【杭州开云集致科技有限公司】

CloudCanal 产品白皮书

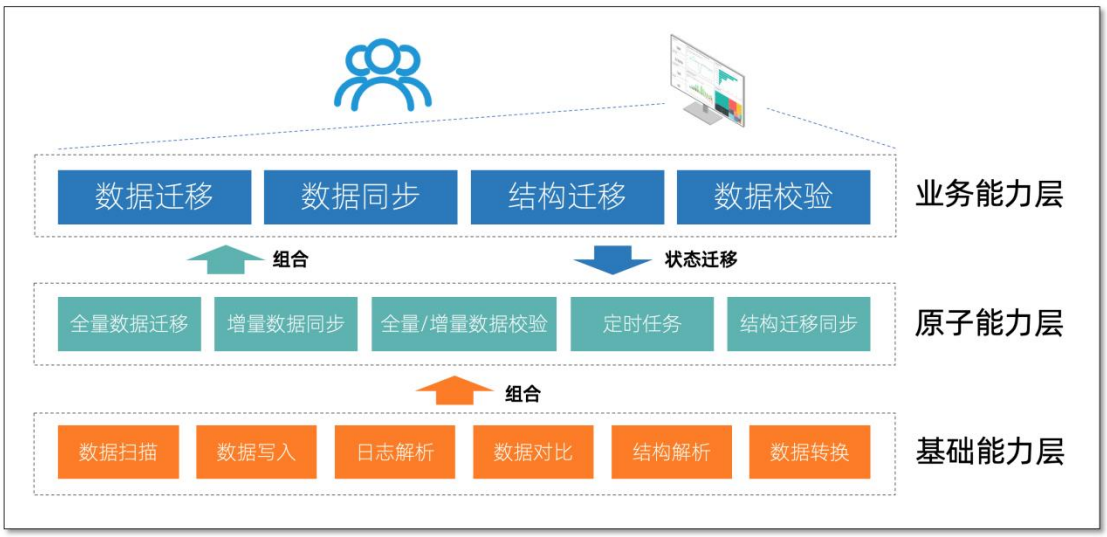
--阐述 CloudCanal 数据同步产品的说明文档

目录

1 产品概述

CloudCanal 是一款数据迁移同步工具，帮助企业快速构建高质量数据流通通道，产品包含 SaaS 模式和私有输出专享模式。开发团队核心成员来自大厂，具备数据库内核、大规模分布式系统、云产品构建背景，懂数据库，懂分布式，懂云产品商业和服务模式。

2 核心能力



2.1 数据迁移

数据迁移 将指定数据源数据全量搬迁到目标数据源，支持多种数据源，具备断点续传、顺序分页扫描、并行扫描、批量写入、并行写入、数据条件过滤等特点，对源端数据源影响小且性能好，同时满足数据轻度处理需求。

数据迁移 可选搭配 **结构迁移**、**迁移后指定时长数据同步**、**数据校验**，满足可能的业务平滑切换需求。

2.2 数据同步

数据同步 通过消费源端数据源增量操作日志，准实时在对端数据源重放，以达到数据同步目的，支持多种数据源，具备断点续传、DDL 同步、边同步边校验、对端事务保持、高性能对端写入、数据条件过滤等特点。

