

Business Intelligence

Assignment 1: Data Platforms – Group 44

Jakob Pinterits

1327264

TU Wien

e1327264@student.tuwien.ac.at

Gent Rexha

11832486

TU Wien

e11832486@student.tuwien.ac.at

Christoph Wimmer

01130693

TU Wien

e1130693@student.tuwien.ac.at

1 Intro

The organization which will be described is Siemens AG Austria. Its mission is to sell and create technology over multiple different sectors i.e: Building Technology, Drive Technology, Energy, Healthineers, Industrial Automation, Mobility, Financing, Services, Software, Consumer Products. It acts as an independent firm inside Austria but is still liable to the Headquarters in Germany, from where all the big decisions come from.

2 OLTP

2.1 Customer Relationship Management (CRM)

2.1.1 Describe the operations that it performs (bullet point list).

- Campaigns
- Sale Orders
- Deliveries
- Invoices
- Leads, etc.

2.1.2 Characterize the data that the system stores.

- Financial data
- Customer data
- Geographic data
- Marketing data

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

- How much are customers buying?
- How long are deliveries taking?
- Which regions are buying more than others?
- Which leads are turning into opportunities?
- Which customers are getting which quotes?

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

- What are the common characteristics of big purchasing customers?
- Which customers are having delivery problems and what's their customers support experience?
- Which customers are getting visited by sales people?

- Which part of the inventory is getting sold and not stocked enough, and which is not getting sold at all, i.e. dynamic inventory management?
- Which sub-segment, relative to the others, is growing the most?

2.2 Customer Relationship Management (CRM)

2.2.1 Describe the operations that it performs (bullet point list).

- Customer Services - Capture and maintain customer relationships, facilitate the use of customer experience.
- Human Resources - Maintain a complete employee database and to optimal utilise of all employees.
- Accounting - Automate any financial operations while ensuring regulatory compliance and gaining real-time insight into overall performance.
- Production - Helps and planning and optimising the manufacturing capacity and material resources.
- Distribution - Control warehouse processes and manage movements in the warehouse and respond faster to challenges and changes in supply and demand.

2.2.2 Characterize the data that the system stores.

- Product data
- Customer services data
- Production Data
- Human Resources data
- Finance and Accounting data

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

- Who is corresponding with which customer?
- Which employee has a certain degree?
- What is the real-time performance of a certain sub-segment?
- How can the manufacturing capacity be optimized?
- How many receivables are outstanding?

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

- How much are customers buying?
- How long take certain deliveries?
- Which regions are buying more than others?

- Which leads are turning into opportunities?
- Which customers are getting which quotes?

2.3 Configurable Parameters & Formulas (CPF)

2.3.1 Describe the operations that it performs.

- Adapt pricing based on customer, sales quantity & rebates
- Automate pricing procedures

2.3.2 Characterize the data that the system stores.

- Financial data
- Conditions
- Parameters
- Formulas

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

- Find out relevant pricing parameters?
- Which products are getting the biggest discounts?
- How many total discounted products are there?
- Which type of sub-segment products are getting the biggest discount?
- Are there any products without discounts being sold?

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

- What is the optimal price for maximizing profits?
- Which products sell best/worst?
- Which customers are buying only discounted products?
- What product should be developed next?
- Is the discount dependent on the purchasing power of the customer?

3 Data Warehouse

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

- Which customers are not getting visited but still buying our products?
- How can discounts be more fine-tuned depending on the salesperson and customer?
- Do visit frequency and length depend on the type of customer (purchase amount, customer class, customer sub-segment)?
- What impact do discounts have on sales numbers?
- Which products require a lot of customer support?

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.

Key subject areas:

- Visits
- Sales
- Customer Information
- Discount
- Segment growth

Segment growth is based on customer data, sales data, marketing data and employee data from which a year to year growth is calculated based on the amount of sales, amount of money invested into marketing, customer growth and customer service experience, and last but not least amount of money spent on employee and services.

