IS5 – DW Assignments

Group 18

Roope Halmineva, Davide Lagano, Anand Gankhuyag, Alketa Bardhoshi, Catherine Wandozereho

Assignment 1: Dimensional Modelling

Task 1: Star-Join Schemas

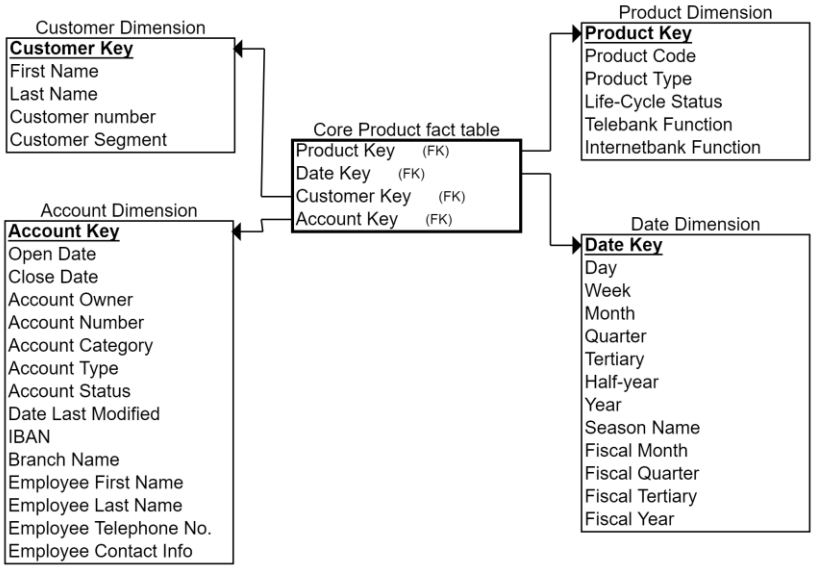


Figure 1: Core Product Fact Table

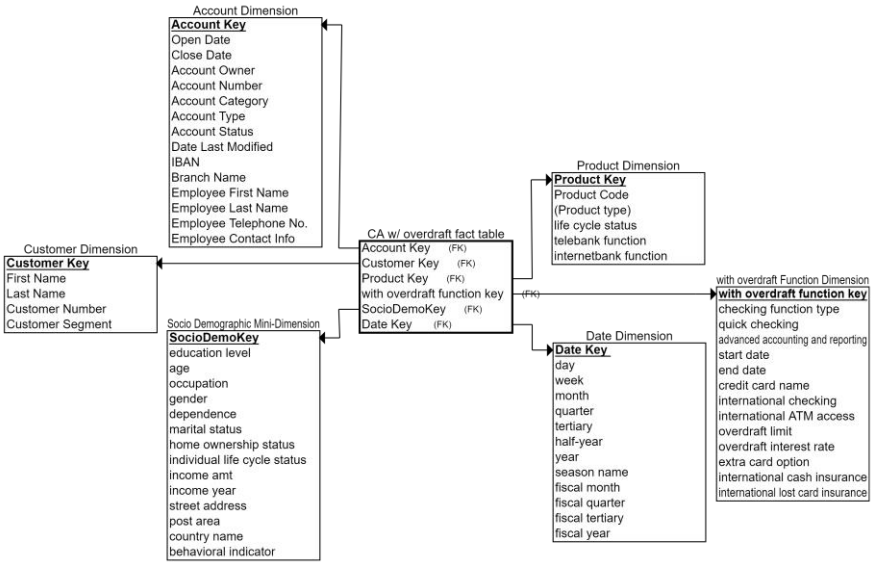


Figure 2: Credit Card Configuration Fact Table

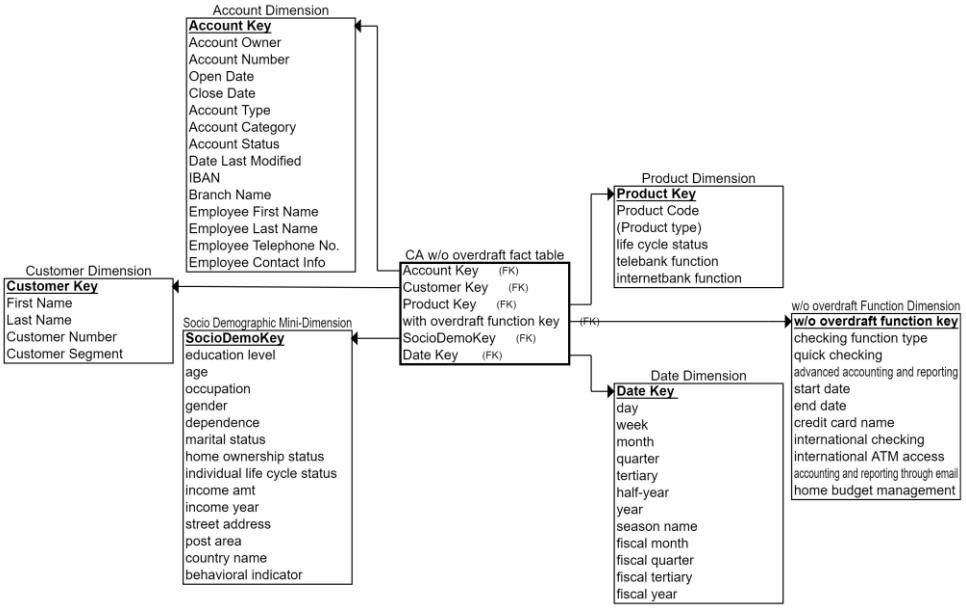


Figure 3: Debit Card Configuration Fact Table

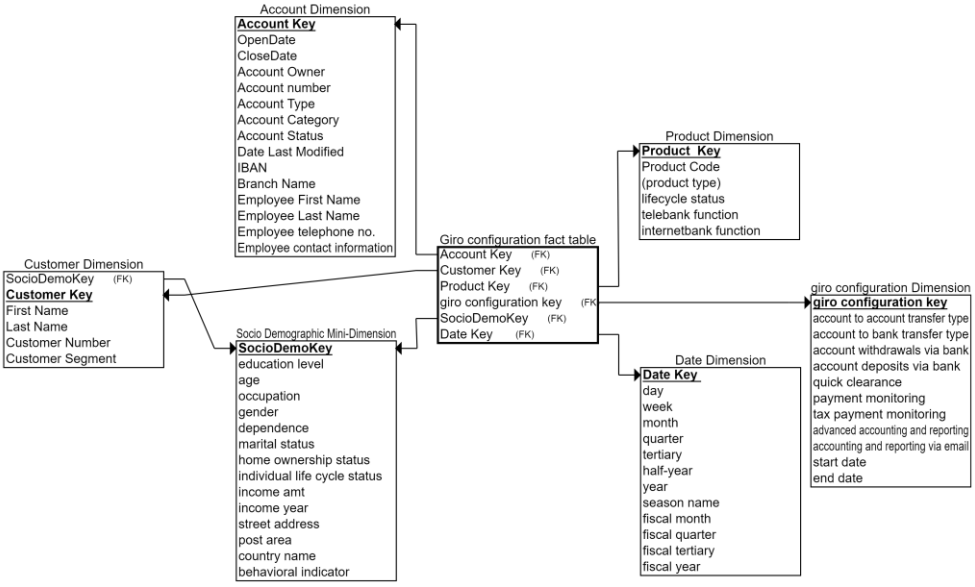


Figure 4: Giro Configuration Fact Table

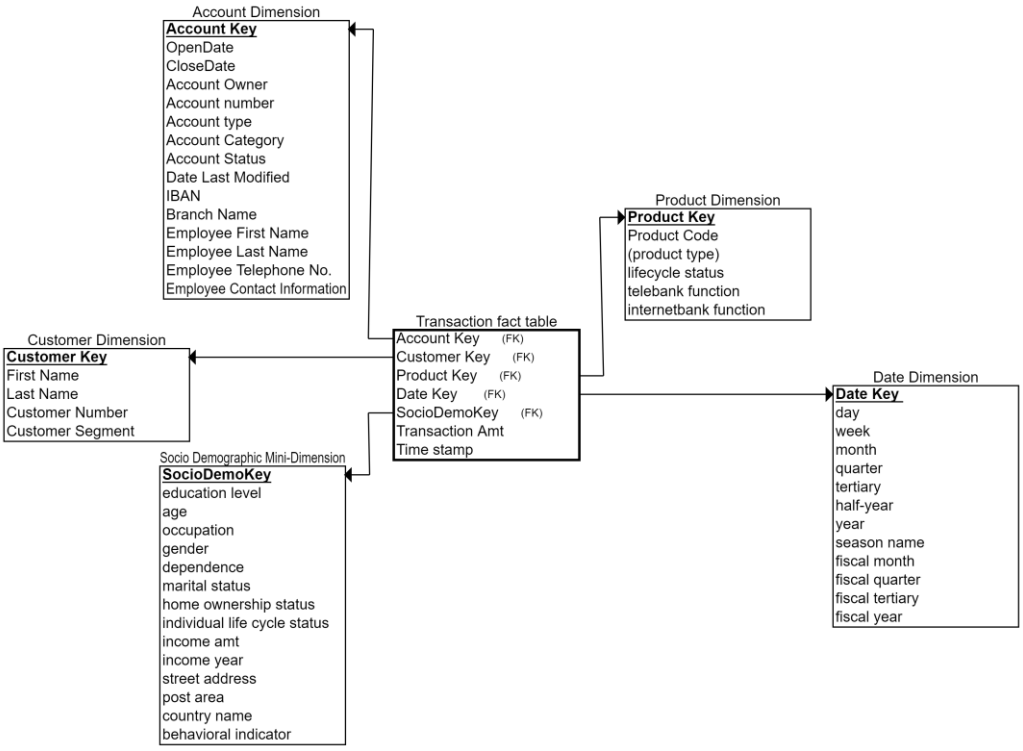


Figure 5: Transaction Fact Table

Task 2: Bus Matrix

Dimensions / Business processes	Date	Product	Account	Customer	Socio Demographic	Giro Functions	Current Account w/ overdraft functions	Current account w/o overdraft functions
