2: Data Science with Python

EN6106 - Emerging Topics in IT

Level III - Semester 6

Overview

2.1 Introduction to Data Science and Data Engineering

2.2 Data Engineering Pipeline and Infrastructure

2.3 How Data Driven Insights can be Applied in Different Fields

2.4 Using Data Science to Extract Meaning from Data

2.5 Data Visualization

2.6 Data Science and Python

Intended Learning Outcomes

At the end of this lesson, you will be able to;

- Identify the range of disciplines to which Information Technology is applicable.
- Develop new skills related to the topics covered in this module.
- Apply newly learned skills and develop IT applications related to those topics.

2.1 Introduction to Data Science and Data Engineering

What is Data?

- Data is a collection of information.

❖ Data can be categorized into two groups:

- Unstructured data - not organized. We must organize the data for analysis purposes.
- Structured data - organized and easier to work with.

2.1 Introduction to Data Science and Data Engineering

What is Data Science?

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

- It involves the use of various techniques from fields such as statistics, machine learning, computer science and domain knowledge to analyze and interpret data.

- The goal of data science is to gain insights and make data driven decisions to solve complex problems in various industries, such as healthcare, finance, and retail.

2.1 Introduction to Data Science and Data Engineering

What is Data Science?

- Data Science is about finding patterns in data, through analysis, and make future predictions.

- By using Data Science, companies are able to make:

  - Better decisions
  - Predictive analysis
  - Pattern discoveries

2.1 Introduction to Data Science and Data Engineering

Where is Data Science Needed?

- Data science is needed in a wide range of industries and fields, including but not limited to:

1. **Healthcare:** Data science is used to analyze patient data to improve diagnosis and treatment.
2. **Finance:** Data science is used to detect fraud, predict stock prices, and manage financial risks.
3. **Retail:** Data science is used to optimize pricing, identify customer buying patterns, and improve marketing strategies.
4. **Manufacturing**: Data science is used to optimize production processes, improve supply chain efficiency and monitor equipment performance.
5. **Media and entertainment:** Data science is used to analyze audience data and improve content recommendations.

2.1 Introduction to Data Science and Data Engineering

Where is Data Science Needed?

6. **Marketing and Advertising:** Data science is used to analyze customer data, improve targeting and personalization, and optimize ROI.
7. **Government:** Data science is used to improve public service delivery and in decision making for policies.
8. **Transportation:** Data science is used to optimize routes, predict maintenance needs, and improve overall efficiency.
9. **Energy:** Data science is used to optimize energy consumption, predict equipment failures and improve overall performance.
10. **Sports:** Data science is used to analyze player performance and game data to gain insights and improve performance.

2.1 Introduction to Data Science and Data Engineering

How Does a Data Scientist Work?

❖ A Data Scientist requires expertise in several backgrounds:

- ❑ Machine Learning
- ❑ Statistics
- ❑ Programming (Python or R)
- ❑ Mathematics
- ❑ Databases

❖ A Data Scientist must find patterns within the data. Before he/she can find the patterns, he/she must organize the data in a standard format.

2.1 Introduction to Data Science and Data Engineering

How Does a Data Scientist Work?

- A data scientist typically follows a process that includes the following steps:

1. **Defining the problem:** Understand the problem that needs to be solved and to define the objectives and goals of the project.
2. **Collecting and cleaning the data:** Collect and clean the data that will be used for the analysis. This may involve accessing different data sources, merging and transforming the data, and removing any errors or inconsistencies.
3. **Exploring the data:** The data scientist will explore the data to gain insights and understand its characteristics. This may involve using techniques such as visualizations, descriptive statistics and correlation analysis.

2.1 Introduction to Data Science and Data Engineering

How Does a Data Scientist Work?

