Data and Information

raw facts that describe the characteristics of an event

Data converted into a meaningful and useful context

Data Base: maintains info about various types of objects, events, people and places

Importance of Data Management

Data Management: structured approach for capturing, storing, processing, integrating, distributing securing and archiving data effectively through its life cycle

Goal of Data Management: Provide infrastructure & tools to transform raw data into useful HQ info.

Data Life Cycle:
1. **Data** Sources and Databases
2. Data **Storage**
3. Data **Analysis**
4. **Results**
5. **Solutions**

Why Data management matters?
Not effective unless HQ data is accessible when needed

Challenges:
- Volume of data
- Data is scattered all over organization
- Data is created and used offline without quality control checks
- Redundancy and out of date data calls for a lot of maintenance

Principles of Data Management:
- **Principle of Diminishing data Value:** the more recent the info, the more valuable .
- **Principle of 90/90 data use:** 90% of data is seldom accessed after 90 days
- **Principle of Data in context:** Investment in DM infrastructure may be huge

Traditional File Organization

Problems
- Data Redundancy
- Data Inconsistency
- Data Isolation, Lack of Data Sharing and Availability
- Poor Security

File Organization
- **File:** Group of records of same type
- **Record:** group of related fields
- **Field:** groups of words or a complete number
- **Byte:** groups of bits that represent a single character
- **Bit:** smaller unit of data (binary digit)

Entity: Person/place/event which info is stored
Entity Class: Collection of similar entities
Attributes: characteristics or properties of an entity class

Database Management Systems

Relational DBMS: represents data as 2D tables

- **Software package to create and maintain databases**
- Acts as an interface between application and data files
- Allows organization to centralized data

Primary Key: field that uniquely IDs entity classes in DB
Foreign Key: primary key of one table that appears as an attribute in another table and acts to provide logical relationship between 2 tables

3 Basic Operations of Relationship DBMS
Select: Creates Subset of rows that meet specific criteria
Join: combines relational tables to provide users with info
Project: Enables users to create new tables containing revelvant info

Functions of DBMS: Data filtering and profiling, Data Quality, Data Synchronization, Data enrichment, Data maintenance

Data Quality

Information Cleansing and Scrubbing
A process that weeds out and fixes or discards inconsistent, incorrect and incomplete information

Business Impacts from low quality data
- Inability to accurately track customers
- Difficulty identifying valuable customers
- Inability to identify selling opportunities
- Marketing to nonexistent customers
- Difficulty tracking revenue due to inaccurate invoices
- Inability to build strong customer relationships

Data Warehouse

Model of an Enterprise Data Warehouse
Data from various sources are extracted, transformed & loaded (ETL) into data warehouse; the used to support functions and apps throughout the enterprise

- For organizational learning purposes, data from **many sources** & **many time periods** must be gathered together and organized in a consistent and useful way
- A data warehouse is a **copy of transaction data specifically structured** for querying analysis reporting and more rigorous data mining
- Note that the data warehouse contains a **copy of the transactions which are not updated or changed** later by the transaction system
- A massive database that store **current and historical** data
- Data are standardized into a common data model
- Data warehousing is a **data store and a process** for bringing together disparate data form throughout an organization for decision-support purposes
- Consolidates data for management analysis and decision making
- RDBMS are often used for data warehousing

Data Centers

Data Centers
- Host and integrate networks, computer systems and storage devices
- High reliability and availability
- High security

Distributed Database
- Database is stored in 1+ physical location
- **Reduces vulnerability**
- **Increase** service and **responsiveness** to local users
- Can often run on **smaller, less expensive computers**
- Depend on a HQ telecommunications lines

Two Types of Databases
- Centralized & Distributed

Data Centers
- **Partitioned Database**
  - Parts of database stored in different physical locations
- **Replicated Database**
  - Duplicate entire database at all remote locations

Data Mart
- Subset of data warehouse
- Contains a summarized or highly focused portion of data for a **specified business unit** or group of users