Giro	X	X	X	X	X	X		
Current account w/overdraft	X	X	X	X	X		X	
Current account w/o overdraft	X	X	X	X	X			X
Transaction	X	X	X	X	X			
Core Product	X	X	X	X				

Figure 6: Bus Matrix

Task 3: Hierarchies

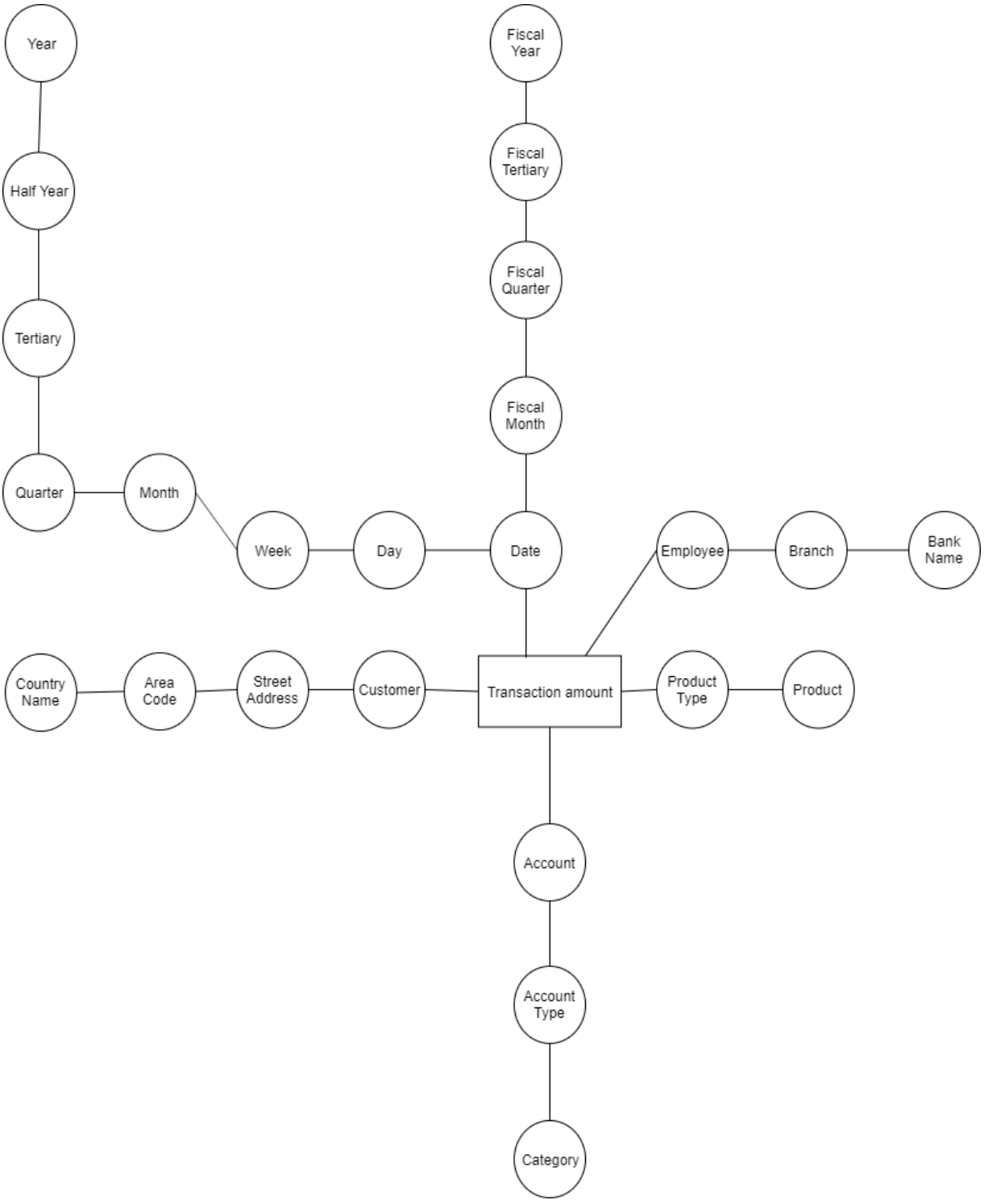


Figure 6: Hierarchies

Task 4: How to Manage Fast Searches on Large Amounts of Data

In order to manage fast searches on large amounts of data we would use OLAP cubes to make aggregations towards dimensional hierarchies. This would mean that 'slicing and dicing' data by product type, by month, or by other dimensions would be much faster and more efficient as the tables of information would be pre-aggregated.

