**Introduction to Data Engineering.**

**Learning Objectives**

* Recall the different entities that form a modern data ecosystem.

Data repositories, data integration platform, data pipeline, data platform, Big data, data stores, data security, data Governance and compliance, data privacy.

* Describe and differentiate between the role and responsibilities of Data

Data Engineer: the role of a data engineer is to create and maintain data architecture for business operations, mined reporting and analysis, they are responsible for collecting, processing, storing, and making data available for end users.

Data Analyst: the role of a DA is to transform data into a plain language for decision. They are responsible for analyzing data to find out insights.

Data Scientist: The role of a DC is to analyze data for actionable insights and they are responsible for detecting actionable insights and building predictive model with machine learning and deep learning.

Business Analyst and Business Intelligence Analyst: They leverage the works of DA, DC to find applicable business decisions such as in marketing.

* Engineers, Data Scientists, Data Analysts, Business Analysts, and Business Intelligence Analysts.
* Explain what Data Engineering is

Data Engineering involves collecting data from disparate sources, cleaning and transforming data to make it useable, and storing data to make it accessible to users.

* List the tasks that need to be performed in a typical data engineering lifecycle.CPSM
* Identify the essential skills and qualities for data engineering as identified by data professionals.

They must have a good knowledge of programming, a sound knowledge of system and technology architectures, they must a deep knowledge of relational databases and NoSQL databases, good work ethics.

* Summarize how Data Engineering has evolved over the past few decades.

Data engineering landscape today is very different from what it was two decades back, the quantity of data that is handled today, was unthinkable two decades agp, variety and formats of data available, big data was unheard, expectation of data engineering have changed a lot, where the introduction automation tool has appeared, cloud computing, they spend more time on setting up and managing systems, two decades ago DE was dealing only RDBMS but today the amount of variety in data, it is expected from DE to know to work with NoSQL data. They need to know how to work with Big data system and pipelines

* Discuss the responsibilities and skillsets of a Data Engineer.
* Recall the various ways that data professionals define data engineering and differentiate it from data analysis and data science.
* Describe what a day in the life of a Data Engineer looks like.

API, Web scrapping, Data Repository, dashboard data pipelines….

WEEK 2

* Describe the elements of a data engineering ecosystem which includes data, data repositories, data integration platforms, data pipelines, languages, and BI and Reporting tools.
* Differentiate between the three structures of data: structured, semi-structured, and unstructured.
* Compare and contrast standard file formats for Data Engineering.
* Describe common sources of data including relational databases; flat files and XML datasets; APIs and web services; web scraping; and data streams and feeds.
* Discuss the characteristics and use of some of the programming, querying, and scripting languages relevant to data professionals.
* Define metadata management and explain its importance.
* Explain what data repositories are and the purpose they serve.
* Describe RDBMSs, list examples of them, and summarize their use cases, advantages, and disadvantages.
* Recall experiences that data professionals have had working with varied data sources and types.
* Describe RDBMSes and NoSQL databases as well as examples of use.
* Define NoSQL databases and list their types.
* Differentiate between RDBMs and NoSQL databases.
* Discuss the characteristics and applications of data warehouses, data marts, and data lakes.
* List essential considerations for choosing a data repository.
* Summarize ETL and ELT process as well as data pipelines.
* Explain the use of Data Integration Platforms and how they relate to data pipelines and the ETL and ELT processes.
* Recognize the data engineering tools, databases, and data repositories with which data professionals work.
* [Optional] Create a Db2 instance on IBM Cloud.
* Summarize what big data is and how it impacts the collection, monitoring, storage, analysis, and reporting of data.
* Discuss the role that Apache Hadoop, Apache Hive, and Apache Spark play in Big Data analytics.
* Recall various data professionals’ viewpoints regarding Big Data’s impact on data engineering.