Big Data Analytics in Power Distribution Systems

Nanpeng Yu†‡, *Member, IEEE*, Sunil Shah‡, Raymond Johnson‡, *Member, IEEE*, Robert Sherick‡, Mingguo Hong*, *Member IEEE* and Kenneth Loparo*, *Fellow, IEEE*

†University of California, Riverside
Department of Electrical and
Computer Engineering, Riverside, CA
92521 USA

‡Southern California Edison
Advanced Technology Lab
Westminster, CA 92683

*Case Western Reserve University
Department of Electrical Engineering
and Computer Science
Cleveland, Ohio, 44106

***Abstract*—Penetration of advanced sensor systems such as advanced metering infrastructure (AMI), high-frequency overhead and underground current and voltage sensors have been increasing significantly in power distribution systems over the past few years. According to U.S. energy information administration (EIA), the aggregated AMI installation experienced a 17 times increase from 2007 to 2012. The AMI usually collects electricity usage data every 15 minute, instead of once a month. This is a 3,000 fold increase in the amount of data utilities would have processed in the past. It is estimated that the electricity usage data collected through AMI in the U.S. amount to well above 100 terabytes in 2012. To unleash full value of the complex data sets, innovative big data algorithms need to be developed to transform the way we operate and plan for the distribution system. This paper not only proposes promising applications but also provides an in-depth discussion of technical and regulatory challenges and risks of big data analytics in power distribution systems. In addition, a flexible system architecture design is proposed to handle heterogeneous big data analysis workloads.**

***Index Terms*—Advanced Metering Infrastructure, Big Data Analytics, Power Distribution Systems, Data Mining, Predictive Analytics**

I. Introduction

According to the Navigant Research Report, the estimated installed base of smart meters worldwide will surpass 1.1 billion by 2022 [1]. AMI typically collects electricity usage data in the range of 15 minutes to 1 hour, instead of once a month. This is up to a three thousand fold increase in the amount of data utilities would have processed in the past. It means that by 2022 the electric utility industry will be swamped by more than 2 petabytes of data annually from smart meters alone. As we move to the so called Internet of Things, and more devices are connected to the electric grid, even more data will be produced. The primary and secondary value imbedded in the complex and heterogeneous data sets from power distribution systems is immense. However, strategies for unlocking the potential of big data in distribution systems are at an early stage of development. Most utilities in the world are underprepared and expected to struggle with the growing volume of data [2]. There are three major obstacles to the development and implementation of big data analytics in power distribution systems. The first barrier to adoption of big data analytics in the distribution grid is the lack of innovative use cases and application proposals that convert big data into valuable operational intelligence. The second hurdle to adoption is insufficient research on big data analytics system architecture design and advanced mathematics for petascale data. The last obstacle to adoption is the risk of failing to adhere to data privacy and data protection standards. This paper addresses all three obstacles to spearhead the advancement of big data analytics in power distribution systems.

A. Big Data Applications in Other Industries

Big data analytics have been revolutionizing many industries ranging from mature industries such as consumer staples to fast-growing industries such as information technology. Swamped by data measured in petabytes, the information technology industry is the first to adopt big data algorithms and processes. Nowadays, most of the popular computer and web applications are powered by big data analytics algorithms. For example, Amazon's recommendation system now contributes to a third of its sales. The system is fed by a large amount of information on users' behaviors and activities. Similarly, by availing itself of billions of pages of translations of widely varying quality, the Google's translation system has become more accurate than those of other systems [3]. The most notable advances in analytics have come from the social media internet companies such as Yahoo, Facebook, LinkedIn and Google, who have been able to advance the technology to the point where they can handle the huge volume of data in real-time to determine the most effective ad placement strategies.

