SALES ANALYSIS REPORT

Case study to explore the relationship between Quantity and Value, detect anomalies, and extract insights in a Sales Dataset.

MORRIS MUSYOKI
JANUARY 30, 2025
0795026559
https://github.com/wayne-cipher
morrismmusyoki254@gmail.com

OVERVIEW

This document provides comprehensive outline of the steps taken to clean and prepare the dataset for exploratory and advanced analysis, including visualizations, insights and recommendations. The case study dataset has 333406 entries and 7 variables (2 numerical and 5 categorical.

SECTION 1: DATA CLEANING AND PREPARATION

Issues identified and mitigations.

The dataset contained 3524 duplicates which were manually inspected to ensure they indeed are real duplicates then subsequently deleted.

Inconsistent data types in Date, Quantity and Unit price columns changed to their correct data types; date/time, whole number and currency respectively.

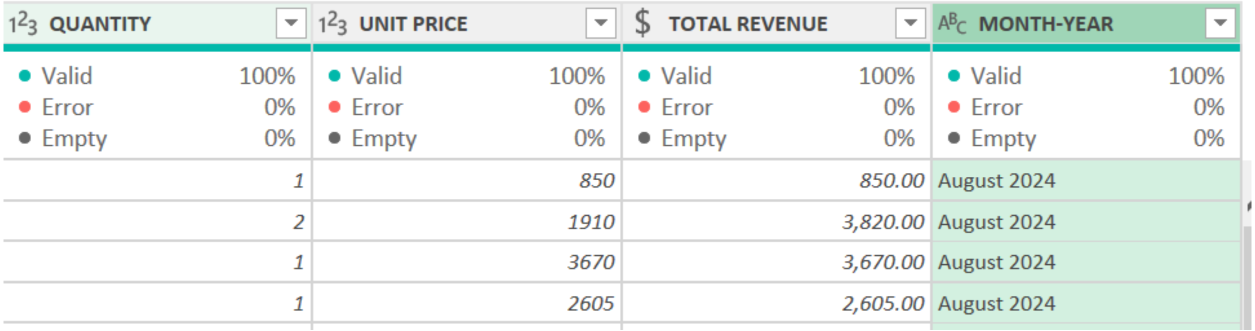
Unit Price column had 8 missing values. They were replaced with the median (1850) of the entire column. It also had 26 rows with input 0 which were assumed to have been promotional products with full discount at that period of time, therefore kept it as zero.

40 rows had Quantity 0 which implies the product was out of stock so no transaction made. This means these rows are irrelevant to the analysis. The rows were dropped.

Adding columns

The "Month-Year" column is introduced to facilitate trend evaluation by using grouping data via particular time durations. This lets in for less complicated identification of seasonal styles and fluctuations in sales interest and we can uncover insights into how sales performance modifications over time. Additionally, this temporal aggregation enables in comparing different intervals for more informed decision-making.

The "TOTAL REVENUE" column is added to the dataset to calculate overall sales performance by multiplying quantity sold by unit price. By studying overall sales, we are able to pick out trends and top performing products, which helps guide strategic enterprise choices.



123 QUANTITY	123 UNIT PRICE	$ TOTAL REVENUE	ABC MONTH-YEAR
● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
1	850	850.00	August 2024
2	1910	3,820.00	August 2024
1	3670	3,670.00	August 2024
1	2605	2,605.00	August 2024

SECTION 2: EXPLORATORY DATA ANALYSIS

Sales Overview

The table below shows the top 10 performing business categories based on the total sales. It also highlights the total Quantity and Total Sales for each Category. Businesses in these categories show a great protentional for growth.

ANONYMYZED CATEGORY	TOTAL QUANTITY	TOTAL SALES
Category-75	152643	$549,509,348.00
Category-76	72928	$351,827,338.00
Category-120	171443	$322,737,950.00
Category-100	77704	$136,417,463.00
Category-119	68615	$103,900,839.00
Category-77	28825	$77,791,642.00
Category-91	21081	$44,700,098.00
Category-101	19803	$36,003,677.00
Category-85	23368	$34,298,630.00
Category-121	14936	$22,677,154.00

