

**DATA WAREHOUSING AND BIG DATA**

**CS7079**

**GROUP COURSEWORK**

**GROUP MEMBERS:**

**OYINKANSOLA AKINOLA –22012591**

**OLISA UNEGBU-22020843**

**TAYYAB JABBAR- 22071836**

**MAMOONA NAJEEB- 22029169**

**MONIKA.- 22041350**



**LONDON  
METROPOLITAN  
UNIVERSITY**

## **Abstract**

*The aim of this coursework is to offer Blue Sky Online, an online retailer of consumer electronics, with the conceptual design of a dimensional data model based on a business case scenario. Blue Sky Online faces challenges in managing its extensive customer base and making well-informed decisions on pricing, customer targeting, and product offerings. A data warehouse may help by combining customer data and sales transactions for better analytics, which can help address these problems. The case study introduction opens the report, which is then followed by a review of related topics in the literature. Next, an overview of the available data sources is done. To fully understand the organisation and usefulness of different sources of data within the framework of the business case, an in-depth review is conducted.*

*Analysing and creating the dimensional data model is the initial step. The granularity, measures, hierarchies, and their reasons are taken into consideration while building the central fact table, which reflects consumer sale transactions. Complementary dimensions are discovered, and attributes and metrics are generated for each dimension, including customer details and transaction time. The report uses a simple star schema to visually depict the connections between the dimensions and the central fact table.*

## **INTRODUCTION**

Businesses encounter decision-making issues especially in the dynamic and highly competitive market of consumer electronics retail. Pricing strategies, attracting high-value users, assuring customer satisfaction and optimising product offers are among the issues. The ability to harness data and use it for educated decision-making is important in such an environment.

Blue Sky Online is an online electronics retailer with a large client base in the UK and Europe. It operates through a variety of channels, including its own website, Amazon, eBay, Tesco, and others. However, the corporation is interested in monitoring and evaluating client interactions to strengthen its business connections with customers, promote customer retention, and drive sales growth.

The objective of this report is to design dimensional data models and a data warehouse. The data warehouse's objective is to provide reports and do analysis to help stakeholders make better decisions.

## **LITERATURE REVIEW**

- **Grain:** In a data warehouse context, the grain determines the amount of detail or granularity with which data is kept in a fact table. It reflects the information included in each row of the fact table. The grain used is important since it influences the extent and depth of the analysis that may be conducted.
- **Central Fact Table:** A central fact table is an essential part of a data warehouse system. It holds quantitative data and is linked to other dimensions using foreign keys.
- **Hierarchies:** In a dimensional data model, hierarchies allow data to be organised and aggregated at various degrees of granularity. Hierarchies are required for digging down or rolling up data for analysis. For example, a time hierarchy can aggregate data from daily to

monthly or annual levels. Hierarchies are employed to enable flexibility in studying data at different levels, allowing users to get insights from both comprehensive and summarised data.

- **Dimensions:** Dimensions provide context for the measures in a fact table. They enrich our central fact table and allow for in-depth analysis and reporting.
- **Attributes:** Attributes are specific characteristics or properties of dimensions and measures. Attributes define the details within dimensions that are used to slice and dice data for analysis. They play a vital role in creating significant reports and insights.
- **Star schema:** Star schema is a sort of dimensional data model in which a central fact table is linked to several dimension tables via foreign keys. The simple star schema is a simplistic and efficient architecture for data warehouses. This schema also facilitates data analysis and reporting by providing a structured framework for understanding and analysing the data.

## DATA SOURCES

The data sources for this data warehousing project consist of six primary datasets, each offering a distinct perspective on Blue Sky Online's operations. These datasets are:

1. **Customer Details 1:** This dataset contains essential customer information, including names, customer codes, birthdates, marital statuses, gender, postcodes, income and cities. It serves as a foundational component for understanding the customer base.
2. **Customer Details 2:** This dataset provides additional customer information like first name, last name, and income and will be joined to the customer details 1.
3. **Customer Sale Transactions:** This dataset contains customer transaction data, such as transaction dates, customer codes, invoice numbers, total costs, total retail prices, payment types, and selling channels. It is important for tracking and analysing customer purchasing behaviour.
4. **Generated Datetime:** The generated datetime dataset has date-related information, including full dates, date names, day of the week, day names of the week, day of the month, and various other temporal attributes. It aids in time-based analysis and reporting.
5. **Payments Data:** This dataset contains information related to payment types and their corresponding codes which may enable categorization and analysis of payment methods.
6. **Selling Channels:** This dataset details selling channels used by Blue Sky Online, including their names, codes, and commission rates, providing insights into the revenue streams.

## ANALYSIS & DESIGN OF DIMENSIONAL DATA MODEL

### 1.1 Identification and definition of the grain of a central fact table and hierarchies

The central fact table captures the main business activity of Blue Sky Online which is Sales of Consumer Electronics. To effectively design this fact table, it is vital to define its grain and consider the presence of hierarchies.

The grain of a central fact table establishes the level of detail at which we record business activities. This fundamental decision profoundly influences the data model's size, performance, and relevance to

our analysis requirements. For Blue Sky Online, identifying the appropriate grain is pivotal in representing customer sale transactions accurately.

A thorough analysis of the "Customer Sale Transactions information" indicates that the grain should accurately reflect individual product transactions. Each record in the core fact database is associated with a specific product purchase transaction, as shown by the unique invoice numbers that are included. The most important details about every transaction, including the date of the transaction, the customer code, the total cost, the total retail price, the mode of payment, and the selling channel, should be recorded.

For data to be organised at different levels of detail within dimensions, hierarchies are required. Dimensions like time and customers may show hierarchies in our dimensional data model, allowing for more flexible analysis. The existence of hierarchies facilitates the provision of data at various granularities of detail to meet various analytical needs.

- Time Dimension Hierarchies: The time dimension will be equipped with hierarchies that include year, quarter, month, and day. These hierarchies enable analysts to drill down or roll up in timebased analysis, accommodating a wide range of time-related questions.
- Customer Dimension Hierarchies: For the customer dimension, hierarchies will be established based on customer demographics. Hierarchies will include age groups, marital status, and gender, allowing for segmented analysis based on these attributes.

These hierarchies were selected in accordance with the analytical specifications listed in the business case scenario. These hierarchies are made to give an all-encompassing view of the customer sales transactions and to enable the company to explore data at different levels of detail as needed.

Since it includes the finest degree of data pertinent to the business activities, the central fact table was designed with the granularity of individual product transactions. This level of detail makes it possible to fully record every product transaction, providing important data for analysis. In order to enable successful data investigation and evaluation, dimension hierarchies are constructed in accordance with the analytical requirements specified in the business case scenario.

## **1.2. Identification of Dimensions for the Central Fact Table**

This section identifies and describes the dimensions that will make up the central fact table. These factors are crucial for giving the data context and detail, which is in line with the reporting and analysis specifications mentioned in the business case scenario.

### **1.2.1. Customer Dimension**

The customer dimension is a foundation in our data model. It provides comprehensive information about Blue Sky Online's customers. This dimension contains attributes such as customer names, codes, birthdates, marital statuses, genders, postcodes, and cities.

The customer dimension is pivotal in understanding and segmenting the customer base. It will enhance the analysis of customer behaviour, preferences, and demographics. Such insights are integral for enhancing customer satisfaction and targeting high value consumers.

### 1.2.2. Time Dimension

The time dimension offers a lot of attributes including full dates, date names, days of the week, day names of the week, day of the month, and much more.

The time dimension is important for time-based analysis. It helps in dissecting customer sale transactions across different time periods, providing insights into trends, seasonality, and performance variations. This dimension supports a range of time-related reporting requirements, as mentioned in the business case scenario. Based on the details of the time data which comprises of day, week, month, quarter, and year, we will be using the monthly details for further analysis of the coursework.

### 1.2.3. Payment Dimension

The payment dimension dataset defines various payment types along with their associated codes.

The payment dimension plays a very important role in categorizing and analysing payment methods used in customer transactions. It enables in the detection of customer payment preferences and offers a thorough view of payment-related decisions and trends.

### 1.2.4. Selling Channel Dimension

The selling channel dimension includes information about the names, codes, and commission rates of different sales channels. It is important for assessing the effectiveness of various sales channels and allows the evaluation of commission rates.

These dimensions are selected to correspond to the reporting and analysis requirements stated in the business case scenario. They give enough context and attributes to enhance the central fact table, converting it into a strong basis for data-driven decision-making.

### 1.2.5 Product Dimension

The Product dimension is added based on discretion and what the Blue Online wants to achieve on the analysis. It will be important for determining the product offerings.

