Data Science || CS321 || Mid Semester Exam 2021

u19cs012@coed.svnit.ac.in Switch account

Draft saved

Your email will be recorded when you submit this form

* Required

Paper

*

Consider the pseudo-code for MapReduce's WordCount example (not shown here). Let's now assume that you want to determine the average amount of words per sentence. Which part of the (pseudo-)code do you need to adapt?

- (●) map() and reduce()
- ( ) Only map()
- ( ) Only reduce()
- ( ) The code does not have to be changed.

*

Which statement is/are correct regarding MapReduce?

- (●) The term MapReduce refers to two separate and distinct tasks
- ( ) Is the core component for data ingestion in the Hadoop framework.
- ( ) Helps to combine the input data set into a number of parts and run a program on all data parts parallel at once.
- ( ) Is the parent project of Apache Hadoop.

*

Consider Hadoop's WordCount program: for a given text, compute the frequency of each word in it. The input is read line by line. As input, you are given one le that contains a single line of text:
A Krishna Radha Radha

How many Mapper objects and Reducer objects are created? How many calls to map() and reduce() are made?

- ( ) 1 Mapper object, 3 Reducer objects, 3 calls of map(), 3 calls to reduce()
- (●) 1 Mapper object, 1 Reducer object, 1 call of map(), 3 calls to reduce()
- ( ) 3 Mapper objects, 1 Reducer object, 3 calls of map(), 1 calls to reduce()
- ( ) 3 Mapper objects, 3 Reducer objects, 1 call of map(), 1 call to reduce()

*

Consider the following statements:
i) In Hadoop, after each map phase the results are written to HDFS
ii) The shuffle operation does not require any communication among the worker nodes
iii) The combiner operation can improve performance of a MR computation
Mention which statement is/are correct.

- ( ) ii) only
- ( ) i) and iii) only
- ( ) i) and ii) and iii) only
- (•) iii) only

*

Which of the following definitions of complex data types in Pig are correct?

1. Tuple: a set of key/value pairs
2. Tuple: an ordered set of fields.
3. Bag: an unordered set of tuples.
4. Bag: an ordered set of fields.
5. Map: an ordered set of fields.
6. Map: a collection of tuples.

- ○ 3 and 6 are correct
- ○ 2 and 4 are correct
- ◉ 2 and 3 are correct
- ○ 2,4 and 6 are correct

*

Point out the incorrect statement.

- ○ Pig is a high level language.
- ◉ Pig can invoke code in language like Java Only
- ○ YARN enhances a Hadoop compute cluster in many ways
- ○ All of the above

*

Consider the following statements:
i) Data locality is not taken into account when a Map job is scheduled
ii) The user provided map function takes as input a key and a list of values
iii) Map and Reduce are higher-level functions that take as input a collection and a user defined function
Mention which statement is/are correct.

- ○ i) and ii) and iii) only
- ○ i) and iii) only
- ◉ iii) only
- ○ ii) only

*

Consider the pseudo-code for MapReduce's WordCount example (not shown here). Let's now assume that you want to determine the frequency of phrases consisting of 2 words each instead of determining the frequency of single words. Which part of the pseudo-code do you need to adapt?

- ◉ Both map() and reduce()
- ○ Only map()
- ○ Only reduce()
- ○ The code does not have to be changed

*

Which of the following statements is correct?

- (●) Pig is an execution engine that compiles Pig Latin scripts into database queries.
- ( ) Pig is an execution engine that compiles Pig Latin scripts into HDFS.
- ( ) Pig is an execution engine that utilizes the MapReduce engine in Hadoop.
- ( ) Pig is an execution engine that replaces the MapReduce engine in Hadoop.

*

If we want to execute pig in batch mode then which mode is correct from the options below.

- (●) Pig Grunt shell command
- ( ) Pig scripts
- ( ) Pig options
- ( ) All of the mentioned

*

Which of the following is/are not correct with respect to Hive?

- ( ) Hive works well on all files stored in HDFS
- (●) Both A and B
- ( ) Hive provides SQL interface to process large amount of data
- ( ) Hive needs a relational database like oracle to perform query operations and store data.

*

The below example falls into the category of which data type?
Where do you live?
1. India
2. Australia
3. Sri Lanka
4. England

- ○ Interval
- ○ Ordinal
- ◉ Nominal
- ○ Ratio

*

True or false: It is possible to start reducers while some mappers are still running.

- ◉ False
- ○ True

*

The below example falls into the category of which data type?
The number of marks you scored in the GRE.
A. 400-500
B. 500-600
C. 600-700
D. More than 700

- ○ Ratio
- ◉ Interval
- ○ Ordinal
- ○ Nominal

*

How input file is passed to the mapper engine?

