Getting ready for data analytics of electric power distribution systems

BENJAMÍN ZAYAS, ALFREDO ESPINOSA, VICKY SANCHEZ AND JAVIER PEREZ
Gerencia de Gestión Integral de Procesos
Instituto Nacional de Electricidad y Energías Limpias
Reforma 113, Col. Palmira, C.P. 62490
CUERNAVACA, MORELOS, MÉXICO
{zayas;aer}@iie.org.mx; vickysanglez@gmail.com; x.a.p.g_@hotmail.com; http://www.ineel.mx

Abstract: - The modernization of power utilities through the deployment of emergent technologies across the grid and advanced information systems are producing large amount of data that have to be managed with new approaches and technologies using existing and new analytical techniques. Data analysis is taking an important role in supporting decision making, big data technologies enable the analysis of data that was not possible using conventional systems for managing data. Analytical models have also evolved from descriptive to advanced analytics. Advanced analytics refer to the analysis of the data that produce results in which traditional business intelligence approaches are not appropriate. Therefore, an analytical platform has to be adopted by the utilities that better fit to the analytics needs in order to extract remarkable values for efficiently and effectively driven decisions while saving implementation cost.

Key-Words: - big data; analytics; power utilities; analytics models

1 Introduction

The advent of emerging technologies that enable the massive deployment of sensors and advanced systems to a lower cost is transforming the operation and decisions making of every industry. Data generation is growing every year by various sources and types of data; these are unstructured, semi-structured and structured data. Managing these data will be a fundamental activity for the power industry. The electric industry is moving to this direction through smart grid initiative. A smart grid is an electricity network that uses digital and emergent technologies to monitor and manage the operational processes to transport electricity from the generation sources to end-users. The modernization of the electricity infrastructure aims to optimize the operational capacity, integrate new applications and renewable energy sources in order to improve efficiency, reliability, safety and reduce environmental impact for economic, environmental and social benefits [1].

The modernization of the electric power system will produce a continuous flood of data from smart meters, server's logs machines, device status, user-system interaction and sensors [2]. This massive amount of data has to be collected, stored and analysed to get insights for driven decision making across the smart grid operational processes. In order to achieve this goal, the utilities must be prepared for this change by restructuring their fragmented legacy systems and adapting operation practices that limit data management and data analysis in a holistic, secure and interoperable manner. In this evolution, data become the most relevant asset of any power utility, which will equip utilities with remarkable value for data driven decision making if the data is used efficiently and effectively.

In this paper an approach for integrating an analytical platform for power utility distribution data is suggested, which combines conventional data managing systems for descriptive analytics and open-source big data software ecosystem for advanced analytics. This includes the major decisions regarding software selection, configuration of a platform, and demonstrations of some analytical applications.

2 A Big Data Analytics Solution

2.1 Big Data and Analytics

According to Stimmel [3], big data analytics are the application of techniques that are designed to reveal insights that facilitate the understanding, prediction, and exposition of hidden information to improve operational and business efficiency and to deliver real-world situational awareness. These techniques cannot be implemented using conventional system for managing data.

Although there is not a concrete definition of big data, most definitions refer to the applications that are able to collect and process a huge volume of heterogeneous data to a high velocity. Unlike traditional systems and techniques for managing relational data, big data applications are able to collect and process a higher volume of relational and non-relational data to a higher velocity in scalable commodity hardware [4 - 7]. The literature has also described this type of applications in terms "value", however value is not a characteristics of processing a huge amount of all type of data, but the ability to implement techniques to extract valuable insights from data for a particular purpose. The distinct characteristics of big data applications such as volume, variety and velocity, do not necessarily have to be present at the same time. In other words, a large amount of data sets that are diverse in nature (structured, semi-structured and unstructured) can be processed in a batch or streaming modalities. We can have also applications for interactive analysis of a large amount of historical data in the form of relational and non-relational data sets. Similarly, we can have automated applications for processing real-time or near real-time streaming data of semi-structured and unstructured data sets.

