Business Intelligence

Assignment 1: Data Warehousing Basics and Modeling

Martí Paulet López
TU Wien
e12305831@student.tuwien.ac.at

1 Intro

The organization which will be described is Amazon. Amazon is a multinational technology and e-commerce B2C company founded in 1994 by Jeff Bezos. Its mission is to continually raise the bar of the customer experience by using the internet and technology to help consumers find, discover, and buy anything, and empower businesses and content creators to maximize their success.

2 OLTP

2.1 Order Management (OM)

2.1.1 List and describe at least three operations that it performs (bullet point list).

- Inventory management: Monitoring stock levels, restocking and managing product availability.
- Shipping: Receiving, processing, and fulfilling the orders.
- Logistics: Package tracking, delivery coordination and management of package returns.
- Order processing: Process incoming orders, process payments, and confirm orders.

2.1.2 Characterize the data that the system stores (bullet point list).

- Order details: order identifier, customer information, product details.
- Inventory data: stock levels, product attributes and characteristics.
- Shipping data: tracking numbers, real-time monitoring.
- Logistics data: package details, dispatching and delivery times.

2.1.3 List 5 typical questions that can be answered using that data.

- What is the remaining quantity of a specific product that is still in stock?
- What is the expected delivery time for customers' orders?
- What products are included in an order?
- What is the status of a particular order?
- What is the average amount of time it takes to process an order?

2.1.4 List 5 strategically important questions that cannot be answered based (only) on the data that this system holds.

- What are the current trends in customer preferences and society that are impacting our product inventory?
- In what ways do customer satisfaction differ between payment methods?
- What steps can we take to improve the supply chain and decrease delivery times?
- How might an error in an order affect the customer?
- What are the common characteristics of big-purchase customers?

2.2 Customer Relationship Management (CRM)

2.2.1 List and describe at least three operations that it performs (bullet point list).

- Customer data management: Store and update customer purchase information.
- Customer service and interaction tracking: Managing customer inquiries and interactions.
- Marketing and sales campaign management: Monitor customer responses to marketing activities.

2.2.2 Characterize the data that the system stores (bullet point list).

- Customer data: contact information, purchase history, preferences.
- Marketing campaign's data: email open rates, click-through rates, increases in demand.

- Geographic data: location history, bestselling areas, areas not actively present.

2.2.3 List 5 typical questions that can be answered using that data.

- Is our customer segmentation strategy effective in tailoring marketing campaigns?
- Which customers are the most valuable to us in terms of their lifetime value?
- What is the response rate to our latest email marketing campaign?
- What are the most common issues that customers reach out to us about?
- Which areas are seeing the highest sales and what are the reasons behind it?

2.2.4 List 5 strategically important questions that cannot be answered based (only) on the data that this system holds.

- Which marketing channels provide the best Return On Investment (ROI) in the long term?
- How can we personalize the shopping experience for individual customers?
- What is the correlation between customer retention and satisfaction?
- Why is there a higher level of activity in certain cities compared to others when we use the same marketing strategies?
- What support can be given to customers who are dissatisfied?

2.3 Warehouse Management (WM)

2.3.1 List and describe at least three operations that it performs (bullet point list).

- Inventory control: Monitoring the movement and storage of products in the warehouse.
- Order packing and picking: Preparing orders for having an efficient shipment.
- Quality control: Ensuring that products meet quality standards before shipping.

2.3.2 Characterize the data that the system stores (bullet point list).

- Warehouse layout and inventory locations.
- Order fulfillment data: packing slips, packing lists.
- Quality control checks and results.

2.3.3 List 5 typical questions that can be answered using that data.

- Where is a specific product located in the warehouse?
- What is the status of a particular order in the packing process.
- Are there any control issues with a batch of products?
- Is it sufficient to continue with the current methodology based on the quality control results?
- Can the product locations be arranged in a way that makes the packaging process easier to execute?

2.3.4 List 5 strategically important questions that cannot be answered based (only) on the data that this system holds.

- What can be done to optimize the warehouse layout to reduce picking times?
- Which packing strategies are the most efficient for different types of products?
- What is the impact of quality control measures on customer satisfaction?
- What steps can we take to detect any potential errors or malfunctions in quality control?
- Are the warehouse's workers and materials sufficient to ensure an efficient packing and shipping process?

3 Data Warehouse

3.1 Provide 5 examples for important strategic questions that will be answered using the data in the DTW.

- Which product categories are the most popular among customers, and how can we make sure inventory and marketing are optimized for them?
- How do customer preferences change over time, and how can we use this data to forecast demand and give better product recommendations?
- Can we improve our logistics operations to meet customer expectations?
- How effective are our marketing campaigns about sales and customer engagement, and how can we improve our marketing strategies?
- What is the relation between customer reviews and product sales, and how can we use this data to improve product quality and customer feedback?

3.2 Define 5 key subject areas. Then pick one of these areas and describe the data that will be stored on it in the DWH. List the source systems that will provide the various pieces of data on the subject area.

- Sales and revenue.
- Customer preferences.
- Inventory and supply chain management.
- Marketing effectiveness.
- Product quality.

The data warehouse for Amazon's "Customer Preferences" subject area would store the following types of data:

- Customer profiles: Information about each customer, including contact details, demographics, and registration history.
- Purchase history: Records of all past purchases made by each customer, including order details, product descriptions, quantities, and prices.
- Product ratings and reviews: Customer-submitted product ratings, reviews, and feedback on product quality and satisfaction.
- Wishlist and cart data: Information about products added to wishlists and shopping carts, providing insights into customer preferences and purchase intents.
- Browsing and clickstream data: Records of customer website interactions, including product views, time spent on pages, and clicks, helping identify interests and behaviors.

