What are the 5 V’s of Big Data?

* 1. Velocity: Speed of Data into system
  2. Volume: Size of Data
  3. Variety: Types of Data
  4. Value: Insight
  5. Veracity: Worthiness of data

What are the Various Tuning Techniques in HIVE?

* 1. Change the execution Engine
     1. Map Reduce <- Default
     2. TEZ
        1. Logical Plan: Predicate push down
     3. Spark
  2. Partitioning the Bucketing
     1. Partitioning Types: Static and Dynamic
  3. File Formats; Sequential file formats: Processed line by line
     1. CSV
     2. Jason: Key-Value Pair
     3. XML: <Parent> <Child>
     4. URL: Http
     5. AVRO
     6. RC (ROW COLUMN)
     7. ORC (OPTIMIZED RC)
     8. PARQUET: handles nested data
  4. Mapside joins/ Bucketed Joins/ Sort Merge Bucketed
     1. Mapside Join: When smaller table is stored in memory
     2. Bucketed Join: Only the required buckets are are fetched on the mapper side, not the whole table
     3. Sort Merge Bucketed: each mapper reads a bucket from the first table and the corresponding bucket from the second table

What are ACID transactions?

* 1. A: Atomicity: All or Nothing
  2. C: Consistency: Values should be Consistent
  3. I: Isolation: Transactions should be visible to other sessions
  4. D: Durability: Committing

That are OLAP and OLTP?

* 1. OLAP: Online Analytical Processing
     1. complex queries to analyze aggregated historical data from OLTP systems
  2. OLTP: Online Transactional Processing
     1. captures, stores, and processes data from transactions in real time

Explain Fault Tolerance in Hadoop?

* 1. Replication
     1. Default replication factor is 3
  2. Rack Awareness
     1. instead of storing data in a single rack, we store it onto multiple racks

How do we decide the number of Mappers in Map Reduce?

* 1. Number of input splits, also number of logs but the latter is preferred

What are the Daemons in Hadoop?

* 1. Namenode
  2. Secondary Name node
  3. DataNode
  4. Resource manager
  5. Node manager

What are the different types of data?

* 1. Structured
  2. Semi-Structured
  3. Unstructured

Wat is the architecture of Hadoop

* 1. Master: Name Node, Resource Manager, Node Manager
  2. Slave: Secondary name Node, Data Node

What is Serialization vs Deserialization

* 1. Serialization: Writing data to disk in the network
     1. Python Term: Pickling
  2. Deserialization: Reading data to a disk from a network
     1. Python term: Depickling

What is Hadoop used for?

* 1. Storage and Processing

What is High Availability?

* 1. This is a hot standby name node that uses a shared edit log
     1. Fail Over Controller
     2. Quarum Journal Nodes

What is the difference between Hive and Hbase

* 1. HIVE
     1. used for batch processing of big data
     2. Analytical use case (OLAP) is accomplished using hive
     3. Apache Hive should be used for data warehousing requirements
  2. HBase
     1. Results of analytics are stored in HBase for random access,(can do queries on OLTP, NoSQL, no analytics)
     2. Crud operations.
     3. HBase is high scalable (scales horizontally using off the shelf region servers), highly available, consistent and low latency NoSQL database.