Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Cloud Based Analytical Framework for Synchrophasor Data Analysis

Pavel Etingov

CIGRE Grid of the Future 2017

October 22-25 2017, Cleveland, Ohio

Project team

- Project is supported by the DOE through the Grid GMLC program

- PNNL
  - Pavel Etingov
  - Jason Hou
  - Heng Wang
  - Huiying Ren
  - Dimitri Zarzhitsky
  - Jacques de Chalendar (Stanford University)
- BPA
  - Dmitry Kosterev
  - Tony Faris
  - Steve Yang

Project background

- Active deployment of phasor measurement units (PMUs), smart meters, and other measurement devices dramatically increased the size of data collected by electrical utilities.
- The volume of the collected information is continuing to grow, that makes it very difficult to process it, run analysis and extract insights
- The collected data enables many insights about power system state and dynamic behavior
- Extracting this information can help:
  - Increase situation awareness.
  - Detect events in the system (e.g. under frequency or voltage events).
  - Detect abnormalities.

Project goals

- Develop a framework for PMU big data analysis
  - Event detection
  - Abnormalities detection
  - Improved situational awareness
  - System identification (learning system dynamic behavior)
  - Advanced visualization
- Framework is based on the cloud technology and distributed computing:
  - PNNL institutional cloud system or Microsoft Azure
  - Apache SPARK for distributed big data analysis and Machine Learning (ML)

PNNL cluster infrastructure

- PNNL cloud is based on OpenStack (a free and open-source software platform for cloud computing)

- Cloudera Apache Hadoop Distribution:

  - Apache Spark (an open source cluster computing framework)

  - Apache Hive (a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis)

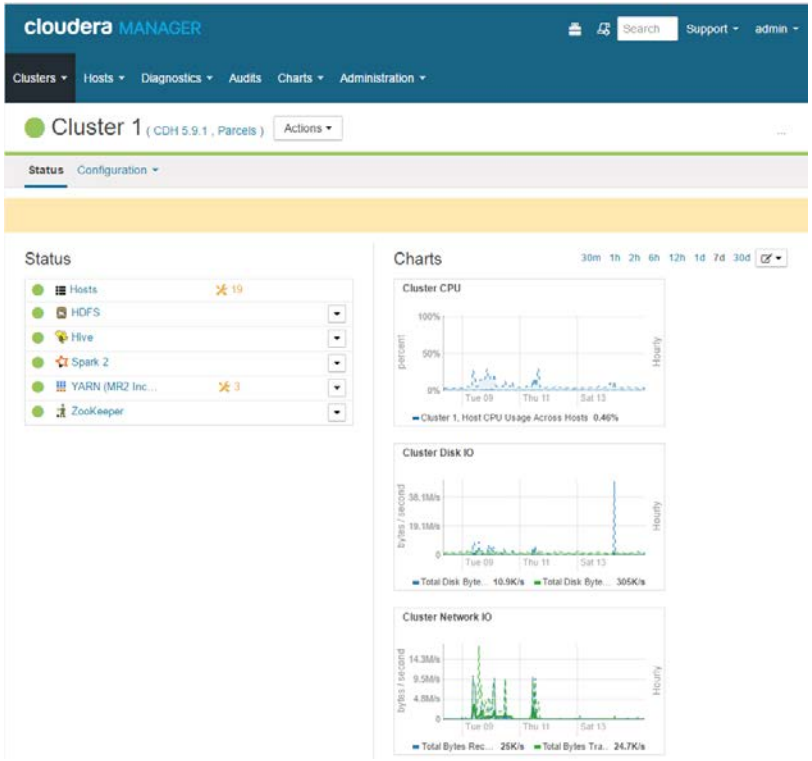
  - HBase (an open source, non-relational, distributed database)

