E-commerce DataWarehouse Project

Objective Statement:

An e-commerce company wants to develop a one-time (only for a specific data source) data warehouse pipeline to store all of its data incrementally from 2016 to 2018.

Company Size: Small - Medium

Limitations:

Constraints	Limitation	Description
Storage Capacity	Yes	Can’t afford multiple staging databases, have to work with as minimum storage as possible
Computational Cost and Resource	Yes	Can’t afford to use multiple computational resources’s cost on each incremental loading. One-time multiple resources are acceptable but must be reduced to a minimum on the second iteration.

Special Requirement: Fast ETL pipeline processing (quicker the process lesser the resource cost).

Available Knowledge:

- Source: Single CSV file.
- Nature of Source: Not changing (data won’t change).
- Source Size: 15MB file.
- Records in Source: Over a million rows.
- Source Data Integrity: Data is very messy.
- Source to Warehouse: Need to Load Incrementally Based on the Years.

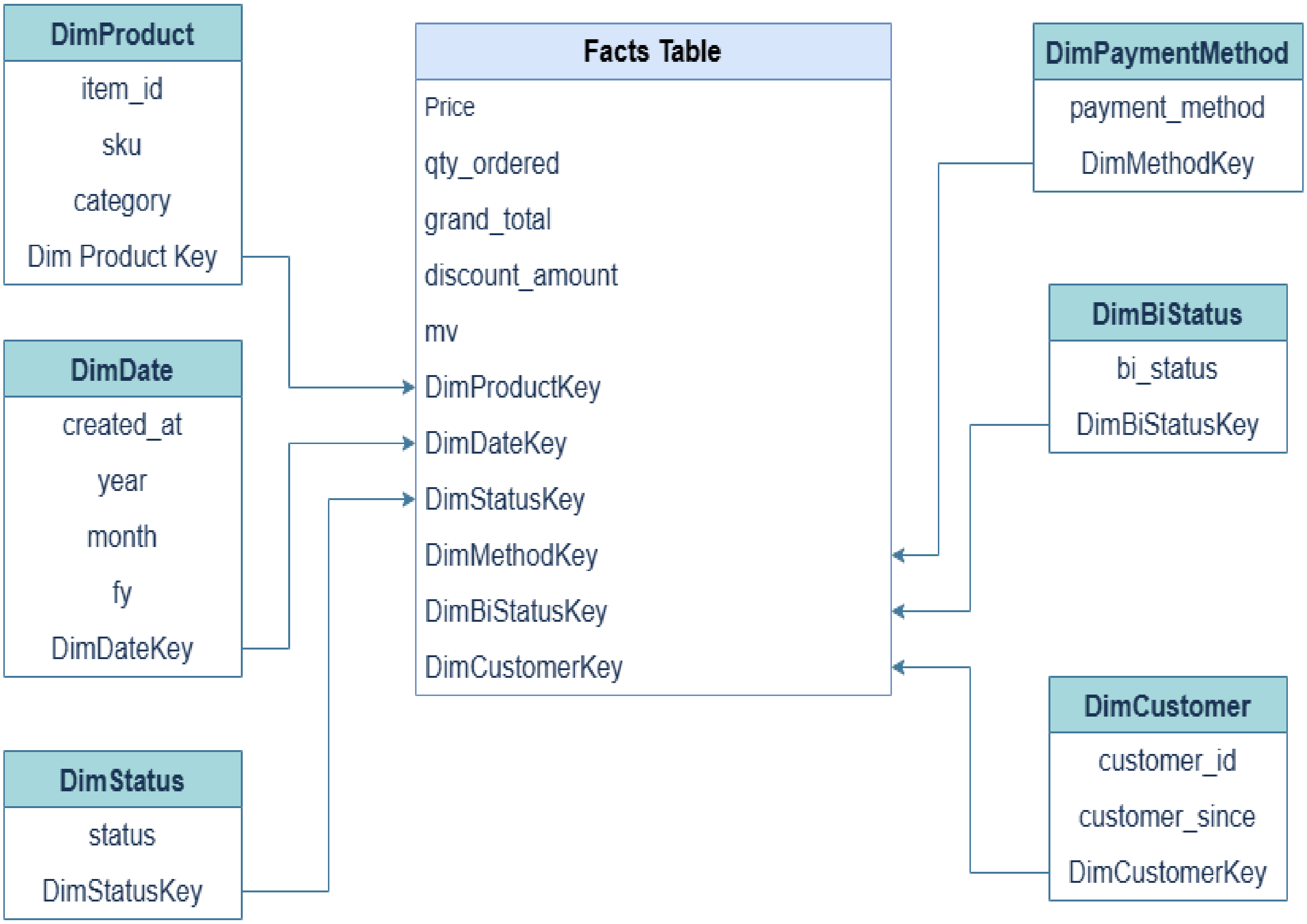
Process Plan:
From the above knowledge, I conclude to design the data warehouse pipeline as follows:

Stages	Extract	Transform & Cleaning	Load
1	Extract from source	Cleansing and transforming on the go	Load one time on a temporary table or file (this will be deleted as soon data is loaded to the staging table)
2	—------	—-----------	Load the clean data to the staging table
3	Extract from the staging table	Transform Data into Facts and Dimensions	Load to final data warehouse tables i.e facts & dimension

