

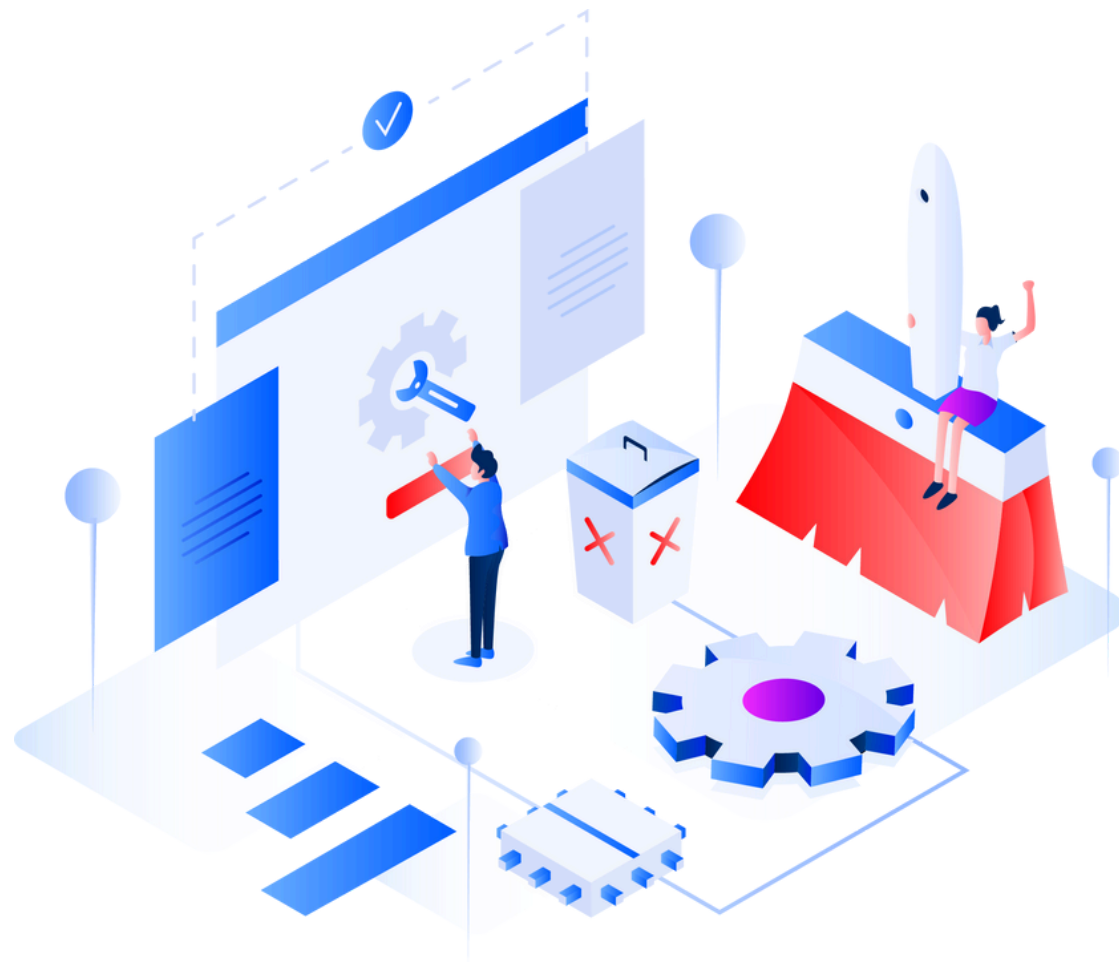
# HIGH QUALITY DATA



وزارة الاتصالات  
وتكنولوجيا المعلومات



# Good Data Sources



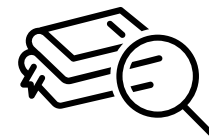
## 01 - Reliable

Accurate, Complete, and unbiased data



## 02 - Original

First or second-party data, or cited third party data



## 03 - Comprehensive

Has all needed knowledge to answer the business questions under study



## 04 - Current

Up to date data

Next Page



# High Quality Data

**1**

**ACCURACY**

The data must conform to actual, real-world scenarios and reflect real-world objects and events.

**2**

**COMPLETENESS**

Completeness measures the data's ability to deliver all the mandatory values that are available successfully.

**3**

**CONSISTENCY**

Data consistency describes the data's uniformity as it moves across applications and networks and when it comes from multiple sources.

# High Quality Data

4

**TIMELINESS**

Timely data is information that is readily available whenever it's needed. This dimension also covers keeping the data current.

5

**VALIDITY**

Data must be collected according to the organization's defined business rules and parameters. The information should also conform to the correct, accepted formats, and all dataset values should fall within the proper range.

# High Quality Data

Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior
AUC2016XC	Aman	30	3.5	Jr

Identify the problems in the dataset





# Data Profiling



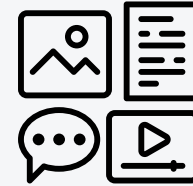
Data profiling is the process of examining, analyzing, and creating useful summaries of data. The process yields a high-level overview which aids in the discovery of data quality issues, risks, and overall trends.

# Types of Data Profiling



## Structure discovery

helps determine whether your data is consistent and formatted correctly. It uses basic statistics to provide information about the validity of data. Sometimes referred to as Data Quality.



## Content discovery

It helps to explore the contents of your data through some basic statistics like MIN, MAX, Range, Quartiles, IQR, Frequency, and distribution.



## Relationship discovery

Relationship discovery identifies connections between different datasets.