NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Digital Analytics: An approach for Data Quality Control

Ana Rita Xavier Marques

Internship report presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics with a major in Data Science

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DIGITAL ANALYTICS: AN APPROACH FOR DATA QUALITY CONTROL

by

Ana Rita Xavier Marques

Internship report presented as a partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics with a major in Data Science

Supervisor: Nuno Miguel da Conceição António

October 2021

To my boyfriend, Diogo,
for the unconditional love, patience, and support

ACKNOWLEDGEMENTS

Professor Nuno António, thank you for all the guidance and support, for being a source of motivation and positivity, and for always providing valuable inputs that were essential along this journey. My sincere gratitude for accepting to supervise me without even knowing me and for the availability to help in each step of the way.

I also want to thank my manager, Paulo Dias Faria, who accepted this challenge, helping and guiding me throughout the project. Who gave me the opportunity to explore a new path and take the lead. To my amazing team, Margarida, Adílio, Álvaro e Lucas, for the support and shared knowledge during the internship, and the valuable brainstormings, one of the main foundations of the project.

To my friends, Sofia, Ana, David, Pedro, and Marisa, who always encourage me to do my best, who stood by my side sharing concerns and questions, always with encouraging words, a smiley face, a tight hug, and, of course, our radish mascot.

Lastly, but not least, to my boyfriend, Diogo, for the unconditional support, love, and patience during critical times, and to my parents, Rita and Jorge, for the endless caring and encouragement to always excel myself and proceed with my ambitions.

To each one of you, thank you!

ABSTRACT

With the emergence of the Digital Era, a new way of analyzing customers' behavior also emerged. It's not only about analyzing data from traditional data warehouses but also about measuring users' digital footprint on websites, mobile applications, and other digital data sources. Nowadays, companies collect data on their digital channels to improve website design and user experience, optimize e-commerce, track, and measure the success of actions and programs, identify problems, and improve the digital channels' performance. But the question that arises is how valid, accurate, and complete the data is. Do digital analysts understand each data point they have at their disposal? In this internship report will be given a detailed view of the critical points of digital analytics data quality, the adjacent problems and a solution will be presented to support and help the digital analysts overcome some of the challenges in this area.

KEYWORDS

Digital Analytics; Data Quality; Data Cataloging; Web Crawler; Tracking

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CRM	Customer Relationship Management
DAA	Digital Analytics Association
DS	Design Science
DSR	Design Science Research
DSRM	Design Science Research Methodology
ETL	Extract, Transform, and Load
GDPR	General Data Protection Regulation
HTML	Hypertext Markup Language
IS	Information Systems
IT	Information Technology
JSON	JavaScript Object Notation
QA	Quality Assurance
TMS	Tag Management System
UDL	Universal Data Layer
URL	Uniform Resource Locator

KEY TERMS

Data Governance
"The process of managing the availability, usability, integrity, and security of the data in enterprise systems, based on internal data standards and policies that also control data usage." (Stedman & Vaughan, 2020).

Digital Analytics
"Encompasses the collection, measurement, analysis, visualization, and interpretation of digital data illustrating user behavior on websites, mobile sites, and mobile applications." (AT Internet, n.d.).

Digital Event
"The user's interaction/activity with a web page element that is tracked." (Sharma, n.d.).

Tag
"In digital analytics, a tag is an element included on each webpage to be measured. The tag is a small piece of code that is inserted into the page's source code. It allows the third-party analytics tool to log connections on its server including analytics platforms (Google Analytics and Adobe Analytics), A/B testing tools (Adobe target, Optimizer), and marketing vendors (DoubleClick Floodlight pixel, Facebook pixel)." (AT Internet, n.d.; Chen, 2017).

Track
In digital analytics, a track is commonly referred to as an ID that uniquely identifies a web/app page or clickable element (Note: this term is not scientifically defined, it is jargon in the business environment).

Universal Data Layer
"A data layer is a JavaScript object that is used to pass information from your website to your Tag Manager container. You can then use that information to populate variables and activate triggers in your tag configurations." (Google, n.d.).

Web Crawler
"A software application that runs automated tasks (scripts) on the web pages, typically for web indexing." ("Web Crawler," 2021).

1. INTRODUCTION

1.1. INTERNSHIP DESCRIPTION

In the second year of the NOVA IMS Master's degree in Data Science and Advanced Analytics with a major in Data Science, I enrolled in a 12-month internship at a telecommunications company. This report explores one of the analytical activities performed and the actions taken to overcome the obstacles, focusing on digital analytics data.

During the internship, I was part of the Digital Analytics Chapter at the Digital Channels Tribe – Consumer Business Unit -, and my main responsibilities were focused on digital data testing, management, and reporting. These three sections are the full spectrum of the digital data life, from its development and cataloging to the valuable insights that can be extracted from it.

The work during this period was focused on two blocks. The first one was to develop a centralized documentation database of the website and mobile application analytical events to keep track of all the functionalities and screens the company offered in these digital channels. The core of this project combines data governance and quality control to support reporting and ensure data quality. The second main block of the developed work was reporting consisting of building Adobe Analytics and Power BI dashboards to support the digital transformation business decisions across the company. These dashboards aimed to analyze the usability of the digital channels and support the Team's decisions about the functionalities and journeys to be integrated on the digital channels. The contextualization of the business objectives behind these tasks will be explained in the following sub-section of the report.

1.2. BUSINESS CONTEXTUALIZATION

The company is a telecommunications services provider from mobile to network services (fixed bundles that include tv, internet, and voice services) and also provides and sells IoT services and products like smartphones and accessories. Additionally, the business focuses on taking care of the customers in a self-care dimension through its digital channels. Furthermore, one of the company's goals is to take advantage of the digital transformation to increase brand awareness, customer loyalty, and satisfaction, reduce costs and enhance the customer experience.

To meet the business objectives mentioned above, it is essential to collect data from the digital channels, process it, and analyze it. This process is where the Digital Analytics Team comes in. It ensures that the analytics tags are working both on the website and mobile application and that in each new implementation, the analytics tags and tracks are integrated. Thereby, it is possible to collect data about which pages customers saw and where they clicked, analyzing their behaviors, the usability of each channel and feature, and the interest or intention to buy a product or service.

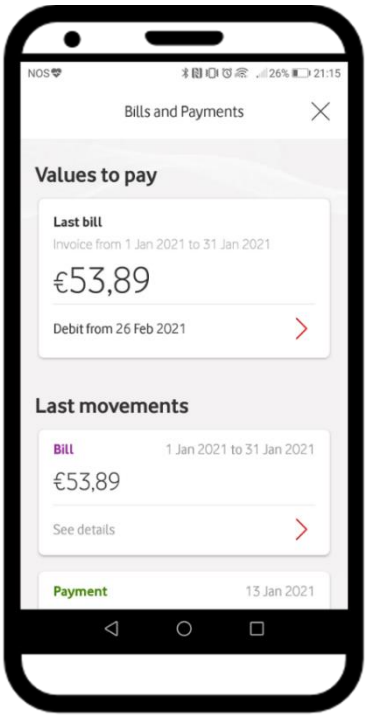
1.3. PROBLEM STATEMENT

During the years the company has been collecting digital data, the world evolved along with the knowledge on how to collect data from digital platforms and on how to ensure data reliability and accuracy. New insights, solutions, and better approaches started to appear, and the way to track digital behavior was adjusted over time, followed by where to store the data that also changed. Moreover, both the company's website and mobile application evolved with new features, pages, and products, and, while new analytics nomenclatures were added, others ceased to exist.

To overcome all the challenges from the last tracking method and follow better practices, in 2019, the analytical tracking was shifted to a new system within a new management framework. This upgrade put a lot of pressure on the Digital Analytics Team, which had to ensure that all the systems integrations worked and that everything in digital platforms was tracked.

This migration is 97% complete on the website, while in the mobile application, only 70% of the features are tracked using the new system, and for now, there is a mix of both systems. Some features are tracked and measured using the new system, while others are still based on the old system. Additionally, focusing on the app, the implementation process took some time. Within that time, a new and completely different version of the mobile application was released - a new design, structure, and experience. Thus, although some features were already shifted and implemented within the new system, many features were replaced, and new analytical nomenclatures were given with the new developments.

Due to the changes mentioned above, in the analytics tool, for different timeframes, the same page can now be associated with three different analytical nomenclatures: the legacy's name, the phase 1 new system's name, and the phase 2 new system's name. A visual example for the section bills and payments of the company's app is given in Figure 1 and Table 1. Since the cataloging of previous analytical data was not managed, there are some difficulties in understanding what exists, what is being measured, and what is the correct terminology of pages and events to perform the analyses. The historical data also constitutes a problem: if we do not have a catalog of today's events, how can we know the name of the fixed bundles' page from two years ago? Is it the same, or is there a different terminology for it? At the same time, there are other factors drifting data quality that need to be addressed. For example, a particular broken tag or track may not have been implemented as intended, thus needing to be fixed.



System	Page Name
Legacy	mcare:bills and payments:overview
Phase 1 New System	Billing
Phase 2 New System	Billing Home

Table 1 - Page Name's Examples during the Migration Phases

Figure 1 – App's Bills and Payments Section

These are common problems in digital analytics "as the very purpose of web analytics is to provide compelling evidence to make website changes and drive continuous optimization" (AT Internet & Digital Analytics Association, 2019). While website/app complexity is added, and upgrades and changes are made, a breach for uncertain digital analytics is opened, whilst more thousands of data are generated, making it hard to keep up.

"Uncertain data quality is a reality in web analytics." (Nelson, 2011)

It was imperative to tackle this problem as the migration process continues and new developments were at the front door. Therefore, this business need set off the work developed during the internship.

1.4. Report's Structure

This report is based on Digital Data Quality. The project developed consists of an artifact that supports a solution for digital analytics data quality control for the company's use case.

Since the knowledge about Digital Analytics was learned on the job, many pieces of the subject were unknown before the start of the project. Therefore, this report also constitutes a learning opportunity about the details, concepts, and best practices that the literature states and recommends for digital analytics and data quality.

In the next chapter, a summary of the literature will be presented, giving an overview of the main concept - Digital Analytics. This chapter summarizes the dimensions for data quality, compiles the problems and challenges faced in this changing environment, and presents some data control procedures to help digital analysts. The digital analytics framework and the surrounding topics, namely

the digital analytics implementation, the tag management system, and the data flow are also detailed in this chapter. Additionally, the methodology used (Design Science Research Methodology - DSRM) is also described.

In Chapter 3 - Methodology, the developed work is fully described at each stage using the DSRM guidelines along with prior and related work that motivated the artifact development. In this description, the obstacles, solutions paths, and limitations are presented. In Chapter 4 - Results and Discussion, the results of the developed work are observed, evaluated, and discussed, sharing a reflection that compares the problem and objectives with the final product, while in Chapter 5, conclusions are drawn from the work developed during the internship, its contribution in the business context and personal development. The last chapter describes the limitations found during the design process and future work paths on the project.

2. LITERATURE REVIEW

This chapter provides a detailed overview of the theoretical background of Digital Analytics and is divided into three main blocks. The first block explains the Digital Analytics concept and describes the digital data - its structure, concepts, and metrics. On the second block, the digital analytics framework is presented which is based on the company's digital analytics framework. The third and final block encompasses the data quality theoretical background, where the dimensions of data quality in digital analytics are detailed along with the challenges faced in the digital environment, possible solutions, and organizational pillars for ensuring data quality.

2.1. DIGITAL ANALYTICS

The term Digital Analytics "encompasses the collection, measurement, analysis, visualization and interpretation of digital data illustrating user behavior on websites, mobile sites, and mobile applications" (AT Internet, n.d.). This concept derived directly from the term Web Analytics around 2012 when web analytics providers realized that the goal was "no longer about measuring only website usage but instead understanding the entire digital footprint of users" (Zheng & Peltsverger, 2015).

Data has incredible potential and can help businesses across every industry. When well-integrated, managed, processed, and analyzed, data can provide essential insights, and currently, more than 50% of companies worldwide use data to support decision-making (Oxford Economics & SAP, 2016). With this in mind, digital analytics can be used for four distinct but interconnected purposes (Zheng & Peltsverger, 2015):

1. Improve the website/application design to enhance customer experience and satisfaction. This improvement consists of optimizing navigation and architecture, improving content presentation and layout, integrating new features and functionalities to more easily retain the user and capture its attention.
2. Optimize e-Commerce and improve e-CRM on customer orientation, acquisition, and retention. Digital analytics allows companies to get the full picture of their digital channels' usability, understand which features add value to customers and which difficulties they face to provide a more personalized approach to the target audience.
3. Track and measure the success of actions and programs (e.g., commercial campaigns, bill payments, services subscription). Analyzing the digital journeys from the very start (e.g., the start of the journey can be "see the detail of the commercial campaign" and the end "the commercial campaign subscription with success") brings value to the business. Sometimes the journey flows between different channels, hence it is crucial to understand which traffic sources are more effective and have a greater impact so that it is possible to allocate resources to the most profitable channels. Understanding the funnel of visitors along the journey can also help the business readjust the flow (objective 1 and 2) or even identify technical problems (objective 4).

4. Identify technical problems and improve website/mobile application performance. Two of the main metrics that drive a good user experience are the time a page takes to load (Responsiveness) and the availability of the pages (Uptime). Digital analytics is the key to get these indicators.

As digital analytics drives a cyclical process of website/mobile application optimization (Bekavac & Garbin Praničević, 2015), the ultimate challenge becomes ensuring high data quality throughout the complete lifecycle of the data, maintaining information as a mirror of reality.

2.2. DIGITAL DATA

Digital data can be very challenging, especially when people are used to structured data and traditional databases. Therefore, perceiving how digital data is structured is crucial to understand all the inherent concepts and the factors affecting its quality.

Digital analytics can be categorized into off-site and on-site digital analytics. Off-site analytics are based on data from external sources (e.g., surveys, market reports, competitor comparison, public information) to analyze the websites or mobile application potential audience, answering "how your website compares to others". On the other hand, on-site analytics are based on data collected directly from the website/mobile application to analyze user interaction and engagement. For example, the collection of data from the usability of each feature and functionality, user journey, interest/intention to buy a product or service to answer "how users are interacting with your website". The two most used on-site digital analytics tools are Google Analytics and Adobe Analytics (Adamiak, 2020). The focus of this report is on on-site analytics.

The data generated from each specific user interaction on a digital platform is gathered into a "hit". A Hit (also known as image request) is a single user-interaction block of data sent to "Analytics". In the business environment, "Analytics" is commonly referred to as the analytics tool that gathers all the digital data.