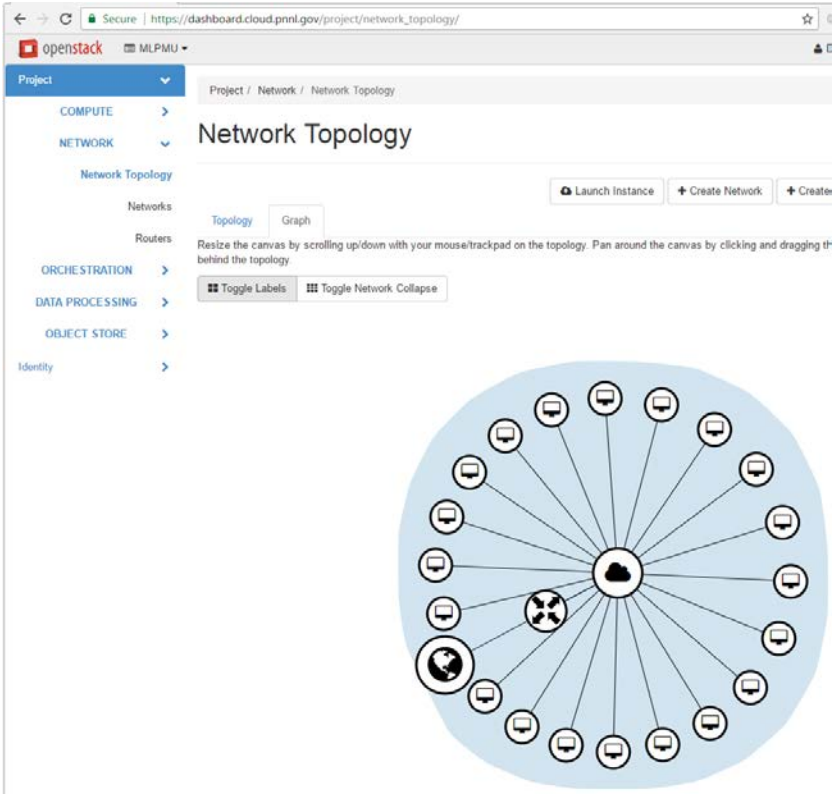
Apache Spark

- Large scale parallel data processing framework
- Extremely powerful (up to 100x faster than Hadoop)
- Large datasets distributed across multiple nodes within a computer cluster
- Support real time data stream
- Built-in Machine Learning library
- Support different languages (Scala, Java, Python, R)
- Support different data sources (SQL, Hive, HBase, Cassandra, Oracle, etc)
- Open source and free
- Available through public cloud services (Amazon AWS, Microsoft Azure, IBM, etc) and through new PNNL institutional cloud system.

Spark research cluster based on PNNL cloud

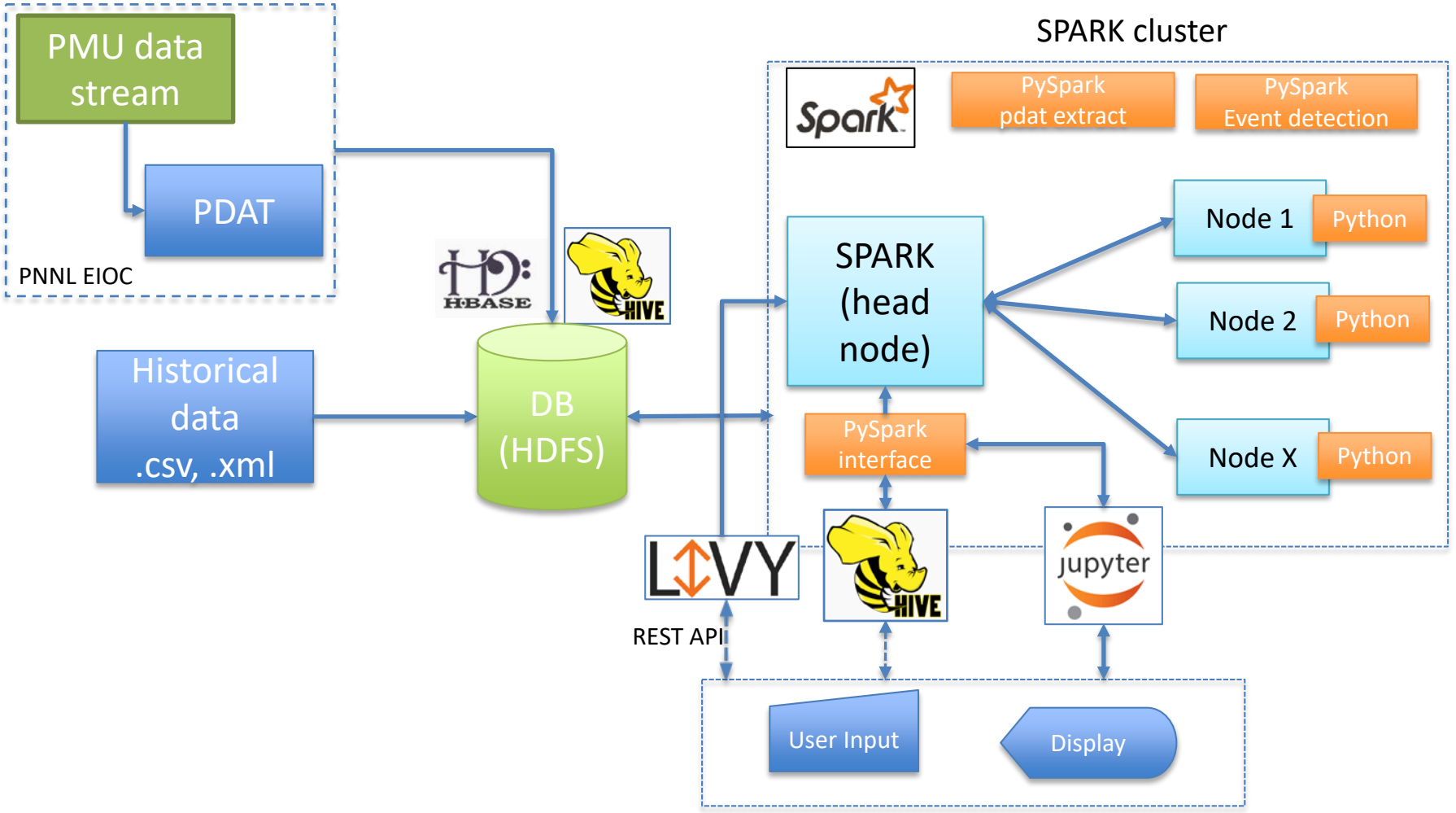
- Current configuration
  - 20 nodes
  - RAM 512 Gb
- Recently upgraded to Spark 2.2





