Building a National Infrastructure for Artificial Intelligence on the Grid

By Mohini Bariya (Berkeley), Theo Laughner (Lifescale Analytics) and Sean Murphy (PingThings)

Introduction

Utilities are unnecessarily limited by their ability (1) to explore and quickly test hypotheses about data and then (2) to move analytic use cases from prototypes to production in operational systems. This arises because utilities are slow to traverse the *analytics pipeline*. The analytics pipeline not only supports the implementation of applications for previously identified use cases, but also enables the discovery of new use cases through exploration of the available measurement data. Utilities are tied to legacy platforms, ill equipped for analytics or the traversal of the analytics pipeline, and need the right tools to process and analyze time series sensor data at scale.

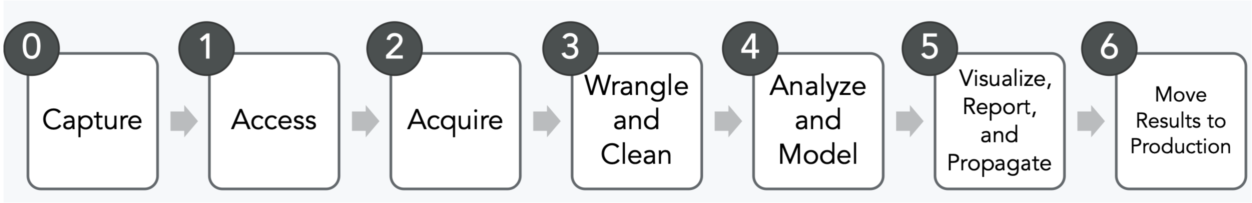


Figure 1 – Analytics Pipeline

The analytics pipeline (Figure 1) captures a process to realize value from data by the development of analytic use cases. Each element of the pipeline is hampered by traditional utility data systems. For example, even if the data is captured (Step 0) from the device, access (Step 1) may be limited to very few people or measurements due to storage space or limited network bandwidth. Even data transferred and stored must be verified (Step 3) before analysis is performed, because bad data will yield bad information. Once the data is cleaned, it can finally be used for exploratory research and analysis (Step 4). If the results are conclusive, then reporting and visualization take place manually (Step 5) and often with desktop-based tools like MATLAB or Excel. Taking these prototype results and migrating them back to production is nearly impossible (Step 6). In some cases, production rollout of an analytic either never occurs or requires significant development effort.

Next generation sensors are being deployed across the grid at an increasing rate and sensors already embedded in smart assets lay dormant and only require software activation. Still, utility data systems are unable to handle the deluge of data. Moreover, even if such legacy platforms could handle the quantities of data, there are few systems that can make the data available for advanced analytical techniques like machine learning and artificial intelligence. Without known use cases and applications with previously identified return on investment (ROI), utilities are unlikely to adopt modern computational paradigms even though the most capital-intensive part of data acquisition—sensor deployment—has already been completed. Thus, the industry faces the classic chicken and egg scenario; which came first, the data platform or the application?

To encourage the utility industry to adopt these techniques, the Department of Energy's Advanced Research Projects Agency (ARPA-E) funded the National Infrastructure for Artificial Intelligence on the Grid. The NI4AI has a 3-fold mission: 1) provide open access to a platform architected to make working with time series data at scale easy, 2) collect and host a variety of open sensor data sets to support research, education, and industry applications for sensor data, and finally 3) to build a community around this ecosystem.

The Problem

Surveying the utility landscape, we see an ever-present chasm in the grid analytics space. This gap prevents ideas and hypotheses about how data from grid sensors can create value from being fully developed as prototype use cases and then deployed into production. Currently, at best, these ideas get implemented in MATLAB and run on a laptop, with results potentially presented at conferences or published in journals. However, even successful prototypes that enhance the grid's stability, resiliency, and/or reliability can take years to be operationally deployed if they ever make it to the production environment.

The slow pace of analytic use case development has allowed high-value data to languish unused at utilities and even to be deleted due to storage space and dated software cost models. This limited use of sensor data inhibits the utility's ability to push the boundaries of their technical capabilities and gain greater insight into the real-time operation of their networks. For example, phasor measurement capabilities are built into many smart assets such as modern smart relays, consequently, the number of PMUs connected to the grid is already in the hundreds of thousands [i]. Many transmission utilities have dozens, hundreds, or even thousands of PMUs deployed but inactive, lacking the capabilities to leverage the vast data volumes that these sensors generate when operational.

