CS –32
Data Warehousing with SQL Server 2012

01. Introduction to

Data

Warehousing

Definition of data and table.

- What is Data?
  - **Data** is a collection of information related any special topic.
- What is Table?
  - A **Table** is a collection of related data held in (રાખવામાં) a structured format within a database.

Definition of Database and RDBMS.

- What is Database?
  - **Database** is collection of information in a specified format.
  - In simple mining : Database is a collection of tables of specified information.
- What is RDBMS?
  - **RDBMS** stands for Relational Database Management System.

Basic Info of Data warehouse.

- What is Data Warehouse?
  - Data (માહિતી) Ware (વેચાણ માટેનો સ્ટોક) House (ઘર)

    (Simple mining is that Data Warehouse is a collection of RDBMS.)

    Note : This is not perfect definition.

Definition of Data Warehouse.

- Definition of Data Warehouse
  - Data warehouse is a Subject Oriented (વિષય લક્ષી) Integrated (એકત્રિત કરવું) Non-volatile (ભૂંસી ના શકાય તેવું) Time variant (ટાઈમ સર) collection of data in support of management`s decisions.
- Example :
  - Big-Bazar Shopping Mall.

Explanation of Data Warehouse.

- Subject oriented (વિષય લક્ષી) :
  - Store particular data (information) like

    1. Sales 2. Purchase

    3. Employee 4. Customer
- Integrated (એકત્રિત કરવું) :
  - All the database have been made in different database platform like (Oracle, SQL, MS-Access etc.)
  - Data warehouse support all platforms.

Explanation of Data Warehouse.

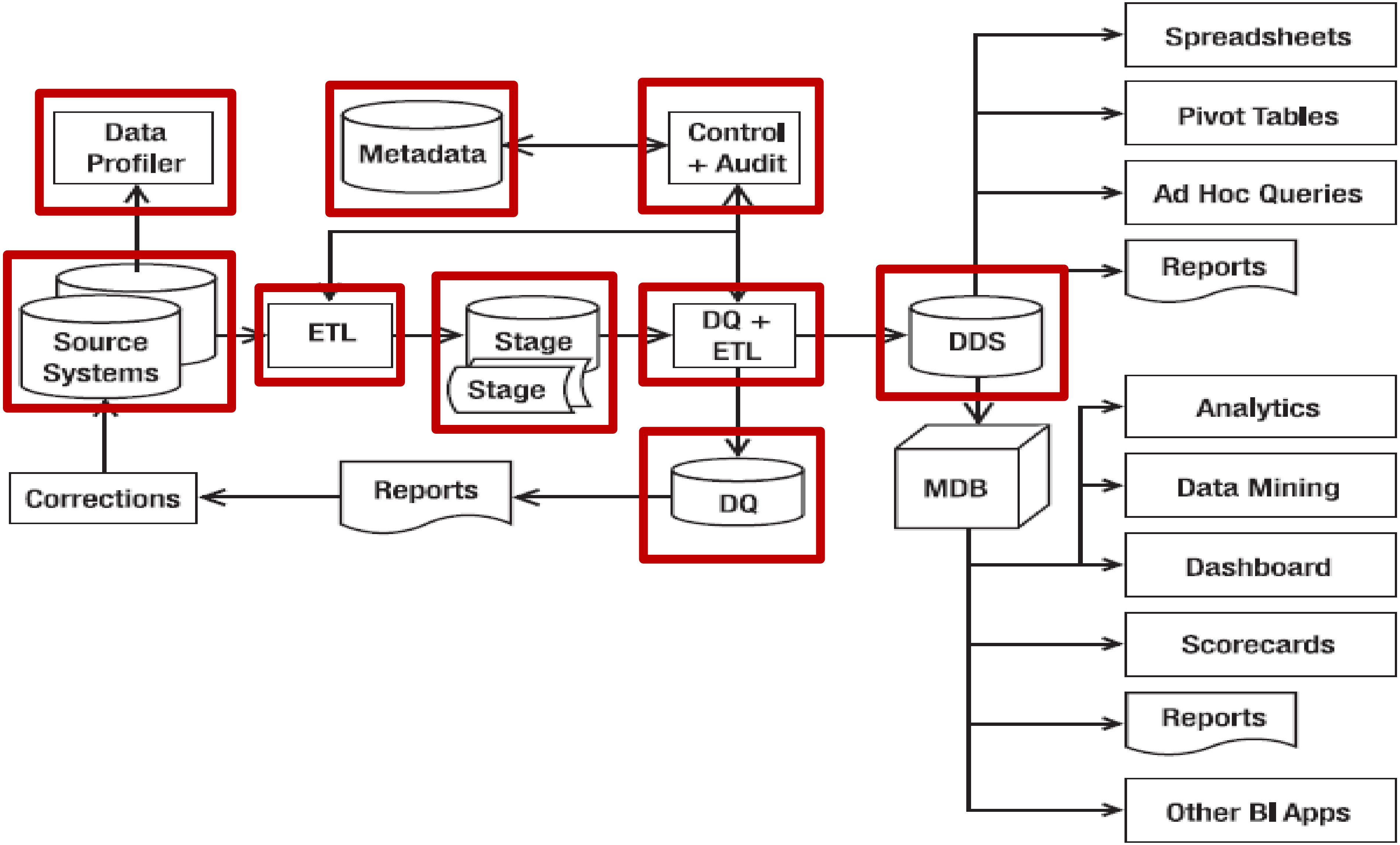
- Non-Volatile (ભૂંસી ના શકાય તેવું) :
  - Once data stored in Data Warehouse then it can not update. It can only readable mode.
- Time variant (ટાઈમ સર) :
  - We can show all the data which is stored in past.

