Date:

Syllabus: BR16 Class: BE Seat No: 7278650
Sem: VII Reg No - 2017C02 Sub: BDA

Q1 MCQ

1) option: B : High level of Communication exists between the various nodes.

2) Option: A : Ordered

3) Option: D : Mongo DB

4) option: C : Data Node

5) Option: B : 8

6) option: B : 2

7) option: A : Cold start

8) Option: D : Biased reservoir Sampling

9) Option: C : Microsoft instant Messenger

10) option: C : Periodic

Date:

Syllabus: R16 Class: BE Seat No - 7278650
Sem: VII Sub: BDA

Q2 A

2) Hadoop Core Components

→ ① Hadoop has a master-slave topology
② In this topology, we have one master node and multiple slave nodes.
③ Master node's function is to assign task to various nodes and manage resources. The slave nodes do the actual computing.
④ Slave node stores the real data whereas on master we have meta data.

Hadoop Distributed file System (HDFS)

① HDFS is based on Google file system (GFS)
② HDFS runs on clusters on commodity hardware.
③ The file system has several similarities with the existing distributed file systems.

HDFS follows the master-slave architecture and it has following core elements.

① ~~Name Node~~ Namenode:
i) It is a deamon which runs on master node of hadoop cluster
ii) There is only one namenode in the cluster.
iii) It ~~Store~~ Contains metadata of all the files stored on HDFS which is known as namespace of HDFS
iv) It maintain two files ie. Edit Log & FsImage.
v) Edit log is used to record every change that occours to file system metadata
vi) FsImage stores entire namespace, mapping of blocks to Files and file system properties.

R16 / BE / sem: VII / Sub: BDA / seat: 72780650

(2) Datanode:

i) It is a deamon which runson slave machines of Hadoop
ii) There are number of data nodes in a Cluster.
iii) It is responsible for serving read/write request from the Clients. It also performs block creation, deletion, and replication upon instructions from the name node.
iv) It also sends a Heartbeat message to the namenode periodically about the blocks it hold.

(3) BLOCK:

i) generally the user data is stored in the files of HDFS
ii) The file in a file system will be divided into one or more segments andor stored in individual data nodes.
iii) This file segments are called as Blocks.

(4) Mapreduce:

i) Map reduce is a Softwar framework
ii) Map reduce is a data processing layer of HADOOP
iii) It is a softwar framework that allows you to write applications for processing a large amount of data.
iv) In map reduce a application is brokendown into number of small parts are also called as fragments.
v) These blocks can run on any node in the cluster.
vi) Data processing is done by mapreduce.
vii) Mapreduce core functions are
   a) Read input
   b) Function mapping
   c) Partition, Compare & sort
   d) Function reducing
   e) Write Output

R16 | BE | sem-VII | Sub: BDA | seat : 7278650

HADOOP Ecosystem Components.

(1) Core Components of Hadoop Ecosystem is nothing but the different Components that are ~~built~~ built on the hadoop platform directly.

(2) Hadoop ecosystems are.

a) HDFS
b) MapReduce
c) HIVE
D) Pig
e) HBase
F) ZooKeeper
g) Sqoop

Date:

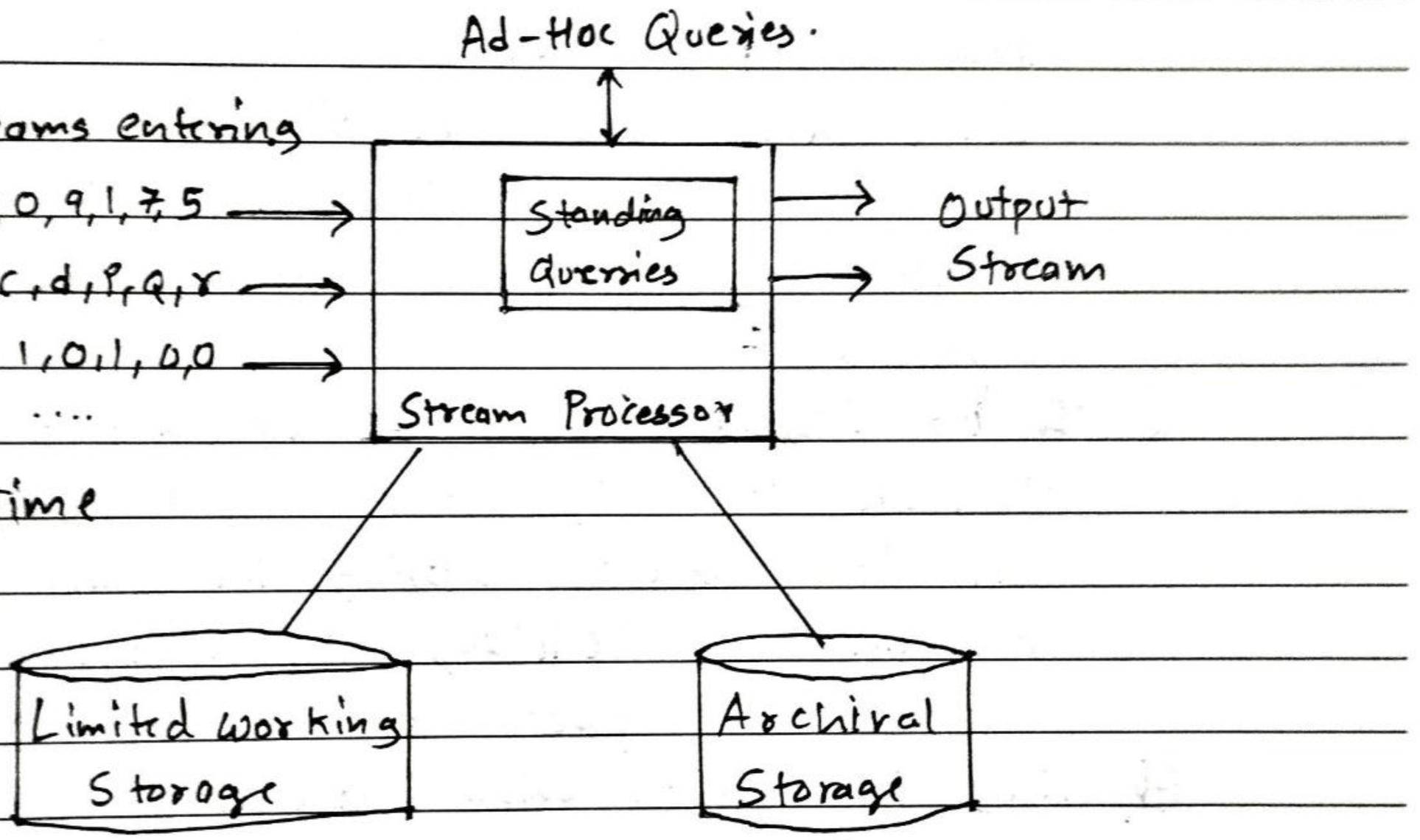
R16/BE/Sem-VII/2017C02/Sub:BDA/Seat:72786850

Q2B

2) Data Stream Management System.

→ ① For handeling and Controlling the data streams, we require a Concrete, Standardized model or framework so that data Stream will be processed in a properly defined manner.

② Block diagram of data stream management System.



③ The data-stream management System architecture is very similar to that of conventional relational database management System architecture.

④ The basic difference is that processor block or more specifically a query processor is replaced with specialized block known as stream processor.

⑤ The first block in the system architecture shows the input.

⑥ The number of data stream generated from different sources will enter into the System.

Date:

R16 | BE | Sem - VII | 2017-02 | Sub: BDA | Seat: 2278650

⑦ Every data stream has its own characteristics such as:
a) Every data stream can schedule and rearrange its own data items
b) Every data stream involved ~~in heterogeneity~~ heterogeneity ie. in each data stream, we can found different kinds of data such as a numerical data, alphabets, alphanumeric data, graphics data, textual data, binary data or any converted transformed or translated data.
c) Every data stream has different input data rate.
d) No uniformity is maintained by the elements of different data streams while entering into the stream processor.

⑧ In the second block of architecture, it is abstracted that there are two different subsystems exist One of which will take care of storing the data stream and other responsible for fetching the data stream from secondary storage and processing it by loading into main memory.

⑨ Hence the rate at which ~~Stre~~ stream enters into the system is not the burden which is involved in the stream processing.

