aHDFS: An Erasure-Coded Data Archival System for Hadoop Clusters

ABSTRACT

In this paper, we propose an erasure-coded data archival system called aHDFS for Hadoop clusters, where RS(k+r; k) codes are employed to archive data replicas in the Hadoop distributed file system or HDFS.

EXISTING SYSTEM

Disk storage (also sometimes called the drive storage) is a general category of storage mechanisms where data are recorded by various electronic, magnetic, optical, or mechanical changes to a surface layer of one or more rotating disks. Many disk-based systems are ill-suited for long-term storage because their high energy demands and management requirements make them cost-ineffective for archival purposes. Existing disk-based archival storage systems are inadequate for Hadoop clusters due to the ignorance of data replicas and the map-reduce programming model

DRAWBACKS

- These are inadequate for Hadoop clusters due to the ignorance of data replicas and the map-reduce programming model.

PROPOSED SYSTEM

In this proposed system, we propose an erasure-coded data archival system called aHDFS for Hadoop clusters, where RS(k+r; k) codes are employed to archive data replicas in the Hadoop distributed file system or HDFS. We develop two archival strategies (i.e., aHDFS-Grouping and aHDFS-Pipeline) in aHDFS to speed up the data archival process. aHDFS-Grouping – a MapReduce-based data archiving scheme - keeps each mapper's intermediate output Key-Value pairs in a local key-value store. With the local store in place, aHDFS-Grouping merges all the intermediate key-value pairs with the same key into one single key-value pair, followed by shuffling the single Key-Value pair to reducers to generate final parity blocks. aHDFS-Pipeline forms a data archival pipeline using multiple data node in a Hadoop cluster. aHDFS-Pipeline

delivers the merged single key-value pair to a subsequent node's local key-value store. Last node in the pipeline is responsible for outputting parity blocks. We implement aHDFS in a real-world Hadoop cluster. The experimental results show that aHDFS-Grouping and aHDFS Pipeline speed up Baseline's shuffle and reduce phases by a factor of 10 and 5, respectively. When block size is larger than 32MB, aHDFS improves the performance of HDFS-RAID and HDFS-EC by approximately 31.8% and 15.7%, respectively.

ADVANTAGES

- It provides minimum storage cost
- It speeds up data archival performance in Hadoop distributed file system (HDFS) on Hadoop clusters.

SYSTEM REQUIREMENTS

H/W System Configuration:-

- Processor - Pentium –IV
- RAM - 4 GB (min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

S/W System Configuration:-

- Operating System : Linux
- Application Server : Tomcat5.0/6.X
- Backend coding : Java
- Tool : Virtual Box
- Environment : Ubuntu
- Technology : Hadoop