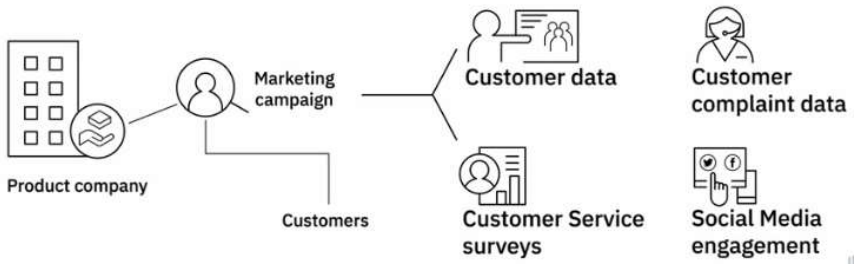
Week 3

Friday, April 23, 2021 7:05 PM

Identifying Data for Analysis

Process for identifying data

- **Step 1**: Determine the information you want to collect
  - The specific information you need
  - The possible sources for this data



- **Step 2**: Define a plan for collecting data
  - Establish a timeframe for collecting data
  - How much data is sufficient for a credible analysis
  - Define dependencies, risks, and mitigation plan

- **Step 3:** Determine your data collection methods
  - Sources of data
  - Type of data
  - Timeframe over which you need the data
  - Volume of data

Key Considerations

- 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.**

Data Quality

- Working with data from disparate sources without considering how it measures against the quality metric can lead to failure. In order to be reliable, data needs to be:
  - Free of errors
  - Accurate
  - Complete
  - Relevant
  - Accessible

Data Governance (Security; Regulation and Compliances)

- Data Governance policies and procedures relate to the usability, integrity, and availability of data. Penalties for non-compliance can run for millions of dollars.

Data Privacy

Identifying the right data is a very important step of the data analysis process. Don right, it will ensure that you are able to look at a problem from multiple perspectives and your findings are credible and reliable.

Data Sources

Data Sources

- **Data sources can be internal or external to the organization**
    - **Primary** - information obtained directly from the source
        - Data from the organization's CRM, HR, or workflow applications
        - Data you gather directly through surveys, interviews, discussions, observations, and focus groups
    - **Secondary -** information retrieved from existing sources.
        - 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
    - **Third-party -** refers to data 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

- **Databases:** can be a source of primary, secondary and third-party data
    - Internal applications for managing processes, workflows and customers
    - External databases available on a subscription basis or for purchase

- **Web**: source of publicly available data that is available to companies and individuals for free or commercial use
    - Textbooks
    - Government records
    - Papers and articles for public consumption

- **Social media sites and interactive platforms**:
    - Facebook
    - Twitter
    - Google
    - Youtube
    - Instagram

- **Sensor Data**
    - Wearable devices
    - Smart buildings

    - Smart cities
    - Smartphones
    - Medical devices

- **Data Exchange**
    - Data from business applications
    - Sensor devices
    - Social media activity

- **Surveys**
- **Interviews**
- **Census**
- **Observations studies**

How to Gather and Import Data

- Gathering data from data sources such as databases, the web, sensor data, data exchanges, and several other sources leveraged for specific data needs. Importing data into different types of data repositories.

SQL, or Structured Query Language (used for extracting information from relational databases)

- Offers simples commands to specify
    - What is to be retrieved from the database
    - Table from which it needs to be extracted
    - Grouping records with matching values
    - Dictating the sequence in which the query results are displayed
    - Limiting the number of results that can be returned by the query

API, or Application Programming Interfaces

- Popularly used for extracting data from a variety of data sources.
- Are invoked from applications that require the data and access an endpoint containing the data. Endpoints can include databases, web services, and data marketplaces.
- Also used for data validation.

Extracting data from the web

Web Scraping (Screen Scraping, Web Harvesting)

- For downloading specific data from web pages based on defined parameters
- For extracting data such as text, contact information, images, videos, podcasts and product items from a web property
- RSS feeds are used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis

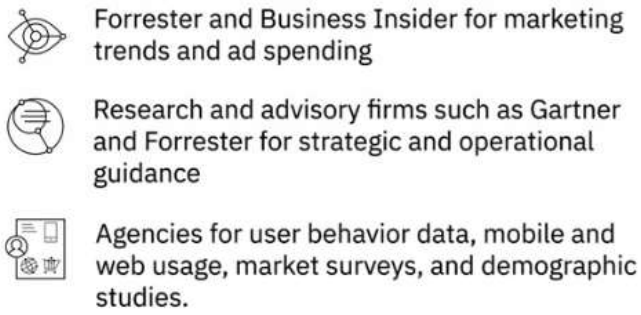
Sensor Data
- Instruments
- IoT devices
- Applications
- GPS data from cars

