Table of Contents

Module 1

1 Introduction to Data Analytics

1.1 The Growing Importance of Data Analytics

Businesses today recognize the untapped value in data and consider data analytics a crucial factor for gaining a competitive advantage. According to a Forrester Consulting report:

"Businesses today recognize the untapped value in data and data analytics as a crucial factor for business competitiveness."

To support data and analytics initiatives, companies are:

- Hiring and upskilling people
- Expanding analytics teams
- Creating centers of excellence
- Establishing multi-pronged data and analytics practices

1.2 High Demand for Data Analysts

There is a significant mismatch between the supply of skilled data analysts and the growing demand, making this profession:

- Highly sought after
- Well-paid

1.3 Career Opportunities in Data Analytics

Mastering data analytics can be a career in itself or a foundation for other roles, including:

- Data Science
- Data Engineering
- Business Analytics
- Business Intelligence Analytics

1.4 Who Should Take This Course

This course is ideal for:

- Fresh graduates from any academic background
- Working professionals seeking a mid-career transition
- Data-driven decision-makers
- Anyone in an analytics-enabled role

1.5 What This Course Covers

This introductory course will help you:

- Understand the data ecosystem
- Learn the fundamentals of data analysis, including:
    - Data gathering
    - Data wrangling
    - Data mining
    - Data analysis
    - Data visualization

1.6 A Glimpse into the Data Analyst Role

You will:

- Experience a "Day in the Life of a Data Analyst"
- Hear practicing data analysts share:
    - How they entered the field
    - Career options and learning paths
    - What employers look for in candidates
    - Best practices in data analysis processes

2 The Modern Data Ecosystem

2.1 The Rise of Data

A 2020 Forbes report highlights the explosive growth of data:

"The constant increase in data processing speeds and bandwidth, the nonstop invention of new tools for creating, sharing, and consuming data, and the steady addition of new data creators and consumers worldwide, ensure that data growth continues unabated. Data begets more data in a constant virtuous cycle."

2.2 Components of a Modern Data Ecosystem

A modern data ecosystem consists of interconnected, independent, and evolving components:

- Integration of data from disparate sources
- Application of different analytical techniques and skill sets
- Collaboration among active stakeholders
- Use of various tools, applications, and infrastructure for data storage, processing, and dissemination

2.3 Data Sources

Data originates from a diverse range of structured and unstructured datasets:

- Text
- Images
- Videos
- Click streams
- User conversations
- Social media platforms
- IoT (Internet of Things) devices
- Real-time streaming events
- Legacy databases
- Professional data providers and agencies

These sources are dynamic and continue to expand in diversity.

2.4 Acquiring Data

When handling varied sources, the first step is to acquire data by pulling copies into a central repository:

- It involves identifying suitable formats, interfaces, and sources
- Key challenges: ensuring reliability, security, and integrity of the data

2.5 Organizing and Preparing Data

Once raw data is centralized:

- It must be organized, cleaned, and optimized for user access
- It must adhere to organizational standards and compliance requirements

2.5.1 Compliance Examples

- Adhering to privacy regulations related to personal, health, biometric, or household data
- Conforming to master data tables for consistency across organizational systems

2.5.2 Challenges

- Managing repositories with high availability, flexibility, accessibility, and security

2.6 Delivering Data to Stakeholders

Stakeholders such as analysts, programmers, and applications interact with the enterprise data repository:

2.6.1 Stakeholder Needs

- **Data Analysts:** Require access to raw data
- **Business Stakeholders:** Rely on reports and dashboards
- **Applications:** Need custom APIs for integration

2.6.2 Challenges

- Designing appropriate interfaces, APIs, and applications tailored to user needs

2.7 Influence of Emerging Technologies

New technologies are reshaping the data landscape:

2.7.1 Cloud Computing

- Provides limitless storage
- Offers high-performance computing
- Enables access to open-source tools and libraries

2.7.2 Machine Learning

- Facilitates predictive modeling using past data

2.7.3 Big Data

- Datasets are massive and highly varied
- Traditional tools are no longer sufficient
- New tools and techniques are required for deeper insights and business impact

3 Understanding Data Roles in Organizations

3.1 Data as a Strategic Asset

Organizations that leverage data to uncover opportunities and differentiate themselves are leading the future. From detecting fraud through financial transaction patterns to personalizing customer offers using behavior analysis, data has become the key to gaining a competitive edge.

3.2 The Data Ecosystem

To derive value from data, a wide array of skill sets is required. Let's explore the primary roles in the data ecosystem:

3.3 Role of Data Engineers

Data engineers are responsible for developing and maintaining data architectures. They ensure data is available for business operations and analysis.

3.3.1 Key Responsibilities

- Extract, integrate, and organize data from various sources

- Clean, transform, and prepare data
- Design, store, and manage data in repositories
- Enable data accessibility for business applications and stakeholders

3.3.2 Required Skills

- Programming knowledge
- Understanding of system and technology architectures
- Expertise in relational and non-relational databases

3.4 Role of Data Analysts

Data analysts translate data into plain language to assist decision-making within organizations.

3.4.1 Key Responsibilities

- Inspect and clean data
- Derive insights and identify patterns
- Apply statistical methods
- Visualize data for interpretation and presentation

3.4.2 Typical Questions They Answer

- Are users satisfied with our site's search functionality?
- What do people think about our rebranding initiative?
- Is there a correlation between sales of two different products?

3.4.3 Required Skills

- Proficiency in spreadsheets and querying databases
- Use of statistical tools for dashboards and charts
- Basic programming skills
- Analytical thinking and storytelling abilities

3.5 Role of Data Scientists

Data scientists build models that create predictive insights from historical data.

3.5.1 Key Responsibilities

- Analyze data for actionable insights
- Build machine learning and deep learning models
- Create predictive models

3.5.2 Typical Questions They Answer

- How many new social media followers are expected next month?

- What percentage of customers might churn in the next quarter?
- Is this financial transaction suspicious for a customer?

3.5.3 Required Skills

- Strong foundation in mathematics and statistics
- Programming and database knowledge
- Experience in building data models
- Domain expertise

3.6 Role of Business Analysts and BI Analysts

Business analysts and BI analysts interpret insights and predictions to drive decision-making.

3.6.1 Business Analysts

- Utilize insights from analysts and scientists
- Identify implications for business decisions

3.6.2 Business Intelligence (BI) Analysts

- Focus on external market forces
- Monitor and organize business function data
- Explore data to extract actionable insights that enhance performance

3.7 Summary

- **Data Engineers**: Convert raw data into usable formats.
- **Data Analysts**: Generate insights from processed data.
- **Data Scientists**: Predict future outcomes using historical data.
- **Business Analysts & BI Analysts**: Use insights and predictions to guide business strategies.

4 Data Analysis

4.1 What is Data Analysis?

Data analysis is the process of:

- Gathering data
- Cleaning data
- Analyzing and mining data
- Interpreting results
- Reporting findings

It involves discovering patterns and correlations in data, which lead to valuable insights and conclusions. This process helps businesses understand past performance and supports data-driven decision-making for future strategies. It also enables organizations to validate decisions before implementation, saving time and resources.

4.2 Types of Data Analysis

There are four primary types of data analysis, each serving a specific purpose in the data analysis process:

4.2.1 Descriptive Analytics

- **Purpose:** Understand what happened over a given time.
- **Method:** Summarizes past data.
- **Use Case:** Provides insights through key performance indicators (KPIs), such as cash flow analysis.

4.2.2 Diagnostic Analytics

- **Purpose:** Understand why something happened.
- **Method:** Explores data deeper using insights from descriptive analytics.
- **Use Case:** Investigates unusual changes like unexpected traffic spikes or sales surges without known triggers.

4.2.3 Predictive Analytics

- **Purpose:** Anticipate what might happen next.
- **Method:** Uses historical data and trends.
- **Use Case:** Applied in risk assessments and sales forecasts.
- **Note:** Predictions are probabilistic, not certain.

4.2.4 Prescriptive Analytics

- **Purpose:** Decide what should be done next.
- **Method:** Analyzes past decisions and potential outcomes.
- **Use Case:**
  - Self-driving cars determine optimal routes and speeds.
  - Airlines adjust ticket prices based on demand, weather, gas prices, and traffic conditions.

4.3 Key Steps in the Data Analysis Process

4.3.1 Understanding the Problem and Desired Result

- Define the problem to be solved.
- Clarify the desired outcomes.
- Establish where you are and where you need to go.

4.3.2 Setting a Clear Metric

- Identify what needs to be measured (e.g., product sales).

- Determine how it will be measured (e.g., quarterly, seasonally).

4.3.3 Gathering Data

- Identify required data and sources.
- Choose the right tools for data collection.

4.3.4 Cleaning Data

- Fix issues that can impact analysis accuracy.
- Address:
    - Missing or incomplete values
    - Outliers (e.g., age value of 150)
- Standardize data from multiple sources.

4.3.5 Analyzing and Mining Data

- Extract data and view it from multiple perspectives.
- Manipulate data to:
    - Identify trends
    - Discover correlations
    - Uncover patterns and variations

4.3.6 Interpreting Results

- Evaluate findings.
- Consider if the results are defensible.
- Acknowledge any limitations or special conditions.

4.3.7 Presenting Your Findings

- Communicate findings clearly and impactfully.
- Use formats such as:
    - Reports
    - Dashboards
    - Charts and graphs
    - Maps
    - Case studies

5 Data Analytics vs. Data Analysis

5.1 Overview

The terms **Data Analysis** and **Data Analytics** are often used interchangeably, including in this course. However, there is a subtle yet noteworthy distinction between the two. Some professionals argue that these terms represent different concepts and should not be used as synonyms.

5.2 Definitions

5.2.1 Analysis

- *Definition*: A detailed examination of the elements or structure of something.
- *Example*: Business analysis, psychoanalysis, etc.
- *Note*: Analysis does not necessarily involve numerical data.

5.2.2 Analytics

- *Definition*: The systematic computational analysis of data or statistics.
- *Example*: Data-driven assessments, predictive modeling, etc.
- *Note*: Analytics almost always involves numbers and data for computational evaluation and inference.

5.3 Technical Differences

- **Analysis** can be qualitative and may not rely on numerical data.
- **Analytics**, on the other hand, generally implies data-driven, quantitative assessment.
- Even when the prefix "Data" is omitted, *analytics* still suggests numerical computation.

5.4 Interpretation and Controversy

Some experts claim:

- **Data Analysis** involves historical data and is inference-based.
- **Data Analytics** focuses on predicting future outcomes.

6 Summary and Highlights

6.1 Overview of a Modern Data Ecosystem

A modern data ecosystem is a network of interconnected and continuously evolving components. It includes:

6.1.1 Data Formats and Sources

- Data is available in various formats, structures, and from numerous sources.

6.1.2 Enterprise Data Environment

- A staging area where raw data is organized, cleaned, and optimized for end-user consumption.

6.1.3 End-Users

- Includes business stakeholders, analysts, and programmers who utilize data for diverse purposes.

6.1.4 Emerging Technologies

- Technologies like Cloud Computing, Machine Learning, and Big Data are reshaping the data landscape and expanding its possibilities.

6.1.5 Key Roles in the Ecosystem

- Data Engineers
- Data Analysts
- Data Scientists
- Business Analysts
- Business Intelligence Analysts

These professionals collaborate to derive insights and deliver business value from data.

6.2 Types of Data Analysis

Based on specific goals and desired outcomes, four primary types of data analysis are identified:

6.2.1 Descriptive Analytics

- Helps answer: **"What happened?"**

6.2.2 Diagnostic Analytics

- Helps answer: **"Why did it happen?"**

6.2.3 Predictive Analytics

- Uses historical data and trends to predict: **"What will happen next?"**

6.2.4 Prescriptive Analytics

- Recommends actions by answering: **"What should be done next?"**

6.3 The Data Analysis Process

The process of analyzing data involves several key steps:

6.3.1 Problem Understanding

- Define the problem clearly and determine the desired outcomes.

6.3.2 Setting Evaluation Metrics

- Establish clear metrics for evaluating success.

6.3.3 Data Handling and Analysis

- Gather, clean, analyze, and mine data to uncover insights.

6.3.4 Communicating Results

- Present findings in a way that effectively informs and influences decision-making.

7 Responsibilities and Skillsets of a Data Analyst

7.1 Typical Responsibilities of a Data Analyst

While the role of a Data Analyst may vary based on the organization and its level of data maturity, there are common responsibilities found across most data analyst roles:

- Acquiring data from primary and secondary sources
- Creating queries to extract data from databases and data collection systems
- Filtering, cleaning, standardizing, and reorganizing data for analysis
- Using statistical tools to interpret data sets
- Identifying patterns and correlations using statistical techniques
- Analyzing complex data sets to interpret trends
- Preparing reports and charts to communicate insights
- Documenting the data analysis process for transparency and reproducibility

7.2 Skillsets Required for a Data Analyst

The role demands a combination of technical, functional, and soft skills.

7.2.1 Technical Skills

- Expertise in spreadsheets (e.g., Microsoft Excel, Google Sheets)
- Proficiency in statistical analysis and visualization tools:
    - IBM Cognos
    - IBM SPSS
    - Oracle Visual Analyzer
    - Microsoft Power BI
    - SAS
    - Tableau
- Programming languages:
    - Python
    - R
    - (In some cases) C++, Java, MATLAB
- Strong knowledge of SQL and experience with relational and NoSQL databases

- Ability to work with data repositories:
    - Data marts
    - Data warehouses
    - Data lakes
    - Data pipelines
- Familiarity with Big Data tools:
    - Hadoop
    - Hive
    - Spark

7.2.2 Functional Skills

- Proficiency in Statistics for analyzing and validating data
- Analytical thinking for interpreting data and making forecasts
- Problem-solving to derive actionable insights
- Probing to understand problems from multiple stakeholder perspectives
- Data visualization skills to effectively present findings
- Project management for handling processes, people, and timelines

7.2.3 Soft Skills

- Collaboration with business and cross-functional teams
- Effective communication for reporting and presenting findings
- Storytelling to make data insights compelling and persuasive
- Curiosity to explore unexpected patterns and question assumptions
- Intuition developed through pattern recognition and experience

8 Generative AI: An Essential Skill for today's Data Analysts

8.1 Introduction

As a beginner in data analytics, you're stepping into a field that's rapidly evolving. Generative AI is becoming an essential tool for data analysts, allowing them to create new content and gain deeper insights. Let's explore what generative AI is and how it can enhance your skills.

8.2 What is generative AI?

Generative AI refers to a class of artificial intelligence models that create new content such as text, images, music, and more by learning patterns from existing data.

Generative AI can respond naturally to human conversation and serve as a tool for customer service and personalization of customer workflows. For example, you can use AI-powered chatbots, voice bots, and virtual assistants that respond more accurately to customers for first-contact resolution.

8.3 How does generative AI work?

Generative AI starts with a prompt that could be in the form of a text, an image, a video, a design, musical notes, or any input that the AI system can process. Various AI algorithms then return new content in response to the prompt. Content can include essays, solutions to problems, or realistic fakes created from pictures or audio of a person.

Early versions of generative AI required submitting data via an API or an otherwise complicated process. Developers had to familiarize themselves with special tools and write applications using languages such as Python.

Now, pioneers in generative AI are developing better user experiences that let you describe a request in plain language. After an initial response, you can also customize the results with feedback about the style, tone, and other elements you want the generated content to reflect.

