

Title: **Big Data model and warehouse for Pestron Foods**

Module: **Big Data (ACUA6002)**

Student Id: **1815743**

Lecturer: **Terry Walcott**

Date: **03.09.2021**

Word count: **4558** *(excluding headers, figures, tables, references, and appendices)*

Contents

Abstract.....	2
Introduction	2
Literature Review	3
Methods.....	6
Results.....	8
Conclusions	18
References	19
APPENDIX A: Big Data Tools advantages and disadvantages	20

List of Figures

Figure 1: KNIME PestronFoods workflow group	8
Figure 2: Sales data set workflow	9
Figure 3: CSV Reader node configuration	9
Figure 4: Column Filter node configuration	9
Figure 5: String Manipulation node for extract year configuration	10
Figure 6: String Manipulation (extract Day) node execution output	10
Figure 7: Column Filter (remove Date) node execution output	10
Figure 8: Graphical representation of sales based on product line	11
Figure 9: Graphical representation of sales based on customer type	11
Figure 10: Stores data set workflow	12
Figure 11: Graphical representation of stores based on the city	12
Figure 12: Predictive model workflow	13
Figure 13: Column name error	13
Figure 14: Columns renaming output	13
Figure 15: Graphical representation of sales based on product line	14
Figure 16: Linear Regression Learner node configuration	14
Figure 17: Linear Regression Learner - coefficients and statistics output	15
Figure 18: Predictive model applied - predicted total based on product line	15
Figure 19: Predictive model applied - predicted total based on customer type	15
Figure 20: CSV Writer node configuration	16
Figure 21: Snowflake Data Warehouse	16
Figure 22: Snowflake database	16
Figure 23: Snowflake JDBC driver configuration	17
Figure 24: Database Connector node configuration	17
Figure 25: Snowflake PESTRONFOODSWH warehouse workload	18
Figure 26: Snowflake PESTRON FOODS DB database interrogation	18

List of Tables

Table 1: Big Data tools	6
Table 2: Pestron Foods sales data set variables	7
Table 3: Pestron Foods stores data set variables	8

Abstract

In the last years, the concept of Big Data was a buzzword, barely understood by the organizations. The studies showed that it was poorly explored especially by the small businesses. It was a myth that only large amount of data can be used and only the big companies have the resources to implement it. However, today more large, medium, or small businesses have an increasing emphasis for Big Data. This report explores the usage of Big Data especially in small businesses. The report presents several history events that are considered the foundation of Big Data, a set of real-life examples where Big Data brought significant advantages, and based on a comprehensive literature review, the main concepts, and models of it. The report also presents the results of implementing, using KNIME Analytics Platform and Snowflake tools, a predictive model and a data warehouse that provides deeper insights on the existing data sets for Pestron Foods business and a vision for creating future business strategies.

The overall aim of this report is to understand the significant advantages that small companies have by using Big Data and how they can mitigate the challenges and choose the right tools.

Keywords: small businesses, Big Data, advantages, challenges

Introduction

In recent years, there has been an increasing interest in using data science for analysing and decision-making especially in large organizations. A considerable number of studies has been published on the benefits that big businesses had by implementing Big Data despite the industry they were part of. According to Marr (2016) Walmart, which has more than 20000 stores worldwide and it is considered one of the largest retailers, used Big Data to determine their customers' preferences and provide them the products they want to buy. Understanding the customers' behaviour wasn't their only scope. By collecting and processing around 2.5 petabytes of information every hour, Walmart was able to decide the best price of the products which increased the sales, simplify the logistical operations, monitor the performance, and create a strategy for the customer services. Walmart had the resources to create their internal Big Data team. It is obviously that Big Data is critical for Walmart as it is for another large organizations as Amazon, Netflix, BBC, or the US Government [1].

However, the world economy is not supported only by large organizations. A report by World Trade Organization (2016) presented that only 10% of the business are large enterprises, the difference of 90% represents small and medium size enterprises [2]. This statistic shows how important are small businesses in the economy, but it also brings a concern on how the enterprises are categorized in small or large. Over the last decades, the definition of small business has been a controversial and disputed subject. The U.S. Securities and Exchange Commission (1981) defined the small businesses based on the income. Ten years later, Ang (1991) offered a more specific definition, where a small business is managed by a sole proprietor rather than a management team. Drinan (1995) stated that the small businesses should be defined based on the industry and size. Now, the size of the organization is an important criterion. U.K. Government (2020) estimated the number of small businesses, based on the size, less than 50 employees, to almost six million. In the same statistic, the number of large businesses, where the number of employees is greater than 250, was 7835. There is a considerable difference between the number of the small and large businesses [3].

In this economical context, how Big Data is used by small businesses is an important subject and this report will focus on the usage of Big Data in small organizations. "Literature review" chapter will present a comprehensive literature review on Big Data definitions, its advantages, and challenges. "Methods" chapter will analyse the Pestron Foods business requirements and data and the needed

steps to successfully accomplish the implementation of a Big Data solution in their organization. Pestron Foods is a mini supermarket, that do electronic products import and export to several markets from Europe and Americas. Pestron Foods is a small business, and their aim is to identify opportunities and adopt strategies for increasing sales and customers loyalty and to optimize the import-export process by understanding their customers preferences and establish the correct price. "Results" chapter will show the output of implementing a Big Data solution that satisfy the company business requirements.

The overall aim of this report is to discuss the benefits of Big Data usage in small businesses and to find an implementation solution for Pestron Foods to mitigate the challenges and accomplish their business target.

Literature Review

This chapter will focus on Big Data definitions and the advantages and challenges small businesses have by implementing Big Data solutions.

In recent years, there has been an increasing focus on Big Data as the companies, despite their size, are constantly looking for growing, bringing value to their businesses, and predicting the future. However, the concept is still not easily understood by everybody, and the confusion is caused by the Big Data definition itself. Even if the term was a buzzword in 2012, previous studies reported a huge history behind the concept of Big Data. According to Southerland, data analysis was to predict mortality rate by John Graunt in 1662 [4]. At that moment, the Big Data term wasn't known, but the data itself was a critical element for doing analysis. In 1865, Richard Millar Devens described how a banker found a solution to collect and analyse the information and get advantage by acting accordingly before his competitors. This process was described by using "business intelligence" term. The evidence suggests that the same period introduced an issue that has grown in importance, and it represents still a challenge now: how fast can data be collected and analysis. This problem was addressed by Herman Hollerith, a businessman, statistician, and inventor, which invented in 1881 the Hollerith Tabulating Machine used by the US Census Bureau to reduce the collection and processing of a large amount of census data from 10 years of work to only three months [5]. Herman started the Tabulating Machine Company, which became later the base for what is now known as IBM company [6]. In 19th century, the engineer Fritz Pfleumer discovered a method to store the information on magnetic tape. So far, those events, drove to the definition of Big Data in the way that it is understood and used now but in 1989, Erik Larson stated "the keepers of big data say they do it for the customer's benefit. But data have a way of being used for purposes other than originally intended". His statement shown the real challenges that collecting, and processing data have. These challenges related to privacy, security and ethics are addressed now by every entity that uses Big Data. However, Big Data was defined from multiple perspectives. IBM (2012) defined Big Data as a massive collection of data collected from different data sources, while other authors were focused on the storage and analysis requirements of using big data [7].

Overall, in most of the literature, the Big Data is defined by its three main components: data, the analytics of data and the interpretation of the data analytics which brings business value. Hurwitz et. al. [8] defines big data in terms of the three Vs: volume, variety, and velocity. Volume refers to the amount of data, velocity measure how fast the data is processed, and variety indicates the types of data. However, this definition can diminish the importance of Big Data, as other aspects as complexity, accuracy, and the context in which Big Data analysis is applied must be considered.

