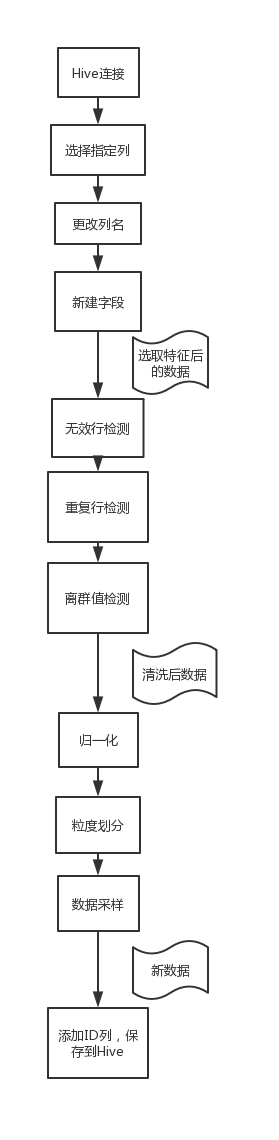
**demo3，L1数据点-特征提取，数据清洗和样本构造**

工业数据L1数据点14a11034616000\_\_\_\_\_\_\_\_00001\_data包含了401列工业处理特征，共2558行数据。本例子拟利用pyspark组件选取其中5列特征值，构造新字段，进行简单的清洗，在归一化后每3条记录取均值改变粒度，最后对数据进行了随机采样。具体过程如下流程图所示：

该流程分为特征选取、数据清洗、样本构造三部分，包含了11次pyspark组件调用，详见代码demo3.py。整个任务耗时约10分钟。

C:\Users\Jerry\AppData\Roaming\Tencent\Users\2250486211\QQ\WinTemp\RichOle\)P{Y[ODWMJ39S14_9GPSP%Y.png提交时的文件结构如下，

​

在该文件结构下运行如下命令，即可正确提交demo1.py

spark-submit --jars /usr/hdp/current/hive\_warehouse\_connector/hive-warehoue-connector-assembly-1.0.0.3.1.0.0-78.jar --py-files DataTool.zip demo3.py

问题：

1，缺失值处理时，当表中某一字段下没有缺失值或者全部为缺失值时，程序抛出错误，无法继续，欠妥。

2, dataTransfer组件中，granularityPartition 和 multiFieldPartition 函数的aggmode参数无法实现std计算，具体改进方法还在研究。

注意：

数据处理的所有功能仅适用于单元格中int，float或者string格式的数据，不适用于dict，set，list等集合型数据

一个bug：

pyspark在输入NaN格式时，是以float('NaN')格式输入，但在对float('NaN')进行值替换操作时，表格中的0也会被替换。但又发现float('NaN')==0结果为False。因此目前能识别NaN格式的操作只有isnan函数，其他方法还不清楚。