Data Warehousing…

- Definitions of Data Warehousing

1. Data warehouse is a Subject Oriented, Integrated, Non-volatile, Time variant collection of data in support of management's decisions.
2. An integrated collection of data about a collection of subject of subjects, which is not volatile in time and can support decision taken by the management.

A diagram of DATA Warehouse System



A diagram introduction

Let’s go through the diagram in Figure, component by component, from left to right.

- Source System (OLTP)
  - The source systems are the OLTP systems that contain the data you want to load into the data warehouse.
  - Online Transaction Processing (OLTP) is a system whose main purpose is to capture and store the business transactions.
  - The source systems’ data is examined using a data profiler to understand the characteristics of the data.

A diagram introduction

- Profiler
  - A data profiler is a tool that has the capability to analyze data, such as finding out how many rows are in each table, how many rows contain NULL values, and so on.

A diagram introduction

- Staging Area
  - Staging (data) A staging area, or landing zone, is an intermediate storage area used for data processing during the extract, transform and load (ETL) process.

A diagram introduction

- ETL (**E**xtract, **T**ransform, **L**oad)
  - The extract, transform, and load (ETL) system then brings data from various source systems into a staging area.
  - ETL is a system that has the capability to connect to the source systems, read the data, transform the data, and load it into a target system.

A diagram introduction

  - The ETL system then integrates, transforms, and loads the data into a dimensional data store (DDS).
- DQ (**D**ata **Q**uality)
  - When the ETL system loads the data into the DDS, the data quality rules do various data quality checks. Bad data is put into the **d**ata **q**uality (DQ) database to be reported and then corrected in the source systems.

A diagram introduction

- DDS (**D**imensional **D**ata **S**tore)
  - A DDS is a database that stores the data warehouse data in a different format than OLTP.

A diagram introduction

- In DDS the data is arranged in a dimensional format that is more suitable for analysis. That is the reason for getting the data from the source system into the DDS and then querying the DDS instead of querying the source system directly.

A diagram introduction

- The second reason is because a DDS contains integrated data from several source systems.
- When the ETL system loads the data into the DDS, the data quality rules do various data quality checks. Bad data is put into the **d**ata **q**uality (DQ) database to be reported and then corrected in the source systems.

A diagram introduction

- Control +Audit
  - Bad data can also be automatically corrected or tolerated if it is within a certain limit.
  - The ETL system is managed and orchestrated by the control system, based on the sequence, rules, and logic stored in the metadata.

A diagram introduction

- Meta Data :
  - Meta Data is data about data.
  - Meta Data definitions used by all processes within the data warehouse.
  - In order to provide universal data access, it is necessary to maintain some form of data directory of Mata data information.

History :

- One of the key differences between a transactional system and a data warehouse system is the capability and capacity to store history.
- Most transactional systems store some history, but data warehouse systems store very long history.
- In my experience, transactional systems store only one to three years of data; beyond that, the data is cleared.

