**Assignment 2: For Fast Learners**

**Level:** Advanced  
**Goal:** To challenge analytical thinking and application of concepts in real-world contexts.  
**Total Marks:** 100  
**Instructions:**

* Attempt **all questions in Section A**.
* Attempt **any three** from Section B.
* Attempt **any one part** from Sections C–G.
* Include diagrams, examples, or case studies to support your answers.

**Section A: Conceptual Insights (2 x 10 = 20 marks)**

1. Identify five big data tools/platforms and describe their primary use cases.
2. Analyze the role of Hadoop in managing large-scale data processing tasks.
3. Discuss three advantages of MapReduce in handling massive data workloads.
4. Define the function of heartbeats in HDFS. Why are they important?
5. Explain how data replication in Hadoop ensures fault tolerance.
6. Contrast Flume and Sqoop based on their architecture and data flow.
7. How do NoSQL databases differ from traditional relational databases?
8. Outline the purpose of schedulers in a distributed system like Hadoop.
9. Discuss the difference in how Pig and MapReduce process data.
10. What is the Hive metastore and how does it support data querying?

**Section B: Analytical Questions (10 x 3 = 30 marks)**

1. Describe in detail the architecture of the Hadoop ecosystem with a diagram.
2. Compare Master-Slave vs. Peer-to-Peer replication in distributed systems.
3. Outline the client-side process of accessing and storing data in HDFS.
4. Explain with syntax how CRUD operations are executed in MongoDB.
5. Illustrate Hive architecture and describe how it processes queries.

**Section C: Big Picture Thinking (10 x 5 = 50 marks)**

**Question 3: Choose any one**

* (a) Compare big data analysis and reporting. How do they serve different business goals?
* (b) Break down the major components of big data architecture and explain their interaction.

**Question 4: Choose any one**

* (a) Describe the MapReduce architecture and how tasks are scheduled and executed.
* (b) Explain YARN architecture. Describe ResourceManager, NodeManager, and ApplicationMaster.

**Question 5: Choose any one**

* (a) Discuss the design principles behind HDFS. Include data block storage and fault tolerance.
* (b) What hardware and software specs are needed to set up a Hadoop cluster? Describe the setup steps.

**Question 6: Choose any one**

* (a) Highlight Spark features and discuss how Spark works with HDFS, YARN, and Hive.
* (b) List key differences between Java and Scala. Why is Scala preferred for Spark?

**Question 7: Choose any one**

* (a) Describe Hive’s architecture and explain the flow of data from input to result.
* (b) Compare:
  + Apache Pig vs MapReduce
  + Pig vs SQL
  + Pig vs Hive