A Method for Improving Business Data Analysis

Pedro Pires
pedro.miguel.p@tecnico.ulisboa.pt
Instituto Superior Técnico
Portugal

Abstract

Businesses are now, more than ever, facing unprecedented challenges. One of these challenges concerns the use of data inside organizations, many of which are known to use spreadsheet programs to manage, store, cleanse and create reports on. This has proven to be far from ideal when taking into account data validity and process efficiency. In this Master thesis we propose a generic solution to address this problem. We followed a Design Science Research Methodology and performed two demonstrations (using two different approaches) to solve this problem. The first demonstration focused on descriptive analytics and the second one on predictive analytics. The evaluation of the system was done both with a DSRM evaluation framework as well as with two metrics for measuring error (Mean Absolute Error and Weighted Absolute Percentage Error). We gathered that the two approaches yielded good results when compared to error-prone traditional spreadsheets and encouraging results in terms of the viability of predictions in this context. The predictions themselves yielded good to poor results depending on the technique used, which we believe was due to the limited amount of data used. This work was developed in a Portuguese public finance organization and used datasets on the same subject.

Keywords: Data Analysis, Business Intelligence, Data Science, Forecasting, Public Finance

1 Introduction

Nowadays, many companies struggle with the data tsunami they have been facing in recent years. In this struggle, data is often wrongfully managed or stored. The problem we address in this article is regarding this data, in particular, when data is stored in spreadsheet programs like Microsoft Excel. This often happens for companies that lose track over how many datasets they have, if they have been validated or even if the information is updated. This is an issue for data-driven companies where their most valuable asset is despised. As such, companies, and specifically older ones are in dire need of modernization which, in this case, is often called digital transformation (also tackled in this work).

1.1 Motivation

The motivation for this work lied deeply in the interest of solving a problem many companies still face nowadays - the incorrect use of spreadsheet programs beyond their capacity. Tools like Microsoft excel are very powerful although, nowadays, we have to our service tools that are more indicated for everyday data analysis tasks. As this type of work has been done for some years now, the novelty in our approach is in its simplicity and ability of adoption for companies/organizations starting their digital transformation.

At the core of this work is the will to solve a common, and also very costly, business problem as backed by our research. By taking a practical approach in this work, we wished to demonstrate the essential steps for the implementation of a BI system from end-to-end.

1.2 Research Problem and Proposal

Briefly put, the research problem we address aims to tackle "spreadsheet hell", meaning the endless, confusing, error-full spreadsheets that are used to store, treat, analyze, cross and visualize data. This problem affects many companies that are heavily dependent on spreadsheets as the lifeblood of their business.

Our research proposal, which will be detailed later on, is based on a full system architecture that aims to solve all the implications of our research problem. This system encompasses the stages from the reception of the data all through to its deliverance and consumption by the target audience whether that is consumers or a board of directors.

1.3 Research Methodology

For this work we decided to follow the Design Science Research Methodology, DSRM. This choice was due to it being an iterative method adaptable to our research problem, helping us to develop a solution for the problem in question.

This methodology gives us a process model to apply in our work and a methodic way to guide or research. Hevner et al. [15], in 2004, proposed 7 guidelines to further understand the requirements for effective design-science research. The reason for this choice was due to the popularity of DSRM in applied resource disciplines as the one of this work. As these disciplines apply theories from many areas the use of a carefully defined methodology was recommended by the literature [15]. It makes particular sense for this project, by belonging to the field of Information Systems and since we want to solve an IT problem [7].

1.4 Organization

This article is organized as follows: section 2 presents the research problem, in section 3 we described the research proposal and then the demonstration, in section 4. Section 5 is dedicated to the evaluation of the work developed and in section 6 we delve on the conclusions.

2 Research problem

Users in business make use of information that is stored in databases and use that data to extract knowledge to make business decisions as informed as possible [9]. Nowadays, databases are part of companies' enterprise architecture model [16]. In theory, these databases are generally simple and easy to use/maintain [6]. The term databases here is to be understood as a tabular design and not relational databases which we will dwell into further on.

This use of tables to store and use information has been a common practice for quite some time and even more the use of spreadsheet programs. A notorious example of a spreadsheet program is Microsoft Excel, in fact, it is considered to be the industry-leading spreadsheet program. It is known to be a number cruncher but is also often used for data visualization and analysis [12].

Spreadsheets are known as the most widely used programming systems in the world, these are used for businesses and personal use for a very wide variety of purposes. From simple calculations to complex financial models[1]. These types of tools although having quite a lot of potential are very error-prone and, depending on the case, the impact can vary from meaningless to being considered as one of the causes for the 2008 financial crisis [3] (although there were more substantial causes [2]). This has led to some research on the subject and the surge of possible solutions to avoid errors [1].

The problem we address is directed to spreadsheet programs, such as Microsoft Excel. When these are being used beyond their capabilities and possibly incur in errors or complications when clearly other tools would be more suitable.

The limitations of these tools has been widely studied [3], with countless examples [4] of error stories with unprecedented consequences. From governments to banks, that were and still are very dependent on such fallible technologies. Yet, these spreadsheets are still pointed out to be "integral to the function and operation of the global financial system".

