Pacific Northwest
NATIONAL LABORATORY

PNNL-32303

FOA 1861 Data Curation Overview

October 2021

Jeffery S Banning
James Follum
Eric S Andersen

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-32303

FOA 1861 Data Curation Overview

October 2021

Jeffery S Banning
James Follum
Eric S Andersen

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

This document describes the process executed to collect, examine, and consolidate Phasor Measurement Unit (PMU) data from multiple transmission operators into a common dataset. The consolidated PMU data set was further anonymized and distributed to the Department of Energy Funding Opportunity Announcement (FOA) 1861 Big Data Analysis of Synchrophasor Data awardees

Acknowledgments

The authors would like to acknowledge Sandra Jenkins of DOE's Office of Electricity and Carol Painter of DOE's National Energy Technology Laboratory (NETL) for supporting and guiding the work presented here. We also thank the organizations that shared their measurement data, along with the individuals from those organizations who took the time to extract it. Finally, we would like to acknowledge Alison Silverstein of Alison Silverstein Consulting for her integral role in conceiving and launching the FOA 1861 effort.

Acronyms and Abbreviations

CII	Critical Infrastructure Information
CSV	comma-separated values
DOE	Department of Energy
ePDC	enhanced Phasor Data Concentrator
EPG	Electric Power Group
ETL	Extract-Transform-Load
FOA	Funding Opportunity Announcement
ID	identification
MW	mega-Watts
NDA	Non-Disclosure Agreement
NERC	North American Electric Reliability Corporation
ORC	Optimized Row Columnar
PMU	Phasor Measurement Unit
PNNL	Pacific Northwest National Laboratory
QC	quality control
RC	reliability coordinator
SEL	Schweitzer Engineering Laboratories
TOPS	Transmission Operators
UTC	Coordinated Universal Time

Contents

Tables

1.0 Background

Over the past two decades, utilities in the North American Bulk Electric System have deployed time-synchronized telemetry creating unprecedented amounts of data to support operations, planning, and research. Some of this data is shared between utilities for shared wide area situation awareness and to support smooth and reliable operations of the electric grid. In order to better gain insights from the large amount of data, sophisticated data analytics/machine learning based strategies may be required.

This time-synchronized synchrophasor data is of great value in the scientific community and can be hard to acquire. The North American Electric Reliability Corporation (NERC) deems its use as critical infrastructure information (CII) when used for real-time operational decision-making, and the data needs to be protected. Furthermore, even if not used for real-time operational decision-making, utility datasets may contain other sensitive operational information that they want to protect for business purposes. Agreements have been established for sharing data between utilities, such as the Western Interconnection Data Sharing Agreement and the agreement for the Eastern Interconnection Data Sharing Network.

Many utilities are willing to contribute data for research purposes but do so with restrictions on how the data can be used, for how long it can be used, and what to do with the data once it's no longer being used. For the most part, stipulations for data use are captured in non-disclosure agreements (NDAs) between the utility and the data user. Pacific Northwest National Laboratory's (PNNL's) Electricity Infrastructure Operations Center serves as a research data repository for utility data and has put significant effort toward managing and curating these types of unique utility datasets.

1.1 Introduction

The Department of Energy (DOE) Office of Electricity's Transmission Reliability Research Program needed three large synchrophasor datasets, one from each of the three U.S.-based interconnections, with multiple utilities to support eight big data analytics, machine learning, and artificial intelligence research projects (Awardees) being authorized under Funding Opportunity Announcement (FOA) 1861. At the onset of negotiating agreements with the data providing utilities, a primary concern was how these large datasets would be protected from revealing sensitive data to the FOA awardees, and if the data could somehow be anonymized as part of the aggregation and assembly process.

In order to create such large Phasor Measurement Unit (PMU) datasets, two years of data from multiple transmission operators (TOPS) across the three interconnections were collected by PNNL. All the data was received in disparate media, binary format, and data structure format. All the data had to be transformed in order to create a common structure to support the research projects and had to be anonymized prior to being released to the awardees. This report will provide an overview of the steps performed to transform the "raw" utility data into three common, aggregated, and anonymized datasets that were distributed to the FOA 1861 awardees.

1.2 Defining the Datasets Requirement

The synchrophasor datasets to support the anticipated machine learning and artificial intelligence projects needed to be from geographically dispersed utilities across each of the

three primary, U.S.-based transmission interconnections and had to be large enough to be considered “big data”.

A set of requirements were prepared to help the participating utilities understand what was being asked for and to help with identifying adequate datasets to support the research.

