Pacific Northwest
NATIONAL LABORATORY

PNNL-SA-161370

Data for FOA 1861 Bringing It All Together

April 13, 2021

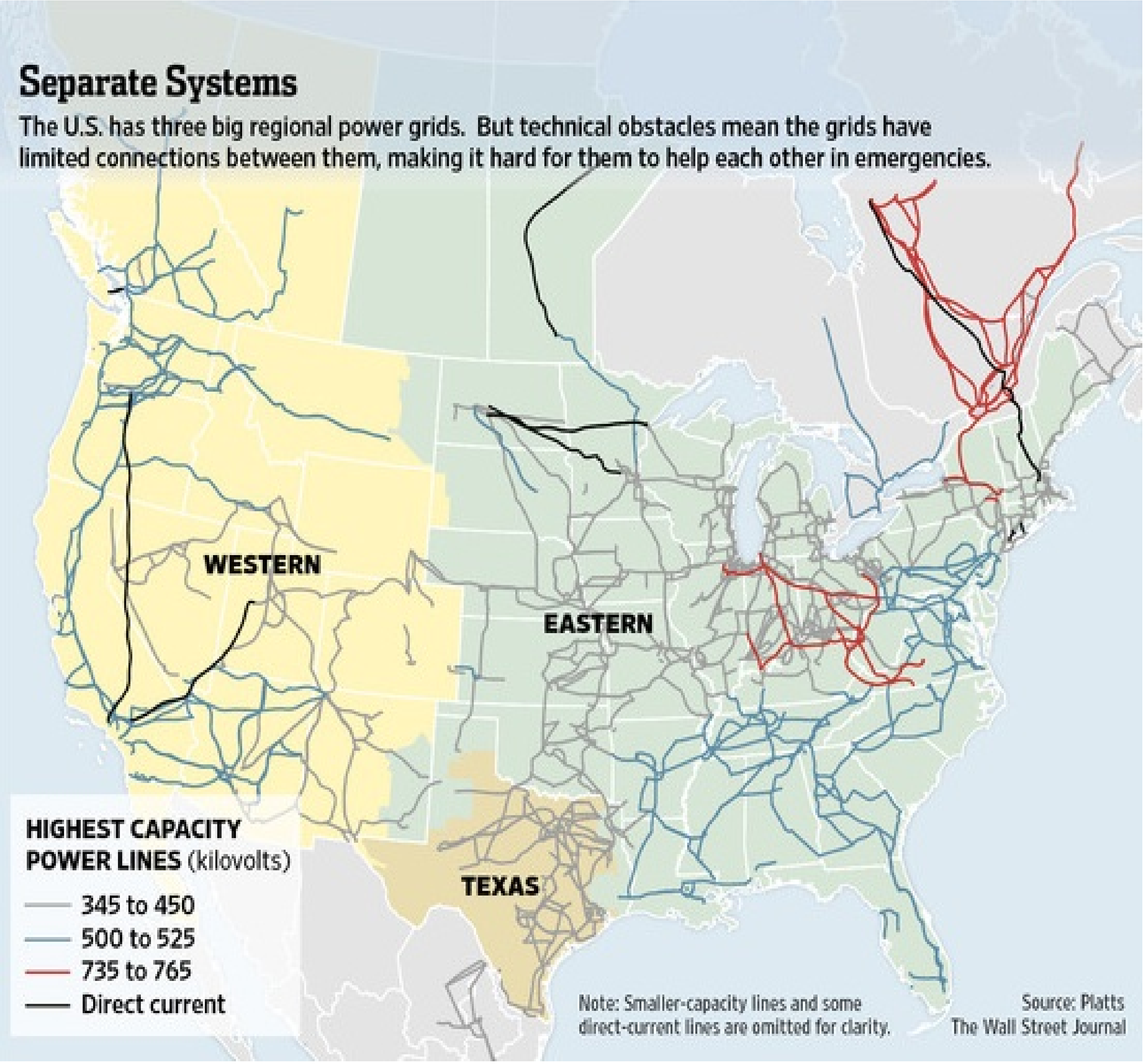
Jeffery Banning
Eric Andersen
Jim Follum

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

The Need for Data

- Dept of Energy FOA (Funding Opportunity Announcement)1861
- Derive additional value from the vast amounts of sensor data already generated
- Real world data from each of the three US interconnections



Separate Systems

The U.S. has three big regional power grids. But technical obstacles mean the grids have limited connections between them, making it hard for them to help each other in emergencies.

