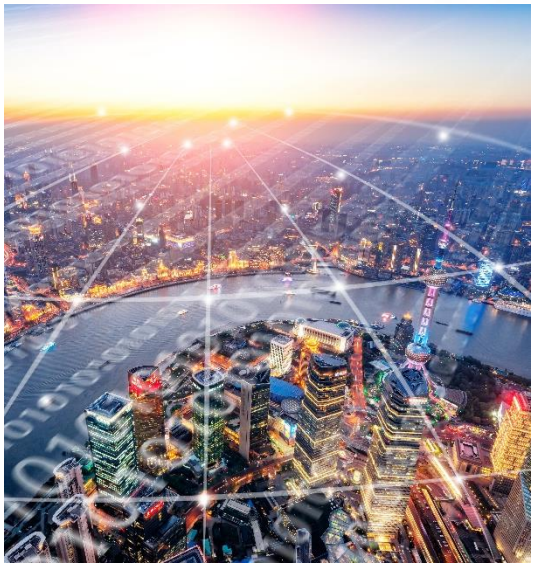
EAC

Electricity Advisory Committee

Big Data Analytics:

Recommendations for the U.S. Department of Energy

A Report by
the Electricity Advisory Committee
February 2021



Big Data Analytics: Recommendations for the U.S. Department of Energy

Introduction

The electric industry sector is facing an "explosion" of data coming from a variety of sources. Some examples include field measurements (e.g., smart meters, synchrophasors, smart sensors), weather measurements (ground stations, radar, satellite, and specialized systems such as the National Lightning Detection Network), asset monitoring (embedded sensors for condition-based monitoring), distributed generation data, data about electric vehicle charging, customer-driven data (e.g., Internet of Things devices, smart meters, demand response devices), and other important data sources for outage management (animal migration, vegetation management, fire detection, and water and gas management). Such data, often called "big data," contains valuable information to improve reliability, cost-effectiveness, energy efficiency, operations, planning, and asset and outage management, as well as the customer experience and grid resilience. This is amplified in a rapidly changing market environment that is transitioning to higher levels of distributed energy resources and the challenges that creates.

The use of big data analytics brings a host of practical challenges associated with insufficient training to perform data collection, curation, cleansing, and feature extraction, as well as challenges around data management policies, including privacy, confidentiality, and security. The market offering of big data analytics products and related services for the utility industry is rather modest, with limited electricity domain support expertise. In addressing such issues, electric sector stakeholders are facing an emerging need to define how to capitalize on big data, as well as information and communication technologies automation to create new business opportunities and manage market-driven change. At the same time, they face challenges in developing a justification for relevant investments and developing a business model to recover costs. There needs to be a greater incentive for the commercial and industrial segment to care about leveraging big data.

Approach

The Electricity Advisory Committee's (EAC) Smart Grid Subcommittee held a series of panel sessions on issues related to big data analytics during the EAC's March and October 2019 meetings and February and October 2020 meetings. Each panel focused on different aspects of big data analytics and engaged different stakeholder groups. The findings from these panel sessions are presented in the section below, followed by recommendations that are drawn from these findings. Summaries of the relevant meetings and associated presentations can be found on the EAC website here.

Findings

The findings presented here derive from the following nine priorities, with more detail provided below.

1. The North American Energy Resilience Model is being developed with substantial reliance on grid data. The relevant data models[1] and associated analytics are yet to be fully defined.
2. The use cases that justify large investments in data analytics infrastructure need to be demonstrated to ensure clear returns on investments.
3. Big data research is in its infancy in the electric utility industry due to lack of resources and expertise, while in other industries it is developing by leaps and bounds. The U.S. Department of Energy's (DOE) research funding will be needed to move the broader utility ecosystem forward.
4. If big data is not readily available, and not accompanied by power system data, the opportunity for advanced data analytics is dramatically reduced and often made infeasible.
5. For data to become available, it must be shared among many parties. Cybersecurity, confidentiality, ownership, and privacy concerns need to be resolved for the sharing to be feasible.
6. Once made available, big data needs to be judiciously integrated to allow for cost-effective management and utilization. Such solutions are not readily available in the utility industry.
7. Data analytics implementation requires expertise and skills currently not widely available in the industry. This shortfall needs to be overcome through expanded education and training.
8. To implement big data analytics, the design of legacy utility solutions that handle large amounts of data needs to be fundamentally changed.
9. The cost recovery model must be considered so that justifiable investments in data analytics can be considered as capital investments with clearly defined customer benefits.

