Hope Foundation's

Finolex Academy of Management and Technology

Ratnagiri, Maharashtra -415639

Big Data Analysis

Assignment I - Open Notebook Assignment

YEAR/SEM: *BE* **IT/VIII**

DATE: 05/02/2020

MARKS: 10

Batch 2

Name: Kazi Jawwad A Rahim

Q1	State different types of NOSQL Databases	[2]	M2	R	CO812
Q2	What is Hadoop? Explain NameNode and Datanode?	[2]	M2	U	CO812
Q3	What database is facebook and Amazon are using?	[2]	M2	U	CO812
Q4	What is the function of Inputformat class?	[2]	M2	A	CO822
Q5	What are the Five V's of Big Data System ?	[2]	M1	R	CO811
Q6	What is commodity Hardware ?	[2]	M2	R	CO812
Q7	Differentiate between BASE and ACID	[2]	M2	A	CO812
Q8	How Hadoop identifies straggling nodes?	[2]	M2	R	CO812
Q9	What is a Document Store? Explain with example.	[2]	M1	R	CO812
Q10	Explain Mapreduce as applied to wordcount	[2]	M3	A	CO813

07/02/2020

Ms. Priyanka Subhash Bandagale

Assistant Professor

10/10

Q.1 →

Types of NoSQL databases -

1) Key-value stores: These are least complex of the NoSQL databases. They are collection of key-value pair.

2) Column-oriented database: These database store each column separately, allowing for quicker scans when only a small number of columns are involved.

3) Document-store: A document store does assume a certain document structure that can be specified with a schema.

4) Graph store: Graph store is geared toward storing relations between entities in an efficient manner. It have two main component
edge - It can have a direction
node - The entities themselves.

Q.2 →

Hadoop - Hadoop is an opensource f/w of the Apache foundation. It is a f/w written in Java It is opensource s/w f/w for big data computation. The hadoop f/w can store huge amount of data by dividing it into blocks and string across multiple computers and computations can be run in parallel across multiple machines.

Name Node: It is a master that contains metadata. In general it maintains directories and files and manages block present on data node.

Data Node: These are the slaves which provides actual storage and deployed on each machine. They are responsible for processing read & write request of the client. It handles block storage and maintains multiple values block integrity.

Q.3 →

The Facebook and Amazon are using key-value store type of NoSQL database. In this type the data is stored in key-value pairs.

Q.4 →

The InputFormat class is one of the fundamental classes in the Hadoop Map Reduce framework. This class is responsible for defining two main things –

a) Data Split: It is fundamental concept in Hadoop Map Reduce f/w which defines both the size of individual map tasks and its potential execution server.

b) Record reader: It is responsible for actual reading records from the input file and submitting them to the mapper.

Q.5 →

5 V's of Big Data:

i) Volume: Volume is a huge amount of data as the name Big Data itself is related to a size which is enormous.

2) Velocity: It refers to the high speed of accumulation of data. In Big Data velocity, data flows in from sources like machines, networks, social media, mobile phones, etc.

3) Variety: It refers to nature of data that is structured, semi-structured and un-structured. It also refers to heterogeneous sources.

4) Veracity: It refers to inconsistencies and uncertainity in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.

5) Value: Data in itself is of no use or importance but it needs to be converted into something valuable to extract information. Hence, you can say that Value! is the most important V of all the 5 V's.

Q.6

→ Commodity Hardware: Computer hardware that is affordable and easy to obtain. Typically it is a low-performance system that is IBM-PC-compatible and is capable of running Microsoft Windows, Linux or Ms-Dos without requiring any special devices or equipment.

Q.7

→ Document Store: A document-store is a type of NoSQL database. It does assume a certain document structure that can be specified with a schema. In key-value

store data can't be indexed. Hence, searching is not possible. This problem is resolved in document store. They uses a tree structure and they are designed to store everyday document as it is ~~simple~~ and they allowed for complex queries and calculations.
e.g: Mongo DB, Couch DB, etc.

