Data Intensive Computing - Review Questions 1

1. Explain how a file region can be left in an inconsistent state in GFS?

   In GFS, failed mutations (writes or record appends) lead to chunks that are undefined and inconsistent. Due to this inconsistent state of the file region, different clients may see different data at different times.

2. Briefly explain how HopsFS uses User Defined Partitioning (UDP) and Distributed Aware Transaction (DAT) to improve the system performance (e.g., CPU usage, network traffic and latency)?

   With the aim of improving the performance of the file system operations, HopFS uses User Defined Partitioning to distribute data across different database nodes: the namespace is partitioned such that the metadata for all immediate descendants of a directory reside on the same database shard for efficient directory listing. This allows Partition Pruned Index Scans, meaning that scan operations are localized to a single database shard. Moreover, HopFS uses Distributed Aware Transaction to choose which Transaction Coordinator handles which file systems operations: the transaction is started on the database shard that stores all/most of the metadata required for the current file system operation.

3. Show with an example that in the CAP theorem, we can have only two properties out of Consistency, Availability, and Partition Tolerance at the same time.

   Answer3

4. How does BigTable provide strong consistency?

   Answer4

5. Write a simple query in Cypher for each of the following requests:

   - Match a Person called "John Doe"

     Answer5
   - Find FRIENDS_OF John Doe"

     Answer6
   - Count "John Doe" 's direct acquaintances

     Answer7