

Data Management II – ETL and Smart Data Project

(Product Sales Data)

Presented by:

Shekhar Singh (11011694)

Content

- Introduction
- Problem Statement
- Data Profiling
- ETL Pipeline
- Star Schema
- Demo

Introduction

- **Dataset** : Product Sales Data
- Fictitious company called Adventure Works Cycle (AWC) , a multinational manufacturer and seller of bicycles and its related accessories
- **Records** : 1048543 ; **Attributes** : 13
- Fails on Completeness, Consistency, Accuracy, Conformity and Validity dimensions of Data Quality

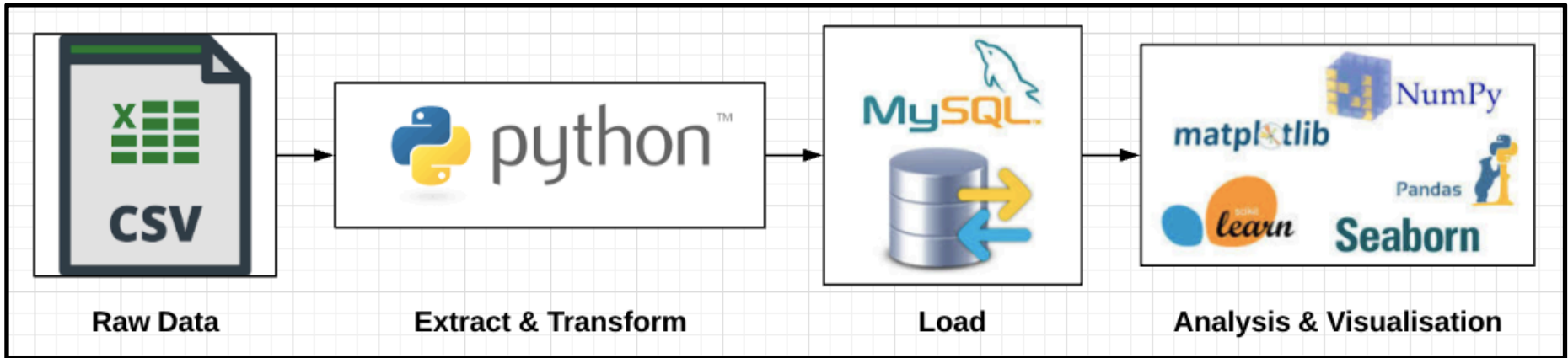
Problem Statement

- Considering total purchase amount, classify customers into 3 categories (High, Mid or Low)
- Classify employees into 3 bonus categories (3% Bonus, 2% Bonus, 1% Bonus) based on the total count & amount of sale provided
- Explore co-relations between employee, customers and products

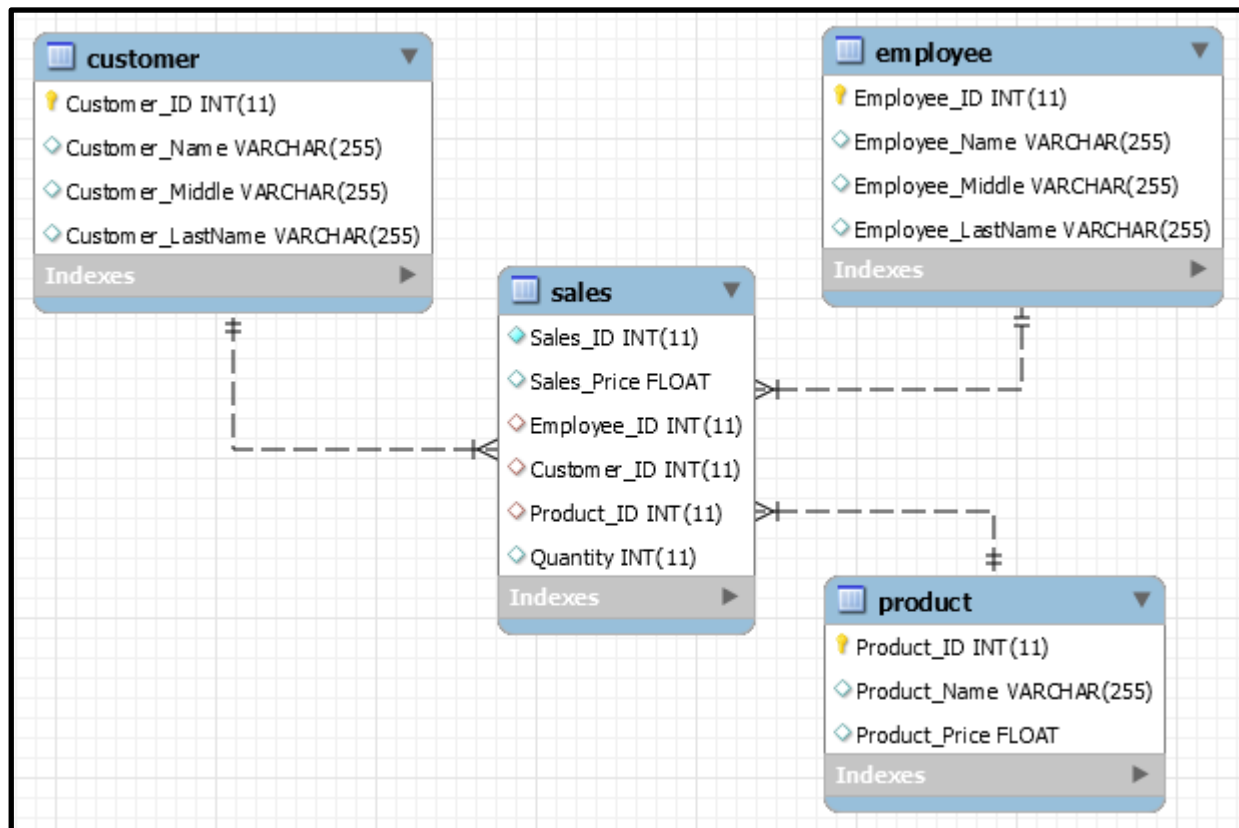
Data Profiling

| Column | Data Profiling Summary |
|-------------------|---|
| SalesID | 'SalesID' column is complete, valid and format consistent but not unique as it has duplicates. |
| Employee ID | 'Employee ID' column is consistent and valid but not complete due to missing values. |
| CustomerID | 'CustomerID' column is complete and consistent but whether its accurate and valid or not will be checked during data cleaning. |
| ProductID | 'ProductID' column is neither complete, as some of the fields are blank, nor accurate as some of them comprise of alphabets as well along with numbers. |
| Quantity | 'Quantity' column is incomplete, inaccurate and inconsistent as well. |
| Product Name | 'Product Name' column is incomplete, inaccurate and inconsistent as well. |
| Product Price | 'Product Price' column is incomplete and inconsistent. |
| Employee_Name | 'Employee_Name' column is incomplete, inconsistent and not unique as it has repetitions. |
| Employee Middle | 'Employee Middle' column is incomplete, inconsistent and invalid. |
| Employee Lastname | 'Employee Lastname' column is incomplete, inconsistent and invalid. |
| Customer Name | 'Customer Name' column is incomplete, inconsistent and invalid. |
| Customer Middle | 'Customer Middle' column is incomplete, inconsistent and invalid. |
| Customer LastName | 'Customer LastName' column is incomplete, inconsistent and invalid. |

ETL Pipeline



Star Schema



Demo

Thank You!