In more detail, the main known risks of spreadsheets can include: human error, fraud, overconfidence, interpretation and archiving [3]. This article indicates that 90% of spreadsheets contain errors mainly because these spreadsheets are rarely tested, even recent studies point to, about 50% of spreadsheet models used in large companies, having defects. Due to the mix of program code and data, spreadsheets appear to be the perfect environment to perpetrate fraud. Once more, due to spreadsheets not being checked for errors these are not found/fixed. This can be due to the overconfidence employees place on it. The translation of a business problem into a spreadsheet can lead to issues regarding the interpretation decision-makers have on said data. An example of problems in archiving is the case of failed Jamaican commercial banks [10], poor archiving can lead to weakness in spreadsheet control which, in turn, can lead to operational risk [3].

The problem is worsened when there is a need to analyze data contained in these spreadsheets, as they are very hard to read and validate. This analysis, which usually leads to the production of reports based on the information in those spreadsheets, can be faulty because there are no mechanisms to automatically ensure data quality.

It's hard to deny that spreadsheet programs can be great tools it only depends on what they're being used for. And here is where the core problem of this Master Thesis lies, when using technologies like spreadsheets for data analysis and validation we are faced with "spreadsheet hell". The term can include poor data management and poor data quality [13]. It can also occur on two levels micro and macro. The micro level refers to "Frankensheets", these are big, ugly spreadsheet monsters that are hard to understand, hard to use and hard to test. On the other hand, at the macro level regardless of the quality (or lack thereof) the problem lies in the ways these spreadsheets are used, shared and replicated.

It is normal for small businesses to prefer simpler and more affordable technologies, this is possibly the reason why most of the spreadsheet-related issues occur and are kept for very long time [6].

One of the issues that tools like Microsoft Excel also have is related to collaboration. Excel is not natively a collaborative tool, and to make it so would imply use of external technologies and qualification for those tools [5].

As highlighted in the present section we provided the theoretical foundations to frame our research problem.

3 Research Proposal

In the following sections we will be presenting our research proposal.

3.1 Objectives

The main objective of this Master Thesis was the development of an architecture to solve the research problem we address. Essentially, we intend to provide an alternative to the problematic data management and analysis of data inside programs like Microsoft Excel. This implies the creation of an artifact - a system that comprises different tools and techniques to improve data analysis. To solve this problem, we propose to make use of new digital technologies and align them in a system architecture so we construct a complete and ready to use/implement system.

3.2 Pipeline

In order to help in the development of our solution we created a pipeline to organize our work and to keep it as close to the methods from the available literature. This pipeline is set to accompany the majority of our work to help ensure no steps are missed or performed in an incorrect order.

This pipeline's simplicity make it so it can be adaptable to virtually any organization that may suffer from the same problem. The particular stages of this pipeline are described in further detail in the following section.

3.3 Description

In order to accomplish the objectives set in the previous section we propose the creation of an artifact to explore the artifact itself and how well it suits the needs required. Many companies face these issues of using Excel beyond the programs capabilities and often to find quite some errors as well as coming at great expense monetary and efficiency-wise. Very often, Excel is used to crunch numbers but also to store, collect, cleanse, operate, visualize and even correct data. Anyone that has used Excel knows that it is simple and intuitive yet also very prone to errors, so, it is not the ideal tool to be using, specially solely for this task.

Our system will be comprised of many-step processes to get from the first stage where data is received and all the way to the last step which includes the analysis or production of reports on that data, never forgetting the validation component.

In practical terms, the system we propose can be separated into three logical parts categorized according to the academic area to which they belong to (from data science to data visualization and database management): ETL, Analyzing the data on a BI tool and Production of reports.

From this list we can detail what our proposal is to include more specifically. The most time consuming, and possibly the most important part, is the ETL step of the process of our system. To solve the problem of timesheets we realized there was a need to look at data in a different way. We believe that with the 34 subsystems of the ETL architecture proposed by the Kimball Group [8] is a great way to approach this step of our system.

The first component of the ETL, will consist in the extraction and aggregation of the data from the varied data sources available, subsequently comes the transformation stage of the ETL, in which we plan on spending the most time since transforming the data is very important to the theoretical concepts that accompany the design of a data warehouse. The transformation stage is also very crucial to an area such as the one of this Master Thesis due to some of the problems usually reported from Excel spreadsheets. Data is saved in a strange fashion, often doesn't comply to any specific rules, just the need for average human understanding. General activities here imply transposition of how the data is saved, dividing data into logical parts or dimensions (and ensure that it is at it's most granular level), removing aggregations from the data (since these usually mean redundant information), creation of hierarchies for attributes, establishing a prior-defined set of business rules to ensure the validity of the results, pivoting tables, etc. These are some of the transformations we expect to perform to ensure that the designed data warehouse is compliant with the rules of this modeling technique, almost guaranteeing the success of analysis later on.