The following is a list of requirements[1] for the data we sought to obtain from our partner utilities:

1) *Types of Data Requested from each entity*
    a) *PMU data from each of the three interconnections*
        (1) *Full fidelity (usually 30 or 60 Hz) time-stamped data from PMUs from 12 or more distinct topological locations across the data provider’s footprint. This can include PMUs that were not on-line over the entire two-year period.*
        (2) *Each data stream should include (at a minimum) measures of frequency, 3-phase/positive sequence voltage, current and phase angle for at least one location (e.g. phasor, site) for each PMU.*
        (3) *PMUs may be selected based on factors including data quality, data completeness, and representative of a wide area (i.e. not signals right next to each other). We would prefer PMUs with complete signal streams rather than PMUs that sent mostly 00s over the two-year period.*
        (4) *Any available information about data quality identification should be included. This could be flags and the definitions of the flag values or NA values (i.e. 0, Na, Nan).*
        (5) *There is no need to clean the data before submitting it.*
    b) *Complementary datasets*
        (1) *SCADA or state estimator data – A few signals from locations proximate to the locations for provided PMU data, if the utility is willing to share the common location with the researchers. (PNNL will be anonymizing location names but will maintain the correspondence between specific locations across multiple data types.) SCADA or state estimator data should include time identifiers to enable correlation with the PMU data.*
        (2) *Event logs -- This data will help the researchers identify certain events, with the understanding that not all events are contained in the logs. Event logs should contain time and type of event. Events could be de-identified if desired by the utility (e.g. similar events may be referred to as Event A-1, Event A-2, etc.). Ideally, the event log could include information on which PMUs detected the event or what rules are used to identify a certain type of event. (Specific PMU and SCADA identities and locations, and corresponding data from event logs, will be anonymized.)*
2) *Data Timeframe -- We will be asking for data covering the period from 1/1/2016 through 1/1/2018 (2 years and a day), to be collected in the fall of 2018.*

[1] PNNL-27857, PNNL Plan for Acquiring, Anonymizing, and Protecting Utility Data

3) *Format and Transportation of Data*
    a) *The data owners should place the data into a commonly used PMU data file format and other common industry data formats for non-PMU data. Data should include time stamps.*
    b) *Data can be stored on portable hard drive(s) and sent to PNNL using FedEx or another agreed upon method (including secure on-line methods).*

Establishing a robust set of requirements up front helped with negotiating with the utilities to provide the data needed for the research.

2.0 Collection and Examination of Data

Determining which organizations were willing and able to contribute PMU data for 2016 and 2017 was an arduous process. Extracting data for such a long time period is a time-intensive task, and not one all organizations were able to commit to. Even for organizations willing to extract the measurements, the requested time period was not always available. Once data was delivered to PNNL, efficiently managing 70TB of raw PMU data from disparate archive systems was a significant task. The process of requesting, acquiring, and managing the raw dataset is described in the following subsections.

2.1 Data Authorization

PNNL has a long-standing reputation with utilities as a neutral party and honest brokers of technology. Members of the FOA 1861 project team used their personal and professional utility contacts to reach out to staff at various TOPS to see if they had archived PMU data available and if they would be willing to provide two years (2016 - 2017 and 2018 - 2019) of that data. Finding utilities with two years' worth of archived data turned out to be somewhat of a challenge, as many TOPS do not store their PMU data beyond what they need to meet regulatory requirements. To compound the difficulty, some TOPS in the west indicated that they were willing to share their data, but their data was archived by Peak Reliability, and Peak was well into the process of shutting down and did not have the resources available to support the retrieval of archived data. And some TOPS simply were not comfortable with sharing their PMU data.

After the list of TOPS was reduced by willingness and availability, the remaining TOPS required an NDA before they would share the data. For some TOPS, the NDA process was simple and fast. But others required quite a few iterations before the NDA language was acceptable by both parties.

2.2 Data Acquisition

Once the NDAs were complete, it was necessary to work with the TOPS primary PMU contact to determine the most efficient and effective way to retrieve data from their respective PMU archives and transfer it to PNNL. Multiple archive strategies are being used by the TOPS, including:

- TOPS-developed proprietary, file-based binary archive system
- Schweitzer Engineering Laboratories (SEL) Synchrowave proprietary, file-based binary archive system
- Electric Power Group (EPG) enhanced Phasor Data Concentrator (ePDC) SQL Server Database archive system
- OSIsoft PI proprietary file-based binary archive system.

Depending on the PMU archive system used to store the historical data, the process used to export the data into a usable format took varying paths. For example:

- Exported data per our specifications from proprietary binary files to comma-separated values (CSV) files. CSV files were sent to PNNL via external USB hard drives.

