Harmony Hub: Data Warehouse for Music Streaming

Team: Inside Data

Abhishek Shinde, Sejal Sardal, Rushikesh Shinde, Shashank Guda

Project Overview

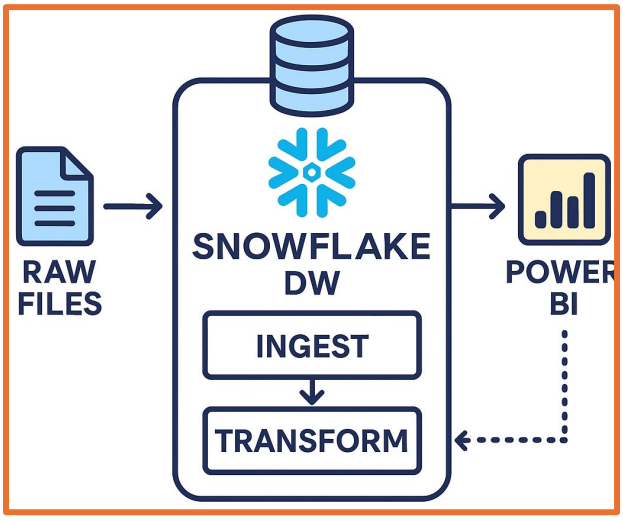
Harmony Hub is a data warehouse solution built on Snowflake for a music streaming platform. The goal of this project is to design, implement, and optimize a centralized data infrastructure to manage and analyze vast volumes of music-related data, such as:

- Tracks and genres
- Artists and albums
- User subscriptions
- Streaming session logs

The platform supports advanced analytical queries for business intelligence, user behavior tracking, subscription modeling, and genre-based insights.

Data Warehouse Pipeline Architecture

The Harmony Hub project follows a streamlined ELT (Extract-Load-Transform) data pipeline architecture built on Snowflake Data Warehouse, enabling efficient ingestion, transformation, and analytics of large-scale music streaming data. The pipeline supports end-to-end data flow from raw file ingestion to visualization in Power BI, ensuring high performance, scalability, and modularity.



Pipeline Components

1. Raw Data Ingestion

The pipeline begins with the ingestion of raw files in formats such as CSV, JSON, or Parquet. These files include critical datasets such as:

- User demographics and subscription details
- Streaming session logs
- Track and genre metadata
- Artist and album information

These files are manually uploaded into Snowflake's staging area.

2. Snowflake Data Warehouse (DW)

The core processing takes place in Snowflake, a cloud-native data warehouse that serves both as the ingestion layer and transformation engine.

- **Ingestion Layer**: Raw files are loaded into raw tables using Snowflake's native COPY INTO operations from internal or external stages.
- **Transformation Layer**: Using SQL-based transformation logic, data is cleaned, joined, enriched, and structured into:
    - **Fact Tables**: Including *fact_sessions*, *fact_user_subscriptions*, and *bridge_track_genres*
    - **Dimension Tables**: Such as *dim_users*, *dim_tracks*, *dim_genres*, *dim_artists*, and *dim_albums*

These transformations implement business rules, generate surrogate keys, and ensure referential integrity across the schema.

3. Analytics & Visualization (Power BI)

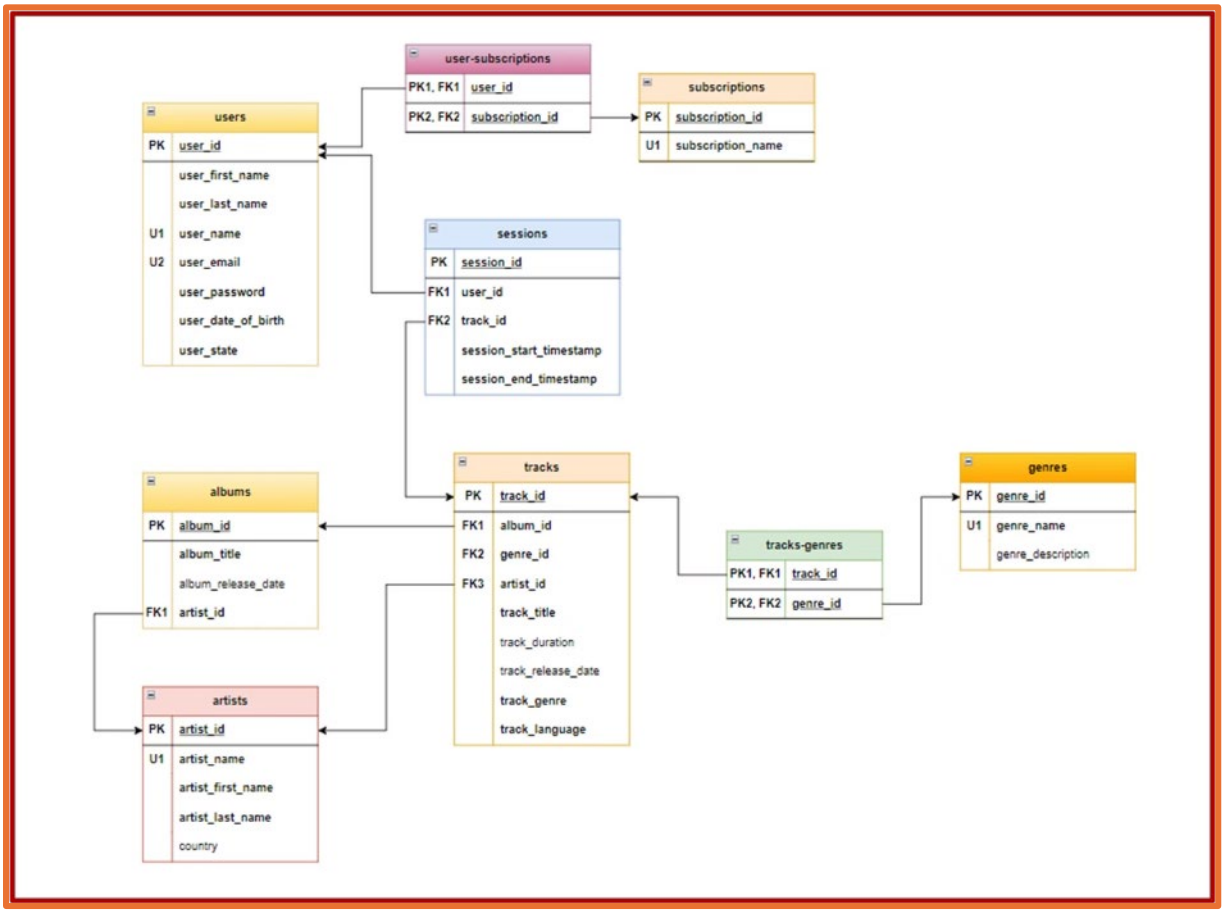
Once data is structured in Snowflake, it is connected to Power BI for advanced reporting and dashboarding. Power BI utilizes Import Mode to fetch data from Snowflake, enabling real-time and historical insights.

Entity Relationship Design

The logical design includes a normalized relational schema to ensure data integrity and reduce redundancy.

Key Entities:

- **Users:** Contains personal details and subscription linkage
- **Tracks:** Captures metadata of each song
- **Sessions:** Logs user activity with timestamps
- **Genres, Artists, Albums:** Support metadata organization
- **User-Subscriptions:** Many-to-many bridge for user plans
- **Tracks-Genres:** Handles the many-to-many mapping between songs and their genres



Dimensional Modeling & Star Schema

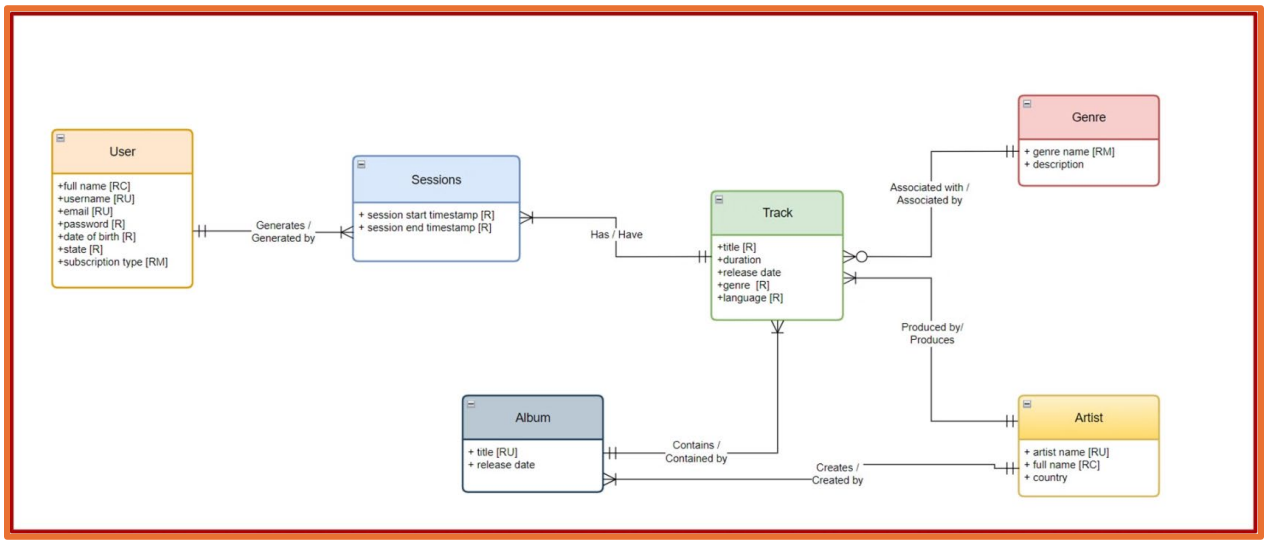
To support analytical workloads, a **star schema** was created with clearly defined fact and dimension tables. This structure enhances query performance for aggregations and time-series analysis.

Dimensions:

- **DIM_USERS** – User information
- **DIM_ARTISTS** – Artist information
- **DIM_ALBUMS** – Album metadata
- **DIM_TRACKS** – Track details
- **DIM_GENRES** – Genre descriptions
- **DIM_SUBSCRIPTIONS** – Subscription type metadata
- **DIM_DATE** – Date dimension for temporal analysis

Fact Tables:

- **FACT_SESSIONS** – Records of user streaming activity
- **FACT_USER_SUBSCRIPTIONS** – User-subscription history
- **BRIDGE_TRACK_GENRES** – Connects tracks and genres (many-to-many)



Bus Matrix Design

The Bus Matrix helps in identifying the analytical capabilities across business processes and aligns them with respective fact and dimension tables.

Business Process	Fact Table	Grain	Type
User Subscription Tracking	user_subscriptions	One row per user per subscription	Transaction
Session Activity	sessions	One row per session	Transaction
Track Info	tracks	One row per track	Periodic Snapshot
Album Release Analysis	albums	One row per album	Periodic Snapshot
Genre Coverage	tracks-genres	One row per track per genre	Factless

The matrix guides which dimensions relate to which fact tables, facilitating reusable dimension modeling.

				Dimension Tables						
Business Process Name	**Fact Table**	**Grainularity**	**Fact Grain Type**	**dim_users**	**dim_tracks**	**dim_albums**	**dim_artists**	**dim_genres**	**dim_date**	**dim_subscriptions**
User Subscription Tracking	user_subscriptions	One row per user per subscription	Transaction	X						X
Session Activity	sessions	One row per session record	Transaction	X	X				X	
Track Information	tracks	One row per track	Periodic Snapshot		X	X	X	X		
Album Release Analysis	albums	One row per album	Periodic Snapshot			X	X		X	
Genre Coverage Analysis	tracks-genres	One row per track per genre	Factless Fact Table		X			X		

Detailed Dimensional Modeling

We conducted attribute-level design for each table. Each dimension was enriched with surrogate keys, relevant metadata, and sourced directly from raw input tables. All date/time fields were normalized for compatibility with the DIM_DATE table.

Highlights:

- Consistent data types and naming conventions
- Use of surrogate keys for dimensional stability
- Cardinality checks for fact-dimension linkage
- Pre-calculated durations and states for faster analysis

Column Name	Description	DataType	P	uniq	not_n	Source
genrekey	dimension key		x	x	x	surrogate key
genre_id	primary key of the source systems (busir	INT	x	x	x	genres.genre_id
genres_genre_name	Genre name	VARCHAR(255)		x	x	genres.genres_genre_name
genres_description	Genre description	VARCHAR(255)				genres.genres_description
subscriptionkey	dimension key		x	x	x	surrogate key
subscription_id	primary key of the source systems (busir	INT	x	x	x	subscriptions.subscription_id
subscriptions_subscription_	Subscription name	STRING		x	x	subscriptions.subscriptions_subscriptior
userkey	dimension key		x	x	x	surrogate key
user_id	primary key of the source systems (busir	INT	x	x	x	users.user_id
users_first_name	User first name	STRING		x	x	users.users_first_name
users_last_name	User last name	STRING		x	x	users.users_last_name
users_user_name	User name	STRING(50)		x	x	users.users_user_name
users_user_email	User email	STRING(255)		x	x	users.users_user_email
users_date_of_birth	User date of birth	DATE		x	x	users.users_date_of_birth
users_state	User state	STRING(50)		x	x	users.users_state
users_subscription_type	User subscription type	STRING(50)		x	x	users.users_subscription_type
artistkey	dimension key		x	x	x	surrogate key
artist_id	primary key of the source systems (busir	INT	x	x	x	artists.artist_id
artists_artist_name	Artist name	STRING		x	x	artists.artists_artist_name
artists_first_name	Artist first name	STRING				artists.artists_first_name
artists_last_name	Artist last name	STRING				artists.artists_last_name
artists_country	Artist country	STRING(50)				artists.artists_country
albumkey	dimension key		x	x	x	surrogate key
album_id	primary key of the source systems (busir	INT	x	x	x	albums.album_id
albums_title	Album title	STRING		x	x	albums.albums_title
albums_release_date	Album release date	DATE				albums.albums_release_date
albums_artist_id	Artist ID (FK)	INT			x	albums.albums_artist_id
trackkey	dimension key		x	x	x	surrogate key
track_id	primary key of the source systems (busir	INT	x	x	x	tracks.track_id
tracks_album_id	Album ID (FK)	INT			x	tracks.tracks_album_id
tracks_genre_id	Genre ID (FK)	INT			x	tracks.tracks_genre_id
tracks_artist_id	Artist ID (FK)	INT			x	tracks.tracks_artist_id
tracks_title	Track title	STRING		x	x	tracks.tracks_title
tracks_duration	Track duration	INT				tracks.tracks_duration
tracks_release_date	Track release date	DATE				tracks.tracks_release_date
tracks_genre	Track genre	STRING(50)		x	x	tracks.tracks_genre
tracks_language	Track language	STRING(50)		x	x	tracks.tracks_language
sessionkey	dimension key		x	x	x	surrogate key
session_id	primary key of the source systems (busir	INT	x	x	x	sessions.session_id
sessions_user_id	User ID (FK)	INT			x	sessions.sessions_user_id
sessions_track_id	Track ID (FK)	INT			x	sessions.sessions_track_id
sessions_start_timestamp	Session start timestamp	TIMESTAMP_NTZ	x		x	sessions.sessions_start_timestamp
sessions_end_timestamp	Session end timestamp	TIMESTAMP_NTZ				sessions.sessions_end_timestamp

Business Problem: Regional Genre Popularity Analysis

In a competitive music streaming industry, delivering personalized and geographically relevant content is essential for user engagement, retention, and monetization. Understanding regional listening preferences enables streaming services, record labels, and advertisers to align their strategies with the cultural and behavioral nuances of users across different states. However, deriving such insights from large-scale user activity data requires a robust and scalable data infrastructure. This business problem is addressed through the implementation of a Snowflake-based data warehouse in the Harmony Hub project, which supports structured analytical workflows and efficient querying.

Role of the Data Warehouse

The Harmony Hub data warehouse is designed to consolidate and model streaming activity data in a way that supports complex, high-value analytical queries. By integrating user demographics, session activity, and track metadata, the warehouse enables organizations to generate actionable insights. Specifically, the warehouse schema includes:

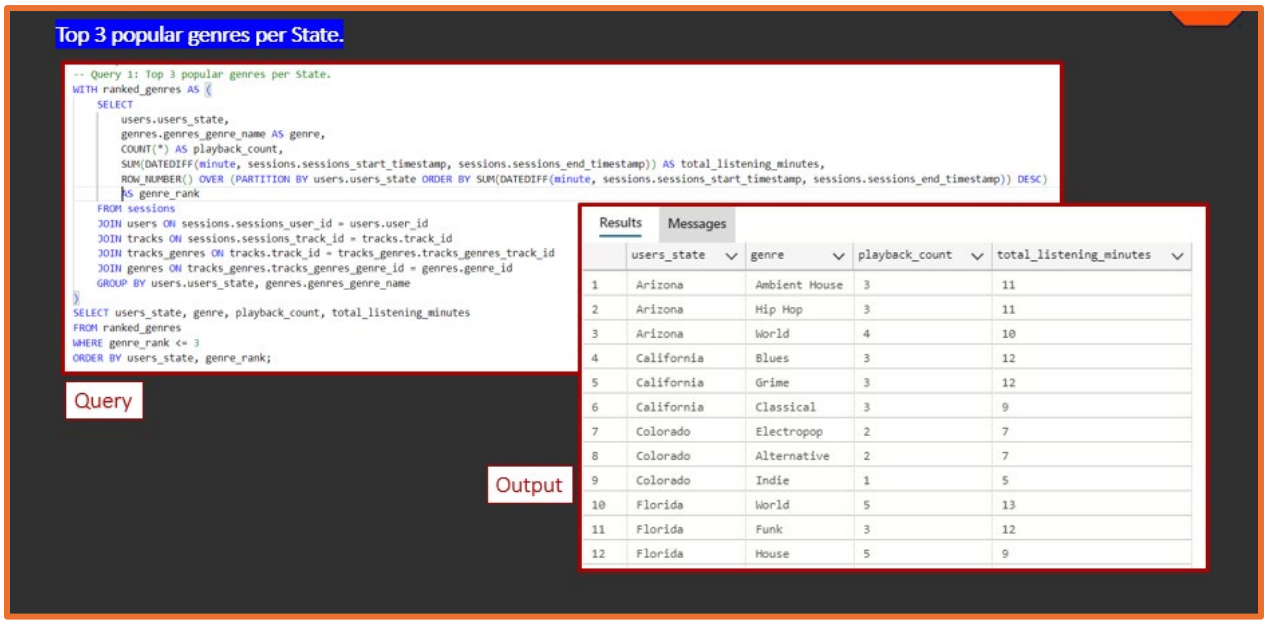
- **User Information** (dim_users) with location and subscription data
- **Streaming Sessions** (fact_sessions) capturing session-level engagement
- **Track and Genre Metadata** (dim_tracks, dim_genres, bridge_track_genres)

Through this architecture, SQL queries can be executed efficiently to join session data with genre classifications and user state information, enabling a detailed breakdown of musical preferences across regions.

Query Insight: Most Popular Genres per State

This specific analytical query provides a state-wise breakdown of the most popular music genres, evaluated through both the number of listening sessions and the total listening duration. The insight produced is highly valuable across multiple business functions:

- **Artists and Record Labels**: Gain visibility into where their genre of music is performing best, aiding in targeted marketing and tour planning.
- **Advertising and Marketing Teams**: Can optimize campaign targeting by aligning messaging with the musical preferences of specific geographic segments.



	users_state	genre	playback_count	total_listening_minutes
1	Arizona	Ambient House	3	11
2	Arizona	Hip Hop	3	11
3	Arizona	World	4	10
4	California	Blues	3	12
5	California	Grime	3	12
6	California	Classical	3	9
7	Colorado	Electropop	2	7
8	Colorado	Alternative	2	7
9	Colorado	Indie	1	5
10	Florida	World	5	13
11	Florida	Funk	3	12
12	Florida	House	5	9

For example, if the analysis shows that 'Ambient House' has the highest listening duration in Arizona, while 'Blues' is dominant in California, it signals strong regional differentiation in musical taste. Such findings inform product, content, and marketing decisions at both strategic and operational levels.

Business Process Context

This analysis aligns with the business process of **"Genre Coverage Analysis"**, as outlined in the project's bus matrix. The process leverages the tracks-genres factless fact table and joins it with relevant dimension tables to derive insights. The use of dimensional modeling ensures scalability and reusability of components across other analytical processes within the platform.

Business Problem: Artist Popularity by Age Demographics

In the highly personalized world of music streaming, understanding how different demographic segments interact with artists is key to effective audience engagement. One of the most valuable dimensions in audience analysis is age group, which often correlates strongly with musical taste, content preferences, and consumption patterns. For artists, record labels, and streaming platforms, insights into generational listening behavior are essential for targeting and outreach.

This business problem is addressed within the Harmony Hub data warehouse through age-based segmentation and user activity analysis, enabled by a robust dimensional model and efficient query design.

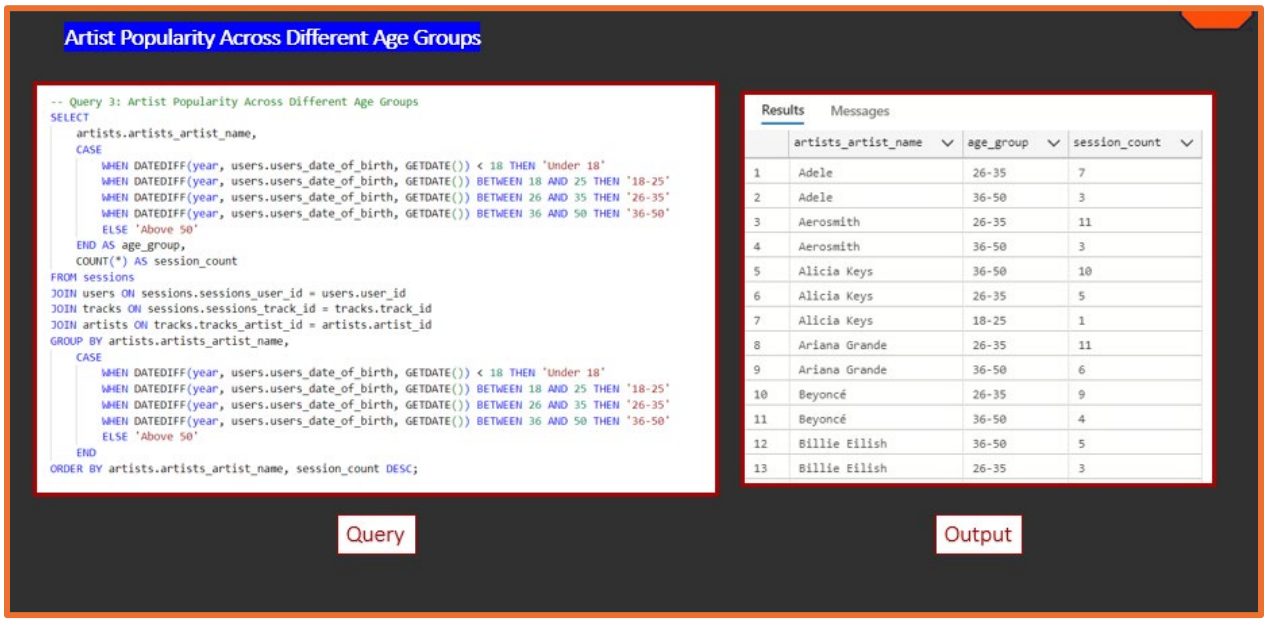
Role of the Data Warehouse

The Harmony Hub data warehouse enables this analysis through the integration of user demographic data with artist-level streaming metrics. The relevant data points include:

- **User Demographics** from the dim_users table (age, gender, location)
- **Streaming Sessions** from the fact_sessions table (track played, duration, timestamp)
- **Artist Metadata** via the dim_artists and dim_tracks tables (track-to-artist mapping)

The data is structured to allow for analytical queries that group session data by artist and user age group, making it possible to identify which artists resonate with which generational cohorts.

Query Insight: Artist Popularity Across Age Groups



This analysis identifies the most popular artists within distinct age brackets, such as “Under 18,” “18–30,” “31–50,” and “Above 50.” It aggregates listening metrics—such as session count and total play duration—by user age and artist.

These insights enable the following strategic outcomes:

- **Artists and Labels**: Can segment their marketing campaigns and tailor content releases to the age groups most aligned with their current fanbase.
- **Streaming Services**: Can improve personalization algorithms and playlist generation by factoring in age-specific preferences.
- **Event Planners and Promoters**: Can use this data to forecast demand for concerts or promotions among specific age demographics.

For example, if an artist is shown to be particularly popular with listeners aged 18–30, their promotional efforts such as social media campaigns or concert tours can be optimized to target that audience. Conversely, if another artist shows a strong following among users over 50, outreach could focus on more traditional channels or curated experiences for mature listeners.

Business Process Context

This analysis supports the business process of **"User Segmentation by Demographics"**, enabling data-driven decision-making for audience targeting. It leverages facts from the fact_sessions table and dimensions from dim_users and dim_artists, demonstrating the power of the warehouse's star schema design in enabling multidimensional insights.

Business Problem: Track Popularity and Listening Behavior by Time of Day

In a data-driven music streaming environment, understanding temporal listening behavior is vital for curating relevant content, improving user experience, and optimizing system performance. Listeners' preferences and engagement levels often vary significantly depending on the time of day, reflecting lifestyle patterns such as work, study, exercise, or relaxation routines.

This business problem is addressed through time-based session analytics within the Harmony Hub data warehouse. The ability to measure and analyze track popularity and session duration across different hours of the day offers key insights for strategic content planning and user engagement.

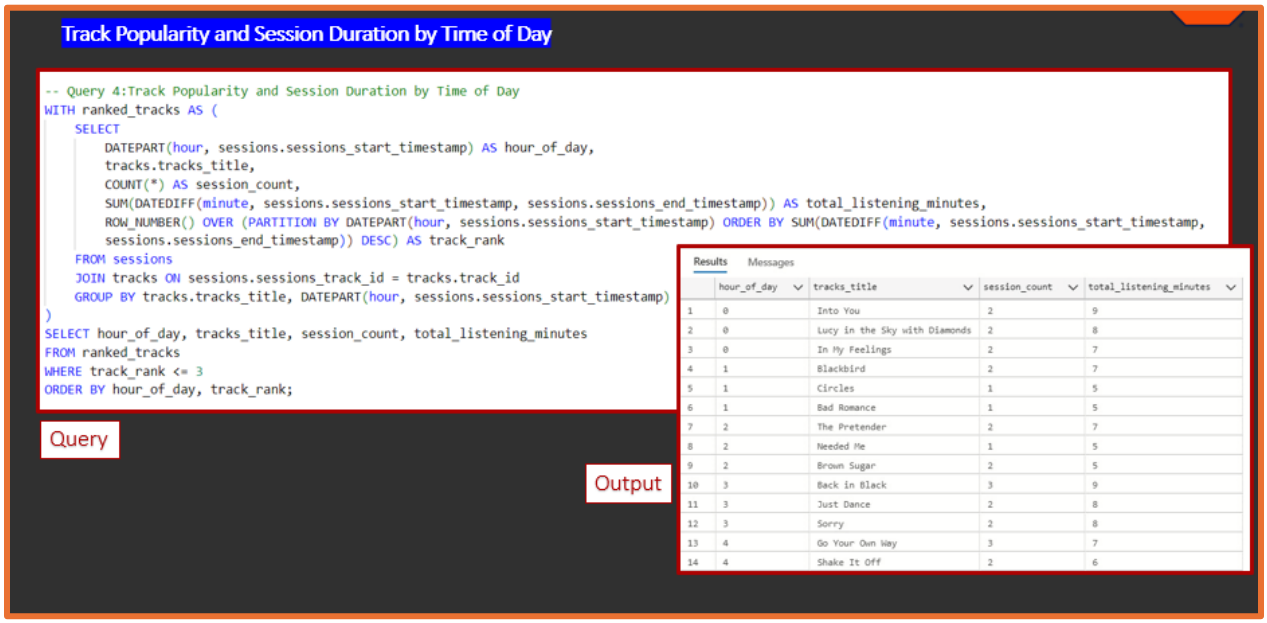
Role of the Data Warehouse

The Harmony Hub data warehouse integrates timestamped session data with track metadata, enabling time-of-day analysis through the following components:

- **Session Logs** from fact_sessions, capturing session start times, end times, and associated track IDs
- **Track Metadata** from dim_tracks for context on each track
- **Date and Time Dimensions** via dim_date and an optional dim_time or derived hour field

The data model supports efficient filtering and aggregation to analyze listening activity across 24-hour intervals, enabling precise insights into user behavior throughout the day.

Query Insight: Track Popularity and Session Duration by Time of Day



This query provides a breakdown of the most popular tracks and average session durations across different hours (e.g., early morning, midday, evening, late night). It enables the identification of **hour-specific trends**, such as:

- Tracks with the highest engagement during **morning commutes** (e.g., energetic or motivational music)
- Relaxing tracks gaining popularity during **evening hours**
- Increased session durations during **late-night streaming**, possibly indicating deep engagement or passive listening

These insights enable several business outcomes:

- **Streaming Platforms**: Can dynamically update homepage recommendations and playlists based on real-time listening trends
- **Content Teams**: Can schedule feature placements or track promotions during peak engagement windows
- **Advertisers**: Can target campaigns with contextually appropriate messaging (e.g., fitness ads during morning sessions, wellness ads in the evening)

For instance, discovering that lo-fi chillhop tracks peak in the 9 PM to 11 PM window can inform content promotion strategies and targeted playlist updates.

Business Process Context

This analysis is aligned with the business process of **"Temporal Listening Analysis"**, which is central to enhancing user engagement through contextual personalization. It utilizes fact_sessions with derived time-of-day groupings and joins with track-level metadata, demonstrating how time-based behavioral analytics are supported by the dimensional warehouse design.

Power BI Dashboards: Business Value Through Data Warehousing

The Power BI dashboards built for the Harmony Hub project serve as a powerful visualization layer atop the Snowflake Data Warehouse. They translate raw streaming, subscription, and user data into actionable insights, addressing key business challenges across user engagement, content strategy, and operational planning.

1. Executive Overview Dashboard: Monitoring Platform Health

This dashboard functions as the centralized reporting layer for key performance indicators. It supports stakeholders in evaluating platform activity, subscription adoption, content performance, and user engagement trends.

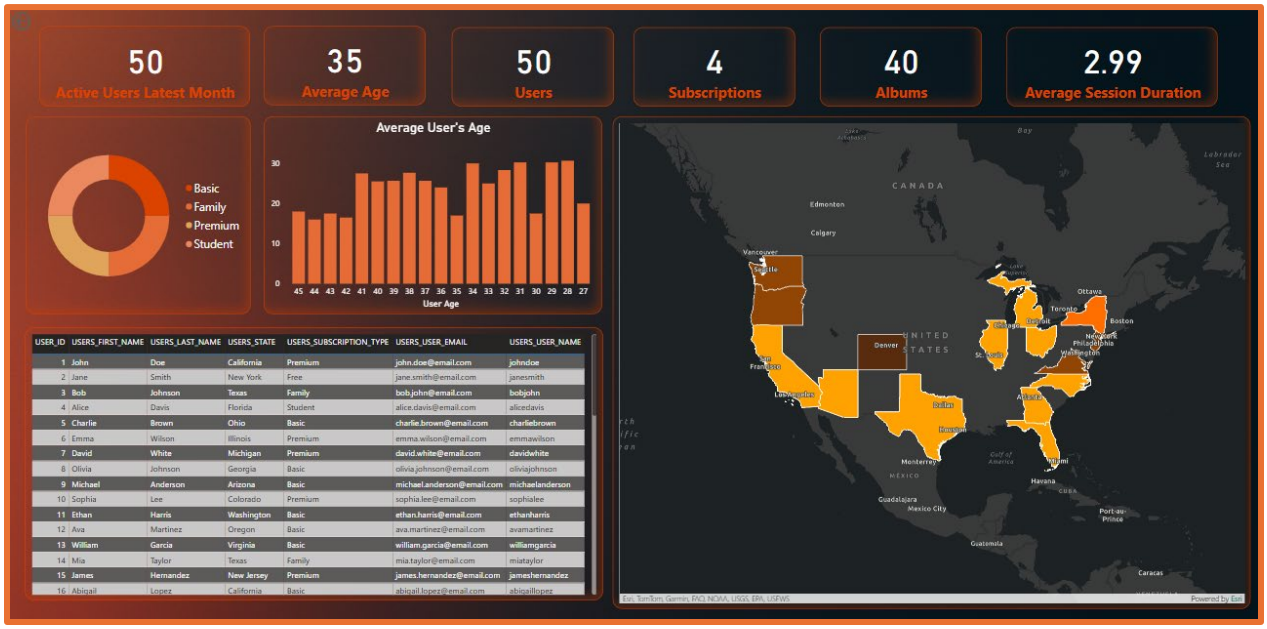
Business Problems Addressed:

- How is the overall user engagement trending over time?
- Are subscription plans being adopted consistently across the platform?
- Which music genres and album releases are gaining traction month-over-month?
- How effective are user retention strategies based on session duration?

The underlying warehouse consolidates user activity logs, genre classifications, album metadata, and subscription types into well-structured dimension and fact tables. This structure enables seamless aggregation, time-series comparison, and cross-category filtering within Power BI.

---

2. Users & Subscriptions Dashboard: Understanding Audience Segments



This dashboard provides a deep dive into user demographics, subscription preferences, and geographic distribution. It facilitates segmentation-based strategy formulation for product, marketing, and customer experience teams.

Business Problems Addressed:

- What are the dominant age groups engaging with the platform?
- How do subscription preferences vary across different demographic or geographic segments?
- Which regions show higher engagement, and where are growth opportunities?

Data from *dim_users*, *dim_subscriptions*, and *fact_sessions* is joined within the warehouse to offer a complete profile for each user. This allows Power BI to filter and visualize user behavior by age, region, and subscription type supporting personalized marketing and regional campaign planning.

---

3. Artists & Tracks Dashboard: Informing Content and Licensing Strategy

This dashboard focuses on artist popularity, genre preferences, and the lifecycle of track engagement. It is especially useful for content managers and record labels aiming to optimize catalog curation and promotional efforts.

Business Problems Addressed:

- Which artists are most influential across different listener segments?
- How does genre popularity shift over time?
- Are track releases aligned with seasonal engagement patterns?

By modeling *dim_artists*, *dim_tracks*, *dim_genres*, and their relationships with *fact_sessions*, the data warehouse enables detailed artist- and genre-level analytics. Power BI consumes this structured data to generate comparative views and performance rankings, supporting smarter content acquisition and promotion.

Across all dashboards, Snowflake's data warehouse architecture ensures:

- **Centralized, accurate, and cleansed data** for trustworthy reporting
- **Optimized schema designs** (star schema and bridge tables) for performance
- **Time-based snapshotting and trend analysis** using dimensional modeling

These dashboards not only answer key business questions but also serve as tools for **continuous optimization** enabling Harmony Hub to evolve its content and service offerings based on rich, data-driven insights.
