AI-Generated Travel Itinerary

INFO 7390 – Advances Data Sci/Architecture

Ram Srikar Putcha (002304724)

Table of Contents

Executive Summary

The project introduces an AI-powered Travel Itinerary Generator which helps in simplifying trip planning. The structured data (e.g., things to do, attractions, hotels), and unstructured data (YouTube transcripts) helps in itinerary generation. The data pipeline is established through Airflow, Snowflake, and Pinecone to store structured and unstructured data. The frontend, built with Streamlit, which offers an interactive platform for users.Streamlit along with FastAPI, LLMs, and CrewAI agents helps in generating customized itineraries based on user preferences. The system aims to reduce planning effort, and overall travel experience.

Project Overview

This project builds an intelligent end to end system that focuses on automating the creation of personalized travel itineraries using LLM and big data. It combines structured data – hotels, attractions in the cities and various day tours available in cities, with unstructured content from YouTube video transcripts to create full-fledged travel plans. The project provides options to users to input their preferences like travel destination, dates, and interests. Once the options are provided the system creates day wise itineraries based on individual travel style.

The unique characteristic of the system is the orchestration of various technologies working together. Apache Airflow handles automated data ingestion into Snowflake to store structured data, and into Pinecone to store unstructured embeddings. LLM is paired with different CrewAI agents which helps in the itinerary generation logic. The frontend of the system is developed using Streamlit for a smooth user experience and FastAPI for the backend communication. The entire data pipeline is designed with the focus on scalability and real-world usability, so the system is both technically sound and actually helpful to a wide range of users.

Problem Statement

Travel itinerary planning is time-consuming, and a frustrating task for travelers. Travelers must sort through numerous sources, both structured data sources (hotels, restaurants, attractions) and unstructured content (guides, blogs, user reviews), and manually craft their own customized plans. This leads to suboptimal decisions, missed local experiences, inaccurate budget projections.

The solution presented is an AI-driven Travel Itinerary Generator based on structured hospitality and tourism database information, as well as knowledge from unstructured YouTube video transcripts. The system focuses on six major US cities: Chicago, New York, San Francisco, Los Angeles, Las Vegas, and Seattle. The solution aims to leverage automation for itinerary creation through the ETL pipelines (Airflow), structured data stored in Snowflake and unstructured data stored in Pinecone, backend APIs (FastAPI), autonomous CrewAI agents, and Large Language Models (LLMs). Users will interact using Streamlit frontend, being presented with personalized itineraries dynamically created according to users' requirements (destination, dates, interests, budget).

By automating and personalizing the process of itinerary creation, the suggested system aims to significantly boost user interaction, reduce travel planning effort, optimize budgets, and offer highly personalized recommendations—overall enhancing the quality and satisfaction of travel experiences.

Project Objectives

1. Integrating Various Sources of Data:

- Integrate structured data sources like hotel directories, tours of various places in cities, attractions with unstructured sources like YouTube transcripts to offer real-time personalized plans

2. Simplifying the Travel Planning Process:

- The project aims to remove the intimidating task of researching cities and aggregating travel information manually by offering a one stop intelligent solution that offers end-to-end day-wise trip plans within seconds

3. Offering Real-Time Personalization:

- Design of the project focuses on an adaptive system that will help in personalizing travel recommendations based on real-time user inputs such as hotels, so each plan is unique and context-based