Consider the measurements from a single PMU which could be used to detect voltage sags on a system. Synchrophasor data is perfect for advanced analytics because its intrinsically high sample rate--30 Hz or greater--allows observation of short-duration, sub-second events that span only a few cycles. However, scanning a year of just a single voltage phasor magnitude signal at 120 Hz requires the processing of 3,784,320,000 data points or approximately 60GB of time series data. Most analytic tools in the utilities' toolbox are ill-equipped to handle this quantity of data.

Approach

It is possible to develop and deploy an analytic for grid sensor data in days rather than months or years by walking through the steps taken to achieve these results. The key to such speed is the use of an open, state-of-the-art platform to ingest, store, clean, visualize, and process time series data from such grid sensors such as PMUs, digital fault recorders, point on wave sensors, smart meters, and power quality meters. This universal sensor analytics platform was designed with a deep understanding of the analytics pipeline to allow utilities and other organizations to traverse it at warp speed. The platform allows authorized utility users to easily access data and enables artificial intelligence and analytics as core components of the platform (instead of bolted on additions), making available best of breed open source data visualization, analysis, and machine learning software libraries. The paper, *A Universal Platform for Utility Sensor Data Analytics and Artificial Intelligence*, details this universal sensor analytics platform, describing the underlying technologies and innovations that enable such capabilities [ii].

Even with a platform in place, the process of exploring, prototyping, deploying, and operationalizing new analytics for utility data is slow. In the interest of catalyzing this process, ARPA-E funded a multi-year project called the National Infrastructure for Artificial Intelligence on the Grid (NI4AI). The project is designed to eliminate barriers to accessing and analyzing grid data by providing widespread access to real-world sensor data and to state-of-the-art data analysis tools. As shown in figure 2, below, the project has a 3-fold mission: provide open access to a cloud-based platform, host a variety of open data sets that can meet the needs of researchers, educators, and industry, and build a community to facilitate collaborations across institutions and among researchers, students, and practitioners using the data and platform to explore new opportunities for extracting value from grid sensor data.

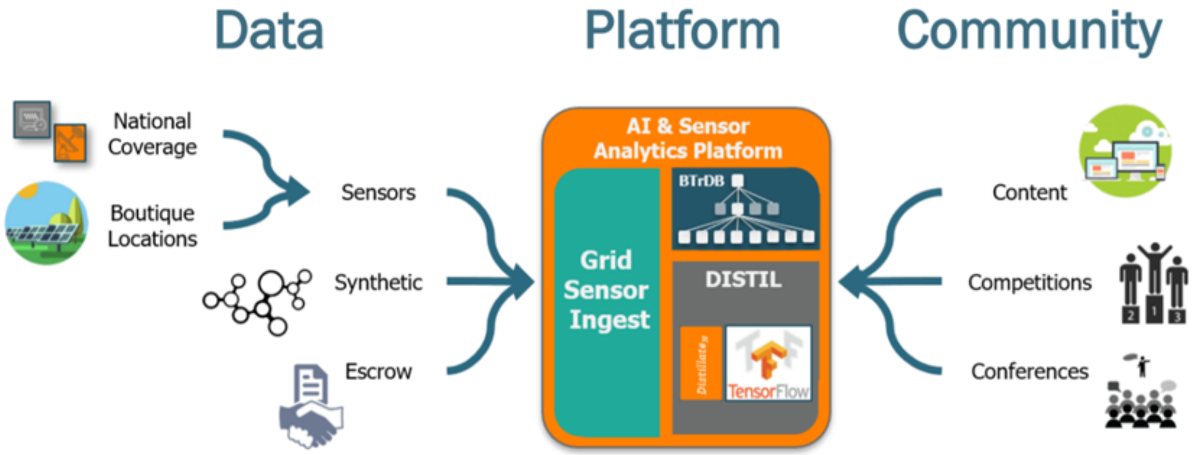


Figure 2 – NI4AI Mission

In phase one, the NI4AI project provided open access to a high-performance platform to host sensor data of all kinds. This type of platform is critical, because as new sensors are deployed on the utility infrastructure, it will be necessary to ingest data from a variety of sensor networks. The ability to synthesize data across multiple and disparate sensor networks is essential to making better decisions. The open platform enables collaboration across institutions and reduce the time required for data contributors or commercial partners to request, vet, and deploy new analytics from users.

