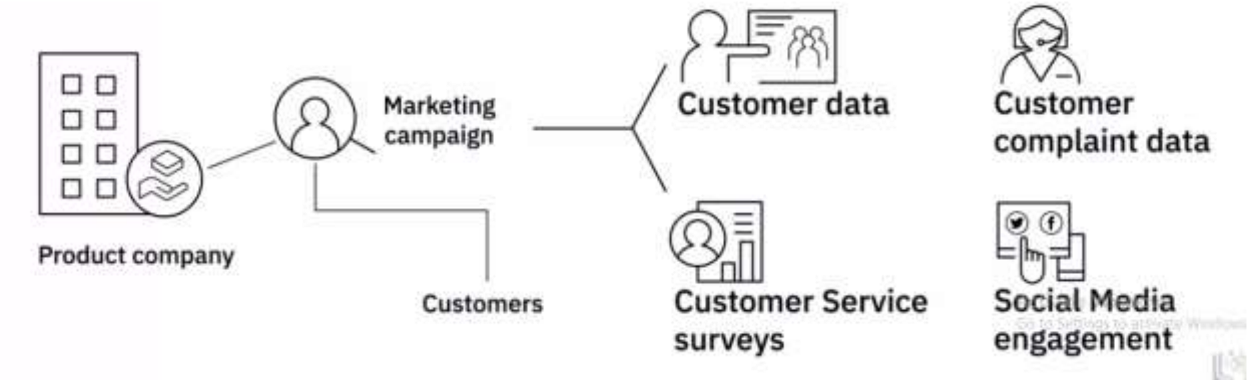
Gathering Data

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 data



- Ex: the example of a product company that wants to create targeted marketing campaigns based on the age group that buys their products the most. Their goal is to the design reach-outs that appeal most to this segment and encourage them to further influence their friends and peers into buying these products. Based on this use case, some of the obvious information that you will identify includes the customer profile, purchase history, location, age, education, profession, income and marital status, for example. To ensure you gain even greater insights into this segment, you may also decide to collect the customer complaint data for this segment to understand the kind of issues they face because this could discourage them from recommending your products. To know how satisfied they were with the resolution of their issues, you could collect the ratings from the customer service surveys. Taking this a step forward, you may want to understand how these customers talk about your products on social media and how many of their connections engage with them in these discussions, for example, the likes, shares, and comments their posts receive.
- Step 2: Define a plan for collecting data
  - You need to establish a timeframe for collecting the data you have identified. Some of the data you need may be required on an ongoing basis and some over a defined period of time. For collecting website data, for example you may need to have the

numbers refreshed in real-time. But if you tracking data for a specific event, you have a definite beginning and end date for collecting the data.

  - You can also define how much data would be sufficient for you to reach a credible analysis. Is the volume defined by the segment, for example, all customers within the age range of 21 to 30 years; or a dataset of a hundred thousand customers within the age range of 21 to 30.
  - You can also use this step to define the dependencies, risks, mitigation plan, and several other such factors that are relevant to your initiative. The purpose of the plan should be to establish the clarity you need for execution.

- Step 3: Determine your data collection methods
  - In this step, you will identify the methods for collection the data you need. You will define how you will collect the data from the data sources you have identified, such as internal systems, social media sites, or third-party data providers. Your methods will depend on the type of data, the timeframe over which you need the data, and the volume of data. Once your plan and data collection methods are finalized, you can implement your data collection strategy and start collecting data. You will be making updates to your plan as you go along because conditions evolve as you implement the plan on the ground.

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. None of these are one-time considerations but are relevant through the life cycle of the data analysis process.

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

- You need to define the quality traits, the metric, and the checkpoints in order to ensure that your analysis is going to be based on quality data.

Data Governance

- Issues pertaining to data governance include:
  - Security
  - Regulation
  - Compliances
- Data Governance policies and procedures relate to the usability, integrity, and availability of data. Penalties for non-compliance can run into millions of dollars and can hurt the credibility of not just your findings, but also your organization.

Data Privacy

- Data privacy includes issues such as:
  - Confidentiality
  - License for use
  - Compliance to mandated regulations
- You need to define:
  - Checks
  - Validations
  - Auditable trail
- Loss of trust in the data used for analysis can compromise the process, result in suspect finding, and invite penalties.