数据同步 可选搭配 **结构迁移**、**数据初始化(全量迁移)**、**单次或定时数据全量校验**，既便利，又能满足业务长周期数据同步对于数据质量的要求。

2.3 结构迁移

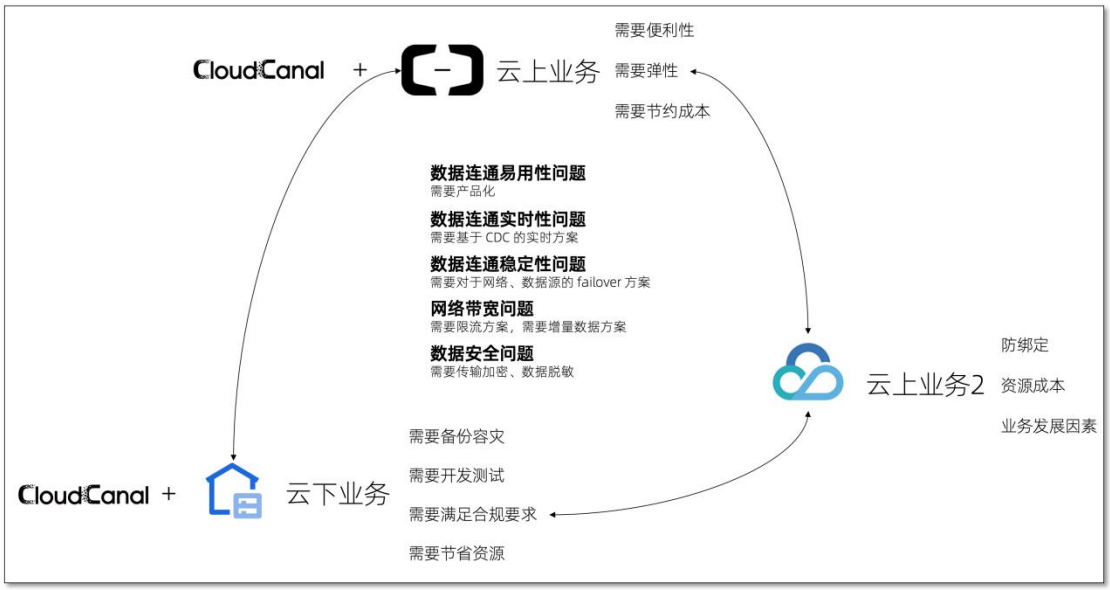
结构迁移 帮助用户快速镜像指定数据源结构，具备类型转换、数据库方言转换、命名映射等特点，可独立使用，也可作为 **数据迁移** 或 **数据同步** 准备步骤，灵活满足新数据构建需求。

2.4 数据校验

数据校验 让数据质量可衡量，可单独使用，也可配合 **数据迁移** 或 **数据同步** 使用，具备全量校验、增量校验、采样率、定时执行、校验数据条件过滤等特性，满足用户灵活的数据质量验证需求。

