A Real-Time Analytics Platform for Large-scale Industrial Process Monitoring and Control
Tao Gong, Shaobo Zheng, Song Han, Mark Nixon

Motivation
In most process monitoring and control systems, data is typically sampled at the workstations and compressed before being archived in historians. Feedback is provided through alarm and control messages, and visualizations to both the system itself and to human operators. Analytics applications typically read data from these historians, process the data, and provide results and visualization off-line or in a time period that is considerably slow in comparison to the performance of the manufacturing process. Because of the focus on control, only a minimal set of process conditions and plant equipment is monitored. There needs to be a way to both improve the performance of the analytics system and scale the system to monitor a much larger set of plant resources in real-time.

Contributions
- Designed an efficient and scalable data management and analytics framework for large-scale industrial process monitoring and control systems.
- Large-volume distributed real-time data flows collected from heterogeneous data sources are streamed, stored, processed, and visualized in the proposed systems in near real-time.
- Designed a generic real-time streaming protocol, and a novel database schema on HBase for efficient data loading and retrieving.
- Implemented various decision making algorithms in process control applications into logical computation topologies on Storm and mapped them to physical machines in the cluster to fully utilize the parallel computational capability.

Data Flow and Real-Time Processing Framework
We used a combination of Kafka and Storm frameworks to achieve “at least once” delivery guarantee, horizontal scalability, and support real-time processing on time series data.

Research Goals
- Elastic computation model;
- Minimum resource utilization;
- End-to-end timing guarantee;

Efficient Time Series Database Design
We designed efficient database schemas for both raw and aggregated unstructured time series data on HBase, including both process measurements and network health information.

From Batch to Continuous Data Analytics
We are working on a data pipeline and toolset designed to specifically work with time-series and process data for product quality prediction and process fault detection.

Real-Time Data Visualization
We implemented a rich set of web services by accessing HBase through Thrift interface. This enables the users to visualize time series data and their statistics in a real-time manner.