**Data Engineering Introduction**

**Introduction to Data Engineering.**

**Learning Objectives**

* Recall the different entities that form a modern data ecosystem.
* Describe and differentiate between the role and responsibilities of Data Engineers, Data Scientists, Data Analysts, Business Analysts, and Business Intelligence Analysts.
* Explain what Data Engineering is.
* List the tasks that need to be performed in a typical data engineering lifecycle.
* Identify the essential skills and qualities for data engineering as identified by data professionals.
* Summarize how Data Engineering has evolved over the past few decades.
* 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.

The value of data depends on two things: **The accuracy of data and the accessibility of data when needed.**

As a Data Engineer, you will know about data, data repositories, data pipelines, data integration platforms, Big data, Data Platforms, data stores, ETL process, ELT process, data security, data privacy, and Governance and compliance.

**Data sources**: structured and unstructured data.

When you’re working with so many different sources of data, the first step is to pull a copy of the data from the original sources into a data repository.

Reliability, security, and integrity of the data being acquired are some of the challenges you work through at this stage (Data Collection).

**Once the raw data is in a commonplace, it needs to get organized, cleaned up, and optimized for access by end-users**. The data will also need to conform to compliances and standards enforced in the organization.

For example, conforming to guidelines that regulate the storage and use of personal data such as health, biometrics, or household data in the case of IoT devices. Adhering to master data tables within the organization, to ensure standardization of master data across all applications and systems of an organization, is another example. The key challenges at this stage could involve **data management and working with data repositories that provide high availability, flexibility, accessibility, and security**.

Finally, we have our business stakeholders, applications, programmers, analysts, and data science use cases all pulling this data from the enterprise data repository. The key challenges at this stage could include **the interfaces, APIs, and applications that can get this data to the end-users in line with their specific needs**. For example, Data Analysts may need the raw data to work with, business stakeholders may need reports and dashboards, and applications may need custom APIs to pull this data.

* K**ey Players in the Data Ecosystem.**

Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts.

**Data Engineers** are people who develop and maintain data architectures and make data available for business operations and analysis.

Data Engineers work within the data ecosystem to extract, integrate and organize data from disparate sources. They clean, transform and prepare data. They design, store and manage data in data repositories.

A data engineer must have a good knowledge of programming, sound knowledge of systems and technology architectures and in-depth understanding of relational databases and non-relational data stores.

**Data analysts** transform data and numbers into plain language to make decisions.

Data analysts are responsible for inspecting and cleaning data **for deriving insights**. They identify correlations, find patterns, and apply statistical methods to analyze and mine data. They visualize data to interpret and present the findings of data analysis.

Data analysts must have a good knowledge of **spreadsheets**, **writing queries**, **and using statistical tools to create charts and dashboards**. Modern data analyst knows **programming skills**. They must have a **strong analytical and story-telling skills.**

Analysts are the people who answer questions such as “Are the users’ search experiences generally good or bad with the search functionality on our site” or “What is the popular perception of people regarding our rebranding initiatives” or “Is there a correlation between sales of one product and another.

**Data scientists** are responsible for analyzing data for **actionable insights** and creating **predictive models** using machine learning and deep learning.

Data Scientists are people who answer questions such as “How many new social media followers am I likely to get next month?” or “What percentage of my customers am I likely to lose to competition in the next quarter” or “Is this financial transaction unusual for this customer?”.

A data scientist must have knowledge of **mathematics and Statistics**, understand **programming languages, databases, and building data models**, and finally they must have **domain knowledge.**

**Business Analysts and Business Intelligence analysts** leverage the work of data analysts and data scientists to look at possible implications for their business and the actions they need to take or recommend. **BI analysts** also focus on market forces and external influences that shape their business, they organize and monitor data on different business functions, and they explore data to extract i**nsights and actionables that improve business performance**.

**Conclusion:**

Data Engineering converts raw data into usable data. Data Analytics uses this data to generate insights. Data Scientists use Data Analytics and Data Engineering to predict **the future using data from the past**. Business Analysts and Business Intelligence Analysts use these insights and predictions **to drive decisions that benefit and grow their business.**

* **What is Data Engineering?**

In the simplest possible terms, the field of Data Engineering concerns itself with the mechanics for the flow and access of data. And its goal is to make quality data available for fact-finding and data-driven decision making.

Data engineering involves: Collecting data, Processing data, Storing data and making data available to users securely. **CPSM** steps

1. **Collecting data**: Extracting, integrating and organizing data from disparate sources. To collect required data you need to develop **tools, workflows, and processes** that help you to acquire data from multiple sources, as well as design, build and maintain scalable data architecture to store data. Data can be stored in databases, data lake, data warehouses or any other type of data repository.
2. **Processing data**: Clean, transform and prepare data to make it usable. You need to implement and maintain distributed systems for large-scale processing of data Design pipelines for the extraction, transformation, and loading of data into data repositories. Design or implement solutions for validating and safeguarding quality, privacy, and security of data. Optimize tools, systems, and workflows for performance, reliability, and scalability.
3. **Storing data:** Storing data for reliable and easy availability of data. For this, you need to: Architect or implement data stores for the storage of processed data. Ensure systems are scalable, keeping in mind the evolving nature of data and business needs. Ensure tools and systems are in place that take care of data privacy, security, compliance, monitoring, backup, and recovery.
4. **Making data available to users securely:** This includes the use of: APIs, services, and programs that retrieve data on defined parameters for use by end-users. Interfaces and dashboards that present data to users so they can derive insights from the data. Ensure the right measures and checks and balances are in place to keep data secure and provide rights-based access to users

**In conclusion:**

data engineering works to provide a robust and scalable structure to make quality data available for decision-making. More than any other data profession, data engineering is about the tools and technologies involved in data manipulation. But it is also about understanding the complexities of data and how it is ultimately leveraged for fact-finding and decision-making.

* **Viewpoint: Defining data engineering?**

Check and remember their definition

**The responsibilities of a data engineer**

The overarching responsibility of a data engineer is to provide analytics-ready data to data.

Data is **analytics-ready** when it is **accurate**, **reliable**, **complies to regulations it is governed by**, and is **accessible to consumers** when they need it.

At a broad level, Data Engineers: Extract, organize, and integrate data from disparate sources **(C)**. Prepare data for analysis and reporting by transforming and cleansing it **(P)**. Design and manage data pipelines that encompass the journey of data from source to destination systems, Setup and manage the infrastructure required for the ingestion, processing, and storage of data **(S & M)**.

A Data engineer's **technical skills** required are:

**Operating systems knowledge**: Unix, Linux, Windows administrative tools and System Utilities and commands.

**Infrastructure components:** Virtual Machine, Networking, Application services and Cloud-services.

**Databases and data Warehouses:** RDBMS, NoSQL, and data warehouses.

**Data Pipeline:** ApacheBeam, Airflow, dataflow.

**ETL** tool: IBM Infosphere, AWS, and Improvado.

**Languages:** Query language (SQL and NoSQL), Programming language (Python, Java), Shell and Scripting language (Unix and Linux shell and PowerShell).

**Big data processing:** Hadoop, Hive, and Apache Spark

A data engineer's **functional skills** required are:

* Convert business requirements into technical specifications.
* Work with software development lifecycle: Ideation, Architecture, design, prototyping, testing, and deployment.
* Understand data’s potential application in businesses
* Understand the risk of poor data management.//Data quality, security and privacy and compliance.

A data engineer’s **soft skills** are:

Interpersonal skills, effective communication, teamwork, and collaboration.

Viewpoint: Skills and qualities to be a successful data engineer.

Good work ethic, passion for learning, knowledge of programming language and love for data.