Smart Meter Data Analytics

Xiufeng Liu
xiufeng.liu@uwaterloo.ca
University of Waterloo, Canada

Outline

- Benchmarking smart meter data analytic technologies
- Smart meter data analytics system (SMAS)
  - Demo

Introduction

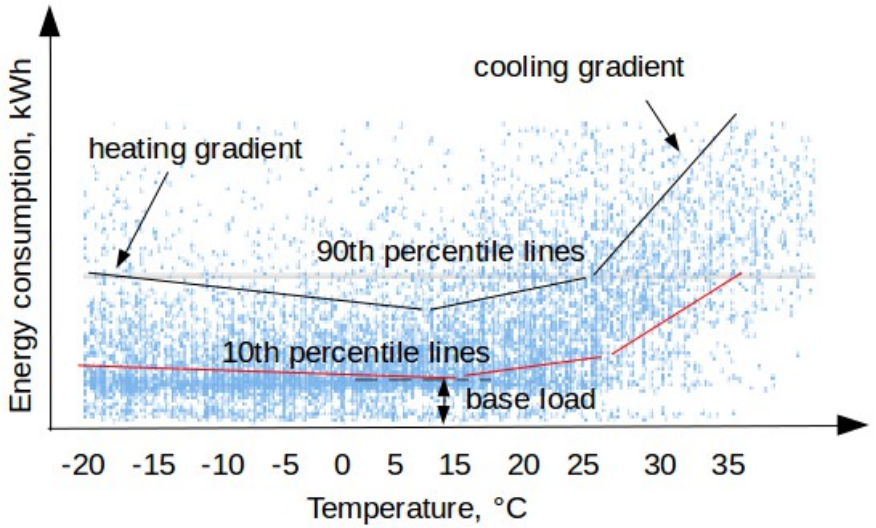
- The widely use of smart meters makes the data analytic possible
- Smart data analytics can help energy providers and consumers understand and reduce energy usage
- Diverse analytic technologies appear
  - What technologies to be used?
  - Best practices?
- The need of data generator

The benchmark technologies

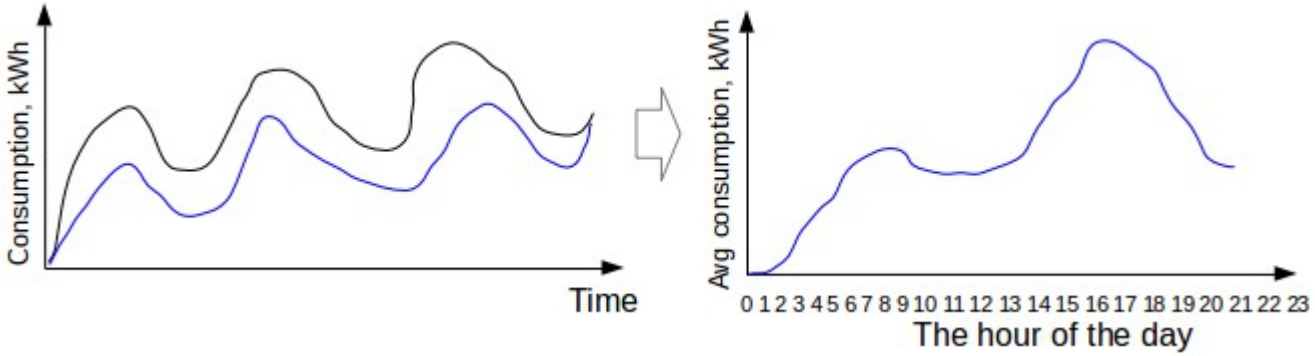
- Centralized Systems:
  - The traditional analytic tool: *Matlab*
  - In-database analytic tool: *MADLib in PostgreSQL*
  - In-memory column store*: System C (KDB)*
- Distributed Computing Systems:
  - Main memory based: *Spark*
  - Hadoop based: *Hive*

The benchmark algorithms

- 3-Line algorithm:

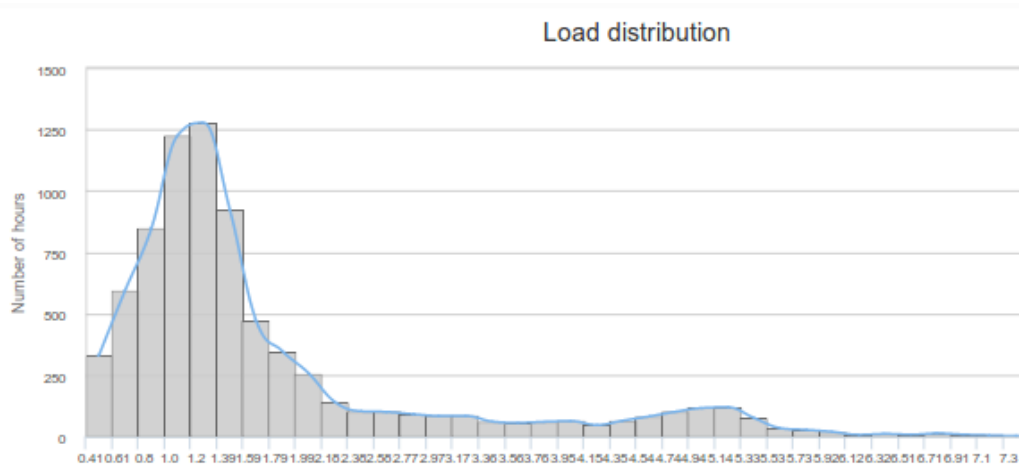


- Periodic auto-regression (PAR):



The benchmark algorithms

- Histogram:



- Cosine similarity: $$\frac{X \cdot Y}{||X|| * ||Y||}$$

Experiments – centralized systems

- Using real-world (Essex 10GB) and synthetic data sets
  - Data loading: partition vs. non-partition
  - Impact of partitioning on performance
  - Cold start vs. warm start

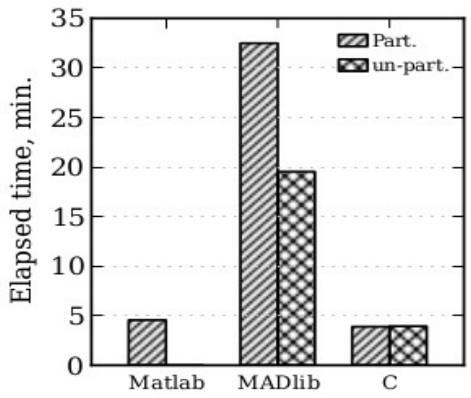


Figure 4: Data loading times, 10GB real dataset.

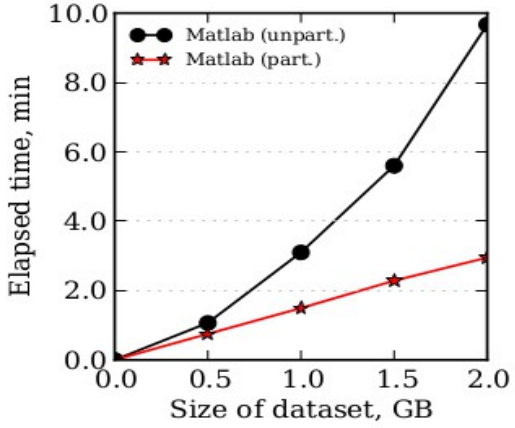


Figure 5: Impact of data partitioning on analytics, 3-line algorithm.

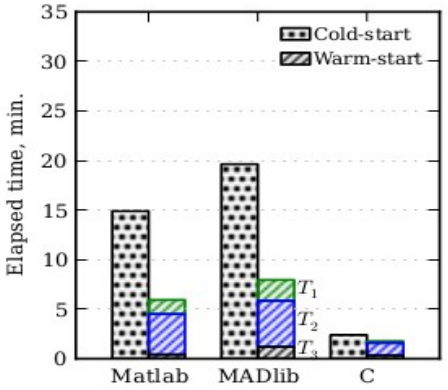


Figure 6: Cold-start vs. warm-start, 3-line algorithm, 10GB real dataset.

