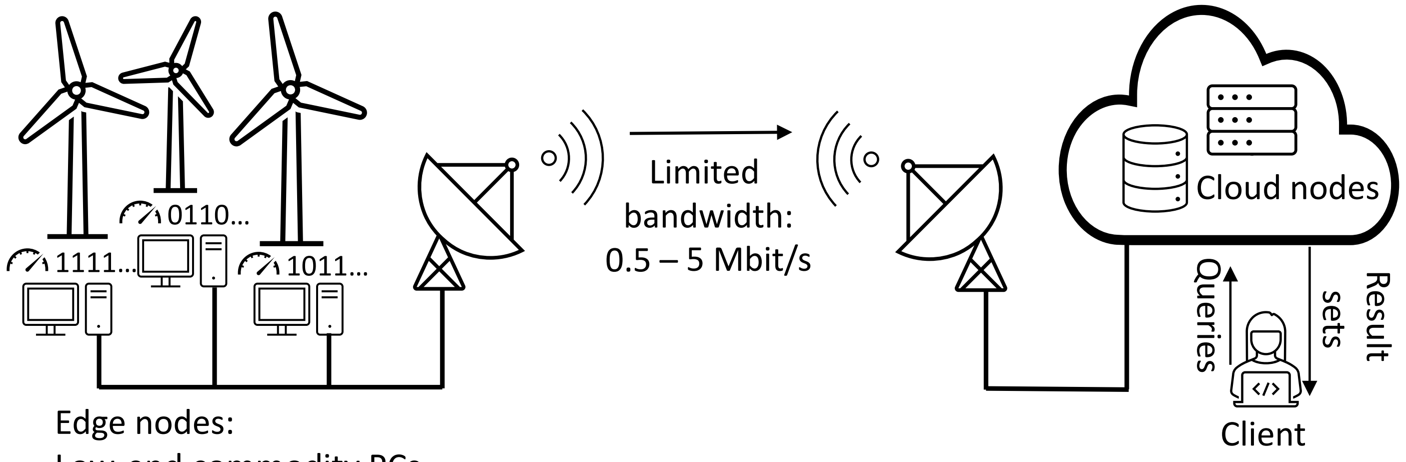
ModelarData

Why Model-Based Lossy Compression is Great for Wind Turbine Analytics

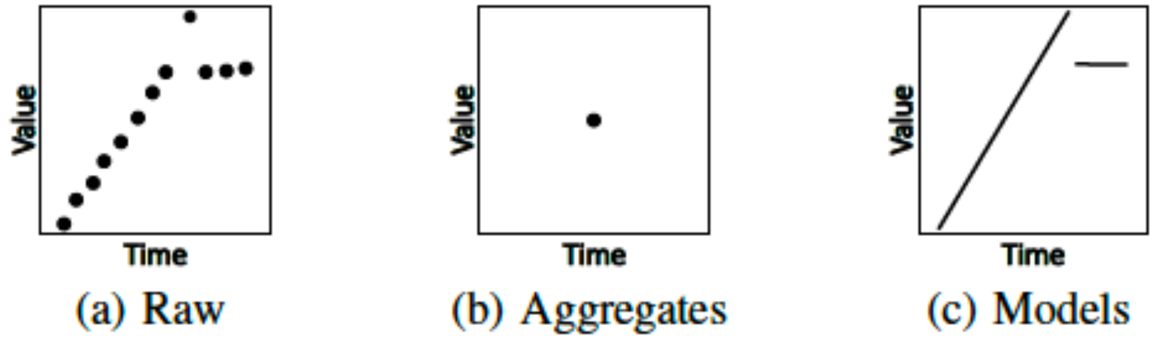
Søren Kejser Jensen, Christian Thomsen,
Torben Bach Pedersen, Carlos Enrique Muniz-Cuza,
Abduvoris Abduvakhobov
Aalborg University

Wind Turbine Analytics Background

- Wind turbines have *100s of sensors*
- 100 turbines at 100 Hz is *> 11PiB/year*
- Data collected by *weak edge devices* + transferred to cloud via *slow connections*
- Raw data too big: *compression needed*
- But which compression?
  - Practioners use *simple aggregates*, e.g., 10 min AVG: loose outliers + fluctuations
  - Lossless compression: *not enough*
  - *Model-Based Lossy compression is better*
- Analytics
  - Time interval aggregates
  - Machine learning (time series forecasting, yaw misalignment,…)