Analytics refers to the process for systematic analysis of data using a variety of techniques to get insights from a set of data. These techniques are based on a combination of business rules, algorithms, machine learning, data mining, statistics analysis, natural language processing, text analytics, artificial intelligence, visualization, and so on. The analytical techniques are applied to different data sets including historical and transactional data, and real-time data feeds [6]. There are three main analytical models that are not necessarily mutually exclusive for analytical purposes, these are: descriptive, predictive, and prescriptive.

Descriptive analytics uses analytical techniques to provide insights into the past. It allows us to learn from historical behaviours and understand what happened or what is happening. Descriptive models identify different relationships and categories between data, summaries and statistic (mean, variance, max, min, percentile, etc.) of single variables or group of variables.

Predictive analytics by contrast afford insights to estimate the likelihood of future outcomes such as events and behaviours. Forecast techniques are applied to expose what might happen in the future. This encompasses a variety of statistical techniques to analyse current and historical facts to make predictions about future events.

Prescriptive analytics use optimization and simulation algorithms to advice on possible outcomes and a plan of actions. Then, actions and resources might be prepared in advance to respond to likely a situation [8]. The goal is to produce information about the most high-value actions for taking preventing measures [3]. Predictive and prescriptive analytics are categorised as advanced analytics or discovery analytics.

Analytical techniques have been developed and used for long time. However, at the present time there are plenty of terabytes of data sets that adapt well to analytical models. Analytics have evolved from merely confirming or identifying facts using traditional analytical methods, to predicting events and behaviour through advanced techniques [9]. The combination of the characteristics of big data and advanced analytic techniques defines big data analytics as illustrated in Fig .

Smart grid data analytics cannot depend only on a category of analytics but a combination of analytical models to obtain useful results at whole domain of utility´s processes. Some utilities have already adopted a strategy for big data analytics, implementing a variety of solutions across the smart grid processes. The relative value of analytics and data analysis evolution has been represented by the Utility Analytics Institute in Fig 1, which was first published in [10].

2.2 Selecting a Big Data Analytics Solution

In selecting a big data analytics solution, utilities have to consider how effectively collect, store and analyse data in order to support data driven decision making while saving implementation cost.

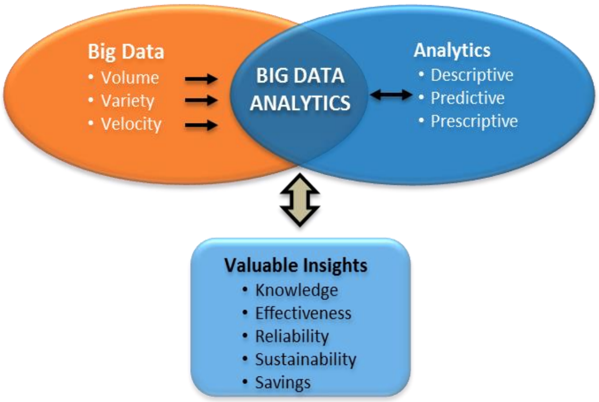

Fig 1. Techniques and technologies of big data analytics lead to knowledge and insights that can be applied for effectiveness, reliability, sustainability and savings for utilities.

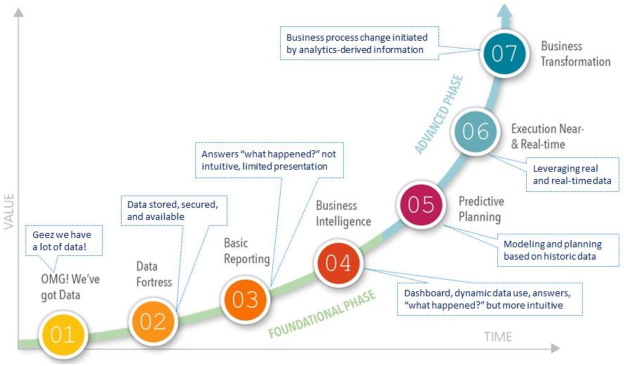
Business process change initiated by analytics-derived information
07
Business Transformation
ADVANCED PHASE
06
Execution Near- & Real-time
Leveraging real and real-time data
Answers "what happened?" not intuitive, limited presentation
Data stored, secured, and available
Business Intelligence
05
Predictive Planning
Modeling and planning based on historic data
VALUE
Geez we have a lot of data!
OMG! We've got Data
Data Fortress
Basic Reporting
04
Dashboard, dynamic data use, answers, "what happened?" but more intuitive
03
FOUNDATIONAL PHASE
02
01
TIME