8.4 Key techniques in generative AI:

Generative adversarial networks (GANs): GANs consist of two neural networks: the generator and the discriminator. The generator creates new data, whereas the discriminator evaluates it. Over time, the generator improves to produce realistic data.

Variational autoencoders (VAEs): VAEs encode input data into a compressed format and then decode it back, generating new data points similar to the input data.

Transformers: Used primarily in natural language processing (NLP), transformers generate human-like text by predicting the next word in a sequence. Generative Pre-trained Transformer 3 (GPT-3) is a notable example.

8.5 Generative AI models

Generative AI models combine various AI algorithms to represent and process content. For example, to generate text, various NLP techniques transform raw characters (e.g., letters, punctuation, and words) into sentences, parts of speech, entities, and actions, which are represented as vectors using multiple encoding techniques. Similarly, images are transformed into various visual elements, also expressed as vectors. One caution is that these techniques can also encode the biases, racism, deception, and puffery contained in the training data.

Once developers settle on a way to represent the world, they apply a particular neural network to generate new content in response to a query or prompt. Techniques such as GANs and VAEs—neural networks with a decoder and encoder—are suitable for generating realistic human faces, synthetic data for AI training, or even facsimiles of particular humans.

Recent progress in transformers, such as Google's Bidirectional Encoder Representations from Transformers (BERT), OpenAI's GPT, and Google AlphaFold, have also resulted in neural networks that can not only encode language, images, and proteins but also generate new content.

8.6 What are the use cases for generative AI?

Generative AI can be applied in various use cases to generate virtually any kind of content. The technology is becoming more accessible to users of all kinds thanks to cutting-edge breakthroughs like GPT that can be tuned for different applications.

Some of the use cases for generative AI include the following:

- Implementing chatbots for customer service and technical support.
- Deploying deepfakes for mimicking people or even specific individuals.
- Improving dubbing for movies and educational content in different languages.
- Writing email responses, dating profiles, resumes, and term papers.
- Creating photorealistic art in a particular style.
- Improving product demonstration videos.
- Suggesting new drug compounds to test.
- Designing physical products and buildings.
- Optimizing new chip designs.
- Writing music in a specific style or tone.

8.7 What are the benefits of generative AI?

Generative AI can be applied extensively across many areas of the business. It can make it easier to interpret and understand existing content and automatically create new content. Developers are exploring ways that generative AI can improve existing workflows, with an eye to adapting workflows entirely to take advantage of the technology. Some of the potential benefits of implementing generative AI include the following:

- Automating the manual process of writing content.
- Reducing the effort of responding to emails.
- Improving the response to specific technical queries.
- Creating realistic representations of people.
- Summarizing complex information into a coherent narrative.
- Simplifying the process of creating content in a particular style

8.8 What are the limitations of generative AI?

Early implementations of generative AI vividly illustrate its many limitations. Some of the challenges generative AI presents result from the specific approaches used to implement particular use cases. For example, a summary of a complex topic is easier to read than an explanation that includes various sources supporting key points. The readability of the summary, however, comes at the expense of a user being able to vet where the information comes from.

Here are some of the limitations to consider when implementing or using a generative AI app:

- It does not always identify the source of content.
- It can be challenging to assess the bias of original sources.
- Realistic-sounding content makes it harder to identify inaccurate information.

- It can be difficult to understand how to tune in to new circumstances.
- Results can gloss over bias, prejudice, and hatred.

8.9 What are the concerns surrounding generative AI?

·The rise of generative AI is also fueling various concerns. These relate to the quality of results, the potential for misuse and abuse, and the potential to disrupt existing business models. Here are some of the specific types of problematic issues posed by the current state of generative AI:

- It can provide inaccurate and misleading information.
- It is more difficult to trust without knowing the source and provenance of information.
- It can promote new kinds of plagiarism that ignore the rights of content creators and artists of original content.
- It might disrupt existing business models built around search engine optimization and advertising.
- It makes it easier to generate fake news.
- It makes it easier to claim that real photographic evidence of wrongdoing was just an AI-generated fake.
- It could impersonate people for more effective social engineering cyberattacks.
- Given the newness of GenAI tools and their rapid adoption, enterprises should prepare for the inevitable "trough of disillusionment" that's part and parcel of emerging technology by adopting sound AI engineering practices and making responsible AI a cornerstone of their GenAI efforts, ensuring transparency, ethical considerations, and long-term sustainability in their AI implementations.

8.10 What are some examples of generative AI tools?

Generative AI tools exist for various modalities, such as text, imagery, music, code, and voices. Some popular AI content generators to explore include the following:

- Text generation tools include GPT, Jasper, AI-Writer, and Lex.
- Image generation tools include Dall-E 2, Midjourney, and Stable Diffusion.
- Music generation tools include Amper, Dadabots, and MuseNet.
- Code generation tools include codeStarter, Codex, GitHub Copilot, and Tabnine.
- Voice synthesis tools include Descript, Listnr, and Podcast.ai.
- AI chip design tool companies include Synopsys, Cadence, Google, and NVIDIA.

8.11 Applications of generative AI in data analytics

Generative AI has many applications that can enhance your data analytics work:

Data augmentation: Create synthetic data to augment existing data sets, which is especially useful when data is scarce or imbalanced. This can improve predictive model performance.

Anomaly Detection: Identify anomalies or outliers by understanding the distribution of normal data. This is valuable in fraud detection, network security, and quality control.

Text and image generation: Generate realistic text and images for marketing, content creation, and customer engagement, such as automatic product descriptions and marketing visuals.

Simulation and forecasting: Simulate scenarios and forecast future events by generating potential outcomes from historical data. This is crucial in financial planning, supply chain management, and strategic decision-making.

8.12 Conclusion

Generative AI is a transformative technology that can significantly enhance your capabilities as a data analyst. By mastering generative AI techniques, you can unlock new possibilities in data augmentation, anomaly detection, content creation, and forecasting. As you embark on this journey, remember to balance innovation with ethical responsibility, ensuring that AI is used positively.

9 A Day in the Life of a Data Analyst

9.1 Overview of Daily Activities

A Data Analyst's day can involve a variety of tasks:

- Acquiring data from different sources
- Writing queries to extract data
- Analyzing data to find insights
- Creating reports and dashboards
- Engaging with stakeholders to gather requirements and present results
- Cleaning and preparing data — a major part of the role

9.2 A Typical Insight-Focused Day

Among the diverse responsibilities, one of the most rewarding aspects is uncovering insights from data.

9.2.1 Analyst Introduction

Sivaram Jaladi, a Data Analyst at **Fluentgrid**, shares his experience. Fluentgrid, based in Vishakhapatnam, India, is an IBM partner and Beacon award winner, specializing in smart grid and smart city solutions.

9.2.2 Problem Identification

A client — a power utility company in South India — is experiencing a surge in overbilling complaints. The pattern suggests it's not random.

9.3 Data Exploration Process

9.3.1 Initial Data Sources

Sivaram begins by reviewing:

- Complaint data
- Subscriber information
- Billing data

9.3.2 Forming Hypotheses

He outlines key questions:

- Are overbilling complaints tied to specific consumption ranges?
- Do certain areas report more complaints?
- Are the same subscribers submitting repeated complaints?
- What's the frequency of these repeated complaints?

9.3.3 Dataset Identification

Datasets are selected to test the hypotheses:

- Average annual, quarterly, and monthly billing amounts
- Geographic location of complainants
- Subscription start date
- Meter make and serial numbers

9.4 Data Analysis and Findings

9.4.1 Billing Range Analysis

- Investigates if complaints correlate with specific billing amounts.

9.4.2 Geographic Concentration

- Identifies zip codes with high complaint frequency.
- This reveals localized patterns worth deeper analysis.

9.4.3 Subscriber Duration

- Over 95% of complainants have been subscribers for more than seven years.
- However, not all long-term subscribers reported issues.

9.4.4 Meter Data Investigation

- Analysis of meter serial numbers and manufacturers shows that:
    - Complaints correlate with a specific batch of meters.
    - These meters were installed in areas showing high complaint volumes.

9.5 Presenting Insights

Sivaram prepares to share:

- Key findings
- Data sources
- Analytical process used to reach conclusions

This thorough documentation adds credibility and transparency.

9.6 Future Considerations

The issue might be resolved, or it could resurface with:

- Similar complaints but different data patterns
- Entirely new types of complaints requiring fresh analysis

10 Applications of Data Analytics

10.1 Data Analytics in Everyday Life

The applications of data analytics are pervasive in daily life. Every commercial seen by a consumer is backed by analytics — determining what information to present, whether it's survey results like "four out of 10 dentists recommend" or nutritional information like calorie counts. Even in personal health scenarios, such as monitoring blood sugar levels in diabetes, analytical processes are at play. These examples show that data analytics is not separate from daily life but an integral part of it.

10.2 Cross-Industry Relevance

Analytics today is widely applicable across every industry, vertical, and organizational function:

10.2.1 Sales and Financials

- Sales pipeline analysis
- Monthly financial reports with predefined formats

10.2.2 Human Resources

- Headcount planning
- Headcount reviews

10.2.3 Industry Examples

- Airlines
- Pharmaceuticals

- Banking

Each function within these sectors benefits significantly from data-driven insights.

10.3 Adaptation During the Pandemic

During the pandemic, companies have closely monitored changing customer buying habits. These behaviors often differed from expectations, making data analytics essential for:

- Adapting business strategies
- Meeting shifting customer demands
- Staying competitive and relevant in real-time

10.4 Applications in Finance

Data analytics is increasingly influential in finance, particularly through the use of alternative data:

10.4.1 Sentiment Analysis

- Analyzing tweets and news stories to enhance traditional financial models
- Informing smarter investment decisions

10.4.2 Satellite Imagery

- Tracking industrial development using visual satellite data

10.4.3 Geolocation Data

- Monitoring foot traffic to stores
- Predicting sales volumes based on location data trends

11 Summary and Highlights

The role of a Data Analyst spans across:

- Acquiring data that best serves the use case.
- Preparing and analyzing data to understand what it represents.
- Interpreting and effectively communicating the message to stakeholders who need to act on the findings.
- Ensuring that the process is documented for future reference and repeatability.

In order to play this role successfully, Data Analysts need a mix of technical, functional, and soft skills.

- Technical Skills include varying levels of proficiency in using spreadsheets, statistical tools, visualization tools, programming, and querying languages, and the ability to work with different types of data repositories and big data platforms.
- An understanding of Statistics, Analytical techniques, problem-solving, the ability to probe a situation from multiple perspectives, data visualization, and project management skills – all of which come under Functional Skills a Data Analyst needs in order to play an effective role.
- Soft Skills include the ability to work collaboratively, communicate effectively, tell a compelling story with data, and garner support and buy-in from stakeholders. Curiosity to explore different pathways and intuition that helps to give a sense of the future based on past experiences are also essential skills for being a good Data Analyst.

Module 2

12 Data Analyst's Ecosystem

12.1 Overview

A data analyst's ecosystem includes the infrastructure, software, tools, frameworks, and processes used to gather, clean, analyze, mine, and visualize data. This section provides a high-level overview before diving into details in later sections.

12.2 Types of Data

Based on how well-defined the structure of the data is, data can be categorized into three types:

12.2.1 Structured Data

- Follows a rigid format
- Can be organized into rows and columns
- Commonly found in databases and spreadsheets

12.2.2 Semi-Structured Data

- Mix of structured and unstructured elements
- Example: Emails (structured fields like sender and recipient + unstructured message content)

12.2.3 Unstructured Data

- Complex, qualitative, and not reducible to rows and columns
- Examples: Photos, videos, text files, PDFs, social media content

12.3 Data Sources and Formats

Data comes in various file formats and is collected from multiple sources, such as:

- Relational and non-relational databases
- APIs and web services
- Data streams
- Social platforms
- Sensor devices

12.4 Data Repositories

The type, format, and source of data influence the choice of data repository. Key types include:

- Databases
- Data warehouses
- Data marts
- Data lakes
- Big data stores

12.4.1 Working with Big Data

- Requires big data warehouses to store and process large-volume, high-velocity data
- Involves frameworks for real-time analytics on complex data

12.5 Languages in the Ecosystem

The ecosystem includes several types of languages:

12.5.1 Query Languages

- SQL: Used for querying and manipulating data

12.5.2 Programming Languages

- Python: Used to develop data applications

12.5.3 Shell and Scripting Languages

- Shell scripts: Automate repetitive operational tasks

12.6 Tools and Frameworks

Various tools and processes support each stage of the analytics workflow:

12.6.1 Data Collection and Processing

- Tools for gathering, extracting, transforming, and loading data

12.6.2 Data Preparation

- Tools for data wrangling and cleaning

12.6.3 Analysis and Visualization

- Tools for mining, analyzing, and visualizing data

12.6.4 Examples

- Spreadsheets
- Jupyter Notebooks
- IBM Cognos

13 Types of Data

13.1 Introduction to Data

Data is unorganized information that is processed to make it meaningful. It comprises facts, observations, perceptions, numbers, characters, symbols, and images that can be interpreted to derive meaning.

One of the ways data can be categorized is by its structure. Data can be:

- Structured
- Semi-structured
- Unstructured

13.2 Structured Data

Structured data has a well-defined structure or adheres to a specified data model. It can be stored in well-defined schemas such as databases and is often represented in a tabular format with rows and columns.

13.2.1 Characteristics of Structured Data

- Contains objective facts and numbers.
- Easily collected, exported, stored, and organized.
- Suitable for typical databases.
- Can be examined using standard data analysis tools and methods.

13.2.2 Sources of Structured Data

- SQL Databases
- Online Transaction Processing (OLTP) Systems
- Spreadsheets (e.g., Excel, Google Spreadsheets)
- Online forms

- Sensors (e.g., GPS, RFID tags)
- Network and Web server logs

13.2.3 Storage of Structured Data

- Typically stored in relational or SQL databases.

13.3 Semi-structured Data

Semi-structured data has some organizational properties but lacks a fixed or rigid schema. It cannot be stored in traditional databases in the form of rows and columns.

13.3.1 Characteristics of Semi-structured Data

- Contains tags, elements, or metadata used to group and organize data in a hierarchy.

13.3.2 Sources of Semi-structured Data

- E-mails
- XML and other markup languages
- Binary executables
- TCP/IP packets
- Zipped files
- Integrated data from different sources

13.3.3 Storage and Formats

- XML and JSON are commonly used to define tags and attributes for storing and exchanging semi-structured data.

13.4 Unstructured Data

Unstructured data lacks an identifiable structure and cannot be organized using conventional relational databases.

13.4.1 Characteristics of Unstructured Data

- Does not follow a specific format, sequence, semantics, or rules.
- Capable of handling heterogeneous sources.
- Widely used in business intelligence and analytics.

13.4.2 Sources of Unstructured Data

- Web pages
- Social media feeds
- Image files (e.g., JPEG, GIF, PNG)
- Video and audio files
- Documents and PDFs

- PowerPoint presentations
- Media logs
- Surveys

13.4.3 Storage of Unstructured Data

- Stored in files and documents (e.g., Word documents) for manual analysis.
- Can also be stored in NoSQL databases with built-in analysis tools.

14 Common Data File Formats

14.1 Importance of Understanding File Formats

As a data professional, you will work with various data file types and formats. Understanding the structure, benefits, and limitations of each format helps you choose the most suitable one for your data and performance needs.