To facilitate future improvements, the details of the findings are grouped in three areas:

Area 1: Existing DOE Big Data Efforts

- There is a lack of awareness in the industry about the research DOE conducts related to grid analytics. Sharing research results from DOE-funded projects and making DOE expertise available for others to leverage are desirable but need to be further articulated.

- DOE could assist by providing data analytics on the data that is already available. Automating data analytics would save operators time and increase overall efficiency. This requires the sharing of data among all interested parties.

- It is widely acknowledged that data analytics are an indispensable part of the North American Energy Resilience Model and will provide the capability to simulate and track electric phenomena, cascading faults, or oscillations as they occur in real time. To be utilized by DOE from a national security standpoint, the model will need to incorporate all data that the federal government is engineering and using in the weapons labs and by the U.S. Department of Defense.

Area 2: Data, Data Availability, and Data Requirements

[1] A data model is an abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities.

- The need for more research is evident, and it should address privacy, confidentiality, and security as major issues. Research needs to be focused on cyber-physical systems and data models.

- Data ownership and privacy are key issues. Restrictions on sharing utility and customer data can be an impediment for future research, and best practices or industry standards are needed. An approach to anonymizing data is also needed.

- Cybersecurity is an issue. Utilities invest a significant amount of money on data security. Sharing data with a third party introduces risks. Partnerships between the private sector and government agencies are needed to resolve the issues of secure data sharing.

- Making operational data available for use while maintaining security is a challenge. The integrity of operational data is addressed by the North American Electric Reliability Corporation's (NERC) Critical Infrastructure Protection requirements.

- Putting data in context is imperative; otherwise, operators face the problem of being data rich but information poor. As an example, synchrophasors provide the means to know how the system is operating so proper action can be taken if necessary. This requires data models and also physical grid models to take full advantage of big data analytics.

- Data comes from a variety of sources, both inside and outside of utilities. Often, one dataset[2] is not sufficient to perform data analytics. Many are needed, and data should be automatically integrated into prediction analytics. Such automated means do not readily exist in the industry.

- Often, the vast majority of the time and effort associated with using big data is spent on processing and preparing it, while considerably less time is spent on developing analytics. Preparing large amounts of data is expensive, and thus the problems that big data may address need to be big enough to provide a reasonable return on investment. The challenge of defining use cases has not been resolved.

- The EAC and broader electric industry sector is interested in DOE's role in data curation and hosting. DOE could create requirements for data quality and promote standards for interoperability along with communication network capacity requirements. DOE has developed several data repositories, including with synthetic datasets. The differentiation of the value of synthetic networks for certain data analytics studies needs to be made.

- NERC currently monitors system resilience using a substantial amount of data. This first step will focus on national security; however, many capabilities from the work DOE is doing will translate to the sector more broadly.

- There is certain data that the industry cannot share with the public. DOE has increased data-sharing capabilities in the Western Interconnection and is increasing coordination efforts across the eastern United States as well. However, data sharing remains a challenge.

[2] A dataset is a collection of data records.

- Data storage is another concern since data can accumulate quickly. Some can be stored on the cloud, and sometimes this is more cost effective. Some utilities may not be concerned about the security of the data stored on the cloud; however, some still are. Such concerns need to be addressed.

- There are challenges regarding the feasibility of integrating data from outside sources into a utility's dataset. Identifying and eliminating gaps and bad data within datasets are additional challenges.

- Data alone cannot explain why equipment fails or system events occur. The physical model of power systems is just as important as the data model to understand the causal relationship between the two models. There is a spatiotemporal dimension to the issue as well, since the risks to and impacts on grid infrastructure vary across parts of the system and across time.