Oracle defines the volume as the amount of data that matters [9]. The volume measures the size of data sets that the organization needs to process and analyse. Over the last year, the volume

significantly increased being larger than terabytes or even petabytes. Another interesting aspect is that the data volume is different based on industry. Marr (2018) published an article about the way that Instagram use Big Data. He mentioned that Instagram has 800 million active users per month and the generated data, which in this example includes mainly photos and videos, is as large as the company chief executive officer declared “We are also going to be a big data company”. However, small companies are not generated such amount of data. There is no need of having terabytes of generated data to achieve the same benefits as the big companies do.

Another important characteristic of Big Data is variety. Hurwitz et. al. (2013) [8] simply defines data variety as the different types of data. He even accentuates its importance as considering it “the spice of life” and classifies it in *structured* and *unstructured* data. Structured data refers to data that follow a format and has a defined length. Structured data is simple, organized, and quantitative and contains searchable data (numbers, date, arrays of characters known as strings) and depends on a data model. It is usually stored in Excel files, Google Sheets, or relational database management systems (RDBMS). In retail industry, structured data is represented by inventory, shop details (name, address, phone number), product hierarchies, shipping data, financial data etc. Unstructured data can be difficult to organize as it doesn’t follow a defined pattern and it has a higher level of complexity. Unstructured data is represented by image, video, log files etc. In retail industry, it can be product image meta data, customer reviews, marketing emails etc [10].

So far, the report describes the Big Data through the amount and types of data. But these two characteristics wouldn’t be enough for bringing value to the businesses without considering the time measure. Velocity measures the speed with which the data, structured and unstructured, is generated. Statistics shown that every minute, Twitter generates 98000 tweets, Facebook has 695000 status updates, Google has 698445 searches and Gmail generates unstructured data from 168 million emails sent. Velocity has a large impact on businesses. However, velocity doesn’t necessary mean productivity, but it is crucial for business decision-making.

It is a fact that large organizations have considerable benefits from using Big Data, but the lack of studies and literature for the small businesses is a concern. An IBM report shown that only 23% of small and medium businesses use big Data. However, in the second part, this chapter will focus on the importance of Big Data for small businesses.

Twiddy & Company is a family-owned small business that manage 998 homes for renting in North Caroline. Their aim is to increase the customers satisfaction and properties owners’ profit. As it is a small company, the data is stored in spreadsheets which makes is difficult to analyse. Their marketing director identified the limitations of making good decisions caused by the incapability of properly analyse their operational data. Twiddy decided to implement a Big Data solution by using a SAS (Statistical Analytical System) which organize the company’s data into a customizable format. This format was shared with the properties’ owners and contractors, and they were able to decide the price of the rent based on market conditions, season, or property characteristics (location, size, number of rooms, etc.) and to make customized recommendations to their customers. Twiddy declared also that the company cut the cost of maintenance by 15% and eliminate invoice processing errors. The benefits brought by the Big Data in such a small company, was for Twiddy as “tasting ice cream for the first time”.

Twiddy & Company is not the only small businesses that took advantage of processing and analysing their data.

Carvana, an online marketplace for cars, founded in 2013 in Phoenix, saw the opportunities in Big Data. The company has only 50 employees, and their aim is to improve customers’ experience, but they also want a better way to predict the cars sales. According to their co-founder they want “to determine which cars up for auction didn’t meet our quality standards”. Carvana decided to work with

a data-expert team which analyse hundreds of data across several databases. The analysis was against the credit reports provided by multiple companies and improve the decision of interest rates for different customers and it also helped to identify the suspect buyers.

Overall, according to McKinsey Global Institute (2015) [11], there are several ways that Big Data can bring value to small businesses:

- It can offer a better understanding of the information provided by the data analysis. The information become transparent, and it can be frequently used in decision-making process. Also, the companies can use it in experiments to simulate different situations and make better predictions.
- It can be used to develop future generation of products and services. If a product is not preferred by the customers, or it doesn't have highly desired features, then it can be improved in a future version.
- It can be used to identify potential markets and to create a segmentation of customers to mitigate the risk and improve the marketing strategy.

Big Data offers to small businesses a huge growth potential, but it also comes with challenges. Large organizations have all the resources, including financial resources, to adopt Big Data, but small companies can meet limitations as lack of knowledge, speed of data growth, confusion generated by the Big Data tools selection, data integration or security.

One of the most common issue is represented by the lack of professional knowledge of what is Big Data and how it can be efficiently used. The employers might not know about the importance of data storage. The data is not stored into databases, there is no backup strategy, and when the data is needed, it can't be easily accessed. Small companies can hire data scientists, analysts, or engineers, but this solution results into an increase cost of recruitment. A better solution is to invest in their existing staff by providing training programs and purchase knowledge analytics solutions, known as Big Data tools. Table 1 shows some of the Big Data tools, classified based on their role (See Appendix A):

Role	Tool
Analytics Frameworks	Apache Hadoop
	Apache Spark
Data Storage	Big Data Sqoop
Data Visualization	IBM Watson Analytics
	D3.js
	Microsoft Power BI
Processing	Google BigQuery
	Amazon Redshift
Testing	Hive
	HBase
Governance	Apache Atlas

Table 1: Big Data tools

The selection of Big Data tools represents a challenge for most of the small companies, where there is no professional data expertise. Instead of losing money and time trying to identify the appropriate technology, the company can hire a data consultant which will be able, based on experience, to recommend the tools that will support the business requirements and scenario. A

professional selection will also resolve the integration issue as data is collecting from different sources and the aggregation is crucial for analysis and business intelligence. Another important aspect is represented by the data security and especially by the vulnerability to fake data generation. Based on IBM report (2009), in the UK retail industry, more than 80% of the companies have inconsistent data which impacts the operations and profit. Hazen et. al. (2014) [12] stated that the insights are correct and concluded only if they are obtain based on correct data. For mitigate this challenge, the companies should use tools for data quality control.

This chapter discussed about the definition of Big Data and the historical events that underlines the concept. Several case studies were also discussed to prevent the advantages and challenges that small companies have by implementing Big Data solutions.

In the next chapter, the report will focus on the steps required to successfully implement a Big Data solution that satisfy the Pestron Foods organization requirements. The data set provided by the company will be analysed. It will introduce literature research on several Big Data algorithms and models.

Methods

This chapter will focus on data analysis methods. The Pestron Foods data set will be described. The tools and algorithms used for implementing a Big Data solution will be decided based on the following literature research.

Extracting value and producing a rigorous analysis of diverse and large data sets is a complex task and requires the analyst to follow principles and scientific methods. In the organizational context, the following methods can be applied:

- Classification

Suthaharan (2012) defines classification as a technique used to classify data based on format, type of applied analysis, processing technique, data source and data storage. Unstructured data is classified into structured classes. Classification technique involves two phases. First phase is the learning process where the training data sets are provided and analysed, and the patterns are created. In the second phase, the data sets are evaluated and tested, and more accurate classification patterns are created. Classification technique is either supervised or unsupervised. In supervised classification, a possible class is known in advanced and Decision Tree (DT) or Support Vector Machine (SVM) algorithm is commonly used [13].

- Regression

Hurwitz et. al. (2013) [8] defines regression as a technique used to estimate the relationship between two or more variables. It is based on a dependent variable (a factor that needs to be understand or predict) and an independent variable (a factor that impact the dependent variable). For example, the sales of a product can depend on weather. One of the most used algorithms is linear regression.

- Similarity matching

Similarity matching is a technique used to identify entities (products, services, customers etc.) based on common characteristics. This technique can bring a huge benefit for targeting marketing campaigns [14].

The process of implementing a Big Data solution for Pestron Foods organization includes data collection, validation, and analysis.

