Instagram - Data Warehouse Implementation

V. Bartolomeu, N. Ivakko, F. Salimi, S. Ferreira
FEUP, Faculty of Engineering, University of Porto
Master in Data Science and Enginnering

***Abstract*—This report outlines a data warehouse project for an instagram dataset. This project includes the creation of dimensional model with dimensions and fact tables, using PostgreSQL as the database engine. The ETL process was applied into the database, followed by creating interactive dashboards using Tableau to visualize the data.**
The main goal of this project was to earn experience in data warehouse building techniques and provide insights into potential application of it in the real world.

***Index Terms*—Data Warehouse, Dimensional Model, ETL process, Business Analytics, Interactive Dashboards.**

I. Introduction

The concept of data warehousing was first introduced in the late 1980s, as a response to the growing need for more efficient data management and analysis in large organizations. The first data warehouses were typically custom-built solutions that required significant upfront investment and expertise. However, with the advent of standardized data warehousing tools and technologies in the 1990s, the process of building and maintaining data warehouses became much more accessible and affordable.

As data warehousing evolved, so did its use cases. Originally designed for reporting and analysis purposes, data warehouses have since been adapted to support a wide range of applications, including customer relationship management, supply chain management, and financial analysis. In particular, the rise of business intelligence (BI) has been a major driver of data warehouse adoption, as BI relies heavily on the availability of accurate and timely data to drive insights and inform decision-making. Nowadays, data warehouses are considered a cornerstone of modern BI infrastructure, providing a foundation for data integration, transformation, and storage that enables organizations to make data-driven decisions with confidence. [1][2]

While data warehouses are typically associated with large-scale enterprise environments, the concepts and principles behind them can be applied in a variety of contexts, including social media platforms such as Instagram. In fact, Instagram is an excellent example of a platform that relies heavily on data warehousing to store and analyze vast amounts of user-generated content. By leveraging the power of data warehousing, Instagram is able to deliver a personalized and engaging user experience, while also providing valuable insights to businesses and advertisers who use the platform to reach their target audiences. In a university project, the principles of data warehousing can be used to build a similar platform or application that leverages the power of data to deliver a compelling user experience and drive valuable insights.

Instagram was launched in 2010 by Kevin Systrom and Mike Krieger as a mobile-only photo-sharing app for iOS devices. In just two months, the app had gained one million users, and by 2012, it had over 30 million active users. In that same year, Facebook acquired Instagram for $1 billion, and since then, the platform has continued to grow and evolve rapidly.

Today, Instagram has expanded far beyond its original photo-sharing roots and includes features such as short-form videos, Stories, Reels, IGTV, and more. The platform has become an essential tool for businesses, influencers, and individuals looking to connect with others and share their content with a global audience. Instagram's growth and success can be attributed to its ability to adapt to changing user needs and preferences, as well as its dedication to user engagement and innovation. [3]

II. Project description

In this report we present an example use-case of data warehousing in the scope of the masters course "Data Warehouses". The project serves the purpose to apply the contents learned in the lecture and to gain practical experience in creating dimensional models for data warehouses. The choice of topic is left to each group, taking into account the following requirements: at least 10000 entries in the dataset, aggregated facts or snapshots with at least one semi-additive measure, and at least four dimensions. The dimensions should include a temporal dimension as well as be common to both facts.

As already indicated in the introduction, our project choice falls on Instagram, or rather Instagram posts and their metadata. According to a report by Statista, as of 2019, the largest age group of Instagram users in the United States were between the ages of 25 and 34, making up approximately 33.1% of the platform's user base. The second-largest age group was between 18 and 24 years old, with 22.7% of Instagram users falling within that range. Users aged 35 to 44 made up 18.8% of the platform's user base, while those aged 45 to 54 made up 9.9%. Users over 55 years old represented just 5.1% of Instagram's users in the United States [4].

In order to make this report understandable for all ages, we would first like to provide a brief overview of Instagram's functionalities. The process of using Instagram typically begins with creating an account, which requires a unique username and password. Once an account is created, users can choose to make their profile public or private. Private profiles require users to approve follower requests, while public profiles allow anyone to follow and view the user's content. After creating an account on Instagram, users can also add additional information to their profile, such as a description and their first and last name. This information can help other users find and connect with them on the platform. Additionally, users can add a profile picture to help personalize their account and make it more easily recognizable to others Users can follow other accounts and gain followers themselves, building a network of connections and expanding their reach on the platform. Instagram allows users to post various types of content, including photos, videos, and short-form content like Reels. Each post can be accompanied by a caption, location tag, and hashtags, which can help increase visibility and engagement. Users can engage with each other's content by leaving comments and likes, creating a sense of community and connection on the platform.

III. Dataset

We have selected a dataset for our project that is readily accessible on the popular data science platform Kaggle. This particular dataset is available for free and can be downloaded without any additional costs or restrictions [5]. The dataset is split into three csv files:

- instagram_locations.csv
- instagram_posts.csv
- instagram_profiles.csv

The raw data contained about 42 million posts, linked to 1.2 million locations and 4.5 million profiles.

1) Locations: The locations file that is included in the Instagram dataset is a comprehensive and detailed dataset that contains a wealth of information about various locations around the world. With 23 columns, this dataset allows for granular exploration of location data, providing information at different levels of detail. Each location entry can include a wide range of information, such as the name of the location, street name, zip code, city, region, and country. Additionally, every location in the dataset is assigned a unique ID, making it easy to track and identify specific locations.

The dataset also includes a number of additional metadata fields, such as URLs, GPS coordinates, and timestamps indicating when the location was visited. This allows for even deeper analysis and exploration of the location data, providing insights into how and when people interact with different locations. It's worth noting that Instagram's location tagging feature is highly adaptable, and can be used to suit a wide range of use cases and scenarios. This includes the ability to create custom or personalized location tags, such as ”my bed” or ”favorite coffee shop”, which can provide a fun and creative way for users to interact with the platform. It should be emphasized here that locations do not have to be described in their full precision.

