Chapter 9
Multidimensional Analysis of Big Data

Salman Ahmed Shaikh
National Institute of Advanced Industrial Science and Technology (AIST), Japan

Kousuke Nakabasami
University of Tsukuba, Japan

Toshiyuki Amagasa
University of Tsukuba, Japan

Hiroyuki Kitagawa
University of Tsukuba, Japan

ABSTRACT

Data warehousing and multidimensional analysis go side by side. Data warehouses provide clean and partially normalized data for fast, consistent, and interactive multidimensional analysis. With the advancement in data generation and collection technologies, businesses and organizations are now generating big data (defined by 3Vs; i.e., volume, variety, and velocity). Since the big data is different from traditional data, it requires different set of tools and techniques for processing and analysis. This chapter discusses multidimensional analysis (also known as on-line analytical processing or OLAP) of big data by focusing particularly on data streams, characterized by huge volume and high velocity. OLAP requires to maintain a number of materialized views corresponding to user queries for interactive analysis. Precisely, this chapter discusses the issues in maintaining the materialized views for data streams, the use of special window for the maintenance of materialized views and the coupling issues of stream processing engine (SPE) with OLAP engine.
Multidimensional Analysis of Big Data

INTRODUCTION

Due to the increase of stream data sources, such as sensors, GPS, micro blogs, etc., the need to aggregate and analyze stream data has increased. Many organizations require instant decisions exploiting the latest information from the data streams. For instance, timely analysis of business data is required for improving profit, network packets need to be monitored in real time for identifying network attacks, etc. Online analytical processing (OLAP) is a well-known and useful approach to analyse data in a multi-dimensional fashion, initially given for disk-based static data. OLAP requires hierarchical arrangement of dimensional attributes (C.E.F. et al, 1993). For the effective OLAP analysis, the data is converted into a multi-dimensional schema, also known as star schema. The data in star schema is represented as a data cube, where each cube cell contains measure across multiple dimensions. A user may be interested in analysing data across different combination of dimensions or examining different views of it. These are often termed as OLAP operations and to support these operations, data is organized as lattice nodes. Each vertex of lattice corresponds to an aggregate query, called (OLAP queries). Materialized views are maintained for the selected lattice vertices. Maintenance of materialized views is handled by OLAP engine, which require clean and structured data stream. Since the raw data stream is inherently unstructured and contains missing values, Stream Processing Engines (SPE) are usually used with OLAP engine to provide it clean and structured stream. There exist many stream processing engines such as STREAM, S4 and Borealis (Arasu et al, 2016; Neumeyer et al, 2010, Cangialosi et al, 2005). These SPEs uses continuous queries to process data streams continuously.

To support OLAP over data streams, SPEs are usually coupled with OLAP engine. J. Han, et al. (2007) in was the first to propose Stream Cube architecture to facilitate OLAP for continuous stream data. In order to reduce the query response time and to reduce the storage cost, Stream Cube keeps the distant data at high granularity level, while only very new data at low granularity level. To further reduce the query response time, the stream cube pre-computes some OLAP query results at coarser, intermediate and finer aggregation levels. However, in their work, the use of SPE is not taken into account and it is not possible to perform the fine-grained analysis of the distant data as only the most recent data is available at finer resolution. Moreover, a few materialized query results between two layers, i.e., observation layer and minimal interesting layer, are available and users cannot obtain the aggregation results beyond the minimal interesting layer.

In section 4 of this chapter we present a stream OLAP architecture consisting of an SPE and an OLAP engine which is based on our research work. To get the required results, the naive approach is to materialize and maintain OLAP query results of all vertices representing the combinations of dimensions and their hierarchies in a lattice. However, this results in a large number of materialized cubes and will also affect the performance of the SPE. Moreover, all the aggregation results are not needed at all the time. Thus, a cost-based optimization algorithm is discussed. The algorithm decides which queries should be materialized in cooperation with the SPE and which query results should be derived on-demand from other materialized query results. The optimization algorithm tries to minimize the query processing cost by keeping in view the available memory (Nakabasmi et al, 2015).

A number of solutions have been proposed for OLAP analysis on data streams in the near past (Han et al, 2005; Duan et al, 2011; Sadoghi et al, 2016). The requirement to produce real time OLAP analysis on fast and evolving data streams is not possible unless the data to be analysed reside on primary memory.