BIG DATA CHALLENGES – A MULTIDISCIPLINARY TEAM APPROACH

Isabel FONSECA
EDP Distribuição – Portugal
isabel.fonseca@edp.pt

João CASTRO
EDP Distribuição – Portugal
joaosaraiva.castro@edp.pt

André ÁGUAS
EDP Distribuição – Portugal
andre.aguas@edp.pt

Pedro GONÇALVES
EDP Distribuição – Portugal
pedrojose.goncalves@edp.pt

Susana MAGALHÃES
EDP Distribuição – Portugal
susana.magalhaes@edp.pt

Joana BRAANCAMP
EDP Distribuição – Portugal
joana.braamcamp@edp.pt

ABSTRACT

The advent of smart grids created a surge in the quantity of data available to the DSO. Data has become an asset as crucial for grid operation as the grid per se. However, if the first concern is to have data, the second one is to have good data.

This paper aims at demonstrating how data validation rules were built to extract value from Big Data using in-house tools and expertise. It presents a true account on a specific project that involved extracting, crunching and analysing data from multiple sources for a large volume of data and still proceeding when there were no guidelines for dealing with unexpected results.

INTRODUCTION

The paradigm is universal: the DSOs business is becoming increasingly data driven, the volume and variety of data is continuously growing, thus information systems need to adapt in a very fast manner. EDP Distribuição, the main Distribution System Operator in Portugal, is no exception and while its information systems are being updated and replaced, management opted for a "fail-fast approach" to cope with an incessantly changing data environment. The idea was to exploit in-house capabilities while gaining more insight to incorporate on the new Big Data-oriented IT systems and saving costs.

MOTIVATION

With the advance of Analytics and Data Science, a large amount of work has been done in the data segments of energy consumption and generation. These analyses include quality assurance models, but also predictive and adaptive models. There was however, another very important data segment with significantly different behaviour in terms of data analysis that still is yet to be fully exploited – electrical power substations.

Being the border points between the grid's different voltage levels, the electrical substations give us the possibility to analyse the energy flows in the grid with higher granularity. This information becomes valuable to provide accurate information to the market; for improvements in the internal data billing; but also for grid planning itself.

In fact, validated datasets are crucial for the calculation of the Energy Balance since the accuracy of this information allows the power loss calculation by voltage level with greater precision. In 2017 EDP Distribuição launched a project on a tight schedule focused on the challenge of determining the losses by voltage level with high accuracy results.

In addition to this, electrical substations data quality is crucial for other business areas: the Revenue Assurance department, for instance, launched a project that aimed at developing an automated calculation of a KPI for losses by geography and by voltage level, using this very data.

Furthermore, the referred data segment is increasingly essential for the planning of power distribution grids. With the high volume of metering data being collected, new challenges arise, and the current focus is on developing the ability to analyse and extract valuable information that allow us to adapt the management and control of distribution grid to the new reality [1]. Metering data plays here a critical role as it describes with accuracy the grid operating conditions supporting network planning and therefore it is crucial to ensure that this information is properly validated.

Finally, it is very important that this data comes from a unique source, to ensure that there is consistency in the information that is provided. In fact, when the same data is requested by different parties at different points in time, it is a challenge to ensure that the same information is given, and that it has the best possible quality.

In 2017 data was already being collected remotely for nearly all the electrical substations, with 6 measurement points being collected every 15 minutes: active power and reactive power (inductive and capacitive), all injected and withdrawn from the grid. However, since electrical substations are not considered in the billing process, there was a large opportunity for improvement in terms of validation and analysis of this data, as well as in terms of the registration of these installations' cadastre in the information systems.

Given the importance of this data and the need to use it in the short term for two main internal studies, a project was launched with the intent of building a validation and estimation engine for electrical substations data.

BUILDING A DATA QUALITY MACHINE

The first step was to choose the software and the process to create the validation machine. Several tools were studied and various scenarios were analysed based on all possible information.

On the one hand, it would be possible to use the telemetering system to create validation and estimation rules. The software currently used by EDP Distribuição allows the activation of some algorithms, which would be a feasible option since it would allow a quick assembly of the intended validation motor. However, activating the algorithms in the telemetering system would not allow the raw data to be stored in case of estimation data. By validating and estimating directly in the telemetering system, it would only be possible to access already estimated data and no more raw data. Since raw data is also required for other analyses, this option was set aside.

On the other hand, there was the possibility of augmenting new features in the corporate validation system, so to have a data quality machine integrated with what already existed in the company. This would be the ideal procedural solution. However, the scheduling of new features in a corporate system demands a time frame that was insupportable, as it requires an iterative process of establishing requirements, briefing development teams, assessing functionalities and applying corrections.