Experiments – centralized systems

- Execution times of using real-world data sets (10GB essex)

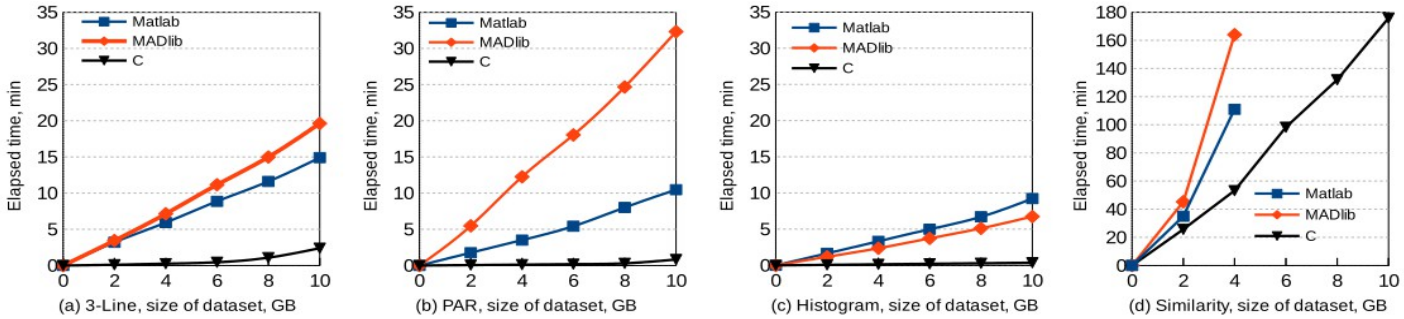


Figure 7: Single-threaded execution times of each algorithm using each system.

- Execution times using large synthetic data sets.

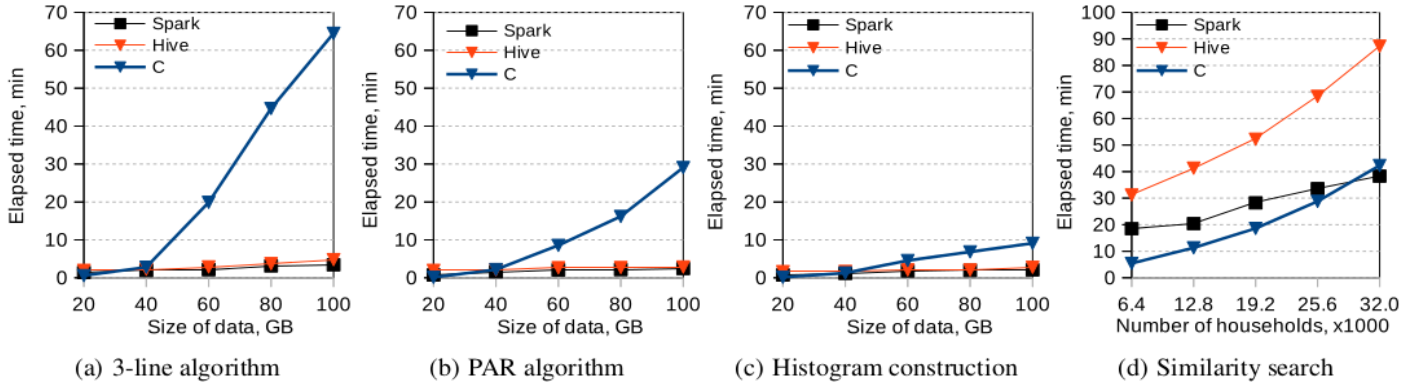


Figure 11: Execution times using large synthetic data sets.

Experiments – distributed systems

- Cluster: one master node + 16 slave nodes
- Test systems: Spark and Hive
- Use three types of data formats:
  - 1st Format: one file (that may be partitioned arbitrarily) with one smart meter reading per line
  - 2nd Format: One file with one household per line (i.e., all the readings from a single household on a single line)
  - 3rd Format: Many files, with one or more households per file (but no household scattered among many files)
- Measure the scalability and speedup

Experiments – distributed systems

- The execution times and speedup of the 1st data format:

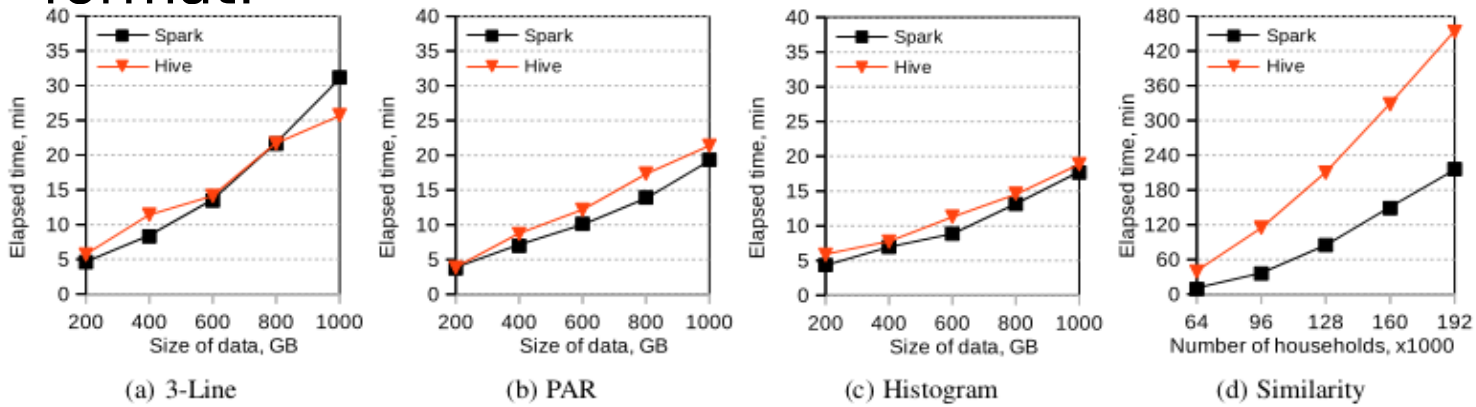


Figure 13: Execution times using the first data format in Spark and Hive.

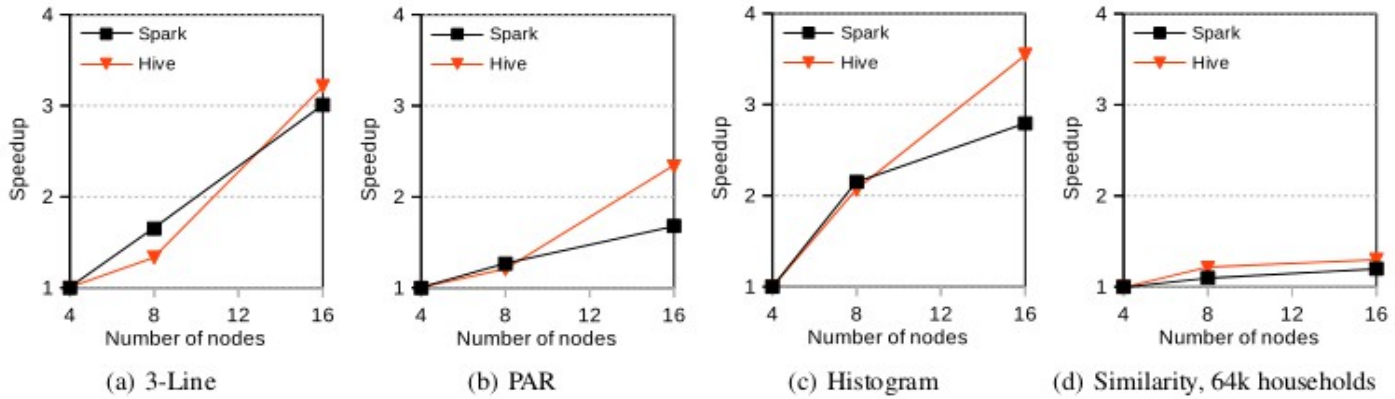


Figure 14: Speedup obtained using the first data format in Spark and Hive.

