBMS
• Mostly centralized • Tightly coupled data storage and processing component • Few mature implementations • Plato, LittleTable, VergeDB, Apache IoTDB	• Predominantly distributed • New processing engine on top of external data store • High number of mature implementations • Apache Druid, Bolt, Gorilla, BTrDB	• Extends popular RDBMs • Predominantly centralized • Few of mature implementations • Chronix, EdgeDB and Heracles

ModelarDB

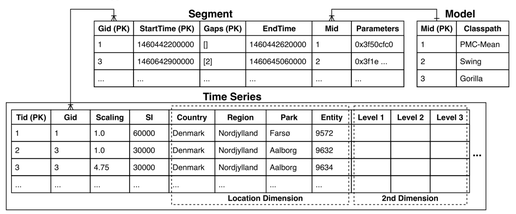
A distributed, parallel processing TSDB with external store built to store regular time series with the help of mathematical models.

- Uses Apache Cassandra for storage and Apache Spark for query processing
- Approximates time series using mathematical functions (models) and stores only model coefficients. Currently uses three model types:



Jensen et al. 2o21.

- Also groups correlated time series together and compresses them as one stream of models to reduce the storage required
- Defines new schema for storing model coefficients and time series groups



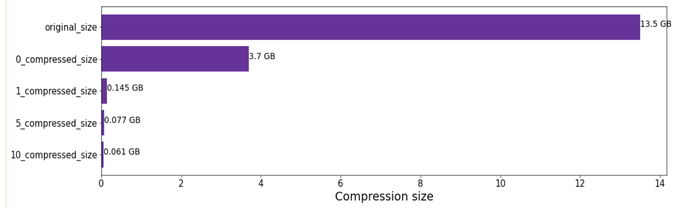
Jensen et al. 2o21.

Methodology

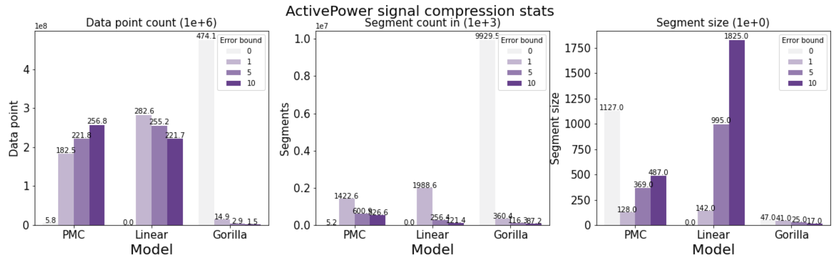
The following methods will be deployed on iterative basis to accomplish established project objectives:

- Literature review
- Problem analysis
- Data collection and analysis
- Solution design
- Feasibility analysis
- Prototype development
- Evaluation and testing

ModelarDB on Data from Siemens Gamesa



- 3 years of measurement from Powerlog controller. Three signals (columns): ActivePower, ActivePower60 and PowerError
- Each signal makes 4,5 GB parquet file
- Signals with high variance mostly require lossless compression for 0 percent error bound
- Some time intervals are irregular and time drifts also occur



Secondment

Secondment partner: Siemens Gamesa Renewable Energy (SGRE)

Start date: March 21, 2022

Collaboration with SGRE

> ~110 TB size of data

> ~800 wind parks around the world

> 90 % data from wind turbines

> 10 % from wind park controllers

> 150 ms sampling interval

> Mainly used for predictive maintenance e.g. anomaly detection, performance monitoring with fixed aggregate levels: 1 week, 1 day, 4-6-12 hour aggregates

Research Questions

- How can we efficiently evaluate the compression performance of model types and the quality of compression to varying error bounds of ModelarDB on different datasets?
- Depending on outcomes of the evalution, what other model types can be implemented to improve the compression and query performance of ModelarDB on real-life RES datasets?
- How can time series automatically be grouped using different correlation statistics and provided heuristics during the ingestion process?
- How can model-based ingestion of time series with a dynamic sampling interval and error bound be supported in ModelarDB?

Bibliography

- Jensen, Søren Kejser, Torben Bach Pedersen, and Christian Thomsen. "Scalable Model-Based Management of Correlated Dimensional Time Series in ModelarDB+." 2021 IEEE 37th International Conference on Data Engineering (ICDE). IEEE, 2021.
- Jensen, Søren Kejser, Torben Bach Pedersen, and Christian Thomsen. "Time series management systems: A survey." IEEE Transactions on Knowledge and Data Engineering 29.11 (2017): 2581-2600.