Hence, to obtain a more agile solution, it was decided to resort to an internal server, isolated from any other system, with immediate availability and easily accessible. The server was explored as an Infrastructure as a Service (IaaS) cloud-provided, meaning that it is fully self-service for accessing and monitoring its performance in terms of computation, networking, storage, and other services. This Cloud Computing model had the advantage of providing complete control of the infrastructure to the data scientist team.

The tool used to store the data and to code the algorithms was SQL Server Management Studio (SSMS). SSMS is an integrated environment for managing any SQL infrastructure, and provides tools to query, design, and manage databases and data warehouses, either in local computers or in the cloud.

To run the project and manage the chain of procedures in a simple and visual way, the SQL Server Integration Services (SSIS) tool was used, since it is a platform for building enterprise-level data integration and data transformations solutions, useful to manage SQL Server objects and data. The SSIS designer provides a visual representation of the work being done. An example of a SSIS visualisation can be seen in Figure 1, where validation modules are scheduled in a flowchart.

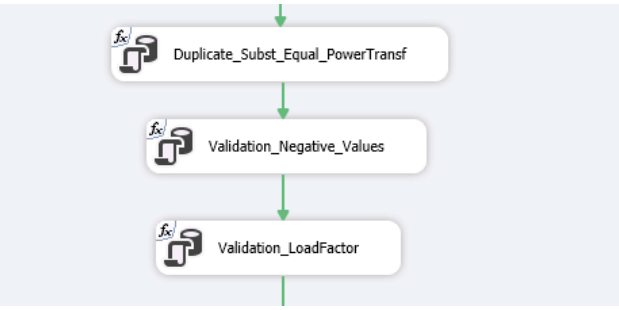

Figure 1 – SSIS Validation Module Flowchart

This server combined with the Microsoft SQL Server tools were then the basis of the in-house built solution for the intended data quality machine.

A MULTI-DISCIPLINARY TEAM WORK

The mining of the data had to be done in an end-to-end approach, to ensure the best possible precision. In fact, as pointed in [2] one of the main steps towards building a sustainable data management program is the integration of data from multiple systems and break organisational silos.

In order to do so, throughout the various phases of the project a distinct set of skills was needed, hence the importance of gathering a multi-disciplinary team of different experts.

Import and Validation of Energy Data

Since the SQL Server is not integrated with the corporate information systems, it was necessary to create a tailor-made data import process. Both metering data and geographic information system (GIS) data were necessary, thus two distinct sources were needed.

On the one hand, interval reads for every 15 minutes of a month were imported to the new data quality machine. This data was extracted from the telemetering system in 6 .txt files, each of these with nearly 4GB. In the import process a primary data treatment was necessary, in order to facilitate later calculations of losses and energy balance analysis. In terms of data quality, the process first would look for duplicate rows and then for empty records.

On the other hand, data from the GIS was necessary, as it was the company's main asset registry. Here, the main challenge was to match the master data from the GIS with the master data from the telemetering system. Since these two systems are not integrated, for some installations the information is not equal on the same precise moment. Thus, an algorithm was developed to match the data of the two sources according to the number of power transformers in each installation.

Finally, validation rules were coded: master data validations, such as the verification of duplicate

installations; and validations of energy values, being the main rules the inspection for negative values, the examination of the load factor and the search for missing values.

Validation of the elements of the GIS's registry

The validation module revealed to be of the upmost importance for grid planning, as discrepancies in the GIS were detected with these rules. Specific algorithms were created for electrical substations data, as internal knowledge about these types of installations grew.

The main validation rules developed to detect duplicate secondary substations are briefly explained below:

- Same identification code except for one digit:

Detects secondary substations for which the identification codes are identic for the same period when the digit that refers to the voltage level is excluded. If for these codes the load curves are the same, the duplicated data from the oldest code is automatically eliminated. If there is no oldest code, the duplicated data is flagged for later analysis and correction.

- Different identification codes but same load curves:

Detects load curves that are potentially duplicated. If in the analysis period the load curves are the same for two different identification codes, there is a high probability that data is duplicated for one substation and missing for the other. For these cases, the correction is always manual.

- Same identification codes but different load curves:

Detects secondary substations with exactly the same identification code but with different load curves nonetheless. For these cases, the correction is always manual.

These validation rules, although very specific, have a significant impact on the quality of the data. In fact, to accomplish a successful data science project, the programming of complex statistical models is not sufficient, as a deep business knowledge is essential.

