Hadoop 3.0纠删码(Erasure Coding)：节省一半存储空间

随着大数据技术的发展，HDFS作为Hadoop的核心模块之一得到了广泛的应用。为了系统的可靠性，HDFS通过复制来实现这种机制。但在HDFS中每一份数据都有两个副本，这也使得存储利用率仅为1/3，每TB数据都需要占用3TB的存储空间。随着数据量的增长，复制的代价也变得越来越明显：传统的3份复制相当于增加了200%的存储开销，给存储空间和网络带宽带来了很大的压力。因此，在保证可靠性的前提下如何提高存储利用率已成为当前HDFS应用的主要问题之一。

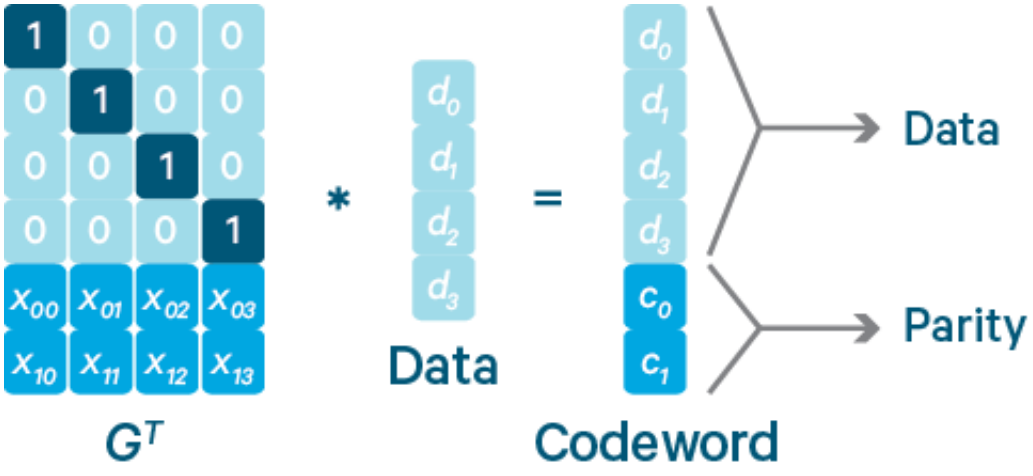
针对这些问题，英特尔、Cloudera、华为以及其他的Apache Hadoop communit共同参与开始引入纠删码（Erasure Coding，EC）技术，在保证数据可靠性的同时大幅降低存储开销。相关代码已经进入trunk，并计划3.0版本中发布。

Erasure coding纠删码技术简称EC，是一种数据保护技术。最早用于通信行业中数据传输中的数据恢复，是一种编码容错技术。他通过在原始数据中加入新的校验数据，使得各个部分的数据产生关联性。在一定范围的数据出错情况下，通过纠删码技术都可以进行恢复。

纠删码（Erasure Code）与 Reed Solomon码

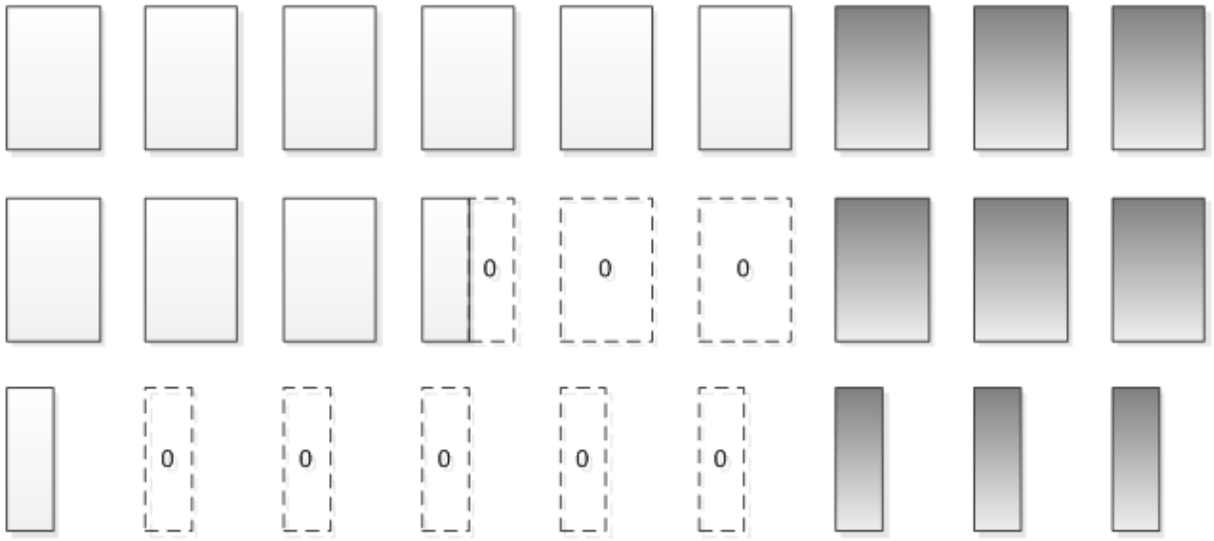
在存储系统中，纠删码技术主要是通过利用纠删码算法将原始的数据进行编码得到校验，并将数据和校验一并存储起来，以达到容错的目的。其基本思想是将 k 块原始的数据元素通过一定的编码计算，得到m块校验元素。对于这 k +m块元素，当其中任意的m块元素出错（包括数据和校验出错），均可以通过对应的重构算法恢复出原来的 k 块数据。生成校验的过程被成为编码（encoding），恢复丢失数据块的过程被称为解码（decoding）。