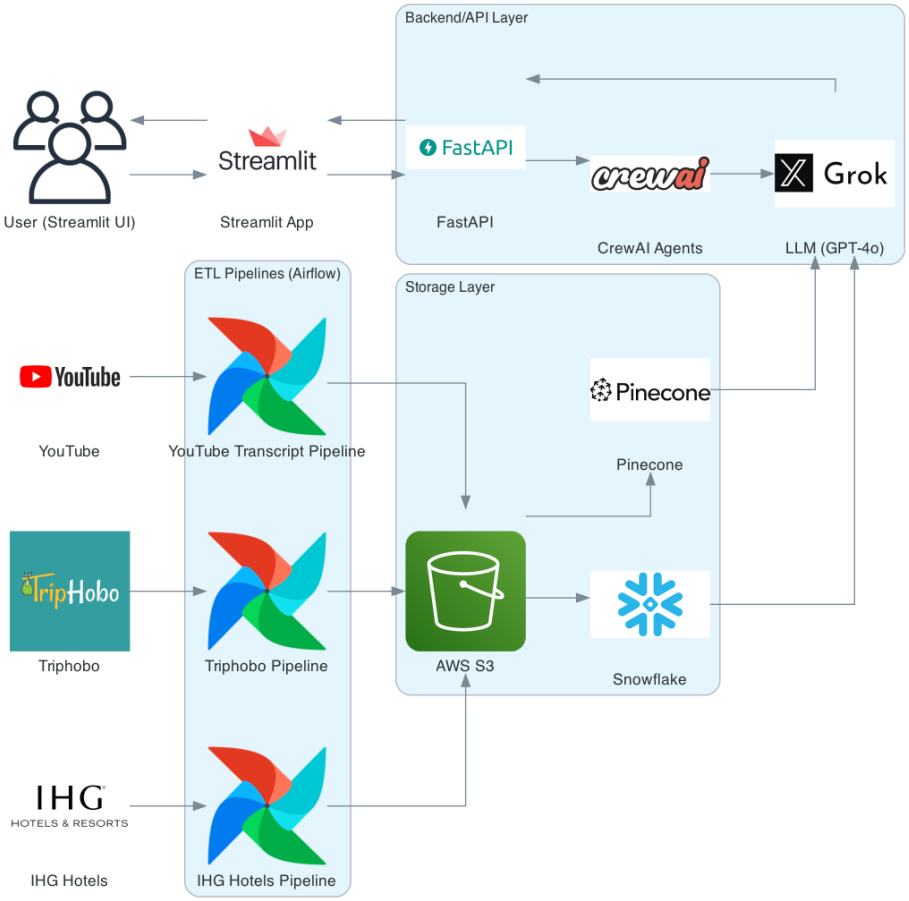
4. Creating a User based Experience:

- Prioritize usability by having an intuitive yet simple Streamlit user interface to make an AI-driven travel itinerary generation that is accessible to tourists, corporate travelers, and agencies on a platform

5. Leveraging Big Data, AI Agents and LLM:

- The project focuses on leveraging the power of big data processing using Airflow, AWS S3, and Snowflake, and pairing it with the potential of LLMs like OpenAI and autonomous CrewAI agents for trip planning

Architecture Diagram



AI Travel Itinerary System Architecture

Methodology

The system is designed using modular components connected through a series of data pipelines, intelligent CrewAI agents, and a Streamlit frontend interface. The development process is broken down into three key stages: Data Collection & Preprocessing, Backend Intelligence & Pipeline Integration, and Frontend Interaction & Itinerary Generation.

1. Data Collection & Preprocessing (Airflow Pipelines)

Three separate Apache Airflow pipelines were created to automate the data collection and storage process in appropriate databases:

- **Attractions Pipeline:**
Scrapes city tours of different places within the city, and popular tourist attractions from the Triphobo website for six cities (Chicago, New York, San Francisco, Los Angeles, Las Vegas, Seattle). The raw data collected is first stored in AWS S3 before preprocessing. Then the data stored in S3 is cleaned, preprocessed, and loaded into Snowflake for structured querying

- **Hotel Data Pipeline:**
Collects daily hotel information from the IHG website for the same six cities. The scraped data is stored in AWS S3, preprocessed, and saved in Snowflake for integration into the itinerary recommendation process

- **YouTube Vlogs Transcript Pipeline:**
Transcripts from YouTube travel vlogs related to the six cities are fetched using YouTube API. The transcripts are then stored in S3, chunked for context management, embedded using vector representations, and stored in Pinecone, enabling semantic retrieval of travel advice and local experiences

2. Backend Intelligence & Agent Orchestration

- The system uses CrewAI agents to gather data from different data sources based on the user query, prompt formatting, and coordination between user queries and the LLM

- FastAPI endpoints are built to connect the frontend Streamlit UI with the CrewAI agents and other backend services

- The Crew AI agents collect user preferences and pass them to a Large Language Model (LLM), which dynamically generates a detailed and structured day-wise itinerary using both

structured and unstructured travel data

3. Frontend Interaction (Streamlit UI)

- A simple and interactive Streamlit interface is developed for users to select the city, travel dates, and custom preferences like what to include in the itinerary

- These inputs are then sent to the FastAPI backend endpoints, and in response, the system displays a personalized, AI powered itinerary based on both the structured data in Snowflake and insights from unstructured content in Pinecone

Technology Stack

- Programming Languages

1. Python: Used for building ETL pipelines, backend and frontend development
2. SQL: To query the structured data stored in Snowflake

- Data Components

1. Apache Airflow: Building ETL pipeline for structured and unstructured data
2. Snowflake: Used to store structured data
3. Pinecone: Used to store embeddings of youtube transcripts

- Frameworks and Python Libraries

1. Streamlit: Used to build interactive user interface
2. FastAPI: Connects streamlit with CrewAI agents, handles API requests
3. CrewAI: Orchestrated autonomous agents and managing tasks
4. BeautifulSoup: Scrape data from website

- APIs:

1. YouTube API: Used to scrape transcripts of YouTube videos related to travelling
2. Grok API: Generates the travel itinerary based on data collected from different sources

- Cloud & Deployment Tools

1. AWS S3: To store the structured and unstructured data during ETL process
2. Google Cloud Platform: Used to host the application
3. GitHub: Collaboration and CI/CD pipeline
4. Docker: To dockerize the application and host it on GCP

Data Sources

To build the AI-powered travel itinerary generation system, a mix of structured and unstructured data from real-world sources was used to make sure that the travel plans suggested are relevant to the user's request and personalized.