4. **Modeling the data:** Use techniques from statistics and machine learning to build models that can be used to make predictions or to understand the relationships between different variables.
5. **Evaluating the models:** Evaluate the performance of the models using techniques such as cross-validation.
6. **Communicating the results:** Communicate the results of the analysis to the relevant stakeholders (business leaders, other data scientists, software engineers).
7. **Deployment:** The model can be deployed to production, where it can be used to make predictions or automate decision-making in real-world scenarios.
8. **Monitoring and maintenance:** Monitoring the performance of the deployed models, and making updates and improvements as necessary.

2.1 Introduction to Data Science and Data Engineering

What is Data Engineering

- Data Engineering is the process of designing, building, maintaining, and troubleshooting the infrastructure and systems that support the collection, storage, and processing of data.

- Data engineers work closely with data scientists and analysts to make sure that the data is accurate, reliable, and available for analysis and decision-making.

2.1 Introduction to Data Science and Data Engineering

What is Data Engineering

The main responsibilities of a data engineer include:

1. **Designing and building data pipelines:** Data engineers design and build the data pipelines that move data from various sources, such as databases, applications, and sensors, into a central data repository or data lake
2. **Storing and managing data:** Data engineers are responsible for storing and managing large amounts of data, which may include both structured and unstructured data. They may use technologies such as databases, data warehousing, and distributed file systems to store and manage the data

2.1 Introduction to Data Science and Data Engineering

What is Data Engineering

The main responsibilities of a data engineer include:

3. **Processing and analyzing data:** Data engineers work with big data technologies such as Apache Hadoop, Apache Spark, and Apache Storm to process and analyze large data sets.
4. **Ensuring data quality and security:** Data engineers ensure that the data is accurate, complete, and consistent and that the data infrastructure is secure and compliant with relevant regulations and standards.
5. **Monitoring and troubleshooting data systems:** Data engineers monitor the performance of data systems and troubleshoot issues as they arise.
6. **Collaborating with data scientists and analysts:** Data engineers work closely with data scientists and analysts to understand their data needs and to make sure that the data is available and accessible for analysis and decision-making.

2.2 Data Engineering Pipeline and Infrastructure

- ❑ A data pipeline is a system that takes data from its various sources and funnels it to its destination. It’s one component of an organization’s data infrastructure.
- ❑ Data infrastructure simply describes the unique combination of data systems, processes, and architecture that allow data to fulfill its function in an organization.
- ❑ Data pipelines are the connective tissue of this infrastructure.
- ❑ Within the data infrastructure context, the sources that feed into data pipelines could be databases, SaaS apps, data streams, or data lakes.
- ❑ Destinations could be another database, data warehouse, or operational systems where it will be analyzed and leveraged to meet business goals.

2.2 Data Engineering Pipeline and Infrastructure

What is a data pipeline?

- A data pipeline is a series of steps that are used to extract, transform, and load (ETL) data from various sources to a central location for storage and further analysis.
- A data pipeline is designed to move data from one place to another in an automated and efficient manner.

2.2 Data Engineering Pipeline and Infrastructure

What is a data pipeline?

- A typical data pipeline includes the following steps:

1. **Extraction:** Data is extracted from various sources such as databases, web APIs, social media, IoT devices, and log files.
2. **Transformation:** The extracted data is transformed and cleaned to ensure consistency and accuracy. This step may include tasks such as data validation, deduplication, and normalization.
3. **Loading:** The transformed data is loaded into a central storage location such as a data lake or data warehouse.
4. **Processing:** After the data is loaded, it can be processed to extract insights, features, and create new datasets.
5. **Analysis:** The final step is data analysis, where data scientists and analysts use the data to make predictions, identify patterns and gain insights.

2.2 Data Engineering Pipeline and Infrastructure

Why do we need data pipelines?

- Data-driven enterprises need to have data efficiently moved from one location to another and turned into actionable information as quickly as possible.
- However, there are many obstacles to clean data flow, such as bottlenecks (which result in latency), data corruption, or multiple data sources producing conflicting or redundant information.
- Data pipelines take all the manual steps needed to solve those problems and turn the process into a smooth, automated workflow.
- Furthermore, data pipelines improve security by restricting access to authorized teams only.

