

Continuous Assessment II

Course code: CAP456

Course Name: INTRODUCTION to Big Data

Section : D2112

NAME : JAYSHRI LAL PANDIT

ROLL NO: RD2112A03

Reg No. : 12111670

① Discuss various Big Data challenges.

Ans:- Various Big Data challenges and their solution are given below:-

(1.) Lack of Proper Understanding of Big Data :-

Companies fail in their Big data initiatives due to insufficient understanding. Employees may not know what data is, its storage, processing, importance, and sources. Data professionals may know what is going on, but others may not have a clear picture. For example, if employees do not understand the importance of data storage, they might not keep the backup of sensitive data. They might not use databases properly for storage. As a result, when this important data is required, it cannot be retrieved easily.

Solution :-

Do Big Data workshops and seminars must be held at companies for everyone. Basic training programs must be arranged for all the employees who are handling data.

regularity and are a part of the Big Data projects. A basic understanding of data concepts must be ~~in~~ inculcated by all levels of the organization.

(2.) Data Growth issues :-

One of the most pressing challenges of Big data is storing all these huge sets of data properly. The amount of data being stored in data centers and databases of companies is increasing rapidly. As these data sets grow exponentially with time, it gets extremely difficult to handle.

Most of the data is unstructured and comes from documents, videos, audios, text files and other sources. This means that you cannot find them in databases.

Solution :-

In order to handle these large data sets, companies are opting for modern techniques, such as compression, ~~to~~ tiering, and deduplication. Compression is used

For reducing the number of bits in the data, thus reducing the number of bits in the data, its overall size. Deduplication is the process of removing duplicate and unwanted data from a data set.

Data tiering allows companies to store data in different storage tiers. It ensures that the data is residing in most appropriate storage space. Data tiers can be public cloud, private cloud, and flash storage, depending on the size and importance.

Companies are also opting for Big Data tools, such as Hadoop, NOSQL and other technologies.

③ Confusion While Big Data tool Selection

Companies often get confused while selecting the best tool for Big Data analysis and storage. Is HBase or Cassandra the best technology for data storage? Is Hadoop MapReduce good enough or will Spark be a better option for data analytics and storage?

These questions bother companies and sometimes they are unable to find the answers. They end up making poor decisions and selecting inappropriate technology. As a result, money, time, efforts and work hours are wasted.

Solution :-

The best way to go about it is to seek professional help. You can either hire experienced professionals who know much more about these tools. Another way is to go for Big Data consulting. Here, consultants will give a recommendation of the best tools, based on your company's scenario. Based on their advice, you can work out a strategy and then select the best tool for you.

④ Lack of data professional :-

To run these modern technologies and Big Data tools, companies need skilled data professionals. These professionals will include data scientists, data analysts and

Data engineers who are experienced in working with the tools and making sense out of huge data sets.

Companies face a problem of lack of Big Data professionals. This is because data handling tools have ~~over~~evolved rapidly, but in most cases, the professionals have not. Actionable steps need to be taken in order to bridge this gap.

Solution

Solution :-

Companies are investing more money in the ~~recruit~~ recruitment of skilled professionals. They also have to offer training programs to the existing staff to get the most out of them.

Another important step taken by organizations is the purchase of data analytics solutions that are powered by artificial intelligence / machine learning. These tools can ~~be~~ be run by professionals who are not data science experts but have basic knowledge. This step helps companies to save a lot of money for ~~recruitment~~ ^{int}

⑤ Securing Data :-

Securing these huge sets of data is one of the daunting challenges of Big Data. Often companies are so busy in understanding, storing and analysing their data sets that they push data security for later stages. But, this is not a smart move as unprotected data repositories can become breeding grounds for malicious hackers.

Solution :-

Companies are ~~recruiting~~ recruiting more cybersecurity professionals to protect their data. Other steps taken for security data include :-

- Data encryption
- Data Segregation
- Identity and access control
- Implementation of endpoint security
- Real-time security monitoring
- Use Big Data security tools, such as IBM ~~Guard~~ Guardian.

⑥ Integrating data from a variety of Sources.

Data in an organization comes from a variety of sources, such as social

media pages, ERP applications, customer logs, financial reports, e-mails, presentations and reports created by employees. Combining all this data to prepare reports is a challenging task.

This is an area often neglected by firms. But, data integration is crucial for analysis, reporting and business intelligence, so it has to be perfect.

Solution :-

Companies have to solve their data integration problems by purchasing the right tools. Some of the best data integration tools are mentioned below :-

- Microsoft SQL
- IBM InfoSphere
- Clover DX
- Oracle Data Service Integration
- Xplenty
- QlikView

② Explain Big Data Analytics Cycle.

Ans:- The Big Data Analytics Life cycle is divided into nine phases

, named as :-

(1) Business Problem Definition Phase :-

In this stage, the team learns about the business domain, which presents the motivation and goals for carrying out the analysis. In this stage, the problem is identified, and assumptions are made that how much potential gain a company will make after carrying out the analysis. Important activities in this step include framing the business problem as an analytics challenge that can be addressed in subsequent phases. It helps the decision-makers understand the business resources that will be required to be utilized thereby determining the underlying budget required to carry out the project.

(2) Data Identification :-

Once the business case is identified, now it's time to find the appropriate ~~databases~~ datasets to work with. In this stage, analysis is done to see what other companies have done for a similar case. Depending on

the business case and the scope of analysis of the project being addressed, the sources of datasets can be either external or internal to the company. In this case of internal datasets, the datasets can include data collected from internal sources, such as feedback forms, from existing software, on the other hand, for external datasets, the list includes datasets from third-party providers.

