Available online at www.sciencedirect.com

ScienceDirect

The Journal of Finance and Data Science 1 (2015) 33–41
http://www.keaipublishing.com/en/journals/jfds/

Latency critical big data computing in finance

Xinhui Tian [a,b,*], Rui Han [a], Lei Wang [a,b], Gang Lu [a,b], Jianfeng Zhan [a]

[a] *State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, China*
[b] *University of Chinese Academy of Sciences, China*

Received 6 March 2015; revised 14 March 2015; accepted 30 March 2015
Available online 28 August 2015

Abstract

Analytics based on big data computing can benefit today's banking and financial organizations on many aspects, and provide much valuable information for organizations to achieve more intelligent trading, which can help them to gain a great competitive advantage. However, the large scale of data and the critical latency analytics requirement in finance poses a great challenge for current system architecture. In this paper, we first analyze the challenges brought by the financial latency critical big data computing, then propose a discussion on how to handle these challenges from a perspective of multi-level system. We also talk about current researches on low latency in different system levels. The discussions and conclusions in the paper can be useful to the banking and financial organizations with the critical latency requirement of big data analytics.

Keywords: Big data; Financial analytics; Latency

1. Introduction

As one of the essential factors of system and network performance, latency indicates how fast a user can get a response after the user sent out a request. Low latency, which means systems response quickly to actions, can make users feel more natural and fluid than long response time.[1] In the financial market, as more and more business trades and banking operations are executed online, lower latency now means more revenues, especially for companies which adopt high frequency trading to earn huge profit. High frequency trading means to rapidly trade large volumes of securities by using automated financial tools.[2] A millisecond decrease in a trade delay may boost a high-speed firm's earnings by about 100 million per year,[3] and also helps a firm to gain great competition advantage.

Traditionally, financial organizations can achieve low latency via adopting high performance computers, which provide great processing capability, especially the capability of floating-point processing. When the processing capability is not enough, high performance computers can also be scaled via two methods, which are scale up (adding more CPUs or memory to a single computer) and scale out (adding more computing nodes, and connecting them with

* Corresponding author. State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, China.

E-mail address: tianxinhui@ict.ac.cn (X. Tian).
Peer review under responsibility of China Science Publishing & Media Ltd.

http://dx.doi.org/10.1016/j.jfds.2015.07.002

high performance interconnects). However, as the size of data needed to be analyzed is growing dramatically in the last few years, the primary bottleneck has shifted to the performance of storage system, and the frequent data movement in traditional high performance computing can significantly impact the latency when the volume of processing data is huge. Therefore, the system architecture for financial computing needs to be improved in such a situation.

Such data explosion problem can also be called as the big data problem, which has been a hot trend in recent years. The big data means that the collected data sets are becoming too large and complex to be processing via traditional data processing applications.[4] IDC forecasts that the world's data will be doubling every two years with 1.8 trillion gigabytes expected to be created,[5] and the global volume of data will increase from 130 to 40,000 exabytes by 2020.[6] Another report from HP also presents that the data size in financial world is really big now. For example, there are more than 10,000 payment card transactions executing per second across the world in 2012, and total number of U.S. online banking households is about 66 million in 2014.[7] The New York Stock Exchange also has to process about 2 TB data daily in 2012, and expects to exceed 10 PB a day by 2015.[8]

The large scale of data contain enormously valuable information, and analytics based on big data can provide financial organizations with more business opportunities and the possibility to gain a more holistic view of both market and customers. Big data analytics can benefit banking and financial market firms in many aspects, such as accurate customer analytics, risk analysis and fraud detection. These approaches can lead to smarter and more intelligent trading, which can help organizations to avoid latent risks and provide more personalized services, thus to get a higher degree of competition advantage.

According to the report from SAP,[3] the profitability keeps falling in recent years, and organizations are now evolving towards smart trading based on big data analytics. Besides designing more complex computing model and system, how to make such large scale computation real time is still an very important problem that is needed to be considered seriously. In fact, many approaches of big data analytics would not be beneficial if the latency were not be controlled in a low level, especially for some high speed analytics such as the risk management of stock trading. Another fact is that, the rapid growing data flood also make the political, social and economic events that can impact the financial markets, now take only a few minutes, which may take days before. Such a challenge requires financial organizations process the events immediately as they are acquired. A complex computation may also demand process the new coming data together with some historical records, which can involve a huge size of input data. For example, NYSE Euronext has employed big data analytics to detect new patterns of illegal trading, and they need to process about two terabytes of data daily and get the analytics result in near real time.[9] Hence, this is difference from traditional data stream problem. Researches on low latency and high throughout data stream processing on a big data store are desired.

Consequently, the latency critical big data analytics in finance, which require the latency to be kept in a critical level, requires the system to satisfy many different types of latency demands. One is the absolute end-to-end latency, which directly decides the speed of market access and real time trading. Another is the new demand on how to process complex analytics on huge amounts of data faster than competitors to capture trading opportunities. As the data size keeps growing rapidly, many organizations require the analytics latency close to near real time, thus to be able to extract valuable information for competitive advantage. This demand poses a great challenge to current system architecture with many difficult problems. For example, how to organize such large scale of data for history analytics, how to process the streaming data immediately, and how to effectively execute jobs with different priorities. All these problems are the hurdles that has to be moved for smarter trading. In this paper, we mainly discuss how to guarantee the second latency demand from a multi-level system perspective. We first give a detailed discussion on the challenges to achieve low latency big data computing in section 2. In Section 3, we give a discussion about a multi-level system solution for this challenge. We also discuss recent researches on latency critical big data computing in both industry and academia in Section 4.

2. Challenges

In this section we try to discuss and summary the challenges to achieve low latency financial big data analytics. The first problem is how to deal with the problem of massive data storage and organization. Many organizations need to keep historic data of many years for trend prediction and other complex analytics, which poses a great challenge for the reliability of the storage system. The rapidly growing data also require the storage architecture to provide good scalability to support scaling out when the data size increases up to the storage boundary. Another problem is the