- Sent PNNL proprietary binary files via external USB hard drives. Included was a software tool to extract CSV files from the binary files.
- Sent PNNL SEL Synchrowave binary files via upload to an Azure Blob storage site. PNNL used a SEL Synchrowave tool to export the data to CSV files.
- Some TOPS relied on their regional reliability coordinator (RC) to archive historical PMU data. PNNL worked with the RC to export data from their archives to CSV files. The export process took up to four weeks, in some cases. Once the export was complete, the CSV files were sent to PNNL on external USB hard drives.
- Sent PNNL a SQL Server backup of their EPG ePDC database via external USB hard drive. The database was restored to a local SQL Server and then the EPG export tool was used to extract the data into CSV files.
- Worked with one TOPS to devise a way to bulk export data from an OSIsoft PI archive server into CSV files. Once the export was complete, the CSV files were sent to PNNL via RDX removable drive cartridges.

The resultant raw data provided is summarized in Table 1, below.

Table 1. Utility Data Summary.

Summary of Utility Synchrophasor Data Contributed			
Interconnection	**Number of Data Providers (Utilities)**	**Total Number of PMUs**	**Size of the Received Raw Data (TB)**
Eastern Interconnection (EI)	5	250	38.6
Electric Reliability Council of Texas (ERCOT)	1	221	10.6
Western Interconnection (WI)	3	43	19.0
Total	9	514	68.2

2.3 Initial Data Structure

All data needed to be in a CSV format as a starting point for the creation of the FOA 1861 PMU dataset. As mentioned earlier, some data was received in a CSV format, and other data was converted to CSV at PNNL. One aspect of the CSV conversion is that the processes that create the CSV files do not change the PMU data. Depending on the export tool used, the individual PMU data fields and the time range of data included could be adjusted, but the raw values of the data are not changed during this process.

Once all the TOPS archived PMU data was in a CSV format, it was necessary to inspect the internal format of each TOPS CSV file. It turns out that no TOPS CSV format was exactly the same. Inspecting each format revealed significant differences.

2.3.1 Number of PMUs Per File

A PMU consists of multiple, independent measurement channels: Frequency, Rate of Change of Frequency, PMU Status, Voltage Magnitude/Angle (Pos Seq, A/B/C Phases), and Current Magnitude/Angle (Pos Seq, A/B/C Phases). In addition, a single PMU may include data from

multiple voltage buses and current lines. The “raw” CSV files came in the following two flavors, with regard to the number of PMUs that exist in one CSV file:

- One set of unique PMU data per CSV file for a specific date range
- More than one set of unique PMU data per CSV file for a specific date range.

The CSV files that contained multiple unique PMU data per file had anywhere from 15 to 100 PMUs per file. This created CSV files from 100 to over 1,000 columns.

2.3.2 Type of Data Fields Per PMU

Each TOPS made independent decisions about substation and PMU configurations. That led to a wide variety of data that was transmitted/archived by the different TOPS. The most common combination of data for a single PMU was frequency, rate of change of frequency, positive sequence voltage angle and magnitude from a single bus, and positive sequence current angle and magnitude from a single line. But there were many PMUs that included the A/B/C phases for both voltage and current, the PMU status flag value, and voltage/current data from multiple busses and lines.

2.3.3 Additional Data Differences

Most PMUs were archived at 30 samples per second, but some TOPS archived the data at 60 samples per second. The Coordinated Universal Time (UTC) timestamp values had some differences in the date format and in how the sub-second values were recorded. The total number of UTC millisecond decimal places differed, as well as slight rounding differences. And the voltage magnitudes could be stored as Volts or kiloVolts.

2.4 Big Data Structure Decisions

There were three primary decisions concerning how the FOA 1861 dataset was to be constructed.

- What PMU data to include
- What file format to store the dataset
- How to anonymize the data

2.4.1 PMU Field Selection

As mentioned earlier, the data collected by a single PMU varies greatly. Some of the factors related to the PMU data received and archived are based on TOPS preferences, substation configuration, and available network bandwidth. Discussions with TOPS contacts and resident PNNL PMU subject matter experts and inspection of sample data provided by TOPS showed that the majority of PMU data archived for a single PMU had the following data:

- UTC Timestamp
- Single Bus Positive Sequence Voltage Magnitude and Angle
- Single Line Positive Sequence Current Magnitude and Angle
- Frequency

- Rate of Change of Frequency.

Roughly 95 percent of all PMU data received included these fields, at a minimum. The primary exception was that one TOPS only recorded the three phase values for voltage and current but not the positive sequence value.

Some data included positive sequence voltage/current, as well as the three phase values. It was decided to include the three phase values for voltage and current, if present, in the FOA data. If the data did not include any three-phase data, that would be recorded as an empty field.

The PMU status value was present in roughly 60 percent of the data received. If present, it was always received as a hexadecimal value. We decided to include the status value in the FOA data. If present, the status value was converted to an integer value. If not present, it was recorded as zero.

