Gepuk

WIE3008

BUSINESS ANALYTICS & INTELLIGENCE

SoleStride

a Shoe E-Commerce company

Agenda

1. Requirement Analysis
2. Data Warehouse & Data Lake Design
3. Data Modelling
4. OLAP (Online Analytical Processing)
5. ETL Processes
6. Implementation Plan
7. Cost Estimation
8. Conclusion
9. Reference

Group Members

Aiman Fatihah (17206983)

PROJECT MANAGER

Wan Suraya (U2005345)

ETL ENGINEER

Izzat Hakeem (17204856)

DATA ARCHITECTURE

Hanan Nizam (17206018)

DATA ENGINEER

Nurul Filzah (17205112)

BUSINESS ANALYST

STRUCTURED

UNSTRUCTURED

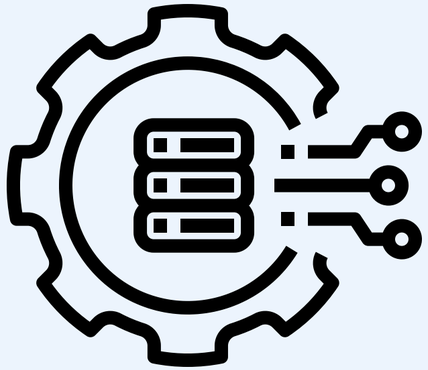
REQUIREMENT ANALYSIS

VARIOUS

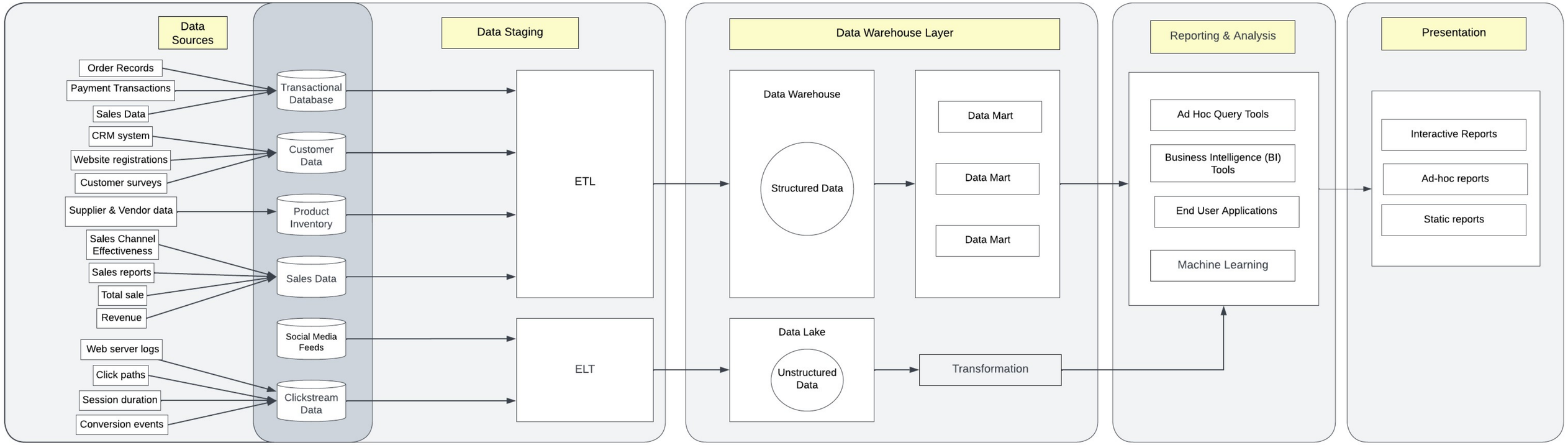
Data Source

Transactional Databases	Customer Data	Sales data	Social Media Feeds	Clickstream data

• Order Records
• Payment transactions
• Shipping and Logistics
	• CRM system
	• Website registrations
	• Customer surveys
		• Sales reports and analytics.	• Data from social media platforms
			  ◦ Twitter
			  ◦ Facebook
			  ◦ Tiktok
			  ◦ Instagram
				• Page views
				• Click paths
				• Session duration
				• Conversion events



Data Warehouse & Data Lake Design



Data Warehouse

Data Sources

Order records, payment transcations, sales data, CRM system, website registartion, customer surveys, supplier and vendor data, sales reports.

Data Staging

Extraction:
Gathers all the relevant information from the data sources.

Transformation:
Cleaning: Standardize and clean data to ensure consistency such as handling missing values and correcting typos.
Integration: Combine data from various sourcesto ensure a unified format and structure.
Aggregation: Summarize and aggregate data for better analysissuch as aggregating sales data by month or region
Validation: Validate data to ensure accuracy and reliability.