## 1.3. Identification of Attributes and Measures

### 1.3.1. Customer Dimension Attributes

Customer Dimension Attributes	Data Type
Customer Name	String
Customer code	Varchar (Primary key)
Birth Date	Date
Marital Status	String
Gender	String
Postcode	Varchar
City	String
Income	Integer

### 1.3.2 Time Dimension Attributes

Time Dimension Attributes	Data Type
DateKey	Integer(Primary Key)
FullDate	Date
DateName	String
DayofWeek	Integer
DayNameofWeek	String
DayofMonth	Integer
DayofYear	Integer
WeekdayWeekend	String
WeekofYear	Integer
MonthName	String
MonthofYear	Integer
IsLastDayofMonth	String
CalendarQuarter	Integer
CalendarYear	Integer
CalendarYearMonth	String
CalendarYearQtr	String
FiscalMonthofYear	Integer
FiscalQuarter	Integer
FiscalYear	Integer
FiscalYearMonth	String
FiscalYearQtr	String

### 1.3.3. Payment Dimension Attributes

Payment Dimension Attributes	Data Type
Name	String
RetailerPaymentTypeID	Integer (Primary key)

### 1.3.4. Selling Channel Dimension Attributes

Selling Channel Dimension Attributes	Data Type
Name	String
Code	String (Primary key)
CommissionRate	Integer

### 1.3.5 Product Dimension Attributes

Product Attributes	Data Type
Productid	Integer (Primary Key)
ProductName	String
ProductType	String

### 1.3.6. Fact Table Measures

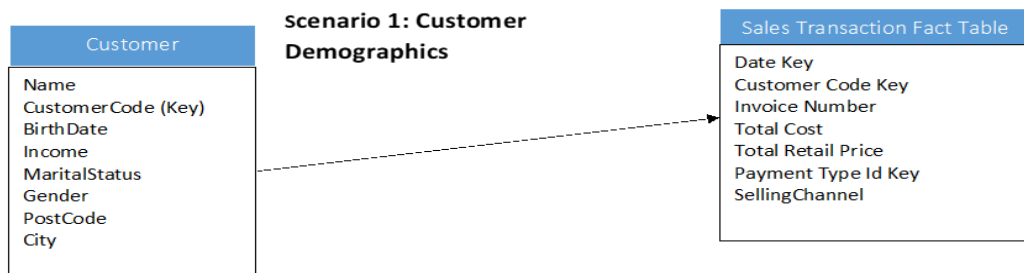
Fact Table Measures	Data Type
Date Key	Integer
CustomerCode	Varchar
Product ID	Integer
InvoiceNumber	Varchar
PaymentTypeID	Integer
SellingChannelCode	string
TotalCost	Float
TotalRetailPrice	Float

## 5.4 Simple Star Schema

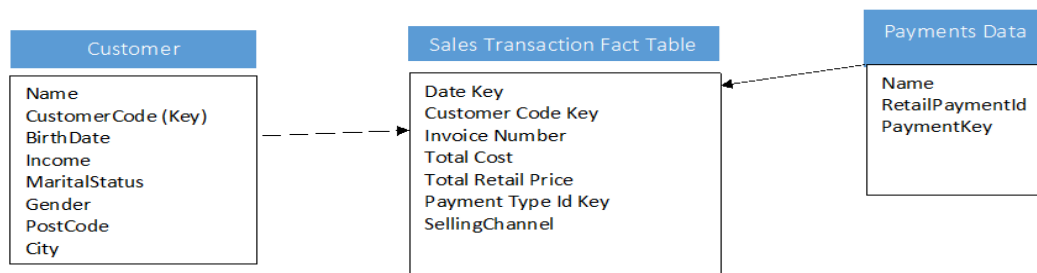
**Scenario 1: Customer Demographics**, to analyse customer behaviour, purchases made.

**Scenario 2: Sales Analysis**, it allows to get sales transaction with customer name, payment details, payment methods etc.

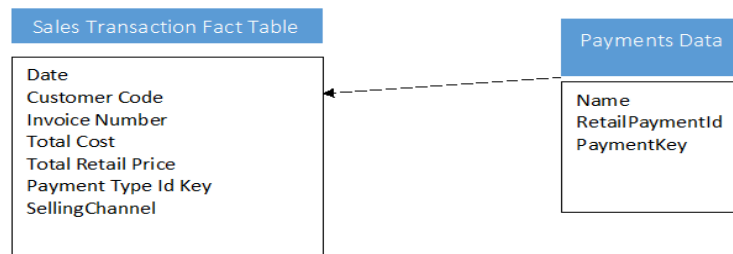
**Scenario 3: Payment Analysis**, used to get payment information, type of method used to perform transaction, maximum payment done using which payment method etc.



**Scenario 2: Sale Analysis**



**Scenario 3: Payment Analysis**



*Figure 1: Business scenarios*

Below is the final simple star schema that links the five-dimension tables to the fact table.



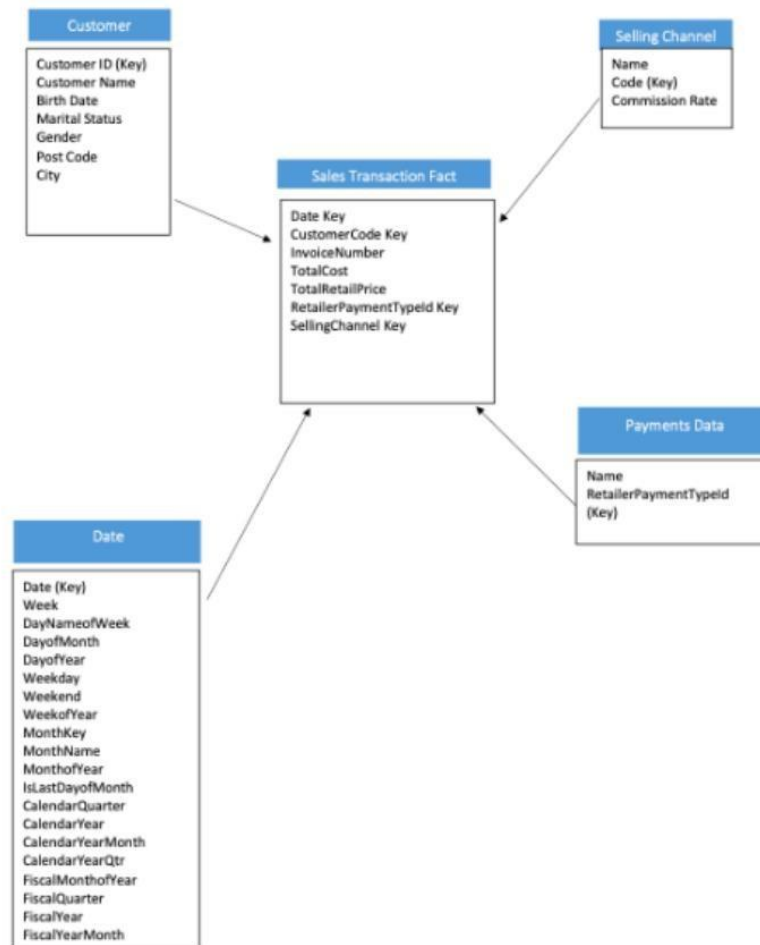


Figure 2: Simple Star Schema for Blue Sky Online

## Conclusion

Blue sky online which is an online customer retailer company needed a perfectly designed data warehouse which can address the challenging analysis like changing customer behaviour based on their demographic data, their preferences, and above all to identify the profitable customers on timely basis. So, to meet these requirements, we have successfully designed and implemented a data warehouse at conceptual level which perfectly fulfils all the reporting and analysis requirements of Blue Sky Online business which can give this company a competitive edge over others.

## PART B (Individual- Olisa Unegbu)

Employing the Simple Star Schema methodology to architect the business infrastructure of Blue Sky Online, an esteemed online consumer electronics retailer established in 1990, involved the creation of a database within the SQL Server Management Studio (SSMS). Within this framework, both the dimension and fact tables were meticulously established and populated. Presented herewith are select snapshots encapsulating the pertinent code segments instrumental in the formulation of these pivotal database components:

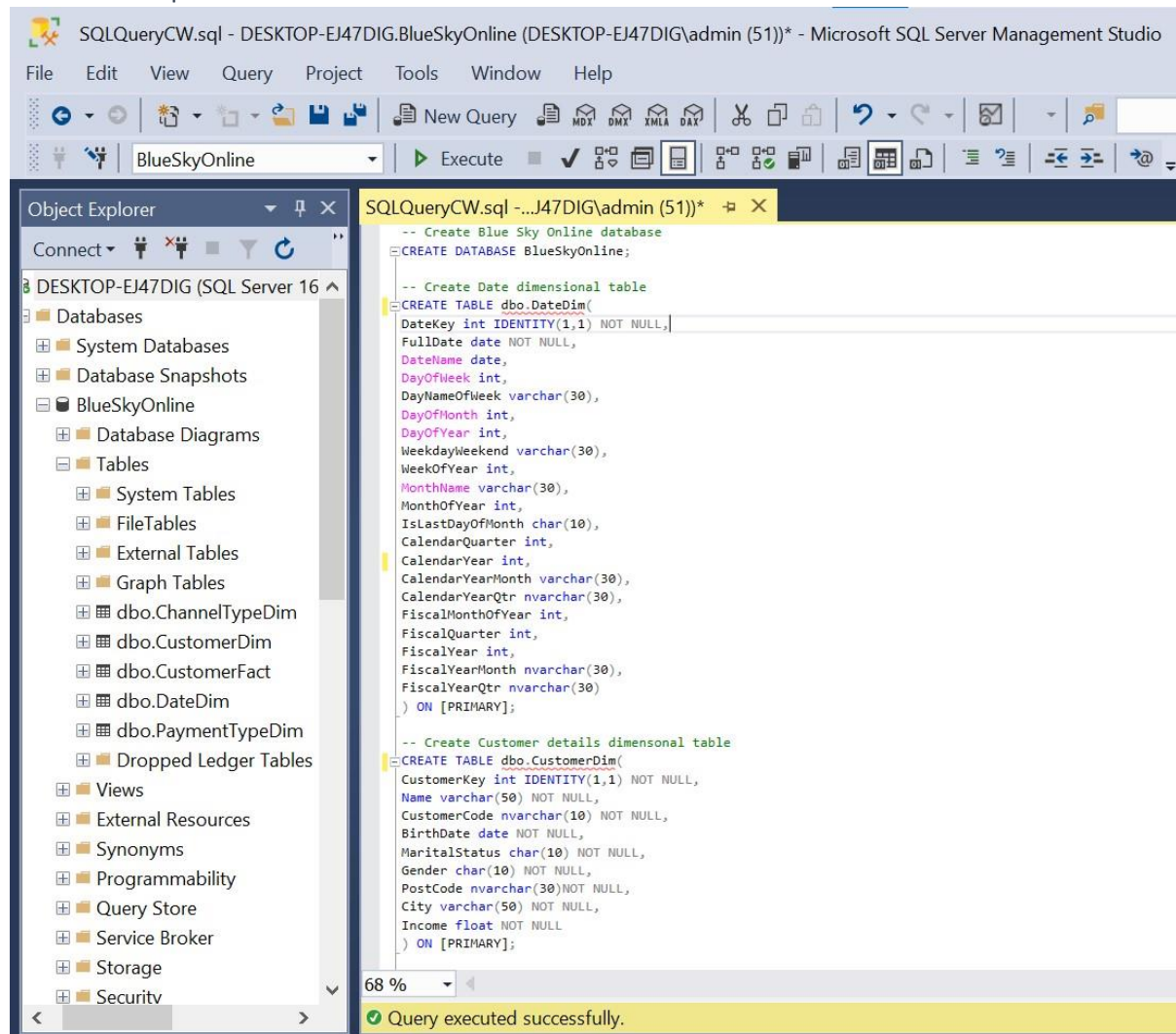


Figure 2.1: showing codes used to create the database and Date and Customer dimension tables.

In Figure 2.1, the database creation process is illustrated, initiated by the execution of the CREATE DATABASE command, establishing the BlueSkyOnline database within the SSMS server (DESKTOPEJ47DIG). Subsequently, the creation of dimension tables commences with the Date dimension table taking precedence. This table, comprised of twenty-one (21) attributes sourced from the provided Generated datetime data, designates the DateKey attribute as its identity key.

Following this, the Customer details dimensional table is generated. Preceding its creation, data manipulation and transformation were executed in Excel to reconcile two distinct sets of customer details: customer1 and customer2 details. Merging the FirstName and LastName attributes from customer2 into the Name attribute of customer1, the income attribute from customer2 was then

integrated into customer1, augmenting its attributes from 8 to 9. The resultant table, referred to as CustomerDim, was crafted using the CREATE TABLE command, wherein the CustomerKey served as the identity key.

Figure 2.2 proceeds to showcase the creation of payment type and selling channel dimension tables, alongside the customer fact table:

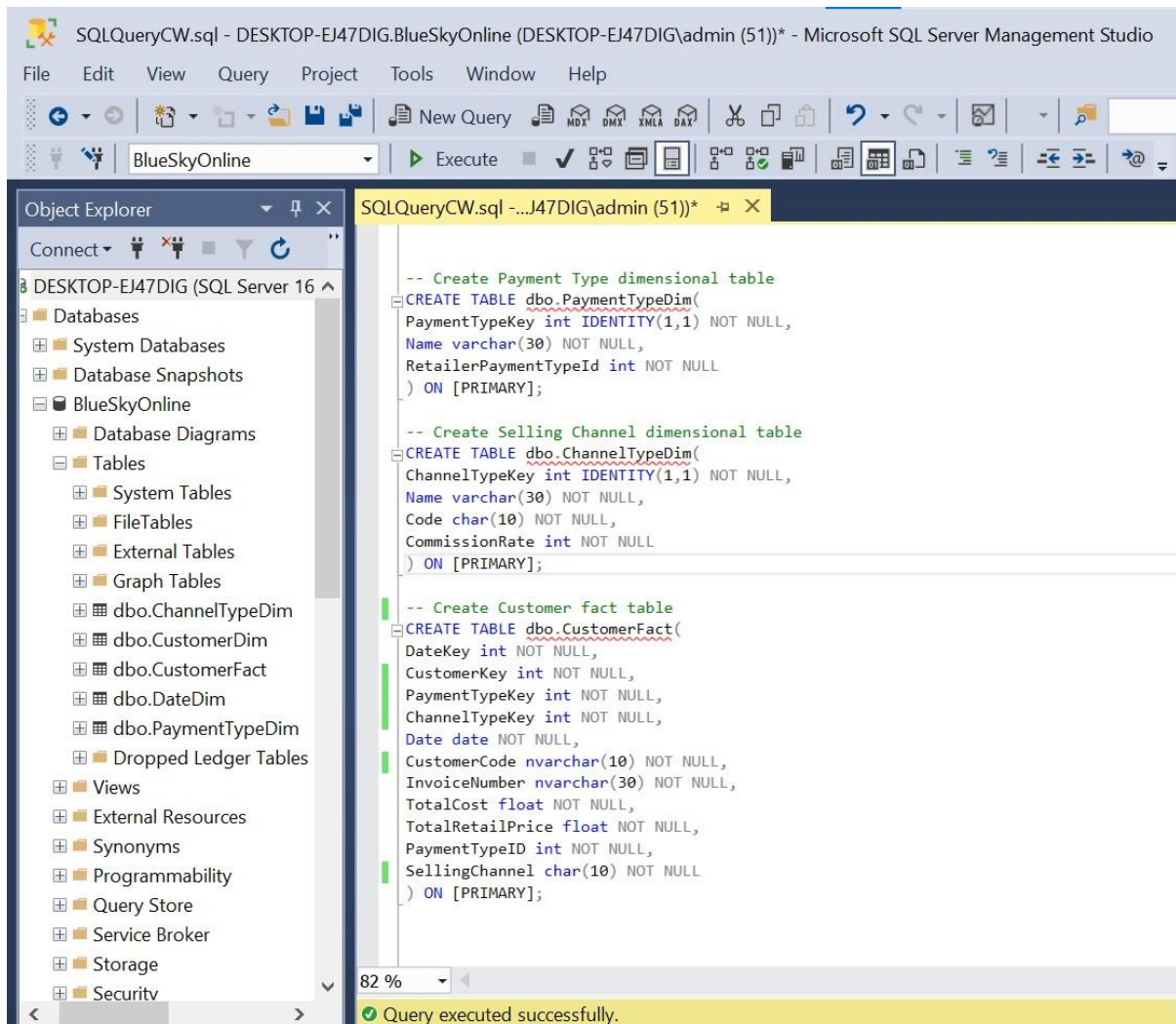


Figure 2.2: showing the SSMS environment codes for payment and selling channel tables and for customer fact table.

In Figure 2.2, the identity keys, PaymentTypeKey, and ChannelTypeKey, were employed in creating the payment type and selling channel dimension tables, serving as unique identifiers. Subsequently, the customer fact table was expressed by incorporating all identity keys from the four previously crafted dimension tables. These identity keys play a crucial role in establishing links between the dimension tables and the fact table, augmenting the attributes of the fact table from 7 to 11.

Following the creation of tables, the establishment of primary and foreign keys ensued. Alterations were made to the initially created tables, introducing primary keys, foreign keys, and constraints. These keys play a pivotal role in a well-designed data model, contributing to the maintenance of data quality, optimization of queries, assurance of data consistency and integrity, and the facilitation of relationships between tables in data warehousing.

Figure 2.3 explains the successfully executed codes for building the primary and foreign keys.

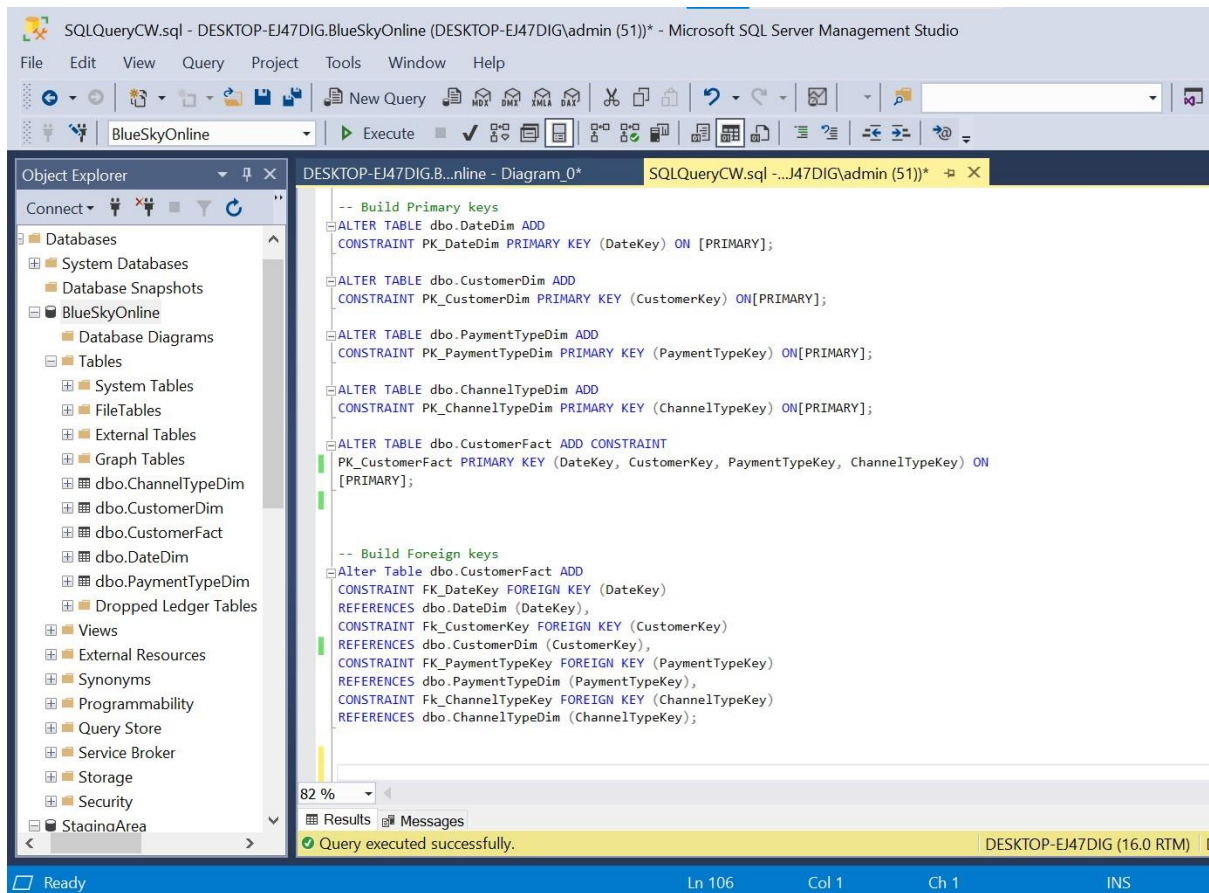


Figure 2.3: showing the primary and foreign keys codes.