Fig 1. The utility analytics value curve according to the Utility Analytics Institute.

In this respect, there are two main approaches, using either a dedicated platform or third party services. Both solutions have advantages and disadvantages. The former involves a big effort and time in order to design, develop, implement and maintain a platform. The latter can be implemented through third-party big data products and services. Although there are a number of different commercial software platforms that offer a big data and analytics solutions, most of them are cloud computing architecture offering Software-as-a-Service and Platform-as-a-Service products for utilities. A number of companies that offer big data analytics solutions and sensing technologies based on intelligent devices are described in [2]. However, preconceived analytical applications require customization to meet utility analytic needs; customization might be inflexible to fit the data models of the utility and likely to demand intensive communication resources. In addition, for these services there are still issues related to security, privacy and loss of control over sensitive data [3].

Regardless of the approach, utilities have to consider the analytics tools that have been used for long time and have been useful to support decision making. Similar techniques of these tools can be implemented in the new analytical models and improve their capabilities. For instance, the architecture of those analytical tools might not be prepared to handle data from multiple processes efficiently and their hardware has to be modernized. In addition, their architecture is most likely to be difficult to scale, costly to adapt and maintain.

According to Zhou et al. [2] and Sing and McCulloch [11], the application of big data analytics in power utilities is in an early development stage and the value of big data and advanced analytics are not fully explored and mature. Therefore, there are still some challenges to be addressed in order to achieve the potential of big data analytics. Nerveless, utilities have seen analytics as a strategic approach to drive operational excellence of their processes and gradually have been applied [12, 13].

3 A Conceptual Architecture for an Analytics Platform

Utilities have long used descriptive analytics tools that have been helpful to answer very specific questions. These include spreadsheets and business intelligences tools. Decision making was based on insights derived from reports and statistics obtained from structured data. Power utilities have found valuable information without the need of a more sophisticated analysis. Therefore, utilities have considered these tools and techniques that are still useful to support operational decisions for specific domains. However, legacy systems capabilities must be augmented with modernized hardware and new technologies in order to interoperate with corporate systems; for instance, to integrate geospatial data and third-party information, such as demographics, meteorological and social media data. Figure 2 shows the conceptual architecture analytics platform for power distribution system data, where the flow of information goes from data sources, to the representation of insights in any visualization device.

3.1 Data sources

Data source encompass grid, customer, business and external data. Although analytics apply across the whole processes of a power utility, the platform proposed in this paper is focused on the energy distribution process. In particular, data that are related to operational distribution systems, customer´s data, power consumption data, business data as well as support systems data such as spatial data. External data includes data that can be correlated with utility data such as meteorological, demographic, geographic and social media data. All these types of data are produced by a diversity of sources.

3.2 Integration

Electric utility data can be integrated by an enterprise service bus (ESB) and the semantic of data by a common information model (CIM) in order to facilitate scalability and to enable semantic interoperability through the whole corporate system.

While CIM adapter servers extract data for each grid systems and load them to the ESB, a client CIM adapter combined with other techniques for data ingestion, load data from ESB to data storage [14, 15]. To some extent, CIM model solves issues regarding duplicated data, the different meaning of the same data in diverse systems, inconsistency, and incompatibility of data [16]. EBS solves the problem of integrating multiple data silos by integrating legacy systems and new systems. Integration involves also techniques such as extract-transform-load, messaging and API's that are used for loading and storing data after transformation rules are applied to data such as cleansing, reformatting, standardization, aggregation, etc. For instance, different time stamps in multiple systems of the same type of data, conversion of units and magnitudes, granularity of data, and so on.

