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.

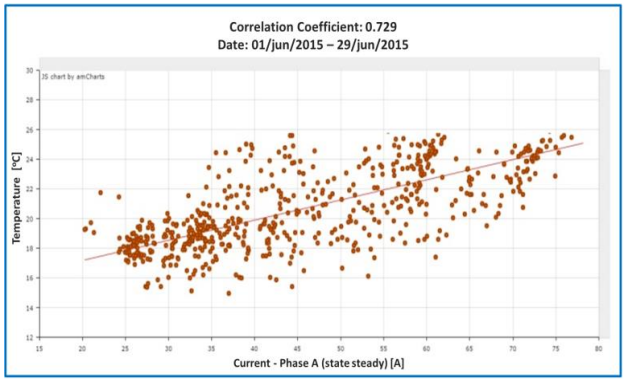

Fig 4. A scatter plot shows the correlation between temperature and average current.

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/.