- Cluster will be upgraded to 1 Tb RAM

Cloud based ML-PMU Framework



EIOC - Electricity Infrastructure Operations Center
HDFS- Hadoop Distributed File System

PMU data stream

- PNNL receives PMU data stream from Bonneville Power Administration
  - 12 PMUs
  - Multiple channels (Voltage and Current Phasors, Frequency, ROCOF)
- PMU Data stored in PDAT format
  - PDAT format developed by BPA
  - Based on IEEE Std. C37.118.2-2011
  - Binary files
  - Each file contains 1 minute of data
  - One file ~ 5 MB

Data frame organization defined by IEEE C37.118.2

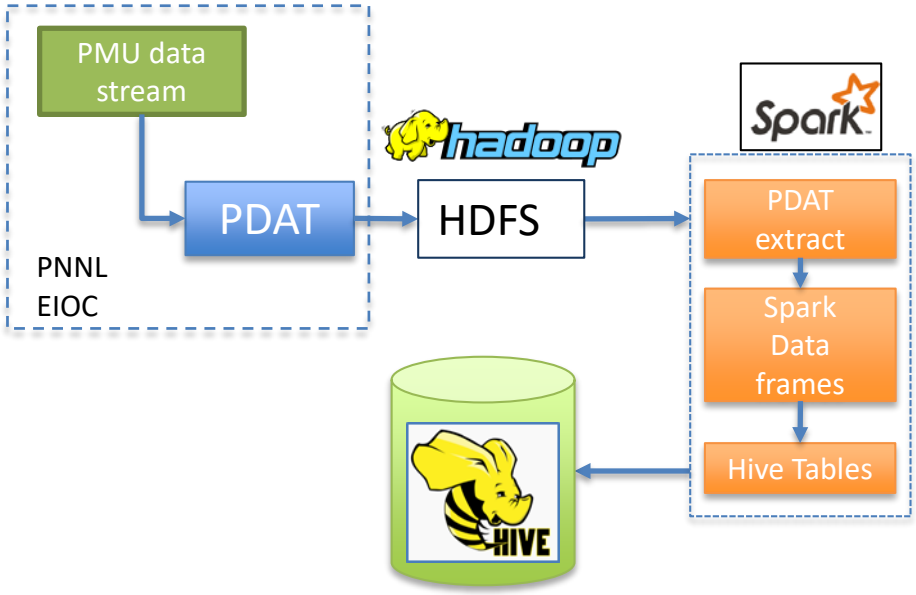
No.	Field	Size (bytes)	Comment
1	SYNC	2	Sync byte followed by frame type and version number.
2	FRAMESIZE	2	Number of bytes in frame, defined in 6.2.
3	IDCODE	2	Stream source ID number, 16-bit integer, defined in 6.2.
4	SOC	4	SOC time stamp, defined in 6.2, for all measurements in frame.
5	FRACSEC	4	Fraction of Second and Time Quality, defined in 6.2, for all measurements in frame.
6	STAT	2	Bit-mapped flags.
7	PHASORS	4 × PHNMR or 8 × PHNMR	Phasor estimates. May be single phase or 3-phase positive, negative, or zero sequence. Four or 8 bytes each depending on the fixed 16-bit or floating-point format used, as indicated by the FORMAT field in the configuration frame. The number of values is determined by the PHNMR field in configuration 1, 2, and 3 frames.
8	FREQ	2 / 4	Frequency (fixed or floating point).
9	DFREQ	2 / 4	ROCOF (fixed or floating point).
10	ANALOG	2 × ANNMR or 4 × ANNMR	Analog data, 2 or 4 bytes per value depending on fixed or floating-point format used, as indicated by the FORMAT field in configuration 1, 2, and 3 frames. The number of values is determined by the ANNMR field in configuration 1, 2, and 3 frames.
11	DIGITAL	2 × DGNMR	Digital data, usually representing 16 digital status points (channels). The number of values is determined by the DGNMR field in configuration 1, 2, and 3 frames.
	Repeat 6–11		Fields 6–11 are repeated for as many PMUs as in NUM_PMU field in configuration frame.
12+	CHK	2	CRC-CCITT

