Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 103 (2017) 536 – 542

Procedia Computer Science

XIIth International Symposium «Intelligent Systems», INTELS'16, 5-7 October 2016, Moscow, Russia

The Big Data approach to collecting and analyzing traffic data in large scale networks

L.U. Laboshin*, A.A. Lukashin, V.S. Zaborovsky

Peter the Great Saint-Petersburg Polytechnic University, 29, Polytechnicheskaya street,, St.Petersburg, 195251, Russia

Abstract

Information processing is currently one of the most vital tasks. With the growth and development of information and telecommunication technologies increased the volume of data transmitted over the Internet. Simultaneously with the processing large amount of information raises the question of its protection. The given paper proposes a distributed cloud-computing framework based on the Big Data approach where both storage and computing resources can be scaled out to collect and process traffic from a large-scale network in a reasonable time.

Peer-review under responsibility of the scientific committee of the XIIth International Symposium "Intelligent Systems"

Keywords: Network traffic analysis; actor model; mapreduce; DFS; akka; scala; netgraph

1. Introduction

Currently, computer technology continues to evolve. The increase of computational power leads to a change of content, for example, recent progress in display technology has resulted in the ultra high-definition video. Actively developing the internet of things (IoT) means more devices utilizing a data networks. Most of these changes lead to an increase in the volume of information transmitted over networks. Throughput is constantly increasing in order to provide an effective service. Currently, the technology GigabitEthernet (1 Gbit/s) has almost completely replaced FastEthernet (100Mbit/s) on personal computers. In ISP networks is widely used 10 Gbit/s, is the transition to 40 Gbit/s and more. The task of information security is still relevant. Malicious software and attacks on information

* Corresponding author.
E-mail address: laboshinl@neva.ru

Peer-review under responsibility of the scientific committee of the XIIth International Symposium "Intelligent Systems"
doi:10.1016/j.procs.2017.01.048

systems can have a major impact not only on the commercial companies but also for public service. And with the growth of information influence will continue to increase. There are many tools for information security: antivirus software, firewalls, intrusion detection system (IDS), and others. And if the tools are working in real-time continue to evolve along with the threats to information security, needing to analyze previously accumulated data, in particular, network intrusion detection system, should be considered separately.

As already noted, such funds shall require the prior accumulation of a data packet. If you install network IDS in a channel with a bandwidth of 1Gbps it turns out that every minute it is necessary to analyze up to 15 GB of data. In the case of channel 10Gbps, this amount rises up to 150 GB of data. And if there is a need to analyze the traffic for an hour, then the volume is 900 and 9000 Gigabytes respectively. Storing and processing it becomes challenging. The analysis time of such a volume of data makes it difficult to quickly identify an attack or a policy violation access. These problems can be solved using new approaches for data storing and processing relating to the “big data" computing task class. The most famous approach is the use of distributed filesystems to store very large amounts of data and use computation models such as MapReduce or BSP for analysis. There are ready software and hardware platforms for these purposes. However, these tools are not suited for collecting and analyzing network traffic.

In this paper, we propose a distributed framework for collecting and analyzing network traffic that provides:

1. Load scalability: The ability to easily expand and contract its resource pool to accommodate heavier or lighter loads or a number of inputs;
2. Reliable storage capacity to handle large amounts of data;
3. Provides analysis of the entire amount of data in reasonable time.

On the basis of this framework, it is possible to build a new type of intrusion detection system which is able to handle very large amounts of data. Recent development in cloud computing enables the ability of such system to scale out if necessary, for example, to reduce the analysis time, by adding the desired number of compute nodes. Due to architecture scalability, the system will be flexible enough to meet the continuous growth of data volumes.

The aim of this work is to develop a system architecture able to retrieve, store, and process network traffic of high-speed data channels in a distributed manner.

2. Recent work

Various tools such as Wireshark network protocol analyzer or CoralReef are available for monitoring, measurement and analysis passive Internet traffic data. However, most of these tools designed to run on single high-performance server which is not capable of handling huge amount of traffic data captured at very high-speed links. Lee et al. [1, 2] proposes a Hadoop-based traffic analysis system with a limitation of the only collecting periodic statistic like total traffic and host/port count. RIPE library does not consider the parallel-processing capability of reading packet records from distributed filesystem which leads to performance degradation. Vierira et al. [3, 4] have developed a parallel packet processing system only for JXTA-based applications. Jethoe [5] in his paper proposed a method for detected DDOS attacks using Hadoop. In our papers [6, 7] we proposed a MapReduce approach to performing deep packet inspection (DPI) and application-aware network content inspection with TCP-flow extraction on large traffic traces.

3. Scalable data processing technologies

3.1. MapReduce

MapReduce is the model of parallel programming developed by the Google company as a solution to the information search tasks. Main opportunities of this technology are

- automatic multi-sequencing of the task on a cluster of standard architecture servers;
- load balancing between cluster nodes;
- protection against failures of the equipment by restarting of the task on the other node of a cluster;
- the distributed filesystem for data storage on internal disks of servers of a cluster.