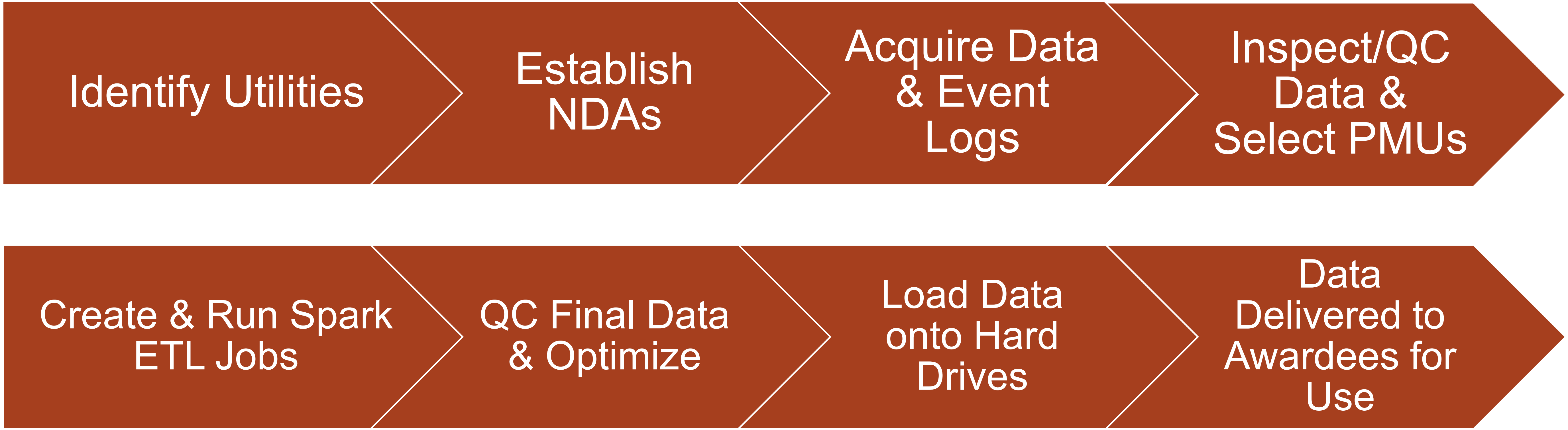
Note: Smaller-capacity lines and some direct-current lines are omitted for clarity.

Source: Platts
The Wall Street Journal

Source identified on image, used without permission.

Pacific Northwest NATIONAL LABORATORY

The High-Level Process



Identify Utilities

Establish NDAs

Acquire Data & Event Logs

Inspect/QC Data & Select PMUs

Create & Run Spark ETL Jobs

QC Final Data & Optimize

Load Data onto Hard Drives

Data Delivered to Awardees for Use

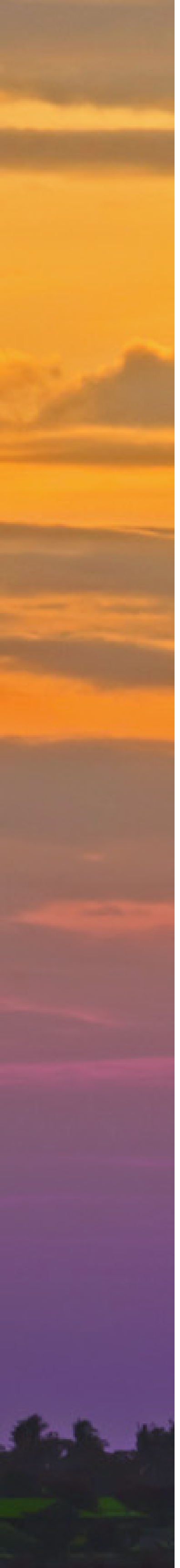
Pacific Northwest
NATIONAL LABORATORY

Obtaining the Data

- Near real-time PMU may fall under Critical Infrastructure Protection (CIP)
- Some utilities were hesitant to contribute data due to Critical Infrastructure Information (CII) and CIP concerns
- Obtaining older data and receiving the data under an NDA helped alleviate concerns
- PNNL anonymized the data

Anonymized Data Set: How It Started

- Every contributor's data was different
- Multiple file types.   All had to be converted to CSV
- Archived frame rates, 30/sec and 60/sec
- Positive sequence, ABC phases, status values, 1 or 2 voltage measurements, 1 to 6 current measurements
- Final schema: Single voltage measurement (Pos, ABC), single current measurement (Pos, ABC), F/dF, Status
- Single PMU/file; multiple PMU/file
- Each PMU assigned random ID value