We should mention that some of the parameters of this file do not give real or meaningful information, especially in the latitude and longitude columns. After a quick analysis we saw that some of the locations that these parameters refer to are remote places in a desert or middle of the ocean. This can be caused by a reading error when gathering the information or when transferring the data into a csv file due to the abnormal syntax of the coordinate system.

2) Posts: The posts file in our Instagram dataset is a valuable resource for analyzing user-generated content on the platform. It contains 10 columns, with each row representing a unique post uploaded by a user. In addition to a unique post ID, the file also includes a profile ID, which identifies the user who created the post. This information can be useful for identifying patterns in posting behavior and preferences across different user groups. Furthermore, the timestamp of when the post was created is also provided, allowing for analysis of post frequency and timing. Other available metadata in the posts file includes the type of post (1: Photo, 2: Video, 3: Multi), which can be helpful in understanding user preferences for different types of content, as well as the number of likes and comments that each post has received. These metrics provide valuable insights into post engagement and user behavior on the platform, and can be used to inform marketing and engagement strategies for businesses and individuals alike.

3) Profiles: The dataset that includes profiles has a total of 11 columns. Each profile is identified with a unique profile ID, which is crucial for easy identification and tracking. The columns contain several fields of information, including a description and full name, which are not mandatory. Additionally, the dataset provides the number of followers, followed profiles, and the number of posts for each profile. Notably, the dataset also differentiates business accounts from personal accounts, and this information is included in the dataset as well. A business account is one that represents a company or an organization, while a personal account belongs to an individual user. Moreover, the dataset includes a timestamp for each entry, indicating the time when the profile's data was recorded. Our initial assumption is that this timestamp enables researchers to track changes in the profile's information and analyze trends over time.

A. Preparation and Insights

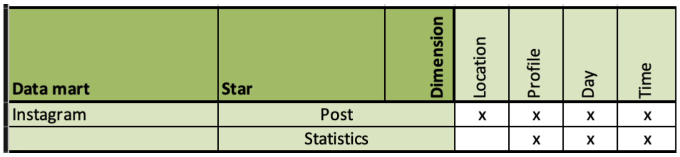
As mentioned previously, the profiles and posts included in the dataset may contain optional fields, such as a location that may only have country information or just a street name. To streamline the data ingestion process, only posts that have complete information were selected for inclusion in the dataset. This means that each post entry in the dataset must contain information about its location, among other details. In the next step, a dimensional model of a data warehouse will be created using the selected posts. To ensure consistency and accuracy in the model, only the profiles that belong to these cleaned post entries will be used. This will enable the data warehouse to provide meaningful insights into the relationships between posts and profiles, allowing for more effective analysis and decision-making. By carefully selecting the most relevant and complete data, the dimensional model will be optimized for efficient and accurate reporting. Contrary to our previous assumptions, the profiles dataset does not track the evolution of profiles over time. Instead, it provides a snapshot of each profile at a specific point in time. However, it's worth noting that some profiles appear multiple times in the dataset, with around 120 insignificant duplicates or more. These duplicates often occur within milliseconds of each other, suggesting that a web crawler may have been used to collect the data. Although there are only a few instances where changes in the number of followers or followed profiles can be observed, they may be a result of initial wrong recordings or the use of bots to make a profile appear to have a higher reach. In such cases, it's difficult to determine the validity of the data, and caution must be exercised when interpreting these results. To avoid duplicate recordings in our use case, we have adopted a strategy whereby the record with the most followers is selected from a duplicate entry, provided that a change in the number of followers is observed. This approach allows us to streamline the dataset and eliminate redundant entries, while ensuring that the most accurate and up-to-date information is used in our analysis. By taking this approach, we can confidently draw conclusions from the data and generate insights that drive informed decision-making.

IV. Data Warehouse Design

When designing the data warehouse, we utilized a dimensional bus matrix, which helped to ensure that the data warehouse was scalable, adaptable, and could meet the evolving needs of the organization. Specifically, we developed a data mart that was focused on profile analysis, including insights into posting patterns and overall high usage times on Instagram.

In order to build the Data Mart of this project we identified the main measures available in our dataset and these are:

- The posts that are an action of a profile associated with a location and a timestamp and itself also measures how the community responded to this post by registering the number of likes and comments;
- The evolution of profile statistics with time, as this fact can measure the indirect impact of the posts in the overall statistics of a profile.



Data mart	Star	Dimension	Location	Profile	Day	Time
Instagram	Post		x	x	x	x
	Statistics			x	x	x

Fig. 1. Dimensional Bus Matrix.

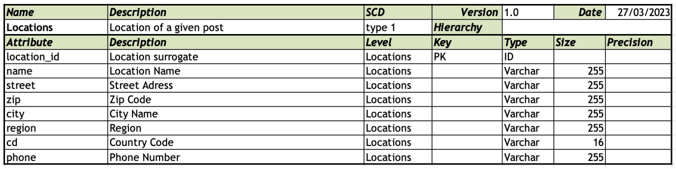
Having defined our analysis goals we can now proceed to design a model that can evaluate the measures we defined. In order to do so for the Post fact table we need to have information regarding the location of the post, who made the post and at what time the post was made. For the Statistics fact table we are need to gather information from all of the tables mentioned previously besides the location. These attribute tables will then be the dimensions of the data warehouse dimensional model.

V. Dimensions Description

In this section we are going to thoroughly define our dimensions and address the parameters they have.

A. Location

The location dimension addresses parameters of different granularity regarding the information of where the post was made and its primary key is the location id. It contains parameters of different granularity, so they are organized by levels with their respective level keys and unique keys. The hierarchy of the levels is done in the following order from lowest to highest granularity: region, country, city, street and zip code. This allows us to make different types of analysis with respect to the location. As was mentioned before but should be reminded again, some of the posts don't have actual real location but descriptions as ”my bed”, ”home”, etc, so these instances will have blank attributes regarding to the levels description and won't be used for the graphical analysis.



