|  |  |  |
| --- | --- | --- |
| **UMBC – CSEE Department** |  | Dr. Waleed Youssef |
| **Data Science Program** |  | youssef1@umbc.edu |
| **Spring 2023** |  |  |
| **DATA 603 – Big Data Platforms** | | |
| **Homework #2** | | |
|  |  |  |

**Questions:**

1. **[10 points]** Watch the OSI video (in the slides) and discuss possible few challenges from your point of view?

Video URL Link: <https://youtu.be/LANW3m7UgWs>

* + As the packet must travel 7layers in source and 7layers in destination the time taken to reach destination will be increased.
  + There is a probability in losing some packets as it must travel a lot of layers.
  + Because of involvement of many layers the complexity to implement is difficult.
  + Because of complex nature, involvement of a lot of equipment cost to implement is more.

1. **[15 points]** Discuss and present different computing models presented in Lecture #2? Present pros and cons for each model?

**Distributed Systems**: It is a collection of independent nodes which perform a single task. Multiple independent nodes are connected through network. Network includes hubs, switches, routers, nodes etc. TCP/IP model is mostly used model for networking in distributed systems.

Pros of Distributed system:

Abstraction: Even though there are multiple nodes are present in the network it apprears to be a single node to end user.

Fault tolerance: Distributed systems can continue to provide service even if some parts of it is failed as load is distributed among other nodes.

Consistency: As it is fault tolerant it provides uninterrupted service.

Scalability: Based on the demand required nodes can be added or removed.

Performance: As single task is divided into parts for answer the time taken to complete a task is greatly reduced.

Cons of distributed systems:

Complexity: They difficult to design and implement as it involves multiple devices.

Latency: As the data as to travel through multiple devices there maybe a latency.

Bandwidth of Network: The bandwidth of network is not reliable all the time and may cause delay in processing.

Security: As there are multiple devices involved it increases the number of entry points for attackers and security becomes complex.

Cost: As there are multiple devices the overall cost to implement drastically increases.

**Cloud Computing Model**: It is on-demand availability of computer resources like storage, computing power.

Pros of Cloud Computing:

High speed of deployment of resources with few clicks.

Cost: No initial investment. Pay only for the resources utilized.

Scalability: Based on the demand resources required can be added or removed and we have to pay only for the resources used.

Accessibility: As cloud is accessed over internet it can accessed anywhere in the world on any device that is connected to internet.

Availability: Most of the cloud services are highly available. AWS IS 99.9% available.

Cons of Cloud Computing:

Dependence on Network: Cloud Computing depends on internet to run so any disruption in network affects its accessibility.

Lack of control: Since data is stored off-site and as we have no control on that company, it will be difficult to control data storage. [1]

Migration: Migration form one cloud provider to another is difficult. As architecture, policies differ from one cloud provider to another.

**Parallel Computing:**

A form of computing where multiple instructions are carried out simultaneously.

Parallel computing can be achieved by any of following Flynn’s Taxanomy:

SISD (Single Instruction Single Data)

SIMD (Single Instruction Multiple Data)

MISD (Multiple Instructions Single Data)

MIMD (Multiple Instruction Multiple Data)

Pros of Parallel Computing:

Time: As single instruction is divided between multiple systems the speed to compute is increased.

Scalability: As the complexity of problem increases number resources to solve it increased.

Efficiency: As most of the resources are used to computing the usage is achieved nearly 100% avoiding wastage of resources. [2]

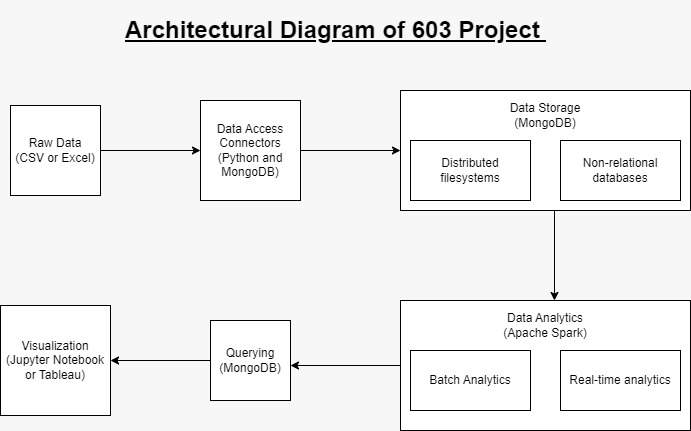
Cons Of Parallel Computing:

Coding Complexity: Code to implement parallel computing is pretty complex and is difficult to implement.

High Cost: As more resources are required to to run parallel computing architecture like data synchronization, data threads, etc. Cost to implement it is high. [3]

As the number of resouces is high there is huge power consumption.

1. **[15 points]** Draw an architecture diagram of the environment you propose to use in your class project. If you are not sure yet about the environment, do you best and just present a proposal that might be changed later.



1. **[10 points]** For your development environment, propose a possible environment you may use in this class, include tools to be used (Apache Spark, Streaming, MongoDB, Hadoop HDFS, Anaconda, Python, etc.), possible stack of architecture, and how tools will be integrated together?

**Raw Data:** Data in either CSV or Excel is gathered from online sources like Kaggle or Google Data Sets.

**Data Access Connectors:** To access this data we will use Python Pandas Data Frame.

**Data Storage:** For data storage we will use a non-relational database MongoDB. Because it has rich querying and analytics.

**Data Analytics:** For data analytics we will be using Apache Spark because it is open source and 100x faster than Hadoop for large scale data processing.

**Visualization:** For rich visualizations we will be using Jupyter Notebook or Tableau.

* + **Notes:**
    - *Do not include tools that you will not be using*
    - *Do not just copy diagrams as is from the Web*

**References:**

[1]Edmondson, J. (2022, March 13). *Pros and Cons of Cloud Storage: Is Cloud Storage right for your data?* Businesstechweekly.com. from <https://www.businesstechweekly.com/operational-efficiency/cloud-computing/pros-and-cons-of-cloud-storage/>

[2] *Parallel Programming: Definition, Benefits and Industry Uses*. (2021, July 21). Indeed. from <https://www.indeed.com/career-advice/career-development/parallel-programming>

[3] Kacker, T. (2022, September 20). *Replace a Character in a String Python*. Scaler. Retrieved February 12, 2023, from <https://www.scaler.com/topics/parallel-operating-system/>