🚀 Ultimate C-CAT Cheat Sheet: Big Data & AI

SECTION 1: BIG DATA FUNDAMENTALS

1. The "5 Vs" of Big Data (Definition)

- **Trick:** Think of a powerful Car Engine (**V5**).
  - **Volume:** Size (Terabytes, Petabytes).
  - **Velocity:** Speed of generation (Streaming, Real-time).
  - **Variety:** Different shapes (Text, Video, XML).
  - **Veracity:** Trustworthiness (Accuracy/Quality). **(Important!)**
  - **Value:** Business usefulness.
  - *Exam Trap:* **Validity** is NOT one of the 5 Vs.

2. Data Types & Storage

Type	Characteristics	Examples	Storage Location
Structured	Fixed Schema, Rows/Cols	SQL Tables, Excel, CSV	Data Warehouse
Unstructured	No Schema, Heavy	Video, Audio, Images, Emails	**Data Lake**
Semi-Structured	Tags/Keys but no strict table	JSON, XML, NoSQL data	NoSQL DBs

- **Key Concept:**
  - **Schema-on-Write:** SQL (Must define table *before* adding data).
  - **Schema-on-Read:** Big Data/Hadoop (Define structure *only when you use* the data).

SECTION 2: HADOOP ECOSYSTEM (The "Zoo")

1. Core Components (HDFS + MapReduce + YARN)

- **HDFS (Storage):** The hard drive of Hadoop.
    - **NameNode (Master):** Stores **Metadata** (file names, permissions). Does NOT store file contents.
    - **DataNode (Slave):** Stores actual **Data Blocks**. Sends "Heartbeats" (3 sec) to Master.
    - **Secondary NameNode:** NOT a backup! It is a **Helper** (does checkpointing).
- **MapReduce (Processing):** The processor.
    - **Mapper:** Processes data $\rightarrow$ Outputs (Key, Value) pairs.
    - **Reducer:** Aggregates the output.
- **YARN (Management):** The Operating System.
    - *Full Form:* **Y**et **A**nother **R**esource **N**egotiator.
    - *Job:* Manages resources (RAM/CPU) for the cluster.

2. The "Must-Memorize" Numbers

- **Default Block Size: 128 MB** (Hadoop 2.x/3.x) or **64 MB** (Old Hadoop 1.x).
- **Default Replication Factor: 3** (Data is copied 3 times for safety).
- **Hardware Type:** Runs on **Commodity Hardware** (Cheap, standard consumer-grade hardware).

3. Ecosystem Tools Match-Up

Tool	Keyword / Trick	Role
Hive	"SQL-like"	Data Warehousing (Uses HQL).
Pig	"Scripting"	Data Flow Language (Pig Latin). ETL.
Spark	"In-Memory" / "Real-Time"	100x faster than MapReduce.
Flume	"Logs"	Ingesting streaming logs.
Sqoop	"SQL + Hadoop"	Transfer data between SQL & Hadoop.
Zookeeper	"Coordinator"	Distributed coordination/synchronization.

SECTION 3: DATABASE CONCEPTS

1. CAP Theorem (For Distributed Systems)

- **Rule:** You can only pick **2 out of 3**.
    1. **C**onsistency (Everyone sees same data).
    2. **A**vailability (System always responds).
    3. **P**artition Tolerance (System handles network breaks).
- **SQL (RDBMS):** Prioritizes **CA**.
- **NoSQL:** Prioritizes **AP** or **CP**.

2. Columnar vs Row-Oriented

- **Row-Oriented:** Standard SQL (Good for writing new records).
- **Column-Oriented:** HBase, Cassandra (Good for **reading/analytics** on Big Data).

SECTION 4: ARTIFICIAL INTELLIGENCE (AI)

1. The Hierarchy

- **AI:** Mimicking human behavior.
- **ML:** Learning from data without explicit programming.
- **DL:** Neural Networks (Brain-like structure).