2.2 Data Engineering Pipeline and Infrastructure

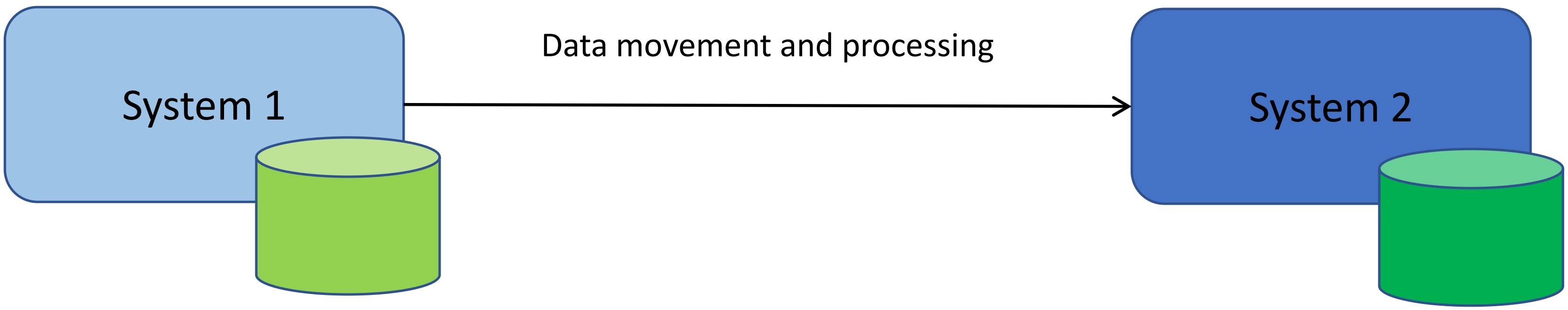
What is Data Pipeline Architecture?

- A data pipeline is a set of tools and activities for moving data from one system with its method of data storage and processing to another system in which it can be stored and managed differently.
- We define data pipeline architecture as the complete system designed to capture, organize, and dispatch data used for accurate, actionable insights.
- The architecture exists to provide the best laid-out design to manage all data events, making analysis, reporting, and usage easier.

2.2 Data Engineering Pipeline and Infrastructure

What is Data Pipeline Architecture?

- The simplest illustration of a data pipeline is given below.



2.2 Data Engineering Pipeline and Infrastructure

What is Data Pipeline Architecture?

- ❑ We break down data pipeline architecture into a series of parts and processes, including:

1. Sources
2. Joins
3. Extraction
4. Standardisation
5. Correction
6. Loads
7. Automation

2.2 Data Engineering Pipeline and Infrastructure

What is Data Pipeline Architecture?

- ❑ **Sources -**This part is where it all begins, where the information comes from. This stage potentially involves different sources, such as application APIs, the cloud, relational databases, NoSQL, and Apache Hadoop.
- ❑ **Joins -** Data from different sources are often combined as it travels through the pipeline. Joins list the criteria and logic for how this data comes together.
- ❑ **Extraction -** Data analysts may want certain specific data found in larger fields, like an area code in a telephone number contact field. Sometimes, a business needs multiple values assembled or extracted.
- ❑ **Standardization -** Say you have some data listed in miles and other data in kilometers. Standardization ensures all data follows the same measurement units and is presented in an acceptable size, font, and color.

2.2 Data Engineering Pipeline and Infrastructure

What is Data Pipeline Architecture?

- □ **Correction -** If you have data, then you will have errors. It could be something as simple as a zip code that doesn’t exist or a confusing acronym. The correction phase also removes corrupt records.
- □ **Loads -** Once the data is cleaned up, it's loaded into the proper analysis system, usually a data warehouse, another relational database, or a Hadoop framework.
- □ **Automation -** Data pipelines employ the automation process either continuously or on a schedule. The automation process handles error detection, status reports, and monitoring.