Load: Load the transformed data into the designated data warehouse

Data Mart

- Subset has been created directly from the dw.
- Utilize OLAP cubes for multidimensional analysis.
- Create cubes based on key business metrics, such as sales, customer satisfaction, and supplier performance.
- Dimensions in the cubes would align with the dimension tables in the data warehouse.

Reporting & Analysis

- Connect reporting tools (e.g., Tableau, Power BI) to the data warehouse.
- Generate reports and dashboards based on user requirements.
- Enable ad-hoc querying for business analysts and decision-makers.

Data Lake

Data Sources

Social Media Feeds: Twitter and Facebook's data, collected through their APIs and web scraping methods.

Clickstream Data: Web server logs and user interactions on the website.

Data Staging

Extraction:

- Real-time extraction of Social Media Feeds and Clickstream Data.
- Capturing brand mentions, sentiment analysis, user engagement, page views, click paths, session duration, and conversion events.

Load:

- Direct the extracted unstructured data to a staging area for initial storage and processing.
- Ensure scalability and resilience for handling large data volumes.
- Implement partitioning for efficient data retrieval during subsequent processing stages.

Transformation:

Quality check: Verify data integrity, addressing anomalies, missing values, and duplicates.

Transformation: structure and standardize data.

- Schema-on-read approaches for dynamic social media and clickstream data.

Data Lake

Storage:

- Flexible schema to accommodate diverse data types.
- Specific storage formats suitable for big data processing.

Integration: Integrate enriched unstructured data with structured data from other sources for comprehensive analysis.

Accessibility:

- Accessibility for data scientists and analysts to explore raw data.
- Collaboration and data exploration in a flexible environment.

Reporting & Analysis

Data Exploration

- Batch processing to analyze the unstructured data in the Data Lake.
- In-depth exploration of raw data to uncover patterns and trends.

Advanced Analytics: Machine learning and predictive modeling.

Integration with Reporting Tools:

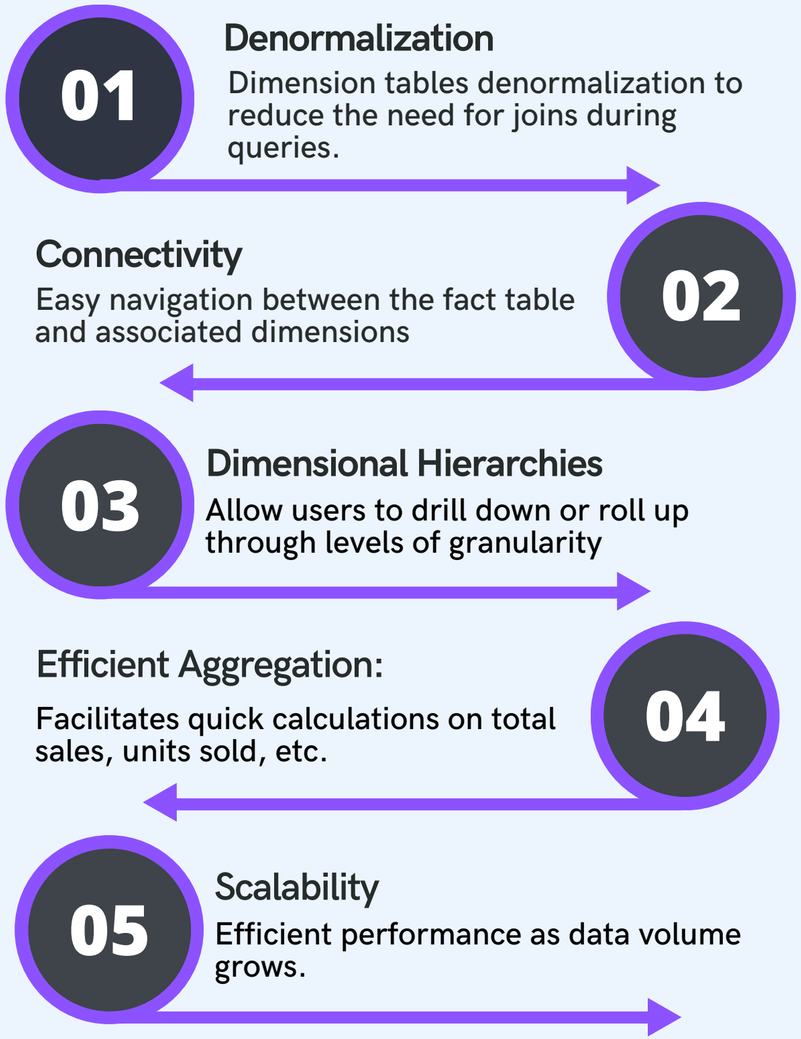
- Connect reporting tools (e.g., Tableau, Power BI) to the Data Lake.
- Enable users to create reports and dashboards based on comprehensive data sets.