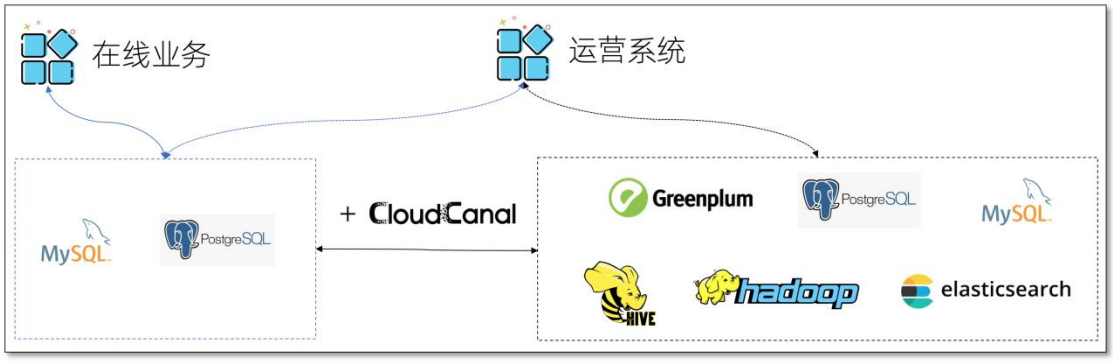
3 使用场景

3.1 云上云下、多云数据生态构建



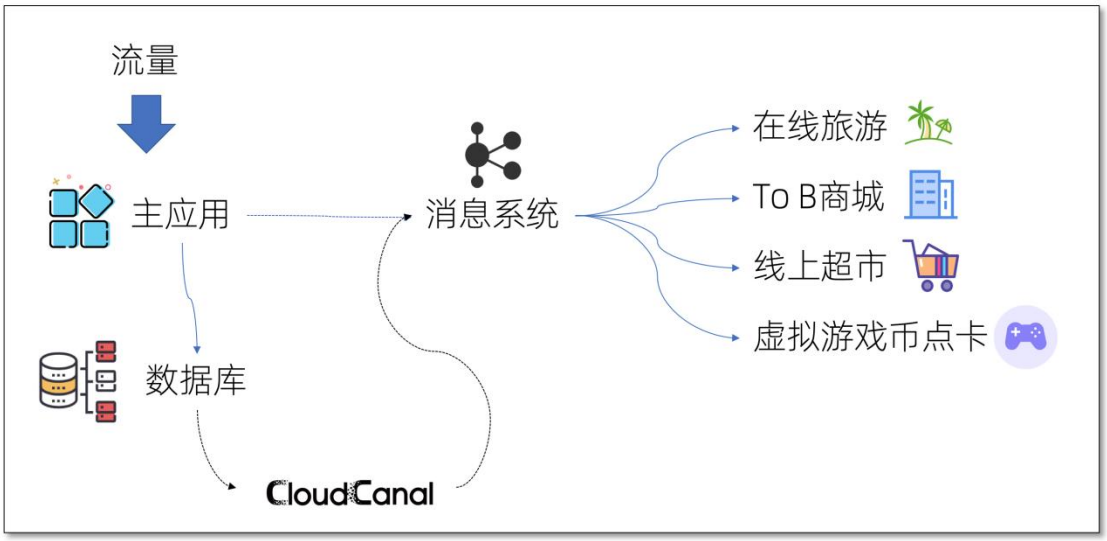
不同类型业务、开发和生产、主数据和数仓等不同维度数据放置于多云或云上云下环境，以满足高弹性、高性价比、可控性、安全合规等需求。CloudCanal 安全通信、稳定性、主流数据源支撑、全面的功能很好地满足此场景要求。

3.2 实时数仓构建



数据实时多维删选、聚合、链接在业务场景中越来越多，对于'快'的诉求永不停歇，找到一个强大的实时数仓同时，如何让主数据流畅、实时到达也成为了一个关键需求，CloudCanal 主流数仓支撑很好满足此类场景需求。

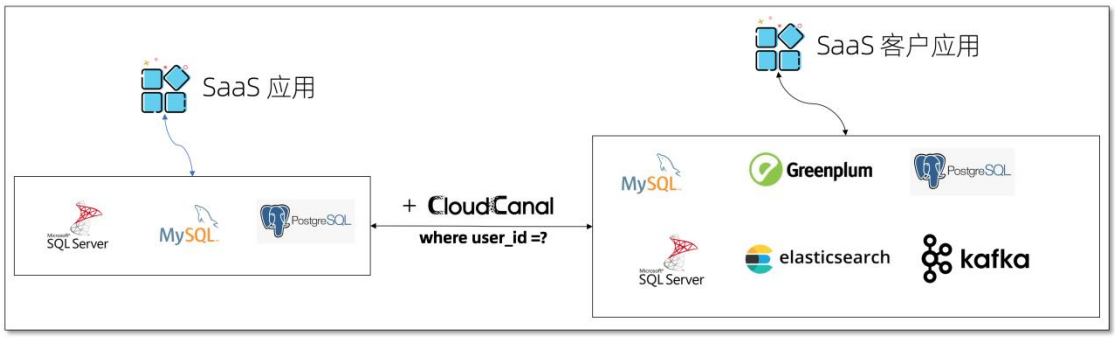
3.3 周边业务异步化



高并发业务的其中一个重要优化即同步操作 **只保留最关键操作** ，其他操作皆 **异步化** ，通过 **消息订阅模式** 补完流程，但写消息中间件有很多细节需要注意，包括如何保持事务，如何规避消息中间件不可用等问题，