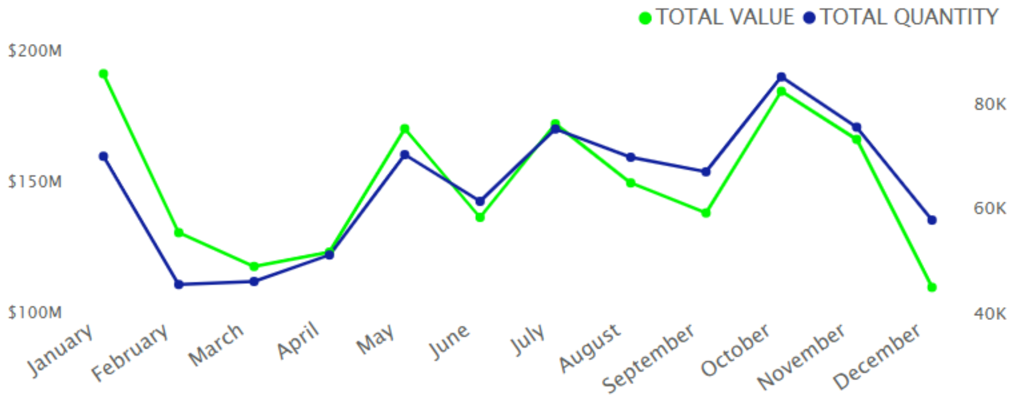
The table below shows the top 10 performing businesses based on the total sales. It also highlights the total Quantity and Total Sales for each Business.

ANONYMYZED BUSINESS	TOTAL QUANTITY	TOTAL SALES
Business-978e	14023	$28,076,363.00
Business-fe7d	6743	$26,997,121.00
Business-6068	8262	$16,542,970.00
Business-07de	6134	$16,414,443.00
Business-7a03	6318	$13,968,451.00
Business-ba13	5545	$13,692,866.00
Business-1e3e	4981	$13,192,967.00
Business-468e	5452	$12,554,997.00
Business-f4f4	3852	$11,952,941.00
Business-5613	4089	$11,895,552.00

P.S: Full interactive tables showing all businesses and categories are fully shown in the dashboard.

Sales and Quantity Trend over Time

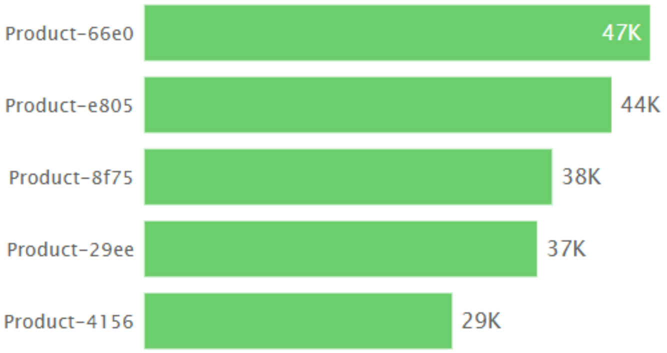
Both the total value and quantity bought exhibit comparable patterns in the course of the 12 months, with each strains rising and falling in tandem. This correlation suggests that as the quantity bought will increase, general income also have a tendency to rise, indicating an instantaneous relationship among the range of products sold and the revenue generated. The constant motion of each tendencies reinforces the concept that adjustments in client demand directly impact basic sales overall performance.



PERFORMANCE ANALYSIS

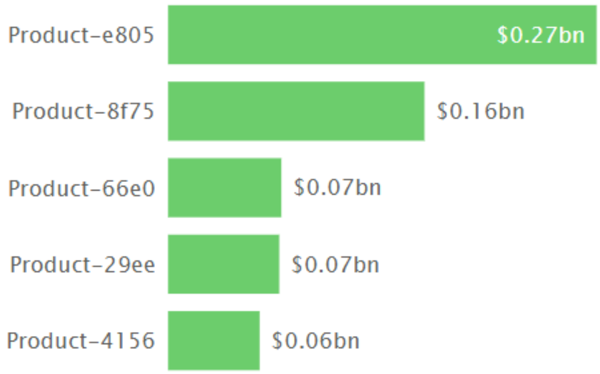
Top 5 frequently purchased products based on quantity

The chart displays the top five frequently purchased products based on quantity sold, highlighting their popularity among consumers. Leading the list is Product-66e0, with an impressive total of 47,000 units sold, followed closely by Product-e805 at 44,000 units. Product-8f75 comes in third with 38,000 units, while Product-29ee and Product-4156 round out the top five with 37,000 and 29,000 units sold, respectively. This data illustrates the demand for these products, offering valuable insights into consumer preferences and purchase behaviors.



