FINAL EXAM 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 Final Exam:

Pre-midterm topics:

- Understand HDFS storage, block size, and calculation
- MapReduce and its application to various scenarios

Post-midterm topics:

1. Apache Spark and Spark SQL
   - RDD and Dataframes
   - Aggregation, grouping, and join operations on Dataframes **

2. MLlib
   - Basic Idea and operations -> transform and fit
   - Concept of transformers and estimators
   - Pipelines - how to create a multi-stage pipelines
   - Concept of Recommender Systems
   - Frequent Pattern Mining and Association Rules concept and calculation

* You will not be asked about specific classification or clustering algorithms.

3. GraphX
   - Basic Concepts - indegree, outdegree, edges, vertices, Pagerank, triangle count
   - GraphFrame operations – select, project, join, aggregate **
   - Graph operations using GraphFrame e.g. PageRank, grouping, etc

4. Spark Structured Streaming
   - Ideas and concepts
   - What is unbounded table
   - Basic operations on streaming Dataframes, window operations on Dataframes.

5. Hive
   - Basic Concepts
   - Data storage and Data Model
   - HQL - advanced queries involving grouping and aggregation

6. NoSQL concepts
- How are NoSQL databases different from RDBMS
- CAP theorem
- Where do various databases fall in terms of CAP

7. MongoDB
- Data Model
- Queries - advanced involving grouping and aggregation

8. HBase
- Know data model, column-oriented properties
- Data structure and indexing
- Basic Queries (not much detail)

9. Cassandra
- Architecture in detail
- Data model
- Query Model
- Partitioning