The first data set provided by Pestron Foods organization is stored in csv file and contains structured data related to their sales and 17 variables with the following meaning:

Variable	Type	Description
InvoiceID	Number (Integer)	Number of the recoded invoice in the company system for a specific sale
Branch	String	Name of the branch that process the sale. The data set contains data about three branches: A, B and C
City	String	City where the branch is located. The data set contains data about three cities: Mandalay, Naypyitaw and Yangon
Customertype	String	Type of the customer. This variable indicates if the customer is a member of the store or not
Gender	String	Customer's gender with two possible values: Female and Male
Productline	String	Product category
Unitprice	Number (Double)	Price for a single unit of measure of the product
Quantity	Integer	Purchased product's quantity
Tax5%	Number (Double)	The amount that represents a five percentage from the product price as tax
Total	Number (Double)	Total amount paid for the purchased products (including tax5%).
Date	Date	Date when the invoice was generated
Time	String	Time when the invoice was generated
Payment	String	Method of payment. This variable indicates if the customer pays using cash, credit cards or E-wallet
cogs	Number (Double)	Cost of goods sold calculated as deducting the tax5% from the total amount
grossmarginpercentage	Number (Double)	The amount of profit made before deducting the sales
grossincome	Number (Double)	Total income before deductions
Rating	Number (Double)	A number that indicates the customer satisfaction

Table 2: Pestron Foods sales data set variables

The second data set provided by Pestron Foods organization is stored in csv file and contains structured data related to their stored existing in Columbus city and nine variables with the following meaning:

Variable	Type	Description
city	String	City where the store is located
columbus_area	String	Geographic area where the store is located within the city
ebt	String	Electronic Benefits Transfer used to accept payments
name	String	Store's name
phone_number	String	Store's phone number
state	String	City's state
street_address	String	Store's address

wic	String	Special Supplemental Nutrition Program also used to accept payments
zip_code	Number (Integer)	Store's zip code

Table 3: Pestron Foods stores data set variables

The next chapter will show the data cleaning as the provided data set can contain irrelevant information, data validation, visualization and analysis steps results using KNIME Analytics Platform which is a free and open-source software for creating data science. KNIME is intuitive and offers features and reusable components accessible for everyone and it doesn't require knowledge of coding. As mentioned in "Introduction" chapter, one of the company's aims is to increase sales and customer loyalty for successfully achieve it, a linear regression model will be created.

Results

This chapter will present the output of initial visualization of the data sets provided by Pestron Foods organization, the output of implementing a linear regression model and a creating a data warehouse.

A workflow group is created for Pestron Foods, and it contains two workflows: Sales workflow and Stores workflow.

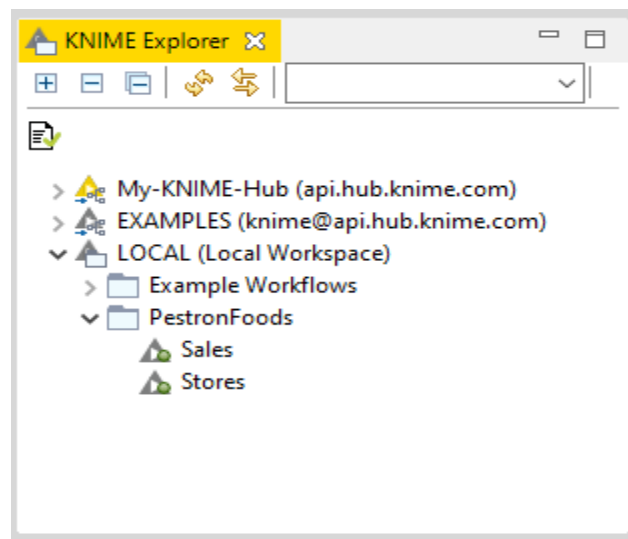


Figure 1: KNIME PestronFoods workflow group

Sales workflow

Sales data set is prepared and visualised as the following workflow shows:

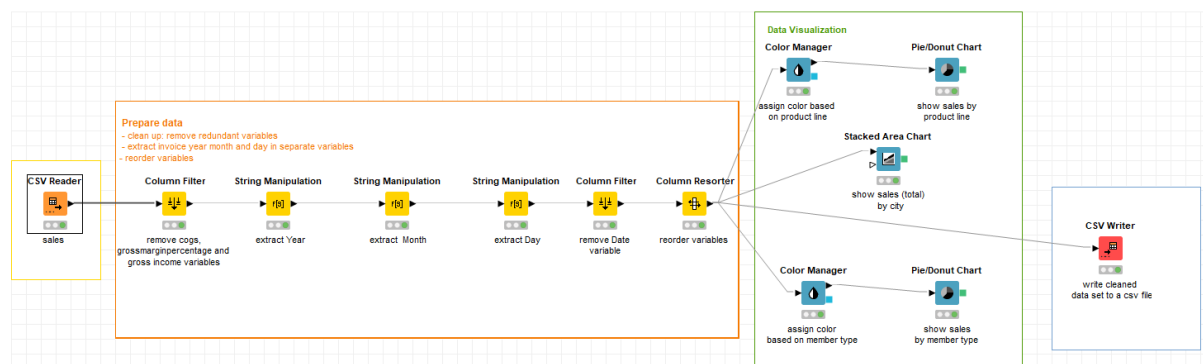


Figure 2: Sales data set workflow

CSV Reader node is used to import the sales data set csv file provided by the company.

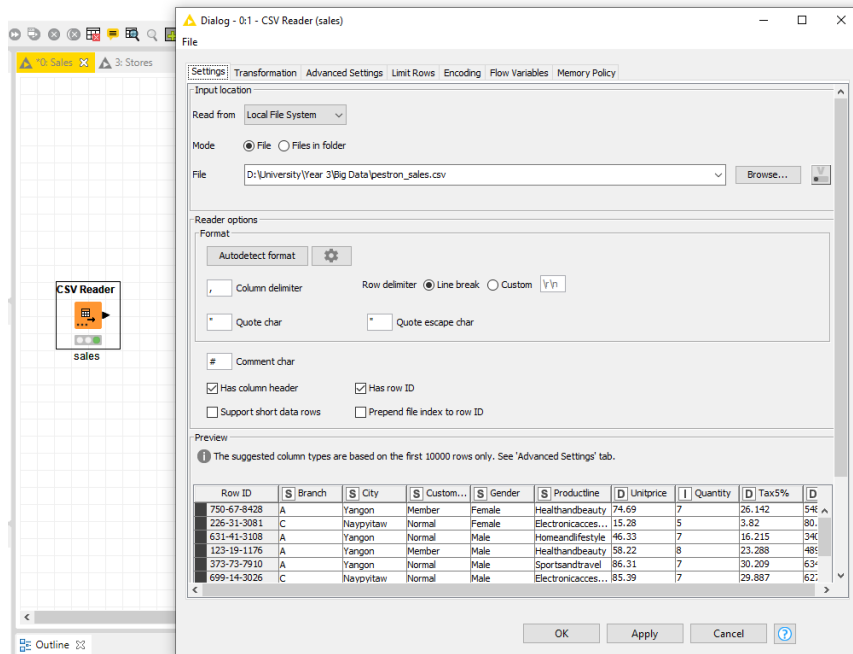


Figure 3: CSV Reader node configuration

An initial analysis is required to exclude unnecessary information. Duplicated rows, if existing are eliminated and redundant columns are removed. Cogs, grossmarginpercentage and grossincome variable will be removed using a Column Filter node, as they can be calculated from total and tax5% variables.