Ad-hoc Querying:

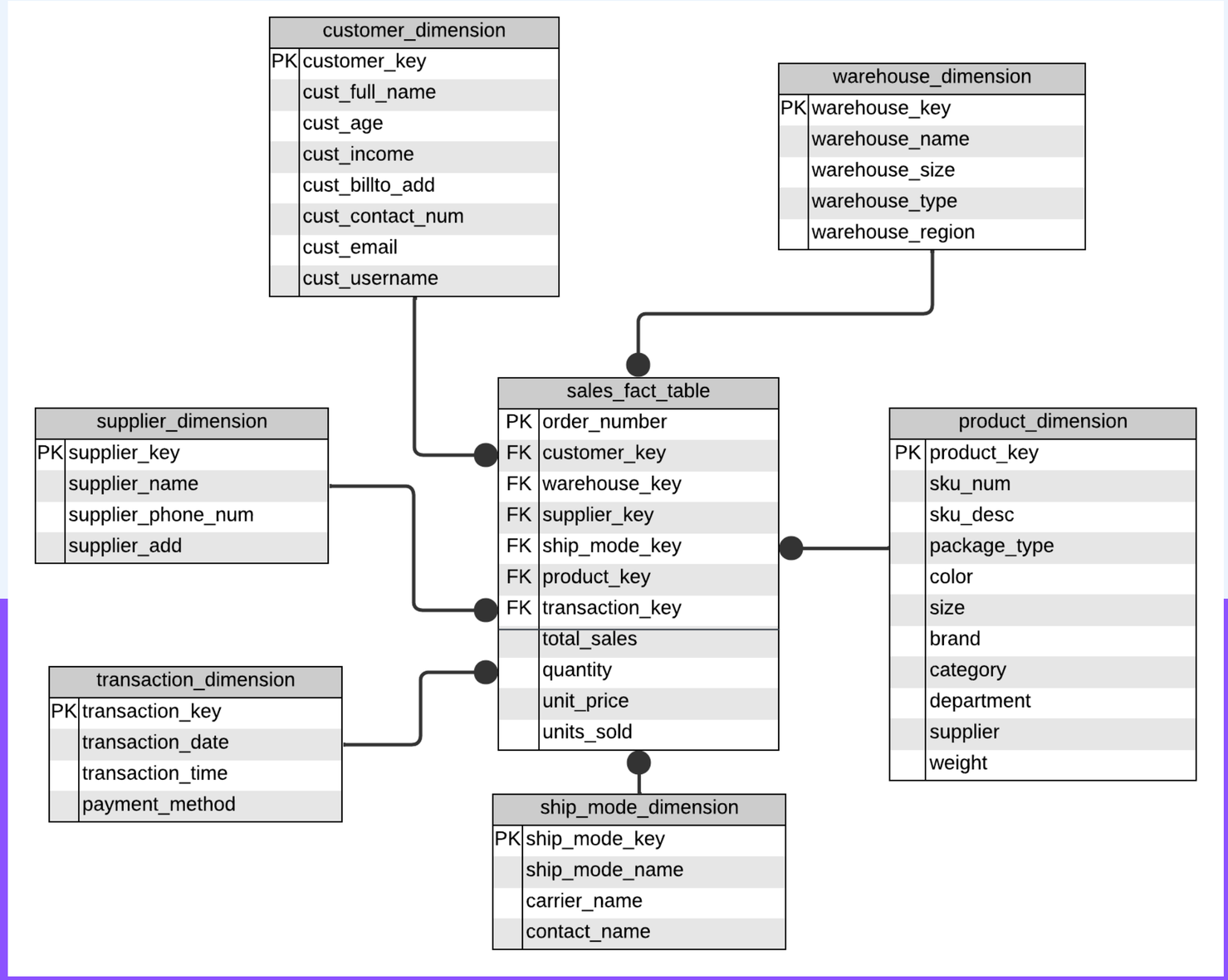
- Ad-hoc querying for data scientists and analysts to extract insights on-demand.
- Flexible exploration of both structured and unstructured data within the Data Lake.

