Sample Exam Questions

The exam would contain close to 10-12 Multiple Choice Questions, 3-4 short answer questions, and 1-2 detailed explanation questions. Duration of the exam would be 1.5 hours and it would be a closed book exam.

The questions would be formed from the lectures and slides, and would test you on general level of Big Data understanding discussed in the classes. Below is a sample of questions that will give you a fair idea of questions to be expected on the exam.

Multiple Choice Questions

1) What is the default block size in HDFS?

2) Which deamon is responsible for splitting and allocating files in HDFS to individual machines?

3) In Apache Spark, what is the default number of partitions created for a RDD?

4) Which of the following is not a deamon process that runs on Hadoop cluster?

   a. TaskTracker

   b. DataNode

   c. TaskNode

   d. JobTracker

5) Which of the following application types can Spark run in addition to batch-processing jobs?

   a. Stream processing

   b. Machine learning

   c. Graph processing

   d. All of the above

6) What is the default replication factor in HDFS?

Short Answer Questions

1) Describe briefly the component of Hadoop and the functions they perform.

2) What is the difference between an action and a transformation in Apache Spark?

3) Briefly explain a DStream and how you can use it for analyzing Streaming data.

4) Briefly provide an overview of AWS (Amazon Web Services) and how is it relevant for Big Data processing?

Long Answer Questions

1) Imagine that you have to develop a sentiment analysis engine for the purpose of trading based on mood/sentiment of people (essentially what people write) on various websites (such as TimesOfIndia, IBNLive) and social media websites (e.g. Twitter) in additional to traditional sources of data such as Stock Prices, Macroeconomic Data. Please lay out a plan on how you would build such a system using the Big Data tools.

   In this particular question, it is not important to be objectively complete. The focus instead is on applying the thought process to use knowledge of tools learnt to implement real-life projects.