Abstract

Nowadays large volumes of data are being generated continuously from different resources in various formats. To efficiently analyze the data, Big Data Processing Platforms have been developed to replace the traditional systems like SQL. The new platforms are usually built upon a cluster of machines, with one machine working as the master and the others working as slaves. The master node, as a central manager, arranges the data to be stored on each slave and coordinates the slaves to process the data in a distributed manner. In such a complex and distributed system, resource management is critical for the system performance and security is another big concern especially when participating nodes are possible to be compromised.

This dissertation mainly focuses on three challenges in resource management of big data processing platforms. First, the status of a cluster is dynamic. The performance of a cluster can be affected by hardware failure and service interference. The prefixed setting cannot arrange the resource dynamically. Second, the data flow is complicated. Many popular systems process an application in multiple stages, and one stage may depend on the result of earlier stages. Traditional scheduling policies such as FIFO cannot handle the resources well. Third, data exchanging is frequent and time-exhausting. Especially in an IoT cluster, the small devices communicate through WIFI, which suffers from signal interference and generates a significant network delay. In addition, machines in a cluster have various processing and storage capabilities, which makes above problems even more challenging. This thesis develops new resource management strategies to solve previous challenges respectively. We create EA²S² that arranges data storage automatically, AutoPath that allocates resources to each stage intelligently and owlBit that transfers data packet efficiently in heterogeneous environments.

This dissertation also studies the security of big data processing platforms from the aspects of resource abuse and data verification. In a cluster of thousands of machines, it is common for several of them to be attacked, and the compromised nodes can launch severe attack without being noticed. These attacks may abuse the resources to degrade system performance or falsify the original data to generate misleading results. In this thesis, we design such kind of attacks, and provide solutions against them.