The last component is fairly simple, it consists of the loading component of data. In our case, in a SQL server database, with the implemented business rules to ensure once more the quality and validity of the data inside the data warehouse.

According to the 34 subsystem ETL process [8] we should have four groups of subsystems, the first three regarding ETL, respectively and the former regarding the active management of the ETL environment.

In the first group, with respect to Extraction, there are three subsystems: data profiling, in which data sources are explored to determine fit as a source and there's a collection of cleaning and conforming requirements; data capture in which changes are isolated from the source system to reduce the process burden; and the extraction and loading of data (into the data warehouse) for further processing.

The second group, about Transformation, focuses on data cleansing and conforming: data is first cleansed and screened for quality, data quality processes are defined to check if business rules are being respected; cleaning control with error event schema and audit dimension is performed; deduplication of data, meaning elimination of redundant members of core dimensions (i.e. customers or products); and data conforming, ensures common dimension attributes in conformed dimensions and common metrics across related fact-tables.

The third group, regarding the preparation for presentation, including: implementation of logic for slowly changing dimensions (SCD) attributes; production of surrogate keys that are independent between dimensions; hierarchy manager, delivering multiple simultaneous, embedded hierarchical structures in a dimensions; special dimensions manager that creates placeholders for repeatable processes supporting the multidimensional design characteristics; fact table builders create the three primary types of fact tables including transaction grain, periodic snapshots and accumulating snapshots; surrogate key pipeline replaces operational keys for the incoming fact table records with appropriate dimension surrogate keys; multi-valued bridge table builder creates and manages bridge tables for multi-valued relationships; late arriving data handler applies special changes to standard procedures due to late arriving fact or dimensional data; dimension manager is a centralized component that prepares and publishes the conformed dimensions to the data warehouse; fact table provider administrates one or more

fact tables being responsible for its creation maintenance and use; aggregate builder builds and maintains aggregates for seamless use with navigation technologies for improved query performance; OLAP Cube builder uses data from the produced schema to populate the OLAP cubes; and data propagation manager prepares the conformed and integrated data from the data warehouse server to be delivered on other environments.

The last group focuses on the management of the ETL environment since the success of the data warehouse depends heavily on the quality of data present there loaded. To achieve this success the ETL system must aim to guarantee three criteria: reliability, for the processes from the system to run consistently to provide data on time and trustworthy at any level of detail; availability, ensuring the needs of its service implying the data warehouse must be available as needed; manageability, a data warehouse is always a work in progress constantly changing and growing with the business, this relies on the correct adaption of the ETL process.

This last group of subsystems includes: a job scheduler that manages the ETL execution strategy; backup system that keeps a backup in the need for recovery restart or archival purposes; recovery and restart the actual process for recovery and restart in the event of failure; version control takes snapshots for archiving and recovering all the logic and metadata from the pipeline; version migration, migrating a complete version of the pipeline from development into test and the production; workflow monitor, guarantees the ETL processes operate efficiently and that the data warehouse is being loaded at the correct times; sorting, serves the ETL processing role; lineage and dependency, identifies the source of a data element and all transformations or vice versa; problem escalation supports the structure that aids the resolution of ETL problems; paralleling and pipelining enables the ETL system to automatically leverage multiple processors or grid computing that respect the schedule needs; security ensures authorized access to all ETL data and metadata by individual and role; compliance manager, supports the organization's compliance with the imposed requirements by maintaining the data chain of custody and by tracking who has authorized access to data; and metadata repository, captures the ETL metadata including the process metadata as well as technical or business metadata.

We decided to opt for a simple and intuitive BI tool so that the produced data warehouse could serve its purpose which, according to Kimball [8], means to have its data easy and fast to access, be labeled meaningfully, consistent.

The BI tool to be used should be able to explore the data warehouse extensively, making use of all the benefits a data warehouse provides, such as cross analysis, slicing and dicing or filtering. This tool should be easy to learn and use and it should facilitate the adoption of prior business uses. It should also prove to be simpler and far more efficient than the old way of doing things. It should have rapid increase learning curve to the users don't get discouraged upon adoption and it should also feel familiar to the user. This should also allow the users to have independence and mastery over the creation of reports without having to rely on IT or data warehouse managers.

In general, the architecture of our system can be described as a multi-platform solution that starts in the business process of an organization to gather the data it uses for analysis from the different data sources. From here, the data is set to follow the pipeline to ensure, completeness, correction and validity. This pipeline was defined with very low specificity to ensure it was adaptable to other contexts, if needed. There was also a strategic choice to not use just one specific tool, since this generic proposal can be applied with the any BI tool. There is also the option to chose whichever tool is preferred for storing the data, here we provided a particular example of Azure SQL Server tool, but this can also be changed.

The advantage of our system lies not only on **posing as a solution to eventually solve "spreadsheet hell" but also on the enabled cross-analysis capacity as well as how efficient the whole process becomes due to our complete end-to-end system**.

4 Demonstration

In this section, we will address the demonstration step of the methodology we chose, DSRM which follows the use of the designed artifact to solve the proposed problem. This will cover experiences and simulations to do so. It is important to know how an artifact is to be used to solve the problem at hand.

