**INTRODUCTION TO DATA ENGINEERING**

**WHAT IS DATA ENGINEERING?**

**Modern Data Ecosystem and role of Data Engineering**

**Modern Data Ecosystem**

Modern data ecosystem includes a network of interconnected and continually evolving entities that include:

* Data, that is available in a host of different formats, structures, and sources.
* Enterprise Data Environment, in which raw data is staged so it can be organized, cleaned, and optimized for use by end-users.
* End-users, such as business stakeholders, analysts, and programmers who consume data for various purposes.

Emerging technologies such as Cloud Computing, Machine Learning, and Big Data, are continually reshaping the data ecosystem and the possibilities it offers.

Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts, all play a vital role in the ecosystem for deriving insights and business results from data.

**Key Players in the Data Ecosystem:**

* Data Engineering converts raw data into usable data.
* Data Analytics uses this data to generate insights. (Visualization reports as well)
* Data Scientists use Data Analytics and Data Engineering to predict the future using data from the past. (Build and deploy ML Algorithms for predictive modelling and augmented decision making).
* Business Analysts and Business Intelligence Analysts use these insights and predictions to drive decisions that benefit and grow their business. (Helps business stakeholders in better decision making I.e., Increase revenue and lower the cost).

**What is Data Engineering?**

* 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.

<https://www.coursera.org/learn/introduction-to-data-engineering/lecture/mAPJb/viewpoints-defining-data-engineering>

Data Engineering involves Designing, Building, maintaining data infrastructures and platforms (DB, Big data repo’s, Data pipelines) and optimizing these systems. Data Engineering makes sure that data is highly available, consistent, secure and recoverable for the users (Data Analyst, Data Scientist etc.). The goal of Data Engineering is to make quality data available for analytics and decision-making.

<https://www.coursera.org/learn/introduction-to-data-engineering/lecture/IDoU2/viewpoints-evolution-of-data-engineering>

**Note**: Data can be stored in different types of data repositories depending on the nature of the data and the intended use case. **Databases** are typically used to store structured data that can be organized into tables, while **data warehouses** are used to store large volumes of structured data for analysis and reporting. **Data Mart** is a sub-section of data warehouse built specifically for particular business function, purpose or community of users. **Data lakes,** on the other hand, are used to store both structured and unstructured data in its native format for processing and analysis.

In addition to these types of data repositories, there are other types of storage solutions, such as NoSQL databases, object stores, and file systems, that can be used to store data extracted from disparate sources. The choice of data repository depends on factors such as the type of data, the volume of data, the processing requirements, and the intended use case.

**Responsibilities and Skillsets of a Data Engineer**

The role of a Data Engineer includes:

* Gathering data from disparate sources.
* Integrating data into a unified view for data consumers.
* Preparing data for analytics and reporting.
* Managing data pipelines for a continuous flow of data from source to destination systems.
* Managing the complete infrastructure for the collection, processing, and storage of data.

To be successful in their role, Data Engineers need a mix of technical, functional, and soft skills.

* Technical Skills include working with different operating systems and infrastructure components such as virtual machines, networks, and application services. It also includes working with databases and data warehouses, data pipelines, ETL tools, big data processing tools, and languages for querying, manipulating, and processing data.
* An understanding of the potential application of data in business is an important skill for a data engineer. Other functional skills include the ability to convert business requirements into technical specifications, an understanding of the software development lifecycle, and the areas of data quality, privacy, security, and governance.
* Soft Skills include interpersonal skills, the ability to work collaboratively, teamwork, and effective communication.

**Metadata**: Metadata is the data about the other data

<https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTS1EQjAxMDBFTi1Ta2lsbHNOZXR3b3JrL3JlYWRpbmdzL1JlYWRpbmdfTWV0YWRhdGFfYW5kX01ldGFkYXRhX01hbmFnZW1lbnQubWQiLCJ0b29sX3R5cGUiOiJpbnN0cnVjdGlvbmFsLWxhYiIsImFkbWluIjpmYWxzZSwiaWF0IjoxNjY4NzU2MDM2fQ.b7sVyhWetNSEC5a11BO5OTyrvfzgGFA5hvb2unrul2Q>

**THE DATA ENGINEERING ECOSYSTEM**

**The Data Ecosystem and Languages for Data Professionals**

A Data Engineer’s ecosystem includes the infrastructure, tools, frameworks, and processes for extracting data, architecting and managing data pipelines and data repositories, managing workflows, developing applications, and managing BI and Reporting tools.

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**, that is data which is partially organized and partially free form.
* **Unstructured data**, that is data which cannot 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 repositories include **Transactional or OLTP** (designed to store high volume of day-to-day operational data) and **Analytical or OLAP** (optimized for conducting complex data analytics).

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 analyse data. These can be classified as:

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

**Data Repositories, Data Pipelines, and Data Integration Platforms**

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 store house.
* **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**, that serve as storage repositories for large amounts of structured, semi-structured, and unstructured data in their native format.
* **Bigdata Stores**, that provide distributed computational and storage infrastructure to store, scale, and process very large data sets.

The **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.

The **ELT**, or Extract Load and Transfer, Process is a variation of the ETL Process. In this process, extracted data is loaded into the target system before the transformations are applied. This process is ideal for Data Lakes and working with Big Data. **ELT Provides greater flexibility for Data Analysts & Data Scientists for Exploratory Data Analysis (EDA).**

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

**Data Integration** Platforms combine disparate sources of data, physically or logically, to provide a unified view of the data for analytics purposes.

In terms of data integration, Data Pipelines are to perform Data Integration, while ETL is a process or activity within data integration.

**Big Data Platforms**

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** (also **Veracity, Value)** of data challenged the tools and systems used for conventional data, leading to the emergence of processing tools and platforms designed specifically for Big Data. Big Data processing technologies help derive value from big data. These include NoSQL databases and Data Lakes and open-source technologies such as Apache Hadoop, Apache Hive, and Apache Spark.

* **Hadoop** provides distributed storage and processing of large datasets across clusters of computers. One of its main components, the Hadoop File Distribution System, or HDFS, is a storage system for big data.
* **Hive** is an open-source data warehouse software for reading, writing, and managing large datasets files that are stored directly either in **HDFS** or **Apache HBase.**
* **Spark** is a general-purpose data processing engine designed to extract and process large volumes of data.

**DATA ENGINEERING LIFE CYCLE**

**Data Platforms, Data Stores, and Security**

The architecture of a data platform can be seen as a set of layers, or functional components, each one performing a set of specific tasks. These layers include:

* **Data Ingestion or Data Collection Layer**, responsible for bringing data from source systems into the data platform.
* **Data Storage and Integration Layer**, responsible for storing and merging extracted data.
* **Data Processing Layer**, responsible for validating, transforming, and applying business rules to data. Analysis and User Interface Layer, responsible for delivering processed data to data consumers.
* **Data Pipeline Layer**, responsible for implementing and maintaining a continuously flowing data pipeline.

A well-designed data repository is essential for building a system that is scalable and capable of performing during high workloads.

The choice or design of a data store is influenced by the type and volume of data that needs to be stored, the intended use of data, and storage considerations. The privacy, security, and governance need of your organization also influence this choice.

The **CIA**, or Confidentiality, Integrity, and Availability triad are three key components of an effective strategy for information security. The CIA triad is applicable to all facets of security, be it infrastructure, network, application, or data security.

**Data Collection and Data Wrangling**

Depending on where the data must be sourced from, there are a number of methods and tools available for gathering data. These include query languages for extracting data from databases, APIs, Web Scraping, Data Streams, RSS Feeds, and Data Exchanges.

Once the data you need has been gathered and imported, your next step is to make it analytics-ready. This is where the process of **Data Wrangling, or Data Munging**, comes in. Data Wrangling involves a whole range of **transformations** and **cleansing** activities performed on the data.

Transformation of raw data includes the tasks you undertake to:

* Structurally manipulate and combine data using Joins and Unions.
* Normalize data, that is, clean the database of unused and redundant data, inconsistance data.
* Denormalize data, that is, combine data from multiple tables into a single table so that it can be queried faster.

Cleansing activities include:

* Profiling data to uncover anomalies and quality issues.
* Visualizing data using statistical methods in order to spot outliers.
* Fixing issues such as missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors, and outliers.

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 features, strengths, limitations, and applications.

**Note**:

**Jupyter Notebook**: An open-source web application that is widely used for data cleaning and transformation, data visualization and statistical modelling.

**Querying Data, Performance Tuning, and Troubleshooting**

* In order for raw data to become analytics-ready, a number of transformation and cleansing tasks need to be performed on raw data. And that requires you to understand your dataset from multiple perspectives. One of the ways in which you can explore your dataset is to query it.
* Basic querying techniques can help you explore your data, such as, counting and aggregating a dataset, identifying extreme values, slicing data, sorting data, filtering patterns, and grouping data.
* In a data engineering lifecycle, the performance of data pipelines, platforms, databases, applications, tools, queries, and scheduled jobs, need to be constantly monitored for performance and availability.
* The performance of a data pipeline can get impacted if the workload increases significantly, or there are application failures, or a scheduled job does not work as expected, or some of the tools in the pipeline run into compatibility issues.
* Databases are susceptible to outages, capacity overutilization, application slowdown, and conflicting activities and queries being executed simultaneously.
* Monitoring and alerting systems collect quantitative data in real time to give visibility into the performance of data pipelines, platforms, databases, applications, tools, queries, scheduled jobs, and more.
* Time-based and condition-based maintenance schedules generate data that helps identify systems and procedures responsible for faults and low availability.

**Governance and Compliance**

Data Governance is a collection of principles, practices, and processes that help maintain the security, privacy, and integrity of data through its lifecycle.

Personal Information and Sensitive Personal Information, that is, data that can be traced back to an individual or can be used to identify or cause harm to an individual, needs to be protected through governance regulations.

General Data Protection Regulation, or GDPR, is one such regulation that protects the personal data and privacy of EU citizens for transactions that occur within EU member states. Regulations, such as HIPAA (Health Insurance Portability and Accountability Act) for Healthcare, PCI DSS (Payment Card Industry Data Security Standard) for retail, and SOX (Sarbanes Oxley) for financial data are some of the industry-specific regulations.

Compliance covers the processes and procedures through which an organization adheres to regulations and conducts its operations in a legal and ethical manner.

Compliance requires organizations to maintain an auditable trail of personal data through its lifecycle, which includes acquisition, processing, storage, sharing, retention, and disposal of data.

Tools and technologies play a critical role in the implementation of a governance framework, offering features such as:

* Authentication and Access Control.
* Encryption and Data Masking.
* Hosting options that comply with requirements and restrictions for international data transfers.
* Monitoring and Alerting functionalities.
* Data erasure tools that ensure deleted data cannot be retrieved.

[**https://blogs.gartner.com/nick-heudecker/hyping-dataops/**](https://blogs.gartner.com/nick-heudecker/hyping-dataops/)

**CAREER OPPORTUNITIES AND DATA ENGINEERING ACTIVITIES**

**Career Opportunities and Learning Paths**

Data Engineering is reported to be one of the top ten jobs experiencing tremendous growth in the U.S. today. It is also reported to be one of the fastest growing tech occupations with year-over-year growth of around 50%.

Currently, the demand for skilled data engineers far outweighs the supply, which means companies are willing to pay a premium to hire skilled data engineers.

Data engineering roles in organizations tend to break the specialization up into Data Architecture, Database Design and Architecture, Data Platforms, Data Pipelines and ETL, Data Warehouses, and Big Data.

Regardless of the niche you choose to specialize in, knowledge of operating systems, languages, databases, and infrastructure components, is essential.

To work your way up from a Junior Data Engineer to a Lead or Principal Data Engineer, you need to continually advance your technical, functional, and soft skills from a foundational level to an expert level. You need to not only expand your skills in your niche area, but also into other areas of data engineering at the same time.

Big Data Engineers and Machine Learning Engineers are some of the emerging roles in this field and they require specialized skills in addition to basic data engineering.

There are several paths you can consider in order to gain entry into the data engineering field.

* An academic degree in Computer Science or engineering qualifies you for an entry-level job.
* If you are not a graduate, or a graduate in a non-relate stream, you can earn professional certifications from online multi-course specializations offered by learning platforms such as Coursera, edX, and Udacity.
* If you have a coding background, or you are an IT Support Specialist, a Software Tester, a Programmer, or a data professional such as a Statistician, Data Analyst, or BI Analyst, you can upskill with the help of online courses to become a Data Engineer.

[**https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTS1EQjAxMDBFTi1Ta2lsbHNOZXR3b3JrL3JlYWRpbmdzL1JlYWRpbmdfRGF0YV9XYXJlaG91c2luZ19TcGVjaWFsaXN0Lm1kIiwidG9vbF90eXBlIjoiaW5zdHJ1Y3Rpb25hbC1sYWIiLCJhZG1pbiI6ZmFsc2UsImlhdCI6MTY2ODc2NzkwN30.kFgDF0JjKnDi\_Riz5kW0EFnpD6YNQNvzQZLi6GjVWMI**](https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTS1EQjAxMDBFTi1Ta2lsbHNOZXR3b3JrL3JlYWRpbmdzL1JlYWRpbmdfRGF0YV9XYXJlaG91c2luZ19TcGVjaWFsaXN0Lm1kIiwidG9vbF90eXBlIjoiaW5zdHJ1Y3Rpb25hbC1sYWIiLCJhZG1pbiI6ZmFsc2UsImlhdCI6MTY2ODc2NzkwN30.kFgDF0JjKnDi_Riz5kW0EFnpD6YNQNvzQZLi6GjVWMI)

**Peer Graded Assignment:**

1. Transactional, or OLTP, systems are designed for handling high-frequency transactional data being generated in real-time. Examples of the type of transactions these systems would be used for, include:

* Online banking transactions
* Online financial transaction
* ATM transaction
* Airline booking
* Hotel booking
* Movie ticket purchases
* eCommerce purchase
* In-store purchases

2. According to the course material, the architecture of a data platform consists of five layers, each layer responsible for a specific set of tasks. These five layers are

1. Data Ingestion or Data Collection Layer

2. Data Storage and Integration Layer

3. Data Processing Layer

4. Analysis and User Interface Layer

5. Data Pipeline Layer

3. According to the course material, these are some of the factors you need to consider while selecting a data store that meets the needs of your organization:

1. **The type of data you need to store.**
2. **The volume of data you need to store.**
3. **Intended use of data.**
4. **Storage considerations.**
5. **Privacy, Security, and Governance considerations.**