> "Each time the tracking code is triggered by a user's behavior (for example, user loads a page on a website or a screen in a mobile app), Analytics records that activity. Each interaction is packaged into a hit and sent (...)*" (Google, n.d.)*

There are various types of hits, but the most common are page view hits and event hits. As the name suggests, a page view hit occurs when an instance of a page is loaded (or reloaded) in a browser. Event hits are user interactions with page elements that can be measured independently from a webpage or screen load (e.g., button clicks, downloads, link clicks, form submissions). Usually, event hits refer to actions performed by the user.

Page view hits usually have the variable page name associated, which identifies the analytics page name. Event hits have four specific components, the variables event category, event action, event label, and event value. The event category identifies events over related user interface

elements/events. The event action is typically the type of event or interaction for a particular object. The event label is usually the associated label of the clicked button or action performed. Lastly, the event value is an integer used to assign a numerical value to a page object. An example of a page view hit and event hit can be found in Section 8.1 and 8.2 of the Appendix Chapter, respectively.

Hit Type	Description	Personalized Variables
Page view Hits	an instance of a page is loaded (or reloaded) in a browser	Page Name
Event Hits	user interactions with page elements	Event Category Event Action Event Label Event Value

Table 2 - Hit Types

To better visualize a page view hit and an event hit, the following example gives a visual of them. Figure 2 represents a page view of the App's homepage, which is the source of two event hits – button 1 and button 2.

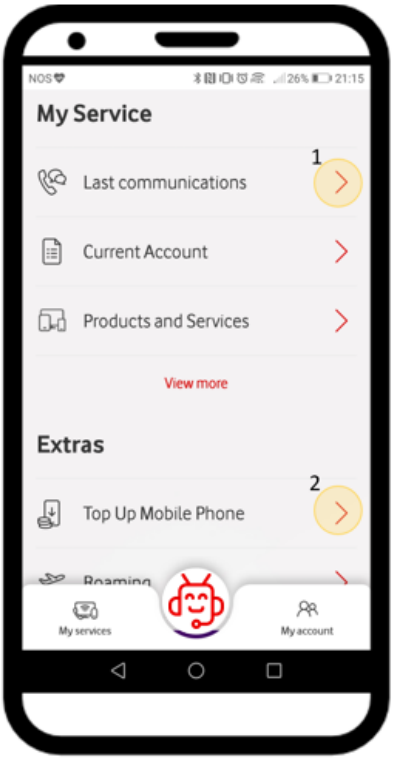


Figure 2 - App's Homepage

Variable	Page	Button 1	Button 2
Page Name	Homepage	-	-
Event Action	-	My Service	Extras
Event Category	-	Homepage	Homepage
Event Label	-	Last Communications	Top Up Mobile Phone
Event Value	-	1	1

Table 3 - Variables Example

The most common types of variables are Props and eVars. Props are considered traffic variables attached to the hit. They do not persist beyond the image they were fired on, and they cannot be associated with other variables that are not in the same image request. eVars, on the other hand, are considered conversion variables used to determine which website dimensions have contributed the most to successful events. They are persistent in preserving the value that was originally fired. (Adobe, 2019).

Props	eVars
Associated with the Hit	Associated with the Session/Success Event
Fired in every single hit	Fired once until is rewritten
Not persistent	Persistent

Table 4 - Props and eVars Comparison

As digital data is unstructured, each hit has a set of variables that can differ from one to another, depending on multiple factors, namely hit type and event scope (e-commerce events, for example, have specific and personalized variables). There are multiple collected variables, and they are usually part of a scope – User data, Hit data, and Product (e-commerce) data. In the following table, there are a few examples of variables.

Scope	Variables	Short Description
User	Marketing Cloud ID	Universal, persistent ID that identifies visitors across all the Adobe solutions
	Visitor ID Account Active	Unique and encrypted visitor ID
	Visitor Customer Type	Consumer or Business
	Visitor Asset Plan Type Active	Mobile Prepay, Mobile Postpay, Fixed Service, etc.
Content	URL	Web page's address
	Page Section	Web page's path (e.g., Loja Online:Telemóveis:Samsung Galaxy A71)
	Site Section	Web page's section (e.g., App, Site, Youth Site, etc.)
	Page Name	
eCommerce	Product Name	
	Product Category	
	Product Brand	
	Cart ID	

Table 5 - Variables Example by Scope

Finally, using the data extracted from the web/app, multiple metrics are available to analyze different perspectives of the user interaction. The three most-used metrics that give an overview of the type of analysis that could be made are listed below:

1. Unique Visitors: the number of uniquely identified users by an ID that is stored in a cookie. This metric depends on cookies, and it is limited when a user deletes their cookies or if he/she uses multiple browsers/devices. If cookies for the current domain can't be found, new cookies with a new ID are set (Adobe, 2020).
2. Page View: an instance of a page being loaded (or reloaded) in a browser (Adobe, 2020).
3. Visit or Session: a series of single-user page views during a period of activity. The session ends after the user either closes the browser, clears cookies, is inactive for 30 minutes, 12 hours of activity, 2500 hits, or 100 hits in 100 seconds (Adobe, 2020).

All the metrics available ensure different types of analyses such as clickstream analysis where the goal is to analyze the navigation path a visitor browsed; engagement analysis measuring, for example, visited pages per session, the average duration of a visit; conversion analysis or funnel analysis where is analyzed from the visitors or visits that started a specific journey how many completed it.

2.3. DIGITAL ANALYTICS FRAMEWORK

Digital Analytics is a dynamic environment in evolution. Different tools communicate with each other, triggering different elements on the digital platforms to deliver the best customer experience with targeted and personalized features. Data is in each point of the big picture. To enter into the dimension of data quality, it is crucial to understand how everything fits together to perform digital analytics. Thus, this sub-chapter will explain the technical details of the digital analytics implementation and the data flow until the data arrives at the endpoint to be analyzed.

This topic is based on how the digital analytics framework is structured and how the processes are designed and executed at the internship's company. In this context, the framework is mainly supported by Tealium, a tag management system, and Adobe Analytics, the digital analytics tool.

2.3.1. Tag Management System

A tag management system (TMS) is a software solution that makes it simple to implement, manage, and maintain tags on digital properties using an intuitive web interface. The tags are a means to collect and move data between the website or mobile app session and the technology vendor such as Facebook, Google, and Adobe Analytics (Tealium, n.d.). The process flow of how tags work can be seen on the diagram shown in Section 8.3 of the Appendix Chapter.

The TMS interacts at the level of the web page where its JavaScript code is implemented. When a user opens the companies' website, the browser[1] requests the page located at the URL[2], and the page content (HTML[3]) is returned. Then, the browser begins to parse the HTML document, processing various elements like images, links, text, or scripts like the tag manager's JavaScript. In turn, this specific JavaScript loads and executes the data layer, load rules, tags, and extensions defined in the Tag Manager. These four components are the essentials for analytics implementation.

[1] "A web browser (commonly referred to as a browser) is a software application for retrieving, presenting and traversing information resources on the World Wide Web." (*Reading: Web Browser | BCIS 1305 Business Computer Applications*, n.d.).
[2] *URL (Uniform Resource Locator)* is "a protocol for specifying addresses on the internet." (*Definition of URL | Dictionary.Com*, n.d.).
[3] *Hypertext Markup Language (HTML)* is "a standardized system for tagging text files to achieve font, color, graphic, and hyperlink effects on World Wide Web pages." (Pournelle, 2004).

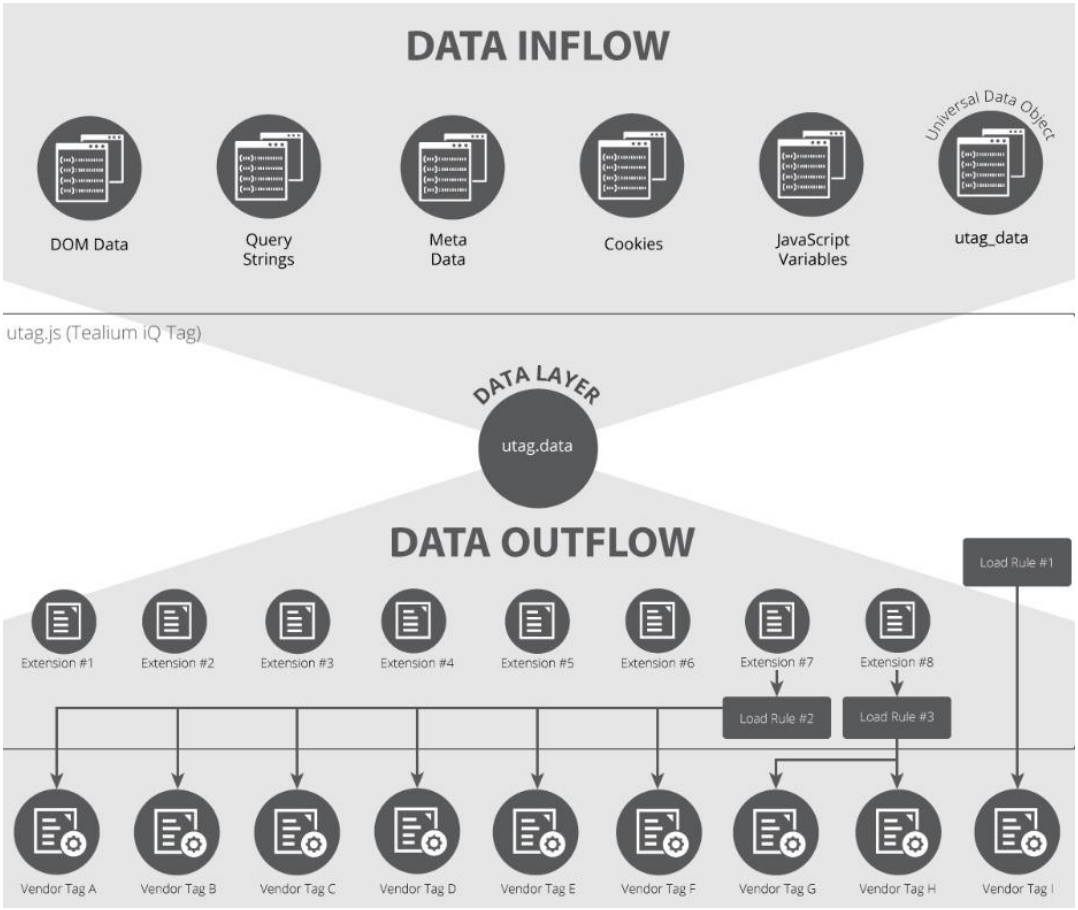


Figure 3 - Tealium Data Flow
Source: Tealium (2016)

A data layer is a JavaScript object used to pass information from the website to the Tag Manager container. The information populates variables and activates triggers that define which tags are going to be processed (Google, n.d.). The Universal Data Layer (UDL) or just Data Layer (utag.data object in Tealium) comprises the statement of all the variables that are collected across the website, the visitor interactions, and events that are tracked and browser data (Tealium, n.d.).

However, the data collected on the website will only be sent, for example, to Adobe Analytics, if they are set up in the UDL to what is called “data points” (Tealium variable) and mapped in the TMS to the vendor (in this case, to Adobe Analytics variable). A trivial example can be the following:

a) The data layer states the data points “page_name” and “event_label”.
b) In the TMS, the “page_name” Tealium variable is mapped to the “Page Name” Adobe Analytics variable.
c) The populated data layer, which overwrites the original data layer for the triggered event, contained the variables “page_name” and “page_section”.

In this case, only the “page_name” value would be sent to Adobe Analytics since the “page_section” is not defined as a Tealium variable and thus is not mapped to the Adobe Analytics variable.

After the data layer has been processed and populated, the tag manager's JavaScript executes the tags and extensions (depending on the order defined in the script) according to the load rules specified. The load rules are conditional statements based on the data layer points that determine precisely when and where a tag loads on the website. The tags at Tag Manager set up the account details and options, data mappings, and load rules for data to be shared with vendors. Finally, the extensions contain data preprocessing rules before data is sent to vendors.

Summing up, the tag management system is the foundation for data collection and governance, operating as an intermediate that manages, standardizes, and gathers all the data in a single place, distributing it to multiple vendors (Figure 4). It maps all the variables collected to a global variable to be used to integrate different types of tags. Without a tag manager, for each vendor implementation, it would be necessary to hardcode it at the website, and the same information would be collected and replicated multiple times. In this case, the TMS was only explained at the level of the website as the objective was to understand its integration and purpose.

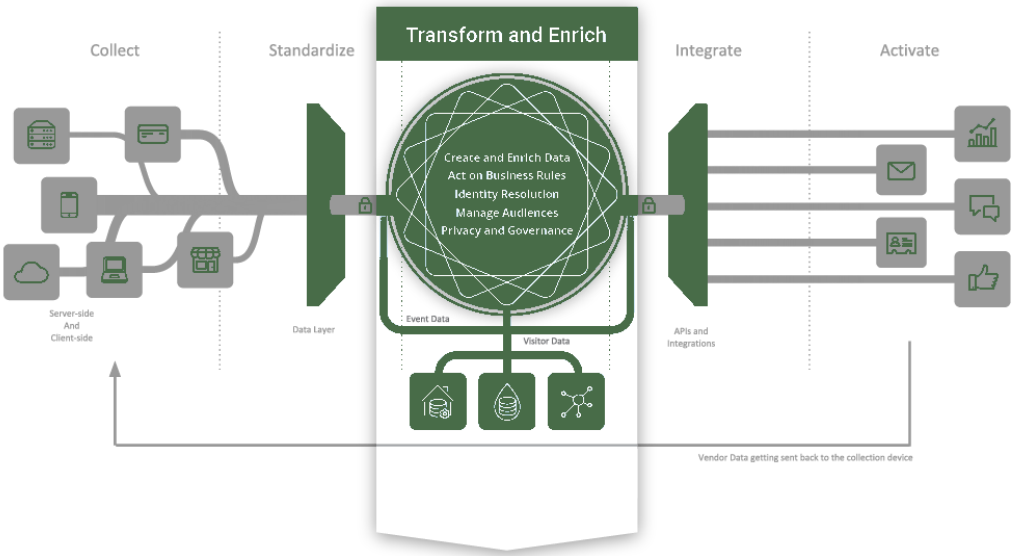


Figure 4 - Tealium Customer Data Hub
Source: Tealium (n.d.)

2.3.2. Digital Analytics Implementation

To collect digital data to perform analytics is necessary to complete a set of developments and integrations at different stages in different tools. Hence, the digital analytics implementation process is composed of two phases: Development and Setup/Configuration.

In the Development process, the first task to take place, based on the analytics tagging/tracking request, is constituted by the following steps (Figure 5) (Analytics Demystified, 2015):

1. Understand and define the business requirements – identify goals and measurement objectives to determine which elements need to be tracked with analytics and which data must be collected.