B. Driving Forces and Enabling Technologies toward Big Data Analytics in Power Distribution Systems

Lower data storage and data collection cost in the power distribution system are the two major driving forces toward big data analytics. The rapidly falling cost of digital storage has spearheaded most industries into the big data era. The advanced sensor system such as AMI and the corresponding communication network is pushing down the data collection cost significantly in the distribution systems. For example, the smart meters with a built in two-way communication system developed in the past 5 to 10 years have allowed easier information sharing between utilities and customers. Both Federal sponsored programs and market forces are facilitating the wide-spread adoption of AMI in the U.S. In addition, because of Smart Grid initiatives, the number and variety of distribution devices and equipment that needs to be monitored and controlled continues to increase. This is mostly driven by state and federal policies that incentivize adoption of distributed generation, energy storage, electric vehicle, direct load control, and distribution system automation systems. As of today, the amount of customer and equipment information being collected in the power distribution systems had utterly swamped the traditional tools used for processing them. In summary, the system operator and planner are now facing a heterogeneous and complex big data set that is growing exponentially. This calls for a dramatic change in the way we handle and analyze the information in the distribution systems.

C. Value Proposition of Big Data Applications in Power Distribution Systems

Utilities use data collected from field devices to inform decisions for various applications including grid operations, electric system planning, wholesale market participation, and policy discussions. Historically the field devices have been capable of power monitoring, and high costs have limited the deployment to larger infrastructure including transmission and distribution substations. It is rare for a utility pre-AMI deployment to have visibility beyond the distribution substation other than monthly billing data. Limited visibility becomes an even larger issue when trying to predict the changes happening at the distribution level such as solar photovoltaic (PV) and electric vehicle (EV) adoption. The amount of data becomes prohibitive to analyze using ubiquitous software such as Microsoft Excel and must be analyzed in a big data environment. Once this challenge is overcome there is an opportunity for utilities to improve decision making. The impacts to utility business operations include: improved equipment diagnosis, state estimation and distribution grid visualization, more granular load and renewable forecasting, device failure prediction, and comprehensive policy impact analysis.

II. Big Data Applications in Distribution Systems

Big data analytics could be applied to improve both short-term distribution system operations and long-term distribution system planning processes. With big data, the sum is always more valuable than parts of the data set. Studying different subsets of the complex distribution system data set leads us to distinct applications. This section proposes promising big data applications for both short-term operations and long-term planning studies. The applications include detection of energy theft, customer consumption behavior modeling, spatial load and renewable forecast, distribution system visualization, state estimation, and distribution system planning.

A. Application for Short-Term System Operations

Detect Energy Theft: The primary use of interval meter data is automation in billing and settlement. Usually, both customer-based electronic meter and concentrator meters are installed in the distribution system. The concentrator meters aggregate and track data from multiple customer-based meters. By analyzing an avalanche of paired data from concentrator meters and customer-based meters, irregular energy loss patterns could be easily identified [4]. If we add in historical energy theft information, it is possible to further filter out reasonable changes in energy consumption trends and detect potential energy theft.

Detect EV and Rooftop Solar Integration: In the past, residential customer load profiles have been dependent upon larger loads, which have been heating & cooling systems. With changes to these characteristic load shapes from new technologies such as electric vehicles and rooftop solar PV there is potential to detect when a customer adopts such a system as well as system performance. Power flow returned to the grid is an easy identifier of a photovoltaic system installation and the magnitude (although the signal is mixed with load) can be identified by recognizing the season and expected load of the customer. Dependent on customer behavior, the EV detection would use fuzzy logic to identify a characteristic change in baseload. Tracking these two adoption trends is critical for structuring power purchase agreements, planning infrastructure upgrades, and informing state policies.

Develop More Granular Load Forecast: The availability of interval meter data also creates the opportunity to develop more accurate and granular load forecast in terms of both location and time. More accurate forecasts will benefit both transmission and distribution system operations. A more accurate spatial forecast in the transmission system has huge impact on the unit commitment and dispatch process. A reduction of 1% in mean absolute percentage error (MAPE) could decrease annual variable generation cost in the United States by approximately $160 million [5]. Currently, the utilities typically submit aggregate load forecast in their service territories to the market/system operator. The market operator would then disaggregate the forecast to the substation level by using weather forecast information and historical load distribution factors. With more granular customer consumption and behavior information, the utilities are in a great position to improve the spatial load forecast accuracy through mining both electricity consumption data and weather information.