3.3 Data storage

Data storage consists of repositories, techniques and software tools for managing relational and no relational data. The former uses conventional relational database management system such as data warehouse and data marts. The latter software involves Apache Hadoop framework and its ecosystem. Hadoop includes Hadoop Distributed File System for data storage, and other software that runs on top of HDFS, like HBase.

3.4 Analytics

Analytics considers descriptive analytics and advanced analytics models. Descriptive analytics is mainly based on OLAP cubes and query based analysis with business intelligence tools. By integrating more information such geospatial data and external data, the capabilities of analytical applications can be augmented for analysing power distribution data. Some examples of application are described below.

- Unbalanced phases. Analytical models can be designed to identify unbalanced phases in asymmetric power distribution feeders produced by unbalanced of current, real and reactive power. By comparing historical behaviour against unbalanced profiles of the current situation, atypical behaviour can be identified. Alerts can also be produced by unbalanced voltage as a consequence of incipient fails or declared fails. This type of applications can contribute to the reduction of technical and non-technical losses.
- Geographical load balancing. Load balancing models can identify the consumption profiles by metering the input-output of power in circuits by geographical zones or electrical regions. Real-time comparison of the inning energy against the sum of power outputs can be very useful to calculate losses. Alerts can be produced when a high loss are present or when unusual consumption behaviour occurs. The model can also be correlated with other variables such as climate, natural phenomena, and even supervising field crew.

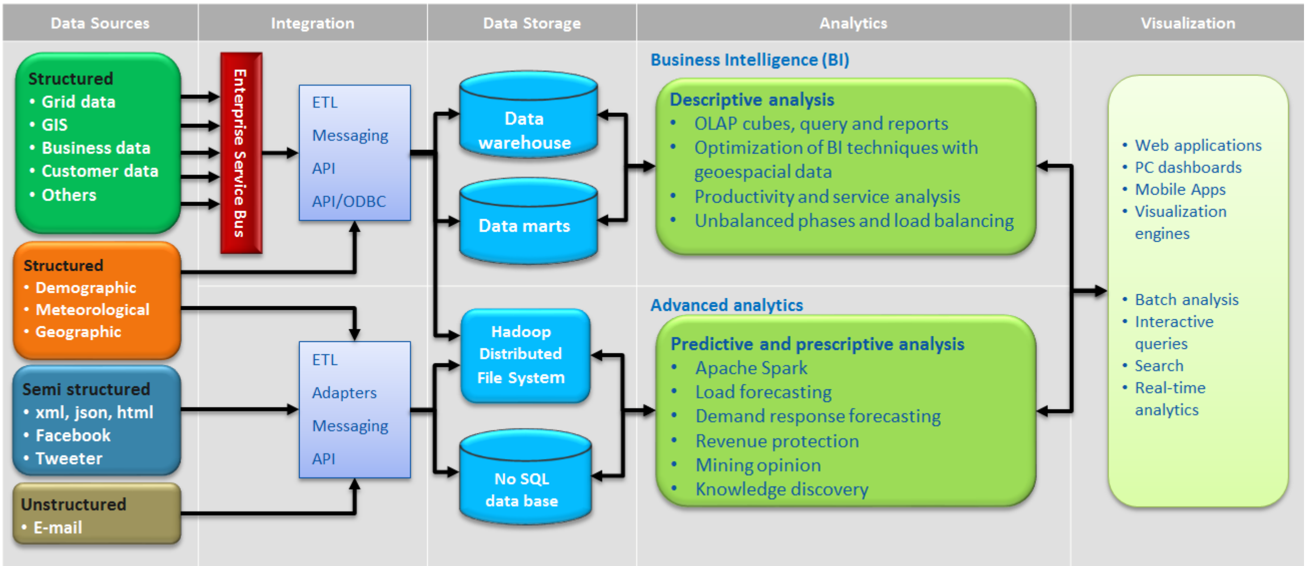

Figure 2. The conceptual architecture of the analytics platform for electric distribution data.

