Paper No: 21ISGT1193

GE Research

Big Data Analysis of Massive PMU Datasets: A Data Platform Perspective

Vijay S. Kumar, Tianyi Wang, Kareem S. Aggour,
Pengyuan Wang, Philip J. Hart, Weizhong Yan
GE Research
v.kumar1@ge.com

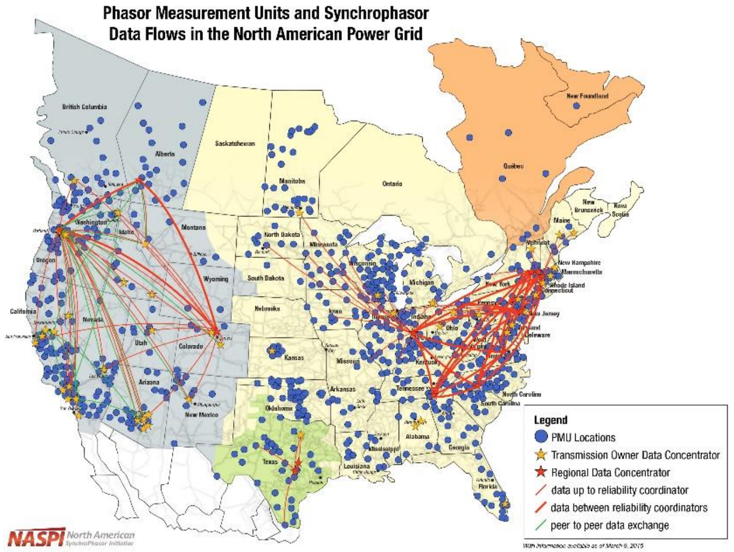
IEEE PES Power & Energy Society®

IEEE

Background

The Department of Energy (DOE) assembled a phasor measurement unit (PMU) dataset from 443 PMUs across Eastern, Western, and Texas interconnects, along with event logs w/ 1000s of recorded events.

- Overall program objectives:
  - Apply big data, AI & ML technology and capabilities to extract new insights, such as validated grid event signatures (generator trip, line fault, etc.),
  - Develop systems and tools for effective grid operation and management.
- This work presents a custom-built **data platform** that we successfully applied to support our **feature-generation-intensive ML** strategy for grid event signature identification.
- Expected outcomes of data platform:
  - Offline analysis (data quality assessment, pre-processing) over massive PMU data
  - Easier feature generation by power systems SMEs and data scientists
  - Scalable and reliable generation and storage of (tens of millions of) features for normality modeling and signature identification



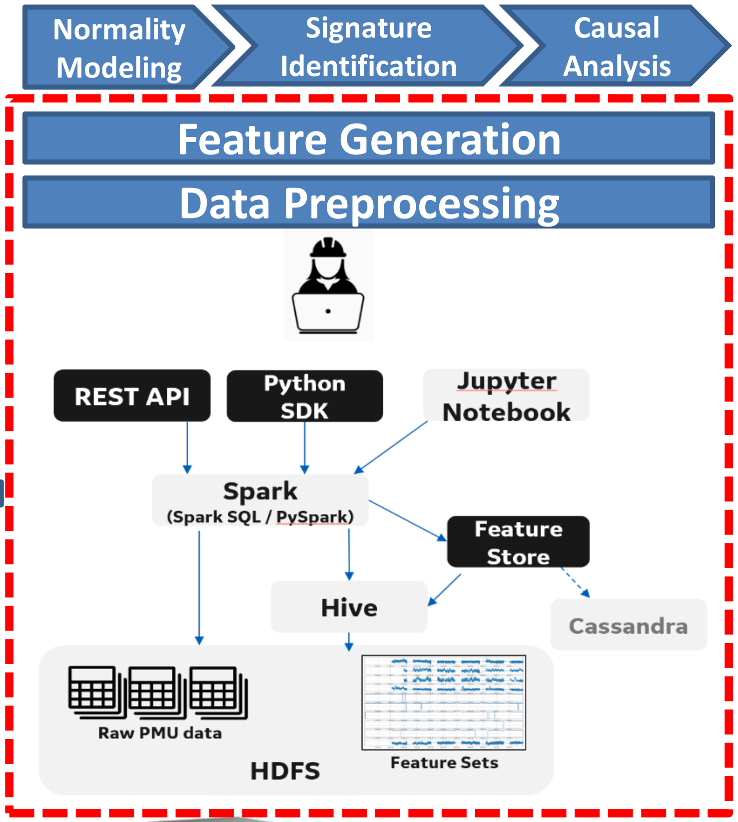
[Image courtesy of NASPI, https://www.naspi.org/]

Interconnect	# PMUs	# records	Compressed data size (Terabytes)
IC_A	212	160,809,031,796	2.9
IC_B	43	93,353,826,102	4.7
IC_C	188	241,437,700,843	11.0
Total	443	495,600,558,741	18.5 TB

Training Dataset size characteristics (two-year data; sampled at 30-60Hz)

Data Platform for PMU Big Data Analysis

- **Data Lake** architecture
  - PMU dataset (Parquet files) loaded into Hadoop; Schema defined in Hive
  - Can serve multiple users and analysis applications at the same time
- Key contributions
  - software abstractions/APIs for easier access to and processing of massive real-world PMU data (**layered feature generation framework; feature store**)
  - targeted performance optimizations for large-scale feature generation



Feature Function
- Computes one or more values over raw data

Feature Wrapper
- Group feature functions together based on commonalities
- Prepare input data; organize results

Feature Gen. Engine
- Raw data loading; pre- and post-processing tasks
- Apply rolling window over time