- ○ In Key - Value Pairs
- ○ All at Once
- ◉ In Chunks based on Cluster Size
- ○ Line by line

*

The below example falls into the category of which data type?
Rate the movie "The Dark Knight" according to your preference.
1. Excellent
2. Good
3. Satisfactory
4. Unsatisfactory

- ○ Ratio
- ○ Interval
- ○ Nominal
- ◉ Ordinal

*

Which of the following is not a feature of HiveQL?

- ○ Supports indexes
- ◉ Supports joins
- ○ Support views
- ○ Support Transactions

*

What are the factors that lead to data quality issues?
A. Manual data entry errors
B. Lack of complete information
C. Aggregating data from various sources

- (●) A, B and C
- ( ) A and B
- ( ) A and C
- ( ) B and C

*

Which of the following operations does not require any communication with Namenode?

- (●) A client reading a block of data from the cluster.
- ( ) A client reading a file from the cluster.
- ( ) A client writing a file to HDFS.
- ( ) A client requesting the filename of a given block of data.

*

True or False: The input to reducers is grouped by key.

- ( ) False
- (●) True

*

Rahul has a Hadoop cluster with 20 machines with the following Hadoop setup: replication factor 2, 128MB input split size. Each machine has 500GB of HDFS disk space. The cluster is currently empty (no job, no data). Rahul intends to upload 4 Terabytes of plain text (in 4 files of approximately 1 Terabyte each), followed by running Hadoop's standard WordCount1 job. What is going to happen?

- ○ The data upload fails at the first file: it is too large to fit onto a DataNode.
- ○ The data upload fails at a later stage: the disks are full.
- ○ WordCount fails: too many input splits to process.
- ◉ WordCount runs successfully.

*

Which of the following is the advantage of a distributed database over a centralized database?

- ◉ Module wise growth
- ○ Simplicity and ease
- ○ Slow Response
- ○ None of the above

*

Can be described as inconsistencies and uncertainty in data

- ○ Volatility
- ○ Value
- ○ Vulnerability
- ○ Variability
- ◉ Veracity
- ○ Variety

*

What are the three important aspects of Netflix's Recommendation Engine?

- ○ History of films and TV Series, History of User on Netflix, Taggers who tag content
- ○ History of films and TV Series, History of User on Netflix, Machine Learning Algorithm
- ○ History of User on Netflix, History of films and TV Series, Taggers who tag content
- ◉ History of User on Netflix, Taggers who tag content, Machine Learning Algorithm

*

Which of the following is an example of qualitative data?

- ○ Apple's market capitalization is $1 trillion
- ○ There is a 10% increase in revenue with the inclusion of a new iPhone
- ○ The new Apple's iPhone costs Rs.100000
- ◉ New iPhone is the best iPhone ever

*

True or False: When $HIVE_HOME/bin/hive is run without -f option, it enters script mode.

- ◉ True
- ○ False

*

Which of the following is not an advantage of replication?

- ○ Reduced network traffic
- ○ Each transaction may proceed without coordination across the network.
- ◉ Each site must have the same storage capacity.
- ○ If the database fails at one site, a copy can be located at another site.

*

Consider the following statements:
i) In HDFS there is a single node storing the file system metadata, but many nodes storing the data (file content)
ii) HDFS provides a random access interface for writing to an existing file, i.e., we can overwrite any part of an existing file in HDFS
iii) Fault tolerance in HDFS is achieved through replication
iv) Reads in HDFS are slower than writes

Mention which statements are correct.

- ( ) ii), iii) and iv) only
- (•) i), iii) and iv) only
- ( ) i) and ii) only
- ( ) i) and iii) only

*

What are the three main concepts of Data Science?

- ( ) Data, Science, and Knowledge
- ( ) Programming languages, Probability, Machine learning concepts
- (•) Mathematics, Computer Science, and Domain Expertise
- ( ) Machine Learning, Data Processing, and Statistical Research

*

LinkedIn uses which recommendation engine to recommend new friends, groups, and other social connections?

- ○ None
- ◉ Collaborative filtering
- ○ Content based filtering
- ○ Mobile Recommender Systems

*

Which of the following statements is not correct with respect to outliers?

- ○ Outliers have an effect on regression parameters.
- ◉ Influential cases will always show up as outliers.
- ○ Outliers have an effect on the mean.
- ○ Outliers are values very different from the rest of the data.

*

A homogenous distributed database considered as?

- [ ] A different DBMS is used at each location and data are not distributed across all nodes
- [x] The same DBMS is used at each location and data are distributed across all nodes.
- [ ] A different DBMS is used at each location and data are distributed across all nodes.
- [ ] The same DBMS is used at each location and data are not distributed across all nodes

*

Which of the following is the feature of distributed databases?

