**Data Flow**

A screenshot of a computer

Description automatically generated

Data Analyst

Data Scientist

Data Engineer

# **Data Engineer**

* Builds and maintains the systems that Data Scientists and Data Analysts use to collect, store, and analyze data.
* Skills to design and build data pipelines, and to ensure that data is stored in a secure and efficient way.
* Main goal is to build a scalable infrastructure that makes data accessible to anyone who needs it.

**Responsibilities**

* + Data Mining for getting insights from data

\*Data mining involves discovering patterns, trends, and valuable information from large datasets

\*A data engineer might analyze customer purchase history to identify trends and patterns, helping the marketing team understand customer preferences and optimize their campaigns.

* + Conversion of erroneous data into a useable form for data analysis

\*Clean and preprocess data to ensure it is accurate and ready for analysis. This involves handling missing values, correcting errors, and transforming data into a standardized format.

(E.g., duplicate records, correcting typos, filling missing values in sales dataset)

* + Writing queries on data

\* Use query languages (e.g., SQL) to extract, filter, and manipulate data from databases. This skill is crucial for retrieving specific information needed for analysis.

* + Develop large data warehouses with the help of Extract, Transform, Load (ETL)

\* ETL is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.

**Extract**

* Raw data is exported from source locations to a staging area
* Data management teams can extract data from a variety of data sources (E.g. SQL, NoSQL, CRM, ERP, Flat files, Email, Web pages)

**Transform**

* Filtering, cleansing, de-duplicating, validating, and authenticating the data.
* Performing calculations, translations, or summarizations based on the raw data.
* Formatting the data into tables or joined tables to match the schema of the target data warehouse.

**Load**

* The transformed data is moved from the staging area into a target data warehouse.
* Initial loading of all data, followed by periodic loading of incremental data changes.
  + Storing data in a secure and efficient way
  + Maintaining data infrastructure
  + Constructing testing and maintaining the architecture implementing new ways to improve the data quality, building data pipeline, creating, and integrating APIs

**Skills**

* Deep understanding of scripting and programming languages such as Java, SQL, SAS, Python etc.
* Handling frameworks as Hadoop, Hive, Apache Spark, Pig, NoSQL, and Data Streaming etc.

# **Data Analyst**

* Pull analyze and interpret data to extract meaningful, actionable insights to solve business critical problems.
* Clean and conversion data
* Can create dashboards and report that can use to drive business actions for stakeholders.
* Translating numeric data into a form that can be understood by everyone in an organization
* Translated data will used by upper management to make informed business decisions.

**Responsibilities**

* + Analyze data from data warehouse which prepared by data engineer
  + Cleaning, transforming, and analysing data insight, performing exploratory,
  + Use basic algorithms like logistic regression, linear regression and so on to analyzing data and identify patterns
  + Creating reports and visualizations like dashboards

**Skills**

* Expertise in statistic tools such as Microsoft Excel, SQL, SAS Miner, SPSS (Statistical Package for the Social Sciences)
* Fundamentals of data handling
* Reporting
* Visualization skills using BI tool (Business Intelligence Tool) like Tableau or Power BI
* Great communication skill as needs to be able to connect with both technical and non-technical audiences
* Additional: Programming languages such as Python, R, and JavaScript will be plus point

# **Data Scientist**

* Senior most in team and have deep expertise in machine learning, statistics, and data handling.
* Knowledge of statistics, machine learning, and programming to extract meaning from data.
* Skills to solve complex problems, identify trends, and make predictions.

**Responsibilities**

* + Develop machine learning algorithms
  + Building deep learning models
  + Building and testing machine learning models that can operate on Big Data
  + Understand and interpret Big Data analysis
  + Analyzing data to identify trends
  + Making predictions based on data
  + Communicating their findings to stakeholders
  + Deliver results that have an impact on business outcomes

**Skills**

* Data techniques such as clustering, neural networks, decision trees, and the like for deriving business insights.
* Deep understanding of programming languages such as Python, Java, SQL, R, SAS etc.
* Knowledge of Big Data Framework like Hadoop, Apache Spark, and Pig.