Big Data Analytics Important Questions PDF

The most important Big Data Analytics questions are the characteristics of big data (the Vs), types of data, the Hadoop ecosystem, HDFS architecture and replication, MapReduce with word count and other examples, YARN, Hive and HiveQL, Pig Latin, HBase and NoSQL databases (key-value, document, column and graph), the CAP theorem, Apache Spark with RDDs, and stream processing with Kafka or Spark Streaming.

Key takeaways

- HDFS architecture, MapReduce and the Hadoop ecosystem are asked in almost every paper.
- Write MapReduce answers with the map, shuffle and reduce stages for a small example.
- Compare Hadoop MapReduce and Spark in a table.
- These are practice questions, not past papers; check your own syllabus and official previous papers.

About these important questions

This page lists the most important Big Data Analytics questions for B.Tech CSE and IT exams, arranged unit by unit. The units follow the standard syllabus that most Indian technical universities share, so the order may differ slightly from yours.

Please note: these are practice questions based on common syllabus topics. They are not verified past papers, and we do not claim that any question will appear in your exam. Check your own university syllabus first, then solve the official previous year papers. AKTU students can download them from the AKTU old question paper archive.

Unit 1: Introduction to Big Data

Covers big data characteristics, data types, sources and analytics.

Short questions (2 marks)

- Define big data.
- List the 5 Vs of big data.
- Differentiate between structured, semi-structured and unstructured data.

Long questions

1. Explain the characteristics of big data with examples.
2. Explain the challenges of traditional systems in handling big data.

3. Explain the types of analytics: descriptive, diagnostic, predictive and prescriptive.
4. Explain big data applications in retail, healthcare, banking and social media.
5. Explain the big data analytics life cycle.
6. Compare a traditional data warehouse with a Hadoop-based data lake, covering schema, cost and the kind of data each stores.

Unit 2: Hadoop and HDFS

Covers Hadoop architecture, HDFS and YARN.

Short questions (2 marks)

- What is the role of the NameNode?
- What is the default replication factor in HDFS?
- What is a block in HDFS?

Long questions

1. Explain the Hadoop ecosystem with a diagram (HDFS, YARN, MapReduce, Hive, Pig, HBase, Sqoop, Flume, ZooKeeper).
2. Explain the HDFS architecture: NameNode, DataNode and Secondary NameNode.
3. Explain how a file is read from and written to HDFS.
4. Explain replication and rack awareness in HDFS.
5. Explain the YARN architecture: ResourceManager, NodeManager and ApplicationMaster.

Unit 3: MapReduce

Covers the MapReduce programming model and its phases. This unit carries the most long questions.

Short questions (2 marks)

- What is a combiner?
- What is a partitioner?
- Differentiate between the map and reduce phases.

Long questions

1. Explain the MapReduce model with a word count example, showing the input, map output, shuffle and sort, and reduce output.
2. Explain the anatomy of a MapReduce job run.
3. Explain the role of combiners and partitioners with examples.
4. Write MapReduce logic to find the maximum temperature per year from weather records.
5. Explain input formats, output formats and counters.
6. Explain the limitations of MapReduce.
7. Explain how MapReduce handles a failed task or a failed node, and what speculative execution means.

Unit 4: Hadoop Tools and NoSQL

Covers Hive, Pig, HBase and NoSQL databases.

Short questions (2 marks)

- What is HiveQL?
- State the CAP theorem.
- Differentiate between SQL and NoSQL databases.

Long questions

1. Explain the Hive architecture and write HiveQL queries to create a table, load data and run aggregates.
2. Explain Pig Latin and its data model, with an example script.
3. Compare Hive and Pig.
4. Explain the HBase data model and architecture.
5. Explain the types of NoSQL databases (key-value, document, column-family and graph) with an example of each.
6. Explain the CAP theorem and the BASE properties.

Unit 5: Spark and Stream Processing

Covers Apache Spark and real-time data processing. Some universities add data visualisation here.

Short questions (2 marks)

- What is an RDD?
- Differentiate between a transformation and an action in Spark.
- What is lazy evaluation?

