

# 主流数据同步/ETL工具的比较

比较维度/产品		TurboDX	Oracle Goldengate	Kettle	DataX	Informatica
设计 及 架 构	适用场景	异构数据库实时复制同步、备份、读写分离(支持视图增量)、实时ETL/ELT、数据汇聚、数据分发、文件同步分发、WS/REST服务对接、大数据MPP/NoSQL加载、数据跨节点安全传输	主要用于数据库复制、备份、容灾，不支持视图	面向数据仓库建模传统ETL工具	面向数据仓库建模传统ETL工具	面向数据仓库建模传统ETL工具
	产品架构	微服务容器架构、内存多线程流式处理、非侵入性架构、高容错机制设计、完全web界面任务配置和监控管理；简单易用、适应变化、灵活性高，可无缝升级为高可用性集群	任务的源端读与目标端写进程分别运行在两个实例进程中，中间通过TCP网络协议传输私有文件格式数据；可做集群部署，规避单点故障，但需依赖于外部环境，如Oracle RAC等	C/S客户端组件流程设计，批处理模式，线上生产环境没有管理界面；主从结构非高可用，扩展性差，架构容错性低，灵活性差	脚本方式执行任务，批处理模式、没有图形开发界面和监控界面；支持单机部署和集群部署两种方式	C/S客户端模式，开发和生产环境需要独立部署；schema mapping非自动；任务可复制性比较差，难于灵活适应数据需求的变化
	使用方式	完全web图形化界面“点击式”任务设计和监控管理，简单易用，不需要额外的开发和生产发布；无需在源库端或目标库端部署代理程序，对源库性能影响几乎为零；高级版支持多租户SaaS服务平台的使用方式	没有图形化的界面，操作皆为命令行方式，可配置能力差。界面配置管理工具需单独购买。	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，再发布到生产环境，线上生产环境没有界面，需要通过日志来调试、debug，效率低，费时费力	DataX是以脚本的方式执行任务的，需要完全吃透源码才可以调用，学习成本高，没有图形化开发配置和监控管理界面，运维成本相高	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，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
	元数据目录及智能分析	支持。具有字段识别、关系分析、主数据梳理等智能元数据分析功能，交换任务基于元数据库配置。	无	无	无	需另购数据目录产品
	任务类型	支持：1.全量任务；2.实时增量任务(日志CDC)；3.全量+增量任务(源库不停服务模式)；4.动态复制任务(DDL+DML)；5.表、视图增量交换整合任务（增量触发方式可选：CDC触发、标识位、时间戳、触发器、全量比对)；6.自定义SQL-EL任务；7.文件交换任务；8.数据文件加载任务；9.WS/REST服务对接任务	只支持CDC增量(日志模式)的复制同步任务，不支持全量任务；按表交换整合的任务(ETL)需另购ODI产品；不支持二进制文件的复制同步任务，不支持数据文件加载数据库/仓库的任务场景；不支持WS/REST服务对接；没有数据比对的功能	支持批处理的任务(ETL)，不支持日志模式的CDC增量复制同步任务；不支持二进制文件的复制同步任务场景；不支持WS/REST服务对接；没有数据比对的任务功能	支持批处理的任务(ETL)，不支持日志模式的CDC增量复制同步任务；不支持二进制文件的复制同步任务场景；不支持WS/REST服务对接；没有数据比对的任务功能	支持批处理的任务(ETL)，支持日志模式的CDC增量复制同步需另购CDC产品模块；不支持二进制文件的复制同步任务场景；不支持WS/REST服务对接；没有数据比对的任务功能

