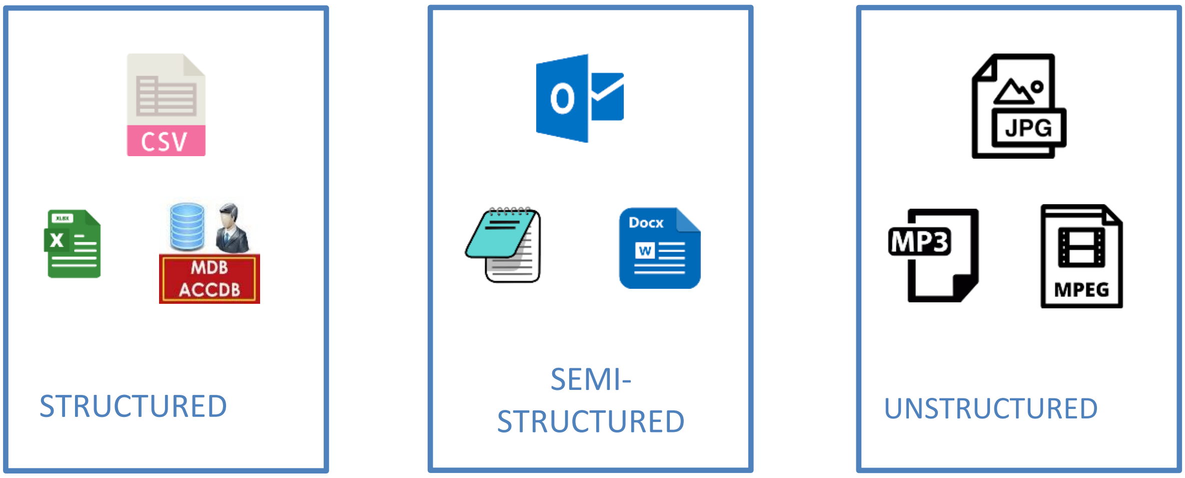
1 BIG DATA

1.1 Introduction and definition of the term Big Data

The amount of data created by humans is increasing rapidly every year due to the introduction of new technologies, various electronic devices, and communication channels such as social networks. Big data is a group of huge data sets that cannot be processed using typical computer methods.

Big Data is a massive collection of data that increases dramatically over time. This is a data set that is so huge and complicated that no typical data management technologies can efficiently store or process it. Big data is like regular data, except much bigger. Big data analytics is the application of advanced analytical techniques to very large, heterogeneous data sets that may contain structured, semi-structured, and unstructured data, as well as data from many sources and sizes ranging from terabytes to zettabytes.

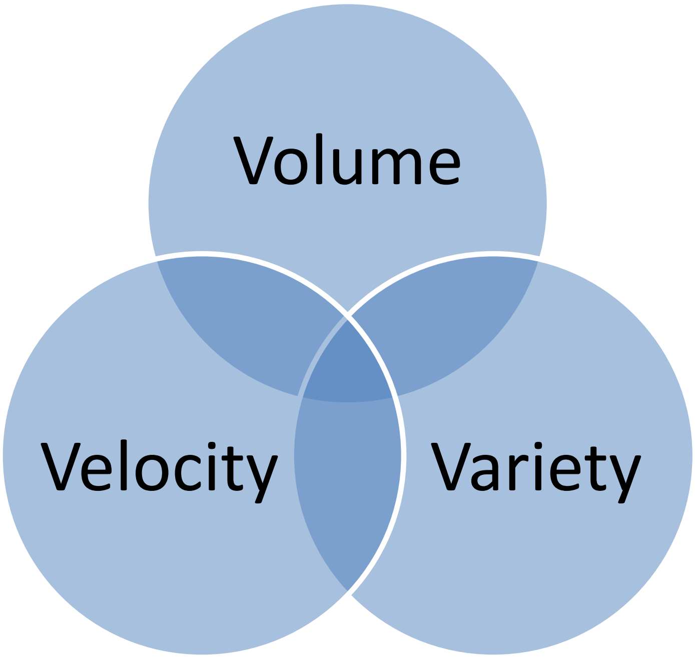


Picture 1 Examples of structured, semi-structured and unstructured files

To put that into perspective, Facebook as a social network produces 4+ petabytes of data every day, which is about a million gigabytes. If we add even more detailed statistics about Facebook, five new profiles are created every second and there are approximately 32 billion active users per day.

