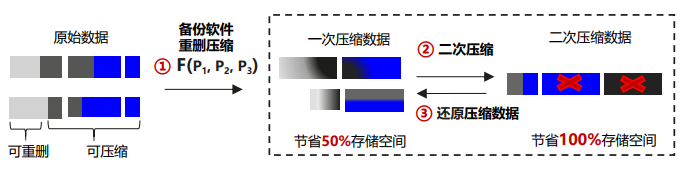
针对已压缩数据的二次压缩算法

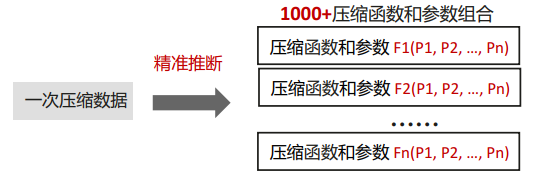
**技术背景**

存量数据中存在很多已压缩数据，通过对这些数据进行重新编码，取得更好的压缩效果，节省存储空间。

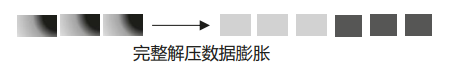


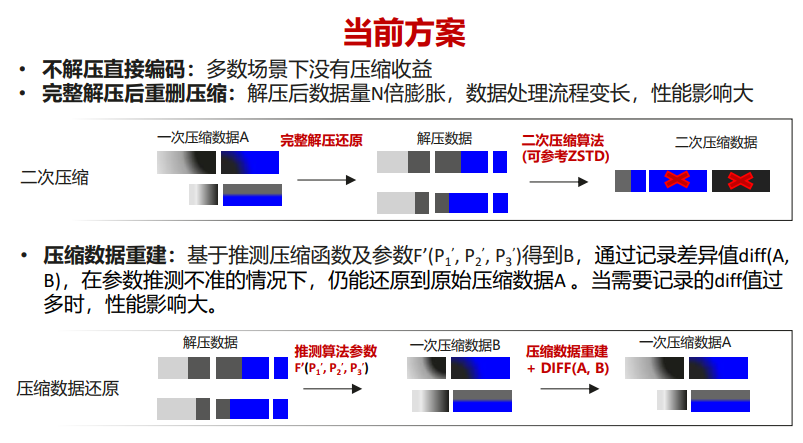
**技术挑战**

• **压缩函数及参数组合多：**压缩算法可调参数类型和个数较多，大量场景下分布和结构特征不明显，难以精确估计。



• **二次压缩难度大：**压缩编码后数据规律丢失，难以直接压缩；完整解压数据膨胀影响性能。





**技术诉求**

基于ARM平台进行原型验证，备份软件已压缩数据的压缩比相比基线算法提升50%，二次压缩性能400MB/s/core。

**参考文献：**

[1] CompressDB: Enabling Efficient Compressed Data Direct Processing for Various Databases, SIGMOD 2022

[2] The Design, Implementation, and Deployment of a System to Transparently Compress Hundreds of Petabytes of Image Files for a File-Storage Service, NSDI 2017