ACADEMIC CITY
UNIVERSITY COLLEGE

BUSINESS INTELLIGENCE WEEK 5

Lecturer:

KWESI MENSA CRANKSON

kwesi.crankson@acity.edu.gh

Content

1. Data Warehousing

2. The ETL Process

3. Selection of a BI project

1. DATA WAREHOUSING

- In the case of a company there is the possibility that relevant information is stored on different platforms in different locations within the company,

- The process of gathering all the information and storing them in one location for processing is referred to as **DATA WAREHOUSING**

- Data warehousing involves the use of one location (Data Warehouse) which is a data-storage system purposely built to store historical operational data.

- This data in the warehouse is used for creating reports and performing different analysis which will generate reports to show the trends among others.

- Doing analysis on current operational data will be expensive and cause problems to the day-to-day activities of the company. Data warehousing separates operational data which is used for reporting and analysis.

- Data warehouse does not necessarily refer to one single database but can refer to several linked databases which are related to store data.

- The finance and accounting departments (for example) may use one data storage system, while the HR and sales teams use their own data-collection mechanisms. Many of the objects and information nuggets contained in these different systems will be the same.

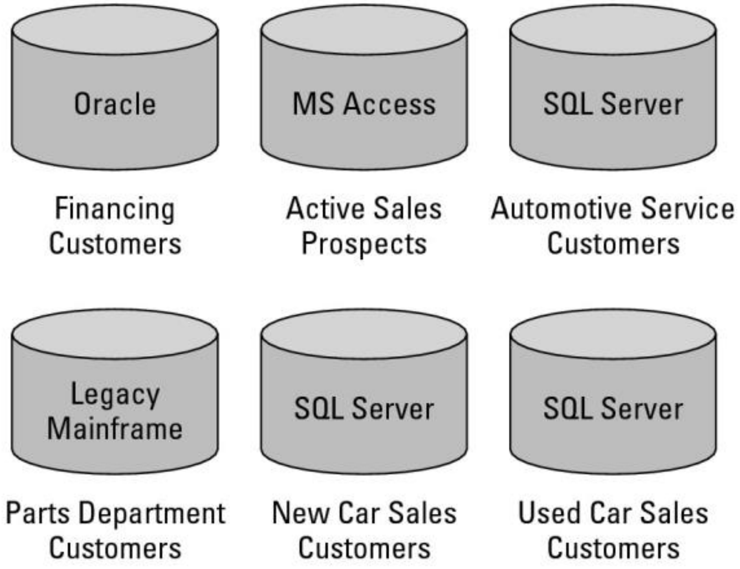
- The data warehouse provides the logical — and, in most cases, physical — link that connects those objects together across departmental lines.

Difference between Data Warehouse and Data Repository

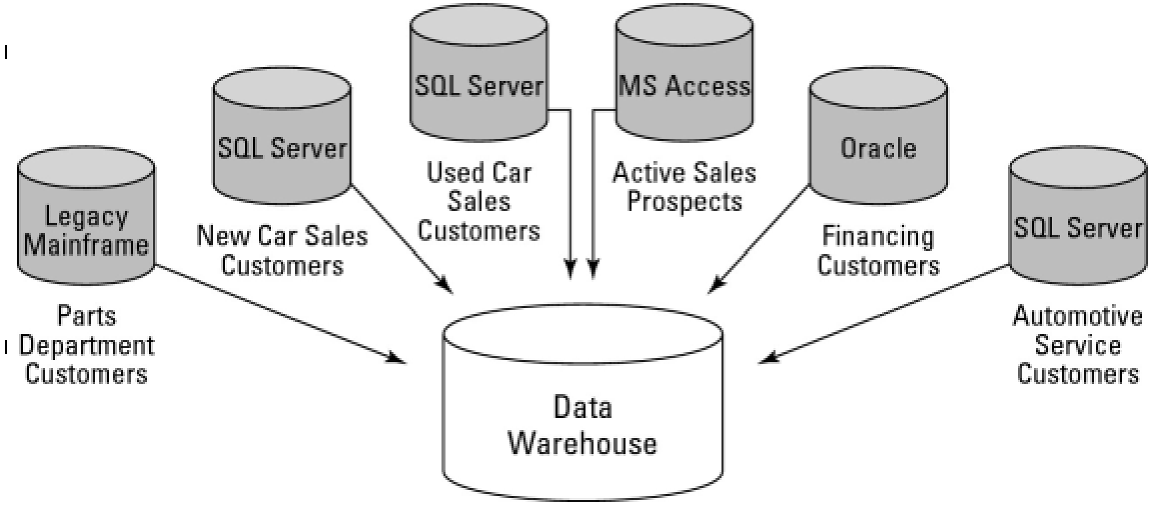
- A DATA WAREHOUSE represents all the transactional events that were made in the past, on which the parent company performs read-only activities such as running reports and statistical analyses.