## 主流数据同步/ETL工具的比较

比较维度/产品		TurboDX	Oracle Goldengate	Kettle	DataX	Informatica
功能	<b>CDC机制</b>	事务增量CDC基于无侵入的日志模式(如 Oracle redo、Mysql binlog) , 按表/视图增量支持CDC触发、标识位、时间戳、触发器、自增序列、全量比对等多种方式可选	主要是基于日志	基于时间戳、触发器等	离线批处理	基于日志、基于时间戳和自增序列等多种方式可选; 日志CDC需另购CDC产品模块
	<b>对数据库的影响</b>	基于日志的采集方式无需在源库端部署任务代理程序(Agent)及建任何表, 对源数据库无侵入和影响压力	源端数据库需要预留额外的缓存空间	对数据库表结构有要求, 存在一定侵入性	通过sql select 采集数据, 对数据源有压力	基于日志的采集方式对数据库无侵入性, 但需另购CDC产品模块
	<b>自动断点续传</b>	所有任务类型均支持断点续传; 且集群版中任务转移后, 任务在新节点会自动从断点续传	支持	不支持	不支持	不支持, 依赖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
	<b>异构转换映射</b>	图形界面化、自动化的schema mapping和智能化的异构数据类型匹配; 支持schema级、表级、字段级的映射和记录级的函数处理	需手动配置异构数据间的映射	手动配置schema mapping及代码逻辑处理	通过编写json脚本进行schema mapping映射及代码函数处理	手动配置schema mapping, 通过编写脚本进行映射及函数处理
	<b>数据过滤、数据审核清洗、数据脱敏、函数处理</b>	支持数据过滤, 支持按规则数据审核和清洗。支持银行级别的数据脱敏和自定义函数处理, 支持源和目标数据比对及回补	轻量清洗	围绕数据仓库的数据需求进行建模计算, 清洗功能相对复杂, 需要手动编程	需要根据自身清晰规则编写清洗脚本, 进行调用(DataX3.0提供的功能)	支持复杂逻辑的清洗和转化
	<b>冲突策略</b>	支持用户勾选: 1.以源为主; 2.以目标为主; 3.自定义策略及智能规则	支持	不支持	不支持	支持

## 主流数据同步/ETL工具的比较

比较维度/产品		TurboDX	Oracle Goldengate	Kettle	DataX	Informatica
功能	流量控制	全量和增量均支持流量调节	不支持	不支持	不支持	不支持
	写端加载优化	支持用户勾选CDC串行、batched、或协同并行加载方式，以提升CDC事务增量的写入目标库的性能	支持事务增量的串行及并行加载方式	不支持按事务增量的加载，不保证表增量的时间次序性	不支持按事务增量的加载，不保证表增量的时间次序性	不支持按事务增量的加载，不保证表增量的时间次序性
	双向双写场景	支持	支持	不支持	不支持	不支持
	支持源库集群日志CDC	支持OracleRAC、DRDS、MySQL、PostgreSQL、MongoDB等源库集群	支持Oracle RAC	不支持	不支持	支持Oracle RAC，但需另购CDC产品模块
	监控预警通知	可视化的过程实时监控，提供多样化的图表，辅助运维，故障问题可实时预警和邮件通知(短信通知接口可定制)；提供对异常数据的回补功能，数据比对功能可生成报告	无图形化的界面预警和通知	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	依赖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	Monitor可以看到报错信息，信息相对笼统，定位问题仍需依赖分析日志
	数据发布/订阅服务	支持、内置	可支持，如通过第三方通道服务如Kafka	不支持	不支持	不支持
	Oracle视图实时增量	支持Oracle视图实时增量同步并保证数据一致性，解决了重型软件的读写分离痛点	不支持	不支持	不支持	不支持
	WS/REST服务双向对接	支持	不支持	不支持	不支持	不支持
	NoSQL, Kafka/RabbitMQ及各厂商大数据平台	支持Hadoop(Hdfs、Hive、HBase、Kudu)、MongoDB、Kafka、RabbitMQ及CDH、HDP、星环TDH、华为FI等各家大数据平台，支持kerberos认证连接	支持Kafka	不支持Kafka	不支持Kafka	不支持Kafka

## 主流数据同步/ETL工具的比较

比较维度/产品		TurboDX	Oracle Goldengate	Kettle	DataX	Informatica
功能	HA高可用性/负载集群	支持。集群任务节点互为热备，实现高可用性和负载均衡，及任务故障转移/自动恢复，任务迁移保持断点续传	需要与OracleRAC集成来支持	不支持	不支持	不支持
	多租户SaaS服务平台使用模式	支持	不支持	不支持	不支持	不支持
	部署位置	本地、云端、跨云	本地	本地	本地、云端	本地
	跨网络节点分布部署	支持，通过内置的数据通道服务	支持	不支持	不支持	不支持
特性	数据实时性	实时，秒级延时	实时	非实时	定时	支持实时，但是主流应用都是基于时间戳等方式做批量处理，实时同步效率未知
	应用难度	低	中	高	高	高
	是否需要开发	否	是	是	是	是
	易用性	高	中	低	低	低
	稳定性	高	高	低	中	中
其他	实施及售后服务	产品简单易用，用户或实施服务商可自我实施，原厂商售后技术支持服务	原厂和第三方的实施和售后服务	开源软件，需自客户自行实施、维护	阿里开源代码，需要客户自动实施、开发、维护	主要为第三方的实施和售后服务
	产地、信创	国产自主研发和知识产权。支持国产芯片、服务器、操作系统，已通过国产化鲲鹏适配认证。	美国	国外开源软件	阿里开源	美国