Develop More Granular Renewable Generation Forecast: With rapid penetration of distributed renewable generation in distribution system, the need for accurate distributed renewable forecast becomes critical. For example, as distributed renewable (mostly solar photovoltaic) penetration levels in distribution circuits reach 15% and beyond in Hawaii and Southern California, the distributed generation starts to have significant impacts on distribution systems planning and operation. An accurate spatial joint load and rooftop solar generation forecast could greatly help distribution system operators manage circuit overloading, address reverse power flow and improve circuit voltage profile and power quality [6]. The challenge of producing granular renewable generation forecast lies in the lack of direct rooftop photovoltaic generation measurement as most of the residential solar panels do not have a separate meter. Hence, the solar generation has to be derived from historical load and net load measurements as well as irradiance data collected from local weather stations.

Visualization in Power Distribution Systems: Visualization of power distribution systems is critical in facilitating system operator's management of distribution systems. The difficulty of distribution system visualization arises from two sources the distributed nature of the data, and the scale of the data set. While the North American power transmission network can be represented by less than 100,000 buses, the number of customer (load node) in the distribution system well exceeds 100 million. In addition, the distribution system has a much larger number of operable equipment such as capacitors, voltage regulators, and transformers compared to the transmission system. To satisfy distribution system operation's real-time requirements in an interactive manner, the distribution system visualization system should be capable of explicitly trading the speed of display against accuracy. In addition, to handle the distributed nature of the data, new database structures and mathematical methodologies such as dimension reduction need to be developed.

State Estimation: Grid operators are dependent on state estimation to predict the impacts of their decisions in day-to-day operations. State estimation is the application of power flow equations, data from field monitors, and heuristics to measure grid conditions. The applications that rely on state estimation for performing analysis of the distribution grid are numerous and include overloading prevention, outage management, DER dispatch, and integration with transmission operations [7]. With the introduction of AMI data the state estimation algorithms can reduce error by allocating load correctly and thus improving the decision-making capability of grid operators.

Equipment Diagnosis: Electric utilities maintain billion dollar asset bases of infrastructure and these systems require regular maintenance. A sample of 34 parent utility companies will require $70 billion in capital expenditures in 2013 [8]. To maximize effective replacement of these systems AMI data can be leveraged to predict failure of distribution transformers, underground cable, overhead lines, and voltage regulation devices. Transformer life can be predicted by observing loading patterns throughout the year. The degradation of transformer life is non-linear as the insulation of the transformer breaks down quicker at higher temperatures [9].

B. *Application for Long-Term System Planning Studies*

Model Customer Consumption Behavior under Various Incentive and Pricing Structures: Another use of interval meter data is to support new retail pricing mechanisms such as time-of-use pricing and critical-peak pricing. With the load metered at 15-minute intervals, customer's response to changes in the transmission system conditions and wholesale electricity prices can be accurately measured and rewarded. Big data analytics could be applied to quantify the benefits of new retail rate design and measure the success of various demand response programs. Using the same data set, researchers could model and study the differences in customer's electric energy consumption behavior under different pricing mechanisms and incentive programs. The analysis results will provide helpful insights on how to quantify, extract and aggregate load flexibility. These insights will also facilitate the utilities' and demand response aggregators' efforts to refine and tailor electricity retail pricing structures and demand response programs.

Transformation of Distribution System Planning Process: Current distribution planning processes for electric utilities already rely upon large data sets. The planning process typically requires a considerable amount of man-hours and a dedicated department for electric system planning. Integrating AMI data to existing planning processes would be of great benefit for three primary reasons. 1) Electric system planning departments rely upon a single metered data point for a single distribution circuit. This singular point is the basis for infrastructure upgrades. AMI data provides extra metering points for spatial awareness of planning requirements. This helps the engineer not only maintain the primary substation gateway, but also plan for the radials at the end of the circuit. 2) Field devices often fail through physical degradation or unreliable communication mediums. If the device fails even for a moment during a critical loading condition the data can be erroneous. AMI data provides added redundancy in load estimation if the primary monitoring point was to fail. 3) Customer behaviors are changing with the adoption of EVs and solar PV. Typically weather regression is done to estimate load during a high loading weather driven condition or peak system condition [10]. As solar PV and EV adoption increases the customer behavior will change and AMI provides insight to the less predictable behavior of these devices. This helps the engineer plan more effectively when predicting load growth and weather sensitivity for a particular distribution circuit. Even with these added benefits it can be difficult to integrate this big data opportunity into the current planning process.

