Student Answer Script View

MIT MPL-BTech-M Sc - MCA - 1st-3rd-5th and 7th Semester - Mid Term Examination - Sep 2024

Answer Sheet

Student Name:	HARSH GONDAL . .
Roll Number:	220962334
Course:	Computer Science and Engineering - Artificial Intelligence and Machine Learning
Year/Sem:	Semester 5
Subject Name:	BIG DATA ANALYTICS
Exam Date:	25-Sep-2024

Q.No : 1)

What does RDD stands for in Apache Spark?

Resilient Distributed Dataset	Reaction Distributed Dataset	Reliable Distributed Dataset	Redundant Data Distribu

Q.No : 2)

Which of the following operations will cause an RDD to be recomputed in Spark?

count() cache() map() filter()

Q.No : 3)

Which component of Hadoop is responsible for resource management?

HBASE | YARN | Pig | Sqoop

Q.No : 4)

Which of the following is a NoSQL database in the Hadoop ecosystem?

HDFS | HIVE | HBASE | Sqoop

Q.No : 5)

Which of the following criteria is used by default to split nodes in PySpark's DecisionTreeClassifier?

Information Gain | Variance | Gini Index | Entrophy

Q.No : 6)

Which method is used to make predictions using a trained decision tree model in PySpark?

transform() fit() predict() forecast()

Q.No : 7)

Which of the following techniques is typically used to evaluate the performance of an ALS model in PySpark?

AUC | MSE | Precision | Recall

Q.No : 8)

For a data to be considered as Big Data following are necessary

Volume, Velocity, Value	Volume, Velocity, Variety	Volume, Velocity, Veracity	Volume, Veracity, Variety

Q.No : 9)

The method wholeTextFile() creates _____ RDD

Double | Pair | Sequence | Union

Q.No : 10)

The function range() creates _____ RDD

Python | Sequence | Union | Parallel

Q.No : 11)

How do you schedule a YARN application? How YARN schedules and executes the application. Draw the YARN architecture.

Page:1

- The user either runs standalone mode or pseudo mode of Hadoop, which launches a JVM a single type for standalone mode, and JVM for every node in pseudo mode.

- This sends the application to the Resource Manager which is the master daemon of the yarn architecture. The Resource Manager directly responsible to client's request.

- The Resource Manager delegates a container the first or one Node Manager, this delegated process is called Application Master. The Application master who figures out how many resources are required for the tasks and negotiate with Resource Manager to allocate resources to Node Managers, which can inside itself.

- The resources are then allocated by Application all other nodes, they starting processing

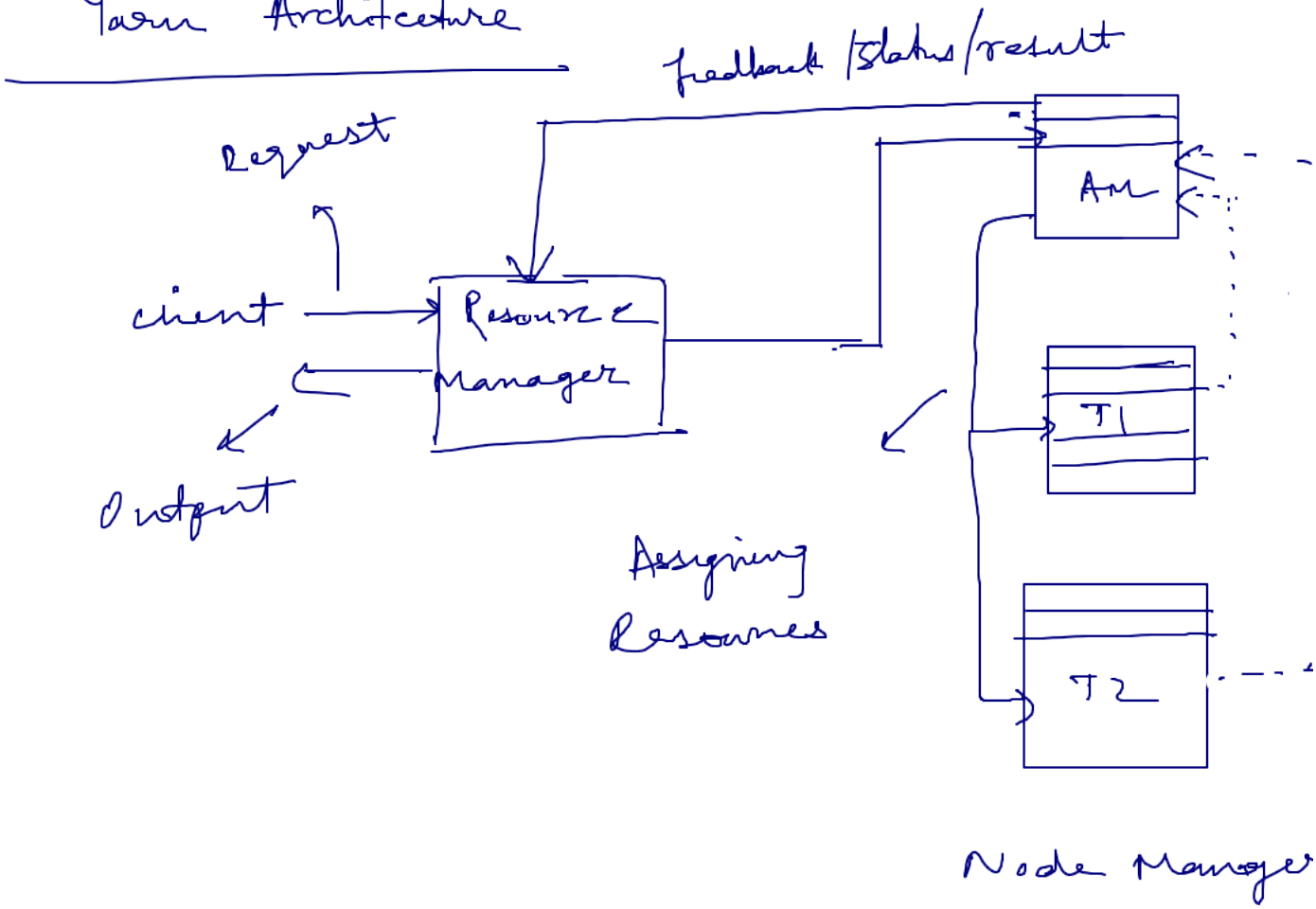
the data. The Application master keeps a ch

