Big Data Analytics in Smart Grids

Filip FEDELEȘ, Ionuț ȚARANU

Data analytics are now playing a more important role in the modern industrial systems. Driven by the development of information and communication technology, an information layer is now added to the conventional electricity transmission and distribution network for data collection, storage and analysis with the help of wide installation of smart meters and sensors.

Big data has a potential to unlock novel groundbreaking opportunities in the power grid sector that enhances a multitude of technical, social, and economic gains. The currently untapped potential of applying the science of big data for better planning and operation of the power grid is a very challenging task and needs significant efforts all-around. As power grid technologies evolve in conjunction with measurement and communication technologies, this results in unprecedented amount of heterogeneous big data sets from diverse sources.

***Keywords**: big data analytics, smart grid*

1 Introduction

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today’s technology, it’s possible to analyze your data and get answers from it almost immediately – an effort that’s slower and less efficient with more traditional business intelligence solutions.

Big data is a term used to describe massive amounts of information (**Figure 1**) that frequently occurs in the form of unstructured data sets that cannot be analyzed with standard database software.

The energy industry has worked with big data for years, regularly processing significant amounts of information produced on an intra-hourly basis.

Markets settle on metered data that measures power in five-minute increments. Utilities use supervisory control and data acquisition (SCADA) systems. Investors and planners run models with full representation of each generating unit, transmission load flow and hourly dispatch. Although other industries are relatively new to big data, they are finding innovative ways to use it. Applying these innovations to the energy industry promises to be transformative.

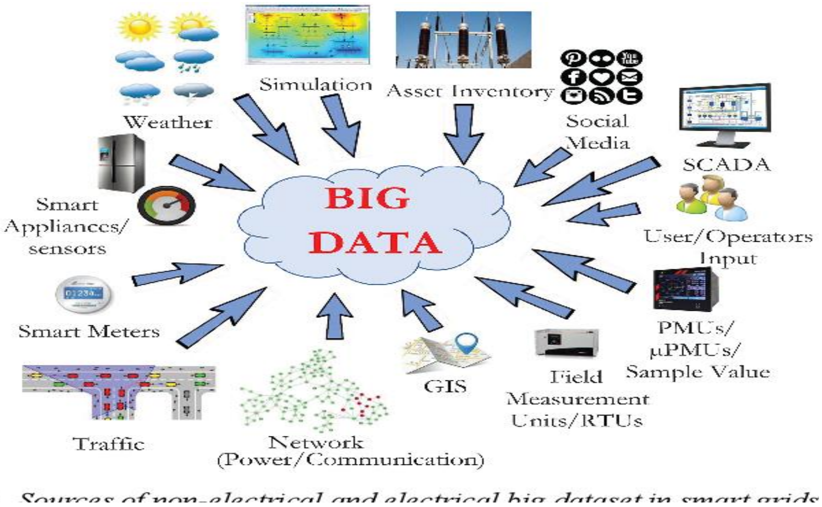
Weather
Simulation
Asset Inventory
Social Media
SCADA
Smart Appliances/ sensors
BIG DATA
User/Operators Input
Smart Meters
PMUs/ µPMUs/ Sample Value
GIS
Field Measurement Units/RTUs
Traffic
Network (Power/Communication)
1.1 Sources of non-electrical and electrical big dataset in smart grids

Fig. 1. Big Data ecosystem

2. Characteristics of Big Data

Big Data refers to the large, diverse sets of information (**Figure 2**) that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered. Big Data often comes from multiple sources and arrives in multiple formats.

Big Data can be categorized as unstructured or structured. Structured data consists of information already managed by the organization in databases and spreadsheets; it is frequently numeric in nature. Unstructured data is information that is unorganized and does not fall into a pre-determined model or format. It includes data gathered from social media sources, which help institutions gather information on customer needs.

The presence of sensors and other inputs in smart devices allows for data to be gathered across a broad spectrum of situations and circumstances.

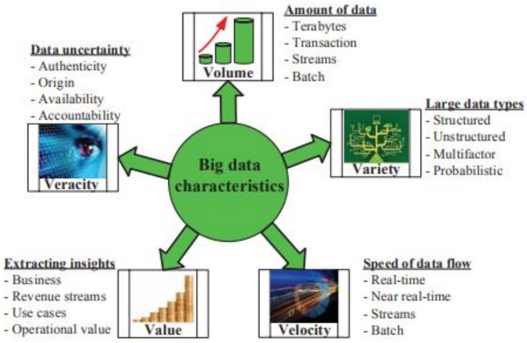

Fig. 2. Big Data characteristics

3.Characteristics of Smart Grids

Smart Grids comprise a broad mix of technologies to ptimize electricity networks, extending from the end user to distribution and transmission.

