Fundamentals of Big Data Analytics

Quiz 2’s Solution

The number of maps is usually driven by the total size of ____

(a) Input

Which of the following is used for the execution of a Mapper or a Reducer on a slice of data?

(a) Task

Which of the following is used to schedule jobs and track the assigned jobs to Task tracker?

(c) Job Tracker

In how many stages does the MapReduce program normally execute?

(b) 3 (Map Shuffle/sort Reduce)

Which of the following platforms does Apache Hadoop run on?

(b) Cross Platform

Which of the following is the correct statement?

(a) Data locality means moving computation to the data, instead of data to the computation

Which type of data can Hadoop deal with?

(d) All of the above

Hadoop Framework is written in:

(a) Java

What is HDFS Federation and how it is different from the earlier version of HDFS. Please highlight the issues and solutions as well. (7)

HDFS (Hadoop Distributed File System) Federation is an extension of the HDFS architecture that enables the scaling of the file system to support a large number of files and larger clusters. The primary goal of HDFS Federation is to enhance the scalability and availability of the HDFS architecture while maintaining backward compatibility with earlier versions.

While HDFS Federation improves the scalability and availability of the HDFS architecture, it also introduces some challenges that need to be addressed.

1. **Tight coupling of Block Storage and Namespace:** Currently the collocation (closeness) of namespace and block management in the namenode has resulted in tight coupling of these two layers. This makes alternate implementations of namenodes challenging and limits other services from using the block storage directly.
2. **Namespace scalability:** While HDFS cluster storage scales horizontally with the addition of datanodes, the namespace does not. Currently the namespace can only be vertically scaled on a single namenode. The namenode stores the entire file system metadata in memory. This limits the number of blocks, files, and directories supported on the file system to what can be accommodated in the memory of a single namenode.
3. **Performance:** File system operations are limited to the throughput of a single namenode, which currently supports 60K tasks. The Next Generation of Apache MapReduce will support more than 100K concurrent tasks, which will require multiple namenodes.

Overall, HDFS Federation improves the scalability and availability of HDFS while maintaining backward compatibility with earlier versions. However, it also introduces new challenges that need to be addressed to ensure that the system operates efficiently and effectively.

Please read the following scenario:

Amazon wants to calculate its total sale city wise for the year of 2020 in USA. A chunk of data is given below:

Sr. No	Date	City	Amount ($)
1	15/1/2020	Chicago	20,000
2	16/4/2020	Boston	10,000

Please solve the problem using the map reduce framework and clearly write your strategy in form of pseudocode for Mapper and Reducer functions. Also please clearly mention the key value pairs as well.

```
# Mapper Function
def mapper(record):
    # Extract the City and Amount from the record
    city = record['City']
    amount = record['Amount ($)']
    # Emit the key-value pair (City, Amount)
    emit(city, amount)
```

```
# Reducer Function
def reducer(key, values):
    # Key: City
    # Values: List of Amounts
    # Calculate the total sale for the City
    total_sale = sum(values)
    # Emit the key-value pair (City, Total Sale)
    emit(key, total_sale)
```

Input:

[{'Sr. No': 1, 'Date': '15/1/2020', 'City': 'Chicago', 'Amount ($)': 20_000}, {'Sr. No': 2, 'Date': '16/4/2020', 'City': 'Boston', 'Amount ($)': 10_000}]

Output:

[('Chicago', 20_000), ('Boston', 10_000)]