- The need for a common vocabulary when making decisions around using big data is paramount because solution providers and users may have different backgrounds and expertise.

- The volume of data is growing rapidly. For example, integrating smart meters and synchrophasor data with supervisory control and data acquisition (SCADA) systems will lead to data volumes that are 100 times larger. If information about the state of the physical system is not time synchronized, root cause analysis may be impossible. The industry requires different levels of data for different use cases. Another example that can add millions of data points is the use of inverter-connected devices (e.g., solar panels) at the grid edge.

Area 3: Data Analytics Capabilities, Use Cases, and Cost Recovery

- The existing architecture for systems supporting utility monitoring, and control and protection functions—including energy management systems (EMS) for transmission, advanced distribution management systems (ADMS), synchrophasor systems, advanced smart metering platforms, asset and outage management solutions, and so forth—is several decades old.

- The legacy EMS, ADMS, market management systems, asset management systems, and back-office systems are not designed with big data analytics requirements in mind. The cost recovery approach for updating such systems to accommodate big data and related analytics needs to rapidly evolve (e.g., capital vs. operations and maintenance expenses).

- Data analytics, particularly focused on big data, is an emerging area. Training and preparation that enables employees to extract value from data analytics are lacking. Such training needs to include academia, as well as consultants, vendors, and the National Laboratories.

- Becoming a utility engineer with a specialization in data analytics is a niche job, and it is an important priority for utilities and other stakeholders to attract the next generation of talent.

- Most utilities do not have their own analytics department, and there is a need to collaborate to determine whether algorithms are reaching the right conclusions. There is a gap between utilities that know their business and data analysts who know the algorithms. All industry

segments would benefit from working together so that algorithms and datasets are not duplicated. Large-scale testbeds are needed to evaluate various solutions.

- From a consumer's point of view, a value proposition that can be asked is "What can consumers get out of and in exchange for their data and other grid data?" Consumers are part of the grid and generate, receive, and utilize data; thus, they can leverage data to meet the needs of the grid, as well as their own needs.

- Electricity consumers own the data collected from smart meters monitoring their energy use; however, it is not clear how revenue streams associated with the monetization of consumer data are allocated. It also is not clear how the cost recovery for data analytics based on smart meter data should be regulated.

- While installation of a single phasor measurement unit (PMU) has relatively low initial costs, there are high maintenance costs. DOE has played a role in pushing down certain capital costs by offering subsidies for PMU installation. However, it is not clear how future cost recovery for synchrophasor data analytics should be regulated.

- Monetization of data is an inevitable step in achieving widespread deployment of data analytics. Licensing approaches for the use of data with mutual benefits for the provider and the user need to be developed.

Recommendations

Recommendations are based on the panels and discussions among EAC members, panel participants, and DOE staff. Recommendations have been prioritized into three areas: (1) Existing DOE Big Data Efforts; (2) Data, Data Availability, and Data Requirements; and (3) Data Analytics Capabilities, Use Cases, and Cost Recovery.

Area 1: Existing DOE Big Data Efforts

1. Grid Resiliency and Modeling Capabilities: DOE's efforts to create data models for grid monitoring and identification of natural and adversarial threats are highly encouraged, and DOE should utilize both physical and data models that can be correlated for improved root cause analysis capabilities. The traditional utility space should leverage best practices in data analytics based on artificial intelligence (AI) and machine learning.

An example use case would be analyzing outage occurrences in both general and catastrophic storm conditions to identify areas highly subject to outages. This could be combined with other analysis, such as vegetation and asset health. A second use case would be using PMU data across a small area or larger region in conjunction with SCADA and other data to assess the impact of system anomalies as the strength (fault current availability) of the system changes with increased inverter-based generation sources.

2. End-User Needs: DOE should fund large data analytics testbeds with utility, university, and vendor participation, where advanced apps can be demonstrated and evaluated when implemented on data analytics platforms.

An example use case would be analyzing data collected on circuits with equipment asset health and locations to predict asset failures before an event causes an outage. The definition of "equipment" could be broadened to include non-serialized items such as arrestors, cutouts, insulators, and so forth.

