NoSQL Tutorial

Map-Reduce

Design Lab (CS69202)
IIT Kharagpur

SQL vs NoSQL

- SQL has tables with fixed rows and columns whereas NoSQL has dynamic data storage models (can be altered anytime).
- SQL is a general purpose database whereas NoSQL is a database used for handling documents, key-value pairs, wide-column data and graphs.
- SQL is fixed whereas NoSQL is flexible. (SQL has RDBMS schema consisting of relationship between tables and constraints defined initially which is not so for NoSQL - can be altered without handling relationships so any data can be handled with ease)

SQL vs NoSQL

- SQL has vertical scaling (increase resource of server where SQL is present - Central). NoSQL has horizontal scaling (data distributed among other local machines - Distributed).
- SQL has ACID properties. NoSQL has CAP properties (Consistency, Availability, Partition Tolerance)
- NoSQL has the power to handle very large amounts of data unlike SQL.

Examples

- SQL - MySQL, PostgreSQL, MS SQL Server
- NoSQL - MongoDB, Cassandra, HBase

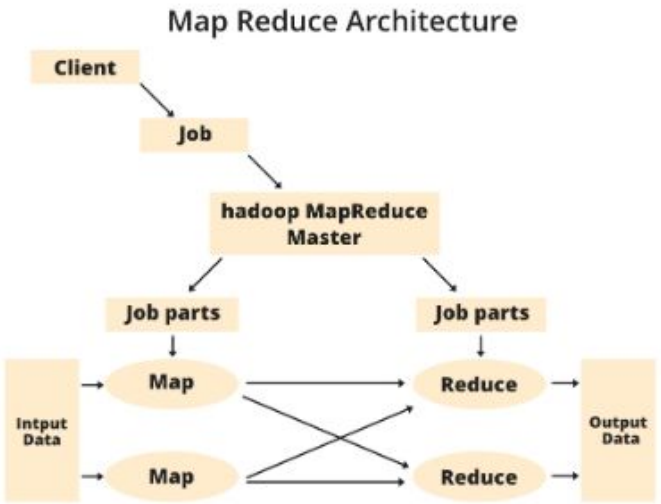
Advantages of NoSQL

- **Flexible Databases :** Allows you to easily make changes to your database as requirements change.
- **Horizontal scaling :** Allow you to scale-out horizontally, meaning you can add cheaper commodity servers whenever you need to.
- **Fast Queries :** NoSQL does not require joins, unlike SQL, making execution of queries faster. Data in SQL databases is typically normalized, so queries for a single object or entity require you to join data from multiple tables. Increasing table size implies faster joins.

MapReduce

- MapReduce is a programming paradigm used for efficient processing in parallel over large data-sets in a distributed manner.
- The data is first split and then combined to produce the final result.