History :

- For example,
  - Let’s have a look at a sales order–processing system.
    - The purpose of this system is to process customer orders.
    - Once an order is dispatched and paid, it is closed, and after two or three years, you want to purge the closed orders out of the active system and archive them to maintain system performance.

History :

- You may want to keep the records for, say, two years, in case the customer queries their orders, but you don’t want to keep ten years worth of data on the active system, because that slows the system down.
- Some regulations (which differ from country to country) require you to keep data for up to five or seven years, such as for tax purposes or to adhere (વફાદાર રહેવું) to stock exchange regulations. But this does not mean you must keep the data on the active system.

History :

- You can archive it to offline media. That's what a typical transaction system does:
  - It keeps only two to three years of data in the active system and archives the rest either to an offline media or to a secondary read-only system/database.

History :

- A data warehouse, on the other hand, stores years and years of history in the active system.
- The amount of historical data to store in the data warehouse depends on the business requirements. As per requirement Data warehouse tables can become very large.

History :

- Imagine a supermarket chain that has 100 stores. Each store welcomes 1,000 customers a day, each purchasing 10 items. This means 100 x 1000 x 10 =1000000 (1 million) sales order item records every day.
- In a year, you will have 365 million records. If you store 10 years of data, you will have 3.65 billion records.

Query

- *Querying* is the process of getting data from a data store, which satisfies certain criteria.
- Here is an example of a simple query:
  - "How many customers do you have?"
- A data warehouse is built to be queried.
- That is the number-one purpose of its existence.

Query

- Users are not allowed to update the data warehouse. Users can only query the data warehouse.
- Only the ETL system is allowed to update the data warehouse.

Types and Info of Systems…

- What is System?
  - System means which kind of Data stored in Data Warehouse.
  - There are two types of systems in Data Warehouse.
    1. Operational System
    2. Informational System

Types and Info of Systems…

- Operational System
  - Operational System means data which update in day-to-day.
  - It means that in the operational system, data are not fixed. Data always in changing mode.
  - In operational system, data can change like,
    - Edit, Delete, Update.

Types and Info of Systems…

- Examples : Order entry, inventory, Employers salary, Customers review, Customers detail etc.
- Because of their importance to the organization, operational systems were almost-always the first part of the enterprise to be computerized.

Def.01 Create a structure of Database and Tables for Informational system. With fields, data types, size, keyField.

- Database Name : SchoolMast
  - Tables : YearMaster
    - UserType
    - User
    - Departments
    - Class
    - Subjects
    - ExamMaster
    - StudentData
    - StaffData
    - Attendance
    - Complaints
    - Activities
    - Reminder

Def.02 Write Queries to add DATA (Insert Query) in given tables from Def.01.

- Write Insert Queries for ALL the Tables.

Def.03 Write Queries to Update DATA (Update Query) in given tables from Def.01.

- Write various Update Queries to update date in tables.
- First note the update details and then write Queries to update that data.

Types and Info of Systems…

- Informational System
  - In this type data stored in strategic (વ્યૂહાત્મક) format.
  - It means that informational system is used to store preplanned data.

Types and Info of Systems…

- Example : Sales, purchase, buffer stock information, seasonal stock information, marketing view information etc.
- Informational system can only store readable data. It means we only can read data.

Business Intelligence :

- Business intelligence is a collection of activities to understand business situations by performing.
- Various types of analysis on the company data as well as on external data from third parties to help make strategic, tactical(વ્યૂહાત્મક), and operational business decisions and take necessary actions for improving business performance.

Business Intelligence :

- This includes gathering, analyzing, understanding, and managing data about operation performance, customer and supplier activities, financial performance, market movements, competition, regulatory compliance (પાલન), and quality controls.

Business Intelligence :

- Examples of business intelligence are the following:
  - Business performance management, including producing key performance indicators such as
    - Daily sales, resource utilization, and main operational costs for each region, product line, and time period, as well as their aggregates, to enable people to take tactical actions to get operational performance on the desired tracks.

Reporting

- In a data warehousing context, a report is a program that retrieves data from the data warehouse and presents it to the users on the screen or on paper.
- Users also can **subscribe** to these reports so that they can be sent to the users automatically by e-mail at certain times or in response to events.

Reporting

- The reports are built according to the functional specifications. They display the DDS data required by the business user to analyze and understand business situations.
- The most common form of report is a tabular form containing simple columns.
- There is another form of report known as cross tab or matrix.