2. Formalize an analytics request defining the key components of the digital analytics implementation – enumerate the elements that should have analytics tracks associated, the variables attached to these elements that have to be filled, and their values.
3. Incorporate code at the website or app source that includes the information defined in the previous step and the tag from the Tag Manager. Developers implement this step.
4. Perform quality assurance (QA) testing to ensure that tags and tracks are correctly implemented and that data appears within the predefined elements with the correct values.
5. Approve the analytics tagging request.

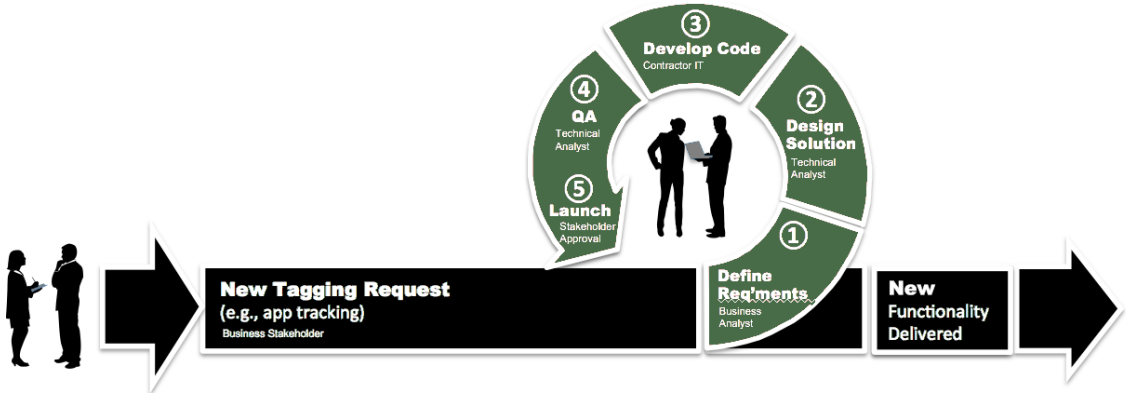


Figure 5 - Analytics Tagging Request Process
Source: Analytics Demystified (2015)

Once the process reaches its final stage, it is ready to be released to the production environment. Except for the third step, digital analysts are involved throughout the process.

After the development phase, a set of configurations must be ensured before the analytics tool starts receiving data. These configurations are essentially associated with the Topic 2.3.1 Tag Management System and the analytics tool.

1. If a new variable is created, add it to the TMS data layer and implement the data mapping between the TMS data point and the vendor's variable.
2. If necessary, apply a preprocessing rule for the data that will be collected.
3. Configure the analytics tool as intended. The analytics tool has a varied range of possible configurations, from traffic management to data configuration. The following topics are an example of some configurations that may be made.
    a. Customize the calendar, setting up the first day of the week.
    b. Configure the internal URLs and IP to exclude internal network traffic.
    c. Configuration of variables that will be used (all the variables need to be listed) - activation, nomenclatures.
    d. Configuration of events: nomenclature, type, polarity (up is good or up is bad), visibility (Always Record Event, Record Once Per Visit or Use Event ID – ties the given event to a custom ID and subsequent counts to the given event with the same event ID are ignored).

e. Exclude bot traffic[4].
f. Data Governance to be compliant with GDPR.

4. Test and validate all the previous configurations.

When the implementation process has been completed, the data flow can begin its designated course.

2.3.3. Data Flow

The data flow for digital analytics is relatively simple and does not differ much from the standard data flow. The user interaction with the website or app triggers the data collection where a long list of parameters is gathered to the hit via Tag Manager. Then, the data is preprocessed on the Tag Manager and is sent to the analytics tool, where it is manipulated - processed, formatted, and organized. Finally, the data is ready to be used to visualize patterns and extract information.

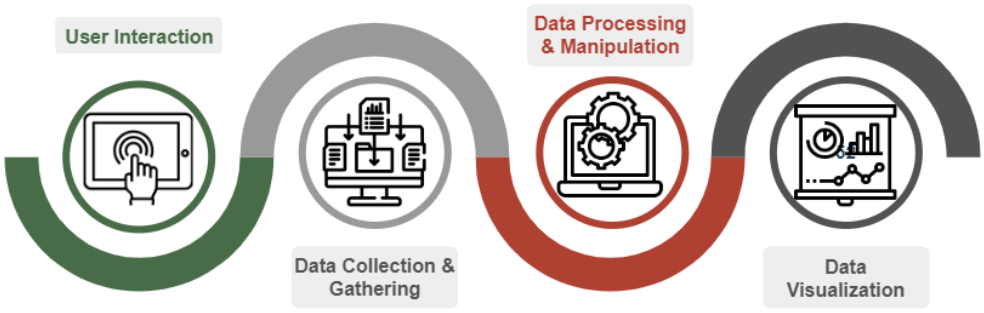


Figure 6 - Digital Analytics Data Flow Diagram
Adapted from Analytics Demystified (2015)

2.3.4. Analytics Incorporated into Development Cycles

Before going deep on the topic, the working model in place at the company must be contextualized – The Agile Working Model.

Agile is an iterative, time-boxed approach for managing projects, and it promotes adaptive planning, evolutionary development, quick delivery, and continuous improvement (Rubio, 2020). The Agile working model empowers teams to quickly deliver increments or new versions, giving flexibility to quickly adapt objectives and change planning to meet customer needs and demands. This flexibility contributes to higher customer value, engagement, and satisfaction, and is seen as a competitive advantage.

During the digital analytics implementation, the development teams must insert analytics tags and tracks at the source as was already described. However, the lineup between digital analytics and development teams has room to grow and to be strengthened in the pursuit of an agile and communicative environment.

Nowadays, analytics is a crucial part of every business since data supports decision-makers. However, to extract insightful information from data is necessary to collect data, and in digital platforms, besides

[4] *Bot traffic* describes "any non-human traffic to a website or an app." (Missulawin, 2019).

analysts, developers are also needed. Considering an Agile working model, every time the development team delivers a new piece of the puzzle, data collection should be part of it too. Hence, the development process should include the implementation of analytics at the product level. To do so, analysts should also be part of the team alongside developers.

The culture of analytics integration in the development process is evolving, and the mindset moves in the collaboration's direction. For now, there are still some constraints at this level. Some teams are more willing to integrate analysts into the process. Others uphold this task as secondary, performing it from time to time or delivering analytics with no feedback expectations.

When analytics are not being delivered as intended, with some errors or incoherencies requiring improvement, is where the process gets stuck. Since development teams usually have their product backlog[5] booked, a minor issue of analytics that does not directly affect the final product will not be a priority. Thus, fixing a simple analytics error will take time just for developers to consider it.

Summing up, fostering collaboration between business stakeholders, analytics, and development teams is crucial to ensure and enhance an agile culture, where communication and agility contribute to a faster and better problem-solving to accomplish business objectives and needs with the right flexibility and adaptation to a growing environment.

2.4. DATA GOVERNANCE

To understand the spectrum of data quality and the mother term giving rise to information assets management and its importance, the concept of Data Governance and its goals must be perceived.

Data Governance is defined in many ways by different authors. According to Weber et al. and paraphrasing Sadiq in *Handbook of Data Quality Research and Practice*, Data Governance specifies a structural framework for decision-making rights and responsibilities regarding the use of data in an enterprise. John Ladley in *Data governance: how to design, deploy, and sustain an effective data governance program* references the *Data Management Body of Knowledge* (DMBOK), where Data Governance is defined as "The exercise of authority, control, and shared decision making (planning, monitoring, and enforcement) over the management of data assets". Data Governance is also defined as "the organization and implementation of policies, procedures, structure, roles, and responsibilities which outline and enforce rules of engagement, decision rights, and accountabilities for the effective management of information assets." (Ladley, 2012). As a more practical and on the job definition, at the micro-level (focusing on an individual company), Data Governance is a "the process of managing the availability, usability, integrity, and security of the data in enterprise systems, based on internal data standards and policies that also control data usage" (Stedman & Vaughan, 2020).

[5] "The Product Backlog is an emergent, ordered list of what is needed to improve the product." (Schwaber & Sutherland, 2020).

In the literature review, it is clearly stated that “ensuring data is managed” and “management of data” are two completely different sides of the coin – “the Governance V” (Figure 7). Although being complements to achieve a common goal, data governance and the people who govern identify the required controls, policies, processes, and develop rules for correct and restrict data management as well as verify their compliance. On the other hand, data management is “the business function that develops and executes plans, policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data and information”, according to the DMBOK as John Ladley states.

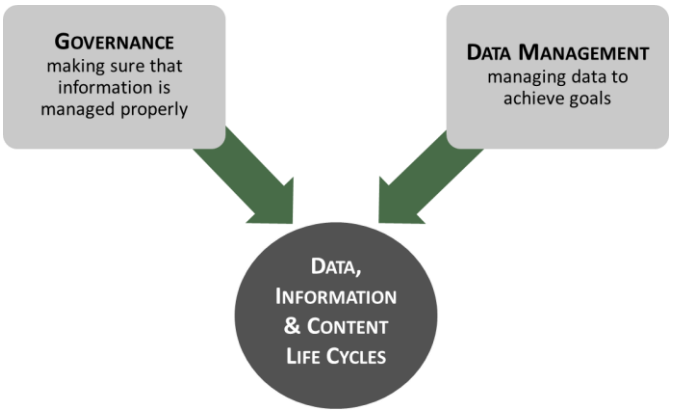


Figure 7 - The Governance V
Adapted from Ladley (2012)

Data governance aims to maximize the value of data assets in companies, while data quality management, a sub-function of data management, aims to maximize data quality.

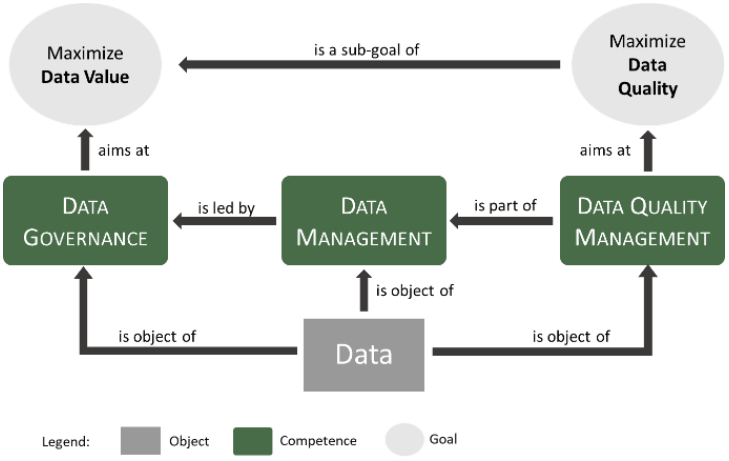


Figure 8 - Data Governance, Data Management, and Data Quality Management
Adapted from Sadiq (2013)

The goals of data governance are multiple, and some of them are listed below:

- Increase consistency and confidence in data-driven decision-making.
- Ensuring that data is used properly.
- Improve data security.
- Create and enforce data distribution within the company preventing data silos.
- Ensuring compliance with regulatory requirements.

Data governance became relevant to the subject since data quality is one of its key drivers. It supports data quality solutions ensuring that data quality standards and rules are defined and integrated into development and in day-to-day operations; ensuring that ongoing evaluation of data quality occurs; and ensuring that organization issues related to changed processes and priorities are addressed, paraphrasing John Ladley in *Data governance: how to design, deploy, and sustain an effective data governance program.* Thus, a data governance program must be set side-by-side with business objectives to ensure data quality.

2.5. DATA QUALITY IN DIGITAL ANALYTICS

The main goal of digital analytics is to drive a continuous process of the website/mobile application optimization, which triggers change, increasing site complexity – new pages, new products, new design, new features, new tools, and systems. Unavoidably, data quality is compromised.

Due to the nature of digital platforms, uncertain data quality is a given and a serious issue (Nelson, 2011). In fact, 44% of the digital professionals identified data accuracy as one of the top 3 data-related challenges (TMMData & Digital Analytics Association, 2017). Moreover, 32% of the digital professionals identified incomplete and inaccurate data as the biggest analytics-specific challenges (TMMData & Digital Analytics Association, 2017)**.** Since data supports business, poor data quality can negatively impact companies and generate serious consequences such as poor-quality decision-making, loss of income and business opportunities, contamination of data repositories (Customer Relationship Management, Data Lakes, Data Warehouses), and hence contamination of future analysis and predictions, decreased internal confidence, and loss of credibility (AT Internet & Digital Analytics Association, 2019). Without preventive action, data becomes dangerous.

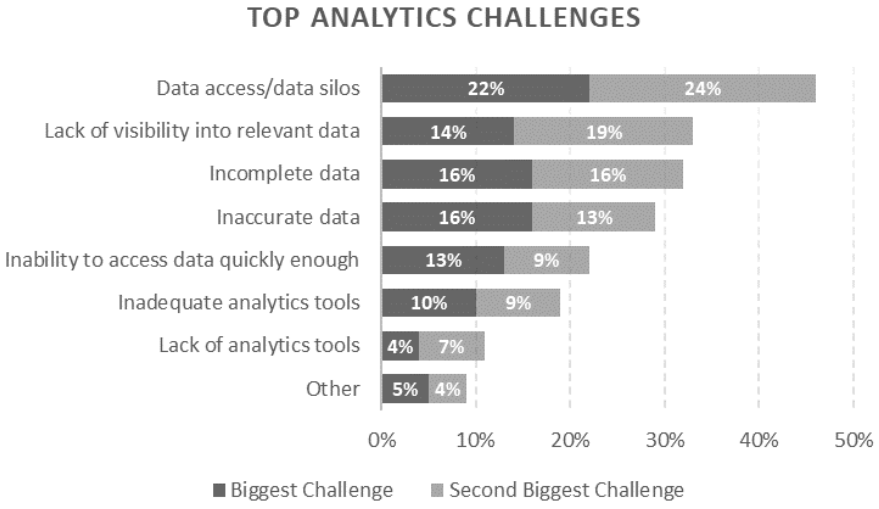


Figure 9 - Top Analytics Specific Challenges
Adapted from TMMData & Digital Analytics Association (2017)

Fortunately, it is possible to effectively implement strategies and procedures to manage and minimize the risks of uncertain data quality.

2.5.1. The Dimensions of Data Quality

To understand what data quality refers to and assess it, we must understand the dimensions that compose it. There are various identified dimensions for Data Quality (Sidi et al., 2013). According to the research, some dimensions such as accuracy, completeness, consistency, and timeliness are more referenced than others. The further detailed quality dimensions in digital analytics were mainly based on *Data Quality in Digital Analytics 2019* by AT Internet & DAA, crossing information from other sources.