In the previous section, we presented the research proposal, here we will demonstrate our solution as well as test it in a Portuguese finance organization. This was the practical component of our work and it will be thoroughly described in the present chapter.

Due to privacy concerns we will not be disclosing the name of the organization. Henceforth, we will the using the name Organization when mentioning the entity.

The organization where this work was developed it is one that specializes in analyzing public finance data being granted with the evaluation of the quality of fiscal policies and executions.

4.1 Previous business processes

This section is dedicated to the description of how their business processes were before the implementation of our proposal. As the present work was tested in a public finance organization some aspects of the implementation of our solution were adapted to fit their particular business models. The Organization is a public institution and their main objective is to ensure the correct application of public finances.

This organization is particular in the sense that their high business value is not in terms of money but in terms of the impact it can have on its community. The goal is to provide a fair and just analysis of public finance.

The maximum value at the Organization is achieved by the accomplishment of its mission in providing reports on public finance in Portugal. Thus the best way to enhance their value is to optimize their business process, which means to optimize how their data is handled.

The work performed is very data-driven, which aligns perfectly with our field of study and our problem definition. The Organization is divided into two main categories of data-related business-roles, that is technical personal and technical coordinators, both of these operate the data directly; being the technical coordinators the ones to be held accountable for the quality of that data.

In the Organization, a usual data exchange starts with the technical staff asking public entities (such as public administrations) for specific data that is needed for analysis deemed necessary. This was the first opportunity we encountered where our solution could serve to optimize their business uses for data and to solve a problem related to the reception of that data.

Then, after asking for the data, typically via email, the technicians are sent a Microsoft Excel file containing the requested data, afterwards the technicians validate the data manually, which can often contain errors, and then is followed another exchange of emails to fix certain errors or ask for clarification of some values that might seem off or raise questions.

After this sometimes lengthy email exchange process only to obtain the data, the technicians perform one last quality check to ensure the data is trusted. All these steps are part of the Organization's business process and they are crucial since the value of their work relies on the quality of the data they work with.

These steps are fundamental to their day to day work and this is where a proposal like ours can help. After these interactions the data is stored in spreadsheets that often may exceed their capacity.

Processes like these are very common across the fields of operation like this one making it an ideal place to test our proposal to solve that same problem.

According to the studied literature, this is somewhat of a common problem since many companies rely solely on Microsoft Excel for their day-to-day activities. Companies sometimes lack the fundamental theoretical concepts that would allow them to develop more complete solutions to somewhat complex problems.

On the subject of Microsoft Excel, just because it is a very powerful tool it doesn't mean it's a suitable tool for any task at hand, as stated before. In fact, that has been proven time and time again. Despite being appointed as an easy program to use for finance it poses many risks and becomes unreliable when it's the only tool being used. The pros are being a low-cost alternative included in the software packages that companies might already have to use (for instance for word processors and email clients), it's fairly easy to use, it's intuitive and from its basic functions to some more complex formulas it's simple and direct to learn, it provides the users with templates and it provides effortless integration in the Office 365 already used by the majority of companies.

However, Excel lacks some useful features such as collaboration in documents, scalability for support of bigger data sets, and proper data visualization tools. Although having some basic features it lacks more complete capabilities competitors have. Situations like these capture precisely our research problem using Excel as the only tool to collect, analyze, validate, and visualize data. Considering the challenges and limitations it may pose, it becomes clear that using this program for data mining and data visualization is far from the ideal solution. In the present section, we present the artifact of this Master Thesis as a system to improve all the faults that the previous system had. While presenting better features in terms of knowledge discovery and potentiating machine learning to meet other business objectives the organization also desired to achieve.

4.2 Tools

The tools used for this work are listed below and are described in detail in the main document.

- First Demonstration
  - SQL Server
  - Power BI
- Second Demonstration
  - python
  - pandas
  - scikit-learn
  - NumPy
  - matplotlib

This demonstration can be divided into two different parts, the one for the first experience and another one for the second. These were planned so that we could test different approaches to solve the same research problem.

4.3 Description of the original datasets

After having decided where to test our solution and after meeting with the organization we concluded to use a representative sample of datasets, also in their interest to test with our proposal. The reason for this alignment in goals was due to their fond interest in developing the organization's use of digital technologies, and with this project, they would be able to delve precisely into digital transformation and data mining.

The datasets they were interested in bringing first into this new business model were of public finance and, to be

more specific, relating to the National Health Service (SNS in its Portuguese abbreviation). This data ranged from the number of employees included in SNS, to how much money was being spent on payments whether to people, medicine, or technicians. It was a very rich dataset containing over 100 attributes relating to the Portuguese Hospitals that are part of the SNS. There were 3 chosen datasets, these were: SNS Accounting Information, HR Information and Public Hospitals' Accounting Information

To describe the original datasets used in this master thesis we decided to divide them into logical categories of the field of public finance. These categories were: entity name and general information, users' reach, primary health care activity, assets and liabilities and estate, pharmaceuticals, human resources (both in number and in financial value), depreciation and EBITDA, and late payments. These datasets when cross-analyzed allow us to have a much more interesting analysis. For instance, by having both the number of people working of a certain category and the value spent on wages we can have the average cost per employee, among other metrics.

