In-Memory Analytics

Jorge Manjarrez Sánchez
Instituto Tecnologico Superior de Jerez, Mexico

INTRODUCTION

Processing speed and data set volume are a challenge for traditional analytics systems. Consider when do you want to make a recommendation for an online customer: right there while she is still browsing your catalogue or next time (if) she returns back? When do you want to detect a credit card is fraudulent: before a transaction is committed or after a client complain some days later? Score or decide when a credit card transaction is fraudulent or an incoming email is spam, is relatively easy from a computationally point of view provided the appropriate algorithms are implemented correctly. But in real-life systems where hundreds of thousands of such transactions arrive simultaneously and the correct decision must be taken instantaneously or in near real-time, speed is crucial. Even though only for the second example question an incorrect false negative, i.e., the no detection of a fraud, can have severe negative consequences, these examples highlight two main challenges in data analytics: the need to process ever larger data volumes and the need of obtaining fast or real time results. These are the concerns of In-memory analytics.

In what follows, we provide a definition and some concepts necessary to understand the problem and solution approach of in-memory analytics. Then we present a review of some important technological proposals to deal with two main challenges in data analytics: speed and scalability. The presentation is made from a data management perspective, independent of their nature (structured or unstructured), because an efficient data management strategy provides the infrastructure to handle and enable fast access to voluminous data sets for their analytical processing.

DOI: 10.4018/978-1-5225-2255-3.ch157

BACKGROUND

Analytics is the processing of data for facts and information discovery. Embedded within data there are facts about some events, perhaps several and a variety of them. These facts are knowledge helpful for taking decisions, recommendations systems, fraud detection, sentiment analysis in social networks and customer identification and many other applications. In a broad sense analytics can provide an answer for questions such as what happened? Why did it happen? What will happen? What should I do? And they correspond to the task of descriptive, diagnostic, predictive and prescriptive analytics respectively. The answer to these questions are obtained by the application of techniques from statistics, machine learning and computer science in the processing of data.

Analytics is the processing of data for making sense of all of it. Data sets are very large and to process them one has to cope with two orthogonal concerns: accuracy and speed of results. Accuracy needs to process as much data as possible, not just random samples. Not using complete datasets leads to loss of information (Wang, Callan, & Zheng, 2015) and obtained responses are approximations with certain level of error. Accuracy of the statistical model and henceforth that of predictions are proportional to samples size and quality. The bigger the better, but even that we can have today larger data sets, their processing takes considerable time. It consumes more processing time and disk input/output (IO) operations because data must be swapped back and forth from hard drives to RAM and CPU caches. But due to speed differences in data access and transfer between these devices, there is a processing bottleneck. One approach to avoid IO processing bottlenecks is to use as

much as possible RAM memory, hence the name In-Memory Analytics, where data is loaded and kept into RAM for their fast consumption by the processing algorithms.

FAST ANALYTICS

In this section we address the speed challenge. When the speed of results is a factor to consider for efficient analytics, and taking into consideration that it must be used as much data as possible for the abovementioned reasons, then one must reduce or eliminate processing bottlenecks. They can be the IO latency and CPU usage. CPU is good at computing mathematical operations and multicore CPU's enhance their performance. IO depends on the type of disk and RAM. Solid state disks are faster than hard disk drives, but random access memory is faster. A typical hard drive has a latency of 5 ms while RAM has only 100 ns, it is 50000 times faster. But also the faster the more expensive. There is one more rapid type of memory: cache memory, which is a small and expensive memory within the CPU die. Main memory or In-memory refers to the processing of data performed in RAM.

The data management infrastructure provides the processing capabilities according to the workload type. The typical workloads in a database, can be Online Analytical Processing (OLAP) or Online Transaction Processing (OLTP). The difference between both are their usage patterns, size of result sets and processing speed. Traditionally OLTP deals with faster small transactions and OLAP seeks for answer precision on very large data sets, but nowadays speed is an important factor to provide for both of them, hence the use of in-memory analytics, the usage of RAM fast data access and to enhance analytics performance. But the issue is that real data sets are usually larger than the memory affordable and available in a system. Also note that not all available memory can be used for data storage, enough space must be allocated to handle temporary objects and intermediate results of computations. To address this concern, one approach is to use compression. For instance, by using compression and column-based storage it is possible to reduce storage space because only required attributes are retrieved from disk. Additionally, compression leverages CPU ability to handle numbers instead of text which combined with columnar storage allows an application to keep large quantities of data in memory. A drawback is that this approach is read optimized, insert or update operations are not easy. Some other approaches to make efficient use of memory while computing intermediate results are discussed in (Duan, Li, Tang, Xiao, & Li, 2013) within the context of Spark, an open source in-memory analytics platform. Their proposal makes automatic management of the memory space used by intermediate results susceptible of being used in further computations instead of letting the full responsibility to the programmer.

The idea of in-memory databases is not new (Lehman & Carey, 1986), it has been studied since more than two decades ago, but nowadays it is possible for real enterprise applications to process larger amounts of data because of RAM price drops. In addition to high prices, there was a technical limitation of the operating systems for the addressing of larger memory beyond some GB. Today these limitations have gone and prices started to decrease and now it is common for a single server to have 1 TB of RAM and a cluster of computers to sum up several Terabytes. The scalability issue to cope with big data sets is the topic of the next section.

In-memory analytics appliances from major vendors integrate an optimized computer with an in-memory database. The main database vendors have developed their in-memory proposal: Microsoft Hekaton (Diaconu et al., 2013) is an engine integrated within SQL Server, the memory optimized part of the database is marked as memory optimization. Code of stored procedures can also be compiled to enhance CPU usage. Oracle TimesTen (Lahiri & Folkman, 2013) is a separate product that can also be used as a fast data cache