③ Data Acquisition and filtration:-

Once the source of data is identified, now it is time to gather the data from such sources. This kind of data is mostly unstructured. Then it is subjected to filtration, such as removal of the corrupt data or irrelevant data, which is of no scope to the analysis objective. Here corrupt data means data that may have missing records, a copy of the filtered data is stored and compressed, as it can be of use in the future, for some analysis.

④ Data Extraction :-

Now the data is filtered, but there might be a possibility that some of the entries of the data might be incomplete, to rectify this issue, a separate phase is created, known as the data extraction phase. In this phase, the data, which don't match with the underlying scope of the analysis, are extracted and transformed in such a form.

⑤ Data Validation & Representation:-

As mention in phase III, the data is collected from various sources, which results in the data being unstructured. There might be a possibility, that the data might have constraints, that are unsuitable, which can lead to false results. Hence there is a need to ~~are~~ clean and validate the data. It includes removing any ~~invalid~~ data and establishing complex validation rules. There are many ways to validate and clean the data. for example,

a dataset might contain few rows, with null entries. If a similar dataset is present, then those entries are copied from that dataset, else those rows are dropped.

⑥ Data Aggregation & Representation:

The data is ~~cle~~ cleansed and validated, against certain rules set by the 'enterprise'. But the data might be spread across multiple datasets, and it is not advisable to work with multiple datasets. Hence, the datasets are joined together. For example: If there are two datasets, namely that of a student Academic Section and Student Personal Details Section, then both can be joined together via common fields, i.e. roll numbers.

⑦ Data Analysis :

Here comes the actual step, the analysis task. Depending on the nature of the big data problem, analysis is carried out. Data analysis can be classified as Confirmatory analysis and

Exploratory analysis. In confirmatory analysis, the cause of a phenomenon is analyzed before. The assumption is called the hypothesis. The data is analyzed to approve or disapprove the hypothesis. This kind of analysis provides definitive answers to some specific questions and confirms whether an assumption was ~~true~~ or not.

(8) Data Visualization Visualization :-

Now we have the answer to some questions, using the information from the data in the datasets. But these answers are still in a form that can't be presented to business users. A sort of representation is required to obtain value or some conclusion from the analysis. Hence, various tools are used to visualize the data in graphic form, which can easily be interpreted by business users.

Visualization is said to influence the interpretation of the results. Moreover, it allows the users

to discover answers to questions that are yet to be formulated.

⑨ Utilization of analysis results:-

The analysis is done, the results are visualized, now it's time for the business users to make decisions to utilize the results. The results can be used for optimization, to refine the business process. It can ~~also~~ also be used as an input for the Systems to enhance performance.

③. Explain Sharding and Replication in details.

Ans:-

Sharding

Sharding is a method for allocating data across multiple machines. MongoDB used Sharding to help deployments with very big data sets and large throughput the operation. By Sharding, you combine more devices to carry data extension ~~to~~ and the needs of read and write operations.

- Why Sharding?

- Database system having big data sets or high throughput requests can doubt the ability of a single server.
- For example, High query flows can drain the CPU limit of the server.
- The working set sizes are larger than the system's Ram to stress the I/O capacity of the disk drive.

Sharding determines the problem with horizontal scaling breaking the system dataset and ~~store~~ store over multiple servers, adding new servers to increase the volume as needed; Now, instead of one signal as primary, we have multiple servers called shard. we have different routing servers that will route data to the shard servers.

Advantages of Sharding :-

- Sharding adds more servers to a data field automatically adjust data loads across various servers.
- The number of operations each shard manage get reduced.
- It also increases the write capacity by splitting the write load over multiple instances.
- It gives high availability due to the deployment of replica servers for shard and config.
- Total capacity will get increased by adding multiple shards.

Replication

Replication stores multiple copies of a dataset, known as replicas, multiple nodes. Replication provides scalability and availability due to the fact that the same data is replicated on various nodes. Fault tolerance is also achieved since data redundancy ensures that data is not lost when an individual node fails.

These are two different methods that have are used to implement replication :-

- Master-slave
- peer-to-peer

Master-slave :-

During master-slave replication, nodes are arranged in a master-slave config. configuration, and all data is written to a master node. Once saved, the data is replicated over to multiple slave nodes. All external write requests, including insert, update and delete, ~~occur~~ occur on the master node, whereas read requests can be fulfilled by any slave node.

Master-slave replication is ideal for read intensive loads rather than write intensive loads since growing read demands can be managed by horizontal scaling to add more slave nodes. Writes are consistent, as all writes are coordinated by master node. The implication

is that write performance will suffer as the amount of writes increases. If the master node fails, reads are still possible via any of the slave nodes. A slave node can be configured as a backup node for the master node. In the event that the master node fails, writes are not supported until a master node is reestablished. The master node is either resurrected from a backup of the master node, or a new master node is chosen from the slave nodes.

Peer-to-Peer replication :-

With peer-to-peer replication, all nodes operate at the same level. In other words, there is not a master-slave relationship between the nodes. Each node, known as a peer, is equally capable of handling reads and writes. ~~Peer~~ Peer-to-peer replication is prone to write inconsistencies that occur as a result of a simultaneous update of the same data across multiple peers. This

can be addressed by implementing either a ~~per~~ pessimistic or optimistic concurrency strategy.

- Pessimistic concurrency is a proactive strategy that prevents inconsistency. It uses locking to ensure that only one update to a record can occur at a time. However, that is detrimental to availability since the database record being updated remains unavailable until all locks are released.

- Optimistic concurrency is a reactive strategy that does not use locking. Instead, it allows inconsistency to occur with knowledge that eventually consistency will be achieved after all updates have propagated.