Homework 5

for the course Parallel and distributed processing

Aleksandar Ivanovski (Student ID: 186063)

20 December 2020

Question 1

What is Big data? Find different definitions and compare them.

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

Big data refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around a long time.

In simple words, big data is larger, more complex data sets, especially from new data sources.

All of the definitions share the following terms but in other words, i.e. Big Data is type of data that can not be dealt with traditional methods, the source and popularity of this data comes from the newly mankind which is purely data-driven.

Question 2

What are VVV in Big data. Give examples for each V.

The tree V's mean, Volume, Variety, Velocity. Each term is self explanatory, but let's define them.

The **Volume** tells how much data is there, the **Variety** tells how diverse are the different types of data, and the **Velocity** tells at what speed the data is generated.

An example for the Volume is: The data that Amazon has is 1000000 TB. An example for the Variety is: Facebook has Media data (videos, photos, GIFs, etc...), Chat data, Financial data (from the paid services), Analytics data (for ads, marketing, etc...), and many more types of data. An example for the Velocity is: Facebook collects 500 TB of data daily.

The numbers are from 20.12.2020

Question 3

What is HDFS? Provide a brief explanation.

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.

Question 4

Why is Hadoop a distributed programming framework?

Because Hadoop consist of many modules which provide, distributed file system, ability for execution of parallel programs, i.e. map-reduce, for job scheduling, node management, job tracker, and many more. In other words Hadoop provides all the things that are needed for a distributed programming.

Question 5

Specify the models of cloud services and give examples of companies who offer services for each model?

The models of cloud services are the following:

IaaS provides cloud based services, and pay as you go model of billing, storage services, network, virtualization and a complete infrastructure which supports horizontal and vertical scaling of the different aspects. Some examples are: DigitalOcean, Linode, Rackspace, Amazon Web Services (AWS), Cisco Metapod, Microsoft Azure, Google Compute Engine (GCE).

PaaS, provides software and hardware which is accessible over the internet and the cloud, it relies on some infrastructure which can be managed by the provider or could be an IaaS, the end user has a complete managed platform, ready for deployment of the customer's applications. Some examples are: AWS Elastic Beanstalk, Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos, OpenShift. Heroku provides PaaS, but relies on AWS for IaaS, which is an interesting example of using the cloud services, worth mentioning.

SaaS, provides software which is accessible over the internet, but it is managed by third parties. Some examples are: Google Workspace, Dropbox, Salesforce, Cisco WebEx, Concur, GoToMeeting.

Question 6

Specify example of cloud computing which provides resources for data storage and warehousing! Elaborate the example.

Firebase's CloudStorage is such an example. It provides practically infinite amount on storage, scalable on demand.

Question 7

Specify some of the challenges from introducing cloud computing.

Some of the challenges surrounding the implementation of cloud computing are: Security issues, Cost management and containment, Lack of resources/expertise, Governance/Control, Compliance, Managing multiple clouds, Performance, Building a private cloud, Segmented usage and adoption, Migration.

There are many examples of inappropriate cost management which lead to near bankruptcy. Recent example of near bankruptcy is elaborated in the article available on the following link.

Question 8

Specify companies which provide private cloud services, and give a short explanation.

HPE - Key leader in private cloud, offers hardware and services as well.

VMware - Arguably the top name in private cloud.

Dell EMC - Deep expertise in storage hardware and services.

Oracle - Leader in databases and enterprise storage.

IBM / Red Hat OpenShift - PaaS offerings, works with other cloud platforms.

Microsoft - Legacy expertise in enterprise data centers.

Cisco - The leader in networking.

NetApp - Strong in storage and backup

AWS - The dominant vendor in cloud computing.