21, rue d'Artois, F-75008 PARIS
http : //www.cigre.org

CIGRE US National Committee
2016 Grid of the Future Symposium

Unlocking the Value of PMU Data for Electric Utilities

S.P. MURPHY, J. SCHUMAN
PingThings, Inc.

SUMMARY

With the advent and subsequent deployment of synchrophasor technology across the transmission portion of the grid, the electric utility industry and federal government have made a significant and public investment toward a more data-oriented future over the last two decades. We will argue that all of the necessary ingredients are now available to leverage the data produced by these new sensors to generate far greater ROI than had originally been anticipated.

To unlock the latent value in this data, utilities need the (1) appropriate raw material in the form of high quality data, (2) suitable financial motivation to use the data to solve known and unknown problems, and (3) capability to realize the use cases in a scalable and cost effective fashion. This paper will show that data quality is a readily solvable issue despite having not been adequately addressed. It is neither a technology nor an algorithm problem, but appears to be one of organizational will. Significant incentive to use PMU and other next generation sensor data exists in many forms. PMU data can be leveraged to reduce unnecessary costs for utilities and enhance operational visibility without the need for additional investment in equipment and systems. We also note that other industry sectors have evolved the state of the art in big data storage and analysis and that the utility industry could capitalize on the value that those investments have produced.

KEYWORDS

data science, Phasor Measurement Unit, synchrophasor, predictive analytics, time series data, real time stream analytics, data quality

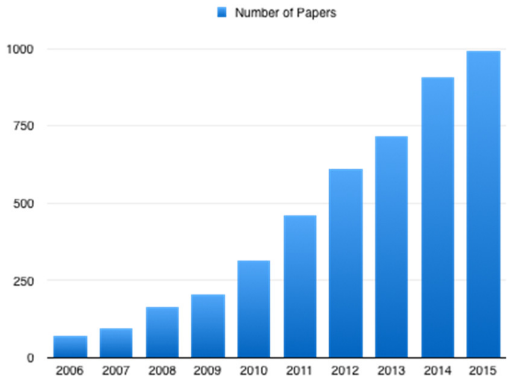
sean@pingthings.io

Introduction

The phasor representation of current and voltage was first published in 1893[1]. Nearly a century later, the phasor measurement unit (PMU) was invented at Virginia Tech and, in 1995, the IEEE published the first PMU data specification[2].The Department of Energy and the Office of Electricity Delivery and Energy Reliability, with the help of the American Recovery and Reinvestment Act of 2009, pushed $357 million to the appropriation and deployment of PMUs across the US power grid. From 2009 to 2015, there number of networked PMUs in the United States of America has grown over 900%, from 166 to over 1,700[3].

While it took two decades for PMUs to go from inception to deployment, the proliferation rate of synchrophasors and other next generation sensors on the grid continues to increase for several reasons. First, major hardware manufacturers, such as Schweitzer and GE, are embedding synchrophasor-type sensing into many standard relay packages and other previously non-instrumented grid components. As old equipment is replaced with new, more capable models, the number of sensors on the transmission grid has, and will continue to increase regardless of utility intention. Second, while synchrophasors have long been associated with the transmission side of the grid, numerous efforts are underway to bring micro-synchrophasors or µPMUs with higher-fidelity monitoring to the distribution side of the grid[4]. Further, numerous other data sources external to utilities have become increasingly relevant to optimal grid function. Sources such as the magnetometer data feeds from the United States Geological Survey, the planetary K index provided by the Space Weather Prediction Center, and the magnetotelluric data provided by Earthscope provide insight into the behavior of the nation's power grid[5].

This increase in sensor penetration on the grid has been met with a corresponding increase in academic attention. Google Scholar currently reports a total of 5,350 papers with references to synchrophasors worldwide. As shown in the bar chart below, patents and papers have been on a steady increase for the past decade as synchrophasor technology matures and manifests itself in different locations within the bulk power system, such as relays and intelligent electrical devices (IEDs).



Similarly, the number of books focused on synchrophasors has seen a corresponding rise; Amazon now reports thirty-seven books available on this subject matter.

Given this growth trajectory, it is natural to ask, where is the power industry now? The 1,700 networked synchrophasors capture a terabyte of data per month describing, in detail, the complex behavior of this critical infrastructure that allows our modern society to function. How is this data being used? Is it being used? While many would not consider a terabyte "web-scale" data, it is larger than the utility industry is accustomed to and taxes the capabilities of current data systems and software.