In phase two, the NI4AI project is deploying sensors to generate data that is openly accessible through the platform. This data set captures a variety of different grid dynamics, including continuous wide-area monitoring data (like PMUs), and boutique data sets targeting local dynamics of interest such as Power Quality or Protection. The sensors will stream data into the platform, providing open access to data which users (e.g., researchers, students, vendors, or practitioners) may leverage in their own work. Stakeholders who are generating data themselves (e.g., utilities, hardware providers, etc.) may also contribute anonymized data for analysts to study use cases of particular interest to them, and to deploy solutions developed via the platform.

Finally, the project fosters collaboration between different types of users across different institutions including researchers, students, and industry. By providing open data and challenging users to use it in answering questions that stakeholders have, the project aims to remove any and all obstacles to the rapid prototyping, deployment, and adoption of new use cases for data analytics, machine learning, and artificial intelligence. Competitions will be hosted to allow analysts to showcase their ideas. Results will be shared at conferences. Finally, an online community will be developed to share and promote ideas developed using the platform.

The result of this investment is to provide widespread access to an ecosystem for quickly exploring, prototyping, and deploying new analytics that will allow the industry to extract more value from time series data. The NI4AI ecosystem is supported by the PredictiveGrid™ Platform shown in Figure 3. The platform uses an innovative database architecture custom-built to streamline workflows for interacting with long time histories of high-frequency and high-density sensor data. The platform offers faster-than-real-time visualization and analysis of large data archives, reducing time and effort it takes for users in any area of expertise – including data analysts, engineers, and grid operators – to explore ideas for translating data into actionable information as problems or questions come up on the job.

