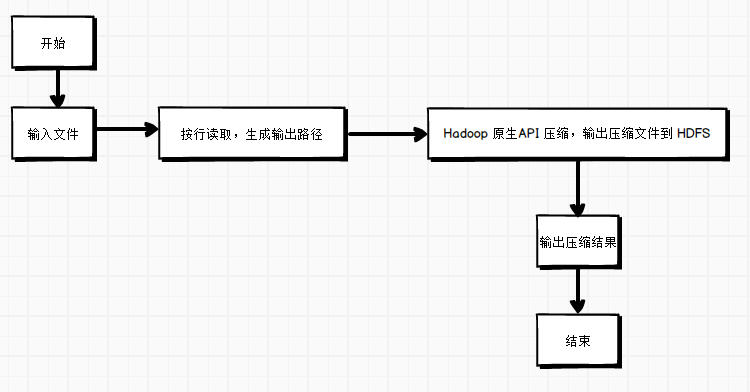
# Hadoop压缩程序设计文档

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **修订版本** | **描述** | **作者** |
| 2019-11-15 | v1.0 | 初始版本 | 范鹏义 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## 概述

Hadoop压缩程序是将线上已经存储好的历史数据按指定压缩格式压缩到指定目录，目的是为了节省磁盘空间，将冷数据压缩归档后，还需要对应的解压缩程序，将压缩好的数据解压缩后应与压缩前数据格式、数据条数一致。

## 流程图



## 压缩数据逻辑描述

### **3.1 文件输入路径获取**

使用 Hadoop FileSystem 获取当前目录下所有路径，循环判断如果是文件则打印，如果是目录则递归调用方法。

对应文件程序为：com.people.util.getFileList ,传入 HDFS 上路径即可获得对应目录下的文件路径，包括子目录的。

### **3.2 输入文件**：

将文件路径输出的文件路径放在当前目录下的 compress.txt 中，一行一个文件路径，作为程序的输入。

### **3.3 Hadoop API 压缩**

根据输入路径获取文件系统 FileSystem ,将原文件读入为输入流，将压缩路径的输入流作为输出，使用 IOUtils.copyBytes(in, compressOut, 4096,true);

### **3.4 压缩文件输出结果**

如果文件压缩成功，则输出对应的结果 ，

输入路径：

hdfs://nserver:9000/data\_center/crawl/new\_article/year=2018/month=05/2018-05-29-18.txt compress ---->

输出路径：hdfs://nserver:9000/data\_center/crawl/new\_article/mycompress/year=2018/month=05/2018-05-29-18.bz2 ,result ----> success!