Reed-Solomon（RS）码是存储系统较为常用的一种纠删码，它有两个参数k和m，记为RS(k，m)。如图1所示，k个数据块组成一个向量被乘上一个生成矩阵（Generator Matrix）GT从而得到一个码字（codeword）向量，该向量由k个数据块和m个校验块构成。如果一个数据块丢失，可以用(GT)-1乘以码字向量来恢复出丢失的数据块。RS(k，m)最多可容忍m个块（包括数据块和校验块）丢失。



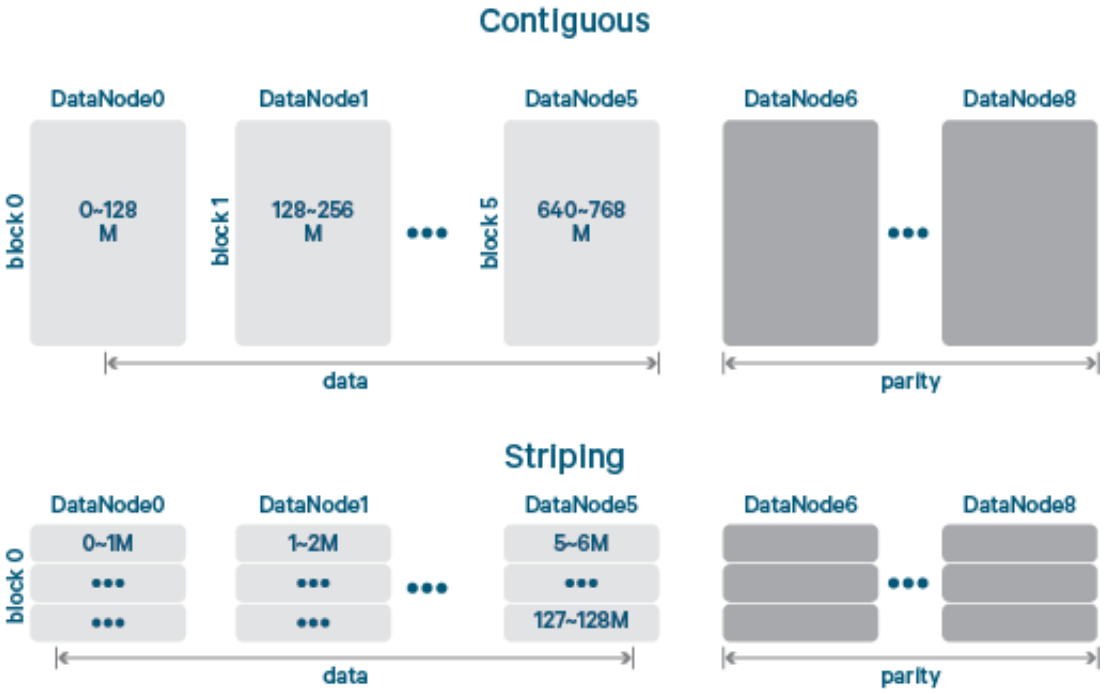
块组（BlockGroup）

对HDFS的一个普通文件来说，构成它的基本单位是块。对于EC模式下的文件，构成它的基本单位为块组。块组由一定数目的数据块加上生成的校验块放一起构成。以RS(6，3)为例，每一个块组包含1-6个数据块，以及3个校验块。进行EC编码的前提是每个块的长度一致。如果不一致，则应填充0。图2给出三种不同类型的块组及其编码。



连续布局（Contiguous Layout）VS 条形布局（Striping Layout）

数据被依次写入一个块中，一个块写满之后再写入下一个块，数据的这种分布方式被称为连续布局。在一些分布式文件系统如QFS和Ceph中，广泛使用另外一种布局：条形布局。条（stripe）是由若干个相同大小单元（cell）构成的序列。在条形布局下，数据被依次写入条的各个单元中，当条被写满之后就写入下一个条，一个条的不同单元位于不同的数据块中。



项目计划

由于HDFS的内部逻辑已经相当复杂，所以整个HDFS EC项目的实现主要分为三个阶段：

1、用户可以读和写一个条形布局（Striping Layout）的文件；如果该文件的一个块丢失，后台能够检查出并恢复；如果在读的过程中发现数据丢失，能够立即解码出丢失的数据从而不影响读操作。
2、支持将一个多备份模式（HDFS原有模式）的文件转换成连续布局（Contiguous Layout，定义见下文），以及从连续布局转换成多备份模式。
3、编解码器将作为插件，用户可指定文件所使用的编解码器。

第一阶段（HDFS-7285）已经实现第1个功能，第二阶段（HDFS-8030）正在进行中，将实现第2和第3个功能。

Erasure Coding技术的优劣势

优势

纠删码技术作为一门数据保护技术，自然有许多的优势，首先可以解决的就是目前分布式系统，云计算中采用副本来防止数据的丢失。副本机制确实可以解决数据丢失的问题，但是翻倍的数据存储空间也必然要被消耗，这一点却是非常致命的。EC技术的运用就可以直接解决这个问题。

劣势

EC技术的优势确实明显，但是他的使用也是需要一些代价的，一旦数据需要恢复，他会造成2大资源的消耗:

1、网络带宽的消耗，因为数据恢复需要去读其他的数据块和校验块
2、进行编码，解码计算需要消耗CPU资源

概况来讲一句话，就是既耗网络又耗CPU，看来代价也不小。所以这么来看，将此计数用于线上服务可能会觉得不够稳定，所以最好的选择是用于冷数据集群，有下面2点原因可以支持这种选择

1、冷数据集群往往有大量的长期没有被访问的数据，体量确实很大，采用EC技术，可以大大减少副本数
2、冷数据集群基本稳定，耗资源量少，所以一旦进行数据恢复，将不会对集群造成大的影响

出于上述2种原因，冷数据集群无非是一个很好的选择。