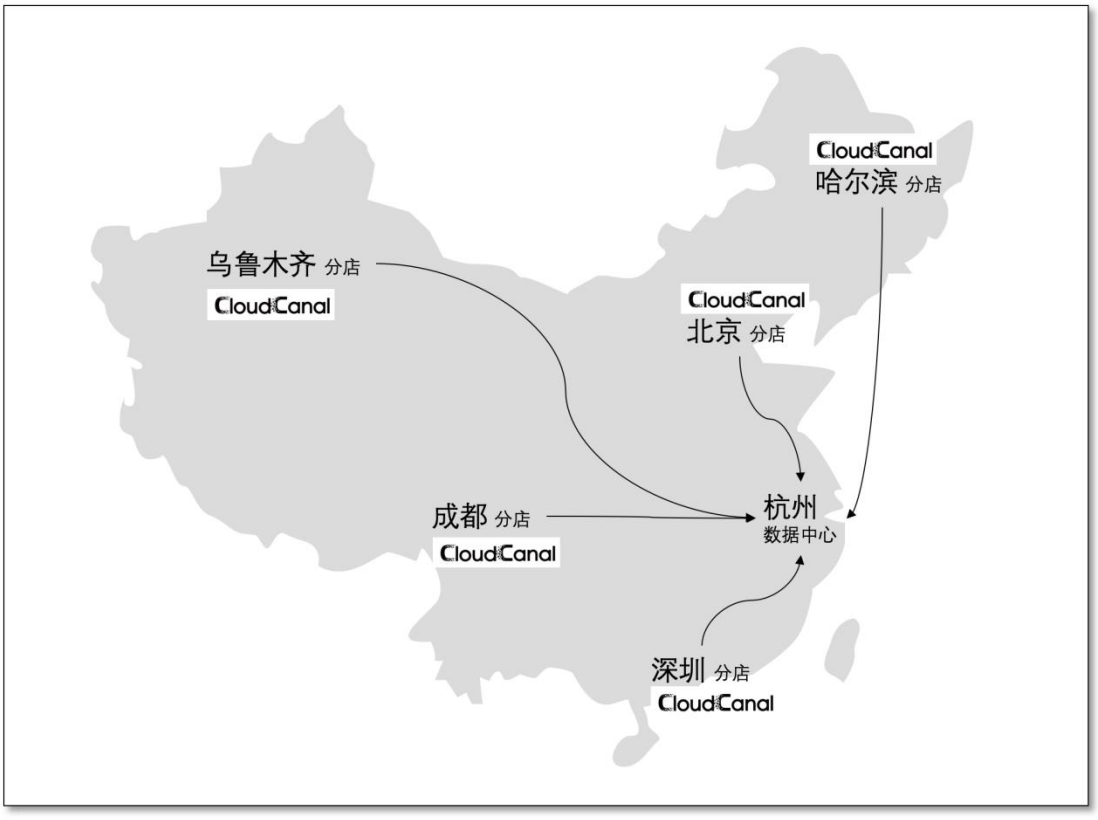
CloudCanal 通过 **链接数据增量变更** 和 **消息中间件**，主业务不需要关注消息中间件即可完成业务的异步化。

3.4 数据按需抽取同步



对于业务型 SaaS 平台，快速抽取同步指定用户数据构建专享服务是一项高价值业务，CloudCanal 数据条件过滤功能让这个工作顺畅进行。

3.5 数据集散



分散于各地的门店、网点产生订单等行为数据，迁移同步到云数据库、云数仓，再将数据归档到云上或自建大数据系统。完整的数据集散生态构建，CloudCanal 跨网络部署、容灾重试策略、主流数据库支撑很好匹配此场景诉求。