1. Accuracy

   Accuracy refers to the degree to which the data measurement conforms to the correct value. The more accurate a measurement is, the closer it is to its true value. An equally important measure, but sometimes confused with accuracy, is precision, which refers to the repeatability or reproducibility of a measurement.

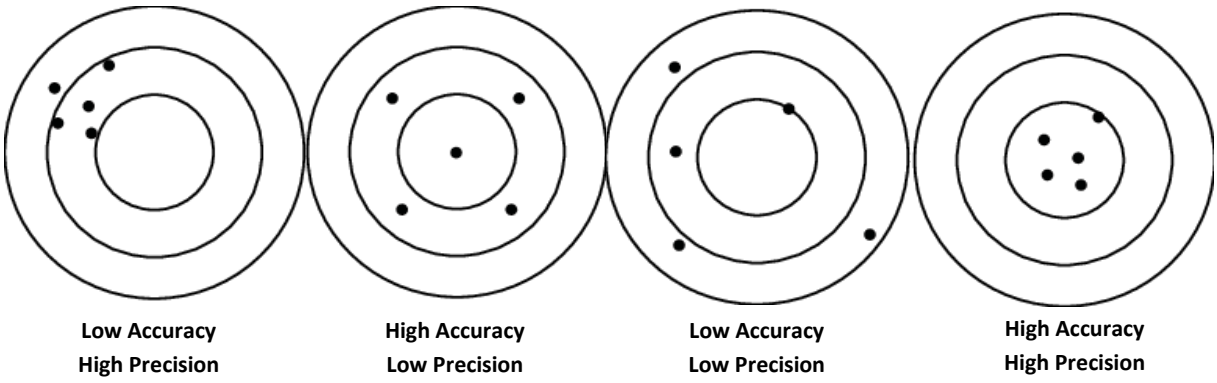


Figure 10 - Accuracy Versus Precision Metrics

   When thinking of this metric, the question is “does my data reflect reality over time?”. The data accuracy can be affected by several triggers, such as digital metrics calculations that depend on the analytics solution used and bot traffic that distort reality and impact the collected data.

2. Completeness

   For data to be truly accurate, it must be complete. Completeness must answer to “Is data sufficient enough to make informed decisions?”, “Is all data being collected as expected?”, “Is data missing or corrupt?”. Digital analytics data completeness can be affected, for example, by the unavailability of the data collection server, missing or broken tags/tracks, and insufficient data collection.

   *“*When data is missing or corrupt, you risk acting based on skewed information that doesn’t fully reflect reality.*”* (AT Internet & Digital Analytics Association, 2019)

   Missing or broken tags/tracks compromise data completeness impacting data accuracy. In digital analytics, this issue is pervasive, especially on large websites with a massive number of

pages that are constantly being upgraded and changed. Unfortunately, if a tag/track is missing, the user behavior for that specific situation is not being measured, constituting a rupture of the user journey. Regarding broken tags/tracks, sometimes, when the severity of the issue is high, the data is so obviously wrong that it is easy to identify and understand that data quality has been compromised (e.g., when the data is consistent from May to September and suddenly the numbers drop). However, these cases are not the most dangerous. The lesser the errors are, the less likely are they to be spotted. In the end, our analyses are being impacted by incomplete and inaccurate data without analysts being aware of the problem. Therefore, tag/track auditing and monitoring are critical – this topic will be explored further in the report.

> "If your data lacks a certain layer of information, such as geolocation data or information about the device used, you're missing out on a valuable piece of the picture."
> (AT Internet & Digital Analytics Association, 2019)

Another common problem that jeopardizes data completeness is the lack of enough data to provide the whole picture. This can result from a digital analytics tool that does not provide enough flexibility and capabilities to collect the needed information or the result of poorly and incomplete requirements.

3. Cleanliness or Validity

Despite all efforts to provide meaningful, consistent, and clean data, polluted data is inevitable due to the nature of the digital environment, which is constantly changing. In digital analytics, a small error in the tracks can drastically create and accumulate impure data with questionable value to the business. For example, in a mobile application feature development, the same attribute in the Android operating system might be defined with values in Portuguese. In contrast, in the iOS operating system, the values are delivered in English. This difference will cause data inconsistency as the same event is measured with different values depending on the operating system. This problem, in particular, can be easily worked around. However, it limits the analysis since the attribute structure was compromised. These errors arising from erroneous or outdated tracking values cause unintelligible or inconsistent formatting across events.

To overcome these problems and ensure clean and valid data, it is vital to define internal standards, rules, and procedures to ensure values conformity and standardization. Secondly, when problems are identified, it is necessary to apply an effective and agile procedure to deliver the correction as soon as possible. Once again, if data control procedures were applied, many problems would be avoided.

4. Timeliness

   Timeliness concerns the accessibility and availability of information when needed. Nowadays, real-time data is crucial as the world is converging to a culture of anytime, anywhere, any place where people consume, react, and decide quickly. What is exciting today, tomorrow is obsolete. If by the time data arrives at decision makers' hands, it is already too late to act upon, data becomes irrelevant. Therefore, markets must analyze data in real-time to identify and solve problems quickly, and access and match behavior between website and time context, analyzing the impact of each business decision to adjust strategies and react quickly.

5. Consistency

   Consistency means that data across all systems reflect the same information and are in synch with each other across the company.

   Nowadays, users' journeys flow across multiple devices and channels. As companies collect and analyze these data, capturing the entire digital footprint of users, they are often caught in having disparate data. Adding to this, companies also have a wide range of tools to calculate the reported metrics, leading to data discrepancies. According to the *Data Managers Feel Overwhelmed by Abundance of Tools* Report by Winterberry Group, the biggest challenge companies worldwide face concerning the day-to-day management of data is using too many different data management tools/systems in use. Some proposed solutions include opting for a single tool to measure user behavior across multiple devices and platforms and free up access to data within the company to avoid data silos[6] and data disparities between different company departments. However, these two solutions are not the responsibility of analysts, and although they can advise it, they cannot be dependent, neither make data consistency depend on these factors.

6. Compliance

   The General Data Protection Regulation (GDPR), in force since 2018, is a legally enforced European Union (EU) regulation on how organizations and companies use and maintain the integrity of personal data. It aims to regulate personal data protection and privacy of European users ("General Data Protection Regulation", 2021). The GDPR states a set of regulations based on informed consent to data processing from a data subject. To what concerns analytics, personal data includes all traffic and user behavior during a user's visit to a digital platform, including IP addresses, cookies IDs, GPS coordinates, and user's navigation flow.

   In the context of this report, compliance affects data quality in two opposite manners. On one hand, it fosters data accuracy, according to Article 5 point 1(d), personal data shall be

[6] *Data silo* "is a collection of information in an organization that is isolated from and not accessible by other parts of the organization." (Alley, 2018).

"accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');" ("Art. 5 GDPR – Principles Relating to Processing of Personal Data," n.d.). On the other hand, users must give consent for companies to collect their data. As some metrics in digital analytics depend on cookies, if users do not give their consent, digital analytics quality will be at stake. As an example, if the cookies were cleaned, a new unique visitor ID is counted, and a false representation of the audience will take place.

2.5.2. Organizational Pillars for High Data Quality

David Nelson in *Your Numbers Are Wrong: Ensuring High Quality Web Analytics Data 2011* defines three prerequisites for any digital analytics organization that are critical for ensuring high data quality.

1. Resources: Do organizations have enough human resources?
2. Skill sets: Do organizations have the right skills?
3. Commitment: Are organizations committed to ensuring data quality?

The absence or deficiency of any of these will ensure substandard data quality and ongoing deterioration. (Nelson, 2011)

These three prerequisites seem pretty simple and basic, but the reality shows that organizations are neglecting them. In the *State of Digital Analytics: The Persistent Challenge of Data Access & Governance,* a survey conducted by TMMData & DAA to 800 digital analytics professionals in July 2017, less than a quarter of the respondents said their organization had a governance program in place (14% didn't know and 19% was in the early stages of a rollout).

Figure 11 - Status of Organizational Data Governance
Adapted from TMMData & Digital Analytics Association (2017)

Although these prerequisites should be a priority to establish a data-quality-driven organization, they are not currently being prioritized, constituting a major barrier for establishing data governance.

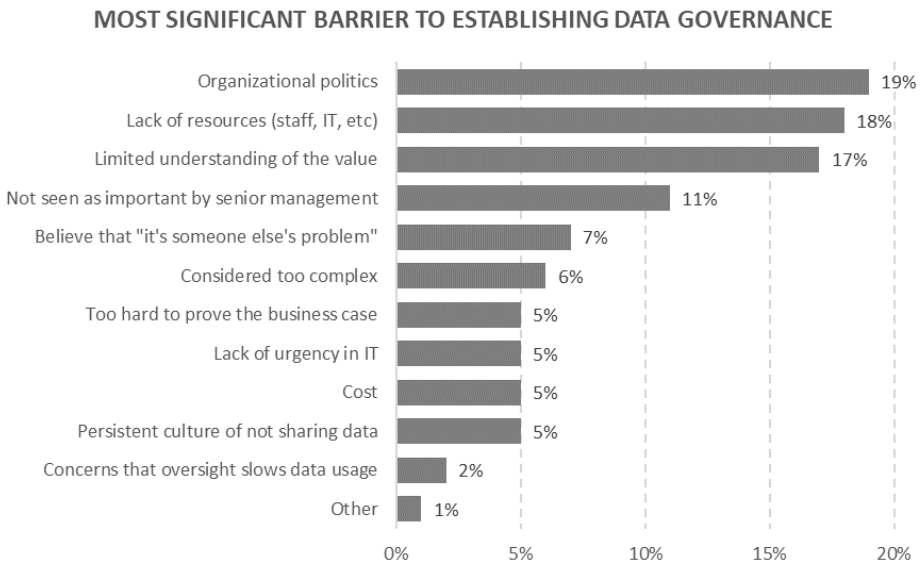


Figure 12 - Most Significant Barrier to Establish Data Governance
Adapted from TMMData & Digital Analytics Association (2017)

However, the survey findings also indicate that in addition to the 37% of organizations preparing to formalize data governance programs (Figure 11), hiring additional staff and providing professional development are among the top 3 investment priorities for the upcoming year.

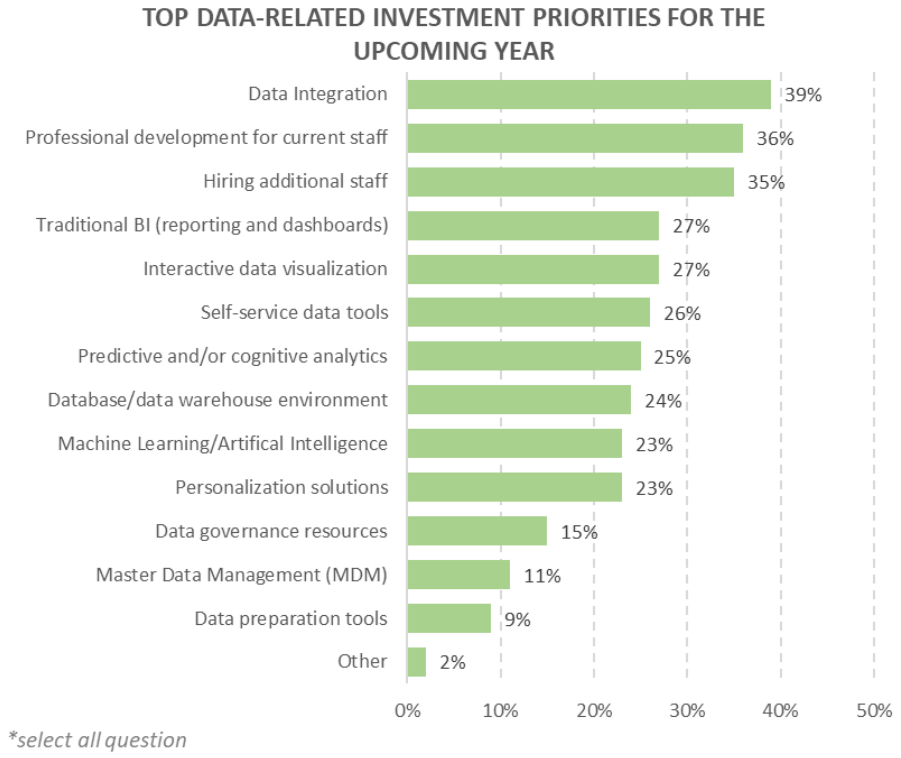


Figure 13 - Top Data-related Investment Priorities for the Upcoming Year
Adapted from TMMData & Digital Analytics Association (2017)

In summary, even though data quality is a main issue and organizations are becoming more data-driven, it is still being disregarded. Companies do not have the required capabilities to ensure it.

2.5.3. Data Control Procedures

To ensure high data quality, it is essential to work hard at each stage of the data process - collection, processing, storage, analysis, and insights. Some quality assurance (QA) processes can provide a solid foundation to drive data quality and value, stated by David Nelson in *Your Numbers are Wrong: Ensuring High Quality Web Analytics Data*.

1. Ongoing tag audits

   A tag audit is a systematic, comprehensive evaluation of a website's current tag configuration and execution. It combines automated site scanning and debugging tools. In a tag audit, a web crawler[7] software crawls the website and scans network requests from tags, recording where tags are and what data is being collected (Jason Call, 2017).

   Tag auditing allows identifying missing, duplicate, or broken tags, uncovering JavaScript errors, and creating a close-to-real-time map of custom variable tags throughout the digital platform. With this QA process, the organization becomes more confident with its tagging accuracy and its digital analytics implementation quality.

2. New tag/tracking implementation QA

   Typically, a software development process is constituted, among others, by the development, testing, and validation phases. There is a step for analytical tracking integration in the development phase in the case of a new website/mobile application feature or functionality development. Testing and validating digital analytics tags and tracking is crucial to ensure appropriate “site” tagging and tracking before new pages or functionalities go alive.

   Ideally, it should be a QA tester to perform the QA process since it will provide a more critical evaluation and avoid bias. If the developer also performs quality assurance, it will create unintended biases that lead to less critical evaluation and misuse of the highly skilled developers’ time and capabilities.

3. Troubleshooting data issues flagged by end-users

   Despite all efforts to ensure accurate tagging and tracking pre-deployment, it is frequent to detect unexpected and unintended data on a daily basis when digital analysts or end users are performing analysis and looking at the data. When this is detected, the bug should be reported to solve it as quickly as possible.

[7] A Web Crawler is a software application that runs automated tasks (scripts) on the web pages, typically for the purpose of web indexing. (“Web Crawler,” 2021).

3. METHODOLOGY

Despite all mechanisms that the literature states to mitigate the data quality problems faced in digital analytics, a question remains to be answered: which mechanism can be applied to have updated documentation and cataloging of all the pages and buttons that trigger events on the digital channels, to support analysts in reporting?

