NEWSLETTER

November 2022

MORE

Management of Real-time Energy Data

Editorial

Dear reader,

We are entering the final year of the project, and we are happy to share with you the progress we had in extracting actionable knowledge from the data of solar and wind parks! We are excited to announce the first integrated version of MORE's platform which is available at https://github.com/MORE-EU. The platform provides scalable data analytics and ML methods which rely on the lossy compression of data and the parallel execution of the algorithms.

In the current newsletter, we highlight some of the most recent developments of MORE. Athena Research Center presents the visualization platform for renewable energy analytics, and the University of Aalborg and ModelarData describe ModelarDB System for time series data. Moreover, we present the outcomes of our first hackathon at Aalborg, the plans for our second hackathon in Athens, and the outstanding achievement of IBM winning the 2nd Position in the Federated Tumour Segmentation Challenge using federated learning techniques developed in the context of MORE. The latter is evidence of the broad impact of MORE algorithms in data analysis.

Until our next newsletter, you can keep in touch with MORE by visiting our website or following us on Twitter (@MOREAnalytic) and LinkedIn (https://www.linkedin.com/company/more_h2020_project/). We look forward to your feedback and participation in MORE's events that we plan for the following months!

Yours sincerely,
Manolis Terrovitis
Principal Researcher, Research Center Athena
Coordinator of the MORE project

1

A Self-Service Visualization Platform for Renewable Energy Analytics

V.Stamatopoulos, S.Maroulis (Athena Research Center)

The sheer volume of the time-series data that are generated from sensors of wind turbines and solar panels makes their exploration and analysis very challenging. This is especially the case for business users in the Renewable Energy Sources (RES) sector, who may have limited skills in data analytics techniques. The Self-Service Visualization Platform for Renewable Energy Analytics targets specifically such users and provides an intuitive visual analytics UI for them to interact with big geo-located time-series data collected from wind turbines and solar panels. The platform offers real-time visual analytics at the macroscale (e.g., park KPIs on a map) and microscale (e.g., analysis of multivariate time series from multiple sensors of a single wind turbine) through a dashboard-like UI and informs users on various KPIs and important events in real time.

More specifically, the UI allows users to first navigate on a map that shows all the available solar or wind parks and view statistics for them, aggregated on the park level. Then, selecting a specific turbine or solar panel, they can start exploring the time series data collected by its mounted sensors. The time-series visualization dashboard offers functionality for interactively visualizing and filtering RES sensor data, highlighting intervals and data points of interest, comparing data from different RES and finding and visualizing patterns and complex events. To empower the visual analysis of the RES data, despite the challenges posed by their volume, the platform adopts innovative data management techniques in the background. First, to enable real-time interactive exploration and analysis of millions of RES data points, an in-memory indexing approach is employed, which is initialized based on the raw time-series data and maintains a hierarchical representation of them. Further, index adaptation and data prefetching techniques are employed, such that the response time of future user operations is kept at interactive rates.

Besides the visualization of historical RES time-series data, the platform provides real-time monitoring functionality through the visualization of various KPIs and the alerting of any detected anomalies or critical events that negatively impact performance. Such events include, for example, the accumulation of dust and debris on the surface of solar panels or yaw misalignment incidents in the case of wind turbines. To further support the monitoring and timely identification of potential problems, users can also enable the visualization of forecasted data for the KPIs they currently view.

See more: https://github.com/MORE-EU/more-visual

Watch the demo:

2

ModelarDB: Managing Extreme-Scale Time Series from Wind and Solar Parks Using Models

Christian Thomsen, Søren Kejser Jensen

ModelarDB is an efficient Time Series Management System (TSMS) for high-frequency sensor time series. ModelarDB provides state-of-the-art lossless and lossy compression and query performance by representing time series using different types of so-called models, such as constant and linear functions. We use the term model for any representation of a time series from which the original time series can be recreated within a known error bound (possibly 0%). For example, the linear function v = a * t + b can represent an increasing, decreasing, or constant time series and reduces storage requirements from one value per data point to only two values: a and b. This is illustrated in the figure below, where the original data points are white circles, the linear function is a black line, and the reconstructed data points are red:

As the structure of the time series being compressed changes over time, ModelarDB automatically and dynamically switches between different model types to accommodate. The compressed time series can be efficiently queried using a relational interface and SQL without any knowledge about the model-based representation. ModelarDB's query optimiser automatically rewrites the queries to exploit the model-based representation.