2.2 Data Engineering Pipeline and Infrastructure

Data Pipeline Tools

- ❑ Data pipelining tools and solutions come in many forms, but they all have the same three requirements:
  - Extract data from multiple relevant data sources.
  - Clean, alter, and enrich the data so it can be ready for analysis.
  - Load the data to a single source of information, usually a data lake or a data warehouse.

2.2 Data Engineering Pipeline and Infrastructure

Data Pipeline Tools

- ❑ **Batch -** Batch processing tools are best suited for moving large amounts of data at regularly scheduled intervals, but you don't require it in real-time. Popular pipeline tools include:

  Informatica PowerCenter

  IBM InfoSphere DataStage

- ❑ **Cloud-native -** These tools are optimized for working with cloud-based data, like Amazon Web Services (AWS) buckets. Since the cloud also hosts the tools, organizations save on in-house infrastructure costs. Cloud-native data pipelining tools include:

  Blendo

  Confluent

2.2 Data Engineering Pipeline and Infrastructure

Data Pipeline Tools

- ❑ **Open-source -** A classic example of "you get what you pay for," open source tools are home-grown resources built or customized by your organization's experienced staff. Open source tools include:

  Apache Kafka

  Apache Airflow

  Talend

- ❑ **Real-time -** As the name suggests, these tools are designed to handle data in real-time. These solutions are perfect for processing data from streaming sources such as telemetry data from connected devices (like the Internet of Things) or financial markets. Real-time data pipeline tools include:

  Confluent

  Hevo Data

  StreamSets

2.3. How Data Driven Insights can be Applied in Different Fields

Data-driven insights can be applied in a wide range of fields.

1. **Marketing:** By analyzing customer data, companies can identify trends and patterns that can inform their marketing strategies and help them target specific segments of their audience more effectively.
2. **Healthcare:** By analyzing patient data, healthcare providers can identify trends and patterns in disease and treatment outcomes, which can inform the development of new treatments and improve patient care.
3. **Finance:** By analyzing financial data, banks and other financial institutions can identify trends and patterns in investment and spending patterns, which can inform their risk management strategies and help them make better investment decisions.

2.3. How Data Driven Insights can be Applied in Different Fields

4. **Manufacturing:** By analyzing data on production processes and equipment performance, manufacturers can identify areas for improvement, optimize production processes, and reduce costs.