To put this into perspective, Google was integrating over 20 petabytes of data (or 20,000 terabytes) in 2012 to provide Google Maps as a *free* service to consumers. Powering Google maps was a wide array of data including satellite imagery from government and third party providers, millions of sensors embedded in roads in each state, and video, LIDAR, and 65-megapixel images collected by its Street View vehicles[6].

Due to its technical and engineering nature, the utility sector as an industry has always embraced data. However, if we define the hackneyed term "big data" as data of volume, velocity, and variety that exceeds normal capabilities, the utilities seem hesitant to take the next step, even in 2016. In contrast, the intrinsic value of big data was being realized and discussed in 2006 by other sectors:

> *"Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.[7]"*

To unlock the value of the data that PMUs and micro PMUs and other sensors collect, utilities need to build applications that leverage this data to solve problems of interest. To enable this capability, there are several fundamental requirements. First, to use data, the utilities must have data and the quality of the data must be good. Second, the industry must see value in the appropriate use cases that drive the development of products or applications. Finally, with sufficient motivation and raw material, utilities must have access to the capability to store, understand, and use data at scale. The goal of this paper is to provide an overview of these requirements and explore each, demonstrating that the time is ripe for data-driven applications to help re-invent the US power grid.

1 Raw Material - Data Quality

Data serves as the raw material needed to build applications. The value that can be created by such data-driven applications is a function of the underlying data quality. Data that accurately describes the system state can be used to power a wide range of use cases that will be discussed later in this paper. At the other end of the spectrum, bad data cannot; as they say in computer science, "garbage in, garbage out." Fortunately for the industry, data does not have to be perfect to be useful but it is critical to understand and quantify the errors and issues it contains so that we understand how confident we can be in whatever decision is being supported by the data.

Currently, there is a debate to understand how varying levels of data quality impact various applications but this effort is somewhat premature. As the United States Supreme Court Justice Potter Stewart said about the threshold for obscenity, the same test holds true for data quality.

> *"I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description, and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that."*

There is good data and then there is bad data and, based on the current state of data quality in the industry, there is little question when data is bad. PMU sensors should not report months of negative, identical floating-point values for the magnitude of the voltage phasor on the high side of an energized transformer. If so, a data quality problem exists and needs to be resolved.

Beyond not enabling value creation, bad data is an organizational liability. Spending significant resources installing PMUs and the data storage systems required to (at least) archive the data and then finding the historian contains ten terabytes of bad data is not a best practice in any industry. Worse, such behavior could open an awkward line of questioning from regulators, policy makers, or even consumers, especially when public funds contributed to the initiative. Once the problems of data quality have been addressed, there is an expectation that good data will be used for applications beyond archiving. Once the data genie is out of the bottle, the industry will have to advance.

1.1 Wrangling Data Quality

Data quality can divided into two components: (1) completeness and (2) accuracy. Data completeness examines if the expected data is present and answers such questions as whether frames were dropped and did frames arrive with a bad CRC. Data accuracy tests the degree to which the values reflect the physical system being measured. For PMU data, this applies to both the time data and the sensor measurements including voltage and current magnitudes and angles and frequencies. Such metrics can be derived from the header information provided by the IEEE C37.118 specification (the STAT flag in particular) or computed from the actual data or a combination of the two. A number of checks exist to test PMU data for accuracy including testing for:

- Null Values – “blank” values that were inserted by software.
- Missing Values – values that were not captured by the sensors and/or the system.
- Zero Values – while zero is a possible value for measurements, the presence of one or more zeroes can indicate bad data.
- Repeated Values – the probability of seeing two sequential and identical 32-bit or 64-bit floating-point values produced by a sensor monitoring a physical process is low so repeated values are problematic.
- Out of Range Values – measurements that are outside the range of possible values as dictated by physics.
  Off Normal Values – numbers that are some number of standard deviations above or below a running mean or median computed from the last set of samples.

More advanced criteria can be used to identify additional data issues as well. PMU-based measurements can be compared to values from SCADA systems. Statistical patterns in both a single time series and across multiple signals can indicate data quality issues.

