新增数据处理流程

新增的数据主要包括论文专利和项目的数据，由于新增的爬虫代码由不同的人员编写，所以预处理时字段的对应会有所不同，但是处理的流程大致相同

一、预处理

首先是对爬取的数据做预处理，这一步工作是在进行新增数据处理的时候都要做的，因为这一步是对爬取的数据进行各个字段的处理和调整(调整字段的位置维持和数据库各字段位置一致)，为了防止后续从不同地方，不同人编写代码导致的预处理过程不一致，所以应该采用统一接口来实现。

输入数据 爬虫的内容

输出数据 数据库成果相应的字段 + 成果类型代码(subject\_code)

成果单位(unit) 单位类型(unit\_code) 省份代码(proveince\_code)

单位代码(unit\_code) 单位等级代码(rakn\_code)

城市代码(city,目前还没用到过用null填充)

一作到四作单位等级代码(有四个)

输出的数据各个字段如果是空的按照null处理

预处理调整字段的位置根据爬取的内容决定

此外，还需要对以下几个字段进行处理

成果都需要处理的字段:

单位(论文unit 专利applicant项目unit )：无论是论文表，专家表还是项目表都需要处理，涉及的单位只要包括成果的单位和各个作者的单位。主要是调用com.unitProcess类的dispose方法进行处理，考虑到新增数据量不大，可以是使用上述的方法，当新增的数量太大导致处理时间超出了预算，这就需要使用到hdfs中的unitMap这个处理前和处理后的单位对照表来关联单位，处理大部分单位，小部分单位表中没有对照的单位再调用unitProcess类来进行处理。成果单位处理完之后，原单位也是需要保存下来，保存在\_source字段。对于每个单位，处理完都会产生处理后的单位名字，单位类型，省份代码，单位等级和城市代码。

单位类型(unit\_code): 这个字段和单位类型(unit\_code)相同

成果类型(subject):成果类型按照各个成果之间的划分不同不同处理。

成果ID:每个成果都需要产生UUID

需论文表处理的字段:

期刊质量(journal\_quality):根据期刊名字和journal\_information进行关联，journal\_information是张小表，直接进内存关联。

注：论文名以“通知”，“召开”结尾或者论文名包括“投稿”和“作者指南”的都舍弃

需要专利处理的字段：

需项目表处理的字段：

项目类型(type):根据爬去的内容得到（有时候需要处理，根据爬的内容而定）

项目类型二级分类(project\_type\_more):根据项目类型(type)和project\_type\_map关联得到，同样是张小表，直接关联得到

项目类型一级分类(project\_type):取项目类型二级分类字段的第一个字符得到

二、补上其余作者的单位

作者若没有单位，则需要和专家表根据专家名字和成果单位进行关联，补上在专家表存在记录的相应的单位。（四个作者，需要进行四次，有待改进）

输入 第一步输出文件

输出 替换相应作者下的单位，字段位置不改变

三、抽取专家

抽取所有带单位的作者，在抽取过程中如果发现一作没有单位，用成果的单位替代。只抽取单位是标准库中的单位、研究所、公司的专家(rank\_code = 01\*|022|032)

注：在这一步要判断专家的名字，包含[“课题组”,” 编辑部”,” 项目组”,” 不公开姓名”,” 发明人”]的都删除(也可以在下一步完成)

注：之后涉及到抽取专家单位的差不多都是这种操作，单位类型也要变成换之后的

输入 第二步的输出表

输出 作者名字 作者单位 成果ID 成果名字 第几作者 作者来源

作者来源:论文的第一作者 4

论文的其它作者 41

专利的第一作者 3

专利的其它作者 31

项目的第一作者 2

项目的其它作者 21

四、补上新抽取的专家UUID

给抽出专家补上uuid， 和专家表根据作者的名字和单位做关联，如果该专家在专家表中，补上专家表中该专家的UUID，如果该专家不存在专家表中，产生新的UUID补上。抽取专家的时候需要和高校更名表(university\_change\_name，可以直接加载到内存)关联，如果存在更名，专家单位变更成更名的专家

输入 第三步输入的文件

输出 专家ID 作者名字 作者单位 成果ID 成果名字 第几作者 作者来源

五、为第二步的输出表关联作者的ID

将第四步的输出和第二步的输出文件根据成果ID做关联，在第二步的输出文件末尾补上相关联的专家ID

六、抽取出增加的需要入库的成果表

将第五步的输出文件抽取出数据库成果表需要的那几个字段

七、提取增加的成果和专家关联记录

所需的字段都在第四步输出的文件中，从中提取即可

八、提取出增加的专家表

将第四步的输出和专家表根据专家ID关联，提取出在第四步的输出中没有出现在专家表的新专家

注：六，七，八步骤字段的长度有限制，不能超出数据库字段的限制长度