Module-3
Gathering and Wrangling Data

In this module, you will learn about the process and steps involved in identifying, gathering, and importing data from disparate sources. You will learn about the tasks involved in wrangling and cleaning data in order to make it ready for analysis. In addition, you will gain an understanding of the different tools that can be used for gathering, importing, wrangling, and cleaning data, along with some of their characteristics, strengths, limitations, and applications.

Learning Objectives

- Explain the steps and processes you need to take to identify, gather, and import data from disparate sources.
- Describe the tools and techniques required for wrangling and cleaning data so as to make it analysis-ready.

Before you start

Practically Speaking: How to Make Data Work for You

🟦 **Welcome and Introduction to Module 3**

- Warm welcome back to the learning series: *Practically Speaking*.
- Recognition of progress through **Modules 1 and 2**.
- Introduction to **Module 3: Gathering and Wrangling Data**.
    - Described as a **pivotal** module.
    - Focus on **core processes** in working with data:
        - Finding data
        - Bringing it together
        - Shaping and cleaning it
        - Laying the groundwork for analysis

🟩 **Importance of Data Preparation**

- Analogy: Laying the **foundation of a house**.
- Skills are crucial for generating **reliable insights**.

- Increasing organizational reliance on **data-driven decisions**.
- Emphasis on:
    - **Data quality** = Insight quality
    - Good preparation prevents flawed results

🟧 **Real-World Example: Healthcare**

- **Scenario:** Healthcare organization improving patient outcomes
- Challenges:
    - Scattered data:
        - Electronic Health Records (EHRs)
        - Spreadsheets
        - Surveys
        - Handwritten notes
    - Various formats and systems
- Application of Module 3 skills:
    - Data is gathered from all sources
    - Inconsistencies resolved:
        - Abbreviations vs. full terms
        - Different date formats
    - Clean, standardized, **unified patient record**
- Outcome:
    - Better analysis of patterns (e.g., treatment effectiveness)
    - Smarter, more informed decisions
    - Efficient use of resources

🟫 **Analogy: Data Preparation = Woodworking**

- Raw data compared to **raw lumber**

- Steps a woodworker takes:
    - Selecting the right pieces
    - Planing, cutting, smoothing
- Poor preparation = fragile, faulty results
- Skilled preparation = strong, functional outcomes
- Parallel in data work:
    - Clean, prepped data enables trustworthy, actionable insights

---

🟪 Practical Tips for Module 3

1. **Don't rush into analysis**
    - Understand the data's **origin**:
        - Who collected it?
        - Why and how?
    - Context reveals biases and limitations.
2. **Expect iteration**
    - Data wrangling is **not linear**.
    - Requires:
        - Exploring
        - Transforming
        - Checking and validating
        - Looping back as needed
    - Analogy: **Sculpting**—refine as you go
3. **Learn the terminology**
    - Understand key technical terms:
        - **Data profiling**
        - **Data imputation**
        - **Schema mapping**
    - Importance:

- Shared language
- Mental map of the process
- Awareness of appropriate tools and steps

🟧 Actionable Next Step

- **Reflect on a real decision** you care about.
- Identify:
  - One **internal** data source (e.g., company records)
  - One **external** source (e.g., public datasets)
- Example: Traffic patterns
  - Internal: Maybe limited or unavailable
  - External: Public transit data, social media mentions

🟥 Reflection Prompt

- Think of a time you organized something messy:
  - Physical clutter or digital files
- Consider:
  - Your strategy
  - Challenges and wins
- Relate that to:
  - Preparing data for analysis
  - Creating order from chaos
  - Satisfaction of clarity and readiness

📌 Summary

Module 3 of *Practically Speaking*—**"Gathering and Wrangling Data"**—emphasizes the foundational importance of data preparation. It explores how collecting, cleaning, and standardizing data ensures reliable insights and better decision-making. Through real-world examples like healthcare and metaphors such as woodworking and sculpting, the module highlights that quality analysis depends on quality data.

Learners are encouraged to take a thoughtful, iterative approach, understand the data's origins, and learn the key terminology. A practical step involves identifying potential data sources for a real decision, and a reflective prompt helps connect personal experience to data organization.

