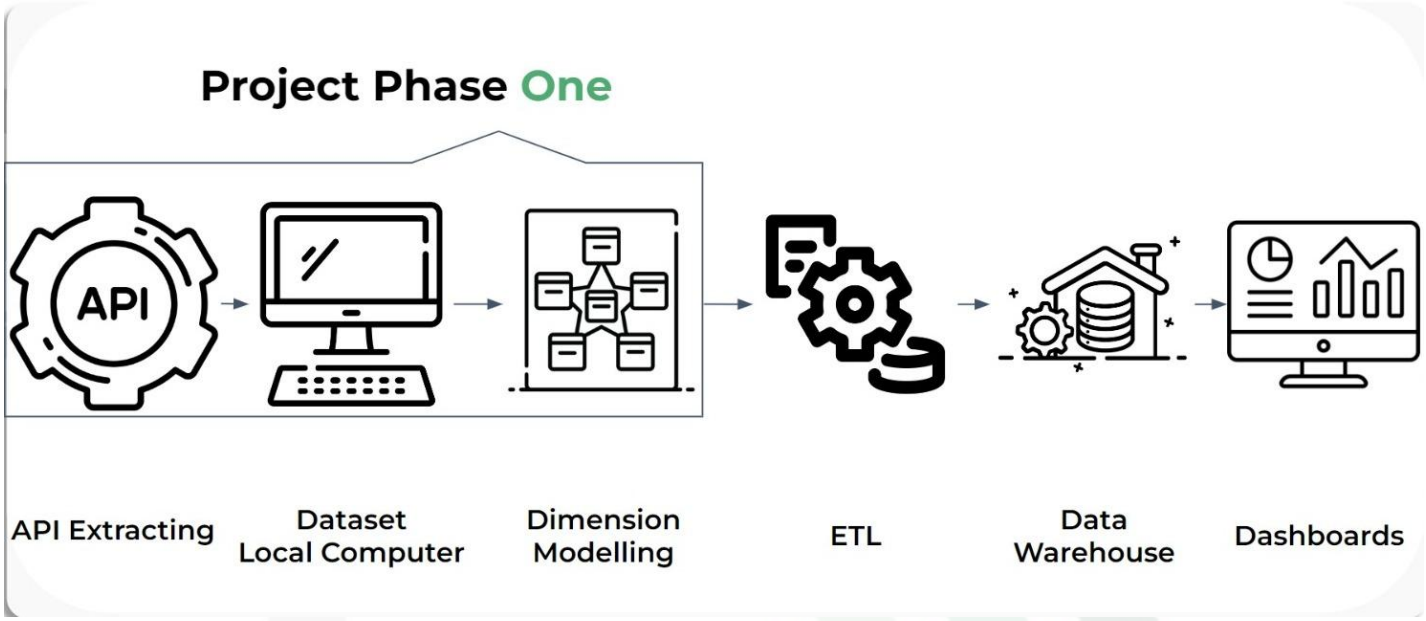
POTENTIA's summer training Project
Phase one

Project pipeline:



Use case:

Movalytics is a movie company that is currently facing challenges with their operations. They find themselves spending a significant amount of money on movie productions, but the total income generated from these movies falls short of their expected targets. In response, the company's managers have decided to conduct a thorough analysis to identify the root cause of this issue.

However, during their analysis, they realized that their existing data is not in a suitable format nor designed for in-depth analysis. To address this, they have made a decision to hire a skilled data engineer. The primary goal is to develop a robust data warehouse tailored to their business needs. This data warehouse will facilitate easy analysis and provide the necessary infrastructure for capturing valuable insights that can help identify the problem and propose effective solutions.

Your role:

As a junior data engineer you are hired to provide a solution for the Movalytics company. First ,you are required to get the data from its API, clean and transform it, then apply the necessary processes to make the data clear, easy to understand and useful. After that you have to put it in a structured format considering the suitable type of schema, dimensions and facts

where the goal is to make the jobs that will be applied fast, simple and clear without any redundancies.

1. Data Extraction using API

We are going to extract data from **OMDB** (Open Movie Database) is a comprehensive online database that provides information about movies, TV shows, and other video content. You can follow the following link to the website: https://www.omdbapi.com/.

This API poses a distinct challenge, unlike the APIs you've

previously interacted with. This marks your initial task, offering a chance to show your abilities and establish your proficiency.

You are required to extract only 100 movies from this API.

Here are some hints to start with:

1. Your endpoint will be (http://www.omdbapi.com/?apikey={yourkey}&t={movieTitle})
2. You are given a csv file containing movie titles : https://drive.google.com/file/d/1eVci9m4LrrOjSRwHbLTG3JPp-OzCfoC_/view?usp=drive_link

   - First, you need to extract the titles from the CSV file and store them in a list.

   - Then, you will use these collected titles to construct URLs (endpoints) for fetching data about each movie through individual API responses.

2. Data Modelling

After extracting the data from the API , you should start data modeling but before that you need to understand the data and define the columns you need and after the analysis the business owner said that there is no need to keep the following information (Poster, DVD, Type, Awards , Ratings , totalseasons , Response, Error , Website , Production). Finally you can start building your schema.

Here are some hints to start with:

1. Consider the main entities involved in the movie data.

2. Identify the primary keys for each entity and if it is not involved in the data you can create a primary key for it on your own.

3. If there are many-to-many relationships (multi-valued attributes) between entities, consider using bridge tables to establish the connections between them.

4. Consider the low cardinality attributes if there are any to build a junk table.