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

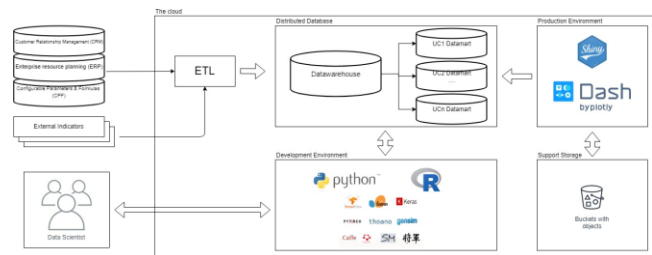


Figure 1: Design of a Data Warehouse based on Siemens AG Austria

- Gain a deep understanding of the business data and develop a data model. Having a common mental model amongst developers is crucial for success.
- Employ agile development. Like other software projects many fast iterations are preferable over one big leap so one can adapt to changing needs.
- Adopt and potentially adapt a recognized reference model. These have been tested in practice and are known to work.

Advantages:

- Having a common mental model amongst developers ensures everybody is working towards a common goal.
- Like any other software projects many fast iterations are preferable over one big leap so one can adapt to changing needs.
- Many models have been used and tested in the industry. Using one of them ensures the model works in practice and

decreases the likelihood of hidden flaws that only become noticeable once deployed.

Disadvantages:

- Agile development incurs more overhead than using waterfall.
- The industry standard models may not be applicable to this particular use case.

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

- Customers firms can be merged with other firms and can be subsided into smaller ones where keeping track of every firm and keeping a linear timescale for each company can be hard because not all the merges are straightforward.
- ETL load times are dependent on other systems where bandwidth can sometimes be limited, external factors are not always available.
- Data has to still to be queried through SQL to the distributed database, which limits the accessibility of data to persons who have an understanding of query languages and the database architecture.
- Some of the OLTP systems are based on user input from various employees i.e. visits and marketing campaigns/events. This results in a very wide range of data quality where depending on the sales person a visit might be described as “Coffee with the customer” up to a one-page report with minutes and everything discussed during the meeting.
- Extracting data from production systems puts additional strain on them, potentially requiring upgrades.

4 Data Mart

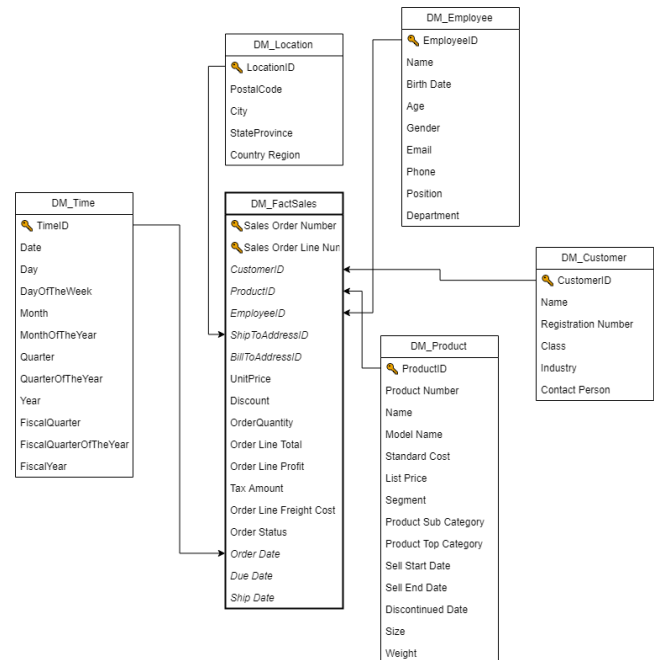


Figure 2: Data mart as a star schema including the fact table (DM_FactSales) and five dimension tables (DM_Time, DM_Location, DM_Customer, DM_Employee, DM_Product), where the key symbol indicates the primary key and italic indicates foreign keys

5 Data Lake

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

It is more cost effective to use a data warehouse because we have a well-defined use case for the data. Data lakes often degenerate into “dumps” where a lot of gathered data is put without ever being analyzed. Warehouses on the other hand can be specifically tailored to the task at hand, making them more efficient.

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.

Data lakes are well suited for future, not yet foreseeable tasks. Future technologies may be able to extract value from data that appears immaterial at this point in time. Data lakes can be used to store any information available, rather than just the one used right now. This will allow future technologies to be used on current data, rather than having to start data collection from scratch.

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

- Make both systems run independently, receiving data from the same (or different) sources.
- Feed all data into the lake first, then periodically run tasks to extract useful information into the warehouse.

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

Data lakes require careful consideration regarding data governance. Under GDPR users have to be able to delete all of their data on request. This is challenging with data lakes as the data is potentially unstructured and scattered over a disorganized file system. Data warehouses don't suffer from this problem due to imposing strict schemas.