Паралелно и дистрибуирано процесирање

Домашна 5

Стефан Милев | 206055

1.

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.

2.

Volume: the quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. The size of big data is usually larger than terabytes and petabytes.

Variety: the type and nature of the data. The earlier technologies like RDBMSs were capable of handling structured data efficiently and effectively. However, the change in type and nature from structured to semi-structured or unstructured challenged the existing tools and technologies. The Big Data technologies evolved with the prime intention to capture, store, and process the semi-structured and unstructured (variety) data generated with high speed(velocity), and huge in size (volume). Later, these tools and technologies were explored and used for handling structured data also but preferable for storage. Eventually, the processing of structured data was still kept as optional, either using big data or traditional RDBMSs. This helps in analyzing data towards effective usage of the hidden insights exposed from the data collected via social media, log files, and sensors, etc. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

Velocity: the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data is produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.

3.

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project.

4.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

5.

SaaS: software as a service is a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted. It is sometimes referred to as "on-demand software", and was formerly referred to as "software plus services" by Microsoft. SaaS applications are also known as Web-based software, on-demand software and hosted software.

IaaS: infrastructure as a service are online services that provide high-level APIs used to dereference various low-level details of underlying network infrastructure like physical computing resources, location, data partitioning, scaling, security, backup etc. A hypervisor, such as Xen, Oracle VirtualBox, Oracle VM, KVM, VMware ESX/ESXi, or Hyper-V, LXD, runs the virtual machines as guests. Pools of hypervisors within the cloud operational system can support large numbers of virtual machines and the ability to scale services up and down according to customers' varying requirements.

PaaS: platform as a service or application platform as a service (aPaaS) or platform-based service is a category of cloud computing services that provides a platform allowing customers to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app. The first public platform as a service was Zimki, launched by Fotango, a London-based company owned by Canon Europe.

6.

Cloud storage is a cloud computing model that stores data on the Internet through a cloud computing provider who manages and operates data storage as a service. It's delivered on demand with just-in-time capacity and costs, and eliminates buying and managing your own data storage infrastructure. This gives you agility, global scale and durability, with "anytime, anywhere" data access.

Amazon S3 or Amazon Simple Storage Service is a service offered by Amazon Web Services (AWS) that provides object storage through a web service interface. Amazon S3 uses the same scalable storage infrastructure that Amazon.com uses to run its global e-commerce network. Amazon S3 can be employed to store any type of object which allows for uses like storage for Internet applications, backup and recovery, disaster recovery, data archives, data lakes for analytics, and hybrid cloud storage.

7.

There are security issues of various kinds related with cloud computing falling into two broad categories: First, the issues related to the cloud security that the cloud providers face (like software provided to the organizations, infrastructure as a service). Secondly, the issues related to the cloud security that the customers experience (organizations who store data on the cloud).

Most issues start from the fact that the user loses control of his or her data, because it is stored on a computer belonging to someone else (the cloud provider). This happens when the owner of the remote servers is a person or organization other than the user; as their interests may point in different directions (for example, the user may wish that his or her information is kept private, but the owner of the remote servers may want to take advantage of it for their own business). Other issues hampering the adoption of cloud technologies include the uncertainties related to guaranteed QoS provisioning, automated management, and remediation in cloud systems.

Several deterrents to the widespread adoption of cloud computing remain. They include:
- Reliability
- Availability of services and data
- Security
- Complexity
- Costs
- regulations and legal issues
- Performance
- Migration
- Reversion
- The lack of standards
- Limited customization
- Issues of privacy

8.

Adobe Connect: Adobe Connect may be purchased for on-premises deployment, multi-tenant hosted deployment, managed private cloud as ACMS, or managed by third-party managed private cloud provider ConnectSolutions.

Microsoft licenses for Lync, Sharepoint and Exchange may be purchased for on-premises deployment, a multi-tenant hosted deployment via Office 365, or managed by third-party cloud hosting from Azaleos, ConnectSolutions and others.

Rackspace.

Indiqus.

CenturyLink.