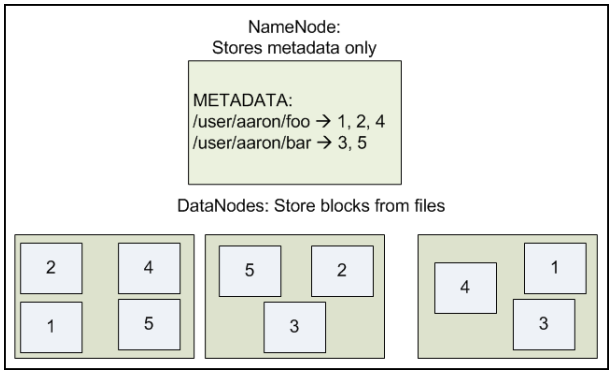
Name: _____ USC ID: _____

Quiz 4: HDFS & File Formats (10 points. 15 minutes)

1. [5 points] Refer to the following diagram on an example HDFS. Answer the following questions.



a. [1 point] What is the replication factor in this HDFS?

Each block has two replicas distributed across three DataNodes.

Thus the replication factor in this HDFS is **2**.

b. [1 point] Which node does the client first contact when reading/writing a file?

Client first contacts **NameNode**, which informs the client of the closest DataNodes storing blocks of the file when reading, and selects DataNodes for holding its replicas when writing.

c. [1 point] What is the typical size of a block in HDFS?

64MB, which is much large than disk block size.

d. [2 points] When writing a file in HDFS, how many packets is each block divided into? What is the size of each packet?

One point for each

One block, which is 64MB, is divided into **1024** packets, each of which is **64KB**.

2. [3 points] Unicode code point for the Chinese character 中(means middle) is U+4E2D. Give its **UTF-8** encoding in both **binary** and **hexadecimal** formats.

U+4E2D is within the range from U+0800 to U+FFFF, denoting that the code sequence length being 3. U+4E2D in binary is **0100 1110 0010 1101**. Encode in the following steps:

1. Take 6 bits at a time backwards from end and add leading **10** to form the last two code units;
2. Add leading **111**, which indicates this code point consists of 3 code units, to the rest 4 bits and 0's to any unfitted spaces (one 0 in this case) to form the first code unit;
3. The binary code will be: **11100100 10111000 10101101**.
4. The hexadecimal code will be: **E4 B8 AD**

0.5 point for each transformation error between binary and hexadecimal formats, and each division and completion errors when forming code units. 2 points for wrong number of code units.

3. [2 points] What is the output of `json.dumps(['foo', {25: ('bar', None, 1.0, 2, False)}])`?

```
["foo", {"25": ["bar", null, 1.0, 2, false]}]
```

0.5 point deducted for each minor error, such as quotation marks and wrong capitalization. 1 point deducted for each wrong data structure.