**Data Management**

**2.1 Introduction :**

**What is Data Management?**

Data management is a platform wherein enterprises are gathering analyzing and storing data for its user.

Data management is a solution that make processing, validation, and other essential functions simpler and less time intensive.

e.g. Organizations and enterprises using Big Data to inform business decisions and gain deep insights into customer behavior, trends, and opportunities for creating extraordinary customer experiences.

**What is data science ?**

**Data science** is a multi-disciplinary field that uses **scientific** methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured **data. data science** is termed as statistics.



**Data Management Challenges :**

1. Simply collecting data is not enough; enterprises and organizations need to understand from the start that data management and [data analytics](http://www.ngdata.com/s/big-data-analytics/) only will be successful
2. Each step of data collection and management must lead toward acquiring the right data and analyzing it in order to get the actionable intelligence necessary for making truly data-driven business decisions.

In an organization data management should be done perfectly as it leads to optimization of various processes in the organization. A few of the best practices that an organization should have for proper data management are :

1. There should be simplified access to the traditional as well emerging data
2. Data should be cleaned or preprocessed in order to use in business decision process.
3. Data should be visualized using advanced visualization techniques.
4. Shape data using flexible manipulation techniques

Data Management is a platform through which organizations gather, sort and store their information, process it using Data Science or Data Analytics techniques and then repackage it in visualized ways that are useful to the marketers.

**2.2 Data collection**

Data collection involves gathering and measuring data on a particular subject in a structured way.

The data collected can be used to answer pre-specified questions

Data collection involves gathering and measuring data on a particular subject in a structured way.

The data collected can be used to answer pre-specified questions

Data collection involves gathering and measuring data on a particular subject in a structured way.

The data collected can be used to answer pre-specified questions

Data collection involves gathering and measuring data on a particular subject in a structured way. The data collected can be used to answer pre-specified questions, test for differences between different groups of respondents and evaluate the outcome of a change.

Examples of data collection in business:

* Manufacturers might collect data to see how well the processes in their factories are running and to identify where improvement could be made.
* Marketing professionals might collect data to identify customer trends and allow them to target customers more effectively.
* Human resource professionals might collect data to help them to understand and evaluate the effect of their staff on the business.
* Finance professionals might collect data to understand and analyses trends in markets.

There are two types of data collection methods

1. Primary data collection methods
2. Secondary data collection methods
3. **Primary data collection methods**

Primary data is also termed as raw data. This type of information is obtained by means of survey, experimentation, observations and not subjected to any processing or manipulation.

Primary data is collected by following methods

1. **Observation method**

In observation method data is collected by watching other persons behavior without controlling it.

Example :Watching the life of street-children provides a detailed description of their social life.

This method is further classified as,

1. **Structured(descriptive) and Unstructured(exploratory) observation-**

For **structured observation**, the researcher specifies in detail what is to be observed and how the measurements are to be recorded, e.g., an auditor performing inventory analysis in a store.

In **Unstructured observation**, the observer does not have any fixed problem definition, the observer monitors all aspects of the phenomenon that seem relevant to the problem at hand ,e.g., observing children playing with new toys.

1. **Participants, Non-participants and disguised observation –**

When the observer observes by making himself more or less, the member of the group he is observing, it is **participant observation**.

When the observer observes by detaching himself from the group under observation it is **non-participant observation**.

If the observer observes in such a manner that his presence is unknown to the people he is observing it is **disguised observation.**

1. **Controlled(laboratory) and uncontrolled(exploratory) observation** –

If the observation takes place in the natural setting it is **uncontrolled observation** but when observation takes place according to some pre-arranged plans, involving experimental procedure it is a **controlled observation**.

Adavantage –

1. Data is not affected by past and future actions.
2. Natural behaviour of the group can be recorded.

Disadvantage –

1. Expressive method
2. Information provided is limited.
3. **Interview method**

An **interview** is a conversation for gathering information. **Interviews** can be conducted face-to-face or over the telephone.

1. **Personal Interview**–It requires a person known an interviewer to ask questions generally in a face to face contact to the other person.

It can be-

Direct personal investigation – The interviewer has to collect the information personally from the services concerned.

Indirect oral examination – The interviewer has to cross examine other persons who are supposed to have a knowledge about the problem.

Structured interviews – Interview involving the use of pre-determined questions and of highly standard techniques of recording.

Unstructured interviews – It does not follow a system of pre-determined questions and is characterized by flexibility of approach to questioning.