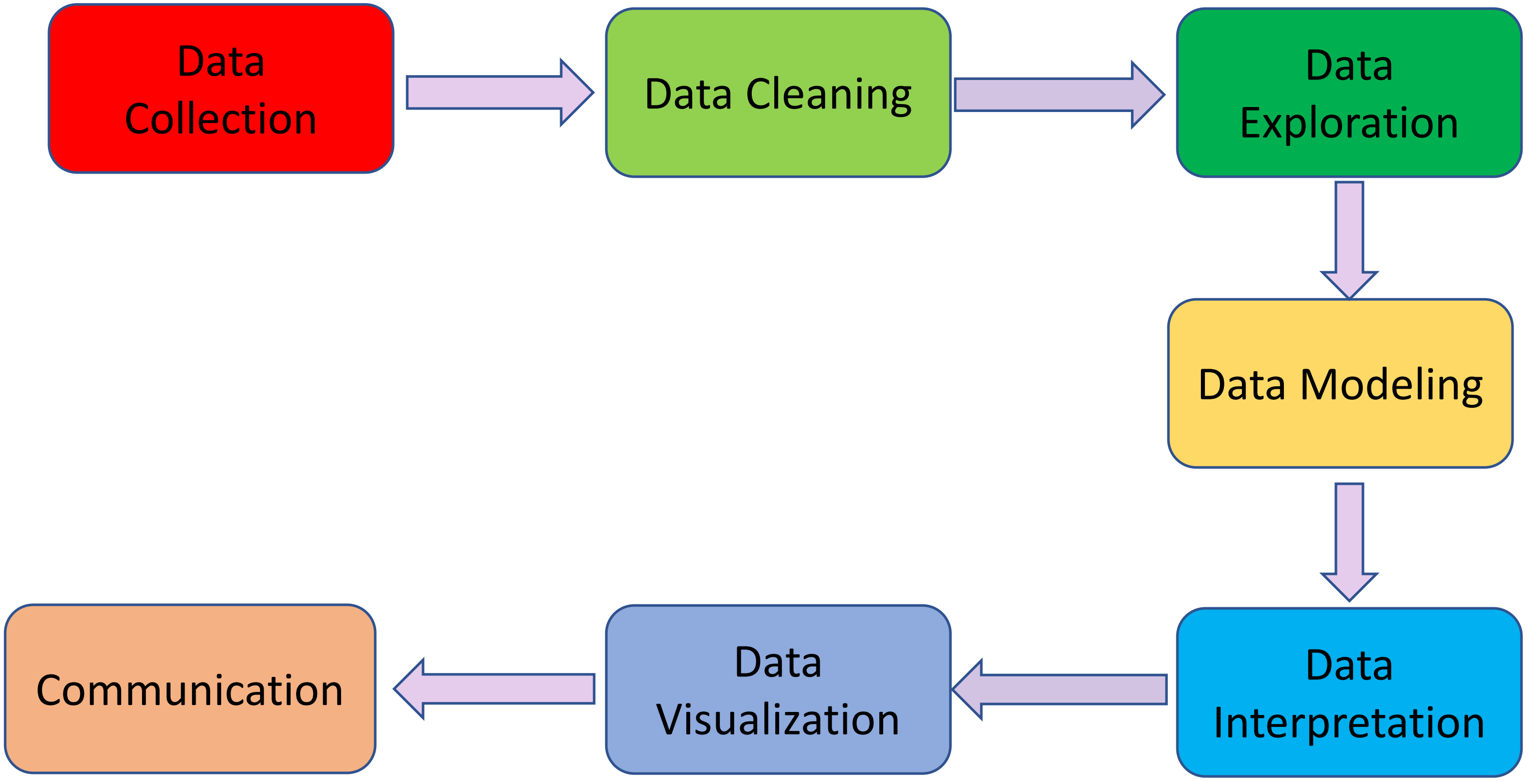
5. **Retail:** By analyzing data on customer behavior and sales, retailers can identify trends and patterns in consumer demand and adjust their inventory and pricing strategies accordingly.

6. **Transportation:** By analyzing data on traffic patterns, transportation companies can optimize routes and schedules, reducing costs and increasing efficiency.

7. **Public Sector:** By analyzing data from various sources, government agencies can identify trends and patterns that inform policy decisions, improve the delivery of services, and enhance the overall effectiveness of government.

2.4. Using Data Science to Extract Meaning from Data

❑ There are several steps that can be taken to use data science to extract meaning from data:



❑ This process often requires iteration between the steps, specially when the insights from the data are not clear or require further exploration.

2.4. Using Data Science to Extract Meaning from Data

❑ There are several steps that can be taken to use data science to extract meaning from data:

1. **Data collection:** The first step is to collect data from various sources. This can include surveys, experiments, sensor data, and other sources.
2. **Data cleaning:** Once the data is collected, it must be cleaned and preprocessed to remove any errors, outliers, or missing values. This step is important to ensure that the data is accurate and ready for analysis.
3. **Data exploration:** After the data has been cleaned, it should be explored to understand its structure and characteristics. This can be done using techniques such as visualizations, summary statistics, and correlation analyses.

2.4. Using Data Science to Extract Meaning from Data

❑ There are several steps that can be taken to use data science to extract meaning from data:

4. Data modeling: Once the data has been explored, models can be built to extract meaning from it. This can include statistical models, machine learning models, and other techniques.

5. Data interpretation: Finally, the results of the data analysis should be interpreted to extract meaning and insights from the data. This can include identifying patterns, making predictions, or drawing conclusions.