2. Search Algorithms (AI)

Search Type	Algorithm	Data Structure Used	Characteristic
Uninformed (Blind)	**BFS** (Breadth-First)	**Queue (FIFO)**	Finds shortest path. Slow.
	DFS (Depth-First)	**Stack (LIFO)**	Goes deep fast. Can get lost.
Informed (Heuristic)	_A (A-Star)*_	Priority Queue	Uses formula $f(n) = g(n) + h(n)$. Best path.

	Hill Climbing	-	Greedy. Can get stuck in "Local Maxima".

SECTION 5: MACHINE LEARNING

1. Types of Learning (The "Student" Trick)

Type	Analogy	Description	Algorithms (Memorize!)
Supervised	**Teacher**	Input + **Labeled Output** is given.	Linear Regression, Logistic Regression, SVM, Decision Trees, Naive Bayes.
Unsupervised	**Self-Study**	Input ONLY (No labels). Find patterns.	**K-Means Clustering**, Apriori (Market Basket), PCA.
Reinforcement	**Gamer**	Learn via **Reward & Penalty**.	Q-Learning.

- *Exam Trap:* **Logistic Regression** is for **Classification** (Yes/No), NOT Regression (Numbers).

2. Confusion Matrix Terms

- **True Positive (TP):** Correctly predicted YES.
- **False Positive (FP):** "False Alarm" (Type I Error).
- **False Negative (FN):** "Missed It" (Type II Error).

3. NLP (Natural Language Processing)

- **Corpus:** The entire collection of text documents.
- **Tokenization:** Chopping text into words.
- **Stop Words:** Useless words removed during cleaning (e.g., "is", "the", "at").

SECTION 6: RAPID FIRE FULL FORMS

- **HDFS:** Hadoop Distributed File System
- **YARN:** Yet Another Resource Negotiator

- **JSON:** JavaScript Object Notation
- **SVM:** Support Vector Machine
- **ANN:** Artificial Neural Network
- **IoT:** Internet of Things
- **SaaS:** Software as a Service

Part 1: Big Data (Expected: 4-5 Questions)

Focus on the "5 Vs", Definitions, and Hadoop Ecosystem basics.

Top Predicted Questions / Concepts:

1. **The 5 Vs of Big Data (Most Repeated)**
   - **Question:** Which "V" refers to the trustworthiness or quality of the data?
   - **Answer: Veracity**.
   - *Must Know:*
     - **Volume:** Size of data.
     - **Velocity:** Speed of generation.
     - **Variety:** Different formats (structured, unstructured).
     - **Veracity:** Uncertainty/Quality.
     - **Value:** Business value.
2. **Structured vs. Unstructured Data**
   - **Question:** Email bodies, videos, and social media posts are examples of which type of data?
   - **Answer: Unstructured Data**.
   - *(Note: RDBMS tables are Structured; XML/JSON is Semi-Structured).*
3. **Hadoop Components (HDFS & MapReduce)**
   - **Question:** In the Hadoop ecosystem, which component is responsible for storage?
   - **Answer: HDFS** (Hadoop Distributed File System).
   - **Question:** What is the programming model used for processing large data sets in parallel?
   - **Answer: MapReduce**.
4. **SQL vs. NoSQL**
   - **Question:** Which of the following is a **NoSQL** database?
   - **Options:** MySQL, Oracle, MongoDB, PostgreSQL.
   - **Answer: MongoDB**.
   - *Concept:* NoSQL is for unstructured data, scales horizontally, and is schema-less.
5. **CAP Theorem (The Concept)**
   - **Question:** In CAP Theorem, what does 'P' stand for?
   - **Answer: Partition Tolerance**.
   - *(C = Consistency, A = Availability).*

Part 2: Artificial Intelligence (Expected: 4-5 Questions)

Focus on definitions of AI vs ML vs DL, and types of Learning.

Top Predicted Questions / Concepts:

6. **AI vs. ML vs. DL Hierarchy**
   - **Question:** Which statement correctly describes the relationship between AI, ML, and DL?

