All things you need to know about Big Data

What is Big Data? Let's TipsMake.com find out what you need to know about Big Data in this article!

1. 10 tips for businesses before deciding to invest in Big Data
2. Is the data structure and algorithm necessary for a Web Developer?
3. Top 10 basic network troubleshooting tools that IT people need to know

Analyzing a lot of data is only part of the Big Data analysis and previous data analysis. Let's TipsMake.com find out **what you need to know about Big Data** in this article!

What is the difference between data (data) and Big Data (big data)?

What is Big Data?

Big Data is a term used to refer to a very large and complex set of data that traditional data processing applications and tools cannot collect, manage, and process data in a reasonable amount of time. physical.

These large data sets may include structured data, unstructured data, and semistructured data, each with a little difference.

In fact, how much data is enough to call 'big' is still controversial, but it can be multiples of petabytes - and with the largest projects within exabyte (multiples of bytes) .

Usually, Big Data has 3 features:

1. Huge amount of data;
2. Various types of data;
3. The speed at which data needs to be processed and analyzed.

Data that makes up large data warehouses can come from sources including websites, social media, desktop applications, mobile applications, and scientific experiments. , increasing sensor devices and other devices in Internet connected devices (IoT- internet of things).

The concept of Big Data comes with related components that allow organizations to put data into practical use and solve some business problems, including the IT infrastructure needed to support Big Data. , the analysis applies to the data and technology needed for Big Data projects, related skills sets and practical cases that make sense for Big Data.

Big Data and Analytics

What really brings value from all the big data organizations collecting is data analysis (Analytics). If not analyzed, it is just a bunch of data with limited use in business.

By applying analytics to big data, companies can see benefits such as increased revenue, improved customer service, higher efficiency and increased competitiveness.

Data analysis involves examining the data set to gather detailed information or draw conclusions about what is included in it, such as trends and predictions about future activity.

By analyzing data, organizations can make better business decisions such as time and place to run marketing campaigns or introduce new products or services.

The analysis can refer to more intelligent or advanced business applications, predictive analysis such as those used by scientific organizations. The most advanced type of data analysis is **data mining** , where analysts evaluate large datasets to determine relationships, patterns and trends.

Data analysis may include analysis of exploration data (to identify patterns and relationships in data) and analysis of validation data (application of statistical techniques to find assumptions about the set). whether the data is correct or not.

Another area is quantitative data analysis (or analysis of digital data with statistically comparable variables) compared to qualitative data analysis (focusing on non-personal data such as videos, pictures and text).

IT infrastructure supports Big Data

For the Big Data concept at work, organizations need to have the infrastructure to collect and store data, provide access and ensure information while storing and transporting.

At a high level, including storage systems and servers designed for Big Data, software for data management and integration, business intelligence software (business intelligence) and data analysis, applications Use Big Data.

Much of this infrastructure will be in place because companies want to continue to make use of their data center investments. However, more and more organizations rely on cloud computing services to handle many of their large data requests.

Data collection requires source. Many of the following applications, such as web apps, social media channels, mobile apps and built-in email archives. But when IoT becomes more popular, companies may need to deploy sensors on all devices, vehicles and products to collect data, as well as new applications that create user data. . (IoT-oriented data analysis has its specific techniques and tools.)

In order to store all incoming data, organizations need to have enough storage on site. Storage options include traditional data warehousing, data lake (huge raw data archive in the original format until business users need data) and storage in the cloud.

Security infrastructure tools include data encryption, user authentication and other access controls, monitoring systems, firewalls, enterprise mobility management and other products to system and data protection.

Specific large data technology (Big-data-specific technologies)

In addition to the IT infrastructure mentioned above used for data in general, there are a number of specific technologies for Big Data that your IT infrastructure should support.

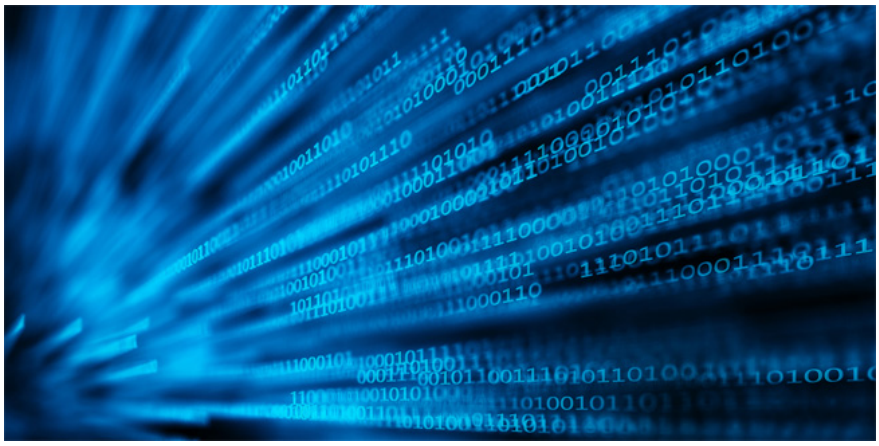
Hadoop ecosystem

Hadoop is one of the technologies closely related to Big Data. The Apache Hadoop project develops open source software for scalable and distributed computers.

Hadoop software library is a framework that allows the distribution of large data sets on computer groups using a simple programming model. It is designed to extend from a single server to thousands of other machines, each providing local computing and storage.

The project includes:

1. **Hadoop Common** , popular utilities that support other Hadoop sections;
2. **Hadoop Distributed File System** , provides high application data access;
3. **Hadoop YARN** , a framework for work planning and resource management;
4. **Hadoop MapReduce** , a YARN-based system for parallel processing of large data sets.