Task 5: How to Manage the Issue that the Product Portfolio Contains Dissimilar Products

Since the product portfolio contains dissimilar products, we designed a core product star schema that has the common attributes shared by all products, and separate configuration schemas for each product type with their unique attributes. Decision makers will be able to compare single product types with each other and can compare the differences in demographics of the customers that purchase the different types of products by drilling across. It also allows for comparison between different configurations of single product types.

Task 6: How to Manage Additional Information on Customers Bought from an External Vendor

If additional information on customers were to be bought from an external vendor, it would be best to implement an outrigger or mini dimension depending on the type of information that is to be bought. Outriggers should be used sparingly since there is the risk of snowflaking, which is then hard for the business decision-makers to understand and navigate, therefore a mini dimension would be best for each external source, connecting to the customer dimension.

Task 7: How to Handle That the Bank Cannot Link Activities on An Account To a Certain Individual Customer

Since the bank cannot always link activities on an account to a certain individual customer, the data warehouse will have drill across capabilities to check which customer carried out which activity. When there is a shared account the customer ID will show null, however for non-shared accounts it will be possible to link the customer to individual activities. If the customer ID shows NULL for shared account activities, however, it cannot be used as a primary key.

Task 8: Strategy for Managing Changing Dimensions

In order to manage changes in source data over time, we would use the slowly changing dimension Type 2: add row. This means creating a new additional dimension record using a new value of the surrogate key, adding a new row in the dimension table. We would use type 2 for all dimensions and attributes in order to capture all historical data in the data warehouse. The trade-off of using Type 2 for all dimensions and attributes is that there is the risk of accumulating very large dimension tables, which would then mean longer query processing times, however since it is impossible to predict future data needs, it makes sense to gather all historical data in case it becomes advantageous to decision-making.

Task 9: What Difference Would Our Model Have If the Company Identified Transactions by Transaction ID?

If our model had a transaction ID, it would be what is called a degenerate dimension. This means that the transaction ID number would be a generated number as a fact in the fact table, without a corresponding transaction dimension. The difference this would make in our model will be seen as an extra row in the transaction fact table, by which it would be possible to uniquely identify a single transaction.

Assignment 2: DW Architecture

1. **Why is it important, according to Kimball & Ross, to separate back room ETL work from the front room presentation area in your data warehouse architecture? (Minimum 100 words)**

The back room gathers information from different sources, ETL processes and data sources while the front room is the way how the data is presented and accessed (Reporting, Dashboard, OLAP). Some of the main operations involved in the back room is including acquiring data information, the extraction of the data in the proper format and after that is delivered in the front room. This architecture is a great way to separate the task from the reporting layer in ETL's complex logic. [1]

The separation of the back room ETL work and the front room is important according to Kimball and Ross because of:

1) Providing detailed security at a row, column, or applications level

2)Building query performance-enhancing indexes and aggregations

3)Providing continuous up-time under service-level agreements

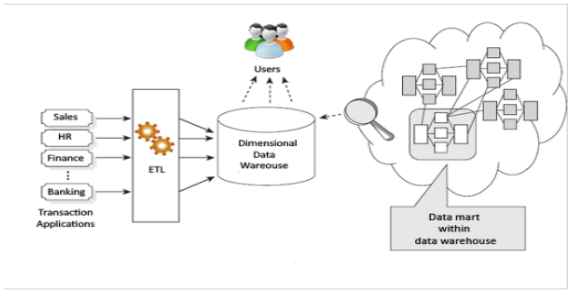
4) Guaranteeing that all data sets are consistent with each other

2. **What are the benefits and drawbacks of the Hub-and-Spoke CIF (Inmon) and the Kimball data warehouse architectures? Support you answer with a diagram. (Minimum 100 words)**

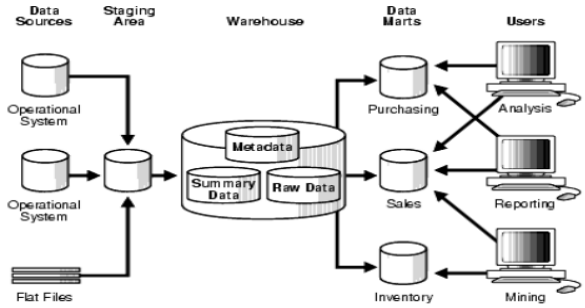
Hub and Spoke architecture is a top down approach because it shows all the complete view of the enterprise data. Data is stored in 3NF to capture historic data changes, and timestamps are added to each table column. Reporting is not supported directly and requires dimensional data marts for it.

Kimball architecture is a bottom up approach because is based on building reporting data marts sequentially based on the business priorities. The data is stored in a star or snowflake structure where large fact tables representing numerical measures and counts are presented in 3NF and the smaller ones in 2NF. Data historic changes are tracked toward changing dimensions.

Kimball approach is less complicated, and it is the preferred one due to the easy and fast implementation and due to rapid performance to the complex queries. It supports iterative agile developments and it is the optimal choice for most of the organizations. Despite the advantage that Inmon’s approach represents a complete view of the enterprise data model, it's more difficult to achieve because it is more expensive, and it requires a long-term commitment. [2]



Kimball architecture



Inmon architecture

3. What are the key organizational factors in data warehouse architecture selection? Describe how these factors influence data warehouse architecture selection. Which hypothesis found strong support in Ariyachandra & Watson (2010)? (Minimum 150 words)

The selection of the data warehouse is based in some key factors which are:

Information Inderdependence - so organizations with higher interdependence are more likely to choose an EDW rather IDM architecture

Urgency – Organization with higher urgency are more likely to choose IDM / DBA rather than EDW

Strategic view –organizations that implement DW as a short term rather than strategic infrastructure are more likely to choose IDM than EDW

Task routiness – organizations with lower task routiness are more likely to choose EDW rather than FED architecture

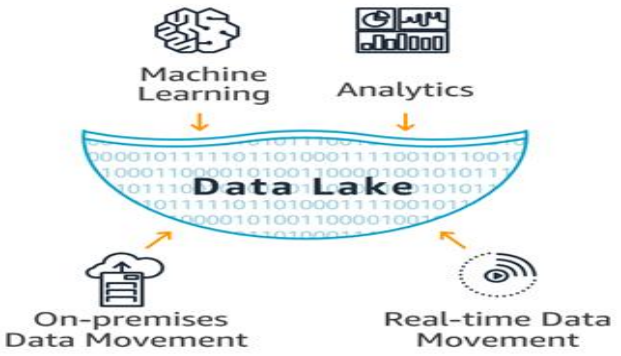
Resource Constrains – organizations with lower resources available are more likely to choose IDM/DBA rather than EDW

Perceived ability of IT staff – organizations with low perceived ability IT staff are likely to choose IDM/ DBA rather than EDW

Sponsorship level – organizations with a higher sponsorship level are likely to choose EDW rather than IDM/DBA

4. **What is a data lake, what is its function, and what are its potential benefits to an organization? If possible, support you answer with a diagram. (Minimum 150 words)**

According to Wikipedia, a Data lake is a system or repository of data stored in its natural/raw format, usually object blobs or files. A data lake is usually a single store of all enterprise data where data is transformed , cleaned and manipulated .