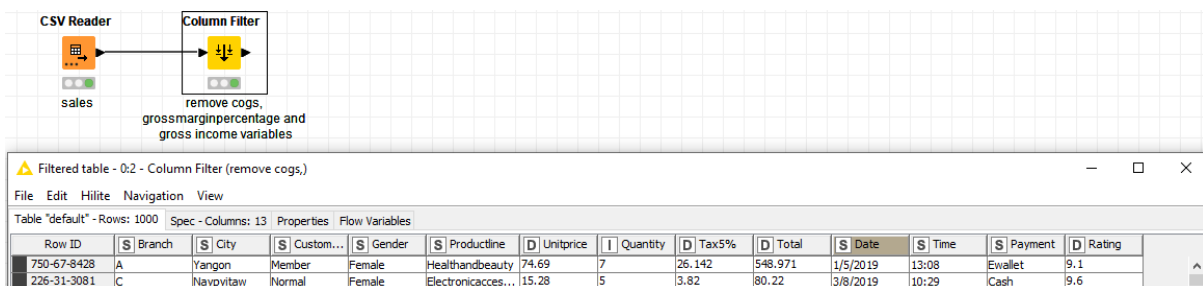


Figure 4: Column Filter node configuration

The Date variable has a string format and for a better analysing process the year, month and day were extracted into separate variables by using String Manipulation nodes.

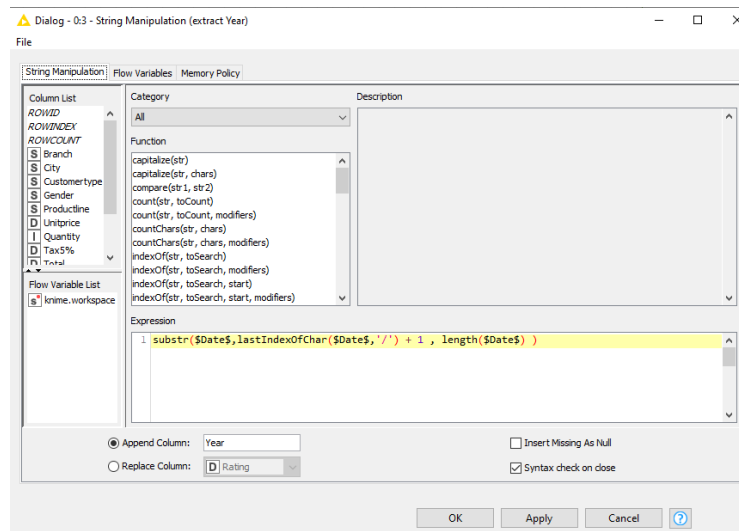


Figure 5: String Manipulation node for extract year configuration

Year variable extraction required an expression to be executed over the Date string and the result is stored into a separate column “Year”. The same mechanism, but with different string manipulation expressions, was applied for Month and Day and the output is reflected in Figure 6.

Appended table - 0-5 - String Manipulation (extract Day)

Table "default" - Rows: 1000 Spec - Columns: 16 Properties Flow Variables

Row ID	S City	S Custom...	S Gender	S Productline	D Unitprice	I Quantity	D Tax5%	D Total	S Date	S Time	S Payment	D Rating	S Year	S Month	S Day
750-67-8428	Yangon	Member	Female	Healthandbeauty	74.69	7	26.142	548.971	1/5/2019	13:08	Ewallet	9.1	2019	1	5
226-31-3081	Naypyitaw	Normal	Female	Electronicces...	15.28	5	3.82	80.22	3/8/2019	10:29	Cash	9.6	2019	3	8
631-41-3108	Yangon	Normal	Male	Homeandlifestyle	46.33	7	16.215	340.526	3/3/2019	13:23	Creditcard	7.4	2019	3	3
123-19-1176	Yangon	Member	Male	Healthandbeauty	58.22	8	23.288	489.048	1/27/2019	20:33	Ewallet	8.4	2019	1	27
373-73-7910	Yangon	Normal	Male	Sportsandtravel	86.31	7	30.209	634.379	2/8/2019	10:37	Ewallet	5.3	2019	2	8
699-14-3026	Naypyitaw	Normal	Male	Electronicces...	85.39	7	29.887	627.616	3/25/2019	18:30	Ewallet	4.1	2019	3	25
355-53-5943	Yangon	Member	Female	Electronicces...	68.84	6	20.652	433.692	2/25/2019	14:36	Ewallet	5.8	2019	2	25
315-22-5665	Naypyitaw	Normal	Female	Homeandlifestyle	73.56	10	36.78	772.38	2/24/2019	11:38	Ewallet	8	2019	2	24
665-32-9167	Yangon	Member	Female	Healthandbeauty	36.26	2	3.626	76.146	1/10/2019	17:15	Creditcard	7.2	2019	1	10
692-92-5582	Mandalay	Member	Female	Foodandbever...	54.84	3	8.226	172.746	2/20/2019	13:27	Creditcard	5.9	2019	2	20
351-62-0822	Mandalay	Member	Female	Fashionaccess...	14.48	4	2.896	60.816	2/6/2019	18:07	Ewallet	4.5	2019	2	6
529-56-3974	Mandalay	Member	Male	Electronicces...	25.51	4	5.102	107.142	3/9/2019	17:03	Cash	6.8	2019	3	9
365-64-0515	Yangon	Normal	Female	Electronicces...	46.95	5	11.738	246.488	2/12/2019	10:25	Ewallet	7.1	2019	2	12

Figure 6: String Manipulation (extract Day) node execution output

Once these data were extracted, the Date variable become redundant, and it was removed from the data set using a Column Filter node. Figure 7 shows the data set after removing the Date variable:

Filtered table - 0-6 - Column Filter (remove Date)

Table "default" - Rows: 1000 Spec - Columns: 15 Properties Flow Variables

Row ID	S Branch	S City	S Custom...	S Gender	S Productline	D Unitprice	I Quantity	D Tax5%	D Total	S Time	S Payment	D Rating	S Year	S Month	S Day
750-67-8428	A	Yangon	Member	Female	Healthandbeauty	74.69	7	26.142	548.971	13:08	Ewallet	9.1	2019	1	5
226-31-3081	C	Naypyitaw	Normal	Female	Electronicces...	15.28	5	3.82	80.22	10:29	Cash	9.6	2019	3	8
631-41-3108	A	Yangon	Normal	Male	Homeandlifestyle	46.33	7	16.215	340.526	13:23	Creditcard	7.4	2019	3	3
123-19-1176	A	Yangon	Member	Male	Healthandbeauty	58.22	8	23.288	489.048	20:33	Ewallet	8.4	2019	1	27
373-73-7910	A	Yangon	Normal	Male	Sportsandtravel	86.31	7	30.209	634.379	10:37	Ewallet	5.3	2019	2	8
699-14-3026	C	Naypyitaw	Normal	Male	Electronicces...	85.39	7	29.887	627.616	18:30	Ewallet	4.1	2019	3	25
355-53-5943	A	Yangon	Member	Female	Electronicces...	68.84	6	20.652	433.692	14:36	Ewallet	5.8	2019	2	25
315-22-5665	C	Naypyitaw	Normal	Female	Homeandlifestyle	73.56	10	36.78	772.38	11:38	Ewallet	8	2019	2	24

Figure 7: Column Filter (remove Date) node execution output

Last step applied, before performing a data visualization, was the reordering of columns using Column Resorter node.

Data visualization section contains several charts that offer a better understanding of the sales data set.

Figure 8 shows a graphical representation of the sales based on the product line. To obtain this chart, a Color Manager node was used and then a Pie/Donut chart node.