Experiments – distributed systems

- The execution times and speedups of the 2nd data format:

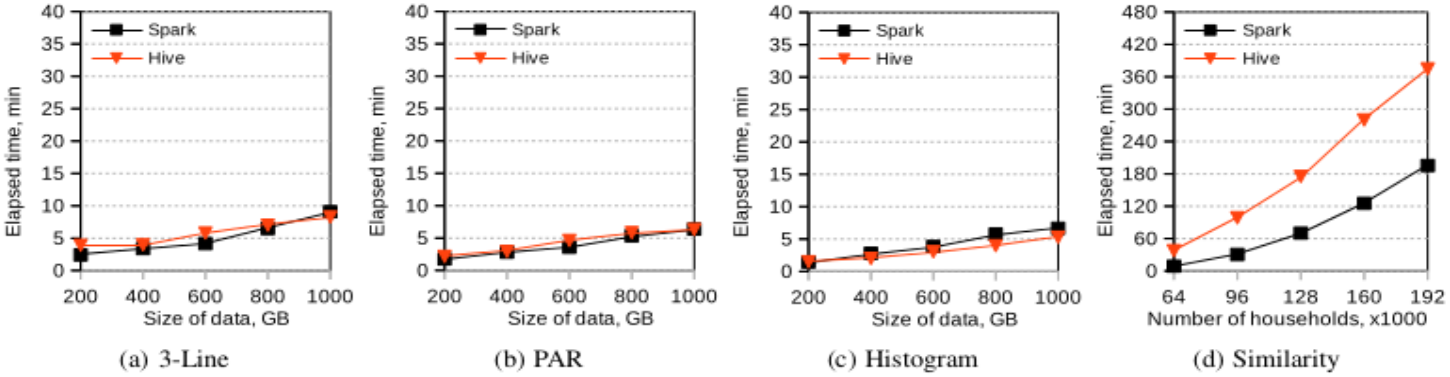


Figure 16: Execution times using the second data format in Spark and Hive.

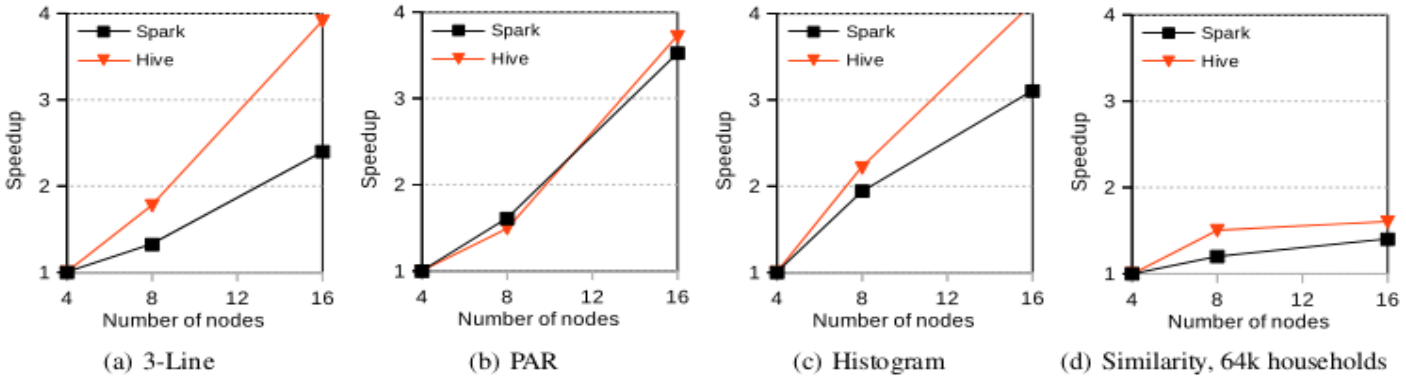


Figure 17: Speedup obtained using the second data format in Spark and Hive.