Reporting

- These matrix reports are like Excel pivot tables, where one data attribute becomes the rows, another data attribute becomes the columns, and each cell on the report contains the value corresponding to the row and column attributes.
- Data warehouse reports are used to present the business data to users, but they are also used for data warehouse administration purposes. They are used to monitor data quality, to monitor the usage of data warehouse applications, and to monitor ETL activities.

Online Analytical Processing (OLAP)

- OLAP is the activity of interactively analyzing business transaction data stored in the dimensional data warehouse to make tactical and strategic business decisions.
- Typical people who do OLAP work are business analysts, business managers, and executives.
- Typical functionality in OLAP includes aggregating (totaling), drilling down (getting the details), and slicing and dicing.

Online Analytical Processing (OLAP)

- OLAP functionality can be delivered using a relational database or using a multidimensional database.
- OLAP that uses a relational database is known as relational online analytical processing (ROLAP).
- OLAP that uses a multidimensional database is known as multidimensional online analytical processing (MOLAP).

Online Analytical Processing (OLAP)

- An example of OLAP is analyzing the effectiveness of a marketing campaign initiative on certain products by measuring sales growth over a certain period.
- Another example is to analyze the impact of a price increase to the product sales in different regions and product groups at the same period of time.

Data Mining

- Data mining is a process to explore data to find the patterns and relationships that describe the data and to predict the unknown or future values of the data.
- The key value in data mining is the ability to understand why some things happened in the past and to predict what will happen in the future.
- When data mining is used to explain the current or past situation, it is called descriptive analytics. When data mining is used to predict the future, it is called predictive analytics.

Data Warehousing Today

- Today most data warehouses are used for business intelligence to enhance CRM (Customer Relationship Management) and for data mining.
- Some are also used for reporting, and some are used for data integration. These usages are all interrelated;
- For example,
  - Business intelligence and CRM use data mining, BI uses reporting, and BI and CRM also use data integration.

Business Intelligence

- It seems that many vendors prefer to use the term business intelligence rather than data warehousing.
- In other words, they are more focused on what a data warehouse can do for a business.

Business Intelligence

- As we discussed previously, many data warehouses today are used for BI. That is the purpose of a data warehouse is to help business users understand their business better; to help them make better operational, tactical, and strategic business decisions; and to help them improve business performance.

Business Intelligence

- Many companies have built business intelligence systems to help these processes, such as understanding business processes, making better decisions (through better use of information and through data-based decision making), and improving business performance (that is, managing business more scientifically and with more information).

Business Intelligence

- These systems help the business users get the information from the huge amount of business data.
- These systems also help business users to understand the pattern of the business data and predict future behavior using data mining.
- Data mining enables the business to find certain patterns in the data and forecast the future values of the data.

Business Intelligence

- Almost every single aspect of business operations now is touched by business intelligence:
  - Call center, supply chain, customer analytics, finance, and workforce.
- Almost every function is covered too:
  - Analysis, reporting, alert, querying, dashboard, and data integration.
  - A lot of business leaders these days make decisions based on data.

Business Intelligence

- A business intelligence tool running and operating on top of a data warehouse could be an invaluable support tool for more information that purpose.
- This is achieved using reports and OLAP. Data warehouse reports are used to present the integrated business data in the data warehouse to the business users. OLAP enables the business to interactively analyze business transaction data stored in the dimensional data warehouse.

CRM (Customer Relationship Management)

- We discussed CRM earlier in this chapter. A *customer* is a person or organization that consumes your products or services.
- In non business organizations, such as universities and government agencies, a customer is the person who the organization serves.
- A CRM system consists of applications that support CRM activities.

Data mining

- Data mining is a field that has been growing fast in the past few years. It is also known as *knowledge discovery*, because it includes trying to find meaningful and useful information from a large amount of data.
- It is an interactive or automated process to find patterns describing the data and to predict the future behavior of the data based on these patterns.

Data mining

- Data mining systems can work with many types of data formats:
  - **Various types of databases** (relational databases, hierarchical databases, dimensional databases, object-oriented databases, and multidimensional databases), **files** (spreadsheet files, XML files, and structured text files), **unstructured or semi structured data** (documents, e-mails, and XML files), [Continue…]

Data mining

- Data mining systems can work with many types of data formats:
  - **stream data** (plant measurements, temperatures and pressures, network traffic, and telecommunication traffic), **multimedia files** (audio, video, images, and speeches), **web sites/pages**, and **web logs**.