6. Data visualization: Visualization of data is the key to present insights and patterns to stakeholders. It helps in understanding the patterns and insights in a simple and effective way.

7. Communication: Communicating the insights and findings to stakeholders is important to drive the business decisions.

2.5 Data Visualization

- Data visualization is the process of creating graphical representations of data to make it more easily understandable and interpretable.
- Data visualization is the graphical representation of information and data.
- The goal of data visualization is to take complex data sets and present them in a way that is easy to understand and interpret.
- This can include creating charts, graphs, maps, and other visual representations of data.
- Data visualization can be used in many different fields such as business, science, engineering, and many more.
- It is a powerful tool to communicate the insights and patterns in data which helps in decision making.

2.5 Data Visualization

❑ There are many different types of data visualizations, each with its own strengths and weaknesses.

- Bar charts: used to compare the values of different categories
- Line charts: used to show changes in data over time
- Pie charts: used to show the proportion of different categories
- Scatter plots: used to show the relationship between two variables
- Heat maps: used to show the distribution of data across a two-dimensional surface
- Geographic maps: used to show data on a geographical map
- Bubble charts: used to show the relationship between three variables
- Treemaps: used to show the hierarchical structure of data
- Word Clouds: used to show the frequency of words in a text

2.5 Data Visualization

❑ It is important to note that when creating data visualizations, it is important to choose the right type of visualization for the data and to use best practices for design and color choices to make the information clear and easy to understand.

❑ By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

❑ Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

❑ In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

2.5 Data Visualization

Advantages

- ❑ Our eyes are drawn to colors and patterns. We can quickly identify red from blue, and squares from circles.
- ❑ Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message.
- ❑ When we see a chart, we quickly see trends and outliers.
- ❑ Some other advantages of data visualization include:
    - Easily sharing information.
    - Interactively explore opportunities.
    - Visualize patterns and relationships.

2.5 Data Visualization

Disadvantages

- ❑ When viewing a visualization with many different data points, it’s easy to make an inaccurate assumption.

- ❑ Some other disadvantages include:
  - Biased or inaccurate information.
  - Correlation doesn’t always mean causation.
  - Core messages can get lost in translation.

2.6 Data Science & Python

Python

- ❑ Python is a programming language widely used by Data Scientists.
- ❑ Python has in-built mathematical libraries and functions, making it easier to calculate mathematical problems and to perform data analysis.
  - Pandas
  - Numpy
  - Matplotlib
  - SciPy

2.6 Data Science & Python

Why we use python for data science?

There are several reasons why Python is a popular choice for data science.

1. Python has a large and active community, which means there are many libraries and frameworks available for data science tasks, such as NumPy, pandas, Matplotlib, scikit-learn, TensorFlow, and Keras.
2. Python is easy to learn and use, even for people with little or no programming experience. Its simple and readable syntax makes it a great choice for data exploration, cleaning, and visualization.
3. Python is a versatile language, which means it can be used for a wide range of tasks, from web development to machine learning to data science. This makes it a great choice for data scientists who need to perform a variety of tasks in their workflow.

2.6 Data Science & Python

Why we use python for data science?

There are several reasons why Python is a popular choice for data science.

4. Python has a large number of libraries for data science and machine learning, which makes it easy to perform complex tasks such as data preprocessing, modeling, and evaluation.

5. Python's libraries and frameworks are well-documented and supported, which means it's easy to find help and resources when you need them.

2.6 Data Science & Python

❖ There are many ways to use Python for data science. Common libraries and frameworks include.

- ❑ **NumPy:** A library for handling numerical data and arrays.
- ❑ **pandas:** A library for handling and manipulating data in a tabular format.
- ❑ **Matplotlib and Seaborn:** Libraries for creating visualizations of data.
- ❑ **scikit-learn:** A library for machine learning tasks.
- ❑ **TensorFlow and Keras:** Libraries for deep learning.