A data lake's functions include extending access to more users, and advanced analytics and data science. The benefits of the data lake are low cost techologies and the improvement of archival , refinement, it involves increased operational efficiencies and the ability to analyze data without having to move your data to a separate analytics system.

Elimination of data silos – supports data use and sharing

Reduce cost of IT – manage and analyze a huge amount of data in a data lake envirnoment

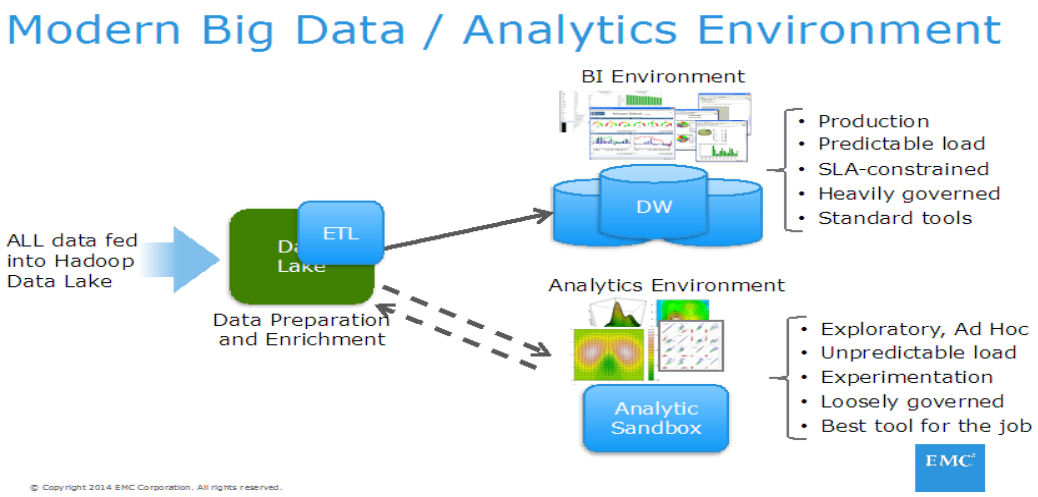
Provide a scalable, flexible, and shared storage- supports BI, analytics

Store all data- it can store a huge amount of the data which can be inserted to data warehouse

5. **How can a data warehouse system make use of a data lake? Support you answer with a diagram. (Minimum 100 words)**

Due to too much data created every day, data warehouse was overwhelmed. This made for the requirement for the data lake to aid in storage repository that holds a vast amount of raw data in their state. Data lake technologies can scale to massive volumes of data and combining datasets as easy with data stored in its state. The data sharing allows patterns to emerge, providing a launching point for data warehousing, data marts, and a wide range of analytics capabilities. The building of a data lake within a Data Warehouse can easily the cost effectiveness in loading, transforming and analyzing unlimited amounts of structured and semi structured data.

A data lake can also act as the data source for a data warehouse. Here the raw data is ingested into the data lake and then transformed into a structured queryable format. This transformation uses an ETL (extract-load-transform) pipeline, where the data is ingested and transformed. Source data that is already relational may go directly into the data warehouse, using an ETL process, skipping the data lake.

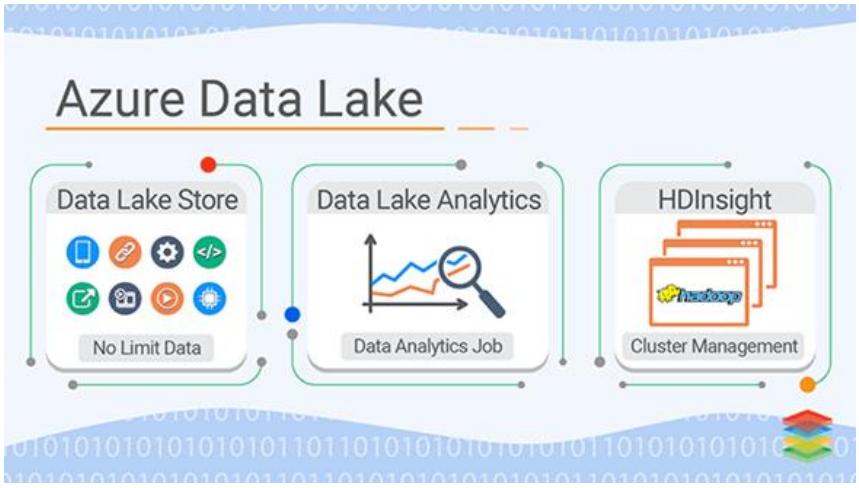


The figure above illustrates that the data lake sits in front of the data warehouse to provide a data repository. Panoply is a cloud-based data warehouse which integrates with S3 data lakes and many other data sources. Panoply allows you to pull large volumes of data from a cloud-based data lake like S3, without having an ETL process in place.

6. **Give an example of a data lake product and describe it briefly. If possible, support you answer with a diagram or picture. (Minimum 150 words)**

Data lake is a repository for raw data in untransformed way, and all data must be retrieved from source data location. We have quite a few examples of data lake products among them are Google cloud,

Amazon S3 or AWS S3 (Simple Storage Service), Azure Blob Storage or Azure cloud to mention.



AZURE Data Lake contains the following components; Data Lake Store that includes the Data in raw object form with no particular schema type defined, based on Apache Hadoop File System (HDFS). This has no limit data store.

Analytics Job Service This provides high throughput on data lake for raw or any other given data format for analytics and real-time reporting and monitoring, its scalable and auto-scalable with the flexibility of payment for processing.

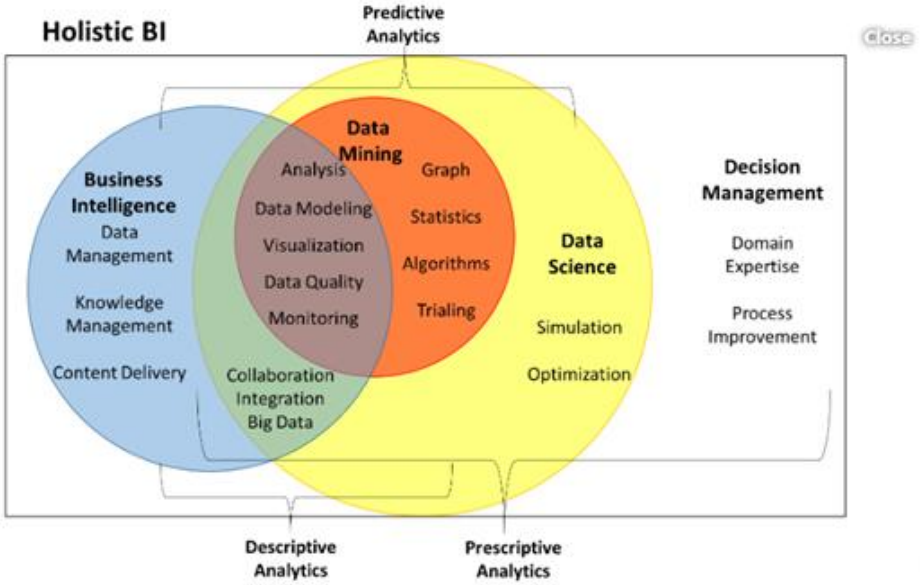
Cluster management, handle analytics via Spark, clusters and U-SQL that can process any data with SQL like syntax and additional ADFS driver functions defined by Azure custom functions with a highly available data warehouse service from premise where many tools can be used to investigate data for analytics, reporting, monitoring and business intelligence.