Q.7

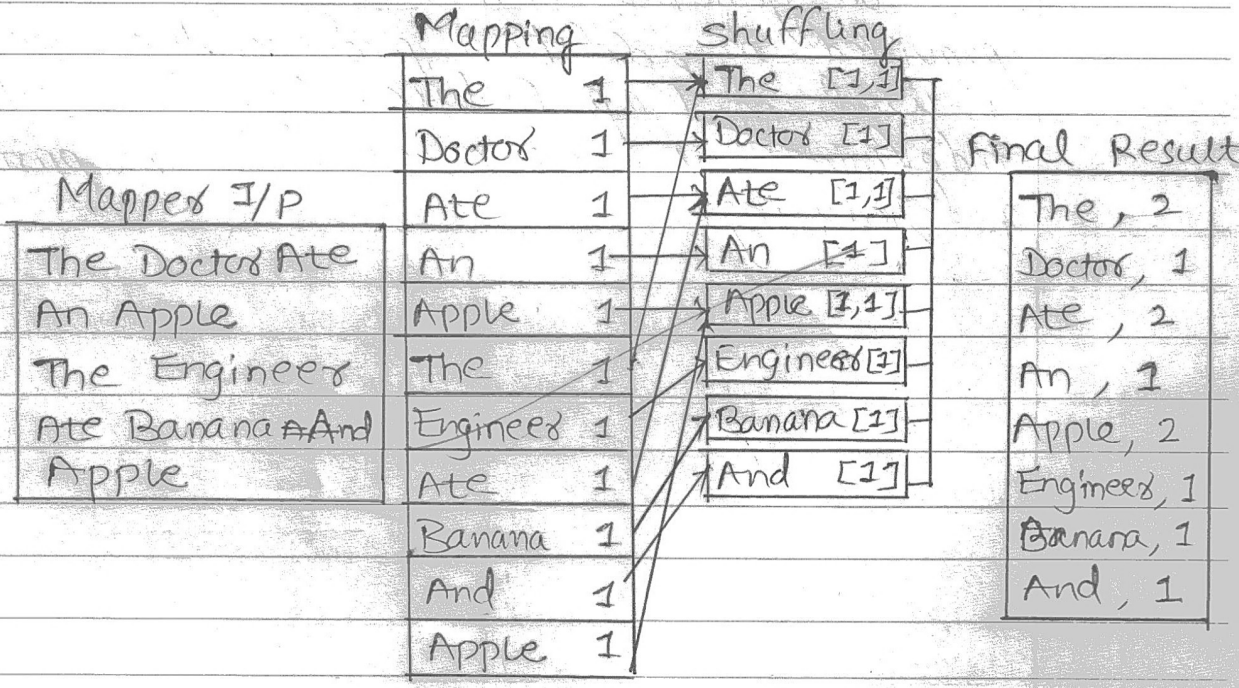
BASE	ACID
1) Weak consistency.	1) Strong Consistency.
2) ~~Isolat~~ Availability first.	2) Isolation
3) Best effort	3) Focus on "commit"
4) Approximate answers.	4) Nested transactions.
5) Aggressive	5) Conservative.
6) Easier evaluation	6) Difficult evaluation.

Q.8

Namenode periodically receives a heartbeat and a block report from each datanode in the cluster. Every datanode sends heartbeat message after every 3 seconds to namenode. The health report is just information about a particular datanode that working properly or not. In the other words we can say that particular data node ~~is~~ that is ~~working prope~~ alive or not. In the other words we can say that particular datanode contains information about all the blocks on that resides on the corresponding datanodes. When namenode doesn't receive any heartbeat message for

10 min from a particular datanode then corresponding datanode is considered dead or failed by namenode. Since, blocks will be under replicated, the system starts the replication process from one node to another by taking all block information from the block report of corresponding datanode. The data for replication transfers directly from one data-node to another without data passing through namenode

Q.10 →



Standard example to obtain MapReduce is to count no. of words in document is an element. The map function has to read document and break it into sequence of words. Each word is counted as 1. I/p can be a repository or collection of document & o/p is no. of occurances of each word, and single Map task can process all the docs in one or more chunks. The reduce function

will add up all values in list. The o/p of reduce task is sequence of word, V where word is the key that appears at least one among all i/p docs & V is total no. of times word has appeared among all those documents.

To optimize map process if particular word w appears k times among all docs assign to process there will be k times (word, 1) key-value pair as a result of Map execution, which can be generated into single pair i.e., (word, k) provided in addition process. The word count is used in Map Reduce algorithm, with all indefiniate values occured in mapping, shuffling and reducing.

★★★