Week 1 Quiz Solution

1. When can data be considered as Big-data?
   a. When the data is so huge that it cannot be stored in RAM
   b. *When the data has expanded to such an extent that a single computing system is unable to store and process it.
   c. When the data cannot be stored in a file
   d. When two files are needed to store the data

2. The default Input-split size in Hadoop 2 is:
   a. 64 MB
   b. *128 MB
   c. 256 MB
   d. 512 MB

   Explanation: Hadoop 2 has 128 MB, Hadoop 1 has 64 MB

3. Which Node holds the metadata?
   a. *Name Node
   b. Data Node
   c. Edge Node
   d. Meta Node

   Explanation: NameNode stored Metadata in form of FSImage and Edit logs.

4. Which Node holds the actual data and in what form?
   a. Name Node in the form of tables
   b. Data Node in the form of tables
   c. Name Node, in the form of blocks
   d. *Data node, in the form of blocks

   Explanation: Actual data is stored on DataNode in form of data block.

5. How do we know about Data Node failure?
   a. Checkpointing
   b. *Failure of Heart beats to NameNode for 30 seconds.
   c. Failure of Heart beats to NameNode for 10 seconds

Week 1 Quiz Solution

d. Secondary Name Node

Explanation: DataNode sends heartbeat to NameNode in every 3 secs. If NameNode does not receive 10 heartbeat (i.e., 3 * 10 secs) then the DataNode is considered failed.

6. Data node failure is handled by?
   a. *Replication-factor.
   b. Checkpointing
   c. Block Report
   d. Secondary NameNode

   Explanation: If the DataNode fails then the data block which it holds are created on another DataNode. So, the replication factor is always maintained. If the failed DataNode comes back then it's treated as fresh node and used for newer data blocks.

7. Why is the NameNode no longer a single-point of failure in Hadoop 2?
   a. Replication-factor.
   b. Block Report
   c. Hear beat
   d. *Checkpointing

   Explanation: Checkpointing is a process that takes an fsimage and edit log and compacts them into a new fsimage. This way, instead of replaying a potentially unbounded edit log, the NameNode can load the final in-memory state directly from the fsimage. This is a far more efficient operation and reduces NameNode startup time.

8. Metadata is also known as?
   a. Actual Data
   b. *Block Mapping Information
   c. Log Data

   Explanation: Metadata stores which data block is stored on which DataNode hence it can be called as block mapping information

Week 1 Quiz Solution

9. The Balanced approach of 2-racks and 3-copies in Rack awareness mechanism is adopted to
   a. Maximize Write-bandwidth and Maximize Redundancy.
   b. *Minimize Write-bandwidth and Maximize Redundancy.

   Explanation: Rack awareness reduces write traffic in between different racks by placing write requests to replicas on the same rack or nearby rack, thus reducing the cost of write.

10. Command to rename a file?
    a. Hadoop fs -cp
    b. Hadoop fs -get
    c. *Hadoop fs -mv
    d. Hadoop fs -rename

11. The alternative command for "Hadoop fs -copyToLocal"
    a. *Hadoop fs -get
    b. Hadoop fs -copyFromLocal
    c. Hadoop fs -moveToLocal
    d. Hadoop fs -getToLocal

12. Which of the statements about Hadoop is true?
    a. Hadoop provides High Latency but Low Throughput
    b. *Hadoop provides High Latency and High Throughput