Data Modelling
Star Schema

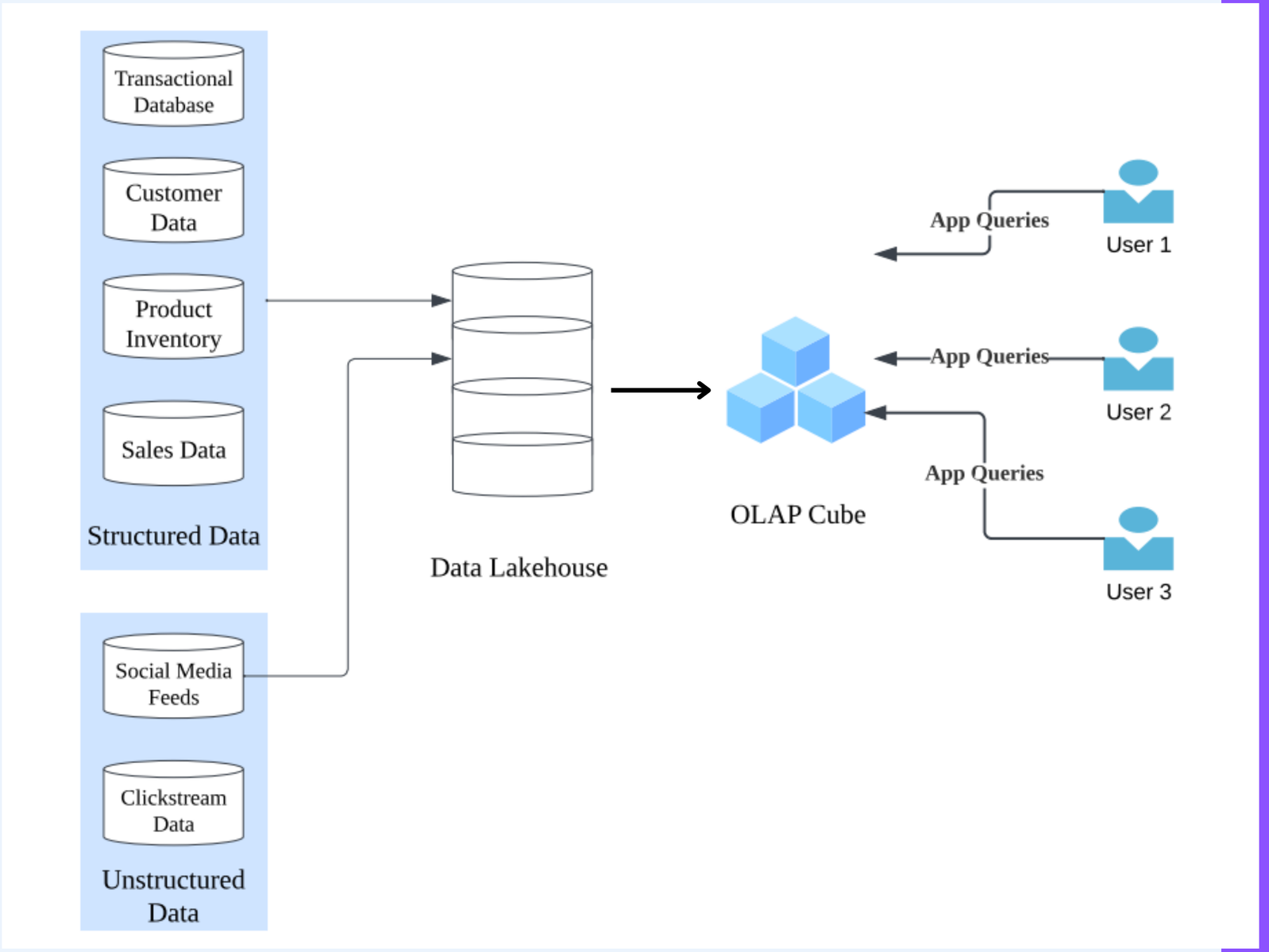
- A star schema is a type of data modeling technique used in data warehousing to represent data in a structured and intuitive way.
- Data is organized into a central fact table, surrounded by dimension tables.