1. City Tours & Attractions (Structured Data):

Travel information like popular city tours and must-see attractions from the Triphobo website. This included details like the name of the city, types of attractions, short descriptions, and how long each visit usually takes. After scraping the data from the triphobo website using playwright the data was cleaned and was stored in Snowflake so it can be easily searched and used to build custom itineraries.

2. Hotel Information (Structured Data):

The hotel listings from the IHG (InterContinental Hotels Group) website were scraped using BeautifulSoup. This gave the information such as hotel names, locations, amenities, ratings, and prices. Just like the attraction data, this hotel information was stored in Snowflake for easy access and retrieval.

3. YouTube Travel Vlogs (Unstructured Data):

To capture real, on-the-ground travel experiences, transcripts from YouTube travel vlogs were fetched using YouTube API. These videos contain detailed information about travel tips, personal stories, and hidden gems which might not be found on traditional travel websites. These YouTube transcripts were turned into searchable text using embeddings and were stored in Pinecone, which helps the system understand and find relevant insights.

By combining travel details with personal experiences from travelers, the system can create smarter, more personalized travel plans that feel both informative and local.

Application Testing and Validation

To make sure the system works reliably and according to the requirements several rounds of testing were performed during the development phase.

1. Unit Testing

I. Individual components of the system were tested in isolation. This included tasks within the Airflow DAG, FastAPI endpoints, and data preprocessing scripts
II. Functions for data extraction, transformation, and loading were verified using sample inputs to ensure correct outputs and consistent schema structure

2. Integration Testing

I. The complete data pipeline was tested end-to-end — from data scraping and storage in S3, to preprocessing, and loading into Snowflake or Pinecone — to confirm smooth and accurate data flow across each stage
II. Connections between the Streamlit frontend, FastAPI backend, and CrewAI agents were validated to ensure seamless interaction and proper response handling

3. Functional Testing

I. The Streamlit interface was tested with a variety of user inputs to evaluate dynamic itinerary generation across all six supported cities
II. Different combinations of travel dates, and preferences were used to ensure that itineraries adapted appropriately to the provided input

4. LLM Output Validation

I. Outputs from the large language model (OpenAI) were manually reviewed to assess relevance, clarity, and overall usefulness of the generated itineraries
II. Suggestions were cross-checked against real-world data, including attractions and accommodations, to confirm accuracy and reliability

Future Scope

The travel itinerary generation system demonstrates the potential of combining LLM, agents, and data from various sources towards intelligent planning of trips. In future iterations, the system can be extended by the inclusion of functionalities like user login, saving previous itineraries, and fine tuning of recommendations. The system can also be modified by the addition of real-time APIs like Skyscanner API or Eventbrite API for fetching data for flights, and local activities respectively. The current version sets the stage for the creation of a more feature-laden travel planning system.

Conclusion

The project aimed to make travel planning easier and more intelligent using GenAI techniques and big data technologies by consolidating multiple categories of travel data—like hotels, popular attractions, city tours, and YouTube travel vlogs transcripts, the system gathers all the necessary information in one place. It then uses tools like Airflow, Snowflake, Pinecone, and FastAPI to create personalized travel itineraries based on user interests.

Key innovations in the system are:

- A robust ETL pipeline created with Apache Airflow for effective data handling
- Backend API-based APIs with modular and scalable deployment using FastAPI
- Incorporating large language models (LLMs) for generating detailed, context-aware itineraries
- Use of agents (CrewAI) in multi-step plan execution (weather + budget + local tips)
- Live user interface with budget management, map displays, and itinerary modification

Working on this project allowed us to apply what we have learned throughout the course to a real issue. It gave us hands-on experience with existing and various emerging technologies and showed us how GenAI can be applied to improve tedious tasks like planning itineraries for vacations. While more can be added later, this version meets the minimum requirements and illustrates the potential of developing intelligent travel systems.

References

- CrewAI documentation
- Airflow documentation
- FastAPI guide
- Streamlit guide
- Pinecone Documentation
- YouTube Transcripts API