HelloBetter Challenge:

Objective:

You are joining the Data team at HelloBetter, a mental health platform providing digital therapeutic solutions. Your role as Data Engineer involves designing, building, and maintaining the data pipelines that extract data from various sources, transform it appropriately, and load it into our Data Warehouse for use by our Data Analysts, Data Scientists, ML Engineers, and stakeholders through Metabase.

HelloBetter has recently launched a chatbot that provides mental health support to users in the form of coaching. This chatbot generates detailed interaction logs, which are stored in a mongoDB database. Our Product team has requested that we integrate this data into our analytics platform and enrich it with weather data to analyze potential correlations between environmental factors and user mental states. For example, do individuals experience a happier mood on a sunny day?

Architecture Overview:

The project includes architecture that:

- Extract data from both the chatbot and weather API
- Load the data into S3 buckets
- Implements data quality checks and validation
- Transforms and joins the data for analytics purposes
- Makes the final datasets available in Metabase

Tools and Technology:

This project is entirely deployed and executed in the cloud, utilizing Amazon Web Services (AWS) as the primary platform.

- **Programming Language:** Python – Used for scripting data ingestion workflows and implementing pipeline logic.
- **Data Analytics:** SQL – Used for data exploration, analysis, and querying within the data warehouse.
- **Processing Engine:** Apache Spark – Used for large-scale data processing and transformation tasks.
- **Dashboarding Tool:** Metabase – Used for creating dashboards, performing analysis, and visualizing insights.

