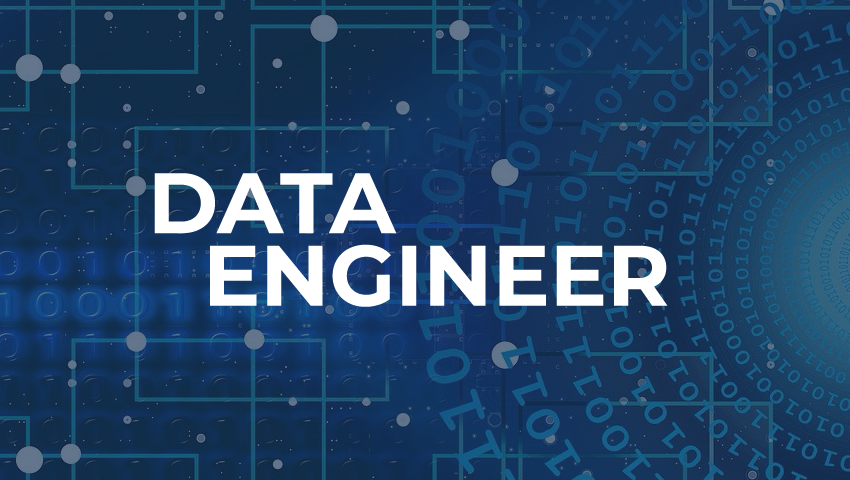
**Data Engineering for Beginners**



*(source: QuickOps Consulting)*

**Data Engineering** is one of the fastest rising career position in the data technology field. This is a complex field with task of making raw data usable to data scientists and groups within an organization. Data engineering encompasses numerous specialties of data science.

A ***data engineer*** is a type of software engineer who creates big data ETL pipelines to manage the flow of data through the organization.

**Responsibilities of a Data Engineer**

* Develops and maintains data pipelines
* Ensures Data quality & accuracy
* Design, develop and maintain database architecture
* More of ETL process – (Extract Transform Load) This is an automated process to get data into a database from the source.
* As long as the data engineer will be getting data from a given source, they will need to ensure data quality and accuracy.

**Skills Required to Become a Data Engineer**

1. ***SQL/NoSQL***

Structured Query Language or simply SQL (‘sequel’) is the cream when it comes to fundamental of skill-set for Data Engineers and any other related field of data and databases. Modern data is greatly stored in Relational Database Management Systems (RDBMS) and hence this demands highly aspiring Data Engineers to master the SQL. There are a lot of Relational Database Systems including ***Oracle 23C,*** ***MySQL, SQL Server, PostgreSQL*** just but to mention a few.

For NoSQL, - This means that data is not stored in table format hence the data is in a non-relational database design does not require a schema, it offers rapid scalability to manage large and typically unstructured data sets. Examples of NoSQL database systems are ***MongoDB*** and ***Cassandra***

1. ***Python/R***

To be able to work with data from source and create automation tasks around data, the data engineer need to have knowledge of a programming language with respective packages or libraries. Python so far is highly considered given to its versatility in the libraries like ***Pandas, Airflow***. For **R,** a Data Engineer have at their disposal ***Tidyverse*** is a collection of R packages, primarily for data engineering and analytics. These packages are ***ggplot2***, **purr**r, ***tibble***, ***dplyr***, ***tidyr***, ***stringr***, ***readr***, and ***forcats***

1. ***Apache Kafka***

This is an **open-source** distributed event streaming platform widely used in data engineering for **building real-time data pipelines** and **streaming applications**. It is designed to handle high-throughput, fault-tolerant, and scalable data streaming, making it a popular choice for processing and managing large volumes of data in real time.

It is used to design and implement robust, real-time data pipelines, enabling the efficient processing, storage, and analysis of streaming data.

1. ***AWS/Azure/GCP***

As a Data Engineer, having a knowledge of Cloud Computing is an important part of you. Big Data which cannot be

1. ***Apache Spark***

Apache Spark is **a powerful open-source** distributed computing system that is widely used in the field of data engineering. It is designed to handle large-scale data processing tasks and is known for its speed, ease of use, and versatility in supporting a wide range of applications.

It serves as a critical tool for performing data transformation, data integration, and data analytics. It provides a unified engine for big data processing that supports ***Java, Scala, Python***, and ***R***. This flexibility makes it easier for data engineers to work with different data formats and leverage their preferred programming languages for data processing tasks.

Apache Spark is particularly valuable in data engineering for the following reasons:

1. ***Speed and efficiency:*** Apache Spark can process large datasets much faster than traditional data processing systems, thanks to its in-memory processing capabilities and optimized execution engine.
2. ***Scalability:*** It can handle massive datasets and scale seamlessly to accommodate growing data volumes, making it suitable for handling big data challenges.
3. ***Fault tolerance:*** It ensures fault tolerance by storing intermediate results during processing, enabling the system to recover from failures without losing data.
4. ***Versatility:*** Apache Spark supports a wide range of data processing tasks, including batch processing, real-time streaming, machine learning, and graph processing, making it a versatile solution for diverse data engineering needs.

Data engineers utilize Apache Spark to build and maintain data pipelines, process and analyze large datasets, and perform complex data transformations. By leveraging its capabilities, data engineers can efficiently manage and process big data, extract valuable insights, and build robust data-driven applications and systems.Top of Form

1. ***Data Structures and Algorithms.***

This is crucial in data engineering, which are fundamental building blocks for the efficient management and processing of data.

***Data structures*** refer to the specific ways data is organized and stored in a memory of a computer, enabling efficient access, modification, and deletion of data. ***Algorithms*** are step-by-step procedures or formulas for performing specific tasks, such as searching, sorting, and data manipulation. Data engineering heavily relies on various algorithms to efficiently handle data operations, such as ***data transformation***, ***data integration***, and ***data aggregation***.