14.2 Standard File Formats

Some of the commonly used file formats include:

- Delimited Text Files
- Microsoft Excel Open XML Spreadsheet (XLSX)
- Extensible Markup Language (XML)
- Portable Document Format (PDF)
- JavaScript Object Notation (JSON)

14.3 Delimited Text File Formats

Delimited text files store data in plain text, where each row contains values separated by a delimiter.

14.3.1 Common Delimiters

- Comma (,)
- Tab (\t)
- Colon (:)
- Vertical Bar (|)
- Space ( )

14.3.2 CSV and TSV Files

- **CSV (Comma-Separated Values):** Uses commas as delimiters.
- **TSV (Tab-Separated Values):** Uses tabs as delimiters and is useful when data contains commas.

14.3.3 Characteristics

- Each row represents a record.
- The first row usually acts as a column header.
- Columns may contain various data types such as dates, strings, or integers.
- Field values can be of any length.
- Supported by most applications.
- Simple and widely used for data interchange.

14.4 Microsoft Excel Open XML Spreadsheet (XLSX)

- XML-based spreadsheet format created by Microsoft.
- Files are also known as workbooks.

14.4.1 Structure

- A workbook can contain multiple worksheets.
- Worksheets consist of rows and columns, with each cell at their intersection containing data.

14.4.2 Features

- Open format: Compatible with many applications.
- Supports all Excel functions.
- Considered secure, as it cannot save malicious code.

14.5 Extensible Markup Language (XML)

- The markup language is used for encoding data.
- Both human-readable and machine-readable.
- Designed for sending information over the internet.

14.5.1 Characteristics

- Self-descriptive structure.
- Does not use predefined tags (unlike HTML).
- Platform and programming language independent.
- Ideal for data sharing across different systems.

14.6 Portable Document Format (PDF)

- Developed by Adobe to present documents consistently across platforms.

14.6.1 Features

- Independent of software, hardware, and operating systems.
- Widely used for legal and financial documents.
- Suitable for forms and data input.

14.7 JavaScript Object Notation (JSON)

- Text-based open standard for structured data transmission.
- Language-independent and readable in any programming language.

14.7.1 Benefits

- Easy to use.
- Compatible with most browsers.
- Ideal for sharing all types and sizes of data, including audio and video.
- Commonly used in APIs and web services.

15 Data Sources for Analysis

15.1 Overview

Data sources today are more dynamic and diverse than ever. This section covers common sources used in data analytics, including:

- Relational Databases
- Flat Files and XML Datasets
- APIs and Web Services
- Web Scraping
- Data Streams and Feeds

15.2 Relational Databases

Organizations use relational databases in internal applications to manage daily business activities such as:

- Customer transactions
- Human resource activities
- Workflow management

Popular relational databases include:

- SQL Server
- Oracle
- MySQL
- IBM DB2

These databases store data in a structured format and are commonly used for analysis, such as:

- Analyzing sales across regions using retail transaction data
- Making sales projections from CRM system data

15.3 External Datasets

In addition to internal data, organizations can use external datasets:

- **Government Datasets**: Demographic and economic data
- **Commercial Datasets**: Point-of-sale, financial, and weather data for strategic decisions

These datasets are commonly shared as:

- Flat files (e.g., CSV)
- Spreadsheet files
- XML documents

15.4 Flat Files

Flat files store data in plain text format:

- One record or row per line
- Values separated by delimiters (commas, semicolons, tabs)
- Typically map to a single table

15.4.1 CSV Files

- The most common flat file format
- Values separated by commas

15.4.2 Spreadsheet Files

Spreadsheets are structured like flat files but offer more features:

- Organized into rows and columns
- Can contain multiple worksheets
- Store data in .XLS or .XLSX format (e.g., Microsoft Excel)
- Include formatting, formulas, and custom storage formats

Other tools include:

- Google Sheets
- Apple Numbers
- LibreOffice

15.5 XML Datasets

XML files store data using markup tags:

- Support hierarchical data structures
- Common use cases:

    - Online surveys
    - Bank statements
    - Other unstructured datasets

15.6 APIs and Web Services

APIs (Application Programming Interfaces) and Web Services allow access to data via network or web requests:

- Return data in formats like plain text, XML, HTML, JSON, or media files

15.6.1 Popular API Use Cases

- **Social Media**: Twitter and Facebook APIs for sentiment analysis and opinion mining
- **Financial Markets**: Stock Market APIs for share prices and trading analysis
- **Validation**: Lookup APIs for data cleaning and geolocation
- **Internal/External Data Access**: Pulling data from databases

15.7 Web Scraping

Web scraping extracts data from unstructured web sources:

- Also called screen scraping, web harvesting, or web data extraction
- Enables extraction of text, images, videos, contacts, product details, etc.

15.7.1 Use Cases

- Price comparisons from retail and e-commerce websites
- Generating sales leads
- Extracting forum and community data
- Creating datasets for machine learning models

15.7.2 Popular Tools

- BeautifulSoup
- Scrapy
- Pandas
- Selenium

15.8 Data Streams

Data streams consist of continuous data flows from sources such as:

- IoT devices
- GPS data from vehicles
- Web programs and social media
- Instruments and sensors

15.8.1 Applications

- Financial trading from stock tickers
- Demand prediction from retail transactions
- Threat detection from surveillance feeds
- Sentiment analysis from social media
- Monitoring industrial/farming equipment
- Web performance tracking from click data
- Flight event monitoring for rescheduling

15.8.2 Tools for Stream Processing

- Apache Kafka
- Apache Spark Streaming
- Apache Storm

15.9 RSS Feeds

RSS (Really Simple Syndication) feeds are used to:

- Capture updated content from forums and news sites
- Provide real-time updates via feed readers that convert RSS into readable streams

16 Languages for Data Professionals

16.1 Overview

Data professionals work with various types of languages that can be categorized into three main groups:

- Query Languages
- Programming Languages
- Shell and Scripting Languages

Proficiency in at least one language from each category is essential.

16.2 Query Languages

16.2.1 SQL (Structured Query Language)

SQL is primarily used to access and manipulate data in relational databases.

16.2.1.1 Uses of SQL

- Insert, update, and delete records
- Create databases, tables, and views
- Write stored procedures for reusable instructions

16.2.1.2 Advantages of SQL

- **Portability**: Platform-independent usage
- **Wide Compatibility**: Works with various databases and repositories
- **Simple Syntax**: Similar to English, enabling easier learning
- **Efficiency**: Retrieves large amounts of data quickly
- **Interpreter System**: Enables quick execution and prototyping
- **Popularity**: Large user base and comprehensive documentation

16.3 Programming Languages

16.3.1 Python

Python is a widely used, open-source, high-level programming language.

16.3.1.1 Key Features

- Expresses concepts with fewer lines of code
- Simple syntax and readability
- Ideal for beginners due to its low learning curve
- Powerful in handling large-scale, high-computational data tasks
- Offers parallel processing for performance efficiency

16.3.1.2 Libraries and Functionalities

- **Data Cleaning and Analysis**: Pandas
- **Statistical Analysis**: Numpy, Scipy
- **Web Scraping**: BeautifulSoup, Scrapy
- **Data Visualization**: Matplotlib, Seaborn
- **Image Processing**: OpenCV

16.3.1.3 Benefits of Python

- Easy to learn
- Open-source and community-driven
- Cross-platform compatibility (Windows and Linux)
- Rich set of analytics libraries
- Supports multiple paradigms: object-oriented, functional, procedural, and imperative

16.3.2 R

R is an open-source programming language ideal for statistics and data analysis.

16.3.2.1 Key Features

- Platform-independent
- Compatible with other programming languages (e.g., Python)
- Highly extensible through user-defined functions
- Capable of handling both structured and unstructured data

16.3.2.2 Libraries and Tools

- **Visualization**: ggplot2, Plotly
- **Reporting**: Embedded scripts and interactive web apps
- **Strength**: Preferred for developing statistical tools

16.3.3 Java

Java is an object-oriented, platform-independent language widely used in data analytics.

16.3.3.1 Uses in Data Analytics

- Data cleaning, import/export, statistical analysis, and visualization
- Supports popular big data frameworks: Hadoop, Hive, Spark
- Suited for performance-critical projects

16.4 Shell and Scripting Languages

16.4.1 Unix/Linux Shell

Shell scripts are collections of UNIX commands for automating tasks.

16.4.1.1 Common Use Cases

- File manipulation
- Program execution
- System administration (e.g., backups, log analysis)
- Installation scripts and batch processing

16.4.2 PowerShell

PowerShell is a Microsoft automation tool and scripting framework.

16.4.2.1 Key Features

- Works with structured data formats (JSON, CSV, XML)
- Includes a command-line shell and scripting language
- Object-based pipeline for advanced data operations

16.4.2.2 Use Cases

- Data mining
- GUI development
- Dashboard and report creation
- Automation of administrative tasks

17 Summary and Highlights

A data analyst ecosystem includes the infrastructure, software, tools, frameworks, and processes used to gather, clean, analyze, mine, and visualize data.

Based on how well-defined the structure of the data is, data can be categorized as:

- Structured Data, that is data which is well organized in formats that can be stored in databases.
- Semi-structured data is data that is partially organized and partially free form.
- Unstructured Data is data that can not be organized conventionally into rows and columns.

Data comes in a wide-ranging variety of file formats, such as delimited text files, spreadsheets, XML, PDF, and JSON, each with its own list of benefits and limitations of use.

Data is extracted from multiple data sources, ranging from relational and non-relational databases to APIs, web services, data streams, social platforms, and sensor devices.

Once the data is identified and gathered from different sources, it needs to be staged in a data repository so that it can be prepared for analysis. The type, format, and sources of data influence the type of data repository that can be used.

Data professionals need a host of languages that can help them extract, prepare, and analyze data. These can be classified as:

- Querying languages, such as SQL, are used for accessing and manipulating data from databases.
- Programming languages such as Python, R, and Java, for developing applications and controlling application behavior.
- Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, for automating repetitive operational tasks.

18 Data Repositories

18.1 Overview

A data repository refers to a collection of data that has been organized and isolated for use in business operations, reporting, or data analysis. It may consist of one or more databases and vary in size from small systems to large-scale infrastructures.

This section provides an overview of different types of data repositories, including:

- Databases
- Data Warehouses
- Big Data Stores

Each of these will be examined in greater detail in upcoming sections.

18.2 Databases

A **database** is a structured collection of data designed for:

- Input
- Storage
- Search and retrieval
- Modification

18.2.1 Database Management System (DBMS)

A **Database Management System (DBMS)** is a suite of programs used to create and maintain a database. Key functionalities include:

- Storing data
- Modifying data
- Extracting information using queries

18.2.1.1 Example

To find customers who have been inactive for over six months, the DBMS uses querying to retrieve relevant records from the database.

Note: Although "database" and "DBMS" are different, they are often used interchangeably.

18.2.2 Types of Databases

The choice of the database depends on factors like:

- Data type and structure
- Querying mechanisms
- Latency requirements
- Transaction speeds
- Intended use of data

The two main types of databases are:

18.2.2.1 Relational Databases

Also known as **RDBMSes**, relational databases:

- Organize data in a tabular format (rows and columns)
- Follow a strict schema
- Are optimized for complex queries and large datasets
- Use **Structured Query Language (SQL)** as the standard query language

18.2.2.2 Non-Relational Databases

Also referred to as **NoSQL** (Not Only SQL), non-relational databases:

- Support schema-less or free-form data storage
- Are ideal for handling large volumes of diverse data
- Emerged due to the rise of cloud computing, IoT, and social media
- Are optimized for speed, scalability, and flexibility
- Are widely used for **Big Data** processing

18.3 Data Warehouses

A **data warehouse** acts as a central repository, integrating information from various sources. It consolidates data through the **Extract, Transform, Load (ETL)** process:

18.3.1 ETL Process

- **Extract**: Gather data from multiple sources
- **Transform**: Clean and prepare the data for use
- **Load**: Import the processed data into the central repository

Data warehouses are typically used for:

- Business intelligence
- Analytical reporting

18.3.2 Related Concepts

- Data Marts and Data Lakes (covered later)
- Traditionally, data warehouses were built on relational databases
- Now, with NoSQL and diverse data sources, non-relational repositories are also used for data warehousing

18.4 Big Data Stores

Big Data Stores are designed to manage extremely large datasets and provide:

- Distributed computational infrastructure
- Scalable storage systems
- High-speed data processing capabilities

They are essential in environments where data volume, velocity, and variety exceed traditional processing limits.

19 Relational Databases

19.1 Structure of Relational Databases

A relational database organizes data into tables, where each table consists of rows (records) and columns (attributes). Tables can be linked based on shared data, allowing relationships between them.

19.1.1 Example: Customer and Transaction Tables

- **Customer Table Attributes**: Customer ID, Customer Name, Customer Address, Customer Primary Phone
- **Transaction Table Attributes**: Transaction Date, Customer ID, Transaction Amount, Payment Method
- Tables are related through the **Customer ID** field.

19.2 Purpose and Advantages of Linking Tables

Linking tables enables:

- Queries that retrieve new tables from existing ones
- Understanding relationships among data
- Generating reports like customer statements
- Making better data-driven decisions

19.3 Querying Data with SQL

- SQL (Structured Query Language) is used for querying relational databases.
- SQL allows for processing millions of records efficiently.

19.4 Comparison with Spreadsheets

- Both use rows and columns, but relational databases:
  - Are optimized for large volumes of data
  - Follow a well-defined schema
  - Support relationships between tables

19.5 Features and Benefits of Relational Databases

19.5.1 Data Integrity and Consistency

- Unique rows and columns per table
- Relationships minimize data redundancy
- Fields can be restricted to specific data types and values

19.5.2 Performance and Security

- Fast processing and data retrieval
- Controlled access with strong security policies

19.5.3 Scalability and Support

- Available as:
  - Open-source (self-supported or commercially supported)
  - Commercial closed-source systems
- Popular systems include:
  - IBM DB2
  - Microsoft SQL Server
  - MySQL
  - Oracle Database
  - PostgreSQL

19.5.4 Cloud-Based Relational Databases

- Known as **Database-as-a-Service (DBaaS)**
- Benefit from cloud computing and storage scalability
- Popular services:
  - Amazon RDS
  - Google Cloud SQL
  - IBM DB2 on Cloud
  - Oracle Cloud
  - SQL Azure

19.5.5 Maturity and Flexibility

- RDBMS is a mature technology with widespread documentation and community support
- Supports live updates: Add columns, rename tables, and more without downtime

19.5.6 Reduced Redundancy

- Centralized data management (e.g., one customer entry, linked across tables)

19.5.7 Backup and Recovery

- Easy export/import
- Continuous mirroring for cloud databases ensures minimal data loss

19.5.8 ACID Compliance

- Ensures reliable and consistent transactions
- ACID: **Atomicity, Consistency, Isolation, Durability**

19.5.9 ACID stands for:

19.5.9.1 Atomicity

- Ensures that **all operations in a transaction are completed successfully**, or **none at all**.
- If one part of the transaction fails, the **entire transaction is rolled back**.
- All or Nothing.

Example:
If you're transferring money between two bank accounts, both the debit and credit operations must occur — or neither.

19.5.9.2 Consistency

- Ensures that a transaction brings the database from **one valid state to another**.
- Any data written to the database must be valid according to **all defined rules**, such as constraints, cascades, and triggers.

Example:
A transaction shouldn't break rules like "a student's age must be > 0".

19.5.9.3 Isolation

- Ensures that **concurrent transactions** do not affect each other.
- Each transaction should execute **independently** as if it's the only one running.

Example:
Two people booking the last seat on a flight at the same time — isolation ensures that only one succeeds.