In Section 1.3 of this report, this problem was already stated "Since the cataloging of previous analytical data was not managed, there are some difficulties in understanding what exists, what is being measured, and what is the correct nomenclature of pages and events to perform the analyses" and it will be the focus of further work.

3.1. PRIOR AND RELATED ORGANIZATION'S WORK

To answer the previously stated question, a manual procedure was implemented in an early stage of the internship. The project was already in progress when the internship started, and it was focused on the company's App.

The solution was only available for the App as the tracking methodology applied is based on a JSON[8] file referencing all the triggered events. Once an event is triggered, the action ID inserted by developers at the app's source code is matched with the action variable of the JSON file, and the respective hit is triggered by the Tag Manager based on the JSON data. Taking as an example the page view of the App's Homepage: the developers at the production stage of the App's Homepage inserted the action ID "Page View Homepage" in the app's source; the JSON file has the event in Figure 14; the Tag Manager triggers the event in Figure 14 and sends it to the Adobe Analytics.


```
{
    "id": "1",
    "track_type": "view",
    "action": "Page View Homepage",
    "event": ["page_view"],
    "page_name": "Homepage App",
    "page_section": "App",
    "is_dynamic_udl": "true",
    "ignore":"true"
}
```


Figure 14 - First Event JSON file

As the JSON file contains all the App events for the new tracking system based on the Universal Data Layer (UDL), it is already a good baseline to document and use for reporting. However, it lacks some essential elements, for example, the respective page screenshot or the respective legacy's tracks that help cross information and get the full picture. Therefore, for complete cataloging, the JSON file must be combined with manual input.

[8] "JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate." (*JSON*, n.d.).

Hence, the solution is provided into a single excel document, “App UDL Catalog” and comprises three inputs: the JSON file, manual screenshots, and manual input/annotations. This workbook has two main worksheets: Data and Journeys. The Data worksheet comprises the JSON file transposed to a table, while the Journeys worksheet consists of the control table that keeps a record of all the available App’s sub journeys (). Each sub journey has its worksheet and its structure can be seen in Figure 16. The data in each sub journey worksheet is linked to the Data and Journeys worksheets using the JSON ID as a reference.

ID	Sorting ID	Journey	Link
1	1	Login	Login
2	2	Primeiro Acesso e Permissões	1st Access & Perm
79	3	Login: Dados Biométricos	Login>Biometrics
77	4	Dados Biométricos: Ativar: Login	Biometrics Activate>Login
80	5	Dados Biométricos: Ativar: Settings	Biometrics Activate>Settings
78	6	Dados Biométricos: Desativar	Biometrics Deactivate
3	7	Homepage Dashboard	HP Dashboard
4	8	Homepage Discover and Other Entries	HP Discover & Other Entries
66	9	Pesquisar	Search
39	10	Notificações	Notifications
6	11	Os meu Produtos e Serviços: Utilização	P&S Usage
7	12	Os meu Produtos e Serviços: Últimas Comunicações	P&S Last Communications
8	13	Os meus Produtos e Serviços: Carregamentos	P&S Topups
9	14	Os meus Prdutos e Serviços: Detalhes do meu Tarifário	P&S My Tariff
10	15	Os meus Produtos e Serviços: Promoções e Destaques	P&S Promo & Highlights

Figure 15 - App UDL Catalog - Journeys Worksheet Preview

	Page View - Produtos e Serviços	Click Event - Tab Detalhes	Page View - Detalhe Meu Tarifário	Click Event - Consultar outros tarifários
id	297	61	187	189
track_type	view	event	view	event
action	Page View My Plan Tabs Container	Event Click My Tariff	Page View My Tariff	Click Event My Tariff See More Tariff
event_action		Tariff Plan		Tariff Plan
event_category		Products and Services		Products and Services
event	page_view	event_instance	page_view	event_instance
event_label		See My Tariff Plan Detail		See More My Tariff Press
event_value	0	1	0	1
page_name	My Plan Tabs Container		My Tariff Plan Detail	
ignore	false	false	false	false

Figure 16 - App UDL Catalog – Products and Services Tariff Plan Sub Journey - Worksheet Preview

This solution became the App's Bible, and it is extremely helpful for analysts to perform reporting, giving a clear view of what is being measured and how can we make use of the collected data – what do the events correspond to, visually; which variables are being collected; historical data for each specific element (page or button). However, this solution requires a considerable amount of maintenance. Every week new pages are developed, and thus, new tracks are implemented, which have to be integrated into the Catalog as well as some annotations and screenshots. Every time a track is deprecated or affected, the Catalog also has to be updated.

Despite being a practical and fast solution in the short term, it does not provide the necessary scalability as new pages are developed, and new events are created. In addition to this, as it is a manual procedure, a manual validation has to be performed to check if an event is triggered, complete, and up to date. Consequently, the Catalog can become outdated without anyone being aware of it.

Nevertheless, the solution proves to be highly valuable - summarizes all the app events in production or deprecated providing a ground truth. It ensures completeness and cleanliness since every event needs to be formally validated before being documented. Consequently, it also contributes to data accuracy. The more validation, the more reliable the data is.

As was previously mentioned, one of the essential procedures to provide a solid foundation to drive data quality and value is to test and validate, which allows identifying if the tracking was correctly implemented. During these QA processes, analysts must critically think if more data should be collected and prevent some future problems of inconsistent formatting. These processes are fostered by the need to maintain the App Data Catalog updated.

3.2. Design Science Research Methodology (DSRM)

Having the App's tracking documented, there was an urge to develop a solution for the company's Website that would provide the same base information – all the web events and the respective collected data across the website. However, due to the website's dimension and volatility, a similar process to the App's procedure was unfeasible, which required more planning and research.

For this purpose, the Design Science Research Methodology (DSRM), although being an intended methodology for research and not usually seen in design in practice - "(…) for design in practice, the DSRM may contain unnecessary elements for some contexts, while being too much general to support design in others" (Peffers et al., 2007) -, was applied since its framework "is consistent with the DS research processes employed in the IS discipline" (Peffers et al., 2007) while providing an accepted and grounded process for the design of IT artifacts[9].

Paraphrasing Peffers et al., a methodology is a system of principles, practices, and procedures applied to a specific branch of knowledge that might help researchers produce and present high-quality

[9] "IT artifacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems)." (Hevner et al., 2004).

research accepted as valuable, rigorous, and publishable. In the context of DSR, a methodology would include three elements: principles to define "design science research", practice rules, and a process for carrying out and presenting the research.

3.2.1. Principles

"Design science creates and evaluates IT artifacts intended to solve identified organizational problems." (Hevner et al., 2004)

IT artifacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems) (Hevner et al., 2004). Summing up, design science is a problem-solving process, and the goal is to create a designed IT object to meet a business need.

3.2.2. Practice Rules

Hevner et al. defined seven guidelines that describe the requirements for effective design science research - "Researchers (…) must use their creative skills and judgment to determine when, where, and how to apply each of the guidelines in a specific research project. (…) each of these guidelines should be addressed in some manner (…)" (Hevner et al., 2004).

Guideline	Description
Guideline 1: Design as an Artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation
Guideline 2: Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems
Guildeline 3: Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
Guildeline 4: Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Guideline 5: Research Rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Guideline 6: Design as a Search Process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Table 6 - Design Science Research Guidelines
Adapted from Hevner et al. (2004)

These guidelines will be covered to support the work developed.

3.2.3. Design Process

"The design process is a sequence of expert activities that produces an innovative product (i.e., the design artifact). The evaluation of the artifact then provides feedback information and a better understanding of the problem in order to improve both the quality of the product and the design process. This build-and-evaluate loop is typically iterated a number of times before the final design artifact is generated (Markus et al. 2002). During this creative process, the design-science researcher must be cognizant of evolving both the design process and the design artifact as part of the research." (Hevner et al., 2004)

Based on Peffers et al., the DSRM process model is composed of six phases:

1. Problem identification and motivation: define the specific research problem and justify the value of a solution. Since the problem definition will be used to develop an artifact that can effectively provide a solution, it may be helpful to atomize the problem conceptually so that the solution can capture its complexity. Justifying the value of a solution accomplishes two things: it motivates the researcher and the research audience to pursue the solution and accept the results, and it helps to understand the reasoning associated with the researcher's understanding of the problem. Resources required for this activity include knowledge of the state of the problem and the importance of its solution.
2. Define the objectives for a solution: infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible. The objectives can be quantitative, e.g., terms in which a desirable solution would be better than current ones, or qualitative, e.g., a description of how a new artifact is expected to support solutions to problems not hitherto addressed. The objectives should be inferred rationally from the problem specification. Resources required for this include knowledge of the state of problems and current solutions, if any, and their efficacy.
3. Design and development: create the artifact. Conceptually, a design research artifact can be any designed object in which a research contribution is embedded in the design. This activity includes determining the artifact's desired functionality and its architecture and then creating the actual artifact. The resources required to move from objectives to design and development include knowledge of theory that can be brought to bear in a solution.
4. Demonstration: demonstrate the use of the artifact to solve one or more instances of the problem. This demonstration could involve its use in experimentation, simulation, case study, proof, or other appropriate activities. The resources required for the demonstration include effective knowledge of how to use the artifact to solve the problem.
5. Evaluation: observe and measure how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution to actual observed results from the use of the artifact in the demonstration. It requires knowledge of relevant metrics and analysis techniques. Depending on the nature of the problem venue and the artifact, evaluation could take many forms. It could include such items as a comparison of the artifact's

functionality with the solution objectives from activity two above, objective quantitative performance measures, such as budgets or items produced, the results of satisfaction surveys, client feedback, or simulations. It could include quantifiable measures of system performance, such as response time or availability. Conceptually, such evaluation could include any appropriate empirical evidence or logical proof. At the end of this activity, the researchers can decide whether to iterate back to step three to improve the artifact's effectiveness or continue communication and leave further improvement to subsequent projects. The nature of the research venue may dictate whether such iteration is feasible or not.

6. Communication: communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practicing professionals, when appropriate. In scholarly research publications, researchers might use the structure of this process to structure the paper, just as the nominal structure of an empirical research process (problem definition, literature review, hypothesis development, data collection, analysis, results, discussion, and conclusion) is a common structure for empirical research papers. Communication requires knowledge of the disciplinary culture.

It is important to note that in the literature, the solutions vary regarding Activity 4. Demonstration and Activity 5. Evaluation. Some authors use only one of the activities to prove the artifact works or formally assess its value, while others include both phases on the methodology.

Although this process is described in nominal sequential order, due to the nature of a design process, it provides the necessary flexibility to perform the design. Researchers can start in almost any step. The Design Process is pictured in Section 8.4 of the Appendix Chapter.

3.3. DEVELOPING A WEB DATA CATALOG PROTOTYPE

The DSRM process model for the Web Data Catalog artifact is fully described in this subchapter.

3.3.1. Problem Identification and Motivation

In every user interaction, tons of data are collected: the page name that the user interacted with, the respective page section, the URL, the browser he is using, the name of the clicked button, among other information. All the data is gathered into the hit and sent to the analytics tool every time the user interacts with a website element that is being tracked. This data is then the source of multiple analyses that provide insights to support decision-making.

In the organization, this cycle has been active for years, and data has been piling. If websites' volatility is added to the cycle, the data situation can get ambiguous. The inherent changing environment causes new pages and new buttons to be added as new products appear or the opposite. Pages improvements/adjustments are also an effect of user experience and interaction design improvements that consequently affect data since it is altered to meet the definition and characteristics of the new components.

With the growing digital world, the increase in collected data has been exponential, and it has come to a point where analysts are disoriented - they do not know where the data comes from, which data they have at their disposal, and how to make use of it thoroughly and insightfully. This problem induces more issues on the data quality dimensions – is the data accurate? Does it reflect reality? Is data missing or corrupt? Is data valid?

Consider the following hypothetical scenario to understand the dimension of the problem. A website has 200 different pages and 100 buttons, each user makes three visits to the website in a month, and in each visit, five pages are visited, and three buttons are clicked. This fictitious website has 300 000 visitors each month, and, in each hit, 20 different variables are collected, which can admit multiple values. Assuming the previous scenario, in each month, the website statistics reveal 300 000 visitors and 900 000 visits, 4 500 000 page views, and 2 700 000 event hits, a total of 7 200 000 hits and 144 000 000 data values (this number has duplicate unique values). These data points are a lot of data to work on and, given the nature of a website, it is even harder to make sense of it as the website grows and changes. To give a hint of what digital analysts are subject to, the company's website does not have 300 thousand visitors, it has, approximately, 1.3 million visitors and 2 million visits, more than 200 pages, and 100 buttons, and the number of variables collected in each hit varies but are for sure more than 20.

To help analysts be more productive and enhance the data quality in this challenging and changing environment, a solution must be developed to understand what the collected data corresponds to on the website and track changes over time.

3.3.2. Objectives Definition

Currently, the primary source of knowledge of the website's digital data is the operation knowledge retained by employees, which is at risk if they leave the company. Besides this, a testing procedure can be performed whenever analysts are in doubt or want to perform some reporting on new features. Still, it is only valid for the test's timeframe and can be cross-validated using the data in the analytics tool to understand when the event has been triggered. The testing involves interacting with the website and checking the event using debugging tools to analyze the data gathered into the hit.

The analytics tool can also be used to understand where the data is being triggered by performing some debugging. Still, it is not as straightforward as it seems because when the webpages' elements are similar, they can derive similar data. Looking only at the data, sometimes it is difficult to understand where it comes from. Therefore, it is quite challenging to come to a conclusion. For instance, the event label of an event is "Close summary", and we want to understand if this event label refers to closing the product summary page. When the event label is checked in the analytics tool, it references two different event categories, "Product Detail" and "Product Summary". In this case, we should perform more debugging to be sure of the root.

As the three solutions described above to help analysts in reporting and to ensure data quality are not robust and can be time-consuming, it is crucial to find a better solution. Given the problem that digital analysts were facing that jeopardizes and limits their daily work, the objective was to develop the scope of DSRM as the artifact, a Web Data Catalog that would combine the visual component on the website and the respective data collected. Every time a new webpage/element is created or an existing webpage/element is improved, it should be included in the documentation. Additionally, considering the website's dimension and volatility, it should be an automated process with minimal user intervention, avoiding manual and recurrent processes to keep the Catalog updated.

The solution should retrieve all the data objects triggered (Tealium and Adobe Analytics) and the visuals associated with each event. All the props and eVars collected should also be matched by a reference table that provides the respective Tealium's and Adobe Analytics' variable to cross-related how the data is transformed and distributed. The database should keep all the records (historical and current events), and they should be assigned to a Journey and Sub-journey (Category and Subcategory) for easy navigation. The final product should be delivered in a way that is easy for end-users to browse through the data using a dynamic and intuitive interface.