MapReduce - Components



Credits : GeeksForGeeks

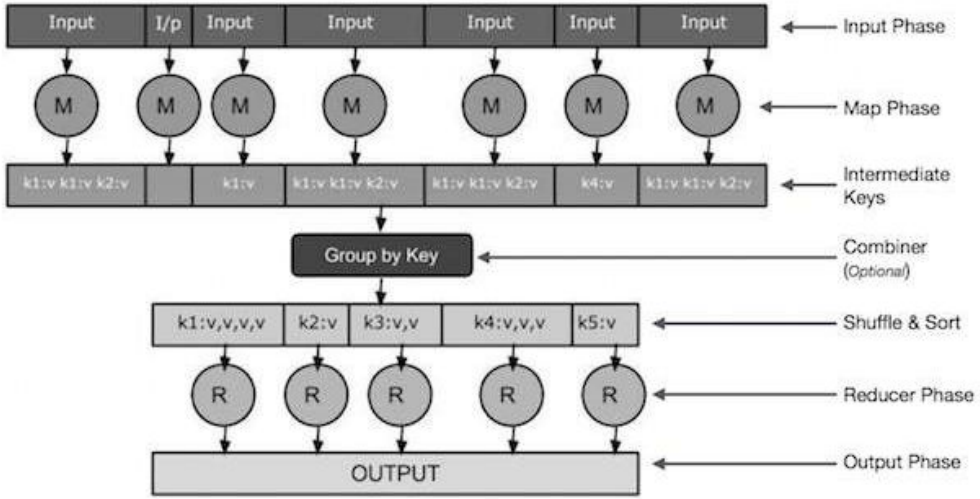
MapReduce - Phases

- Its main use is to map the input data in key-value pairs.
- The input to the map may be a key-value pair where the key can be the id of some kind of address and value is the actual value that it keeps.
- The *Map()* function will be executed in its memory repository on each of these input key-value pairs and generates the intermediate key-value pair which works as input for the Reducer or *Reduce()* function.

MapReduce - Phases

- The intermediate key-value pairs that work as input for Reducer are shuffled and sort and send to the *Reduce()* function.
- Reducer aggregate or group the data based on its key-value pair as per the reducer algorithm written by the developer.

MapReduce Working



Credits : Tutorialspoint

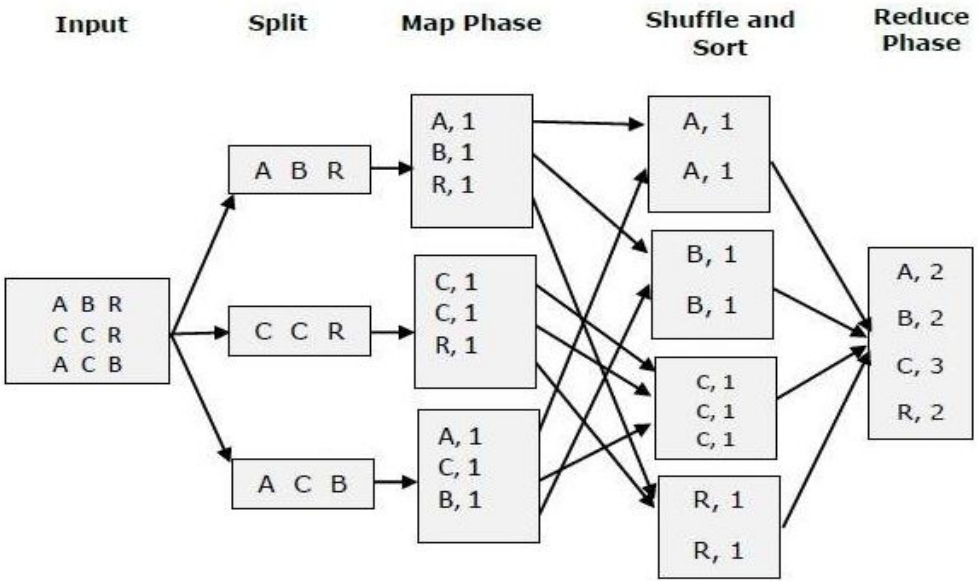
MapReduce Working

- **Input Phase** – Here we have a Record Reader that translates each record in an input file and sends the parsed data to the mapper in the form of key-value pairs.
- **Map** – Map is a user-defined function, which takes a series of key-value pairs and processes each one of them to generate zero or more key-value pairs.
- **Intermediate Keys** – They key-value pairs generated by the mapper are known as intermediate keys.
- **Shuffle and Sort** – The Reducer task starts with the Shuffle and Sort step. It downloads the grouped key-value pairs onto the local machine, where the Reducer is running. The individual key-value pairs are sorted by key into a larger data list. The data list groups the equivalent keys together so that their values can be iterated easily in the Reducer task.

MapReduce Working

- **Reducer** – The Reducer takes the grouped key-value paired data as input and runs a Reducer function on each one of them. Here, the data can be aggregated, filtered, and combined in a number of ways, and it requires a wide range of processing. Once the execution is over, it gives zero or more key-value pairs to the final step.
- **Output Phase** – In the output phase, we have an output formatter that translates the final key-value pairs from the Reducer function and writes them onto a file using a record writer.