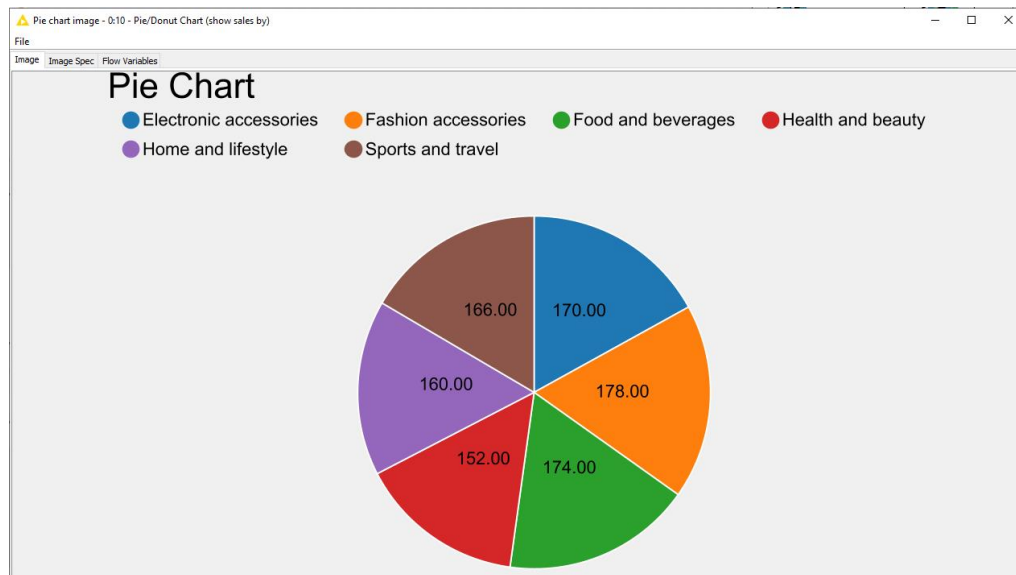


Figure 8: Graphical representation of sales based on product line

Products from the “Fashion accessories” categories are the most sold ones, while the “Health and beauty” category has the lowest number. Understanding the customers’ loyalty is also an important aspect and Figure 9 illustrates it.



Figure 9: Graphical representation of sales based on customer type

It is imperative to notice that there is almost an equality between the member and non-member customers. The target will be to increase the number of members.

After performing an initial data visualization, the cleaned data set was exported into a csv file using the CSV Writer and this file will be used for defining the model and creating the Data Warehouse for Pestron Foods organization.

Stores workflow

Stores data set contains a very low number of records, and it is very specific to the details about the stores that the organization has in Columbus city. For this data set only an initial simple data visualization was performed.

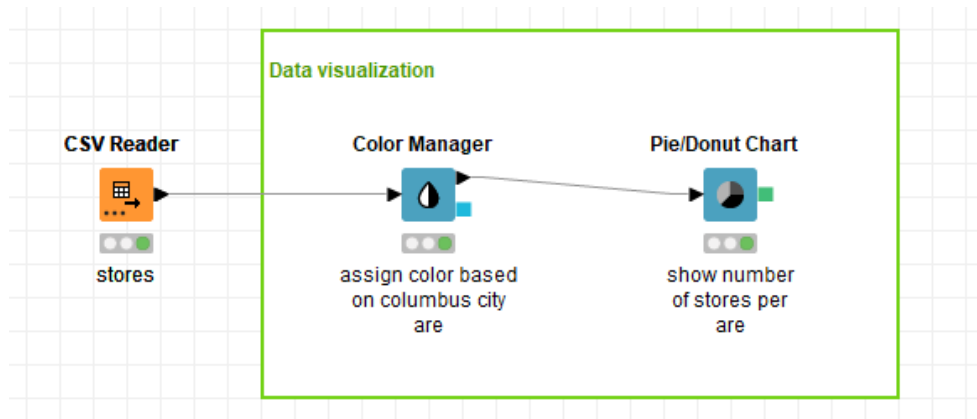


Figure 10: Stores data set workflow

The data set was imported in KNIME using the CSV Reader and the number of stored per each area of the city were displayed in a graphical representation.

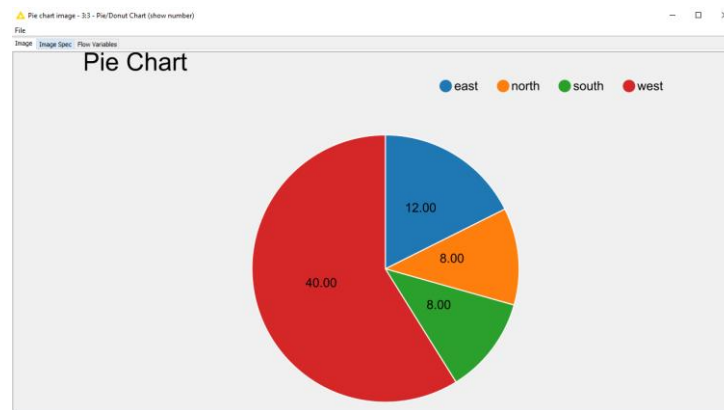


Figure 11: Graphical representation of stores based on the city

As Figure 11 shows, the Pestron Foods organization has most of the stores, existing in Columbus city, located in the west area of the city.

Next step is to create a predictive model using the linear regression method. A new workflow, called PredictiveModel was created. Pestron Foods collected 1000 sales for six product categories across three branches in different cities. The aim is to build a predictive model and find out the sales for each product category for a particular customer type (member or normal customer).

Figure 11 shows the steps required to obtain the predictive model.

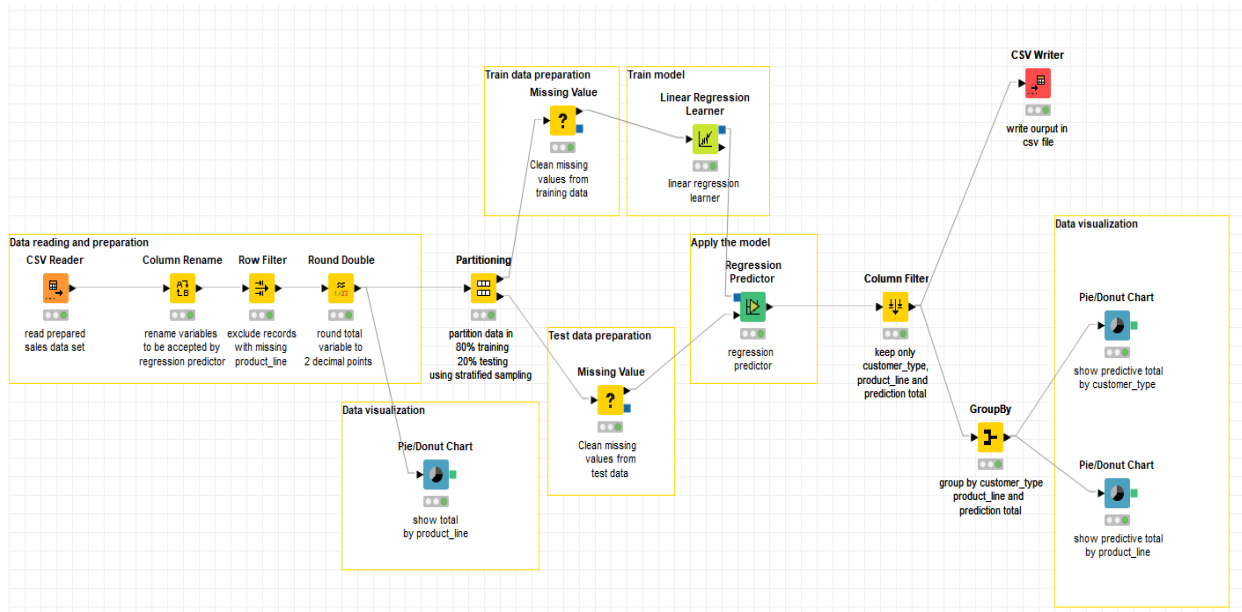


Figure 12: Predictive model workflow

Linear regression model was chosen for a better understanding of the relationship between the total and the type of customer and product line. Prepared data set, obtain in Sales workflow was imported in KNIME using a CSV Reader node. The variable's names were modified to mitigate an issue exposed by the regression predictor. For example, customer type variable was initially called "Customertype" which results in the error shown by Figure 13. It was renamed to "customer_tpe" and for consistency all variables followed the same naming convention as shown in Figure 14.