Page:2

of how much data is processed and feedback is given time to time to Application Master

- The Application Master on receiving feedback propogates it to Resource Manager which in turn to the client.

Yarn Architecture



Q.No : 12)

Write the code example for closure. How closure works?.

Page:1

Example of closure

```
def call_func ( ):
    def ret_message ( concept ):
        return "This is a" + concept
    return concept
```

closure is a python object in which scope is enclosed at the time the object is inst

So, when I call call-func (closure), i will be printed this is a Closure.

Q.No : 13)

Summarize different components of Hadoop Ecosystem.

Page:1

Different components of Hadoop Ecosystem are –

1) Mahout – A python library used for traini machine learning algorithms

2) Pig – A high level language, which is analyze large dataset in its own lan

3) Hive – It is a data warehousing system which facilitates storing data summarization ad hoc queries and analyzing large dataset It has its own database structure lang called HQL (Hive Query Language)

4) Hbase – It is distributed, column orient database, where data is stored in H HDFS supports both map reduce and A queries

5) Sqoop – It is tool used for trans bulk data into HDFS

Page:2

6) Zoo keeper – It is a centralised service information configuration, naming and group various services

7) Ambari – It is a web based tool for provisional, monitor of various hadoop clusters.

Q.No : 14)

Write a PySpark program to count the number of sentences in a text file. The sentences can be descriptive, interrogative, or exclamatory.

Page:1

```
from pyspark import SparkContext

sc = SparkContext()

sc.textFile(file_path).
flatMap(lambda x : x.split('.')).
flatMap(lambda x : x.split('?')).
flatMap(lambda x : x.split('!'))
filter(lambda x : x=='.' || x=='?'

sc.collect()
```

Q.No : 15)

How do you ensure that all the Hadoop environment variables are available for Hadoop installation? How do you create the index for HDFS? How do you check all Ha
running?.

Page:1

to ensure hadoop environment variables a available for hadoop installation. you can the psendo mode. In the psendo mod JVM is created for all modes ensures en variables are running;

The index of Hdfs is created data is fi loaded into the namenode.

The slave daemons keeping updating their respective progress to the master node wh checks they are runing or not, where the master nodes are checked, on the OnAp received by the client and FsImage keep on updating

Q.No : 16)

Write the code snippets for training the als model implicitly and explicitly.

Page:1

Code Snippets for ALS model are —

Explicitly —

```
als = ALS {
        user : 'user_id'
        item : 'item_id'
        ratings : 'ratings_col' }
als.fit()
```

Implicitly

```
als = ALS {
        user : 'user_id'
        item : 'item_id'
      }
als.fit()
```

Q.No : 17)

Differentiate between top(5) and take(5) methods.

Page:1

- The top(5) method is a rdd method that collects the 5 elements from an in decreasing order and prints it

whereas take(5) collects the first 5 ele in an RDD and prints it.

Q.No : 18)

Summarize cogrouped RDD.

Page:1

Cogrouped RDD is a RDD formed by child and Parent RDD. Since RDD's are resilient that is there lineage is which refers to all the transformations in directed Acyclic graph. The Cogrouped is a partioned according to its lineages.

Q.No : 19)

What is hyper parameter tuning? How do you achieve it in Decision trees? .

Hyper parameter tuning refers to making grid and tune hyperparameter to machine le models. Hyperparameters consists of functions sets constraints on training and evaluating mod It is done by ParamGridBuilder() which in conjunction with CrossValidation() or TrainValidat plots the grid.

In Decision Trees, the hyperparamaters a

- max depth
- max bins
- Information Gain
- Impurity measure

So to tune these parameters, following code snippets-

```
from pyspark.ml.hyperparameter.tuning import
    ParamGridBuilder

c = ParamGridBuilder()

c.addGrid('max-depth' : [4,5]),
```

C. addGrid ('max-bins' : [20, 50],

Page:2

c. addCol('impurity' : [a, b]),

c. addCol('infogain' : [c, d]).

in such a way, the hyperparameters
tuned to particular which play a signific
role, in training and evaluating model.

^ Top