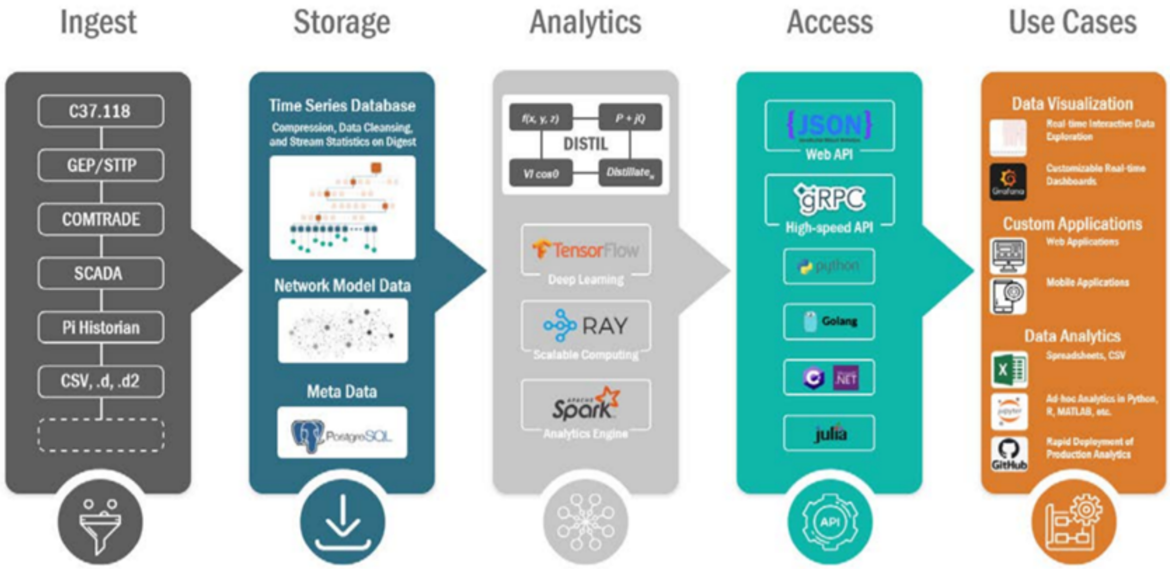
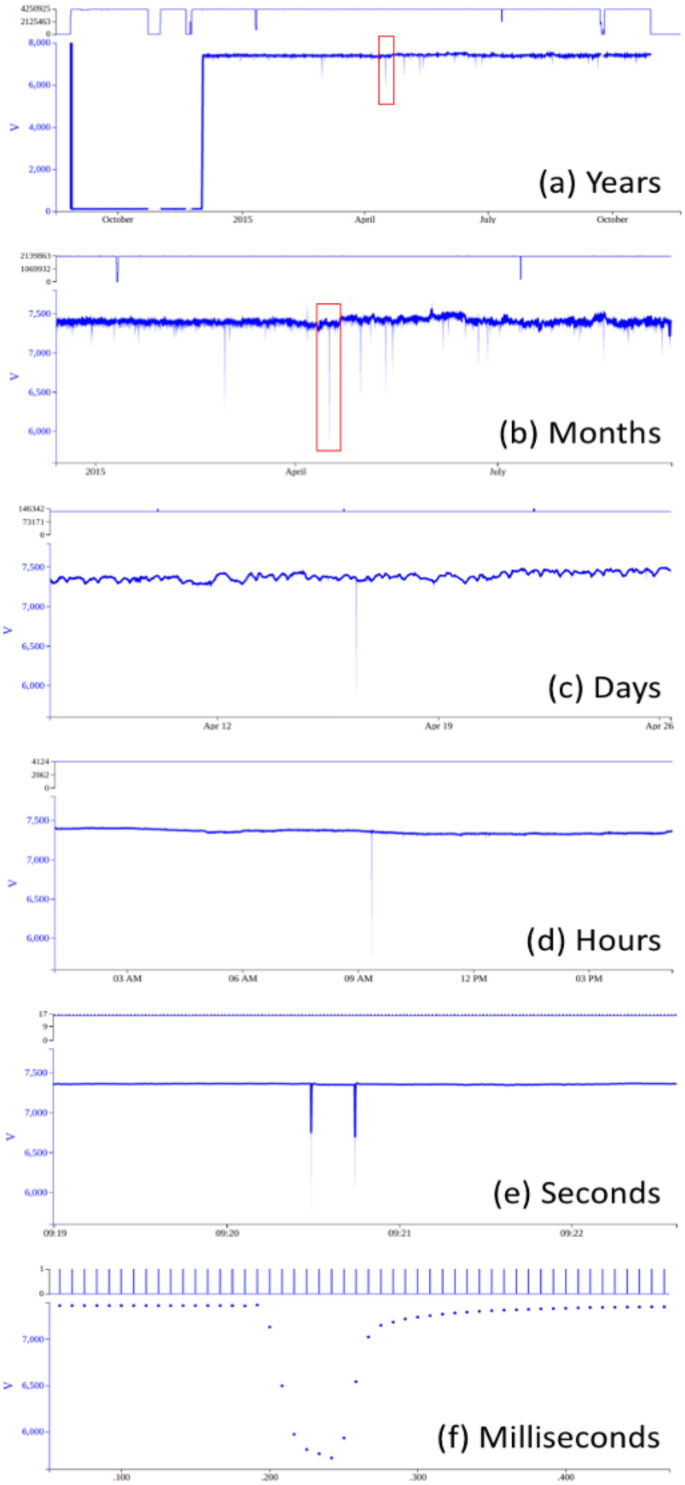