19.5.9.4 Durability

- Once a transaction is committed, its changes are **permanently saved** to the database.
- Even in case of a system crash, the changes won't be lost.

Example:
If you place an order online and the server crashes right after — your order still exists in the database.

19.6 Use Cases of Relational Databases

19.6.1 Online Transaction Processing (OLTP)

- Supports high-speed, transaction-oriented operations
- Handles concurrent user access efficiently

19.6.2 Data Warehousing and OLAP

- Optimized for analyzing historical data
- Used for business intelligence and insights

19.6.3 IoT Solutions

- Suitable for lightweight, high-speed data collection and processing from edge devices

19.7 Limitations of Relational Databases

- Not ideal for semi-structured or unstructured data
- Schema and data types must match for migration
- Field length limitations can result in data loss if exceeded

20 Introduction to NoSQL

20.1 What is NoSQL?

NoSQL stands for “Not Only SQL” and refers to non-relational database systems that allow flexible schemas for storing and retrieving data. Unlike traditional relational databases, NoSQL databases are built to scale efficiently, offer high performance, and support developers' ease of use.

- "No" in NoSQL emphasizes "Not Only" SQL, not a complete rejection of SQL.
- These databases support various data models and do not rely on the rigid table-based format of relational databases.
- NoSQL allows for schema-less storage, making it ideal for structured, semi-structured, or unstructured data.

20.2 Rise of NoSQL

While NoSQL databases have existed for years, they have gained popularity recently with the rise of:

- Cloud computing
- Big data
- High-volume web and mobile applications

20.3 Types of NoSQL Databases

There are four primary types of NoSQL databases, each serving different data modeling needs:

20.3.1 Key-Value Store

- **Structure**: Data is stored as key-value pairs.

- **Use Cases**:
  - User session storage
  - Real-time recommendations
  - Targeted advertising
  - In-memory data caching
- **Limitations**:
  - Not ideal for querying based on values or complex relationships
- **Examples**:
  - Redis
  - Memcached
  - DynamoDB

20.3.2 Document-Based Databases

- **Structure**: Store data as individual documents (often JSON or BSON).
- **Benefits**:
  - Flexible indexing
  - Powerful ad hoc queries
  - Analytics over document collections
- **Use Cases**:
  - eCommerce platforms
  - Medical records
  - CRM systems
  - Analytics platforms
- **Limitations**:
  - Not ideal for complex search queries or multi-operation transactions
- **Examples**:
  - MongoDB
  - DocumentDB
  - CouchDB
  - Cloudant

20.3.3 Column-Based Databases

- **Structure**: Data is stored in columns instead of rows.
- **Concept**: Related columns are grouped into a *column family* (e.g., customer profile vs. purchase hist
- Fast accessory).
- **Advantages**: due to columnar storage
  - Ideal for write-heavy systems
- **Use Cases**:
  - Time-series data
  - Weather data
  - IoT applications
- **Limitations**:
  - Not ideal for frequent query pattern changes or complex queries
- **Examples**:
  - Cassandra
  - HBase

20.3.4 Graph-Based Databases

- **Structure**: Use nodes (data) and edges (relationships) to store data in a graph model.
- **Advantages**:
  - Great for finding relationships and patterns
  - Useful for visualizing complex networks
- **Use Cases**:
  - Social networks
  - Fraud detection
  - Access management
  - Real-time recommendations
- **Limitations**:
  - Not optimized for large-scale analytics or high-volume transactions
- **Examples**:
  - Neo4J
  - CosmosDB

20.4 Advantages of NoSQL

- **Scalability**: Easily scales across multiple data centers and cloud infrastructure.
- **Flexibility**: Supports unstructured and semi-structured data with schema-less design.
- **Performance**: Optimized for high-throughput workloads.
- **Cost-Effective**: Runs on low-cost, commodity hardware.
- **Agility**: Easier to iterate and innovate with flexible architecture.

20.5 NoSQL vs. Relational Databases

Feature	Relational Databases (RDBMS)	NoSQL Databases
Schema	Rigid, predefined	Flexible, schema-less
Data Types	Structured	Structured, semi-structured, unstructured
ACID Compliance	Full support	Limited or eventual consistency
Cost	Expensive hardware/software	Commodity hardware, cost-efficient
Maturity	Mature and well-documented	Newer, evolving ecosystem

21 Exploring Data Warehouses, Data Marts, Data Lakes, ETL, and Data Pipelines

21.1 Introduction

Earlier in the course, we explored databases, data warehouses, and big data stores. This section dives deeper into data warehouses, data marts, and data lakes, and explains the ETL process and data pipelines.

21.2 Data Warehouses

A data warehouse is a multi-purpose storage system where data is already modeled and structured for analysis. Organizations use data warehouses to handle massive amounts of data from operational systems, making it analysis-ready and readily available for reporting.

21.2.1 Purpose and Functionality

- Acts as the single source of truth
- Stores current and historical data
- Data is cleansed, conformed, and categorized
- Enables operational and performance analytics

21.3 Data Marts

A data mart is a subset of a data warehouse, designed for specific business functions or user communities.

21.3.1 Characteristics

- Tailored for stakeholders with specific data needs
- Used for departmental analytics (e.g., sales, finance)
- Offers isolated security and performance
- Main purpose: business-specific reporting and analytics

21.4 Data Lakes

A data lake is a storage repository that holds vast amounts of raw data in its native format.

21.4.1 Key Features

- Stores structured, semi-structured, and unstructured data
- Each data element is uniquely identified and tagged with metadata
- Retains all source data without exclusions
- Ideal for predictive and advanced analytics

21.4.2 Comparison with Data Warehouses

- **Data Warehouse**: Stores processed, structured data for analysis
- **Data Lake**: Stores raw data for flexible use cases
- Can act as a staging area for data warehouses

21.5 The ETL Process (Extract, Transform, Load)

ETL is the automated process used to convert raw data into analysis-ready data.

21.5.1 Overview

- **Extract**: Gather data from source locations
- **Transform**: Clean and convert data into usable format
- **Load**: Insert the transformed data into a data repository

21.5.2 Extract

Methods of extraction:

- **Batch Processing**: Large chunks of data moved at scheduled intervals (Tools: Stitch, Blendo)
- **Stream Processing**: Real-time data movement and transformation (Tools: Apache Samza, Apache Storm, Apache Kafka)

21.5.3 Transform

Data transformation involves:

- Standardizing formats (e.g., dates, units)
- Removing duplicates
- Filtering unnecessary data
- Enriching data (e.g., splitting full names)
- Establishing relationships across tables
- Applying business rules and validations

21.5.4 Load

Processed data is moved to the destination system via:

- **Initial Loading:** Loading the entire dataset for the first time into the destination system.
- **Incremental Loading:** Loading only the new or changed data since the last load.
- **Full Refresh:** Delete all existing data in the destination and reload everything from scratch

21.5.4.1 Load Verification

- Check for missing/null values

- Monitor server performance
- Handle load failures with proper recovery mechanisms

21.6 Data Pipelines

Data pipelines manage the entire journey of moving data from source to destination, including ETL as a subset.

21.6.1 Features and Architecture

- Support batch, streaming, or hybrid data processing
- Enable continuous transformation for real-time data (e.g., sensor data)
- The destination is typically a data lake or another application/tool

21.6.2 Popular Tools

- Apache Beam
- DataFlow

21.6.3 Use Cases

- Long-running batch queries
- Interactive queries
- Real-time streaming data processing

22 Introduction to Big Data

22.1 What is Big Data

In today's digital world, everyone leaves a data trail. From travel habits to workouts and entertainment, the increasing number of internet-connected devices record vast amounts of data. This phenomenon is known as **Big Data**.

Ernst & Young defines big data as:

"Dynamic, large, and disparate volumes of data created by people, tools, and machines. It requires innovative and scalable technology to collect, host, and analytically process data to drive real-time business insights related to consumers, risk, profit, performance, productivity management, and shareholder value."

While there's no single definition of big data, most definitions share five common elements known as the **5 V's of Big Data**:

22.2 The 5 V's of Big Data

22.2.1 Velocity

- Refers to the speed at which data accumulates.

- Data is generated at an extremely fast pace.
- Real-time and -near-real-time streaming technologies can process data quickly.

22.2.2 Volume

- Refers to the scale or amount of data being stored.
- Driven by:
  - Increased number of data sources
  - High-resolution sensors
  - Scalable storage infrastructure

22.2.3 Variety

- Refers to the diversity of data types:
  - **Structured Data**: Fits neatly into rows and columns (e.g., relational databases)
  - **Unstructured Data**: Tweets, blog posts, images, videos, etc.
- Comes from multiple sources: machines, people, internal and external processes
- Driven by mobile tech, social media, wearables, geo-technologies, and more

22.2.4 Veracity

- Refers to the quality and origin of data, including:
  - Consistency
  - Completeness
  - Integrity
  - Ambiguity
- Challenges include determining if the information is accurate or false

22.2.5 Value

- Refers to the ability to turn data into meaningful insights
- Value can be medical, social, financial, or personal
- The key goal is to **derive actionable insights** from the data

22.3 Real-world examples of the V's

22.3.1 Velocity

- Every 60 seconds, hours of video are uploaded to YouTube.
- This demonstrates the rapid pace at which data is generated.

22.3.2 Volume

- With over 7 billion people worldwide, the majority use digital devices.
- These devices generate ~2.5 quintillion bytes of data every day (equivalent to 10 million Blu-ray DVDs).

22.3.3 Variety

- Examples include:
  - Text
  - Pictures
  - Video
  - Sound
  - Health data from wearables
  - Data from Internet of Things (IoT) devices

22.3.4 Veracity

- 80% of data is unstructured.
- Requires classification, analysis, and visualization for reliable insights.

22.4 Tools and Technologies for Big Data

Traditional data analysis tools are inadequate for handling massive datasets. Instead, distributed computing tools are used, such as:

22.4.1 Big Data Tools

- **Apache Spark**
- **Hadoop and its ecosystem**

These tools help extract, load, analyze, and process data across distributed resources to uncover new insights.

22.5 Impact of Big Data

Big data enables organizations to:

- Better connect with customers
- Offer more personalized services
- Make data-driven decisions

22.6 Final Thoughts

Every interaction with digital devices—whether unlocking a smartphone or tracking a workout—contributes to the world of big data. That data travels far and wide through analysis pipelines and returns as insights that influence our lives and the services we use.

23 Big Data Processing Technologies

23.1 Overview

Big Data processing technologies enable working with large volumes of structured, semi-structured, and unstructured data to extract value. This section discusses three key open-source technologies: **Apache Hadoop**, **Apache Hive**, and **Apache Spark**, and their roles in big data analytics.

23.2 Apache Hadoop

23.2.1 What is Hadoop?

Apache Hadoop is a Java-based open-source framework designed for distributed storage and processing of large datasets across clusters of computers. It offers a reliable, scalable, and cost-effective solution for storing and processing data with no strict format requirements.

23.2.2 Features of Hadoop

- Distributed system: Nodes (individual computers) are organized into clusters.
- Scalable: Can scale from a single node to thousands.
- Format flexibility: Supports structured, semi-structured, and unstructured data.
- Real-time, self-service access.
- Enterprise cost optimization: Moves infrequently used "cold" data to a Hadoop system.

23.2.3 Hadoop Distributed File System (HDFS)

HDFS is a key component of Hadoop, designed for scalable and reliable data storage across multiple hardware units.

23.2.3.1 Key Benefits:

- **File Partitioning:** Splits large files across multiple nodes for parallel access.
- **Fault Tolerance:** Replicates file blocks across nodes to prevent data loss.
- **Parallel Processing:** Computations run on nodes where data resides.
- **Data Locality:** Brings computation closer to data to minimize network congestion.

23.2.3.2 Example:

If a phonebook is stored using Hadoop, names starting with A might be on server 1, B on server 2, and so on. Hadoop stores and replicates these pieces across the cluster to ensure reliability and availability.

23.2.3.3 Additional Benefits:

- Fast recovery from hardware failures.
- High throughput access to streaming data.

- Scales to hundreds of nodes.
- Portable across different hardware and OS platforms.

23.3 Apache Hive

23.3.1 What is Hive?

Hive is an open-source data warehouse software built on top of Hadoop. It facilitates reading, writing, and managing large datasets stored in HDFS or systems like Apache HBase.

23.3.2 Use Cases:

- Best suited for data warehousing tasks: ETL, reporting, and data analysis.
- Supports SQL-based tools for easy data access.

23.3.3 Limitations:

- Not ideal for transaction processing due to high latency.
- Less suitable for applications needing fast response times.

23.4 Apache Spark

23.4.1 What is Spark?

Apache Spark is a general-purpose distributed data processing engine that performs large-scale data analytics in real time.

23.4.2 Key Features:

- **In-Memory Processing:** Increases computation speed; writes to disk only when necessary.
- **Multi-language Support:** Interfaces available for Java, Scala, Python, R, and SQL.
- **Flexible Deployment:** Can run standalone or on top of Hadoop.
- **Diverse Data Source Access:** Supports HDFS, Hive, and more.

23.4.3 Use Cases:

- Interactive Analytics
- Stream Processing
- Machine Learning
- Data Integration
- ETL (Extract, Transform, Load)

23.4.4 Core Benefit:

Spark excels in the fast processing of streaming data and performing real-time complex analytics, making it a powerful tool in modern data environments.

24 Summary and Highlights

A Data Repository is a general term that refers to data that has been collected, organized, and isolated so that it can be used for reporting, analytics, and also for archival purposes.

The different types of Data Repositories include:

- Databases, which can be relational or non-relational, each following a set of organizational principles, the types of data they can store, and the tools that can be used to query, organize, and retrieve data.
- Data Warehouses, that consolidate incoming data into one comprehensive storehouse.
- Data Marts, that are essentially sub-sections of a data warehouse, built to isolate data for a particular business function or use case.
- Data Lakes, serve as storage repositories for large amounts of structured, semi-structured, and unstructured data in their native format.
- Big Data Stores, provide distributed computational and storage infrastructure to store, scale, and process very large data sets.

ETL, or Extract Transform and Load, Process is an automated process that converts raw data into analysis-ready data by:

- Extracting data from source locations.
- Transforming raw data by cleaning, enriching, standardizing, and validating it.
- Loading the processed data into a destination system or data repository.

Data Pipeline, sometimes used interchangeably with ETL, encompasses the entire journey of moving data from the source to a destination data lake or application, using the ETL process.

Big Data refers to the vast amounts of data that is being produced each moment of every day, by people, tools, and machines. The sheer velocity, volume, and variety of data challenge the tools and systems used for conventional data. These challenges led to the emergence of processing tools and platforms designed specifically for Big Data, such as Apache Hadoop, Apache Hive, and Apache Spark.

Module 3

25 Identifying the Right Data for Your Use Case

25.1 Understanding the Current and Desired States

At this stage, you understand where you currently stand and where you aim to go. You also have a clear metric in place:

- **What will be measured**
- **How it will be measured**

The next logical step is to identify the data required for your use case.

25.2 Identifying Required Data

The data identification process begins by outlining the specific information needed and the possible sources. Your objectives will guide these decisions.

25.2.1 Example Use Case: Targeted Marketing by a Product Company

Goal: Create targeted marketing campaigns for the age group that purchases the most.