A variety of products in the marketplace offer different solutions to the issue of data quality assurance. Some products are open source and free while others are commercial. Some offer snapshot assessments of data quality and others provide continuous, real-time monitoring. One-off reports can provide a snap-shot of data quality that help to identify and locate the initial set of problems that plague data quality. However, the grid is dynamic; equipment fails, topologies change, and events occur. With near certainty, data quality will fluctuate and bad data will infiltrate the system, impacting downstream dependencies. The only way to accommodate this reality and minimize its impact is to monitor sensor data quality in real-time.

Real-time applications using data also require data quality to be continuously monitored. In the best case, the applications’ underlying algorithms can adapt to differing levels of noise and data drops accordingly. If this is not possible, the applications must at least notify the user of the confidence in the provided results. Finally, any meaningful solution will not only describe and quantify the current issues with data but also help diagnose where the potential problem exists.

2 Motivation

Data will become an asset for the industry once the quality issue is solved. As sensor measurements are just another type of data, collecting power grid data is comparable to accruing assets. With enough measurements, an organization enters “big data” territory. The three defining characteristics of big data are the three V’s - **V**olume, **V**elocity, and **V**ariety.

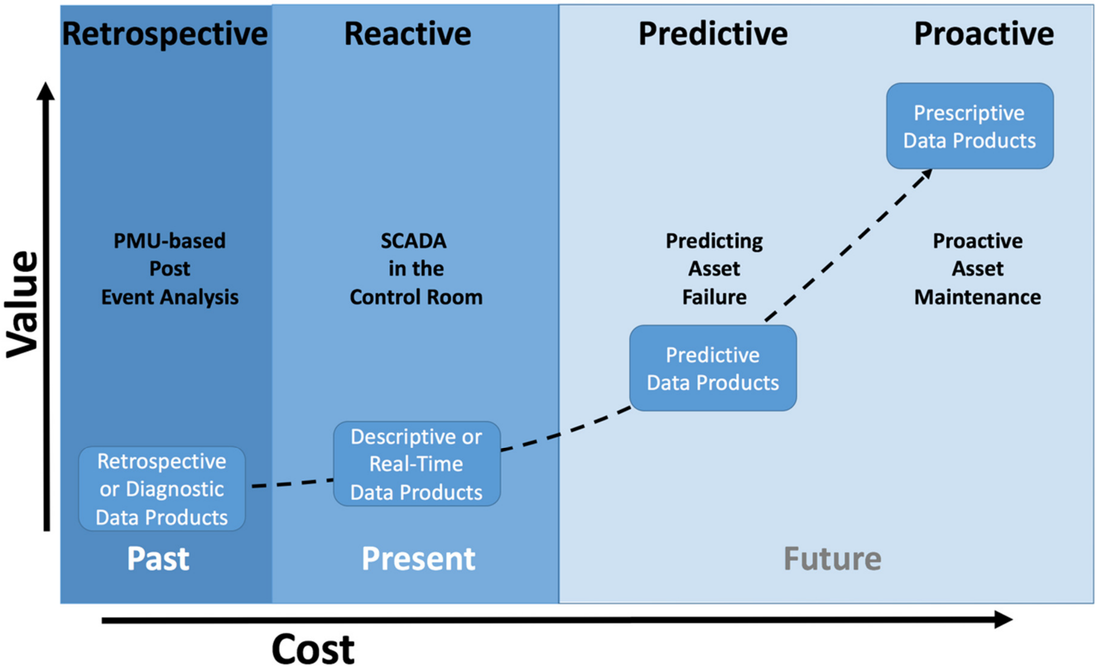
1. Volume - the amount of data as measured in bytes, with terms such as Exabyte (1,000,000,000,000,000,000 bytes or 10^{18} bytes) entering the vernacular.
2. Velocity - the rate at which new data is created. This aspect is particularly relevant to utilities as next generation PMUs move from 30Hz to 60Hz, 120Hz, or 240Hz sampling and beyond.
3. Variety - data comes in many forms beyond what can easily be stored in a spreadsheet, including video, audio, and natural language.

Some add **V**eracity as the fourth “V,” which partially refers to data quality.

While “big” data is large in amount, size is relative to the available tool and skill sets. The term “big data” has resonated with society because data has become a “big” deal. Where scientists and engineers have long known that data was useful, the rest of the world has realized that data can create many types of value—financial, environmental, and social value—and the only V that matters is “**V**alue.”