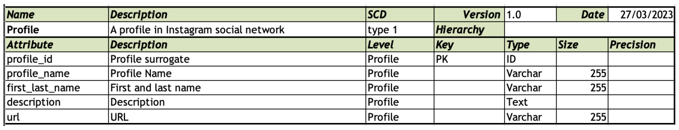
Name	Description	SCD	Version	1.0	Date	27/03/2023
Locations	Location of a given post	type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precision
location_id	Location surrogate	Locations	PK	ID		
name	Location Name	Locations		Varchar	255	
street	Street Adress	Locations		Varchar	255	
zip	Zip Code	Locations		Varchar	255	
city	City Name	Locations		Varchar	255	
region	Region	Locations		Varchar	255	
cd	Country Code	Locations		Varchar	16	
phone	Phone Number	Locations		Varchar	255	

Fig. 2. Location dimension description.

B. Profile

The profile dimension is responsible to describe the profile that made the post. It stores information about the name of the profiles name, profile description, link to the profile and its primary key is the profile id.

Regarding the timestamp dimensions we wanted to simplify the queries and analysis so we separated the date and time into two different tables and defined a column for each attribute. These tables could be merged into one and organize the granularity in levels, we opted for the other approach since one of the project requirements is to have at least 4 dimensions.

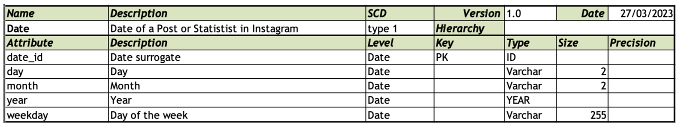


Name	Description	SCD	Version	1.0	Date	27/03/2023
Profile	A profile in Instagram social network	type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precision
profile_id	Profile surrogate	Profile	PK	ID		
profile_name	Profile Name	Profile		Varchar	255	
first_last_name	First and last name	Profile		Varchar	255	
description	Description	Profile		Text		
url	URL	Profile		Varchar	255	

Fig. 3. Profile dimension description.

C. Date

The date dimension stores all the attributes of a timestamp: day, month, year and the weekday. Its primary key is the date id.

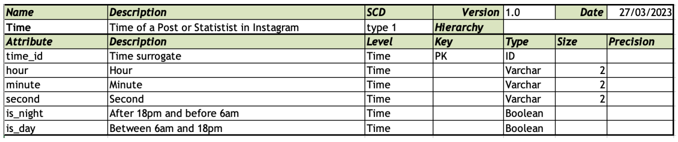


Name	Description	SCD	Version	1.0	Date	27/03/2023
Date	Date of a Post or Statistist in Instagram	type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precision
date_id	Date surrogate	Date	PK	ID		
day	Day	Date		Varchar	2	
month	Month	Date		Varchar	2	
year	Year	Date		YEAR		
weekday	Day of the week	Date		Varchar	255	

Fig. 4. Date dimension description.

D. Time

The time dimension stores information about the time attribute: hour, minute, second and it has another two attributes that will be more meaningful in the analysis part that are a boolean that defines if the instance is at day or night time. Night time is defined between 18pm and 6am and day time is the rest, from 6am to 18pm. Its primary key is the time id.



Name	Description	SCD	Version	1.0	Date	27/03/2023
Time	Time of a Post or Statistist in Instagram	type 1	Hierarchy			
Attribute	Description	Level	Key	Type	Size	Precision
time_id	Time surrogate	Time	PK	ID		
hour	Hour	Time		Varchar	2	
minute	Minute	Time		Varchar	2	
second	Second	Time		Varchar	2	
is_night	After 18pm and before 6am	Time		Boolean		
is_day	Between 6am and 18pm	Time		Boolean		

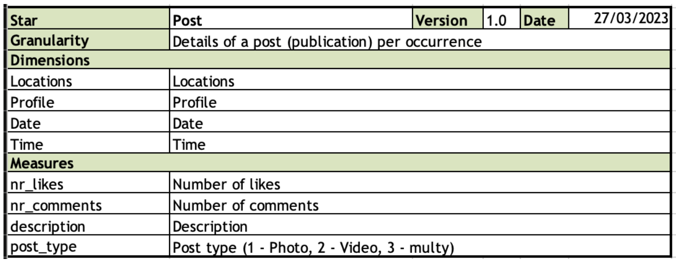
Fig. 5. Time dimension description.

VI. Facts Description

As described previously in Section IV in this data warehouse project there are two different events available, the post and profile statistics.

A. Post

The post fact table is responsible for registering the action of posting a publication in a certain profile and evaluating the response from the community to that post. It's stored information regarding the location associated with the post, the profile that published the post, the date and time of when the data was collected. These attributes are foreign keys from the Location, Profile, Date and Time dimensions respectively. For each post is also recorded the number of likes and comments it had when the data was collected, the type of the post (1 - Photo, 2 - Video, 3 - multi) and the post description.

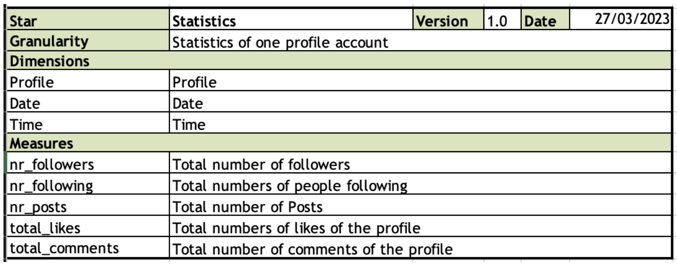


Star	Post	Version	1.0	Date	27/03/2023
Granularity	Details of a post (publication) per occurrence				
Dimensions					
Locations	Locations				
Profile	Profile				
Date	Date				
Time	Time				
Measures					
nr_likes	Number of likes				
nr_comments	Number of comments				
description	Description				
post_type	Post type (1 - Photo, 2 - Video, 3 - multy)				

Fig. 6. Post fact table.

B. Statistics

The statistics fact table measures the profile attributes at a given point in time. It's consequently connected to the profile, date and time dimensions by profile, date and time ids. Its measures are the total number of likes and comments of the given profile, the number of followers, number of following, the profile description and finally the number of posts.