The first dataset included attributes like the values for the SNS account, ranging from revenue and expense, balance as well as the particular values that allow for the breakdown of the revenue and expense. There are also some variations in this dataset, parallel to the real and observed values we have the variation and the value that was predicted to be spent.

The HR dataset is fairly simple, including the information for each of the entities, the year, and the value by category. The category list is quite comprehensive, it ranges from interns and doctors to nurses, to all the technical staff that makes up a Hospital.

The final dataset, and possibly the richest one, is the one regarding the accounting information of all the Hospitals that belong to the SNS. For each of the entities, we have information regarding the group of those entities, the year and 77 different attributes regarding the accounting information, the patient reach and types of primary health care activity performed by each entity.

These datasets were stored in an Excel file with the following structure: in the first column the names of all the entities, followed by a column with the year to which that row information belongs to, and followed by several columns containing the values for the attributes in the column above. Naturally, this would pose a problem in terms of data representation since certain visualizations and programs require something similar to a star schema. This even allows for a richer analysis of the data since the information is properly organized. But we'll cover this more in-depth in the ETL and Data Warehouse sections.

Just from a first analysis, these datasets had many problems, and many more we only found in the ETL stage of this work. The first issues were the organization of the information, empty values marked as "NA", "n/d" or "n.d." and changes in representation of values from year to year.

The dataset chosen for the second demonstration contained information from the Social Security accounting information. Its previous business process followed an Excel spreadsheet and calculations made on it. As stated in our proposal, we intend to solve our research problem by creating a system architecture that comprises all the steps from data extraction all the way to report creation.

This dataset although more simple in terms of number of dimensions was more complex in its own structure, it was organized through different levels of specialization.

For the ETL stage we used a scripting language to perform the manipulation of our data which was python.

The extraction phase was quite simple containing just the collection of a large dataset from the Social Security Account. This dataset belongs to just one entity and its validation is more simple than that of the previous demonstration, since there are no business rules to check for, at least to be implemented onto the data warehouse.

This is also a particular dataset since some missing values can be the norm. This is due to some attributes of social security well fare having existed in the past but no longer exist in the present, yet these have to be maintained to preserve past information.

The last subsystem in the extraction phase, according to the Kimball Group, is importing the source data into the data warehouse environment for further manipulation.

The following step is the transformation phase in which we performed the majority of the manipulations from the ETL stage. For this transformation we created a python script that would read the Excel file provided by the public entity responsible for these data and the only source of data for this demonstration.

In python we made use of a python library called pandas, which is very popular and commonly used for data manipulation and wrangling. Some essential concepts were obtained from the book "Python for Data Analysis" [11].

This script after reading the Excel provided by the public entity, which always follows a pre-determined format, proceeds to convert the data into a pandas DataFrame and is from this data structure that the rest of our work for this section revolves. After creating it we separate this accounting data into expense and revenue. These are the two separate DataFrames that we will be managing henceforth.

Parallel to this work we were given another Excel sheet that would be the target manipulation of this original dataset. This is the in-house state for this data to be analyzed for whenever there is a need to produce a new report this manipulation needs to happen. This is also a different view on the data making use of only the needed attributes from the entity that is responsible for providing the data.

After analyzing this file we noticed that there were quite a few attribute aggregations and quite a few hierarchies, also, one of the reasons for the creation of a data warehouse. From the theory we studied, we know that there is no need to store anything but the values at their most atomic levels since the rest are just aggregations of those very values.

As such, from the excel file we decided to encode the logical hierarchies and, to later be able to construct our data warehouse, we decided to create a python dictionary with tuples for the keys and values. The reason behind this was to encode the hierarchies needed for this data and to later use that dictionary to build the hierarchies in our data warehouse.

After creating the dictionary with the connections we went back to the two DataFrames mentioned before regarding the expense and revenue for the SS account. The first step after separating the two datasets was removing the values that were not at the most atomic level, instead we decided to create a list of the name and level number for the most granular data. This meant to use the dictionary to check what were the values that respected this condition. We then created a new Dataframe (one for the expense and another for the revenue) that only contained these values. We used a function from the pandas library called melt. This transforms the data from a wide format (values encoded through the columns) to a long format (values encoded with two columns one for the name of the attribute and another for the corresponding value). This operation is also often called "unpivoting" in the BI community. This long format, and theory and practice tell us so, will allow us in the future to perform drill downs or rollups.

As of now, these manipulations have led us to having three columns one for the dates, another for the attributes of the most granular level and another one for the corresponding values.

After this we needed to construct the rest of the dataset from the information encoded in the dictionary. In terms of the revenue, as an example, the most granular level happened to be just up to level five, so to be clear we renamed the column attribute to level five which is the corresponding level in this case for the revenue. The remaining levels were created from the mapping of that level five to level four and then level three and so on.