Focused interviews- It is meant to focus attention on the given experience of the respondent and its effect. The interviewer may ask questions in any manner or sequence with the aim to explore reasons and motives of the respondent.

Clinical interviews – It is concerned with broad underlying feeling and motives or individual’s life experience which are used as method to ellict information under this method at the interviewer direction.

Non directive interview – The interviewer’s function is to encourage the respondent to talk about the given topic with a bare minimum of direct questioning.

Advantages –

1. More depth information can be obtained.
2. This method provides greater flexibility.
3. Personal information obtained in a well manner.

Disadvantages –

1. Expensive method.
2. More time consuming.
3. High skilled interviewer is required.
4. **Telephonic Interviews**

It requires the interviewer to collect information by contacting respondents on telephone and asking questions or opinion orally.

Advantage –

1. It is flexible, fast and cheaper than other methods.
2. Recall is easy and there is a higher rate of response.
3. No field staff is required.

Disadvantage –

1. Interview period exceed five minutes maximum which is less.
2. Questions have to be short and to the point.
3. Less information can be collected.
4. **Questionnaire**

In this method, a questionnaire is sent(mailed) to the concerned respondents who are expected to read, understand and reply on their own and return the questionnaire.

Essentials of a good questionnaire :

* It should be short and simple
* Questions should proceed in a logical sequence.
* Technical terms and vague expressions must be avoided.
* Adequate space for answers must be provided.
* The physical appearances – quality of paper, colour etc must be good to attract the attention of the respondent.

Advantage –

1. Respondents have adequate time to give answers.
2. Respondents are easily and conveniently approachable
3. Large samples can be used to be more reliable.

Disadvantage –

1. Control over questions is lost once it is sent.
2. It is inflexible once sent.
3. Time taking and slow process.
4. Possibility of ambiguous replies.
5. **Secondary data collection methods**

Secondary data is collected by someone else other than user i.e. data is already analysed by someone else. Common source of secondary data include various published or unpublished data, books, newspapers, magazines, trade journals etc.

A researcher can obtain secondary data from various sources.

Published data are available in :

1. Publications of government.
2. Technical and trade journals.
3. Reports of various business, banks etc.
4. Public records.
5. Statistical or historical documents.

Unpublished data may be found in letters, diaries, unpublished biographies or work.

**2.3 Data cleaning/extraction**

**2.3.1 Data cleaning**

Raw Data -- Data that has been collected but not cleaned. Also called source, primary, or atomic data.

Data Cleaning -- The process of fixing or removing incorrect, incomplete, and irrelevant data from a dataset. Also called data cleansing, preparing, or scrubbing.

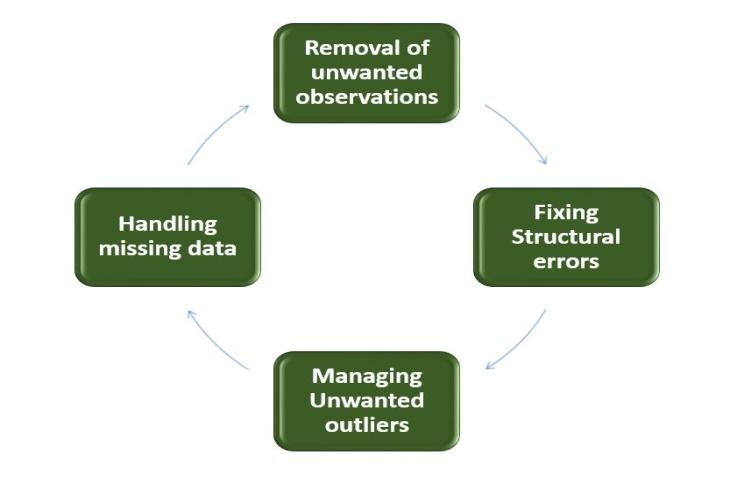
Data cleaning is one of the important parts of machine learning. It plays an important role in building a model. Data cleaning is a form of data management.

Though data cleansing does and can involve deleting information, it is focused more on updating, correcting, and consolidating data to ensure your system is as effective as possible

A very simple algorithm gives desired result if we have a cleaned dataset.

Different types of data will require different types of cleaning.

**Steps involved in Data Cleaning**



1. **Removal of unwanted observations**  
   While collecting data from various data sources, it might happen we will get duplicate data and irrelevant observations that don’t actually fit the specific problem that we are trying to solve, so this step come into existence which includes deleting duplicate/redundant or irrelevant values from your dataset.

* Redundant observations does not provide efficient and faithful results as the data repeats and as well alter the efficiency by a great extent.
* Irrelevant observations are any type of data that is of no use to us and can be removed directly.