Star	Statistics	Version	1.0	Date	27/03/2023
Granularity	Statistics of one profile account				
Dimensions					
Profile	Profile				
Date	Date				
Time	Time				
Measures					
nr_followers	Total number of followers				
nr_following	Total numbers of people following				
nr_posts	Total number of Posts				
total_likes	Total numbers of likes of the profile				
total_comments	Total number of comments of the profile				

Fig. 7. Statistics fact table.

Unfortunately we don't have multiple entries of the same post so we can't evaluate the evolution of the post statistics over time. The same issue happens in the profile statistics fact table because the multiple records for a profile were all taken milliseconds apart from each other so it's not possible to either, evaluate the impact of a post in a profiles' followers and following record or evaluate the evolution in its statistics in a period of time.

VII. Dimensional Model

After defining all of the measures and attributes that our model should address we now have the resulting dimensional model.

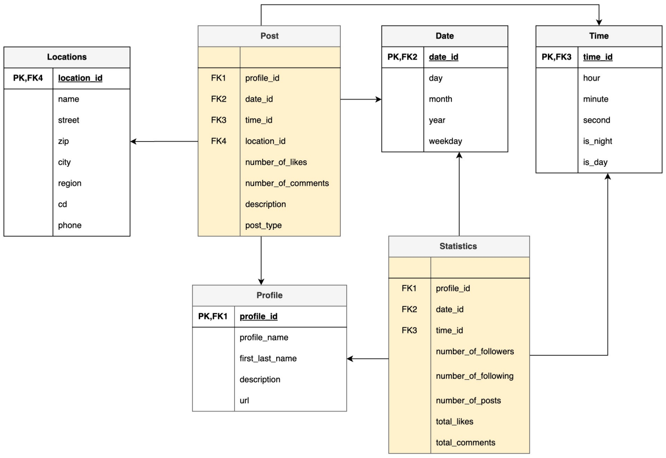


Fig. 8. Dimensional Model

VIII. ETL Process

A. Extraction

Our initial step towards building a data warehouse for the Instagram dataset was the Extraction process. We obtained the dataset from the Kaggle website, as mentioned in section III. Kaggle is a well-known platform that provides publicly available datasets for various purposes. It is a valuable resource for researchers and data scientists who require large datasets to conduct their analyses.

B. Transformation

The preparation of the dataset was carried out using the Python programming language with the aid of the Pandas module. As described in section III-A, we restricted the dataset to retain solely complete data from the posts and profiles tables, as the majority of the analysis will be conducted on the data available in these tables. To achieve this, we eliminated duplicates from the profiles table and ensured that the posts table contained complete data.

Both the posts and profiles tables had a ”description” column that could contain multiple emojis, hashtags as well as regular expressions such as '\n', '\r', or '\t'. In order to load the data into a Postgres database, these expressions needed to be rectified using double backslashes. Additionally, the CSV files required reformatting to conform to the UTF-8 format, which simplified data ingestion into the database.

C. Loading

After preparing and cleaning our data, we uploaded the updated dataset to new CSV files and then onto a Postgres server. We utilized Postgres to create dimensions and query them to form the fact tables of our model. To begin, we created empty tables for the raw data on Postgresql. These empty tables were created by matching the column names of the CSV files from the previous step, and we used the copy command to transfer all the data from those files to the raw tables on Postgresql. To prevent confusion, we separated the schemas for the raw data and for the dimensional model, naming them ”dwproject” and ”instagram,” respectively. We also paid close attention to the format of each column, as we used specific queries with EXTRACT to obtain data from the timestamp format when loading the dimensional tables. Using the correct format was crucial to our loading process.

After transferring the raw data to the server, we proceeded to create empty tables for each dimension. It is worth noting that we had to create the dimensions of the dimensional model before the fact tables since the fact tables contain foreign keys from the dimensions. We created these empty dimension tables in the ”instagram” schema and defined a primary key for each. We also ensured that each column had the appropriate format, such as integer, biginteger, varchar, and text. Overall, we created four empty tables for each dimension. After this step, we were ready to create the empty fact tables. These tables included the foreign keys from the dimensions based on the dimensional bus matrix and dimensional model explained earlier. We also added the necessary columns for the measures of each fact table. By creating the empty tables for the dimensions and fact tables, we set up the base tables for loading the raw data into them.

As previously mentioned, we uploaded the raw data into our database in a new schema, then created the empty dimension and fact tables for our dimensional model in a separate schema. To load the dimensional model from the raw data, we first loaded the dimensions since we needed their primary keys as foreign keys for the fact tables. Each of these dimension tables required specific crafting and queries since they had different natures and sources. We began with the query for loading the date table, which should include all the different dates from our raw data. Since the ”profile.csv” and ”post.csv” tables both had timestamps, we needed to gather all the dates in these two tables. Therefore, we used the UNION query to gather the dates from the timestamps of both tables and insert them into the date table. Our date table had separate columns for the day, month, and year, and we used the EXTRACT query to obtain each of these items from the timestamps of the raw data tables. Overall, we loaded the date dimension with all the existing days, months, and years in both the profiles and posts raw data. For loading the time dimension, we followed the same approach as the date dimension and loaded this table by extracting the hour, minute, and second from the timestamps of the profiles and posts raw data. After carefully analyzing the times in both raw tables, we discovered that we had almost 86000 distinct records for it, indicating that we had at least one record for every second of every minute of every hour in a day. Thus, we defined a complete time table with 86400 rows and included all the possible hours, minutes, and seconds of a day.

The next dimension to load was the profile dimension, which should include the unique profiles in each row. However, the raw data of the profiles table had duplicates and inconsistencies, so we had to organize our queries and insert a single row for each profile in the profile dimension. Our fourth dimension was the location dimension, which was loaded from the raw data and included information related to the locations that we had. Loading the fact tables is the most significant step in loading the dimensional model after loading the dimension tables. There are two fact tables, each with its own unique features. The statistics fact table is more complex to load, as it requires the total number of likes and comments a profile has, as well as other information such as the total number of followers and following, in addition to foreign keys. To load this table, a query was developed that selects information related to the total number of likes and comments for each profile using the WITH clause and creates a temporary named result set from the raw posts table. This result set is then joined with multiple dimension tables and the raw profiles table, and the data is grouped by profile ID. To prevent duplicates, the MAX function was used during the join process. The post fact table was loaded by joining the location, time, and date dimensions, and all foreign keys and measures were inserted into this table. This table contains the most records, with around 4 million taking around 10 minutes to load.