Feature Gen. Executor
- Distribute execution across cluster
- Parallelize by time and/or feature batch

Feature Store
- Common APIs to store, read, update, delete batches of features
- Extract specific features for given PMUs and time range

Layered feature generation framework & feature store dedicated to PMU data

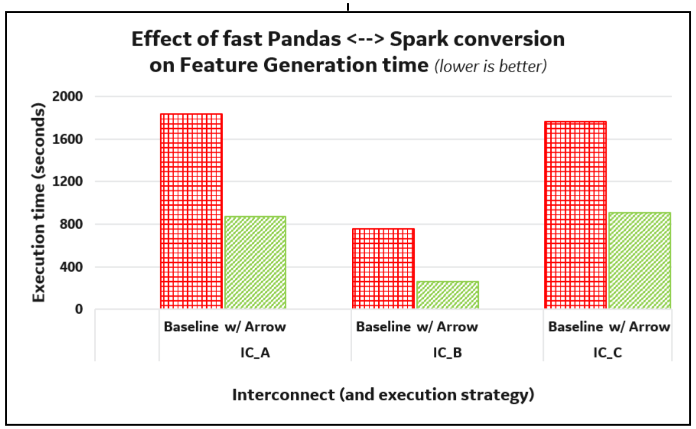
On-premise cluster:
- *2000+ CPU cores*
- *23 TB RAM*
- *370 TB DAS*

Results

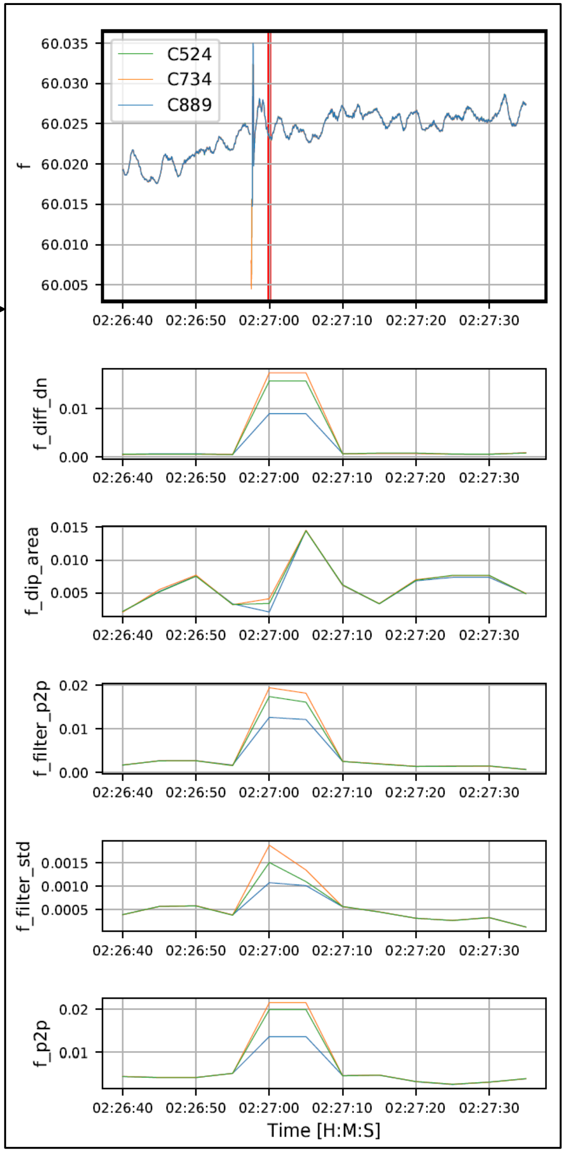
Feature name	Raw signal channel	Description
f_diff_dn	**f** *(frequency)*	Maximum step down in 0.1 second
f_filter_p2p	**f** *(frequency)*	Peak-to-peak value after filtering out 1st principal component among all PMUs; used to characterize asynchronization with peers.
vm_diff_dn	**vp_m** *(voltage magnitude)*	Maximum step down in 0.1 second
vm_diff_up	**vp_m** *(voltage magnitude)*	Maximum step up in 0.1 second
vm_p2p	**vp_m** *(voltage magnitude)*	Peak-to-peak value
im_std	**ip_m** *(current magnitude)*	Standard deviation
im_diff_dn	**ip_m** *(current magnitude)*	Maximum step down in 0.1 second
im_RP	**ip_m** *(current magnitude)*	Exhibition of strong frequency components in the signal; used to characterize oscillations.
p_diff_up	**p** *(active power)*	Maximum step down in 0.1 second

Sample features

- *Overall: 60+ feature functions to be calculated per every 5 seconds of raw data*
- *Across all PMUs in an interconnect; grouped into 7+ feature batches*



Targeted Performance Optimizations



Resulting **productivity gains** offer power and grid systems researchers significant advantages

- e.g., 89 million feature values per PMU in IC_B (23.5 GB) in ~ 50 minutes
- flexible feature store (add, update, delete, query feature batches)

Conclusions/Recommendations

- Analysis of real-world PMU datasets has associated practical challenges
  - Data volume, spatiotemporal/heterogeneous nature, data & label quality
  - Feature-engineering based ML strategy needed to uncover grid event signatures
- Foundational data platform on which to build advanced tools for grid systems operation and management
  - Successfully applied to offline analysis of a massive real-world (DOE) PMU dataset
  - Bad data analysis, massively-parallel feature generation, scalable feature storage
- Custom-built data platform dedicated to grid data (e.g., PMU) can outperform generic, turn-key Big Data tools and service offerings.

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000915.