Wrangling Data

What is Data Wrangling

- Data wrangling, also known as data munging, is an iterative process that involves data exploration, transformation, validation, and making it available for a credible and meaningful analysis.
    - It includes a range of tasks involved in preparing raw data for a clearly defined purpose, where raw data at this stage is data that has been collated through various data sources in a data repository.
- The Data Wrangling Process: Data wrangling captures a range of tasks involved in preparing data for analysis. Typically, it is a 4-step process that involves—Discovery, Transformation, Validation, and Publishing.
- The Discovery phase, also known as the Exploration phase, is about understanding your data better with respect to your use case. The objective is to figure out specifically how best you can clean, structure, organize, and map the data you have for your use case.
- The next phase, which is the Transformation phase, forms the bulk of the data wrangling process. It involves the tasks you undertake to transform the data, such as structuring, normalizing, denormalizing, cleaning, and enriching the data.
    - Let's begin with the first transformation task – Structuring. This task includes actions that change the form and schema of your data. The incoming data can be in varied formats. You might, for example, have some data coming from a relational database and some data from Web APIs.
    - In order to merge them, you will need to change the form or schema of your data. This change may be as simple as changing the order of fields within a record or dataset or as complex as combining fields into complex structures.
    - Joins and Unions are the most common structural transformations used to combine data from one or more tables. How they combine the data is different. Joins combine columns. When two tables are joined together, columns from the first source table are combined with columns from the second source table—in the same row. So, each row in the resultant table contains columns from both tables. Unions combine rows. Rows of data from the first source table are combined with rows of

data from the second source table into a single table. Each row in the resultant table is from one source table or another.

- Transformation can also include normalization and denormalization of data.
  - Normalization focuses on cleaning the database of unused data and reducing redundancy and inconsistency. Data coming from transactional systems, for eample, where a number of insert, update, and delete operations are performed on an ongoing basis, are highly normalized.
  - Denormalization is used to combine data from multiple tables into a single table so that it can be queried faster. For example, normalized data coming from transactional systems is typically denormalized before running queries for reporting and analysis.
- Another transformation type is Cleaning.
  - Cleaning tasks are actions that fix 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. For example, you may want to find out the demographic information on the sale of a certain product, but the data you have received does not capture the gender. You either need to source this data point and merge it with your existing dataset, or you may need to remove, and not consider the records with this field missing. We will explore many more examples of data cleaning further on in the course.
- Enriching the data—is the fourth type of transformation.
  - When you consider the data you have, to look at additional data points that could make your analysis more meaningful, you are looking at enriching your data. For example, in a large organization with information fragmented across systems, you may need to enrich the dataset provided by one system with information available in other systems, or even public datasets.
  - Consider a scenario where you sell IT peripherals to businesses and want to analyze the buying patterns of your customers over the last five years. You have the

customer master and transaction tables from where you've captured the customer information and purchase history.

- Supplementing your dataset with the performance data of these businesses, possibly available as a public dataset, could be valuable for you to understand factors influencing their purchase decisions.
- Inserting metadata also enriches data. For example, computing a sentiment score from a customer feedback log, collecting geo-based weather data from a resorts location to analyze occupancy trends, or capturing published time and tags for a blog post.

- After transformation, the next phase in Data Wrangling is Validation.
  - This is where you check the quality of the data post structuring, normalizing, cleaning, and enriching.
  - Validation rules refer to repetitive programming steps used to verify the consistency, quality, and security of the data you have.
- This brings us to Publishing—the fourth phase of the data wrangling process.
  - 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.
  - Lastly, it is important to note the criticality of documenting the steps and considerations you have taken to convert the raw data to analysis-ready data. All phases of data wrangling are iterative in nature. In order to replicate the steps and to revisit your considerations for performing these steps, it is vital that you document all considerations and actions.

Tools for Data Wrangling

- In this video, we will look at some of the popularly used data wrangling software and tools, such as: Excel Power Query / Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python and R.
- Let's begin with the most basic software used for manual wrangling—Spreadsheets.
  - Spreadsheets such as Microsoft Excel and Google Sheets have a host of features and in-built formulae that can help you identify issues, clean, and transform data.