- Productivity. In order to know the performance of the utility is necessary to maintain the constant analysis of productivity. For this purpose, data from different systems can be integrated to analyse key business performance indicators. Metrics can be based on the following information: data operational performance (quality and reliability of the power distribution system, operational efficiency of each power distribution feeder, maintenance), operational and management costs (distribution costs, service costs), financial performance (indicator of cost recovery, sales revenue will be integrated energy), quality of customer service (service performance, reconnection performance, handling complaints, performance call center), energy efficiency and management of demand (annual reduction of kWh in total energy use), data of accidents of field crews and offices.
- Management Service. In order to improve the management of service and customer satisfaction, analytical algorithms can provide insights of outages and interruptions of the service. These algorithms will allow the optimization of resources, dispatching services, tracking of tasks and geospatial location of field crews and support to locate faults, disconnections and reconnections in the grid. For this purpose, it is also necessary to correlate information from different systems to measure profitability, productivity and effectiveness of each technician, field crew and shifts. Other information must be obtained regarding the types of customers (residential, commercial, industrial, government, agriculture, etc.), service quality and possibility of retention of customers. This activity might facilitate the identification and visualization of geographical regions/crews with higher or lower level of profitability to reduce cost and improve customer satisfaction.

Advanced analytics models are focused mainly in applications in which prediction is foremost objective to support decision making. Some examples of applications where predictive analytics can be applied to support the distribution process are described below.

- Load and demand response forecasting. Load forecasting requires the identification of typical and atypical profiles related to consumption and energy demand by customers in specific geographic areas. These profiles can be correlated with data from different sources of information, such as demographic, economic, financial, meteorological phenomena and even demonstrations to integrate results and identify profiles of typical and atypical behaviour. In this way, a bank of knowledge can be created to generate future data of the behaviour of load under real-time conditions of electrical and non-electrical variables. To predict future load conditions, peak demand must be also identified in order to provide recommendations to handle the load more efficiently; reduce the stress of the grid; and improve network operation.
- Revenue protection. These applications are focused on detecting illegal actions such as energy theft, tampering, bypassing, and malfunctioning meters in order to reduce non-technical losses. To identify theft of energy, historical data of energy consumption can be analysed and correlated with geographic and demographic data to recognize patterns and establish behaviours to identify atypical consumption and illegal consumption. With these profiles alteration, bridging and faults meter can also be detected.
- Mining opinion. Mining opinion is also known as sentiment analysis, the objective is to know the opinion of customers through natural language processing and text analysis. It can be used to identify patterns and generate conjectures related to opinion and customer satisfaction about disturbances, faults, intermittency, low or high voltage, etc. By this manner, the utility can understand customer interests, motivations, and behaviours related to energy. These insights are not only valuable for decision-making, but also to design information campaigns, marketing offers, rates and new services, etc. These factors allow the utility to incorporate customer feedback on issues related directly to business requirements, new services and even technology designs.
- Knowledge discovery. Since predictive analytical models generate knowledge from a dataset and an explicit goal, models for new knowledge discovery can be based on explicit knowledge and a non-explicit goal. New discovery knowledge algorithms can

explore massive datasets, identifying information and discovering new knowledge which has not yet been formulated or even when we do not know is in the data.

3.5 Visualization

Visualization is an essential component of analytics, in which the representation of the results has to facilitate the interpretation of the analysis. Visualization incorporates human-computer techniques to represent knowledge, user-system interactions models and visualization engines. Web applications, dashboards and mobile devices enabling interactive and batch analytics.