1.1.1 Characteristic properties of Big Data

Big data can be described by three basic characteristics - quantity (Volume), speed of its creation (Velocity) and diversity (Variety). These characteristics are usually shown in the following figure:



Picture 2 Characteristic properties of Big Data

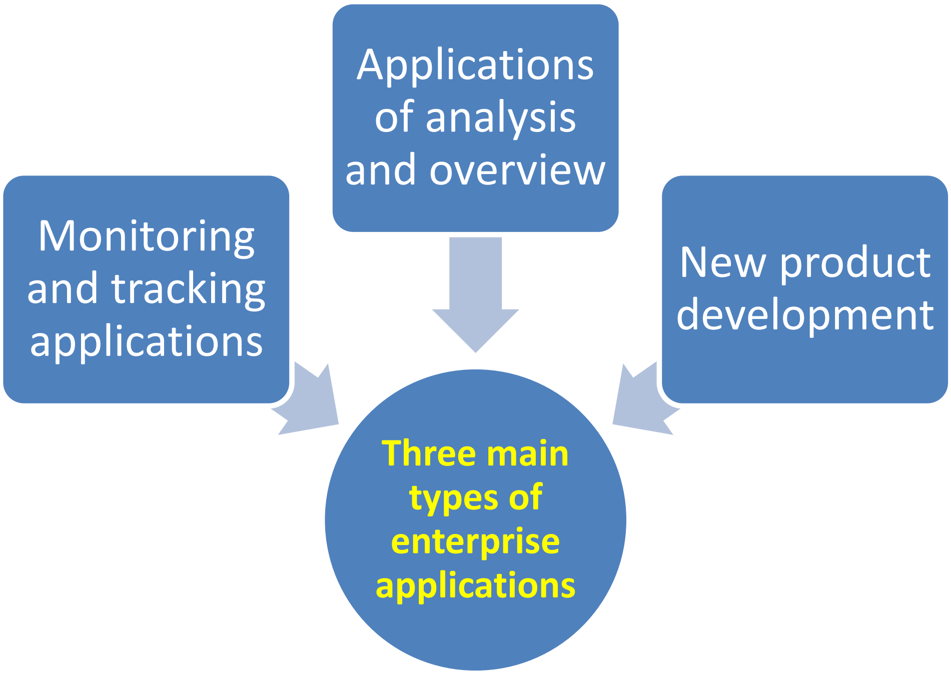
- ✓ ***Volume –*** large amount of data. Data size plays a critical role in determining its value.
- ✓ ***Velocity –*** *d*ata comes in at high speed from machines, networks, social media, mobile phones, and other sources at the speed of big data. There is a large and constant influx of data. This affects the potential of data, or how quickly data is created and processed to meet needs.
- ✓ ***Variety –*** this includes the division of data into structured, semi-structured and unstructured. Structured data is data that is organized. Semi-structured data is data that has an unconventional structure. Unstructured data is data that is not organized.

The advantages that flow from the processing of big data are as follows:

- ✓ Businesses can use external information when making decisions.
- ✓ Access to social data from search engines and sites like Facebook, Twitter allow organizations to fine-tune their business strategies.
- ✓ Improved customer service.
- ✓ Timely identification of risk to the product/service, if any.
- ✓ Better operational efficiency.

1.2 Applications related to Big Data

All data must be recorded and processed, which requires a lot of expertise, resources, and time. Data can be creatively and meaningfully used to deliver business benefits. There are three kinds of enterprise applications, each with varying degrees of revolutionary potential.


Applications of analysis and overview
Monitoring and tracking applications
New product development
Three main types of enterprise applications

Picture 3 Applications related to Big Data

Monitoring and tracking applications:

- ✓ public health monitoring,
- ✓ asset tracking,
- ✓ supply chain monitoring,
- ✓ preventive maintenance of machines and equipment.

Applications of analysis and overview:

- ✓ predictive police - crime control,
- ✓ victory in political elections,
- ✓ personal health.

New product development:

- ✓ flexible car insurance – data obtained from GPS can be used to predict the risk of accidents,
- ✓ retail promotion by location,
- ✓ recommendation of services and goods.

1.3 Nástroje používané pre spracovanie Big Data

Apache Hadoop

It enables distributed processing of large data sets across clusters of computers. It is one of the most powerful big data technologies with the ability to grow from a single server to thousands of computers.

HPCC Systems

HPCC is a big data tool developed by LexisNexis Risk Solution. It brings data processing on a single platform, a single architecture, and a single programming language.