Finally, a decision needed to be made on how to deal with multiple voltage buses and current lines included with a single PMU. Discussions with PNNL power systems engineers indicated that voltage bus pairs for a single PMU tend to be mostly in sync with each other. It was concluded that there was not much to gain by including multiple voltage buses in the FOA data. The decision was made to only include a single voltage bus.

Dealing with multiple current lines was a bit more difficult. Roughly half the PMUs included multiple current lines. If multiple lines were present, there could be from two to six lines. The decision was made to only include a single current line in the FOA data, primarily to provide consistency for the FOA awardees. If multiple current lines were included, each awardee could be using different currents in their research. By providing a single current per PMU, all researchers will be using the same data.

The final FOA 1861 data schema includes these PMU data fields:

- UTC: UTC timestamp of data (String)
- VP_M: Voltage Magnitude (volts), Positive Sequence (String)
- VA_M: Voltage Magnitude (volts), A Phase (String)
- VB_M: Voltage Magnitude (volts), B Phase (String)
- VC_M: Voltage Magnitude (volts), C Phase (String)
- VP_A: Voltage Angle (degrees), Positive Sequence (String)
- VA_A: Voltage Angle (degrees), A Phase (String)
- VB_A: Voltage Angle (degrees), B Phase (String)
- VC_A: Voltage Angle (degrees), C Phase (String)
- IP_M: Current Magnitude (amps), Positive Sequence (String)
- IA_M: Current Magnitude (amps), A Phase (String)
- IB_M: Current Magnitude (amps), B Phase (String)
- IC_M: Current Magnitude (amps), C Phase (String)
- IP_A: Current Angle (degrees), Positive Sequence (String)

- IA_A: Current Angle (degrees), A Phase (String)
- IB_A: Current Angle (degrees), B Phase (String)
- IC_A: Current Angle (degrees), C Phase (String)
- F: Frequency (Hz) (String)
- DF: Rate of Change of Frequency (ROCOF) (String)
- STATUS: PMU Status Value (Integer).

2.4.2 Big Data File Type

For the amount of data that would be included in the FOA data, using CSV files would not be a good choice. Research into the most common “big data” file formats led us to Parquet, AVRO, and Optimized Row Columnar (ORC). Parquet was chosen for these reasons:

- Compression: Parquet files sizes can be up to 80 percent less than the source CSV files.
- Read Performance: Parquet files tend to be read optimized compared to AVRO and ORC. Parquet does lag in write performance, but the FOA awardees would only be reading the data.
- Columnar Storage Format: Parquet reads even more efficient by not requiring a full record read. If researchers are only working on three data fields, only those three fields can be read from the Parquet file. AVRO and ORC require the reading of all fields.

2.4.3 Data Anonymization

The TOPS were all interested in our plans to anonymize the PMU data to prevent the FOA awardees from easily identifying where the PMU data originated. Due to the nature of the North American Bulk Electric System, it was necessary to separate the PMU data into the three interconnections: ERCOT, WI, and EI. Those were labeled interconnections A, B, and C.

For every PMU in each interconnection, a random integer value from 100 to 999 was assigned, along with the interconnection letter prefix (for example, A123, B234, C345). When data from each PMU was ingested, transformed, and then written to the Parquet files, the randomized PMU identification (ID) value was assigned.

We understand that this provides a simple anonymization feature. By just looking at the PMU ID values, you would not be able to discern which group of PMUs belong to a specific TOPS. But due to the nature of electric transmission, it would not be too difficult for a researcher with basic knowledge of the electric transmission grid to determine which group of PMUs share similar data characteristics and are probably related by TOPS or geography. We asked the FOA awardees to refrain from attempting to create a topological map from the data or attempting to identify which TOPS a specific PMU belongs to.

2.4.4 FOA Data Platform

A big data platform would be required to ingest multiple terabytes/billions of rows of PMU data and transform it all into a common Parquet schema. We did some very basic cost estimates for building an internal Spark cluster in our group data center. The capital cost and administration overhead required to build and maintain a Spark cluster for this project would not have been cost-efficient.

PNNL has contracts with Amazon AWS and Microsoft Azure for their cloud environments. We demoed both environments and worked with Amazon and Microsoft reps to better understand each cloud environments' offerings. After working with both environments' tools, we decided that the Databricks application would be the best option for the FOA project. Both AWS and Azure offered Databricks, but Azure did not require a minimum usage contract. Azure Databricks could be billed strictly by usage, so we decided to host the FOA data on Microsoft Azure.

3.0 Extract-Transform-Load (ETL) Design

The ETL process has the following three distinct steps when used for the FOA project:

1. Extract (read) the data from the CSV files.
2. Transform the data to a common FOA Data schema and common field formats.
3. Load the transformed data into the final Parquet dataset.