25.2.2 Required Information

- Customer profile
- Purchase history
- Location
- Age
- Education
- Profession
- Income
- Marital status

To further refine insights:

- **Customer complaints**: Understand recurring issues.
- **Customer service survey ratings**: Assess satisfaction levels.
- **Social media engagement**: Analyze likes, shares, and comments to gauge public sentiment and influence.

25.3 Defining a Data Collection Plan

Once the data is identified, the next step is to create a collection plan.

25.3.1 Key Considerations

- **Timeframe**: Define whether data is needed in real-time or over a fixed period.
  - Real-time: e.g., website visitor data

    - Fixed period: e.g., specific events
- **Volume of Data**: Define the quantity required for a credible analysis.
    - Example: All customers aged 21–30 or 100,000 customers in that range.
- **Dependencies and Risks**:
    - Identify any project dependencies.
    - Outline potential risks and mitigation strategies.

The goal is to establish clarity for effective execution.

25.4 Choosing Data Collection Methods

Determine how to collect the data from the identified sources:

25.4.1 Sources

- Internal systems
- Social media platforms
- Third-party data providers

25.4.2 Factors Influencing Collection Methods

- Type of data
- Timeframe
- Volume

Once finalized, implement your strategy and begin data collection. Be prepared to update the plan as new challenges or insights emerge.

25.5 Ensuring Data Quality, Security, and Privacy

The source, method, and practice of collecting data affect:

- **Quality**
- **Security**
- **Privacy**

These considerations span the entire data analysis lifecycle.

25.5.1 Traits of Quality Data

- Error-free
- Accurate
- Complete
- Relevant
- Accessible

Define:

- Quality traits
- Metrics
- Checkpoints to maintain high-quality analysis

25.5.2 Data Governance

Focus on:

- **Usability**
- **Integrity**
- **Availability**

Be mindful of:

- Regulatory compliance
- Penalties for non-compliance

25.5.3 Data Privacy

Ensure:

- Confidentiality
- Proper licensing
- Compliance with regulations

Establish:

- Validation checks
- Auditable data trails

Failure to maintain trust in data can lead to flawed analysis and legal consequences.

25.6 Conclusion

Identifying the right data is crucial to credible and reliable analysis. A well-structured approach ensures multiple perspectives are considered and that findings are grounded in quality data.

26 Data Sources

26.1 Types of Data Sources

Data sources can be internal or external to an organization and are typically classified as primary, secondary, or third-party sources.

26.1.1 Primary Data

Primary data is information collected directly by you from the original source. It includes:

- Internal sources such as:
  - CRM systems
  - HR systems
  - Workflow applications
- Direct methods such as:
  - Surveys
  - Interviews
  - Discussions
  - Observations
  - Focus groups

26.1.2 Secondary Data

Secondary data is information retrieved from pre-existing sources. These sources include:

- External databases
- Research articles and publications
- Training materials
- Internet searches
- Public financial records
- Data collected through externally conducted:
  - Surveys
  - Interviews
  - Observations
  - Focus groups

26.1.3 Third-Party Data

Third-party data is purchased from aggregators who collect and compile data from various sources. These datasets are sold for business use and decision-making.

26.2 Common Data Sources

There are various sources from which data can be gathered.

26.2.1 Databases

Databases can hold primary, secondary, and third-party data:

- Internal databases: Support organizational processes, workflows, and customer management.
- External databases: Available via subscriptions or purchase.

26.2.2 Cloud Platforms

With the shift to cloud computing, organizations can now access real-time data and insights on-demand.

26.2.3 Web Data

The web provides publicly available data for both individual and commercial use. This includes:

- Textbooks
- Government records
- Research papers
- Articles
- Social media platforms (e.g., Facebook, Twitter, Google, YouTube, Instagram)
    - Used for gathering user opinions and behavior

26.2.4 Sensor Data

Collected from devices such as:

- Wearables
- Smart buildings
- Smart cities
- Smartphones
- Medical devices
- Household appliances

26.2.5 Data Exchange

Involves voluntary sharing of data between:

- Individuals
- Organizations
- Governments

Exchanged data may include:

- Business application data
- Sensor data
- Social media activity
- Location data
- Consumer behavior data

26.2.6 Surveys

Surveys collect data using structured questionnaires distributed to targeted individuals. For example:

- Understanding customer interest in an upgraded product version
- Can be web-based or paper-based

26.2.7 Census Data

Used to gather household and demographic information such as:

- Wealth and income
- Population statistics

26.2.8 Interviews

Used to collect qualitative data, such as:

- Participant opinions
- Personal experiences

Examples:

- Interviewing a customer service executive about work challenges
- Formats include telephone, web-based, or face-to-face

26.2.9 Observation Studies

Involve monitoring participants in specific scenarios. For example:

- Watching users navigate an e-commerce site to analyze ease of use during product discovery and purchase

26.3 Evolving Nature of Data Sources

Data sources today are more dynamic and diverse than ever before. Continuously evolving, they provide more robust and meaningful insights when primary data is supplemented with secondary and third-party data sources.

27 Data Gathering and Importing Methods

27.1 Overview

There are various methods and tools used to gather data from different sources such as databases, the web, sensors, data exchanges, and more. We will learn about importing this data into various types of data repositories.

27.2 Data Gathering Techniques

27.2.1 SQL for Relational Databases

SQL (Structured Query Language) is used to extract data from relational databases. Key features include:

- Specifying data to retrieve
- Selecting tables
- Grouping records
- Ordering results
- Limiting returned results

27.2.2 Querying Non-Relational Databases

Non-relational databases may use SQL or specific tools such as:

- **CQL** (Cassandra Query Language)
- **GraphQL** (Neo4J)

27.2.3 Using APIs

APIs (Application Programming Interfaces) are widely used for data extraction and validation. APIs access endpoints such as:

- Databases
- Web services
- Data marketplaces

Use Case: Validating postal addresses and zip codes.

27.2.4 Web Scraping

Web scraping (screen scraping or web harvesting) extracts defined data from web pages such as:

- Text
- Contact info
- Images
- Videos
- Podcasts
- Product details

27.2.5 RSS Feeds

RSS feeds capture frequently updated data from sources like forums and news websites.

27.2.6 Data Streams

Data streams aggregate real-time data from sources such as:

- IoT devices
- Instruments
- GPS trackers
- Social media platforms

27.2.7 Data Exchange Platforms

These platforms facilitate secure data sharing between providers and consumers. They offer:

- Exchange standards and protocols
- Security and governance
- Licensing workflows
- Data protection
- Legal frameworks
- Quarantined analytics environments

Popular Platforms:

- AWS Data Exchange
- Crunchbase
- Lotame
- Snowflake

27.2.8 Other Data Sources

Additional data sources include:

- **Marketing Trends:** Forrester, Business Insider
- **Strategic Insights:** Gartner, Forrester
- **User Behavior & Demographics:** Trusted market research firms

27.3 Importing Data into Repositories

27.3.1 Purpose of Importing

After data is gathered, it must be imported into repositories for analysis. The importing process:

- Combines data from different sources
- Provides a unified view
- Allows querying and manipulation

27.3.2 Choosing the Right Repository

The destination repository depends on the type and volume of data:

27.3.2.1 Relational Databases

- Store structured data with a defined schema
- Suitable for data from OLTP systems, spreadsheets, forms, sensors, and logs

27.3.2.2 NoSQL Databases

- Support structured and semi-structured data
- Store data from forms, XML, JSON, and zipped files

27.3.2.3 Semi-Structured Data

- Includes data with some structure but not a fixed schema
- Examples: Emails, XML, binary files, TCP/IP protocols
- XML and JSON are commonly used formats

27.3.2.4 Unstructured Data

- Lacks predefined structure or schema
- Includes social media, images, videos, documents, surveys
- Stored in NoSQL databases and Data Lakes

27.3.2.5 Data Lakes

- Accommodate all data types and schema
- Ideal for large-scale storage and analysis

27.4 Tools and Languages for Importing Data

27.4.1 ETL Tools

- **Talend**
- **Informatica**

27.4.2 Programming Languages

- **Python** and its libraries
- **R** and its libraries

These tools and languages automate the importing and transformation of data into the desired repository for analysis.

28 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 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.

29 Data Wrangling

29.1 Overview

Data wrangling, also known as data munging, is an iterative process that involves data exploration, transformation, validation, and publishing to enable meaningful analysis. It focuses on preparing raw data—often sourced from multiple repositories—for a clearly defined purpose.

29.2 The Four-Step Process of Data Wrangling

Data wrangling typically includes four main phases:

- Discovery
- Transformation
- Validation
- Publishing

29.3 Discovery (Exploration Phase)

The Discovery phase involves understanding the data in the context of its intended use. The goal is to determine the best ways to clean, structure, organize, and map the data for analysis.

29.4 Transformation

The Transformation phase is the core of the data-wrangling process. It consists of several tasks that change the form, format, and content of the data.

29.4.1 Structuring

Structuring involves altering the schema and format of the data to facilitate integration from multiple sources. Data may arrive from different formats such as relational databases and Web APIs.

29.4.1.1 Common Structural Transformations:

- **Joins**: Combine columns from two tables into the same row.
- **Unions**: Combine rows from two tables into one.

29.4.2 Normalization and Denormalization

- **Normalization**: Cleans databases by removing redundancy and inconsistency. Common in transactional systems.
- **Denormalization**: Merges data from multiple tables into a single table to allow faster querying and analysis.

29.4.3 Cleaning

Cleaning aims to fix irregularities and inaccuracies in the data to ensure reliable analysis. Issues addressed include:

- Missing or incomplete data
- Biased or null values
- Outliers

Example: If demographic information is missing for a product sale dataset (e.g., missing gender), you may need to:

- Source and merge missing data
- Or remove records missing this information

29.4.4 Enriching

Enriching involves supplementing the existing dataset with additional data points to add depth to the analysis.

Examples:

- Adding business performance data to customer purchase records
- Computing sentiment scores from customer feedback
- Integrating weather data to analyze resort occupancy trends
- Capturing metadata such as publication time and tags for blog analysis

29.5 Validation

After transformation, the next step is to validate the data to ensure quality. Validation checks the consistency, accuracy, and security of the data using predefined rules.

29.6 Publishing

Publishing involves delivering the final, validated dataset along with its metadata for downstream project needs.

29.7 Importance of Documentation

Since data wrangling is iterative, it's critical to document each step and decision. This documentation supports reproducibility, transparency, and revision of the wrangling process.

30 Popular Data Wrangling Tools

30.1 Overview

These tools help clean, transform, and prepare data for analysis. Examples include:

- Excel Power Query / Spreadsheets
- OpenRefine
- Google DataPrep
- Watson Studio Refinery
- Trifacta Wrangler
- Python
- R

30.2 Spreadsheets

30.2.1 Microsoft Excel and Google Sheets

Spreadsheets are commonly used for manual data wrangling. They include features and built-in formulae that help:

- Identify issues
- Clean data
- Transform data

30.2.2 Add-ins

- **Microsoft Power Query** (for Excel): Allows importing and transforming data from various sources.
- **Google Sheets Query Function**: Offers similar capabilities within Google Sheets.

30.3 OpenRefine

OpenRefine is an open-source tool with the following features:

- Supports multiple formats: TSV, CSV, XLS, XML, and JSON
- Clean and transform data
- Extend data with web services

- Offers an easy-to-use menu-based interface

30.4 Google DataPrep

Google DataPrep is an intelligent, cloud-based data-wrangling tool that:

- Works with structured and unstructured data
- Provides visual data exploration
- Automatically detects schemas, data types, and anomalies
- Offers suggestions for the next best steps
- Is fully managed, requiring no installation or setup

30.5 Watson Studio Refinery

Part of IBM Watson Studio, this tool helps:

- Discover, cleanse, and transform large datasets
- Automatically detect data types and apply governance policies
- Explore data from multiple sources

30.6 Trifacta Wrangler

Trifacta Wrangler is a cloud-based, interactive tool that:

- Cleans and transforms messy, real-world data
- Exports cleaned data to Excel, Tableau, or R
- Supports real-time collaboration among team members

30.7 Python

Python offers a wide range of libraries and tools for data manipulation:

30.7.1 Jupyter Notebook

- Open-source web application
- Used for data cleaning, transformation, visualization, and modeling

30.7.2 NumPy (Numerical Python)

- Provides support for multi-dimensional arrays and matrices
- Offers high-level mathematical operations

30.7.3 Pandas

- Designed for fast and efficient data analysis
- Simplifies complex operations like merging, joining, and transforming data
- Reduces common alignment errors from data coming from multiple sources

30.8 R

R provides several packages specifically designed for data wrangling:

30.8.1 Dplyr

- Simple and precise syntax
- Powerful for manipulating data

30.8.2 Data.table

- Efficient in aggregating large datasets

30.8.3 Jsonlite

- Robust for parsing JSON data
- Useful for working with web APIs

30.9 Choosing the Right Tool

The best tool for data wrangling depends on your specific needs:

- Supported data size and structures
- Cleaning and transformation features
- Infrastructure requirements
- Ease of use and learning curve

31 Data Cleaning and Its Importance

31.1 Importance of Data Quality

According to a Gartner report, poor data quality undermines an organization's competitive standing and weakens business objectives. Missing, inconsistent, or incorrect data can lead to false conclusions and ineffective decisions—often at a significant cost.

31.2 Common Data Issues

Data collected from multiple sources may suffer from the following issues:

- Missing values
- Inaccuracies
- Duplicates
- Incorrect or missing delimiters
- Inconsistent records
- Insufficient parameters

31.2.1 Handling Faulty Data

- Some data can be corrected manually or with data-wrangling tools and scripts.
- If data cannot be repaired, it must be removed from the dataset.

31.3 Data Cleaning vs. Data Wrangling

Although the terms are sometimes used interchangeably, it's important to note:

- **Data Cleaning** is a subset of the **Data Wrangling** process.
- It plays a key role in the **Transformation phase** of the data-wrangling workflow.

31.4 Data Cleaning Workflow

A typical data-cleaning workflow includes:

1. **Inspection**
2. **Cleaning**
3. **Verification**

31.4.1 1. Inspection

This phase identifies issues in the dataset.

- Use tools and scripts to define and validate rules and constraints.
- Use **data profiling** to understand data structure, content, and relationships.
    - Detect anomalies like null values, duplicates, and out-of-range values.
- Use **data visualization** to detect outliers (e.g., plotting average income).

31.4.2 2. Cleaning

Techniques depend on the use case and data issues.

31.4.2.1 Missing Values

- **Filter out** records with missing values.
- **Source missing information** is critical to the use case.
- **Imputation**: Replace missing values with statistically derived values.

31.4.2.2 Duplicate Data

- Remove repeated data points.

31.4.2.3 Irrelevant Data

- Exclude data not relevant to the context of analysis (e.g., contact numbers in a health study).

31.4.2.4 Data Type Conversion

- Ensure field values match expected data types (e.g., numbers as numeric types, and dates as date types).

31.4.2.5 Standardization

- Convert all strings to lowercase.
- Standardize data formats and units of measurement.

31.4.2.6 Syntax Errors

- Remove leading/trailing white spaces.
- Fix typos and inconsistent formatting (e.g., "New York" vs. "NY").

31.4.2.7 Outliers

- Outliers may or may not be incorrect.
- Example of incorrect: Age 5 in a voter database.
- Example of a valid outlier: Annual income of $1 million in a group with typical incomes between $100,000–$200,000.
- Decide on inclusion based on the potential to skew results.

31.4.3 3. Verification

- Inspect results to ensure cleaning has been effective.
- Re-validate the dataset against rules and constraints.

