Big Data Management Systems

Assignment 2

Hadoop

Negin Baghbanzadeh

Question 1

A colleague has suggested using Hadoop to build a banking system for tracking customers' deposits. Give three good reasons why this is a *very bad* idea.

1. One of the main advantages of Hadoop is processing large amounts of data; but tracking customer's deposits does not have that much processing.
2. In Hadoop data is replicated across nodes so availability increases but consistency reduces while consistency is an important factor for customer's deposit tracking.
3. Costumer's deposit data is very sensitive data and needs high security. It may not be possible to achieve needed security with Hadoop.

Question 2

Your colleague was impressed by your arguments above and has come to seek your counsel on another matter. He has heard that graph databases are really cool and wants to know if an analytics system his department is planning should use one as its datastore. Give three good questions you could ask to assess whether this would be an appropriate technology for your associate's application.

1. Is the data of their analytics system complexly-related? (For example the data of social networks is complexly-related) (If yes, graph database is well suited)
2. Does this analytics system use any graph based machine learning systems? (Such as Google's machine learning platform) (If yes, graph database is well suited)
3. Is this analytics system for a recommender system that models relationships between customers and data? (If yes, graph database is well suited)
4. Is this analytics system used for a transactional system? (If yes, graph database isn't a good idea)

Question 3

Your reputation is spreading. A department head wants to know if they should switch from a relational database management system to a column-oriented one such as HBase. You have some questions. Three to be exact.

1. Is their data used for analytics? (for example, business analytics) (If yes, column-oriented database is preferred)
2. How much do you care about using less space while storing the same amount of data? (If less space is important, column-oriented database is preferred)
3. Is their system a transactional system? (If yes, relational database is preferred)

Question 4

A friend is now asking about whether MongoDB would be a good choice for a datastore for a customer relationship management system they're planning to build. Three more good questions, please.

1. Does their data have a uniform structure? (If no, MongoDB is a good choice)
2. How much availability is important to them? (If a lot, MongoDB is a good choice)
3. How much consistency is important to them? (If a lot, MongoDB is not a good idea)

Question 5

Whenever you have replicated data it is impossible to keep the copies exactly synchronized. Why and why is this a problem?

Since we have our data in more than one location, updating all of the copies is expensive and could lead to the database performing poorly.

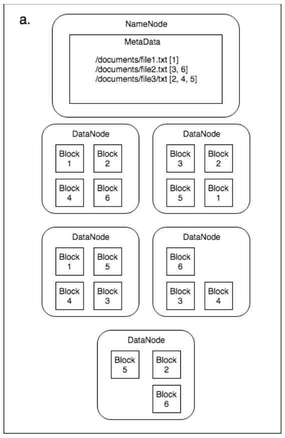
Question 6

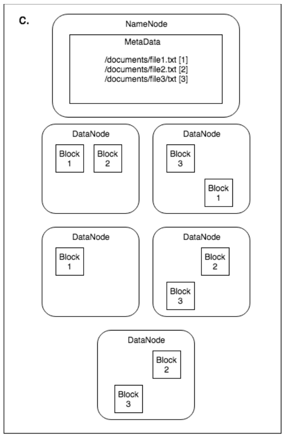
Which diagram(s) best illustrates a HDFS cluster configuration that has a replication factor of 3 for 3 files.

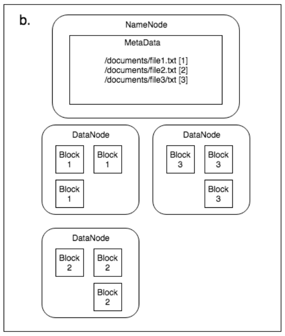
In all the HDFS clustered, there are 3 files but their sizes are different. For example, in diagram ‘a’, the ‘file3’ is split into three blocks.

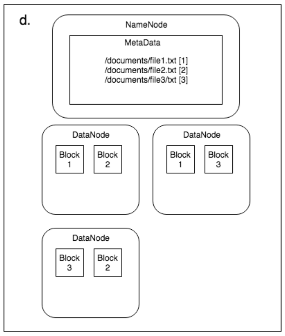
We can see in options ‘a’ and ‘c’, there are 3 copies of each file spread across the DataNodes so their replication factor is 3. In option 2 the replication factor is 2. In option ‘b’ each block has 3 copies but all of them are in one DataNode so the replication factor is not 3.