Data Exchange Platforms (AWS DataExchange, Crunchbase, Lotame, Snowflake)

- Allow the exchange of data between data providers and data consumers
- Have a set of well-defined exchange standards, protocols, and formats relevant for exchanging data
- Facilitate the exchange of data
- Ensure that security and governance are maintained

- Provide data licensing workflows, de-identification and protection of personal information, legal frameworks, and a quarantined analytics environment

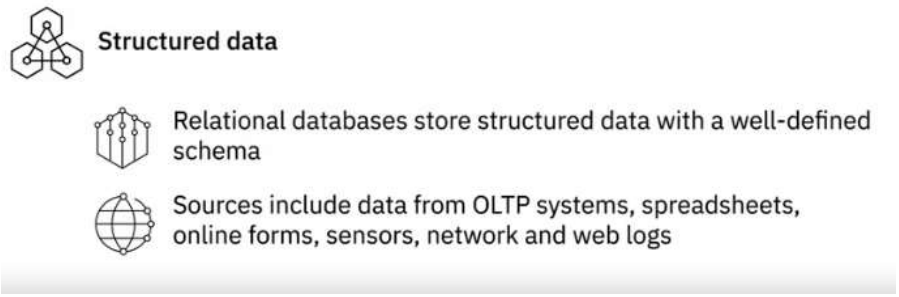
Other Sources

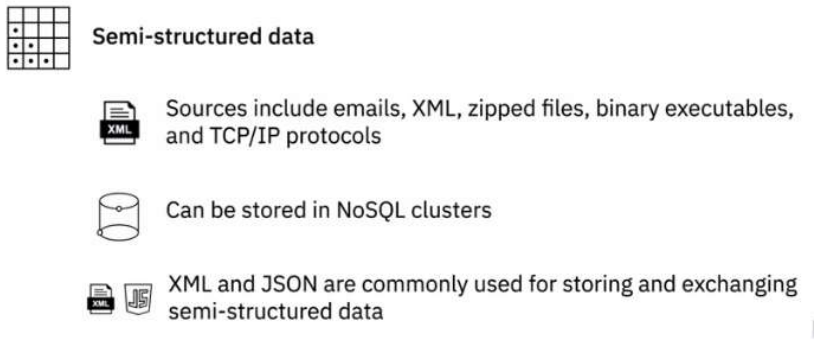


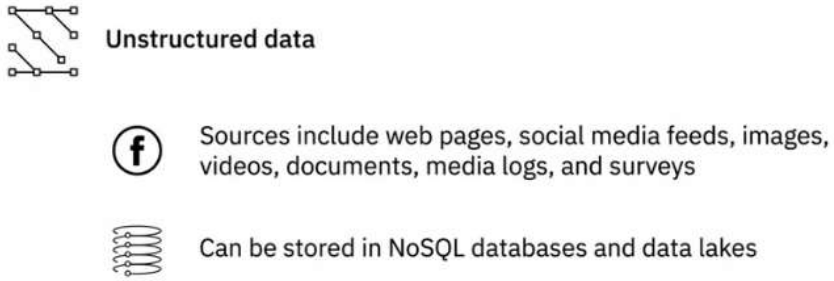
Importing Data

Data that has been identified and gathered from the various data sources now needs to be loaded or imported into a data repository before it can be wrangled, mined, and analyzed.

The importing process involves combining data from different sources to provide a combined view and a single interface using which you can query and manipulate the data.







ETL tools and data pipelines provide automated functions that facilitate the process of importing data.
Tools such as:

- Talend
- Informatica
- Python
- R

Summary and Highlights

- 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

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. Process of preparing raw data.