Top 5 most valuable products

In this evaluation of product overall performance, the top five maximum treasured merchandise are identified based on general sales generated. Leading the list is Product-e805, which has carried out an extremely good value of $0.27 billion, making it the standout performer. Following intently is Product-8f75 with a cost of $0.16 billion. Both Product-66e0 and Product-29ee share a price of $0.07 billion, highlighting their competitive positioning inside the marketplace. Finally, Product-4156 rounds out the listing with a value of $0.06 billion. These records underscore the vast contributions of these merchandise to our overall sales stream.



SECTION 3: ADVANCED ANALYSIS

Customer Segmentation

A new table was generated from the main Case Study dataset called Segmentation Table with the main column ANONYMIZED BUSINESS, Total Quantity, Total Value and Frequency of Transactions. All these was done a DAX function in Power BI.

The DAX Function:

```
SegmentationTable =
SUMMARIZE(
    Case_Study_Data,
    Case_Study_Data[ANONYMIZED BUSINESS],
    "Total Quantity", SUM(Case_Study_Data[QUANTITY]),
    "Total Value", SUM(Case_Study_Data[ TOTAL REVENUE ]),
    "Frequency of Transactions", COUNTROWS('Case_Study_Data')
)
```

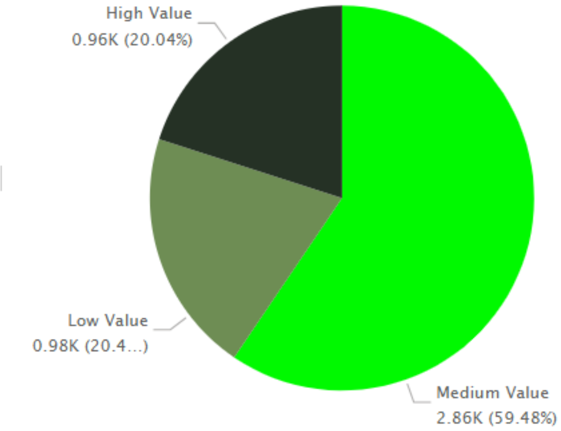
After creating these table, one more column had to be added; Value Segment based on the Total Value column of the Segmentation Table. A threshold to determine where each business should fall is created using the following DAX functions:

```
HighValueThreshold = PERCENTILEX.INC(SegmentationTable, [Total Value], 0.8)
MediumValueThreshold = PERCENTILEX.INC(SegmentationTable, [Total Value], 0.2)
```

The DAX Function for creating the column:

```
Value Segment =
SWITCH(
    TRUE(),
    SegmentationTable[Total Value] > 355000, "High Value",
    SegmentationTable[Total Value] > 7000, "Medium Value",
    "Low Value"
)
```

As shown in the code above, the businesses are classified in three classes; High value, Medium value and Low value. From the figure below it clearly shows the number of businesses with high value are as much as those with low value and most businesses lie in the medium class.



Recommendations on Business Segmentation

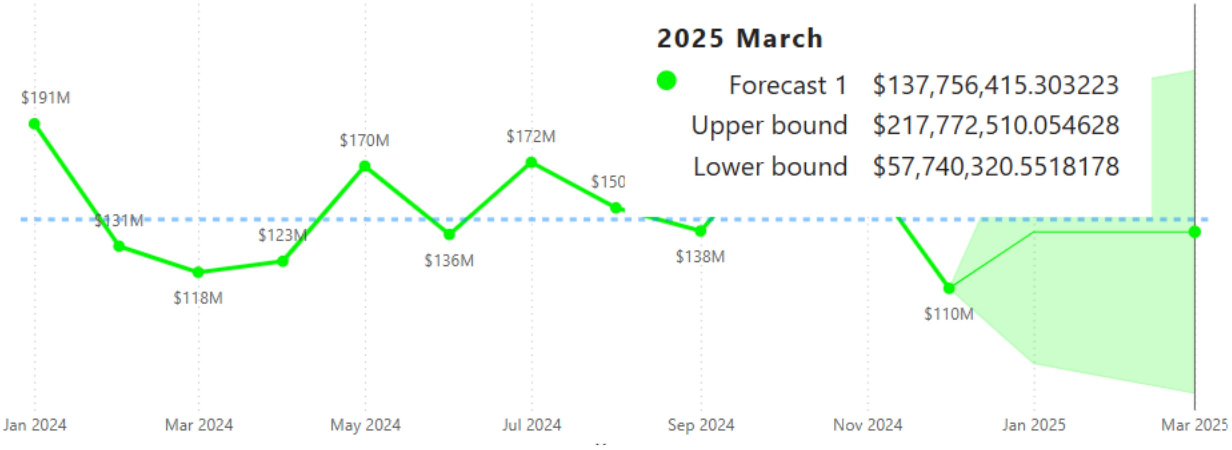
High level: Assign dedicated managers to foster strong, long-term relationships, and high-level service to maintain these businesses on top.