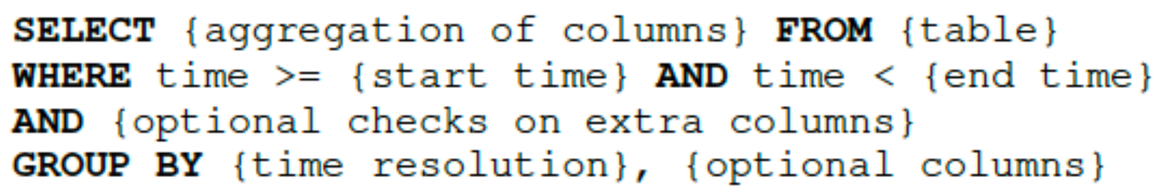


Edge nodes:
Low-end commodity PCs,
4 CPU cores, 4 GiB RAM, and an HDD



(a) Raw (b) Aggregates (c) Models

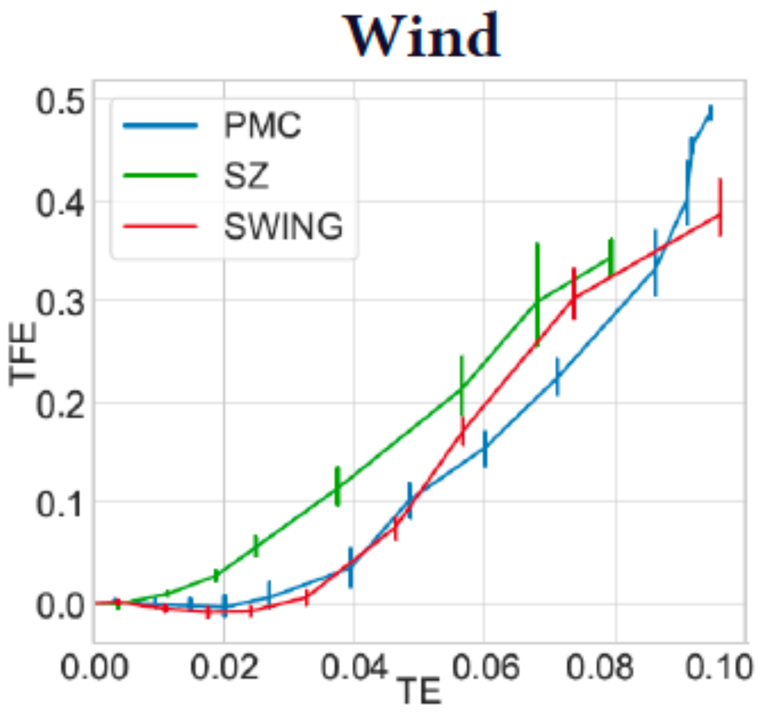
Fig. 1. Representations of long high-quality high-frequency time series



```
SELECT {aggregation of columns} FROM {table}
WHERE time >= {start time} AND time < {end time}
AND {optional checks on extra columns}
GROUP BY {time resolution}, {optional columns}
```

Results for ModelarDB Legacy Lossy Compression

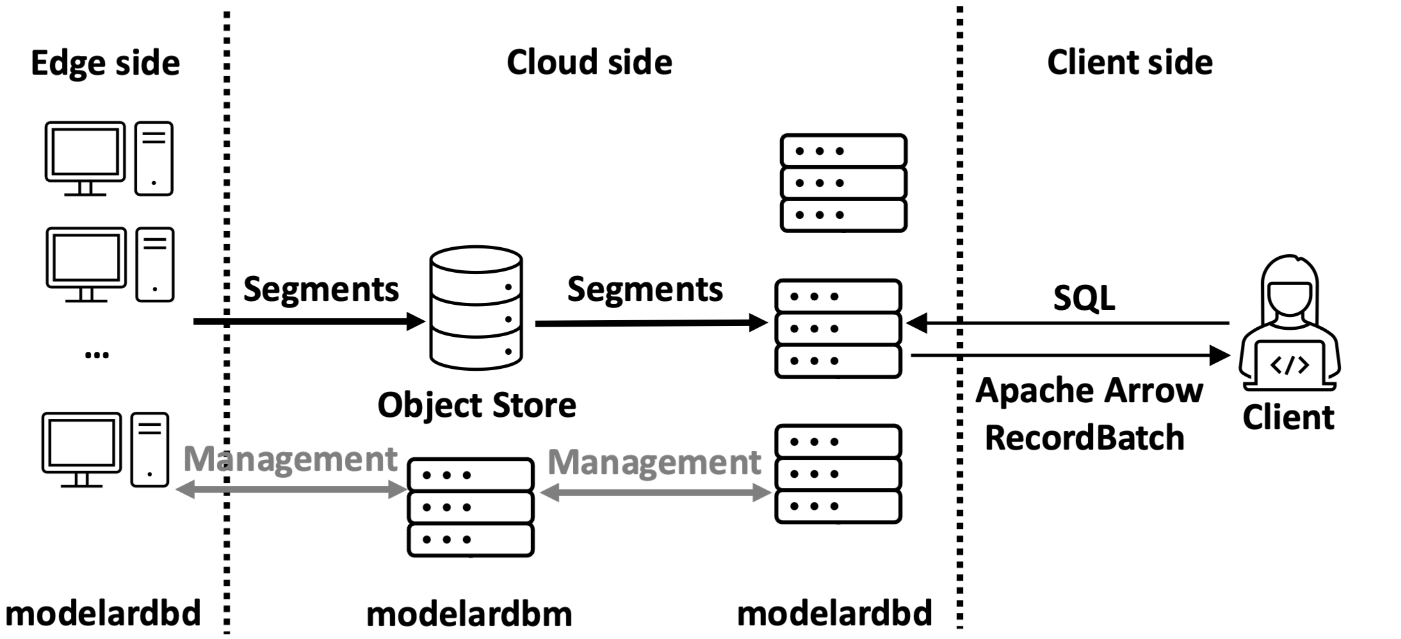
- ModelarDB Legacy Time Series Management System
  - JVM-based, on top of Apache Spark and Cassandra
  - Compresses time series using simple models (constant, line, XOR)
  - Per-value error bound, possibly 0%
- Compression results
  - 1.53x (0%) to 48.89x (10%) less storage than Apache ORC
    - Even more for Parquet
  - Up to 573x faster aggregate queries
  - Similar compression to simple aggregates, but up ~17(!) orders of magnitude less error