3.1 Inspect/Create CSV Files

Before data can be extracted from the CSV files, it is necessary to understand what data is contained in each CSV file. As mentioned earlier, each TOPS provided data in a format that is unique; therefore, every TOPS CSV file has a unique extraction process. The steps below provide an overview of the steps needed to determine the data each TOPS has submitted, and which data are to be included in the FOA data.

1. Get all data received from TOPS into a CSV format that can be consumed by Databricks
    a. Data received in CSV files
        i. Inspect CSV files to determine what data is included.
        ii. Create schema to define the data contained in the CSV files.
        iii. Determine which bus/lines to include, if more than one of each are present.
    b. Data received in proprietary binary files
        i. Use provided tools to export from binary to CSV.
        ii. Inspect CSV files to determine what data is included.
        iii. Create schema to define the data contained in the CSV files.
        iv. Determine which bus/lines to include, if more than one of each are present.
    c. Data received in SQL Server backup file
        i. Restore database to local SQL Server.
        ii. Work with vendor to get a demo copy of extraction tool.
        iii. Extract data from SQL database into CSV files.
        iv. Create schema to define the data contained in the CSV files.
        v. Determine which bus/lines to include, if more than one of each are present.
2. Single-PMU CSV vs. multi-PMU CSV considerations
    a. The strategy to consume CSV files was very different depending on the number of PMUs that were stored in a single CSV file.
    b. Single-PMU CSV files are relatively easy. The CSV schema is common for all files from single TOPS.
    c. Usually less than 20 data columns in a single-PMU CSV.
    d. Single-PMU CSV files are segmented by folder structure. Top-level folder designates the unique PMU. Subfolders segment the data by year/month/day.

e. Multi-PMU CSV files are much more difficult to work with.

f. Depending on the number of PMUs provided by the TOPS, there could be over 1,000 data columns in a CSV file.

g. Building the schema requires a manual inspection of all columns. The header rows provided enough information to determine which columns of data belong to a single PMU.

h. The multi-PMU CSV file schema was not static over time. If a TOPS added/removed PMUs, the schema would change. That had to be accounted for.

3. Initial quality control (QC) checks

   a. Generally speaking, the single-PMU CSV files provided better data. They seemed to be more consistent. Multi-PMU CSV files had many more data quality issues.

   b. It may be that the TOPS providing single-PMU CSV files selected which data to provide to PNNL. They may have avoided PMUs that were not of good data quality.

   c. Multi-PMU CSV files seemed to be a selection of all data from a single TOPS with no pre-export QC.

   d. Multi-PMU CSV files required a more extensive QC check.

      i. Load a small subset of the data (all months, six non-contiguous days, six non-contiguous hours).

      ii. Run some queries comparing total records with good records per PMU. Repeat for a different set of months/days/hours. NOTE: Data was considered “good” if the values stored in each field were within expected ranges and did not have extensive missing or out-of-range values. Data was considered “bad” if values were missing, out of expected range, or had the archive systems common value for errors.

      iii. Usually, the ratio of good records to total records was comparable between data subsets.

      iv. Usually, if the PMU provided good data, it was in the 90 percent or higher ratio of good to bad data or was close to 0 percent for bad PMUs.

      v. A small number of PMUs were somewhere in between. A cutoff of 50 percent was used to determine if the PMU would be included in the dataset.

4. Multibus/Line Selection

   a. Using the description of the signals, data quality, and power systems engineer input, a single voltage bus and current line for each PMU were selected to include in the FOA data.

3.1.1 Assign Anonymized ID Values

An Excel spreadsheet was used to create the list of anonymized IDs that would be assigned to each PMU. For each set of TOPS PMUs, a text file had been created that maps the anonymized ID to the TOPS PMU identity. The map files were used during the ETL runs.

3.2 ETL Execution

Azure Databricks clusters, notebooks, and jobs were created to execute the ETL process for the various TOPS PMU data. During the development phase, Databricks notebooks were run interactively with small subsets of the data to test and refine the process. Once the process was complete, jobs were created to run the ETL process on large clusters and reading in the entire raw dataset.

The two types of CSV files received, single-PMU and multi-PMU, required two distinct processes during the ETL execution. And within those two distinct processes, each ETL process needed to be modified based on the data received from each TOPS.

3.2.1 Single-PMU ETL Process

All single-PMU CSV files are segmented by folder structure. For example, TOPS-1 provided data for 10 PMUs. That data is stored in folder TOPS-1. The data for each PMU is stored in a unique folder under TOPS-1: TOPS-1\PMU1, TOPS-1\PMU2, etc. Under each PMU folder is where the CSV files would reside. Depending on how the data was delivered, there could be additional folder layers based on year/month/day. And there was variation on the time span contained in a single file. It varied from one full day per CSV file down to one minute per CSV file.