Not only can better technologies for monitoring, control and automation stimulate the development of new business models, they can unlock system-wide benefits including reduced outages, shorter response times, deferral of investments to the grids themselves and distributed energy resource integration.

At the end-user level, smart grids can enable demand flexibility and consumer participation in the energy system, including through demand response, electric vehicle (EV) charging and self-produced distributed generation and storage.

Demand flexibility can increase the overall capacity of the system to integrate variable renewables while accelerating the electrification of heating, cooling and industry at a lower cost. Deploying a physical layer of smart-grid infrastructure – underpinned by smart meters – can help unlock these benefits.

Electric power plants are generally dispatched so that the plants with the lowest operating costs (baseload plants) come on first, followed by more expensive plants when load increases, and finally, the most expensive plants during times of peak load1 . Very little electricity is stored for future use because storage is typically too costly. For this reason the marginal cost of supplying electricity is much higher during times of peak load. However, most electricity consumers are charged the same price for every kWh they consume. This is economically inefficient as the prices consumers pay do not reflect the true costs of production. Advanced electricity pricing refers to a broad range of approaches and pricing programmes that try to make consumer prices more accurately reflect real-time production costs so that customers shift consumption toward times when electricity is less expensive. Advanced pricing can also shift consumption to times when RE is available. Three representative advanced pricing schemes are described further.

Electricity usage typically peaks around the same time every day in a given area. The simplest method of discouraging electricity use during peak times is to institute a time-of-use (TOU) price schedule, under which electricity is least expensive when loads are low (typically at night) and most expensive during peak times (usually afternoons).

Customers paying TOU rates may adjust loads manually or use building or home energy management systems (BEMS/HEMS) to control their loads. TOU pricing schemes may vary with the season but are generally set far in advance. This means TOU pricing does not help much on the few days per year when load approaches its annual peak. TOU pricing programmes are becoming common. TOU pricing is typically advantageous for solar PV, which produces power during the daytime, when the price is usually high.

Wind plant power forecasting has become a priority for grid operators as utility-scale wind plants have come to make up a significant portion of grid capacity in some areas. With wind penetrations around 25%, studies have shown that wind forecasting can save tens to hundreds of millions of dollars per year in operating costs over several states in the U.S. (Lew, et al., 2011). When NWP power forecasts for regional aggregations of wind plants are compared to actual wind power output for those aggregations, error rates of 5% are typical. Error rates for single locations are two to four times higher. Current day-ahead NWP error rates are not expected to drop significantly. Wind plants may also use very short-term (millisecond scale) wind nowcasting to optimize power output by dynamically adjusting the pitch of turbine blades (Madrigal, 2010). Light Detection and Ranging (LIDAR) and Sonic Detection and Ranging (SODAR) wind sensors located on turbines are used for this purpose. This technology is experimental.

4. Big Data applications in power distribution systems

The carbon emission reduction and sustainability of environment are the driving force and construction purpose of smart grid, which is designed in a decentralized structure. The employment of distributed generator units in modern power distribution system now provides an effective means for the utilization of widespread renewable energy such as wind and solar energy. These emerging microgrids are vital for the expectation of a low-carbon society. Moreover, the close distance between the generator and loads in microgrid improves the reliability of power delivery and reduces the power transmission loss. The ability to operate in an island mode also protects the load from damages caused by power system including voltage fluctuation, frequency deviation, etc.

Distribution automation (DA) is a concept of smart grid which focuses on the operation and system reliability at the distribution level. A successful DA has the capability to localize and isolate the faults in distribution system with a reduced restoration time and improved customer satisfaction. Under the concept of DA, increasing volume of operational data have been collected from supervisory control and data acquisition (SCADA) or advanced metering infrastructure (AMI) for state monitoring and fault diagnosis.

