**Checklist of ETL Pipeline for Cancer Clinical Trials from Clinicaltrials.gov**

**☑ Understand the project objective and scope.**

**☑ Design the database schema with tables:** Trials, Conditions, Interventions, Trial\_conditions, Trial\_interventions, Sponsors, Locations, and Log\_pages.

**☑ Set up a PostgreSQL database.**

**☑ Develop the Extract phase:**

☑ Interact with ClinicalTrials.gov API to retrieve data in JSON format.

☑ Handle pagination for large datasets.

☑ Extract relevant fields like nct\_id, title, status, phase, conditions, interventions, sponsor, and location.

**☑ Develop the Transform phase:**

☑ Map data to database schema.

☑ Handle duplicate entries and maintain data integrity.

☑ Ensure missing or incomplete data is addressed.

**☑ Develop the Load phase:**

☑ Connect to PostgreSQL database.

☑ Insert data into appropriate tables while ensuring relationships are preserved.

**☑ Schedule the ETL pipeline to run periodically.**

**☑ Validate the ETL process by querying the database and checking data accuracy.**

**☑ Document the ETL process, schema design, and usage instructions.**