⑩ the third block represents the active storage or working storage for processing the different data streams.

⑪ The working storage area may also contain sub-streams which are integral part of main core stream to generate result for a given query.

⑫ Working storage basically a main memory but situation demands then data items within stream can be fetched from secondary storage. Working storage has limited size.

⑬ The fourth block of system architecture is known as archival ~~storag~~. storage

Date:

R16 | BE | Sem-VII | 2017(02) | Sub: BDA | Seat-72778650

(14) As name indicates, the block is responsible for maintaining the details of every transactions within the System.
(15) It is also responsible to maintain edit logs.
(16) Edit logs are nothing but updation of data
(17) The fifth block is responsible for displaying or delivering the Output stream generated as a result of processing done by the stream processor. Usually by taking the support of working Storage and Occasionally by taking Support of archival Storage.

	DBMS	DSMS
(1)	DBMS reffers to Database management System	DSMS reffers to Data Stream management System
(2)	DBMS deals with persistant data	DSMS deals with Stream data.
(3)	In DBMS random data access takes place	In DSMS Sequential data access takes place
(4)	It is based on Query driven processing model ie. pull based model	It is based on Data driven processing modele ie. push based model
(5)	The data Update rate is low	The data update rate is high
(6)	It dose not provide real time service	It provides real time Servics.
(7)	DBMS uses Unbounded disk Store, means Unlimited Secondary Storage	DSMS uses bounded main memory means limited main memory.

Date:

Q3A

1) Hadoop is an open source Software programming framework for storing large amount of data and performing Computation.

① Features of Hadoop
a) Low cost
b) High Computing power
c) Scalability
d) Huge & Flexible Storage
e) Fault Tolerance & Data Protection.

② Limitations of Hadoop
a) issue with small files.
b) Slow processing Speed
c) Latency
d) No real-time data processing
e) Support for batch processing only
f) Not ~~cost-effective~~ Caching.
g) Not ease of Use
h) No Delta iteration.

3) Key Value Store Database

→ ① It is one of the most basic types of NO SQL databases.
② The kind of NoSQL database is used as a collection, dictionaries, associative arrays, etc.
③ Data is stored in ~~key~~ key/value pairs.
④ It is designed in such a way to handel lots of data and heavy load
⑤ Key-Value pair Storage database store data as a hash table where each key is unique, and the value can be a JSON BLOB, String etc.

(6) Example : ~~MBase~~ ~~Cassandra~~ ~~Hypertable~~

a) ~~Ms~~ Azure Table Storage (ATS)

b) Dyanmo DB

Document Database.

(1) Document oriented No-SQL database stores data as key-value pair but the value part is stored as a document

(2) The document is stored in JSON or XML formats.

(3) ~~B~~ Every document contains a ~~pair~~ unique key used to retrive the document.

(4) Key is used for storing, retrieving and managing document oriented information also known as semi structured data.

(5) Examples:

a) MongoDB

b) Couch DB.

Date:

R16|BE|sem-VII|Sub:BDA| Scat No- 7278650 Pg no

Q3 B

2) How dead ends are handled in Page rank

→ A dead end is a web page with no links out. The presence of dead ends will cause the Page Rank of some or all the pages to go to 0 in the iterative computation, including page that are not dead ends.

(2) Dead end can be eliminated befor undertaking a page Rank calculation by recursively dropping nodes with no arcs out.

(3) Note that Dropping one node can cause another which linked only to it to become a dead end So the process must be recursive.

(4) There are two approaches to deal with dead ends.

a) We can drop the dead ends from the graph and also drop their incoming arcs. Doing so may creat more dead ends which also have to be dropped. recursively. However eventualy, we wind up with a Strongly Connected Component (SCC) none of those nodes are dead ends.

Recursive deletion of ~~dat~~ dead ends will remove part of the out-components, tendril's, and tubes but leave the SCC and in-component, as well as parts of any small isolated components.

b) We can modify the process by which random Surfers are assumed to move about the web. This method which we refer to as 'taxation' also Solves the problem of spider traps.

