DATA ENGINEERING BASICS

MARCH 14, 2023

BYTEWISE LIMITED

Contents

1. Big data

- Big data is combination of **structured, unstructured and semi structured data collected from forms or organizations, predictive modeling and other use.**
- **Big data can come from business transaction systems, customer databases social networks etc.**
- Big data refers to massive, complex data sets that are rapidly generated and transmitted from a wide variety of sources.
- We generate data whenever we go online, when we carry our GPS-equipped smartphones, when we communicate with our friends through social media or chat applications, and when we shop.
- You could say we leave digital footprints with everything we do that involves a digital action, which is almost everything. On top of this, the amount of machine-generated data is rapidly growing too.
- Data is generated and shared when our "smart" home devices communicate with each other or with their home servers. Industrial machinery in plants and factories around the world are increasingly equipped with sensors that gather and transmit data.
- The term "Big Data" refers to the collection of all this data
- Big Data projects often use cutting-edge analytics involving artificial intelligence and machine learning. Big data is stored in petabytes and zettabytes.

- Improving healthcare.
- Predicting and responding to natural and man-made disasters
- Preventing crime

5V's:

- ➢ **Volume: amount of data generated**
- ➢ **Velocity: speed at which data is generated**
- ➢ **Veracity: accuracy and trustworthiness**
- ➢ **Value: benefits of data to company**
- ➢ **Variety: structured, unstructured and semi structured data**

2. Data Lake

- Fundamentally, a data lake holds data in its rawest form, without the need for it to have been processed or analyzed. Data lakes.
- They allow you to dump data in its original format and can become a sandbox in which analysts and developers can play.
- A data lake is a storage repository that can rapidly ingest large amounts of raw data in its native format.
- As a result, business users can quickly access it whenever needed and data scientists can apply analytics to get insights. Unlike its older cousin – the data warehouse – a data lake is ideal for storing unstructured big data like tweets, images, voice and streaming data. But it can be used to store all types of data – any source, any size, any speed, and any structure.
- The term "data lake" refers to the ad hoc nature of data in a data lake, as opposed to the clean and processed data stored in traditional data warehouse systems.
- A data lake is an unstructured repository of unprocessed data, stored without organization or hierarchy.
- A data lake works on a principle called schema on read. This means that there is no predefined schema into which data needs to be fitted before storage.

- **Flexibility**, as data scientists can quickly and easily configure queries
- **Accessibility**, as all users can access all data
- **Affordability**, as many data lake technologies are open source

3. Database

- A database is information that is set up for easy access, management and updating.
- Databases are used for storing, maintaining and accessing any sort of data. They collect information on people, places or things.
- That information is gathered in one place so that it can be observed and analyzed. Databases can be thought of as an organized collection of information.
- A database is an organized collection of structured information, or data, typically stored electronically in a computer system. A database is usually controlled by a database management system (DBMS).
- Databases, on the other hand, are designed to hold much larger collections of organized information — massive amounts, sometimes.
- Databases allow multiple users at the same time to quickly and securely access and query the data using highly complex logic and language.

4. Data warehouse

- Data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose.

- A well-designed data warehouse will perform queries very quickly, deliver high data throughput, and provide enough flexibility for end users to “slice and dice” or reduce the volume of data for closer examination to meet a variety of demands—whether at a high level or at a very fine, detailed level.
- The data warehouse serves as the functional foundation for middleware BI environments that provide end users with reports, dashboards, and other interfaces.
- A data warehouse is the storage of information over time by a business or other organization.
- New data is periodically added by people in various key departments such as marketing and sales.
- The warehouse becomes a library of historical data that can be retrieved and analyzed in order to inform decision-making in the business.
- The key factors in building an effective data warehouse include defining the information that is critical to the organization and identifying the sources of the information.
- A data warehouse is designed to allow its users to run queries and analyses on historical data derived from transactional sources.
- Data added to the warehouse does not change and cannot be altered. The warehouse is the source that is used to run analytics on past events
- A data warehouse is a computer system designed to store and analyze large amounts of structured or semi-structured data.
- Data engineers and scientists, business analysts and decision-makers access the data using BI tools,