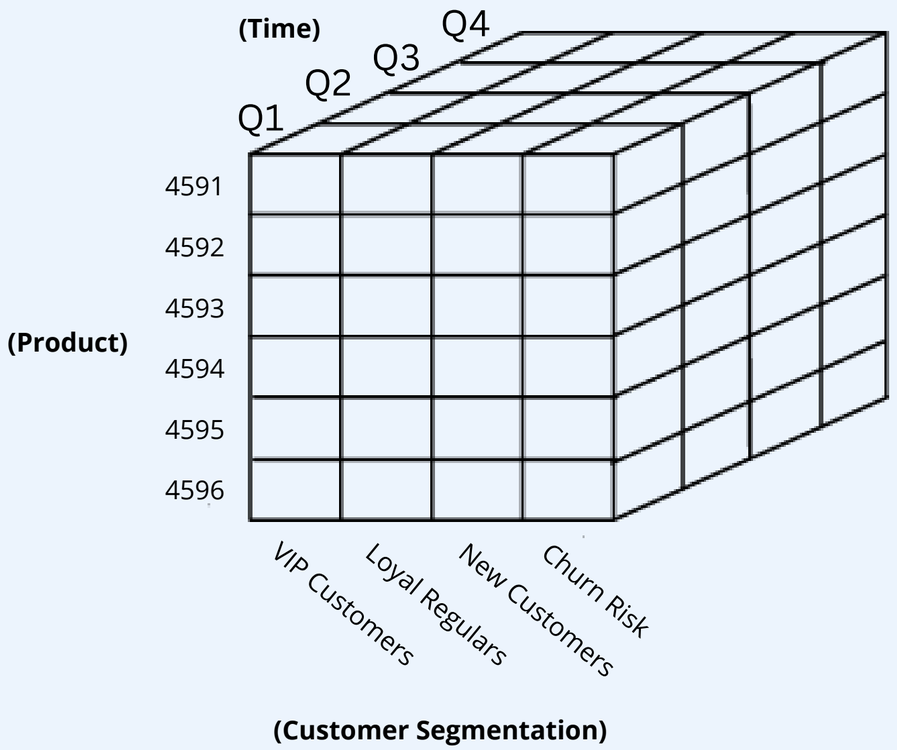
Data Modelling



OLAP Solutions



OLAP Cubes

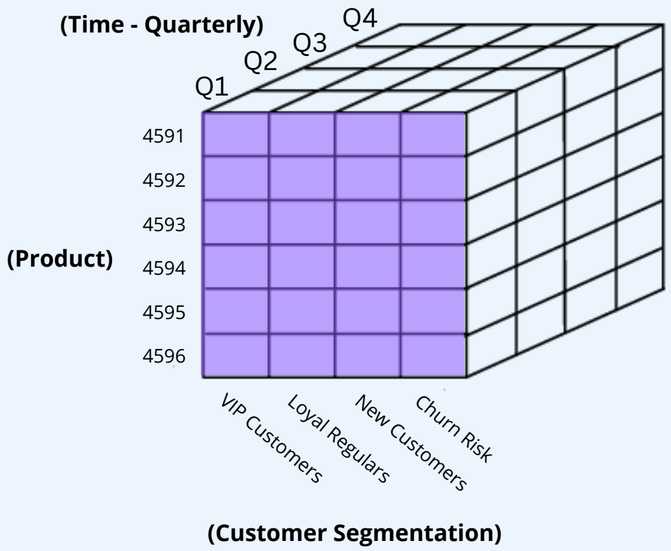


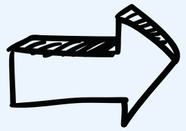
What?:

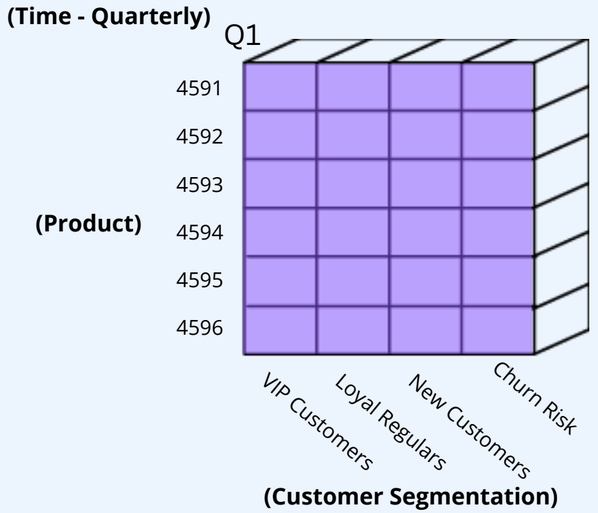
OLAP Cube aggregates total sales amount across time dimension, product dimension and customer segmentation dimension

