PNNL-27857

Pacific Northwest
NATIONAL LABORATORY

PNNL Plan for Acquiring, Anonymizing, and Protecting Utility Data

August 2018

ES Andersen
BG Amidan
JS Banning
A Silverstein

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
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-27857

PNNL Plan for Acquiring, Anonymizing, and Protecting Utility Data

August 2018

ES Andersen
BG Amidan
JS Banning
A Silverstein

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

Pacific Northwest National Laboratory
Richland, Washington 99352

Abstract

Objective

This plan details the processes that PNNL will use for acquiring, anonymizing and protecting utility data offered for use in research projects affiliated with the U.S. Department of Energy's Office of Electric Reliability and Resilience (OE). PNNL has been tasked with obtaining agreements with utilities to provide PMU and related data for research purposes, and to assemble anonymized practice and test datasets for the three primary U.S. electric grid interconnections. DOE intends to provide anonymized PMU data to selected researchers to explore the capability of artificial intelligence tools to identify and improve existing knowledge, and to discover new insights and tools for better grid operation and management.

Introduction

ACQUIRING DATA – Utility to PNNL

Data to be requested
The DOE research projects will require large datasets. We are presently targeting collaborative data partnerships with eight to ten utilities and RTOs spanning all three interconnections.

We seek these data 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 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.
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).

Data Provision Agreements

PNNL seeks agreement from the data-providing utilities that PNNL's anonymization process and protection schemes will adequately protect the utility from any release of sensitive information that could potentially damage or increase power system operational risks. That said, once the anonymization is complete and protection schemes are in place, PNNL and DOE will move forward under the expectation that the anonymized datasets are inherently safe to be shared with the researchers and their research audience and there will be no need for extensive data protections or restrictions on use of the anonymized datasets.

Data Formatting

Each PMU owner collects its PMU data based on its specific needs for safe, effective grid operation. Standards like C37.118 are intentionally customizable to allow utilities the freedom to define their data streams to meet their needs. If the raw PMU data is provided in inconsistent formats, PNNL will convert the raw data into a common format and data structure to enable combination of the contributed datasets into interconnection-specific datasets.

Data Anonymization

Data anonymization is the process of removing identifying information to reduce the risk that the data will expose security-sensitive information (and the risk that the anonymized data can be re-identified. To anonymize the contributed PMU data, PNNL intends to anonymize all PMU, SCADA and event log site names and locations across each dataset using general naming conventions and randomization. For example, if a data stream has the name "FarRidge_NorthBus_Voltage", it would be changed to "PMU068_5_Voltage". (The number "068" was randomly selected under the assumption there are 200 different PMUs numbered PMU001 to PMU200, and the number "5" was randomly selected under the assumption there are less than 9 different phasor locations). PNNL intends to keep intact both the time-stamps and the type of variable (i.e. frequency, voltage, current, etc.) for each record in the database (because removing those metadata would destroy the correlation value lying within the PMU data).

PNNL wants to work with the data owners on how to handle high-profile grid events that may have received significant press or industry attention, because these events could reveal information about the location where the event occurred. We hope to retain data on high-profile events to facilitate identification of precursor events, but we welcome suggestions on whether and how to modify the data.

Non-Disclosure Agreements

If needed, PNNL will establish non-disclosure agreements (NDAs) to obtain the data from the utilities in order to protect the raw PMU and associated data as PNNL implements the data anonymization process.

We anticipate that the NDAs will only be needed for transmittal and handling of the raw PMU data provided by the utilities to PNNL; additional NDAs covering down-stream data users should not be needed because the original raw data will not be shared with anyone outside of PNNL.

If utility data owners have concerns about the use and security vulnerability of anonymized datasets, PNNL and DOE seek to understand and address these concerns as early in the process as possible.

Data Hosting in PNNL's Electricity Infrastructure Operations Center (EIOC)

Like all governmental agencies, DOE must be and is compliant with the Federal Information Security Management Act (FISMA), and requires such compliance from all DOE labs and contractors.