- A data *repository* is a snapshot of the *present* (and perhaps the recent past); the company’s transactional systems regularly update the data in the repository for specific tactical decision-support tasks.

1. Data storage for an automotive company using individual databases



2. Using a data warehouse to give access to all databases in the automotive



- It allows for better operational decisions by the sales people when a customer walks in the door.

- Storing it in a data warehouse that’s customized for reporting purposes will mean managers can build meaningful customer and sales reports in a fraction of the time it would take if anyone attempted to bridge the six operational systems for the same purpose.

Challenges of Data Warehousing

- Since data is initially stored in separate databases there is the possibility of having duplicates. So part of building the data warehouse involves identifying situations where that single person is stored in more than one operational system.

- To accomplish unification in the automotive data warehouse, you have to compare the customer records in each database, combining those where you're sure the duplicate data relates to a single person. To start with, you have to sort the records into categories.

The different categories for sorting records can be:

- Two or more identical records, where each of the data fields are exactly alike and likely relate to a single customer.
- Records where you can make an educated guess (or apply certain rules to come to a conclusion) that they relate to the same customer.
- Records that share a few characteristics and *might* be the same, but you can’t be sure without further investigation.

Data Quality and Data Integrity in Data Warehousing

- It is important for all the data found in the Data Warehouse to conform to the rules of Quality and Integrity.

1. Data quality: Is the data usable and complete?

It is important to clean up the data and ensure that it's all in a uniform, usable format.

2. Data integrity: Is the information *correct*? Does it represent what everyone believes it does?

The Architecture used for Data Warehouses

- This is mostly dependent on the specific needs of the organization. Some of the common examples are:

1. **Simple.** In this architecture all the metadata, summary data, and raw data are stored within the central repository of the warehouse. The repository gains data from data sources on one end and accessed by end users for analysis, reporting, and mining on the other end.
2. **Simple with a staging area.** The operational data is cleaned and processed before storing it in the warehouse. Even though this can be done easily and straightforward there is the addition of a staging area for most data warehouses to simplify data preparation.

3. **Sandboxes.** These are private, secure, safe areas which gives the chance to companies to quickly explore new datasets or means to analyze data without having to follow the formal rules and protocol of the data warehouse.

4. **Hub and spoke.** With the incorporation of a data mart the organization is able to serve various lines of business as it can customize its functions. When the data is ready for use, it is moved to the appropriate data mart.

2. The ETL Process

- ETL processes (Extract, Transform and Load) are a series of steps which are utilized by the system to move information from the source data systems to the target database.

- The step taken in the ETL process of every organization goes a long way to influence the success or otherwise of the BI functions in that organization.

EXTRACT

- The first stage is to EXTRACT the data from the various sources such as transactional systems, spreadsheets, and flat files.

- This step involves reading data from the source systems and storing it in a staging area.