OLAP Cubes

SLICE

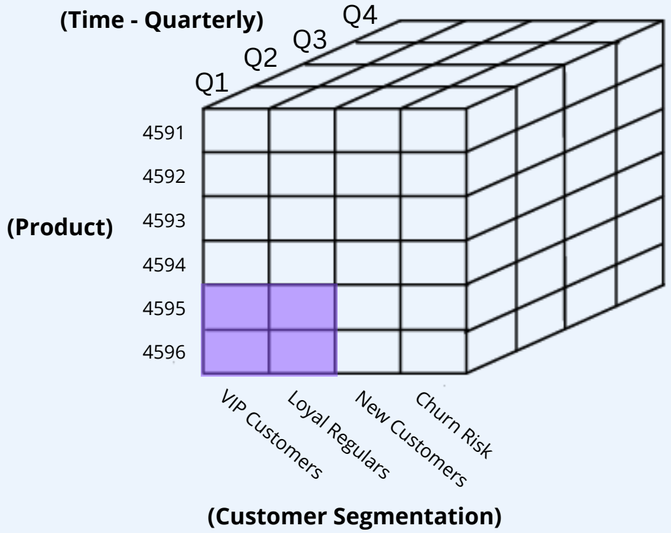


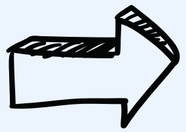


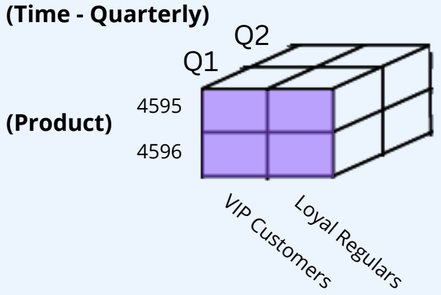


Provide 2-dimensional view from OLAP cube

DICE



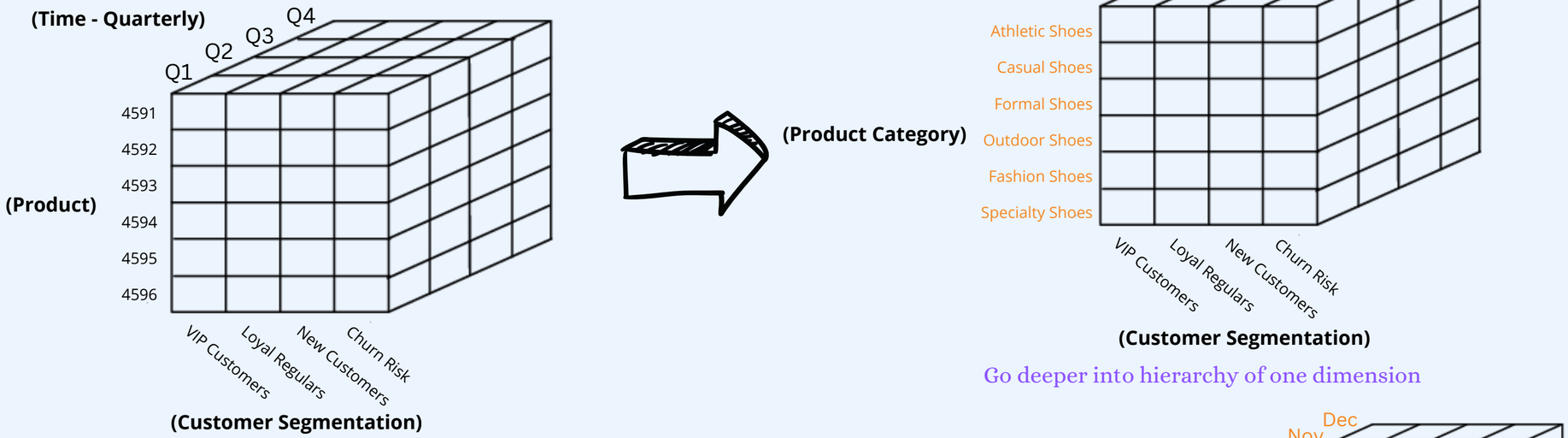




Provide smaller subcube from an OLAP cube

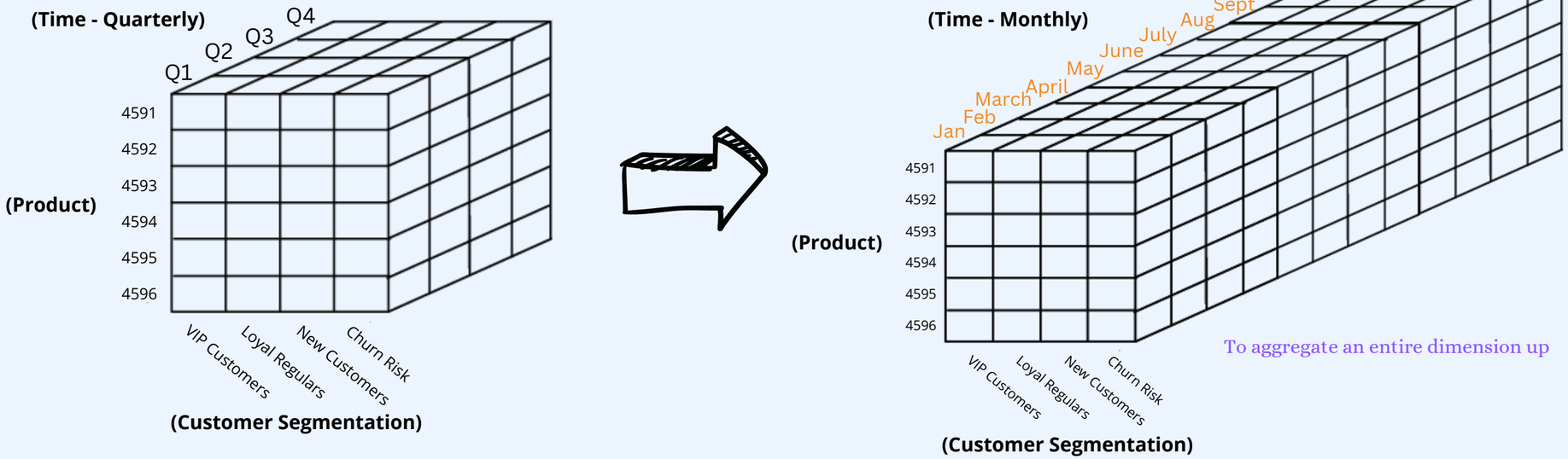
OLAP Cubes

DRILL-DOWN



Go deeper into hierarchy of one dimension

ROLL-UP



To aggregate an entire dimension up

OLAP Tools

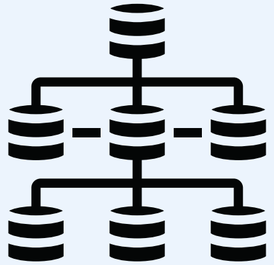
- Provides **online** analytical processing **(OLAP)** and **data mining** functionalities
- Allows users to create and manage **multidimensional** data models.
- Enables the creation of **OLAP cubes.**

- Used to build reports, dashboards and data data visualization.
- Able to integrate with **Microsoft SQL Server SSAS**.
- Supports decision making process.

ETL Processes

Data Sources

Structured Data



ETL

Unstructured Data

ELT

ETL Processes (for Structured data)

Extract

Extract data from various sources

- Transactional Database
- Customer data from CRM systems, website registrations, and surveys.
- Product inventory data.

Tools used: kafka

Transform

Tools used: talend

Modifies data to adhere to the format required by the data warehouse

- Cleaning: Eliminate errors, inconsistencies and format transformation
- Integration: Reconsiles data originating from various sources
- Aggregation: Summarizes data based on granularity/level of detail required

Load

- Load the ***transformed data*** into the data warehouse