Medium level: Provide personalized marketing strategies that address specific points effectively highlighting your value propositions to enhance loyalty and engagement.

Low level: Focus on cheap marketing strategies like automated emails and ads to maintain a presence without significant investment. Try another different type of business brand.

Forecasting

Based on the 12 months data provided, 3 months forecast in the future showed that the Total Sales would be Approximately 138M by March 2025 as shown on the graph below (forecast shaded area). This is slightly below the median but it’s a positive sign that there’ll be some progress considering the fluctuation in sales in December 2024.



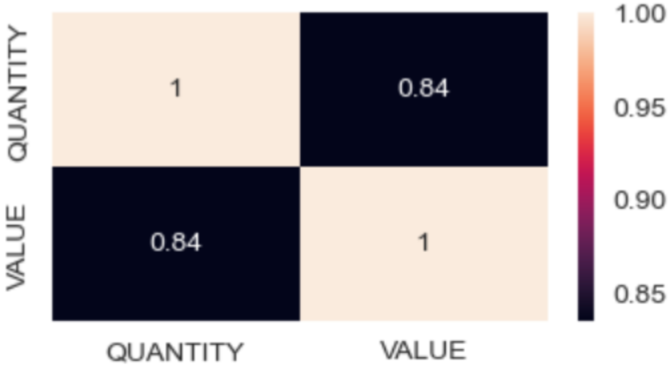
Anomalies

The months May – July and December there are high drops in sales. In December I'd suppose customers have gone to different locations to be with there families and this would explain why the sales are too low.

The same year in January sales spike show high returns to the business owners, which can be explained by customers trying to plan for the year so expect them to do more purchases.

Correlation

To better understand correlation between Quantity and Value a heat map is generated using the Python scripting icon in Power BI. The heatmap shows positive value (0.84) implying positive perfect correlation (as one variable increases, the other also increases).



Factors driving sales performance

This correlation shows that the volume of inventory (Quantity) is key to overall business performance and revenue generation. This means efficient inventory management is key to ensuring high demand products are always available.

Enhancing customer engagement by bringing from time-to-time loyalty programs and promotional products.

Monitoring market trends and technology advancements in business.

SECTION 4: STRATEGIC INSIGHTS AND RECOMMENDATIONS

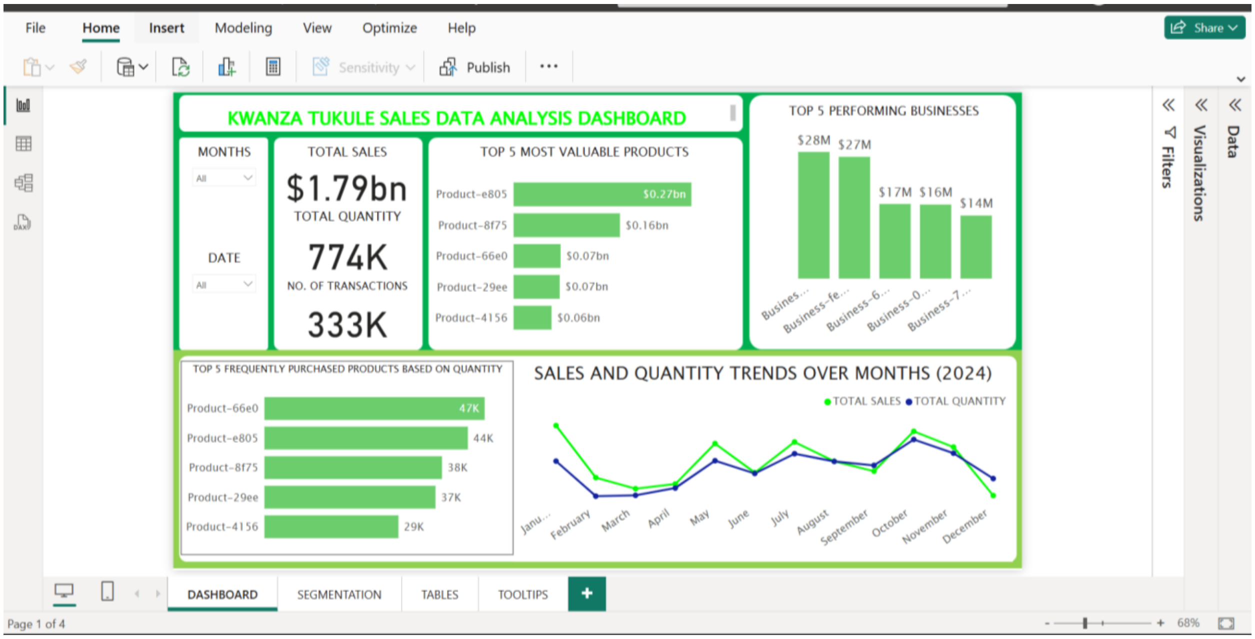
1. **Category-79**: All products in this category require priority marketing campaigns. This category in a period of 12 months it generated approximately $550M with the second category having $200M less Total Sales compared to category-79. This category is in high demand that means faster ROI. Most times customers are attracted to products which other customers by which can drive the sales through social proof.

