**Delta Report**

Name: Atharva Khaire

BUID: U15258568

Course: Advanced Database Management (CS 779 A1)

How MongoDB utilizes MapReduce?

When processing large amounts of data in MongoDB, MapReduce splits the data into smaller chunks, processes them in parallel, and then combines the results. Map and Reduce are the two processes involved.

The Map phase applies a map function to every document in the input collection of documents. From the document, the map function extracts key-value pairs and outputs them. A group of intermediate key-value pairs is the result of the Map phase.

Each key-value pair in each output from the Map phase is then subjected to a reduce function in the Reduce phase. The reduce function creates a final output by combining the values for each key.

Here is a detailed explanation of how MongoDB employs MapReduce:

* The data that was provided is kept in a MongoDB collection.
* Each document in the collection is given the Map function definition and application. From the document, the Map function extracts key-value pairs and outputs them.
* The Reduce function receives the output of the Map function that has been grouped by key.
* The final output of the Reduce function is the sum of the values for each key.
* The result is either returned or the final output is saved in a fresh MongoDB collection.

Additionally, MongoDB offers features that improve the MapReduce procedure. It permits the employment of a "finalize" function, for instance, to carry out further processing on the results of the Reduce function. Additionally, it offers replication and sharding options to support distributed data processing across multiple servers.

There aren’t any additional code or commands in python to perform MapReduce procedure. MongoDB automatically implements it regardless of the size of data. Although, MapReduce is supposed to speed up the process, yet with python it is really slow during data extraction. MapReduce is primarily important for analytics but MongoDB is more of a storage base. Even in my project, it can’t be used for analytics as during extraction of data into a dataframe, it took ages.