Data Engineering for Everyone - Lexicon

Data Engineering for Everyone - Lexicon

- **Airflow**: an open-source workflow management platform used to schedule data engineering tasks. Started at Airbnb, now maintained by the Apache foundation.
- **AWS**: Amazon Web Services. Amazon's cloud computing services.
- **Azure**: Microsoft's cloud services.
- **Big data**: the systematic storage, management and analysis of datasets that are too large or complex to be dealt with by traditional data-processing application software. Big Data revolves around 4 Vs: volume, variety, velocity, and veracity.
- **Cloud computing**: the use of a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
- **Database schema**: the skeleton structure representing the logical view of the entire database: it defines how the data is organized, how the relations among them are associated, and formulates all the constraints that are to be applied on the data.
- **Data catalog**: a data catalog is a metadata management tool that companies use to inventory and organize the data within their systems. Typical benefits include improvements to data discovery, governance, and access.
- **Data engineering**: the process of transforming data into a format suited for analysis.
- **Data ingestion**: the process of obtaining and importing data for immediate use or storage in a database.
- **Data lake**: a repository of data stored in its natural/raw format.
- **Data pipeline**: a series of data processing steps. They are built and maintained by Data Engineers, and used by Data Scientists and Analysts.
- **Data processing**: the collection and manipulation of items of data to produce meaningful information.
- **Data science**: the process of extracting knowledge from data.
- **Data warehouse**: a central repository of integrated data from one or more disparate sources.
- **ETL**: Extract, Transform, Load, the process of pulling data from a database to move it to another database.
- **Google Cloud**: Google's cloud services.
- **Luigi**: an open-source workflow management platform used to schedule data engineering tasks.
- **NoSQL**: Non-SQL. A NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.
- **Parallel computing**: simultaneous, rather than sequential, use of multiple compute resources to solve a computational problem. Several instructions are executed concurrently on different

processors, rather than sequentially on one processor. An overall control/coordination mechanism is employed.

- **Query**: a request for information from a database.
- **Redshift**: Amazon's data warehouse service.
- **S3**: Amazon's object storage service.
- **Scheduling**: the glue of a data engineering system, holding each small piece together and organizing how they work together, by running jobs in a specific order and resolving all dependencies correctly.
- **SQL**: Structured Query Language, the standard language to communicate with relational database management systems.
- **Structured data**: data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis.
- **Unstructured data**: information that either does not have a pre-defined data model or is not organized in a pre-defined manner.
- **View**: output of a stored, frequent query on the data