Data mining

- You can use data mining for various business and non business applications including the following:
- Finding out which products are likely to be purchased together, either by analyzing the shopping data and taking into account the purchase probability or by analyzing order data.
- Shopping (browsing) data is specific to the online industry, whilst (જે સમયે) order data is generic to all industries.

Data mining

- In the railway or telecommunications area, predicting which tracks or networks of cables and switches are likely to have problems this year, so you can allocate resources (technician, monitoring, and alert systems, and so on) on those areas of the network.

Master Data Management (MDM)

- To understand what master data management is, first we need to understand what master data is.
- In OLTP systems, there are two categories of data:
  - Transaction data and
  - Master data

Master Data Management (MDM)

- Transaction data
  - Transaction data consists of business entities in OLTP systems that record business transactions consisting of identity, value, and attribute columns.
- Master data.
  - Master data consists of the business entities in the OLTP systems that describe business transactions consisting of identity and attribute columns.

Master Data Management (MDM)

- Transaction data is linked to master data so that master data describes the business transaction.
- Let’s take the classic example of sales order–processing first and then look at another example in public transport.

Master Data Management (MDM)

- An online music shop with three brands has about 80,000 songs.
- Each brand has its own web store:
  - Energize(પ્રોત્સાહિત કરવું) is aimed at young people.
  - Ranch is aimed at men, and
  - Essence is aimed at women.
- Every day, thousands of customers purchase and download thousands of different songs.
- Every time a customer purchases a song, a transaction happens. All the entities in this event are master data.

Customer Data Integration

- Customer data integration (CDI) is the MDM for customer data. CDI is the process of retrieving, cleaning, storing, maintaining, and distributing customer data.
- A CDI system retrieves customer data from OLTP systems, cleans it, stores it in a customer master data store, maintains the customer data, keeps it up-to-date, and distributes the customer data to other systems.

Future Trends in Data Warehousing

- Several future trends in data warehousing today. They are unstructured data, search, service oriented Architecture, and real-time data warehousing.

Unstructured Data

- Data that is in databases is structured;
  - It is organized in rows and columns. Structured data; The source system is a database.
  - It can be a relational database (tables, rows, and columns), and it may be an object-oriented database (classes and types) or a hierarchical database (a tree-like structure). However, they all have data structure.

Unstructured Data

- Unstructured data,
  - On the other hand, does not have a data structure such as rows and columns, a tree-like structure, or classes and types.
- Examples of unstructured data are,
  - Documents, images (photos, diagrams, and pictures), audio (songs, speeches, and sounds), video (films, animations), streaming data, text, e-mails, and Internet web sites.

Unstructured Data

- Arguably,
  - Some people say this kind of data is semi structured data, with the argument that there is some structure, so it has attributes.
  - For example,
    - An e-mail has attributes such as from, to, sent date, date of creation, date of received, subject, and body; a document has attributes such as title, subject, author, number of pages, number of words, and last-modified date.

Unstructured Data

- For example,
  - Let’s say that you have 1 million e-mails as your unstructured data.
  - They have attributes, such as from, to, cc, bcc, subject, date created, date sent, attachments, number of words in the body, host address, originator address, recipient address, and so on.
  - You then store these attributes in a relational table, and then e-mails are stored as files with the file name and location stored in the table.

Search

- This section answers the second question, how do you get the information out?
- The answer is by searching. To get the information out of structured data, provided that you know the structure,
- You can do a select query, whether using a static report or manual interactive ad hoc queries.

Search

- If you use a BI application, the application can go through the metadata and display the structure of the data and then assist you in navigating through the data to retrieve the information you need.
- To get the information out of unstructured data, especially text data such as documents, e-mails, and web pages, you do a search.

Search

- Like on the Internet, the search engine has already crawled the data warehouse and indexed the unstructured data.
- The search engine has categorized the unstructured data based on their types and their properties and, in the case of web pages, their links.

Search

- You can now type what you want to find in a search box, and the search engine will go through its index, find the locations of the information, and display the results.
- It can also offer predefined searches, wrapped in a nice hierarchical structure for you to navigate and choose. It can also memorize user searches that could assist you in defining what to type when searching.

Service-Oriented Architecture (SOA)

- SOA is a method of building an application using a number of smaller, independent components that talk to each other by offering and consuming their services.
- These components can be distributed; in fact, they can be located on different sides of the world.

