Chapter 15

Data Streams Processing Techniques Data Streams Processing Techniques

Fatma Mohamed
Ain Shams University, Egypt

Nagwa L. Badr
Ain Shams University, Egypt

Rasha M. Ismail
Ain Shams University, Egypt

Mohamed F. Tolba
Ain Shams University, Egypt

ABSTRACT

Many modern applications in several domains such as sensor networks, financial applications, web logs and click-streams operate on continuous, unbounded, rapid, time-varying streams of data elements. These applications present new challenges that are not addressed by traditional data management techniques. For the query processing of continuous data streams, we consider in particular continuous queries which are evaluated continuously as data streams continue to arrive. The answer to a continuous query is produced over time, always reflecting the stream data seen so far. One of the most critical requirements of stream processing is fast processing. So, parallel and distributed processing would be good solutions. This paper gives (1) analysis to the different continuous query processing techniques; (2) a comparative study for the data streams execution environments; and (3) finally, we propose an integrated system for processing data streams based on cloud computing which apply continuous query optimization technique on cloud environment.

INTRODUCTION

Recently a new class of data-intensive applications has become widely recognized: applications in which the data are modeled best not as persistent relations but as transient data streams. However, their continuous arrival in multiple, rapid, time-varying, possibly unpredictable and unbounded streams appear to yield some fundamentally new research problems. These applications also have inherent real-time requirements, and queries on the streaming data should be finished within their respective deadlines (Kapitanova, Son, Kang & Kim, 2011; Lijie & Yaxuan, 2010). In this context, researchers have proposed

DOI: 10.4018/978-1-5225-2229-4.ch015

a new computing paradigm based on Stream Processing Engines (SPEs). SPEs are computing systems designed to process continuous streams of data with minimal delay. Data streams are not stored, but are processed on-the-fly using continuous queries. The latter differs from queries in traditional database systems because a continuous query is constantly "standing" over the streaming tuples and results are continuously output. In the last few years, there have been substantial advancements in the field of data stream processing. From centralized SPEs, to Distributed Stream Processing Engines (DSPEs), which distribute different queries among a cluster of nodes (interquery parallelism) or even distributing different operators of a query across different nodes (interoperator parallelism). However, some applications have reached the limits of current distributed data streaming infrastructures (Gulisano, Jimenez-Peris, Patino-Martinez, Soriente & Valduriez, 2012).

Because of the continuous changes in input rates, DSPSs need techniques for adjusting resources dynamically with workload changes. Making decisions when to update resource allocation in response to workload changes and how, is an important issue. Effective algorithms for elastic resource management and load balancing were proposed, where resizes the number of VMs in a DSPS deployment in response to workload demands by taking throughput measurements of each involved VM (Cerviño, Kalyvianaki, Salvachúa & Pietzuch, 2012; Fernandez, Migliavacca, Kalyvianaki & Pietzuch, 2013; Gulisano et al., 2012). Thus, cloud computing has emerged as a flexible for facilitating resource management for elastic application deployments at unprecedented scale. Cloud providers offer a shared set of machines to cloud tenants, often following an Infrastructure-as-a-Service (IaaS) model. Tenants create their own virtual infrastructures on top of physical resources through virtualization. Virtual machines (VMs) then act as execution environments for applications (Cerviño et al., 2012).

Thus, we categorize research challenges in data streams to: 1) Continuous queries processing which focus on continuous queries optimization, how to provide real time answering of continuous queries, how to process different typed of continuous queries, and how efficiently process multiple continuous queries. 2) Data streams execution environments where different environments were proposed to execute data streams such as parallel, distributed, and cloud environments. Where, exploiting parallelism and distribution techniques to fast data streams processing, also exploiting virtualization strategies in cloud to provide elastic processing environment in response to workload demands.

The rest of the chapter is organized as follows: related background is introduced first, and then efficient processing techniques for continuous queries, including different algorithms for effective continuous query optimization are presented. Then, we present different execution environments for data streams, which include parallel, distributed, and cloud environments. Then, we present the related research issues. And then our proposed system for data streams processing over cloud computing is presented. Then future research directions are presented. Finally, we present the conclusion.

BACKGROUND

Data Streams

Data streams are the data which generated from most of the recent applications such as sensor networks, real-time internet traffic analysis, and on-line financial trading. This data has a continuous, unbounded, rapid and time-varying nature rather than finite stored data sets which generated from the traditional applications. Thus the traditional database management systems (DBMSs) are not suitable with this