Figure 3 –PredictiveGrid ™ System Diagram

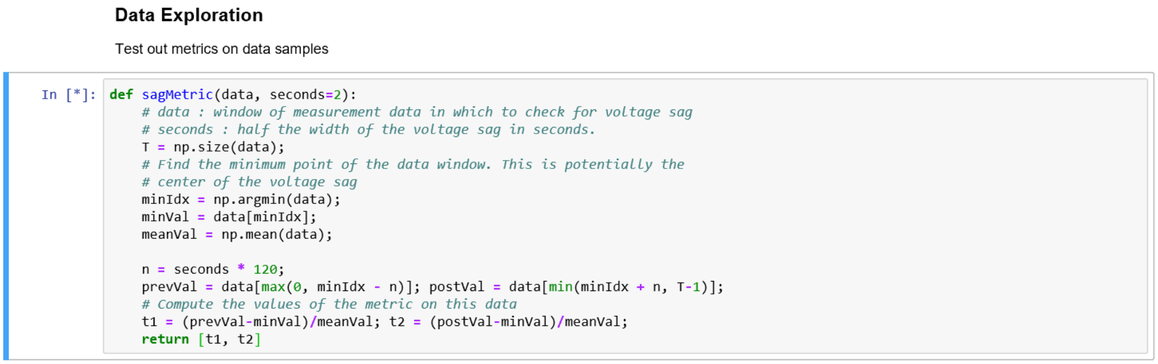
Example – Voltage Sag Identification
Analytic development often begins with data exploration, a step whose value is difficult to overemphasize, especially when the user is unfamiliar with the measurement data or the phenomenon of interest. Pervasive sensor data, especially continuous, high frequency measurements of the electric grid, is relatively new, and some sensors, such as distribution synchrophasors (µPMUs), are novel. Many utilities and academic researchers are unfamiliar with the sensor measurements or only have experience with simulated versions different from real data. For example, there is some uncertainty about the meaning of µPMU measurements [iii]. During transient events such as faults or when the system frequency is not constant, the voltage and current do not follow the perfect sinusoidal model--with a 60Hz frequency and fixed amplitude and phase. However, the PMU always outputs a magnitude and angle measurement that implicitly assumes a perfect sinusoid. Therefore, the physical interpretation of the returned magnitude and angle measurements can be ambiguous.   Voltage sags have a very distinctive appearance in PMU voltage magnitude measurements. However, users familiar with point-on-wave or lower resolution measurements may not immediately recognize the right metric to isolate these events in PMU magnitude data. Therefore, exploring the data and gaining familiarity with these novel data sets is a vital first step in developing new applications.



The NI4AI enables *exploration* of high volume, high resolution, multi-modal grid sensor data via rapid data access and interactive, multi-stream visualization across time scales [iv]. Given this capability, voltage sags can be identified at the lowest temporal resolution, where months or years of data are visible (top panel, Figure 4). The visualization not only shows the average value for the time period represented by a single pixel column but also shows the minimum and maximum values as a shaded region. Thus, even at this resolution, voltage decreases are visible as fine spikes. Traversing the panels in Figure 4 from top to bottom, each shows an increased level of “zoom” or finer temporal resolution and required a separate query of the platform. Each query completed in less than 200 milliseconds making truly interactive data exploration possible [v]. With the platform, the user can select an area of interest, shown in red, and smoothly zoom in to resolve individual 120Hz measurements. At this level of zoom, the exact shape of the transient voltage decrease is evident. After examining numerous such events, the user is armed with intuition for what makes these events unique and can start to prototype a potential detection approach.

The general approach to voltage sag detection is to compute a metric on a window of data that indicates the presence of a voltage sag. Based on the data exploration step, the voltage sag’s shape suggests several potential metrics. The first possibility, Voltage Sag Metric 1 or VSM-1, finds the minimum value within the signal segment and computes the differences between this minimum value and the measurements shortly preceding and following it (both differences are expected to be large). The second possibility, VSM-2, calculates the difference between the window’s mean and minimum (expected to be large) and the difference between the window’s maximum and mean (expected to be small). Metric two ensures that the voltage sag consists of a narrow spike that drops significantly below an otherwise predominantly flat signal.

Figure 4 – Panels (a) – (f) show plots of an event in interest, highlighted by a red box, at increasing temporal resolutions. Each query requesting PMU data from the platform took less than 250 milliseconds to complete, allowing for interactive data exploration.



Data Exploration

Test out metrics on data samples

In [*]:

```
def sagMetric(data, seconds=2):
    # data : window of measurement data in which to check for voltage sag
    # seconds : half the width of the voltage sag in seconds.
    T = np.size(data);
    # Find the minimum point of the data window. This is potentially the
    # center of the voltage sag
    minIdx = np.argmin(data);
    minVal = data[minIdx];
    meanVal = np.mean(data);

    n = seconds * 120;
    prevVal = data[max(0, minIdx - n)]; postVal = data[min(minIdx + n, T-1)];
    # Compute the values of the metric on this data
    t1 = (prevVal-minVal)/meanVal; t2 = (postVal-minVal)/meanVal;
    return [t1, t2]
```

Figure 5 - A sample input cell from a Jupyter Notebook, written in Python, showing the code that implements the first Voltage Sag Metric. Output from this code would be displayed immediately below the cell in the same document, enabling a literate programming style.