- **Data Wrangling involves a four-step process:**

- **Discovery**
    - Examining and understanding your data with respect to your use case. Creating a plan for cleaning, structuring, organizing, and mapping your data.
- **Transformation**
    - **Structuring Data** (Joins or Unions of data)
        - This task includes actions that change the form and schema of your data as the incoming data can be in varied formats.
    - **Normalizing and Denormalizing Data**
        - Normalizing data includes
            - Cleaning unused data
            - Reducing redundancy
            - Reducing inconsistency
        - Denormalizing data

        - Combining data from multiples tables into a single table for faster querying of data for reports and analysis
    - **Cleaning Data**
        - Cleaning tasks are actions that fix irregularities in data in order to produce a credible and accurate analysis.
    - **Enriching Data**
        - 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.
- **Validation**
    - Checking the quality of data after structuring, normalizing, denormalizing, cleaning, and enriching of data
    - Verifying consistency, quality, and security of data
- **Publishing**
    - Delivering the output of the wrangled data for downstream project needs

Popular tools for Data Wrangling

- Excel Power Query / Spreadsheets

- OpenRefine
    - 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.

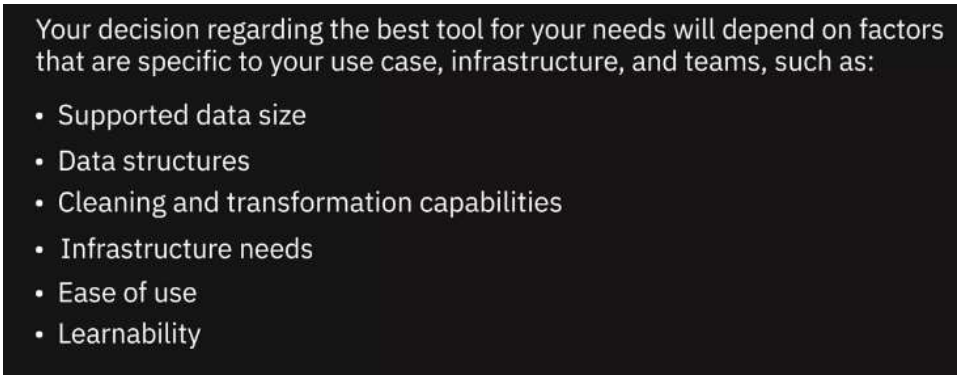
- Google DataPrep
    - Google DataPrep is an intelligent cloud data service that allows you to visually explore, clean, and prepare both structured and unstructured data for analysis.
    - DataPrep can automatically detect schemas, data types, and anomalies.

- Watson Studio Refinery
    - Watson Studio Refinery, available via IBM Watson Studio, allows you to discover, cleanse, and transform data with built-in operations.
    - It detects data types and classifications automatically and also enforces applicable data governance policies automatically.

- Trifacta Wrangler
    - 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.
    - Allows team members to work simultaneously

- Python - has a huge library and set of packages that offer powerful data manipulation capabilities.
    - **Jupyter Notebook** - An open-source web application widely used for data cleaning and transformation, statistical modeling, and data visualization
    - **Numpy (Numerical Python) -** It provides support for large multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these arrays
    - **Pandas** - Allows complex operations such as merging, joining, and transforming huge chunks of data using simples, single-line commands

- R - Offers a series of libraries and packages that are explicitly created for wrangling messy data.
    - Dplyr - A powerful library for data wrangling with a precise and straightforward syntax
    - Data.table - Helps aggregate large data sets quickly
    - Jsonlite - A robust JSON parsing tool, great for interacting with web API's



Data Cleaning

- Poor quality data weakens an organization's competitive standing and undermines critical business objectives.
- Data sets picked up from disparate sources could have a number of issues, including missing values, inaccuracies, duplicates, incorrect or missing delimiter, inconsistent records, and insufficient parameters.

Data Cleaning Workflow includes:

- **Inspection** - helps you to inspect the source data to understand the structure, content, and interrelationships in your data. It uncovers anomalies and data quality issues.
    - Detecting issues and errors
    - Validating against rules and constraints
    - Profiling data to inspect source data
    - Visualizing data using statistical methods

- **Cleaning** - the techniques applied for cleaning your dataset will depend on your use case and the type of issues you encounter.
    - Missing values can cause unexpected or biased results
    - Duplicate data are data points that are repeated in your dataset
    - Irrelevant data is data that is not contextual to your use case
    - Data type conversion is needed to ensure that values in a field are stored as the data type of that field
    - Standardizing data is needed to ensure date-time formats and units of measurement are standard across the dataset
    - Syntax errors, such as white spaces, extra spaces, typos, and formats need to be fixed
    - Outliers need to be examined for accuracy and inclusion in the dataset

- **Verification -** Inspecting results to establish effectiveness and accuracy achieved as a result of the data cleaning

It is important to document:
- Changes undertaken as part of the data cleaning operation
- Reasons for undertaking these changes
- Quality of currently stored data

Summary and Highlights

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.