# Data Processing

Prastuti Gupta

September 2021

## Introduction

- Data processing occurs when data is collected and translated into usable information.

However, the processing of data largely depends on the following

- The volume of data that need to be processed.
- The complexity of data processing operations.
- Capacity and inbuilt technology of respective computer system.
- Technical skills.
- Time constraints.

## Different kinds of Data Sources in the world

- Big Data
- Structured, unstructured, semi-structured data
- Time-stamped data
- Machine data
- Spatiotemporal data
- Open data
- Dark data
- Real time data
- Genomics data
- Operational data
- High-dimensional data
- Unverified outdated data
- Translytic Data

## Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. Data lake uses a flat architecture to store data. Each data element in a lake is assigned a unique identifier and tagged with a set of extended metadata tags.

- Source of the data - Data lakes typically receive both relational and non-relational data from IoT devices, social media, mobile apps and corporate applications.
- Users - Data lakes are more useful when an organization needs a large repository of data, but does not have a purpose for all of it and can afford to apply a schema to it upon access.
- Data quality - The data in a data lake is less reliable because it could be arriving from any source in any state. It may be curated, and it may not be, depending on the source.

- Processing - The schema for a data lake is on-read, meaning it doesn't exist until the data has been accessed and someone chooses to use it for something.
- Performance/cost - Data lakes are designed with low cost in mind, but query results are improving as the concept and surrounding technologies mature.

## Different stages of Data Processing

Six stages of data processing -

- Data collection : Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

## Different stages of Data Processing

Six stages of data processing -

- Data collection : Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.
- Data preparation : Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as preprocessing is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data).

- Data input : Data input is the first stage in which raw data begins to take the form of usable information.

- Data input : Data input is the first stage in which raw data begins to take the form of usable information.
- Processing : During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation.

- Data input : Data input is the first stage in which raw data begins to take the form of usable information.
- Processing : During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation.
- Data output/interpretation : It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytic projects.

- Data input : Data input is the first stage in which raw data begins to take the form of usable information.
- Processing : During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation.
- Data output/interpretation : It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytic projects.
- Data storage : After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on.