3.2.1.1 Extract

- The Databricks notebook creates a connection to the specific TOPS raw CSV storage folder.
- A TOPS-specific CSV schema was created to match up the data delivered with the corresponding data fields in FOA common schema. In many cases, the TOPS CSV files did not include data for all the FOA common schema fields, so those fields were skipped.
- The CSV files are read and data extracted into a Spark dataframe.

3.2.1.2 Transform

- The transformation process knows which activities to perform on the dataframe based on the name of the fields assigned in the schema.
- Filter the dataframe to exclude any data with known bad data. For example, one TOPS CSV export used a specific string value pattern for the UTC timestamp value indicating a that a particular row of data is “bad.” All data containing that string pattern in the UTC field is filtered out.
- Modify the UTC timestamp format, if necessary. Not all UTC timestamps were delivered in common schema format. Some needed to be modified to common YYYY-MM-DD HH:MM:SS.sss format.
- Create meta data fields to aid in partitioning and searching. Five additional fields were created based on the UTC timestamp value, theYear, theMonth, theDay, theHour, and theMinute.
- Assign the interconnection value (A, B, or C).

- Create the PMU ID field based on the folder structure of the CSV file that was read into the dataframe. The folder structure names were mapped to the randomly assigned PMU ID values.
- Filter the dataframe based on “theYear” value. Anything other than 2016 or 2017 is dropped.
- Drop any the temporary fields created to reconstruct the UTC format.
- Each TOPS archive used a set of specific string/numeric values to indicate “bad” data. Search all the PMU data fields in the dataframe and replace all instances of those bad data markers with an empty field.
- Convert kilovolts to volts.

3.2.1.3 Load

- Once the transformation is complete, begin the load into the dataset.
- The load process creates a folder hierarchy of Year/Month/Day based on the UTC timestamp and creates the individual Parquet files that contain the final data.

3.3 Multi-PMU ETL Process

The multi-PMU CSV files were much more difficult to work with. The primary reason was because these types of CSV files had hundreds to over 1,000 fields. A manual inspection of each TOPS CSV format was needed to determine the total number of individual PMUs contained in the file and the fields associated with each PMU. Most of the time, the number of fields and field order remained constant for all TOPS CSV files, but there were a few instances where the number and order of the CSV fields changed over time due to the TOPS adding or retiring PMUs in their transmission network. The general format was field 1 (UTC timestamp value), fields 2–10 (PMU1), 11–20 (PMU2), etc. The related PMU data fields were usually contiguous. The PMU/field location had to be mapped out to ensure that the expected data was stored in the final PMU dataset.

3.3.1.1 Extract

- The Databricks notebook creates a connection to the specific TOPS raw CSV storage folder.
- A TOPS-specific CSV schema was created identifying the individual PMU’s data fields and then matching up each PMU’s data with the corresponding data field in FOA common schema. In many cases, the TOPS CSV files did not include data for all the FOA common schema fields, so those fields were skipped.
- The CSV files are read and data extracted into a Spark dataframe.

3.3.1.2 Transform

- The transformation process ran in loop, working on a single PMU data at a time. This essentially converted the multi-PMU to a single-PMU for the transformation process.
- The transformation process knows which activities to perform on the dataframe based on the name of the fields assigned in the schema and the PMU loop sequence.
- For each PMU in the loop, the transformation steps were the same as the single-PMU process.

3.3.1.3 Load

- Once the transformation is complete, begin the load into the dataset.
- The load process creates a folder hierarchy of Year/Month/Day based on the UTC timestamp and creates the individual Parquet files that contain the final data.

3.3.1.4 Repeat

- Repeat the Transformation and Load steps for each PMU in the file.

4.0 Data Partition for Training and Testing Datasets

Furthermore, the data not only had to be anonymized but also had to be split so that a portion of it could be used for a training dataset (to train the machine learning/artificial intelligence algorithms) and a test dataset (to verify that the training algorithms work correctly).

Partitioning the data into a training dataset and a testing dataset allows the researcher an opportunity to create supervised, learning-based models using the training data and determine the effectiveness of their predictions on the testing data. This helps address the model overfitting issue that will occur if partitioning is not used. Overfitting occurs when a model becomes so detailed and dependent on the training dataset that it models the random noise of the training dataset more so than the real effects within the data system. This will negatively impact the model performance when applied to new data.

Cross validation methods are a common way to combat this issue, while assessing the effectiveness of a model. There are many ways to apply cross validation, but the step they have in common is dividing the dataset into training and testing (sometimes called validation or held out) datasets. The proportions of these splits tend to range between 60/40 (60 percent training, 40 percent testing) and 90/10 (90 percent training, 10 percent testing). There needs to be enough data and events in the training set so that the testing can be effective.