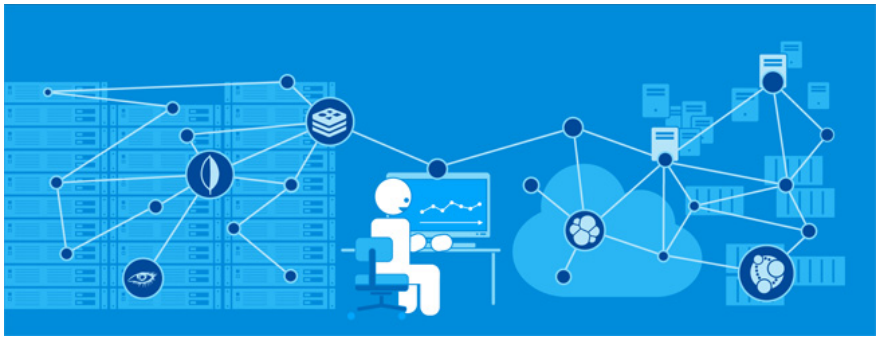


Apache Spark

Part of the Hadoop ecosystem, Apache Spark is an open source cluster computing framework used as a Big Data processing tool in Hadoop. Spark has become one of the important Big Data processing frameworks and can be deployed in many different ways. It provides native constraints for Java, Scala, Python (especially the Anaconda Python distro) and the R programming language (especially for Big Data R) and supports SQL, streaming data, machine learning and graph processing.

Data lakes

Data lakes are repositories that store huge amounts of raw data in the original format until business users need data. The factors that help increase data lake are the digital transformation initiatives and the development of IoT. The data lake is designed to help users easily access a large amount of data when needed.

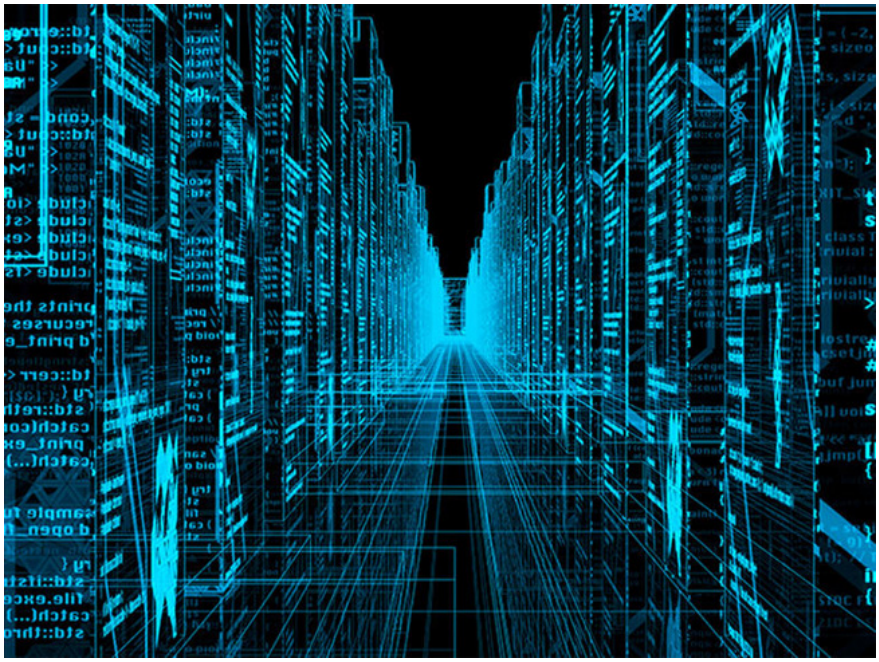


NoSQL database

Common SQL databases are designed for reliable transactions and random queries, but they also have limitations such as the rigid schema that makes it unsuitable for some types of applications. NoSQL database outlines the limitations, storage and management of data in ways that enable high speed operation and flexibility. Many databases have been developed by companies, looking for better ways to store content or handle data for large websites. Unlike SQL databases, many NoSQL databases can be expanded horizontally across hundreds or thousands of servers.

Database in memory

In-memory database (IMDB - In-memory databases) is a database management system that relies heavily on main memory to store data, instead of disk. The memory database is faster than the optimized databases in the disk, an important point to use Big Data analysis, creating data warehouses and metadata.



Big Data skills

Big Data and efforts to analyze Big Data require specific skills, whether from within the organization or through external experts.

Many skills are related to important data technology components such as Hadoop, Spark, NoSQL, database in memory and analysis software.

Other areas are about principles such as data science, data mining, statistical analysis and quantification, data visualization, general-purpose programming, and data structures and Data Structure and algorithms. In addition, there should be people with overall management skills to manage the progress of Big Data projects.

With the popularity of data analysis projects and the shortage of personnel on these skills, finding experienced professionals is one of the biggest challenges for organizations.

Cases of using Big Data

Big Data and analytics can be applied in many business issues and various use cases. Below are a few examples:

1. **Customer analysis** . Companies can check customer data to improve the user experience, improve conversion rates and keep customers better.
2. **Analysis of activities** . Improving operational efficiency and using assets better is the goal of many companies. Analyzing Big Data can help businesses operate more efficiently and improve performance.
3. **Fraud prevention** . Data analysis can help organizations identify suspicious activities and patterns that can indicate fraudulent behavior and help minimize risks.
4. **Price optimization** . Companies can use Big Data analysis to optimize prices for products and services, helping to increase revenue.

Refer to some more articles:

1. 12 extremely useful tricks for JavaScript programmers
2. If you want a successful career, find out about the five 2018 technology trends!
3. Why are there many Microsoft Visual C ++ Redistributable versions installed on the computer?

Having fun!