Data factory to manage the data on Azure Data Storage properly.

In conclusion, AZURE can store almost unlimited data, does instant job processing, business Intelligence and Analytics.

7. What are the commonalities and differences between business intelligence and data science? (Minimum 150 words)

The most important commonality between business intelligence and data science is that both are focused on data and their goal is to analyze it in order to achieve a profit for the organization. Several years ago who worked with data in the company were known as data analysts. Nowadays business moved from reporting past performance to predicting future trends, known as data science.

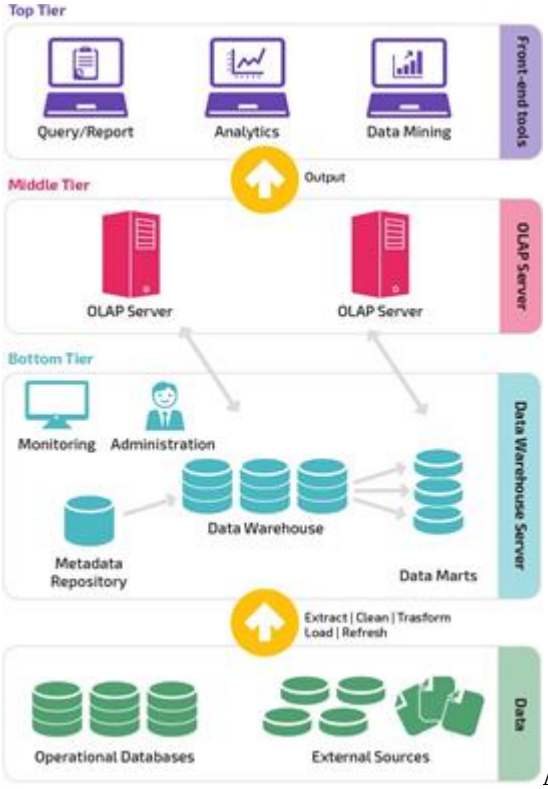


As we can see in the image above, BI and Data science has a lot of similarities. For example, the analysis data modeling, data management, collaboration integration big data…

According to *Mike Merritt-Holmes,* the differences between BI and data science are principally 10:
Focus: the focus of BI is at the present, data science at the future.
Process: BI process is static and comparative. Data science process are dynamic.
Data source: BI worked with pre-planned data that are added slowly. Data science fits better with Data Lake and works also with unstructured data.
Transform: BI answer question you know, data science finds new questions.
Storage: BI worked prevalently with DW; data science can be distributed real time.
Data Quality: BI provides SVOT, data science offers precision, confidence level and so on.
IT owned vs business owned: Business intelligence owned by IT department. Data science owned by analysts.

Analysis: BI works is focused on prescriptive or retrospective. Data science on predictive.
Business value: higher in data science because it's able to work based on future prediction, not just on the past.

8. **How can DW/BI and big data / data science initiatives be combined in an organization? If possible, support you answer with a diagram or picture. (Minimum 150 words)**



As we can see in the image on the previous question, Data science is a 'big umbrella' that contain a lot of fields as data modelling, visualization, statistics, simulation, optimization…
BI is a data analysis and insight process that helps businesses make decisions. Data scientists find meaningful hypotheses in a BI system and can respond to them using available data. It could be considered as a synonym for data analytics for business data-analysts and data scientists discover meaningful hypotheses in an effective BI process and can respond to them using the data. The image on the left fully explains how DW/BI and big data/data science initiatives can be combined in an organization.
We start from data derived from operational db and external sources, after extracting, clean, transform, load, refresh data, we pass to the DW Server, which provides to serve the necessary data that the data science needs.

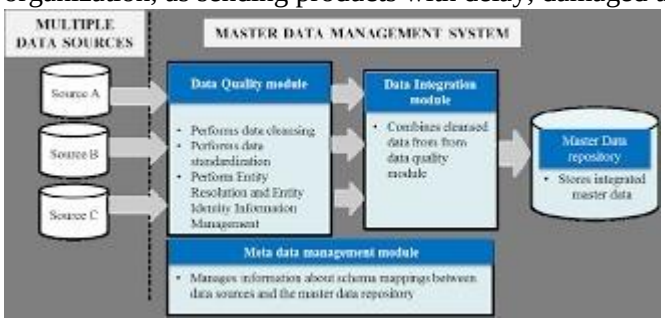
9. **What is a master data management system, what is its function, and what are its benefits to an organization? If possible, support you answer with a diagram. (Minimum 150 words)**

A master data management system is a data management function referring to the management of shared master data within an organization. It can enable an organization to associate all of its critical data to a single reference platform (Haneem et al., 2017). This system contains information about different fields as customers, products. The key benefits of such a system include improved efficiency by eliminating unreadable and inaccessible data, improved decision-making through better insights by having all critical data in one place, and effective prioritizing, as a data management system can allow

for a unified view of strategic customer relationships ("Master Data Management for Small Business and Large Corporations Alike", 2019).