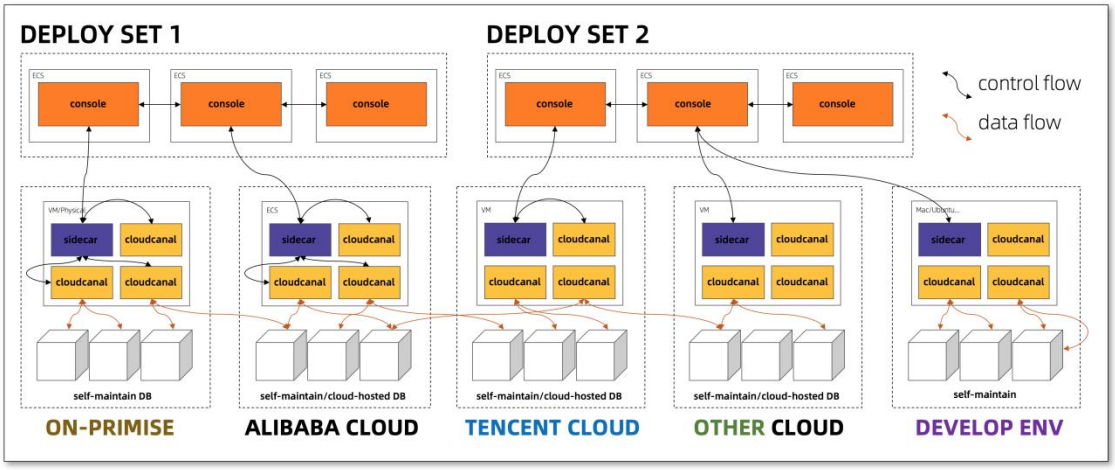
3.6 更多场景

数据迁移同步工具能够极大丰富业务的数据使用场景，让业务更好的使用数据，充分发挥数据本来的价值。更多场景欢迎交流。

4 产品架构

CloudCanal 产品架构整体体现为多租户、分布式的特点，目标是解决用户在各个网络环境下、数据在各个数据库或数据源间流动的需求。
本文从整体架构、容灾方案、网络方案简要介绍 CloudCanal 是如何达成这个目标。

4.1 整体架构



CloudCanal 组件分为 **Console**，**Sidecar** 和 **Server**，其中 Console 是集中化的管控服务, 以集群方式存在。
Sidecar 和 Server 部署具体迁移同步节点上，一个节点通常为一台 VM，或物理机，或云托管主机(ECS,EC2 等)，一个节点上只会有一个
Sidecar 进程，而 Server 进程有 0~n 个。

Console 包含了 CloudCanal 所有的产品化服务，包括生命周期管理、容灾调度、监控告警、流程状态机流转、机器和数据源管理、用户权限等，
其中用户资源被很好地隔离，所有操作被鉴权、审计记录。

Sidecar 进程单向访问 Console,其职责包括获取本用户需要运行的任务配置、收集和上报运行中任务的状态、执行任务的健康检查等工作。

Server 进程具体执行数据流动任务,根据不同类型的任务,Server 将运行不同类型的任务逻辑,包括全量数据迁移、增量数据同步、数据校验等，
同时它也会不断上报任务的执行位点、资源使用状况等。