    Explanation: Lets us first understand the term latency and throughput

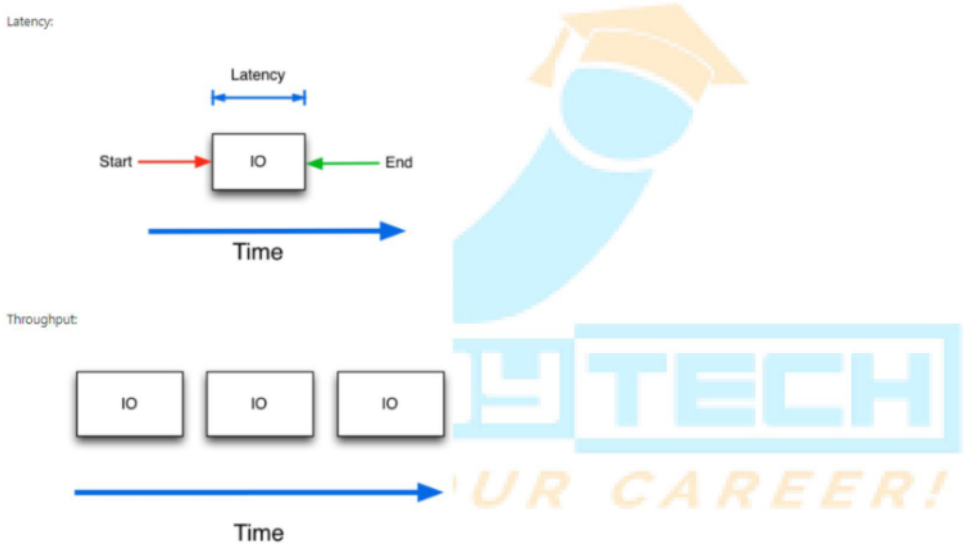
    - Latency is the time required to perform some action or to produce some result. Latency is measured in units of time -- hours, minutes, seconds, nanoseconds or clock periods.
    - Throughput is the number of such actions executed or results produced per unit of time. This is measured in units of whatever is being produced (cars, motorcycles, I/O samples, memory words, iterations etc.) per unit of time.

Week 1 Quiz Solution

Let's us understand using an example

- It takes eight hours to manufacture a car and that the factory produces one hundred and twenty cars per day.
- The latency is: 8 hours.
- The throughput is: 120 cars / day or 5 cars / hour.

Let's us understand using picture



Let's us understand in terms of Hadoop

- High Latency: In case of HDFS, since the request first goes to NameNode and then goes to DataNodes, there is a delay in getting the first byte of data. Therefore, there is high latency in accessing data from HDFS.
- High Throughput: HDFS is optimized for delivering high throughput by using write-once-read-many principle which simplifies data coherency issues and tries to process data locally.

Week 1 Quiz Solution

13. What will happen if the block size in Hadoop cluster is set to 4KB?
    a. Under utilization of cluster
    b. Lesser number of blocks are created
    c. *Over burdening of NameNode
    d. Better parallelism will be achieved

    Explanation: If the block size is set to 4KB then will lead to many data blocks. NameNode stores the metadata and DataNode will always ask NameNode to get next data block address which will over burden the NameNode.

14. Scenario where Hadoop can be a good fit?
    a. *Small number of very big files
    b. Large number of small files

    Explanation: Hadoop stores data in form of data blocks and every file is given new data block.
    For e.g.: 4 files of 20MB each
    4*20=80MB but it will occupy 4 data blocks 4*128=512 MB
    So, for 80MB data it occupies 512 MB which is not good for Hadoop.
    Hence small number of files with large size is good.

15. Suppose, in Hadoop2.0 we have a 750 MB of input file and there are 3 nodes in the cluster, with default replication factor, what will be the total number of blocks generated in HDFS for that file?
    a. 10
    b. 24
    c. 12
    d. *18

    Explanation: File size/ data block size = number of data blocks

    750/128=6

    Default replication factor is 3, 6*3 =18.

Week 1 Quiz Solution

Conclusion: we have total 18 data blocks which can be stored on 3 nodes.

16. How NameNode gets to know if a data block is corrupted?
    a. Heartbeat
    b. *Block Report
    c. Secondary NameNode sends Notification
    d. Metadata

Explanation: All DataNode send a heartbeat and block report to the NameNode in the Hadoop cluster. Heartbeat ensures that the DataNode are alive. A block report contains a list of all blocks on a DataNode.

17. Four important configuration files we have talked about?
    a. CORE-SITE.XML, HIVE-SITE.XML, MAPRED-SITE.XML, YARN-SITE.XML
    b. HIVE-SITE.XML, HDFS-MAPRED.XML, CORE-SITE.XML, MAPRED-DEFAULT.XML
    c. MAPRED-DEFAULT. XML, CORE-YARN.XML, HIVE-CORE.XML, CORE-SITE.XML
    d. *HDFS-SITE.XML, CORE-SITE.XML, MAPRED-SITE.XML, YARN-SITE.XML

18. On which node do you login in Hadoop cluster?
    a. Name Node
    b. Data Node
    c. *Edge Node
    d. Slave Node

Explanation: Edge node is a client-facing machine that has all client tools to operate on a cluster.