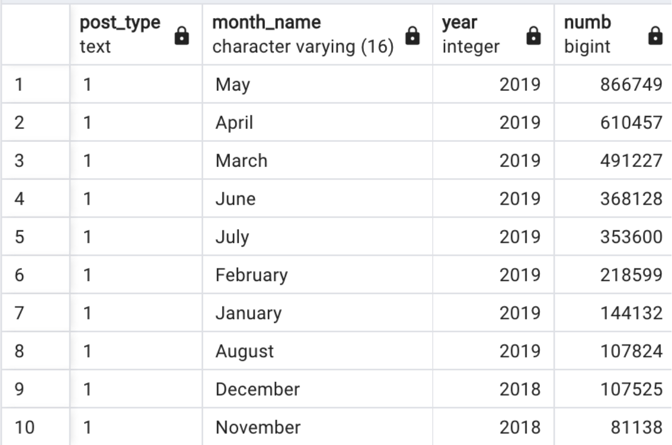
IX. Queries

In our project, we used basic SQL queries to validate that our data was correctly loaded into the data warehouse. These queries provided a simple way to check the accuracy of our data and ensure that our dimensional model was functioning as expected. By using SQL queries to validate our data, we were able to quickly identify any errors or issues with our data and make necessary corrections. These queries also provided a proof of concept that our data warehouse was useful for easy querying and information extraction.

As a result of our validation process, we were confident in the accuracy of our data and the effectiveness of our data warehouse. This gave us the assurance we needed to proceed with our analysis and reporting, knowing that we had a solid foundation of accurate data to work with. By using SQL queries to validate our data, we were able to ensure the success of our project and provide valuable insights for further analysis.

- Post Type Evolution Analysis through Yearly and Monthly Sorting of Posts

```
SELECT post.post_type,
    date.month_name, date.year,
    COUNT(*) as numb
FROM post JOIN date ON post.date_id =
    date.date_id
GROUP BY post.post_type,
    date.month_name, date.year
ORDER BY numb DESC LIMIT 10
```

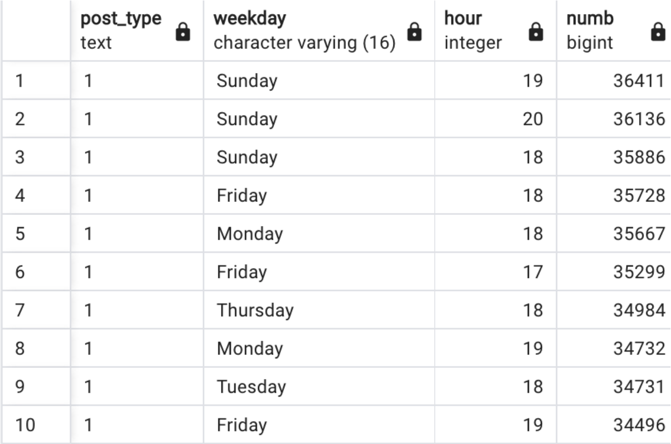


	post_type	month_name	year	numb
	text	character varying (16)	integer	bigint
1	1	May	2019	866749
2	1	April	2019	610457
3	1	March	2019	491227
4	1	June	2019	368128
5	1	July	2019	353600
6	1	February	2019	218599
7	1	January	2019	144132
8	1	August	2019	107824
9	1	December	2018	107525
10	1	November	2018	81138

Fig. 9. Query 1.

- Analyzing Posting Trends by Weekday and Hour to Identify High Demand Intervals:

```
SELECT post.post_type, date.weekday,
    time.hour, COUNT(*) as numb
FROM post JOIN date ON post.date_id =
    date.date_id JOIN time ON
    post.time_id = time.time_id
GROUP BY post.post_type,
    date.weekday, time.hour
ORDER BY numb DESC
```

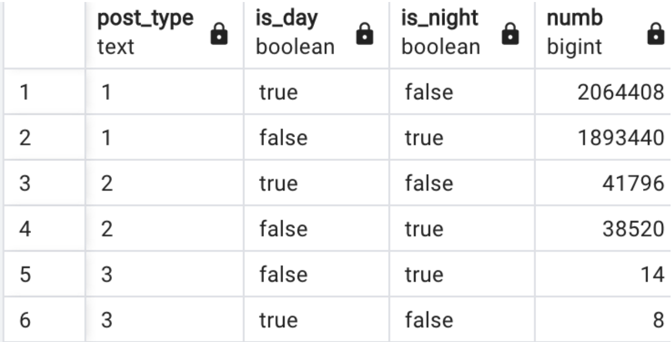


	post_type	weekday	hour	numb
	text	character varying (16)	integer	bigint
1	1	Sunday	19	36411
2	1	Sunday	20	36136
3	1	Sunday	18	35886
4	1	Friday	18	35728
5	1	Monday	18	35667
6	1	Friday	17	35299
7	1	Thursday	18	34984
8	1	Monday	19	34732
9	1	Tuesday	18	34731
10	1	Friday	19	34496

Fig. 10. Query 2.

- Night/daytime activity of postings:

```
SELECT post.post_type, time.is_day,
    time.is_night, COUNT(*) as numb
FROM post JOIN time ON post.time_id =
    time.time_id
GROUP BY post.post_type, time.is_day,
    time.is_night
ORDER BY numb DESC
```



	post_type	is_day	is_night	numb
	text	boolean	boolean	bigint
1	1	true	false	2064408
2	1	false	true	1893440
3	2	true	false	41796
4	2	false	true	38520
5	3	false	true	14
6	3	true	false	8

Fig. 11. Query 3.

X. Graphical Analysis

In order to gain valuable insights from the data, we decided to create graphs and visualizations from our database. As large companies often use powerful tools such as Power BI for this purpose, we initially explored using this software. However, since we were working on a macOS system, we were unable to use Power BI.