The solution developed should support analysts in reporting, being easier to understand which dimensions and values to use, and keeping a record of the historical data when changes are faced. It must also become a quality assurance source that will enhance data accuracy, completeness, and validity.

3.3.3. Design and Development

3.3.3.1. Generic Considerations

Taking into account the problem, motivation, and objectives, there were several solutions on the table.

The first solution considered but immediately excluded at the beginning of the discussion was to perform a manual procedure similar to the App Data Catalog. However, adding to the drawbacks of the solution already exposed in Section 3.1, the Website does not have a JSON file referencing all the triggered events. The App and Web tracking frameworks are slightly different. While the App tracking is based on the JSON file and the developers, only have to hard code the track ID and some context-specific variables. All the Web tracking is hard-coded by developers, which limits the production of a similar solution.

The second solution was to extract the raw data from Adobe Analytics using the Data Feeds functionality, load this data into an ETL framework, and build a structure that would feed a web application where end-users could rapidly access and cross information with minimum knowledge on databases queries. This process would be implemented by scheduling data feeds extractions. The Data Feeds pull hit-level data from Adobe Analytics without processing it.

The third and last solution is constituted by a web crawler that would navigate through the primary/main domain, collecting all the data objects and screenshots of the web pages as previously defined in the Objectives Section, storing the information retrieved in a database at the hit-level (one event for each page or click). Then, the data would be processed and cross-related to understand if the events reflect the same action (i.e., the same page view in different timeframes with different page names) being aggregated in an easy-to-read and interpret form. This data would be the source of an interactive web application to navigate over the data.

Between the Data Feeds solution and the Web Crawler solution, there were some pros and cons. The Data Feeds solution would allow having every single event across the website over all the customer profiles (business, consumer, prepaid, postpaid, fixed). However, the extractions from the Data Feeds would be delayed one week or more. They would contain an extreme amount of data that would have to be stored in a robust Data Warehouse and have a considerable processing power to perform the ETL. Additionally, the data would be duplicated multiple times. The data feed extracts all the hit-level data regardless of two or more hits being exactly the same from different users and timeframes. On the other hand, the Web Crawler would crawl the website at scheduled times and retrieve all the data objects one time, allowing to check for errors and navigate through the data with low latency. Nonetheless, this method also has limitations regarding logged areas that would require user input to solve reCAPTCHA[10] and features that depend on customer profile, requiring multiple logins to collect all the data views.

Considering the pros and cons, the Web Crawler solution was chosen since the Data Feeds require a framework that the Team does not possess, and the Web Crawler prompts an artifact design and development to prove its feasibility.

3.3.3.2. Artifact Design and Development

Although the Web Crawler was the adopted solution, the development process comprises more than the web crawler development, as the final goal was to produce a Web Data Catalog.

As it is known, there are two main types of hits - page views and event hits. For the objective at hand, to develop a complete Data Catalog, all the events must be accounted for and retrieved. However, due to the complexity of replicating all the click events across the website, and collecting the respective data objects, only page views were accounted for in this prototype.

The development process comprises three main objects/processes – the Web Crawler, the aggregation process, and the Web Application to easily visualize and navigate through the data. The programming language used along the project was Python.

[10] "reCAPTCHA is a free service from Google that helps protect websites from spam and abuse. A "CAPTCHA" is a turing test to tell human and bots apart." (Google, n.d.).

Regarding the Web Crawler, some research was conducted to understand how it should be defined and constructed for the objective at hand. Using as reference *Testing Adobe Analytics with Python* 2017 by John Simmons, the process was based on the Selenium package (Bowen, 2020) to instantiate a headless web driver (the browser object without the user interface) and the BeautifulSoup package (Richardson, 2021) to parse the HTML page allowing to retrieve the absolute[11] and relative links[12] . After getting the links on the first page (i.e., the page of the primary domain), a recursive crawl is performed to retrieve all the links on every page of the company's website. Then, the links retrieved pass through a verification process to ensure they are within the domain, they are HTML links and not ZIPs, images or PDFs, and other particular business conditions. At the end of the crawler's first phase, a list of links from where the data will be collected is saved on a JSON file and other JSON documents for debugging.

In the web crawler's second phase, another web driver is instantiated to navigate through the retrieved links one at a time, collecting the Tealium and Adobe Analytics data objects and taking the web page's screenshots. For each event, there is a composite primary key constituted by the canonical URL ID that is incremented for each new canonical URL[13], the timestamp of the crawling process, the respective page name, and an aggregation ID that initially is equal to the canonical URL ID. Still, it can be changed manually to ensure a correct aggregation, as will be further explained. Before starting the navigation, when the web driver is instantiated and initialized with graphical display, the web driver interacts with the page accepting the cookies layer to proceed to the website, and the user manually logs in, solving reCAPTCHA. On each page, during data collection, the data is cross-referenced with a data dictionary to match the props and eVars with their respective Adobe Analytics names (an example of the matching props and eVars Adobe and Tealium names can be seen in Figure 17). The data collected is saved in a JSON file. The canonical URLs dictionary (Figure 18) and the dictionary that provides the match between canonical URL ID and page name (Figure 19) are also extracted to a JSON file to serve future recurring cataloging.

```
{
    "eVar1": {
        "tealium": "search_terms",
        "adobe": "Search Terms"
    },
    "eVar2": {
        "tealium": "search_results",
        "adobe": "Search Results"
    }
}
```

Figure 17 - Adobe Analytics and Tealium's variables mappings Snippet

[11] Absolute links provide the full website address always including the domain name of the website. (CoffeeCup Software, 2017).

[12] Relative links do not provide the full website address, instead it is the location based on the page where the link appears. (CoffeeCup Software, 2017).

[13] "A *canonical URL* is the URL of the page that Google thinks is most representative from a set of duplicate pages on your site." (Google, 2021).

```
{
    "1": "https://www.telcocompany.pt",
    "2": "https://www.telcocompany.pt/5g.html",
    "3": "https://www.telcocompany.pt/pacotes/televisao/lista-canais.html",
    "4": "https://www.telcocompany.pt/pacotes.html",
    "5": "https://www.telcocompany.pt/loja/acessorios.html"
    "6": "https://www.telcocompany.pt/loja/cartoes-sim.html",
    "7": "https://www.telcocompany.pt/loja/internet-movel.html",
    "8": "https://www.telcocompany.pt/loja/smart-home.html"
}
```


Figure 18 - Canonical URLs Dictionary Snippet


```
{
    "1_Consumer Homepage": "https://www.telcocompany.pt",
    "2_5G": "https://www.telcocompany.pt/5g.html",
    "3_Channel List": "https://www.telcocompany.pt/pacotes/televisao/lista-canais.html",
    "4_Fixed Bundles": "https://www.telcocompany.pt/pacotes.html",
    "5_Accessories": "https://www.telcocompany.pt/loja/acessorios.html",
    "6_SIM Cards": "https://www.telcocompany.pt/loja/cartoes-sim.html"
    "7_Mobile Internet": "https://www.telcocompany.pt/loja/internet-movel.html",
    "8_Smart Home": "https://www.telcocompany.pt/loja/smart-home.html"
}
```


Figure 19 - Canonical URLs & Page Name Dictionary Snippet

It should be noted that in the web crawling phase, two different crawlers are instantiated at different stages. Although it could have been instantiated one single crawler processing them as consecutive processes, we found it inadequate to use this approach for multiple reasons:

1. The processes require different web driver conditions that directly affect performance. While data collection requires a graphical user interface for the user to log in, the links extraction process only needs a headless web driver, i.e., a web browser without the graphical user interface, which improves speed and performance.
2. Separating the links extraction from the data collection provides the necessary flexibility for debugging, understanding step by step the outputs from each one.
3. With consecutive processes, they would directly affect each other, and as the links extraction output is the input of the data collection, reorganizing them as subsequent processes avoid bottlenecks.

The next step in the development was the aggregation process. One of the Data Catalog objectives is to track changes and keep a chronological record of the events. To meet this objective, after collecting the raw data, the data must be aggregated to easily interpret the changes that reflect the same action (i.e., the same page view in different timeframes with different page names).

Before going into detail in this process, it must be understood what the expected output of the aggregation is and its assumptions. For a prototype and considering the business needs, in the aggregation, the page name is the only variable where a historical record is kept, while the other variables always correspond to the first canonical URL's event. Subsequently, more complexity layers can be added, but we proceeded with this method, due to time restrictions. In Figure 20, a trivial

example is given to better understand how the aggregation process works. The aggregation process is performed by the aggregation ID variable and this decision was heavily discussed, as further exposed.


```
{
    "1_Consumer Homepage": {
        "_id": {
                "canonical_url_id": 1,
                "page_name": "Consumer Homepage",
                "timestamp": "2021-05-19 11:28",
                "aggregation_id": 1
        },
        "canonical_url": "https://www.telcocompany.pt",
        "category": "Homepage",
        "page_section": "Consumer Homepage"
    },
    "1_Homepage Vodafone Site": {
        "_id": {
                "canonical_url_id": 1,
                "page_name": "Homepage Site",
                "timestamp": "2021-06-01 09:43",
                "aggregation_id": 1
        },
        "canonical_url": "https://www.telcocompany.pt",
        "category": "Homepage",
        "page_section_full": "Homepage Site"
    },
}
```

```
{
    "1": {
        "page_name": {
            "2021-05-19 11:28": "Consumer Homepage",
            "2021-06-01 09:43": "Homepage Site"
        },
        "canonical_url": "https://www.telcocompany.pt",
        "category": "Homepage",
        "page_section_full": "Consumer Homepage"
    }
}
```


Figure 20 - Trivial Example of the Aggregation Process Input (Left) and Output (Right) Data

Initially, the outputs from the web crawler would be stored in a MongoDB database, as the data is stored as documents like JSON files, supporting key-value pairs. However, as the aggregation process is extremely business-oriented and not a real “group by" as can be seen in Figure 20, a MongoDB database would limit it. Since we wanted to continue to explore the design process and study the artifact’s feasibility, and a system of files and folders is the least of our problems, we decided to store the data in JSON files.

When the Team arrived at the aggregation step, we found that this process was more complex than initially thought, and some adjustments had to be made to meet the business needs. On the website, concerning page views, the unique identifiers of pages are the canonical URL and the non-existent variable aggregation ID. Still, at first, we made the wrong judgment of considering only the canonical URL. While different pages within the same canonical URL are possible, especially in logged areas and forms, the aggregation cannot be only based on the canonical URL since it would lead to aggregating pages that do not reference the same object. The solution found to overcome the lack of an identifier was to create the aggregation ID variable that is initially equal to the canonical URL ID but can be edited to correctly specify the group where the event belongs. Using this system, the only situation is the need to break down the aggregated object into smaller objects, edit the aggregation ID, and reprocess the aggregation. Hence, two events from different canonical URLs cannot reference the same object.

Finally, after the final aggregation’s output, it is time for the web application to take place. To develop the web application, the Streamlit package was used (Streamlit Inc, 2021). Streamlit is a relatively new open-source app framework for Machine Learning and Data Science teams and focuses on rapid prototyping. It is simple and intuitive to work with, although there is still a lack of some design/product

functionalities. Among the available Python options, Streamlit seemed more promising and adequate for the use case at hand.

The web application reads and outputs the aggregation JSON file and the pages' screenshots. To navigate through the data, there are multiple filters: the Category, and Subcategory of the journey, the URL, the page name, and a selector box that identifies the added or changed events since the last crawl. As the application is an internal document of the company's Digital Analytics Team, the Streamlit app was not published in production. Instead, each time the users desire to navigate the data, they have to run the app using a batch file[14] (the only requirement is to install the needed Python packages placed into a text file). The web application contains some references to the company and logo, and, therefore, the final interface system can not be depicted in this report.

3.3.3.3. Recurring Cataloging Process

After the first version of the Web Data Catalog, the entire process has to be performed from time to time to check for changes and record them in the Web Data Catalog, keeping it updated. However, the recurring cataloging process requires additional steps in the previously explained procedure. The web crawling's first phase – the links extraction – remains unchanged, as well as the data collection in the case where new page names are observed.

In each run of the complete process, two JSON files are stored, as seen above in Figure 18 and Figure 19. The canonical URLs and page name JSON provide the unique key that determines whether a new event was created or not. Based on the relation between these variables, the crawler logic defines if the data should be collected or skipped. If the combination "<<canonical URL ID>>_<<page name>>" does not exist, the data is collected; if the combination exists, the crawler checks if the URL is already accounted for in the aggregation data for this combination - if the URL was not included, collects a simple document as the "5_Accessories" item in Figure 21; otherwise, the crawler continues its path going to the following URL without collecting any data.

[14] A *batch file* is "a computer file with sequential commands to be executed when the file is read." (*Batch file definição e significado | Dicionário Inglês Collins*).

```
{
    "5_Accessories": {
        "_id": {
            "canonical_url_id": 5,
            "page_name": "Accessories",
            "timestamp": "2021-06-30 16:19",
            "aggregation_id": 5
        },
        "canonical_url":"https://www.telcocompany.pt/loja/acessorios.html",
        "urls": [
            "https://www.telcocompany.pt/loja/acessorios.html?segment=consu
            mer&category=Smart%20Home"
        ]
    },
    "193_Apple AirTag Radish Loop em Pele - Loja Online": {
        "_id": {
            "canonical_url_id": 193,
            "page_name": "Apple AirTag Radish Loop em Pele - Loja Online",
            "timestamp": "2021-06-30 16:09",
            "aggregation_id": 193
        },
        "category": "Loja",
        "subcategory": "Acessorios",
        "canonical_url": "https://www.telcocompany.pt/loja/acessorios/
                          outros/apple-airtag-radish-loop-em-pele.html",
        "image_link": "193_AppleAirTagRadishLoopemPeleLojaOnline.png",
        "events": "prodView,event10",
        "urls": [
            "https://www.telcocompany.pt/loja/acessorios/outros/apple-
            airtag-radishloop-em-
            pele.html?color=castanho&paymentType=pvp&segment=consumer"
        ],
        "props": { (...) },
        "evars": { (...) },
        "utag_data": { (...) }
    }
}
```


Figure 21 - Output Data Collection Snippet (Input to the Aggregation Process)

After the data collection, the aggregation process has to be rearranged due to the documents that differ in structure. In this case, a document is a digital event – using Figure 21, “5_Accessories” and "193_Apple AirTag Radish Loop em Pele - Loja Online" are two different documents. For each document that owns the structure of the “5_Accessories”, the aggregation process retrieves the URLs list and appends it to the URLs list of the aggregation document for the respective key “<<canonical URL ID>>_<<page name>>”. The remaining documents are processed as previously explained, aggregating by aggregation ID.

