Identifying Data for Analysis

* Process
  + Step 1: Determine information you want
    - The specific info
    - Diagram

      Description automatically generatedPossible sources.
  + A picture containing diagram

    Description automatically generatedStep 2: Define a plan for collection
  + Graphical user interface

    Description automatically generatedStep 3: Determine collection method
  + Key considerations

A picture containing icon

Description automatically generated

* + Reliability depends on
    - Free of errors
    - Accurate
    - Complete
    - Relevant
    - Accessible
  + Data governance
    - Security
    - Regulation
    - Compliances
  + Data Privacy
    - Confidentiality
    - Complain to mandated regulations
    - License for use

**DATA SOURCES**

* Types (Databases can be sources of all data types)
  + Primary
    - Information obtained directly from the source.
    - Data from CRM, HR, or workflow apps
    - Data gathered from surveys, interview, discussions observations and focus groups.
  + Secondary
    - Information retrieved from existing sources
    - External databases
    - Research articles, publications, training material, internet searches or financial records
    - Survey, interview, discussion, observation, and focus groups.
  + Third-Party
    - Data purchased from aggregators
* Data exchanges
  + Third party data involving the voluntary sharing of data between data providers and consumers.

**How to Gather and Import Data**

* SQL Extracts data with simple commands.
* APIs used for extracting data from variety of data sources
  + Invoked from apps that require the data and access an endpoint containing the data.
  + Also used for data validation.
* Web scraping
  + Downloading specific data from web pages
  + Extract data such as text, contact info, images, video, podcast, and product items.
* Sensor data
  + Data streams are popular source for aggregating constant stream from data
  + GPS etc
* Data Exchanges
  + Exchange between data provider and consumers
  + Have set of well-defined exchange standards, protocols, and formats relevant for exchanging data.
  + Ensure security and governance.
  + Platforms
    - AWS DataExchange, Crunchbase, Lotame, Snowflake
* Importing data
  + Data identified and gathered -> data repository.
  + Destinations
    - Structured data
      * Stored in Well-defined schema
      * Sources include data from oltp systems, spreadsheets, online forms, sensor, network and web logs.
      * Can be store in NoSQL
    - Semi-structured data
      * XML, Zipped files, binary executables and TCP/IP protocols
      * Can be stored in NoSQL clusters
      * XML and JSON are commonly used for storing and exchanging semi-structured data.
    - Unstructured Data
      * Web pages, social

**Summary**

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.

**Data Wrangling**

* AKA data munging
  + Process
    - Data Exploration
      * Discovery
      * Transformation
        + Structing Data

Joins

Combine Columns

Unions

Combine Rows

* + - * + Normalizing and Denormalizing Data

Cleaning unused Data

Reducing Redundancy

Reducing Inconsistency

Denormalizing means to combine data from multiple tables into one for faster querying.

* + - * + Cleaning Data

Fixing irregularities in data to produce a credible and accurate analysis.

* + - * + Enriching Data

Adding data points

* + - * Validation
        + Checking the quality of data
        + Verifying consistency, quality and security of data.
      * Publishing
        + Delivering the output of the wrangled data for downstream projects needs.
      * Documentation
        + Document the steps you took, and you considerations for taking those steps, to converting raw data.
    - Transformation
    - Validation
    - Making it Credible
    - Meaningful Analysis.

**Tools for Data Wrangling**

* Excel Power Query / Spreadsheets
  + Have in built formulae to identify issues, clean, and transform data.
* OpenRefine
  + Open-source tools to import and export data in TSV, CSV, XLS, XNL, and JSON
  + Can clean data, transform, format, and extend data with web services
  + Easy to learn
  + Easy to use
* Google DataPrep
  + Cloud data service
  + Explore clean and prepare both structured and unstructured data
  + Managed service
  + Extremely easy to use
  + Offers suggestions on ideal next steps
  + Detects schemas, data types and anomalies
* Watson Studio Refinery
  + IBM Watson studio
  + Discover, cleanse and transforms data with built in operations
  + Transforms large amount of raw data into consumable, quakily information
  + Flexibility of exploring data from various sources.
  + Detects data types and classification automatically.
  + Enforces data governance.
* Trifactra Wrangler
  + Interactive cloud-based service for cleaning and transforming data.
  + Rearranges unorganized data into data tables.
  + Exportabled to excel, tableau and r
  + Has collaboration features.
* Python
  + Has a huge library and set of packages that offer powerful data manipulation.
  + Jupiter notebook
  + NumPy (Support for large multi-dimensional arrays and matrices and high-level mathematical functions.)
  + Pandas (Fast and easy data analysis operation)
    - Helps prevent common errors that result from misaligned data coming in from different sources.
* R
  + Libraires and packages
  + Dplyr (a powerful library for data wrangling with precise and straightforward syntax.)
  + Data.Table (help aggregate large data sets quickly)
  + Jsonlite (A robust JSON parsing tool, great for interacting with web APIs.)

**Data Cleaning**

* Data sets from disparate sources could have issues with
  + Missing values
  + Inaccuracies
  + Duplicates
  + Incorrect or missing delimiters
  + Inconsistent records
  + Insufficient parameters.
* Data cleaning workflow
  + Inspection
    - Detect issues and errors
    - Validate against rules and constrain
    - Profiling data to inspect source data.
      * Understanding structure, content, inter-relationships for anomalies and data quality issues.
    - Visualizing data using statistical methods.
  + Cleaning
    - Missing values (Unexpected or biased results)
      * Filter out
      * Source missing infor
      * Imputate or calculate missing values.
    - Duplicate data
      * Need to be removed
    - Irrelevant data
      * Data that is not contextual to your use case.
    - Data type conversion
      * Ensure values in a field are stored as the data type of that field.
    - Standardizing data
      * Sure data-time formats, units of measurement
      * Making sure everything is lower case EXP
    - Syntax Errors
      * White spaces
      * Extra spaces
      * Typos
      * Formats
    - Outliers
      * Need to be examined for accuracy and inclusion in the dataset.
  + Verification
    - Inspect results for effectiveness and accuracy
* \*\*\*\*Important to document\*\*\*\*
  + All changes undertaken from the cleaning process
  + Reasons for undertaking.
  + Quality of currently stored data.

**Summary**

In this lesson, you have learned the following information:

Once the data you identified is gathered and imported, your next step is to make it analysis-ready. This is where the process of Data Wrangling, or Data Munging, comes in. Data Wrangling is an iterative process that involves data exploration, transformation, and validation.

Transformation of raw data includes the tasks you undertake to:

* Structurally manipulate and combine the data using Joins and Unions.
* Normalize data, that is, clean the database of unused and redundant data.
* Denormalize data, that is, combine data from multiple tables into a single table so that it can be queried faster.
* Clean data, which involves profiling data to uncover quality issues, visualizing data to spot outliers, and fixing issues such as missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors, and outliers.
* Enrich data, which involves considering additional data points that could add value to the existing data set and lead to a more meaningful analysis.

A variety of software and tools are available for the Data Wrangling process. Some of the popularly used ones include Excel Power Query, Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python, and R, each with their own set of characteristics, strengths, limitations, and applications.