基于 FME 的数据检查和大数据量分析方法

李晓铃

（成都市国土资源信息中心，四川成都，610042）

摘 要：FME 具有数据检查、叠加分析、属性处理、坐标转换、批处理等强大的数据流程化分析处理能力。本文以 WP 文件和 Oracle 空间数据库为数据源，探讨如何利用 FME 进行成都市耕保基金图形数据检查和千万级大数据量分析。实践表明，采用 FME 的流程化处理可以保障耕保数据的有效性、完整性，减少重复操作，大大提高工作效率，实现耕保基金年度变更图形数据处理的智能化、流程化。

0 引言

空间数据库是 GIS 系统开发与应用的基础，空间数据的质量将直接影响 GIS 系统应用、分析、决策的正确性和可靠性，因此有必要在空间数据入库前进行质量检查。同时随着信息化程度的不断深入，成都市国土资源数据已达千万级甚至以亿计，如何快速有效地进行大数据量分析，为领导决策提供数据支撑是至关重要的。

本文以成都市耕保基金图形自检和年度变更分析为例，主要探讨了基于 FME 的空间数据检查与大数据量分析的方法，该方法已运用到实际工作中，取得了一定的成果和经验，在相关行业领域具有推广应用价值。

1 FME 与应用模式

FME（Feature Manipulate Engine，即要素操纵引擎），是加拿大 Safe Software 公司开发的空间数据转换处理系统。可自定义图形化界面，实现超过 270 多种 GIS 及 CAD 空间数据格式的相互转换，具有丰富的 GIS 数据处理功能（如坐标转换、叠加分析、相互运算、构造闭合多边形、属性合并等），通过编写脚本及批处理模式支持海量数据高效处理。

在成都市耕保基金管理工作中，FME 的主要作用在于对耕保基金空间图形的处理：1）前期耕保地块图形数据库的建立，实现了图形自检、复检、入库和删除等；2）年度变更，实现了耕保图形与土地执法、现状数据叠加分析，区县耕保图形数据审核与入库。

由于业务需求调整，对技术流程不断优化和完善，针对不同级别工作人员逐渐形成了比较成熟的 FME 应用模式（图 1），实现了耕保基金年度变更图形数据处理的智能化和流程化。1）市局工作人员，基于 FME desktop 进行全市耕保变更分析，制作自检和入库 fmw 模板通过 FME Server 发布为 Web 服务；2）区（市）县工作人员，根据变更分析结果进行核查，形成以组为单位的耕保图形 WP 文件。采用市局下发的客户端程序进行单个文件或文件夹批量自检：通过自检的组，耕保图形自动入库；未通过，返回错误信息，修改后再自检。

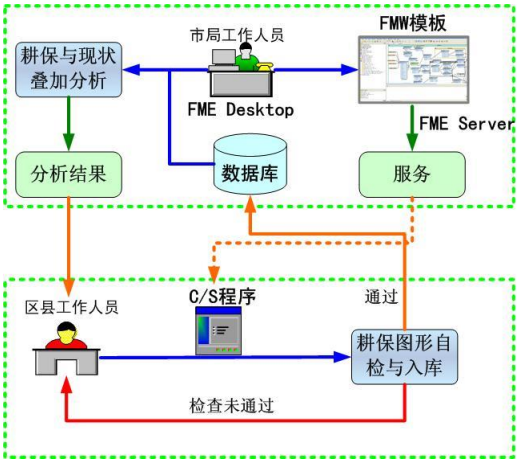

图 1 技术流程

2 GIS 空间数据检查

GIS 系统设计时需制定数据库标准，包括要素分类标准、要素类属性表结构、坐标系统、数据格式等，以此作为数据入库前质量检查项的参照依据。耕保图形自检作为数据入库的质量控制，主要从属性、空间两方面进行数据检查，结果以.TXT 文本方式输出。

2.1 基本功能实现

2.1.1 属性检查

（1）数据结构完整性

检查源数据中是否存在必要字段。例如必须存在 NYDSYQDJH 和 GBDKBH 字段。

通常添加读模块 Schema (Any Format)读取任意格式文件的数据结构，与标准表结构文件对比检查出缺少的字段。对于必要字段较少的情况，可采用 AttributeExploder+ DuplicateRemover 获取源文件的所有字段名称，StringSearcher 正则表达式（如^gbdkbh$）查找是否存在必要字段。

