Review on Data Partitioning Strategies in Big Data Environment

Haneen A. A[1,2], A. Noraziah[1, 2,] Ritu Gupta[1,2]

[1]Faculty of Computer System and Software Engineering, University Malaysia Pahang,
26300, Kuantan, Pahang, Malaysia.

[2]IBM Center of Excellence, University Malaysia Pahang,
26300, Kuantan, Pahang, Malaysia.

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Therefore, applications have to be confronted with the challenges of big data. Thus, data partitioning strategy plays an important role in the database. There are several data partitioning strategies that solve some problems such as low scalability, hot spot and low performance and so on. In this paper we discuss advanced partitioning strategies, their implementation.

Keywords: Big data, Data Partition, Partitions strategies.

1. INTRODUCTION

The world has been in the midst of an extraordinary information explosion over the past decade, spurred by the rapid growth in the use of the Internet and the number of connected devices worldwide. Most of applications will face with the challenges of Big Data. Databases are designed to manage large quantities of data, allowing users to query and update the information they contain. The database community has been developing algorithms to support fast or even real-time queries over relational databases, and, as data sizes grow, they increasingly opt to partition the data for faster subsequent processing. One of the optimized solution for database problems in big data is Partitioning. Partitioning strategy plays an important role in the database. Therefore, different types of partitioning are designed to manage database because of the characteristics of scalability, availability and fault-tolerance. The existing data partitioning strategies will cause some problems such as low scalability, hot spot and low performance and so on. For that creating and maintaining multiple data copies has become a key computing system We present the following contributions: Section 2 introduces some related researches on data partition strategies. In Section 3, we define some of those strategies. Section 4 presents our review. The last section summarizes the paper and gives a short overview of future works

2. BACKGROUND

The topic of big data based databases has emerged