Summing up, in each recurring cataloging process, the web crawler itself just outputs the data that brings changes to the stored data. Then, the data is processed and aggregated, meeting the structure and objectives defined. Finally, the web application reflects the last aggregation data.

3.3.4. Evaluation

The DSRM Evaluation phase will be explained in the following section - Section 4.

3.3.5. Communication

Following the DSRM, the communication is reflected on this internship report, where the problem is communicated, the artifact is explained, and its development process described, along with its utility, drawbacks, and limitations.

4. RESULTS AND DISCUSSION

Once the artifact was fully developed, the complete process was explained and tested by the Team's members. They suggested some add-ins in the web application just to complement the available filters facilitating navigation – a search box that would allow searching by page name and canonical URL, and also a filter that would allow identifying the events that were added or changed since the last crawl.

Their feedback regarding the artifact, which includes the process, and the web application, was positive. The Team considered it a significant step towards the future, being able to test, audit, and document digital events automatically. It provides a foundation on how to start and what to use, and more layers can and should be added to enrich the process and make the most complete and robust solution.

The artifact is a partial solution to the problem providing a Web Data Catalog in an easy-to-read and intuitive interface where the events are organized by Category, Subcategory, and Canonical URL. There is a match between the props and eVars IDs and the respective Adobe Analytics variables, and the page names in page views, historical and current, are being kept. Additionally, there is an exact match between the events and the visual component, and there are no duplicate events in the data. Hence, the artifact only records data changes.

> "Artifacts constructed in design science research are rarely full-grown information systems that are used in practice. Instead, artifacts are innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, (..)."
> (Hevner et al., 2004)

The main objective of the design process was to define and test the idea, procedure, and technical capabilities. Accordingly, to get a minimal viable product at the end of the design process, some decisions were made during the development to narrow the search space:

1. From all the available digital events, only the page views were accounted for in this artifact, excluding a large part of the digital events space (event hits and others).
2. From all the collected variables, only the historical records of the page name variable are documented. The user has to search on the output JSON files from the web crawler for the remaining variables, which is time-consuming.

The second point above is a consequence of the data structure and the aggregation process. This point was one of the difficulties faced during the design process: the two-dimensionality (variables/values and timestamp) complicated the data structure to adopt. There is still some research to be performed on whether to restructure the aggregation process or rethink the data structure.

Additionally, some defects were found:

1. The links extracted from the website's HTML do not include every possible page on the website, which narrows the search space and does not provide the full picture of the website.

2. The process development does not restrictively retrieve page views. In fact, it retrieves the last triggered event on the page. In typical scenarios, only a page view is triggered when accessing an URL, but sometimes multiple events can be triggered. In the first case, the artifact retrieves the page view, but it might retrieve another hit type in the second case.

Regarding the system's performance, the links extraction takes about 1 hour, while the data collection takes approximately one hour and a half. For now, it does not take too long to execute, but when more layers are added, the panorama can change. The web crawler can still be optimized in the future.

The web crawler is vulnerable to the network connection and the website availability, which sometimes can be a constraint. Furthermore, it is also vulnerable to website modifications regarding the cookies layer and the login as the web crawler interacts with these elements using their encoded address – if the element changes, usually its encoded address also changes. Despite these minor concerns, they derive from the nature of the problem and should not constitute an obstacle to disregard this solution.

Finally, regarding the web application, configuring the Python for the batch file to work can also be a slight drawback because, although it is conceptually simple, installing and configuring systems sometimes can be quite challenging. Nevertheless, from the moment the configurations are set, the response time and availability of the web application are almost immediate.

> "Constructing a system instantiation that automates a process demonstrates that the process can, in fact, be automated. It provides "proof by construction"." (Hevner et al., 2004)

Considering the artifact issues found, some adjustments and improvements are to be made on this prototype to provide the expected outcome at this stage. These further improvements will be left to subsequent projects as the project's main objective was accomplished – study the feasibility of a solution to support digital analysts in data quality control through Data Cataloging.

5. CONCLUSIONS

The main objective of this project was to research data quality good practices, find procedures to enhance it, and, ultimately, develop a proof-of-concept about digital data cataloging.

The research contributed to discovering strategies to enhance data quality such as tag/tracking audit and monitoring; the definition of internal standards, rules, and quality assurance procedures to ensure tracking quality and values conformity and standardization; and perceiving the importance of collaborative communication across the organization. It also constituted a learning opportunity, consolidating digital analytics concepts and deepening the technical details on the digital analytics framework and implementation, which facilitated the daily internship's tasks and fostered the Team's debates towards better communication and growth.

On the data quality side, the research provided the fundamental push to enhance it, providing clear strategies that were applied or that are being considered from now on such as exclude bot traffic cleaning the old IPs and inserting new ones; rigorous tracking requests and validations; tag/tracking monitoring procedures more regularly, reporting the issues found right away. Based on the final meeting review, the general Team's opinion was that, during the past year, the data quality and reliability increased substantially, as well as data control procedures that contributed to data consolidation. Consequently, Teams' organization grew, and data silos within the Team decreased.

The developed project provided a proof-of-concept on automated data cataloging, opening up horizons to a wide range of tools and mechanisms that should be explored to complement and strengthen data quality. Even though the data catalog system is fully functional and can be a fundamental support document for digital analysts and other functional teams, it is yet an embryonic information system that needs further work to become sufficiently complete and robust. Hence, the problem described at the beginning of this report – difficulties in understanding which analytical data exists, its source, and terminology to perform analyses - was partially covered as it was provided a prototype of which path to take to achieve this goal. Furthermore, the methodology adopted provided a strong foundation to the artifact's development contributing to narrow the scope of the project and strictly define the requirements (Phase 1 and 2), and to assess the artifact's utility, quality, and efficacy with rigor to provide essential feedback and improve the product quality and the design process.

Regarding data quality, the developed prototype works as an ongoing tag and track auditing system, which systematically collects the data on the website providing a base view of the changes along the time in the data and the errors identified (e.g., "Error 404 Page Not Found"). This contributes to data completeness and accuracy as missing or broken tags and tracks monitoring becomes easier when applying this system. Additionally, having the digital events gathered in one place cross-referenced with their visual element helps analysts to spot potential issues when the data at hand is not enough to provide the full picture or when it is unintelligible or inconsistent formatted, which in turn, fosters data completeness, cleanliness, and validity and, hence, data quality.

This project required a set of skills and knowledge that were not mastered when the project started, namely, JavaScript, HTML, Web Crawling, and Web Applications Development. The acquisition of these skills and knowledge constituted a challenge and a learning opportunity. Additionally, the nature of the data and the business conditions were and are two other main challenges since the data did not and does not possess a unique identifier (single or composite primary key) to facilitate the events cross-relationship and, consequently, the customized aggregation, which in turn, must meet the events cross-relationship and the business needs to provide a historical documentation structure crossing the timestamp and each variable.

In summary, it turned out to be a challenging and rewarding project that stimulated the Team's growth and personal development at the level of work pace, research, autonomy, and resilience to keep looking for solutions and alternatives that could fit the objective and problems at hand.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Limitations were found during the artifact's development, and some decisions were taken that narrowed the development space and the project's dimension.

Due to the website's dimensionality, and since the design process must produce a viable artifact, we had to restrict the development space narrowing it only to include the page view hits on the Data Catalog, which constituted a constraint for the design process. By making this decision, a considerable amount of the total hits triggered on the website were left out. Adding to this, the page name is the only variable where historical values can directly be seen on the web application. In theory, historical values should be easily seen through all the variables, but in practice is much more complex to create the right data structure to visualize this data as previously explained. Accordingly, one of the key points to explore in the future should be to research a data structure that can support the custom aggregation of events allowing historical data, or rethink the aggregation process in a way digital analysts can easily read the data on the web application.

One of the main limitations is the data storage adopted – a system of files and folders – that, due to the lack of support tools and time restrictions, was the easiest way out, but not the best long-term solution. The main reason for it was the need to develop an artifact that the Team could easily use without additional tools requiring licenses and budget. This limitation should be one of the key points to solve in further improvements along with the data structure issue.

Besides the future work listed above, other issues must be addressed in the future to improve the web data catalog. Below future improvements and some recommendations can be found.

1. Comprise every available URL on the website to get the full picture of the website. The analytics tool can be used to extract all the possible links and feed the web crawler's data collection combining the web crawler's output with the analytics tool.
2. Automate the login using a trusted cookie ID that does not ask for reCAPTCHA. Consequently, the data collection could use a headless web driver that would improve performance.
3. Retrieve every event when loading the page, be it a page view or any other event type (in this case, the events will be collected when an instance of a page is being loaded; it still excludes click events). To do so, the solution hypothesis will be to queue the triggered events in a data object explicitly established for the web crawler (using a specific cookie or a custom user agent[15]) being able to collect this data object afterward.
4. Collect all the user interactions events (page views, event hits). This improvement is complex and should be evaluated using a cost-benefit ratio to understand if it is crucial to collect this data since, generally, at the company, data from page views is most commonly used than the remaining hit types. The solution to explore would be similar to the previous point.

[15] "Any software that retrieves, renders and facilitates end user interaction with Web content, or whose user interface is implemented using Web technologies." (*Definition of User Agent - WAI UA Wiki*, 2011).

7. BIBLIOGRAPHY

Adamiak, A. (2020, April 22). 2020 Web Analytics Survey Results | Business Ahead. *Https://Businessahead.Co.Uk/*. https://businessahead.co.uk/2020-web-analytics-survey-results/

Adobe. (2019, June 26). *Compare Props and eVars*. https://helpx.adobe.com/pt/analytics/kb/compare-props-evars.html

Adobe. (2020a, 7). *Page views | Adobe Analytics*. https://experienceleague.adobe.com/docs/analytics/components/metrics/page-views.html?lang=en#metrics

Adobe. (2020b, October 5). *Unique visitors | Adobe Analytics*. https://experienceleague.adobe.com/docs/analytics/components/metrics/unique-visitors.html?lang=en#metrics

Adobe. (2020c, November 5). *Visits | Adobe Analytics*. https://experienceleague.adobe.com/docs/analytics/components/metrics/visits.html?lang=en#behavior-that-affects-visits

Alley, G. (2018, July 20). *What are Data Silos? | Alooma*. https://www.alooma.com/blog/what-are-data-silos

Analytics Demystified. (2015, June 24). *A Framework for Digital Analytics Process*. Analytics Demystified. https://analyticsdemystified.com/analytics-strategy/framework-digital-analytics-process/

Analytics Market. (2019, January 26). *Google Analytics Definitions*. AnalyticsMarket. https://www.analyticsmarket.com/blog/google-analytics-definitions/

Art. 5 GDPR – Principles relating to processing of personal data. (n.d.). *General Data Protection Regulation (GDPR)*. Retrieved January 26, 2021, from https://gdpr-info.eu/art-5-gdpr/

AT Internet. (n.d.). *Digital analytics definitions: AT Internet glossary*. AT Internet. Retrieved January 10, 2021, from https://www.atinternet.com/en/glossary/

AT Internet & Digital Analytics Association, 2019. Data Quality In Digital Analytics (38). Retrieved December 8, 2020, from https://content.atinternet.com/data-quality-in-digital-analytics/

Batch file definição e significado | Dicionário Inglês Collins. (n.d.). Retrieved August 5, 2021, from https://www.collinsdictionary.com/pt/dictionary/english/batch-file

Bekavac, I., & Garbin Praničević, D. (2015). Web analytics tools and web metrics tools: An overview and comparative analysis. *Croatian Operational Research Review*, *6*(2), 373–386. https://doi.org/10.17535/crorr.2015.0029

Bowen, D. (2020). *selenium-tools: Datetime and range slider tools for python selenium* (0.0.2) [Python; OS Independent]. https://dsbowen.github.io/selenium-tools

Chen, I. (2017, August 5). *What Is Tagging In Digital Analytics?* https://www.linkedin.com/pulse/what-tagging-digital-analytics-ivan-chen

CoffeeCup Software. (2017, September 6). *Absolute vs. Relative Paths/Links*. CoffeeCup Software. https://www.coffeecup.com/help/articles/absolute-vs-relative-pathslinks/

Definition of URL | Dictionary.com. (n.d.). Www.Dictionary.Com. Retrieved August 11, 2021, from https://www.dictionary.com/browse/url

Definition of User Agent—WAI UA Wiki. (2011, June 16). https://www.w3.org/WAI/UA/work/wiki/Definition_of_User_Agent

Fleckenstein, M., & Fellows, L. (2018). *Modern Data Strategy*. Springer.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19. https://doi.org/10.1016/0306-4573(94)90020-5

Fürber, C. (2015). *Data Quality Management with Semantic Technologies*. Springer.

General Data Protection Regulation. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=General_Data_Protection_Regulation&oldid=1000699084

Google. (n.d.-a). *Hit—Analytics Help*. Retrieved January 30, 2021, from https://support.google.com/analytics/answer/6086082?hl=en

Google. (n.d.-b). *The data layer—Tag Manager Help*. Retrieved August 11, 2021, from https://support.google.com/tagmanager/answer/6164391?hl=en

Google. (n.d.-c). *What is reCAPTCHA? - reCAPTCHA Help*. Retrieved August 6, 2021, from https://support.google.com/recaptcha/answer/6080904?hl=en

Google. (2021, July 27). *Consolidate Duplicate URLs with Canonicals | Google Search Central*. Google Developers. https://developers.google.com/search/docs/advanced/crawling/consolidate-duplicate-urls

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer Science & Business Media.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Jansen, B. J. (Jim). (2009). Understanding User-Web Interactions via Web Analytics. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *1*(1), 1–102. https://doi.org/10.2200/S00191ED1V01Y200904ICR006

Jason Call. (2017, January 12). *What Is Tag Auditing?* https://resources.observepoint.com/blog/what-is-tag-auditing

JSON. (n.d.). Retrieved August 11, 2021, from https://www.json.org/json-en.html

Jyothi, P. (2017). A Study on Raise of Web Analytics and its Benefits. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING*, *5*, 61–66.

Ladley, J. (2012). *Data governance: How to design, deploy, and sustain an effective data governance program*. Morgan Kaufmann.

Missulawin, I. (2019, January 9). *What is Bot Traffic?* The Click Fraud Blog | ClickCease. https://www.clickcease.com/blog/what-is-bot-traffic/

Nelson, D., 2011. Your Numbers Are Wrong: Ensuring High Quality Web Analytics Data. eClerx (10). Retrieved December 1, 2020, from https://www.digitalanalyticsassociation.org/Files/white_papers/eClerx_Web_Analytics_Data_Qu.pdf

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pournelle, J. (2004). *1001 Computer Words You Need to Know*. Oxford University Press.

