Module name: Business Intelligence

Module code: CST3145

Case Study: Spotify

Submission Date: 4/12/2020

Module Coordinator: Ms. Geethu Joy

School of Science & Technology

Department of Computer Engineering and Informatics

TABLE OF CONTENTS

COURSEWORK 1

CASE STUDY

Spotify is a Swedish based audio streaming and media services provider that offers digital copyright restricted recorded music and podcasts, including more than 60 million songs, from record labels and media companies. (WIkipedia, 2020). Spotify monetises its streaming features through two principal business fragments, Premium Service and Ad-Supported Service. (Johnston, 2020).

Premium Subscribers can appreciate boundless online and offline admittance to Spotify's whole music index without regular interruptions from advertisements. At the same time, clients who do not pay for a premium membership can get access to Spotify's Ad-Supported Service. Such clients have restricted online admittance to the organisation's music catalogue, and their streaming experience is mixed with ads. Spotify generates revenue from its Premium Service segment through the sale of a variety of subscription pricing plans. Simultaneously, their income from the Ad-Supported Service fragment is generated through the offer of sound and video publicising space on its non-premium streaming platform. (Johnston, 2020) . Toward the finish of March 2020, Spotify had around 286 million monthly active users, including 130 million premium subscribers and 156 million free subscribers. (Dredge, 2020).

As should be obvious, there is a difference of 26 million users between the Premium and Ad-based Services offered by Spotify. Therefore, this report aims to show how using a data warehouse and subsequent analysis of the warehouse data can increase clients paying for the Premium Service provided by Spotify.

DATA REQUIREMENTS

INTERNAL DATA

Spotify collects clients' personal data when they sign up for the **mobile** or **web app's** service. (Spotify USA Inc, 2020). This data is **structured** and may include:

- Name
- Age
- Email address

- Phone number
- Birthdate
- Gender
- Street address
- Country

Spotify may also collect Usage data of their clients (Data generated as the customer makes use of the service through the **mobile or web app**). This may include:

- **Information about Service plan type** (Individual, Duo, Family, or Student). (Spotify USA Inc, 2020). This data is **structured** and is generated when the user signs up for the service.
- **Clients' communications with the Spotify Service:** This data is **unstructured** and may include their search queries (counting the date and time of any requests made), streaming history, playlists and libraries made (which are **structured**), browsing history, and their connections with the Spotify Service, content, and other Spotify clients. Likewise, this data may incorporate details of clients' utilisation of third-party applications connected with the Spotify Service. (Spotify USA Inc, 2020). Most of this data can be classified as **clickstream data.**
- **User-generated content** (e.g., photos and interactions with the Customer Service Team). (Spotify USA Inc, 2020). This data is **unstructured.**
- **Certain technical data** like:
  - Online identifiers including cookies and IP addresses, which are **structured** data;
  - Data about the sorts of gadgets clients are utilising, network connection type (e.g., Wi-Fi, 3G, LTE, Bluetooth), ISP (Internet Service Provider), browser type, language, operating system, and Spotify application version. All of this data is **structured**;
  - Clients' non-precise location, which might be gotten or gathered from specific technical data (e.g., clients' IP address, the language setting of their device, or payment currency), to conform to geographic prerequisites in Spotify's licensing arrangements, and convey customised content and advertising to clients. This data is **structured**;
  - movement or orientation-generated portable sensor data (e.g., an accelerometer or gyroscope) needed to provide specific features of the Spotify Service to clients. (Spotify USA Inc, 2020). This data is **unstructured.**

- **Voice data**: Spotify will gather clients' voice information with their authorisation to give them extra highlights and functionalities, e.g., communicating with the Spotify Service aurally. (Spotify USA Inc, 2020). This data is **unstructured**.
- **Payment and Purchase data**: Spotify may gather specific personal data if clients pursue a Trial or buy any of their Paid Subscriptions or make different buys through the Spotify Service. The detailed personal data collected will fluctuate contingent upon the payment method (e.g., direct through clients' cell phone carrier or by invoice), and will incorporate data such as:
    - Name;
    - Date of Birth;
    - Credit or debit card type, expiration date, and certain digits of client's card number;
    - Postal code;
    - Mobile phone number; and
    - Details of clients' purchase and payment history. (Spotify USA Inc, 2020)

    All of this data is **structured.**

- **Surveys and Sweepstakes data:** At the point when clients complete any surveys, react to a review or poll, or take an interest in a challenge, Spotify will gather the personal data clients give. (Spotify USA Inc, 2020). This data is classified as **structured**.

EXTERNAL DATA

Spotify may also collect personal data of clients and other data from various third parties. This data may include:

- **Authentication partner data:** In the event that clients register for or sign in to Spotify Services utilising third-party credentials (e.g., Facebook), Spotify will import clients' data from such third parties to help create clients' accounts. (Spotify USA Inc, 2020). This data will be **structured**.
- **Payment partner data:** On the off chance that clients decide to pay for a feature or service by invoice, Spotify may get data from their payment partners (e.g., PayPal, Stripe, et al.) to enable them to send clients receipts, process their payment, and give them what they have paid for. (Spotify USA Inc, 2020). This data is **structured**.

- **Advertisement partner data:** Spotify may acquire certain data about clients, e.g., cookie id, cell phone id, or email address, and inferences about clients' preferences and interests (derived after **analysis**) from specific sponsors and advertising partners (e.g., Google AdSense) that permit them to convey more significant promotions and measure their viability. (Spotify USA Inc, 2020). This data is **structured**.
- **Record label and Music distributor data:** Spotify works with record labels and music distributors to provide the music available in their streaming catalogue. (Spotify AB, 2020). This data is **structured**.
- **Music blogs and Music Cataloguing service data:** Spotify will also work with music blogs and music cataloguing services like Billboard to obtain data about current music charts and trends. (Lucero, 2020). This data will also be **structured**.
- **Weather data:** Spotify may collect data about the weather from third-party weather service providers (e.g., BBC Weather) to provide song or playlist recommendations to clients. This data will be **unstructured**.
- **Social media data:** Spotify may collect relevant social media data from platforms like Facebook, Twitter, or Instagram, which may provide insight into clients' satisfaction or lack thereof with the Spotify Service. This data will also be **unstructured**.

DATA WAREHOUSE

A data warehouse is an essential tool for an organisation like Spotify to possess because it can give them insight into customer behaviour and enable them to fulfil company objectives (The objective, in this case, is increasing premium clients).

A discussion of data identified which would be suitable for storage in a data warehouse.

The data that would be suitable for storage in the data warehouse is the **structured data**, i.e., highly organised data, and has clear relationships. This will include:

- All client personal data, e.g., Name, Age, Email address, Phone number, Birthdate, Gender, Street address, and Country.

- Clients' Service plan type, e.g., Family Plan or Duo Plan.
- Clients' playlist and music libraries.
- Technical data such as Clients' IP addresses, data about the sorts of gadgets clients are utilising, ISP (Internet Service Provider), browser type, language, operating system, and Spotify application version.
- Clients' non-precise location.
- Clients' payment and purchase data, including credit or debit card type, expiration date, card number, Postal code, and phone number.
- All the music data from record labels and music distributors.

DATA SCHEMA

A suitable data schema will be the **snowflake schema**. A snowflake schema is an extension of the star schema, which gives a multidimensional view of the data. The snowflake schema is utilised to replace existing dimensions with smaller dimensions, giving more information about existing dimensions. This can lead to reduced storage space as the dimension tables will be smaller. A significant disadvantage of the snowflake schema is that it requires more table joins due to the newly created dimensions, thereby increasing the processing time required for analysis.

DETERMINING THE GRAIN

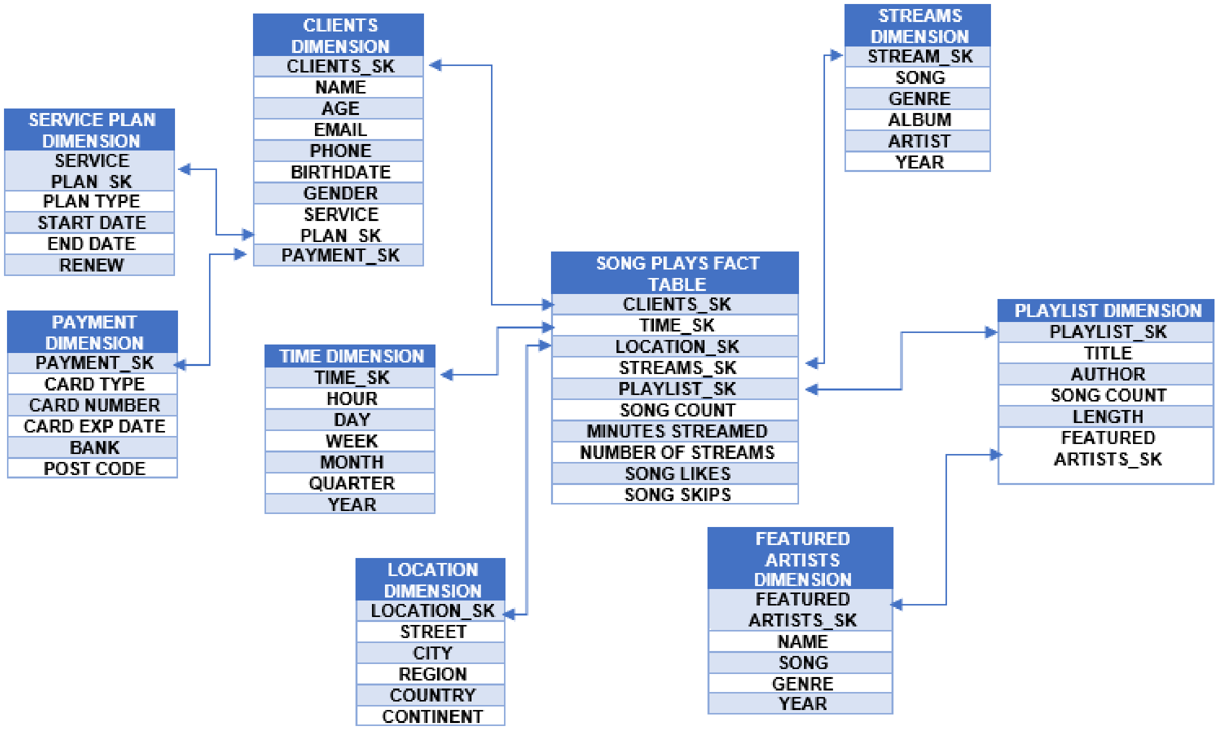
The grain is the level of aggregation of the data in the fact table. It is defined by the lowest level of detail available in the dimension tables.

Following the explanation above, the level of detail can be determined in the dimension tables below;

Client: Name; **Service Plan:** Plan type; **Payment:** Card type; **Time:** Hour; **Location:** Street; **Streams:** Song; **Playlist:** title; **Featured Artists:** Name

SNOWFLAKE SCHEMA

Below is an example of a snowflake schema used for the Spotify business process (**Customer Relationship Management**).



CLIENTS DIMENSION
CLIENTS_SK
NAME
AGE
EMAIL
PHONE
BIRTHDATE
GENDER
SERVICE PLAN_SK
PAYMENT_SK

STREAMS DIMENSION
STREAM_SK
SONG
GENRE
ALBUM
ARTIST
YEAR

SERVICE PLAN DIMENSION
SERVICE PLAN_SK
PLAN TYPE
START DATE
END DATE
RENEW

PAYMENT DIMENSION
PAYMENT_SK
CARD TYPE
CARD NUMBER
CARD EXP DATE
BANK
POST CODE

TIME DIMENSION
TIME_SK
HOUR
DAY
WEEK
MONTH
QUARTER
YEAR

SONG PLAYS FACT TABLE
CLIENTS_SK
TIME_SK
LOCATION_SK
STREAMS_SK
PLAYLIST_SK
SONG COUNT
MINUTES STREAMED
NUMBER OF STREAMS
SONG LIKES
SONG SKIPS

PLAYLIST DIMENSION
PLAYLIST_SK
TITLE
AUTHOR
SONG COUNT
LENGTH
FEATURED ARTISTS_SK

LOCATION DIMENSION
LOCATION_SK
STREET
CITY
REGION
COUNTRY
CONTINENT

FEATURED ARTISTS DIMENSION
FEATURED ARTISTS_SK
NAME
SONG
GENRE
YEAR

EXAMPLE QUERIES

- How many Hip-Hop songs were streamed by clients in Africa?
- How many minutes of music does Chubi Adejoh stream in a week?
- How many clients in Dubai are on the Premium service?

ETL PROCESS

ETL stands for **Extraction, Transformation, Loading.** The ETL process takes data from the operational databases, transforms it into a suitable format, and loads it into the data warehouse. It is the foundation of any data warehouse system because, without it, the warehouse would not exist.

EXTRACTION

In the case of Spotify, there is data coming in from the mobile and web app signup pages, their **Cassandra database** (Cody, 2015), from third parties, and various other sources. Data will have to be extracted from these sources to be loaded into the data warehouse.

The extraction method that will be suitable for Spotify is Partial extraction **with update notification.** It is when the OLTP notifies the manager when there is a change in data. This extraction method will be suitable for Spotify because they generate many data regularly, so there has to be a notification of the frequent updates from their operational systems.

VALIDATION DURING EXTRACTION

- **Check data types match and are relevant**: Checking that clients' phone numbers are **int** and not **float** type**.**
- **Remove redundant and fragmented data:** If a client fills a survey but does not include necessary fields like email or phone number, the data will be removed.
- **Implementing integrity constraints:** Here, domain, entity integrity, referential, and key constraints are established on the tables to ensure data consistency. For example, making sure the age clients input when signing up to Spotify does not exceed a certain threshold.

- **Check keys are relevant and complete:** Ensure the primary key for service plan information matches the foreign key in clients' information.

PRE-TRANSFORMATION: DATA CLEANING

- **Find and remove duplicate tuples:** If a client fills a survey twice, the duplicate data is removed.
- **Delete wrong and inconsistent data:** If a client inputs an age that does not correspond with their birthdate, it is deleted.
- **Attribute mismatch:** The date format of different regions Spotify operates in differs, so it will be made consistent.
- **Combining data from different sources to enrich data:** If a client signs up on the mobile or web app and fills a survey, the data they input will be combined.

TRANSFORMATION

As mentioned earlier, the data extracted from Spotify's operational systems will have to undergo some transformation processes before it can be loaded into the data warehouse. Some of these processes include:

- **Recoding of Records from Multiple Sources**: The format of data collected from the Spotify mobile and web app signup pages might differ from the format of the data gotten from their signup third parties. Also, there is a difference in time zones and payment currencies between the regions Spotify operates in. This type of incompatibility is handled by implementing one consistent format.
- **Mapping data values to coded meanings:** Mapping the gender of clients to "M" or "F" instead of "Male" and "Female".
- **Ensure data cleanliness:** Here, verifications are made to ensure that data in the warehouse conforms to a basic level of consistency and cleanliness, preventing the introduction of dirty data.

LOADING

This is the process of loading transformed data into the data warehouse. The method of loading which will be suitable for Spotify is the **incremental load**. In this method, there is a periodical loading of new, updated data into the warehouse.

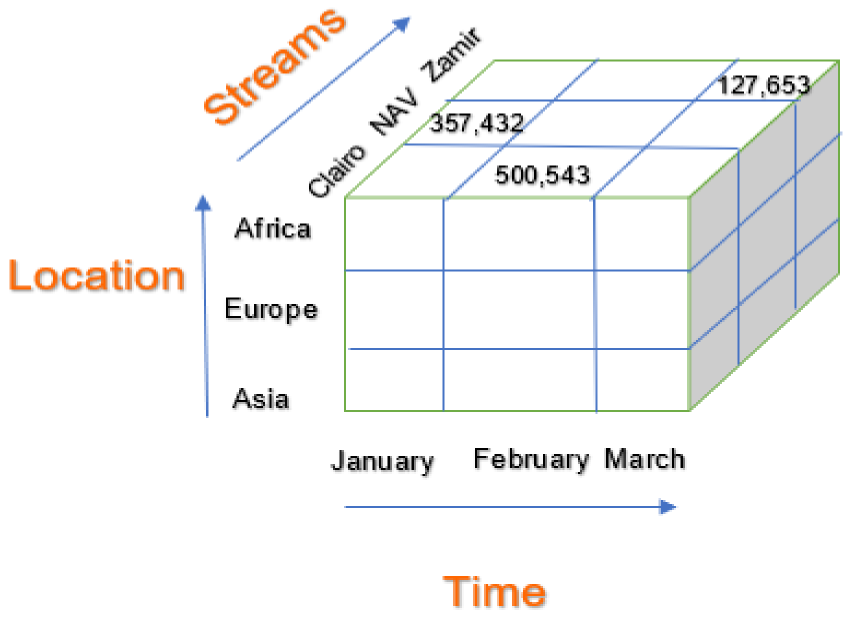
However, **Full refresh** might be used from time to time when specific data warehouse tables need to be deleted and completely reloaded. For example, when customers' service plan is changed from free to premium, the service plan dimension will be deleted and reloaded with the new data.

ONLINE ANALYTICAL PROCESSING (OLAP)

The most suitable OLAP for Spotify would be the **Multidimensional OLAP** (MOLAP). Here, data is stored in multidimensional arrays separate from the data warehouse and is retrieved through the utilisation of **data cubes.**

The data cubes are designed for fast and efficient retrieval of data. Also, because complex calculations are optimised, many calculations are generated as the cube is formed.

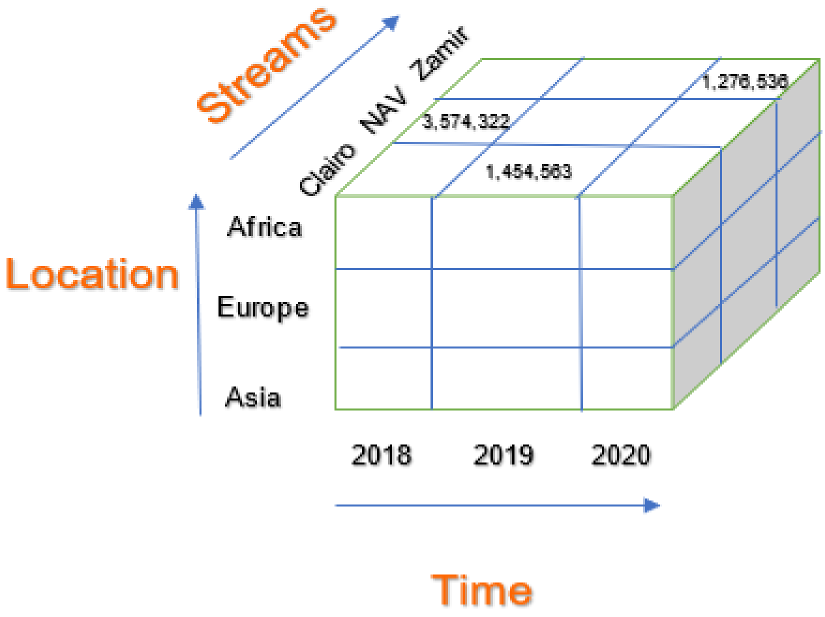
Below is an image of a three-dimensional data cube of the Spotify business process, using the **Location, Time,** and **Streams** dimensions, for the **Number of streams**.



- **Clairo** was streamed 500,543 times in **Africa** in **February**.
- **NAV** was streamed 357,432 times in **Africa** in **January.**
- **Zamir** was streamed 127,653 times in **Africa** in **March.**

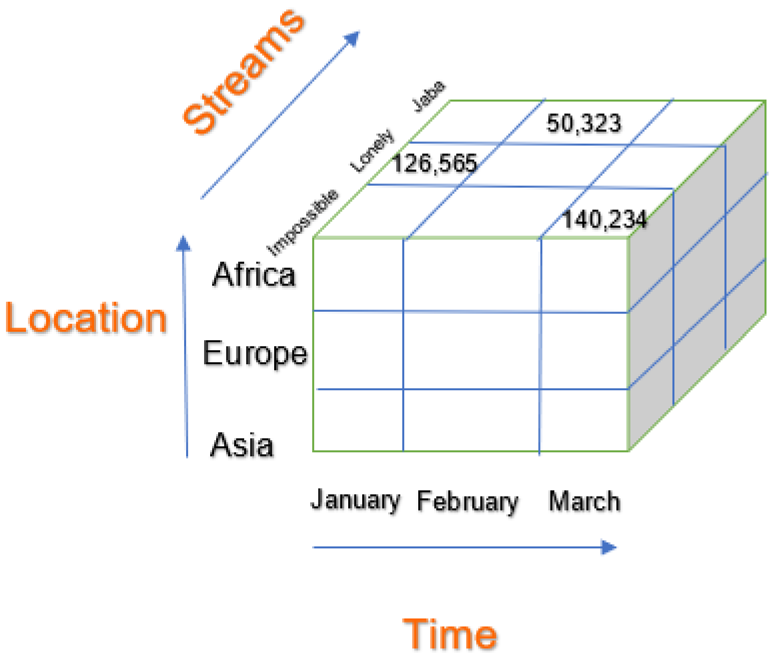
OLAP OPERATIONS

<u>Roll-up</u>: On **Time** from Months to Years.



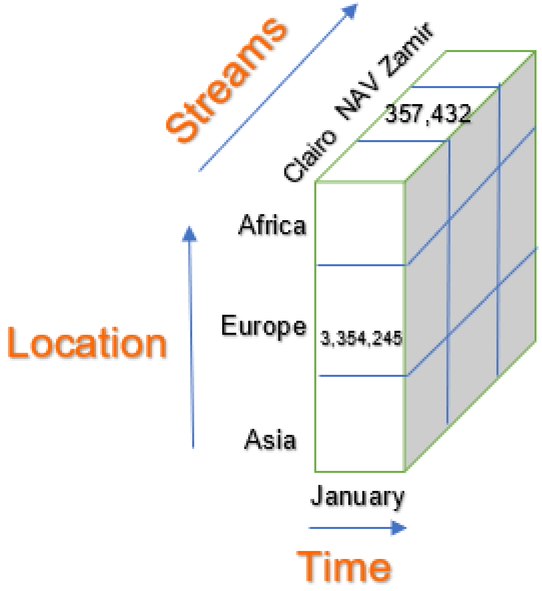
- **Clairo** was streamed 1,454,563 times in **Africa** in **2019**.
- **NAV** was streamed 3,574,322 times in **Africa** in **2018.**
- **Zamir** was streamed 1,276,536 times in **Africa** in **2020.**

Drill-down: On **Streams** from Artist to Song.



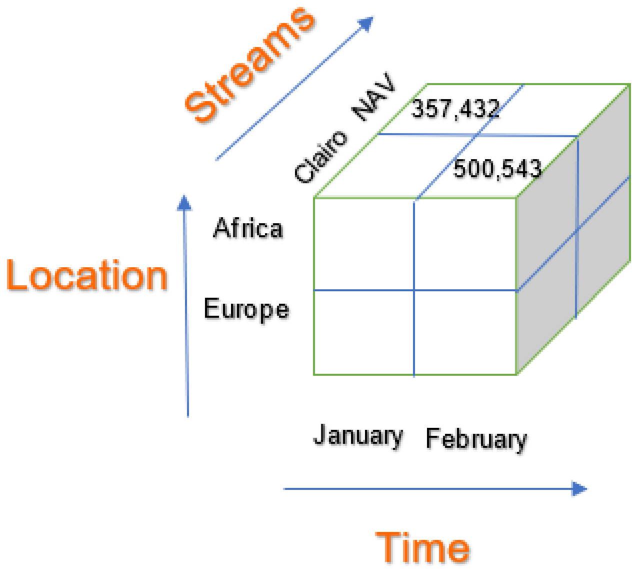
- **Impossible** was streamed 140,234 times in **Africa** in **March**
- **Jaba** was streamed 50,323 times in **Africa** in **January.**
- **Lonely** was streamed 126,565 times in **Africa** in **February**.

Slice: for when **Time** is "January"



