



# Week 1 - Introduction to Data Engineering

## What is Data Engineering?

### Modern Data Ecosystem

To quote a Forbes 2020 report on data in the coming decade

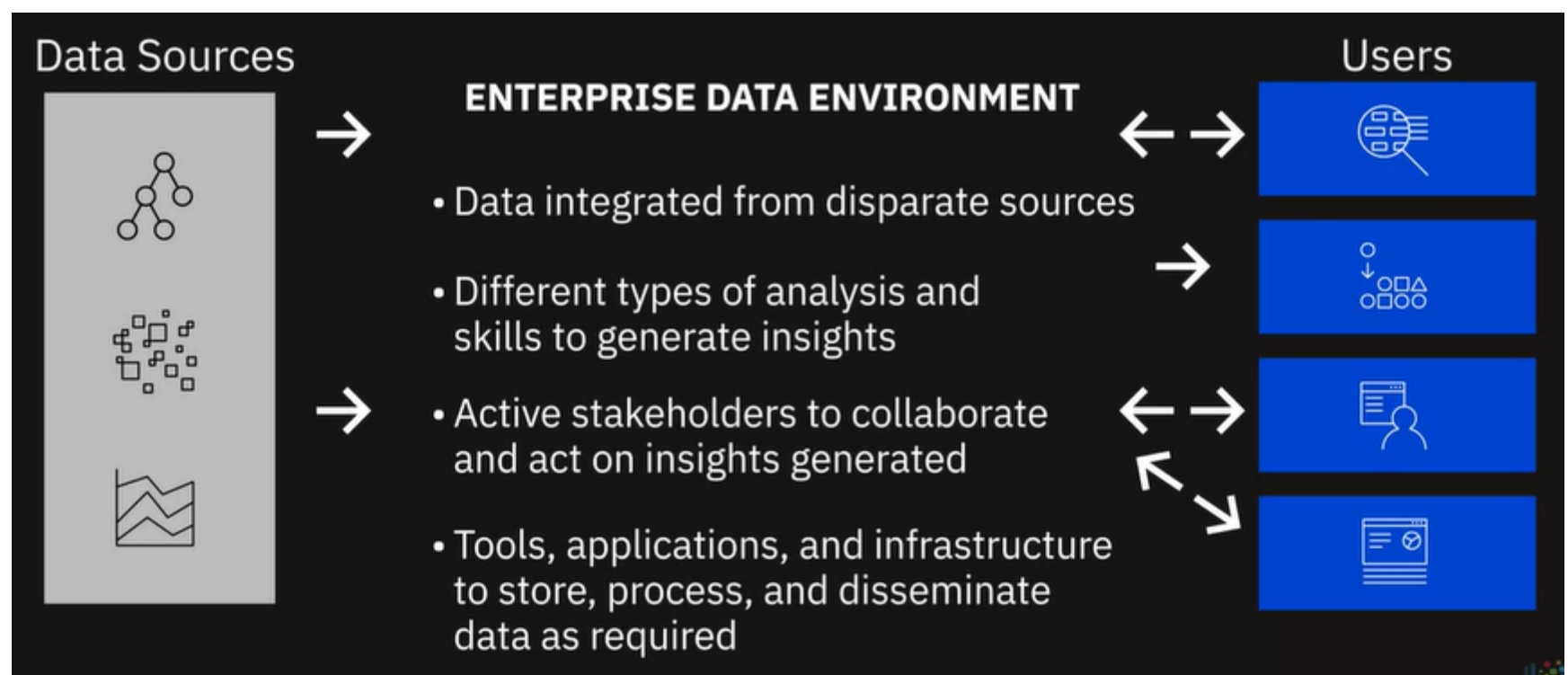
The constant increase in data processing speeds and bandwidth, the nonstop invention of new tools for creating, sharing, and consuming data, and the steady addition of new data creators and consumers around the world, ensure that data growth continues unabated. Data begets more data in a constant virtuous cycle.

A modern data ecosystem includes a whole network of:

- interconnected
- independent,
- continually evolving entities

It includes:

- data that has to be integrated from disparate sources;
- different types of analysis and skills to generate insights;
- active stakeholders to collaborate and act on insights generated;
- and tools, applications, and infrastructure to store, process, and disseminate data as required.



### Data sources

Data is available in a variety of **structured** and **unstructured datasets**,

residing in:

- text
- images
- videos
- clickstreams
- user conversations
- social media platforms
- the Internet of Things (or IoT) devices
- real-time events that stream data

- legacy databases
- sourced from professional data providers and agencies.

When you're working with so many different sources of data

**Pull a copy of the data from the original sources into a data repository.**

At this stage, you're only looking at acquiring the data you need - working with data formats, sources, and interfaces through which this data can be pulled in.

Challenges

- Reliability
- Security
- Integrity of the data

**Once the raw data is in a common place, 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.

Challenges

- data management
- working with data repositories that provide high availability, flexibility, accessibility, and security.

**We have our business stakeholders, applications, programmers, analysts, and data science use cases all pulling this data from the enterprise data repository.**

Challenges

- the interfaces,
- APIs,
- 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, applications may need custom APIs to pull this data.

Emerging technologies that are shaping today's data ecosystem and its possibilities.

For example, **Cloud Computing**, **Machine Learning**, and **Big Data**, to name a few.

## **Key Players in the Data Ecosystem**

# Overview

Organizations are using data to uncover opportunities and applying that knowledge to differentiate themselves from their competition.

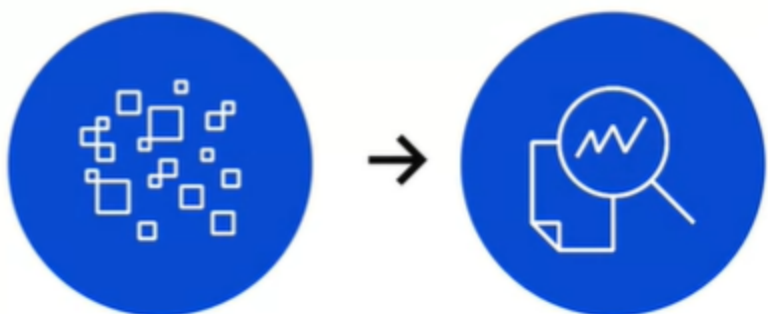
- Identifying patterns in financial data to detect fraud
- Using recommendation engines to drive conversion
- Mining social media posts for customer voice
- Analyzing customer behavior for personalizing offers

## Data Professionals



### Data Professionals:

- Data Engineers
- Data Analysts
- Data Scientists
- Business Analysts
- Business Intelligence Analysts



### Data Engineer

## Data Engineer



Data architectures



Business operations



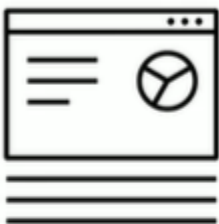
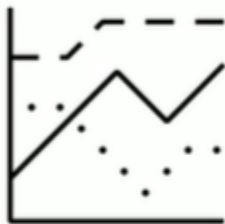
Analysis

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
- Clean, transform, and prepare data
- Design, store, and manage data in data repositories



Business Applications



Data Analysts and  
Data Scientists

# Data Engineer



## Skills:

- Good knowledge of programming
- Sound knowledge of systems and technology architectures
- In-depth understanding of relational databases and non-relational data stores

## Data Analyst

# Data Analyst



## Responsibilities of a Data Analyst:

- Inspect and clean data for deriving insights
- Identify correlations, find patterns, and apply statistical methods to analyze and mine data
- Visualize data to interpret and present the findings of data analysis

# Data Analyst



“Are the users’ search experiences generally good or bad with the search functionality on our site?”

“What is the popular perception of people regarding our rebranding initiatives?”

“Is there a co-relation between sales of one product and another?”



# Data Analyst



## Skills:

- Good knowledge of spreadsheets, writing queries, and using statistical tools to create charts and dashboards
- Programming skills
- Strong analytical and story-telling skills

## Data Scientist

# Data Scientist



## Responsibilities of a Data Scientist:

- Analyze data for actionable insights
- Create predictive models using Machine Learning and Deep Learning



# Data Scientist



“How many new social media followers am I likely to get next month?”

“What percentage of my customers am I likely to lose to competition in the next quarter?”

“Is this financial transaction unusual for this customer?”

### Skills:

- Knowledge of Mathematics and Statistics
- Understanding of programming languages, databases, and building data models
- Domain knowledge

### Business Analyst and BI Analyst

Business 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.



Data Analysts



Data Scientists



### BI Analysts

- Focus on market forces and external influences that shape their business
- Organize and monitor data on different business functions
- Explore data to extract insights and actionables that improve business performance

### To summarize

Data Engineering converts raw data into usable data.

Data Analytics uses this data to generate insights.

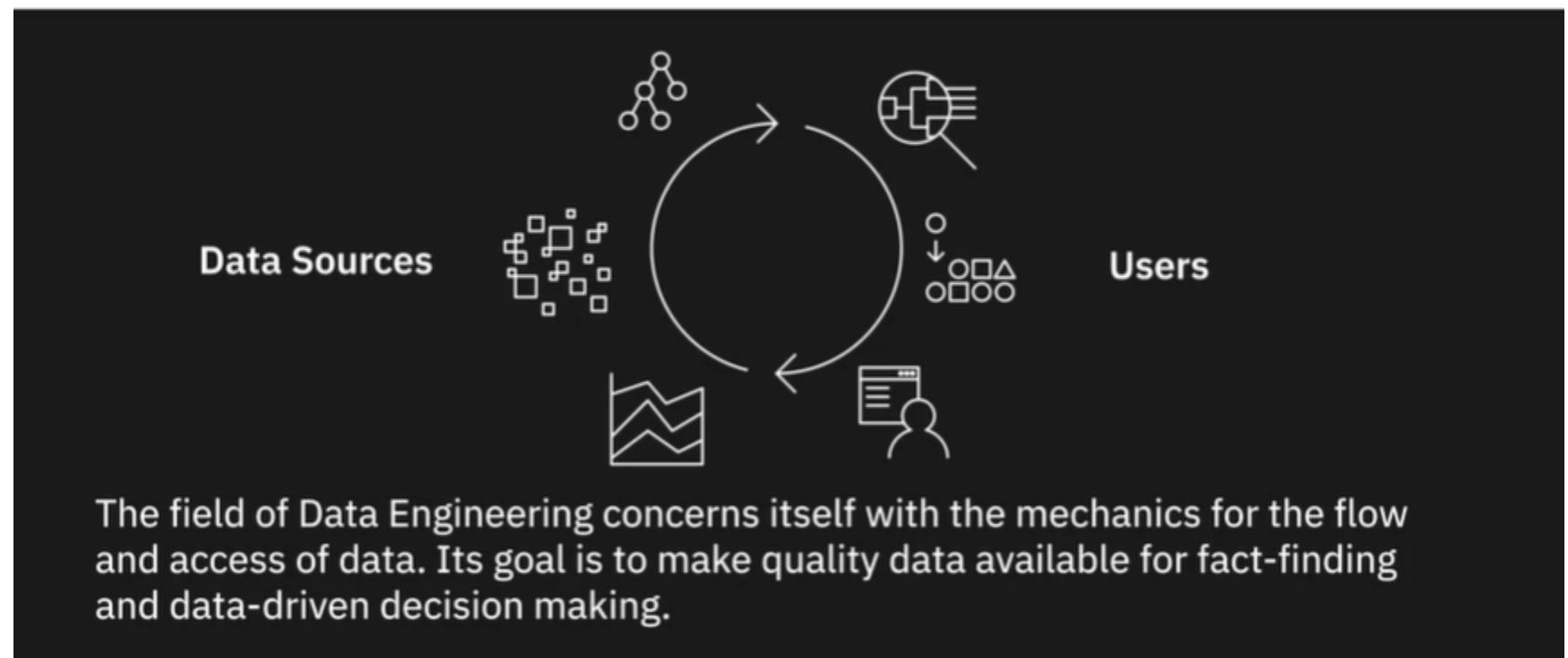
Data Scientists use Data Analytics and Data Engineering to predict the future using data from the past.

Business Analysts and Business Intelligence Analysts use these insights and predictions to drive decisions that benefit and grow their business.

Interestingly, it's not uncommon for data professionals to start their career in one of the data roles and transition to another role within the data ecosystem by supplementing their skills.

## What is Data Engineer?

# Introduction



# What is Data Engineering?





# What is Data Engineering?

The field of Data Engineering involves:



Extracting, integrating, and organizing data from disparate sources

- Data acquisition from multiple sources
- Data architecture for storing source data

# What is Data Engineering?

The field of Data Engineering involves:



Cleaning, transforming, and preparing data to make it usable

- Distributed systems for processing data
- Pipelines for extracting, transforming, and loading data
- Solutions for safeguarding quality, privacy, and security of data
- Performance optimization
- Adherence to compliance guidelines

# What is Data Engineering?

The field of Data Engineering involves:



Storing data for reliability and easy availability of data

- Data stores for storage of processed data
- Scalable systems
- Ensuring data privacy, security, compliance, monitoring, backup, and recovery

# What is Data Engineering?

The field of Data Engineering involves:



Making data available to users securely

- APIs, services, and programs for retrieving data for end-users
- User access through interfaces and dashboards
- Checks and balances to ensure data security

## Data Engineering is a team sport



### Conclusion

- provide a robust and scalable structure to make quality data available for decision-making.
- 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.

### Viewpoints: Defining Data Engineering

# Modern Data Ecosystem and role of Data Engineering

## Summary and Highlights

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.

The goal of Data Engineering is to make quality data available for analytics and decision-making. And it does this by collecting raw source data, processing data so it becomes usable, storing data, and making quality data available to users securely.

## Quiz

### Practice Quiz

#### Question 1

Which emerging technology has made it possible for every enterprise to have access to limitless storage and high-performance computing?

- **Cloud Computing**
- Internet of Things
- Machine Learning
- Big Data

Cloud technologies has made it possible for every enterprise, regardless of its size, to have access to limitless storage and high-performance computing at nominal costs.

#### Question 2

Which of the data roles is responsible for extracting, integrating, and organizing data into data repositories?

- Data Scientist
- Business Intelligence Analyst
- **Data Engineer**
- Data Analyst

Data Engineers are responsible for extracting, integrating, and organizing data into data repositories.

#### Question 3

The field of data engineering concerns itself with the mechanics for the flow and access of data. Which one of the following statements captures the goal of data engineering?

- **Make quality data available for fact-finding and business decision-making**
- Architect data stores for the storage of processed data
- Design pipelines for extracting, transforming, and loading data into data repositories
- Maintain distributed systems for large-scale processing of data

Data engineering is the process of collecting raw data and converting it into analytics-ready data by cleaning, transforming, and preparing data so that it is reliable.

## Graded Quiz

### Question 1

A modern data ecosystem includes a network of continually evolving entities. It includes:

- **Data sources, enterprise data repository, business stakeholders, and tools, applications, and infrastructure to manage data**
- Data sources, databases, and programming languages
- Social media sources, data repositories, and APIs
- Data providers, databases, and programming languages

These are the key entities of a modern data ecosystem.

### Question 2

Data Engineers work within the data ecosystem to:

- Provide business intelligence solutions by monitoring data on different business functions
- Analyze data for deriving insights
- **Develop and maintain data architectures**
- Analyze data for actionable insights

One of the responsibilities of a Data Engineer in a data ecosystem is to develop and maintain data architectures so that data is available for business operations and analysis.

### Question 3

The goal of data engineering is to make quality data available for fact-finding and decision-making. Which one of these statements captures the process of data engineering?

- Processing data and making it available to users securely
- Collecting, processing, and storing data
- **Collecting, processing, storing, and making data available to users securely**
- Collecting, processing, and making data available to users securely

Data engineering includes the collection of data from disparate sources, processing data so that it is usable, storing processed data, and making it available to users securely.

### Question 4

Data extracted from disparate sources can be stored in:

- Databases only
- Data Lakes only
- **Databases, data warehouses, data lakes, or any other type of data repository**
- Data Warehouses only

Data extracted from multiple sources can be stored in any type of data repository, such as, databases, data warehouses, and data lakes.

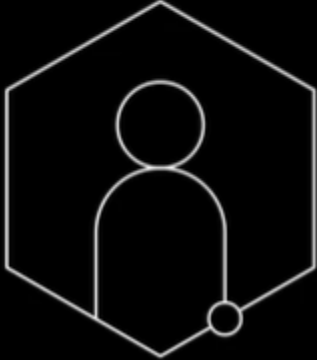
### Question 5

From the provided list, select the three emerging technologies that are shaping today's data ecosystem.

- Big Data, Internet of Things, and Dashboarding
- Machine Language, Cloud Computing, and Internet of Things
- **Cloud Computing, Machine Learning, and Big Data**
- Cloud Computing, Internet of Things, and Dashboarding

Emerging technologies such as Cloud Computing, Machine Learning, and Big Data are shaping today's data ecosystem and its possibilities.





# Responsibilities of a Data Engineer



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

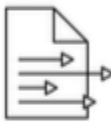

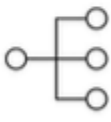
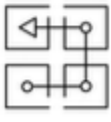
## Responsibilities of a Data Engineer

Data is analytics-ready when it is

-  Accurate
-  Reliable
-  Complies to Regulations
-  Accessible to consumers when they need it

## Responsibilities of a Data Engineer

At a broad level, Data Engineers:

-  Extract, organize, and integrate data from disparate sources
-  Prepare data for analysis and reporting by transforming and cleansing it
-  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  
Data Platforms, Data Stores, Distributed Systems, Data Repositories



# Technical Skills

## Operating Systems



UNIX



Linux



Windows  
Administrative Tools



System Utilities  
& Commands

## Infrastructure Components

Virtual Machines, Networking, Application Services, Cloud-based Services



## Databases and Data Warehouses

### RDBMS



IBM DB2, MySQL,  
Oracle Database,  
PostgreSQL

### NoSQL



Redis, MongoDB,  
Cassandra, Neo4J

### Data Warehouses



Oracle Exadata, IBM  
Db2 Warehouse on Cloud,  
IBM Netezza Performance  
Server, Amazon RedShift

## Data Pipelines



Apache Beam



AirFlow



DataFlow

ETL Tools



Languages



Query languages

SQL for relational databases and SQL-like query languages for NoSQL databases



Programming languages

Python, R, Java



Shell and Scripting languages

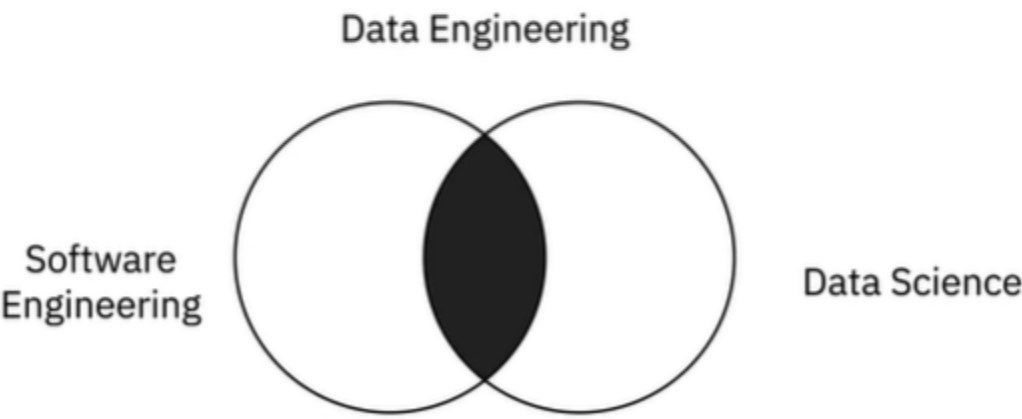
Unix/Linux Shell and PowerShell

Big Data Processing Tools



Functional Skills

Data Engineering is at the intersection of Software engineering and Data Science.



## Functional Skills of a Data Engineer:



Convert business requirements into technical specifications



Work with the complete software development lifecycle  
Ideation -> Architecture -> Design -> Prototyping -> Testing -> Deployment -> Monitoring



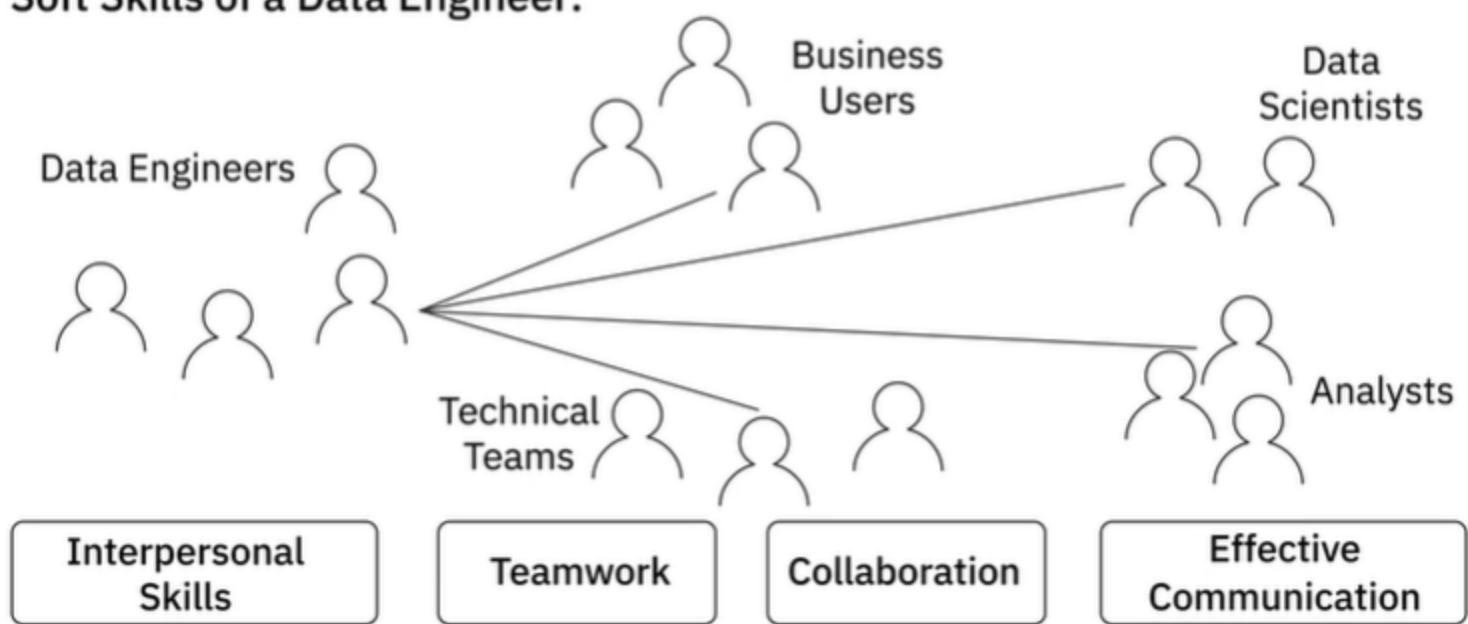
Understand data's potential application in business



Understand risks of poor data management  
Data Quality | Data Privacy | Security | Compliance

## Soft Skills

### Soft Skills of a Data Engineer:



### Conclusion

Data engineering requires a broad set of skillsets.

No one data engineer can possibly master each one of these skills, which means you essentially need to select one or more specialization areas, but have a good understanding of all areas so that you can make more informed decisions.

Your skills will grow over time with experience, the areas you choose to focus on, and the time you invest in upskilling yourself.

### Summary and Highlights

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.

## Quiz

### Practice Quiz

#### Question 1

Which one of these skills is essential to the role of a Data Engineer?

- Proficiency in Statistics
- To inspect analytics-ready data for deriving insights
- Proficiency in creating Deep Learning models
- **To setup and manage the infrastructure required for the ingestion, processing, and storage of data**

Data Engineers are responsible for setting up and managing the infrastructure required for ingesting raw data, processing it, and storing it so that it is available for analytics.

#### Question 2

What, according to Sarah Flinch, needs to be tracked and analyzed in order to keep business updated on the overall sentiment of the consumers?

- Blogging sites
- eCommerce platforms
- Social media sites
- **Social media posts, customer reviews and ratings on eCommerce platforms, and product reviews on blogging sites**

How a product gets talked about on social media, eCommerce platforms, and blogging sites has an immediate impact on sales numbers and brand perception.

## Graded Quiz

#### Question 1

Which one of these functional skills is essential to the role of a Data Engineer?

- Proficiency in working with ETL Tools
- Proficiency in Mathematics
- Inspect analytics-ready data for deriving insights
- **The ability to work with the software development lifecycle**

As a Data Engineer, you will be required to work through different phases of the software development lifecycle, which includes, ideation, architecture, design, prototyping, testing, deployment, and monitoring.

#### Question 2

Oracle Exadata, IBM Db2 Warehouse on Cloud, IBM Netezza Performance Server, and Amazon RedShift are some of the popular \_\_\_\_\_ in use today.

- NoSQL Databases
- **Data Warehouses**
- ETL Tools
- Big Data Platforms

These are some of the popularly used data warehouses.

### Question 3

Data Engineers manage the infrastructure required for the ingestion, processing, and storage of data.

- **True**
- False

This is one of the primary responsibilities of a Data Engineer.