Due to the way our script was written it makes it adaptable to any problem being able to change the way data is organized to one that follows data warehousing modeling.

After all these manipulations we obtained our next-to-final versions of the data warehouse in which we had information of the dates, the levels and the corresponding values.

To ensure the validity of our data warehouse we performed cross validation with one report already produced by the organization. Here we found some issues but after backtracking the wrong attributes into the script we were able to quickly correct them. These were issues in the creation of the dictionary (the data structure where the hierarchies were encoded).

The final step in this ETL process was to load the data onto a data warehouse which was also included in our python script.

As we stated, our approach was consummated with the realization of two different demonstrations. The first demonstration focused on following a pipeline and architecture defined precisely to solve the research problem at hands culminating in the use of a BI tool. The second demonstration also focused on following the same pipeline and architecture to solve the research problem, although the tools used to achieve this were different.

In the first demonstration, we used a very popular BI tool called Power BI, as requested by the organization in which this work was developed. For the second demonstration, we delved into business analytics using tools such as python, pandas, and matplotlib. For this, we covered what is commonly referred to as predictive analytics, also of the interest of the organization. From predictive analytics, we focused essentially on forecasting which we found interesting and also met the requirements of the organization.

More specifically, the first demonstration followed our defined pipeline all the way from source selection to cleansing, to load the data onto the warehouse and finally analyzing it or working with it inside the BI tool of choice. For this demonstration to be successful we would have to be able to fulfill the process of the previous business model. We achieved just that and even more efficiently. With our data warehouse modeling we were able to introduce even more analysis that could have not been done before. According to DSRM, we produced an artifact to solve our research problem which materialized in the form of a system. To complement the demonstration we also produced reports and visuals for the data needs of the organization. These were also used in the official report published by the organization.

In terms of the second demonstration, we set to solve the same research problem, using slightly different techniques and converging on a pure data science technique - forecasting. The data went through all the steps in the pipeline that was entirely manipulated using python script, guaranteeing the automation of the whole process. Furthermore, after passing the data through all the ETL subsystems we used machine learning to make predictions for a few of the attributes of the dataset. In the process of developing this demonstration, we also found that it could be interesting to have the cumulative values (which is the normal form for the dataset) but also the non-cumulative values to see the real value of an attribute in a month. This could maybe allow us to see patterns/friends that are implicit to the data.

For the forecasts we used two different methods: ARIMA and SARIMA, which are, respectively, Autoregressive Integrated Moving Average and Seasonal Autoregressive Integrated Moving Average. These were chosen according to

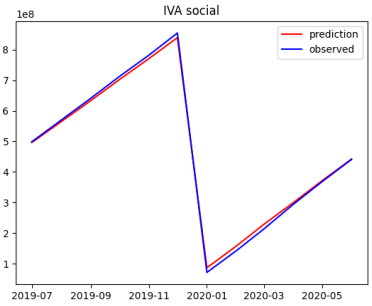

Figure 1. SARIMA technique prediction for the attribute IVA Social.

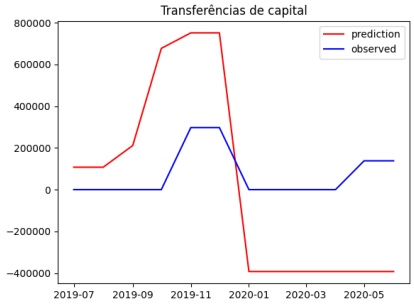

Figure 2. SARIMA technique prediction for the attribute Transferências de Capital.

their popularity in the field. We tested different values to try to find which would be the best parameters for these two techniques (by trying to minimize AIC and BIC).

The following conclusions and Figures are just a select few as the remaining techniques and results can be found on the main document.

We see very encouraging results for the prediction in Figure 1, where there is just a slight deviation from the observed value.

On the other hand we also had predictions that were very poor as is the case for Figure 2.

The results for this forecasting experience were measured using two metrics that are commonly used for this type of work, these are MAE (Mean Absolute Error) and WAPE (Weighted Average Percentage Error). These metrics yielded satisfactory results, although, one could argue these results may not be very reliable due to the interpretation of the metrics and the limited time series we had access to in our dataset.

The palpable results of the first demonstration, besides the system itself, were some visuals produced inside the chosen BI tool, these can be seen in the main document.

The results of the second demonstration are also the system as well as some interesting plots for the predictions we

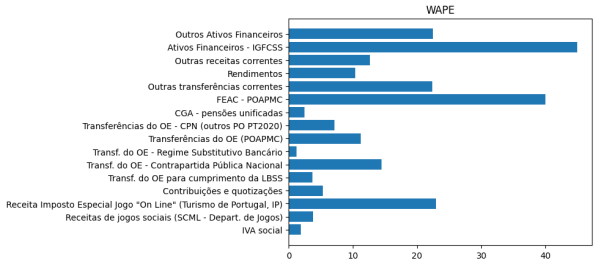

Figure 3. SARIMA-associated error plot for the WAPE measure.

made. For these we used the two very popular techniques mentioned, ARIMA and SARIMA. In Figure 3 we can observe the percentage of error obtained using the WAPE measure. From this figure we can observe that this model performed rather well for a few attributes such as 'IVA social', 'Receitas de jogos sociais', 'Transf. do OE - Regime Substituivo bancário' and 'Transf. do OE para cumprimento da LBSS' which had an error percentage of under 5%, implying an accuracy of the forecast of 95%-99%.