MapReduce Working

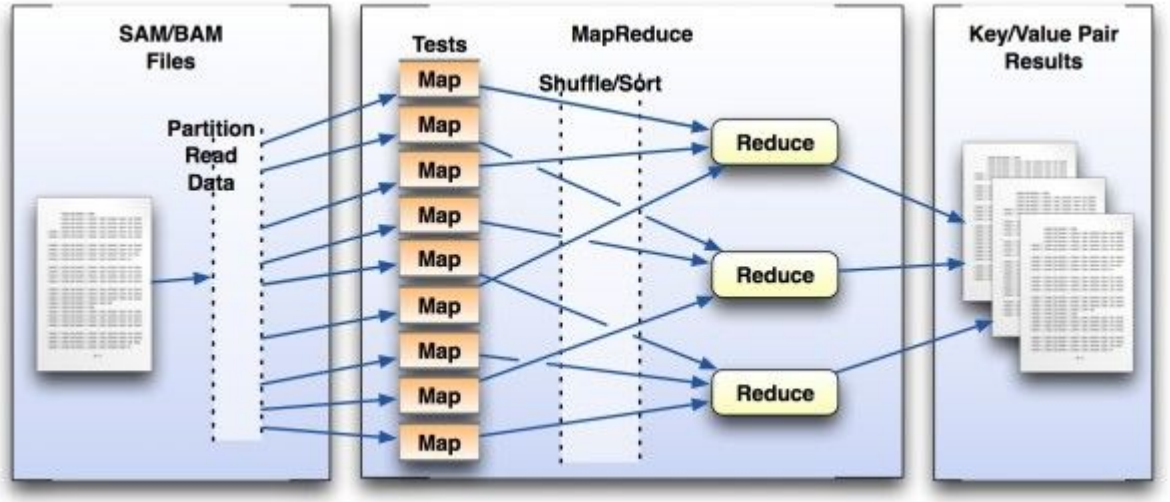


Credits : Tutorialspoint

MapReduce Working - Genome Analysis

Credits : ZymoResearch

MapReduce Working - Genome Analysis

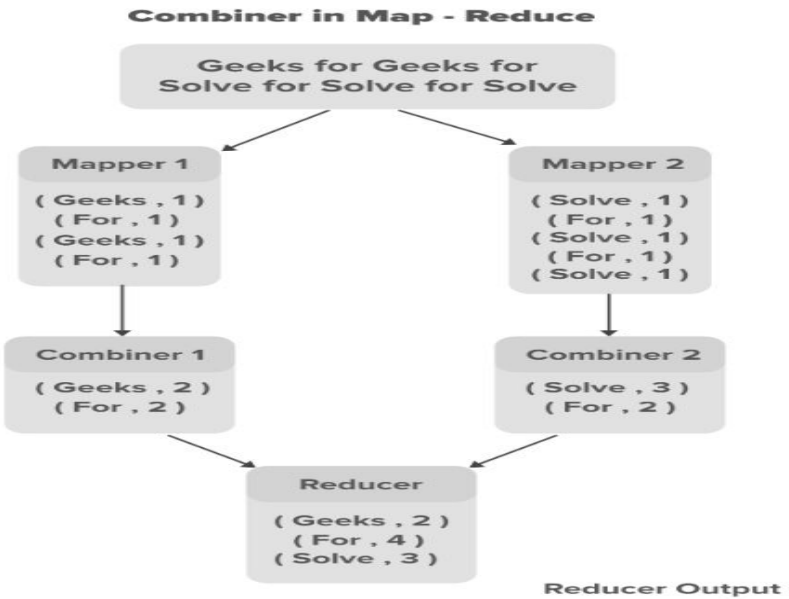


Credits : ResearchGate

MapReduce - Concept of Combiner

- **Combiner** – A combiner is a type of local Reducer that groups similar data from the map phase into identifiable sets. It takes the intermediate keys from the mapper as input and applies a user-defined code to aggregate the values in a small scope of one mapper. It is not a part of the main MapReduce algorithm; it is optional.

MapReduce - Concept of Combiner



Credits : GeeksForGeeks

Thank You