From a user's perspective, ModelarDB stores time series as tables where each row contains a timestamp, fields (i.e., measured values such as temperature, pressure, and humidity), and tags (i.e., describing attributes such as location, manufacturer, and manufacturing year). At the physical level, ModelarDB stores the time series in a very efficient manner by using models as described above. These models can recreate the original values so arbitrary queries can be answered, and many queries can be answered more efficiently directly from the models. For example, for the time series shown in the figure above, a query could ask for the SUM of the values. A naïve way to do that requires the reconstruction of every data point. For regular times series, ModelarDB, however, only reconstructs the first and last data point, as illustrated in the figure below, and can then compute the SUM directly from those. This is much more efficient and contributes to ModelarDB's high query performance.

To further reduce the amount of storage required, the user can optionally specify an error bound for each field. This allows ModelarDB to perform lossy compression such that the models approximate the values within the error bound. The error bound can also be set to 0%, in which case ModelarDB does lossless compression. A 0% error bound is the default.

The first research prototype of ModelarDB already demonstrated significantly better compression than systems widely used in the industry (see the table below).

Since then, more functionality has been added to ModelarDB to provide even better compression. For example, compression of similar time series as one stream of models and avoiding storing derivable time series physically. Thus, compared to widely used formats, ModelarDB currently provides **up to 13.7x faster ingestion, 113x better compression, and 573x faster aggregate queries.** Further, ModelarDB scales close to linearly as more data and nodes are added.

ModelarDB is designed to be deployed both on the edge (e.g., on wind turbines) and on the cloud (e.g., on Microsoft Azure). Data is efficiently ingested and compressed using the different types of models already on the edge and can immediately be queried for low-latency analytics. Depending on the available bandwidth, data amounts, and analytical needs, the compressed data is later transferred to the cloud, where further compression can take place by exploiting similarities between data from multiple edge nodes. This combination of edge and cloud deployment makes it possible for ModelarDB to provide both low-latency queries and practically unlimited scalability.

The latest version of ModelarDB is designed to be cross-platform and is continuously tested on Microsoft Windows, macOS, and Ubuntu. It is implemented in Rust and uses Apache Arrow Flight for communicating with clients, Apache Arrow DataFusion as its query engine, Apache Arrow as its in-memory data format, and Apache Parquet as its on-disk data format.

ModelarDB is an open-source project and is licensed under version 2.0 of the Apache License. The source code, test cases, and documentation are available at https://github.com/ModelarData/ModelarDB-RS

3

MORE's Hackathons

MORE planned two hackathons by the end of 2022, one to be hosted by Aalborg University and one by Athena Research Center. In both hackathons, the participants utilize MORE's components resources that are available at https://github.com/MORE-EU:

- https://github.com/MORE-EU/ModelarDB
  Model-Based Time Series Management.
- https://github.com/MORE-EU/more-edge-analytics
  Lightweight Analytics on the Edge.
- https://github.com/MORE-EU/more_api
  More RESTful API services.
- https://github.com/MORE-EU/sail
  Library for incremental learning algorithms.
- https://github.com/MORE-EU/more-visual
  Visualization of Time Series.
- https://github.com/MORE-EU/more-visual-index
- https://github.com/MORE-EU/more-pattern-extraction
  Time Series Pattern Extraction.
- https://github.com/MORE-EU/complex-event-detection
  Complex Event Detection.
- https://github.com/MORE-EU/matrixprofile
  Accessible computation of Matrix Profile.

1st MORE Hackathon – Analyze Time Series Using the MORE Platform

(Where: Aalborg – When: 2020-10-13 – 08:30-16:15)

Søren Kejser Jensen, Christian Thomsen, Torben Bach Pedersen

The hackathon was held on Thursday, October 13, 8-16, co-located with the MORE physical plenary meeting in Aalborg, October 13-14. The participants were five 9th-semester (pre-Master thesis) students from Software Engineering doing a master project on compression and satellite-based transmission of time series from commercial ships, where they are trying to extend ModelarDB to their particular use cases. They were given the following intro document:

"Renewable Energy Systems (RES) are critical infrastructure that is heavily monitored through high-quality sensors. These sensors can produce data points at a sub-second sampling interval. However, ingesting, managing, and analysing such vast amounts of sensor data is currently infeasible. For example, permanently storing all of the data points is cost prohibitive, and even transferring the data points over the 500 Kbits/s to 5 Mbits/s connection available at some RES installations is not possible. A common solution to this problem is simply collecting aggregates, e.g., 10-minute averages, instead of the raw data points. Thus, valuable anomalies and outliers are lost. To remedy this, tools that ingest, manage, and analyze these time series are required.