III. Technical Challenges in Developing Big Data Applications in Power Distribution Systems

The challenges in developing big data applications in power distribution system are two-fold: first, to design a flexible system architecture that accommodates and optimizes big data analytic workloads, and second, to develop scalable mathematical tools capable of processing distributed data.

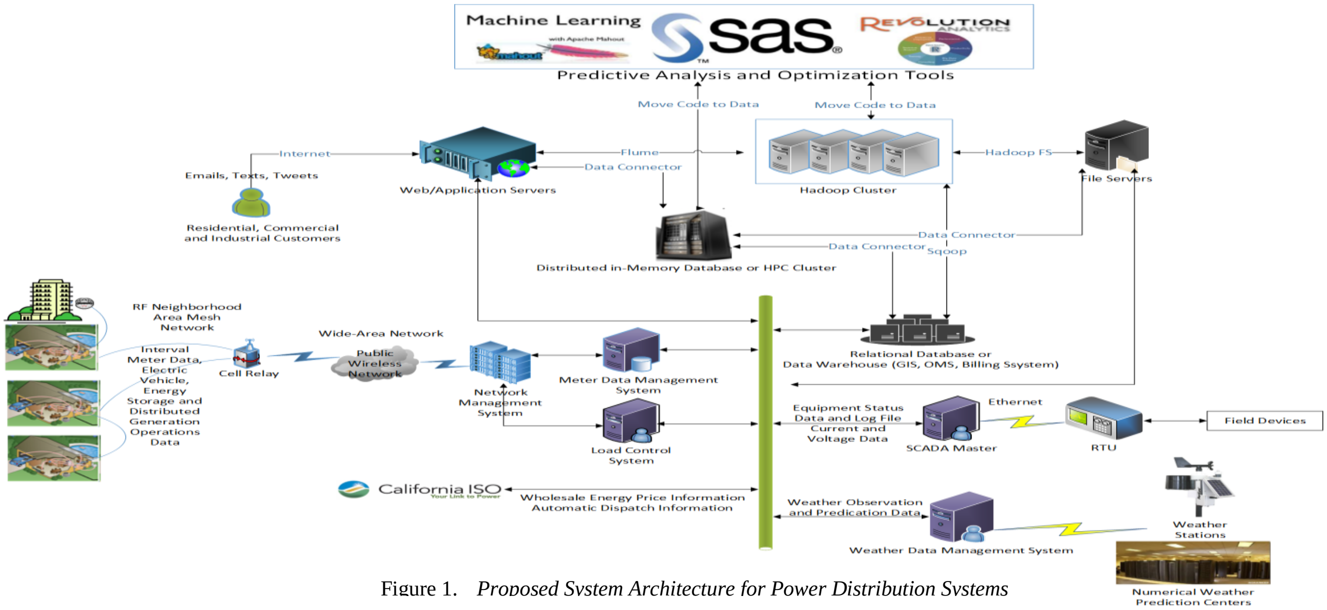
Machine Learning
with Apache Mahout
sas
Revolution Analytics
Predictive Analysis and Optimization Tools
Move Code to Data
Move Code to Data
Internet
Emails, Texts, Tweets
Residential, Commercial and Industrial Customers
Web/Application Servers
Flume
Data Connector
Hadoop Cluster
Hadoop FS
File Servers
Data Connector
Data Connector
Sqoop
Distributed in-Memory Database or HPC Cluster
RF Neighborhood Area Mesh Network
Interval Meter Data, Electric Vehicle, Energy Storage and Distributed Generation Operations Data
Cell Relay
Wide-Area Network
Public Wireless Network
Network Management System
Meter Data Management System
Load Control System
Relational Database or Data Warehouse (GIS, OMS, Billing Ssystem)
Equipment Status Data and Log File Current and Voltage Data
Ethernet
SCADA Master
RTU
Field Devices
California ISO
Your Link to Power
Wholesale Energy Price Information
Automatic Dispatch Information
Weather Observation and Predication Data
Weather Data Management System
Weather Stations
Numerical Weather Prediction Centers

Figure 1. *Proposed System Architecture for Power Distribution Systems*

In this paper, we propose a flexible system architecture design capable of handling heterogeneous workloads with structured, semi-structured and unstructured data from various sources. In addition, scalable mathematical tools such as machine learning, dimension reduction and streaming data analysis that are applicable to the unique complex data sets in power distribution systems are discussed in detail.

