yyo6lpuyg

June 9, 2025

1 Big Data Q&A

1.1 a) List two primary differences between Hadoop version-2 and Hadoop version-3

Feature	Hadoop 2	Hadoop 3
Storage Optimization	Replication-based (3x replication)	Erasure Coding (more space-efficient)
Support for GPUs	Not supported	GPU-aware scheduling support

1.2 b) What is Hive Metastore? Can NoSQL Database HBase be configured as Hive Metastore?

Hive Metastore is a central repository that stores metadata about Hive tables, databases, schemas, and columns. It helps Hive execute SQL-like queries using this metadata.

Can HBase be used as Hive Metastore?

No, HBase cannot be directly configured as Hive Metastore. Hive requires a traditional RDBMS (e.g., MySQL, PostgreSQL) that supports ACID transactions and JDBC connectivity, which HBase lacks.

1.3 c) Using an example, depict how MapReduce computes word count

Input:
`"hello world hello"`

Map Phase Output:

```
<hello, 1>
<world, 1>
<hello, 1>
```

Shuffle & Sort Phase:

```
<hello, [1, 1]>
```

```
<world, [1]>
```

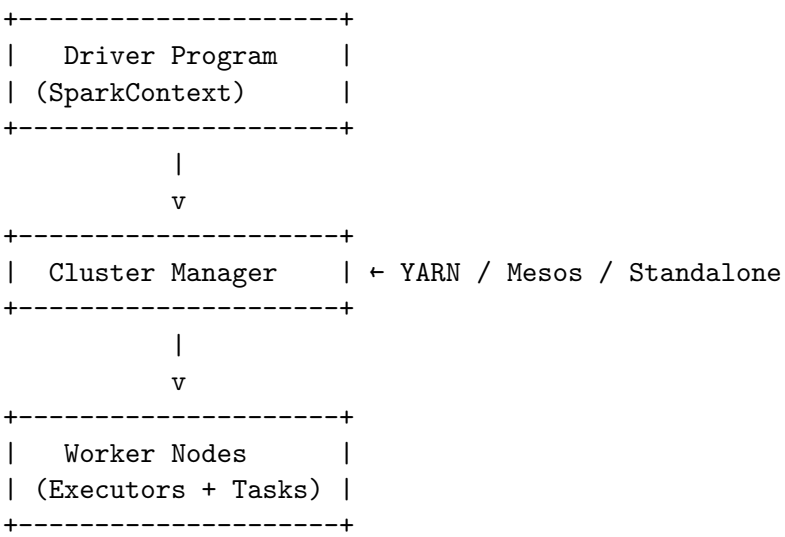
Reduce Phase:

```
<hello, 2>
<world, 1>
```

Final Output:

```
hello 2
world 1
```

1.4 d) Draw Spark Architecture and explain its various components



1.4.1 Components:

- **Driver Program**: Launches the Spark job and maintains `SparkContext`.
- **Cluster Manager**: Allocates resources (can be YARN, Mesos, or Standalone).
- **Worker Nodes**: Actual nodes in the cluster that run computations.
- **Executors**: Processes launched on worker nodes to run tasks.
- **Tasks**: Smallest unit of work executed in parallel across executors.

1.5 e) What is CAP theorem? Where does MongoDB stand in CAP theorem?

CAP Theorem states that in a distributed system, only two out of the following three guarantees can be fully achieved at once: 1. **Consistency** – All nodes return the most recent data. 2. **Availability** – Every request receives a response. 3. **Partition Tolerance** – The system continues operating even with network partitions.

MongoDB is generally a **CP (Consistency + Partition Tolerance)** system by default, but can be configured for **AP** trade-offs using tunable consistency levels (e.g., read preferences, write concerns).

1.6 f) List any 4 differences between Data Lake and Data Warehouse

Feature	Data Lake	Data Warehouse
Data Type	Structured, semi-structured, raw	Structured data only
Storage Cost	Low (e.g., HDFS, S3)	High (proprietary storage formats)
Schema	Schema-on-read	Schema-on-write
Use Case	Big Data, ML, Data Exploration	Business Intelligence, Reporting

---

1.7 g) Explain Partitioning in Hive with an example

Partitioning is a technique in Hive to divide a table into smaller parts based on the value of a column, improving query performance.

1.7.1 Example:

```
CREATE TABLE sales (
    item STRING,
    amount INT
)
PARTITIONED BY (region STRING);
```

1.7.2 Insert Data:

```
INSERT INTO TABLE sales PARTITION(region='east') VALUES ('pen', 10);
```

This stores the data physically in:

`/warehouse/sales/region=east/`

Querying with a filter on the partition column (e.g., `region = 'east'`) is faster because Hive skips irrelevant partitions (called *partition pruning*).

1.8 h) List any 2 differences between Coalesce and Repartition (in Spark)

Feature	Coalesce	Repartition
Data Movement	Avoids full shuffle; merges partitions	Performs full shuffle across cluster
Use Case	Efficiently reduce partitions	Increase/decrease partitions evenly