We decided to use Tableau for our visualization needs. This software allowed us to easily connect to our database and create visually appealing and interactive graphs. Tableau is a powerful data visualization tool that allows users to create interactive and dynamic graphs from various data sources. One of the main features of Tableau is the ability to create dashboards, which are collections of visualizations and analyses that can be organized in a single view. Dashboards are highly customizable and can be used to display multiple graphs, charts, and tables, making it easier to gain insights from complex data sets. Tableau offers a variety of visualization options to display data extracted and queried from a database. Figure 12 and 13 presents an example of how Tableau can be used to visualize a streamgraph that displays how the data is combined to build the graphs for different visualizations.

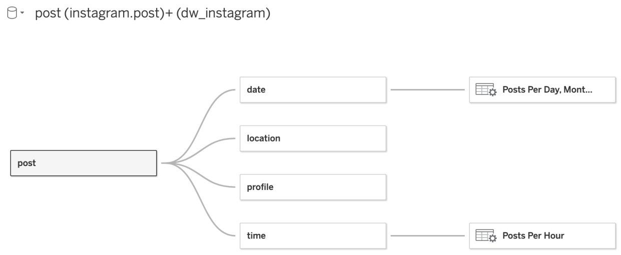


Fig. 12. Tableau - Data Source Posts

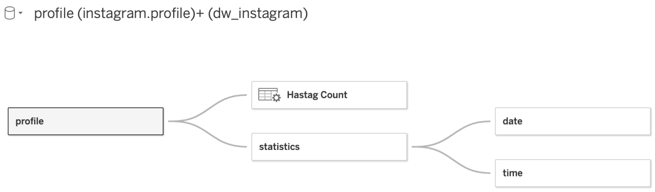


Fig. 13. Tableau - Data Source Statistics

The graphs included in this section are intended for demonstration and explanatory purposes only. For a more comprehensive view of these graphs, please refer to https://github.com/soniaferreira-pires/DW_proj.

A. Trend by Location

To ensure that the level implementation of the posting locations was accurate, we adopted a top-down approach in analyzing the data. This involved creating increasingly abstract graphs that followed the hierarchical structure of locations. Since the dataset contained posts from various locations around the world, it was important to establish a clear and comprehensive hierarchy to enable easy analysis and visualization of the data. By using a top-down approach, we were able to build a solid foundation for our analysis and gain valuable insights into the posting behavior of different regions and locations.

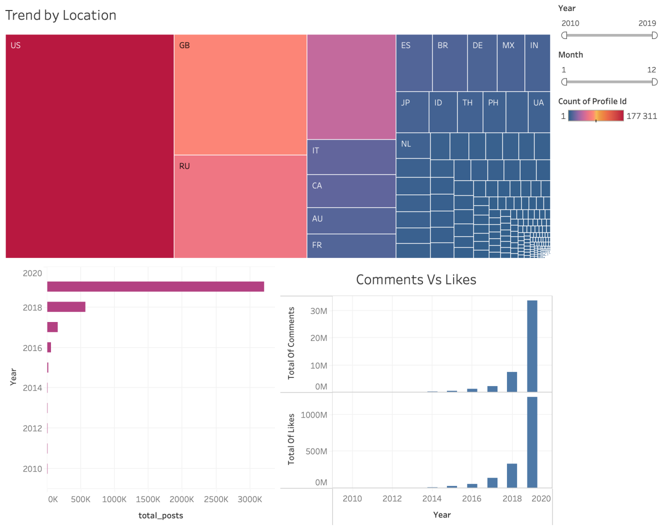


Fig. 14. Trend per location

The distribution of profiles across all locations is illustrated in Figure 14 at the top, while the bottom section displays the progression of total posts, likes, and comments for all locations over the course of a year.

Tableau is a powerful tool that allows for interactive data visualization, and provides features such as slide bars for refining the data and selecting specific locations to gain detailed insights, as demonstrated in Figure 15. The ability to interact with the data in this way provides a more detailed and nuanced understanding of the patterns and trends present in the dataset. Such insights can be valuable in a range of contexts, from identifying areas of high demand or interest to understanding the factors that drive user engagement

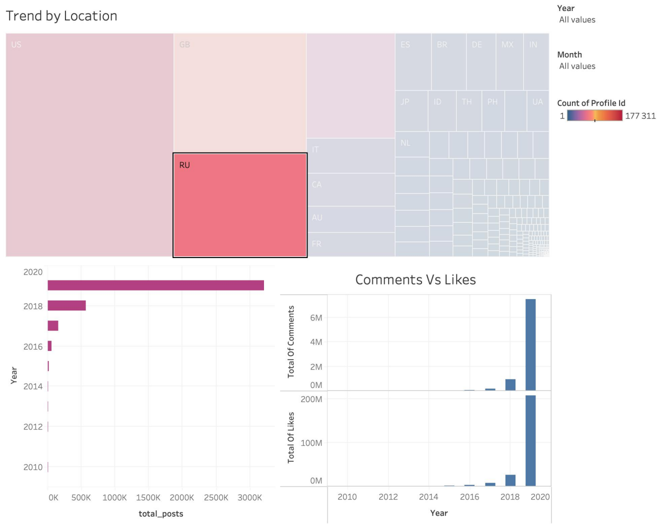


Fig. 15. Trend per location

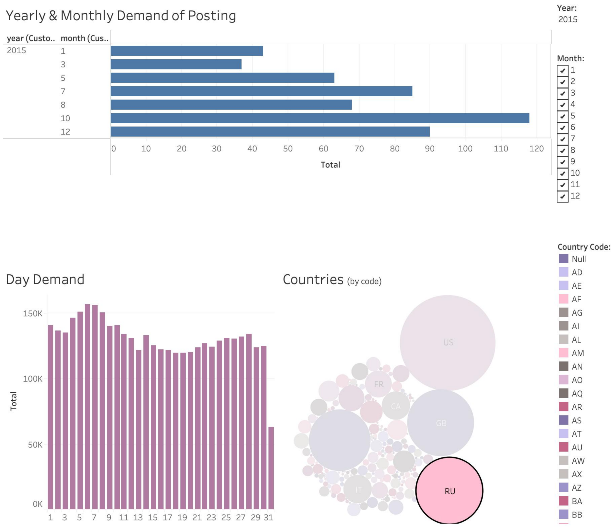


Fig. 17. Trend 2015 per location

B. *Posting activity by time*

The dimensional model used in our data analysis provides the ability to examine posting activities in general, enabling the identification of significant posting events over the year. Our analysis led to the creation of a graph that showcases a detailed and refined study of posting activities throughout the year, with a particular focus on identifying trends. For instance, we present an example of such analysis for the year 2015 as seen in Figure 16. The resulting graph enables us to identify trends and patterns, providing valuable information about the behavior of our data over time.