Experiments – distributed systems

- The execution times and speedups of the 3rd data format:

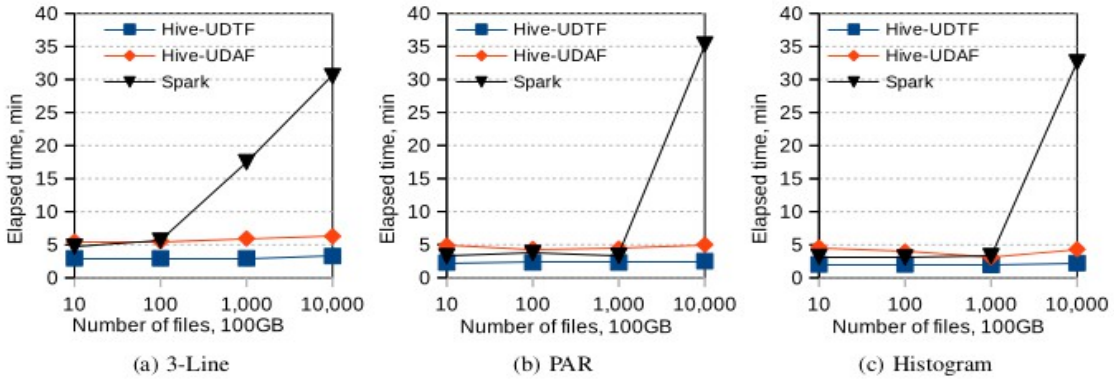


Figure 18: Execution times using the third data format in Spark and Hive.

