IGNACIO CARREON CARRASCO

ASSESSMENT 1 DESIGN DATA PIPELINE

BDA601

LECTURER: DR SHAHRZAD SAREMI

Academic Integrity Declaration

I declare that except where referenced, the work I am submitting for this assessment task is my own work. I have read and am aware of the Academic Integrity Policy and Procedure of Torrens University, Australia, viewable online at http://www.torrens.edu.au/policies-and-forms. I am also aware that I need to keep a copy of all submitted material and any drafts and I agree to do so.

TABLE OF CONTENTS

INTRODUCTION …..................................................................................................................3

DATA SOURCES ….................................................................................................................4

INTEGRATION CHALLENGES …........................................................................................10

DATA LAKE ….............................................................................................................14

DATA PIPELINE ….............................................................................................................15

REFERENCES .…............................................................................................................16

INTRODUCTION

In the following work, the case of the company Big Retail, a company dedicated to online retail sales in Adelaide, Australia, will be presented.

For purely commercial reasons, Big Retail requires updating its internal processes in order to adopt a strategy focused on data. The objective of this document is to provide different information regarding the data strategy that is intended to be implemented, beginning with the process of integrating all sources of information, whether internal or external.

At the end of this project, diagrams will be shown in which the data lake and the data pipeline that are recommended to be used within the company as a starting point are exemplified.

DATA SOURCES

A wealth of external data can help organizations plan and respond at a granular level. Although external data sources offer immense potential, they also present several practical challenges. Aaser, M., & McElhaney, D. (2021)

The internal and external data sources chosen for this case study are those that are relevant to the business and the client's needs.

INTERNAL DATA

|  |  |
| --- | --- |
| Source | Databases |
| Description | Databases from different areas of the company such as Sales, Marketing, Customer Service, Finance, etc. |
| Format | Structured |
| Assumption | Including internal company information that comes from internal databases strengthens and lends credibility to the strategy to be followed, since this type of data is the core of what is intended to be achieved. |

|  |  |
| --- | --- |
| Source | Website data |
| Description | Related to our users. Connection time, shopping tendencies, time spent on the website as well as best-selling products during certain times. The cookies used by the website will be of significant help in complementing the sales strategy and customer profiling. |
| Format | Structured |
| Assumption | Big Retail business is based 100% on its website, for this reason knowing perfectly all the information related to what happens before, during, and after the purchase is essential. |

|  |  |
| --- | --- |
| Source | Files (Sales, Marketing, Finance, Inventory, etc.) |
| Description | All the documents (Word, Excel, PDF, etc.) that the company has from different areas. |
| Format | Unstructured |
| Assumption | The internal documents of the company contain valuable information that is often not used to the maximum for various reasons. |

|  |  |
| --- | --- |
| Source | Media channels |
| Description | Data from the different digital platforms that the company has, be it opinions, messages, photos and videos of both clients and posts published by the organization |
| Format | Unstructured |
| Assumption | The use of information found in our digital media can help us better understand our customers, their taste for the brand and the areas to improve. |

|  |  |
| --- | --- |
| Source | Saas Platforms |
| Description | Saas services used by the company for accounting, finance, marketing activities, among others, such as CRM and ERP |
| Format | Structured |
| Assumption | Software with which the company works have relevant information from each of the company's areas of operation, integrating them into the data strategy is essential to thoroughly understand daily operations, |

EXTERNAL DATA

|  |  |
| --- | --- |
| Source | Competitors |
| Description | Analysis of competitors where information can be consulted such as: approximate digital marketing budget, products, pricing, seasonal promotions, sales channels, number, and location of competitors, etc. |
| Format | Structured |
| Assumption | Having a broader picture regarding our competitors helps to create more reliable, structured, and complete strategies. |

|  |  |
| --- | --- |
| Source | Social Media |
| Description | Information from social networks in general, public information that refers to the company, industry, trends or potential competitors. |
| Format | Unstructured |
| Assumption | The use of information found in public digital media can help us better understand our customers, the industry their taste for the brand and the areas to improve. |

