End Semester Examination

Name: Mohnish Devaraj

Reg No: 39110636

Roll No: 19S115398

Date: 07 May 2022

Subject Name: Big Data Analytics

Subject Code: SCSA1603

No. of pages: 6

3604153987397915

(31) List out and discuss the characteristic of big data.

Ans:

Big Data:

Big Data refers to the huge collections of data that are structured and unstructured. This data may be sourced from servers, Customer profile information, order and purchase data, financial transactions, ledgers, search, history, and employee records. In large companies, this data collection is continuously growing with time.

Types of Big Data

There are three basic types:

1. Structured Data

As the name suggests, this kind of data is structured and is well-defined. It has a consistent order that can be easily understood by a computer or a human. This data

can be stored, analyzed, and processed using a fixed format. Usually, this kind of data has its own data model.

2. Unstructured data

Any set of data that is not structured or well defined is called unstructured data. This kind of data is unorganized and difficult to handle, understand and analyze. It does not follow a consistent format and may vary at different points.

3. Semi-structured data

This kind of data is somewhat structured but not completely. This may seem to be unstructured at first and does not obey any formal structures of data model such as RDBMS.

There are 5 characteristics of Big Data:

i; Volume:

The name 'Big Data' itself is related to a size which is enormous. To determine the value of data, size of data plays a crucial role. If the volume of data is very large then it is actually considered a 'Big Data'. This means whether a particular data can be considered as a Big Data or not is dependent upon the volume of data.

ii, Velocity

Velocity refers to the highest speed of accumulation of data. In Big data velocity data flows in form of sources like machines, network, social media, mobile phones, etc. There is a massive and continous flow of data.

iii, Variety:

It refers to nature of data that is structured, semi-structured and unstrudured data. It also refers to heterog-enous ~~recursive~~ source. Variety is basically the arrival of data from new source that are both inside and outside of an enterprice.

iv, Veracity:

It refers to inconsistency and uncertainity in data, that is data which is available can sometimes get messy and quality and accurate are difficult to control.

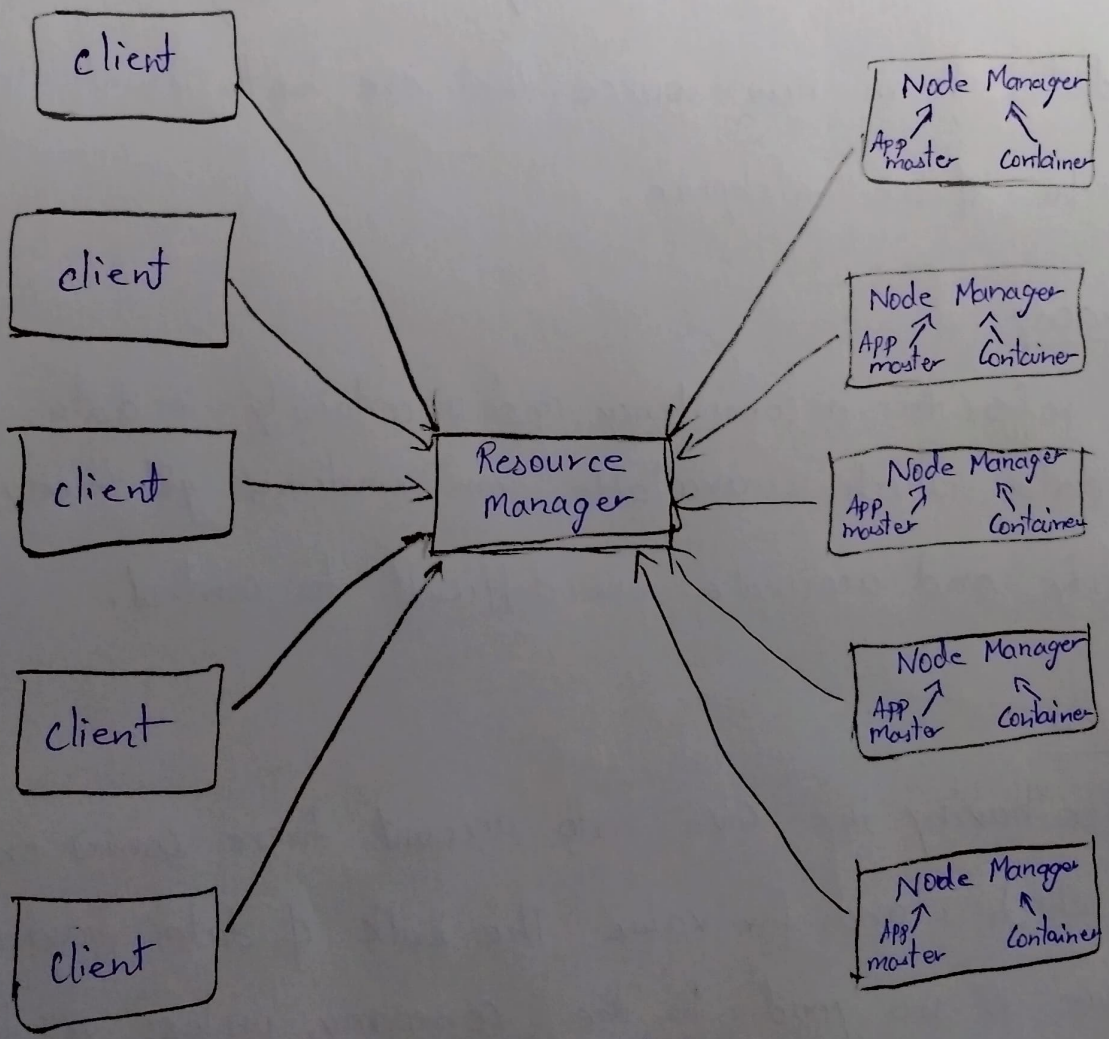
v, Value:

After having the 4Vs into account there comes one more V which stands for value. The bulk of data having no value is no good to the company, unless you turn it into something usefull. Data in itself is of no use or ~~implement~~ importance but it needs to be something valuable.

(34) Summarize the yarn architecture and its components.

Ans:

YARN stands for "Yet Another ~~Rou~~ Resoure Negotiater". YARN architecture basically seperates resource management layer from the processing layer. The responsibility of Job tracker is split between the resource manager and application manager.



Hadoop YARN Architecture

The main components of YARN architecture include:

a) Client:

It submits map-reduce jobs.

b) Resource Manager:

It is the master daemon of YARN and is responsible for resource assignment and management among all the application. Whenever it receives a processing request, it forwards it to the corresponding node manager and allocates resources for the completion of the request accordingly. It has two major components.

i) Scheduler:

It performs scheduling based on the allocated application and available resources. It is a pure scheduler, means it does not perform other tasks such as monitoring or tracking and does not guarantee a restart if a task fails.

ii) Application Manager:

It is responsible for accepting the application and negotiating the first container from the resource manager. It also restarts the Application Master container if a task fails.

c) Node Manager

It take care of individual node on hadoop cluster and manages application and work flow and that particular node. Its primary job is to keep-up with the resource manager. It registers with the resource manager and sends heartbeats with the health status of the node.

d) Application Master

An application is a single job submitted to a framework. The application master is responsible for negotiating resources manager, tracking the status and monitoring progress of a single application. The application master requests the contained from the node manager by sending a Container Launch Context (CLC)

e) Container:

It is a collection of physical resources such as RAM, CPU cores and disk on a single node. The containers are invoked by Containe Launch Context (CLC) which is a record that contains information such as enviroment variable, security tokens, dependencies, etc.