- [ ] Totally centralized at one location and accessed by many sites
- [ ] Totally or partially at one location and distributed at many sites
- [x] Partitioned into segments at different sites
- [ ] All of the above

*

Consider the following statement and identify whether that statement is correct or not.
Statement: The main duties of task tracker are to break down the receive job that is big computations in small parts, allocate the partial computations that are tasks to the slave nodes monitoring the progress and report of task execution from the slave.

- ○ False
- ◉ True

*

True or false: Each mapper/reducer must generate the same number of output key/value pairs as it receives on the input.

- ○ True
- ◉ False

*

Improving the quality of data is required before analysis because

A. Bad quality data leads to misleading information
B. To improve the accuracy of the model
C. It is the second step in the data science process

- ○ A, B and C
- ○ B and C
- ○ A and C
- ◉ A and B

Numerical answer type: Write your answer in Digit (Numeric) format. *

Write down output in numeric form.
Rahul has got the data of runs scored by a batsman as 1, 12, 29, 8, 13, 13, 14, 66, 28, 20, 32, and 38. Can you help Rahul to find the outlier?

38

*

The time it takes for a Hadoop job's Map task to finish mostly depends on:

- ○ the duration of the job's Reduce task
- ○ the duration of the job's shuffle & sort phase
- ◉ the placement of the blocks required for the Map task
- ○ the placement of the NameNode in the cluster

*

True/False: When we read a file for processing big data through mapreduce then The splitting parameter can be anything, e.g. splitting by space, comma, semicolon, or even by a new line ('\n').

- ◉ True
- ○ False

*

Can be described with extracting useful data

- ○ Volume
- ○ Veracity
- ○ Variety
- ◉ Value

*

Let's find a course that generates a huge amount of data. They are generating huge amounts of sensor data from different courses which were unstructured in form. They moved to the Hadoop framework for storing and analyzing data. What technology in the Hadoop framework can they use to analyse this unstructured data?

- ○ RDBMS
- ○ None of these
- ◉ MapReduce programming
- ○ Hive

*

Which of the following is quantitative data?

A. Sergio Agüero scored 21 goals in the last premier league

B. Sergio Agüero hair colour is black

C. Sergio Agüero won Golden Boot only once

- ◉ A and C
- ○ A, B and C
- ○ A and B
- ○ Only C
- ○ B and C

*

From the options below, which one is an example of primary data?

- ○ News paper
- ◉ Census report
- ○ Journal
- ○ Book

*

An autonomous homogenous environment is considered as

- ○ The same DBMS is at each site and a central DBMS coordinates database access
- ◉ The same DBMS is used at each site and each DBMS works independently
- ○ A different DBMS is at each node and each DBMS works independently
- ○ None of these

*

Consider the following statements:
i) Apache Pig was developed as a research project at Facebook.
ii) Task tracker is the MapReduce component on the slave machine as there are multiple slave machines.
iii) YARN enhances a Hadoop compute cluster in many ways

- [ ] i), ii) and iii) are false
- [x] Only ii) and iii) are true
- [ ] Only ii) is true
- [ ] i), ii) and iii) are true
- [ ] Only i) and ii) are true

*

Raw data can be cleaned only one time?

- [ ] False
- [x] True

*

Employee ID	Name	Gender	DOB	Team	Salary
QI08	Nitin	Male	15-10-1990	Marketing	$68000
QI52	Rakesh	Male	07-03-1991	Marketing	$45000
QI52	Rakesh	Male	07-03-1991	Sales	$65000
QI52	Rakesh	Male	07-03-1990	Sales	$65000
QI41	Rohan	Male	05-05-1988	Sales	$70000
QI70	Jay	Male	03-08-1990	Product	$70000
QI20	Kshitij	Male	15-05-1988	Marketing	$72000
QI65	Madhuri	Female	24-04-1995	Finance	$80000
QI85	Prachi	Female	10-05-1998	Sales	$60000

Which of the given options is the best approach to deal with duplicate records in the above data set?

- ○ None of these
- ◉ If available cross verify with the alternate source of data and keep the correct one
- ○ Remove all duplicate values
- ○ Randomly remove two of them

*

Which of the accompanying methodology ought to be utilized to ask a Data Analysis inquiry?

- ○ None of the mentioned
- ○ Find out answer from dataset without asking question
- ◉ Find out the question which is to be answered
- ○ Find only one solution for particular problem

*

In the case where data contains a lot of outliers in a particular range, then which of the following is the best option for imputing missing values.