Week 1 Quiz Solution

19. What kind of scaling does HDFS supports primarily?
    a. Vertical
    b. *Horizontal
    c. Adaptive
    d. Diagonal

    Explanation: Hadoop adds more nodes to existing cluster which can be done without stopping the existing system

20. The core components of Hadoop are?
    a. *HDFS, MapReduce and YARN
    b. HDFS, HIVE and SQOOP
    c. HDFS, HIVE, PIG and YARN
    d. HDFS, MapReduce, YARN and SPARK

21. How to quickly create an empty file in Linux?
    a. Using cat Command
    b. *Using touch Command
    c. Using ls Command
    d. Using create Command

22. When running on a pseudo distributed mode the replication factor is set to
    a. 0
    b. *1
    c. 2
    d. 3

    Explanation: In pseudo distributed mode we have one machine and hence replication factor is set to 1 because even you set to 3, all 3 copies will be stored on same machine and if that one machine fails, then you can't get the data anyways.

Week 1 Quiz Solution

23. For reading/writing data to/from HDFS, clients first connect to
    a. *NameNode
    b. DataNode
    c. Secondary NameNode
    d. Slave Node

24. When a machine is declared as a DataNode, the disk space in it
    a. Can be used only for HDFS storage
    b. *Can be used for both HDFS and non-HDFS storage
    c. Cannot be accessed by non-Hadoop commands
    d. cannot store text files.

    Explanation: HDFS is on top of LFS hence the disk space can be used by both LFS and HDFS.

25. Which utility is used for checking the health of an HDFS file system?
    a. *fsck
    b. fchk
    c. fsch
    d. fcks

26. Is Apache Spark a replacement of Hadoop?
    a. Yes
    b. *No

    Explanation: Spark is just a computation framework whereas Hadoop provides HDFS(Storage), MapReduce(Computation) and YARN(Resource Management)

27. What does Apache Spark provide?
    a. storage + computation

Week 1 Quiz Solution

b. all things whatever Hadoop core provides
c. *only computation
d. only storage

28. What is the main purpose of Sqoop?
a. importing data from RDBMS to HDFS
b. exporting data from HDFS to RDBMS
c. processing data
d. *both importing & exporting data from RDBMS to HDFS and vice versa

29. What is the role of Secondary NameNode?
a. scalability
b. *fault tolerance
c. both scalability & fault tolerance
d. data backup

Explanation: If NameNode fails then SNN can be made active NameNode. SNN provides fault tolerance for NameNode

30. Command which will take you to your home directory
a. cd /
b. cd -
c. *cd ~
d. cd home

Explanation: cd ~ takes you to home directory and cd / takes you to root directory

31. Your home directory in cloudera local & HDFS respectively
a. /user/cloudera & /home/cloudera
b. /home/cloudera & /user/HDFS
c. */home/cloudera & /user/cloudera
d. /home/cloudera & /home/HDFS

Explanation: Cross check using below commands

Week 1 Quiz Solution

Linux command

cd ~

pwd

Hadoop command

Hadoop fs -ls /

Hadoop fs -ls /user/cloudera

32. Consider you are in your home directory in cloudera local. Then what will this command do

ls ./../cloudera/./../cloudera/../../home/cloudera/./.

a. it will error out
b. *the output is same as ls
c. it will list the files of my HDFS directory
d. it will list the files of both my HDFS & local directory

Explanation: default prompt of home directory is considered.

./ means the current directory
../ means the parent of the current directory, not the root directory
/ is the root directory

33. What will this command do? ls -ltr | tail

a. it will error out
b. display 10 oldest files/folders
c. *display 10 newest files/folders
d. display 20 newest files/folders

Explanation: tail
-t option will sort the entries by modification date (with newest first)
-r will reverse the sorting order.
As -t will sort by modification time with newest first, -r will cause the reverse i.e., oldest entries will be shown first now.

34. let's say I create a file with name file.txt, if I execute wc file.txt what will be the output

a. *number of lines, number of words, number of characters
b. number of lines

Week 1 Quiz Solution

c. number of words
d. None

Explanation: By default wc will give number of lines, number of words, number of characters. If you want specific then mention the option.

35. How to find out how many lines in file contains the word hello?
a. ls file.txt | grep hello | wc -l
b. *cat file.txt | grep hello | wc -l
c. grep hello | wc -l | cat file.txt
d. it will error out