🔑 **Key Takeaways:**

- Don’t skip the groundwork—clean data is essential.
- Understand data context and origin.
- Data wrangling is iterative—expect revisions.
- Learn and use precise data-related terminology.
- Begin identifying internal and external data sources for real-world decisions.

📊 **Table: What We Learnt in the Video**

Topic/Section	What We Learnt
Module Purpose	Module 3 focuses on data gathering and wrangling as the base for strong analysis.
Real-world Example (Healthcare)	Unified patient records improve insights and decisions in health care.
Woodworking Analogy	Raw data needs shaping, just like wood, to become useful and reliable.
Tips for Data Wrangling	Understand source, expect iteration, and master terminology.
Action Step	Think of a decision and identify internal/external data sources.
Reflection Exercise	Organizing messy things parallels preparing messy data.

Gathering Data

Identifying Data for Analysis

🔍 Understanding the Context: Where You Are & Where You Want to Be

- **Problem Understanding**
    - Clear understanding of the current situation (Where you are)
    - Clear vision of the desired outcome (Where you want to be)
- **Defined Metrics**
    - Know **what will be measured**
    - Understand **how it will be measured**

🧩 Identifying the Data You Need

a) Specific Information Required

- Determined by your **goals**
- Example: Targeted marketing campaign
    - Customer profile
    - Purchase history
    - Location
    - Age
    - Education
    - Profession
    - Income
    - Marital status

b) Possible Sources of Data

- Customer complaint data to identify:
    - Common issues faced
    - Their effect on customer advocacy
- Customer service survey ratings:

    - To assess satisfaction with issue resolution
- Social media activity:
    - How customers talk about products
    - Level of peer engagement (likes, shares, comments)

🗓️ Planning the Data Collection

- **Establish a timeframe**:
    - **Real-time data**: e.g., website visitors
    - **Event-specific data**: fixed start and end date
- **Volume of Data Needed**:
    - Defined by segment (e.g., age 21–30)
    - Could be specific numbers (e.g., 100,000 users in that age group)
- **Plan Should Include**:
    - Dependencies
    - Risks
    - Mitigation strategies
    - Other relevant factors for execution

📥 Determining Data Collection Methods

- Identify **how to collect data** from:
    - Internal systems
    - Social media
    - Third-party data providers
- Based on:
    - Type of data
    - Timeframe
    - Volume

- **Implement the plan**, but be ready to **update** as conditions evolve

---

📊 Data Quality Considerations

- **Essential traits for data**:
  - Error-free
  - Accurate
  - Complete
  - Relevant
  - Accessible
- Define:
  - Quality traits
  - Metrics
  - Checkpoints

🔐 Data Governance & Security

- **Data Governance** includes:
  - Usability
  - Integrity
  - Availability
- **Non-compliance risks**:
  - Legal penalties
  - Damaged credibility
  - Unreliable analysis

🛡️ Data Privacy

- Must ensure:
  - Confidentiality

    - Licensed usage
    - Compliance with regulations
- Requires:
    - Checks and validations
    - Auditable data trails
- **Loss of trust in data** can lead to:
    - Compromised processes
    - Untrustworthy results
    - Legal & reputational consequences

📌 Summary

The process of identifying and collecting data is critical to the success of any data analysis initiative. Starting from a clear understanding of goals and metrics, the next steps involve identifying the right data, planning how to collect it, and determining the methods for doing so. Data should be accurate, relevant, and gathered from reliable sources. A detailed plan, with consideration for governance, privacy, and quality, ensures trust in the final analysis. Done right, this process enables insights that are both credible and actionable.

Key Takeaways:

- Start with clarity on goals and metrics
- Identify data based on goals and potential sources
- Plan for how, when, and how much data to collect
- Ensure data quality, governance, and privacy throughout

📚 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Problem and Outcome Clarity	Know your current state and desired end goal
Metric Definition	Define what and how things will be measured
Information Identification	Pinpoint data needs (e.g., customer profile, complaints, social engagement)
Planning Data Collection	Set timeframe, data volume, risks, and dependencies
Data Collection Methods	Choose sources and collection techniques based on data type and need

Data Quality	Ensure accuracy, completeness, and relevance of the data
Data Governance	Follow protocols for data integrity, usability, and legal compliance
Data Privacy	Adhere to privacy rules and secure, licensed usage

Data Sources

1. Types of Data Sources

Data sources can be categorized in two ways:

- **By origin:**
  - **Internal**: Within the organization
  - **External**: Outside the organization
- **By type:**
  - **Primary Data**
  - **Secondary Data**
  - **Third-party Data**

---

2. Primary Data

Definition: Data obtained *directly* by the individual or organization from the original source.