Ongoing work

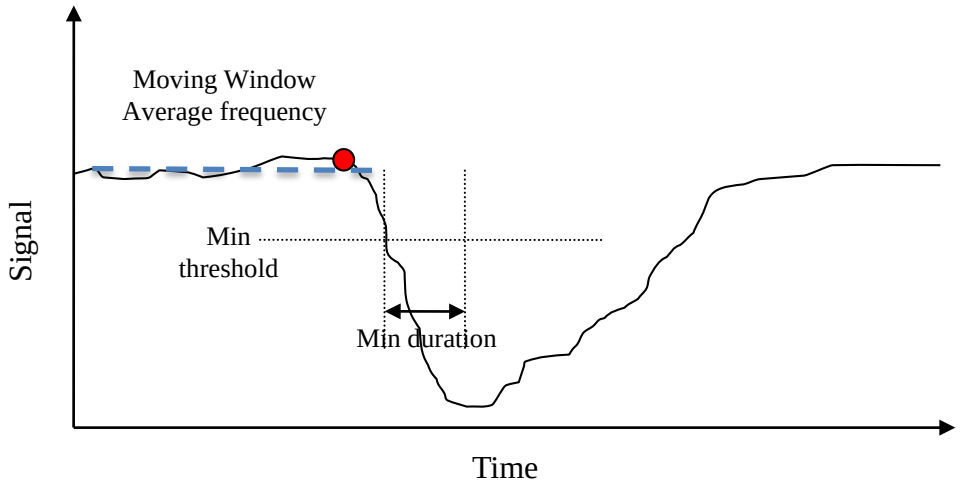
- Python (PySpark) modules:
  - PDAT data extraction
  - Data processing
    - Bad data
    - Missing points
    - Outliers
  - Event detection
    - Frequency events
    - Voltage events
  - Features extraction and analysis
    - Wavelet
    - K-mean Clustering
    - Principal component analysis

PDAT data extraction

- Read information from PDAT and creates SPARK data frames
- Store information in Hive tables
- Implemented in PySpark that allows parallel processing of multiple PDAT files
- Significantly increased performance
  - To read information for 1 hour takes about 20 seconds (20 nodes cluster)



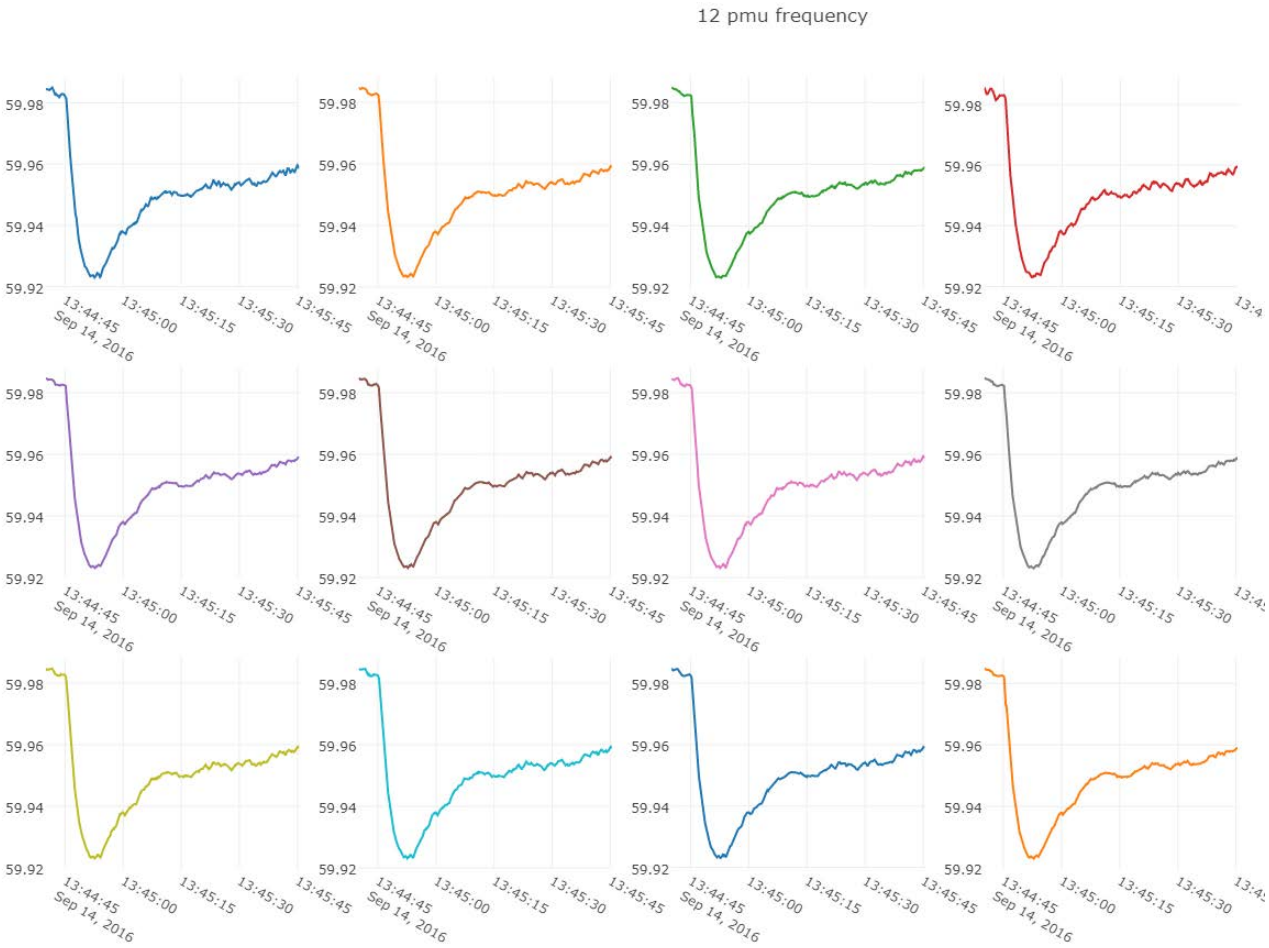
Event detection (threshold based)