Service-Oriented Architecture (SOA)

- Almost every large application can benefit from an SOA approach.
- You don’t build one giant application anymore. Instead, you build many smaller pieces that talk to each other.
- It is the nature of the IT industry that applications will need to be replaced every several years. It could be because of obsolete technology or because of the functionality. Bankruptcy, mergers, and takeovers are also the other drivers to this.

Real-Time Data Warehouse

- A data warehouse, a few years ago, was usually updated every day or every week.
- In the past two to three years, there has been more and more demand to increase the frequency of updates. The users want to see the data in the data warehouse updated every two minutes or even in real time.
- A real-time data warehouse is a data warehouse that is updated (by the ETL - Extract, Transform and Load) the moment the transaction happens in the source system.

Real-Time Data Warehouse

- For example,
  - You can put triggers on the sales transaction table in the source system so that whenever there is a transaction inserted into the database, the trigger fires and sends the new record to the data warehouse as a message.
  - The data warehouse has an active listener that captures the message the moment it arrives, cleanses it, DQs it, transforms it, and inserts it into the fact table immediately.

Data Warehouse Architecture

Data Warehouse Architecture

- A data warehouse system has two main architectures:
  - The data flow architecture and
  - The system architecture.
  - The data flow architecture is about how the data stores are arranged within a data warehouse and how the data flows from the source systems to the users through these data stores.
  - The system architecture is about the physical configuration of the servers, network, software, storage, and clients.

Data flow architecture

- In data warehousing, the data flow architecture is a configuration of data stores within a data warehouse system, along with the arrangement of how the data flows from the source systems through these data stores to the applications used by the end users.
- This includes how the data flows are controlled, logged, and monitored, as well as the mechanism to ensure the quality of the data in the data stores.

Data flow architecture

- Data stores are important components of data flow architecture. We’ll begin the discussion about the data flow architecture by explaining what a data store is.
- A data store is one or more databases or files containing data warehouse data, arranged in a particular format and involved in data warehouse processes.

Data flow architecture

- Based on the user accessibility, you can classify data warehouse data stores into three types:
  - A user-facing data store is a data store that is available to end users and is queried by the end users and end-user applications.
  - An internal data store is a data store that is used internally by data warehouse components for the purpose of integrating, cleansing, logging, and preparing data, and it is not open for query by the end users and end-user applications.