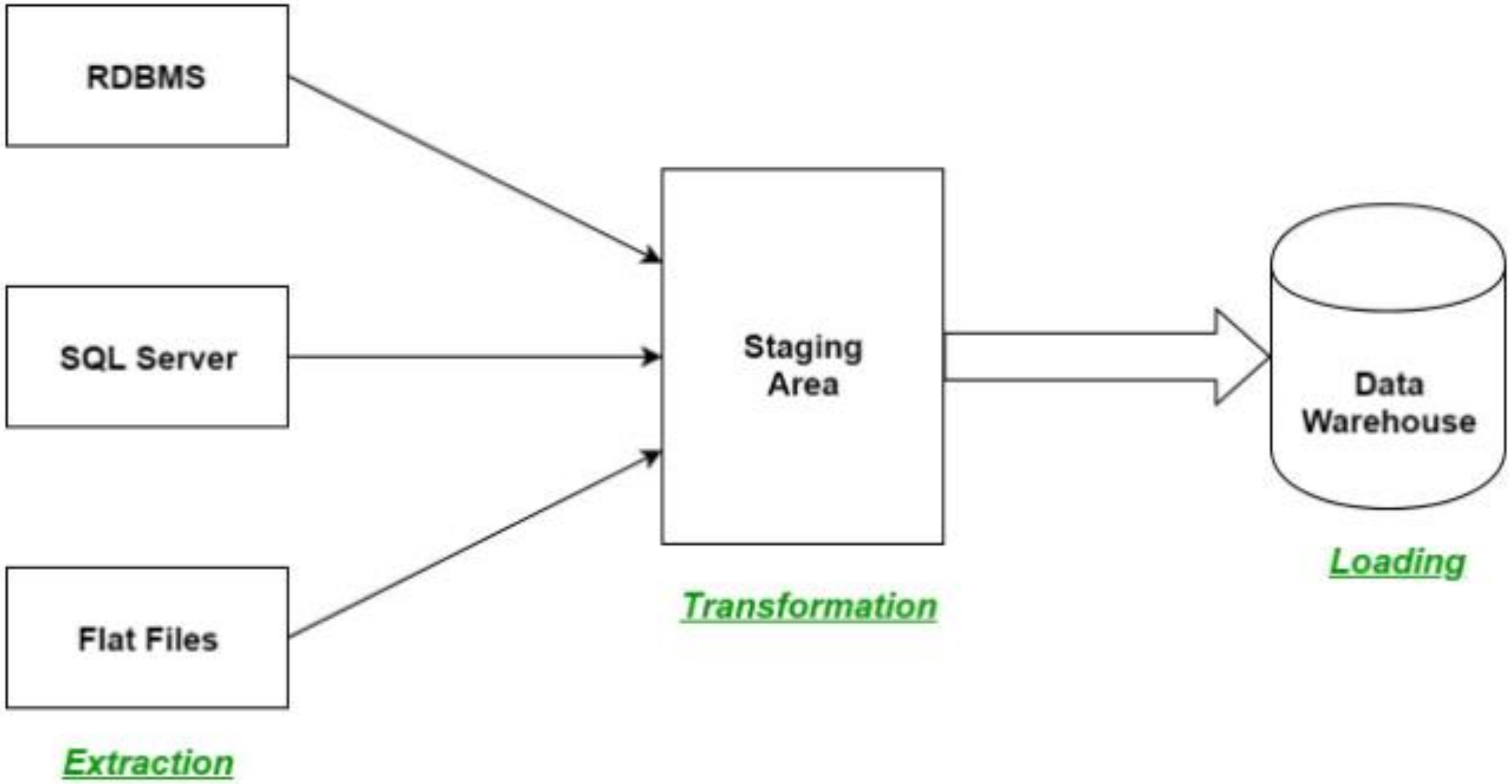
- To create a data warehouse, extraction typically involves combining data from these various sources into a single data set and then validating the data with invalid data flagged or removed.

TRANSFORM

- This stage is for the transformation of the data which has been extracted into a format which will be used for loading into the data warehouse.
- This stage also entails the cleaning and validation of data, converting data types, creating new fields among others.
- There are various rules and functions used which ensures that bad or non-matching data is NOT added to the destination repository

LOADING

- After transforming the data it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.
- In this stage data is delivered and secured for sharing, making business-ready data available to other users and departments, both within the organization and externally.