2. **Customer Retention**: In business sometimes, they depreciate instead of appreciating eg business-108f with $92k returns in January but dropping to $21k in just 5 months. Other businesses include: business-12e1 ($190k to $3k), business-0354 ($78k to $6k) business-0c77 ($104k to $19k) etc.
   Note: To re-engage these customers, we can send personalized emails, provide offers like exclusive discounts and promotions, launch campaigns to specifically re-engage lapsed customers and provide surveys and feedback forms.

3. **Operational Efficiency:** Given the monthly trends, it would be wise to increase inventory primarily at the beginning of the year for a duration of four months. Additionally, the last five months of the year also show significant customer purchases, indicating that businesses generate substantial income during these periods.

DASHBOARD

The screenshots below the dashboard visualizations. To access dashboard file tap https://github.com/wayne-cipher/Assessment



KWANZA TUKULE SALES DATA ANALYSIS DASHBOARD

MONTHS: All

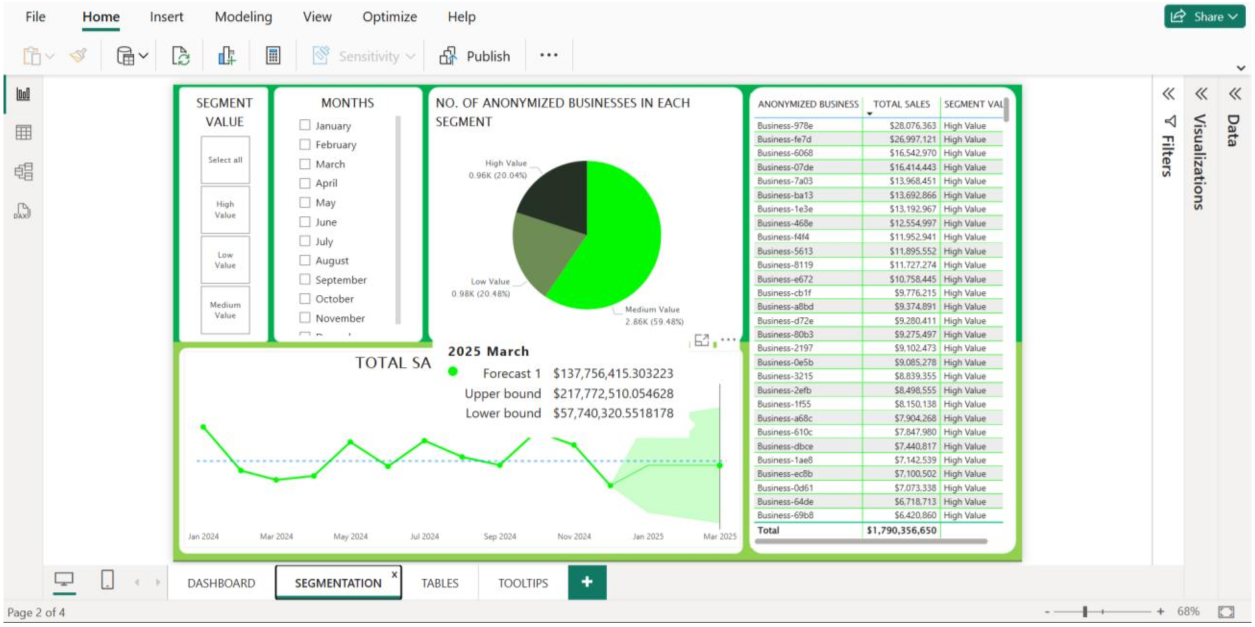
DATE: All

TOTAL SALES: $1.79bn

TOTAL QUANTITY: 774K

NO. OF TRANSACTIONS: 333K

Anonymized business groups segmentation visual and table.



2025 March

● Forecast 1	$137,756,415.303223
Upper bound	$217,772,510.054628
Lower bound	$57,740,320.5518178

ANONYMIZED BUSINESS	TOTAL SALES	SEGMENT VAL
Business-978e	$28.076.363	High Value
Business-fe7d	$26.997.121	High Value
Business-6068	$16.542.970	High Value
Business-07de	$16.414.443	High Value
Business-7a03	$13.968.451	High Value
Business-ba13	$13.692.866	High Value
Business-1e3e	$13.192.967	High Value
Business-468e	$12.554.997	High Value
Business-f4f4	$11.952.941	High Value
Business-5613	$11.895.552	High Value
Business-8119	$11.727.274	High Value
Business-e672	$10.758.445	High Value
Business-cb1f	$9.776.215	High Value
Business-a8bd	$9.374.891	High Value
Business-d72e	$9.280.411	High Value
Business-80b3	$9.275.497	High Value
Business-2197	$9.102.473	High Value
Business-0e5b	$9.085.278	High Value
Business-3215	$8.839.355	High Value
Business-2efb	$8.498.555	High Value
Business-1f55	$8.150.138	High Value
Business-a68c	$7.904.268	High Value
Business-610c	$7.847.980	High Value
Business-dbce	$7.440.817	High Value
Business-1ae8	$7.142.539	High Value
Business-ec8b	$7.100.502	High Value
Business-0d61	$7.073.338	High Value
Business-64de	$6.718.713	High Value
Business-69b8	$6.420.860	High Value
Total	**$1,790,356,650**	

BONUS SECTION: OPEN-ENDED PROBLEM