Thanks to the development of ICT technology in power systems, a huge volume of data can be collected via AMI and communication infrastructures. Power system operating data, weather information and log data of relay protection devices are processed as the input of a one class classification system, which is a data-driven model of fault phenomena based on a hybridization of evolutionary learning and clustering techniques. This fault recognition system is validated in the medium voltage power grid in Rome. The traditional statistical methods such as linear discriminant analysis (LDA) and logistic regression are discussed for mining the relation between power system faults and the features extracted from raw data.

Big data applications in distribution system planning can be divided into two categories

- Short term operations
- Long term planning studies

Short term applications are detection of energy theft, outage detection, peak load monitoring, customer consumption behavior modeling, special load and renewable forecast, distribution system visualization, state estimation and distribution system planning, in which the first three applications are qualified to be very short term applications. Applications in Long term system planning studies include modeling customer consumption behavior under various incentives and pricing structures, transformation of distribution system planning process.

5. Implementation on a cloud computing platform

Cloud computing can be deployed as the infrastructure layer for big data systems to meet certain infrastructure requirements, such as cost-effectiveness, improved accessibility, and scalability. Based on the requirements of the proposed framework, Infrastructure as a Service (IaaS) clouds are appropriate to use to implement the smart grid big data framework. Cloud service providers such as, Amazon AWS and Google can be utilized to build a cluster that will host the framework. In this implementation, a Google cloud platform cluster with six machines is used.

As smart grid data increases exponentially in the future, utilities must envision ever-increasing challenges on data storage, data processing, and data analytics. Even though many electric utilities have realized that deployment of big data analytics is a must and not a choice, for future business growth and efficient operation, implementation of big data analytics in utility framework is lagging. Therefore, there is a need of comprehensive study to investigate current challenges, value proposition to stakeholders (e.g., consumers, utilities, system operators), operational benefits, and potential path forward to deploy big data analytics in power grids.

The high volume data gather in smart grid is similar in size and characteristics to the concept of big data. Big data is defined as data with high volume, velocity, and variety. The sampling frequency from perception devices can make the data size very large. Data velocity reflects the required speed for collecting and processing the data. Hence, big data management and processing techniques (hardware, software, algorithms, AI, etc) can be borrowed and applied in the domain of IoT. In addition, some applications of smart grid can perform their tasks only at specific time a day, such as weather forecasting and one-day ahead of time energy distribution, which can be performed at the night of every day. However some other applications perform their tasks all day round, such as real-time applications that monitor the power grid components. This is needed to speed up energy outage recovery process and real-time response to emergent behaviors in power demands. Even with today’s development in big data processing techniques, managing of data in the smart power grid poses new challenges that are based upon the criticality of power systems, real-time response, proactive solutions, accurate predictions, and security. Hence, we address first the question of where to store the smart grid big data.

The increasing number in services and capabilities of cloud computing make it a good candidate to host SCADA systems. Cloud computing is a model that enables a convenient on-demand access to a shared pool of computing resources such as network, storage, servers, applications, and services. Cloud computing enterprises deliver their services to end users in three models namely, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS provides end users with operating systems, storage, network, and database services deployed within the cloud. PaaS provides end users with capabilities to deploy their applications such as programming languages and libraries that are available within the cloud. SaaS cloud provides a ready to use application for end users.

6. Key Challenges for Big Data Analytics

Table 1

Challenges	**Possible Impact**	**Potential Solution**
Data Volume	Need of increased storage and computing resources	Dimensionality reduction, Parallel computing, Edge computing, Cloud computing, pay-per use
Data Quality	Lack of complete information, misleading decision	Probabilistic and storchastic analysis, data cleaning (e.g. dealing with missing values, smooth out noises, outliers, and inconsistent data)
Data Security	Vulnerable to malacious attack, compromise consumer privacy and integrity, mislead operational decision and financial transactions	Data anonymization (e.g. data aggregation, data encryption, P2DA)
Time Synchronization	Mislead operational decision, wrong interpretation of data, bad diagnostic of past events	Synchronize devices based on same radio clocks or satellite receivers
Data Indexing	Computational complexity, long processing time	Deploy new indexing techniques such as R-trees, Btrees, Quad-trees
Value Proposition	Non-acceptance by stakeholder, delay deployment of big data	Quantifying both technical and economic values to key stakeholders, namely consumer, system operator, utility.
Standards and Regulation	Interface challenges among various computing, storage, and processing platforms, delayed deployment	Regulatory entity define guidelines about data sharing/exchange, and standards should technically ensure regulatory aspects

Data volume

Before we start to build any data processes, we need to know the data volume we are working with: what will be the data volume to start with, and what the data volume will be growing into. If the data size is always small, design and implementation can be much more straightforward and faster. If the data start with being large, or start with being small but will grow fast, the design needs to take performance optimization into consideration. The applications and processes that perform well for big data usually incur too much overhead for small data and cause adverse impact to slow down the process. On the other hand, an application designed for small data would take too long for big data to complete. In other words, an application or process should be designed differently for small data vs. big data.

