

# Unit Testing, MRUnit & MapReduce – Detailed Answers

## 1. Unit Testing Techniques

Unit testing is a software testing technique in which individual units or components of an application are tested independently to ensure correct functionality.

- White Box Testing – Focuses on internal logic, code paths, and conditions.
- Black Box Testing – Tests functionality without knowing internal code.
- Gray Box Testing – Combination of white and black box testing.
- Manual Unit Testing – Tests written and executed manually.
- Automated Unit Testing – Uses tools like JUnit to automate tests.
- Test-Driven Development (TDD) – Tests are written before actual code.
- Mock Testing – Uses mock objects to isolate components.

## 2. Three Core Classes of MRUnit

MRUnit is a testing library used to test Hadoop MapReduce jobs.

1. **MapDriver:** Used to test Mapper logic independently by providing input key-value pairs and verifying output.
2. **ReduceDriver:** Used to test Reducer logic independently with grouped input values.
3. **MapReduceDriver:** Used to test both Mapper and Reducer together in an integrated manner.

## 3. Developing and Testing MapReduce Jobs with MRUnit

Developing MapReduce jobs involves writing Mapper, Reducer, and Driver classes. MRUnit helps in testing these components without running a Hadoop cluster.

- Write Mapper and Reducer classes.
- Add MRUnit dependency to the project.
- Create MapDriver or ReduceDriver test cases.
- Provide input key-value pairs.
- Define expected output.
- Run tests using JUnit framework.
- Validate results and debug logic.

## 4. Shuffle and Sort Phase in MapReduce

Shuffle and Sort is a crucial phase between Map and Reduce operations. It transfers intermediate data from mappers to reducers.

Shuffle groups mapper output by key and transfers it across the network. Sort sorts the data by key before sending it to reducers.

Diagram Explanation:  
Mapper Output → Partitioning → Shuffle → Sorting → Reducer Input

## 5. Types of Failures in MapReduce

MapReduce is designed to handle failures automatically.

- Mapper Failure – Occurs when a map task crashes or times out.
- Reducer Failure – Occurs due to reducer task crashes.
- Node Failure – Happens when a DataNode or TaskTracker fails.

Hadoop handles failures by re-executing failed tasks on other nodes, ensuring fault tolerance.