- User specified
  - Delta frequency
  - Event duration
- Cross validation signal checks to avoid false alarms
- Spark usage significantly increases the computational throughput of the application
- Processing of 1 day takes about 5-7 minutes (processing the same dataset using a PC takes about 1 hour)

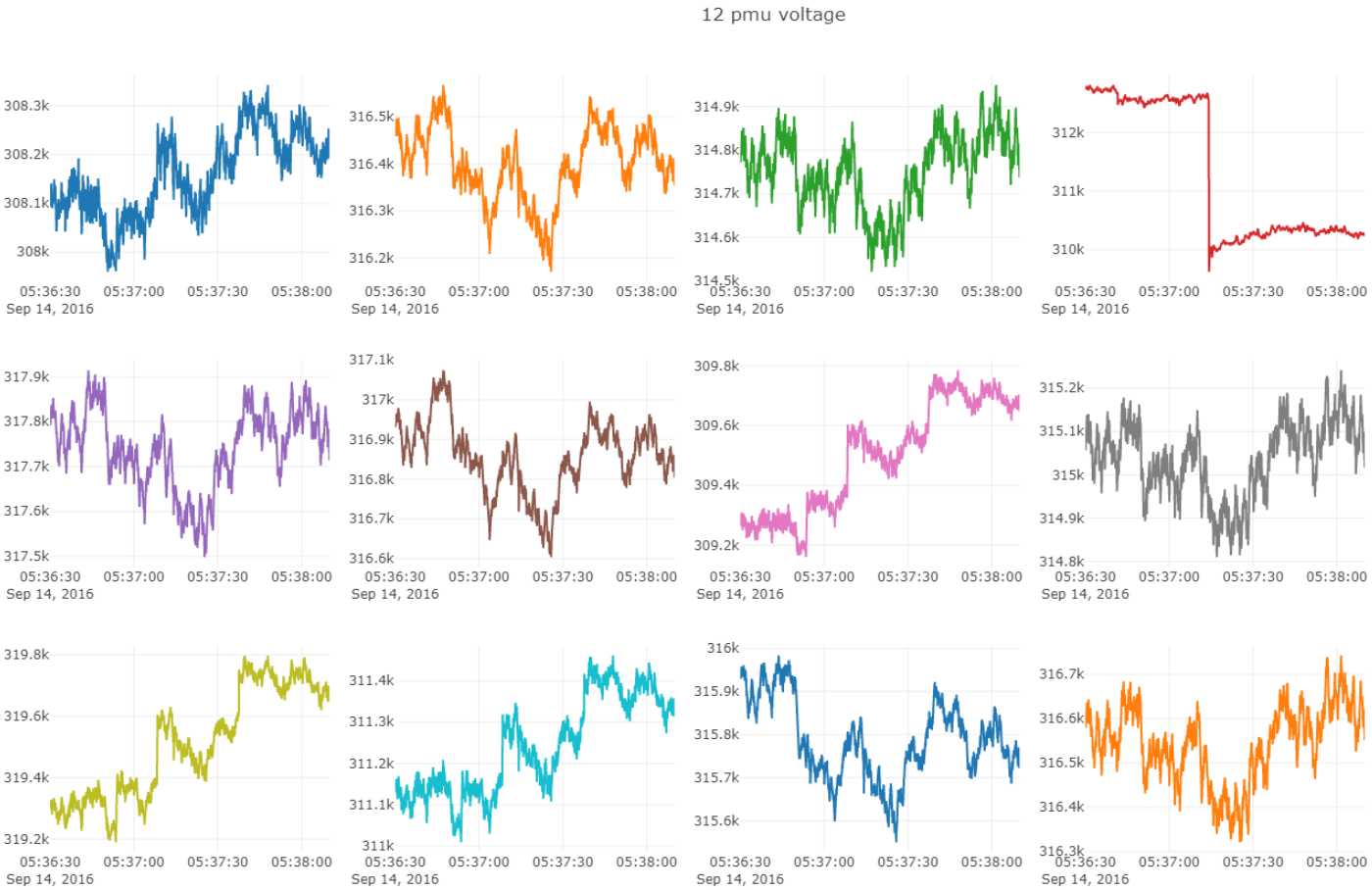
Examples of Detected events

- Frequency events



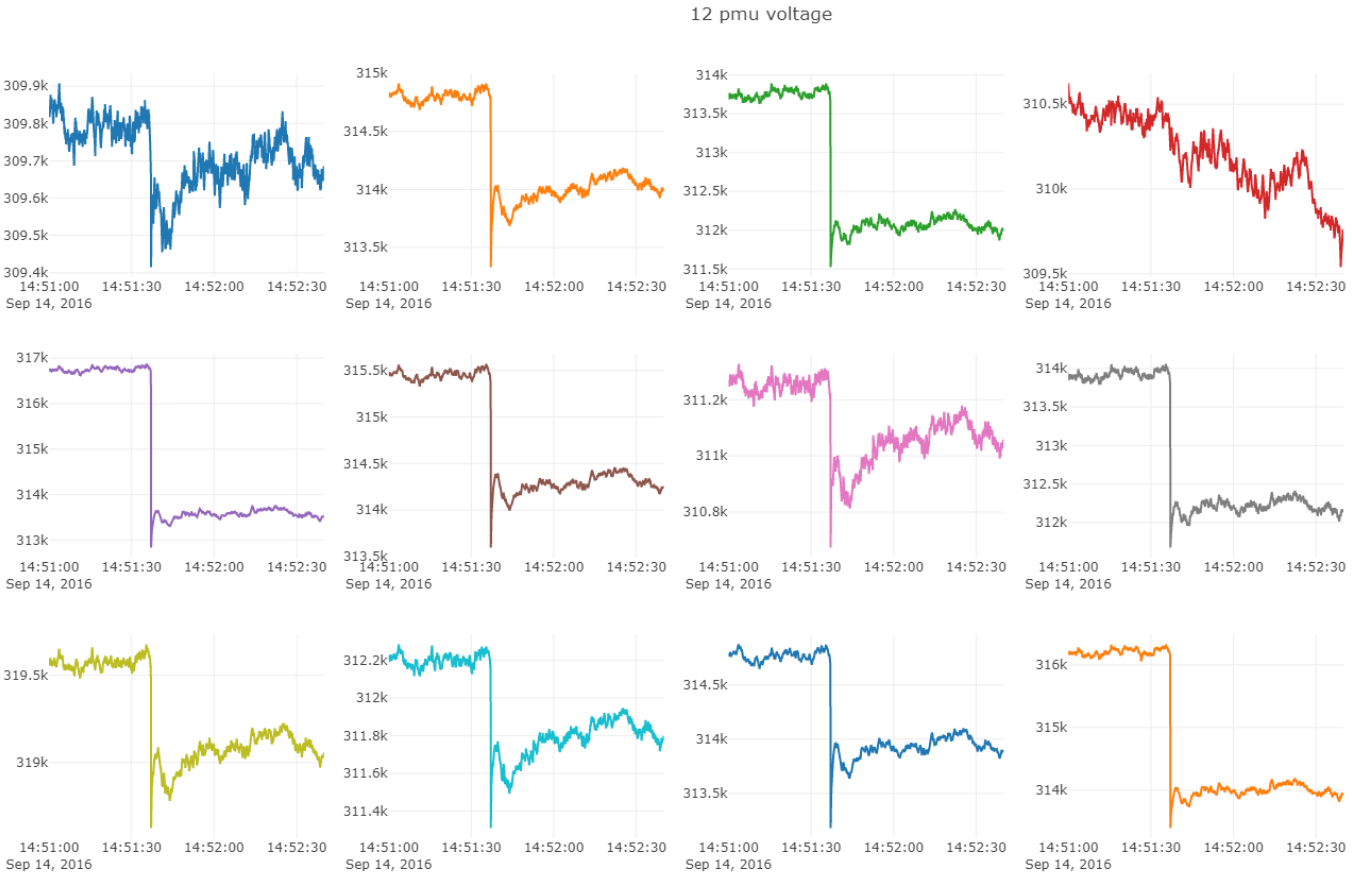
Examples of Detected events

- Voltage event



Examples of Detected events

- Voltage event



Anomaly Detection based on Wavelet Analysis

PMU Signals

Voltage phasors

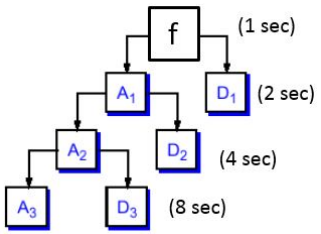
Frequency

ROCOF

Wavelet-based multi-resolution analysis (MRA)

Signals are transformed into time-frequency domain

MRA decomposes signals into approximation (A) and detail (D) components



Wavelet coefficients extraction

Extract the detail wavelet coefficients at scales matching the durations of events

Moving window based anomaly candidates

Identify anomaly candidates at multiresolution levels with moving-window-based anomaly detection

5-min bandwidth of moving window

Anomaly scoring

Anomalous score is set to be 1 when an anomaly is detected at each resolution level.