Figure 2.4 below shows the final analysis and design of the Blue Sky Online data model.

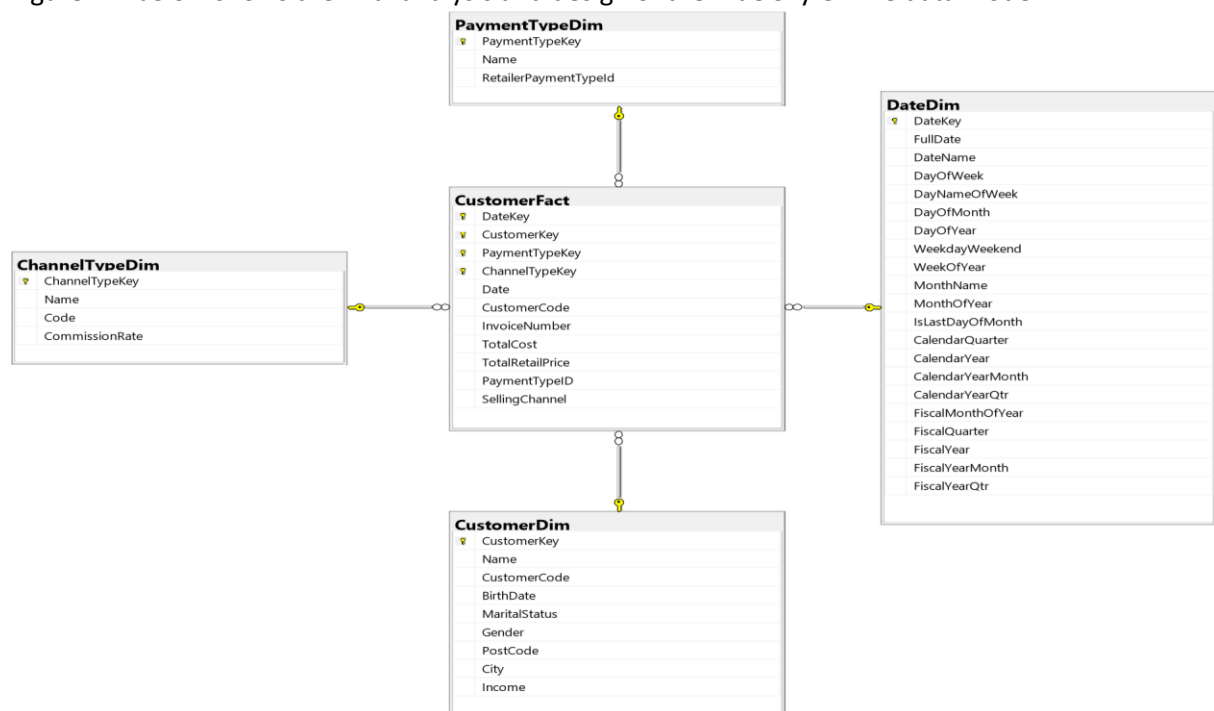


Figure 2.4: showing the database diagram for Blue Sky Online



Upon the creation of the dimension tables and the fact table, subsequent modifications were applied to incorporate essential elements such as primary keys, foreign keys, and constraints. Following this refinement, the SQL Server Import and Export Wizard, illustrated in Figure 3.1, was employed to populate all dimension tables effectively, ensuring the alignment of each attribute with its corresponding data types.

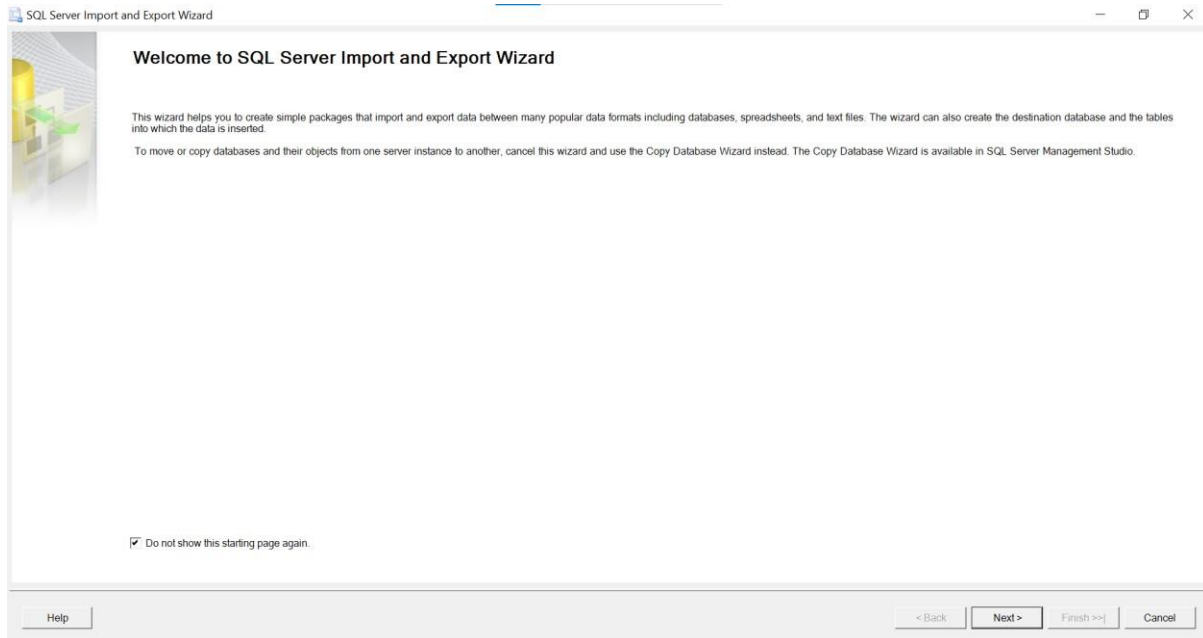


Figure 3.1: SQL server import and export wizard.

The SQL Server Import and Export Wizard facilitated the extraction of data from an external flat file source to populate the dimension tables stored in SSMS. Figures 3.1.1 and 3.1.2 serve as illustrative examples showcasing the outcomes of the populated Customer Details and Selling Channel dimension tables, respectively.

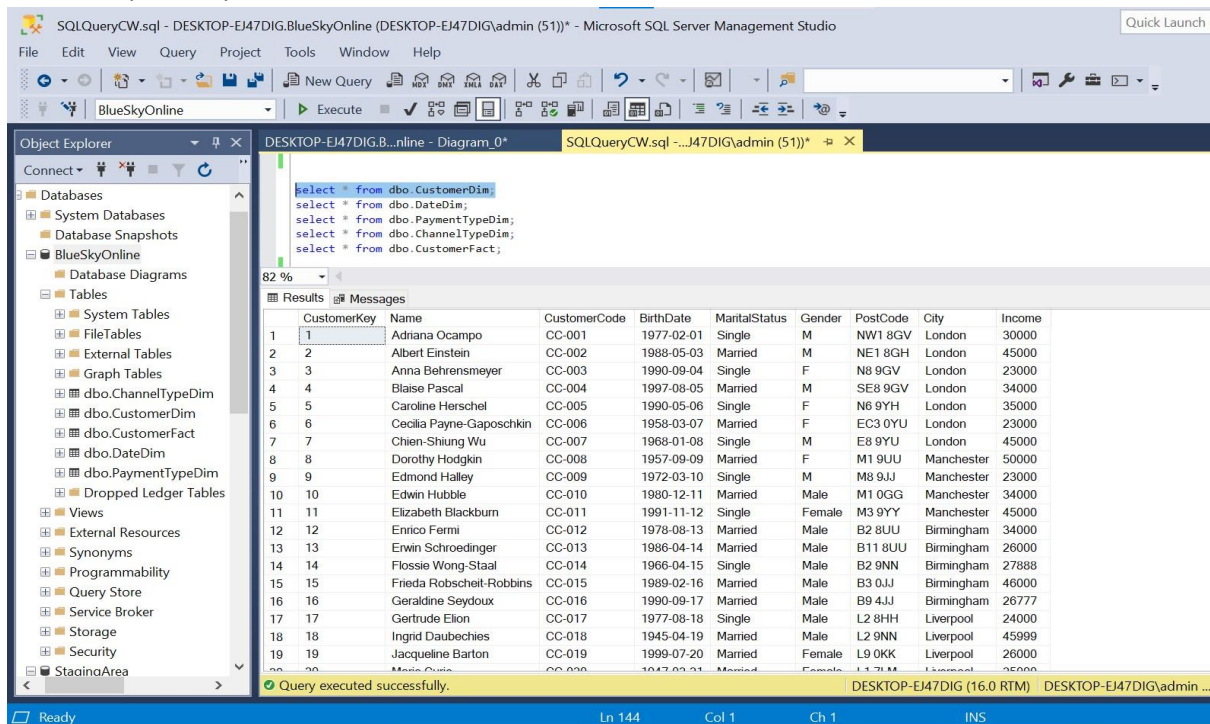


Figure 3.1.1: showing the result of the populated customer details dimension table.