Long questions

1. Explain the Spark architecture: driver, cluster manager and executors.
2. Explain RDDs, their properties and how they provide fault tolerance through lineage.
3. Compare Hadoop MapReduce and Spark.
4. Explain Spark SQL and DataFrames.
5. Explain stream processing with Kafka and Spark Streaming.
6. Explain the role of data visualisation in big data analytics.

Sample answer outline: MapReduce word count

Word count is the classic MapReduce question. A full answer has these parts:

1. **Define MapReduce.** It is a programming model for processing large data in parallel across a cluster. The programmer writes a map function and a reduce function; the framework handles splitting, scheduling, shuffling and fault tolerance.
2. **Take a small input.** For example, two lines: "deer bear river" and "car car river".

3. **Show the map phase.** Each mapper reads a line and emits (word, 1) for every word: (deer, 1), (bear, 1), (river, 1), (car, 1), (car, 1), (river, 1).
4. **Show shuffle and sort.** The framework groups values by key: bear → [1], car → [1, 1], deer → [1], river → [1, 1].
5. **Show the reduce phase.** Each reducer sums the list: (bear, 1), (car, 2), (deer, 1), (river, 2).
6. **Draw the data flow diagram.** Input splits, mappers, shuffle, reducers and output in HDFS.
7. **Mention the combiner.** A combiner can sum counts on each mapper before the shuffle, which reduces network traffic.

Write the map and reduce functions as short pseudocode as well. Some examiners expect both the diagram and the code.

Marks weightage and how to write answers

A common pattern, used by AKTU and many other universities, has a compulsory short-answer section that covers every unit, followed by long-answer questions with a choice inside each unit. Marks per question change between universities and years, so confirm the scheme on a recent official paper.

- **Short answers (about 2 marks):** a crisp definition, one example or diagram, and nothing more. Three to five lines is enough.
- **Long answers (7 to 10 marks):** start with a definition, add a labelled diagram or table, explain the steps, and finish with an example, advantages or a comparison.
- Draw the HDFS and YARN architecture diagrams with labelled components and arrows for data flow.
- For tool questions (Hive, Pig, Spark), include a short query or script. It shows practical knowledge.

A 7-day revision plan for Big Data Analytics

1. **Days 1 to 5:** one unit a day. Read your notes, then write the long answers for that unit from memory and check them against your notes.
2. **Day 6:** draw the Hadoop ecosystem, HDFS and Spark architectures from memory, and redo the word count and maximum temperature MapReduce examples.
3. **Day 7:** solve one full official previous year paper in three hours, then check which units cost you time.

Keep a single sheet of formulas, definitions and diagrams for the last hour before the exam.

Frequently asked questions

Which Big Data Analytics topics are most important?

The Vs of big data, the Hadoop ecosystem, HDFS architecture, MapReduce with examples, YARN, Hive, Pig, HBase, NoSQL types, the CAP theorem and Spark RDDs are asked most often.

Is Hadoop still relevant if companies use Spark?

Yes, for the exam and for understanding. HDFS, YARN and the MapReduce model explain how distributed processing works, and Spark often runs on the same clusters.

Is Big Data Analytics useful for jobs?

Yes. Data engineering and analytics roles use Spark, SQL, NoSQL and cloud data tools, which build on the ideas in this subject.

Are these Big Data Analytics questions from previous year papers?

No. They are practice questions built from topics that appear in most B.Tech Big Data Analytics syllabi. Use them with your own university's official previous papers, for example the AKTU archive at https://aktu.ac.in/old-question-paper.html.

Is this PDF free?

Yes. Fill in the short form on this page and the PDF downloads right away.

Sources and further reading

- AKTU: old question papers (official archive)
- Apache Hadoop documentation
- Apache Spark documentation

About this PDF

These are practice questions for revision, prepared by the Uncodemy trainer team from common B.Tech syllabus topics. They are not past papers, a certificate or a promise of marks. Spot an error? Write to info@uncodemy.com and the team will review it.

More free resources and courses: uncodemy.com | info@uncodemy.com | +91 98183 66550