Another validation that is key to ensure the quality of the grid's registry is the verification of the load factor. If the data indicates that a power transformer is overcharged or undercharged, there is a good possibility that the information residing in the GIS of either the nominal power or the current transformer ratio has not been updated.

Task force on the field

During this study, the team could find some inaccurate data that needed to be confirmed at the source: the smart meter. As there are more than 60K smart meters installed and limited resources, the team had to develop algorithms to identify and prioritise the most critical situations by areas of operation, therefore helping the teams on the ground to focus their attention on the most pressing cases. One of the most important algorithms developed compared the current values from primary and secondary windings with the current transformer ratio.

In order to evaluate the efficiency of the developed algorithms, the analysis on the field started with a 50 installations sample. Each case required an electrical technician in the substation's smart meter and a back-office operator. Electrical technicians should be prepared to measure current and voltages with proper tools and fix connections, if needed. Back-office operators supervised the technician by collecting data remotely and by setting the smart meter's configuration to the correct current transformer ratio. Back-office operators should also collect vector diagrams of voltages and currents as evidence that everything was in order.

As the sample study results exceeded expectations, a task force was created on the field to expand the study to other geographies. During the project's task force, around 600 smart meters were analysed, of which about 80% had in fact at least one field in the registry that needed to be updated.

These works were highlighted as an innovative area of this project, since the development of validation rules for electrical substations led to a closer collaboration between data scientists, electrical engineers and field technicians to update the registry of the network elements.

Energy Data Estimation Methods

The algorithms for the estimation of missing and invalid energy data were coded accordingly to the rules defined by the National Regulatory Authority (NRA) [3]. Essentially, these rules predict a power value as a function of the installation's past load curve.

For secondary substations, however, there are some meters without long historical data. Mainly due to persistent communication difficulties. These occur in remote, inaccessible or signal interference areas: secondary substations located at very high altitude, blocked by obstacles or near the border subject to foreign signal interference. All are situations with difficult telecommunication's network access.

For these cases, an estimation rule was developed based on the premise that a power transformer would have similar load curves to other power transformers that have the same total contracted power. Therefore, clusters of secondary substations were created, classifying secondary substations according to their total contracted power. When a meter of a substation in a given class has no energy data, the power value is estimated with the average between the power values, for the same 15min period, of the meters of substations in the same class of total contracted power.

Calculation of Transformation Losses

After the validation and estimation of metering data, the values of the transformation losses were calculated, in order to have a greater visibility on their impact on total grid losses.

The transformation losses for secondary substations were estimated by consulting a table for copper losses and another table for iron losses, both provided by the NRA [4].

For primary substations, however, transformation losses were calculated based on technical data available: resistance on all wirings and the nominal voltage of each winding, following the formulas for heat losses on power transformers. During acceptance testing, iron losses are measured on site and since these do not vary with the load, these values were used. When it comes to copper losses, calculation is necessary using the resistance values of the windings of each transformer, depending on the number of windings and the side of the metering.

In fact, for primary substations the measurement can be either on the high voltage or the medium voltage side, whereby a certain amount of losses is added or subtracted depending on the metering side. In addition to this, the transformers can have either two or three windings, varying the calculation for each type of transformer.

This operational context meant that the program for calculating the transformation losses of primary substations had a specific structure in order to be computationally efficient: while secondary substation's transformation losses are determined by looking up copper and iron losses tables, primary substations require a whole different algorithm that depends on the characteristics of each transformer and its metering typology.

DEALING WITH BIG DATA

From a very early stage of the project, the challenges of working with Big Data started to reveal themselves: there are around 70,000 electrical substations in EDP Distribuição, and in each of them 6 measurement points are collected. Since the power values are collected with a quarter-hour resolution and there was a need of data from the beginning of 2016, the project's main table started with a size of about 2.24×10^{10} records.

Volume: The need for optimisation

The initial complexity of working with Big Data is the volume of data that needs to be processed and stored. Soon the database had dozens of tables and thousands of code lines.

The first challenge that arose from the large volume of data was in the import and treatment of the monthly data, which quickly exceeded the log memory. Another main struggle appeared when processing historical data, which led to memory issues and was estimated to run for weeks.

Moreover, the need to keep in store historical data, meant an additional effort on code optimisation. A program may be optimised so that it becomes a smaller size, consumes less memory, executes more rapidly, or performs fewer input/output operations. In this project, it was necessary to consider all these aspects and hence the support from an experienced programmer was essential.

Optimisation solutions have essentially gone through less data processing in each cycle and a more efficiently use of keys and indexes. In addition, a functionality of the SQL Server interpreter was used which recommends optimising some specific queries that for some reason may be consuming more resources.