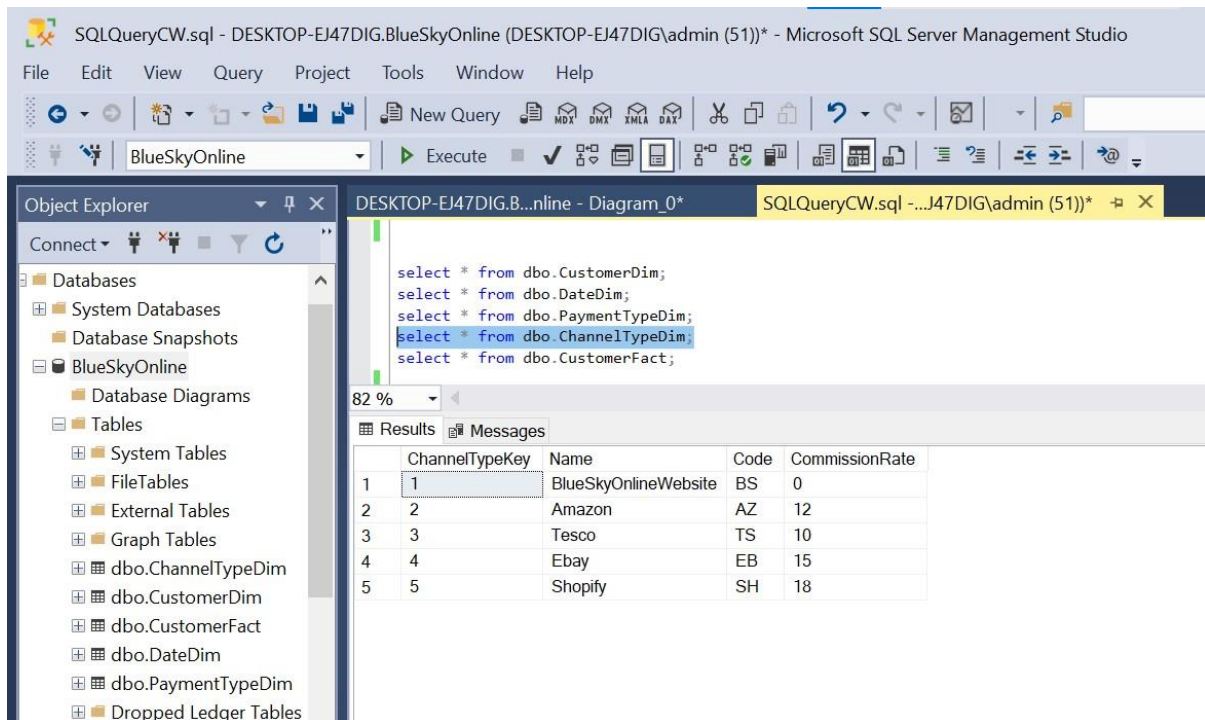


Figure 3.1.2: showing the code and result of the populated selling channel dimension table.

Figure 3.2 illustrates the creation of the StagingArea database along with the corresponding code. This database was established due to the complexity associated with populating the fact table in a manner like the dimension tables. Additionally, the code includes the creation of lookup tables and the customer sales transaction table, essential components for the staging area.

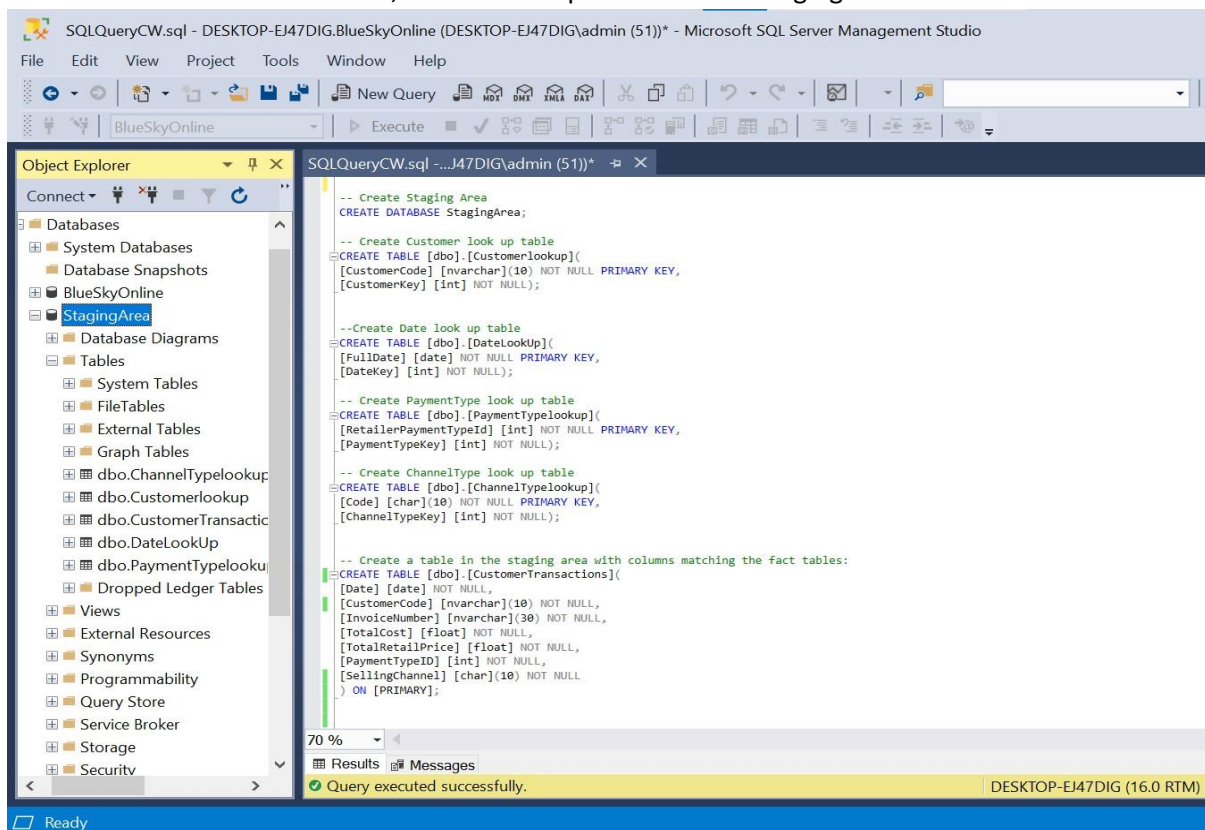


Figure 3.2: showing codes used to create the staging area database, lookup and transaction tables.

The staging area plays a crucial role in the Extract, Transform, and Load (ETL) process within a data warehouse. It serves as a pivotal component by ensuring data quality, facilitating data transformation, integrating data from diverse sources, and providing a controlled environment for managing data flow into the data warehouse. This controlled environment simplifies the process of populating the fact table.

In figure 3.2, the lookup tables are created with primary keys sourced from the dimension tables. These primary keys establish connections between the fact table and the identity keys of the dimension tables. Simultaneously, a sales transaction table is generated with attributes mirroring those of the fact table. The lookup tables are populated using the SQL Server Import and Export Wizard, like the process for the dimension tables, but this time the data is sourced directly from the BlueSkyOnline database within SSMS. The sales transaction table is populated in a manner akin to the dimension tables in the BlueSkyOnline database.

Figure 3.2.1 and 3.2.2 showcase the outcomes of the populated DateLookUp table in the staging area, serving as an illustrative example, and the results of the populated customer sale transactions, respectively.

The screenshot displays the SQL Server Enterprise Manager interface. The left pane shows the 'Object Explorer' with the 'StagingArea' database selected. The right pane shows the 'SQLQueryCW.sql' query window with the following SQL code:

```
select * from dbo.DateDim;
select * from dbo.PaymentTypeDim;
select * from dbo.ChannelTypeDim;
select * from dbo.CustomerFact;
select * from dbo.DateLookUp;
select * from dbo.CustomerTransactions;
```

The 'Results' pane shows the output of the query, displaying a table with two columns: 'FullDate' and 'DateKey'. The data is as follows:

	FullDate	DateKey
1	2006-01-01	1
2	2006-01-02	2
3	2006-01-03	3
4	2006-01-04	4
5	2006-01-05	5
6	2006-01-06	6
7	2006-01-07	7
8	2006-01-08	8
9	2006-01-09	9
10	2006-01-10	10
11	2006-01-11	11
12	2006-01-12	12
13	2006-01-13	13
14	2006-01-14	14
15	2006-01-15	15
16	2006-01-16	16
17	2006-01-17	17
18	2006-01-18	18
19	2006-01-19	19
20	2006-01-20	20
21	2006-01-21	21

The status bar at the bottom indicates 'Query executed successfully.'



Figure 3.2.1: showing the code and result of the populated datelookup in the staging area.