On the other hand, for other attributes it was not so good even reaching a 45% of error in the worst case for Figure 3. There are also other attributes for which the predictions are poor. These are 'Transferências de capital', 'Outras receitas de capital - Ativos Financeiros - Garantias' and also 'Ativos Financeiros - IGFCSS' and 'FEAC - POAPMC'.

As a whole we can gather that roughly, the values of WAPE for this method are around 15% which lead us to an accuracy of 85% for the majority of the forecast attributes shown in Figure 3.

5 Evaluation

Artifacts in the IS field are often considered systems, such as the case for our own artifact. These are created to solve a real life problem having functions and objectives for its purpose.

For the evaluation of the artifact we will be taking a IS approach following our choice for the methodology, using techniques commonly used to evaluate these artifacts. According to a paper on the subject [14] there are many ways to evaluate them, yet there is a need for adapting which ones make sense for the scope of the artifact itself.

Prat et al. suggest that the evaluation criteria should follow a hierarchy based on the theory of justification, conforming into three interrelated levels. These levels are system dimensions, evaluation criteria and sub-criteria.

According to their proposed hierarchy there are five system dimensions, 20 evaluation criteria and 12 sub-criteria [14]. From these we believe that our system relies most on two out of the five system dimensions, these being goal and environment.

Goal, as with any artifact from the DSRM being very goal-driven it was essential that our system were evaluated in terms of resolution of the problem for which it was created. For this Master Thesis, the goal was to solve the many problems associated on the use of spreadsheet programs for storing, cleaning, shaping, transforming, visualizing and crossing data. There are technologies designed specifically for this which we have studied and used to solve our research problem. The fulfillment of the goals should be evaluated following three criteria efficacy, validity and generality [14].

Efficacy is the extent to which the goals were met, in this criteria our system was able to comply with all the demands of the organization in terms of structuring of the system and organization as well. The data was stored in a similar tabular form but, for our system it complied with the design of a data warehouse that allowed for more flexible manipulation of the data.

For this system we designed a more complex data warehouse that would encompass all the dimensions and facts relevant to the production of reports regarding the data being analyzed.

Validity means the degree to which the artifact is able to work correctly, or performs reaches the goal correctly. An example of how our system meets these goals is the implementation of the integrity constraints into our data warehouse that would only accept valid data, in terms of complying with prior defined and in place business rules. This can also encompass reliability which after validating the data that went through our system we can confidently say that it is, indeed, reliable producing consistent and correct results.

The generality of an artifact implies how applicable it is to the broader problem. A broader goal for the artifact means a more general artifact. So much so, that we even applied our system to another problem different from the first one. Due to our research problem, and consequently our goal, being defined to solve the widespread problem of "spreadsheet hell" we can say that it is also a general artifact.

As for the other system dimension relevant for our work, environment, Prat et al. suggest this should be evaluated using the following criteria: consistency with people, consistency with organization and consistency with technology.

This is a very relevant system because, the environment of IS artifacts includes people, organization and technology. It also made particular sense for our system because being an all-rounded solution for such a big problem its environment is also broad and need to be sound for all the smaller components to work correctly.

There are some limitations to this component of evaluation because our solution although complete is only partial to the whole transformation that is to happen inside the organization which was set in motion with some of the work produced for this Master Thesis.

The consistency of the environment for either people, organization and technology encompasses 10 sub-criteria. Utility measures the quality of the artifact in use, our system, for this criterion was met with excitement for this new technology as well as promises since after validating the data (common practice even in the previous business model) all that was left to do was to build reports and share the findings, which is part of the mission of the organization.

Understandability, can also mean ease of use of which our system can take a little time getting used to the learning curve is definitely worth the extra effort. This evaluation is also made in a point of view that the people who are going to be using this system have no previous knowledge of computer science fields such as databases or programming, this could be the reason why the system may seem a little less easy to use than a simple Excel spreadsheet. The only compromise to the ease of use of our system lies only on the multiple parts that compose it. From the data warehouse validation rules, to the process of loading and transforming the data for the data warehouse and also the learning curve for the new Business Intelligence software. Although we firmly believe that once we're past this first period our system can become easier to use and even more reliable.

Ethicality means for the system to not put animals, people, organizations or the public at risk, this sub-criteria does not directly apply to our line of work since the only risk would be the leak of information that is in general, not the case since almost all of it is public domain or publicly accessible.

In terms of the fit of our system with the organization we believe that it is ideal since the work developed there benefices greatly of the use of our system. Considering the line of work of the organization and since business intelligence and business analytics are at the core of its mission, our system was precisely designed to meet those needs.

When it comes to the criterion of consistency with technology the value of the produced artifact lies on it being a new layer built on new IT artifacts, which is also the case of our system. We make use of some of the most recent technologies to harness their potential and elevate our work.

