DSCI-551 Midterm1

Chaoyu Li, 96.5/100

QUESTION 1 [JSON & Firebase]

Consider using Firebase to manage student data. Suppose all student data will be stored in ‘https://dsci551.firebaseio.com/students.json’.

(1.a.) Write a curl command to add a new student with id = “100” and name = “john”.

A: curl -X PUT 'https://dsci551.firebaseio.com/students/100.json' -d '{"name": "john"}'

(1.b.) Write a curl command to add a new attribute age = 25 for student 100. Your command should not remove/overwrite existing data.

A: curl -X PATCH -d '{"age": 25}' 'https://dsci551.firebaseio.com/students/100.json'

(1.c) Write a curl command to add a new student with id = “200”, name = “mary”, age = 28.

A: curl -X PUT 'https://dsci551.firebaseio.com/students/200.json' -d '{"name": "mary", "age": 28}'

(1.d.) Write a curl command to find all students who are at least 25 years old. You may assume the index has been created.

A: curl 'https://dsci551.firebaseio.com/students.json?orderBy="age"&startAt=25&print=pretty'

(1.e.) Show the content of students.json in the JSON format.

A:
```
{"students": {
        "100": {
            "name": "john",
            "age": 25},
        "200": {
            "name": "mary",
            "age": 28}}}
```

QUESTION 2 [Storage systems]

Consider a hard disk drive with the following specifications:

- Rotation speed: 10,000 RPM
- Maximum bandwidth: 100MB/sec
- Maximum seek time: 15ms
- Block size: 4KB

(2.a.) Calculate the completion time for reading of 100MB of sequential data (located on the same track). Show your work. (Note: Your answer should be in *milliseconds* unit.)

A: Rotation Time = 60,000ms/10,000RPM = 6ms

Rotational Latency = 6ms/2 = 3ms

Avg Seek Time = 1/3 * 15ms = 5ms

Transfer Time = 100MB/(100MB/sec) = 1000ms

Completion Time = Rotational Latency + Avg Seek Time + Transfer Time = 1000ms + 3ms + 5ms = 1008ms

(2.b.) What is the actual bandwidth of the drive for the above sequential workload? (answer unit => MB/sec)

A: Actual bandwidth = |w|/completion time = 100MB/1008ms = 99.206MB/sec

(2.c.) Calculate the completion time for reading 8MB of random data. Show your work. (answer unit => seconds)

Number of blocks to be transferred = 8MB/4KB = 2048

A: Completion Time = 2048 * (3ms + 5ms +4KB/(100MB/sec)) = 16.46s

(2.d.) What is the actual bandwidth of the drive for the above random workload? (answer unit => MB/sec)

A: Actual bandwidth = 8MB/16.46s = 0.486MB/sec

QUESTION 3 [SSD]

Consider an SSD block with a single block that contains 3 pages: p1, p2, and p3. P1 contains stale data, p2 has valid data, while p3 is free (no data stored there). Now suppose two pages of new data need to be added to the block.

(3.a.) How should the SSD controller proceed? State the sequence of steps the controller takes to complete the writing of the two pages of new data. Assume the controller only read valid pages into buffer when erasing the block.

A: 1. Read 1 page, rearrange page order and remove stale pages(p1)

2. Erase all pages

3. Write 3 new pages

(3.b.) Suppose read latency is 20 microseconds, write latency is 200 microseconds, and erase latency is 2000 microseconds. Compute the total latency of the above workload (adding of two pages of new data). Show your derivation.

A: There are 1 read, 3 writes and 1 erase.

Total Latency = 20µs + 3 * 200µs + 2000µs = 2.62ms

QUESTION 4 [HDFS]

Consider reading a file '/usr/john/students.json' stored with HDFS. Suppose the size of the file is 500MB, and a client needs to read the last 200MB of the file. Suppose the replication factor of HDFS is 2.

(4.a.) What is the default block size of HDFS?

A: 128MB

(4.b.) How many HDFS blocks need to be allocated to store the file?

A: Number of blocks = 500MB/128MB(ceiling) = 4

Because the replication factor is 2, Total number of blocks = 4 * 2 = 8

(4.c.) Which node/server should the client first talk to when reading the file?