- ○ Mean
- ○ Mode
- ◉ Median
- ○ Any of them

*

Which node is used to arbitrate resources among all the applications in the system.

- ○ NodeManager
- ◉ ResourceManager
- ○ ApplicationMaster
- ○ All of the above

*

Which of the accompanying organizations would be least use of the Recommendation Engine?

- ( ) Ola
- ( ) Instagram
- (•) WhatsApp
- ( ) Facebook

*

Consider the search engine log. We want to learn how much time on average users spend on clicked URLs. We use a resolution of one minute (i.e. all URLs viewed between 1 and 60 seconds by a user are counted together, all URLs viewed between 61 seconds and 120 seconds are counted together, etc.). Dwell times above 24 hours are counted together (as data noise). A Hadoop job is written: the mapper outputs as key/value pair (*,[dwell-time]) for each query log line that contains a click (the value is the actual dwell time). The reducer uses local aggregation:

```
setup():
--- H = associative_array;
reduce(key k, values v):
--- foreach value v in values:
------ H{v}=H{v}+1;
cleanup():
--- foreach value v in H:
------ EmitIntermediate(v,count H{v});
```

What happens if this Hadoop job is started with a query log containing 10 billion lines?

- ○ The Hadoop job crashes and reports an out-of-memory-error.
- ○ The Hadoop job does not compile - hashmaps cannot be used in the Reducer.
- ◉ The Hadoop job runs without an error and outputs the expected results.
- ○ The Hadoop job runs without an error but outputs nothing.

*

Consider the following statements:
i) More complex computations in MR can be written by connecting multiple MR jobs into a more complex workflow
ii) No more than one map job will run on each node of a Hadoop cluster

- (●) i) is correct
- ( ) i) and ii) both are incorrect
- ( ) ii) is correct
- ( ) i) and ii) both are correct

*

If you are predicting the winning team of an IPL game, what would be the first solution step you would start with? Assume you have collected and pre-processed the data.

- ( ) None of these
- (●) Analyse the past performance of the teams
- ( ) Analyse the players' skills in each team
- ( ) Analyse the player's past performance

*

You have a google chrome search engine's query log of the form:
[userid],[timestamp],[query],[url],[click],[dwell-time]
where the userid is the IP address of the user and the timestamp is the time at which the action took place (in epoch). The user either submits a query to the search engine (in which case the query contains the query string submitted, otherwise this value is empty) or views a result list and clicks on a URL (stored in click, otherwise this value is empty). If the user clicked a URL, the log also shows the amount of time in seconds the user spent on the URL (the dwell-time). A mobile example of a query log is the following:
100000,1417789177,The Godfather,,,,
100000,1417789245,The Godfather,,,,
101112,1417712245,The Godfather: Part II,,,,
100000,1417789247,,https://en.wikipedia.org/wiki/The_Godfather
101112,1417712111,The Godfather: Part II,,,,
101112,1417712121,,https://en.wikipedia.org/wiki/The_Godfather_Part_II,1
101112,1417712240,,https://en.wikipedia.org/wiki/The_Godfather_(novel),987

Here, user 100000 searches twice for English movie "The Godfather" and later clicks on the Wikipedia link and spends 25 seconds on it. The user 101112 searches for "The Godfather: Part II" and visits Wikipedia for 1 second and then spends a lot of time on the official Tom Jones website. The log is neither sorted by userid nor by time. We want to know the number of unique queries submitted. To do this, we need to write a Hadoop job that contains:

- ○ only a Reducer and a Counter
- ○ only a Mapper
- ◉ only a Mapper and a Reducer
- ○ only a Mapper and a Counter

*

A bowler had the following scores after 5 games:
296, 305, 297, 395, and 302.
How much does the bowler's mean score increase if the outlier is considered, compared to if the outlier is not considered?

- ○ 13
- ◉ 19
- ○ 16
- ○ 22

Write your answer in small letter and without special characters and spaces. *

How does HDFS detect that a data node has crashed? Just name the mechanism that is used.

heartbeat

*

Consider the following statements:
Statement 1: The Job Tracker is hosted inside the master and it receives the job execution request from the client.
Statement 2: YARN is highly scalable

- ○ Only statement 1 is true
- ○ Only statement 2 is true
- ◉ Both statements are false
- ○ Both statements are true

*

Apache Hadoop YARN stands for:

- (●) Yet Another Resource Negotiator
- ( ) Yet Another Resource Manager
- ( ) Yet Another Reserve Negotiator
- ( ) Yet Another Resource Network

*

What data is stored in a HDFS NameNode?

- ( ) Blocks and heartbeat messages
- ( ) Blocks and block locations
- (●) Filenames, block locations
- ( ) Filenames, blocks and checksums

Back Submit Clear form