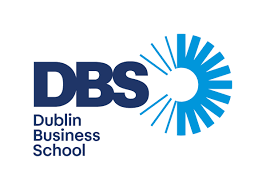
INDIVIDUAL CONTRIBUTION REPORT

NAme: Ajit Singh

Student Id: 10556972

ETL with AMAZON Orders Dataset

Neo4j database building with Ryan School Dataset



Contents

[Finding a Dataset 1](#_Toc325634774)

[Dimensional Modelling for Data warehousing 2](#_Toc325634776)

[Implementation of Source Relational Schema 3](#_Toc325634777)

[Implementaion of Data Warehouse Relational Schema 3](#_Toc325634778)

[ETL with Visual Studio(SSIS) 3](#_Toc325634779)

[Report Generation with Visual Studio(SSRS) 4](#_Toc325634780)

[Tableau 4](#_Toc325634780)

[Neo4j vs SQL 5](#_Toc325634780)

[Individual Contribution 6](#_Toc325634780)

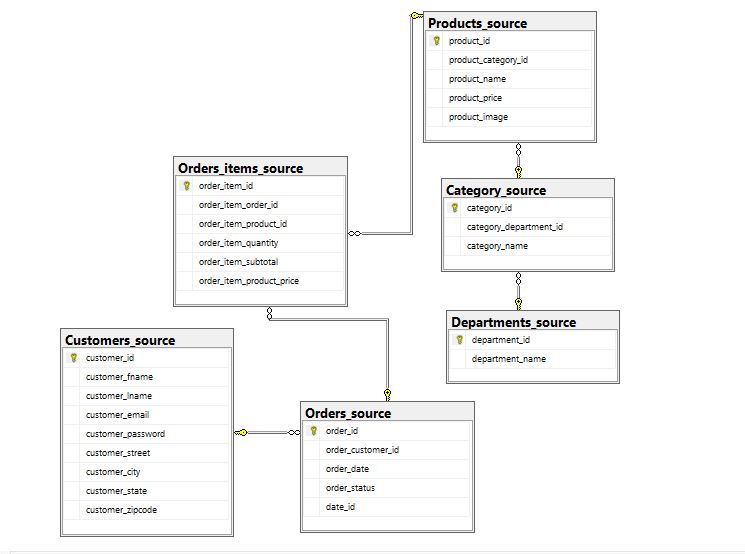
FINDING A DATASET

## Strategic Highlights

Keeping ETL process in mind having data extraction from different information sources and loading into Data warehouse tables, we needed sources with different attributes. And distinct keys to uniquely identify and relate the data for relational schema.

Amazon e-retail dataset consist of csv files, categories, departments, customers, orders, order items, products, represent different files as a separate source system and having a relational attributes which can be used to create Data warehouse schema.

Database Diagram of Source schema, diagram 1.1

diagram 1.1: Source Database diagram.

CLEANING THE AMAZON RETAIL DATASET

## Removing unwanTED characters and symbols:

Transformation performed in Excel after converting csv into an .xl file by eradicating null values, unidentified symbols (“, ‘’’’ etc.). Excel formulas like LEFT and REPLACE is used to replace symbols with blank space.

## Appropriate data conversion:

Converting INT values from STRING into INT and DATE TIMESTAMPS INTO appropriate date format using datatype formula or format option from the ribbon.

Dimensional Modelling for DATA WAREHOUSING

## Modelling objective:

With the dataset in place we can generate the Business insights like business operations in terms of order statuses by date and customer purchase interests in products or respective categories, monthly expenditure of customers and month wise comparison to forecast dip or higher expenditure for the coming months, Top buying product categories etc.

## DATA INCLUSION:

Including the appropriate data in the dimension modelling so that meaningful insights can be generated with the minimum data which is the whole motive of Data warehousing thereby decreasing query time there improving lowering processing time and improving memory utilization.

## Dimension tables and its Attributes:

Required attributes to build a data warehouse in order to meet our objectives can be divided into four dimension tables,

1. Product dimension: Product source table can be included in addition with Category and department table attributes as these data are relevant.
2. Customer dimension: Customer source table can be included with complete details.
3. Calendar dimension: Order date is utmost important to generate reports in between dates and month year and weekday part can be extracted from it to generate reports.
4. Order dimension: Order status should be included in the dimension table as it helps business to gain insights of their operations etc.