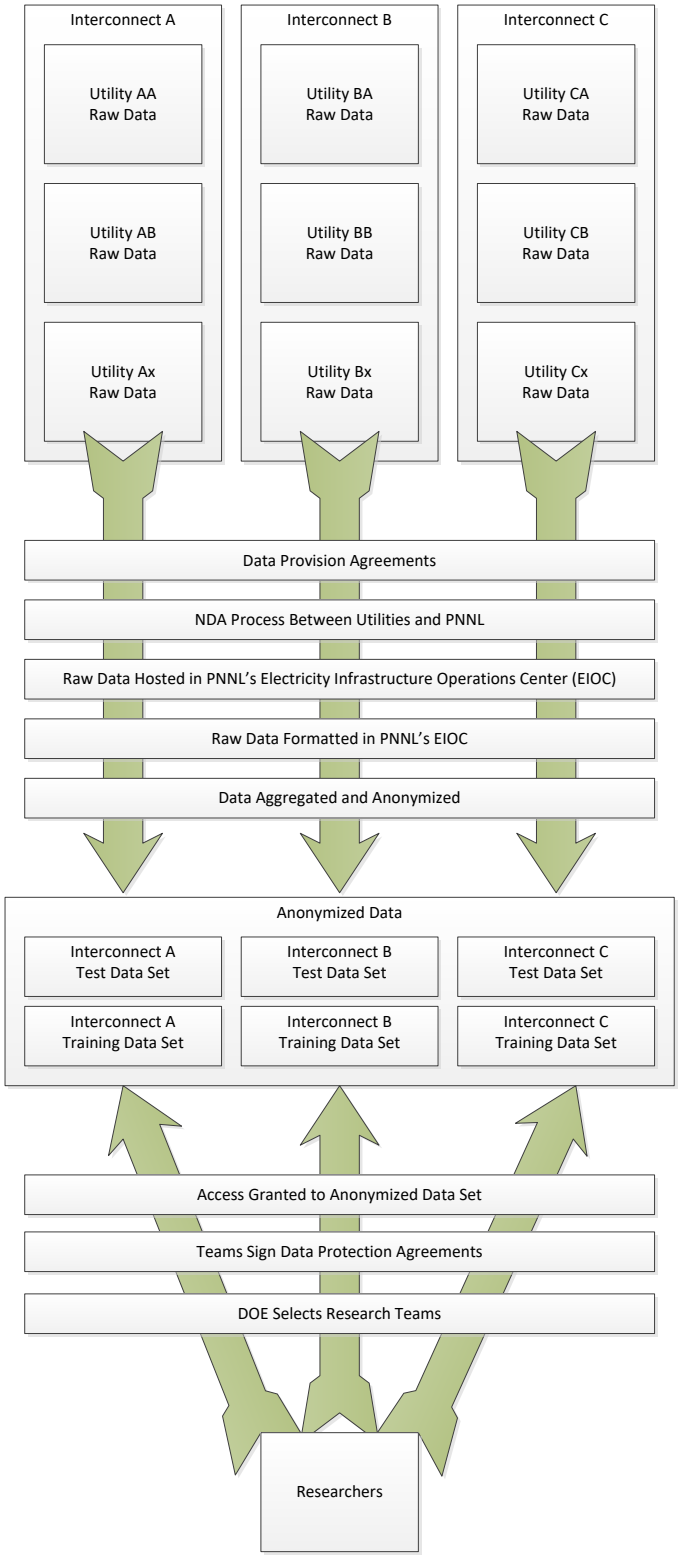
PNNL maintains a baseline security profile for any system managed or owned by PNNL, which is identified in the Site Security Plan (SSP) and approved by DOE as part of the Authorization To Operate (ATO) package. PNNL's minimum level of protection assumes OUO/FOUO/PII data is processed and or stored on the system. PNNL has implemented extensive system-wide tools for staff and system administrators to ensure security is part of their system's lifecycle. For example, a Public Key Infrastructure (PKI) joined with DOE is embedded in desktop products, continuous monitoring of all systems, regular proactive patching, malware signature updates and mandatory annual operational security (OpSec) and security training for all staff and collaborators.

PNNL is looking into whether to host the data collected for this project within PNNL's EIOC[1] or using secure cloud-based storage.

After completion of the data anonymization process and issuance of the datasets to the researchers, PNNL staff will destroy the raw utility data according to the requirements stipulated in the NDA between PNNL and the data provide

[1] Access to the PNNL EIOC server farm is restricted via a Cisco firewall utilizing RSA tokens for two-factor authentication (2FA). All datasets (original data and anonymized datasets) will be hosted in isolated subnets, where all user connections are restricted to the specific subnet with the data that they have permissions to access. Firewall logs and Netflow data will be actively monitored to ensure excessive data exfiltration is not occurring. Furthermore, we have the ability to encrypt data on disk to prevent local data exfiltration, as well as monitoring of data file integrit

Appendix A
Data Management and Anonymization Process Summary



Appendix B
Data Anonymization Process

Data anonymization and masking – To lessen the likelihood that the collected PMU datasets can be used to identify power system and asset vulnerabilities, PNNL staff will use sophisticated data anonymization and masking techniques to mask the identity of the contributing utilities and the name and location of the PMUs and substations or assets they cover.

1) Data will be aggregated by interconnection. Within each interconnection, data contributors and their footprints will be identified by generic names (Blue, Orange, X).
2) Actual PMU names will be changed to generic, random names.
3) The type of data variable (i.e. frequency, voltage) will be apparent in the data.
4) PNNL staff will work with data contributors to identify data protection concerns and recommend additional data anonymization and masking techniques to meet those concerns.
5) PNNL staff will apply the appropriate measures consistently to all datasets collected before providing the collected data to researchers.
6) PNNL will use a consistent method across all three interconnection datasets to divide each dataset between a training dataset (distributed to the researchers in the first phase of the work for initial analysis and correlation development) and the test dataset (distributed to the researchers following submittal of the first analytical report, to be used for analysis in the second research phase).

Appendix C
Data Handling Requirements

Data handling requirements

(1) PNNL will collect the source data from all contributors and process it for anonymization as discussed above.
(2) PNNL staff working on this project have long histories with the National Lab system, have security screenings, and have performed similar tasks on other projects with secure results.
(3) PNNL's data storage systems meet federal FISMA security standards. PNNL and its staff use industry standard practices and protocols to monitor and protect data storage and manage assets.
(4) Restrictions on researchers -- DOE plans to place the following qualifications and restrictions upon the researchers and firms that are allowed to access and work with these PMU datasets.
    a) Demonstrated commercial expertise in big data analysis (machine learning, etc.) in energy or other industry verticals. However, no prior work with PMU data or utility clients is required to qualify for this task.
    b) Commercial (non-academic) researchers should NOT team with an electric industry partner because: having inside information can both expose real data sources and info; grid event expertise will change the analytical results and keep us from being able to tell how good the basic machine learning tools are w/o inside information; and to ease entry of new players and tools into this field.

Pacific Northwest
NATIONAL LABORATORY

www.pnnl.gov

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)

U.S. DEPARTMENT OF
ENERGY