The right answer is **option “e: a and c”**









e. a and c

f. c and d

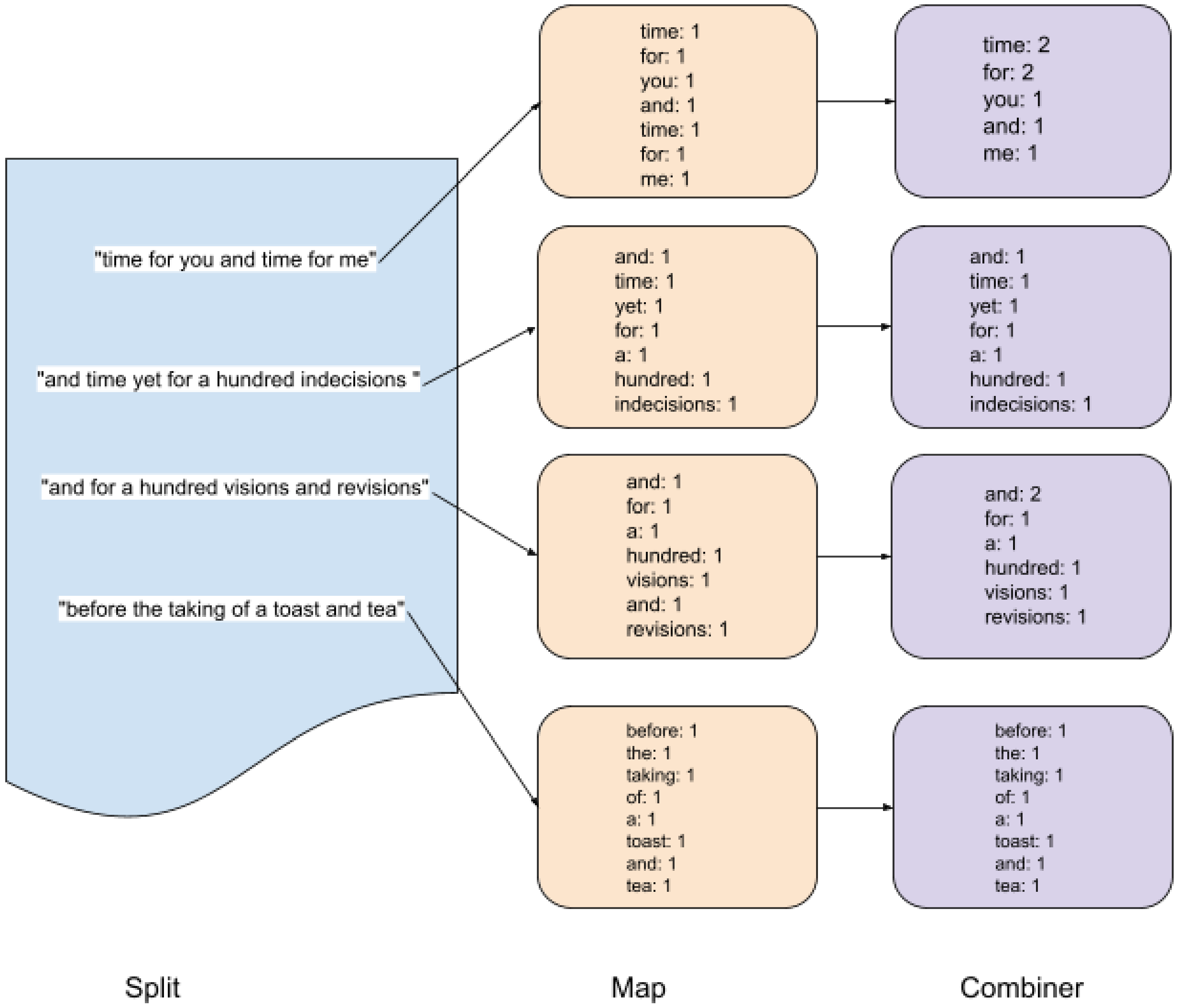
Question 7

Consider the following input text that contains part of a poem called The Love Song of J. Alfred Prufrock BY T. S. ELIOT:

> Time for you and time for me,
>
> And time yet for a hundred indecisions,
>
> And for a hundred visions and revisions,
>
> Before the taking of a toast and tea.

Let’s say we want a word count of this file by creating a MapReduce application. We would split this file into the following four input pairs to four Mapper tasks to consume.

What will the input pairs look like for the Reducer tasks? Assume that combiner tasks are used in the MapReduce application.



The output of the 'combiner', goes into the 'partition' phase and the output of this phase, which is **option 'c'**, will be the input of the 'reducer'.

(time, [2, 1]) (for, [2, 1, 1]) (you, [1])

(and, [1, 1, 2, 1]) (me, [1]) (yet, [1]) (a, [1, 1, 1]) (hundred, [1, 1])

(indecisions, [1]) (visions, [1]) (revisions, [1]) (before, [1])

(the, [1]) (taking, [1]) (of, [1]) (toast, [1]) (tea, [1])