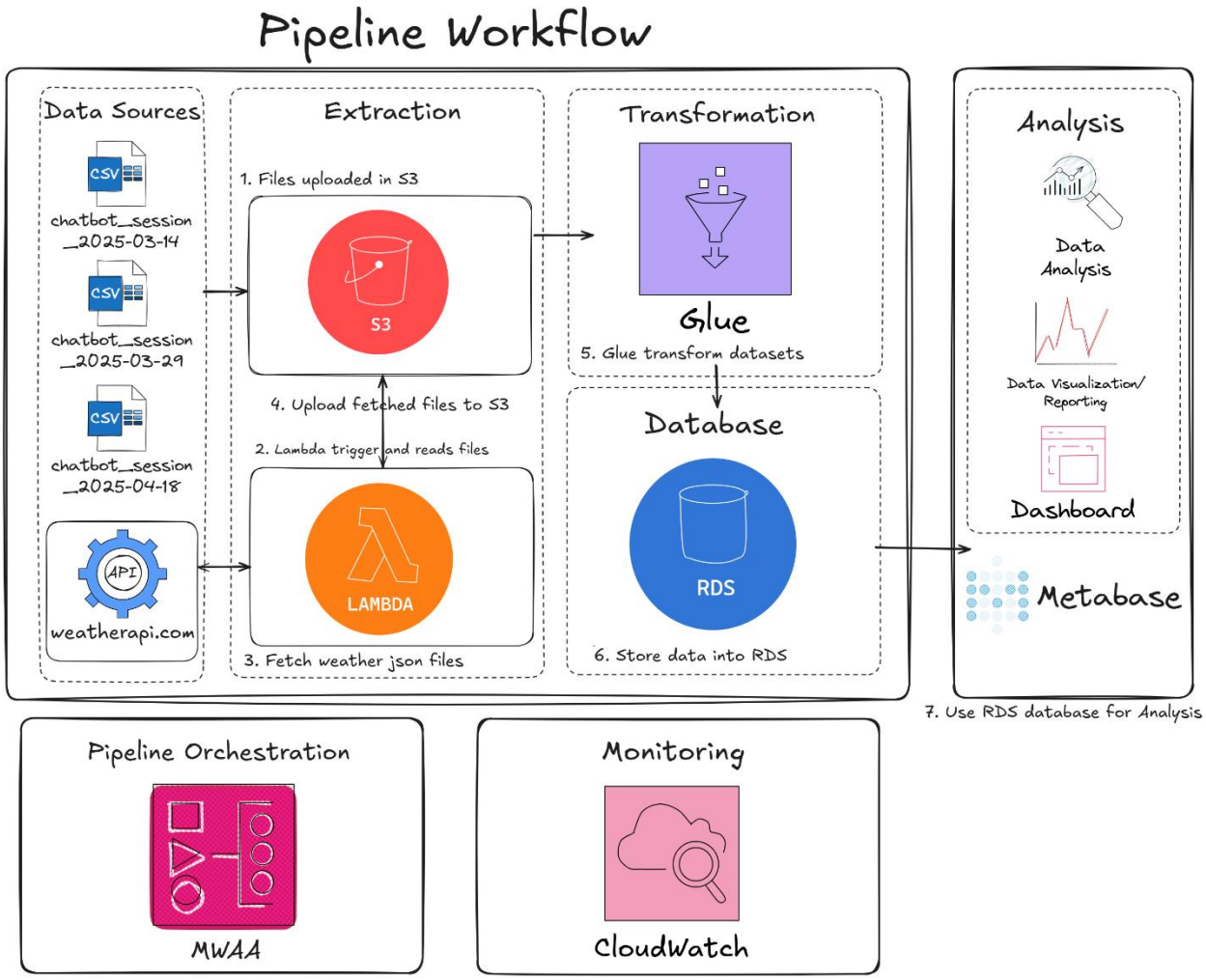


Figure 1: The diagram illustrates the end-to-end data pipeline architecture and processing stages from ingestion to reporting.

Data Extraction:

The Hellobetter datasets was sourced from HelloBetter. Under the **Data Source** section, a ZIP file was available for download containing three CSV files:

- chatbot_sessions_2025-03-14
- chatbot_sessions_2025-03-29
- chatbot_sessions_2025-04-18

These files were uploaded to an **Amazon S3 bucket** (hellobetter-S3-bucket). The upload event triggered an **AWS Lambda function** (extract-weather-data), which extracted the location and date fields from chatbot_sessions_* files. The Lambda function then sent requests to WeatherAPI.com to retrieve the corresponding weather data for each location and date.

Once the weather data was received, the function stored the API responses back into the same S3 bucket as csv files.

- weather_2025-03-14
- weather_2025-03-29
- weather_2025-04-18

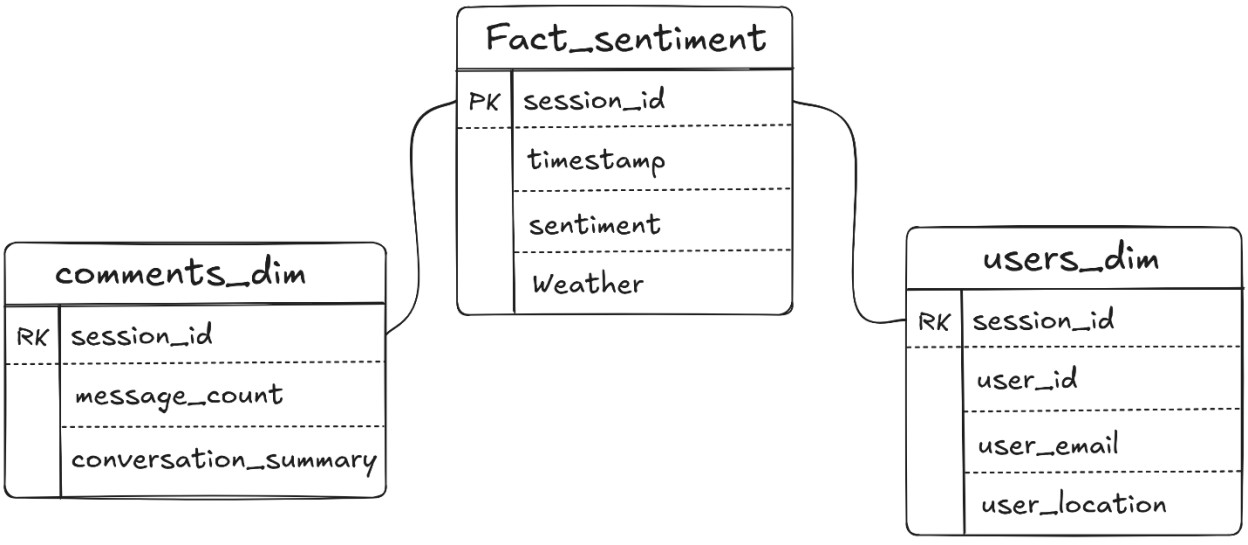
Now, both the chatbot session data and the corresponding weather data were available, the next stage is data transformation.

Transformation:

Data transformation was carried out using **AWS Glue**. An AWS Glue ETL job processed the data by performing the following steps:

1. Data Merging: Matched and merged the chatbot session data with the corresponding weather data for each date.
2. File Consolidation: Since the datasets contained similar structures across different dates, all processed date files were combined into a single dataset.
3. Data Modeling: Designed a schema optimized for analytical queries. This included creating two source tables and one analytics-oriented fact table containing relevant metrics to answer key questions, such as:

Do individuals experience a happier mood on a sunny day?