Apache STORM

Storm is a free and open-source big data computing system. It is one of the best big data tools that offers a fault-tolerant, real-time distributed processing system. With real-time computing capabilities.

Qubole

Qubole Data is an autonomous big data management platform. It is an open source big data tool that is self-managing, self-optimizing, and empowers the data team to focus on business outcomes.

Statwing

Statwing is an easy-to-use statistical tool. It was built by and for big data analysts. Its modern interface selects statistical tests automatically.

The above-mentioned applications are only a selection from a number of other available applications for Big Data processing.

1.4 Challenges in Big Data processing

Lack of proper understanding of Big Data

Companies fail in their big data initiatives due to lack of understanding. Employees may not know what data is, its storage, processing, importance, and sources. Data professionals may know what's going on, but others may not have a clear picture. For example, if employees do not understand the importance of data storage, they may not keep a backup of sensitive data. They may not be using the correct databases for storage. As a result, when this important data is needed, it cannot be easily obtained.

Data growth issues

One of the most pressing big data challenges is properly storing all these huge data sets. The amount of data stored in company data centres and databases is growing rapidly. As these datasets grow exponentially over time, they are extremely difficult to manage. Most data are unstructured and comes from documents, videos, audio, text files and other sources. That means you can't find them in databases.

Confusion when choosing a Big Data processing tool

Companies are often confused when choosing the best big data analytics and storage tool. Is HBase or Cassandra the best technology for data storage? Is Hadoop MapReduce good enough or will Spark be a better choice for data analysis and storage? These questions bother companies and sometimes they cannot find the answers. They end up making bad decisions and choosing the wrong technology. The result is a waste of money, time, effort, and man-hours.

Lack of data professionals

Companies need experienced data professionals to operate these modern technologies and big data tools. These professionals will include data scientists, data analysts, and data engineers who are experienced in working with tools and making sense of huge data sets. Companies face the problem of a shortage of big data experts. This is because data processing tools have evolved rapidly, but professionals in most cases have not.

Data security

Securing these huge data sets is one of the daunting challenges of big data. Companies are often so busy understanding, storing, and analysing their data sets that they push data security at later stages. However, this is not a smart move, as unprotected data stores can become a breeding ground for hackers.

Integration of data from different sources

Data in an organization comes from various sources such as social media sites, ERP applications, customer logs, financial reports, emails, presentations, and reports created by employees. Combining all this data to prepare reports is a challenging task. This is an area that is often neglected by businesses. However, data integration is crucial for analysis, reporting and business intelligence, so it must be perfect.

1.5 Storage of Big Data

The first storage mechanism used by computers to store data was punched cards. Each group of related punch cards (Punch cards related to the same program) used to be stored in a filing cabinet and the files were stored in warehouses. This is very similar to what we currently do when archiving documents in government institutions that still use paperwork daily. This is where the word "File System" comes from. Computer systems have evolved, but the concept remains the same.

Instead of storing information on punched cards, we can now store information/data in a digital format on a digital storage device such as a hard disk, flash drive, etc. Related data is still categorized as files, related groups of files are stored in folders. Each file has a name, an extension, and an icon. The file name indicates the content it has, while the file extension indicates the type of information stored in that file. For example, the extension exe indicates executable files, txt indicates text files, etc. A file management system is used by the operating system to access files and folders stored on the computer or any external storage device.

In Big Data, we often encounter multiple clusters (computers). One of the main advantages of Big Data is that it goes beyond the capabilities of a single super-powerful server with extremely high computing power. The whole idea of Big Data consists in distributing data to several clusters and using the computing power of each cluster (node) to process information. A distributed file system is a system that can handle data access across multiple clusters (nodes). In the next section, we will learn more about how it works. A distributed file system works as follows:

- ✓ **Distribution -** distribution of blocks of data sets among multiple nodes. Each node has its own computing power, which gives the system the ability to process data blocks in parallel.

- ✓ **Replication -** distributed file systems will also replicate data blocks on different clusters by copying the same pieces of information to multiple clusters on different racks.

Data is measured in bits and bytes (pronounced byte). One bit can have the value 0 or 1. Eight bits make up a Byte. Then we have Kilobytes (1000 Bytes), Megabytes (1000^2 Bytes), Gigabytes (1000^3 Bytes), Terabytes (1000^4 Bytes), Petabytes (1000^5 Bytes), Exabytes (1000^6 Bytes) and Zettabytes (1000^7 Bytes).