4 Implementation of an Analytics Platform

The implementation of an analytical platform begins with the configuration of a cluster of servers and the installation and configuration of software, including operative system and software tools for managing data and development. The cluster described in this paper consists of six severs, one for data warehouse with SQL applications to host legacy system with new capabilities. The others servers are configured with one master node and four slave nodes for big data analytics applications with the Hadoop framework and its ecosystem. The Apache Hadoop software library is a framework for distributed storage and processing of large datasets on a cluster of computers. It is designed to scale up from a single computer to thousands of machines. This includes the Hadoop Map-Reduce paradigm for distributed processing, the Hadoop Distributed File System (HDFS) and a framework for job scheduling and cluster resource management (YARN) [17].

4.1 Configuring a cluster

When configuring Hadoop on Ubuntu operating system, the following considerations have to be taken into account. Due to the HDFS replication of data that guaranties fail-tolerance, is not necessary a RAID or logical volume management configuration [18]. While some authors suggest the virtualization of data nodes, others observe that input and output performance of discs and network are penalized with virtual machines. We clarified this issue by performing a simple benchmark with data nodes with virtual machines and physical nodes. After carrying out the benchmark with different combinations between kernel-based virtual machines and physical nodes, we found that physical nodes performed better than nodes with virtual machines. This operation was tested by processing statistical values of 5,000,000 registers of data related to renewal and clean energy. The Hadoop cluster consisted of a master node and four slave nodes with a combined capacity of 128 cores, 512 GB of RAM and 48 TB of hard drives for storage. Each node with 2 Intel Xeon E5-2640 v3 2.6 GHz.

4.2 Developing big data applications

An application to find the correlation between metering data and meteorological data was developed. Although it was a simple application, it aimed to explore the stages for developing big data applications, from integration data to visualization of information as is shown in Fig 3. It was also aimed to create the basis for more sophisticate applications such a streaming integration of data and visualization with mobile apps. Fig 3 also shows the iterative analytics process to answer questions for this particular application.

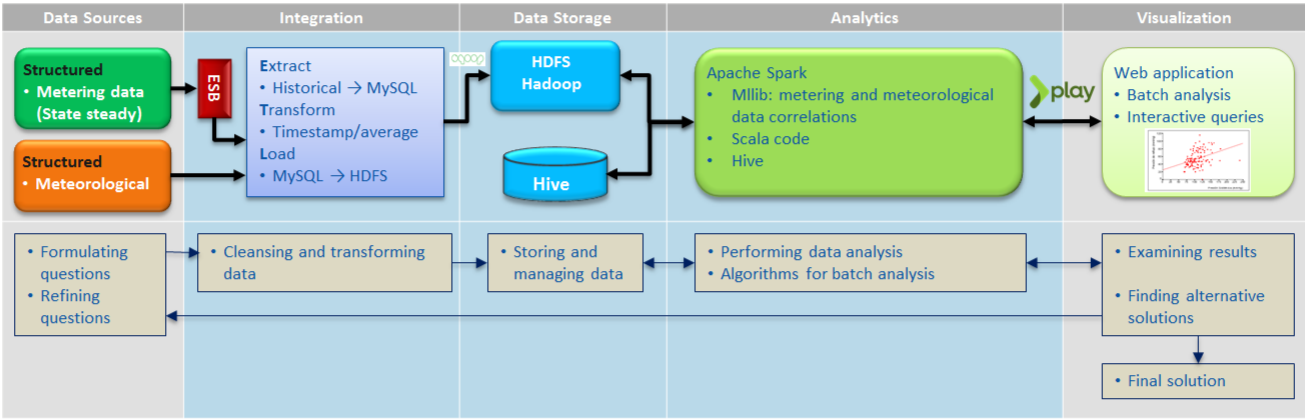

Fig 3. Flow of information and the iterative analytical process to answer questions.