A: NameNode

(4.d.) Which RPC call does the client need to make? Describe the input arguments of the call.

A: getBlockLocations(string src, long offset, long length)

The first argument "src" represents the File name(path).

The second argument "offset" represents the offset to start reading.

The third argument "length" represents how much data need to be read.

(4.e.) What are the values of the arguments for reading the last 200MB of the file?

A: getBlockLocations('/usr/john/students.json', 300MB, 200MB);

(4.f.) What does the RPC call return?

It will return an object that represents the located blocks, and it contains data nodes and offsets.

And it will return the location of the replication blocks.

(4.g.) How many HDFS blocks does the client need to read the requested data from?

A: In the best case, the client needs to read two HDFS blocks because the data requested starts from the third block and the client only needs to read the two nearest blocks to get 200MB data.

In the worst case, the client needs to read four HDFS blocks. If the two nearest blocks are broken during the reading, the client needs to read two more blocks to get 200MB because the replication factor of HDFS is 2.

QUESTION 5 [Miscellaneous]

(5.a.) **Explain** 3V's of big data.

A: Volume, Velocity, Variety

Volume:

The increase in data volume comes from many sources including the clinic [imaging files, genomics/proteomics and other "omics" datasets, biosignal data sets (solid and liquid tissue and cellular analysis), electronic health records], patient (i.e., wearables, biosensors, symptoms, adverse events) sources and third-party sources such as insurance claims data and published literature. For example, one whole genome binary alignment map file typically exceed 90 gigabytes.

Velocity:

With overall study complexity on the rise and the need to process more clinical data points in the same or less amount of time, the velocity at which this volume of data is handled is a critical factor. The intention is to use divergent and large data inputs to more rapidly uncover efficacy and safety signals. This output may enable faster submissions to regulatory authorities for promising new drug candidates and earlier decisions regarding go/no go decisions if toxicities are uncovered or help to bracket the scope of an efficacy or safety signal to particular comorbidities/drug interactions/demographics.

Variety:

There is an increasing variety of data sources that are analyzed independently, cross-compared and used for aggregate comparisons across clinical trials. These include structured and unstructured inputs. The added value of collecting and/or combining long-term follow-up data, incorporation of real-world data, clinical testing/diagnostics, lab results, patient apps, literature and available datasets from healthcare, industry and agencies has become evident. The ability of review available data using traditional methodologies in many cases has become antiquated and inefficient. For example, every year, there are over one million biomedical articles published (~2 papers per minute), evidencing the importance of having effective machine-driven systems and technology platforms for both ingesting and digesting this volume of data.

(5.b.) Name three areas/fields that tend to generate a large amount of data.

A: Financial services, healthcare, social media.

(5.c.) Explain atime, mtime, and ctime of a file in a file system.

A: atime stands for access time. It means the timestamps tell when the last accessed.

ctime stands for inode change time. It means the timestamps tell when the inode itself was last modified.

mtime stands for modification time. It means the timestamps tell when the file content was last modified.

(5.d.) Give a Linux command that only changes ctime of a file (no change to atime and mtime). Explain your answer.

A: Command: chmod 400 filename

Explain: "chmod" command only changes the read and write permissions of the file "filename", and neither accesses the file nor modifies the file content. Therefore, only ctime will change.

(5.e.) If a hard disk drive has 16 cylinder, 8 heads, 32 sectors per track, and section size is 4KB, what is the capacity of the drive? Show your derivation.

A: Capacity = Cylinders × Heads × Sectors × sector_size = 16 * 8 * 32 * 4KB = 16MB

(5.f.) Explain why writing in SSD takes much longer than reading?

A: Because writing needs to find sufficient space and allocate it, then the necessary bits need to be forced on by driving a high voltage across the barrier that holds the charge. However, the reading just finds the file location and gets data from each block. Therefore, writing is much slower than reading.

Also need to explain in terms of electron movement near the floating gate and voltage requirements.

(5.g.) Is [{}] a valid JSON value? Explain your answer

A: Yes. Because the JSON value can be an array[], and in an array, it is valid to contain an object{}.

(5.h.) Explain the difference between PUT and POST in the curl commands to Firebase

A: PUT command will write or overwrite data.

POST command will automatically generate a new key then stores the value, so it will not overwrite existing data.