31.5 Documentation and Reporting

- Document all changes made during cleaning.
- Include the rationale behind each change.
- Report on the quality of the cleaned data.
- Reporting data health is a critical part of the cleaning process.

32 Expert Viewpoints

Data professionals discuss the importance of gathering, cleaning, and preparing data for analysis. Key points include:

- A significant portion of their job involves understanding the source and quality of data, as no dataset is perfect.
- Data professionals emphasize the need to ensure data reliability by running summary statistics and performing logic checks to identify inconsistencies or errors.
- For financial analysis, the data must be accurate, unbiased, and free from errors, allowing for meaningful insights and conclusions.

33 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 problems such as missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors, and outliers.
- Enriching data 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.

Module 4

34 Understanding Statistical Analysis

34.1 What is Statistics?

Statistics is a branch of mathematics that involves the collection, analysis, interpretation, and presentation of numerical or quantitative data. It is used extensively in various aspects of life, such as calculating average income, analyzing age distributions, and identifying top-paying professions.

34.1.1 Applications of Statistics

Statistics is used across industries for data-driven decision-making. Examples include:

- Researchers using statistical methods to ensure vaccine safety and efficacy
- Companies analyzing customer behavior to reduce churn

34.2 What is Statistical Analysis?

Statistical Analysis is the application of statistical methods to a sample of data to understand what the data represents. It involves collecting and scrutinizing every data sample from a population.

34.2.1 Key Concepts

- **Sample**: A representative subset drawn from a total population
- **Population**: A group of people or items sharing at least one common characteristic

Example:

- **Population**: All licensed drivers in a state
- **Sample**: Male drivers over the age of 50

Statistical methods help ensure accurate data interpretation and validate relationships within the data.

34.3 Types of Statistical Analysis

There are two main types of statistical analysis:

34.3.1 Descriptive Statistics

Descriptive Statistics summarize information about a sample without concluding the larger population.

34.3.1.1 Features of Descriptive Statistics

- Use of summary charts, tables, and graphs
- Simplifies raw data interpretation

34.3.1.2 Common Measures

34.3.1.2.1 Central Tendency

Measures that identify the center of a dataset:

- **Mean**: The average value
- **Median**: The middle value in an ordered dataset
- **Mode**: The most frequently occurring value

Example:

- A class of 25 students has test scores. The mean is calculated by dividing the total of all scores by 25. The median is the score at position 13 when scores are ordered. The mode is the score that appears most often.

34.3.1.2.2 Dispersion

Measures the spread or variability in a dataset:

- **Variance**: Indicates the spread from the mean
- **Standard Deviation**: Measures clustering around the mean

- **Range**: Difference between the highest and lowest values

Understanding dispersion helps predict the likelihood of certain events.

34.3.1.2.3 Skewness

Measures the asymmetry of data distribution:

- Determines if data is symmetrically distributed or skewed
- Affects the validity of statistical analyses

34.3.1.2.4 Other Tools

- **Correlation and Scatterplots**: Used to assess relationships between paired data

34.3.2 Inferential Statistics

Inferential Statistics draw conclusions about a population based on sample data.

34.3.2.1 Key Methods

34.3.2.1.1 Hypothesis Testing

- Determines the effectiveness of treatments, such as vaccines, by comparing control and treatment groups

34.3.2.1.2 Confidence Intervals

- Provides a range of values within which the true population value is likely to fall

34.3.2.1.3 Regression Analysis

- Uses hypothesis testing to confirm relationships found in sample data also exist in the population

34.4 Statistical Software Tools

Several tools are used for statistical analysis:

- **SAS** (Statistical Analysis System)
- **SPSS** (Statistical Package for the Social Sciences)
- **Stat Soft**

34.5 Statistics and Data Mining

Statistics is integral to data mining:

- Provides methods and measures for data mining
- Helps distinguish between random noise and meaningful patterns

Both statistics and data mining are essential techniques in data analysis, enabling better decision-making through insights derived from data.

35 Data Mining

35.1 Introduction to Data Mining

Data mining, or the process of extracting knowledge from data, is central to the data analysis process. It is an interdisciplinary field involving:

- Pattern recognition technologies
- Statistical analysis
- Mathematical techniques

35.1.1 Objectives of Data Mining

- Identify correlations and variations in data
- Find patterns and understand trends
- Predict probabilities and future outcomes

35.2 Understanding Patterns and Trends

35.2.1 Pattern Recognition

Pattern recognition involves discovering regularities or commonalities in data. For instance:

- Analyzing login data in an organization can reveal user behaviors such as:
  - Peak login times
  - Roles with the longest session durations
  - Frequently used modules in an application

This involves manual or tool-based analysis to uncover hidden patterns.

35.2.2 Trends

A trend is a general direction in which data tends to move over time. Example:

- Global warming: While annual temperatures may fluctuate slightly, the long-term data shows a consistent increase in global temperature.

35.3 Applications of Data Mining

Data mining is applied across a wide range of industries:

- **Marketing**: Profiling customer behaviors and income to run targeted campaigns
- **Finance**: Monitoring transactions to detect fraudulent activity
- **Healthcare**: Predicting patient risks for specific conditions
- **Education**: Analyzing student performance to provide targeted support

- **Law Enforcement**: Forecasting crime-prone areas for resource deployment
- **Logistics**: Aligning supply chains with demand forecasts

35.4 Techniques in Data Mining

Several techniques are used to detect patterns and build discovery models. These include descriptive, diagnostic, predictive, and prescriptive models:

35.4.1 Classification

- Categories attribute into target groups
- Example: Classifying customers as low, medium, or high spenders based on income

35.4.2 Clustering

- Groups data into clusters for collective analysis
- Example: Grouping customers based on geographic location

35.4.3 Anomaly Detection (Outlier Detection)

- Identifies data points that deviate from the norm
- Example: Detecting unusual credit card activity that may indicate misuse

35.4.4 Association Rule Mining

- Establishes relationships between data events
- Example: Customers who buy laptops often also buy cooling pads

35.4.5 Sequential Patterns

- Traces sequences of events
- Example: Mapping a customer's journey from login to checkout on an e-commerce site

35.4.6 Affinity Grouping

- Discovers co-occurrence relationships
- Example: Recommending products based on the purchase history of similar customers (used for cross-selling and up-selling)

35.4.7 Decision Trees

- Builds classification models in a tree-like structure
- Each branch represents a possible outcome based on input conditions
- Helps visualize the relationship between input and output variables

35.4.8 Regression

- Determines relationships between variables (causal or correlational)
- Example: Predicting house prices based on location and area

35.5 Conclusion

Data mining helps filter out noise from relevant information, allowing organizations to focus efforts and resources more effectively on meaningful data insights.

36 Commonly Used Software and Tools for Data Mining

36.1 Spreadsheets

Spreadsheets like **Microsoft Excel** and **Google Sheets** are commonly used for basic data mining tasks. They are user-friendly and accessible, making them ideal for handling exported data from other systems.

36.1.1 Features

- Host data in an easy-to-read format
- Pivot tables for showcasing specific data aspects
- Easy comparison between different datasets

36.1.2 Add-ins and Tools

- **Microsoft Excel Add-ins**:
  - Data Mining Client for Excel
  - XLMiner
  - KnowledgeMiner
- **Google Sheets Add-ons**:
  - Text Analysis
  - Text Mining
  - Google Analytics

36.2 R Language

R is a leading language for statistical modeling and data mining, widely used by statisticians and data miners.

36.2.1 Capabilities

- Regression
- Classification
- Clustering
- Association Rule Mining
- Text Mining
- Outlier Detection

- Social Network Analysis

36.2.2 Popular R Packages

- **tm**: Text mining framework
- **twitteR**: Mining tweets from Twitter

36.2.3 R Studio

- Open-source IDE for R
- Popular among data mining professionals

36.3 Python

Python is one of the most popular programming languages for data mining due to its simplicity and powerful libraries.

36.3.1 Libraries

- **Pandas**:
  - Handles various data formats
  - Supports data organization, sorting, and manipulation
  - Performs computations like mean, median, mode, range
  - Answers statistical questions
  - Visualizes data with other libraries
- **NumPy**:
  - Mathematical computing and data preparation
  - Built-in functions for data mining

36.3.2 Jupyter Notebooks

- Preferred tool for data scientists using Python
- Ideal for performing data mining and statistical analysis

36.4 IBM SPSS Statistics

SPSS stands for Statistical Package for Social Sciences. Though originally for social science research, it's now widely used in business analytics.

36.4.1 Features

- Advanced analytics and trend analysis
- Text analytics
- Assumption validation
- Business problem translation into data science solutions
- Minimal coding interface
- Efficient data management tools
- Reliable and accurate results

36.5 IBM Watson Studio

Watson Studio is a comprehensive platform included in IBM Cloud Pak for Data.

36.5.1 Key Features

- Leverages open-source tools like Jupyter Notebooks
- Enhances with IBM’s proprietary tools
- Accessible via web browser on cloud or desktop
- A collaborative environment for teams
- Supports tasks from data exploration to AI model building
- Includes SPSS Modeller for predictive modeling

36.6 SAS Enterprise Miner

SAS is a graphical workbench for powerful data mining.

36.6.1 Capabilities

- Interactive data exploration
- Relationship identification within data
- Manages and analyzes data from various sources
- GUI for non-technical users

36.6.2 Features

- Modeling techniques for pattern recognition
- Anomaly detection
- Big data analysis
- Reliability validation
- Easy-to-learn syntax
- Handles large databases
- Offers high security

36.7 Summary

The choice of tool depends on factors such as:

- Data size and structure
- Tool features and visualization capabilities
- Infrastructure requirements
- Ease of use and learnability

37 Summary and Highlights

37.1 Introduction to Statistics

Statistics is a branch of mathematics that focuses on the collection, analysis, interpretation, and presentation of numerical or quantitative data.

37.2 Statistical Analysis

Statistical Analysis involves using statistical methods to understand and interpret data.

37.2.1 Types of Statistical Analysis

37.2.1.1 Descriptive Statistics

Descriptive analysis summarizes what the data represents. Common measures include:

- Central Tendency
- Dispersion
- Skewness

37.2.1.2 Inferential Statistics

Inferential analysis makes generalizations or inferences about data. Common measures include:

- Hypothesis Testing
- Confidence Intervals
- Regression Analysis

37.3 Data Mining

Data Mining is the process of extracting knowledge from data using various techniques.

37.3.1 Techniques in Data Mining

- Classifying attributes of data
- Clustering data into groups
- Establishing relationships between:
    - Events
    - Variables
    - Inputs and outputs

37.4 Tools for Data Analysis and Mining

Various software and tools are used for data analysis and mining. Popular ones include:

- Spreadsheets
- R-Language
- Python
- IBM SPSS Statistics
- IBM Watson Studio
- SAS

Each of these tools has its own characteristics, strengths, limitations, and specific applications.

38 Communicating Data Insights Effectively

38.1 Understanding the Problem and Outcome

The data analysis process begins with clearly understanding the problem to be solved and the desired outcome to be achieved. It concludes with effectively communicating the findings in a way that influences decision-making.

38.2 Collaborative Nature of Data Projects

Data projects typically involve collaboration across various business functions. They require multidisciplinary skills, and the resulting insights are often incorporated into broader business initiatives.

38.3 Importance of Effective Communication

The success of communication depends on how well others understand and trust your insights. As data analysts, it is crucial to tell a compelling story with data by:

- Visualizing insights clearly
- Structuring narratives
- Targeting messages specifically for the audience

38.4 Understanding Your Audience

Before communicating your insights, reconnect with your audience by asking:

- Who is my audience?
- What is important to them?
- What will help them trust me?

38.4.1 Audience Diversity

Your audience may include individuals from various business functions, roles, and levels of impact. Consider:

- Their current knowledge of the problem

- Their operational or strategic role
- How they are affected by the problem

38.4.2 Tailoring the Presentation

Present only the essential information needed to address the business problem. Avoid data overload—facts and figures alone are not persuasive. Instead, craft a focused narrative that supports understanding and decision-making.

38.5 Demonstrating Understanding of the Business Problem

Start your presentation by showing your understanding of:

- The problem to be solved
- The outcome to be achieved

Reflecting this understanding builds trust and secures your audience's attention from the beginning.

38.5.1 Speak the Language of the Business

Use terminology and concepts familiar to your organization's business domain to establish a connection with your audience.

38.6 Structuring the Communication

Organize your content for clarity and impact. Reference the data you've collected, and aim to establish its credibility:

38.6.1 Establishing Credibility

- Share your data sources
- State your hypotheses
- Explain validations
- Clarify assumptions

38.6.2 Organizing Information

Group information into logical categories. For instance:

- Qualitative vs. Quantitative data
- Use top-down or bottom-up narrative structures based on your audience and use case
- Stay consistent throughout your communication

38.7 Choosing the Right Communication Format

Determine the most appropriate formats based on your audience's needs. Consider whether they require:

- Executive summaries
- Fact sheets
- Detailed reports

38.8 Inspiring Action Through Insights

Insights must be communicated in a way that inspires action. If your audience doesn’t grasp or value your findings, the insight becomes ineffective.

38.9 Role of Data Visualization

A strong visualization creates clear mental images and tells a story through graphical representation. Use visual tools to show:

- Comparisons
- Relationships
- Distributions
- Compositions

38.9.1 Visualization Tools

Graphs, charts, and diagrams help reveal patterns and support hypotheses.

38.10 Building Trust and Relatability

To drive value from data:

- Build credibility
- Create a structured narrative
- Use visuals to support your story

This approach enables your audience to trust, understand, and relate to your insights, empowering them to take meaningful action.

39 The Role of Storytelling in the Life of a Data Analyst

39.1 Importance of Storytelling in Data Analysis

Storytelling plays a crucial role in the daily work of data analysts. It is not just about analyzing numbers, but about communicating insights effectively.

- Humans naturally understand the world through stories.
- To influence decisions with data, telling a clear, concise, and compelling story is essential.

39.2 Storytelling Enhances Understanding

Creating a narrative around data helps analysts themselves better understand the dataset:

- A story can uncover the underlying patterns and behavior within data.
- It supports a deeper analysis by providing context and structure.

39.3 Balancing Simplicity and Complexity

Finding the right balance in storytelling is a critical skill:

- It’s important to present a clear and coherent story.
- At the same time, one must ensure complex insights are not oversimplified.

39.4 Communicating with the Audience

No matter how insightful the analysis is, it needs to be communicated effectively:

- If the audience doesn’t understand the message, the insights lose their value.
- Communicating through visuals and storytelling helps convey the usefulness of information to different audiences, from consumers to executives.

39.5 Storytelling as a Critical Skill

Storytelling is considered the “last mile” in delivering data value:

- Technical skills can be learned relatively quickly.
- The ability to extract value from data and communicate it effectively remains rare and highly valuable.

39.6 Emotional Connection and Action

A compelling story creates emotional resonance and inspires action:

- Simply presenting data is not enough; it must be wrapped in a meaningful narrative.
- At Stanford, a study showed that audiences remembered stories more than standalone statistics.
- Embedding facts and figures within a story drives the message home.

39.7 Conclusion

Storytelling is an essential part of data and analytics. It enables analysts to:

- Deliver insights effectively.
- Foster understanding and emotional connection.
- Influence decisions and prompt actions.

Without storytelling, even the most valuable data can fail to make an impact.