（2）属性的有效性

检查属性是否为空、唯一、在规定的值域内。例如 NYDSYQDJH 必须为 26 位数字，GBDKBH 不允许为空、有空格、存在重复。

Tester 可判断属性是否为空，DuplicateRemover 检查属性是否重复，属性值是否符合规范则可灵活运用 StringSearcher 中的正则表达式，如^[0-9]{26}$可检查属性是否为 26 位数字，.* .*可检查空格的存在。

（3）属性逻辑一致性

检查源数据属性与扩展表的相关属性逻辑关系是否一致。例如耕保地块图形的地块数量、地块编号与耕保基金管理系统基础数据中的地块数量、地块编号必须一致。

与其他数据库表属性进行逻辑判断时，首先通过 OracleQuerier 查询相关属性表，FeatureMerger 将源数据与查询结果进行比对，其 INCOMPLETE 和 UNREFERENCED 端口输出要素为不一致的数据。若查询一个属性在其他表中是否存在，通过 SQLExecutor 对外部数据库表执行 SQL 查询，ListElementCounter 对列表元素统计，等于 0 表示不存在；或采用 ListRangeExtractor 获取列表属性的最大值和最小值来判断。

2.1.2 空间检查

（1）几何类型检查

判断耕保图形是否为面状。利用 AttributeExposer 暴露 fme_geometry 格式属性，等于 fme_polygon 即为面状；或通过 GeometryFilter 过滤出不为 AREA（面状）的图形。

（2）图形有效性

对于面要素，检查同一组的耕保地块是否存在自相交、自压盖等。通过 GeometryOGCValidator 判断图形有效性可检查存在自相交、嵌套的洞或重复环等无效的面要素；利用 AreaOnAreaOverlayer 可检查其自压盖的情况。

（3）空间关系

检查图形是否在规定的范围内、与相关要素图形是否存在压盖等。例如耕保地块必须在成都市行政区界线内。耕保地块分别与土地利用现状的非耕（园）地和国有土地、土地执法数据进行叠加分析，压盖面积大于等于阈值认为压盖。

主要利用 FME 的空间叠加分析功能。为减少源文件的加载，待分析的数据都存储在 Oracle 数据库中，可通过 OracleQuerier 空间查询待叠加图层，SpatialRelator 执行空间关系判断图形是否在规定的范围内，Clipper 或 AreaOnAreaOverlayer 进行压盖分析。

2.2 关键处理

（1）功能模块之间的衔接

在一个模板中需要较多转换器实现上述功能（见图 2），可用书签（BookMark）将每个

检查功能模块化，并添加必要的注释或创建自定义函数增加模板的易读性。首先对各检查项进行合理排序，先属性后空间，先耕保图形本身后与其他图层，然后将各功能模块衔接起来，注意模块之间的 3 种关系：

1）条件关系，如必须满足数据结构完整，才能进行属性有效性检查。采用 FeatureMerger（属性合并）选择关联字段，属性一致（COMPLETE）时进入下一个流程。为防止过程中无要素流转，可采用 Creator 创建一个要素和 AttributeCreator 创建一个关联字段。

2）承接关系，如属性有效性正确的要素进行属性逻辑一致性检查，错误的要素信息写入结果文本中。假定 DuplicateRemover 作为属性有效性检查的最后一个转换器，那么 UNIQUE 端口输出要素即为下一步属性逻辑性检查模块的输入。

3）并列关系，如耕保图形在成都市行政区内和耕保图形与土地利用现状压盖分析，两个检查模块互不干扰，同时进行。只需将上一个检查模块的有效输出要素分别作为此两项模块的输入即可。

（2）不同坐标系数据转换

由于各区县作业单位不同，提交的耕保图形 WP 文件坐标系可能存在无带号或有带号的西安 80-34 度带、35 度带的情况，而 Oracle 中存储的土地利用现状等空间数据皆为有带号的西安 80-18 度带。要进行空间分析，需通过 CoordinateExtractor+Tester 根据 X 坐标的值域判断其坐标系，然后利用 Reprojector 将耕保图形投影转换为西安 80-18 度带。

（3）空间查询

对于检查中涉及的数据库表，若添加读模块直接加载，会大大增加数据量。因此使用 OracleQuerier 或 SQLExecutor 转换器，以源数据的属性或空间范围动态查询相关的那部分数据，减少分析的数据量。为避免查询结果重复，执行属性查询涉及源数据的相关属性值必须唯一，源数据不是单一几何对象时，尽量以所有对象的一个边界框作为空间查询的范围。

