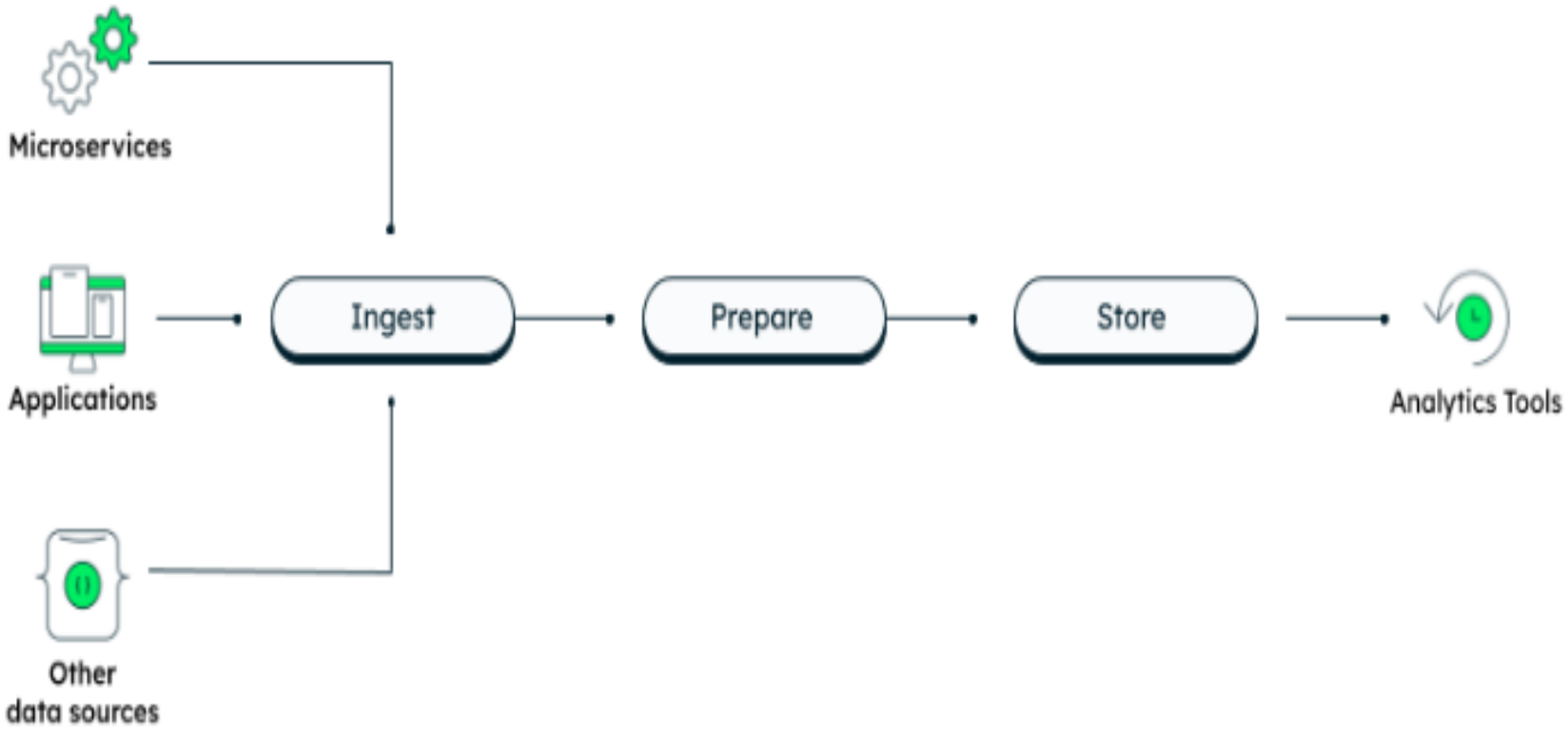
Big Databases

By Mrs. Ankita Joshi

Introduction

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.
- Big data databases rapidly ingest, prepare, and store large amounts of diverse data. They are responsible for converting unstructured and semi-structured data into a format that analytics tools can use. Because of these distinctive requirements, NoSQL (non-relational) databases, such as MongoDB, are a powerful choice for storing big data.

Big Data


Microservices
Applications
Other
data sources
Ingest
Prepare
Store
Analytics Tools

Example of big data

- The **New York Stock Exchange** is an example of Big Data that generates about ***one terabyte*** of new trade data per day.
- The statistic shows that ***500+ terabytes*** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.
- A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time. With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

Types Of Big Data

- **Structured**
- Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data.
- Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.
- However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes(*one billion terabytes* forms *a zettabyte*).

Unstructured

- Any data with unknown form or the structure is classified as unstructured data.
- In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don’t know how to derive value out of it since this data is in its raw form or unstructured format.

Semi-structured

- Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.
- Example of semi-structured data is a data represented in an XML file.

Four V's of big data

- Big data can be described by the following characteristics:
  - Volume
  - Variety
  - Velocity
  - Variability

Volume

- The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence, **‘Volume’** is one characteristic which needs to be considered while dealing with Big Data solutions.
- Just one cross-country airline trip can generate 240 terabytes of flight data.
- IoT sensors on a single factory shop floor can produce thousands of simultaneous data feeds every day.
- Other common examples of big data are Twitter data feeds, webpage clickstreams, and mobile apps.

Variety

- Big data comes in many forms, such as text, audio, video, geospatial, and 3D, none of which can be addressed by highly formatted traditional relational databases. These older systems were designed for smaller volumes of structured data and to run on just a single server, imposing real limitations on speed and capacity.
- Modern big data databases such as MongoDB are engineered to readily accommodate the need for variety—not just multiple data types, but a wide range of enabling infrastructure, including scale-out storage architecture and concurrent processing environments.

Velocity

- The term **‘velocity’** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.
- Accordingly, stock-trading software is designed to log market changes within microseconds. Internet-enabled games serve millions of users simultaneously, each of them generating several actions every second. And IoT devices stream enormous quantities of event data in real time.

Variability

- This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.