For this DOE FOA, it was decided to provide researchers a guide for how the data will be partitioned. This will allow for a consistent partitioning of the data across the many research groups, which will allow for more consistent comparisons of the methods used.

It was decided that the data be split 75/25 to create the training and testing datasets, respectively. This would amount to approximately 78 weeks of training data and 26 weeks of testing data. Mathematical models tend to do better when predicting or identifying events that are hours or days out, instead of months out. For this reason, a simple split of the first 78 weeks training and last 26 weeks testing is not ideal. Also, splitting it this way, with a contiguous six-month period, would not allow for testing across data from all four seasons. A better way to split it would be to divide the data into 13 consecutive time periods of eight weeks each. Within each eight-week period, use the first six weeks as training data and the next two weeks as testing data. Doing these 13 different times across the two-year period will allow for testing across all four seasons. Previous PNNL research with Eastern Interconnect PMU data has shown that baselines of six to eight weeks are more stable and found events better than baselines built on one to four weeks.

5.0 Event Logs

The machine learning methods used by several of the awardees to analyze the synchrophasor measurements require training. In the case of supervised machine learning, the training data needs to be labeled. To effectively train these methods for detecting and categorizing power system disturbances, the awardees required a set of labeled disturbances. These labels were provided to PNNL by the synchrophasor data owners in the form of event logs.

Event logs are maintained by TOPS and RCs to record significant disturbances to the bulk power system occurring within their footprint. Among other uses, the logs can support compliance with NERC's Transmission Availability Data System (TADS). [2] The data is confidential and required anonymization before being delivered to awardees. This section describes PNNL's process of combining the event logs into a single anonymized list.

The syntax used in the anonymized event logs was common across the three interconnections. This required consolidating information from several different data contributors that varied in format, syntax, and level of detail. Where data contributors provided descriptions of events, staff at PNNL manually translated the descriptions to a set of standard labels. Where data contributors used codes, the codes were mapped to the standard labels and automatically translated.

As much as possible, staff at PNNL avoided making inferences about the cause of events. For example, the *weather* label was not recorded in the consolidated logs unless weather was specifically listed by the data contributor, even if several surrounding events were caused by weather. However, where written descriptions made it clear that separate events were linked, a common set of labels were used in the consolidated event logs because award recipients would not have enough information to make this connection without the original logs.

In the syntax developed by PNNL, events are described by three levels of information: category, cause, and descriptors. The event *category* specifies either a) the disrupted equipment (generator, line, transformer, bus, capacitor) or b) the type of event (frequency deviation, oscillation). The event *cause* provides a general description of what caused the disturbance. Further details are provided as *descriptors*.

The event logs were provided to awardees in CSV files with columns for the category, cause, and descriptor labels. Where multiple labels were relevant, they were separated by a comma. For example, a line outage caused by a tree that caused a phase A to ground fault, then a three-phase fault, and finally downed the conductor would be described as shown in Table 2. Additional details are provided in the following subsections.

Table 2. Event Log Example.

Category	Cause	Descriptor
Line	Tree, Equipment	A-G, 3P, Conductor

[2] https://www.nerc.com/pa/RAPA/tads/Pages/default.aspx

5.1 Generator, Line, Transformer, Bus, Capacitor, and Unspecified Categories

The Generator, Line, Transformer, Bus, and Capacitor labels each specify a piece of equipment impacted by the event. In addition, the Unspecified label is used when the impacted equipment was not noted in the provided event log. These categories share a common set of causes and descriptors, which are detailed in Tables 3 through 6, below.

Table 3. List of Cause labels associated with the Generator, Line, Transformer, Bus, Capacitor, and Unspecified Categories.

Cause Label	Description
Trip	This label is used when little information about the event was available. In many cases, the trip was probably related to a fault, but equipment may trip off for a variety of reasons, so this cannot be known for sure.
Fault	This label is more specific than Trip because a fault was known to occur, but the exact reason for the fault was not listed.
Planned Service	Includes maintenance, construction, work on communication networks, etc. Disruptions may be intentional to allow the service to take place but may also occur due to problems during the service. In the latter case, the Human Error descriptor is common.
Planned Operations	Intentional changes that occur as part of power system operation. These changes are not necessarily planned far in advance and may be made in response to urgent requirements.
Planned Testing	Disruptions may be intentional to allow the testing or unintentional due to problems during testing. In the latter case, the Human Error descriptor is common.
Equipment	Indicates the disturbance was caused by mis-operating or damaged equipment.
Weather	A generic label when the specific weather phenomenon was not listed.
Not Lightning	Some event logs do not specify the weather phenomenon that caused the disturbance but do indicate if it was not lightning. This label is used in these circumstances.
Lightning	Indicates a fault or equipment damage due to lightning.
Fire/Smoke	Indicates a fault or equipment damage due to fires.
Ice	Indicates a fault or equipment damage due to icing.
Wind	Indicates a fault or equipment damage due to wind.
Tree	Indicates a fault or equipment damage due to trees contacting or falling on equipment, typically transmission lines.
Earthquake	Indicates a fault or equipment damage due to an earthquake.
Contamination	Indicates a fault due to buildup of bird droppings, dust, corrosion, etc.
Animal	Indicates that contact by an animal (birds, squirrels, and snakes are common) caused the fault. The equipment the animal contacted may differ from the equipment listed in the Category column. For example, if a line outage was caused by a snake on a substation transformer, the category would be Line, the cause would be Animal, and the descriptor would be Transformer.
Galloping	A fault created by high-amplitude, low-frequency oscillation of transmission line conductors due to wind.