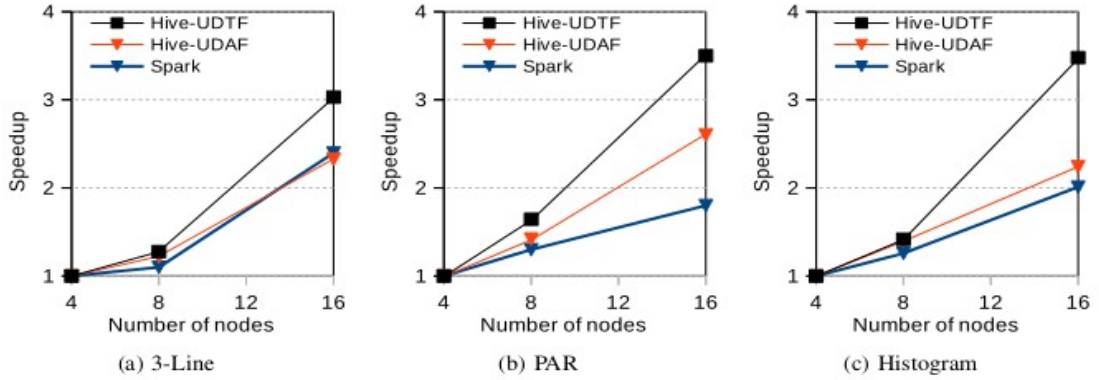


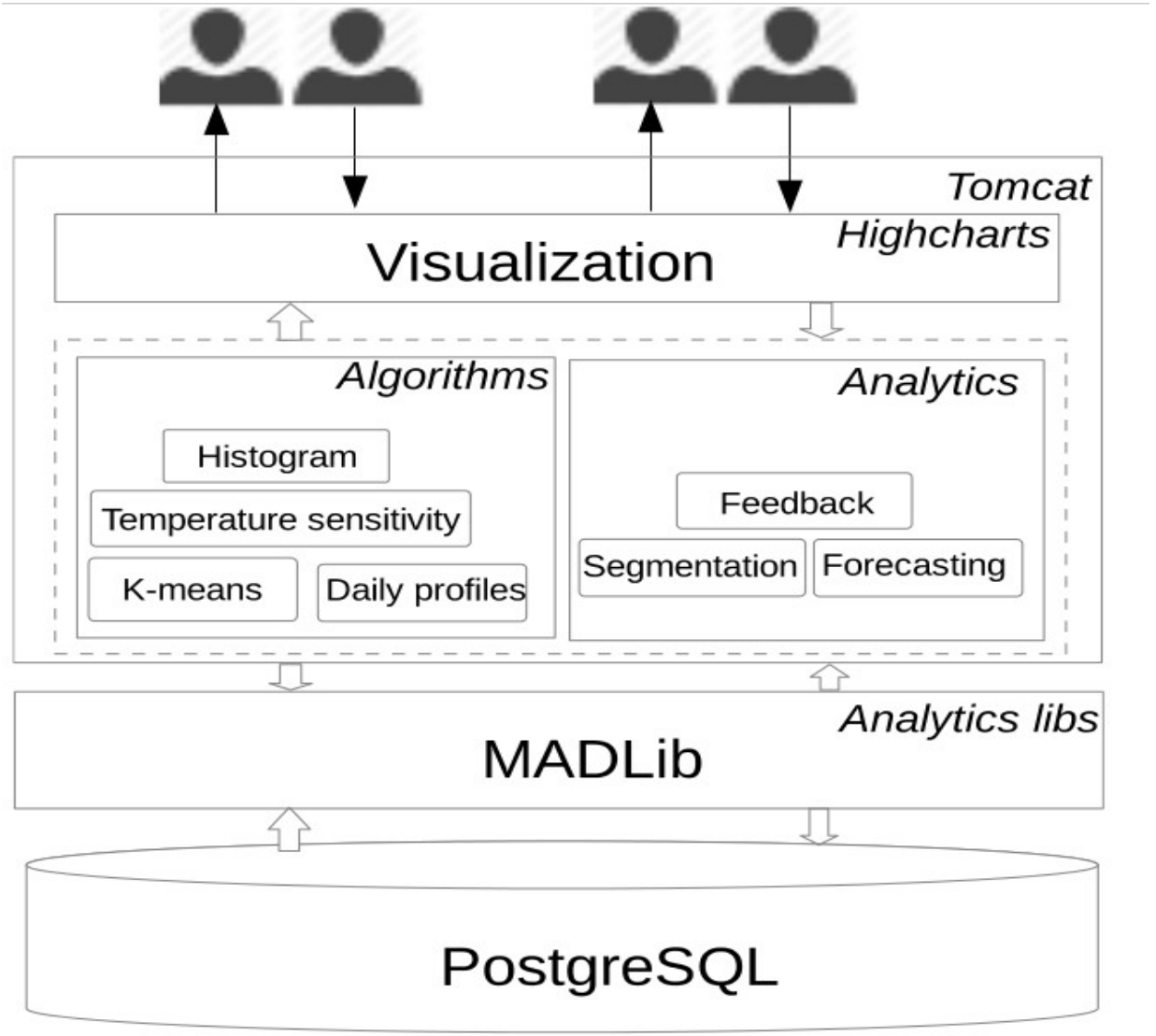
Figure 19: Speedup obtained using the third data format in Spark and Hive, 100 files, 1GB per file.

Summary

- Centralized Systems:
  - System C is the best choices for smart meter data analytics
  - Matlab and MADlib are more programmer-friendly but slower
  - Matlab works better for each time-series in a separate file
- Distributed Computing Systems:
  - Suitable for the analytics of large-scale data sets
  - Spark is faster than Hive, but Hive scale slightly better, and is easier to implement
- The data format does matter with the implementation, and the performance

Smart Meter Data Analytics System (SMAS)

System Architecture



Roles of the use

1. Utilities
2. Energy consultants
3. Energy consumers

Functionalities

- Energy consumption time series analytics
  - Time and location dimensions
  - Different granularities
- Segmentation analytics
  - Cluster customers with similar consumption patters
  - Show on Google map

Functionalities

- Energy demand forecasting
- Pattern discovery
  - Load profiling
  - Load distribution
  - Load disaggregation
- Consumption comparison
- Customer feedback

Demo

Open source - https://github.com/xiufengliu

Search GitHub Explore Gist Blog Help xiufengliu

Smart Meter Data Analytics

xiufengliu

University of Waterloo
Canada
groupme@gmail.com
Joined on 9 Sep 2009

0	0	0
Followers	Starred	Following

Contributions Repositories Public activity Edit profile

Popular repositories

Benchmark
The benchmark of smart meter data analytic technologies 0 ★

DataGenerator
Smart meter synthetic data generator 0 ★

SMAS
Smart Meter Data Analysis System 0 ★

Contributions

Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct
M
W
F

Summary of Pull Requests, issues opened, and commits. Learn more. Less More

Year of contributions	Longest streak	Current streak
74 total	3 days	0 days
Oct 27, 2013 – Oct 27, 2014	October 14 – October 16	Rock – Hard Place

Questions?