Data Storage Solutions - using daily product sales and returns dataset

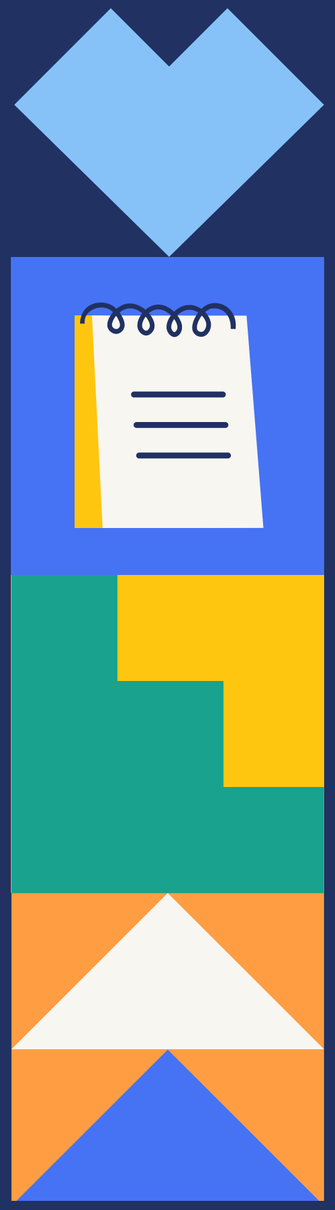
Explore how to organize and present data

Presented to : Dr Anesu Nyabadza

Presented by : Group 1

Sourav Basu 20031652
Kapil Sharma 20030912
Raghava Poral Ramamuthy 20032079
Dilsha Manjeesh 20028151
Aaria Mary 20029035

Date : 08. 12. 2024



Aim & Objectives

Aim

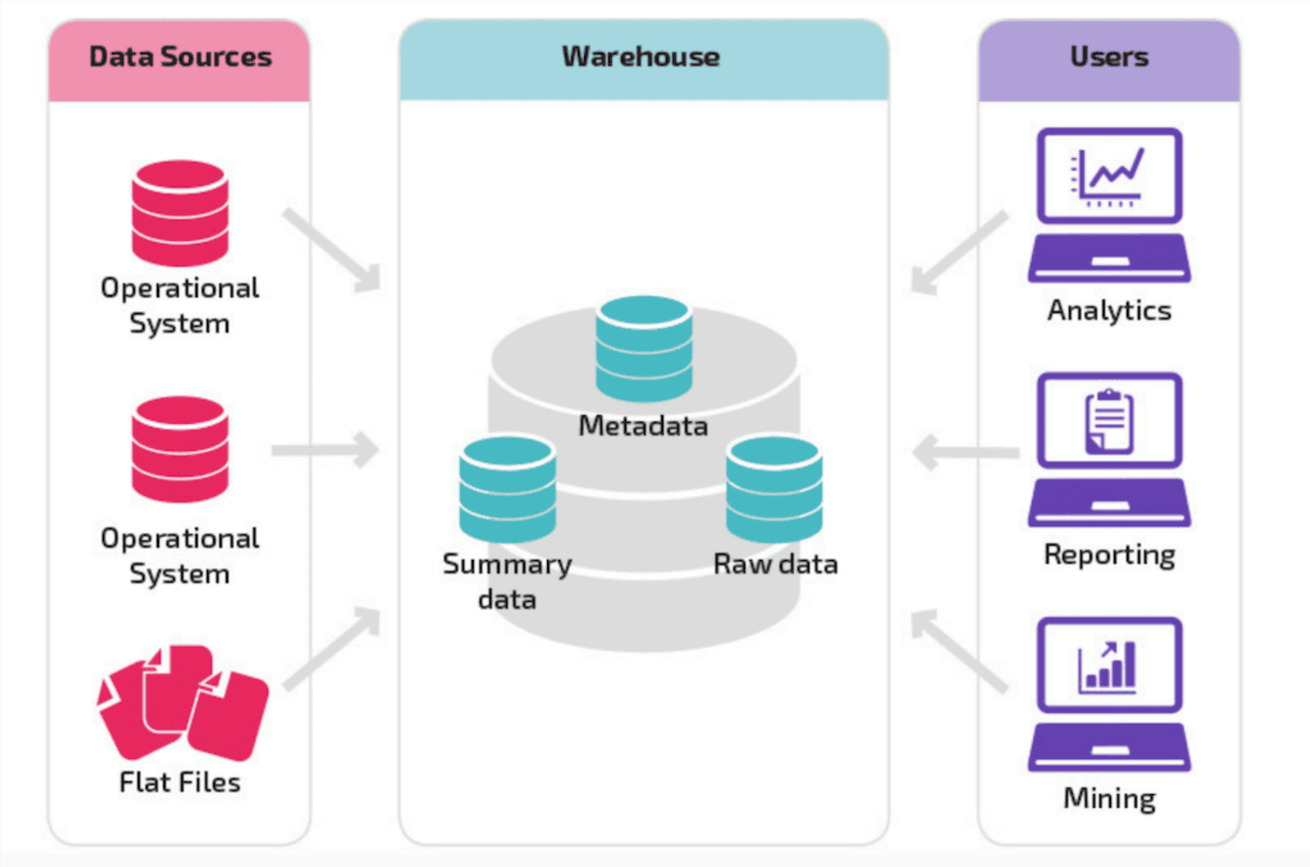
Build a scalable, secure data warehouse to support business intelligence and operational decision-making.

Objectives:

- Normalize raw transactional data.
- Implement ETL workflows for data integration.
- Generate actionable insights using visualizations and reports.
- Compare relational databases and graph databases for complex queries.

