Process for identifying data :

1. Determining the information you want to collect (specific information you need, and possible sources for this data).
2. Define a plan for collecting data.
3. Determine your data collection methods ,depends on sources on data, type of data, volume of data).

Reliable data means : Free of errors, accurate, Complete, Relevant, and accessible.

Identifying the right data is a very important step of the data analysis process.

It will ensure that you are able to look at a problem from multiple perspectives.

Data sources can be :

1. Primary

* Refers to information obtained **directly f**rom the source ie. Data from organization HR, CRM or workflow application.
* Data you gather **directly** through surveys, interviews, discussions, observations and focus groups.

1. Secondary

* Refers to information retrieved from existing sources such as external databases, research articles, publications, training material, internet searches, or financial records available as public data.
* Data collected through externally conducted surveys, interviews, discussions, observations and focus groups

1. Third-Party

* Data you purchased from aggregators who collect data from various sources and combine it into comprehensive datasets for purpose of selling the data.

Sources for gathering data :

1. Databases (internal, external)

* Querying language used for extracting information from relational databases.
* Non relational databases can be queried using SQL or SQL-like query tools.
* Structured data can also be stored in NoSQL.
* Semi structured data such as XML, Zipped files, binary executables and TCP/IP Protocols, can be stored in NoSQL clusters.
* XML and JSON(data types for web services) are commonly used for storing and exchanging semi-structured data.

1. APIs

* Are invoked from applications that require the data and access and end-point containing the data.End point can include databases, web service and data marketplaces.
* APIs are also used for data validation (example to validate postal address and zip codes)

1. Website

* Web Scraping, also known as screen scraping or web harvesting is used for downloading specific data from web pages based on defined parameters.
* Web scraping is used to extract data such as text, contact information, , images, videos, podcasts, and product items from a web property.
* RSS feeds are another source typically used for capturing updated data from online forums and news site where data is refreshed on an ongoing basis.

1. Social media and interactive platforms (such as Facebook, Twitter, Google,etc..)

* Unstructured data include from social media, images, videos, documents, media logs, and surveys.
* NoSQL Databases and datalakes provide a good option to store and manipulate large volumes of unstructured data.
* Data lakes can accommodate all data types and schema.
* ETL Tools and data pipeline provide automated functions that facilitate the processing of of importing data.
* Tools suchs as Talend, Informatica, Python and R.

1. Sensor data

* Are popular souce for aggregating constant streams of data flowing from sources such as instruments, IoT devices, Applications and GPS data from cars.
* Data streams and feeds are also used for extracting data from social media sites and interactive platforms.

1. Data exchange (is source of third party data ie. from business applications, location data, social media activity, consumer behavior data)
   * Data exchange platforms allows the exchange of data between data providers and data consumers.
   * Data exchanges have a set of well-defined exchange standards, protocols and formats relevant for exchanging data.
   * Not only facilitate the exchange of data, they alo ensure that security and governance are maintained.
   * They provided data licensing workflows, de-identification and protection of personal information, legal frameworks and a quarantined analytics environment.
   * Examples AWS DataExchange, Snowflake.
2. Surveys (information to a select group of people)
3. Census data
4. Interviews (qualitative data ie. Opinions and experiences)
5. Observation studies (monitoring in a specific environment or while performing a particular task).

Supplementing your primary data with secondary and 3rd party data sources can help you explore problems and solutions in new & meaningful ways.

A diagram of steps and steps

Description automatically generated

In this lesson, you have learned:

* The process of identifying data begins by determining the information that needs to be collected, which in turn is determined by the goal you seek to achieve.
* Having identified the data, your next step is to identify the sources from which you will extract the required data and define a plan for data collection. Decisions regarding the timeframe over which you need your data set, and how much data would suffice for arriving at a credible analysis also weigh in at this stage.
* Data Sources can be internal or external to the organization, and they can be primary, secondary, or third-party, depending on whether you are obtaining the data directly from the original source, retrieving it from externally available data sources, or purchasing it from data aggregators.
* Some of the data sources from which you could be gathering data include databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys and observation studies.
* Data that has been identified and gathered from the various data sources is combined using a variety of tools and methods to provide a single interface using which data can be queried and manipulated.
* The data you identify, the source of that data, and the practices you employ for gathering the data have implications for quality, security, and privacy, which need to be considered at this stage.