- This process occurs on a regular schedule to keep data warehouse up-to-date

Tools used: talend

ELT Processes (for Unstructured data)

Extract

Tools used:

kafka

Extract data from various sources

- Social Media Feeds (Twitter, Facebook, Tiktok and Instagram
- Clickstream data (Page views, Click paths, Session duration, Conversion events

Load

Tools used:

APACHE nifi

- Raw data is loaded directly into the data lake
- Takes faster time than ETL

Transform

Tools used:

talend

- Raw data is transformed inside the data lake
- Involves cleaning, restructuring, and enriching the data to prepare it for analysis.

CONFIGURATION WITHIN ETL PROCESS

Real-Time Processing

Social Media feeds and Clickstream Data

Using Apache Kafka

Real-time data streaming

Configure **Kafka topics** for each data source to enable parallel processing.

Using Apache NiFi

Real-time data processing and transformation.

Use **Apache NiFi processors** to consume and process messages from Kafka topics in near real-time.

Result

Seamless integration with the Data Lake and Data Warehouse for real-time updates.

CONFIGURATION WITHIN ETL PROCESS

Batch Processing

Transactional databases, customer data, product inventory and sales data

Using Talend

- **Schedule batch jobs** to run at specific intervals → process large volumes of data efficiently
- **Implement incremental loading** strategies → update only the new or modified data since the last extraction
- **Parallel processing** capabilities → optimized batch performance

Result → **Well-monitored** batch jobs for completion and data accuracy.