A. *Data Science Challenge*

Data collected by power systems suffer from three primary issues. 1) They are incomplete in nature. 2) They are heterogeneous and therefore difficult to merge. 3) Systems update or make their data available at different rates. Since the electric grid is operating continuously there is monitor failure during grid operation because of degradation or communication breakdown. Heterogeneity in power system data exists because often the data was intended for a specific application and not collected for a holistic purpose. For example grid operations through the SCADA system operate in the four second range whereas AMI systems collect data in the minute to hourly range. Data mining and heuristics can be used to correct these problems, but must be applied carefully.

B. *Mathematical Challenges*

Traditional mathematical methods are not adequate in handling petabyte scale, high-dimensional, distributed data sets in situ. To address high-dimensionality, machine learning, statistics and optimization algorithms such as classification, clustering, sampling, and linear/nonlinear optimization algorithms need to be easily scalable. Alternatively, scalable and flexible dimension reduction techniques are needed to extract latent features and relevant subsets while balancing accuracy and degree of reduction according to user specification [11].

Large scale data sets in power distribution systems tend to be inherently heterogeneous and distributed. To efficiently analyze the distributed data, the algorithms need to come to the data rather than moving data set to the algorithm. For example, many mathematically algorithms could be modified to accommodate parallelization such as Map-Reduce system. The main idea is to first divide a job into many tasks that run in parallel and only access local information and then shuffle and sort individual task results, and reduce intermediate outcome into final results.

With high velocity of streaming and spatial distribution system data, near real-time analysis and control applications are needed such as anomaly detection and emergency control. In addition, it may not be feasible or economical to storage all raw data and process them later. Therefore, it is desirable to develop algorithms that can operate with a single pass though the data [11].

C. *Proposed System Architecture for Power Distribution Systems Analytics*

The major classes of data in the proposed system architecture are 1) customer data measured using smart meters, 2) grid data measured on the distribution and transmission network using SCADA, 3) market data such as prices, and 4) other data such as weather, macro-economic or publicly collected census and text, tweets data. As shown in Figure 1, through various types of communication networks, the heterogeneous and complex data sets are transmitted and stored in traditional relational database, data warehouse, web servers, application servers and file servers. These data sets are then loaded into Hadoop clusters and/or distributed in-memory databases depending on the size, usefulness of the data, and importance of downstream applications. The Hadoop cluster runs on commodity hardware and provides a cost effective solution to analyze big data sets. The in-memory database or HPC cluster are typically more expensive yet powerful proprietary offerings from vendors. To perform predictive analytics, optimization and control, code developed in analytical tools such as SAS, Revolution R and Mahout will be moved to Hadoop cluster or HPC cluster to perform short-term operations and long-term planning studies.

IV. Risks in Deploying Big Data Applications in Power Distribution Systems

One of the most critical risks in adopting big data applications in the power distribution system is poor quality data impacting the decision-making without the knowledge of the operator. This could happen if critical sensors failures are not detected by the data cleansing routines and the state estimation applications. Another critical risk involves data privacy and data protection. For instance, leaked interval meter data could be exploited to identify when people are not at home, whether they own an electric vehicle and/or rooftop solar panel, or in some cases find out the brand of appliance someone is using. The cornerstone privacy principle “notice and consent” is no longer applicable in the big data era. Hence, to avoid backlash from customers about big data analytics, not only do we need advanced cyber-security but also enhanced laws and regulations to protect data privacy in electric utilities industry.