- OpenRefine is an open-source tool that allows you to import and export data in a wide variety of formats, such as TSV, CSV, XLS, XML, and JSON.
    - Using OpenRefine, you can clean data, transform it from one format to another, and extend data with web services and external data.
    - OpenRefine is easy to learn and easy to use. It offers menu-based operations, which means you don't need to memorize commands or syntax.
- Google DataPrep is an intelligent cloud data service that allows you to visually explore, clean, and prepare both structured and unstructured data for analysis. It is a fully managed service, which means you don't need to install or manage the software or the infrastructure.
    - DataPrep is extremely easy to use. With every action that you take, you get suggestions on what your ideal next step should be. DataPrep can automatically detect schemas, data types, and anomalies.
- Watson Studio Refinery, available via IBM Watson Studio, allows you to discover, cleanse, and transform data with built-in operations.
    - It transforms large amounts of raw data into consumable, quality information that's ready for analytics.
    - Data Refinery offers the flexibility of exploring data residing in a spectrum of data sources. It detects data types and classifications automatically and also enforces applicable data governance policies automatically.
- Trifacta Wrangler is an interactive cloud-based service for cleaning and transforming data. It takes messy, real-world data and cleans and rearranges it into data tables, which can then be exported to Excel, Tableau, and R. It is known for its collaboration features, allowing multiple team members to work simultaneously.
- Python has a huge library and set of packages that offer powerful data manipulation capabilities. Let's look at a few of these libraries and packages.
    - Jupyter Notebook is an open-source web application widely used for data cleaning and transformation, statistical modeling, also data visualization.
    - Numpy, or Numerical Python, is the most basic package that Python offers. It is fast, versatile, interoperable, and easy to use. It provides support for large, multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these arrays.

- Pandas is designed for fast and easy data analysis operations. It allows complex operations such as merging, joining, and transforming huge chunks of data, performed using simple, single-line commands. Using Pandas, you can prevent common errors that result from misaligned data coming in from different sources.

- R, also offers a series of libraries and packages that are explicitly created for wrangling messy data—such as Dplyr, Data.table, and Jsonlite.
  - Using these libraries, you can investigate, manipulate, and analyze data. Dplyr is a powerful library for data wrangling. It has a precise and straightforward syntax.
  - Data.table helps to aggregate large data sets quickly.
  - Jsonlite is a robust JSON parsing tool, great for interacting with web APIs. Tools for data wrangling come with varying capabilities and dimensions.
  - Your decision regarding the best tool for your needs will depend on factors that are specific to your use case, infrastructure, and teams—such as supported data size, data structures, cleaning and transformation capabilities, infrastructure needs, ease of use, and learnability.

Data Cleaning

- Quality of Data
  - According to a Gartner report on data quality, poor quality data weakens an organization's competitive standing and undermines critical business objectives.
  - Missing, inconsistent, or incorrect data can lead to false conclusions and therefore ineffective decisions. And in the business world, that can be costly.
  - Data sets picked up from disparate sources could have a number of issues, including missing values, inaccuracies, duplicates, incorrect or missing delimiters, inconsistent records, and insufficient parameters. In some cases, data can be corrected manually or automatically with the help of data wrangling tools and scripts, but if it cannot be repaired, it must be removed from the dataset.
- Data Wrangling Process
  - Although the terms Data Cleaning and Data Wrangling are sometimes used interchangeably, it is important to keep in mind that data cleaning is only a subset of the entire Data Wrangling process.

- Data Cleaning forms a very significant and integral part of the Transformation phase in a data wrangling workflow. A typical data cleaning workflow includes: Inspection, Cleaning, and Verification.

- Data Cleaning Workflow
  - The first step in the data cleaning workflow is to detect the different types of issues and errors that your dataset may have. You can use scripts and tools that allow you to define specific rules and constraints and validate your data against these rules and constraints. You can also use data profiling and data visualization tools for inspection.
  - Data profiling helps you to inspect the source data to understand the structure, content, and interrelationships in your data. It uncovers anomalies and data quality issues. For example, blank or null values, duplicate data, or whether the value of a field falls within the expected range.
  - Visualizing the data using statistical methods can help you to spot outliers. For example, plotting the average income in a demographic dataset can help you spot outliers.
- Cleaning
  - That brings us to the actual cleaning of the data. The techniques you apply for cleaning your dataset will depend on your use case and the type of issues you encounter. Let's look at some of the more common data issues.
  - Let's start with missing values. Missing values are very important to deal with as they can cause unexpected or biased results. You can choose to filter out the records with missing values or find a way to source that information in case it is intrinsic to your use case. For example, missing age data from a demographics study.
  - A third option is a method known as imputation, which calculates the missing value based on statistical values. Your decision on the course of action you choose needs to be anchored in what's best for your use case. You may also come across duplicate data, data points that are repeated in your dataset. These need to be removed. Another type of issue you may encounter is that of irrelevant data. Data that does not fit within the context of your use case can be considered irrelevant data. For

example, if you are analyzing data about the general health of a segment of the population, their contact numbers may not be relevant for you.

