🚀 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