## FACT TABLE AND ITS ATTRIUTES:

It is the central table in a star schema of a data warehouse. It is used to store quantitative information of product price, total quantity of product per customers,

1. All the primary key of dimension tables will be included in this Fact table
2. Calculation details of product price and sum of product quantities is stored as attributes.

IMPLEMENTATION OF SOURCE RELATIONAL SCHEMA

## Creating TABLE:

Individual table has been created for each csv file with primary key references to other table foreign keys.

## Importing Data:

Used Import/Export wizard to import the data into source tables. Integrated visual studio import wizard was not working, i had to download external wizard to load the data into tables.

Data conversion issue experienced even after changing data type in Excel files. Int values could not be accepted in the source table with primary keys. Primary keys for the source table has been modified to

Decimal which will be converted later on while loading into Data warehouse tables.

Data should be loaded into source tables in such a way that tables with foreign key reference should be loaded after the referenced table is loaded, else error prompted. So import of data should be done sequentially.

IMPLEMENTATION OF DATAWAREHOUSE RELATIONAL SCHEMA

## CREATING TABLES:

Data warehouse has been built with a separate database as AMAZON ORDERS DW.

Tables are created as per the Dimension model with the fact referencing all the primary keys of dimension tables

Customer\_key, Product\_key, Order\_key, calendar\_key are create as a composite key for a non-duplicate data insertion in the fact table. See diagram 1.2 for reference,

ETL with Visual Studio(SSIS)

## Control Flow Task:

## Separate Flow task has been created for each dimension and fact table. Comprising of Source and Destination ADO Data Flow Task

## Data FLOW Task:

Data Flow Task consist of ADO Net Source and ADO NET Destination, helps to perform ETL process,

Extraction and Transformation:

Source is edited with Build query and all the attributes will be added in it. Multiple tables will be joined to get the desired attributes as per the dimension model.

Load:

Flow is connected with the ADO Net destination and linked with the source attributes then loaded into the destination.

Lookup:

Lookups have been used with Fact table to acquire surrogate keys from other dimension tables

Sort and Derived column:

Sort and derived column function is used with the Calendar ETL as Week, month, year, days part needs to be extracted.

Report Generation with Visual Studio(SSRS)

Four reports were generated for gaining customer insights. Either drag and drop or attribute selection can be used to generate following reports,

1. Total expenditure with customer details and category linked with it
2. Department wise Order status for each customer
3. Category wise quantity and price subtotal for each customer
4. Date wise order statuses for each Customer

## Challenges faced:

Report preview issue due to ‘memory exception’ error by while using SSRS. This issue has been rectified limiting the selection of data up to 500 rows.

Tableau

## Visuals generation:

1. Order Price Subtotal by Customers using Horizontal Bar visual representation.
2. Category wise most popular purchase using Pie Chart visual representation.
3. Orders quantity with their statuses by Product using Vertical Bar visual representation.
4. Department wise total expenditure by Customers using Circle view representation.

Neo4J vs SQL

## Graph Database Neo4j:

Nodes, relations and properties are used to store and represent data. Separate nodes have been created for each entity using ‘Ryan School’ dataset.

Relations are built using adjoining keys among the dataset. No schema creation and data type taken into account. Queries and Insertion of data are done with no SQL commands called Cypher.

Properties and relations are represented in a visualization form, gives a better visualization of data storage.

## Relational Database SQL:

Schema structure is taken into account. Format of the tables are fixed and should be defined prior to actually storing it. Data type and primary, foreign key selection should be done forehand.

Multiple queries are used within a query to join multiple table which makes it more complex to use.

Individual Contribution

* Finding a Dataset for the Data warehousing including cleaning and modifying the data type.
* Creating Star schema for the Dimensions modelling and designing the fact table keeping Business objective in mind.
* Creating Source Database schema with the source tables referencing uniquely identifying keys of each table.
* Building Data warehouse tables as per the planned dimension model.
* Using Visual studio SSIS for the ETL process using lookup and derived column function for the respective tables in the Data warehouse.
* Finding Dataset for Neo4j and sql database creation to find useful information out of it.
* Building Neo4j using Cypher Query language including creating labels and loading data in the Graph database, defining unique constraint for each label and making relations with uniquely identifying keys to query the data using Ryan School Dataset.
* Building Relation database using sql including creation of tables, inserting the data and querying the tables to get the desired result.