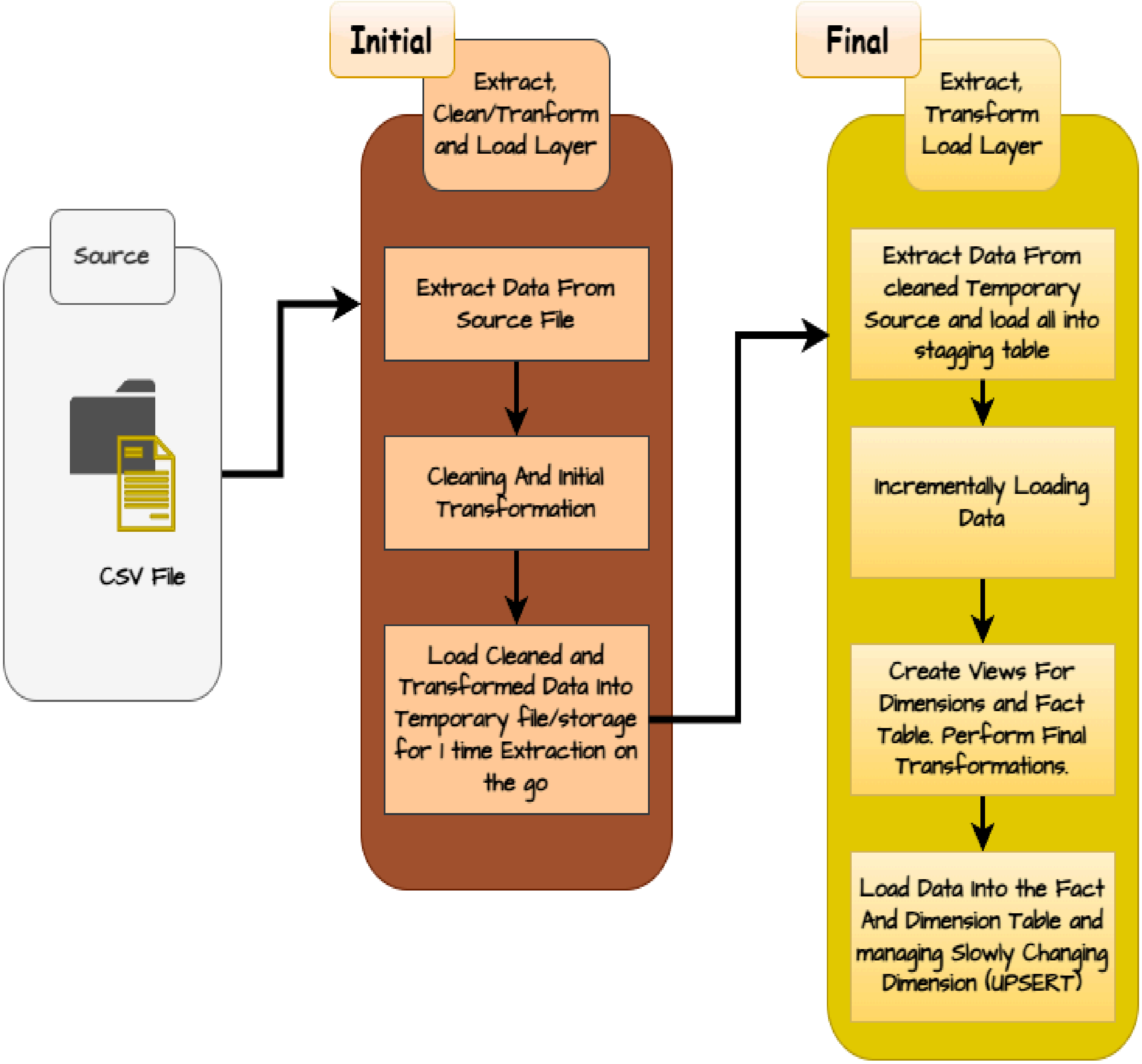
Thought Process

Stages	Process	Reason
1	Extract from the source and not dump raw data into any pre-processing table	1. Less storage resource 2. Minimizing overheads 3. Possible because of static data in the source
1	Cleaning and transformation on the go	1. Computational resource constraints of cleaning data on each increment. 2. Minimizing overheads 3. No cleaning is required in the data warehouse transformation stage, making the process fast. 4. Possible because of static data in the source
2	Loading of cleaned and transformed data into the data warehouse staging table	1. The staging table will now serve as a main source for incremental loading having cleaned data.
3	Extracting, Transforming into facts and dimensions,s and final loading	1. Need to perform incremental loading that's why it is best to transform data into facts and dimensions in this stage. 2. Managing Slowly changing dimensions is easily possible here.

FACT AND DIMENSIONS TABLES



Work Flow Diagram


Source
CSV File
Initial
Extract, Clean/Tranform and Load Layer
Extract Data From Source File
Cleaning And Initial Transformation
Load Cleaned and Transformed Data Into Temporary file/storage for 1 time Extraction on the go
Final
Extract, Transform Load Layer
Extract Data From cleaned Temporary Source and load all into stagging table
Incrementally Loading Data
Create Views For Dimensions and Fact Table. Perform Final Transformations.
Load Data Into the Fact And Dimension Table and managing Slowly Changing Dimension (UPSERT)