1. Given a Hadoop cluster consisting of 1 namenode and 4 datanodes with a default configuration (chunk size: 128MB, each chunk with 3 replicas), to store a file with the size of 256MB.

a. Explain how the file is divided into chunks, how many chunks does this file have?

→ The file is divided into 2 chunks of 128MB each since the default chunk size is 128MB.

b. How are these chunks written to the datanodes? After the file is stored, show the metadata information of the file on the namenode.

→ Each chunk of 128MB is then replicated 3 times and distributed across the 4 datanodes for reliability and data locality.

→ The metadata on the namenode includes:

- File and chunk namespaces
- File size: 256MB
- Number of chunks: 2
- The locations of each chunk replica on the datanodes:
    - Chunk 1 is located on Datanode 1, Datanode 2, and Datanode 3.
    - Chunk 2 is located on Datanode 2, Datanode 3, and Datanode 4.

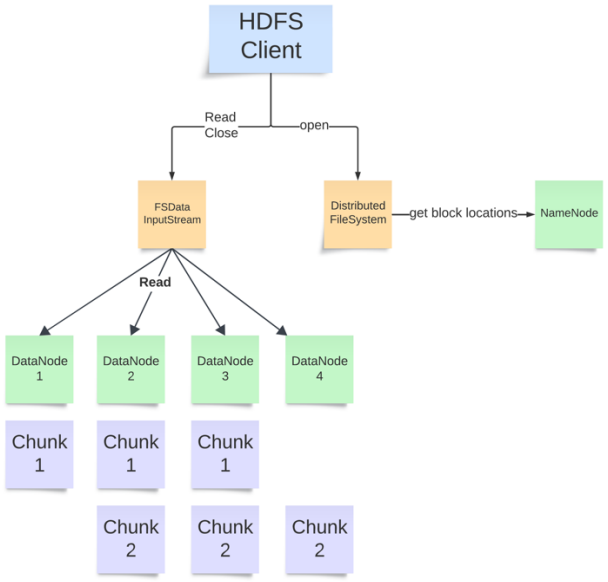
c. How to read the file from the HDFS cluster? Please draw a figure of the above cluster with the namenode and datanodes, then use this figure to answer the above questions.

- The client initiates the file read by calling the open() method which returns a DistributedFileSystem object.
- The DistributedFileSystem asks the NameNode for the block locations of the file. Since the file is 256MB, the file is split into 2 chunks (each 128MB), and each chunk has 3 replicas stored across the 4 DataNodes. The NameNode returns an ordered list of the DataNodes storing these chunks, prioritizing the ones closest to the client. For example:

    Chunk 1: Stored on Datanode 1, Datanode 2, Datanode 3.

    Chunk 2: Stored on Datanode 2, Datanode 3, Datanode 4.

- The DistributedFileSystem provides the client with a FSDataInputStream to read the file.
- The client calls read() on the FSDataInputStream, which reads data from Chunk 1 from one of its replica locations (for example: Datanode 1) and continues reading until Chunk 1 is fully read. Once Chunk 1 is read, the FSDataInputStream switches to Chunk 2, transparently retrieving it from Datanode 2, Datanode 3, or Datanode 4.
- The FSDataInputStream handles the connection switching, data validation (checksum), and error recovery. If an issue occurs while reading from one replica, the stream automatically switches to another replica of the chunk.
- When the client has finished reading, it calls close() on the FSDataInputStream.



2. Mapreduce: Given the following text from Pride and Prejudice, by Jane Austen with 3 sentences (marked by different colors). Please use 3 mappers (one mapper for each sentence) and 1 reducer to count the occurrence of each word in the text (ignore the punctuation signs). Please show your detailed work and results.

"To some the delightful freshness and humour of Northanger Abbey, its completeness, finish, and entrain, obscure the undoubted critical facts that its scale is small, and its scheme, after all, that of burlesque or parody, a kind in which the first rank is reached with difficulty. Persuasion, relatively faint in tone, and not enthralling in interest, has devotees who exalt above all the others its exquisite delicacy and keeping. The catastrophe of Mansfield Park is admittedly theatrical, the hero and heroine are insipid, and the author has almost wickedly destroyed all romantic interest by expressly admitting that Edmund only took Fanny because Mary shocked him, and that Fanny might very likely have taken Crawford if he had been a little more assiduous."

Map 1		**Map 2**	**Map 3**	
(to, 1)	(scale, 1)	(persuasion, 1)	(the, 1)	(by, 1)
(some, 1)	(is, 1)	(relatively, 1)	(catastrophe, 1)	(expressly, 1)
(the, 1)	(small, 1)	(faint, 1)	(of, 1)	(admitting, 1)
(delightful, 1)	(and, 1)	(in, 1)	(mansfield, 1)	(that, 1)
(freshness, 1)	(its, 1)	(tone, 1)	(park, 1)	(edmund, 1)
(and, 1)	(scheme, 1)	(and, 1)	(is, 1)	(only, 1)
(humour, 1)	(after, 1)	(not, 1)	(admittedly, 1)	(took, 1)
(of, 1)	(all, 1)	(enthralling, 1)	(theatrical, 1)	(fanny, 1)
(northanger, 1)	(that, 1)	(in, 1)	(the, 1)	(because, 1)
(abbey, 1)	(of, 1)	(interest, 1)	(hero, 1)	(mary, 1)
(its, 1)	(burlesque, 1)	(has, 1)	(and, 1)	(shocked, 1)
(completeness, 1)	(or, 1)	(devotees, 1)	(heroine, 1)	(him, 1)
(finish, 1)	(parody, 1)	(who, 1)	(are, 1)	(and, 1)
(and, 1)	(a, 1)	(exalt, 1)	(insipid, 1)	(that, 1)

