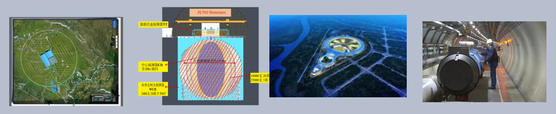
HDTFS： Cost-effective Hadoop Distributed & Tiered File System for High Energy Physics

LIU Xiaoyu (liuxiaoyu@ihep.ac.cn)
Computing Center, Institute of High Energy Physics

Introduction

Many types of high-energy physics experiments.



LHAASO JUNO HEPS LHC

Massive amounts of high-energy physics data.

Name of experiment	Data/Year	Estimated start time
LHAASO	10PB+	-
JUNO	~2PB	2022
HEPS	200/600PB	2025
HL-LHC	600PB	2026

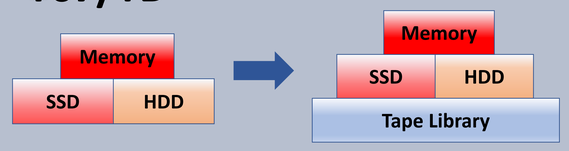
The development of big data technology in Internet area:

- A shift in the mindset of data processing computing-centric to data-centric
- Complete, mature software ecosystem
- Fast deployment, efficient and easy to use, high throughput, good scalability.

Objectives

HDD: ¥343/TB
Tape: ¥67/TB

Problems 1: Converged tape storage.

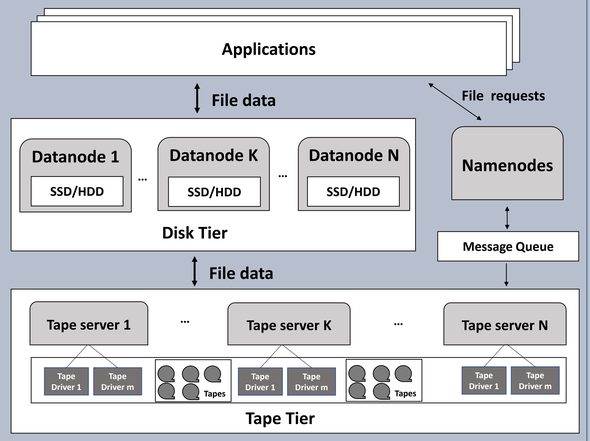


Problems 2: Unified file access interface.

- Java API
  - fs.copyFromLocalFile()
  - fs.copyToLocalFile()
  - …
- CLI
  - hdfs dfs -cmd file
  - hadoop fs -cmd file

Methods

A HDFS cluster typically consi-sts of a set of nodes and pro-vides file system namespace to applications by Namenode. To provide tape-tier storage, we add messaging middlewa-re between the upper HDFS system and the lower-level ta-pe storage management syst-em.



Applications can access files t-hrough a unified interface, and Namenodes is responsible for forwarding requests for tape-layer files.

- Archive
  Namenode sends the request to the request queue, tape li-brary management transfer fi-le data.
- Retrieve
  The processing is similar to archive, except that the requests sent are different.

Results

The disk storage capacity occ-upied by the test directory before and after the archive operation is shown in the figure.

```
[hadoop@helion01 ~]$ hdfs dfs -du -h /user/xy/test
5.0 G  14.9 G  /user/xy/test/data031322
3.0 G  8.9 G   /user/xy/test/data031422
980 M  2.9 G   /user/xy/test/data031522
4.0 G  12.0 G  /user/xy/test/data031622
2.0 G  6.0 G   /user/xy/test/data031722
[hadoop@helion01 ~]$ hdfs dfs -du -h -s /user/xy/test
14.9 G  44.6 G  /user/xy/test
```

capacity occupied before archive

```
[hadoop@helion01 ~]$ hdfs dfs -du -h /user/xy/test
0  0  /user/xy/test/data031322
0  0  /user/xy/test/data031422
0  0  /user/xy/test/data031522
0  0  /user/xy/test/data031622
0  0  /user/xy/test/data031722
[hadoop@helion01 ~]$ hdfs dfs -du -h -s /user/xy/test
0  0  /user/xy/test
```

capacity occupied after archive

Conclusions

HDTFS can utilize tape resou-rces to realize the massive data storage requirements of the high-energy physical Had-oop ecosystem.

References

1. Chen Gang. Data and computing for high energy physics experiments[J]. Scientia Sinica: Physica, Mechanica & Astronomica, 2021, 51(9): 14-23.
2. I. Koltsidas et al., Seamlessly integrating disk and tape in a multi-tiered distributed file system[C]// 2015 IEEE 31st International Conference on Data Engineering, 2015, pp. 1328-1339.
3. H. Dai, Y. Wang, K. B. Kent, et al. The State of the Art of Metadata Managements in Large-Scale Distributed File Systems — Scalability, Performance and Availability[C] IEEE Transactions on Parallel and Distributed Systems, 2022, 33(12): 3850-3869.
4. Qiulan Huang, Zhanchen Wei, Gongxing Sun, et al. Using Hadoop for High Energy Physics Data Analysis[C]// International Conference on Big Scientific Data Management(BigSDM 2018): Beijing, China. Springer Cham, 2018.
5. Jingwei Li, Suyu Huang, Yanjing Ren, et al. Enabling Secure and Space-Efficient Metadata Management in Encrypted Deduplication[J]. IEEE Transactions on Computers, 2022, 71(4): 959-970.
6. Apache Software Foundation. Enable support for heterogeneous storages in HDFS [EB/OL]. [2021-12-26]. https://issues.apache.org/jira/browse/HDFS-2832.
7. CERN. EOSCTA Docs[EB/OL]. [2021-12-26]. https://eoscta.docs.cern.ch.