40 Data Visualization

40.1 What is Data Visualization?

Data visualization is the discipline of communicating information through visual elements such as graphs, charts, and maps. The primary goal is to make information easy to:

- Comprehend
- Interpret
- Retain

Rather than analyzing thousands of rows of raw data, visualization summarizes relationships, trends, and patterns in an understandable format.

40.2 Importance of Choosing the Right Visualization

To effectively deliver your message, it's crucial to select the appropriate visualization. Begin by asking:

- What relationship am I trying to establish?
- Do I want to compare proportions within a whole?
- Am I comparing multiple values over time?
- Do I want to analyze a single value across different timeframes?
- Is the correlation between two variables important?
- Am I detecting anomalies that could affect conclusions?

Always consider:

- The question you want to answer
- Whether your audience needs a static or interactive visualization
- What the key takeaway for your audience should be

Interactive visualizations, for instance, can allow users to modify inputs and observe real-time effects on related variables.

40.3 Types of Charts for Data Visualization

40.3.1 Bar Charts

- Useful for comparing related data sets or components of a whole
- Example: Comparing population numbers of 10 different countries

40.3.2 Column Charts

- Compare values side-by-side to show change over time
- More suitable for displaying both negative and positive values
- Example: Monthly changes in website page views and session durations

40.3.3 Pie Charts

- Show proportional breakdowns of categories that sum to 100%
- Example: Lead generation per marketing channel in a campaign

40.3.4 Line Charts

- Display trends over a continuous variable such as time
- Useful for identifying patterns and comparing multiple series
- Example: Tracking sales changes of one or more products over time

40.4 Dashboards

Dashboards combine multiple data visualizations and reports into a single interface. They help monitor:

- Daily operations
- The health of a business function or process
- Real-time campaign performance

40.4.1 Benefits of Dashboards

- Present a comprehensive overview while enabling detailed drill-downs
- Easy for average users to understand
- Facilitate team collaboration
- Allow on-the-go report generation and evaluation from multiple perspectives

40.4.2 Example Use Case

A marketing dashboard might display:

- Current campaign performance (reach-outs, queries, conversions)
- Comparison with past successful campaigns

Dashboards help you respond quickly to changing data without restarting the analysis process.

41 Data Visualization Tools and Software

41.1 Overview

In this section, we explore some of the most commonly used data visualization software and tools. These range from free and open-source tools to comprehensive commercial analytics solutions:

- Spreadsheets (Microsoft Excel, Google Sheets)
- Jupyter Notebook and Python libraries

- RStudio and Shiny
- IBM Cognos Analytics
- Tableau
- Microsoft Power BI

41.2 Spreadsheets

41.2.1 Microsoft Excel

- Offers a variety of charts: bar, line, pie, pivot, scatter, trend lines, Gantt, Waterfall, and combination charts.
- Provides recommended chart types based on your dataset.
- Allows customization: add chart titles, change element colors, and apply labels.

41.2.2 Google Sheets

- Similar visualization capabilities as Excel.
- Excel has more inbuilt formula-based options.
- Charts auto-update with data changes.
- Preferred for collaboration due to its cloud-based nature.

41.3 Jupyter Notebook and Python Libraries

41.3.1 Jupyter Notebook

- Open-source web application for data exploration and visualization.
- Suitable for beginners and advanced users alike.

41.3.2 Python Visualization Libraries

41.3.2.1 Matplotlib

- Widely used for creating 2D and 3D plots.
- Offers flexibility and high-quality visuals with few lines of code.
- Supported across platforms and has a strong community.

41.3.2.2 Bokeh

- Creates interactive charts and plots.
- Optimized for large or streaming datasets.
- Integrates with other libraries like Matplotlib, Seaborn, and Ggplot.

41.3.2.3 Dash

- Framework for building web-based interactive visualizations using Python.
- Mobile-ready and cross-platform.
- Does not require HTML or JavaScript knowledge.

41.4 RStudio and Shiny

41.4.1 RStudio

- Used to create both basic (histograms, bar, line, scatter plots) and advanced (heat maps, mosaic maps, 3D graphs, correlograms) visualizations.

41.4.2 Shiny

- R package for building interactive web apps.
- Supports embedding R objects such as plots and tables.
- Enables dashboard creation and easy sharing of live apps.

41.5 IBM Cognos Analytics

- End-to-end analytics solution.
- Features include:
  - Custom visual imports
  - Forecasting with time series modeling
  - Visualization recommendations
  - Conditional formatting for highlighting exceptional data
  - Geospatial data overlay capabilities

41.6 Tableau

- Used for creating dashboards and interactive visualizations via drag-and-drop.
- Supports publishing results as visual stories.
- Allows importing R and Python scripts.
- Compatible with:
  - Excel and text files
  - Relational and cloud databases (e.g., Google Analytics, Amazon Redshift)
- Known for its intuitive user interface and superior visualization capabilities.

41.7 Microsoft Power BI

- Cloud-based analytics service by Microsoft.
- Allows report and dashboard creation using drag-and-drop tools.
- Compatible with Excel, SQL Server, and various cloud data repositories.
- Supports collaboration and secure sharing of interactive reports.
- Dashboards consist of interactive visual elements (tiles) that respond to user input.

41.8 Choosing the Right Tool

When selecting a data visualization tool, consider:

- The purpose of the visualization
- Ease of use
- Available features and compatibility with data sources

If you can visualize it, you can create it.

42 Summary and Highlights

42.1 Importance of Data Storytelling

Data has value through the stories that it tells. To communicate findings effectively, it is essential to:

- Ensure the audience can trust, understand, and relate to the insights.
- Establish the credibility of the findings.
- Present data within a structured narrative.
- Support communication with strong visualizations for clarity and actionability.

42.2 What is Data Visualization?

Data visualization is the discipline of communicating information using visual elements such as graphs, charts, and maps. The main goal is to make information easy to:

- Comprehend
- Interpret
- Retain

42.3 Creating Valuable Visualizations

To ensure that data visualization is impactful:

- Focus on the key takeaway for the audience.
- Anticipate their information needs and questions.
- Plan visualizations that deliver the message clearly and effectively.

42.4 Types of Graphs and Charts

There are various types of graphs and charts that can be used depending on the data:

- Bar Charts
- Column Charts
- Pie Charts
- Line Charts

42.5 Dashboards

Dashboards are tools that organize and display visualizations and reports from multiple data sources in one interface. They:

- Are easy to understand
- Enable report generation on the go

42.6 Tools for Data Visualization

When choosing tools for data visualization, consider their ease of use and intended purpose. Popular tools include:

- Spreadsheets
- Jupyter Notebook
- Python Libraries
- R-Studio and R-Shiny
- IBM Cognos Analytics
- Tableau
- Power BI

Module 5

43 Data Analyst Career Landscape

43.1 Data Analyst Job Market

Data analyst job openings are available across industries, government, and academia. Industries such as banking and finance, insurance, healthcare, retail, and information technology highly demand skilled data analysts. Both large corporations and startups actively seek these professionals.

According to **Forbes**, the global big data analytics market was valued at **$37.34 billion in 2018** and is projected to grow at a **CAGR of 12.3% from 2019 to 2027**, reaching **$105.08 billion by 2027**. Currently, the **demand for skilled data analysts exceeds supply**, leading to competitive salaries and increased job opportunities.

43.2 Career Paths in Data Analysis

43.2.1 Broad Classifications

There are three broad categories of roles for data analysts:

- **Data Analyst Specialist Roles**
- **Domain Specialist Roles**
- **Analytics-Enabled Roles**

43.3 Data Analyst Specialist Roles

These roles are ideal for professionals who want to specialize in the technical and functional areas of data analysis.

43.3.1 Career Progression

You may start as an:

- Associate/Junior Data Analyst
- Analyst
- Senior Analyst
- Lead Analyst
- Principal Analyst

43.3.2 Work Environment Variability

- In **small teams**, analysts may handle the entire data analysis process end-to-end.
- In **larger organizations**, roles are often segmented by activity phase, allowing focused skill development in specific areas.

43.3.3 Skill Development

As you grow in your role:

- Advance **technical, statistical, and analytical** skills from basic to expert level
- Master a range of **tools, languages, data repositories, and visualization platforms**
- Improve **communication, stakeholder management, presentation, and project management** skills

43.3.4 Leadership Responsibilities

At senior levels, such as lead or principal analyst, you may:

- Define processes
- Recommend tools and software
- Lead team upskilling
- Participate in hiring and team expansion

Some organizations may delegate these responsibilities to **managers** who have advanced from analyst roles.

43.4 Domain Specialist Roles

Also known as **functional analysts**, these professionals focus on specific industries or domains.

43.4.1 Examples of Domain Specialists

- HR Analyst
- Marketing Analyst
- Sales Analyst
- Healthcare Analyst
- Social Media Analyst

They may not be deeply technical but are recognized authorities in their respective domains.

43.5 Analytics-Enabled Job Roles

These roles involve the use of analytics to enhance decision-making and operational efficiency.

43.5.1 Examples

- Project Managers
- Marketing Managers
- HR Managers

Many job openings in analytics today fall into this category as organizations increasingly rely on data-driven decisions.

43.6 Career Mobility and Growth

As a data analyst, you can explore related career paths by gaining new skills:

43.6.1 Transitioning into Other Data Professions

- **Big Data Engineer**: For those interested in big data and data lakes
- **Business Analytics or BI Analyst**: For those interested in the business side of data

43.7 Continuous Learning and Opportunity

The data analyst field is vast and full of opportunities. With abundant resources available, success depends on:

- Taking initiative
- Seizing opportunities
- Continuously learning and growing in the profession

44 What Employers Look for in a Data Analyst

44.1 Integrity

Employers highly value integrity in Data Analysts. A common interview question is:

"If you had to choose just one, would you rather meet a deadline or get the right answer?"

The ideal response is prioritizing accuracy over deadlines. This is because wrong data can lead to poor multi-million dollar decisions or job losses due to incorrect reporting.

44.2 Communication Skills

The ability to communicate insights is considered one of the most important skills. Even the best analysis is ineffective if it can't be communicated effectively to stakeholders.

44.3 Analytical and Technical Skills

44.3.1 Numerical Fluency and Analytical Thinking

- Strong understanding of complex analysis
- Understanding of AB testing and interpreting results

44.3.2 Technical Skills

- Strong SQL skills
- Programming knowledge in:
    - Python
    - R
    - SQL

44.4 Growth Mindset and Adaptability

Employers look for individuals who are:

- Willing to learn
- Able to keep up with a rapidly changing industry

44.5 Personal Attributes

44.5.1 Detail-Oriented and Ambitious

- Go beyond basic expectations
- Provide alternatives and solutions proactively

44.5.2 Problem-Solving and Troubleshooting

- Capable of independently identifying problems and proposing solutions
- Think outside the box

44.6 Understanding and Working with Data

Employers expect Data Analysts to:

- Be comfortable handling various data formats
- Know what data is needed to solve specific problems
- Analyze and present insights clearly

44.7 Dynamic and Fast Learner

A good Data Analyst should be:

- Adaptable to different datasets and environments
- Capable of quickly learning new tools and technologies, such as switching from Python to RStudio or adapting to different SQL environments

45 Pathways into the Data Analyst Field

45.1 Academic Pathways

An academic degree in fields such as Data Analytics, Statistics, Computer Science, Management Information Systems, or Information Technology Management offers a strong foundation for entering the data analysis field.

45.2 Online Learning Programs

If you don't have an academic qualification, online learning programs can provide an excellent alternative:

- Offered by platforms like **Coursera**, **edX**, and **Udacity**
- Designed and delivered by top domain experts
- Include hands-on assignments and real-world projects
- Projects can be added to your portfolio for job applications

45.2.1 Choosing the Right Path

With a clear understanding of the required technical, functional, and soft skills, selecting a suitable learning path becomes more straightforward.

45.2.2 Key Skills to Develop

- Statistics
- Spreadsheets
- SQL
- Python
- Data Visualization
- Problem-Solving
- Storytelling
- Making Impactful Presentations

45.3 Transitioning from a Different Career

45.3.1 With Non-Technical Experience

If you have experience in a non-technical field, you can still make a successful transition:

- Research required skills and job roles
- Connect with professionals in the field through forums, online communities, and your network
- Explore roles such as **Domain Specialist** or **Functional Analyst**
- If you're from Sales, consider becoming a **Sales Analyst** by building on your industry knowledge and learning analytical skills

45.3.2 With Technical Experience

If you're already in a technical role:

- You can easily adapt to tools and software used in data analysis
- You likely have strong domain knowledge relevant to your industry
- You may already possess transferable skills like:
    - Problem-solving
    - Project management
    - Communication
    - Storytelling

These can be further enhanced through:

- Online courses
- Communities of practice
- Professional forums

45.4 Building a Successful Career

Data analysis is a dynamic and fast-evolving field. To thrive:

- Stay curious
- Be open to continuous learning
- Embrace new challenges

Even without formal qualifications, your motivation, planning, and commitment can help you build a successful career in data analysis.

46 Career Paths in the Data Profession

46.1 Overview

The profession has evolved significantly and now presents numerous opportunities for individuals with varying interests and skill sets.

46.2 Core Career Tracks

46.2.1 Data Analyst to Data Scientist Track

- **Data Analyst**: A common entry point into the data field.

- **Data Scientist**: An advanced role involving deeper analysis and machine learning.
- **Statistician**: A specialization focused on statistical methods, often a starting role.

46.2.2 BI Analyst to Data Engineer Track

- **BI Analyst/Specialist**: Focused on business intelligence and reporting.
- **Data Engineer**: A technical role involving data infrastructure and pipelines.

These two tracks—Data Analyst to Data Scientist and BI Analyst to Data Engineer—represent parallel paths within the data profession.

46.2.3 Advanced Specializations

- **Machine Learning Engineer**
- **AI Engineer**

These roles are suitable for those interested in artificial intelligence and advanced machine learning modeling.

46.3 Evolving Career Opportunities

46.3.1 Machine Learning and Modeling

- Data Analysts can transition into roles focused on building and implementing machine learning models.

46.3.2 Business Strategy

- Analysts can move into strategic roles, using data insights to guide top-level company decisions.

46.3.3 Data Management

- Transitioning into a **Data Manager** role involves overseeing teams of analysts and prioritizing projects.
- This role is crucial due to the high volume of data-related questions versus the available resources to answer them.

46.4 Related Professions for Data-Oriented Individuals

- Bookkeeper
- Accountant
- Certified Public Accountant (CPA)
- Stockbroker
- Financial Analyst
- Real Estate Broker

These roles benefit from strong analytical and numerical skills and often overlap with data analysis responsibilities.

46.5 Suitability for the Role

Being a Data Analyst requires a strong affinity for numbers and detail-oriented thinking. If working with numbers does not come naturally or does not interest you, data analysis may not be the right career path.

47 Advice for Aspiring Data Analysts

Key points include:

- **Continuous Learning**: It's important to keep learning and not get discouraged, as there's always more to learn in analytics.
- **T-Shaped Skills**: Develop broad knowledge across various areas (like A/B testing, machine learning, SQL, etc.) while also gaining deep expertise in one specific area.
- **Utilize Every Experience**: Leverage all job experiences, even non-analytical roles, to understand and analyze data.
- **Prepare Examples**: Be ready to discuss personal and professional experiences related to analytics when speaking with potential employers.
- **Build a Portfolio**: Create a professional portfolio by analyzing fun datasets or finding opportunities to work with data in your current job.
- **Follow Your Passion**: Seek a job that aligns with your interests and brings you joy, as there are many opportunities in data analysis across various industries.