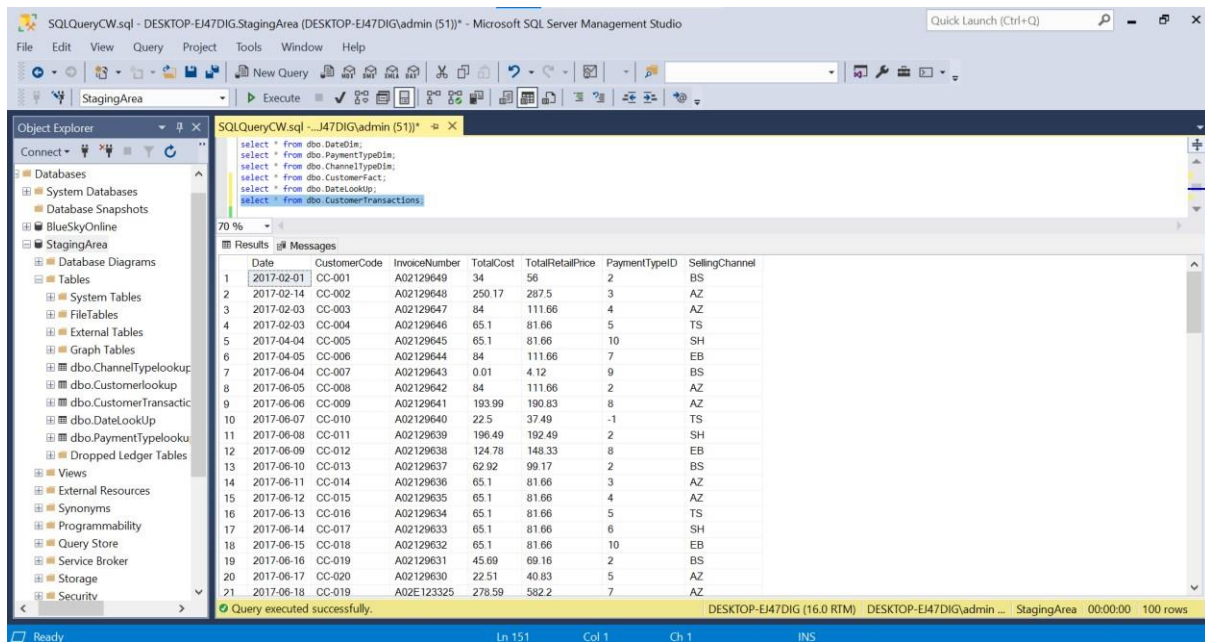


Figure 3.2.2: showing the code and result of the populated customer sales transaction in the staging area.

Following the successful population of the lookup tables and the customer sales transaction table in the staging area database, the fact table within the BlueSkyOnline database underwent population using the SQL Server Import and Export Wizard. In this process, a meticulously crafted query was formulated to establish links between the primary keys in the lookup tables and their corresponding attributes in the fact table, as depicted in figure 3.3 below:

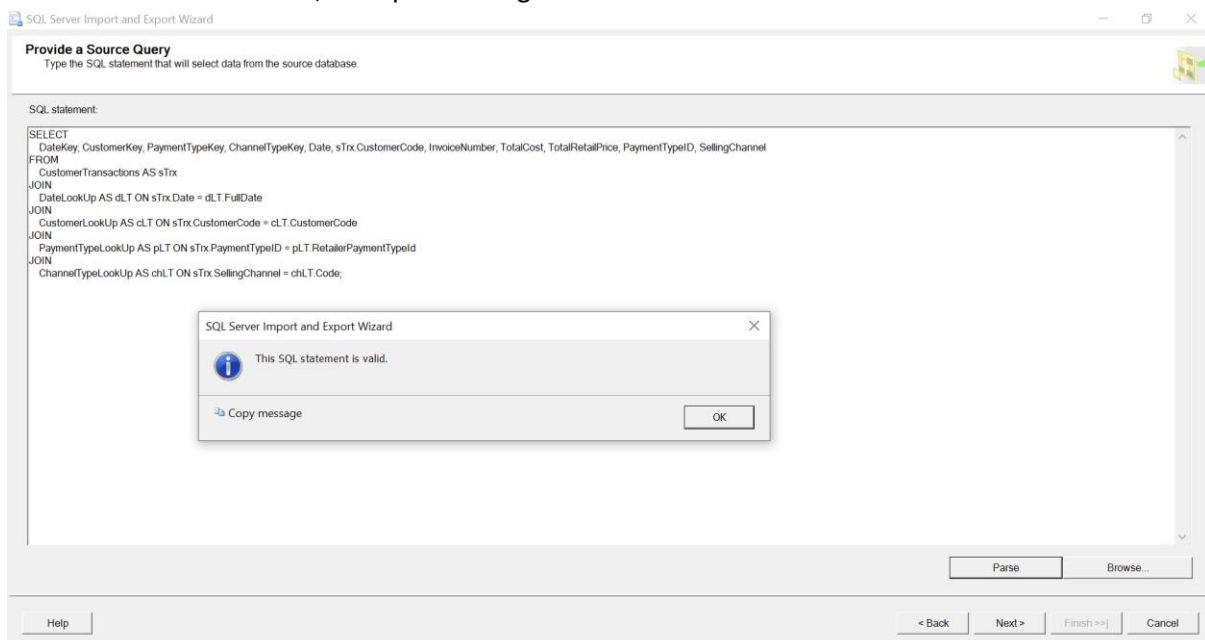


Figure 3.3: showing the codes used to populate the fact table in the BlueSkyOnline database.

The query as shown in figure 3.3 above when parsed showed the SQL statement is valid. This strategic linkage between the primary keys and attributes ensures a seamless and accurate transfer of data, aligning the dimension tables' information with the fact table's structure. The SQL Server Import and



Export Wizard plays a pivotal role in orchestrating this data flow, ensuring the successful integration of valuable information into the fact table. The query provides a more understandable view of customer transactions by replacing the keys in the CustomerTransaction table with descriptive information from associated lookup tables.

Figure 3.4 as shown below shows the code and result of the populated fact table on the BlueSkyOnline database:

SQLQueryCW.sql - J47DIGadmin (511)\*

```

select * from dbo.DateDim;
select * from dbo.PaymentTypeDim;
select * from dbo.ChannelTypeDim;
select * from dbo.CustomerFact;
select * from dbo.DateLookup;
select * from dbo.CustomerTransactions;

```

DateKey	CustomerKey	PaymentTypeKey	ChannelTypeKey	Date	CustomerCode	InvoiceNumber	TotalCost	TotalRetailPrice	PaymentTypeID	SellingChannel
4050	1	3	2	2017-02-01	CC-001	A02129649	34	56	2	BS
4052	3	5	1	2017-02-03	CC-003	A02129647	84	111.66	4	AZ
4052	4	6	3	2017-02-03	CC-004	A02129646	65.1	81.66	5	TS
4063	2	4	2	2017-02-14	CC-002	A02129648	250.17	287.5	3	AZ
4112	5	11	5	2017-04-04	CC-005	A02129645	65.1	81.66	10	SH
4113	6	8	4	2017-04-05	CC-006	A02129644	84	111.66	7	EB
4173	7	10	1	2017-06-04	CC-007	A02129643	0.01	4.12	9	BS
4174	8	3	2	2017-06-05	CC-008	A02129642	84	111.66	2	AZ
4175	9	9	2	2017-06-06	CC-009	A02129641	193.99	190.83	8	AZ
4176	10	1	3	2017-06-07	CC-010	A02129640	22.5	37.49	-1	TS
4177	11	3	5	2017-06-08	CC-011	A02129639	196.49	192.49	2	SH
4178	12	9	4	2017-06-09	CC-012	A02129638	124.78	148.33	8	EB
4179	13	3	1	2017-06-10	CC-013	A02129637	62.92	99.17	2	BS
4180	14	4	2	2017-06-11	CC-014	A02129636	65.1	81.66	3	AZ
4181	15	5	2	2017-06-12	CC-015	A02129635	65.1	81.66	4	AZ
4182	16	6	3	2017-06-13	CC-016	A02129634	65.1	81.66	5	TS
4183	17	7	5	2017-06-14	CC-017	A02129633	65.1	81.66	6	SH
4184	18	11	4	2017-06-15	CC-018	A02129632	65.1	81.66	10	EB
4185	19	3	1	2017-06-16	CC-019	A02129631	45.09	69.16	2	BS
4186	20	6	2	2017-06-17	CC-020	A02129630	22.51	40.83	5	AZ
4187	19	8	2	2017-06-18	CC-019	A02E123325	278.59	582.2	7	AZ