Kilobyte	1 000 bytes
Megabytes	1 000 000 bytes
Gigabytes	1 000 000 000 bytes
Terabytes	1 000 000 000 000 bytes
Petabytes	1 000 000 000 000 000 bytes
Exabytes	1 000 000 000 000 000 000 bytes
Zettabytes	1 000 000 000 000 000 000 000 bytes

1.6 Scalable computing technology via the Internet

With cloud hosting, it's easy to grow and shrink the number and size of servers as needed. This is achieved by either increasing or decreasing cloud resources. This ability to change plans due to fluctuations in business size and needs is a great advantage of cloud computing, especially when there is a sudden increase in demand. This is also called as scalability. In other words, it is the ability of a process, network, software, or organization to grow and manage increased demand. This process has the following attributes:

- ✓ **The age of internet computing**

Billions of people use the internet every day. As a result, supercomputer sites and large data centers must provide high-performance computing services to large numbers of Internet users simultaneously. We need to update data centers with fast servers, storage systems and broadband networks. The purpose is to advance network computing and web services with new technologies.

- ✓ **High Performance Computing (High Performance Computing)**

For many years, HPC systems have emphasized performance in raw speed. The speed of HPC systems has increased from Gflops in the early 1990s to Pflops today. This improvement was mainly driven by the demands of the scientific, engineering and manufacturing communities. However, the number of supercomputer users is limited to less than 10% of all computer users. Today, most computer users use desktop computers or large servers for Internet searches and market-driven computing tasks.

- ✓ **Three new computing paradigms**

Advances in virtualization make it possible to see the growth of Internet clouds as a new computing paradigm. The maturity of radio frequency identification (RFID), Global Positioning System (GPS) and sensor technologies has fuelled the development of the Internet of Things (IoT).

- ✓ **Computational paradigm differences**

For many years, the high-tech community has argued over the precise definitions of centralized computing, parallel computing, distributed computing, and cloud computing. Distributed computing is the opposite of centralized computing. The field of parallel computing largely overlaps with distributed computing, and cloud computing overlaps with distributed, centralized, and parallel computing. This is a computing paradigm in which all computing resources are centralized in one physical system. All resources (processors, memory, and storage) are fully shared and closely interconnected within one integrated operating system. Many data centres and supercomputers are centralized systems, but are used in parallel, distributed, and cloud applications.

1.7 New Data Dimensions and Data Stores

Big data often starts a discussion about new dimensions defined for data. They need to be treated in a different way than just big data. These new challenges are:

- ✓ **Real-time data -** this data is different from the traditional form of data that we store on our servers. It doesn't matter if it falls under structured or unstructured data type. The key aspect is that this is "current data", not old data. They enable situational awareness of what is happening right now. Real-time data raises the issue of perishable and orphaned data that no longer has a valid use but continues to be used, nonetheless.
- ✓ **Shared data -** this is information that is shared within the organization. This includes sharing information between different applications and data sources.

To share information effectively, businesses must ensure that data is consistent, actionable, and extensible. An important aspect is that information sharing complicates the task of determining information authority.

- ✓ **Linked data -** this comes from different data sources that have relationships with each other and maintain context to make it useful for humans and computers. When a user links data, the relationship in that data persists from that point.
- ✓ **Highly confidential data -** this data preserves the context, details, relationships, and identities of important business information. This is largely done through embedded metadata. Highly confidential data allows new meaning to be added without destroying the previous meaning of the data.

The data is stored in the so-called Big Data repositories. This technology was first developed in the early 2000s when companies were faced with storing huge amounts of data that they could not keep on their servers. The problem was that traditional storage methods couldn't handle all this data, so companies had to find new ways to keep it. That's when the Big Data repository was created. It's a way for companies to store large amounts of data without worrying about running out of space.

Warehouse storage and cloud storage are two of the most popular options for storing big data. Warehouse storage is usually on-site, while cloud storage involves storing data off-site in a secure location:

- ✓ **Warehouse storage -** is one of the most common ways to store large amounts of data, but it has its drawbacks. For example, if we need immediate access to our data and want to avoid delays or problems accessing it over the Internet, there may be better options. Storage can also be expensive if we are looking for long-term contracts or need additional staff to manage the warehouse space.
- ✓ **Cloud storage -** is an increasingly popular option as it is easier than ever to use this method thanks to advances in technology such as Amazon Web Services (AWS).