（4）结果冗余信息处理

具有共性的错误要素全部输出会造成信息冗余。例如检查结果信息“WP 文件中耕保地块编号为@Value(GBDKBH)的图形不在成都市范围内”，常规方法是有多个错误要素，便输出多个相似信息。要消除冗余信息，可以通过 ListBuilder 创建列表，ListConcatenator 将列表属性 GBDKBH 连接为一个属性，StringConcatenator 文本编辑即可仅输出一条信息。

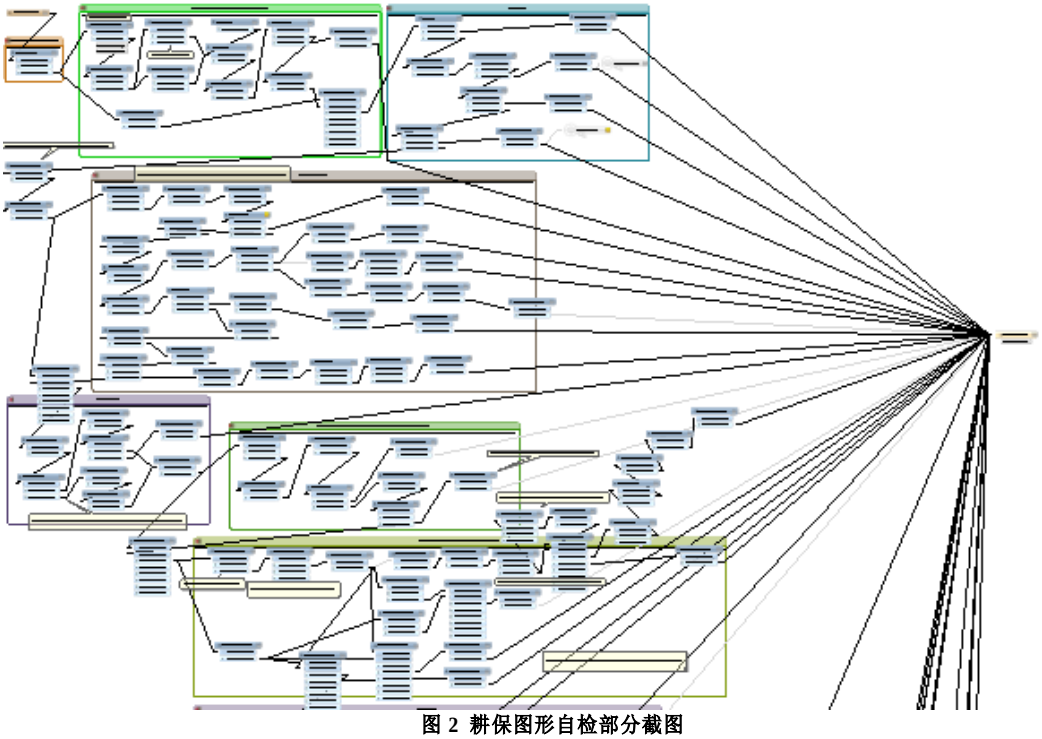

图 2 耕保图形自检部分截图

3 大数据量分析

耕保变更分析是将上一年度的耕保基金图形数据与当年度的土地执法数据以及土地利用现状数据进行叠加分析，其涉及的数据总量达千万条，本文将从数据库层面以及处理工具和方式探讨如何进行大数据量分析。

3.1 数据库层面

（1）大数据量对所使用的数据库工具要求较高，分析的源数据均存储在 Oracle 中，便于集中管理与处理。

（2）对大数据量进行分区操作十分必要，国土资源年度变更数据如耕保基金图形和土地利用现状，均按年进行分区。

（3）在对大数据量进行查询处理过程中，对大表的关键字段建立必要的索引，可大大提高查询的效率。针对两类大数据量的耕保基金图形和土地利用现状数据，为查询中涉及的关键字段 GBDKBH 和 ND、DLBM 和 QSXZ 分别建立索引。

3.2 处理工具与方式

（1）对大数据量分批处理。通过 FME 制作小数据量的分析模板（子模板），然后在批处理模板中采用 WorkspaceRunner 调用子模板，实现分批处理。

以 2011 年耕保图形为例，其数据总量为 10062034 条，根据表 1 行政区个数和最大记录数，选择以村为单位分批进行叠加分析较为合适。

表 1 耕保基金图形数据量情况

县级		乡级		村级		组级	
行政区个数	最大记录数	行政区个数	最大记录数	行政区个数	最大记录数	行政区个数	最大记录数
20	1886741	251	134893	2672	28319	32479	4461