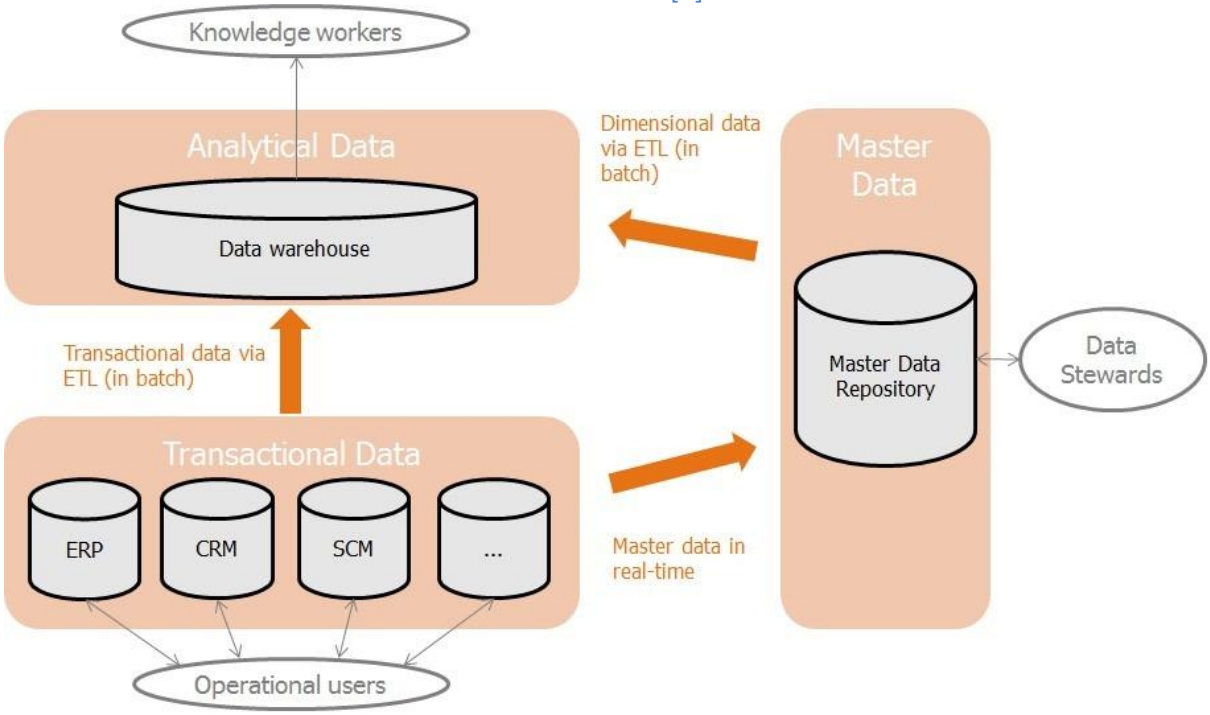
If a company has multiple IT systems, it's possible to have data that are not correct regarding fields previously indicated. The applications are several: we can automatically correct the misspelled words (address of customers, name of customers…) or find allowed categorization and hierarchies of categorizations.

Using the master data management system it's possible to avoid problems that can be harmful for an organization, as sending products with delay, damaged trust of the company.



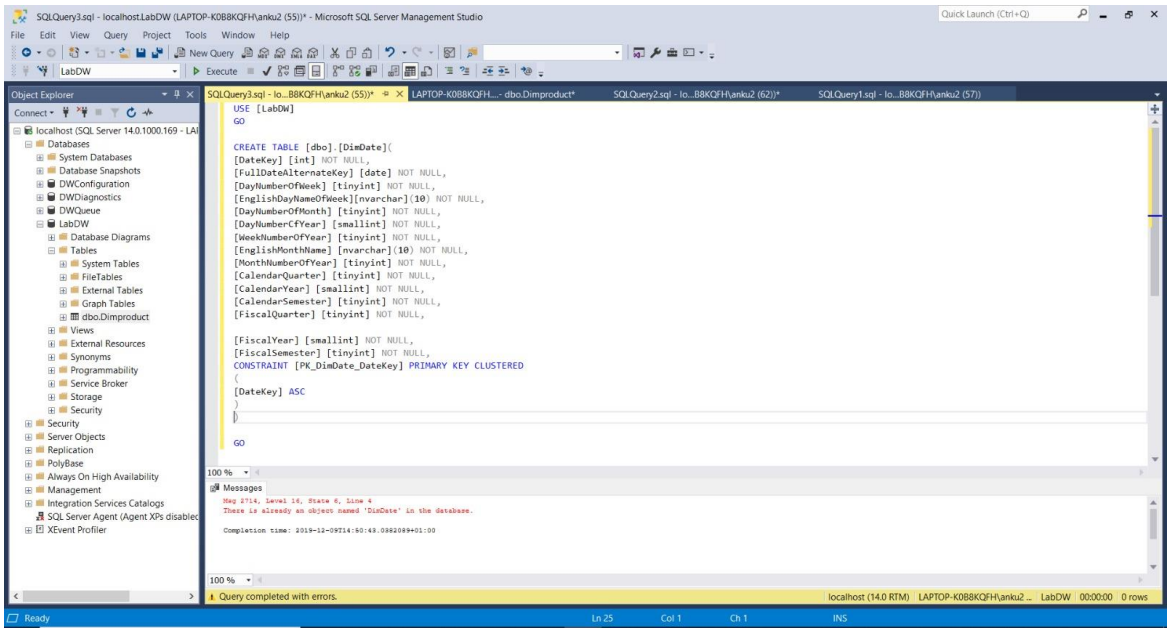
10. How can a data warehouse system make use of a master data management system? Support you answer with a diagram. (Minimum 100 words)

We can divide the architecture in this way: DW that includes Analytical Data ("Data that is calculated and/or derived from transactional information to support the decision making of the organization"). ERP, CRM, SCM … *that includes Transactional Data* ("Data that is being generated by applications in supporting business processes of the organization") and master data repository that includes *Master Data* ("represents business objects upon which transactions are done and the dimensions on which analysis is conducted"). These 3 environments communicate. Transactional data are sent in the DW via the ETL process. But transactional data are also sent in real-time to the Master Data Repository that transform them in dimensional data and send them via ETL at DW. [3]

