# ETL设计文档

## 概述

ETL系统的核心功能就是按照本设计说明书的架构，将数据由数据源系统加载到数据仓库中。其实现的困难在于ETL系统将面临复杂的源数据环境，包括多种多样的数据源平台、繁多的数据种类、巨大的加载数据量、错综复杂的数据关系和参差不齐的数据质量，这些都使ETL的架构和应用设计面临相当的挑战。

通过高效的ETL系统结构、层次化的应用功能划分和标准的程序模板，ETL系统和应用架构设计需要能够达到以下目标：

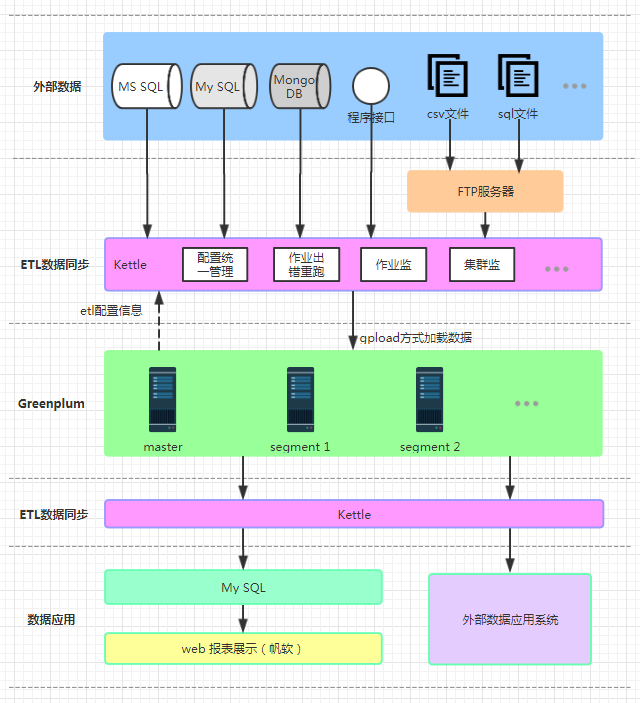
1. 支持各种数据源数据加载到数据仓库的功能；
2. 支持在规定的批处理时间窗口内能够完成数据加载工作，即需要满足日常数据加载的性能需求；
3. 支持作业依赖关系配置，保障作业的执行先后顺序；
4. 支持作业出错自动重试，降低由于短暂网络中断或者源系统升级等不稳定因素的影响，保障系统的健壮性；
5. 减少系统维护的复杂性，支持后续增加新数据或功能的开发工作。

本设计说明书包含以下部分：

1. ETL系统架构设计；
2. ETL应用架构设计；
3. ETL作业配置；
4. ETL数据处理方式；
5. ETL监控与错误作业自动重试；
6. 错误作业日常维护。

## ETL系统架构设计

系统架构设计包括ETL需要关联到的相关系统和系统模块，包括多样的数据源（包括各种数据库、文件方式和程序接口方式等）、ETL相关配置信息、数据存储系统和数据应用系统等。



## ETL应用架构设计

ETL系统部分包括以下三部分：

1. 定时任务启动

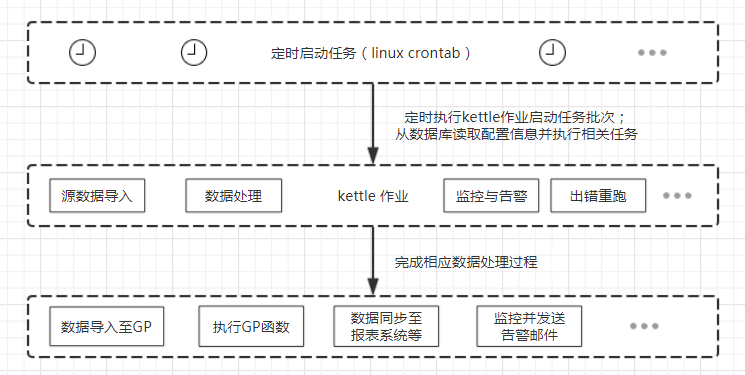
任务调度部分中，使用linux系统的crontab设置定时任务启动任务批次，启动相应的批次之后会按优先级执行前置作业已经完成的作业。

1. 任务配置

在执行作业时用到的相关信息配置在GP库的相应表中，比如数据库连接信息、源表信息、目标表信息、执行作业类型和告警邮件接收人信息等。

1. ETL功能模块

按照作业类型分多功能模块，如：数据导入至GP、执行GP函数完成相关数据汇总和数据处理、从GP将数据同步至其它系统（报表系统等）和监控告警类作业。



作业流程图如下：



## ETL作业配置

所有etl作业通过读取配置表相关信息来确定需要执行哪个kettle子作业、源数据和目标表等信息，这些配置信息表存放在GP的etl模式下面。相关配置信息表说明如下：

etl.etl\_conf\_batch：保存批次的信息，每个批次会开启一个jvm来执行该批次下面的所有作业，目前基本按照每个数据源每一层处理一个单独的批次。

etl.etl\_conf\_job：保存作业信息，作业执行的kettle子作业信息，源数据信息和目标表信息，还有作业运行状态和时间，并通过batch\_id来确定在哪个批次下面执行该作业。

etl.etl\_conf\_job\_table\_field：保存作业处理的表对应字段信息（目前主要针对文件到ods层）

etl.etl\_conf\_kettle\_job：保存kettle资源库中子作业对应的目录和作业名。

etl.etl\_conf\_precondition：保存作业之间的依赖关系，作业需要等前置作业跑完才能执行

etl.etl\_receive\_mail\_address：保存需要接收告警邮件的email地址

etl.etl\_send\_mail\_account：用来设置发送邮件的邮箱的地址、密码、SMTP信息，其中密码需要通过kettle资源库下面的/public/tr\_encode\_password先加密，将密文存放到密码字段中。

## ETL数据处理方式

## ETL监控与错误作业自动重试

## 错误作业手动重跑与日常维护