External factors that could influence sales

i. Economic Conditions: Economic indicators inclusive of GDP growth, unemployment charges, and patron self-belief can considerably affect consumer spending and buying conduct.
ii. Competitor Actions: Changes in competition' pricing techniques, product launches, and marketing campaigns can immediately affect sales performance.
iii. Seasonality: Seasonal tendencies can affect patron demand, making it vital to account for versions all through the yr.
iv. New policies or changes in alternate regulations can have an effect on marketplace dynamics and income.
v. Innovations in technology can create new market opportunities or disrupt existing ones.

Proposed Methodology

i. Collect historical sales data with external data sets, including economic indicators, competing data and seasonal trends. Sources may include government reports, market research firms and industry publications.
ii. Function Technique: Create relevant features from the data collected, such as interval variables for economic indicators, competitive price matrix and seasonal dummy - variable trends.
iii. Conduct the correlation analysis to understand the relationship between sales and external factors, and identify which factors have the most important impact on sales performance.

Scalability for large datasets

Storage

Use a strong database management system (DBMS) as postgreSQL, or cloud -based solutions to handle large data sets effectively (Amazon Redshift or Google BigQuery).

Use data compression techniques to reduce storage requirements and improve recovery time without giving up data integrity.

Partition: The division of large data sets in small, managed segments based on the most important variables (eg date, category) to increase query performance and data management.

Processing

Instead of processing the entire dataset at a time, use batch treatment to handle the small part of data, reduce memory use and adapt to resource allocation.

Distributing data processing tasks in many cores or machines, parallel processing frameworks, utilizing parallel processing structures, data management significantly sharp.

Analysis

Sampling: For seeking/exploratory analysis, consider using representative sampling techniques to gain insight from a small mastery of data, which is quickly and more handled.

Utilizing Cloud Computing Platforms (eg AWS, Azure, Google Cloud) that offer scalable resources and services allow for dynamic allocation of calculation power as needed.