Implementation Plan

Implementation Plan	January					February					March					April				May				
TASKS	W1	W2	W3	W4	W5	W1	W2	W3	W4	W5	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4	W5
Phase 1: Project Kickoff and Team Formation																								
Define roles and responsibilities																								
Define business requirements and objectives																								
Gather feedback from stakeholder																								
Phase 2: Requirement Analysis																								
Analyze existing data sources																								
Understand the data structure, quality, and contents																								
Define data sources																								
Phase 3: Data Warehouse and Data Lake Design																								
Data solution architecture design																								
Explanation of the design																								
Phase 4: Data Modeling																								
Develop data models (Star Schemas)																								

Implementation Plan	June					July				August					September					October					November					December				
TASKS	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4	W5	W1	W2	W3	W4	W5	W1	W2	W3	W4	W5	W1	W2	W3	W4	W5	W1	W2	W3	W4	W5
Phase 5: OLAP Integration																																		
Design the OLAP solution and cubes																																		
Select appropriate OLAP tools																																		
Phase 6: ETL Processes																																		
Tool selection and configuration																																		
Data solution implementation																																		
Phase 7: Deployment & Monitoring																																		
Quality assurance and testing																																		
Monitor the performance																																		
Phase 8: Documentation & Reporting																																		
Project review and reporting																																		

Challenges & Mitigations

Phase	Challenges	Mitigation Strategies
Requirement Analysis	• Incomplete or inconsistent data from various sources.	• Plan for data cleansing and normalization techniques to address inconsistencies.
Data Warehouse and Data Lake Design	• Balancing storage costs and performance requirements. • Integrating and managing both structured and unstructured data efficiently.	• Conduct thorough capacity planning to optimize storage costs. • Choose a flexible architecture that supports both structured and unstructured data.
Data Modelling	• Ensuring scalability and adaptability of the chosen model.	• Choose a data model that can evolve with changing business requirements.
OLAP	• Selecting an OLAP solution that doesn't align with business needs.	• Involve business users in the selection process to ensure the chosen OLAP solution meets their needs.
ETL Processes	• Ensuring data consistency and accuracy during ETL processes.	• Implement data validation and reconciliation checks within ETL workflows.

Cost Estimation

Resources	Cost (Monthly)	Cost (Yearly)
Software		
Talend	RM 5,497.83	RM 65,973.96
Apache Kafka	RM 1,170.05	RM 14,040.60
Apache NiFi	RM 8,928.10	RM 107,137.20
Microsot SQL Server	RM 1,287.53	RM 15,450.36
Microsoft Power BI	RM 23,471.51	RM 281,658.12
Hardware		
Storage	RM 1,127.76	RM 13,533.12
Manpower		
Project Manager	RM 9000	RM 108,000
Data Architecture	RM 11,000	RM 132,000
Data Engineer	RM 8000	RM 96,000
ETL Developer	RM 9000	RM 108,000
Business Analyst	RM 6000	RM 72,000
Total Estimation	**RM 295,711.78**	**RM 1,013,793.36**

Conclusion

Benefits	Long-term Gains
1. Unified Data Management	Centralizing data from diverse sources into a Data Warehouse and Data Lake promotes streamlined operations
2. Enhanced Decision-Making	Implementing OLAP solutions enables quick decision-making through multi-dimensional data analysis fostering a more agile and responsive organization.
3. Improved Customer Relationships	Leveraging ETL processes for customer data integration enhances the understanding of customer preferences and behaviors hence, the organization can personalize marketing campaigns, leading to increased customer satisfaction and loyalty.

References

Kagujral, P. (2023, April 22). Data Wrehouse Architecture. GeeksForGeeks.
https://www.geeksforgeeks.org/data-warehouse-architecture/

Pademkar, P. (2023, March 18). Data Wrehouse Architecture. *Educba*.
https://www.educba.com/data-warehouse-architecture/

Team, D. (2023, July 24). A list of the 18 best ETL tools and why to choose them.
https://www.datacamp.com/blog/a-list-of-the-18-best-etl-tools-and-why-to-choose-them

Hughes, A. (2020). Microsoft SQL Server. TechTarget.
https://www.techtarget.com/searchdatamanagement/definition/SQL-Server

Amazon Web Services. (n.d.). What is OLAP (Online Analytical Processing)? AWS.
https://aws.amazon.com/what-is/olap/

Keboola. (2022, March 4). Understanding OLAP Cubes - A guide for the perplexed.
https://www.keboola.com/blog/olap-cubes

Simplilearn. (2022, December 28). Top ELT tools for modern data stack in 2023: Simplilearn. Simplilearn.com.
https://www.simplilearn.com/elt-tools-article

Gepuk

Thank You