- Cleaning can involve data type conversion as well. This is needed to ensure that values in a field are stored as the data type of that field—for example, numbers stored as numerical data type or date stored as a date data type.

- You may also need to clean your data in order to standardize it. For example, for strings, you may want all values to be in lower case. Similarly, date formats and units of measurement need to be standardized.
- Then there are syntax errors. For example, white spaces, or extra spaces at the beginning or end of a string is a syntax error that needs to be rectified. This can also include fixing typos or format, for example, the state name being entered as a full form such as New York versus an abbreviated form such as NY in some records.
- Data can also have outliers, or values that are vastly different from other observations in the dataset. Outliers may, or may not, be incorrect. For example, when an age field in a voters database has the value 5, you know it is incorrect data and needs to be corrected.
- Now let's consider a group of people where the annual income is in the range of one hundred thousand to two hundred thousand dollars—except for that one person who earns a million dollars a year. While this data point is not incorrect, it is an outlier, and needs to be looked at. Depending on your use case, you may need to decide if including this data will skew the results in a way that does not serve your use case.

- Verification
  - This brings us to the next step in the data cleaning workflow—Verification. In this step, you inspect the results to establish effectiveness and accuracy achieved as a result of the data cleaning operation. You need to re-inspect the data to make sure the rules and constraints applicable on the data still hold after the corrections you made. And in the end, it is important to note that all changes undertaken as part of the data cleaning operation need to be documented. Not just the changes, but also the reasons behind making those changes, and the quality of the currently stored data. Reporting how healthy the data is, is a very crucial step.

Viewpoints: Data Preparation and Reliability

- In this segment, data professional share what portion of their job involves gathering, cleaning, and preparing data for analysis.
- I would say, a relatively big proportion of my job involves gathering, preparing, and cleaning data for analysis. I work at a company with a really great data engineering team. So I don't have to do this kind of work as much as some other data scientists do. But still, any person that is working closely with data, be they're a data scientist, a data analyst, machine learning engineer, really needs to get comfortable understanding where the data comes from. Inevitably, no dataset is perfect. There's always going to be compromises or small errors. o it's really important to spend a significant portion of your time, understanding the underlined data that was used to generate the dataset and what some potential problems might be with that data.
- My job as a CPA involves a lot of analysis. Financial statements, account activity, assessing processes, and controls. The gathering piece can be pretty simple as long as, the accounting information resides in a general ledger system or a central repository where the data is easy to gather.
- Probably, about 30 percent of the job is laying everything out. So when you get into analytics of it, you can just dive right into the meat and potatoes of it. So you need to track the data, make sure it's accurate, make sure things are adding up. Make sure you have all mumps of information. So for example, on financial statements, I need to make sure that people have given me 12 months of [inaudible] statements, I'm not missing any data and that if I am, that I have enough information to be able to project or to forecast or even look back to estimate what was done in the [inaudible] based on what I have.
- That is definitely helpful. In this segment, data professionals talk about the steps they take to ensure data is reliable.
- One of the essential steps to making sure your data is reliable, is to run summary statistics on individual columns in your data and make sure that they're consistent with reality. For example, if you have a column somewhere that records visits per month to a website and you run summary statistics on that column, you get the minimum, the mean, the median, the max, and you see something funky like, one month there's negative visits or something like this. You know, that data isn't reliable.

- Financial information in particular must be reliable. It must be non-bias. It must be free from error. Those are just a few of the many attributes that are necessary for data to be relied upon. So doing what I call a logic check before you get into the details of a transaction. Does it make sense at a high level? If you expected top-line revenue to increase, but you see that it has drastically decreased, then figure that part out first. Is my source correct? Am I running a query in the right period? Am I pulling the right general ledger account? So start there, make sure that basic data integrity questions have been addressed first. Once we know that the data is reliable, then we can start to deep dive into the reviews and form conclusions about the financial performance based on our analysis of the data.

Summary and Highlights

- 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.