The Background Story

The workflow of the entire process can be depicted as follows



the pictorial representation of what is about to happen in a lucid way

Dataset Overview

Source

Source: Product Sales and Returns Dataset (Kaggle)

Attributes

- Item Details: Item_Name, Category, Version, Item_Code.
- Transactions: Buyer_ID, Transaction_ID, Date.
- Revenue Metrics: Total Revenue, Price Reductions, Refunds, Sales Tax.

Dimensional Model design -The schema overview

Fact Table: Fact_Sales

- Metrics: Revenue, Refunds, Tax.

Dimension Tables: Dim_Item, Dim_Buyer, Dim_Date. Flowchart:

Include a simple star schema diagram.

ETL Process

Extract

Source: CSV files.
Tools: Flat file connection in SSIS.

Transform

Data cleansing, lookups for dimension keys, handling nulls.

Load

Fact_Sales and Dimension Tables.
Flowchart:

SSIS Workflow

- error handling with redirection for invalid rows
- create a proper data flow and control flow for pushing parent data from flat files to the database after mapping using correct datatypes

Data Flow Tasks:

Flat File Source → Lookups for Dim_Item, Dim_Buyer, Dim_Date

Error handling with redirection for invalid rows

SSRS Reports

Reports Ctreated

- **Sales Summary Report.**
- **Customer Purchase Behavior Report**
- **Monthly Sales Performance Report**
- **Refund Analysis Report**

All of these reports showers clarity on the business which can enhance the prospect of the business in the long run

Data Visualization in Tableau

Visualizations

- Revenue trends over time
- Top Products by Final Revenue
- Sales and Returns Comparison by Category
- Customer Segmentation by Purchase Behavior

Neo4j Fraud Detection Database

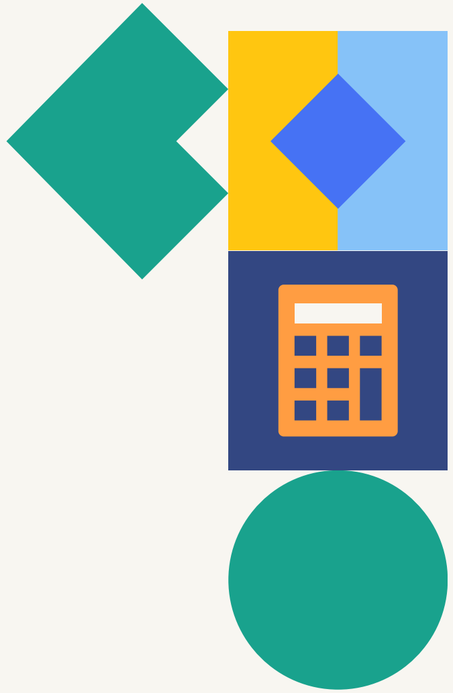
Use Case

A Detect relationships between buyers, items, and transactions

B Identify buyers who made high-value transactions - Comparison

C Neo4j vs Relational Databases

D Faster traversal for relationship-heavy queries



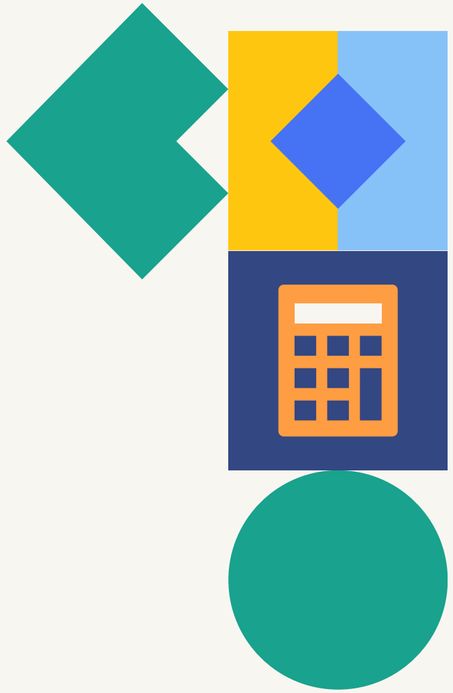
Security

Which two data sets have the same mean?

A Data Encryption: Enabled TDE for sensitive tables.

B Access Controls: Role-based access. Multi-factor authentication.

C Backup and Recovery: Certificates and keys securely stored



Conclusion

A. Centralized data warehouse built on a dimensional model.

B. Insights generated through SSRS and Tableau.

C. Performance analysis comparing Neo4j with SQL for graph-based queries

D. Future Work: Scale the warehouse for larger datasets

References

1. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Wiley.
2. Neo4j, Inc. (2024). Graph Databases for Beginners. Neo4j.
3. "Product Sales and Returns Dataset" (2024). Kaggle. Retrieved from https://www.kaggle.com.
4. Malhotra, Y. (2000). Knowledge Management and Virtual Organizations. Idea Group Publishing.
5. Angles, R., & Gutierrez, C. (2008). Survey of Graph Database Models. ACM Computing Surveys, 40(1), 1-39.
6. Tableau Software. (2024). Tableau for Business Intelligence. Tableau. Retrieved from https://www.tableau.com.
7. Microsoft Corporation. (2024). AdventureWorks Database Documentation. Retrieved from https://learn.microsoft.com.
8. Harrison, T. M., & Zmud, R. W. (1990). Information Systems Design: Theory and Methods. MIS Quarterly.
9. Fowler, M. (2002). Patterns of Enterprise Application Architecture. Addison-Wesley.
10. Sadalage, P., & Fowler, M. (2012). NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley.
11. Elmasri, R., & Navathe, S. (2015). Fundamentals of Database Systems. Pearson.
12. Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, 13(6), 377–387.
13. Kuper, G. M., & Vardi, M. Y. (1993). On the Complexity of Queries in the Graph Model. Theoretical Computer Science, 116(1), 29-50.