Cloud Analytics: End-to-End Data Engineering & Analytics Project

1. Project Overview

The Yelp dataset is a large, real-world dataset containing reviews, ratings, businesses, and user information. The dataset used in this project is approximately **5 GB in JSON format**. Processing such a massive semi-structured file locally is inefficient and slow.

To address this challenge, I built a **cloud-based analytics pipeline** that enables scalable storage, transformation, and analysis of Yelp data.

The workflow followed was:

1. **Split the large JSON file** into 10 smaller files using Python for easier ingestion.
2. **Upload the files to AWS S3** for cloud-based storage.
3. **Ingest data into Snowflake** from S3 staging.
4. **Flatten the JSON** into structured tables using Snowflake SQL.
5. **Apply a UDF (User-Defined Function)** in Snowflake for sentiment analysis of review text.
6. **Perform SQL-based analytics** to answer business questions and generate insights.

This project demonstrates a complete **end-to-end data pipeline** using modern cloud technologies.

---

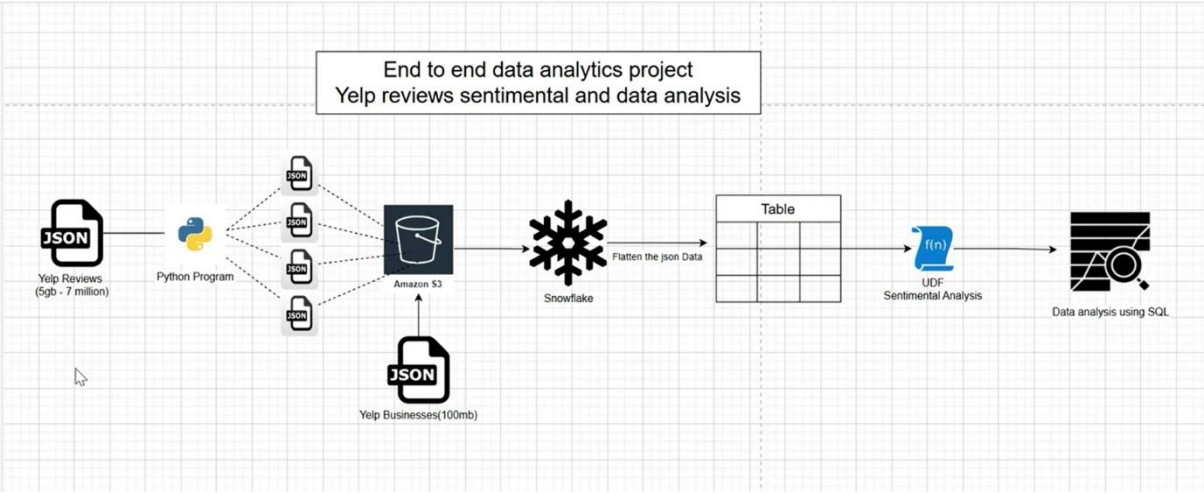
2. Problem Statement

- The Yelp dataset is **too large to analyze locally**. Handling a 5GB JSON file on a personal machine is impractical.
- The data is **semi-structured JSON**, making direct querying and aggregation inefficient.
- Businesses require **structured, query-ready datasets** to derive meaningful insights into customer behavior, trends, and performance.
- The challenge is to design a **scalable and automated workflow** that transforms raw, unstructured data into actionable insights.

---

3. Architecture Diagram

Pipeline Flow:



---

4. Technology Stack

- **Python** → Data preprocessing, splitting large JSON file into manageable chunks.
- **AWS S3** → Cloud storage for raw and staged datasets.
- **Snowflake** → Cloud data warehouse for loading, storing, and querying data at scale.
- **SQL (Snowflake-specific)** → Used for table creation, transformations, UDF application, and analytics.
- **Snowflake UDF (Python-based)** → Applied sentiment analysis directly inside Snowflake for textual reviews.

---

5. Conclusion

This project demonstrates how **cloud-based data engineering and analytics** can be applied to large, real-world datasets like Yelp.

Key takeaways:

- **Scalability**: Cloud storage (S3) and Snowflake allow processing of massive JSON files.
- **Flexibility**: JSON data was flattened into structured, query-ready tables.
- **Innovation**: Sentiment analysis via Snowflake UDF added valuable context to reviews.
- **Business Value**: Analytical queries produced insights into customer behaviour, business performance, and engagement trends.