Conclusions

- Identifying the right data is a very important step of the data analysis process. Done 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, and they can be primary, secondary or third party sources of data.

Primary Data

- The term primary data refers to information obtain directly by you from the sources. This could be from internal sources such as
    - Data from the organization's CRM, HR, or Workflow applications
    - Data you gather directly through surveys, interviews, discussions, observations, and focus groups

Secondary Data

- Secondary data refers to information retrieved from existing sources.
    - External databases
    - Research articles, publications, training material, internet searches, or financial records available as public data
    - Data collected through external conducted surveys, interviews, discussions, observations, and focus groups.

Third-party Data

- Third-party data 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 sources 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.
    - A significant number of business have or are currently moving to the cloud, which is increasingly becoming a source for accessing real time information and on demand insights.

Web

- Web is a sources of publicly available data that is available to companies and individuals for free or commercial use. The web is a rich source of data available in the public domain.
    - textbooks, government records, papers, and articles

Social media sites, and interactive platforms

- Such as Facebook, Twitter, Google, YouTube. An Instagram are increasingly being used to source user data and opinions. Businesses are using these data sources for quantitative and qualitative insights. An existing and potential customers.

Sensor Data

- Sensor data produced by wearable devices, smart buildings, smart cities, smart phones, medical devices, even household appliances is a widely used source of data.

Data Exchange

- Data exchange is a source of 3rd party data that involves the voluntary sharing of data between data providers and data consumers, individuals, organizations and governments could be both data providers and data consumers.
    - The data that is exchanged could include data coming from business applications, sensor devices, social media activity, location data, or consumer behavior data.

Surveys

- Surveys gather information through questionnaires distributed to a select group of people.
    - For example, gauging the interest of existing customers in spending on an updated version of a product. Surveys can be web or paper based.

Census

- Census data is also a commonly used source for gathering household data, such as wealth and income or population data
    - Interviews are source for gathering qualitative data, such as the participants opinions and experiences.
    - For example, an interview conducted to understand the day-to-day challenges faced by a customer service executive.
    - Interviews could be telephonic over the Web or face to face observation.

Observation Studies

- Studies include monitoring participants in a specific environment or while performing a particular task.
    - For example, observing users navigate an E Commerce site to assess
    - Ease with which they are able to find products and make a purchase data from surveys, interviews, an observation.
    - Studies could be available as primary, secondary and 3rd party data.

Sources for Gathering Data

- Dynamic
- Diverse
- Continuously evolving

How to Gather and Import Data

Overview

- Gathering data from the data sources discussed earlier in the course—such as databases, the web, sensor data, data exchanges, and several other sources leveraged for specific data needs.

Using queries to extract data from SQL databases

- SQL, or Structured Query Language, is a querying language used for extracting information from relational databases.
- SQL offers simple commands to specify what is to be retrieved from the database, the table from which it needs to be extracted, grouping records with matching values, dictating the sequence in which the query results are displayed, and limiting the number of results that can be returned by the query, amongst a host of other features and functionalities.
- Non-relational databases can be queried using SQL or SQL-like query tools. Some non-relational databases come with their own querying tools such as CQL for Cassandra and GraphQL for Neo4J.

APIs

- Application Programming Interfaces (or APIs) are also popularly used for extracting data from a variety of data sources.
- APIs are invoked from applications that require the data and access an end-point containing the data. End-points can include databases, web services, and data marketplaces.
- APIs are also used for data validation. For example, a data analyst may utilize an API to validate postal addresses and zip codes.

Web Scrapping (Screen Scrapping, Web Harvesting)

- Web scraping, also known as screen scraping or web harvesting, is used for downloading specific data from web pages based on defined parameters.
- Among other things, 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 sites where data is refreshed on an ongoing basis.

Sensor Data

- Data streams are a popular source for aggregating constant streams of data flowing from sources such as instruments, IoT devices and applications, and GPS data from cars.
- Data streams and feeds are also used for extracting data from social media sites and interactive platforms.