The anomaly score matrices are the summation of scores at multi-resolution levels across all PMUs

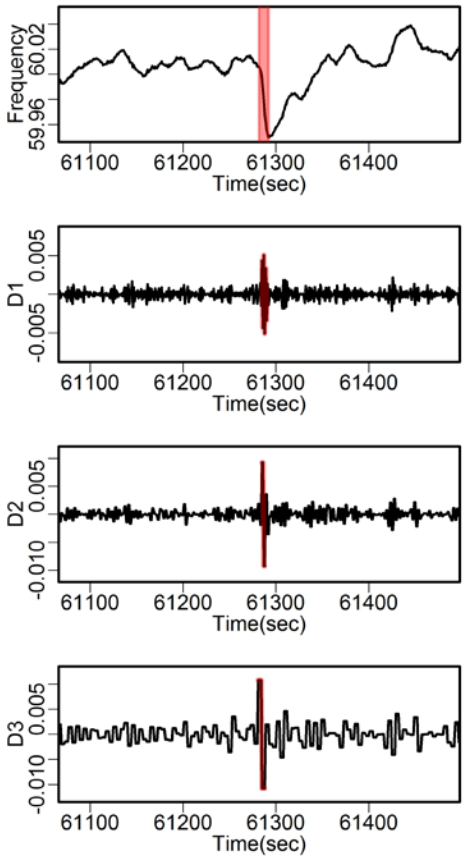
Spatiotemporal correlation analysis

False alarms or localized events corresponds to relatively weak spatiotemporal correlations

Final list of detected anomalies

Anomaly Scoring and Verification

- The anomaly score matrices were calculated across 12 PMUs at multiresolution levels for each PMU attribute.

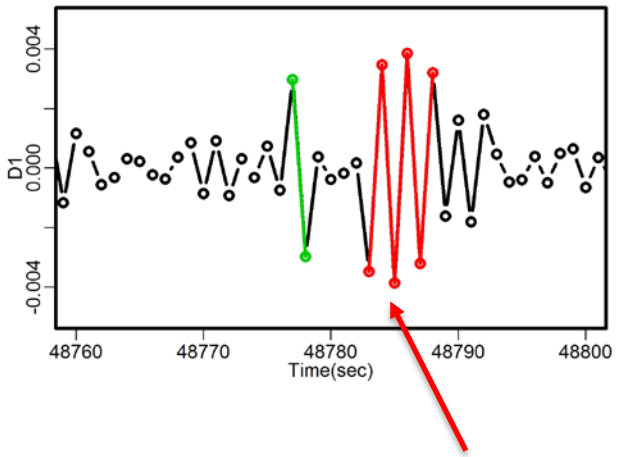


(a) Frequency signal

(b) MRA wavelet coefficient at D1;

(c) MRA wavelet coefficient at D2;

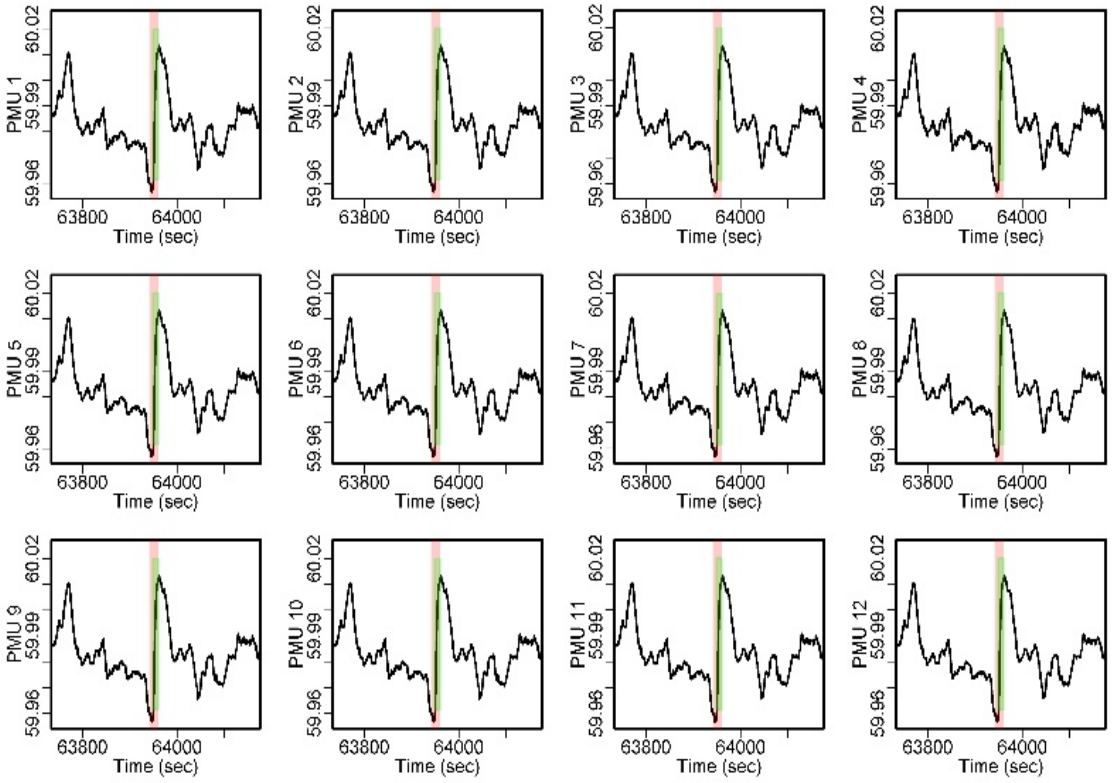
(d) MRA wavelet coefficient at D3.