48 Women in the Data Field

Key points include:

- **Stereotypes**: Women often face stereotypes in data roles, but they can prove these wrong through their skills and work.
- **Finding Your Voice**: Women need to speak up and be heard in meetings, using data to support their ideas.
- **Continuous Learning**: Women are encouraged to keep upskilling in programming and data analysis to strengthen their presence in the field.
- **Gender Neutrality in Roles**: There are no specific roles reserved for any gender; anyone with the right skills can succeed.

49 Generative AI in Data Analytics

49.1 Introduction

Welcome to a new era of data analytics where the fusion of artificial intelligence and data exploration is revolutionizing how information is understood and interpreted.

49.2 What is Generative AI?

Generative Artificial Intelligence, or generative AI, is a category of AI focused on creating new, synthetic data. Unlike traditional AI models that primarily predict or classify, generative models generate entirely new data points, unlocking new possibilities for data analytics.

49.3 Applications of Generative AI in Data Analytics

Generative AI has numerous practical applications that enhance and innovate the data analytics workflow.

49.3.1 1. Synthetic Data Generation

- Solves the challenge of limited data availability.
- Augments existing datasets for diverse and enriched analysis.
- Enables robust model training with enhanced data diversity.

49.3.2 2. Data Imputation

- Fills in missing data points.
- Provides a more complete and accurate picture for analysts and data scientists.

49.3.3 3. Data Representation Transformation

- Converts data between different formats (e.g., text to images and vice versa).
- Offers new perspectives for representing complex information creatively.

49.3.4 4. Data Preparation and Cleaning

- Automates and enhances:
    - Data cleaning
    - Normalization
    - Transformation processes
- Streamlines the journey from raw data to actionable insights.

49.3.5 5. Intelligent Querying and Q&A Models

- Assists in formulating complex queries.
- Optimizes interactions with databases.
- Adapts to evolving data structures.

- Empowers users to ask questions in plain language and get meaningful responses.
    - **Example Models**:
        - **OpenAI GPT**: Fine-tunable language model for Q&A.
        - **Google BERT**: Excels in contextual understanding and user query response.

49.3.6 6. Enhanced Data Visualization

- Improves quality and aesthetics of data visualizations.
- Enables interactive and adaptive visual elements.
- Generates appealing representations that make data more accessible and engaging.
- **Example Tools**:
    - Tableau AI
    - IBM Cognos AI Assistant
    - Google's Looker AI

49.3.7 7. Dashboard Creation

- Simplifies and accelerates the creation process.
- Offers:
    - Dynamic layouts
    - Insightful widgets
    - Personalized user experiences
- Enhances data communication and usability.

49.3.8 8. Data Storytelling

- Generates narrative elements around data.
- Highlights key insights.
- Provides cohesive structure to raw data.
- Transforms analytics into compelling stories.

50 Summary and Highlights

50.1 Data Analyst Roles Across Industries

Data Analyst roles are in demand across multiple industries, including:

- Banking and Finance
- Insurance
- Healthcare
- Retail
- Information Technology

50.2 High Demand and Competitive Salaries

- The demand for skilled data analysts currently exceeds the supply.
- Companies are willing to pay a premium to attract and retain skilled analysts.

50.3 Categories of Data Analyst Job Roles

50.3.1 Data Analyst Specialist Roles

- Start as a **Junior Data Analyst**.
- Progress through levels by enhancing your technical, statistical, and analytical skills.
- Reach advanced roles like **Principal Analyst** with continued skill development.

50.3.2 Domain Specialist Roles

- Suitable for individuals with specialization in a particular domain.
- Allows you to become a recognized authority in that domain.

50.3.3 Analytics-enabled Job Roles

- These roles benefit from analytical skills that can enhance performance and set you apart from peers.

50.3.4 Other Data Professions

- Related roles in the data ecosystem include:
  - Data Engineer
  - Big Data Engineer
  - Data Scientist
  - Business Analyst
  - Business Intelligence Analyst
- With the right skills, you can transition into these positions.

50.4 Pathways to Enter the Data Analyst Field

50.4.1 Academic Education

- Degrees in Data Analytics, Statistics, or Computer Science.

50.4.2 Online Learning

- Specializations and certificate programs on platforms like:
  - Coursera
  - edX
  - Udacity

50.4.3 Mid-career Transitions

- Upskill based on your background:
  - **Technical background**: Focus on refining data analysis-specific technical skills.
  - **Non-technical background**: Start with basic technologies and build your way up gradually.

Module 6

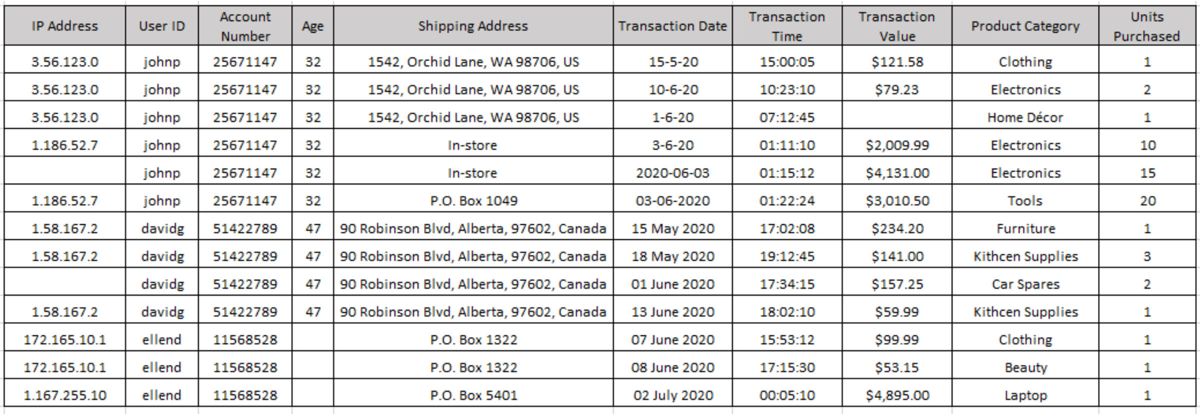
51 Using Data Analysis for Detecting Credit Card Fraud

Today's companies employ analytical techniques for the early detection of credit card frauds, a key factor in mitigating fraud damage. The most common type of credit card fraud does not involve the physical stealing of the card, but that of credit card credentials, which are then used for online purchases.

Imagine that you have been hired as a Data Analyst to work in the Credit Card Division of a bank. And your first assignment is to join your team in using data analysis for the early detection and mitigation of credit card fraud.

In order to prescribe a way forward, that is, suggest what should be done in order for fraud to get detected early on, you need to understand what a fraudulent transaction looks like. And for that you need to start by looking at historical data.

51.1 Here is a sample data set that captures the credit card transaction details for a few users.



IP Address	User ID	Account Number	Age	Shipping Address	Transaction Date	Transaction Time	Transaction Value	Product Category	Units Purchased
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	15-5-20	15:00:05	$121.58	Clothing	1
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	10-6-20	10:23:10	$79.23	Electronics	2
3.56.123.0	johnp	25671147	32	1542, Orchid Lane, WA 98706, US	1-6-20	07:12:45		Home Décor	1
1.186.52.7	johnp	25671147	32	In-store	3-6-20	01:11:10	$2,009.99	Electronics	10
	johnp	25671147	32	In-store	2020-06-03	01:15:12	$4,131.00	Electronics	15
1.186.52.7	johnp	25671147	32	P.O. Box 1049	03-06-2020	01:22:24	$3,010.50	Tools	20
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	15 May 2020	17:02:08	$234.20	Furniture	1
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	18 May 2020	19:12:45	$141.00	Kithcen Supplies	3
	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	01 June 2020	17:34:15	$157.25	Car Spares	2
1.58.167.2	davidg	51422789	47	90 Robinson Blvd, Alberta, 97602, Canada	13 June 2020	18:02:10	$59.99	Kithcen Supplies	1
172.165.10.1	ellend	11568528		P.O. Box 1322	07 June 2020	15:53:12	$99.99	Clothing	1
172.165.10.1	ellend	11568528		P.O. Box 1322	08 June 2020	17:15:30	$53.15	Beauty	1
1.167.255.10	ellend	11568528		P.O. Box 5401	02 July 2020	00:05:10	$4,895.00	Laptop	1

Descriptive techniques of analysis, that is, techniques that help you gain an understanding of what happened, include the identification of patterns and anomalies in data. Anomalies signify a variation in a pattern that seems uncharacteristic, or, out of the ordinary. Anomalies may occur for perfectly valid and genuine reasons, but they do warrant an evaluation because they can be a sign of fraudulent activity.

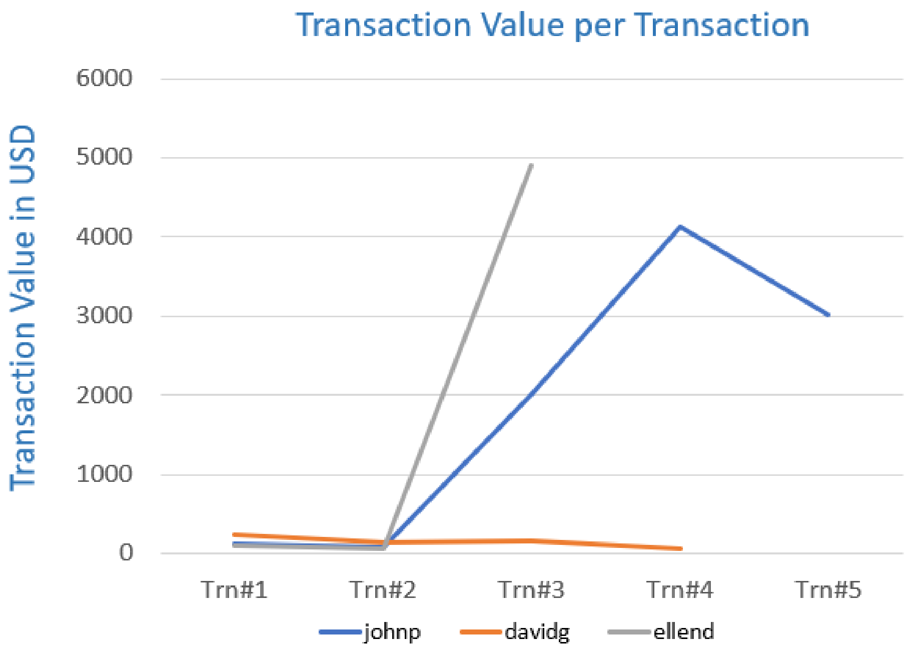
Past studies have suggested that some of the common events that you may need to watch out for include:

- A change in frequency of orders placed, for example, a customer who typically places a couple of orders a month, suddenly makes numerous transactions within a short span of time, sometimes within minutes of the previous order.
- Orders that are significantly higher than a user's average transaction.
- Bulk orders of the same item with slight variations such as color or size—especially if this is atypical of the user's transaction history.
- A sudden change in delivery preference, for example, a change from home or office delivery address to in-store, warehouse, or PO Box delivery.
- A mismatched IP Address, or an IP Address that is not from the general location or area of the billing address.

Before you can analyze the data for patterns and anomalies, you need to:

- **Identify and gather all data points that can be of relevance to your use case.** For example, the card holder's details, transaction details, delivery details, location, and network are some of the data points that could be explored.
- **Clean the data.** You need to identify and fix issues in the data that can lead to false or incomplete findings, su
- **";**ch as missing data values and incorrect data. You may also need to standardize data formats in some cases, for example, the date fields.

Finally, when you arrive at the findings, you will create appropriate visualizations that communicate your findings to your audience. The graph below samples one such visualization that you would use to capture a trend hidden in the sample data set shared earlier on in the case study.



In the next section you will be asked to answer the following 5 (five) questions based on this case study:

1. List at least 5 (five) data points that are required for the analysis and detection of a credit card fraud. (3 marks)
2. Identify 3 (three) errors/issues that could impact the accuracy of your findings, based on a data table provided. (3 marks)
3. Identify 2 (two) anomalies, or unexpected behaviors, that would lead you to believe the transaction may be suspect, based on a data table provided. (2 marks)
4. Briefly explain your key take-away from the provided data visualization chart. (1 mark)
5. Identify the type of analysis that you are performing when you are analyzing historical credit card data to understand what a fraudulent transaction looks like. [Hint: The four types of Analytics include: Descriptive, Diagnostic, Predictive, Prescriptive] (1 mark)

51.2 Answers:

51.2.1 Q1:

My Version

1. A change in the frequency of orders placed.
2. Orders that are significantly higher than a user's average transaction.
3. Bulk orders of the same item with slight variations such as color or size—especially if this is atypical of the user's transaction history.

4. A sudden change in delivery location preferences.
5. A mismatched IP Address, or an IP Address not from the billing address's general location or area.

Correct Version on Coursera:

1. Cardholder / Customer ID
2. Transaction date
3. Transaction time
4. Transaction value
5. Shipping address
6. IP address
7. Device model
8. Location

51.2.2 Q2:

1. **Missing Data values:**

Missing Values in columns such as IP address, Age, and Transaction Value may cause inaccurate findings.

2. **Incorrect Data**

Incorrect Data, if present in the data set may cause inaccurate outputs.

3. **Standardize Date Formats**

Different date formats in the "Transaction date" column may cause inaccuracy in the findings.

51.2.3 Q3:

4. The first unexpected behavior that I have noticed is the In-store transaction of "John" with a different IP Address, "1.186.52.7." This transaction is more than the average spending limit of John for almost the same products related to electronics and in higher quantities.
5. The second unexpected behavior that I have noticed is the transaction of "Ellen" with a different IP Address "1.167.155.10" and an unexpected address "P.O. Box 5401" different from the previous addresses and for an unexpected amount of 4895 dollars.

51.2.4 Q4:

My version:

The Data visualization chart is a Line Graph. This graph shows the relationship of each transaction of John, David, and Ellend with the value of each transaction in US Dollars. The transaction number is on the x-axis of the graph and the value/amount of the transaction is on the y-axis of the graph.

Coursera Version:

The visualization depicts the transaction values per transaction for all three users. The key take-away from this visualization is the sharp rise in the transaction values for users johnp and ellend, which may be indicative of an anomaly.

51.2.5 Q5:

The type of analysis that we are performing is “Descriptive Analysis”. The descriptive analysis helps us to gain an understanding of what happened including the identification of patterns and anomalies in the given data.

52 Congratulations and Next Steps

Congratulations on completing the course! We hope you enjoyed it.

This course is part of:

- IBM Data Analysis and Visualization Foundations Specialization
- IBM Data Analyst Professional Certificate

As a next step, you can explore other courses in these programs, starting with Excel Basics for Data Analysis.

If you are looking to understand and practice the basics of Data Analysis without any programming, we encourage you to complete the 3-course IBM Data Analyst Foundations Specialization. But if you are looking to start a career as a Data Analyst, the 8- course IBM Data Analyst Professional Certificate will empower you with the skills to become job-ready in this field.

We encourage you to leave your feedback and rate the course.

53 Course Credits and Acknowledgements

Primary Instructor Rav Ahuja

Other Contributors & Staff

1. Project Lead: Rav Ahuja Instructional
2. Designer/Writer: Priya Kapoor

Production Team

1. Publishing: Grace Barker, Eboney Hinds

2. Project Coordinator: Heather Vaughan
3. Narration: Bella West
4. Video Production: Simer Preet

Teaching Assistants and Forum Moderators

1. Lakshmi Holla
2. Malika Singla
3. Pratiksha Verma