Figure 3.4: showing the code and the result of the populated fact table in the BlueSkyOnline database. Subsequently, the CustomerFact table underwent exportation to a flat file destination in the form of an external file through the SQL Server Import and Export Wizard. The resultant output was saved in CSV format, as illustrated in figure 4.1 below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	DateKey	CustomerKey	PaymentTypeKey	ChannelTypeKey	Date	CustomerCode	InvoiceNumber	TotalCost	TotalRetailPrice	PaymentTypeID	SellingChannel				
2	4050	1	3	2	01/02/2017	CC-001	A02129649	34	56	2	BS				
3	4052	3	5	1	03/02/2017	CC-003	A02129647	84	111.66	4	AZ				
4	4052	4	6	3	03/02/2017	CC-004	A02129646	65.1	81.66	5	TS				
5	4063	2	4	2	14/02/2017	CC-002	A02129648	250.17	287.5	3	AZ				
6	4112	5	11	5	04/04/2017	CC-005	A02129645	65.1	81.66	10	SH				
7	4113	6	8	4	05/04/2017	CC-006	A02129644	84	111.66	7	EB				
8	4173	7	10	1	04/06/2017	CC-007	A02129643	0.01	4.12	9	BS				
9	4174	8	3	2	05/06/2017	CC-008	A02129642	84	111.66	2	AZ				
10	4175	9	9	2	06/06/2017	CC-009	A02129641	193.99	190.83	8	AZ				
11	4176	10	1	3	07/06/2017	CC-010	A02129640	22.5	37.49	-1	TS				
12	4177	11	3	5	08/06/2017	CC-011	A02129639	196.49	192.49	2	SH				
13	4178	12	9	4	09/06/2017	CC-012	A02129638	124.78	148.33	8	EB				
14	4179	13	3	1	10/06/2017	CC-013	A02129637	62.92	99.17	2	BS				
15	4180	14	4	2	11/06/2017	CC-014	A02129636	65.1	81.66	3	AZ				
16	4181	15	5	2	12/06/2017	CC-015	A02129635	65.1	81.66	4	AZ				
17	4182	16	6	3	13/06/2017	CC-016	A02129634	65.1	81.66	5	TS				
18	4183	17	7	5	14/06/2017	CC-017	A02129633	65.1	81.66	6	SH				
19	4184	18	11	4	15/06/2017	CC-018	A02129632	65.1	81.66	10	EB				

Figure 4.1: showing the exported fact table.

This step marks the extraction of valuable data from the database, offering the flexibility to utilize the information in various applications such as the Apache HDFS (Hadoop Distributed File System) to conduct further analysis for the coursework requirements.

The customer details table and the customer fact table were successfully loaded under the Ambari environment in the '/user/maria\_dev/' path as shown in figure 4.2 below:

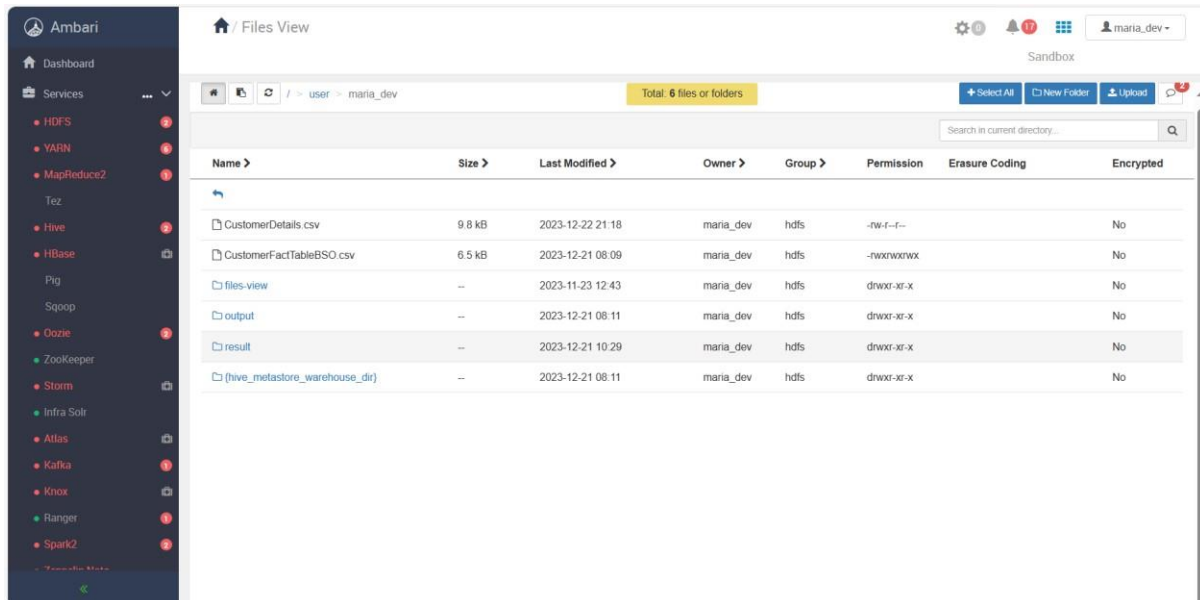


Figure 4.2: showing the customer details and customer fact table in the Ambari environment.

As shown in figure 4.3 below is a query used to create a suitable data structure for loading the CustomerFactTableBSO into HIVE.

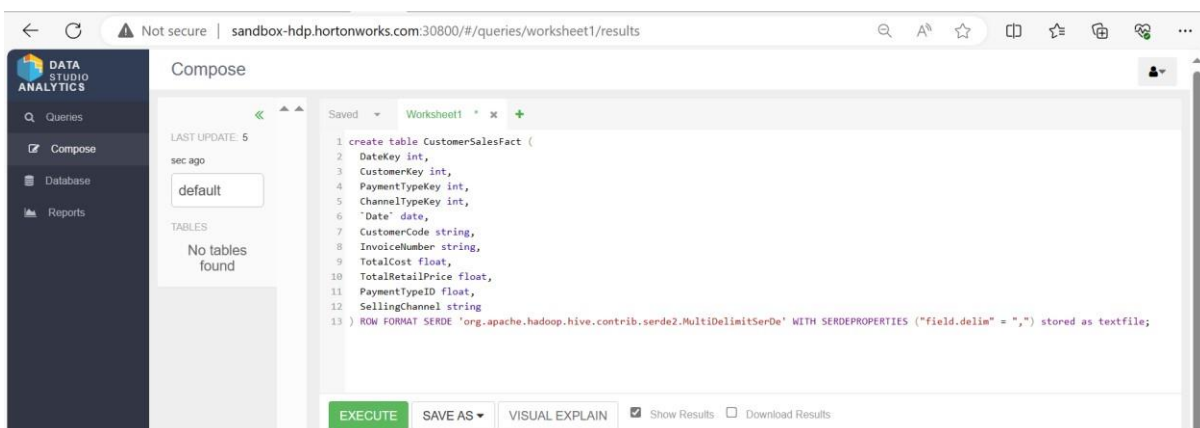


Figure 4.3: showing codes used to create suitable data structure for loading data file into HIVE.

The structure as shown in figure 4.3 above is suitable for loading a CSV-like data file into Hive. The column types match the expected data types, and the SerDe configuration ensures proper handling of the data format. The table is stored as a text file, which aligns with the assumption that the data is in a CSV-like format.

Figure 4.4.1 and 4.4.2 below show the result of the above query:

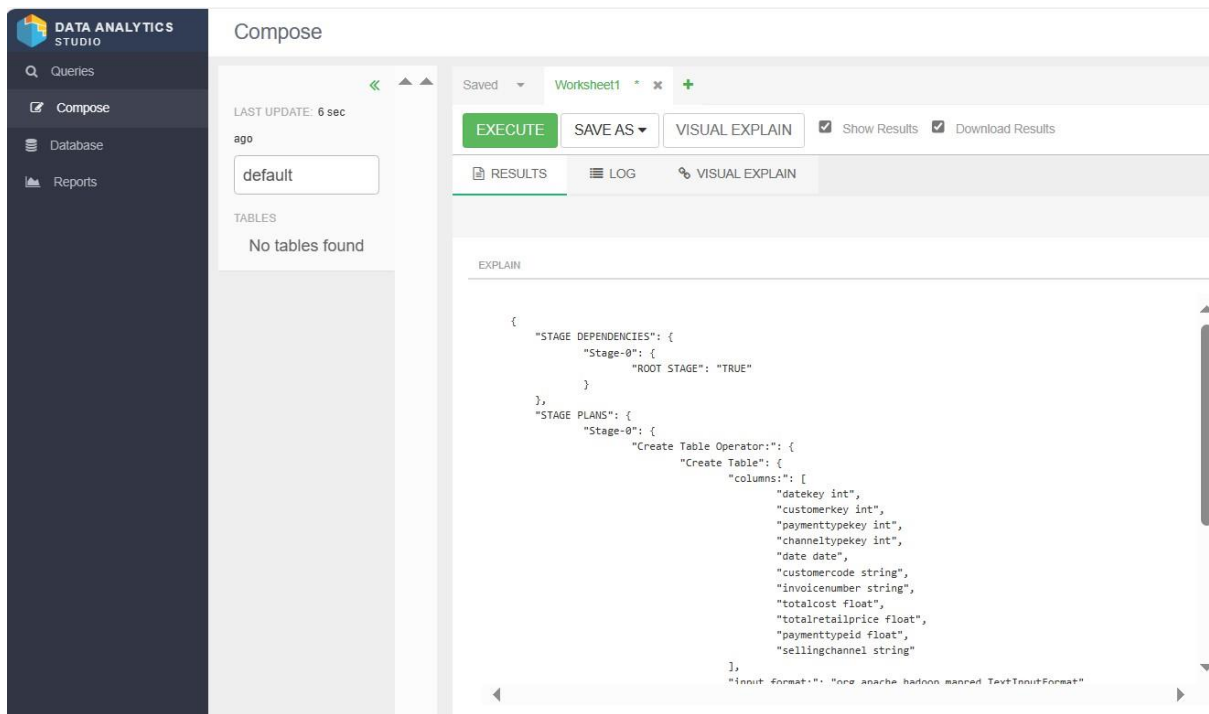


Figure 4.4.1: showing result of the data structure.