ERROR Regression Predictor 11:4 Execute failed: The column "Customer_type" is in the model but not in given table.

Figure 13: Column name error

Row ID	S row_id	S branch	S city	S custom...	S gender	S product_line	D unit_price	I quantity	D tax_5...	D total	I year	I month	I day	S time	S payme...
Row0	750-67-8428	A	Yangon	Member	Female	Healthandbeauty	74.69	7	26.142	548.971	2019	1	5	13:08	Ewallet
Row1	226-31-3081	C	Naayprytaw	Normal	Female	Electronics...	15.28	5	3.82	80.22	2019	3	8	10:29	Cash
Row2	631-41-3108	A	Yangon	Normal	Male	Homeandlifestyle	46.23	7	16.215	340.528	2019	3	3	13:23	Creditcard
Row3	175-12-1176	A	Yangon	Member	Male	Healthandbeauty	58.22	8	23.288	489.048	2019	1	27	20:13	Ewallet

Figure 14: Columns renaming output

In the next node, Row Filter, the data set was cleaned by excluding the records with missing values for product_line variable. Also, the total variable was rounded to two decimal points numbers. Before creating the linear regression model a chart was created for showing total amount generated by each product category. This chart doesn't include the customer category variable.

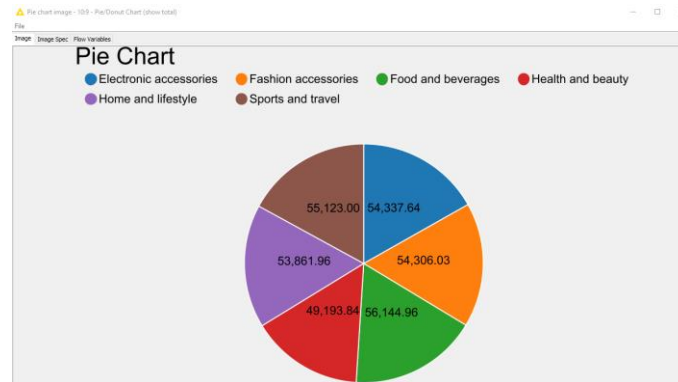


Figure 15: Graphical representation of sales based on product line

Linear regression model needs two data sets, a data set for training and another one for testing. Partitioning model was splitting the initial data set into 80% data for training purpose and 20% data for testing using stratified sampling. Missing values were treated in both training and testing data.

Linear regression learner receives training data as input and its configuration is shown in Figure 16.

Figure 16: Linear Regression Learner node configuration

The output is represented by a model for predictor and coefficients and statistics.

Coefficients and Statistics - 10:5 - Linear Regression Learner (linear regression)					
File Edit Hilite Navigation View					
Table "Coefficients and Statistics" - Rows: 7 Spec - Columns: 5 Properties Flow Variables					
Row ID	Variable	Coeff.	Std. Err.	t-value	P> t
Row1	customer_type=Normal	5.76	17.159	0.336	0.737
Row2	product_line=Fashion...	-12.652	29.062	-0.435	0.663
Row3	product_line=Foodan...	24.928	29.197	0.854	0.393
Row4	product_line=Healtha...	4.654	30.082	0.155	0.877
Row5	product_line=Homea...	30.747	29.597	1.039	0.299
Row6	product_line=Sportsa...	13.545	29.741	0.455	0.649
Row7	Intercept	302.549	22.721	13.316	0

Figure 17: Linear Regression Learner - coefficients and statistics output

These coefficients are essential for Pestron Foods understanding. An additional customer who is not a member increase the total amount by an average amount of 5.76. The low P values (0.299 and 0.393) for Homeandlifestyle and Foodandbeverages product lines indicates that these two categories are statistically significant. An additional unit of Homeandlifestyle increase the average total amount with 30.747.

The predictive model was applied against the testing data set using a Regression Predictor node. The needed columns were filtered and grouped to obtain a data visualization as shown in Figures 18 and 19.

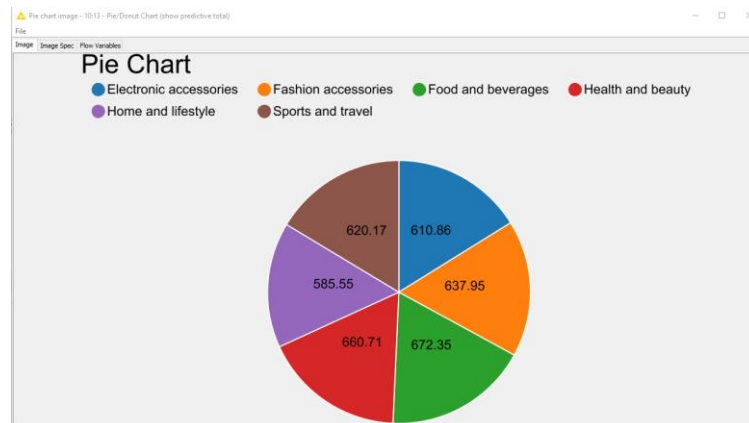


Figure 18: Predictive model applied - predicted total based on product line

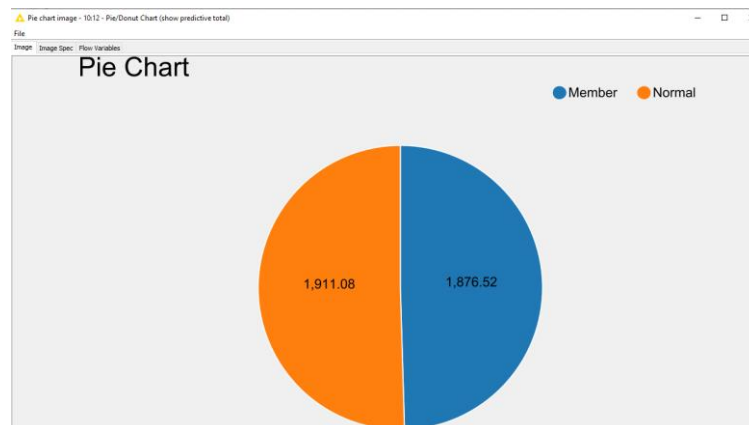


Figure 19: Predictive model applied - predicted total based on customer type

The output of the applied model was written in csv file using a CSV Writer node.

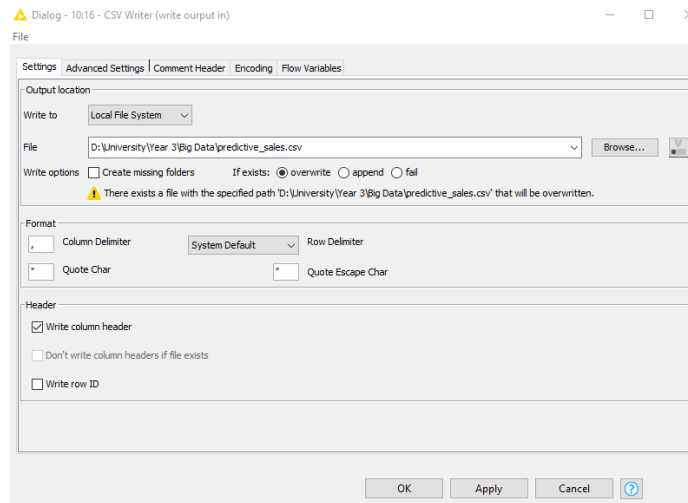


Figure 20: CSV Writer node configuration

Data Warehouse

For the Data Warehouse creation, Snowflake was used. Snowflake architecture contains three layers: storage, compute, and services. Storage layer store all the data, structured or unstructured and has the advantage or automatically management of files, size, metadata or statistics. Compute layer keeps virtual warehouses responsible for processing the queries. Cloud services layer includes authentication, infrastructure and metadata management, query optimization and access control and coordinates the entire system. This architecture brings huge benefits in terms of performance, speed, availability, accessibility, and security and represents an ideal solution for Pestron Foods organization.

