Question Bank for Big Data Analytics

Unit-1: Importance of Big Data

1. Explain classification of digital data.
2. What is structured data? Why it is easy to work with structured data?
3. What is semi-structured data? What are the sources of semi-structured data?
4. What is unstructured data? What are issues with unstructured data? How to deal with unstructured data?
5. Define big data. What are the characteristics of big data? What are the challenges with big data?
6. Explain big data with reference to volume, velocity and variety. What are the other characteristics of big data which are not definitional traits of big data?
7. How is traditional Business Intelligence (BI) environment different from big data environment?
8. What is big data analytics? What big data analytics is not?
9. Explain classification of big data analytics.
10. What are top challenges facing big data? Why big data analytics is important?

Unit -2: Hadoop Architecture

1. What is Hadoop? Explain Hadoop ecosystem.
2. Explain design of Hadoop distributed File System (HDFS).
3. Explain use of Namenodes and Datanodes in Hadoop distributed File System (HDFS).
4. Explain anatomy of a file read and write in Hadoop.
5. Explain Block Caching, HDFS Federation, HDFS High Availability, Failover and fencing in Hadoop distributed File System (HDFS).
6. Explain Coherency Model in Hadoop distributed File System (HDFS).
7. Explain Parallel Copying with distcp in Hadoop distributed File System (HDFS)
8. What is MapReduce in Hadoop? Explain MapReduce architecture.
9. How MapReduce Organizes Work?
10. How MapReduce can be used with National Climatic Data Center (NCDC) dataset to find highest recorded global temperature for each year?
11. Explain MapReduce data flow with a single and multiple reduce task.
12. Explain Anatomy of a MapReduce Job Run.
13. What are the possible Failures in Classic MapReduce? How failures in Classic MapReduce are handled?

Unit-3: Hadoop I/O

1. Explain components of Yet Another Resource Negotiator (YARN).
2. Explain anatomy of a Yet Another Resource Negotiator (YARN) application run.
3. Compare of MapReduce 1 and YARN components. How Yet Another Resource Negotiator (YARN) is better than MapReduce 1.
4. Explain Benefits of using Yet Another Resource Negotiator YARN.
5. Explain scheduling in Yet Another Resource Negotiator (YARN).
6. Explain Capacity Scheduler Configuration in Yet Another Resource Negotiator (YARN).
7. Explain Fair Scheduler Configuration in Yet Another Resource Negotiator (YARN).
8. How Data Integrity is implemented in Hadoop?
9. What are advantages of compressing data stored on Hadoop? What is codec in Hadoop?
10. What is Serialization and Deserialization? What are desirables that an RPC serialization format must satisfy?
11. Explain Writable Interface in Hadoop. Which are Writable classes available in Hadooop?
12. Explain Serialization Frameworks in Hadoop
13. With suitable diagram explain SequenceFile format.

Unit 4: NoSQL Management

1. Explain benefits of Relational databases.
2. What is impedance mismatch? How impedance mismatch has been dealt?
3. What are reasons for Emergence of NoSQL databases?
4. Explain aggregate data models with suitable example.
5. What are consequences of aggregate orientation? What are drawbacks of aggregate orientation?
6. Explain Key-value, document databases and Column-Family Stores.
7. What is Graph database? What are advantages of using Graph databases?
8. What are advantages and disadvantages of Schemaless Databases?
9. What is Materialized Views? What are strategies for building a materialized view?
10. Explain distribution techniques single-server, master-slave replication, sharding, and peer-to-peer replication.
11. What is Update and Read Consistency? What are approaches for maintaining consistency?
12. Explain the CAP theorem in NoSQL world.

Unit-5 : Analytics Framework

1. What is Pig? Compare Pig and MapReduce? What are advantages of using Pig?
2. Explain architecture of Apache Pig with neat diagram.
3. Explain Pig Latin relational operators with suitable example.
4. Explain Pig Eval function, Filter function, Load function, and Store function.
5. Explain Pig Latin Data Processing Operators.
6. What is Apache Hive? What are features of Hive?
7. What is Apache Hive? Compare Hive and MapReduce.
8. Explain architecture of Apache Hive with neat diagram. How Hive works?
9. Explain Apache Hive Metastore configurations with neat diagram.
10. Explain Apache Hive Operators and Functions.
11. Explain Managed Tables and External Tables in Apache Hive.
12. How to import data in Apache Hive? Explain multitable insert statement with suitable example.
13. What is HBase? Explain Data Model of Hbase.
14. Explain concept of regions in Hbase? What are its advantages?

Unit 6: Securing Ecosystem

1. Why do we need to secure Hadoop? What are Challenges for securing the Hadoop ecosystem?
2. What are key security considerations while securing Hadoop-based Big Data ecosystem?
3. Explain reference architecture for Big Data security.
4. What is Kerberos? How Kerberos works?
5. Explain how to set up a secured Hadoop cluster using Kerberos.
6. How to secure Hive interactions in Hadoop?
7. Explain steps are followed to set up a secured Oozie in the Hadoop cluster
8. Explain how to secure Flume.
9. What are the best practices for securing the Hadoop ecosystem components?