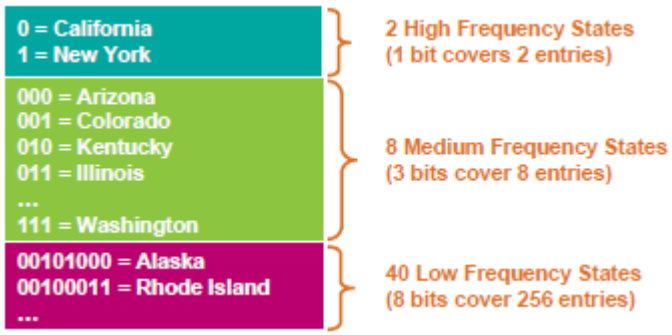
DB2 V10.5 BLU学习系列-03.BLU的压缩

作者：bigdata_lyn@126.com
主页:bigdatalyn.github.io
时间：2016/03

BLU使用多种压缩技术

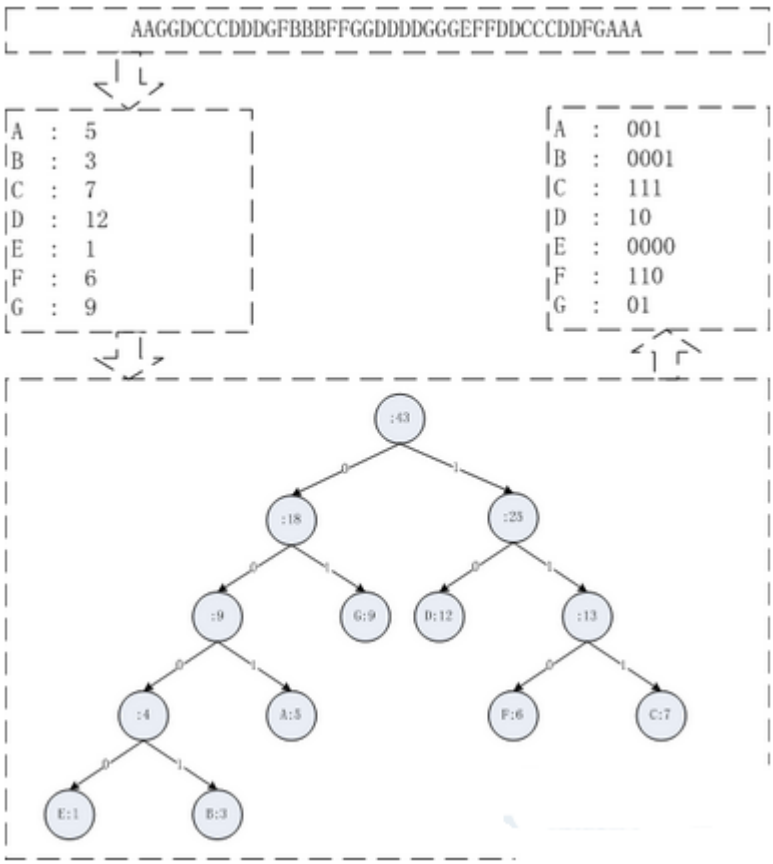
可操作的压缩：支持以压缩格式分析数据

对于出现频率高的值压缩后会用比较少的bit来表示，而且在一个区块内是保序的。结合页级别的压缩，有可能根据列级别的压缩成为3 bits的，在当前页里面出现频率比较高有可能就会压缩成2 bits. 这种压缩算法大大提高了压缩效率，减少了存储和I/O开销。数据页里保存了当前页的压缩字典。



1.接近哈夫曼编码

哈夫曼编码:



2.Offset coding 抵消编码-差值压缩

例如：1002, 1003, 1019, 1045, 1005
数据存储： 1002, 3, 19, 45 5

3.Prefix Compression-前缀压缩

BLU压缩字典里是排序的

例如：
US State North Carolina is stored as 10
South Carolina is stored as 20
20 > 10: Hence South Carolina > North Carolina for dictionary match

DB2在BLU的压缩中不仅有列级别的压缩，而且在页级别的字典会再进一步压缩。

列级别字典：

页级别字典：

当数据被Load的时候，当你使用LOAD REPLACE, LOAD REPLACE RESETDICTIONARY, LOAD REPLACE RESETDICTIONARYONLY或者LOAD INSERT的时候，列级别的压缩字典就会被创建。当新的数据被加入的时候，数据会用表级别的压缩字典进行压缩。BLU还会用页级别的压缩字典来压缩新的时候。这种两层的数据压缩字典，提高了数据的压缩效率。页级别的压缩字典就存在于每个数据页中。每个数据页包含了该页的起始TSN以及总共有多少个TSN。

压缩比较：（大概有十倍左右）