Data Exchange

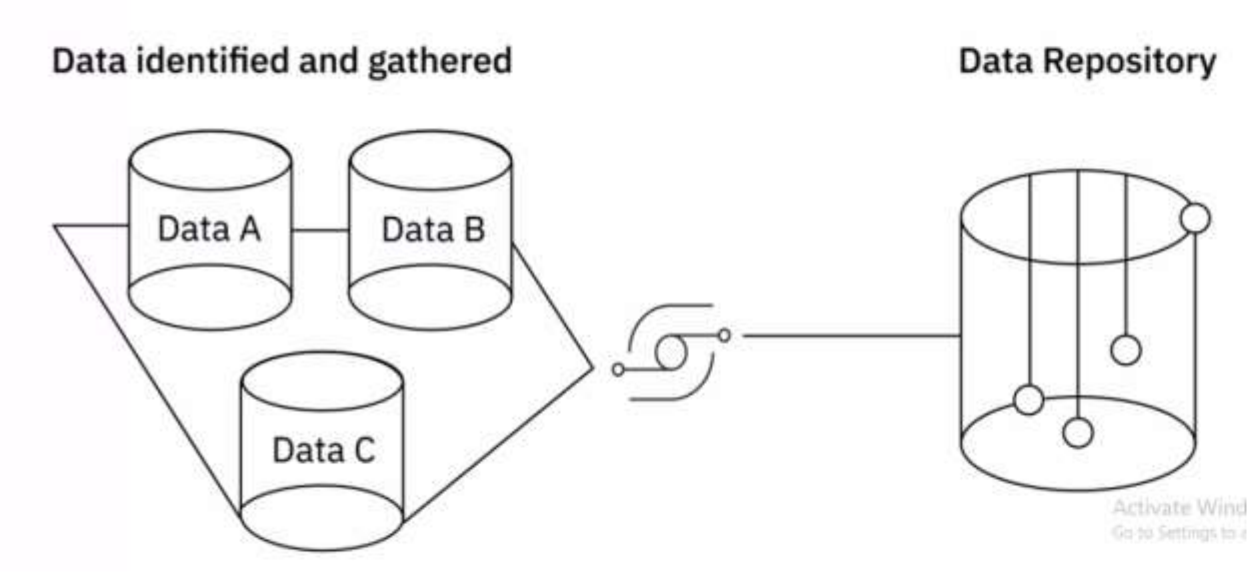
- Data Exchange platforms allow the exchange of data between data providers and data consumers.
- Data Exchanges have a set of well-defined exchange standards, protocols, and formats relevant for exchanging data.
- These platforms not only facilitate the exchange of data, they also ensure that security and governance are maintained.
- They provide data licensing workflows, de-identification and protection of personal information, legal frameworks, and a quarantined analytics environment.
- Examples of popular data exchange platforms include AWS Data Exchange, Crunchbase, Lotame, and Snowflake.

Other Sources

- Numerous other data sources can be tapped into for specific data needs.
- For marketing trends and ad spending, for example, research firms like Forrester and Business Insider are known to provide reliable data.
- Research and advisory firms such as Gartner and Forrester are widely trusted sources for strategic and operational guidance.
- Similarly, there are many trusted names in the areas of user behavior data, mobile and web usage, market surveys, and demographic studies.

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.



- Depending on the data type, the volume of data, and the type of destination repository, you may need varying tools and methods.

Data types and destination repositories

- Specific data repositories are optimized for certain types of data.
    - Relational databases store structured data with a well-defined schema.
    - If you’re using a relational database as the destination system, you will only be able to store structured data, such as data from OLTP systems, spreadsheets, online forms, sensors, network and web logs.
    - Structured data can also be stored in NoSQL.
    - Semi-structured data is data that has some organizational properties but not a rigid schema, such as, data from emails, XML, zipped files, binary executables, and TCP/IP protocols.
    - Unstructured data is data that does not have a structure and cannot be organized into a schema, such as data from web pages, social media feeds, images, videos, documents, media logs, and surveys. NoSQL databases and Data Lakes 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 pipelines provide automated functions that facilitate the process of importing data. Tools such as Talend and Informatica, and programming languages such as Python and R, and their libraries, are widely used for importing data.

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.