Examples:

- **Internal sources**:
  - CRM systems
  - HR applications
  - Workflow management tools
- **Direct methods**:
  - Surveys
  - Interviews
  - Discussions
  - Observations

    - Focus groups

---

3. Secondary Data

Definition: Data retrieved from *existing sources*, not collected firsthand by the user.

Examples:

- External databases
- Research articles & academic publications
- Training materials
- Internet search data
- Public financial records
- Externally conducted:
    - Surveys
    - Interviews
    - Discussions
    - Observations
    - Focus groups

---

4. Third-party Data

Definition: Data purchased from data aggregators who collect and compile data from various sources for resale.

Common Characteristics:

- Often compiled into *comprehensive datasets*
- Used for enhancing decision-making and customer insights

---

5. Common Data Sources

📦 **Databases**

- Can provide **primary**, **secondary**, or **third-party** data

- Types:
    - Internal applications
    - External (subscription-based or purchasable)
    - Cloud-based systems (real-time insights)

🌐 Web

- Source of **publicly available data**
- Includes:
    - Textbooks
    - Government records
    - Research papers and articles
    - Social media platforms (Facebook, Twitter, YouTube, etc.)

📱 Sensor Data

- Generated by:
    - Wearables
    - Smart cities/buildings
    - Smartphones
    - Medical devices
    - Home appliances
- Used extensively for IoT-based analytics

🔁 Data Exchange

- Involves voluntary **sharing** of data between providers and consumers
- Stakeholders: individuals, organizations, governments
- Exchanged data types:
    - Business applications
    - Sensor/device outputs
    - Social media activity

- Location and consumer behavior data

---

6. Survey-Based Sources

📝 Surveys

- Gather structured information through questionnaires
- Formats:
  - Web-based
  - Paper-based
- Example: Customer interest in a product upgrade

🧮 Census Data

- Provides demographic and household insights
- Examples: Income, population stats, wealth

---

7. Qualitative Research Methods

🎙️ Interviews

- Help collect *opinions and experiences*
- Formats:
  - Telephonic
  - Web-based
  - Face-to-face
- Example: Challenges faced by a customer service executive

👀 Observation Studies

- Monitor participants performing tasks
- Example: Observing customers on an e-commerce site to assess usability

---

8. Key Insight

- The combination of **primary**, **secondary**, and **third-party** data allows organizations to:
    - Discover new insights
    - Explore problems from multiple perspectives
    - Make more informed decisions

🗒️ Summary

This video explores the dynamic world of data sources, classifying them by origin (internal/external) and type (primary, secondary, third-party). Primary data is collected directly by the user; secondary data comes from pre-existing sources; third-party data is bought from aggregators. Data sources include internal databases, the web, cloud platforms, sensor devices, surveys, interviews, and data exchanges. By combining these diverse types of data, organizations can enhance their decision-making, develop meaningful insights, and address business challenges with greater precision.

Key Takeaways:

- Understand the **3 main types of data** (Primary, Secondary, Third-party)
- Identify the **source** of each type of data
- Recognize the **tools and methods** used to gather each type
- Learn how to use multiple data types for better insights

📊 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Primary Data	Collected directly through internal systems or user interaction (surveys, etc.)
Secondary Data	Comes from existing external resources like research papers and databases
Third-party Data	Purchased data from aggregators who compile from various sources
Database Sources	Can serve as sources for all three types of data
Cloud & Web Sources	Provide real-time and publicly available data
Sensor Data	Generated by devices (wearables, smart homes, etc.)
Data Exchange	Voluntary sharing of various types of data across entities
Surveys	Structured data collection for customer preferences or feedback
Census Data	Offers large-scale demographic insights
Interviews & Observations	Provide qualitative data about user experiences and behaviors