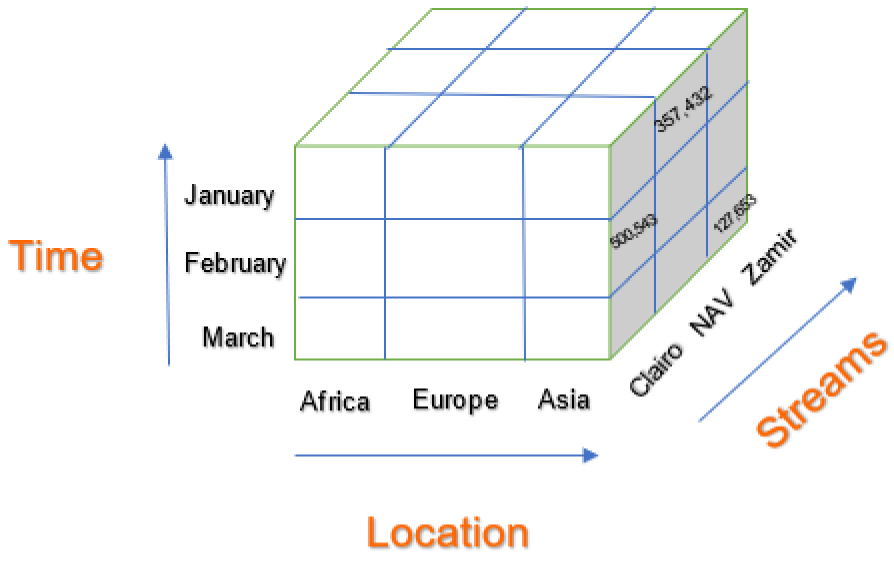
- **Clairo** was streamed 3,354,245 times in **Europe** in **January**.
- **NAV** was streamed 357,432 times in **Africa** in **January.**

Dice: With **Time** as "January" and "February", **Location** as "Africa" and "Europe", **Streams** as "Clairo" and "NAV".



- **Clairo** was streamed 500,543 times in **Africa** in **February**.
- **NAV** was streamed 357,432 times in **Africa** in **January.**

Pivot: Changing the position of all dimensions.



- **Clairo** was streamed 500,543 times in **Africa** in **February**.
- **NAV** was streamed 357,432 times in **Africa** in **January.**
- **Zamir** was streamed 127,653 times in **Africa** in **March.**

BIG DATA

A discussion of data identified which would not **be suitable for storage in a data warehouse.**

- **Data about clients' communications with the Spotify Service:** This may include their search queries, streaming history, browsing history.
  - User-generated content (e.g., Photos and interactions with the Customer Service Team).
  - Online identifiers including cookies and IP addresses;
  - Movement or orientation-generated portable sensor information (e.g., an accelerometer or gyroscope) is needed to provide specific features of the Spotify Service to clients.
- **Voice data**: Spotify will gather clients' voice information with their authorisation to give them extra highlights and functionalities.
- **Weather data:** Spotify may collect weather data from third-party weather service providers (e.g., BBC Weather).
- **Social media data:** Spotify may collect relevant social media data from platforms like Facebook, Twitter, or Instagram, which may provide insight into clients' satisfaction or lack thereof with the Spotify Service.

A discussion of reasons the data is not suitable for storage in a data warehouse.

The data mentioned above will not be suitable for storage in the data warehouse because they are **unstructured**. Unstructured data is data that either does not have a pre-defined data model or is not coordinated in a pre-defined way. As a result, this kind of data will not fit into the data warehouse's data schemas.

Using examples from your scenario, discuss a framework used to collect, store, and analyse this data.

Although this data might not be suitable for storage in a traditional data warehouse, it will become possible by integrating a big data framework like **Hadoop**.
Hadoop is a framework that is used to store and analyse unstructured data. It can work irrespective of a data warehouse and can be located either on the cloud or local servers. Hadoop makes use of HDFS (Hadoop File Distributed File System) to store data.

According to (Bie, 2013), Spotify generated 1 TB of compressed data from users per day,400 GB of data from services per day, 61 TB of data in Hadoop each day. They utilised a 328 node Hadoop cluster, which handled 6500 jobs/day (192.000/month) with 10 PB of storage capacity.
At first, they started with a small (scrap metal) cluster of 37 servers. Then moved to Amazon Elastic Map/Reduce (EMR) and S3 to quickly scale. Then they later built an in-house cluster of 60 nodes because of EMR costs.

Spotify currently has a 2500 node on-premise Hadoop cluster that handles more than 20,000 jobs a day (Li, 2017), a number that is undoubtedly expected to rise.