A Snowflake account is required and a warehouse and database for Pestron Foods organization were created as Figures 21 and 22 show.

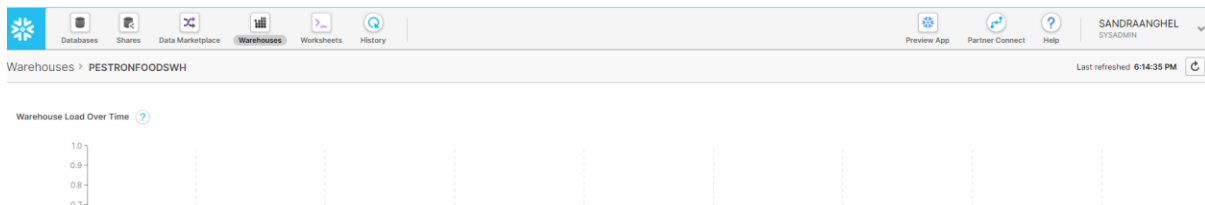


Figure 21: Snowflake Data Warehouse

Table Name	Schema	Creation Time	Owner	Rows	Size	Comment
------------	--------	---------------	-------	------	------	---------

Figure 22: Snowflake database

KNIME requires a Snowflake driver which is easy to install and configured. Based on the official documentation, Snowflake JDBC driver requires at least Java 1.8. Once downloaded the driver was added into KNIME Preferences under Database section.

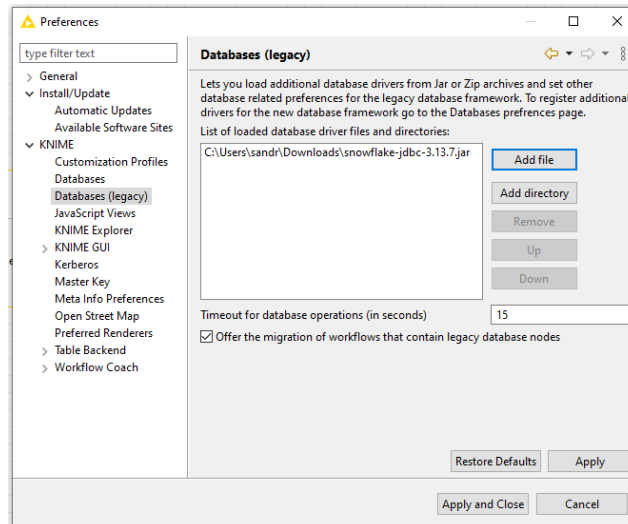


Figure 23: Snowflake JDBC driver configuration

A Database Connector node was used to connect to Snowflake Warehouse.

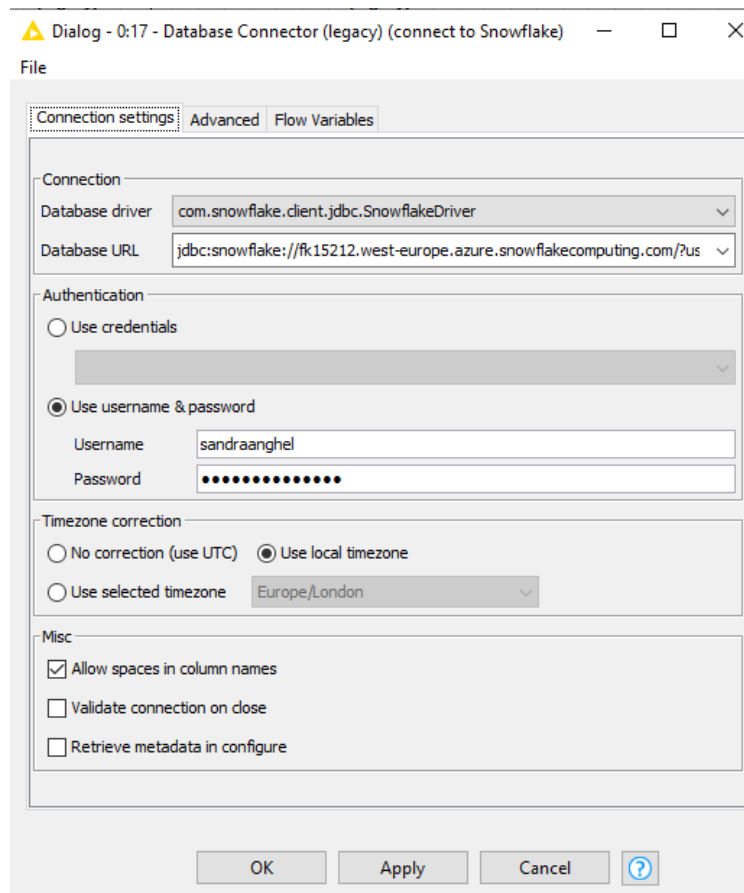


Figure 24: Database Connector node configuration

After the connection was established, the data set is imported into data warehouse by using a Database Writer and the output can be seen in the data warehouse. Complex queries can be used to interrogate the data set.

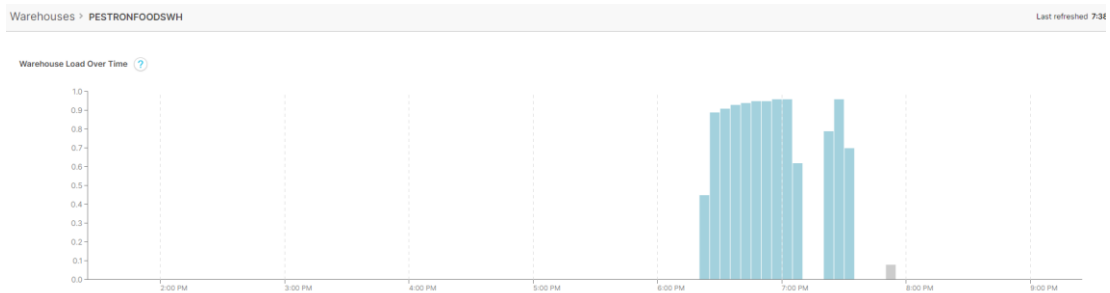


Figure 25: Snowflake PESTRONFOODSWH warehouse workload

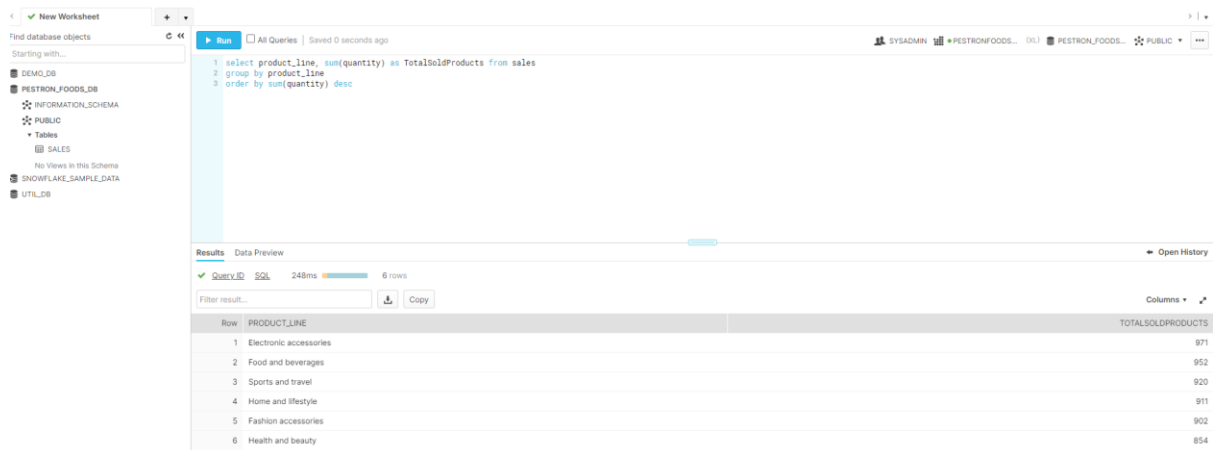


Figure 26: Snowflake PESTRON_FOODS_DB database interrogation

In this chapter, KNIME Analytics Platform was used to create an initial data visualization of the two data sets and based on sales data set, a linear regression model. Based on the predictive result, the company can increase the sales of products from Homeandlifestyle and Foodandbeverages categories by creating a marketing strategy which will send personalized offers to their member customers. Also, Healthandbeauty category requires a detailed analysis as it produces the lowest number of sold products. The data set was imported in a newly created data warehouse. Snowflake was used for the data warehouse creation.

