后端系统作业——综述类

ZY2006109-姬轶

对比分析国内（如 863，核高基成果）、国际（主要公司）的中间件

（1）通用中间件
（2）消息中间件
（3）数据处理中间件（ETL）
（4）Web 应用（至少有容器）
（5）Web GIS 中间件
（6）报表处理中间件
（7）数据共享和交换中间件
（8）统一访问控制中间件

简介：消息处理中间件，即 ETL(Extract-Transform-Load, 数据抽取、转换、装载的过程)，

选择 DataPipline、Kettle、Talend、Informatica、Datax、Oracle Goldengate 六家国际主要公司做数据中间件对比。

DataPipeline 数据质量平台整合了数据质量分析、质量校验、质量监控等多方面特性， 以保证数据质量的完整性、一致性、准确性及唯一性，彻底解决数据孤岛和数据定义进化的问题。

Kettle 中文名称叫水壶，该项目的主程序员 MATT 希望把各种数据放到一个壶里，然后以一种指定的格式流出。Kettle 家族目前包括 4 个产品：Spoon、Pan、CHEF、Kitchen。SPOON 允许你通过图形界面来设计 ETL 转换过程(Transformation)。PAN 允许你批量运行由 Spoon 设计的 ETL 转换 (例如使用一个时间调度器)。Pan 是一个后台执行的程序，没有图形界面。CHEF 允许你创建任务(Job)。 任务通过允许每个转换，任务，脚本等等，更有利于自动化更新数据仓库的复杂工作。任务通过允许每个转换，任务，脚本等等。任务将会被检查，看看是否正确地运行了。KITCHEN 允许你批量使用由 Chef 设计的任务 (例如使用一个时间调度器)。KITCHE 也是一个后台运行的程序。

Talend，是一家专业的开源集成软件公司，为企业提供开源的中间件解决方案，从而让企业能够在他们的应用，系统以及数据库中赢取更大的价值。 在传统软件公司提供封闭、私有的解决方案的领域 Talend 系列软件以开源的形式进

行开发。Talend，可运行于 Hadoop 集群之间，直接生成 MapReduce 代码供 Hadoop 运行，从而可以降低部署难度和成本，加快分析速度。而且 Talend 还支持可进行并发事务处理的 Hadoop2.0。

Informatica 是全球领先的数据管理软件提供商。

DataX 是阿里巴巴集团内被广泛使用的离线数据同步工具/平台，实现包括 MySQL、Oracle、SqlServer、Postgre、HDFS、Hive、ADS、HBase、TableStore(OTS)、MaxCompute(ODPS)、DRDS 等各种异构数据源之间高效的数据同步功能。

GoldenGate 软件是一种基于日志的结构化数据复制软件。GoldenGate 能够实现大量交易数据的实时捕捉、变换和投递，实现源数据库与目标数据库的数据同步，保持亚秒级的数据延迟。源端通过抽取进程提取 redo log 或 archive log 日志内容，通过 pump 进程(TCP/IP 协议)发送到目标端，最后目标端的 rep 进程接收日志、解析并应用到目标端，进而完成数据同步。

这些公司在 ETL 的设计及架构上就存在着较大的差异，具体分析则可以看适用场景、使用方式和底层架构三方面：

适用场景：

DataPipline：主要用于数据融合、数据交换的场景，为超大数据量、搞复杂度数据链路设计的灵活、可扩展的数据交换平台。

Kettle：面向数据仓库建模传统 ETL 工具。

Oracle Goldengate：主要用于数据备份、容灾。

Informatica：面向数据仓库建模传统 ETL 工具。

Talend：面向数据仓库建模传统 ETL 工具。

Datax：面向数据仓库建模传统 ETL 工具。

使用方式：

DataPipline：全流程图形化界面，应用端采用 B/S 架构，Cloud Native 为云而生，所有操作在浏览器内就可以完成，不需要额外的开发和生产发布。

Kettle：不同于 DataPipline，Kettle 采取 C/S 客户端模式，开发和生产环境需要独立部署，任务的编写、调试、修改都在本地完成，需要发布到生产环境，线上生产环境没有界面，需要通过日志来调试、Debug，效率很低，费时费力。

Oracle Goldengate：没有图形化界面，操作都需要命令行方式完成，可配置能力差。

Informatica：C/S 客户端模式，开发和生产环境需要独立部署，任务的编写、调试、修改都在本地完成，需要发布到生产环境，学习成本较高，一般需要受过专业培训的工程师才能使用。。