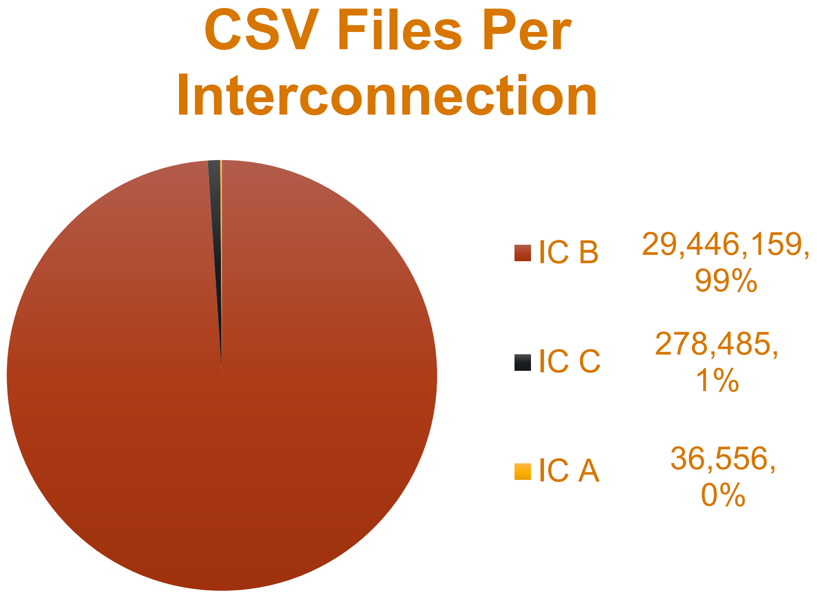


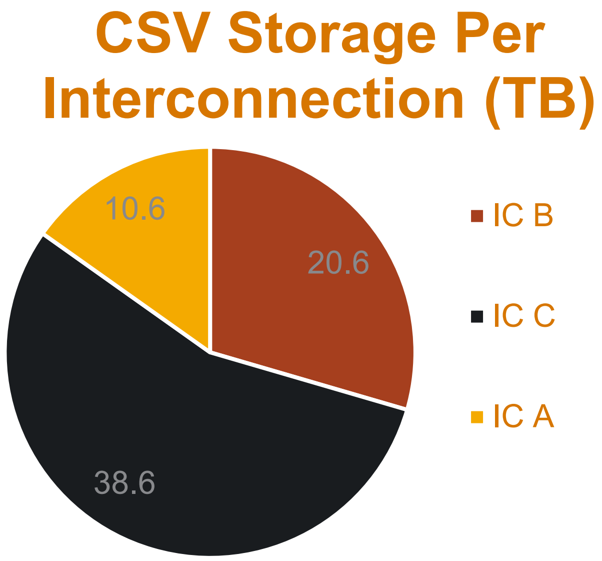
Pacific Northwest
NATIONAL LABORATORY

Source Data Snapshot

	IC C	IC B	IC A	Total
Received Files	278,485	29,446,159	2	29,724,646
PMUs	250	43	221	514
CSV Files	334,315	29,446,159	36,556	29,817,030
CSV Storage	38.6 TB	20.6 TB	10.6 TB	69.8 TB

- Most data covers 2016/2017
- One contributor only had 2018/2019
- Lots of small files in IC B



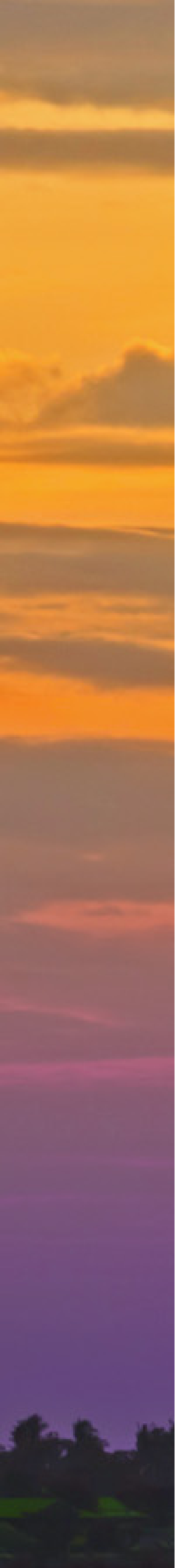


Event logs

- Utilities provided event logs to supplement their PMU data: over 9000 entries
- Several benefits for research teams
  - Indicated events of interest for utilities
  - Supported development of event detection and classification algorithms
  - Provided a means for training supervised learning methods
- Several challenges
  - Anonymization did not allow for detailed event descriptions
  - Syntax varied among data contributors
- Conversion to common syntax
  - Automated conversion of keywords
  - Manual conversion of long-form event descriptions
  - Event descriptions included up to three levels of detail

Anonymized Data Set: How We Got There

- Used Apache Spark for ETL (Extract, Transform, Load)
  - Extracted data from CSV files
  - Transformed data to common schema and field patterns
    - ✓ Map source fields to correct common schema fields
    - ✓ Modify UTC timestamp format
    - ✓ Filter known bad data
    - ✓ Convert to volts, if necessary
    - ✓ Assign anonymized ID value
  - Loaded data into Parquet files
- Dataset Partition
  - Training: Year/Month/Day
  - Test: Added Year/Month/Day/ID option



Pacific Northwest
NATIONAL LABORATORY

Anonymized Data Set: Training and Test Data Sets

- Created two distinct datasets: Training and Test
- Split in repeating 6-week / 2-week pattern for duration of data provided
- Attempting to ensure all FOA awardees are working with the same data during the training and testing phases of their research

Dataset	Total Size	Total Records
Training	20.4 TB	495.6 Billion
Test	7.1 TB	168.3 Billion
Total	27.5 TB	663.9 Billion

Conclusion
What’s Next – Some things to think about

- Utilities should start thinking about how PMU data is archived, and how to make it more accessible for research purposes
- Event logs are as critically important to researchers as the data itself, and the event logs need to be detailed, accurate, and use a common taxonomy between utilities
- Utilities need to communicate to researchers what is important in the data
- Researchers need to communicate back to data providers the kinds of detail they need in the event logs for training ML/AI algorithms

Pacific Northwest
NATIONAL LABORATORY

Thank you