How to Gather and Import Data

1. Methods and Tools for Data Gathering

- **Primary Data Sources Covered:**
    - Databases (Relational and Non-relational)
    - Web
    - Sensor data
    - Data exchanges
    - Streams and feeds
    - APIs

---

2. Using SQL for Data Extraction

- **SQL (Structured Query Language):**
    - Used for querying relational databases.
    - Capabilities:
        - Extracting specific data
        - Specifying tables and fields
        - Grouping data
        - Sorting results
        - Limiting the number of results
    - **Relational Databases**: Require structured data with a predefined schema.

---

3. Querying Non-relational Databases

- **Querying Tools:**
    - SQL or SQL-like tools
    - Examples:
        - **CQL** (Cassandra Query Language) for **Cassandra**
        - **GraphQL** for **Neo4j**

- Support semi-structured or unstructured data

---

4. APIs (Application Programming Interfaces)

- **Purpose:**
    - Extract data from various sources
    - Validate data (e.g., postal address, zip codes)
- **Sources Accessed by APIs:**
    - Databases
    - Web services
    - Data marketplaces

---

5. Web Scraping

- Also known as:
    - Screen scraping
    - Web harvesting
- **Used to Extract:**
    - Text
    - Contact info
    - Images, videos, podcasts
    - Product listings
- **Based on defined parameters**

---

6. RSS Feeds

- **Used For:**
    - Capturing updated data
    - Particularly from:
        - News sites

        - Online forums

---

7. Data Streams

- **Examples of Streaming Sources:**
    - IoT devices
    - Instruments
    - GPS devices (e.g., cars)
    - Social media platforms
- **Use Case:**
    - Continuous data aggregation

---

8. Data Exchange Platforms

- **Functions:**
    - Facilitate secure data exchange between providers and consumers
- **Key Features:**
    - Exchange standards & protocols
    - Legal frameworks
    - De-identification and data protection
    - Analytics environment
    - Licensing workflows
- **Popular Platforms:**
    - AWS Data Exchange
    - Crunchbase
    - Lotame
    - Snowflake

---

9. Specialized Data Sources

- **Examples:**
    - **Marketing trends & ad spending:** Business Insider, Forrester
    - **Strategic & operational data:** Gartner, Forrester
    - **Demographic/user behavior/market surveys:** Various research firms

---

10. Data Importing and Repositories

- **Importing Data:**
    - Combines data from multiple sources
    - Presents a unified interface for querying/manipulating data
- **Factors Influencing Import Method:**
    - Data type
    - Data volume
    - Destination repository

---

11. Types of Data and Storage Repositories

Data Type	Characteristics	Suitable Repositories
Structured	Defined schema	Relational DBs, NoSQL
Semi-structured	Partial organization (e.g., XML, JSON)	NoSQL clusters
Unstructured	No predefined format (e.g., media, social feeds)	NoSQL, Data Lakes

- **Data Formats:**
    - **XML, JSON**: Common for semi-structured data
    - **JSON**: Preferred for web services

---

12. ETL Tools and Programming Languages

- **ETL Tools:**
    - Talend

    - Informatica
- **Programming Languages:**
    - Python (e.g., pandas, requests, BeautifulSoup)
    - R

📝 Summary

This video outlines various methods and tools used for data collection and import. It explains how data can be sourced from relational and non-relational databases, web scraping, APIs, RSS feeds, data streams, and data exchanges. Tools like SQL, GraphQL, and CQL are used for querying, while APIs and ETL tools enable data retrieval and transformation. It also covers the categorization of data into structured, semi-structured, and unstructured formats, and the suitable repositories for each type such as relational databases, NoSQL databases, and data lakes.

Key Takeaways:

- SQL is essential for structured data querying.
- Non-relational DBs often require special querying tools.
- APIs and web scraping are critical for online data access.
- RSS feeds and data streams offer continuous, updated data.
- Data exchanges enable secure, standardized data sharing.
- Data must be properly categorized before choosing import tools and repositories.