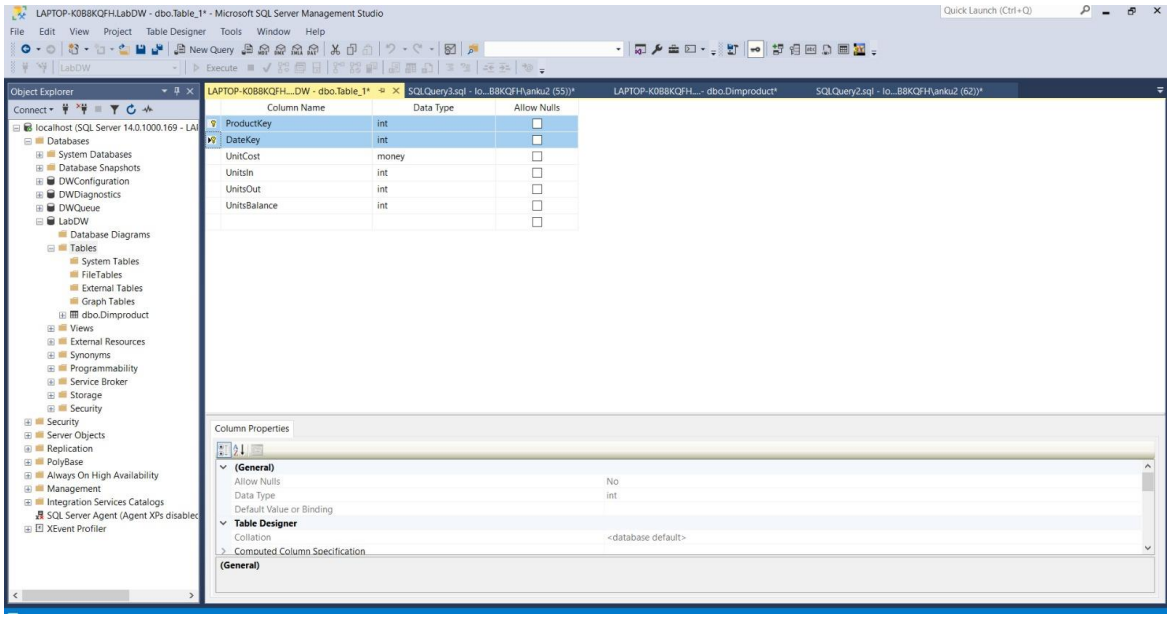


Tool Tutorial/Lab Exercise

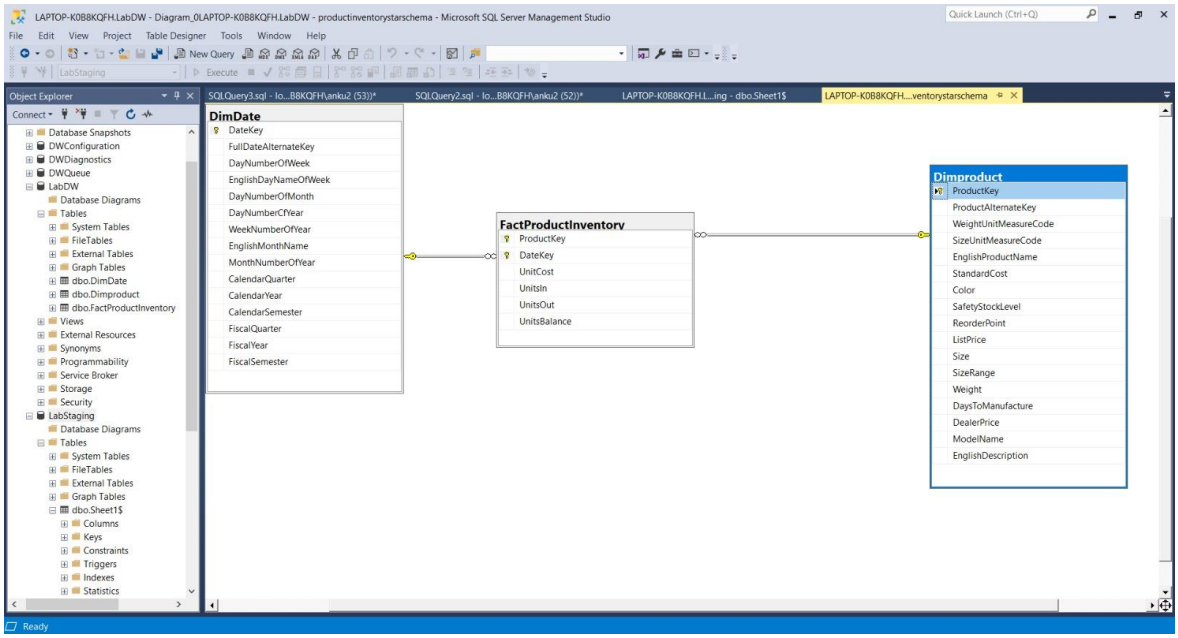
Here we uploaded some screenshots of our steps through Lab/Tutorial.



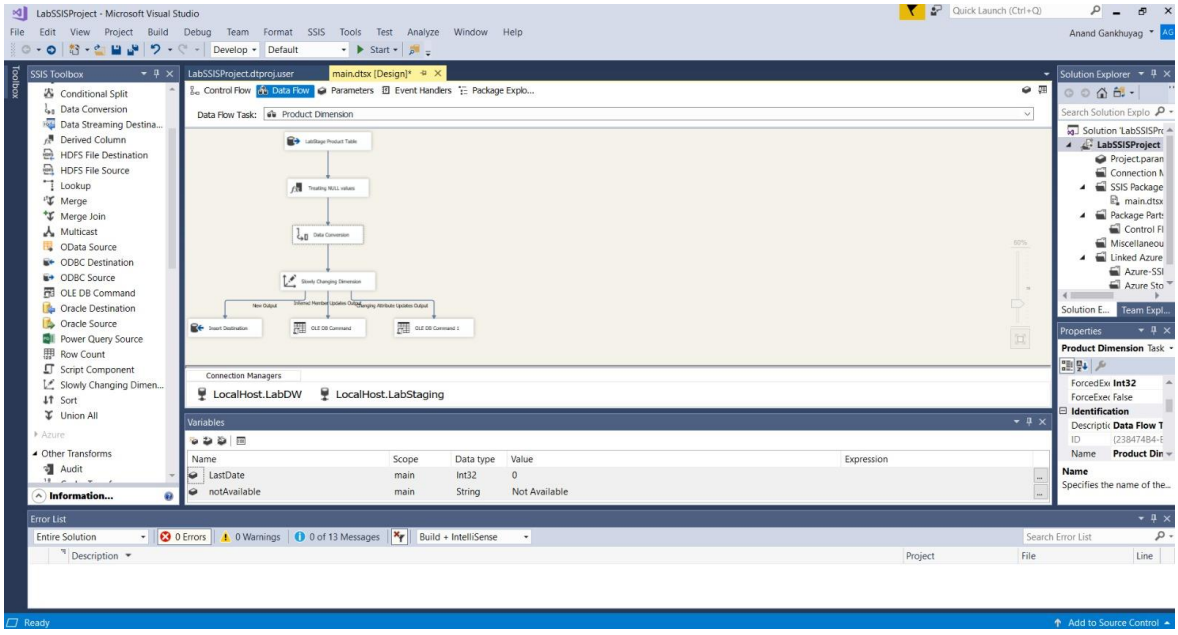
Screenshot 1: Creating Date Dimension



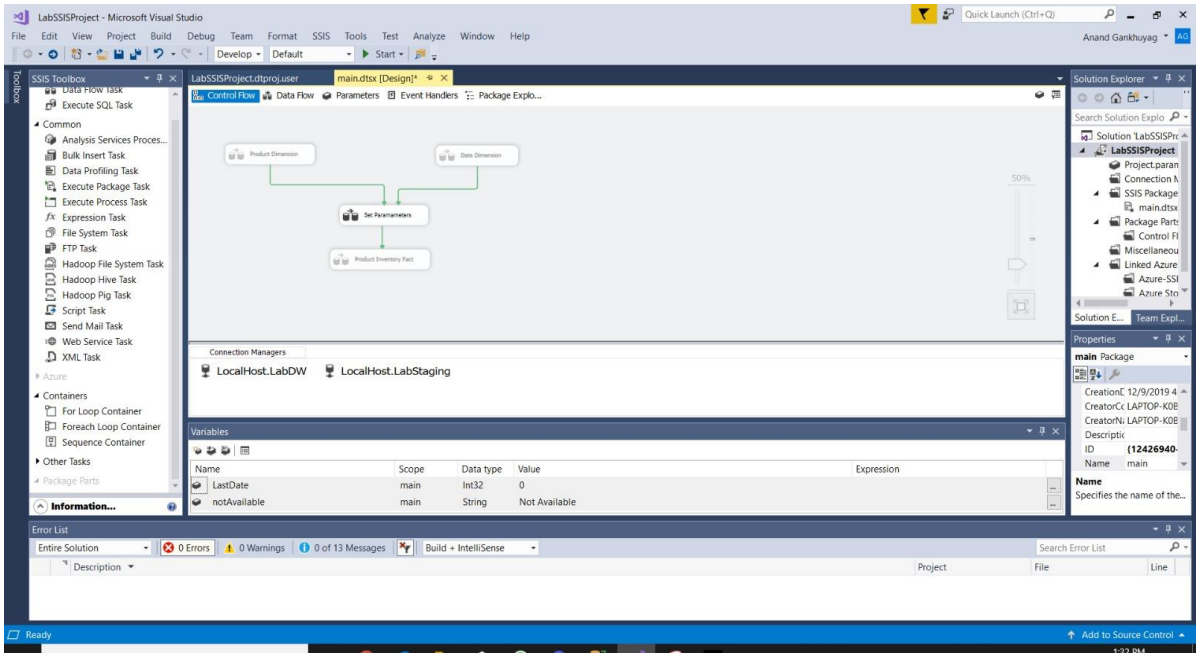
Screenshot 2: Setting Primary Key using the Design Window