According to the article[14], the evaluation should also consider the side effects that this system might have in its environment. From what we studied and, from a digital transformations stance, we know that these fundamental changes to an organization's business process can be met with some resistance from the people. Although from the literature this only happens when it is done incorrectly, there is a need to educate and empower the employees and helping them see the potential these new tools have for their work. By helping them achieve mastery we almost guarantee the success of our system in the long term, according to the literature.

As a matter of fact, our system not only meets the previous requirements it also surpasses some expectations in the field of technology particularly, the way the system is built allows

to use the most modern technologies to treat, analyze and mine the data involved.

From the demonstration in the previous section, the interest behind some of the visualizations isn't on the visuals themselves but more on the efficiency that our system provides. By using our scripts, the work only needs to be done once and all the visualizations are readily available and ready to be updated with just one click with our artifact. Instead of having to manually use the values for the creation of new visualizations.

We believe this is a work with a lot of potential, that has been done in many other fields and is upcoming in many more. This digital transition has to be done mindfully and has to comply with norms and the evolution of the technologies themselves.

6 Conclusion

In this section, we will be presenting the conclusions we were able to draw from this body of work. As we have shown with our evaluation, our system architecture is able to reach its goals with the minor inconvenience of the learning curve for the organizations that use older technologies. Although we firmly believe the pros more than surpass the cons. Also with the extra added benefit of the improvement of process efficiency reducing pointless computations (as aggregations) already implemented into our data warehouse's logical multidimensional model. Precisely due to this fact there is only the need to validate the finest granularity data once, at the moment they are uploaded onto the database.

Although the results of this particular forecasting experience were not the most encouraging (despite having some forecasts that appear to be very close to the observed values as shown in Figure 1) the results of our artifacts were very promising indeed. With our work, we have proven that there can be a quick and easy implementation of our system to solve a problem many organizations face. With the added benefit of time and also cost-efficiency.

One takeaway here is that to ensure that a system is viable to make fair predictions we need a broader time series. This would allow to better train a model and possibly obtain better results.

References

[1] Robin Abraham and Margaret Burnett. 2006. Spreadsheet Programming.

[2] Viral V. Acharya and Matthew Richardson. 2009. Causes of the financial crisis. *Critical Review* 21, 2-3 (6 2009), 195–210. https://doi.org/10.1080/08913810902952903

[3] Grenville J Croll. 2009. Spreadsheets and the Financial Collapse. *arXiv preprint arXiv:0908.4420* (2009). www.eusprig.org

[4] European Spreadsheet Risk Interest Group. 2020. Spreadsheet mistakes - news stories. http://www.eusprig.org/horror-stories.htm

[5] Jörn Freiheit, Ramona Görner, Jost Becker, and Frank Fuchs-Kittowski. 2014. Collaborative Environmental Data Management Framework for Microsoft Excel. *Proceedings of the 28th EnviroInfo 2014 Conference, Oldenburg, Germany* (2014).

[6] Oksana Y Iliashenko and Svetlana V Shirokova. 2014. Application of Database Technology to Improve the Efficiency of Inventory Management for Small Businesses. *WSEAS Transactions on Business and Economics* 11(1) (2014), 810–818. http://www.isem-fem.spb.ru

[7] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* 24, 3 (2007), 45–77.

[8] Ralph Kimball and Margy Ross. 2011. *The data warehouse toolkit: the complete guide to dimensional modeling.* John Wiley & Sons.

[9] David M. Kroenke and David J. Auer. 2012. *Database processing : fundamentals, design, and implementation.* Pearson. 612 pages.

[10] Victoria Lemieux. 2008. Archiving: The Overlooked Spreadsheet Risk. *arXiv preprint arXiv:0803.3231* (2008).

[11] Wes McKinney. 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* " O'Reilly Media, Inc.".

[12] Daniel Z. Meyer and Leanne M. Avery. 2009. Excel as a qualitative data analysis tool. *Field Methods* 21, 1 (2009), 91–112. https://doi.org/10.1177/1525822X08323985

[13] Simon Murphy. 2005. Comparison of Spreadsheets with other development tools (limitations, solutions, workarounds and alternatives). *European Spreadhseet Risk Interest Group* (2005). http://arxiv.org/abs/0801.3853

[14] Nicolas Prat and Jacky Akoka. 2014. ARTIFACT EVALUATION IN INFORMATION SYSTEMS DESIGN-SCIENCE RESEARCH-A HOLISTIC VIEW. *PACIS* (2014), 23–undefined.

[15] Von Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. 2004. Design science in information systems research. *MIS quarterly* 28, 1 (2004), 75–105.

[16] Yuri B. Senichenkov, International Conference on Mathematical Models, Russia) Methods in Applied Sciences (2014 : Saint Petersburg, International Conference on Economics, and Russia) Applied Statistics (2014 : Saint Petersburg. 2014. Recent advances in mathematical methods in applied sciences : Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14) : Proceedings of the 2014 International Conference on Economics and Applied St. In *Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14).* 438.