📊 Table: What We Learnt in the Video

Topic/Section	What We Learnt
SQL for Relational Databases	Used to extract, group, sort, and limit structured data
Non-relational DB Querying	Uses tools like CQL and GraphQL for semi/unstructured data
APIs	Access data from endpoints; used for validation and integration
Web Scraping	Collects specific content from web pages like text, images, products
RSS Feeds	Capture real-time updates from online platforms
Data Streams	Handle continuous data from IoT, GPS, apps, social media
Data Exchange Platforms	Ensure secure, governed data exchange with standards and legal frameworks

Specialized Data Providers	Firms like Gartner and Forrester provide niche market and research data
Data Importing	Combines data from diverse sources into unified repositories
Types of Data	Structured, semi-structured, and unstructured — each with appropriate repositories
ETL & Programming Tools	Talend, Informatica, Python, and R support data import and transformation

Summary and Highlights

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.

Wrangling Data

What is Data Wrangling?

📌 Introduction to Data Wrangling (Data Munging)

- **Definition**: An **iterative process** that prepares raw data for meaningful analysis.
- **Purpose**: To clean, transform, validate, and publish data from various sources.
- **Nature**: Involves multiple tasks tailored for a **clearly defined analytical purpose**.

🔄 The 4-Step Process of Data Wrangling

1. **Discovery (Exploration)**
    - Objective: **Understand the data** in the context of the intended analysis.
    - Activities:
        - Identify the **structure, quality, and relevance** of the data.
        - Analyze how to **clean, organize, and map** the data for the use case.
2. **Transformation**
    - **Core of the wrangling process** – turns raw data into usable form.
    - Involves multiple tasks:

🔧 **Structuring**

- Changes the **form/schema** of the data to make it compatible.
- Deals with **heterogeneous data sources** (e.g., DBs, Web APIs).

- Examples of structural transformations:

Method	Description
Join	Combines columns from two tables into one row
Union	Combines rows from two tables into one dataset

📐 **Normalization**

- **Reduces redundancy** and ensures **data integrity**.
- Common in **transactional systems** with frequent updates.

- 🧱 **Denormalization**
  - Combines data from **multiple tables into one**.
  - Goal: **Improved query performance** (esp. for reports).
- 🧹 **Cleaning**
  - Fixes **inaccurate, missing, or inconsistent** data.
  - Deals with:
    - Null or missing values
    - Bias or outliers
    - Unavailable fields
  - Actions:
    - Sourcing missing data
    - Removing flawed records
- ➕ **Enriching**
  - Adds **external or supplementary data** to enhance analysis.
  - Examples:
    - Add public business performance data to customer transactions
    - Add metadata: sentiment scores, weather data, blog tags

3. **Validation**
   - Ensures the **quality and accuracy** of transformed data.
   - Uses **repetitive programming rules** to:
     - Check consistency
     - Enforce data integrity
     - Ensure data is secure and valid
4. **Publishing**
   - Final phase: **delivers the wrangled dataset** for downstream use.
   - Includes:
     - **Transformed and validated dataset**
     - **Metadata** about the data

- Supports analytics, reporting, and machine learning workflows

---

📝 Importance of Documentation

- Every step and decision in wrangling should be **well-documented**.
- Reasons:
  - Enables **replication** of steps
  - Supports **future revisions**
  - Helps in **auditing and tracking** decisions made during the wrangling process

---

Summary

Data wrangling, or data munging, is a systematic and iterative process to prepare raw data for analysis. It involves four key stages—Discovery, Transformation, Validation, and Publishing. The process begins with understanding the data and determining how best to prepare it for a given use case. Transformation, the most involved phase, includes structuring, normalizing, denormalizing, cleaning, and enriching the data. This is followed by Validation, where the quality and reliability of the data are tested. The final step, Publishing, involves delivering the refined dataset for use. Documentation throughout the process is critical for transparency, replication, and future use.