2.1 Value Creation

Data can be used to quantify aspects of the real world, creating highly detailed models capable of producing insights, understandings, or predictions of the past, present, or future behaviors, respectively, of the system. These views create increasing value, as shown in the figure below as a function of cost.


Retrospective
Reactive
Predictive
Proactive
Prescriptive Data Products
PMU-based Post Event Analysis
SCADA in the Control Room
Predicting Asset Failure
Proactive Asset Maintenance
Predictive Data Products
Descriptive or Real-Time Data Products
Retrospective or Diagnostic Data Products
Past
Present
Future
Value
Cost

The first wave of data-driven applications tend to focus on the past, exploring and understanding events that have already impacted the system. Utilities are no stranger to forensic event analysis and understand the associated value proposition. Studies have shown that PMU data can accelerate these investigations, shrinking the time required from months to weeks, reducing the associated costs and resources consumed[8]. The southwest blackout in 2011 was one of the first examples where synchrophasor data was used to determine the cause of the outage. From a technical standpoint, this is the easiest time period to address given data of sufficient quality has been archived.

The contemporary control room is an excellent example of a data application that describe the present system state. Such applications often require large volumes of data be compressed and summarized for human consumption and response. To date, control rooms are dominated by data from SCADA systems.

The most value can be created by focusing on the prediction and prescription of the future as shown in the right most shaded regions of the figure above and this is where PMU data will shine. Undiscovered patterns predictive of crucial system behaviors are much more likely to be discovered in PMU data, where the actual behavior of the system is being record 30 to 60 times a second, than in SCADA data, with sensor measurements once every few seconds. It is clear that prescient knowledge about asset condition or imminent failure is worth billions of dollars to the utility industry.

When discussing value creation from data, there exists a large difference between traditional reporting methodologies and real time data-driven applications. Reports are a familiar product that are also built from data; data that has been collected is collapsed into meaningful prose and visualizations to convey a message from one person or group to another. Typically, each stage of the process is manual and the end result is static. Academic journal articles are another such example of traditional reporting. Data-driven applications lay at the other end of the spectrum of scalability. In the data-driven application, the result or conclusion is repeatedly created on demand via an automated system. Compared to the traditional report, all steps that were required to generate results now live within computer code with all data sources programmatically accessible.

2.2 Use Cases of Interest

Legitimate use cases for data must be found to justify the cost of building data-driven applications. Fortunately, the majority of the expense for data-driven applications is the deployment of the sensor system and the collection of the data; leveraging the existing data is at least an order of magnitude less costly. As an example, we look at the Waze navigation application for smartphones. With $67 million in investment, Waze built the data infrastructure and analytics to manage an ad-hoc network of 50 million sensors reporting continuous traffic data, whereas the sensors themselves (smartphones) had a combined value of nearly ten billion US dollars.

Some use cases for data are apparent before any data is collected and some become apparent once the data is available. Model validation is an example of the former. It is clear that the 30 Hz-sampled PMU data provides more temporal information than existing SCADA systems and can help improve the accuracy of system models. Additional often-cited uses cases for PMU data that form the basis for data-driven applications include:

- Wide area situational awareness
- Phase angle monitoring
- Real-time visualization
- Operator training and event simulation
- Forensic event analysis
- Oscillation detection
- Voltage stability monitoring and management[9].

Once data is available, additional use cases become apparent. PMU data offers a second source of “truth” for the current operational state of the grid. As a result, this data can be used for cyber security applications to detect external control and operation of critical assets. Second, PMU data can capture very rapid transients on the grid thereby identifying and verifying the operation of switches and breakers (independent confirmation of what is seen by SCADA).

However, the most interesting and valuable use cases for data are often those that cannot be clearly discerned at the time of data collection (or when the collection was first conceived). As an example, consider Netflix, now the movie and television streaming service. Founded in August 1997, Netflix launched its DVD by mail service in April of the following year[10]. It became apparent that the economics of the business model did not work; the newest DVD releases, which were the least cost effective, were the only movies customers selected. In 1999, Netflix started repurposing previously collected data from customers—movie ratings—and an understanding of each movie’s content to provide individualized content recommendations for each user. This was possible using a number of data science techniques such as collaborative filtering and grew into the company’s proprietary Cinematch algorithm. Netflix’s recommendations successfully drove customers to DVDs other than new releases, increasing profitability. This aspect of the business has become so crucial that Netflix’s annual spend on content discovery exceeded $150M annually in 2014[11]. Netflix has taken this quantified understanding of customer preferences and used it to construct new content, such as “House of Cards," moving from predicting individual customer desires to prescribing hits shows for an entire population[12].