The MORE project has developed a collection of tools for efficiently ingesting, managing, transferring, and analysing high-frequency time series. Together these tools constitute the MORE platform. The purpose of this hackathon is to try to ingest, manage, and/or analyse regular time series by either using these components directly or by using the components as part of your own application or data processing pipeline. The key outcomes from the Hackathon will be discussed at the end of the day, so please prepare a 5-minute presentation to start the discussion. So, use your imagination, explore the world of time series, and happy hacking :) "

As a start, a small set of regular time series with energy measurements from houses is provided as REDD-Cleaned, and a Python script for extending it is provided as REDD-Cleaned-Extender.py.
More data sets with regular time series are available online, so the following are just examples:

- https://github.com/zhouhaoyi/ETDataset
- https://github.com/laiguokun/multivariate-time-series-data
- https://github.com/lixus7/Time-Series-Works-Conferences

A VM is available through VNC at [hidden URL]. Remember to first connect to AAU's VPN before connecting to the VM, as it otherwise will fail.
During the day, the participating students could talk to the MORE researchers and discuss their solutions and get help with the use of the MORE platform components. This worked well, and after the event, we got the following feedback from the participants:

"Overall, it was a really nice experience. Super cool both to dive into the systems, but also to meet the various participants who work in the field.

It was fine that there were no fixed tasks, but it would have been nice if we had gotten a few suggestions/examples of how the individual components could be used and combined. Additionally, we spent quite a bit of time understanding the fundamentals behind the components, which gave us less time to get it running. Here it would have been an advantage if we had received information + possibly relevant articles a few days in advance so that we could form an overview before we arrived. Alternatively, you could have combined the hackathon with some different presentations about the components, or you could have run it over two days.

However, we got a good discussion out of it, which in the long term may also turn out to be interesting in relation to our thesis project. We talked about how you could combine ModelarDB with the pattern extraction component and thus use matrix profiling to find patterns in the models in real-time and compare those patterns with raw data. In our case, this is interesting as we do not have access to the raw data on the server after it has been compressed on the edge."

We find this feedback very relevant and will take it into account when designing our 2nd hackathon.

(Upcoming) 2nd MORE Hackathon – Renewable Energy Hackathon
(Where: Athens – When: 2020-12-15 – 10.00)
Danae Pla Karidi

The hackathon will be held on December 15th in Athens at Athena Research Center, and participation in the event is open by completing this form: https://docs.google.com/forms/d/e/1FAIpQLSc4nx3n7cACQbyw1Pw5vNQVjljK2Lt5IDgv7_MyS01VAJ97Qw/viewform

The hackathon will focus on MORE's streaming tools and methods for solar-park optimisation, wind turbine boosting, and visual exploration and will help participants to upgrade their AI and machine learning skills.
Take a look at the hackathon's flyer *(view next page)*:

REAL-TIME RENEWABLE ENERGY HACKATHON

Learn how to process data from large solar parks and wind farms in real time!

MORE PROJECT

MANAGING REAL TIME ENERGY DATA

Upgrade your AI and machine learning skills by using tools for increasing the efficiency of renewable energy systems.

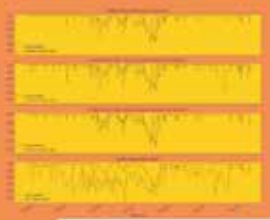
SOLAR PARK OPTIMIZATION

Get to know the new toolkit for real-time soiling detection on streaming solar-park data.

WIND TURBINE BOOSTING

Grasp new machine learning techniques for detecting yaw misalignment on data from large wind farms.

VISUAL EXPLORATION



Visually explore and perform outstanding RES analytics on streaming energy data by training with our Visualization Platform.

WHEN	15/12/2022, 10.00 AM
WHERE	Athena Research Center
REGISTER	Hackathon form

Learn directly from Inaccess RES industry experts!
Find out the modern RES strategies by Perception Dynamics executives!