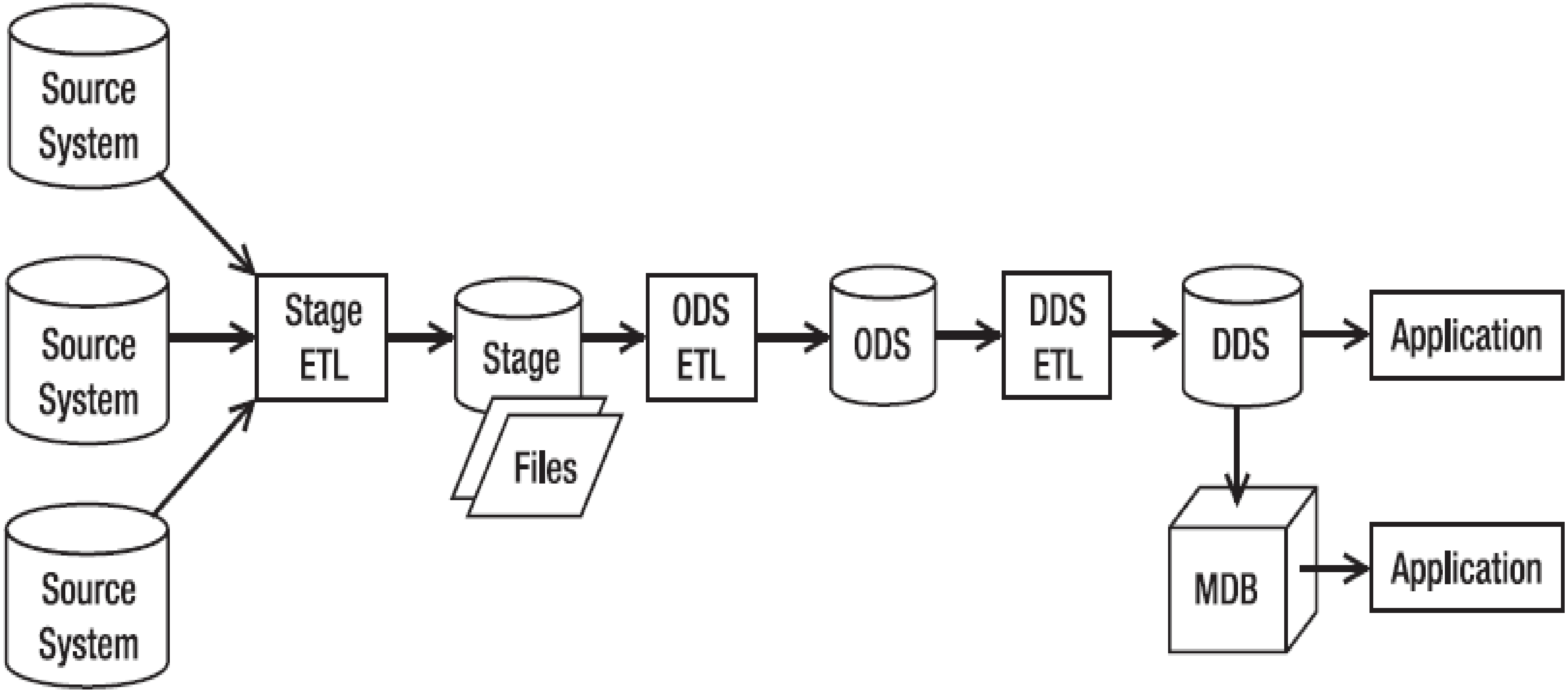
Data flow architecture

- A hybrid data store is used for both internal data warehouse mechanisms and for query by the end users and end-user applications.
- A normalized data store (NDS) is an internal master data store in the form of one or more normalized relational databases for the purpose of integrating data from various source systems captured in a stage, before the data is loaded to a user-facing data store.

Data flow architecture

- An operational data store (ODS) is a hybrid data store in the form of one or more normalized relational databases, containing the transaction data and the most recent version of master data, for the purpose of supporting operational applications.
- A dimensional data store (DDS) is a user-facing data store, in the form of one or more relational databases, where the data is arranged in dimensional format for the purpose of supporting analytical queries.

Data flow architecture with four data stores: stage, ODS, DDS, and MDB.



Data flow architecture with four data stores: stage, ODS, DDS, and MDB.

- A data flow architecture is one of the first things you need to decide when building a data warehouse system because the data flow architecture determines what components need to be built and therefore affects the project plan and costs.
- The data flow architecture shows how the data flows through the data stores within a data warehouse.

Data flow architecture with four data stores: stage, ODS, DDS, and MDB.

- The data flow architecture is designed based on the data requirements from the applications, including the data quality requirements.
- Data warehouse applications require data in different formats. These formats dictate the data stores you need to have.
- If the applications require dimensional format, then you need to build a DDS.

Data flow architecture with four data stores: stage, ODS, DDS, and MDB.

- If the applications require a normalized format for operational purposes, then you need to build an ODS.
- If the application requires multidimensional format, then you need to build an MDB.
  - Once you determine the data stores you need to build, you can design the ETL to populate those data stores. Then you build a data quality mechanism to make sure the data in the data warehouse is correct and complete.

