Contribution ID: **2** Type: **not specified**

Research on EOS File storage Strategy Based on Access Characteristics Using Machine Learning Method

Machine learning has been an attractive topic in high-energy physics field for many years. For example, machine learning algorithms devoted to the reconstruction of particle tracks or jets in high energy physics experiments. EOS is an open source parallel distributed file system. It has been generally used in large scale cluster computing for both physics and user use cases at IHEP, like LHAASO and CEPC. The EOS design has included a split of hot and cold storage by defining groups. It records the frequency of file visits in the past period. Nevertheless, EOS cannot pick the most valuable data for users. Some files are frequently accessed since their creation and they'd better be placed at hot storage. Some files are accessed by a single user and used not so often. They'd better be placed at cold storage. Different files may have common access characteristics. For example, system log files are continuously written. Their pointers always seek for last byte of the files. We assume that other file clusters have such unique read or write characteristics too. So all files could be divided into several categories based on different access characteristics.We used some clustering algorithms, like Birch and Mini-batch K-means, to mine users'file access mode. We communicated with EOS users about file categories. After that, a supervised Machine Learning classification was built. Given a file, it searched this file's access log from EOS FST, compared its access features with different file categories, and predicted which category it belonged to. We wrote MapReduce tasks to process file access characteristics and saved them in HBase. In this paper we discussed two classify algorithms: RandomForest and LSTM. Then several data storage strategies were designed for each file category. The strategies included EOS storage group selection, the number of file copies and redundancy level. In addition, EOS's file scheduler would be redesigned as a plug-in to support file classification and category forecast. This paper will also present several test cases and LHAASO's sample files. More functional and performance tests are in progress.

Intended contribution length

20 minutes

Primary authors: CHENG, Zhenjing (INSTITUE OF HIGH ENERGY PHYSICS); Mr HU, Qingbao (IHEP); LI, Haibo (Institute of High Energy Physics Chinese Academy of Science); CHENG, Yaodong (IHEP); CHEN, Gang (INSTITUTE OF HIGH ENERGY PHYSICS)

Presenter: CHENG, Zhenjing (INSTITUE OF HIGH ENERGY PHYSICS)

Session Classification: Session 1