- **Answer: DL $\subset$ ML $\subset$ AI** (Deep Learning is a subset of Machine Learning, which is a subset of AI).

7. **Types of Machine Learning (High Priority)**
    - **Question:** Training a model using "labeled" data (input-output pairs) is called?
    - **Answer: Supervised Learning**.
    - **Question:** Grouping customers based on purchasing behavior *without* predefined labels (Clustering) is?
    - **Answer: Unsupervised Learning**.
    - **Question:** Learning through rewards and punishments (e.g., a robot learning to walk) is?
    - **Answer: Reinforcement Learning**.
8. **Turing Test**
    - **Question:** The Turing Test is designed to determine what?
    - **Answer: Whether a machine can exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human.**
9. **Basic Terminology (NLP / Vision)**
    - **Question:** What is the field of AI focused on the interaction between computers and human language?
    - **Answer: NLP (Natural Language Processing)**.
10. **Intelligent Agents**
    - **Question:** An agent that perceives its environment through sensors and acts upon it through actuators is called?
    - **Answer: Intelligent Agent** (or Rational Agent).

Quick Recap for the Exam (Day After Tomorrow)

For **Section B**, do not spend hours on complex AI math. Just memorize these distinctions:

- **HDFS** = Storage; **MapReduce** = Processing.
- **Supervised** = Teacher present (Labels); **Unsupervised** = No teacher (Patterns); **Reinforcement** = Reward/Punishment.
- **Veracity** = Trust/Quality of data.
- **Clustering** = Unsupervised; **Classification/Regression** = Supervised.

1. Cloud Computing Models (High Probability)

They often group Cloud basics with Big Data. Just memorize these acronyms:

- **IaaS (Infrastructure as a Service):** You rent raw hardware (e.g., AWS EC2).
- **PaaS (Platform as a Service):** You rent a platform to build apps without worrying about OS/Updates (e.g., Google App Engine, Heroku).
- **SaaS (Software as a Service):** You use finished software over the internet (e.g., Gmail, Google Drive, Zoom).
    - *Trick Question:* "Is Gmail IaaS or SaaS?" $\rightarrow$ **SaaS**.

2. Hadoop Ecosystem (One-Liners)

If they ask about specific tools, it will be simple matching:

- **Hive:** SQL-like queries on Hadoop (Think: "Data Warehousing").
- **Pig:** Scripting language for Hadoop (Think: "Data Flow").
- **Spark:** Fast, **in-memory** data processing (faster than MapReduce).
- **HBase:** A NoSQL database on top of Hadoop.

3. Data Warehouse vs. Data Lake

- **Data Warehouse:** Stores **Structured**, processed data (ready for analysis).
- **Data Lake:** Stores **Raw** data (Structured + Unstructured/Messy data).
- **OLTP vs. OLAP:**
    - **OLTP (Transactional):** Day-to-day operations (e.g., ATM transaction).
    - **OLAP (Analytical):** Historical analysis (e.g., Yearly sales report).

Here are the **7 Vs of Big Data** (The standard 5 + the new 2):

The Standard 5 Vs (You already know these)

1. **Volume:** The size of the data (TB, PB, ZB).
2. **Velocity:** The speed at which data is generated/processed.
3. **Variety:** Different forms (Text, Video, Audio).
4. **Veracity:** Uncertainty/Trustworthiness (Is the data accurate?).
5. **Value:** The business usage/profit derived from data.

The 2 New Vs (Add these to your notes)

6. **Variability:**
    - *Definition:* Refers to the **inconsistency** in the data flow.
    - *Example:* A hashtag trending on Twitter spikes the data flow for 2 hours, then drops. That inconsistent speed/flow is "Variability."
    - *Don't confuse with Variety:* Variety = different *types* (jpg, txt). Variability = changing *speed/meaning* over time.
      +1
7. **Visualization:**
    - *Definition:* The ability to represent complex data in **readable graphs/charts** so humans can understand it.

- *Why it's a V:* If you can't see/read the patterns, the Big Data is useless.