3. Big Data R&D Efforts and Outreach: DOE should facilitate collaboration among the National Laboratories, academia, regional transmission operators (RTOs), vendors, and utilities to develop big data analytics that will lead to advanced methodologies, data structures[3], and use cases that will, in turn, lead to higher reliability, resiliency, security, and efficiency of the electric system.

An example would be the formation of small, collaborative teams, with each team including one or more National Laboratories, universities, RTOs, vendors, and utilities that will first develop a use case with expected outcomes, then develop analytics to support that use case. The teams should coordinate to ensure that a broad range of use cases are covered and to minimize duplication. The developments may have to be RTO- or utility-specific early on but should allow for adoption of the methodology by other RTOs, utilities, or third parties with a minimal need for data structure changes.

Area 2: Data, Data Availability, and Data Requirements

4. Availability of Datasets: DOE should continue funding projects that will generate synthetic data, as well as field demonstrations with utility participation that will use actual utility data. DOE should make available publicly accessible datasets to assist in advancing algorithmic developments.

An example of data that should be shared with the research community is the data to be collected through DOE's Grid Modernization Laboratory Consortium sensors program, and data collected by Oak Ridge National Laboratory's Environment for Analysis of Geo-Located Energy Information (EAGLE-I) real-time situational analysis tool.

5. Cybersecurity: DOE should explore data availability as a means to prevent, detect, and predict cybersecurity threats using non-traditional approaches with AI and machine learning.

An example would be using PMU data in conjunction with other data to detect signatures of concern in a predictive manner. Such an approach to identify and monitor critical infrastructure in a major metropolitan area would facilitate the detection of multiple intrusions in a specific time window to alert authorities.

6. Data Privacy and Confidentiality: DOE should find an effective way to address the anonymization of data, which is an indispensable enabler of data analytics deployment.

An example would be utilizing a customer metadata approach to analyze energy efficiency and customer benefits.

7. Database Development: DOE should (1) examine which databases owned and maintained by the government have value in developing utility data analytics applications, and (2) have these databases

[3] Data structure is a data organization, management, and storage format that enables efficient access and modification.

offer pre-processed data for automated support of data analytics. The tools developed, including those to fill data gaps, should be made readily available to others.

> An example is data offered by the National Oceanic and Atmospheric Administration, which contains relevant weather forecasts for the analysis of weather impacts on grid resilience. Such data is hard to analyze and collect for grid data analytics striving for real-time applications.

Area 3: Data Analytics Capabilities, Use Cases, and Cost Recovery

8. Workforce Development: DOE should team with universities and others in the education sector to offer the services and training needed to expand the number of data analytics experts in the utility and other sectors.

> An example would be the creation of webinar series, tutorials, and short courses with a special emphasis on data analytics and big data utilization in different applications of the electricity sector.

9. Next-Generation Control Systems: DOE should continue examining the future needs for these systems and engage in defining the next generation of control systems.

> An example would be a study to define the next generation of EMS/ADMS systems that will merge SCADA and synchrophasor technology by having PMUs as substation/feeder remote terminal units, and elaborate substation-based processing capabilities. Other substation and line data may be included as well.

10. Cost Recovery Issues: DOE should extend its services to state and federal agencies to enhance the understanding of big data analytics and facilitate valuation. DOE should also coordinate with stakeholders to address cost models for future deployment of data analytics. Such efforts should consider the impact of revenue streams associated with big data monetization.

> An example would be a cost model for predicting outages that incorporates data model-based maintenance and asset management capabilities that can directly improve customer reliability indicators, such as the System Average Interruption Duration Index and System Average Interruption Frequency Index.

Conclusion

While “big data” and “data analytics” are expansive terms, through its 10 recommendations above, the EAC is proposing a concerted effort for DOE to focus the resources and capabilities at its disposal to provide value for many stakeholders in the energy space. The greatest way to add value may be to cross-pollinate traditional utility and other sectors’ data analytics capabilities, such as the broader sectors of the Internet of Things, fintech, and others that employ data analytical techniques.