Developing big data applications requires using all the available tools in the Hadoop ecosystem, including Hadoop Distributed File System, Hadoop Map Reduce, Apache Sqoop, Spark, Hive, Zeppelin, Play Framework, HTML and web based libraries for visualization such as Amcharts.js. Sqoop is a tool designed for efficiently transferring bulk data between relational databases and Apache Hadoop [19]. Spark is a general engine for large-scale data processing. The main advantage Spark has over other processing engines is the capability of in memory computations. Spark provides four main libraries than can be combined in the same application. SQL and DataFrames for data manipulation. MLlib for machine learning algorithms. GraphX for graph computations. Spark Streaming for streaming workloads. Spark offers developer API's in Java, Python, Scala and R [20]. Apache Hive is a data warehouse infrastructure that provides data summarization, querying, and analysis [21]. Zeppelin is a web based notebook that enables interactive data analytics and can be used for data ingestion, discovery, analytics, visualization, and collaboration, all in one unified platform. Play Framework is a tool for developing web applications based on Java and Scala programming languages and uses the model-view-controller approach. Play facilitates the interaction with Spark. HTML is the standard mark-up language used to create web applications and its elements form the building blocks of all related pages. Libraries like Amcharts.js and HighCharts were used for visualization in Web applications or Zeppelin with its Angular API.

After the data was extracted from the utility metering data systems and meteorological data sources, these were cleaned, transformed from row data to data ready for storage. Null data was filtered or filled applying analytical rules. Average of current, voltage, and active and reactive power values were calculated.

The datasets timestamps were matched before they were loaded into the HDFS, either by csv files or using Sqoop. The datasets timestamps were matched before they were loaded into the HDFS, either by cvs files or using Sqoop. Data was loaded from HDFS by Spark where it was processed. Spark was used to calculate the correlation between variables that were stored in Hive. While the analytical process was coded using the Play Framework and Scala, the visualization process was coded with HTML and Amcharts.js.

The application allows interactive and batch analysis of data by a range of dates. It is also possible to find automatically correlations across tables and variables that are stored in Hive. Fig 4 shows a scatter plot of the correlation between temperature in a geographic area and the average of the current of a power distribution circuit located in that area.

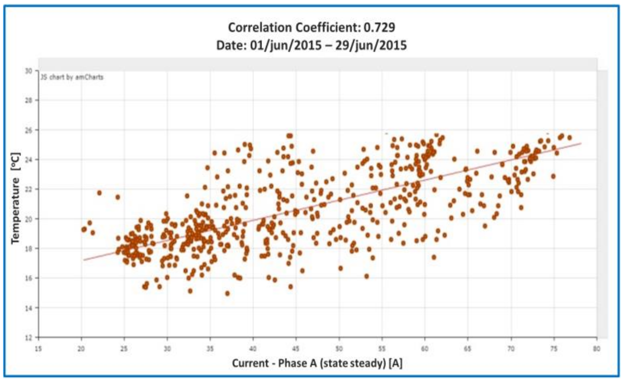

Fig 4. A scatter plot shows the correlation between temperature and average current.

5 Conclusion

Big data analytics will equip electric utilities with the solution for the analysis of huge amount of data but there are also processes where conventional methods will still be useful for answering questions. However, conventional models have to be updated with new technologies and techniques, either using a service approach for an analytics platform or developing a dedicated platform. Advantages and disadvantages of the former option has not been analysed completely in this paper. While exploring the latter approach, we found that developing an analytical platform requires time, knowledge and resources to define analytical models; to design, implement and maintain the platform; and to develop analytics tools.

In so doing, knowledge and experience of a multi-disciplinary team is required, including information technology staff, data scientists, and utility engineers. Apache Hadoop framework, its ecosystem and Apache Spark are open-source tools that provide all functionality to develop big data analytics applications. However, complementary software tools have to be developed in order to implement big data solutions, mainly at the extremes of the flow of information, integration and visualization stages. The integration stage is facilitated when the utility has already been modelled for interoperability based on common information model and enterprise service bus. At the visualization stage, further development of tools is needed in order to facilitate the implementation of applications to visualize analytical results. In addressing analytics and data processing requirements of the power utilities, a number of applications have been suggested in this paper.