Using the Python API, test segments of data that include voltage drops as well as other notable changes that are *not* voltage drops found via data exploration are pulled into the Jupyter Notebook. Next, the two proposed voltage sag metrics are computed over the samples to get a sense of their efficacy. The implementation of the first metric in the notebook is shown in Figure 5. Each proposed metric consists of two values and can be quickly evaluated using a scatter plot Figure 6. This preliminary result suggests that both metrics are effective for detecting voltage sags.

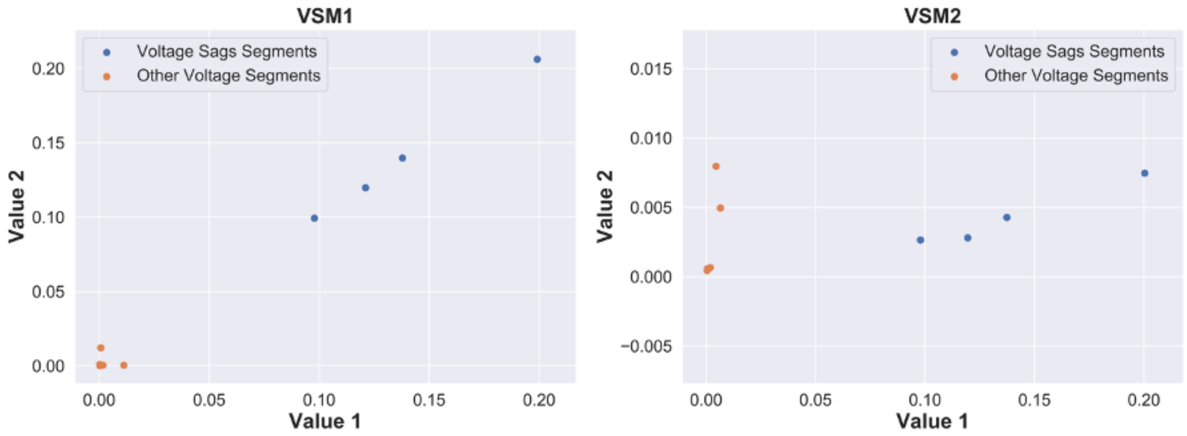


Figure 6 - Scatter plots demonstrating the effectiveness of both voltage sag metrics. The (x,y) pair capture the computed values for each voltage sag metric with the blue dots representing visually confirmed voltage sags and the red dots representing data from other periods of time without voltage sags.

After verifying the promise of the two approaches on a few samples, the speed of the platform enables testing over a much larger data set before being deployed to production. VSM-1 requires scanning through all of the data, querying the database for full resolution measurements (120Hz). Implementing this is as simple as putting our scripts from the exploratory phase into a for loop. Over one day of voltage magnitude measurements, the first approach runs in **23 minutes** or **63x** real time. The algorithm detected six suspected voltage sags.

Metric two can be similarly implemented by querying the data at full resolution within a for loop. However, the formulation of the metric allows us to leverage an important aspect of the platform to achieve even faster performance. In addition to raw values, the universal sensor platform stores summary statistics at the internal nodes of its tree structure. At a particular internal node, the summary statistics consist of the mean, minimum, and maximum over all values "below" that node. These summary statistics can be queried more rapidly than the raw values. Since metric two is defined only in terms of the mean, minimum, and maximum over a window, we can compute it by querying the summary statistics for the window. Over the same day of voltage magnitude measurements, the approach runs in **1.28 seconds** or **67750x** real time. This algorithm detected the same six suspected voltage sags as VSM-1.

In this test case, the two approaches have dramatically different runtimes but detect the same events. This may not always be the case and using summary statistics inherently limits the range of analytics possible compared to using the raw, full resolution data. However, as this example demonstrates, the power of the platform is that both approaches can be prototyped and tested rapidly so the appropriate tradeoffs between speed and accuracy can be chosen.