村级分析子模板关键处理流程为（图 3）：1）创建外部参数 Input_XZQDM，作为动态读取 Oracle Spatial Object 耕保基金图形数据的关联属性，OracleQuerie 空间查询出单个村级耕保地块，并获取其外接矩形；2）以外接矩形 OracleQuerie 空间查询出该范围内的土地执法与土地现状数据；3）将耕保地块分别与执法和现状数据进行 Clipper 叠加分析。

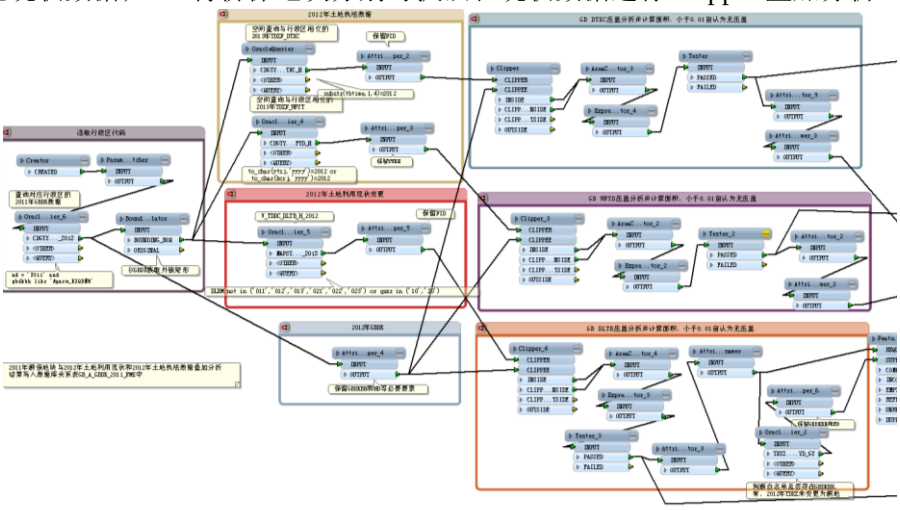

图 3 关键处理流程

以成都市耕保地块的所有村级代码为输入，使用 WorkspaceRunner 调用村级分析子模板实现变更分析的批处理（如图 4）。

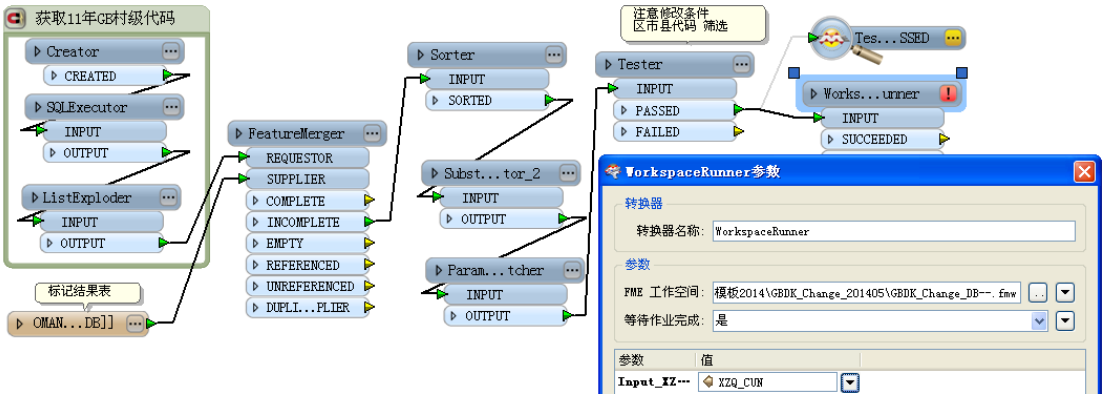

图 4 批处理流程

（2）定制出错处理机制，便于问题的查找与修复。由于网络中断、数据错误或服务等问题可能造成批处理的子程序意外中断或退出，因此需要建立容错机制：1）在村级耕保数据叠加分析子模板中增加运行成功的标记，并将 Input_XZQDM 属性写入标记结果表中；2）在批处理模板中，增加与标记结果表的比对（FeatureMerger），直至 INCOMPLETE 输出端口无要素分析结束，若多次运行某村级子程序仍失败，需在子模板中进行单个测试查找问题。

4 结论

FME 在耕保基金数据年度变更工作的应用实践表明：千万级大数据量的综合分析应采用合适的分级分批处理，批处理采用 WorkspaceRunner 调用子模板的方式更为灵活；需建立容错机制便于网络、服务及数据等不确定因素造成进程失败后的错误定位与处理；书签模块化功能可以化繁为简，但需注意各模块之间的衔接；针对不同级别的工作人员采用不同的应用模式。