As an unexpected use case that could be addressed with PMU data, PingThings demonstrated last year that the reactive power consumed by an auto transformer during a geo-magnetic disturbance could be determined from PMU data. This simply was not possible using SCADA data; although GIC's are considered quasi-DC currents, the causal geo-magnetic disturbances often manifest fast and large transient behavior that would be missed by SCADA. This creates the very real possibility of monitoring the power grid for geomagnetically induced currents without the need for installing costly sensors but by leveraging PMUs.

Beyond unexpected use cases, utilities are faced with additional challenges, such as the need to improve operational and capital costs, facilitate intermittent renewables, and meet more stringent reliability standards in the face of disruptive loads and changing weather patterns. To that end, utilities are becoming interested in applying cross-departmental approaches to gathering information to develop a more holistic asset management and condition monitoring (AMCM) strategy. These approaches entail a better understanding of asset states and using this knowledge to develop a predictive risk-based management strategy—as opposed to responsive. According to Navigant Research, global revenue for power AMCM devices and solutions is expected to grow from $2.6 billion in 2016 to $6.5 billion in 2025.

3 Capability - Ability to Access and Use Big Data

While successful data capture and storage are the first steps to creating data-driven applications, they are not the last. To create new value, utilities and utility partners must be able to access and use the PMU data. For traditional analyses, this often requires the utility to provide a data "dump" to a third party. Current industry data systems, built on older, relational database technology, cannot suitably handle this simple requirement. Anecdotal information suggests that exporting a month of PMU data can take a week or more.

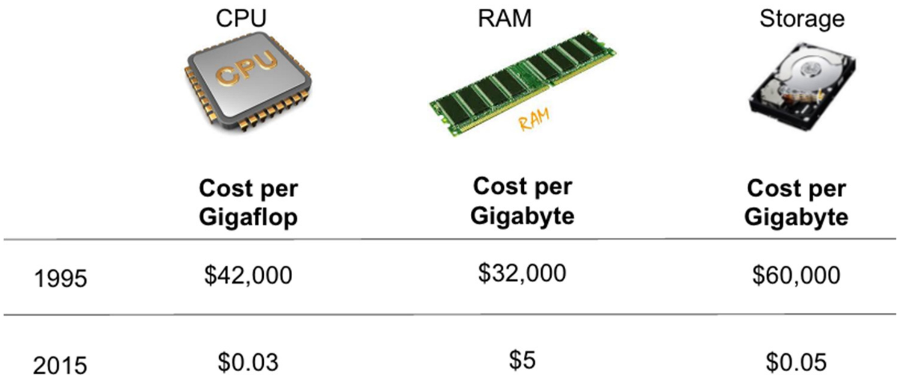
If the utility wants to prototype, build, test, and deploy data-driven applications, the requirements for the underlying data storage systems increases significantly beyond current solutions. Further, synchrophasors generate more data than can be comprehended by humans, either in real-time or retrospectively. Big data by itself lacks value without the approaches necessary to use it and that critical approach is machine learning. Machine learning offers a fundamental paradigm shift for computing. Previous models of programming had humans craft instructions so that computers could execute tasks. In machine learning, the data actually programs the computer to achieve the desired outcome. Thus, all PMU data streaming into phasor data concentrators is not a cost center or burden but an opportunity. This data can be used to train computers to identify events of interest to a utility and much more.

To do this requires data storage and software frameworks for processing data at scale and while streaming. Fortunately for the power industry, other industries have advanced related technologies making these requirements possible and both the hardware and software costs have rapidly plummeted over the last two decades.

3.1 Moore's Law

Gordon Moore, one of the co-founders of Intel, observed that the number of components per integrated chip would double every year in an appropriately titled magazine article: Cramming more components onto integrated circuits. This doubling of component density has allowed processors to double in performance approximately every two years for the last half century, enabling everything from the Internet to smartphones[13].