Conclusions

In conclusion, this report highlights the benefits and challenges those small businesses will face in the Big Data adoption process. The business environment is permanently changing, and, in this context, digitalisation becomes one of the most important driving forces. Small businesses are in competition with large organization and the traditional methods seem to be deprecated. However, small businesses have some challenges that must be handled.

Based on this report findings, the significant challenges that small businesses handle, inclusive the Pestron Foods organization, are as follows:

- Reduce understanding of Big Data concepts
- Lack of infrastructure to process high volume data quickly
- Insufficient in-house data analytic expertise
- Concerns about data protection and data privacy

However, the report determined the major benefits that Pestron Foods organization have by following a linear regression model and the right tools for understanding customers preferences in terms of product lines and predicting the sales based on changes and marketing strategies that they adopt.

As already mentioned, far too little attention has been paid to advantages and challenges exposed by Big Data in small businesses context. Although this report provided an overview of the Big Data concepts and how it can be successfully implemented especially by small companies. Having more time, a detailed analysis and data visualization can be performed.

References

- [1] B. Marr, *Big Data in practice: How 45 successful companies used Big Data Analytics to deliver extraordinary results*, Chichester, West Sussex: John Wiley & Sons, Inc., 2016.
- [2] World Trade Organization, "Levelling the trading field for SMEs," World Trade Organization, 2016.
- [3] U.K. Government, "National statistics: Business population estimates for the UK and regions 2020," 08 10 2020. [Online]. Available: <https://www.gov.uk/government/statistics/business-population-estimates-2020/business-population-estimates-for-the-uk-and-regions-2020-statistical-release-html>. [Accessed 09 08 2021].
- [4] I. Sutherland, "John Graunt: A Tercentenary Tribute," John Wiley & Sons, Inc..
- [5] IBM, "The Punched Card Tabulator," IBM. [Online]. [Accessed 09 08 2021].
- [6] IBM, "Chronological History of IBM," [Online]. Available: https://www.ibm.com/ibm/history/history/history_intro.html. [Accessed 12 08 2021].
- [7] W. Shahriar Akter, "Big data analytics in e-commerce: A systematic review and agenda for future research," Faculty of Business - Papers (Archive), 2016.
- [8] J. Hurwitz, *Big Data for Dummies*, Hoboken, New Jersey: John Wiley & Sons, Inc., 2013.
- [9] Oracle, "What is Big Data?" [Online]. Available: <https://www.oracle.com/uk/big-data/what-is-big-data/>. [Accessed 14 08 2021].
- [10] M. Iqbal, "A Study of Big Data for Business Growth in SMEs: Opportunities Challenges," in *International Conference on Computing, Mathematics and Engineering Technologies – iCoMET*, 2018.
- [11] McKinsey&Company, "Marketing & Sales: Big Data, Analytics and the Future of Marketing & Sales," McKinsey&Company, 2015.
- [12] B. Hazen, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics*, vol. 154, 2014.
- [13] S. Suthaharan, "Big Data Classification: Problems and Challenges in Network Intrusion," Greensboro, 2012.
- [14] E. C. Aggarwal, *Data Classification: Algorithms and Applications*, CRC Press, 2014.

APPENDIX A: Big Data Tools advantages and disadvantages

Apache Hadoop

Official url: [Apache Hadoop](#)

- Advantages
 - Scalability
 - Cost effective
 - Flexibility
 - Resilient to failure
- Disadvantages
 - Security concerns
 - Inefficient support for small data
 - Stability issues

Apache Spark

Official url: [Apache Spark™ - Unified Analytics Engine for Big Data](#)

- Advantages
 - Speed
 - Advanced analytics
 - Easy to use
 - Multilingual
 - Increased access to Big Data
- Disadvantages
 - No support for multi-user environment
 - No automatic optimization processes
 - Reduced number of algorithms

Big Data Sqoop

Official url: [Sqoop - \(apache.org\)](#)

- Advantages
 - Cost effective
 - Optimized ETL (Extract Load, Transform) process
 - Parallel import/export
 - Variety of structured data sources
- Disadvantages
 - Less performant for JDBC connections
 - Slow
 - Special handle requirements for failures
 - No GUI (graphic user interface) provided

IBM Watson Analytics

Official url: [Analytics | IBM](#)

- Advantages
 - Processes unstructured data
 - Performance
 - Support for huge quantities of data
- Disadvantages
 - No multilanguage support (only English)
 - Maintenance
 - No direct support for structured data

D3.js

Official url: [D3.js - Data-Driven Documents \(d3js.org\)](https://d3js.org)

- Advantages
 - Open source
 - Easy to use
 - Integrations with JSON and CSV files
 - Large collection for data visualization
 - Easy to customize
 - Great documentation
- Disadvantages
 - Hard to learn
 - Sower for large datasets
 - Web development experience required

Microsoft Power BI

Official url: [Data Visualization | Microsoft Power BI](https://powerbi.microsoft.com/en-us/data-visualization/)

- Advantages
 - Affordability
 - Custom and interactive visualizations
 - Excel integration
 - Data connectivity
 - Historical data comparison
 - Process tracking
- Disadvantages
 - Complex to understand and learn
 - Pricing tiers based on data size
 - Crowded User Interface

Google BigQuery

Official url: [BigQuery: Cloud Data Warehouse | Google Cloud](https://cloud.google.com/bigquery/)

- Advantages
 - Ready to use without complex configurations
 - Pay-for-use pricing model
 - Serverless architecture

- Automated backup
- Disadvantages
 - SQL experience required
 - Reduced number of partitions
 - Dependent on Google Cloud Platform
 - Hard to learn

Amazon Redshift

Official url: [Amazon Redshift - Cloud Data Warehouse - Amazon Web Services](https://aws.amazon.com/redshift/)

- Advantages
 - Widely adopted
 - Easy to query or administrate
 - Performance
 - Scalability
 - Pricing
 - Security
- Disadvantages
 - Not a serverless architecture
 - Not a multi-cloud solution (unlike Snowflake)
 - Limited support for parallel upload
 - Uniqueness not enforced

Hive

Official url: [Manage projects faster & collaborate better | Hive](https://hive.apache.org/)

- Advantages
 - Centralized communication
 - Transparency
 - Suitable for complex projects
 - Monitoring
- Disadvantages
 - File deletion is permanent
 - Unorganized notifications
 - Lack of search functions

HBase

Official url: [Apache HBase – Apache HBase™ Home](https://hbase.apache.org/)

- Advantages
 - Licence free
 - Fault tolerance
 - Suitable for large volumes of data
 - Easy user interface
 - Scalability
 - Integration with Hive
- Disadvantages
 - No support for transactions
 - Single point of failure

Apache Atlas

Official url: [Apache Atlas – Data Governance and Metadata framework for Hadoop](#)

- Advantages
 - High availability
 - Fault tolerance
 - Security
 - Centralized auditing
 - Multiple ingestion sources supported (Hive, HBase, Sqoop)
- Disadvantages
 - Lack of support/help
 - Lack of TLS certificates validation on Firefox
 - Incorrect behaviour/redirects for unauthenticated users