All this data would be sourced from various operational systems within Amazon, including:

- Customer Relationship Management System: Customer profile and contact information, along with registration and demographic data.
- Order Management System: Data on purchase history, order details, and transaction records.
- Product Reviews and ratings Database: Customer-submitted product reviews, ratings, and feedback.
- Website Clickstream Data: Website interactions and user behaviors as customer browse the site.

3.3 Propose an architecture and development approach for the DWH project. Motivate your choice and highlight the advantages and drawbacks of your proposal.

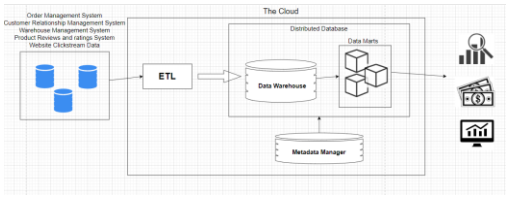

Figure 1: Design of a Cloud-based data warehouse for Amazon.

The motivation for designing a cloud-based data architecture for Amazon lies in its need for scalability, flexibility, and cost-efficiency to manage the diverse data generated by its different operational systems.

Using agile development, like other software projects, is more advantageous than a big leap, as it allows for multiple fast iterations to adapt to changing requirements.

Advantages:

- Scalability: On-demand scalability, the system can easily scale up or down based on data volume and processing requirements.
- Flexibility: (In terms of data integration) It can accommodate structured and unstructured data, streaming data, and various data sources.
- Integration: Cloud providers offer a range of tools and services for data integration, including ETL capabilities.
- Cost-Efficiency: Pay-as-you-go pricing can help manage cost effectively.

Drawbacks:

- Data transfer costs: Costs associated with data transfer to and from the cloud.
- Latency: Real-time data processing may face latency due to data transfer to the cloud.
- Complexity: Managing a cloud based DWH requires expertise in cloud services, which may necessitate additional training or hiring.

3.4 List 5 potential challenges that you foresee in the development process (with respect to data governance, data quality, ETL procedures, value delivery etc.).

- The load times of ETL depend on other systems, as bandwidth can be limited, and external factors may not always be available.
- Some of the OLTP systems are based on user input from various employees and customers i.e., marketing

campaigns, customer reviews, customer interactions through the online services, etc.

- SQL limits the accessibility of data to people who don't understand query languages and the database architecture.
- Ensuring data quality and consistency across all source systems is a common challenge. Data may be incomplete, inaccurate, or inconsistent, which requires effective data cleansing and transformation strategies to maintain the integrity of data in the data warehouse.
- Ensuring data privacy regulations and compliance when working with different source systems.

4 Data Mart

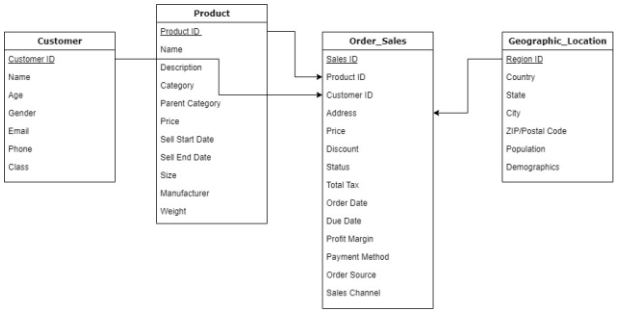

Figure 2: Star Schema design of the Sales and Revenue Data Mart. Underlined attributes indicate primary keys.

This data mart could give information about:

- **Sales Trends and Patterns:** Monthly, quarterly, and yearly sales trends, comparison of sales between different regions or customer segments, peak sales months, etc.
- **Geographic Expansion Strategy:** Analyzing the whole impact of geographic location on sales, decision-making for expanding or focusing on some specific locations.
- **Customer Segmentation:** Analyzing customers preferences and their impacts on some regions, segmenting customers based on their purchase history.

5 Data Lake

5.1 Whether (or not) a data lake would be a cost-effective alternative to the data warehouse solution you proposed.

The storage of raw and unstructured data in a data lake can result in higher storage costs due to the volume and diversity of data. Therefore, the cost-effectiveness of the data lake comes from its role as a complementary solution, not a replacement.

5.2 What use cases (business questions or entirely new business opportunities) a data lake could address that the proposed DWH cannot handle and what benefits you foresee.

A data lake would enable the analysis of unstructured data, such as customer reviews, providing valuable insights into customer sentiments and brand perception. It would support real-time data processing, facilitating applications like real-time inventory tracking and personalized recommendations. Additionally, the data lake's ability to handle diverse data sources would support Amazon's scalability, enabling the company to explore emerging markets and expand product lines more effectively.

5.3 How a data lake would fit into your proposed architecture.

- Data lakes, such as Amazon S3, would serve as a repository for raw and unstructured data, integrated into the data integration and transformation layer.
- Accessing the data lake for analysis is available to data analysts and scientists, who can then use these insights to enhance their analytics in the data warehouse.

5.4 Whether the introduction of a data lake would require any organizational changes.

Data governance, quality and privacy must be carefully considered when dealing with data lakes.

Under GDPR, users must be able to delete all their data on request. The challenge with data lakes lies in the potential unstructured and disorganized data scattered over a disorganized file system.

Considering this fact, Amazon would need to strengthen its data governance practices for both the data lake and the data warehouse. In addition, to fully use the potential of a data lake, Amazon could invest in data engineering and data science capabilities.