This large amount of data exceeds the amount of data that can be stored and computed, as well as retrieved. The challenge is not so much the availability, but the management of this data. With statistics claiming that data would increase 6.6 times the distance between earth and moon by 2020, this is definitely a challenge.

Some of the newest ways developed to manage this data are a hybrid of relational databases combined with NoSQL databases. An example of this is MongoDB, which is an inherent part of the MEAN stack. There are also distributed computing systems like Hadoop to help manage Big Data volumes.

Data Quality

Veracity, one of the most overlooked Big Data characteristics, is directly related to data quality, as it refers to the inherent biases, noise and abnormality in data. Because of veracity, the data values might not be exact real values, rather they might be approximations. In other words, the data might have some inherent impreciseness and uncertainty. Besides data inaccuracies, Veracity also includes data consistency (defined by the statistical reliability of data) and data trustworthiness (based on data origin, data collection and processing methods, security infrastructure, etc.). These data quality issues in turn impact data integrity and data accountability.

While the other V's are relatively well-defined and can be easily measured, Veracity is a complex theoretical construct with no standard approach for measurement. In a way this reflects how complex the topic of "data quality" is within the Big Data context.

Data users and data providers are often different organizations with very different goals and operational procedures. Thus, it is no surprise that their notions of data quality are very different. In many cases, the data providers have no clue about the business use cases of data users (data providers might not even care about it, unless they are getting paid for the data). This disconnect between data source and data use is one of the prime reasons behind the data quality issues symbolized by Veracity.

Data veracity, in general, is how accurate or truthful a data set may be. In the context of big data, however, it takes on a bit more meaning. More specifically, when it comes to the accuracy of big data, it's not just the quality of the data itself but how trustworthy the data source, type, and processing of it is. Removing things like bias, abnormalities or inconsistencies, duplication, and volatility are just a few aspects that factor into improving the accuracy of big data.

Unfortunately, sometimes volatility isn't within our control. The volatility, sometimes referred to as another "V" of big data, is the rate of change and lifetime of the data. An example of highly volatile data includes social media, where sentiments and trending topics change quickly and often. Less volatile data would look something more like weather trends that change less frequently and are easier to predict and track.

The second side of data veracity entails ensuring the processing method of the actual data makes sense based on business needs and the output is pertinent to objectives. Obviously, this is especially important when incorporating primary market research with big data. Interpreting big data in the right way ensures results are relevant and actionable. Further, access to big data means you could spend months sorting through information without focus and a without a method of identifying what data points are relevant. As a result, data should be analyzed in a timely manner, as is difficult with big data, otherwise the insights would fail to be useful.

Data Security

When producing information for big data, organizations have to ensure they have the right balance between utility of the data and privacy. Before the data is stored it should be adequately anonymized, removing any unique identifier for a user. This in itself can be a security challenge as removing unique identifiers might not be enough to guarantee the data will remain anonymous. The anonymized data could be cross-referenced with other available data following de-anonymization techniques.

When storing the data, organizations will face the problem of encryption. Data can't be sent encrypted by the users if the cloud needs to perform operations over the data. A solution for this is to use "Fully Homomorphic Encryption" (FHE), which allows data stored in the cloud to perform operations over the encrypted data so new encrypted data will be created. When the data's decrypted, the results will be as if the operations were carried out over plain text data. So the cloud will be able to perform operations over encrypted data without knowledge of the underlying plain text data.

A significant challenge while using big data is establishing ownership of information. If the data's stored in the cloud, a trust boundary should be established between the data owners and the data storage owners.

Adequate access control mechanisms are key in protecting the data. Access control's traditionally been provided by operating systems or applications restricting access to the information - this typically exposes all the information if the system or application is hacked.

A better approach is to protect the information using encryption that only allows decryption if the entity trying to access the information is authorized by an access control policy.