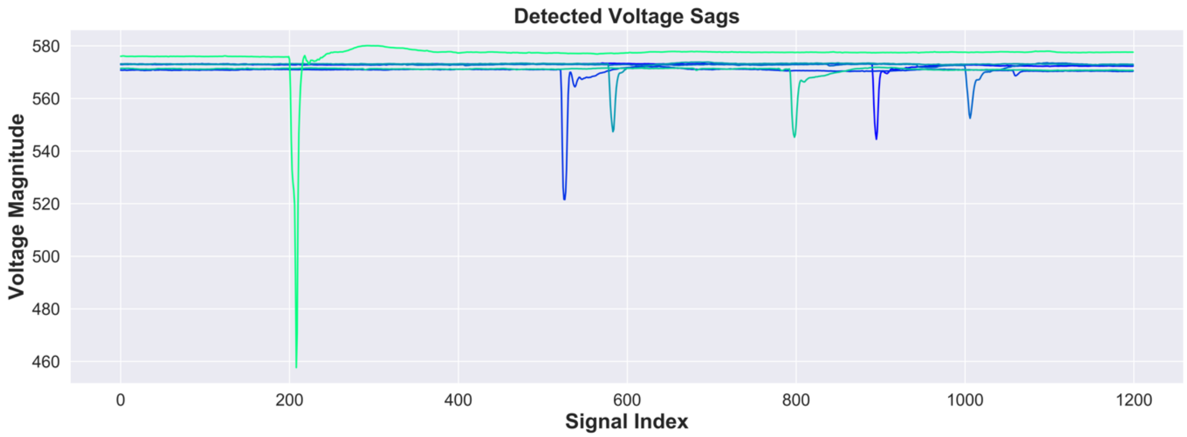


Figure 7 - Visualization showing the actual phasor magnitude data for each detected voltage sag.

Based on the visualization in Figure 6, the detected voltage sags are indeed real voltage sags with a distinctive and consistent shape, validating this metric for a larger sample of data. Additional analyses can be easily visualized to offer more insights into the voltage sags and the event detection approach. A heat map (Figure 8) indicates the sensor locations where voltage sags occur over a month and a histogram (Figure 9) demonstrates the size distribution of the sags. From the heatmap, we see that sensor 1 captures many voltage sags throughout the first part of the month. This may indicate the presence of a periodic load near that particular PMU which becomes inactive during the second part of the month.

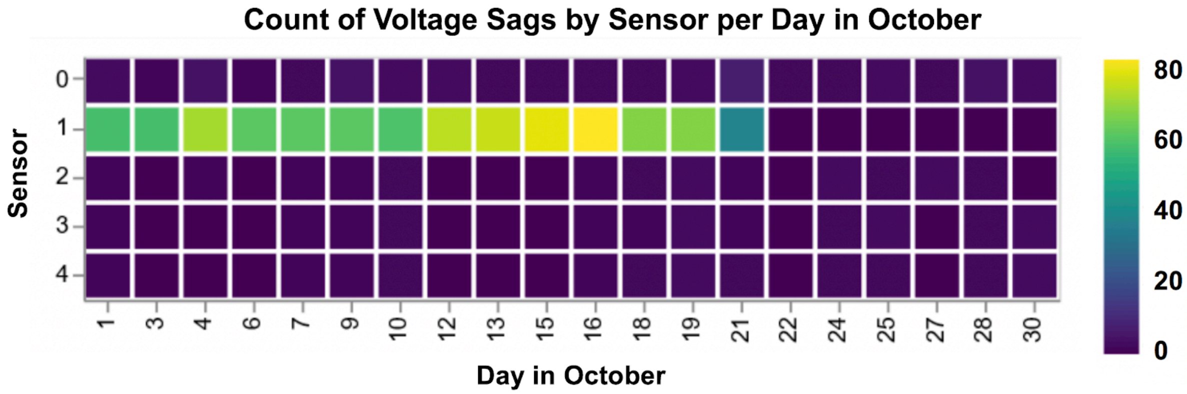


Figure 8 - Histogram of the number of detected voltage sags per day per sensor to give a larger view of total system behavior over time.

