Abstract

Manipulating big data distributed over a cluster is one of the big challenges which most of the current big data oriented companies face. This is evident by the popularity of MapReduce and Hadoop, and most recently Apache Spark, a fast, in-memory distributed collections framework which caters to provide solution for big data management. This paper, present a discussion on how technically Apache Spark help us in Big Data Analysis and Management. The paper aims to provide the conclusion stating apache Spark is more beneficial by almost 50 percent while working on big data. As when data size was increased to 5*10^6 the time taken was drastically reduced by around 50 percent compared to when queried Cassandra without Spark. Cassandra is used as Data Source for conducting our experiment. For this, a experiment is conducted comparing spark with normal Cassandra DataSet or ResultSet. Gradually increased the number of records in Cassandra table and time taken to fetch the records from Cassandra using Spark and traditional Java ResultSet was compared. For the initial stages, when data size was less than 10 percent, Spark showed almost an average response time which was almost equal to the time taken without the use of Spark. As the data size exceeded beyond 10 percent of
records Spark response time dropped by almost 50 percent as compared to querying Cassandra without Spark. Final record was analyzed at $5 \times 10^6$ records. As the data size was increased, Spark was proved better than the traditional Cassandra ResultSet approach by almost reducing the time taken by 50 percent for really big dataset as our case of $5 \times 10^6$ records.

References


Index Terms

Computer Science

Information Sciences

Keywords

Spark, RDD, MapReduce, Hadoop, Cassandra