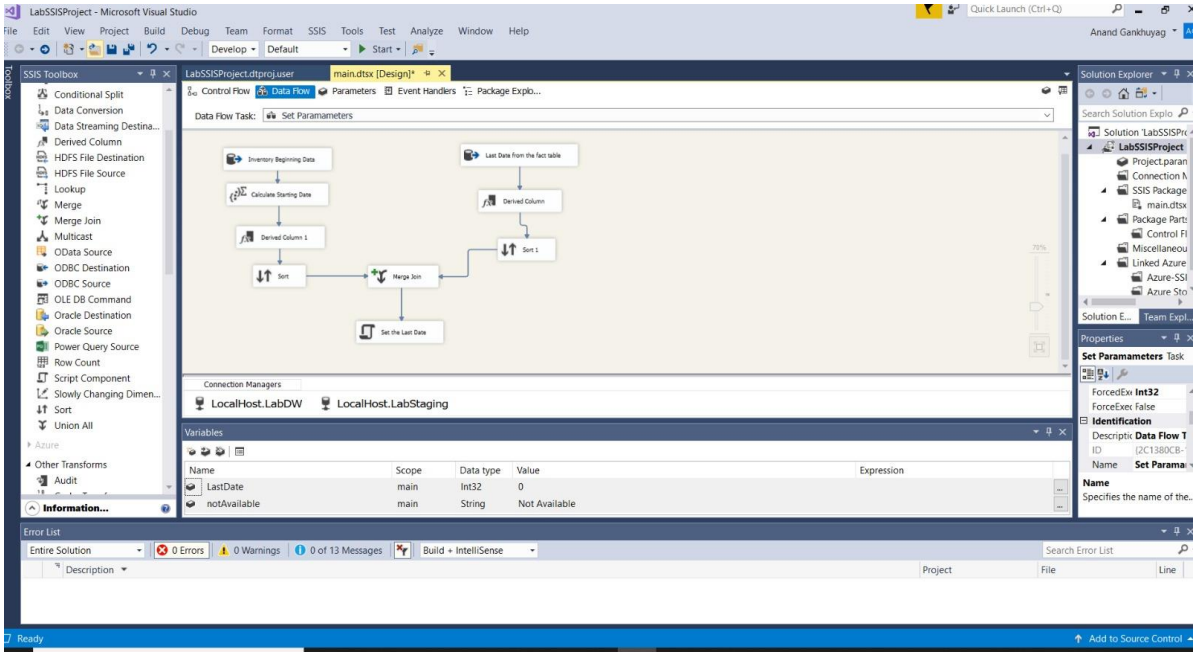
Screenshot 3: The relationship between tables



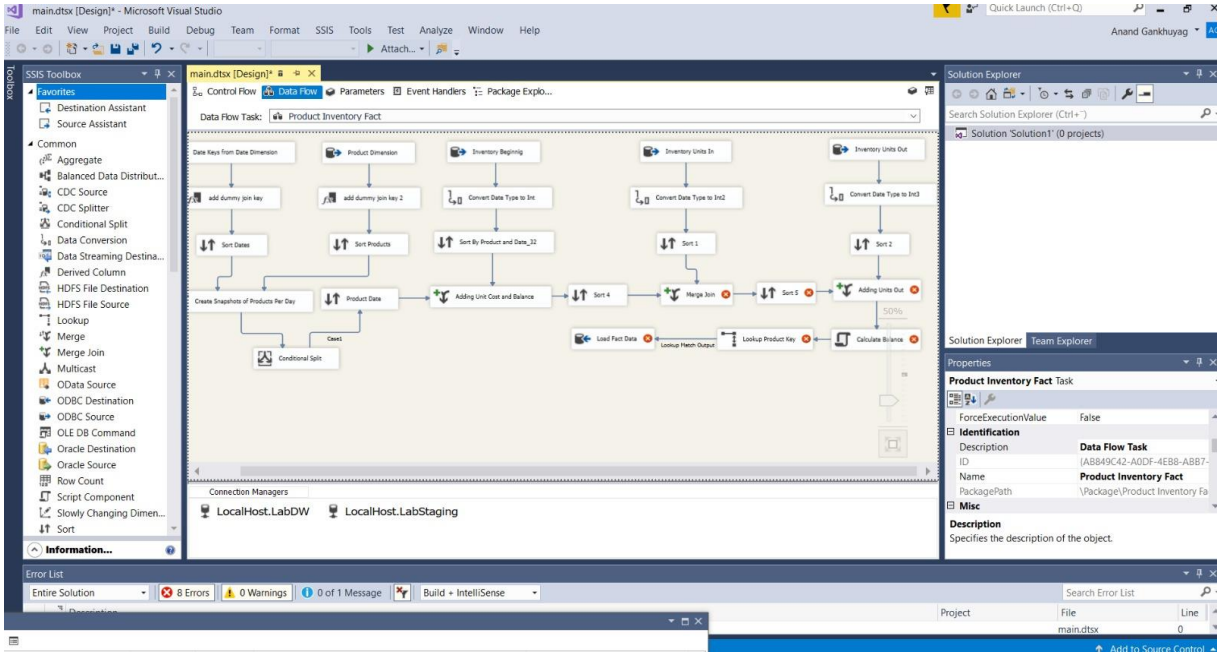
Screenshot 4: The implemented Slowly Changing Scenario



Screenshot 5: Designing the Data flow



Screenshot 6: Setting Parameters flows



Screenshot 7: The Grain has already created, Inventory_Beginnig has retrieved from Staging Area

But we are having trouble on "Converting Date Type to Integer". According to this situation, we could not go further.

References

CaptainStrawWallaby2072. (n.d.). prohibited in the back room and therefore the front room is dedicated to just. Retrieved from https://www.coursehero.com/file/p3qmpru/prohibited-in-the-back-room-and-therefore-the-front-room-is-dedicated-to-just/

Orlov, V., Orlov, V., Vadim, & West Monroe Partners' Technology. (2019, September 12). Data Warehouse Architecture: Inmon CIF, Kimball Dimensional or Linstedt Data Vault? Retrieved from https://blog.westmonroepartners.com/data-warehouse-architecture-inmon-cif-kimball-dimensional-or-linstedt-data-vault/

Master Data Management (MDM) : Architecture & Technology. (2018, August 14). Retrieved from https://www.element61.be/en/resource/master-data-management-mdm-architecture-technology

Department of Computer and Systems Sciences
Stockholm University
Forum 100
SE-164 40 Kista
Phone: 08 – 16 20 00
www.su.se

Stockholm University