Cause Label	Description
Proximity	Indicates that something came close enough to equipment to cause a fault and/or damage (e.g., construction equipment touching an overhead transmission line). Impacted equipment is specified using the equipment descriptors.
Remedial Action Scheme	According to NERC, a Remedial Action Scheme is, "… a scheme designed to detect predetermined System conditions and automatically take corrective actions that may include, but are not limited to, curtailing or tripping generation or other sources, curtailing or tripping load, or reconfiguring a System(s)."

Table 4. List of Descriptor labels associated with the Generator, Line, Transformer, Bus, Capacitor, and Unspecified Categories.

Descriptor Label	Description
Fault Type	Combination of P (generic phase), A (A-phase), B (B-phase), C (C-phase), and G (ground) indicating the type of fault. For example, the entry A-B indicates a phase-to-phase fault between phases A and B. The label P is used when the specific phase was not listed. Three-phase to ground faults are indicated as 3P. Faults on DC lines are indicated as DC.
Human Error	Indicates that a mistake led to the event. This label is limited to errors that occurred at the time of the event, such as mistakenly sending a trip signal to a circuit breaker. Errors made far in advance of the event, such as improper protection settings, are not included.
Power System Condition	Indicates that the outage occurred due to conditions such as instability, abnormal voltage, abnormal frequency, or unique system configurations.
Sympathetic	Indicates that the listed disruption was related to a disturbance to a different piece of equipment. For example, the event log may indicate that the specified transmission line tripped due to a fault on a different line connected to the same substation. Typically, sympathetic tripping can be considered unnecessary and undesirable. The exception is when this descriptor is used with the Planned Operations cause to indicate that a change in service to one piece of equipment required an outage on another piece of equipment.

Table 5. List of Descriptor labels associated only with the Generator Category.

Descriptor Label	Description
Plant	Used exclusively in conjunction with the Equipment cause. Indicates that the mis-operating or damaged equipment that caused the generator trip was located within the generating facility.
Wind Farm	The generation source that tripped was a wind farm. The absence of this descriptor does not indicate that the generation source was not a wind farm.
Dropped MW	Amount of generation that tripped in terms of mega-Watts (MW). Example entry: 1400 MW.

Table 6. List of Descriptor labels used with the Equipment, Animal, and Proximity Causes to indicate the type of equipment that mis-operated, was damaged, or was contacted.

Descriptor Label	Description
Protection	Relays, circuit breakers, etc. Protection equipment is involved in many events, but it was only listed when mis-operation or damage occurred. Note that protection equipment typically mis-operates following a fault, so the Fault cause is not included when this descriptor is used.
Lightning Arrester	A specific piece of equipment used to protect substation equipment from lightning.
Transformer	When used as a descriptor, indicates that a transformer was the cause of the problem. Note that the event category may not be Transformer.
Conductor	Indicates the event was caused by damage to the transmission line's conductor. Note that the event category may not be Line.
Structure	Indicates damage to transmission towers, insulators, jumpers, etc.
Substation	Indicates damage to bus work, switches, insulators, etc. Excludes protection equipment and transformers.
Phase	Indicates the phase that was involved in the equipment damage. Possible entries are A, B, and C.

5.2 Frequency Category

This category includes events that resulted in significant deviation of the system's synchronous frequency from the nominal 60 Hz. These events are typically due to sudden loss of generation but can also result from other significant disturbances. This category does not contain any causes or descriptors.

5.3 Oscillation Category

Oscillations observed by the data provider, whether they originated inside or outside of their territory, are included in this category. The *Causes* used for this category are listed in Table 7, below. The sole descriptor for the oscillation category specified the frequency of the oscillation in Hz.

Table 7. List of Cause labels associated with the Oscillation Category.

Cause Label	Description
Generator	Indicates that a generator was the source of the oscillation.
Wind Farm	Indicates that a wind farm was the source of the oscillation.
Fault	Indicates that a fault initiated the oscillation.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov