UNIT – 1

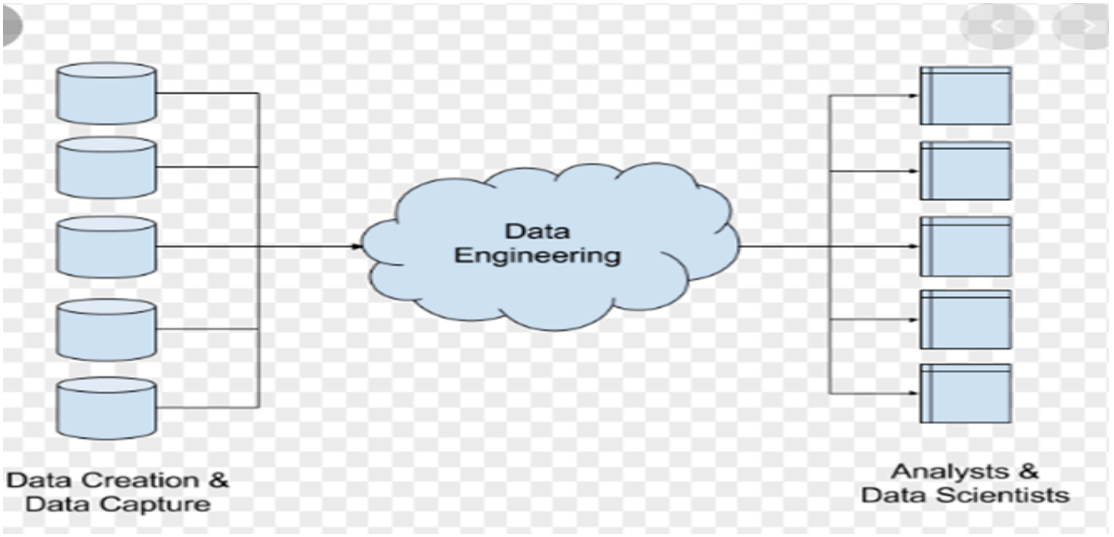
Data Engineering Defined

What Is a Data Engineer?

- Data engineering is the practice of designing and building systems for collecting, storing, and analyzing data at scale.
- It is a broad field with applications in just about every industry.
- Organizations have the ability to collect massive amounts of data, and they need the right people and technology to ensure it is in a highly usable state by the time it reaches data scientists and analysts.

What is data engineering with example?

- Data engineering **helps make data more useful and accessible for consumers of data**.
- For example, data stored in a relational database is managed as tables, like a Microsoft Excel spreadsheet.



What does a data engineer do?

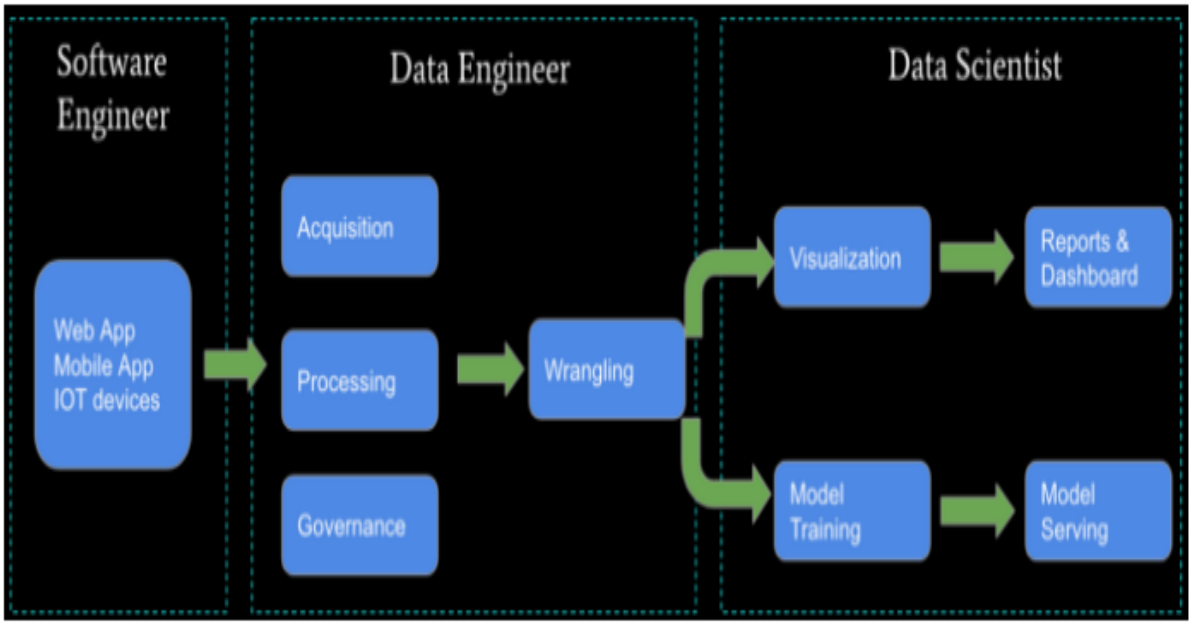
- Data engineers work in a variety of settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret.
- Their ultimate goal is to make data accessible so that organizations can use it to evaluate and optimize their performance.

Common tasks: working with data

- Acquire datasets that align with business needs
- Develop algorithms to transform data into useful, actionable information
- Build, test, and maintain database pipeline architectures

- ❖ Collaborate with management to understand company objectives
- ❖ Create new data validation methods and data analysis tools
- ❖ Ensure compliance with data governance and security policies

Diagram



Data Engineer: Acquisition

- ❖ Data acquisition **focuses on generated data and captures data into the system for processing**.
- ❖ Data should be processed in a real-time manner without any time delay.

Two types:

- Analog Data Acquisition Systems

- Digital Data Acquisition Systems

Data Engineer: Processing

- Data processing is **the method of collecting raw data and translating it into usable information**.
- It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization.
- **Example:** data processing are calculation of satellite orbits, weather forecasting, A stock trading software that converts millions of stock data into a simple graph.

Data Engineer: Data Governance

- Data governance (DG) is the process of managing the availability, usability, integrity and security of the data in enterprise systems, based on internal data standards and policies that also control data usage.
- Effective data governance ensures that data is consistent and trustworthy and doesn't get misused.

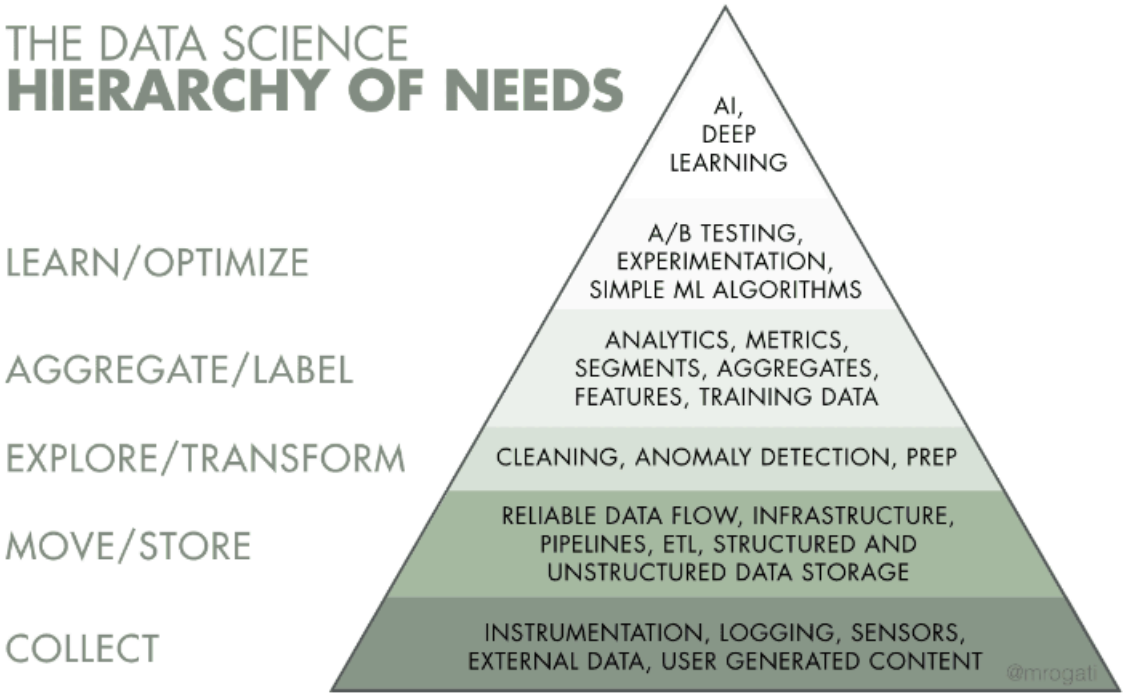
Data Engineer: Wrangling

- Data wrangling is the act of and mapping raw data into another format suitable for another purpose.

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

- With the amount of data and data sources rapidly growing and expanding, it is getting increasingly essential for large amounts of available data to be organized for analysis.

Overview of Data Engineering



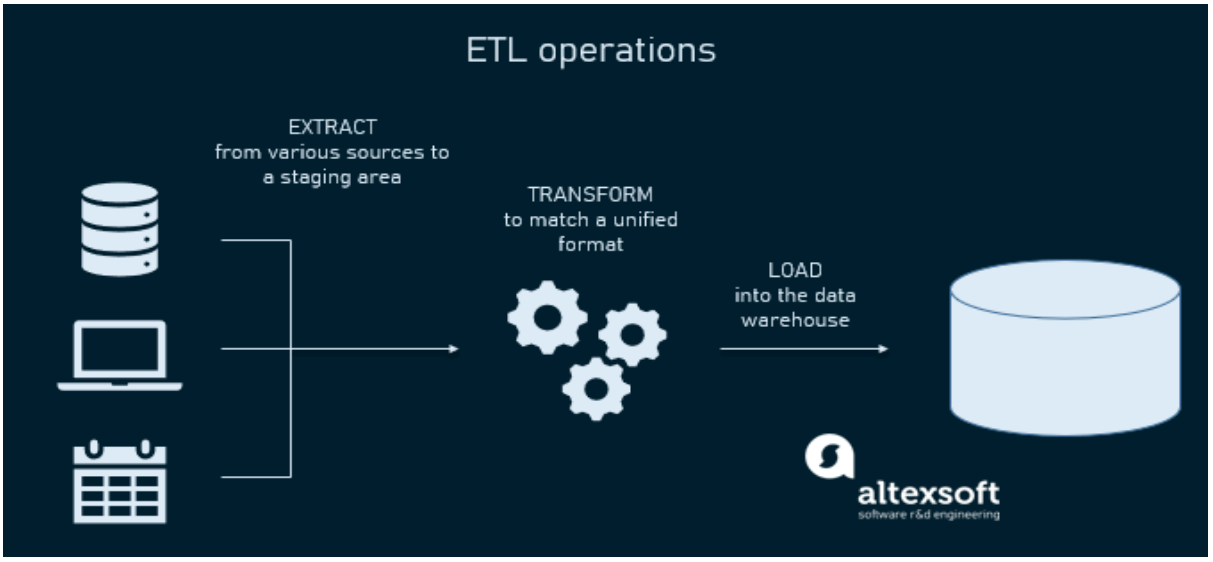
- **Data engineering** is a set of operations aimed at creating interfaces and mechanisms for the flow and access of information.

- It takes dedicated specialists – *data engineers* – to maintain data so that it remains available and usable by others.

- ❖ Data engineers set up and operate the organization's data infrastructure preparing it for further analysis by data analysts and scientists.

What is ETL?

1. **E**xtracting data from source databases,
2. **T**ransforming data to match a unified format for specific business purposes, and
3. **L**oading reformatted data to the storage (mainly, data warehouses).



1. Extract — retrieving incoming data. At the start of the pipeline, we're dealing with raw data from numerous separate sources. Data engineers write pieces of code – *jobs* – that run on a schedule extracting all the data gathered during a certain period.

2. Transform — standardizing data. Data from disparate sources is often inconsistent. So, for efficient querying and analysis, it must be modified. Having data extracted, engineers execute another set of jobs that transforms it to

meet the format requirements (e.g., units of measure, dates, attributes like color or size.) Data transformation is a critical function, as it significantly improves data discoverability and usability.

3. Load — saving data to a new destination.

After bringing data into a usable state, engineers can load it to the destination that typically is a relational database management system (RDBMS), a data warehouse, or Hadoop. Each destination has its specific practices to follow for performance and reliability.

Data pipeline challenges

Two major pitfalls in building data pipelines:

- ✔ lacking relevant metrics
- ✔ underestimating data load

3. Load — saving data to a new destination.

After bringing data into a usable state, engineers can load it to the destination that typically is a relational database management system (RDBMS), a data warehouse, or Hadoop. Each destination has its specific practices to follow for performance and reliability.

Data pipeline challenges

Two major pitfalls in building data pipelines:

- ✔ lacking relevant metrics

✔ underestimating data load

3. Load — saving data to a new destination.

After bringing data into a usable state, engineers can load it to the destination that typically is a relational database management system (RDBMS), a data warehouse, or Hadoop. Each destination has its specific practices to follow for performance and reliability.

Data pipeline challenges

Two major pitfalls in building data pipelines:

✔ lacking relevant metrics

✔ underestimating data load

ETL data pipeline

❖ A **Data pipeline** is basically a set of tools and processes for moving data from one system to another for storage and further handling.

❖ It captures datasets from multiple sources and inserts them into some form of database, another tool, or app, providing quick and reliable access to this combined data for the teams of data scientists, BI engineers, data analysts, etc.

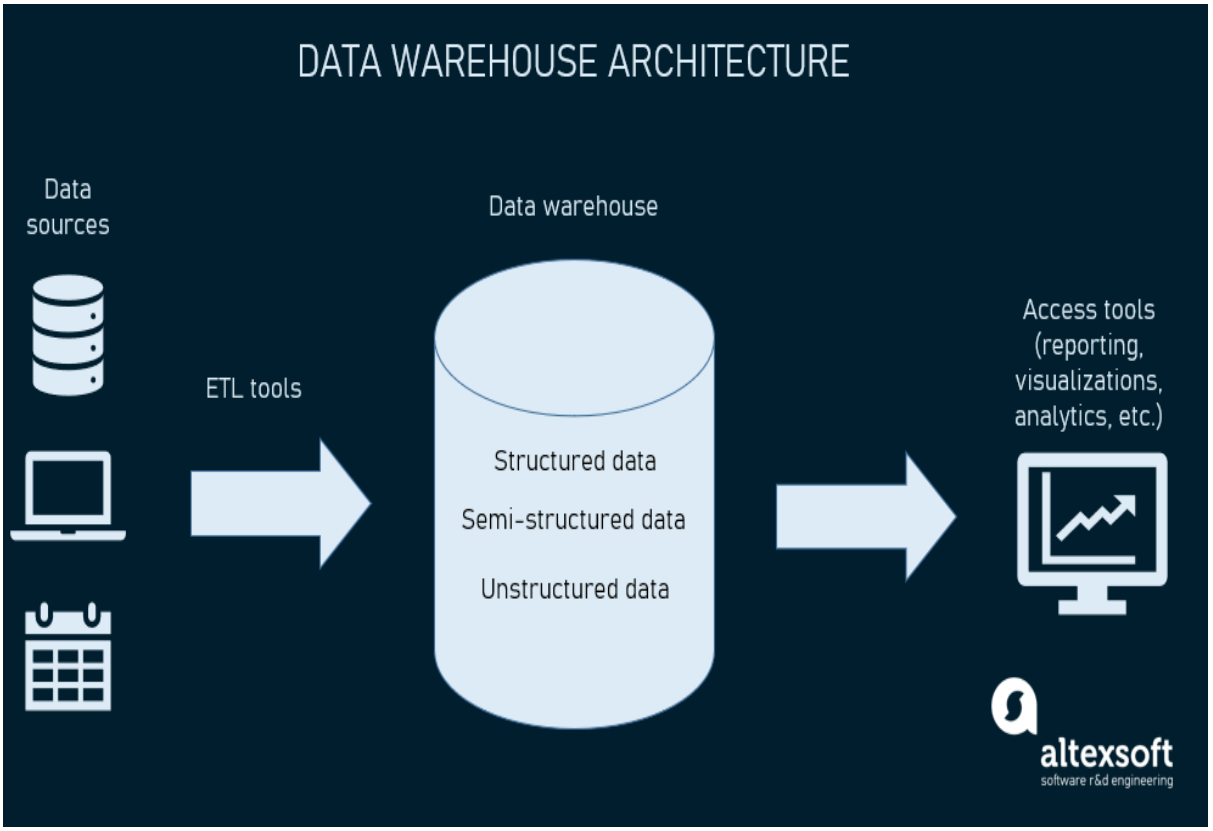
❖ Constructing data pipelines is the core responsibility of data engineering.

A data pipeline is commonly used for,

- moving data to the cloud or to a data warehouse,
- wrangling the data into a single location for convenience in machine learning projects,
- integrating data from various connected devices and systems in IoT,
- copying databases into a cloud data warehouse, and
- bringing data to one place in BI for informed business decisions.

Data warehouse

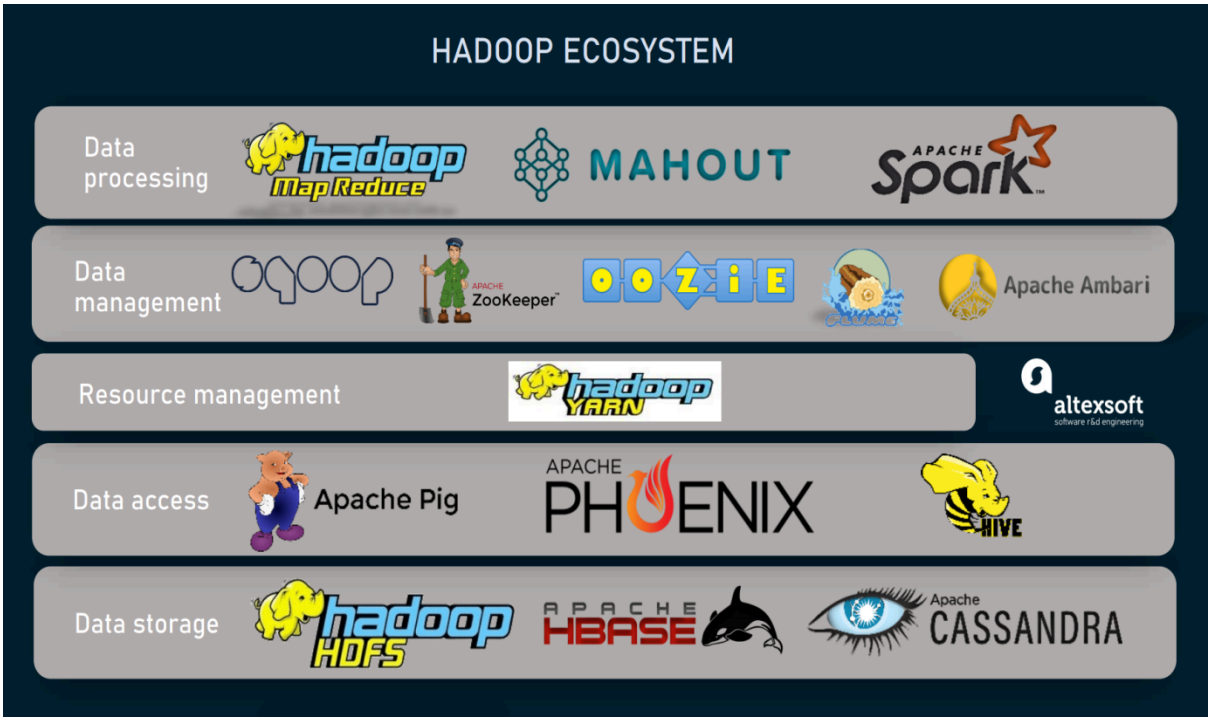
- A **data warehouse** (DW) is a central repository where data is stored in query-able forms.
- From a technical standpoint, a data warehouse is a relational database optimized for reading, aggregating, and querying large volumes of data.
- DWs can combine both structured and unstructured data where unstructured refers to a wide variety of forms (such as images, pdf files, audio formats, etc.) that are harder to categorize and process.



Hadoop

- ❖ Hadoop is a large-scale data processing framework based on Java.
- ❖ This software project is capable of structuring various big data types for further analysis.
- ❖ The platform allows for splitting data analysis jobs across various computers and processing them in parallel.

Hadoop Tools



- **Hadoop Distributed File System (HDFS).** Java-based big data storage system, HDFS includes two components: NameNode stores metadata while DataNode is responsible for actual data and performs operations according to NameNode.

- **MapReduce.** It's a framework for writing applications that process the data stored in HDFS. Parallel in nature, MapReduce programs are effective for performing big data analysis using multiple machines in the cluster.

- **YARN.** As the operating system of Hadoop, YARN helps manage and monitor workloads.

- **Hive.** A system for summarizing, querying and analyzing large datasets, Hive uses its own language – HQL- which is similar to SQL. HiveQL

automatically translates SQL-like queries into MapReduce jobs for execution on Hadoop.

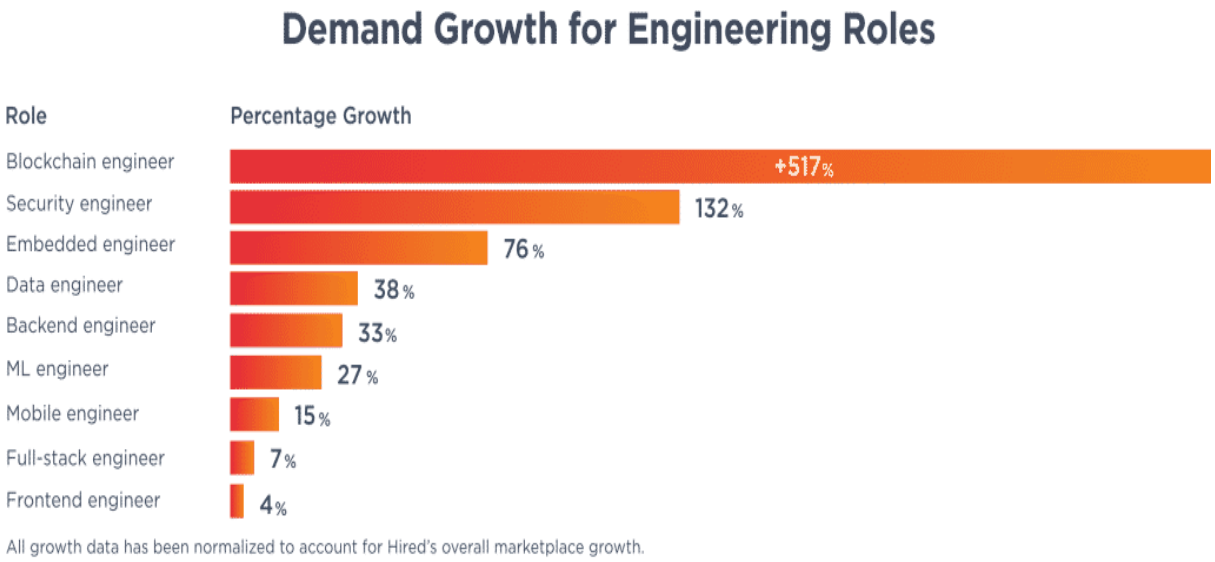
- **Pig.** Having similar goals with Hive, Pig also has its own language – PigLatin. When to use Pig and when to use Hive is the question. Pig is a better option for programming purposes, while Hive is mainly used by data analysts for creating reports.

- **HBase.** A NoSQL database built on top of HDFS that provides real-time access to read or write data.

Existence of Data Engineering

(Big) Data Engineers are in Demand

- Data Scientists as a professional group get a lot of attention and hype.



Why the Critical Need for Data Engineering Now?

- In 2017 that 85% of big data projects fail. This was largely due to a lack of reliable data infrastructures. Fast forward to 2019 and things had not improved.
- The CTO of IBM said that 87% of data science projects never make it into production.
- Gartner reiterated its prediction that now just 80% of projects would fail early days of big data analytics, Data Scientists were very often expected to build the necessary infrastructure and data pipelines to do their work.
- This was not necessarily in their skill sets or expectations for the job. The result was that data modelling would not be done correctly.
- There would be redundant work and inconsistency in the use of data among Data Scientists.
- These kinds of issues prevented companies from being able to extract optimal value from their data projects, so they failed.
- This is why will we continue to see the role of Data Engineers grow in importance and breadth.
- Companies need teams of people whose sole focus is to process data in a way that allows them to extract value from it.

Difference between Data Scientist and Data Engineer

1. **Data Engineers** collect relevant Data. They move and transform this Data into “pipelines” for the Data Science team. They could use programming languages such as Java, Scala, C++ or Python depending on their task.
2. **Data Scientists** analyze, test, aggregate, optimize the data and present it for the company.

What Skills do Data Engineers Need?

- **Foundation software engineering** – Agile, devOps, architecture design, service oriented architecture.
- **Distributed systems** – This would include software engineer skills and software architect skills.
- **Open Frameworks** – Apache Spark, Hadoop, perhaps Hive, MapReduce, Kafka and others…
- **SQL** – This is a database staple and remains that way.
- **Programming - *Python*** has become the favored language for working with data. ***Java*** on the other hand, while still widely sought has fallen out of favor with most data scientists and engineers. ***Scala*** is another language that Apache Spark and Kafka are based on.
- **Pandas** – a Python library for cleaning and manipulating data.

- **Cloud platforms** – AWS is probably the most prevalent cloud skill set for Data Engineers to know. Google Cloud Data Engineering and Microsoft Azure are right behind.

- **Analytics** – While mainly the realm of data scientists, statistical analysis skills or understanding of some of the different mathematical principles or probabilistic principles are necessary for being able to properly manipulate the data so that it is in a shape that is accessible for the people who are doing the end analysis on it.

- **Data modelling** – Data modelling knowledge is quite important now in the sense that a Data Engineer needs to know how they are going to **structure tables, partitions**, where to ***normalize and denormalize*** data in the warehouse, etc. and ***how to think about retrieving certain attributes***.

Need for Data Engineering

What is the need for data engineering?

- Data engineering is important because it allows businesses to optimize data towards usability.

What Is Data Engineering?

- Data engineering, sometimes called information engineering, is a software approach to developing information systems.

- Data engineering encompasses sourcing, transforming, and managing data from various systems.
- This process ensures that data is useful and accessible.
- data engineering emphasizes the practical applications of data collection and analysis.
- data engineering employs intricate methodologies for gathering and authenticating data that range from data integration tool to artificial intelligence.

Why Is Data Engineering Important?

Data engineering is important because it allows businesses to optimize data towards usability. For example, data engineering plays a large role in the following pursuits:

- **Finding** the best practices for refining your **software development life cycle**
- Tightening information security and **protecting your business from cyber attacks**
- Increasing your understanding of business domain knowledge
- Bringing data together into one place via data integration tools

Essential data engineer skills

- **Coding-** Coding is a highly valued skill that is a requirement for a majority of data engineering positions.

- **Data warehousing** - Data engineers are charged with storing and analyzing an incredible amount of data.
- **Knowledge of operating systems** - As a data engineer, possessing an intimate understanding of operating systems like Apple macOS, Microsoft Windows, Linux, Solaris and Unix is vital
- **Database systems** - Data engineers should have a deep understanding of database management
- **Data analysis** - Most employers expect data engineer candidates to have a strong understanding of analytics software, specifically Apache Hadoop-based solutions like MapReduce, Hive, Pig and HBase.
- **Critical thinking skills** - Data engineers need to be able to evaluate issues and then develop solutions that are both creative and effective
- **Basic understanding of machine learning** - machine learning is primarily the focus of data scientists, it can be helpful for data engineers to have at least a basic understanding of using this type of data.
- **Communication skills** - As a data engineer, you have to be able to collaborate with colleagues with and without technical expertise, which is why possessing great communication skills is so important

Benefits of Data Engineering

Top 4 Benefits of Data Engineering

- **Helping Make Better Decisions**: Companies may leverage data-driven insights to better influence their decisions, resulting in improved outcomes.
- **Checking the Outcomes of Decisions**: Self-improvement is an on-going process in data science. This results in reflecting the impact of prior decisions. Without self-reflection, no process is complete

- **Predicting the User Story to Improve the User Experience Predictors are one of the most powerful aspects of machine learning**: Organizations may now develop procedures to track consumer feedback, product success, and what their competitors are doing with so much data to work with.
- **New Business Opportunities Identification**:: Companies can stay competitive if they can anticipate what the market wants and deliver the product before it is needed. In today's economy, a company can no longer rely on instinct to be competitive

Data Engineering Vs Data Science

The main difference between these two data professionals is that data engineers build and maintain the systems and structures that store, extract, and organize data, while data scientists analyze that data to predict trends, glean business insights, and answer questions that are relevant to the organization.

Role and Responsibilities

It helps to think of data engineers and data scientists as having complementary roles. Data engineers build and optimize the systems that allow data scientists to do their job. Data scientists, meanwhile, find meaning in the troves of data that data engineers manage.

What Does a Data Engineer Do?

A data engineer is a data professional who prepares the data infrastructure for analysis. They are focused on the production readiness of raw data and elements such as formats, resilience, scaling, data storage, and security. Data engineers are tasked with designing, building, testing, integrating, managing, and optimizing data from a variety of sources. They also build the infrastructure and architectures that enable data generation.

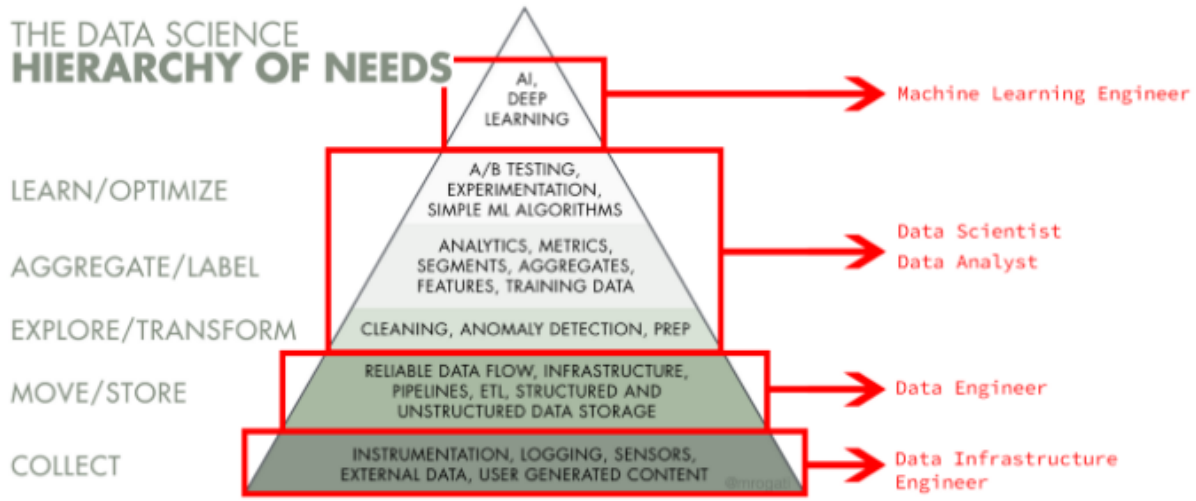
What Does a Data Scientist Do?

Data scientists concentrate on finding new insights from the data that was prepared for them by data engineers. As part of their job, they conduct online experiments, develop hypotheses, and use their knowledge of statistics, data analytics, data visualization, and machine learning algorithms to identify trends and create forecasts for the business.

Education and Requirements

I — what are the differences between a Data Engineer and a Data Scientist?

1- Understand the hierarchy of the Data Process.

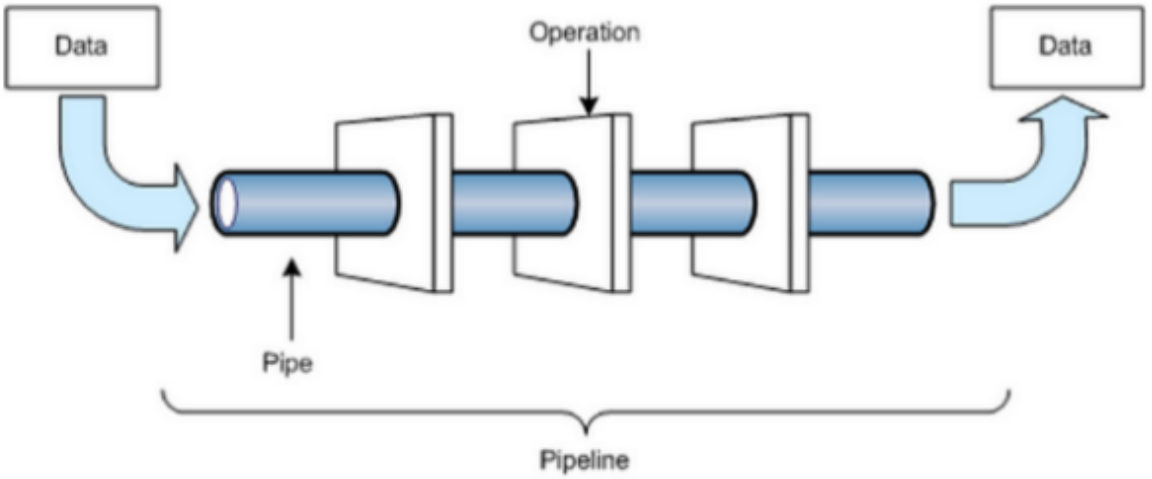


Data Engineers collect relevant Data. They move and transform this Data into “pipelines” for the Data Science team. They could use programming languages such as **Java, Scala, C++ or Python** depending on their task.

Data Scientists analyze, test, aggregate, optimize the data and present it for the company.

2 — Data Engineer — The technical part of data — Design — Build — Arange.

Data Engineers are specialized in 3 main data actions: to design, build and arrange Data "pipelines**".** *Data pipelines are sequences of processing and analysis steps applied to data for a specific purpose.*



What tasks have a Data Engineer in a company?

- Design the big data infrastructure and prepare it to be analyzed.
- Build complex queries to create "pipelines".
- Arrange any problems in the programmed system.

What competencies wait from a Data Engineer?

- Logical mind
- Knowing what data to extract
- Management and organizational skills

- Working with cross-functional teams

3 — Data Scientist — Analyse — test — create — present

- ❖ Data Scientists have normally 4 main tasks in a company.
- ❖ He analyses, tests, creates and presents them to the team.
- ❖ Data Scientists have a math and statistical background.
- ❖ They are also comfortable with creating machine learning and artificial intelligence models.

What tasks have a Data Scientist in a company?

- Work on clean data
- Find solutions with the data available
- Communicate analyzes with the team
- Work onto solution problem and get some

What competencies wait from a Data Engineer?

- Good communication skills.
- Good analysis.
- Good hypothesis.
- Broad knowledge in different techniques in machine learning, data mining, statistics, and big data infrastructures.

- Be a problem solver.

4 — What about job openings and salaries in all that?

- ❖ The number of job openings for data engineers is almost five times higher than the number of job openings for data scientists.
- ❖ This makes sense as most organizations need more data engineers than data scientists on their team