|  |  |
| --- | --- |
| Source | Web traffic data |
| Description | Data available over the internet with detailed information about the different browsing patterns on the web, whether it is the most visited pages, number of visits, website classification, etc. |
| Format | Structured |
| Assumption | Because it is a digital business, it must be important for Big Retail to know the behavior of customers and potential customers within websites similar to the business. |

|  |  |
| --- | --- |
| Source | Macroeconomic/Microeconomic indicators |
| Description | Public databases referring to different topics related to the industry and the economy of the country or region, such as population consumption data, growth of the retail industry by region, online sales vs sales in physical format, industry sales, and use of different payment options, etc. |
| Format | Structured |
| Assumption | This type of information related to the state of the economy in general and to the sector in particular in which the company competes helps us understand market trends that would otherwise be difficult to detect. |

|  |  |
| --- | --- |
| Source | Search trends data |
| Description | Tools like Google Trends collect the information that people consult on the internet. Taking advantage of these searches brings us closer to current market trends |
| Format | Structured |
| Assumption | Web search engines like Google are extremely popular, using information available from these sources helps us understand everything related to our future customers. |

INTEGRATION CHALLENGES

Integration of multiple information systems generally aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system. Ziegler, P., & Dittrich, K. R. (2007).

The successful integration of the information in a company allows to carry out the data processing in a better way to be able to start with the analysis process.

The challenges presented during data integration are described below:

|  |  |
| --- | --- |
| Challenge | Data Sources/Formats |
| Description | Today, digital businesses handle different types of software to carry out their daily activities, these programs have different formats that are handled by different teams within the organization, when moving to an integration process, they represent a challenge for the actors involved. |
| Solution/Action plan | As far as possible, define general formats within the organization depending on the area where it is located  Use software that has formats that are easy to homogenize.  For future integrations, save relevant information according to past experiences. |

|  |  |
| --- | --- |
| Challenge | Availability |
| Description | Each area within business organizations handles diverse types of data in different ways. The availability of this information to other departments is crucial for the proper performance of the company. The lack of communication and exchange of information is essential between areas since it helps the business to increase its productivity. |
| Solution/Action plan | Choose a technological solution that from the beginning unifies all data management.  Create rules in each area in which it is indicated that the corresponding data in each one will be open to collaborators.  Reserve an online space with public data aimed at high-ranking employees from different areas (managers) with datasets from each area. |

|  |  |
| --- | --- |
| Challenge | Software compatibility |
| Description | Within an organization, the use of different software is common, however, it is common for each device to have different versions of the same program, which can cause compatibility problems when sending files or other work activities. |
| Solution/Action plan | Create internal policies to unify the versions of the software used within the company. Perform quarterly reviews to ensure that all areas have the appropriate software.  Talk to software providers to ensure that equipment is updated from time to time. |

|  |  |
| --- | --- |
| Challenge | Volume |
| Description | 100% of digital businesses generate a large amount of information from various sources. Big retail, as it is a retail company focused on online sales, adheres to this principle, for this reason, it is of the utmost importance to create filters in which the data that is important for the business is analyzed. Determining what information is important is paramount to getting the most out of this resource. |
| Solution/Action plan | Acquire data storage technology to structure information in a better way |

|  |  |
| --- | --- |
| Challenge | Time consuming |
| Description | The time required for the integration of the various sources of information can vary depending on the work and the size of the company. This time invested in the process keeps specialized workers away from other types of functions that could be more productive for the company. Trying to minimize the time of this work can help make the operation of the different areas involved more efficient. |
| Solution/Action plan | Define detailed plans with specific schedules for the treatment of information  Design balanced work teams in terms of experience and ability to avoid mistakes  Write rules related to the integration of the data where the objectives of each collaborator are specified. |

|  |  |
| --- | --- |
| Challenge | Quality of data |
| Description | Having a quality database with updated and automatable information represents a competitive advantage in the market. Maintaining the quality of the information is important because it allows organizations to save time and money, the challenges that commonly arise in relation to this issue are due to inaccuracies, outdated or duplicate information. |
| Solution/Action plan | Acquire specialized software for managing duplicate information (Deduplication technology)  Provide constant maintenance (updating) of the company's databases |