Variety: Continuously improving data registry

The second challenge of working with Big Data is the variety of distinct types and sources of data. As already mentioned, having data from distinct sources created the need to standardise and update the information in various systems. However, this update is not a single point in time, it is constantly done for all installations and for the various required fields such as identifiers, power and current ratio data of transformers, electrical quantities required for losses calculations, etc.

In order to overcome the variety's obstacle, algorithm flows were created and the results sent directly to key users for system's registry update.

However, algorithms are not always ready for all situations that can occur on the field. For instance, a power transformer may only have one smart meter, which measures the total power of every output line. However, in Figure 2 one can realise that in this installation there is only one transformer but with two smart meters.

Figure 2 – Power transformer with two smart meters

Similarly, during the project, cases were found where there is a single smart meter for two power transformers connected in parallel.

All these new types of situation have been incorporated in

the systems and added to the validations. In fact, ensuring data quality and updating the grid's registry becomes a never-ending project, where data and systems are iteratively being improved with new business knowledge.

Veracity and Value: Making data quality smarter

Analogously, to get estimates that are closer to real values and filters of ever-finer validations, it is necessary to adjust the rules iteratively. A very important and substantial part has already been done, now the challenge is to continue to incorporate ever more advanced models, combining past experiences and develop a robust data governance culture.

Another crucial aspect of ensuring the veracity of values is to continuously reduce the manual component of validation. Some validations are done manually because there are exceptions that must be analysed more carefully by an operator. However, the more human work the more likely there is an error. Thus, along with the business team analysing energy data and the GIS, programming skills are always needed to codify the analyses that can be automated.

Velocity: Processing data in real time

The main challenge of using non-corporate tools to solve a business requirement is that the velocity vector is somewhat compromised. This solution has fulfilled its purpose and greatly improved the quality and knowledge of the electrical substations data, but to make fast processing and, eventually, process in real time, other types of tools must be used.

The traditional IT infrastructure of any utility company is not prepared to handle the large amount of data generated by smart grids. Traditionally, nightly processing cycles are programmed so as not to impact business processes, but eventually lead to the information being provided hours or days late.

The solution to this new reality is Big Data technologies, which prove to be capable of processing large volumes of data in real time.

EDP Distribuição is already taking the necessary steps to fully embrace the Big Data ecosystem: with the conclusion of this project, the team involved is now assign with the task of migrating all the lessons learned to a new infrastructure based on Hadoop. Converting all existing rules from SQL to Spark and reshaping the data processing flows into Python scripts. The knowledge acquired is now being translated into far more resilient systems, leaving space for further analyses that will create real value from the validated data.

OUTCOMES AND CONCLUSION

The increased penetration of distributed generation and flexible demand technologies has driven the evolution of traditional passive and robust grids into active and smartly controlled ones. Such change of paradigm has altered the planning and operation of distribution systems that are increasingly relying upon the high volume of metering data being collected.

With this project, EDP Distribuição improved the quality of the electrical substations data and developed a laboratorial tool to test new rules, that is used as an intermediate tool for data processing and settlement.

The project's team comprised all sorts of know-how necessary to address the challenge: programming, grid and business skills together defined and coded the validation rules and estimation methods for electrical substations data. All those different areas of knowledge were essential as they contributed with different expertise into the multidimensional blend necessary to compute the validation and estimation for the quality of the data. Data is about knowledge after all.

In order to develop even more advanced algorithms and most of all to increase the velocity of data processing the future is the adoption of Big Data technologies, together with a shift in corporate culture towards a Big Data approach.

With this paper, EDP Distribuição raises the multidisciplinarity of the referred project to an international level, sharing the company's experience with other utilities addressing similar issues of working with Big Data.

REFERENCES

[1] A. Águas, V. Pereira, L. Jorge, R. Bento, R. Prata, J. Machado, P. Carvalho, L. Ferreira, June 2018, *Data analytics and stochastic simulation methods for risk-controlled network planning*, CIRED Workshop, Ljubljana.

[2] Rosa, DlC, May 2016, *Data Quality Management For Electric Utilities*, DNV-GL.

[3] ENTIDADE REGULADORA DOS SERVIÇOS ENERGÉTICOS, 2016, *Guia de Medição, Leitura e Disponibilização de Dados*, Lisboa, Portugal, 55-56.

[4] ENTIDADE REGULADORA DOS SERVIÇOS ENERGÉTICOS, 2016, *Guia de Medição, Leitura e Disponibilização de Dados*, Lisboa, Portugal, 78-79.