**DATA ENGINEERING**

4 general steps through which data flows:

1. Collect and ingest data, from web traffic, surveys, or media consumption.
2. Prepare it, which includes cleaning data e.g. finding missing or duplicate values and converting data into a more organized format.
3. Data is exploited. We explore it, visualize it, build dashboards to track changes. Or Compare two sets data.
4. Run Experiments like which article title gets the most hits, or to build models for example, to forecast Stock prices.

**Data Engineers are Responsible for first step of the process:**

***“INGESTING COLLECTED DATA AND STORING IT”***

**Data for Engineers lay the ground work. Data analysts, Data Scientists and ML Engineers.**

**Data Engineers Deliver:**

1. The correct Data
2. IN The right from form.
3. To the right people
4. As Efficiently as possible.

**DE IS Responsible for:**

1. They Ingest data from different sources.
2. Optimize databases for analysis.
3. Remove Corrupted data.
4. Develop, construct, test and maintain data architectures. Buch as databases and large-scale processing systems to process and handle massive amounts of data.

**DATA WAREHOUSE**

A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence.

**How does a data warehouse work?**

A data warehouse may contain multiple databases. Within each database, data is organized into tables and columns. Within each column, you can define a description of the data, such as integer, data field, or string. Tables can be organized inside of schemas, which you can think of as folders. When data is ingested, it is stored in various tables described by the schema. Query tools use the schema to determine which data tables to access and analyze.

**Benefits of a data warehouse include the following:**

* Informed decision making
* Consolidated data from many sources
* Historical data analysis
* Data quality, consistency, and accuracy
* Separation of analytics processing from transactional databases, which improves performance of both systems

**Data warehouse vs database**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | | **Characteristics** | **Data Warehouse** | **Transactional Database** | | Suitable workloads | Analytics, reporting, big data | Transaction processing | | Data source | Data collected and normalized from many sources | Data captured as-is from a single source, such as a transactional system | | Data capture | Bulk write operations typically on a predetermined batch schedule | Optimized for continuous write operations as new data is available to maximize transaction throughput | | Data normalization | Denormalized schemas, such as the Star schema or Snowflake schema | Highly normalized, static schemas | | Data storage | Optimized for simplicity of access and high-speed query performance using columnar storage | Optimized for high throughout write operations to a single row-oriented physical block | | Data access | Optimized to minimize I/O and maximize data throughput | High volumes of small read operations | |  |