Talend：C/S 客户端模式，开发和生产环境需要独立部署，任务的编写、调试、修改都在本地完成，需要发布到生产环境。

Datax：DataX 十一脚本方式执行任务的，需要完全吃透源码才能调用，学习成本较高，没有图形开发花介面和监控界面，运维成本相对高。

底层架构：

DataPipline：分布式集群高可用架构，可以水平扩展到多节点支持超大数据量，架构容错性高，可以自动调节任务在结点之间分配，适用于大数据场景。

Kettle：主从结构非高可用，扩展性差，架构容错性低，不适用大数据场景。

Oracle Goldengate：可做集群部署，规避单点故障，依赖于外部环境，如 Oracle RAC 等。

Informatica：schma mapping 非自动，可复制性比较差，更新换代不是很强。

Talend：支持分布式部署。

Datax：支持单机部署和集群部署两种方式。

各中间件的功能上也存在一定的差距，CDC 机制、对数据库的影响、自动断点续传、监控预告、数据清洗、数据转换这些功能都有不同的方式。

CDC 机制：

DataPipline：基于日志、基于时间戳和自增系列等多种方式可供选择。

Kettle：基于时间戳、触发器等方式。

Oracle Goldengate：主要是基于日志。

Informatica：基于日志、基于时间戳和自增系列等多种方式可供选择。

Talend：基于触发器、基于时间戳和自增系列等多种方式可供选择。

Datax：离线批处理。

对数据库的影响：

DataPipline：基于日志的采集方式对数据库无侵入性。

Kettle：对数据库表结构有要求，存在一定的侵入性。

Oracle Goldengate：源端数据库需要预留额外的缓存空间。

Informatica：基于日志的采集方式对数据库无侵入性。

Talend：有侵入性。

Datax：通过 sql select 采集数据，对数据源没有侵入性。

自动断点续传：

DataPipline：支持。

Kettle：不支持。

Oracle Goldengate：支持。

Informatica：不支持，她依赖于 ETL 设计的合理性（例如 T-1），指定续读某个时间点的数据，非自动。

Talend：不支持，她依赖于 ETL 设计的合理性（例如 T-1），指定续读某个时间点的数据，非自动。

Datax：不支持。

监控预告：

DataPipline：可视化的过程监控，提供多样化的图表，辅助运维，故障问题可以实时预警。

Kettle：依赖日志定位故障问题，往往只能是后处理的方式，缺少过程预警。

Oracle Goldengate：我图形化的界面预警。

Informatica：monitor 可以看到模糊信息，信息相对笼统，定位问题仍需要依赖分析日志。

Talend：有问题预警，定位问题仍需依赖日志。

Datax：以来工具日志定位故障问题，没有图形化运维界面和预警机制，需要自定义开发。

数据清洗：

DataPipline：围绕数据质量做轻量清洗。

Kettle：围绕数据仓库的数据希求进行建模计算，清洗功能相对复杂，需要手动编程。

Oracle Goldengate：轻量清洗。

Informatica：支持复杂逻辑的清洗和转化。

Talend：支持复杂逻辑的清洗和转化。

Datax：需要根据自己本身的清洗规则编写清洗脚本，进行调用 DataX3.0 所提供的功能。

数据转换：

DataPipline：自动化的 schema mapping。

Kettle：手动配置 schema mapping。

Oracle Goldengate：需要手动配置异构数据间的映射。

Informatica：手动配置 schema mapping。

Talend：手动配置 schema mapping。

Datax：通过编写 json 脚本进行配置 schema mapping。

不同的数据处理中间件因为其组成结构与框架的不同，实现方式也存在差异，所以拥有不同的特性，在数据实时性、应用难度、是否需要开发、易用性、稳定性上都有所不同。

数据实时性：

DataPipline：实时。

Kettle：非实时。

Oracle Goldengate：实时。

Informatica：支持实时，但现在主流应用都是基于时间戳等方式的批量处理，实时同步效率未知。

Talend：实时。

Datax：实时。

应用难度：

DataPipline：低。

Kettle：高。

Oracle Goldengate：中。

Informatica： 高。

Talend：中。

Datax：高。

是否需要开发：

DataPipline：否。

Kettle：是。

Oracle Goldengate：是。

Informatica：是。

Talend：是。

Datax：是。

易用性：

DataPipline：高。

Kettle：低。

Oracle Goldengate：中。

Informatica：低。

Talend：低。

Datax：低。

稳定性：

DataPipline：高。

Kettle：低。

Oracle Goldengate：高。

Informatica：中。

Talend：中。

Datax：中。

最后在其他方面，有一些细微的差别：

DataPipline：原厂实施和售后服务。

Kettle：开源软件，需要客户自行实施、维护。

Oracle Goldengate：原厂和第三方进实施和售后服务。

Informatica：主要为第三方的试试和售后服务。

Talend：分为开源版和企业便，企业版可提供相应的服务。

Datax：阿里开源代码，需要客户自动实施。