The analytical platform in conjunction with these applications should capable of processing data in batch, streaming and interactively. Although the value of big data analytics depends on the ability to design analytic models and an iterative process to extract valuable insights and situational awareness to support decision making, this has to be complemented with engineering for improving electric utility processes.

References:

[1] International Energy Agency, "Technology Roadmap Smart Grids", IEA, Paris, 2011.

[2] K. Zhou, C. Fu and S. Yang, "Big data driven smart energy management: from big data to big insights", *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 215-225, 2016.

[3] C. L. Stimmel, *Big Data Analytics Strategies for The Smart Grid*, London: CRS Press, 2015.

[4] D. Loshin, *Big Data Analytics*, London: Elsevier, 2013.

[5] P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch y G. Lapis, *Understanding Big Data*, London: McGraw-Hill, 2012.

[6] P. Russom, "Big Data Analytics", *The Data Warehousing Institute*, Renton, WA, 2011.

[7] O'Reilly Media, Inc., *Big Data Now*, Cambridge: O'Reilly Media, 2012.

[8] Halo, "Descriptive, Predictive, and Prescriptive Analytics Explained". Accessed 29 06 2016. [Online]. Available: https://halobi.com/2014/10/descriptive-predictive-and-prescriptive-analytics-explained/.

[9] EPRI, "Program on Technology Innovation: Data Analytics and Customer Insights", *The Electric Power Research Institute*, Palo Alto, 2014.

[10] Utility Analytics Institute, "Customer Analytics Report," Aurora, CO, 2014.

[11] C. Sing Lai y M. D. McCulloch, "Big Data Analytics for Smart Grid" IEEE Smart Grid Newsletter, 2015.

[12] S. Callaghan y G. Gauthier, "Driving Operational Excellence Using Analytics "Apps" on a Common Foundation" Utility Analytics Summit Conference, Phoenix, 2015.

[13] M. Angalakudati, "Optimize Asset Maintenance Risk Models Enable Better Decisions" The Utility Analytics Institute, Phoenix, 2015.

[14] A. Espinosa-Reza, M. L. Torres Espíndola, M. Molina-Marín, E. Granados-Gómez y H. R. Aguilar-Valenzuela, "Semantic Interoperability for Historical and Real Time Data Using CIM and OPC-UA for the Smart Grid in Mexico" X, Unpublished.

[15] M. Molina-Marín, E. Granados-Gómez y H. R. Aguilar-Valenzuela, "CIM-Based System for Implementing a Dynamic Dashboard and Analysis Tool for Losses Reduction in the Distribution Power Systems in México" XX, Unpublished.

[16] I. Parra, A. Espinosa, G. Arroyo y S. González, "Innovative Architecture for Information Systems for a Mexican Electricity Utility" de CIGRE 2012 *General Meeting*, Paris, 2012.

[17] The Apache Software Foundation. (21 06 2016) "Welcome to Apache™ Hadoop®!", The Apache Software Foundation, 2016. Accessed 15 07 2016. [Online]. Available: http://hadoop.apache.org/.

[18] T. White, in *Hadoop*: *The Definitive Guide*, Sebastopol, O'Reilly Media Inc., 2012, p. 657.

[19] The Apache Software Foundation, (21 06 2016) "Apache Sqoop," The Apache Software Foundation, 2016. Accessed 16 07 2016 [Online]. Available: http://sqoop.apache.org/.

[20] The Apache Software Foundation, (25 07 2016.) "Apache Spark: Lightning-fast cluster computing" 2016, Accessed 16 07 2016 [Online]. Available: http://spark.apache.org/.

[21] The Apache Software Foundation, "APACHE HIVE TM" The Apache Software Foundation, 2014. Accessed 16 07 2016 [Online]. Available: https://hive.apache.org/.