RDBMS

SQL Server

Flat Files

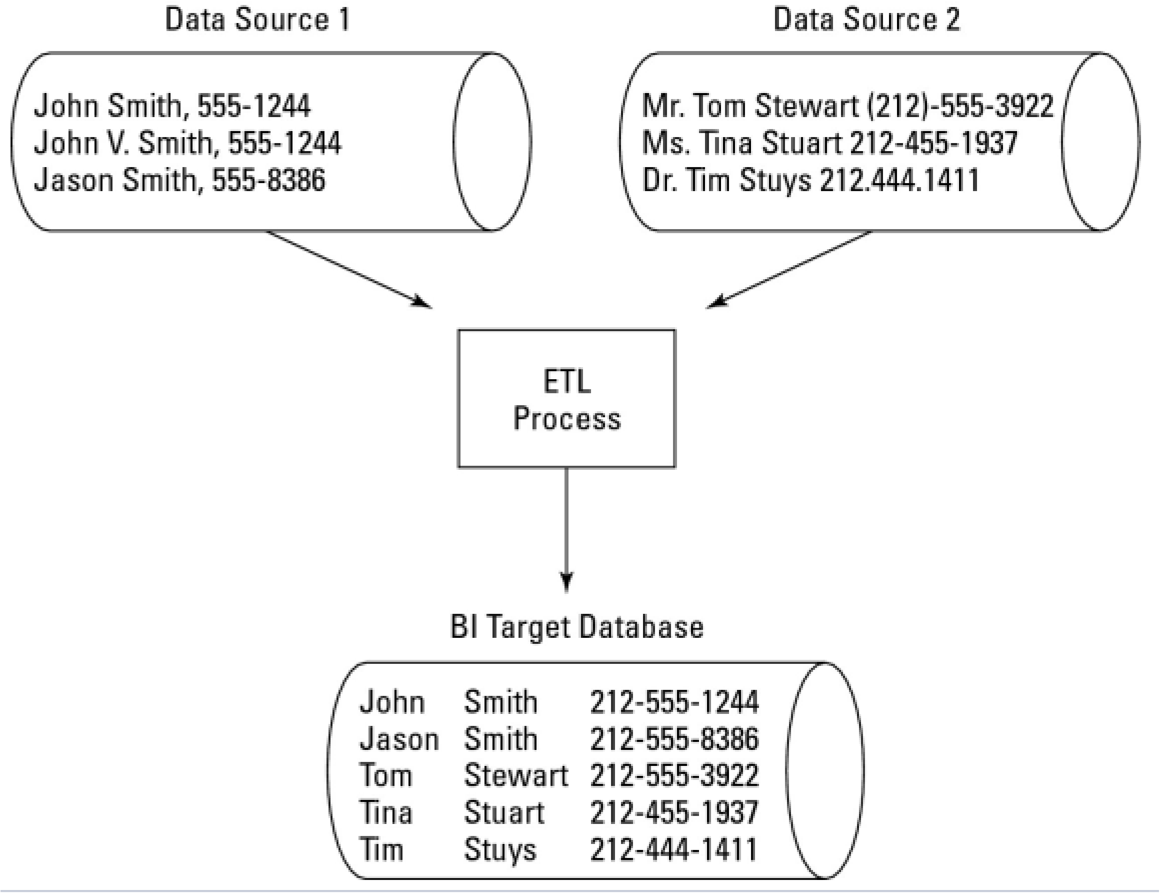
Staging Area

Data Warehouse

Extraction

Transformation

Loading



Data Source 1

John Smith, 555-1244
John V. Smith, 555-1244
Jason Smith, 555-8386

Data Source 2

Mr. Tom Stewart (212)-555-3922
Ms. Tina Stuart 212-455-1937
Dr. Tim Stuys 212.444.1411

ETL Process

BI Target Database

John	Smith	212-555-1244
Jason	Smith	212-555-8386
Tom	Stewart	212-555-3922
Tina	Stuart	212-455-1937
Tim	Stuys	212-444-1411

A Conceptual View of the ETL process.

Advantages and Disadvantages of ETL

ADVANTAGES	
Improved Data Quality	The ETL ensures that all the data which comes into the warehouse is of good quality, accurate and complete
Better Data Integration	ETL helps to transform data from multiple sources and makes it accessible and easy to use.
Increased Data Security	Controls access to the data warehouse and ensures that only authorized users access the data.
Improved Scalability	Provides a means to analyse large volumes of data
Increased Automation	ETL creates more automation that reduces the time and effort required to load and update data in the warehouse.

DISADVANTAGES	
High cost	ETL can prove to be a costly process for companies with little budgets.
Complexity	The processes can be complex and difficult to implement as they may lack the needed expertise.
Limited flexibility	Difficulty in handling unstructured data and real time data streams.
Limited scalability	ETL process can be limited in terms of scalability, as it may not be able to handle very large amounts of data.
Data privacy concerns	The different process can sometimes bring up issues about data privacy, as large amounts of data are collected, stored, and analyzed.

Benefit of ETL to Business Intelligence

- Providing Historical Context

1. ETL provides the organization’s data with a rich historical context.

2. Data from fresh platforms and apps can be integrated with old data in an organization. A long-term picture of the data is provided by the ability to see older datasets with more recent information.

- Consolidated Data view

1. There is a combination of multiple datasets or databases into one unified view for easy analysis.

2. This process ensures that the data is of good quality ad reduces the time needed for moving, categorization and standardization of data.

3. This makes it easier to analyze, visualize, and make sense of large datasets.

- Accurate Data Analysis

1. The ETL process helps provide reliable data analysis to help the organization meet the different compliance and regulatory standards.

2. There is the integrating of ETL tools with data quality tools and this is helpful in the audit, and cleaning of data, ensuring that the data is trustworthy.

- **Task Automation**

1. ETL is able to automate the processes which needs to be continuously performed to ensure efficient analysis.

2. This provides the avenue for data engineers to devote more time to performing more important tasks and less time for easy tasks such as moving and formatting data.

