Data Lakes and Data Intelligence for Future Energy Systems

Professor Torben Bach Pedersen
Center for Data-intensive Systemer (Daisy)
Institut for Datalogi. Aalborg Universitet

tbp@cs.aau.dk
What is a Data Lake?

How do Data Lakes Work?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

1. The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.

2. The reservoir of water is a dataset, where you run analytics on all the data.

3. The outflow of water is the analyzed data.

4. Through this process, you are able to "sift" through all the data quickly to gain key business insights.

[https://www.linkedin.com/pulse/building-data-lake-using-open-source-technologies-aneel/ ]
# DW Versus Data Lake

<table>
<thead>
<tr>
<th>DATA WAREHOUSE</th>
<th>vs.</th>
<th>DATA LAKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>structured, processed</td>
<td>DATA</td>
<td>structured / semi-structured / unstructured, raw</td>
</tr>
<tr>
<td>schema-on-write</td>
<td>PROCESSING</td>
<td>schema-on-read</td>
</tr>
<tr>
<td>expensive for large data volumes</td>
<td>STORAGE</td>
<td>designed for low-cost storage</td>
</tr>
<tr>
<td>less agile, fixed configuration</td>
<td>AGILITY</td>
<td>highly agile, configure and reconfigure as needed</td>
</tr>
<tr>
<td>mature</td>
<td>SECURITY</td>
<td>maturing</td>
</tr>
<tr>
<td>business professionals</td>
<td>USERS</td>
<td>data scientists et. al.</td>
</tr>
</tbody>
</table>

[https://media.lqedn.com/mpr/mpr/shrinknp_800_800/AEEAAQAAAAAARFAAAAAJGfI4WeyOGU4LWJhMWYtNGYwYi1hZjk0LTfIZWNiMml3NjNkOQ.png ]

CITIES Annual Meeting, September 21, 2018
Data Intelligence: Energy Example

Flexible IoT devices → Data Collection → Database → Preprocessing → Remove Outliers, Fill Missing Data, Aggregates → Context Information

Data Intelligence: Energy Example

Schedules → Trading/Energy Management → Flex-Offer Generation Model → Flex-Offer Generation Model → Demand Prognosis

CITIES Annual Meeting, September 21, 2018

https://www.canstockphoto.com/blue-electric-car-31049084.html
https://www.canstockphoto.com/oil-electric-heater-on-wheels-icon-white-49108545.html
https://www.123rf.com/photo_12496301_abstract-cognitive-intelligence.html
https://clip-art-guru.info/weather-clip-art-for-preschoolers/weather-clip-art-for-preschoolers-clip-art-weather-forecast-clipart-clipart-kid/
Data Intelligence: Many Steps…

- **Model** the data
  - What data is available, how is it structured and connected?
- **Collect** the data
  - Using IoT (Internet of Things) devices, etc.
- **Cleanse** the data
  - Correct/tag errors, missing data, outliers
- **Integrate** the data
  - Connect different datasets, match (pseudo) keys, achieve consensus…
- **Store** the data
  - In a **data lake**…
- **Understand** the data
  - Query, analyze, visualize,
- **Predict** the data
  - Build+test machine learning models to predict the future
- **Optimize** given **certain** objectives
  - Based on data and predictions
- **Make decisions and act**
  - Manual or (semi-)automated, what-if,
- **Control/Actuate** to put the decision into action
  - IoT works both ways
What are the challenges?

- **Many** challenges in each step
- I will talk about 2 specific data management challenges

- Storing and querying massive amounts of fast sensor data
  - Model-based storage and querying
  - ModelarDB

- Integrating data+models+optimization
  - Prescriptive analytics
  - Tool that does this *efficiently* and with high programmer productivity
  - SolveDB
Massive, Fast Sensor Data

- Wind turbines/solar panels: 100s of sensors read subsec
  - A modern wind turbine has up to 15,000 streams
- This generates a lot of data
  - 10 reads/second, 4 bytes, 15000 streams $\Rightarrow$ \( \sim 50 \text{ GiB/day/turbine.} \)
    - 100s in a park, 1000s of parks, 20 years...
- Currently not fully exploited/stored
- Typical setup:
  - Consider few (~100) streams
  - Store single avg value every 5-15 min
- Fast variations not seen
Model-Based Data Storage