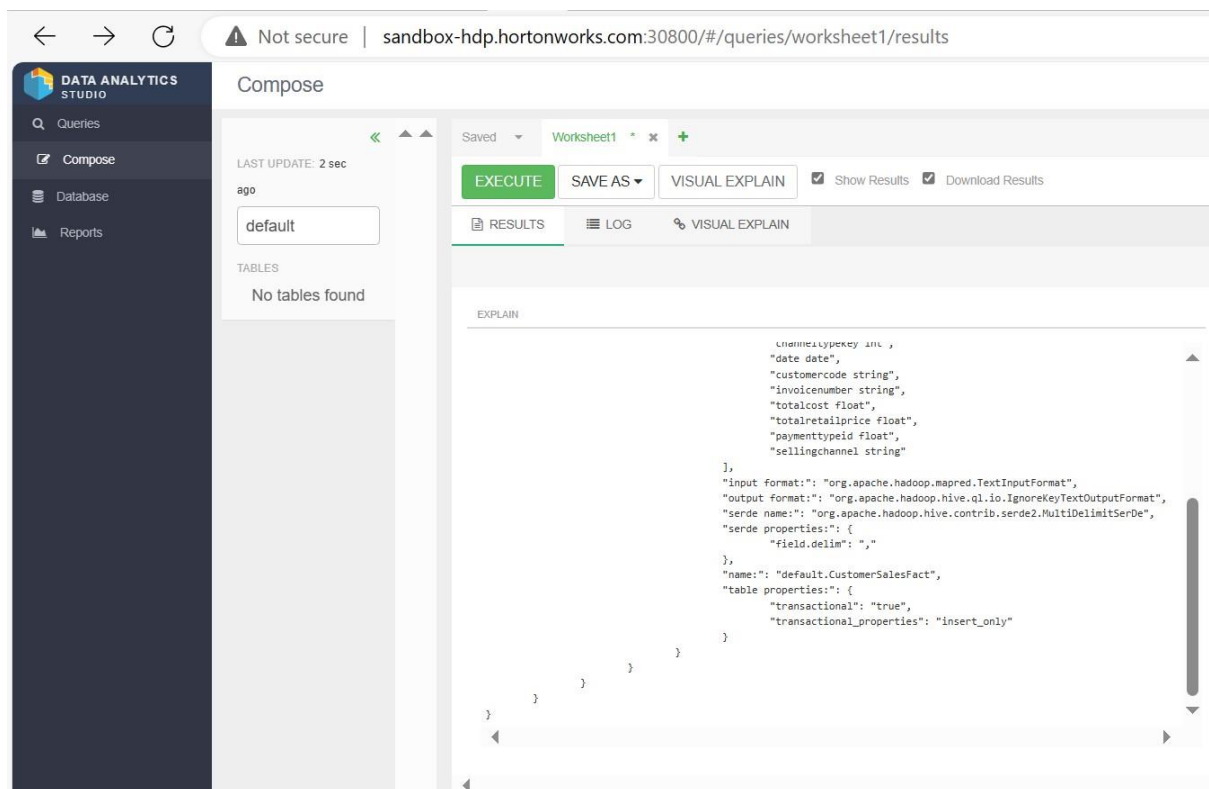
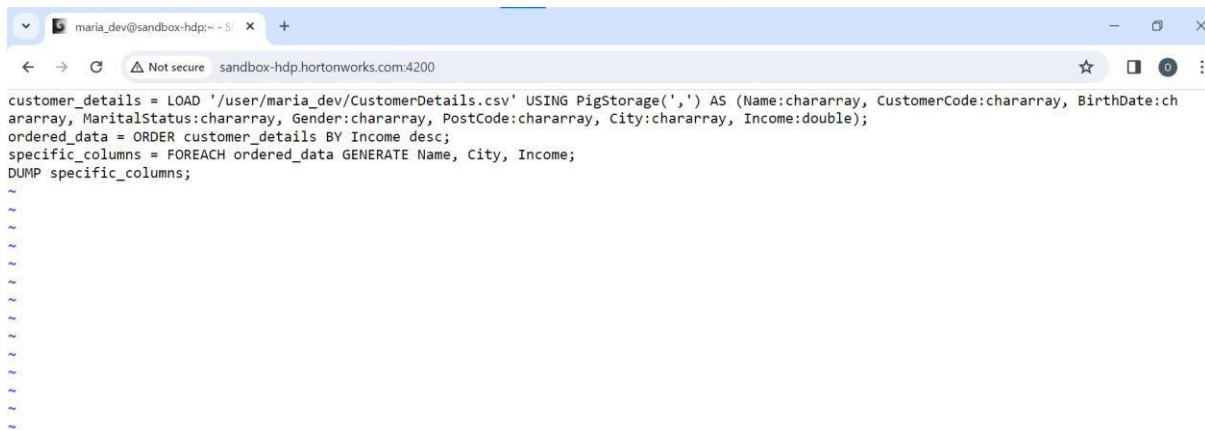


Figure 4.4.2: showing the continuation of the result of the data structure.

The Apache Pig environment was utilized to showcase the manipulation of customer details data, loaded into Ambari as illustrated in figure 4.5 below:

A screenshot of a web browser window with the address bar showing 'sandbox-hdp.hortonworks.com:4200'. The browser displays a Pig script in a text area. The script starts with a LOAD statement for 'CustomerDetails.csv', followed by an ORDER statement to sort by 'Income desc', and a DUMP statement to output the results. The script is as follows:

```
customer_details = LOAD '/user/maria_dev/CustomerDetails.csv' USING PigStorage(',') AS (Name:chararray, CustomerCode:chararray, BirthDate:chararray, MaritalStatus:chararray, Gender:chararray, PostCode:chararray, City:chararray, Income:double);
ordered_data = ORDER customer_details BY Income desc;
specific_columns = FOREACH ordered_data GENERATE Name, City, Income;
DUMP specific_columns;
```

Figure 4.5.1: showing the query for data manipulation on Apache Pig environment.

The script depicted in figure 4.5.1 represents a Pig script designed to process the data file named CustomerDetails through Apache Pig. The script, shown in figure 4.5.2, orchestrates the ordering of customer\_details based on the Income attribute in descending order. It subsequently filters three specific columns from the result set.

A screenshot of a web browser window showing the output of a Pig query. The output is a list of customer names, cities, and incomes, sorted in descending order of income. The first three results are: (Dorothy Hodgkin, Manchester, 50000.0), (Frieda Robscheit-Robbins, Birmingham, 46000.0), and (Ingrid Daubechies, Liverpool, 45999.0). The output is as follows:

```
2023-12-22 22:05:18,412 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-12-22 22:05:18,412 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Dorothy Hodgkin,Manchester,50000.0)
(Frieda Robscheit-Robbins,Birmingham,46000.0)
(Ingrid Daubechies,Liverpool,45999.0)
(Albert Einstein,London,45000.0)
(Chien-Shiung Wu,London,45000.0)
(Elizabeth Blackburn,Manchester,45000.0)
(Caroline Herschel,London,35000.0)
(Enrico Fermi,Birmingham,34000.0)
(Blaise Pascal,London,34000.0)
(Edwin Hubble,Manchester,34000.0)
(Adriana Ocampo ,London,30000.0)
(Flossie Wong-Staal,Birmingham,27888.0)
(Geraldine Seydoux,Birmingham,26777.0)
(Erwin Schroedinger,Birmingham,26000.0)
(Jacqueline Barton,Liverpool,26000.0)
(Marie Curie,Liverpool,25000.0)
(Gertrude Elion,Liverpool,24000.0)
(Anna Behrensmeyer,London,23000.0)
(Cecilia Payne-Gaposchkin,London,23000.0)
(Edmond Halley,Manchester,23000.0)
```

Figure 4.5.2: showing the result of the previous query on Apache Pig.

The outcome as illustrated in figure 4.5.2, reveals that Dorothy Hodgkin, residing in Manchester, holds the highest income at £50,000. Following closely is Frieda Robscheit-Robbins from Birmingham, boasting an annual income of £46,000. Ingrid Daubechies, residing in Liverpool, secures the third position with an income of £45,999.

While reviewing the assessment and curriculum for the MSc. Data Analytics program, I encountered Hadoop as one of the tools to be covered. As someone unfamiliar with Hadoop but eager to explore new technologies, this discovery influenced my decision to choose LondonMet over other universities that extended offers. Working within an agile environment and delving into the world of Hadoop for coursework has proven to be a rewarding experience. Although it poses challenges and demands significant time investment, especially when navigating platforms like Azure Lab, the learning journey is immensely valuable. Hadoop, being an integral component of data analytics, becomes more captivating as one grasps its foundational principles and essential aspects of analysis.

## REFERENCES

- IBM (2021-03-08) Data grain. Retrieved from <https://www.ibm.com/docs/en/ida/9.1.1?topic=phase-step-identify-grain>
- Simplilearn Fact Table vs. Dimension Table - Differences Between The Two. <https://www.simplilearn.com/fact-table-vs-dimension-table-article>
- Microsoft - Understand star schema and the importance for Power BI (Article 02/27/202) Retrieved from <https://learn.microsoft.com/en-us/power-bi/guidance/star-schema>.
- DBT Labs. (2023, 10 28). *Data Grain*. Retrieved from <https://docs.getdbt.com/terms/grain> Microsoft. (2023, 10 29).
- *Hierarchical Data*. Retrieved from <https://learn.microsoft.com/en-us/sql/relationaldatabases/hierarchical-data-sql-server?view=sql-server-ver16>
- Tech Target. (2023, 10 28). *Fact Table*. Retrieved from <https://www.techtarget.com/searchdatamanagement/definition/fact-table>