How is MongoDB used for Big Data and what are the associated data governance and data quality issues?

W1856656 - Sohag Noman

What is MongoDB?

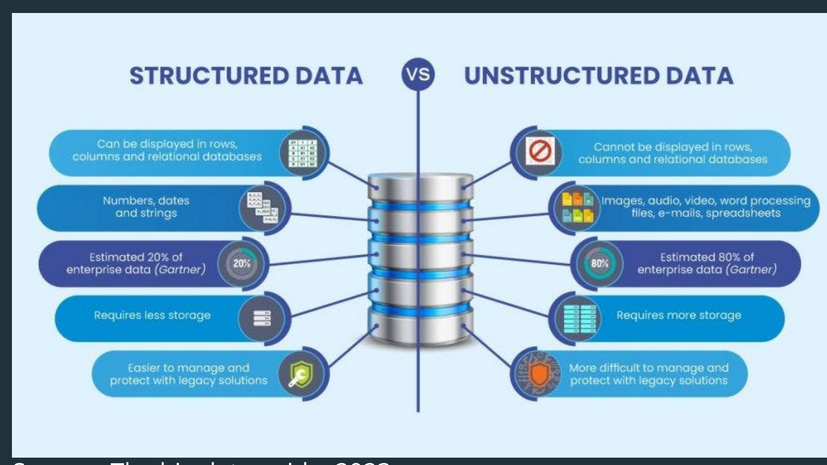
MongoDB is a NoSQL database that is commonly used in Big Data applications due to its flexibility, scalability, and ability to handle large volumes of unstructured or semi-structured data. MongoDB's features and capabilities make it suitable for Big Data use cases in various industries, such as e-commerce, social media, IoT, and more.

What is Big Data?

Although there are many definitions of big data. The primary idea behind big data is that everything we do nowadays leaves a digital footprint or traces (i.e. data) which can be used for analysis.

Source : European Commission 2020

Type of data



Source : The big data guide 2023

Characteristics of big data:

- ***Volume*** - Volume is the size of Big Data involved, covering from gigabytes to terabytes or more.
- ***Velocity*** - The large volume of big data being generated and collected means data processing has to occur with high speed and low latency.
- ***Variety*** - Big data covers a range of formats outside normal financial or transactional data, including text, audio, video, geographical, and 3D – none of which can be addressed by a traditional relational database.

Problems with big data:

- Big data is not manually organised
- It can be time-consuming to process a big volume of data
- It can be hard to find patterns or valuable insights into the data without processing it.

How is MongoDB used for big data?

MongoDB is a proposed big data choice because of its ability to easily handle a broad variety of data types, support for real-time analysis, high-speed data intake, low-latency performance, flexible data model, easy horizontal scale-out, and sophisticated query language.

How is MongoDB helps?

MongoDB stores data in a JSON document format. This allows us to dump whatever data we want, without bothering about the relations of the data. The fact that it is in a JSON format, also implies that information can be obtained very quickly. Things that would take minutes for relational databases to find can be found with MongoDB in mere seconds. Basically, in any circumstance where the data isn't relational and needs to be handled quickly, MongoDB is the most ideal database.

Big Data challenges and quality issues

1. Security
Big data is frequently housed in a centralized data lake. Robust security is necessary to ensure your data stays safe from attack and theft. With its large quantity of valuable, confidential information, big data environments are especially attractive for hackers and cybercriminals. With that in mind, it's important to build a robust security system at an early stage of architecture planning.

2. Complexity
Big Data generally includes numerous data sources with independent data-ingestion components. Building, testing, and troubleshooting Big Data processes are tasks that need high degrees of knowledge and ability.

3. Evolving technologies
It's crucial to choose the proper solutions and components to satisfy the business objectives. Many Big Data technologies, techniques, and standards are relatively new and currently in the process of evolution. Core Hadoop components such as Hive and Pig have reached a level of stability, but other technologies and services remain immature and are expected to change over time.

4. Specialized skill sets
Big Data APIs built on popular languages are gradually coming into use. Nevertheless, Big Data architectures and solutions do often involve unconventional, highly specialized languages and frameworks that demand a steep learning curve for developers and data analysts alike.

Big data governance

Governance is about validating data: making sure records reconcile and that they are usable, accurate, and secure. However, the integration of many sources can make this process complex, and reconciling data from disparate systems that should be agreed upon is a necessary but potentially challenging effort.
Typically, organizations have an internal body responsible for writing governance policies and processes. There should be a good investment also in data management systems, which are required for data cleansing, integration, quality assurance, and integrity management. For this reason, it's crucial to examine at which stage the organization is in its big data journey and how to align its data governance approach with best practices for your industry.

References

-Big data: definition, benefits, challenges (infographics) | News | European Parliament. (2023, March 16)
-The Big Data Guide. (n.d.). MongoDB. https://www.mongodb.com/basics/big-data-explained
-L. (2017, April 11). Why MongoDB is so important for big data | learntek.org. Why MongoDB Is so Important for Big Data | learntek.org. https://www.learntek.org/blog/mongodb-important-big-data/
-Henry, O. B. (2019, October 23). Factors to Consider When Choosing MongoDB for Big Data Applications | Severalnines. Severalnines. https://severalnines.com/blog/factors-consider-when-choosing-mongodb-big-data-applications/
-What Is Big Data Architecture? (n.d.). MongoDB. https://www.mongodb.com/big-data-explained/architecture
-An introduction to MongoDB. (n.d.). An Introduction to MongoDB | PPT. https://www.slideshare.net/CesarTrigo/an-introduction-to-mongodb-36429852