3. Selection of a BI project

- The BI project plan serves as the main tracking and control mechanism for the actual implementation of the Business Intelligence.
- This details the exact moves of the team and the different tasks or milestones clearly spelt out for all members of the team.
- The different resources for the project are also spelt out.

- The BI project plan serves two purposes:

1. A schedule of the various tasks to be performed and by whom. This schedule is for current and future tasks.

2. Serves as the main tool for organizing various tasks in the team. This is ensures that there is common goal or target for all processes being undertaken.

- The project plan is used by project managers to establish expectations for members in and out of the team.

- It helps the team to have a view of all their different tasks and how they influence the work of other. (helps with team building)

- A project plan is never really *complete* in the traditional sense of the word as it will be continuously feed with information from the current scenario to help indicate the future trends.

Essential steps in a BI project

1. Definition of the Scope and Objectives
2. Assessment of the current system and gaps
3. Design the solution and architecture
4. Develop and implement the solution
5. Validate and evaluate the results
6. Manage and maintain the solution

1. Definition of the Scope and Objectives

- Having a clear view of the different challenges or opportunities, the different stakeholders and users, the data requirements and the expected outcomes.

- The scope and objectives needs to be clear, specific, measurable, achievable, relevant, and time-bound

- These three questions also need to be answered during in the definition
  - ❑ Why is the organization undertaking this venture?
  - ❑ What are they expecting to get from it?
  - ❑ Who needs to know the project plan?

2. Assessment of the current system and gaps

- There is the need to assess the current state of the data and analytics capability of the organization.
- This step can be by performing a data audit, a data quality assessment, a data governance review, and a technology assessment.
- Identifying the gaps will help the team identify the different means to fill those gaps in order to have an effective BI project.

3. Design the solution and architecture

- At this stage there needs to be a clear definition of the data model, data integration, the data warehouse, the data visualization, and the analytics tools and methods.

- All these tools and methods should be in line with the organization's overall BI strategy.

- Compatibility with the existing system will be a very important part of this step.

4. Develop and implement the solution

- This stage will have all the coding, testing, debugging, deploying and documentation of the solution provided.
- It involves translating the designed architecture into tangible BI components which can include establishing the needed data pipelines etc.
- Developing the BI solution in accordance with the designed architecture ensures data accuracy, security, and accessibility for stakeholders.

5. Validate and evaluate the results

- At this stage we compare the actual outcomes with the expected outcomes, collect feedback from users or stakeholders and measure key performance indicators (KPIs).

- This needs to be done based on the scope and objectives of the project whiles considering the quality, accuracy, relevance, timeliness, and usability of the data and analytics.

- The validation and evaluation should also identify any gaps, issues, or opportunities for improvement in the solution.

6. Manage and maintain the solution

- Several activities are performed in this stage such as:
  1. Monitoring
  2. Updating
  3. optimizing the solution,
  4. ensuring the data quality, security, and governance.

- There needs to be a process for ongoing monitoring and maintenance

Various Roles in the BI Project Team

1. **Project manager:** He or she is responsible for the coordination and management of all activities within the project.

2. **Business analysts:** The role of the BA is to codify the requirements of the projects which will be gained from the stakeholders. They are also responsible for the requirement documentation and overseeing the change control process.

3. **BI developers:** This group handles all that pertains to technology in the project. They are responsible for the end-user environment (reporting, analysis, visualization, or other tool).

Persons found in this bracket will include application developers, data-mining experts, or report developers.

4. **Database administrator:** The individual handling this process is responsible for designing and maintaining the target data repository that is the destination for operational data feeds.

5. **Data administrator:** This is an important part of the BI project where the administrator is responsible for identifying and analyzing operational data sources throughout an organization, and developing a process to bring the data together.

6. **ETL developers:** Basically for *e*xtracting, *t*ransforming, and *l*oading data. Other responsibilities includes all processes that affect the data as it moves from the operational sources to one of two destinations:

- The BI data repository
- The end-user environment

7. **Testers:** Since there are different parts of a BI project there will always be a lot of quality assurance roles. Some projects may require all of these specialty testers:

- Front-end testers for the user-facing software
- A data-quality analyst who can monitor the data itself to ensure that it’s performing as advertised
- Integration testers who will make sure that applications are working together properly, and that all the pieces have come together as designed

QUESTIONS

- What will be the key considerations in creating a Data Warehouse for a University?

- How does ETL affect the quality of data which is used in an organization and is there a need for a more refined process.