📊 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Data Wrangling Definition	An iterative process for preparing raw data for analysis
Discovery Phase	Understanding data structure, quality, and use-case fit
Transformation	Involves structuring, cleaning, normalizing, denormalizing, and enriching data
Joins vs Unions	Joins combine columns from tables; Unions combine rows
Normalization	Reduces redundancy, ideal for transactional systems
Denormalization	Combines data for faster querying, suitable for reporting
Cleaning	Fixes irregularities like missing, biased, or inaccurate data
Enriching	Supplements data with external/public/metadata for more meaningful insights
Validation	Applies rules to ensure data is consistent, secure, and high-quality
Publishing	Shares final dataset and metadata for downstream needs

Documentation	Crucial for process transparency and repeatability

Tools for Data Wrangling

🛠️ Popular Data Wrangling Tools and Software

📊 Spreadsheets (Excel, Google Sheets)

Basic Tool for Manual Wrangling

- **Common Tools**: Microsoft Excel, Google Sheets
- **Features**:
    - In-built formulas for cleaning and transforming data
    - Ability to identify issues in data manually
- **Add-ins/Functions**:
    - **Microsoft Excel**: Power Query (used for importing, cleaning, and transforming data)
    - **Google Sheets**: QUERY function for data manipulation and filtering

---

🧰 OpenRefine

Open-source Data Cleaning Tool

- **Supported Formats**: TSV, CSV, XLS, XML, JSON
- **Capabilities**:
    - Data cleaning and transformation
    - Format conversion
    - Data extension using web services and external data
- **User-Friendly**:
    - Menu-based operations (no need to memorize commands or syntax)
    - Easy to learn and use

☁️ Google DataPrep

Cloud-based Data Preparation Tool

- **Features**:
    - Visual exploration and preparation of data (structured & unstructured)
    - Fully managed (no installation or infrastructure required)
- **Intelligent Suggestions**:
    - Recommends next actions during wrangling
- **Smart Capabilities**:
    - Auto-detection of schemas, data types, and anomalies

🧠 Watson Studio Refinery (IBM)

Enterprise-Level Data Wrangling in IBM Cloud

- **Key Features**:
    - Cleansing and transforming large datasets
    - Data ready for analytics
- **Advantages**:
    - Explores data from a wide range of sources
    - Auto-detects data types and classifications
    - Automatically enforces data governance policies

🤝 Trifacta Wrangler

Collaborative Cloud-Based Data Cleaning Tool

- **Functions**:
    - Cleans messy real-world data
    - Rearranges data into tables
- **Export Support**: Excel, Tableau, R
- **Highlight**: Real-time collaboration by multiple users

Python for Data Wrangling

Programming-Based Approach with Powerful Libraries

- **Jupyter Notebook**
  - Web-based interactive computing environment
  - Ideal for:
    - Data cleaning and transformation
    - Statistical modeling
    - Visualization

- **Numpy (Numerical Python)**
  - Core package for numerical computation
  - Supports:
    - Multi-dimensional arrays/matrices
    - Mathematical operations on arrays

- **Pandas**
  - Designed for fast and efficient data manipulation
  - Capabilities:
    - Merge, join, filter, transform large datasets
    - Error prevention in handling misaligned data from multiple sources

---

📈 **R for Data Wrangling**

Another Programming Approach for Statistical and Data Manipulation

- **Dplyr**
  - Powerful and simple syntax for data manipulation
  - Supports tasks like filtering, grouping, and summarizing

- **Data.table**
  - Efficient tool for working with large datasets
  - Offers fast aggregation and reshaping

- **Jsonlite**

- Ideal for JSON parsing
- Useful when interacting with web APIs

⚖️ Choosing the Right Tool

- **Factors to Consider**:
    - Supported data size
    - Data structures (structured/unstructured)
    - Cleaning and transformation capabilities
    - Infrastructure and software dependencies
    - Ease of use and learning curve

🗒️ Summary

This video explores **various data wrangling tools and software**, from basic spreadsheet applications like Excel and Google Sheets to advanced cloud-based and programming tools like Trifacta, DataPrep, Python, and R. Each tool has unique strengths, such as automation, visualization, collaboration, or scalability. The right choice depends on the **data size, complexity, infrastructure needs**, and the **user's technical expertise**.

Key Takeaways:

- Spreadsheets are good for beginners and manual data wrangling.
- OpenRefine, DataPrep, and Trifacta are excellent for interactive and visual wrangling.
- Python and R offer powerful programmatic data wrangling through libraries like Pandas, Dplyr, and Jsonlite.
- Cloud tools like Google DataPrep and Watson Refinery reduce infrastructure management and provide scalability.