Data Warehouse Fundamentals
- Extraction, Transformation and Loading (ETL)
  - Process that extracts info from databases, transforms the information using a common set of enterprise definitions, and loads the information into a data warehouse
- Data mart

Suitability of Data Warehouse
- Data warehousing is most appropriate for organizations with the following characteristics
  - End users need access to large amount of data
  - Operational data are stored in different systems
  - Large client base
  - Same data represented differently in different systems
  - Extensive end-user computing is performed

Document Management

Business Records
- Contracts, research and development accounting, memos, meeting minutes, customer communications

Document Management
- Automated control of imaged and electronic documents, spreadsheet, emails, word processing docs, voice and other documents through their life cycle from initial creation to final archiving or destruction

Document Management Systems
- Hardware and software to archive e-documents, convert paper documents to e-documents, then index and store them according to company policy
- DMS increases productivity and efficiency by:
  - Cutting labour costs by automating business processes
  - Faster search in documents to support decision making
  - Minimizing the cost associated with printing and storing documents
  - Improving security of the contents

Enterprise Content Management

ECM includes:
- Electronic document management
- Web content Management
- Digital Asset Management
- Electronic Records Management (ERM)

ERM Vendors
- Oracle
- AccuTrac
- Hummingbirg

Unstructured Business Records:
- Business generate volumes of documents, messages, and memos that, by their nature, contain unstructured context that cannot be put into a database
- Many materials are business records that must be retained for legal auditing
- Must be organized and indexed for easy retrieval
- Records are not needed for current operations or decisions are archived, moved into longer term storage

Business Value of E-Records Management
- Companies need to be prepared to respond to an audit, investigation, etc
- **E-Discovery**
  - The process of gathering electronically stored information in preparation for trial, legal or regulatory investigation.
- **E-Request**
  - When received, a company must produce what is requested or will face charges of obstructing justice or being in contempt of court.

ECM Vendors
- IBM
- Oracle
- EMC

Data Usage and Life Cycle

- Data Sources and Databases
- Data Storage
- Data Analysis
- Results
- Solutions

Multidimensional Data Analysis

- Also called Online Analytical Processing
- Interactive, Exploratory analysis of multidimensional data from multiple dimensions

OLAP Cube
- Cubes are stored in relational databases

Data Mining

- Tools for **deep down** analysis of large pools of data
  - To find **hidden patterns**
  - To **predict** future behaviour
  - To **infer rules** to guide decision making
- Data mining techniques make use of data in a Data Warehouse
- Concern of Privacy

Association Detection
- Reveals the degree to which variables are related and the nature and frequency of these relationships in the information
- **Market Basket Analysis**
  - Analyzes such items as Web sites and checkout scanner information to detect customers' buying behaviour and predict future behaviour by identifying affinities among customers' choices of products and services

Data Mining Techniques
- Uses a variety of techniques to find patterns and relationships in large volumes of information and infer rules that predict future behaviour and guide decision making
- Common forms of data-mining analysis capabilities include
  - Classification
  - Cluster Analysis
  - Association Detection

Classification
- Classes are pre-defined
- Assign each data point to one class
- Examples
  - Grades
  - Consumer Credit Assessment
  - Fraud detection
  - Medical Treatment

Clustering
- A technique to divide an info set into mutually exclusively groups such that the members of each group are as close together as possible to one another and the different groups as far apart as possible
- CRM systems depend on cluster analysis to segment customer information and identify behavioural trait

Data Mining and Target Marketing
- Exclusion according to geography, age, ability to pay, need for product, etc.
- Data mining can help in prospecting:
  - Identifying goods and prospects
  - Choosing appropriate communication channels
  - Picking suitable messages

1. Exploration (word counts, creating topics)
2. Pre-processing (abbreviations, errors)
3. Modeling (neural networks, decision trees)

- Interpreting words and concepts in context
- Discover of trends and patterns from textual info

Text Mining

Data Mining and Customer Relationship Management
- Matching campaigns to customers
  - Cross sell campaign, upsell , loyalty program
- Customer Segmentation
  - Finding behavioural segments
- Reducing exposure to credit risk
  - Predicting who will default