An additional problem is that software commonly used to store big data, such as Hadoop, doesn't always come with user authentication by default. This makes the problem of access control worse, as a default installation would leave the information open to unauthenticated users.

Big data solutions often rely on traditional firewalls or implementations at the application layer to restrict access to the information. The main solution to ensuring data remains protected is the adequate use of encryption. For example, Attribute-Based Encryption can help in providing fine-grained access control of encrypted data.

Anonymizing the data's also important to making sure privacy concerns are addressed. It should be ensured that all sensitive information is removed from the set of records collected.

Real-time security monitoring is also a key security component for a big data project. It's important organizations monitor access to make sure there's no unauthorized access. It's also important threat intelligence is in place to guarantee more sophisticated attacks are detected and the organizations can react to threats accordingly.

For example, many big data solutions look for emergent patterns in real time, whereas data warehouses often focused on infrequent batch runs. How do these different usage models impact security issues and compliance risk?

In the past, large data sets were stored in highly structured relational databases. If you wanted to look for sensitive data such as health records of a patient, you knew exactly where to look and how to access the data.

Removing any identifiable information was also easier in relational databases. Big data makes this a more complex process, especially if the data is unstructured. Organizations will have to track down what pieces of information in their big data are sensitive and then carefully isolate this information to ensure compliance.

Another challenge with big data is that you can have a big variety of users each needing access to a particular subset of information. This means the encryption solution you choose to protect the data has to reflect this new reality. Access control to the data will also need to be more granular to ensure people can only access information they are authorized to see.

Conclusion

In this paper we explained every separate concept for big data, smart grid and cloud computing and how we can get all of them to work together for optimal end results.

We discussed the implementation of cloud energy storage devices, and cloud data storage mechanisms for the smart grid architecture. Using cloud computing applications, energy management techniques in smart grid can be evaluated within the cloud, instead of between the end-user's devices. This architecture gives more memory and storage to evaluate computing mechanism for energy management, and cost-

Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CCCDI – UEFISCDI, project title "Multi-layer aggregator solutions to facilitate optimum demand response and grid flexibility", contract number 71/2018, code: COFUND-ERANET-SMARTGRIDPLUS-SMART-MLA-1, within PNCDI III.

References

[1] H. Hu, Y. Wen, T.-S. Chua, X. Li, Toward scalable systems for big data analytics:
a technology tutorial, IEEE Access 2 (May) (2014) 652–687

[2] P. Mirowski, S. Chen, T. K. Ho, C.-N. Yu, Demand forecasting in smart grids, Bell Labs Tech.J. 18 (4) (2014) 135–158.

[3] Xinghuo Yu, C. Cecati, T. Dillon, Simões, The New Frontier of Smart Grids, , M.G., *IEEE Industrial Electronics Magazine* (2011)

[4] S. Callahan. Big data: The future of energy and utilities (https://www.rdmag.com/article/2015/10/big-data-future-energy-and-utilities)

[5] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, A break in the clouds: Towards a cloud definition, J. ACM SIG-COMM Comput. Commun. Rev., vol. 39, no. 1, pp. 50–55, (2014).

[6] P. V. Krishna, S. Misra, D. Joshi, and M. S. Obaidat, Learning automata based sentiment analysis for recommender system on cloud, J.IEEE Int. Conf. Comput, Inform Telecommun Syst., (2013), pp. 1–5.

[7] Yang Zhang, Tao Huang & Ettore Francesco Bompard, Big data analytics in smart grids: a review (https://energyinformatics.springeropen.com/articles/10.1186/s42162-018-0007-5)

[8] J.N. Bharothu, M. Sridhar, and R.S. Rao, "A literature survey report on Smart Grid technologies", Proc. 2014 International Conference on Smart Electric Grid (ISEG), pp. 1-8.

Ionuț ȚARANU (b. April 28, 1975) has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies in 1999. He followed a master's degree in Databases for Business Application, within the same faculty. As founder of STIMA SOFT, Ionut is an experienced professional in custom software development, with focus on implementation and management of complex software solutions.

Filip FEDELEȘ (b. November 29, 1982) has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies in 2006. He followed a master's degree in Public Management, within the same faculty. As a senior developer at STIMA SOFT, Filip is interested in data structures, new technologies and developing complex software solutions.