(entrain, 1) (obscure, 1) (the, 1) (undoubted, 1) (critical, 1) (facts, 1) (that, 1) (its, 1)	(kind, 1) (in, 1) (which, 1) (the, 1) (first, 1) (rank, 1) (is, 1) (reached, 1) (with, 1) (difficulty, 1)	(above, 1) (all, 1) (the, 1) (others, 1) (its, 1) (exquisite, 1) (delicacy, 1) (and, 1) (keeping, 1)	(and, 1) (the, 1) (author, 1) (has, 1) (almost, 1) (wickedly, 1) (destroyed, 1) (all, 1) (romantic, 1) (interest, 1)	(fanny, 1) (might, 1) (very, 1) (likely, 1) (have, 1) (taken, 1) (crawford, 1) (if, 1) (he, 1) (had, 1) (been, 1) (a, 1) (little, 1) (more, 1) (assiduous, 1)

Group by Key					
(to, 1) (some, 1) (the, 1) (the, 1) (the, 1) (the, 1) (the, 1) (the, 1) (delightful, 1) (freshness, 1) (and, 1) (and, 1) (and, 1) (and, 1) (and, 1) (and, 1) (and, 1) (humour, 1)	(of, 1) (of, 1) (of, 1) (northanger, 1) (abbey, 1) (its, 1) (its, 1) (its, 1) (its, 1) (completeness, 1) (finish, 1) (entrain, 1) (obscure, 1) (undoubted, 1) (critical, 1) (facts, 1) (that, 1) (that, 1) (that, 1) (that, 1) (scale, 1)	(is, 1) (is, 1) (is, 1) (small, 1) (scheme, 1) (after, 1) (all, 1) (all, 1) (all, 1) (burlesque, 1) (or, 1) (parody, 1) (a, 1) (a, 1) (kind, 1) (in, 1) (in, 1) (in, 1) (which, 1) (first, 1) (rank, 1)	(reached, 1) (with, 1) (difficulty, 1) (persuasion, 1) (relatively, 1) (faint, 1) (tone, 1) (not, 1) (enthralling, 1) (interest, 1) (interest, 1) (has, 1) (has, 1) (devotees, 1) (who, 1) (exalt, 1) (above, 1) (others, 1) (exquisite, 1)	(delicacy, 1) (keeping, 1) (catastrophe, 1) (mansfield, 1) (park, 1) (admittedly, 1) (theatrical, 1) (hero, 1) (heroine, 1) (are, 1) (insipid, 1) (author, 1) (almost, 1) (wickedly, 1) (destroyed, 1) (romantic, 1) (by, 1) (expressly, 1) (admitting, 1) (edmund, 1) (only, 1)	(took, 1) (fanny, 1) (fanny, 1) (because, 1) (mary, 1) (shocked, 1) (him, 1) (might, 1) (very, 1) (likely, 1) (have, 1) (taken, 1) (crawford, 1) (if, 1) (he, 1) (had, 1) (been, 1) (little, 1) (more, 1) (assiduous, 1)

Reducer				
(to, 1) (some, 1) (the, 7) (delightful, 1) (freshness, 1) (and, 8)	(that, 4) (scale, 1) (is, 3) (small, 1) (scheme, 1) (after, 1)	(difficulty, 1) (persuasion, 1) (relatively, 1) (faint, 1) (tone, 1) (not, 1)	(park, 1) (admittedly, 1) (theatrical, 1) (hero, 1) (heroine, 1) (are, 1)	(because, 1) (mary, 1) (shocked, 1) (him, 1) (might, 1) (very, 1)

(humour, 1) (of, 3) (northanger, 1) (abbey, 1) (its, 4) (completeness, 1) (finish, 1) (entrain, 1) (obscure, 1) (undoubted, 1) (critical, 1) (facts, 1)	(all, 3) (burlesque, 1) (or, 1) (parody, 1) (a, 2) (kind, 1) (in, 3) (which, 1) (first, 1) (rank, 1) (reached, 1) (with, 1)	(enthralling, 1) (interest, 2) (has, 2) (devotees, 1) (who, 1) (exalt, 1) (above, 1) (others, 1) (exquisite, 1) (delicacy, 1) (keeping, 1) (catastrophe, 1) (mansfield, 1)	(insipid, 1) (author, 1) (almost, 1) (wickedly, 1) (destroyed, 1) (romantic, 1) (by, 1) (expressly, 1) (admitting, 1) (edmund, 1) (only, 1) (took, 1) (fanny, 2)	(likely, 1) (have, 1) (taken, 1) (crawford, 1) (if, 1) (he, 1) (had, 1) (been, 1) (little, 1) (more, 1) (assiduous, 1)