📚 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Spreadsheets	Basic tool; supports formula-based cleaning; Power Query (Excel), Query (Sheets)
OpenRefine	Open-source tool; supports multiple formats; menu-based UI

Google DataPrep	Cloud-based; auto-detects data types/schemas; suggests next steps
Watson Studio Refinery	IBM product; large-scale data transformation with governance enforcement
Trifacta Wrangler	Interactive, collaborative; exports to Excel/Tableau/R
Python	Offers Jupyter, Numpy, Pandas for versatile and scalable wrangling
R	Provides Dplyr, Data.table, Jsonlite for powerful data manipulation
Choosing the Right Tool	Depends on data size, structure, transformation needs, and user skill level

Data Cleaning

1. Importance of Data Quality

- Poor quality data weakens an organization's:
    - Competitive standing
    - Critical business objectives
- Common problems:
    - False conclusions
    - Ineffective decisions
    - Financial losses

---

2. Common Data Quality Issues

- Data from disparate sources may suffer from:
    - Missing values
    - Inaccuracies
    - Duplicates
    - Incorrect or missing delimiters
    - Inconsistent records
    - Insufficient parameters

---

3. Data Cleaning vs. Data Wrangling

- **Data Cleaning** is a **subset** of Data Wrangling.
- Cleaning occurs during the **Transformation Phase** of Data Wrangling.
- Steps in Data Cleaning Workflow:
    1. **Inspection**
    2. **Cleaning**
    3. **Verification**

---

4. Inspection Phase

- Goal: Detect issues and errors in data.
- Tools/Techniques:
    - **Rule-based validation** (scripts, constraints)
    - **Data profiling**:
        - Understand structure, content, relationships
        - Uncover anomalies (nulls, duplicates, range issues)
    - **Data visualization**:
        - Use statistical charts (e.g., average income plots) to spot outliers

---

5. Cleaning Phase

A. Missing Values

- **Strategies**:
    - Remove records
    - Source missing data (if intrinsic)
    - **Imputation**: Fill in missing data based on statistical methods
        - Example: Estimating missing ages in demographic data

B. Duplicate Data

- Identify and remove duplicates

C. Irrelevant Data

- Remove contextually unnecessary data
    - Example: Contact numbers in a general health study

D. Data Type Conversion

- Ensure values are stored in correct formats:
    - Numbers as numerical
    - Dates as date types

E. Standardization

- String formats (e.g., all lowercase)
- Date formats
- Units of measurement

F. Syntax Errors

- Fix:
    - Extra spaces
    - Typos
    - Format inconsistencies
        - Example: “New York” vs. “NY”

G. Outliers

- Outliers may or may not be incorrect:
    - **Incorrect**: Age = 5 in a voters database
    - **Valid but extreme**: Income = $1,000,000 in a group with avg. $100k–$200k
- Decision to retain/remove depends on **use case**

---

6. Verification Phase

- Re-inspect cleaned data:
    - Confirm data integrity post-cleaning
    - Ensure rules and constraints are still valid

---

7. Documentation

- Track:
    - All changes made
    - Reasons behind changes
    - Final data quality status
- Crucial for:
    - Transparency
    - Reporting data health

📝 Summary

This piece highlights the critical role of **data quality** in business success and provides a comprehensive overview of the **data cleaning workflow**, which is a crucial part of the broader **data wrangling process**. It breaks down the workflow into three key steps: **Inspection, Cleaning, and Verification**. Each step involves techniques like data profiling, imputation, type conversion, and standardization to address issues such as missing values, duplicates, syntax errors, irrelevant data, and outliers. The process concludes with **verification and documentation**, ensuring accuracy, rule compliance, and traceability.

Key Takeaways:

- Data cleaning is essential for informed business decisions.
- Inspection tools and techniques help identify data issues early.
- Cleaning methods must be tailored to the dataset and use case.
- Outliers should be carefully assessed, not automatically removed.
- Verification and thorough documentation ensure long-term data health.