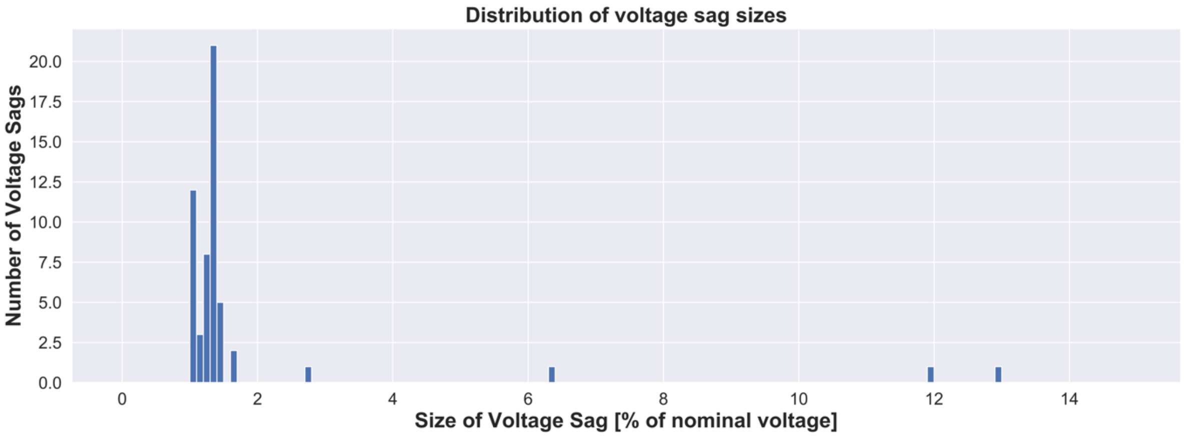


Figure 9 - A histogram showing the distribution of voltage sags detected for all available data.

Conclusion

Utilities are dramatically limited in their ability both to test hypotheses and use cases leveraging data and to move prototype analytics into full deployment in production systems. To show that the status quo is not some fundamental limitation, this paper demonstrates not only the rapid development of a use case of interest using high density PMU data but also the deployment of this use case to a production big data system with operational data. This rapid traversal of the analytics pipeline was made possible through the use of the NI4AI.

To learn more, visit the National Infrastructure for Artificial Intelligence website https://ni4ai.org/ .

Authors

Mohini Bariya is a graduate student at the University of California Berkeley in the Electrical Engineering & Computer Sciences department. Her research interests include the electric Grid, Energy, Smart Grid, data analytics, and synchrophasors.

Theo Laughner is the Director of Engineering at Lifescale Analytics, where he leads the energy practice. He has decades of utility experience integrating data from disparate monitoring systems. He enjoys engaging with utilities and researchers via CIGRE, EPRI, and IEEE to advance the state of the art for grid analytics.

Sean Patrick Murphy is the CEO of PingThings, Inc., creators of the world's fastest time series data management, analytics, and AI platform. PingThings is his fourth company. Previously, he served as a senior scientist at the JHU Applied Physics Laboratory for over a decade, where he focused on machine learning, high-performance and cloud-based computing, and anomaly detection. He has degrees in Mathematics (UMCP), electrical engineering (UMCP), biomedical engineering (JHU), and business (Oxford), all with high honors or distinction.

References

[i] Schweitzer EO, Keynote Presentation, 2017 Reliability Leadership Summit, Washington, DC, March 21, 2017.

[ii] Murphy SP, Schuman J, Jones KD, Bariya M, Andersen M, A Universal Platform for Utility Sensor Analytics and Artificial Intelligence, 2018 Grid of the Future Symposium, Dulles, VA, Oct, 2018, Paper Submitted.

[iii] Kirkham, Harold, and Jeff Dagle. "Synchronous phasor-like measurements." *Innovative Smart Grid Technologies Conference (ISGT), 2014 IEEE PES*. IEEE, 2014.

[iv] Kumar S, Michael P Andersen, and David E. Culler, Unifying data reduction in storage and visualization systems, SIGMOD'18, June 2018, Houston, Texas, USA

[v] Andersen M and Culler D, BTrDB: Optimizing Storage System Design for Timeseries Processing, Fast '16 *14th USENIX Conference on File and Storage Technologies*, Feb 2016.