Storage:

Once the data was processed in the AWS Glue ETL job, it needed to be stored in a structured and easily accessible format. Given the relational nature of the data, **Amazon RDS** was selected as the storage solution. An RDS instance was created, and its endpoint was used to connect to the database.

After the data was loaded into RDS, it became accessible for querying and analysis. The RDS instance could also be connected from a local environment using tools such as **pgAdmin**, enabling efficient data exploration and verification before moving to the analytics stage.

Dashboard and Analysis:

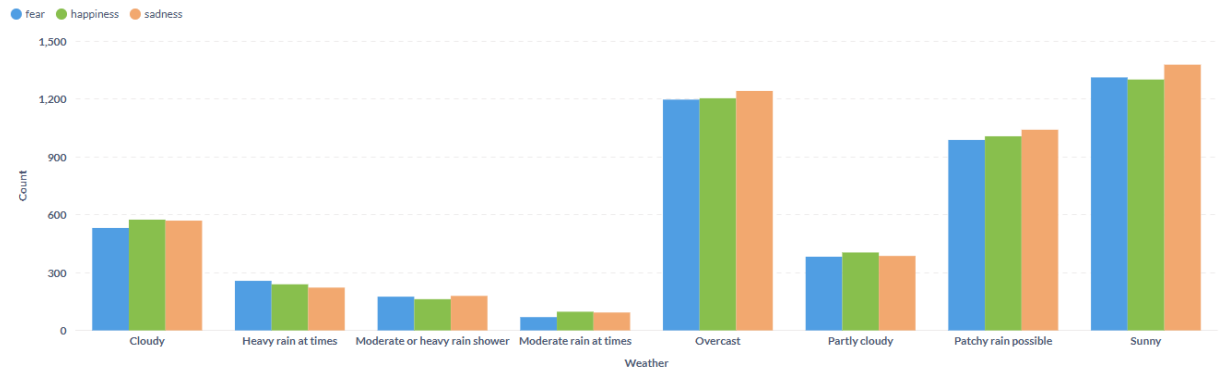
With all data successfully stored in Amazon RDS, it was time to address the primary business question:

Do individuals experience a happier mood on a sunny day?

For this purpose, **Metabase** was used as the dashboarding and visualization tool. Metabase was connected directly to the RDS instance, enabling interactive querying and real-time visualization. A dedicated dashboard was created to present key insights, allowing stakeholders to explore trends, compare sentiment across different weather conditions, and answer business questions effectively.

Business Questions:

How does weather affect sentiment?



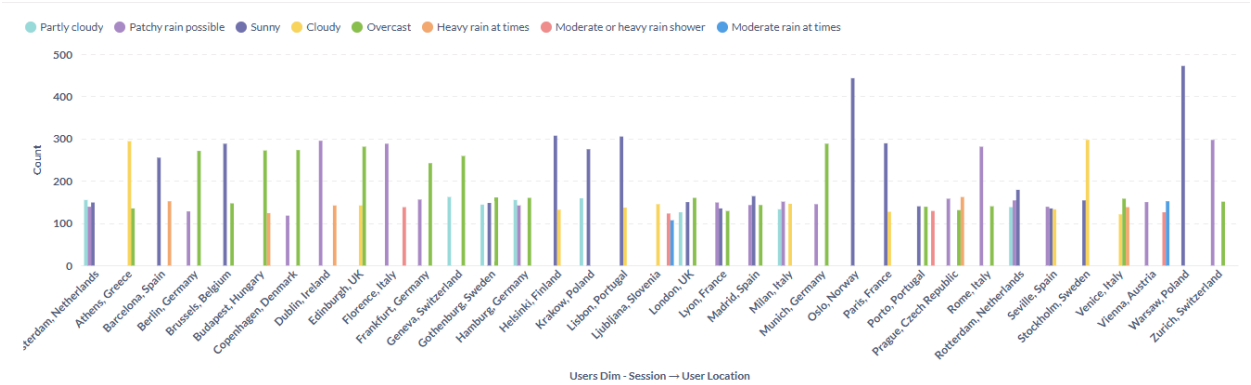
How does sentiment change over time (daily)?

In which month were people happiest?

Which location has the highest percentage of “happy” sentiment?

Sentiment patterns by location:

Weather patterns by location:



Sentiment pattern by Weather in all Month:

Sunny	26.59%
Overcast	24.27%
Patchy rain possible	20.22%
Cloudy	11.16%
Partly cloudy	7.81%
Heavy rain at times	4.79%
Moderate or heavy rain shower	3.44%
Moderate rain at times	1.73%