BIG DATA & AI Question

1. What does Big Data primarily refer to?

- A) Small datasets
- **B) Large volumes of data**
- C) Data stored in Excel sheets
- D) Archived emails only

2. Which of the following is NOT a characteristic of Big Data?

- A) Volume
- B) Velocity
- **C) Validity**
- D) Variety

3. What is Hadoop primarily used for?

- A) Real-time gaming
- **B) Data storage and processing**
- C) Image editing
- D) Web browsing

4. Which component of Hadoop is responsible for storage?

- A) MapReduce
- B) YARN
- C**) HDFS (Hadoop Distributed File System)**
- D) Spark

5. What does HDFS stand for?

- A) High Data File System
- **B) Hadoop Distributed File System**
- C) Huge Data File Store
- D) Hyper Data File Source

6. What is MapReduce used for?

- A) Storing data
- **B) Processing data**
- C) Managing resources
- D) Visualizing data

7. Which of these is a common NoSQL database?

- A) MySQL
- B) PostgreSQL
- **C) MongoDB**
- D) Oracle

8. What is the primary benefit of NoSQL databases?

- A) Strict schema
- **B) Scalability and flexibility**
- C) Slow performance
- D) Complex joins

9. What is a Data Lake?

- A) A repository for structured data only
- **B) A repository for raw data (structured & unstructured)**
- C) A cleaned database
- D) A small storage unit

10. Which of the following is a data visualization tool?

- **A) Tableau**
- B) HDFS
- C) YARN
- D) MongoDB

11. What is Apache Spark primarily used for?

- A) Data Storage
- **B) Real-time processing**
- C) Sending Emails
- D) File Compression

12. What language is Pig Latin associated with?

- A) Java
- **B) Apache Pig**
- C) Python
- D) C++

13. Which of these is a Hadoop-based data warehousing tool?

- A) MongoDB
- **B) Hive**
- C) Cassandra
- D) Redis

14. What is the role of YARN in Hadoop?

- A) Data Processing
- B) Storage
- **C) Resource Management**
- D) Querying

15. What does the term Veracity in Big Data refer to?

- A) Speed of data
- **B) Uncertainty/Accuracy/Quality of data**
- C) Volume of data
- D) Variety of data

16. Which language is primarily used for data analysis?

- A) C++
- **B) R**
- C) HTML
- D) Assembly

17. Which of the following is an ETL tool?

- **A) Talend**
- B) HDFS
- C) Git
- D) Docker

18. What is an example of unstructured data?

- A) SQL Tables
- B) CSV Files
- **C) Emails / Videos / Images**
- D) Excel Spreadsheets

19. What is sentiment analysis used for?

- A) Calculating sales
- **B) Analyzing opinions and emotions**
- C) Storing passwords
- D) Network security

20. Which Big Data tool uses SQL-like queries?

- A) MapReduce
- **B) Hive**
- C) MongoDB
- D) Flume

21. Which programming language is widely used in Big Data analytics for statistical analysis?

- A) C#
- **B) Python**
- C) Swift
- D) Kotlin

22. Which of these is NOT a Big Data processing framework?

- A) Hadoop
- B) Spark
- **C) NoSQL**
- D) Flink

23. ***Which term refers to the transformation of raw data into meaningful information?***

- *A) Data Storage*
- ***B) Data Mining***
- *C) Data Ingestion*
- *D) Data Entry*

24. What is the main advantage of using cloud storage for Big Data?

- A) Fixed capacity
- **B) Scalability and Flexibility**
- C) High cost
- D) Difficulty of use

25. What does data velocity refer to?

- A) Size of data
- **B) Speed of data generation**
- C) Accuracy of data
- D) Format of data

26. What is a key feature of Big Data analytics tools?

- **A) Data Integration**
- B) Data Loss
- C) Single-user access
- D) Manual processing

27. Which tool is commonly used for distributed data processing?

- A) Excel
- **B) Apache Spark / Hadoop**
- C) Notepad
- D) Calculator

28. What type of data does a Data Warehouse primarily store?

- ***A****) Raw, unstructured data*
- *B) Processed, structured data*
- *C) Real-time streams*
- *D) Temporary files*

29. What does ETL stand for in data processing?

- A) Extract, Transfer, Load

- **B) Extract, Transform, Load**
- C) Enter, Test, Log
- D) Exit, Track, Load

30. Which of the following best describes Big Data analytics?

- A) Storing small data
- **B) Analyzing and interpreting large datasets**
- C) Retrieving deleted files
- D) Formatting hard drives

31. Which data format is suitable for storing unstructured data?

- A) SQL Tables
- **B) JSON**
- C) CSV
- D) Fixed-width text

32. What is the purpose of a Data Warehouse?

- A) To delete old data
- **B) To store data for analysis**
- C) To act as RAM
- D) To host websites

33. What is Data Ingestion?

- **A) Importing data for immediate use or storage**
- B) Deleting data
- C) Encrypting data
- D) Visualizing data

Part 2: Additional C-CAT Standard / PYQ Questions (34–40)

34. In the Hadoop architecture, which node is considered the "Master" node?

- A) DataNode
- **B) NameNode**
- C) TaskTracker
- D) SlaveNode

35. HDFS is NOT suitable for which of the following scenarios?

- *A) Storing very large files*
- *B) Batch processing*
- *C) Low-latency data access*
- *D) High throughput access*