Here we modify the Calculation of page rank by allowing each ~~random~~ random Surffer a smal probability of teleporting to a random page rather than

following an out-link from their current page.
The iterative step where we compute a new vector estimate of Page Rank V' from the current page Rank estimate V and the transition matrix 'M' is

$$V' = \beta M v + \frac{(1-\beta)e}{n}$$

Where, β is the Chosen Constant usually in the range of 0.8 to 0.9.
e is a vector of all 1's with the appropriate number of nodes in web graph.
n is the number of nodes in webgraph.

If we donot dead ends in the graph then probability of introduction of New = Probability of not to user to a given web page Choose out-link for Current page by same user.

Another possibility is user will not be able to move to any page as $(1-\beta)e/n$ term is independent of ΣV ie. when we don't encountered with dead ends then

$\Sigma V < 1$ but $\Sigma V = 0$.

Date:

Pg - 12

R16| BE| Sem-VII| Sub: BDA| Seat No- 7278650

Q4 B

2) Stream Clustering Algorithm

→ ① Stream Clustering Algorithm is also known as BDMO Algo.

② BDMO Algo is follows the concept of 'Counting Ones' method. which means that there is a window of length N on binary stream and it Counts the number of 1s that comes in the last k bits where k ≤ N.

③ The BDMO algorithm uses the bucket with allowable bucket sizes that forms a sequence where each size is twice of the previous size.

④ In the algorithm the number of points represents the size of bucket.

⑤ It doesn't consider that the sequence of allowable bucket size start with 1 but Consider only forming a sequence Such as 2, 4, 6, 8, Where each size is twice the previous size.

⑥ For maintaining the buckets, the algorithm considers the size of bucket with the power of two.

⑦ In addition, the number of buckets of each size is either one or two that form a sequence of non decreasing size.

⑧ The buckets that are used in the algorithm, Contains the size and timestamp of the most recent points of the stream.

⑨ Along with this bucket also contains a Collections of records that represents the clusters into which the points of the buckets have been partitioned.

⑩ This record Contains the number of points in the Cluster, the Centroid or clustroid of the Cluster, and other parameters that are reqired to merge and maintain the Clusters.

Date:

R16 | BE | Sem-VII | Sub: BDA | SeatNo-7278650

(11) The major steps of the BDMO Algorithm are as follows.
a) Initialising buckets.
b) Merging buckets
c) Answering queries.

(12) Initialising Buckets.
1) The algorithm uses the smallest bucket size that is p with power of two
2) It creates new bucket with the most recent p points. for p stream elements.
3) The time stamp of the most recent point in the bucket is the time stamp of the new bucket.
4) After this, we may choose to leave every point in a cluster by itself or performing clustring using an appropriate clustering method.
5) For the initialisation of the bucket using selected clustring methods, it calculates the centroid or clusteroid for the clusters and counts the points in each cluster.
6) All this information is stored and becomes a record for each cluster
7) The algorithm also calculates the other required parameters for the merging process.

(13) Merging Buckets.
1) After the initialisation of the bucket, the algorithm needs to review the sequence of a bucket.
2) If their happens to be a bucket in time stamp or more than N time units prior to the current time then nothing of that bucket is in window.

3) In such case algorithm drapits from the list.
4) In case we have crrated 3 bucket size of p, then we must merge the oldest two of the three buckets
5) In case the merger can create two buckets of size 2p. this may require us to merge bucket of increasing size recurssively.
6) For merging two consecutive buckets, the algorithm needs to perform following steps.
   a) For merging the size of bucket should be twice the sizes of the two buckets to be merged.
   b) The timestamp of the merged bucket is the timestamp of the more recent of the two consecutive buckets.
   c) In addition it is necessary to calculate the parameters of the merged clusters.

(14) Answering Queries.

1) A query in the stream-computing model is a length of a Suffix of the Sliding window.
2) Any algo takes all the clusters in all the buckets that are at least partially within the suffix and then merges them using some method.
3) The answer of query is the resulting clusters.
4) For the clustering of the streams the stream-computing model finds out the answer to the query.
5) During the initialisation, the K-means method is used and for merging the buckets the timestam is used hence the algo is unable to find the set of ~~bucke~~ buckets that covers the last in point.
6) After this the algo genrates the answer ~~to~~ in response to query as the centroids or clusteroids of all the points in selected buckets.