MIDTERM TOPICS

Please note that the topics below are some key topics that will be covered. There could be some questions outside these also at the discretion of the instructor. You are responsible for knowing all the topics covered in class.

Topics Covered for Midterm:

1. Basics of Big Data
   - 3V concept
   - What constitutes BD

2. HDFS Characteristics
   - Distributed, Fault tolerant, etc
   - Functions of NameNode and DataNodes
   - Block size concept - calculating number of blocks, total space taken on HDFS, etc
   - Unix commands to interact with HDFS

3. MapReduce
   - Conceptual understanding of map and reduce phases
   - Examples using functions for map and reduce
   - Stages of MapReduce

4. Apache Spark
   - Basic Idea
   - Concept of RDDs including Transformation and Action
   - Spark code using PySpark (Map function, Reduce Function, etc)
   - Pair RDDs
   - Dataframe operations

5. Machine Learning
   - Basics of Machine Learning, conceptual ideas of classification, regression, clustering
   - Classification models such as Logistic Regression, model creation using .fit, application using .transform
   - Feature extraction, transformation, and selection