Reading: Web Browser | BCIS 1305 Business Computer Applications. (n.d.). Retrieved August 2, 2021, from https://courses.lumenlearning.com/sanjacinto-computerapps-v2/chapter/reading-web-browser/

Richardson, L. (2021). *beautifulsoup4: Screen-scraping library* (4.10.0) [Python]. http://www.crummy.com/software/BeautifulSoup/bs4/

Rubio, M., 2020. The Mini Book of Agile (50). Retrieved January 28, 2021, from https://www.mauriciorubio.com/portfolio-items/the-mini-book-of-agile/

Sadiq, S. (2013). *Handbook of Data Quality Research and Practice*. https://doi.org/10.1007/978-3-642-36257-6

Sharma, A. (2018, March 26). *Python for Adobe Analytics audit and regression testing*. https://www.linkedin.com/pulse/python-adobe-analytics-audit-regression-testing-abhinav-sharma

Sharma, H. (n.d.). *Ultimate Guide to Event Tracking in Google Analytics*. Optimize Smart. Retrieved January 10, 2021, from https://www.optimizesmart.com/event-tracking-guide-google-analytics-simplified-version/

Sidi, F., Hassany Shariat Panahy, P., Affendey, L., A. Jabar, M., Ibrahim, H., & Mustapha, A. (2013). *Data quality: A survey of data quality dimensions*. https://doi.org/10.1109/InfRKM.2012.6204995

Simmons, J. (2017, February 1). Testing Adobe Analytics with Python. *Medium*. https://medium.com/@johndavidsimmons/testing-adobe-analytics-with-python-368752a39cc2

Snapshot. (n.d.). Retrieved August 5, 2021, from https://developers.google.com/search/docs/advanced/crawling/consolidate-duplicate-urls

Schwaber, K. and Sutherland, J., 2020. The Scrum Guide. 1st ed. Creative Commons (14). Retrieved January 19, 2021, from https://www.scrum.org/resources/scrum-guide

Stedman, C., & Vaughan, J. (2020, February). *What Is Data Governance and Why Does It Matter?* SearchDataManagement. https://searchdatamanagement.techtarget.com/definition/data-governance

Streamlit Inc. (2021). *streamlit: The fastest way to build data apps in Python* (0.89.0) [Computer software]. https://streamlit.io

Tealium. (n.d.-a). Customer Data Hub—Manage Data Across Various Touchpoints. *Tealium*. Retrieved March 30, 2021, from https://tealium.com/products/

Tealium. (n.d.-b). *Universal Data Object (utag_data) | Tealium for JavaScript (utag.js) |Tealium Developer Docs*. Retrieved February 13, 2021, from https://docs.tealium.com/platforms/javascript/universal-data-object/

Tealium. (n.d.-c). What is Tag Management? | Basic Tag Management. *Tealium*. Retrieved February 9, 2021, from https://tealium.com/resource/fundamentals/what-is-tag-management/

Tealium. (2015a, March 5). *Order of Operations*. https://community.tealiumiq.com/t5/iQ-Tag-Management/Order-of-Operations/ta-p/326

Tealium. (2015b, September 29). *Data Layer Variables*. https://community.tealiumiq.com/t5/iQ-Tag-Management/Data-Layer-Variables/ta-p/9427

Tealium. (2016, April 21). *How the Data Layer Works for Websites*. https://community.tealiumiq.com/t5/Data-Layer/How-the-Data-Layer-Works-for-Websites/ta-p/13618

TMMData & Digital Analytics Association. (2017). *State of Digital Analytics: The Persistent Challenge of Data Access & Governance*. https://content.tmmdata.com/analytics-survey-whitepaper?hsCtaTracking=a1826298-af83-4039-bf28-f7f25bc4d0ef%7Cea3c1050-cb80-4070-a9b8-36ed7e51ba29

Watts, S. (2020, April 16). *What Is Data Governance? Why Do I Need It? – BMC Blogs*. https://www.bmc.com/blogs/data-governance/

Web crawler. (2021). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Web_crawler&oldid=998001592

Zheng, J., & Peltsverger, S. (2015). *Web Analytics Overview*.

8. APPENDIX

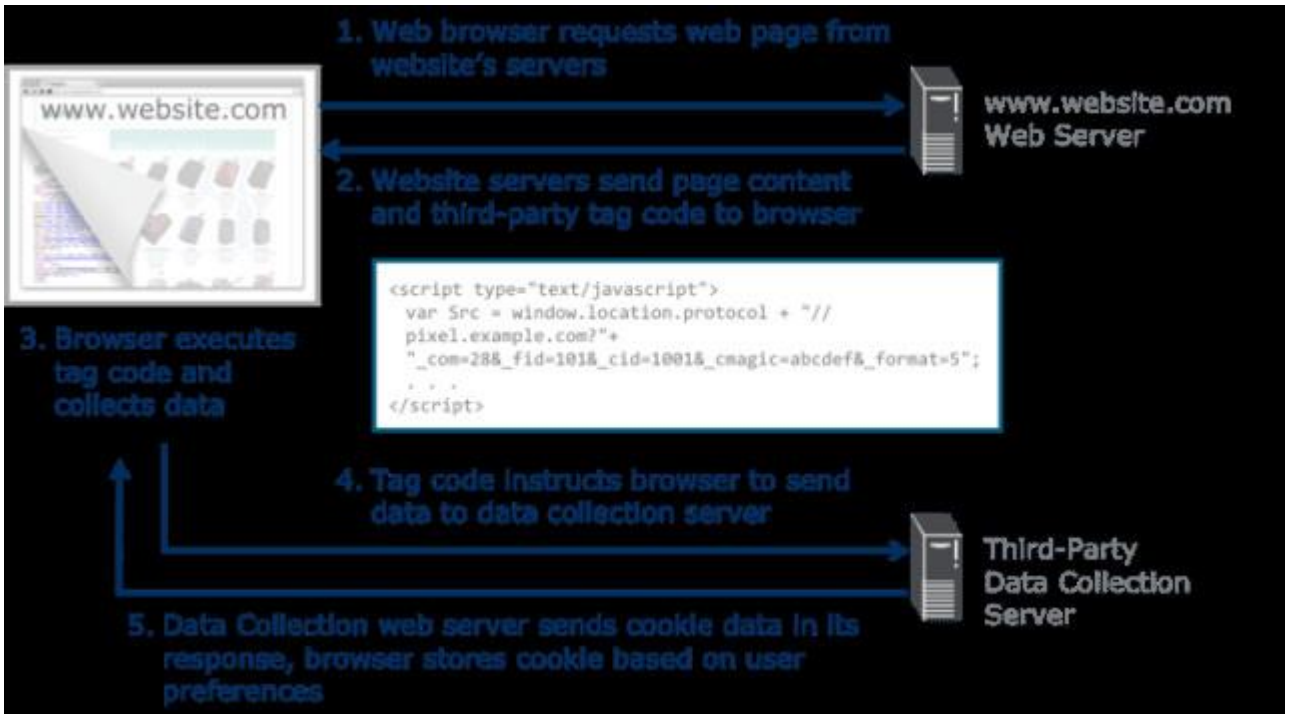
8.1. PAGE VIEW HIT

Adobe Analytics Server Call #1 (1360 chars)	
Report Suite ID	: telcocompanypt
Page Name	: Consumer Homepage
Site Section	: TelcoCompany Site
Current URL	: https://www.telcocompany.pt
Events	: event3
eVar6	: PT
eVar13	: 2021-02-20 19:20:29
eVar15	: N3VxJFAVwaHWq/cKXagrFw==
eVar25	: Logged out
eVar182	: N3VxJFAVwaHWq/cKXagrFw==
eVar183	: 1tUR5BhbmKFYn6hOiTlTfg==
eVar186	: 9EgfrEd5YQkqn9LM7WBaVw==
eVar187	: oxBA9RKHIL8Fka19n+aqFg==
eVar188	: Consumer
eVar190	: Account
eVar192	: Account
eVar195	: Enabled
eVar196	: Enabled
eVar197	: Enabled
eVar198	: Enabled
prop6	: PT
prop9	: Consumer Homepage
prop14	: Consumer Homepage:Consumer Homepage
prop15	: N3VxJFAVwaHWq/cKXagrFw==
prop16	: https://www.telcocompany.pt
prop20	: 260378676585851486439379329520727203 46
prop25	: Logged out
prop30	: Consumer Homepage
prop37	: https://www.telcocompany.pt
prop38	: Telco Company PT -- Telemóveis, Internet, Televisão
prop39	: pt_PT
prop45	: Regular Web
prop46	: Production
prop47	: at_2.1.1
prop52	: 6862604809240385
Currency Code	: EUR
Char Set	: UTF-8
Version of Code	: JS-2.15.0
Data Centre	: swa.telcocompany.pt
Organisation ID	:
Visitor ID	:

8.2. EVENT HIT

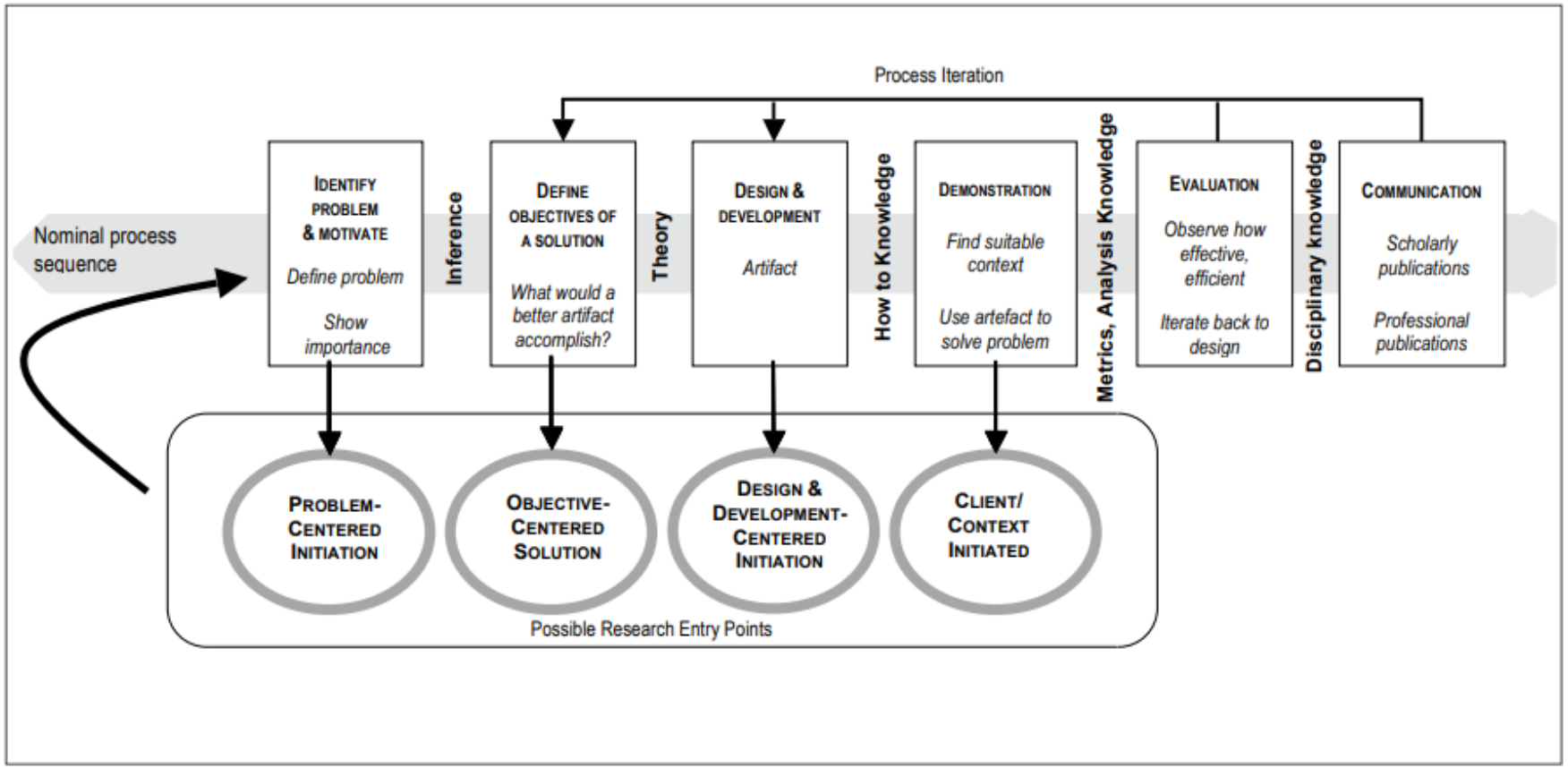
Adobe Analytics Server Call #2 (2047 chars)	
CUSTOM LINK	: no link_name
Report Suite ID	: telcocompanypt
Page Name	: Fixed Bundles
Site Section	: TelcoCompany Site
Current URL	: https://www.telcocompany.pt/pacotes.html?i=quicklinks-tvnetvoz-1#3p
Events	: event55
eVar6	: PT
eVar25	: Logged out
eVar55	: Subscribe Fixed Bundles 3p
eVar56	: Fixed Bundles Online Subscription
eVar57	: Subscribe Fixed Bundles 3p
eVar58	: 1
eVar96	: Subscribe Fixed Bundles Layer
eVar160	: quicklinks-tvnetvoz-1
eVar182	: N3VxJFAVwaHWq/cKXagrFw==
eVar183	: 1tUR5BhbmKFYn6hOiTlTfg==
eVar186	: 9EgfrEd5YQkqn9LM7WBaVw==
eVar187	: oxBA9RKHIL8Fka19n+aqFg==
eVar188	: Consumer
eVar190	: Account
eVar192	: Account
eVar195	: Enabled
eVar196	: Enabled
eVar197	: Enabled
eVar198	: Enabled
prop6	: PT
prop14	: Fixed Bundles
prop16	: https://www.telcocompany.pt/pacotes.html?i=quicklinks-tvnetvoz-1#3p
prop25	: Logged out
prop30	: Fixed Bundles
prop37	: https://www.telcocompany.pt/pacotes.html
prop38	: Pacotes Fibra - Tv Net Voz Portugal
prop39	: pt_PT
prop40	: ,Desconto €2/mês,140 canais,500/100 Mbps,Voz Fixa - Chamadas incluídas
prop41	: binding,offer,television,internet,phone
prop45	: Regular Web
prop46	: Production
prop47	: at_2.1.1
prop52	: 1296629165052912
prop55	: Subscribe Fixed Bundles 3p
prop56	: Fixed Bundles Online Subscription
prop57	: Subscribe Fixed Bundles 3p
Currency Code	: EUR

8.3. TAG PROCESS FLOW



Source: Chen (2017)

8.4. Design Science Research Methodology Process Model



Source: Peffers et al. (2007)