BIG DATA

Assignment-1

SUBMITTED BY:

AYUSH KUMAR JHA
500086400
B.C.A-IOT (2020-2023)

SUBMITTED TO:

SHELLY MA'AM

Q2.What are the challenges of distributed computing ?

Q1.What according to you are the challenges for start-ups and SMEs in Migrating to Big Data Platform ?

1. Lack of knowledge Professionals

To run these modern technologies and large Data tools, companies need skilled data professionals. These professionals will include data scientists, data analysts, and data engineers to work with the tools and make sense of giant data sets. One of the Big Data Challenges that any Company faces is a drag of lack of massive Data professionals. This is often because data handling tools have evolved rapidly, but in most cases, the professionals haven't. Actionable steps have to be taken to bridge this gap.

2. Lack of proper understanding of Massive Data

Companies fail in their Big Data initiatives, all thanks to insufficient understanding. Employees might not know what data is, its storage, processing, importance, and sources. Data professionals may know what's happening, but others might not have a transparent picture. For example, if employees don't understand the importance of knowledge storage, they could not keep the backup of sensitive data. They could not use databases properly for storage. As a result, when this important data is required, it can't be retrieved easily.

3. Data Growth Issues

One of the foremost pressing challenges of massive Data is storing these huge sets of knowledge properly. the quantity of knowledge being stored in data centers and databases of companies is increasing rapidly. As these data sets grow exponentially with time, it gets challenging to handle. Most of the info is unstructured and comes from documents, videos, audio, text files, and other sources. This suggests that you cannot find them in the database.

4. Confusion while Big Data Tool selection

Companies often get confused while selecting the simplest tool for giant Data analysis and storage. Is HBase or Cassandra the simplest technology for data storage? Is Hadoop MapReduce ok, or will Spark be a far better option for data analytics and storage? These questions bother companies, and sometimes they're unable to seek out the answers. They find themselves making poor decisions and selecting inappropriate technology. As a result, money, time, effort, and work hours are wasted.

5. Integrating Data from a Spread of Sources

Data in a corporation comes from various sources, like social media pages, ERP applications, customer logs, financial reports, e-mails, presentations, and reports created by employees. Combining all this data to organize reports may be a challenging task. This is a neighborhood often neglected by firms. Data integration is crucial for analysis, reporting, and business intelligence, so it's perfect.

6. Securing Data

Securing these huge sets of knowledge is one of the daunting challenges of massive Data. Often companies are so busy in understanding, storing, and analyzing their data sets that they push data security for later stages. This is often not a sensible move as unprotected data repositories can become breeding grounds for malicious hackers. Companies can lose up to $3.7 million for a stolen record or a knowledge breach.

Q2.What are the challenges of distributed computing ?

Designing a distributed system does not come as easy and straightforward. A number of challenges need to be overcome in order to get the ideal system. The major challenges in distributed systems are listed below:

Heterogeneity:

Heterogeneity – "Describes a system consisting of multiple distinct components"

Of course, heterogeneity applies to pretty much anything which is made up of many different items or objects including Food! (Okay so that might be a bad analogy but you get the idea)

Anyway, in many systems in order to overcome heterogeneity, a software layer known as Middleware is often used to hide the differences amongst the components underlying layers.

Openness:

Openness –"Property of each subsystem to be open for interaction with other systems"

So once something has been published it cannot be taken back or reversed. Furthermore, in open distributed systems there is often no central authority, as different systems may have their own intermediary.

Security:

The issues surrounding security are those of

Confidentiality

Integration

Availability

To combat these issues encryption techniques such as those of cryptography can help but they are still not absolute. Denial of Service attacks can still occur, where a server or service is bombarded with false requests usually by botnets (zombie computers).

Scalability:

What is scalability?! Okay so basically a system is described as scalable if:

“As the system, number of resources, or users increase the performance of the system is not lost and remains effective in accomplishing its goals”

That’s a fairly self-explanatory description, but there are a number of important issues that arise as a result of increased scalability, such as an increase in cost and physical resources. It is also important to avoid performance bottlenecks by using caching and replication.

Fault handling:

Failures are inevitable in any system; some components may stop functioning while others continue running normally. So naturally, we need a way to:

Detect Failures – Various mechanisms can be employed such as checksums.

Mask Failures – retransmit upon failure to receive acknowledgment

Recover from failures – if a server crashes roll back to the previous state

Build Redundancy – Redundancy is the best way to deal with failures. It is achieved by replicating data so that if one sub-system crashes another may still be able to provide the required information.

Concurrency:

Concurrency issues arise when several clients attempt to request a shared resource at the same time. This is problematic as the outcome of any such data may depend on the execution order, and so synchronization is required.

Transparency:

A distributed system must be able to offer transparency to its users. As a user of a distributed system, you do not care if we are using 20 or 100’s of machines, so we hide this information, presenting the structure as a normal centralized system.

Access Transparency – where resources are accessed in a uniform manner regardless of location

Location Transparency – the physical location of a resource is hidden from the user

Failure Transparency – Always try and Hide failures from users