5 容灾方案

CloudCanal 容灾主要包含两个方面:**管控容灾**和**任务容灾**。

5.1 管控容灾

CloudCanal 主要是通过集群化部署解决，有状态部分交由管控数据库解决。

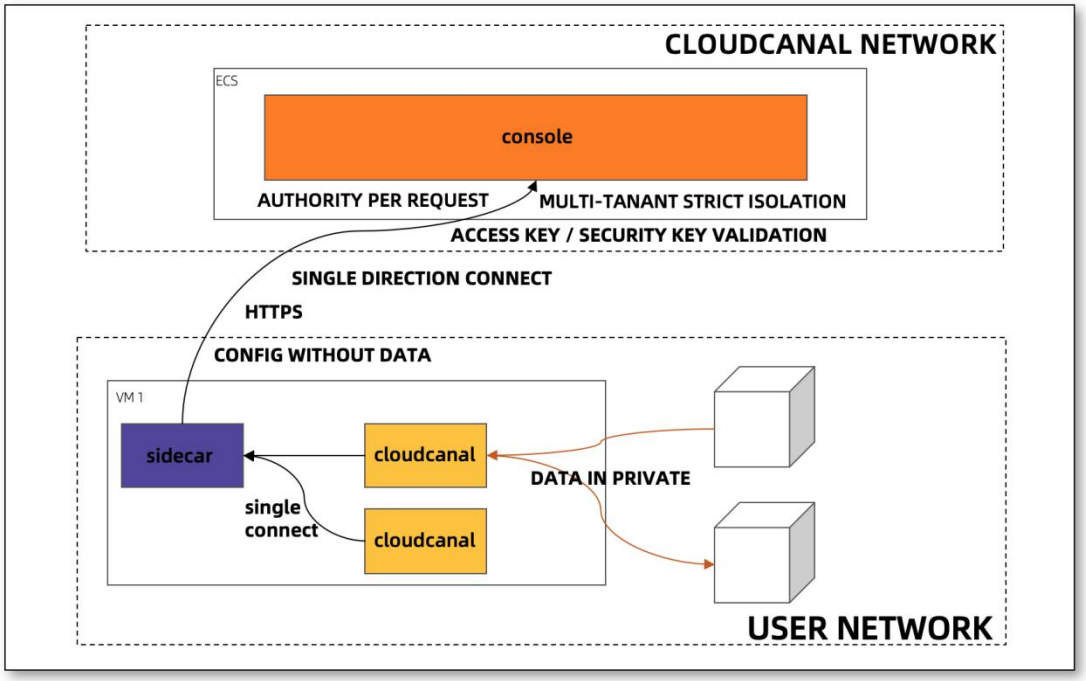
5.2 任务容灾

CloudCanal 任务容灾具备 2 级容灾，在节点资源充足的情况下，不强依赖底层操作系统或物理机器容灾措施。

1 级容灾：Sidecar 进程退出或机器不正常以及网络隔离情况下，Console 根据租期和 Sidecar 链接状态，进行主动容灾调度，

2 级容灾：Sidecar 进程正常，任务进程不正常，Sidecar 通过健康监测保障其负责的任务按照管控指定的状态运行，保活或保死，

6 网络方案(安全措施)



CloudCanal 为了适应多租户、分布式系统部署要求，采用了多种网络安全措施，确保用户数据和信息安全。

6.1 单向链接

用户节点反向链接管控方式进行服务。用户节点不主动暴露网络信息到开放网络，只会链接 CloudCanal 集中化管控服务。

6.2 HTTPS 协议

用户节点和 CloudCanal 通信链路采用 HTTPS 协议，防止盗取并篡改信息。

6.3 数据不出网络

所有数据流转均发生在用户网络环境，数据不流出泄漏。CloudCanal 所有针对数据源的动作均发生在用户网络环境。

6.4 长连接 AccessKey SecurityKey 认证

通信链路采用 TCP 长连接，每一次连接经过用户独有的 AccessKey 和 SecurityKey 认证。

6.5 请求验证

每一次请求主体信息，都经过资源归属验证，防止跨用户获取信息。

6.6 操作审计

控制台操作，用户节点请求均进行操作审计记录，可追踪溯源。

7 产品功能

CloudCanal 具备丰富的功能以满足业务数据流通场景的需求，本篇文章简要介绍该产品具备的主要能力。

7.1 功能介绍

功能名称	功能类型	功能介绍
数据同步	内核	利用源端数据源变更日志，准实时采集并写入对端数据源，对端数据源具备实时写入能力，延迟亚秒级
数据迁移	内核	通过扫描源端数据源数据，批量、多线程写入对端数据源，通常需要秒~小时级别时间完成此类操作
结构迁移	产品化	针对结构化数据源(如关系型数据库)，扫描源端表结构定义，将其应用到对端数据源，对于异构数据源，通常存在类型或特定方言转换
结构同步	内核	针对结构化数据源(如关系型数据库)，接收结构变更日志，将相应变更语句应用到对端，对于异构数据源，通常存在类型或者特定方言转换
全量数据校验	内核	通过扫描源端数据源数据，批量、多线程对比对端数据源内对应数据,并实时报告数据缺失、不一致等情况
增量数据校验	内核	利用源端数据源变更日志，准实时对比对端数据源内对应数据
定时任务	产品化	支持按小时、天、周频率，定时周期性运行数据迁移和数据校验任务
流量控制	内核	支持基于流量的速度控制(增量同步)和基于对对端写入 RPS 的速度控制(全量迁移)
数据条件过滤	内核	设定 SQL where 条件(=、in、>、<)过滤数据，全量迁移和数据同步皆支持
操作过滤	内核	支持操作过滤(默认全同步)，包括 INSERT、UPDATE、DELETE
事务模式	内核	如果两端都支持事务，则可按事务粒度对端写入，即对端保持和源端一致的事务性
高性能优化	内核	非事务模式下，支持按 pk 或 table HASH 归组并行写入
图形化配置	产品化	数据流转任务、机器资源、数据源、监控告警等图形化配置
自动流程	产品化	多阶段数据流转任务流程全自动化支持
环境预检测	产品化	数十项涉及数据源、网络等环境预检查
实体名称映射	产品化	支持已有实体名称映射，比如关系型数据库的库表映射，消息的 topic 映射等
黑白名单	产品化	支持已有实体名称黑白名单，比如关系型数据库库表、消息 topic 等黑白名单
位点回溯	产品化	支持按时间进行位点回溯，重新消费一段时间增量日志
任务调度	产品化	基于租期的任务存活检测和自动调度，也支持手动调度任务到对应机器
多维监控	产品化	支持机器和任务多维度指标监控和展现
多级告警	产品化	支持邮件、钉钉、短信告警
用户权限	产品化	支持用户权限，资源、操作、页面全鉴权，用户间资源、操作严格隔离
全局异常上报	产品化	产品化流程、调度、数据任务操作异常上报并展现，可审计任务运行状