CONCLUSION

This report aimed to mention data, including internal and external data, which could be used to aid Spotify in their decision-making processes. It also explains the data collected by Spotify that will be suitable to be stored in a data warehouse and gives a suitable schema in which the data will be stored. Some example queries are also identified, which can be run on the schema specified. Furthermore, The ETL process is explained concerning the Spotify business process, stating relevant extraction and loading methods and the relevant transformations made on data before loading. Also, it explains the type of OLAP server (MOLAP) that may be used by Spotify, with a diagram of a data cube, and gives examples of the different OLAP operations that can be carried out on the cube. Lastly, the report identifies the data collected by Spotify, which will not be suitable for storage in a data warehouse. It goes ahead to identify a framework (Hadoop) that will enable them to store that data. It also gives a history of Spotify's use of the Hadoop framework and its current infrastructure.

According to (Bie, 2013), Spotify uses Apache Kafka for its data collection. This open-source distributed event streaming platform provides high-performance data pipelines, streaming analytics, and data integration (Apache Software Foundation, 2017), Since Spotify's primary data collection and analysis tools are Kafka and Hadoop, it can be inferred that most of the data that is collected and used by Spotify is unstructured. Therefore, the Spotify data strategy would be heavily reliant on **big data** and its insights.

REFERENCES

Apache Software Foundation, 2017. *APACHE KAFKA.* [Online]
Available at: https://kafka.apache.org/
[Accessed 3 December 2020].

Bie, W. d., 2013. *Big Data Infrastructure at Spotify.* [Online]
Available at:
http://files.meetup.com/4533812/Munchen%20HUG%20Big%20Data%20Infrastructre%20at%20Spotify%2020130925-%20Wouter%20d.pdf
[Accessed 29 November 2020].

Cody, N., 2015. *Cassandra: Data-Driven Configuration.* [Online]
Available at: https://engineering.atspotify.com/2015/09/21/cassandra-data-driven-configuration/
[Accessed 17 November 2020].

Dredge, S., 2020. *How many users do Spotify, Apple Music and other big music streaming services have?.* [Online]
Available at: https://musically.com/2020/02/19/spotify-apple-how-many-users-big-music-streaming-services/#:~:text=At%20the%20end%20of%20March,supported%20(i.e.%20free)%20listeners.
[Accessed 1 November 2020].

Johnston, M., 2020. *How Spotify Makes Money.* [Online]
Available at: https://www.investopedia.com/articles/investing/120314/spotify-makes-internet-music-make-money.asp
[Accessed 1 November 2020].

Li, N., 2017. *Big Data Processing at Spotify: The Road to Scio (Part 1).* [Online]
Available at: https://engineering.atspotify.com/2017/10/16/big-data-processing-at-spotify-the-road-to-scio-part-1/
[Accessed 30 November 2020].

Lucero, M. J., 2020. *Quartz.com.* [Online]
Available at: https://qz.com/1773480/the-problem-with-how-the-music-streaming-industry-handles-data/
[Accessed 2 November 2020].

Spotify AB, 2020. *FAQ.* [Online]
Available at: https://artists.spotify.com/faq/music#how-do-i-get-my-music-on-spotify
[Accessed 2 November 2020].

Spotify USA Inc, 2020. *Spotify Premium.* [Online]
Available at: https://www.spotify.com/us/premium/
[Accessed 1 November 2020].

Spotify USA Inc, 2020. *Spotify Privacy Policy.* [Online]
Available at: https://www.spotify.com/us/legal/privacy-policy/#s4
[Accessed 1 November 2020].

Tutorialspoint, 2020. *Data Warehousing - Schemas.* [Online]
Available at:
https://www.tutorialspoint.com/dwh/dwh_schemas.htm#:~:text=Schema%20is%20a%20logical%20description,associated%20data%2Ditems%20and%20aggregates.&text=A%20database%20uses%20relational%20model,Snowflake%2C%20and%20Fact%20Constellation%20schema.
[Accessed 15 November 2020].

WIkipedia, 2020. *Spotify wiki.* [Online]
Available at: https://en.wikipedia.org/wiki/Spotify
[Accessed 26 October 2020].

Wikipedia, 2020. *Unstructured data.* [Online]
Available at:
https://en.wikipedia.org/wiki/Unstructured_data#:~:text=Unstructured%20data%20(or%20unstructured%20information,numbers%2C%20and%20facts%20as%20well.
[Accessed 18 November 2020].