1. For data to be considered reliable, it needs to be free of all errors and it needs to be relevant to the use case.
2. Emails are considered a source of semi-structured data as they have some organizational properties but do not follow a rigid schema.
3. Defining a plan for collecting the data you have identified is one of the steps in the “Identifying Data” workflow.
4. Determining the information you need to collect is one of the steps in the “Identifying Data” workflow.

**DATA WRAGLING / DATA MUNGING** is an iterative process that involve data exploration, transformation, validation and making it available for a credible and meaningful analysis. It includes in preparing raw data for a clearly defined purposes, where raw data at this stage is data has been collated through various data sources in a data repository. Data wrangling captures a range of task involved in preparing data for analysis. 4-step process that involves :

1. Discovery / Exploration

Understanding your data better with respect and map the data you have for your case. The objective is to figure out specifically how best you can clean, structure, organize, and mapping your data.

1. Transformation (Data cleaning)

It involves the tasks you undertake to transform the data, such as

* Structuring : Change the form and schema of your data. Incoming data can be in varied formats. JOINS (combine columns) and UNIONS(combine rows) are the most common structural transformations used to combine data from one/more tables.

A table and result of a table

Description automatically generated with medium confidence A diagram of a table and table

Description automatically generated

* Normalizing: focuses on cleaning the database of unused data and reducing redundancy and reducing inconsistencies. Data coming from transactional systems ex. Where a number of insert, update, and deleted operations are performed on an ongoing basis, are highly normalized.
* Denormalizing : include combining data from multiple tables into a single table for faster querying of data for reports and analysis.
* Cleaning data ; Fixing irregularities in data in order to produce a credible and accurate analysis. Data that is inaccurate, missing, or incomplete can skew the results of your analysis and need to be considered. It could also be that the data is biased , or has null values in relevant fields, or have outliers.
* Enriching the data : adding data points that could make your analysis more meaningful.

1. Validation

Where you check the quality of the data post structuring, normalizing, cleaning, and enriching of data. Validation to verifying consistency, quality and security of data you have.

1. Publishing : involves delivering the output of the wrangled data for downstream project needs. What is published is the transformed and validated version of the input dataset along with the metadata about the dataset. It is important to document the steps you took and your consideration for taking those steps to convert the raw data to analysis ready data.

Tools for data Wrangling :

1. MS Excel Power Query / Spreadsheet : Add-ins are available that allow you to import data from several different types of sources and cleaning and transforming data as needed.
2. OpenRefine : Open source tools to import/export data in wide variety formats (CSV,XLS,XML, JSON)
3. Google DataPrep : intelligent cloud data that allows you to visually explore, clean and prepare both structure and unstructured data for analysis. You don’t need to install the software.
4. Watson Studio Refinery : IBM Watson studio, allows you to discover, cleanse and transform data with built-in operations.
5. Python : Numpy, Pandas, Scikit-learn
6. R.

Poor quality data weakens an organization’s competitive standing and undermines critical business objectives. **Missing data/values, inconsistent data and incorrect data** can lead to false conclusions and therefore ineffective decisions, and that can be Costly. Datasets pickup from disaparate sources could have a number of issues, including missing values, inaccuracies, duplicates, incorrect or missing delimiters, inconsistent records, and insufficient parameters. If it cannot be repaired it must be removed from the dataset. **REMEMBER** data cleaning is only a subset of the entire DATA WRANGLING process. But data cleaning forms a very significant and integral part of the transformation phase in data wrangling workflow.

Data cleaning workflow includes : (All the changes undertaken as the part of the data cleaning need to be documented).

* Inspection : To detect the different types of issues and errors that your dataset may have. We can use scripts and tools to define specific rules and constraints and Validate your data against these rules&constraints. Can use data visualization tools for inspection and profiling data to inspect source data (structure, content, and inter relationships).
* Cleaning : Techniques you use, depends on your use case and the type of issues you encounter. Missing value, duplicate data (need to be removed), irrelevant data (not contextual yo your case) , data type conversion (to ensure that values as the data type of that field) , standardize data (to ensure date formats and unit of measurement are standard across the dataset), syntax errors (white space, extra spaces, typos, format need to be fixed), Outliers (values that are different from other observations and need to be examined for accuracy and inclusion in the dataset) – outliers is not about correct/incorrect, but needs to be looked at.
* Verification : Inspect the results to establish effectiveness and accuracy achieved as a result of the data cleaning operation. Need to re-inspect the data to make sure the rules and constraints applicable on the data still hold after the corrections you made.