1.7.1 Where are all data from the Internet stored?

Most digital information is stored in three types of places. The first is a global collection of so-called end points (End Points), which include all Internet of Things devices, computers, smartphones, and all other information storage devices. The second is the so-called edge, which includes infrastructure such as cell towers, institutional servers, and offices such as universities, government offices, banks, and factories. The third is the core (Core), most of the data is stored in what are known as traditional data servers and cloud data centres.

There are around 600 hyperscale data centres in the world - with more than 5,000 servers. About 39% of these are in the US, while China, Japan, the UK, Germany, and Australia account for about 30% of the total.

The largest data centers in the world are the China Telecom Data Center in Hohhot, China, which occupies 10.7 million square feet, and The Citadel in Tahoe Reno, Nevada, which occupies 7.2 million square feet and uses 815 megawatts of power.

Picture 4 Data warehouses in China and the USA

1.8 Real examples of Big Data

Big Data in marketing and advertising

- ✓ **Netflix -** Netflix has over 150 million subscribers and collects data on all of them. They track what people watch, when they watch it, what device is being used, whether a show is paused, and how quickly a user finishes watching a series. They even take screenshots of scenes that people watch twice. Why? Because by feeding all this information into its algorithms, Netflix can create its own user profiles. These allow them to personalize the experience by recommending movies and TV shows with impressive accuracy.
- ✓ **Amazon -** Amazon collects a huge amount of data about its users. They track what users buy, how often (and how long) they stay online, and even things like product reviews. Amazon can even estimate people's income based on their billing address. By collecting all this data across millions of users, Amazon can create highly specialized segmented user profiles. Using predictive analytics, they can then target their marketing based on users' browsing habits. It's used to suggest what you might want to buy next, but also for things like grouping products together to make shopping easier.

Big Data in healthcare

- ✓ **Electronic Health Records -** Our medical records include everything from our personal information to family history, allergies and more. For decades, this information was in paper format, limiting its usefulness. However, healthcare systems around the world are now digitizing this data and creating a substantial body of electronic health records.

- ✓ **Big Data and devices for everyday wear -** smart watches and bracelets can collect information about our activities, but also about heart rate or stress level, for example.

Big Data in travel and logistics

- ✓ **Logistics -** logistics companies that track inventory status, shipping reports, product orders, and more use big data to streamline their operations. A good example is UPS. By tracking weather data and truck sensor data, UPS learned the fastest routes for its drivers.
- ✓ **Urban mobility -** Big data is big business in urban mobility, from car rentals to e-bike and e-scooter rentals. Uber is an excellent example of a company that has fully exploited the potential of big data. First, because they have a large database of drivers, they can match users to the nearest driver within seconds. But it doesn't end there. Uber also stores data about every trip you make. This allows them to predict when the service will be busiest, allowing them to adjust their fares accordingly.

Big Data in agriculture

- ✓ **Precision Agriculture -** Farmers use big data to make more informed decisions about their crops. Sensors placed in fields measure moisture levels, temperature, and soil conditions, as well as on tractors and other agricultural machinery. Speaking of agricultural machinery, here's an unusual example: drones. Equipping drones with cameras can provide detailed aerial views of crops, helping to detect diseases or pests.

1.9 Big Data and the collection of privacy information

Online privacy refers to your ability and right to keep private information to yourself. While half of the equation has to do with what you post online and how the services you use share your information, the rest is laws and policies designed to protect consumers. Unfortunately, few regions have strict rules in place regarding how technology companies, advertisers, and online services protect your privacy online. And while there are laws to protect online privacy, not all companies follow them.

One of the reasons online privacies is so complicated is that even if you do everything you can to protect your personal information online, companies can be hacked, collect more data than is necessary, or what they know about you. they can use for dubious or even fraudulent purposes.

Below are some options to protect your privacy and private information online:

1. **Share less information with apps and services -** all social media platforms and apps collect data about who we are, what we're interested in, and what we do online. All these shares and data points make up our online footprint.
2. **Use strong and unique passwords with two-factor authentication - s**trong passwords are the most important, and sometimes the only, protection against identity theft and hackers.
3. **Tighten the privacy settings on your social media accounts -** we don't have to delete your social media accounts to improve your online privacy. Instead, it may be enough to simply check the privacy settings of the online accounts we regularly use.
4. **Remove unused mobile apps and browser extensions -** apps and browser extensions can change their security and privacy policies at any time. If we are not actively using the tool, it is best to remove it.
5. **Don't ignore software or operating system updates -** whenever a software or operating system prompts us to update, it means that some part is broken and needs to be fixed.

In general, it is difficult to prevent data collection in the online space. However, we need to be careful what we share online.