From the electric utility's perspective, Moore's law has unlocked incredible computational possibilities at exponentially decreasing cost. The chart below compares this decrease in cost for three key computer characteristics: processing power, memory size, and non-volatile storage size.



	CPU	RAM	Storage
	Cost per Gigaflop	Cost per Gigabyte	Cost per Gigabyte
1995	$42,000	$32,000	$60,000
2015	$0.03	$5	$0.05

The cost of the microprocessor's computing power has dropped by 1.4 million times. The cost of main memory has decreased by a factor of 6,400 and the cost of permanent storage has fallen by a factor of 1.5 million.

To put these numbers in perspective, it is now possible to purchase enough system memory (RAM) to hold a month of data from every PMU in North America for approximately $10,000. In comparison, the world's fastest supercomputer in 1995, Fujitsu's Numerical Wind Tunnel, did not have this much main memory[14]. Further emphasizing this point, consumer game consoles available today have sufficient hard drive storage space to archive several months of PMU data for the United States of America. These hardware cost decreases are available to utilities both as hardware systems installed locally and via externally managed Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) cloud solutions.

3.2 The Evolution of Big Data Software

While Moore's law enabled the hardware foundation needed for handling big data to be built, the open source software community kept pace and built a vast ecosystem of software for both data storage and analysis at scale over the last two decades. The seminal flow of big data-related papers from Google provides an overview of this rapid evolution. From this stream, one can see the sequence of big data challenges that Google faced and the technical approaches used to surmount them.

Starting in 2003, Google implemented a distributed file system (GFS) and a software infrastructure (MapReduce, the basis for Hadoop) to process data spread across thousands or millions of commodity, inexpensive computer hardware with relative transparency to the end user[15, 16]. This approach enabled orders of magnitude cost reductions when processing massive amounts of data. Per a conversation with a former Google VP, any software engineer at Google could run a job across 50,000 machines without manager approval in 2007.

Next, Google describes how it created numerous data storage and processing systems to handle different types of data – structured, unstructured, graph, and even globally distributed yet synchronized database – at Internet-scale[17, 18, 19]. The key to this innovation is that instead of requiring larger, more expensive hardware to handle more data as was the paradigm at the time, these systems could accommodate more by horizontally scaling across additional machines. This, again, radically changed the cost curve for handling big data.

Finally, Google discusses its framework and processing model for streaming computations at scale[20]. Historically, most data analyses were done on finite or bounded data sets. For example, a utility exports a large csv file containing PMU data that can be loaded in Excel for an academic project. Even in older big data processing frameworks like Hadoop, processing is done in batches, on finite chunks of data. In the real world, many data sets are infinite with new data constantly arriving. This is especially true for utilities and other industries where sensor data is continuously generated. Google's dataflow framework analyzes and processes data streams with the assumption that new data will always arrive, that old data may be retracted, and that batch data processing is just a special case.

Where Google has led the way, open source software solutions have followed. Today, there is a rich ecosystem available that can accommodate the utilities' big data needs and this software can be freely downloaded and used. Just as Moore's Law has helped hardware costs drop exponentially, we have seen even more radical cost reductions in the big data software space. However, these technologies represent a departure from the utility industry's customary methodologies and also require a different mentality and skill set. Further, the big data software revolution has occurred on open source operating systems, primarily different flavors of Linux, and is not Microsoft Windows compatible.

Conclusion

The electric utility industry has made a significant and public investment in the deployment of PMUs. All of the necessary ingredients are available to leverage this data to produce far greater ROI than had originally been anticipated. Data quality is a solvable issue; it is neither a technology nor an algorithm problem, but one of organizational will. Further, significant motivation exists in many forms. PMU data can be used to reduce unnecessary costs for utilities and even previously unknown use cases such as GMD monitoring, are possible. Finally, other industries have evolved the state of the art in distributed compute and storage systems to surpass the current needs of the utility industry.

Some may say that data quality is a catch 22 for the industry. Without obvious benefits, there are no reasons to resolve data quality problems but, without good data, it is not possible to build applications that solve relevant issues. This is incorrect. Let's for a moment, ignore the current PMU application areas. Let's also ignore the tantalizing prospects of unsolved but valuable contributions to grid reliability that could be made. Finally, let's forget that the most valuable use cases for next generation sensor data have probably not even been conceived of yet. There is still the issue of liability. Data, on its own, is a liability; if it is not cleaned and used, then utilities open themselves to uncomfortable questions from regulators, policy makers, and the general public.