1. **Fixing Structural errors**  
   The errors that arise during measurement, transfer of data or other similar situations are called structural errors. Structural errors include typos in the name of features, same attribute with different name, mislabeled classes, i.e. separate classes that should really be the same or inconsistent capitalization.

* For example, the model will treat America and merica as different classes or values, though they represent the same value or red, yellow and red-yellow as different classes or attributes, though one class can be included in other two classes. So, these are some structural errors that make our model inefficient and gives poor quality results.

1. **Managing Unwanted outliers**  
   Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models. Generally, we should not remove outliers until we have a legitimate reason to remove them. Sometimes, removing them improves performance, sometimes not. So, one must have a good reason to remove the outlier, such as suspicious measurements that are unlikely to be the part of real data.
2. **Handling missing data**  
   Missing data is one of the big issue in machine learning. We cannot just ignore or remove the missing observation and must be handled carefully as they can be an indication of something important. We can deal with missing data as follows:
3. Missing values or observations can be dropped.
4. Missing values can be predicted from past and current observations.

**Advantages of Data cleaning :**

1. Improves the Efficiency of Customer Acquisition Activities

### **Improves Decision Making Process**

### **Increases Productivity**

### **Increases Revenue**

### **Streamlines Business Practices**

**2.3.2 Data Extraction**

Data Extraction is the process involves extracting the data from the various source systems.

Common data source formats are:

1. Relational databases and flat files(MS-Word, MS-Excel etc.).
2. non-relational database structures such as Information Management System (IMS)
3. data structures such as Virtual Storage Access Method (VSAM)or Indexed Sequential Access Method (ISAM)
4. even fetching from outside sources such as through web spidering or screen-scraping

**Extraction methods**

The extraction method we choose is highly dependent on the source system and also from the business needs in the target data warehouse environment.

Basically, there are two types of extraction methods :

### Logical Extraction Methods

logical extraction are of two types:

* Full Extraction

The data is extracted completely from the source system. Since this extraction reflects all the data currently available on the source system, there’s no need to keep track of changes to the data source since the last successful extraction.

* Incremental Extraction

At a specific point in time, only the data that has changed since a well-defined event back in history will be extracted.

### Physical Extraction Methods

Depending on the chosen logical extraction method and the capabilities and restrictions on the source side, the extracted data can be physically extracted by two mechanisms. The data can either be extracted online from the source system or from an offline structure.

There are the following methods of physical extraction:

* Online Extraction
* Offline Extraction

#### Online Extraction

The data is extracted directly from the source system itself. The extraction process can connect directly to the source system to access the source tables themselves.

#### Offline Extraction

The data is not extracted directly from the source system but is staged explicitly outside the original source system.

**2.4 Data analysis & Modeling**

**2.4.1 Data Analysis**

Data Analysis is a process of inspecting, cleaning, transforming and modeling data in order to use for discovering useful information, suggesting conclusions and supporting decision-making.

## Types of Data Analysis

There are several data analysis techniques exist which includes various domains such as business, science, social science, etc. with a variety of names. The major data analysis approaches are −

* Data Mining

Data Mining is the process of analysis of large amount of data to extract previously unknown, interesting patterns of data, unusual data and the dependencies. The goal of data mining is the extraction of patterns, knowledge and hidden information from large amounts of data and not the extraction of data itself.

Data mining analysis involves computer science methods at the intersection of the artificial intelligence, machine learning, statistics, and database systems.

Data mining analysis uses machine learning techniques like classification, clustering, Association rule, Time series analysis, Prediction etc.

* Business Intelligence

Business Intelligence is an umbrella that combines architecture tools, databases, analytical tools, applications and methodologies.

Business Intelligence techniques and tools are for acquisition and transformation of large amounts of unstructured business data to help identify, develop and create new strategic business opportunities.

The objective of Business Intelligence is to give business managers and analyst, the ability to conduct appropriate analysis for business purpose.

* Statistical Analysis

Statistics is the study of collection, analysis, interpretation, presentation, and organization of data. In data analysis, two main statistical methodologies are used −

* **Descriptive statistics** − In descriptive statistics, data from the entire population or a sample is summarized with numerical descriptors such as
* Mean, Standard Deviation for Continuous Data
* Frequency, Percentage for Categorical Data
* **Inferential statistics** − It uses patterns in the sample data to draw inferences about the represented population or accounting for randomness. These inferences can be −
* answering yes/no questions about the data (hypothesis testing)
* estimating numerical characteristics of the data (estimation)
* describing associations within the data (correlation)
* modeling relationships within the data (E.g. regression analysis)
* Predictive Analytics

Predictive Analytics use statistical models to analyze current and historical data for forecasting (predictions) about future or otherwise unknown events. In business, predictive analytics is used to identify risks and opportunities that aid in decision-making.

