International Journal of Data Analytics
Volume 2 • Issue 1 • January-June 2021

View Materialization Over Big Data

Akshay Kumar, Jawaharlal Nehru University, India
T. V. Vijay Kumar, Jawaharlal Nehru University, India

ABSTRACT

Advances in technology have resulted in the generation of a large volume of heterogeneous big data for large enterprises engaged in e-commerce, healthcare, education, etc. This is being created at a rapid rate but is low in its veracity. This big data includes large sets of semi-structured and unstructured data and is stored over a distributed file system (DFS). This data can be processed in a fault tolerant manner using several frameworks, tools, and advanced database technologies. Big data can provide important information, which can be used for business decision making. View materialization, which has been widely studied for structured databases or data warehouse, has been extended to big data to enhance efficiency of big data query processing. This paper focuses on the selection of big data views for materialization. The big data views can be identified by extracting a set of query attributes from the set of query workload of an enterprise. The query attributes are interrelated resulting in the creation of alternate access paths for query evaluation. The cost of query processing using big data views involves the integrity of different data types of heterogeneous big data, frequency of queries, change in the size of big data, selected sets of big data materialized views, and updates on big data and these sets of materialized views. The cost of query processing is computed using the stored size of big data views on the DFS system, which is a consistent processing framework of DFS. A big data view selection algorithm that is capable of selecting views from structured, semi-structured, and unstructured data has been proposed in this paper. The proposed algorithm would select big data views that would result in faster processing of most user queries resulting in efficient decision making.

KEYWORDS

Big Data, Database Systems, Decision Making, View Materialization

1 INTRODUCTION

Information and knowledge are the pre-eminent constituents of present day society. The commencement of information societies can be attributed to Internet technologies, especially Web 2.0, which brought disruptive changes in data generation, storage and communication capabilities. Applications of the World Wide Web (*WWW*) on computers and smart devices resulted in data capture, storage, display and processing of large volumes of structured, semi-structured and unstructured data for making informed business decisions. This data had large volume, variety (heterogeneity) and velocity (continuity) but low veracity (trustworthiness), also called "The Four *V*s of Big Data" (Zikopoulos et al. 2011; Gupta et al., 2012; Kumar & Vijay Kumar, 2015). This data is required to be processed at a high speed, so that the value or knowledge is generated in a timely manner in order to facilitate efficient decisions.

A large number of applications, such as social media, education, e-commerce, healthcare, scientific, *IoT*, space, meteorological, mobile applications etc. produce Big Data. Since Big data has a large volume and variety, it cannot be efficiently handled using already existing database tools

DOI: 10.4018/IJDA.2021010103

and techniques. This resulted in the development and enhancement of relational databases and data warehouses, *NoSQL* databases, Graph based data models and distributed file systems (Manyika et al., 2011).

Historically, research related to management of large data can be traced to the relational database model, which is based on the mathematical concept of relations (Codd 1970). The relational model required query processing involving select, project, aggregation, join and other operations. Amongst these, the join operation is the costliest operation for query evaluation. To address this, several query optimization techniques were designed (Chaudhari, 1998) of which view materialization is one such technique. A materialized view has been defined, as the evaluated query that has been cached for future query evaluation (Gupta, 1996). Given a set of relations and materialized views, a user query is required to be reformulated to utilize materialized views for query evaluation. View Materialization is concerned with selecting views, based on the query workload, that conforms to storage space constraints (Chirkova et al. 2001; Mami & Bellahsene, 2012) with the purpose of minimizing the query processing cost. The problem is to find an appropriate subset of views that can improve query response time subject to data updates and storage space constraints. This problem is referred to as the View Selection Problem, which is shown to be a *NP-Hard* problem (Harinarayan et al., 1996; Chirkova et al. 2001).

View materialization has also been studied in the context of data warehouses (Harinarayan et al., 1996; Gupta, 1996; Roussopoulos, 1998). (Harinarayan et al., 1996) modeled the view materialization problem for a data warehouse using a lattice structure, grouping data over the dimensions, whereas, (Gupta H.,1996) has studied the problem using the AND-OR graph structure of the views. One of the key difference between the two approaches is that the latter approach also considers the updates on the data. Both of these solutions have been studied further by many researchers (Kumar & Vijay Kumar, 2018; Vijay Kumar & Haider, 2010)

This paper is concerned with view materialization in the context of Big data. Section 2 discusses the Big data architecture and related Big data stores. Section 3 lists the issues of Big data materialization and the motivation of the work presented in this paper. Section 4 defines the selection of views for materialization in the context of Big data. Section 5 presents the experimental results.

2 THE BIG DATA ARCHITECTURE FOR LARGE DATA STORES

Big data is not just the data of multiple thousand entities but the repetitive data generated by such entities that make data Big (Jacobs, 2009). For example, the number of users, products, vendors etc. of e-commerce web sites may not make the data Big, but storing the number of transactions and other actions made by these users over the e-commerce application is what that makes the data Big. (Jacobs, 2009) suggested that a database of size more than 100 *GB* involving joins on non-key attributes, will require potentially very large computing resources and, therefore, cannot be considered as small data. Big data is also heterogeneous, as it includes structured data (Relational data); semi-structured data (*XML* or similar object data); unstructured data (text, voice, audio), data from web (social media, blogs, web logs, click streams); spatial data (coordinates, *GPS* data); data from sensors and *RFID*; and scientific data.

A Big data application involves the capture, integration and analysis of data from a mix of heterogeneous sources including an organization's databases, data warehouse, semi-structured or *XML* data, data from the web, scientific data and other unstructured data. Big data analytics is useful in social applications, medical science applications, educational application, e-commerce applications etc. (Global Pulse, 2012) illustrates the importance of Big data applications to help the weaker and vulnerable sections of society against the extreme effects of recession and climatic catastrophes. Other major applications of Big data are in e-business, healthcare, traffic management and web analytics. (Paul A. et al. 2016) has used Big data analytics to model human behavior using inputs from various