SUPPORTED BY MORE, A HORIZON 2020 PROJECT FUNDED UNDER GRANT AGREEMENT NO. 957345

MORE
Management of Real-time Energy Data

MORE gets 2nd Position in Federated Tumour Segmentation Challenge

Ambrish Rawat, Seshu Tirupathi

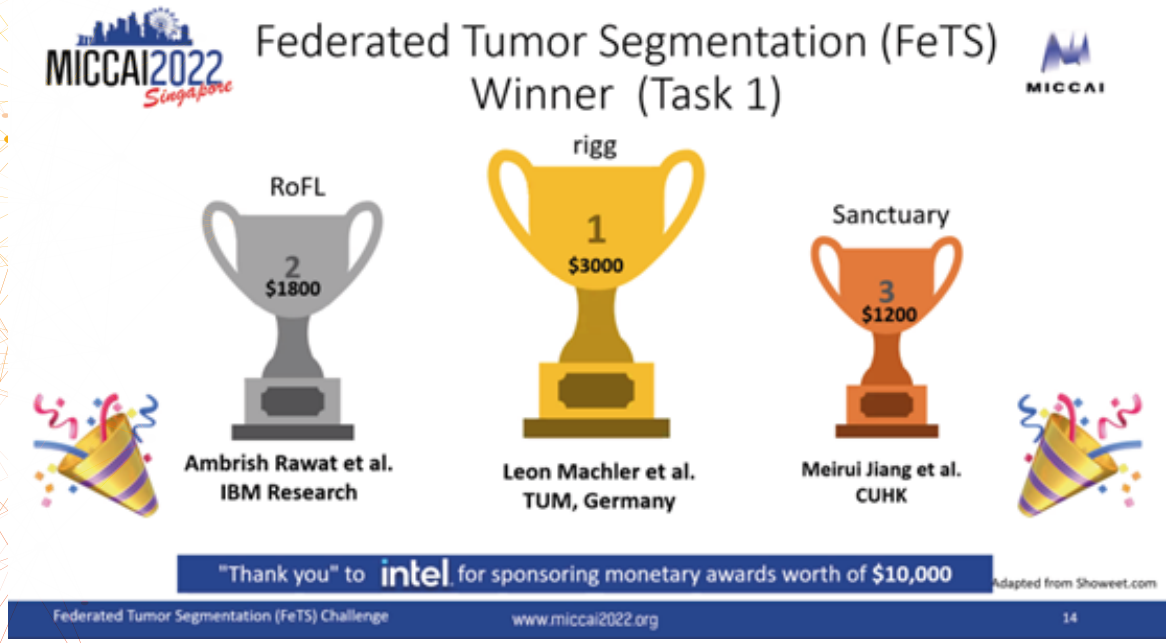
FeTS is one of the largest Federated Learning initiatives (https://www.med.upenn.edu/cbica/fets/), where organisations with access to large databases of MRI scans seek to learn a model collaboratively. There are many challenges for learning a common model across such large databases; for instance, the datasets across the organisations are obtained under varying circumstances resulting in widely different distributions and sizes. Moreover, even the models trained in controlled settings may fail when tested in the wild due to drifts or distributional changes.

To this end, FeTS 2022 at MICCAI presented an open challenge to devise algorithms that can learn useful models in real-life settings with data distributed across 23 different institutions. IBM Research Europe (Ireland) participated in this challenge and devised a novel algorithm for robust federated learning across large datasets with non-iid distributions and secured 2nd rank in the competition. Their algorithm uses carefully designed optimisation methods that help arrive at effective models under computational constraints.

Team Members - Ambrish Rawat, Giulio Zizzo, Swanand Kadhe, Jonathan Epperlein, Stefano Braghin (Ambrish and Giulio were supported by MORE)

- FeTS event coordinators - Ujjwal Baid, Spyridon Bakas
- FeTS 2022 Competition - http://miccai2022.fets.ai
- Coverage in MICCAI Digest - ttps://www.rsipvision.com/MICCAI2022-Tuesday/26

Read the paper, which will be part of the MICCAI proceedings.



Newsletter editing by Danae Pla - Karidi

MORE

Management of Real-time Energy Data

Thanks for subscribing to our newsletter.
Our newsletter is intended to inform our subscribers about our overall performance trajectory.

We'll keep you posted every six months on the latest news on MORE's achievements, results, trends, and interesting news.

For more information, visit:
www.more2020.eu

MORE receives funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 957345.