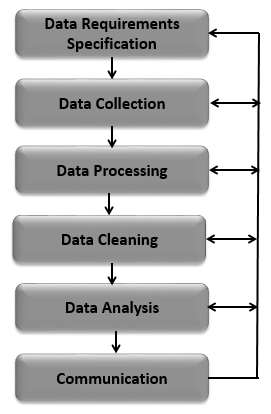
* Text Analytics

Text Analytics, also referred to as Text Mining or as Text Data Mining is the process of deriving high-quality information from text.

**Text analytics** is the process of drawing meaning out of written communication. In a customer experience context, **text analytics** means examining **text** that was written by, or about, customers. You find patterns and topics of interest, and then take practical action based on what you learn.

Data Analysis Process consists of the following phases −

* Data Requirements Specification
* Data Collection
* Data Processing
* Data Cleaning
* Data Analysis
* Communication



## Data Requirements Specification

The data required for analysis is based on a question or an experiment.

Based on the analysis requirement, the input data has been selected (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical.

## Data Collection

Data Collection is the process of collecting information about targeted variables identified as data requirements. The data collected need to be accurate, pure and consistent data, which ensures that related decisions are valid.

Data is collected from various operational and organizational data sources.

The data collected is not in pure format, may not be structured and may contain irrelevant, missing as well redundant data. Hence, the collected data is need to be subjected to data processing and data cleaning.

## Data Processing

In order to use collected data for analysis purpose, data need to go under processing. This includes :

1. Structuring the data as required for analysis.
2. Remove redundancy.
3. Missing data must be supplied or predicted or removed.
4. Erroneous data may be corrected or removed.
5. Remove false information.

## Data Cleaning

The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors. There are several types of Data Cleaning that depend on the type of data. For example, while cleaning the financial data, certain totals might be compared against reliable published numbers or defined thresholds. Likewise, quantitative data methods can be used for outlier detection that would be subsequently excluded in analysis.

## Data Analysis

Data that is processed, organized and cleaned is ready for the analysis. Various data analysis techniques are available to understand, interpret, and derive conclusions based on the requirements. Data Analysis Techniques includes :

* Clustering analysis

Clustering Analysis refers to the process of detecting data sets with similar attributes to learn their similarities as well as differences in the data.

* Association analysis

Association Analysis will allow the business to discover relevant associations between different variables in a large-scale database.

* Regression Analysis

Regression Analysis determine the dependency between attributes. There is an assumption of a single-way causal effect from one attribute to the response of another attribute.

* Classification Analysis

It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

Data Visualization

How the data analysis results are presented to the users, is extremely important because usefulness of the result is dependent on it. For this, Data visualization and GUI strategy are used.

Data Visualization used to examine the data in graphical format.

**Data Visualization** is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts. It helps users in analyzing a large amount of data in a simpler way. It makes complex data more accessible, understandable, and usable.

Visualization techniques includes :

* Graphical
* Icon-based
* Pixel-based
* Hierarchical
* Hybrid
* Line Chart
* Bar Chart
* Pie Chart
* Pivot table
* Scatter Chart

## Communication

The results of the data analysis are to be reported in a format as required by the users to support their decisions and further action. The feedback from the users might result in additional analysis.

**2.4.2 Data Modeling**

Data modeling is the process of creation of data model which specifies how data to be stored in a database. It is used to represent

* Data objects
* The association between different data objects
* The rules

**Types of Data Models**

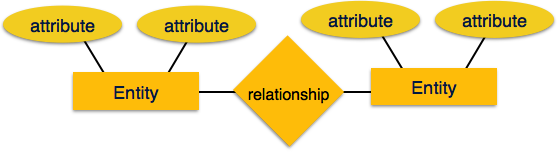
1. **E-R Model**

Entity relationship model is based on real-world entities and relationships among them.

E-R Model is used for conceptual design of database.

ER Model is based on −

* **Entities** and their attributes.
* **Relationships** among entities.



1. **Relational Model**

It is also referred as physical model. This model is used to store data in tables(relations).

Tables are normalized. In normalized tables, values saved are atomic values. Each row in table contains unique values. Each column in a relation contains values from same domain.

