

Continuous Assessment II

Course code: CAP456

Course Name: INTRODUCTION to Big Dala

Section : D2112

NAME: JAYSHRI LAL PANDIT

ROLL NO: RD2112 A03

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. bata 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 :-Data Big Data workshops and seminass must be held at companies for everyone - Basic training programs must be assenged for all the

employees whom are handing data

Page No.

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

(2) Data Growth issues :-

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

Most of the data is unstructured and comes from documents, videos, and other sources. This means that you connot 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 bumber of bits in the data, its overall size. Deduplication is the process of removing duplicate and unwanted data from a data set.

Dota tiering allows companies to

Store day 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 impostance.

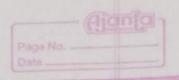
Companies are also opting for

Big Data tools, such as Hadoop,

NOSAL 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 cassardra the best technology for data storage? Is Hadoop MapRedyce 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 roaking poor decisions and selecting inappropriate texthology. As a result, money, time, efforts and work hours are

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 Dalg consulting. Here, consultants will give a recommendation of the best tools, based on your company's scenario. Based on their advise, 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 Dato tools, companies need skilled data professionals.

These professionals will include data scientists, data analysts and

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

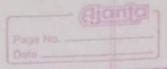
Componies face a problem of lack of Big Data prefessionals - This is because data handling tools have overlived evolved rapidly, but in most cases, the professionak have not actionable steps need to be taken in order to bridge this gap.

Sotion

Solution:

companies are investing more money in the recursive 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 sun by professionals who
are not data science experts but
have basic knowledge. This step helps
companies to save a lot of money too require



Securing Data:

Securing these huge sets of

dout of is one of the dounting

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 rectust 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 Golf 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,
customes logs, financial reports, e-majle
, 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 the intelligence
, so it has to be perfect.

Solution:

Communicate laws to be perfect.

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

- · MicroSoft SQL
- · IBM InfoSphere
- 6 Clover DX
- · Oracle pata service integration
- · Xplenty
- · Qlikview

2) Explain Big Data Analytics cycle.

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

8 of 18

, named as :-

(1) Business Protem Definition / fase: 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 indude framing the business problem as an analytics challenge that can be addressed in subsequent phases: It helps the decision -makers understand the bussiness resources that will be required to be utilized thereby determining the underlying budget required to carry out the project.

(2.) Data Identification: once the butiness case is identified , now it's time to find the appropriate databates datasets to work with . In this stage, analysis is done to see what other companies have done for a similar cose. Depending on

the business code and the scope of analysis of the project being addressed, the Sources of datasets can be either external as infernal to the company. In this case of internal datasets, the datasets can include data collected from internal sources, such as feedback forms, from existing seftware, on the other hand, for external datasets, the list includes datasets from third-pasty 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 may have missing records, as a copy of the filtered data is stored and compressed, as it can be of use in the future, for some analysis.

A) Data Extraction:

Now the data is filtered, but

there might be a possibility

that some of the entries

of the data might be incomplete

of the data might be incomplete

separate phase is created,

known as the data extraction

phase of the data extraction

phase of the data extraction

the underlying scope of the

analysis, are extracted and

transformed in such a form.

(5) Data Valldation & 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 once unsuitable, which can lead to false results. Hence there is a need to receler and validate the data. It included removing any prinvalidate data and establing complex validation rules. There are many ways to validate and clean the data. for example,

a dataset might contain few rows, with null entries off 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 to cleansed and validates, against Certain rules set by the 'enterprise'. But the data might be spread across multiple datasets, and it is not advisible to cook 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 number.

For Data Analysis :
Here Comes the autual 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 Visuation Visualization :=

Now we have the answer to

Some questions, using the
information from the data
in the datasete. But these
answers are still in a form
that can't be presented to
business wers. A sort of
representation is required to
obtains 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 wers,

visualization is said to influence the juterpretation of the results. Moreover, it allows the were to discover answers to questions that are yet to be formulated.

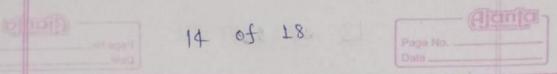
(9) Utilization of analysis results:

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 be also be used as an input for the Systems to enhance performance.

3) Explain sharding and Replication of in details

Ans:- Sharding

Sharding is a method for allocating data across multiple machines. Mango DB wed Sharding to help deprograe with very big data sets and large throughput the operation. By Sharding, you combine more devices to coordy data extension to and the needs of read and curite operation.



- Why Sharding?

- Databale system having big data sets or hight throughput requests can doubt the ability of 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 scalling breaking the system dataset and sotome of Store over multiple servers, adding new servere to increase the volume as needed, Now, instead of one signal as primary, cere have multiple servess called Shard. we have different routing servess that will routed deater to the shard Servers .

Advantages of Sharding :-

- · Sharding adds more serves to a data field automatically adjust data loads across various servers.
- · The number of operations each shard manage got reduced.
- o It also increases the write capacity by splitting the curite load over multiple instances
- of the deployment of replica servers for shard and config.
- 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 indivisual node fails.

P. T. D

Page No.

There 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 congi configuration,

and all data is consisten to

a master node once saved,

the data is replicated over

to multiple slave nodes. All

external corite requests, including

insert, update and delete,

paux occur on the master

node, tuherasas read requests

can be fulfilled by any

slave node.

Moster-slave replication is ideal for read intensive loads rather than corrite intensive loads bracks since growing read demands can be managed by horizontal scalling to add more slave nodes. Corrites are consistent, as all corrites are coordinated by moster node. The implication

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

peer-to-peer replication:

with peer-to-peer replication,

all nodes apcrate at the Same

level. In other woords, there

is not a master-slave

relationship between the nodes.

Each node, known as a peer,

is equally capable of hondling

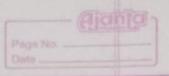
reads and curitel peer peer to - peer replication is prone to

cosite inconsistencies that occurs

as a result of a simultaneous

update of the same data

across multiple peers. This



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

proactive strategy that prevents in consistency. 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 9s a reactive strategy that does not use locking. Instead, it allows in consistency to occur with knowledge that eventually consistency will be achieved after all updates have propagated.