While this paper focused primarily on PMU data-driven applications, there is no reason to limit ourselves to just synchrophasors. Phasors approximate the waveform being measured. Internally, PMUs are sampling the waveform several dozen times per cycle to extrapolate a simplified model of the current and voltage waveforms. From a data perspective, this represents an order of magnitude down sampling of the original data. The relevant question is why? Why not just sample current and voltage waveforms directly at several kilohertz. This is already done by the Digital Fault Recorder (DFR), which samples up to 10 KHz when a particular threshold is exceeded or condition is met. In current devices, only a very brief time period of data is captured due to the relatively high data rate. However, this is a technical limitation only. Despite being historically cost prohibitive, the ability to store, analyze, and use this volume of data is now possible.

BIBLIOGRAPHY

[1] Steinmetz, C., (1893). "Complex Quantities and Their Use in Electrical Engineering".Proceedings of the International Electrical Congress, Chicago. Chicago, Illinois 1893 conference of the AIEE:American Institute of Electrical Engineers Proceedings: 33–74.

[2] 1344-1995 IEEE Standards for Synchrophasers for Power Systems, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=943067&isnumber=20419 (1995).

[3] Advancement of Synchrophasor Technology in Projects Funded by the American Recovery and Reinvestment Act of 2009, March 2016, US Department of Energy

[4] A. von Meier, D. Culler, A. McEachern and R. Arghandeh, “Micro-Synchrophasors for Distribution Systems.” IEEE Power & Energy Society Innovative Smart Grid Technologies Conference, Washington DC, Feb 2014.

[5] Schultz, A., G. D. Egbert, A. Kelbert, T. Peery, V. Clote, B. Fry, S. Erofeeva and staff of the National Geoelectromagnetic Facility and their contractors (2006-2018). "USArray TA Magnetotelluric Transfer Functions". doi:10.17611/DP/EMTF/USARRAY/TA. Retrieved from the IRIS database on Oct 21, 2016

[6] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, Josh Weaver, "Google Street View: Capturing the World at Street Level", *Computer*, vol.43, no. 6, pp. 32-38, June 2010, doi:10.1109/MC.2010.170

[7] M. Palmer, “Data is the new oil,” http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

[8] New technology can improve electric power system efficiency and reliability, March 30, 2012, http://www.eia.gov/todayinenergy/detail.cfm?id=5630

[9] The Value Proposition for Synchrophasor Technology, Itemizing and Calculating the Benefits from Synchrophasor Technology Use, Version 1.0, North American Synchrophasor Initiative NASPI Technical Report, October 2015

[10] https://media.netflix.com/en/about-netflix

[11] N. Hunt, “Quantifying the value of better recommendations” (RecSys 2014, Keynote, https://www.youtube.com/watch?v=lYcDR8z-rRY)

[12] https://www.wired.com/2012/11/netflix-data-gamble/

[13] G. E. Moore, "Cramming more components onto integrated circuits" (Electronics Magazine, 1965. http://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf)

[14] https://www.top500.org/lists/1995/11/

[15] Ghemawat, S., Gobioff, H., and Leung, S. "The Google File System. "*Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles - SOSP '03*(2003).

[16] Dean, J. and Ghemawat, S.. 2004. MapReduce: simplified data processing on large clusters. In *Proc. of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6* (OSDI'04), Vol. 6. USENIX Association, Berkeley, CA, USA, 10-10.

[17] Chang, Fay; Dean, Jeffrey; Ghemawat, Sanjay; Hsieh, Wilson C; Wallach, Deborah A; Burrows, Michael ‘Mike’; Chandra, Tushar; Fikes, Andrew; Gruber, Robert E (2006), "Bigtable: A Distributed Storage System for Structured Data", *Research* (PDF), Google.

[18] Malewicz, Grzegorz, et al. "Pregel: a system for large-scale graph processing." *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010.

[19] James C. Corbett, Jeffrey Dean, Michael Epstein, et al . 2013. Spanner: Google’s Globally Distributed Database. *ACM Trans. Comput. Syst.* 31, 3, Article 8 (Aug 2013).

[20] Akidau, Tyler, et al. "The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing." *Proceedings of the VLDB Endowment* 8.12 (2015): 1792-1803.