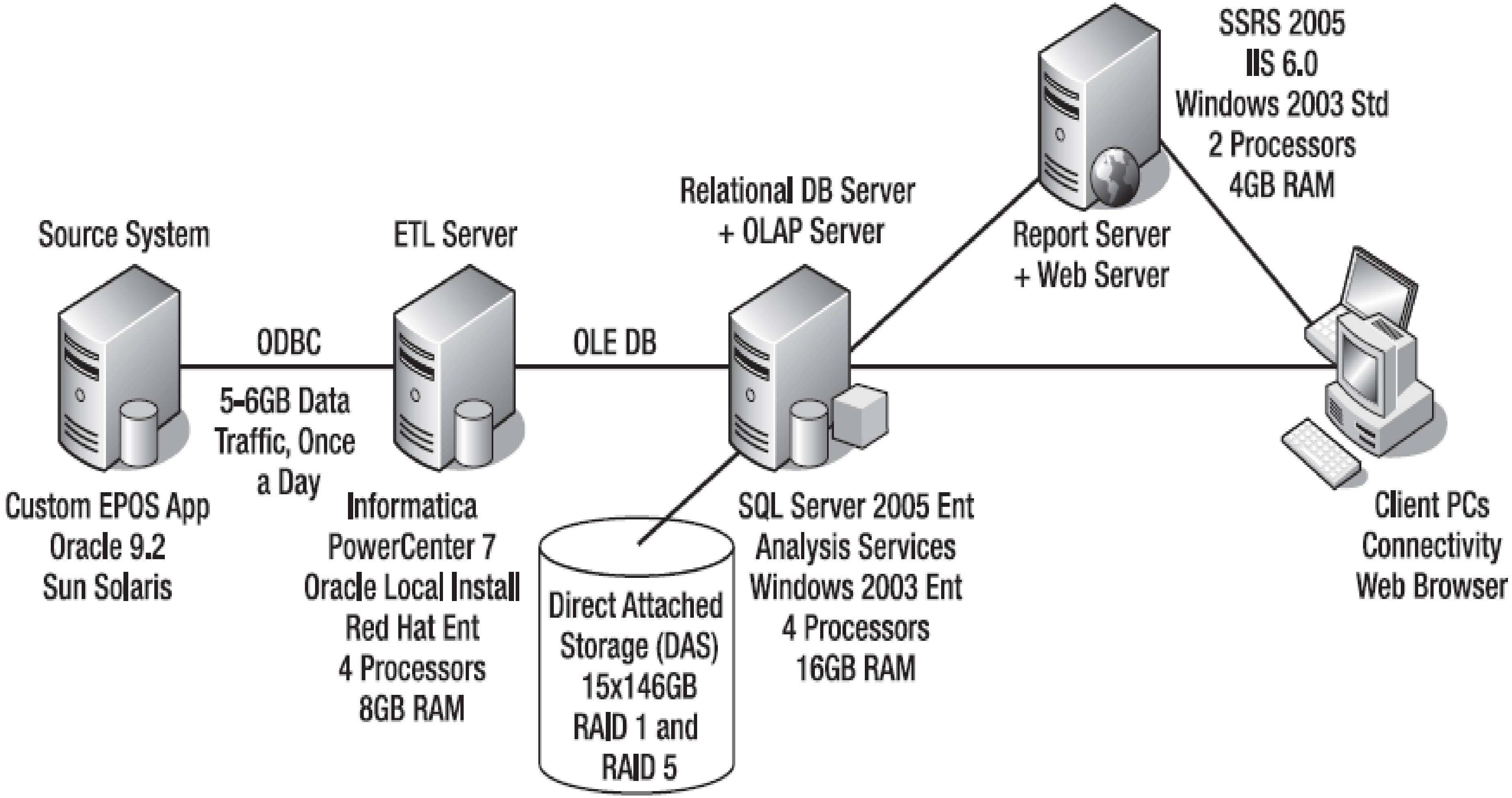
System Architecture

- Data flow architecture show how the data is arranged in the data stores and how the data flows within the data warehouse system.
- Once you have chosen a certain data flow architecture, then we need to design the system architecture, which is the physical arrangement and connections between the servers, network, software, storage system, and clients.

System Architecture

- Designing a system architecture requires knowledge about hardware (especially servers), networking (especially with regard to security and performance and in the last few years also fiber networks), and storage (especially storage area networks [SANs], redundant array of inexpensive disks [RAID], and automated tape backup solutions).

System Architecture DIAGRAM



System Architecture DIAGRAM

- We choose this example because it represents a typical architecture for a medium system.
- We have a dedicated ETL server, separated from the database server.
- It’s a medium size of data; the raw capacity of 2TB is about 400GB to 500GB final usable database space, assuming, we have both development and production environments.

System Architecture DIAGRAM

- The platform is a bit of a mixture, as typically found in organizations: the source system and ETL are not Microsoft.
- The Informatica was probably already there when the data warehouse project started, so they have to use what they already have. Therefore, you can create a system architecture with different platforms.

Case Study :

- The dev environment can be of lower specifications than production.
- For example,
  - In the Amadeus Entertainment case study, dev can be just one server; in other words, the database engine, Integration Services, Analysis Services, and Reporting Services are all installed in one server.

Case Study :

- The most difficult questions to answer when designing the system architecture are about sizing and/or capacity. In particular, infrastructure suppliers usually ask these two questions:
  - How much disk space do you require (how big is your data)?
  - How much processing power do you need (in terms of processor and memory)?