📊 Table: What We Learnt in the Video

Topic/Section	What We Learnt

Importance of Data Quality	Poor data leads to bad decisions, affecting competitiveness and business goals
Common Data Issues	Missing values, duplicates, syntax errors, outliers, irrelevant data
Data Wrangling vs. Cleaning	Cleaning is a subset of wrangling, mainly in the transformation phase
Inspection	Use profiling, visualization, and rules to detect problems
Cleaning Techniques	Imputation, filtering, type conversion, standardization, removing duplicates
Handling Outliers	Assess based on context—may be valid or indicate data entry errors
Verification	Confirm data validity and integrity after cleaning
Documentation	Record all changes, reasons, and final data health for transparency

Viewpoints: Data Preparation and Reliability

🔍 **Portion of the Job Involving Data Gathering, Cleaning & Preparation**

Data Scientists / Data Analysts / ML Engineers:

- **Significant Time Investment:**
    - A large part of their job involves **gathering, cleaning, and preparing data**.
- **Company Dependence:**
    - The need for this work varies depending on how mature the company's **data engineering infrastructure** is.
    - Example: One data scientist highlights having a strong data engineering team, reducing their need to do this task.
- **Universal Importance:**
    - Regardless of role (data scientist, analyst, ML engineer), it's crucial to:
        - Understand **where the data comes from**
        - Recognize that **no dataset is perfect**
        - Be aware of **compromises, missing data, or errors** in the data

Key Point:

- Data professionals must **understand the underlying data** — its origins, structure, and limitations — to ensure quality analysis.

💼 Data Gathering & Preparation in Accounting (CPA Perspective)

Data Access:

- **Easier When Centralized:**
    - If data resides in a **general ledger system** or **central repository**, gathering is straightforward.

Time Allocation:

- **~30% of the Job** involves:
    - Laying out the data before analysis begins
    - Tracking data, ensuring **accuracy** and **completeness**

Importance of Accuracy:

- Must verify:
    - All relevant data (e.g., **12 months of statements**) is received
    - If not available, sufficient data must be present to:
        - **Project**
        - **Forecast**
        - **Estimate** missing information

✅ Ensuring Data Reliability

For General Data Analysis:

- **Run Summary Statistics:**
    - On individual columns to verify they make **real-world sense**
    - Example:
        - If a column for "monthly website visits" has a **negative number**, that indicates bad data

For Financial Data:

- **Must be:**
    - **Reliable**

    - **Non-biased**
    - **Free from error**

Key Practices:

- **Logic Checks:**
    - Start with high-level logic: *Does this data make sense?*
        - e.g., Expecting revenue to increase but it decreases — triggers investigation
- **Source & Query Validation:**
    - Confirm:
        - Correct **data source**
        - Proper **time period**
        - Accurate **GL accounts**

Once Validated:

- Only after confirming data reliability can one **dive deeper** into analytics and derive insights.

Summary

This segment explores how data professionals — from data scientists to CPAs — handle the crucial phase of **data preparation and reliability**. While the intensity of data wrangling varies by role and infrastructure, all experts agree on its **importance for accurate analysis**. They emphasize the need to understand the origin of data, run logic and reliability checks, and ensure data completeness before beginning any meaningful analysis. For financial data, this is even more critical due to the high stakes of misinterpretation.

Key Takeaways:

- Data cleaning and preparation are fundamental and time-consuming.
- Understanding data origin and structure is non-negotiable.
- Summary statistics and logic checks are essential for ensuring reliability.
- Financial data requires extra scrutiny due to its impact and legal implications.

📊 **Table: What We Learnt in the Video**

Topic/Section	What We Learnt

Data Prep in Data Science Roles	A major portion of work involves gathering and cleaning data. Data origin, structure, and quality must be deeply understood.
Role of Data Engineering	A strong data engineering team reduces the manual burden of data prep for data scientists.
CPA's Approach to Data Gathering	Gathering is easier with centralized systems. Roughly 30% of the job involves laying out and validating data.
Importance of Accuracy in Accounting	All statements (e.g., 12 months) must be accounted for. Estimations or projections are needed if data is missing.
Checking Data Reliability	Run summary statistics and high-level checks to ensure data consistency with expectations and reality.
Financial Data Integrity	Must be non-biased and error-free. Basic integrity checks must be performed before analysis begins.

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.