- Downstream ML results
  - Yaw Misaligment: *same* accuracy on MDB-compressed data as raw data
  - Time series forecasting: up to *1.8% better accuracy* than on raw data!
    - U-curve: *some* amount of the *right* compression is *good* for accuracy, but *not too much*

ModelarDB Future

- ModelarDB Legacy Lessons
  - Systems research is time intensive and reusing *components limits* optimizations
  - Modularity *adds complexity* and *limits* optimizations
  - Code generation enables optimizations but trades *latency* for throughput and adds complexity
  - *Pull-based* data ingestion improves performance but *increases complexity*



- Lessons+feedback → ModelarDB (Rust)
  - Architecture
    - modelardbd: ingest + process queries, disk/object store
    - modelardbm: manage clusters, assign queries
  - Open libraries + frameworks + formats
    - Apache Arrow Flight (Communication)
    - Apache Arrow DataFusion (Query Processing)
    - Apache Parquet (Storage)
    - (De)compression in library: easier re-use
  - Up to 2.14x better compression and faster queries

Wind Turbine Analytics Conclusion

- ModelarDB’s model-based error-bounded lossy compression: *sweet spot* of:
  - Great compression ratios
  - Great query performance
  - Great error-bounded data quality
  - Great analytics accuracy
- References

  - S. K. Jensen et al., *Time Series Management Systems: A Survey*, TKDE 29(11), 2017
  - S. K. Jensen et al. *ModelarDB: Modular Model-Based Time Series Management with Spark and Cassandra*, PVLDB 11(11), 2018
  - ——, *Demonstration of ModelarDB: Model-Based Management of Dimensional Time Series*,” SIGMOD, 2019
  - ——, *Scalable Model-Based Management of Correlated Dimensional Time Series in ModelarDB+,* ICDE, 2021
  - S. Tirupathi et al., *Machine Learning Platform for Extreme Scale Computing on Compressed IoT Data*, IEEE BigData, 2022
  - S. K. Jensen et al. *ModelarDB: Integrated Model-Based Management of Time Series from Edge to Cloud*, TLDKS (53), 2023
  - S. K. Jensen et al, *Holistic Analytics of Sensor Data from Renewable Energy Sources: A Vision Paper*, ADBIS (Short Papers), 2023
  - C. E. Muniz-Cuza et al. *Evaluating the Impact of Error-Bounded Lossy Compression on Time Series Forecasting*, EDBT, 2024
  - S. K. Jensen at al., *Time Series Management Systems: A 2022 Survey*, in Data Series Management+Analytics (forthcoming), Palpanas+Zoumpatianos (Eds). Preprint available at: https://vbn.aau.dk/da/publications/time-series-management-systems-a-2022-survey
  - A. Abduvakhobov et al. *Scalable Model-Based Management of Massive High Frequency Wind Turbine Data with ModelarDB*, submitted
  - *Official ModelarDB Legacy Repository*. [Online]. Available: https://github.com/ModelarData/ModelarDB
  - *Official ModelarDB Repository*. [Online]. Available: https://github.com/ModelarData/ModelarDB-RS