36. The multidimensional data model used in Data Warehousing is often referred to as a:

- *A) Star Schema*
- ***B) Data Cube***
- *C) Relation*
- *D) Flat file*

37. Which of the following best describes "Columnar Storage" (used in HBase/Cassandra)?

- *A) Data is stored row by row*
- ***B) Data is stored column by column***
- *C) Data is stored as a graph*
- *D) Data is stored as a document*

38. Which command is used to list files in the Hadoop Distributed File System?

- A) ls -la
- **B) hdfs dfs -ls**
- C) hadoop list
- D) show files

39. Which of the 5 Vs represents the "trustworthiness" of data?

- A) Volume
- B) Variety
- **C) Veracity**
- D) Value

40. Apache Kafka is primarily used for:

- A) Batch Processing
- **B) Real-time Data Streaming**
- C) Data Warehousing
- D) Machine Learning

-----**C-CAT Section B: AI & ML Previous Year MCQs**

(Note: The options for the AI/ML questions were not present in the original document, so only the question is provided below.)

Topic 1: Foundations & Definitions

1. What is the "Turing Test" designed to assess?

2. Who is known as the "Father of Artificial Intelligence"?

3. Which of the following is a component of an Expert System?

4. Strong AI refers to:

Topic 2: Search Algorithms (Most Important for C-CAT)

5. Which search algorithm uses a "Queue" (FIFO) data structure?

6. Which of the following is an "Informed" (Heuristic) search algorithm?

*7. In the A algorithm, the function** ***f(n)*** **is defined as:**

8. What is the main problem with the "Hill Climbing" search algorithm?

9. In a game like Chess or Tic-Tac-Toe, which algorithm is commonly used for the adversary's move?

Topic 3: Machine Learning & Logic

10. "Spam Email Detection" is a classic example of which type of learning?

11. Which of the following is NOT a Supervised Learning algorithm?

12. In Propositional Logic, the symbol $\vee$ represents:

13. Which logic allows the use of quantifiers like "For all" ($\forall$) and "There exists" ($\exists$)?

14. What is "Overfitting" in Machine Learning?

Topic 4: Neural Networks & NLP

15. What is the fundamental unit of an Artificial Neural Network (ANN)?

16. Which process reduces a word to its root form (e.g., "Running" $\to$ "Run")?

17. What does "NLP" stand for in AI?

18. In Fuzzy Logic, a value can be:

Topic 5: Miscellaneous / Tricky PYQs

19. Which of the following agents operates based on the condition "If condition, then action"?

20. What is the "Heuristic Function" $h(n)$?

21. Alpha-Beta pruning is an optimization technique used in:

22. Which language was traditionally preferred for AI programming in the early days?-----Answer KeyPart 1 & 2: Big Data & PYQ Answers (Q1 - Q40)

You are absolutely right to ask about this. **"Basics of Big Data & AI"** is a relatively recent addition to **Section B** of the CDAC C-CAT syllabus, carrying about **8–10 questions** (approx. 24–30 marks).
Since this section is newer, you won't find many "old" PYQs (from 2015–2021). However, based on the recent 2024–2025 trends and the official syllabus scope, the questions are strictly **conceptual and definition-based**. They do not ask for coding or complex derivations.

Here is the **"Cheat Sheet" of Most Likely Questions** for your exam day after tomorrow (Jan 10/11), covering the topics C-CAT is currently asking.

○

Q No.	Answer
1	B) Large volumes of data
2	C) Validity
3	B) Data storage and processing
4	C) HDFS (Hadoop Distributed File Sy…
5	B) Hadoop Distributed File System
6	B) Processing data
7	C) MongoDB
8	B) Scalability and flexibility
9	B) A repository for raw data (structure…
10	A) Tableau
11	B) Real-time processing
12	B) Apache Pig
13	B) Hive
14	C) Resource Management
15	B) Uncertainty/Accuracy/Quality of data

16	B) R
17	A) Talend
18	C) Emails / Videos / Images
19	B) Analyzing opinions and emotions
20	B) Hive
21	B) Python
22	C) NoSQL
23	B) Data Mining
24	B) Scalability and Flexibility
25	B) Speed of data generation
26	A) Data Integration
27	B) Apache Spark / Hadoop
28	B) Processed, structured data
29	B) Extract, Transform, Load
30	B) Analyzing and interpreting large da…
31	B) JSON
32	B) To store data for analysis
33	A) Importing data for immediate use o…
34	B) NameNode
35	C) Low-latency data access
36	B) Data Cube
37	B) Data is stored column by column
38	B) hdfs dfs -ls
39	C) Veracity
40	B) Real-time Data Streaming

Part 3: AI & ML Answers (C-CAT Section B)

Q No.	Answer
1	Whether a machine can exhibit intelli…
2	John McCarthy
3	All of the above
4	AI that possesses consciousness and…
5	Breadth-First Search (BFS)
6	A* (A-Star) Search
7	$f(n) = g(n) + h(n)$
8	It can get stuck in Local Maxima
9	Minimax Algorithm
10	Supervised Learning (Classification)
11	K-Means Clustering
12	OR (Disjunction)
13	First-Order Predicate Logic (FOPL)
14	When the model learns the training d…
15	Perceptron (Neuron)
16	Stemming
17	Natural Language Processing
18	Any value between 0 and 1 (inclusive)
19	Simple Reflex Agent
20	The estimated cost from node n to …
21	Minimax Algorithm
22	LISP (or Prolog)