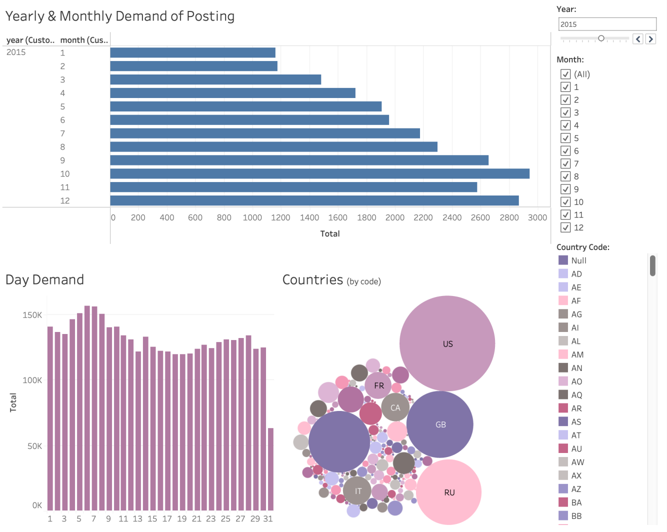


Fig. 16. Trend 2015

The top graph illustrates a steady increase in the number of postings over the year 2015, indicating a rise in the usage of Instagram during that period. The bottom right graph shows the average number of postings per day throughout the month, providing insights into which days have the highest and lowest posting activity.

The bottom left graph displays the overall posting activity distributed across various countries, allowing users to select a specific country for further analysis. Figure 17 provides a more detailed view of the selected country. The bar charts provide a convenient way to select specific months or days of the year for cross analysis, as demonstrated in Figure 18. This feature allows for a more detailed examination of the posting activities during a certain time period and can provide valuable insights into the behavior of Instagram users.

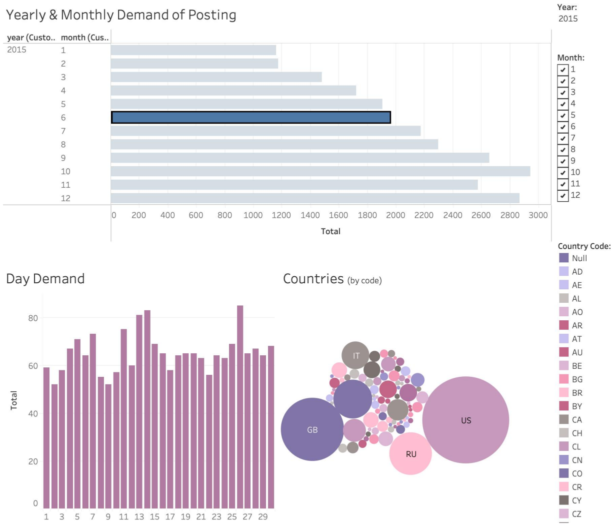


Fig. 18. Trend 2015 per month

C. Analysis per profile

Understanding user behavior can be an important aspect of social media analysis, particularly for companies looking to engage with their target audience. By analyzing user behavior, companies can identify potential influencers for marketing campaigns, and even partner with them for product placements. However, it's important to note that high posting activity does not necessarily equate to high reach. Some profiles might engage in spam-like activities, which could ultimately harm a company or brand if they are selected as ambassadors. Therefore, it's crucial to not only analyze posting activity but also the quality of engagement, and other factors such as audience demographics and interests, to identify the most effective influencers for a particular campaign.

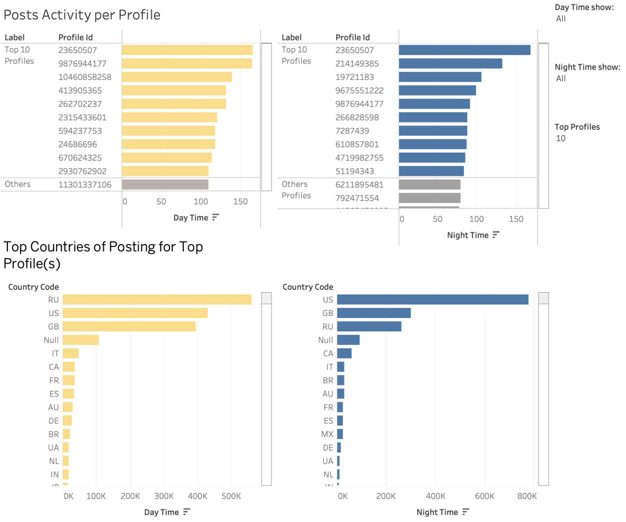


Fig. 19. Profile posting activity

The location tags associated with social media postings provide valuable comprehension into a user's origin and travel behavior. Figure 20 illustrates a sample profile that is primarily located in the Dominican Republic. However, it is important to consider time zones when analyzing nighttime and daytime activity. The web crawler used to collect the data only accounted for the local time zone, leading to potential inaccuracies in the analysis of posting activity in countries like Italy, Spain, and India. These issues could be addressed in the design of the data warehouse to improve the accuracy of location-based analysis. Nonetheless, such location-based perceptivity can be valuable for businesses in identifying potential markets and target audiences.

