**1. Data Ingestion & Parsing**

* **Streaming Algorithms**
  + *Reservoir Sampling*: Select a random sample from an unbounded data stream.
  + *Count-Min Sketch*: Estimate frequency of elements in streaming data with low memory.
  + *Bloom Filter*: Probabilistic membership checking (is this element seen before?).
* **Parsing & Serialization**
  + JSON/XML/Avro/Parquet compression/decompression algorithms (Snappy, Zstandard).

**2. Data Cleaning & Deduplication**

* **String Similarity Matching**
  + *Levenshtein Distance* (edit distance for fuzzy matching).
  + *Jaccard Similarity* (for comparing sets, e.g., tokenized text).
* **Deduplication**
  + *MinHash* (efficient similarity detection).
  + *Locality Sensitive Hashing (LSH)* (bucket similar data together for near-duplicate removal).

**3. Data Transformation**

* **Sorting & Grouping**
  + External Merge Sort (used in distributed systems like Hadoop, Spark).
  + Hash Partitioning (for shuffles, joins, and groupBy).
* **Joins**
  + Hash Join (fast when memory fits).
  + Sort-Merge Join (used in distributed environments).
  + Broadcast Join (small dataset replicated across nodes).
* **Windowing & Aggregation**
  + Sliding Window algorithms (used in streaming frameworks).

**4. Data Storage & Indexing**

* **Indexing Structures**
  + B-Tree / B+Tree (used in databases).
  + LSM Trees (used in Cassandra, RocksDB, HBase for high write throughput).
* **Compression Algorithms**
  + Run-Length Encoding (RLE).
  + Dictionary Encoding.
  + Delta Encoding.

**5. Data Quality & Profiling**

* **Anomaly Detection** (on ingestion pipelines)
  + Z-Score, MAD (Median Absolute Deviation).
  + Streaming threshold-based algorithms.
* **Schema Matching / Validation**
  + Hash-based validation.
  + Constraint satisfaction algorithms.

**6. Optimization & Resource Management**

* **Query Optimization**
  + Cost-Based Optimizers (Catalyst in Spark, Calcite in Flink).
  + Dynamic Programming (used in join order selection).
* **Scheduling Algorithms**
  + DAG scheduling (Apache Airflow, Spark execution planner).
  + Fair scheduling / FIFO in cluster resource managers (YARN, Kubernetes).

**7. Scalability & Fault Tolerance**

* **Checkpointing & Recovery**
  + Chandy-Lamport Snapshot algorithm.
* **Consensus Algorithms**
  + Paxos, Raft (used in distributed databases and metadata management).