A. *Privacy and Information Security Issues*

Power System and Other System Operation Data: While using big data there is the potential to identify patterns and draw conclusions that can be used to harm the grid. There are two agencies responsible for settings standards to protect infrastructure data. The North American Electric Reliability Corporation (NERC) is the regulatory agency responsible for reliability of the bulk power system in North America. NERC has set forward Critical Infrastructure Protection (CIP) standards to protect the public release and potential abuse of assets considered critical [12]. The other agency is the Federal Energy Regulatory Commission (FERC) and has defined Critical Energy Infrastructure Information (CEII) as specific engineering information about proposed or existing critical infrastructure [13]. This includes virtual systems that can be considered critical. Utilities currently adhere to these standards and if handled properly can mitigate any risk proposed from assembling big data applications.

Customer Data: Local government entities, researchers, state and federal agencies and third parties need to access customer electricity usage and usage-related data to advance smart grid related research and energy policies. However, there are many privacy and information security issues related to use of customer electricity usage data. Without appropriate aggregation or anonymization, mining of raw smart meter data combined with other customer related data may allow re-identification of the data with individuals or individual firms and entities. In addition, a breach of the information security system of electrical utilities or a third-party granted data access authority, could lead to unauthorized disclosure, use and modification of a customer’s unencrypted electrical consumption data.

B. *How to Control Privacy Risks in Deployment of Big Data Analytics in Power Distribution Systems*

An array of procedures should be considered to control privacy risks associated with analyzing customer data. First, notice and consent process needs to be established to grant various third parties authority to access customer usage data for specified purpose and not secondary use. Second, anonymize usage data sets by removing identifiable characteristics and information, such as, but not limited to, name, address, and account number based on use cases. Third, adopt a data aggregation standard to prevent identification of data on individual. Fourth, develop non-disclosure agreements with data security protocols to regulate provision of customer data to eligible participants. Similarly, proper security protocol and standards need to be set up for different types of customer data.

V. Conclusions

This paper advances the field of big data analytics in power distribution systems by addressing three major obstacles. Firstly, we not only propose big data applications in distribution systems but also revealed their immense value in the system planning and operation processes. Secondly, to handle heterogeneous big data analytic workloads, a flexible system architecture design that seamlessly integrates Hadoop cluster, in-memory Database, existing relational databases, file servers and web servers is proposed and developed. At last, data privacy and data protection issues involving big data applications in distribution system are tackled.

References

[1] "Smart electric meters, Advanced metering infrastructure, and meter communications: Global market analysis and forecasts," Navigant Research, 2013. Available:www.navigantresearch.com/research/smart-meters

[2] "Power from big data – Are Europe’s utilities ready for the age of data?" Economist Intelligence Unit, 2013.

[3] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live Work, and Think*, Eamon Dolan/Houghton Mifflin, 2013.

[4] "Combatting energy theft with the smart grid," KEMA, Available: www.dnvkema.com/Images/Energy%20Theft-D1V4.pdf.

[5] B. Hobbs, S. Jitprapaikulsarn, S. Konda, V. Chankong, K. Loparo, and D. Maratukulam, "Analysis of the value for unit commitment of improved load forecasts," *IEEE Transactions on Power Systems*, vol.14, no.4, pp.1342,1348, Nov 1999.

[6] "Capacity of distribution feeders for hosting distributed energy resources (DER)," WG C6.24, No 275, August 2014, ELECTRA.

[7] M. Baran and T. McDermott, "Distribution system state estimation using AMI data," *Power Systems Conference and Exposition*, pp. 1-3, Mar, 2009.

[8] "Moody’s: Capital expenditures of US regulated utilities to fall in 2015 and Beyond," Global Credit Research, Oct, 2013. Available:https://www.moodys.com/research/Moodys-Capital-expenditures-of-US-regulated-utilities-to-fall-in--PR_284956.

[9] "Life management techniques for power transformer," CIGRE WG A2.18, Jan, 2003.

[10] "Transmission and distribution planning criteria and guidelines," Long Island Power Authority, Sep, 2010. Available: http://www.lipower.org-/pdfs/company/projects/energyplan10/energyplan10-e6.pdf.

[11] Mathematics for Analysis of Petascale Data, Report on a Department of Energy Workshop, Jun, 2008. Available:http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/Peta_scaled_at_a_workshop_report.pdf

[12] http://www.nerc.com/AboutNERC/Pages/default.aspx

[13] http://www.ferc.gov/legal/ceii-foia/ceii.asp