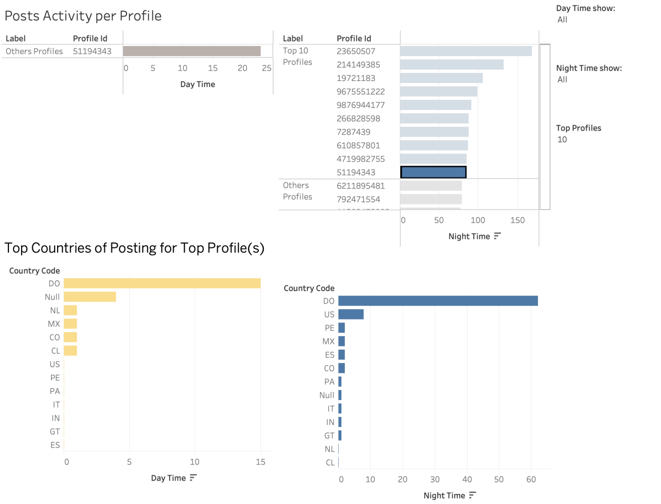


Fig. 20. Profile posting activity

D. Profiles insights

Extracting information about the profiles with most followers, comments, likes can be important to understand audience engagement, identify top performers or identify influencers as previously mentioned.

The dashboard on figure 21 combines a set of top insights related with Profiles, being able to cross-referencing information about the number of followers, comments, likes and hastags, can be of value to understand how certain profiles reach more or less people than others. This can allow to understand how to tailor content strategy or even how to tailor profiles to reach more audience, for instance.

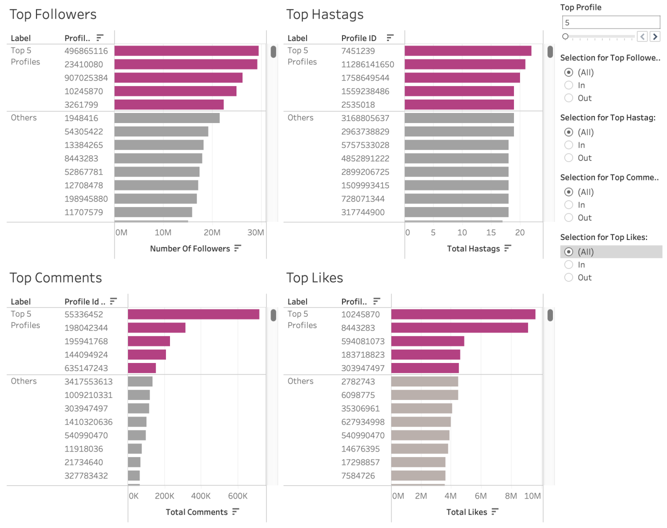


Fig. 21. Top Profiles insights

XI. Conclusions

The main goal of this project was to gain practical experience in creating dimensional models for data warehouses and ETL process. To achieve this goal first we chose the dataset for the instagram, including three raw dataset with overall more than fourthy million records. Based on this data and the available columns we developed the dimensional bus matrix with 4 dimensions of date, time, location, and profile, as well as two fact tables of post, and statistics. After this stage we defined each of these tables in details and created the dimensional model to cover all aspects of the data in hand. After that we cleaned and organised the raw data in Python, as well as creating the proper formatted tables in PostgreSQL, and eventually uploaded the raw data into three tables there. In the next step we created the dimension and fact tables in the server and loaded those tables by queries to finalise the loading stage of our dimensional model. Finally, by using Tableau we imported the tables of our dimensional model to this platform and created the dashboard and visualisations.

The implementation of the data warehouse developed in this project, aimed to perform as a operational database. Operational databases were created to manage and store data used in daily business operations, such as inventory management, customer orders, and human resources information. They have become a key component for any business, however, like any technology, there are both advantages and shortcomings related with them.

On the advantages side, can be listed that operational databases offer several benefits, as data integration, real-time data access, data consistency, scalability, and security. These advantages improve the operational efficiency of businesses, by ensuring that they can quickly access, process, and secure their data. This will allow to take informed decision-making and have a better customer service, which is a key competitive advantage for business.

As for shortcomings, operational databases have several significant ones that require some thoughtfulness. Maybe the most relevant one, or at least one that represents a big challenge is the limited analytical capabilities, which will limit businesses ability to analyze data and have insights that can help to improve and innovate in their business. Performance limitation and cost are another challenge, as high volumes of data processing can lead to slow response times, that will lead to a bad productivity. Furthermore, the cost of implementing and maintaining these databases can be high, especially when the database is large or requires complex integration with other systems, which can be very difficult to support for small and medium enterprises, that usually have limited resources and budgets. Data redundancy is other issue, that can lead to errors and inconsistencies.

In conclusion, while operational databases play a vital role in managing transactional data, it is crucial to recognize that they are not without their limitations. Businesses need to critically examine the advantages and shortcomings of operational databases to determine the best approach for their needs. By carefully considering these factors, businesses can make informed decisions that allow them to optimize their data management strategy and leverage their data effectively, gaining a competitive edge in today's dynamic business landscape. This ensures that they are utilizing the most effective data management strategy that can provide long-term competitive advantages.

Finally, in this project we were able to get the better understanding of the importance of proper data modelling and its role in creating the effective data warehouse. Additionally, we increased our knowledge and skills of visualisation and communicating behaviours/trends through the dashboards. This project provided a valuable experience in real world application of data warehouses and modern business analytics.

References

1 Kimball, R. and Ross, M., *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley, 2013.

2 Gartner, "Magic quadrant for data management solutions for analytics," 2019.

3 Bruns, A. and Burgess, J. E., "The use of twitter hashtags in the formation of ad hoc publics," in *Proceedings of the 26th Australian Conference on Computer-Human Interaction*. ACM, 2015, pp. 448–451.

4 Statista, "Distribution of instagram users in the united states as of august 2019, by age group," 2021, accessed on April 5, 2023. [Online]. Available: https://www.statista.com/statistics/248769/age-distribution-of-us-instagram-users/

5 Shmalex, "Instagram dataset," https://www.kaggle.com/datasets/shmalex/instagram-dataset?select=instagram_locations.csv, 2023, accessed on April 5, 2023.