- Want store and use **all** available sensor data
- Support efficient aggregate queries on historical data
- Streaming analysis of data while it is being ingested
  - Detect underperformance and other problems immediately
  - Enable predictive maintenance
  - Detect and fix a problem before the wind turbine breaks
- Represent data by **models**
  - Store model **parameters** rather than data values
  - User-defined error (0, 1, 5%) allowed
ModelarDB

- **ModelarDB** uses models to store time series data
- Time series functionality in a system-agnostic library
- Some built-in model types, user can add user-def model
- ModelarDB will automatically pick the best
- Query processing and storage by (Spark & Cassandra)
Performance/storage results

- High ingestion rates
  - Allows **online** analytics unlike ORC/Parquet
- Better compression than competitors, even InfluxDB/ORC
  - Achieved by **adaptively using multiple models**
- Low average errors, much lower than error bounds
  - Degrades gracefully as outliers are added
- Close to linear scalability up to 32 nodes
- Effective query optimizations
- Better performance for large aggregate queries
  - Comparable for small aggregate and point-range
- Provides unique **“sweet spot”**
  - Fast ingestion+good compression+fast, scalable online aggregate query processing
ModelarDB Evaluation Summary

- Provides unique “sweet spot”
  - Best compression
  - Fast ingestion
  - Fast, scalable online aggregate query processing
What is Prescriptive Analytics?

Figure 1: Three stages of business analytics (source "Gartner Inc.")
An Example from Smart Energy

- Balancing demand and supply

**Step 1:** Forecast RES production

**Step 2:** Generate demand FlexOffers
Traditional Solution

- Implementation based on RDBMS + R + Java solver

```
#!/bin/bash
export mirabelDbUrl="jdbc:postgresql://localhost/postgres";
export mirabelDbUsername="postgres";
export mirabelDbPassword="";
export mirabelRoot='pwd';
export CLASSPATH="$CLASSPATH:../FlexOffers/Implementation/SDB-scheduling-experiments/target/sdb-scheduling-experiments-0.9.9-SNAPSHOT.jar:../FlexOffers/Implementation/SDB-scheduling-experiments/target/dependency/**"

# Run forecasting
(time Rscript ergv_forecast.R) &> output_fc_r.txt

# Check for the forecasting error
fcerror=`psql -d postgres -c "SELECT * FROM forecast_error" -Ptuples_only`

echo "Forecasting error (MAPE): " $fcerror

# Run scheduling
(time java -Xmx1000m -Xss10m org.mirabel.aggregation.experiments.AggSchExperiment) &> output_sch_c.txt

# Check for imbalance
schImb=`psql -d postgres -c "SELECT * FROM scheduling_imbalances;" -Ptuples_only`

echo "Scheduling remaining imbalance (kWh): " $schImb
```

- Problems:
  - Non-integrated, specialized glued together in ad-hoc fashion
  - Labor-intensive, error-prone, inefficient
SolveDB Solution [SSDBM16,ICDE17]

- SolveDB (PostgreSQL + solvers)
  - Integrates RDBMS and LP/MIP, BB, and specialized solvers
  - Uses SQL extension for problems
  - Lines of code: 3571 versus 237
    - 1-2 orders of magnitude less code for most cases
  - I/O time: 7.3 versus 0.8 secs
    - up to 2 orders of magnitude faster I/O for other cases

CITIES Annual Meeting, September 21, 2018
Conclusion

• Future Energy Systems
  - Lots of data from many new sensors and devices
  - Many types of data, very large volumes, very fast data

• Data Lake
  - Central repository for all the data
  - Challenge: store big and fast sensor streams
  - ModelarDB: model-based storage + querying of sensor streams

• Data Intelligence
  - Add **prediction and optimization** on top of the data
  - Challenge: **ease-of-use/productivitty and efficiency**
  - SolveDB: integrate data, prediction, and optimization with SQL

• Future work:
  - Data lakes: compression of multi time series, indexing
  - Data intelligence: SolveDB +physical models, MPC in the database
References

- Siksnys et al. *Aggregating and Disaggregating Flexibility Objects*. TKDE 2015
Links

• http://people.cs.aau.dk/~tbp
• http://www.dicyps.dk
• https://github.com/skejserjensen/ModelarDB
• http://daisy.aau.dk/solvedb

• http://goflex-project.eu
• http://goflex-community.eu
• http://goflex-community.eu/PlayVideo.asp?Video=1505_BAUM_GOFLEX_Final.mp4
• http://www.flexshape.dk
• http://www.totalflex.dk
• http://www.arrowhead.eu