Contribution ID: **338**   Type: **Poster**

HDTFS：Cost-effective Hadoop Distributed & Tiered File System for High Energy Physics

Monday 24 October 2022 16:10 (30 minutes)

With the scale and complexity of High Energy Physics(HEP) experiments increase, researchers are facing the challenge of large-scale data processing. In terms of storage, HDFS, a distributed file system that supports the "data-centric" processing model, has been widely used in academia and industry. This file system can support Spark and other distributed data localization calculations, researching the application of Hadoop Distributed File System(HDFS) in the field of HEP is the basis for ensuring the application of upper-layer computing in this field. However, HDFS expand the cluster capacity by adding cluster nodes, this way cannot meet the high cost-effective system requirements for the persistence and backup process of massive HEP experimental data. In response to the above problems, researching Hadoop Distributed & Tiered File System(HDTFS) that supports disk-tape storage, taking full advantage of the fast disk access speed and the advantages of large tape storage capacity, low price, and long storage period, to solve the high cost of horizontal expansion of HDFS clusters. The system provides users with a single global namespace, and avoids dependence on external metadata servers to access the data stored on tape. In addition, tape layer resources are managed internally so that users do not have to deal with complex tape storage. The experimental results show that this method can effectively solve the massive data storage of HEP Hadoop cluster.

Experiment context, if any

References

Significance

Primary authors: LIU, Xiaoyu (IHEP); XIA, Libin (IHEP); Mr JIANG, Xiaowei (IHEP（中国科学院高能物理研究所）); SUN, Gongxing (IHEP)

Presenter: LIU, Xiaoyu (IHEP)

Session Classification: Poster session with coffee break