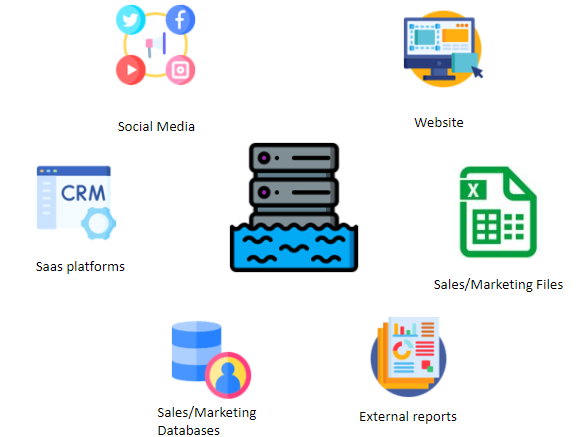
|  |  |
| --- | --- |
| Challenge | Cost |
| Description | An obvious challenge when integrating data is the cost associated with this process. The processing of information, depending on its complexity, can take considerable time and requires qualified workers who are highly sought after in the labor market, this translates into working hours with salaries that include several people. Additionally, there is always the risk of mistakes being made which can cause failures in the service or a change in the company's internal processes. |
| Solution/Action plan | Outsourcing. For small and medium-sized companies that cannot afford excessive costs, professionals hired by project can be used.  Design balanced work teams in terms of experience and ability to avoid mistakes  Write rules related to the integration of the data where the objectives of each collaborator are specified. |

DATA LAKE

A data lake is a massive collection of datasets that: (1) may be hosted in different storage systems; (2) may vary in their formats; (3) may not be accompanied by any useful metadata or may use different formats to describe their metadata; and (4) may change autonomously over time. Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019)

The data lake does not require a rigid schema or manipulation of the data of all shapes and sizes, but it requires maintaining the order of the data arrival. Miloslavskaya, N., & Tolstoy, A. (2016)

Below is an image with the data lake proposal for the Big Retail company, the descriptions of each of the sources can be found in the previous point.

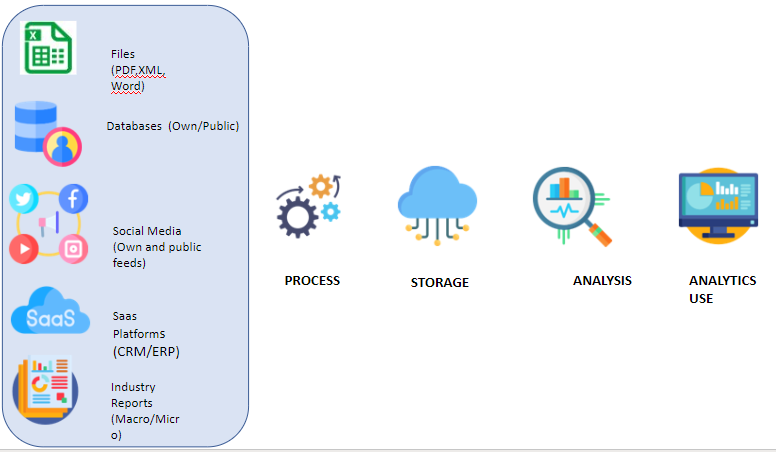


Data lake proposal Big Retail

DATA PIPELINE

As Quemy, A. (2019) mentions raw data are rarely ready to be consumed and must be transformed by a succession of operations usually referred as data pipeline.

The following is a sample of the suggested data pipeline for the Big Retail company.



Big Retail data pipeline proposal

REFERENCES

Ziegler, P., & Dittrich, K. R. (2007). Data integration—problems, approaches, and perspectives. In *Conceptual modelling in information systems engineering* (pp. 39-58). Springer, Berlin, Heidelberg.

Aaser, M., & McElhaney, D. (2021). Harnessing the power of external data.

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment*, *12*(12), 1986-1989.

Miloslavskaya, N., & Tolstoy, A. (2016). Big data, fast data and data lake concepts. *Procedia Computer Science*, *88*, 300-305.

Quemy, A. (2019). Data Pipeline Selection and Optimization. In *DOLAP.*