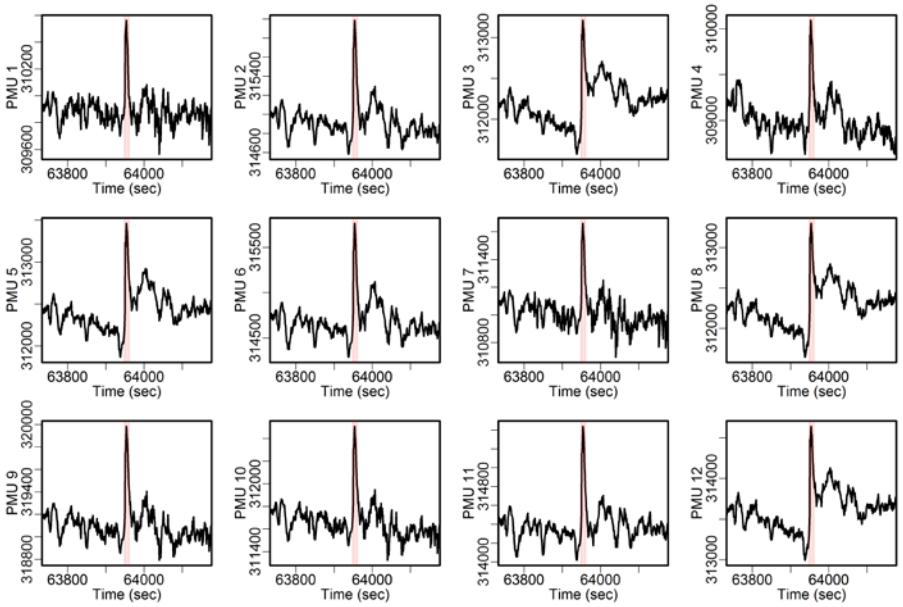
More than 3 sequential points exceeded the threshold and counted as an event. +1 added to the anomaly score matrices.

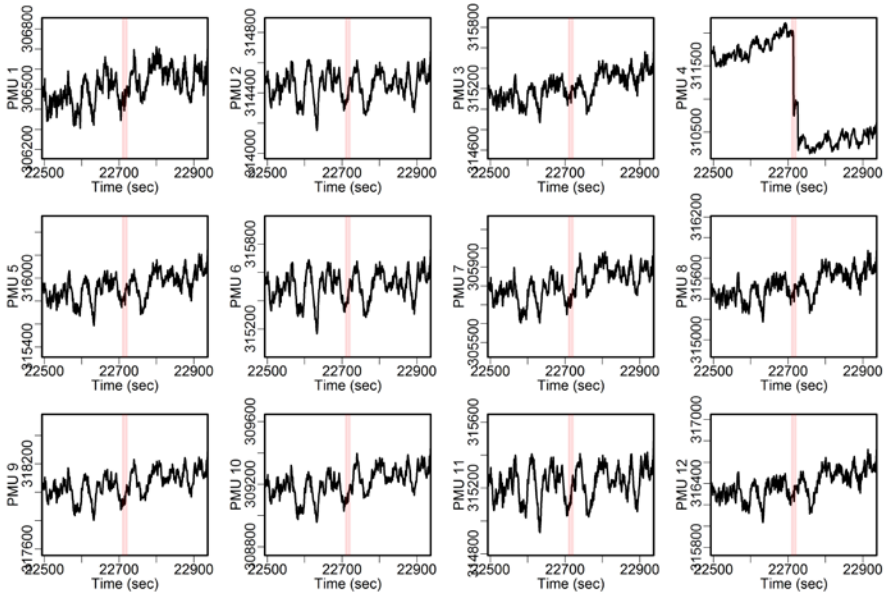
Frequency Anomalies

- Frequency event
  - False alarm (bad data) weak spatiotemporal correlation (weak continuity, weak correlation across units, but note local event is possible)
  - Real event → strong correlation across units, relatively weak temporal correlation with a bandwidth of about 5~20seconds



Voltage Anomalies





Preliminary Results

- Spark cluster for ML and PMU (big data) analysis was deployed. It is based on the PNNL institution cloud system
- PMU data has been collecting in PDAT format (PMU data stream from PBA to PNNL EIOC)
- Methodology for event detection based on wavelet analysis has been developed
  - Enhanced robustness to bad data
- Python (PySpark) modules are under development
  - PDAT data extraction
  - Event detection (based on thresholds)
  - Wavelet anomaly detection
  - Event classification based on PCA and clustering