7.2 核心能力数据源支持

核心能力指**迁移**（单次或定时迁移）、**同步**（增量数据同步）、**校验**（单次或定时数据校验）、**结构**（结构迁移或同步）、**订正**（全量或者增量数据订正

源/目标	MySQL	ORACLE	PostgreSQL	Greenplum	SQLServer	Kafka	RocketMQ	RabbitMQ	ElasticSearch	PolarDB-MySQL
MySQL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ORACLE	✓	✓	✓	✓		✓				
PostgreSQL	✓		✓	✓		✓				
SQLServer	✓				✓	✓				✓
PolarDB-MySQL	✓		✓	✓		✓	✓			✓
MongoDB	✓					✓				✓
Kafka	✓	✓	✓	✓	✓				✓	✓
OceanBase	✓		✓	✓						

源/目标	ClickHouse	StarRocks	Redis	MongoDB	OceanBase	TiDB	Doris	PolarDB-X	Hive	Kudu	ADB for MySQL
MySQL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ORACLE	✓	✓			✓	✓				✓	
PostgreSQL	✓	✓			✓		✓			✓	
SQLServer											
PolarDB-MySQL		✓					✓				
MongoDB						✓					
Kafka	✓	✓		✓							
OceanBase		✓			✓						

核心能力 数据迁移 数据同步 结构迁移 数据校验 数据订正

7.3 数据源支持计划

- 打通大部分数据源
- 细化、丰富支持的数据源版本
  - ORACLE
  - SQLServer
  - PostgreSQL
- 源端
  - API
- 目标端
  - Aliyun OSS
  - 达梦数据库
  - 高斯数据库
  - API
- DataSource SDK
  - 数据源规范,方便用户自行开发数据源插件并接入

7.4 云原生支持计划

CloudCanal 将支持云原生托管数据库服务包含，1.通过云服务商开放 API 对接全自动流程.2.对于云厂商数据源差异进行细致适配

- 腾讯云
- 华为云
- AWS

7.5 产品化能力迭代计划

- 账号授权体系(包含类 LDAP 服务)
- 链路拓扑管理

8 产品优势

CloudCanal 具备安全、便利、中立、稳定、全面等特点，和市场主流产品对比，也具备不错的竞争力，并且快速成长中。

8.1 核心优势

8.1.1 更安全

CloudCanal 将数据流动动作完全放置于用户私有环境下，数据不出该网络，无论是本地开发环境、云上 VPC 、自建机房环境都能很好使用。
客户端单向(出)访问管控,通信链路通过 tls 加密,并采用标准 AccessKey 和 SecurityKey 认证与鉴权,用户所涉及资源均具备良好隔离。

8.1.2 更便利

对于云托管 VM (如阿里云 ECS),全自动安装客户端即可运行数据流转任务，对于自有机器(如本地服务器),只需下载客户端并填充节点身份信息,即可运行。
无论是多阶段复杂数据流转，还是简单结构迁移，无论是长周期数据同步，还是单次数据全量迁移，CloudCanal 都能够流畅配置，5 分钟内运行。

8.1.3 更中立

对于 CloudCanal 已经支持的数据源，还是即将支持的数据源，CloudCanal 秉承中立的原则

- 开源与托管数据源并重，让用户自由选择、搭配风险
- 支持数据进与出，让业务尝试数据库无后顾之忧(vendor lock-in avoiding)
- 运行环境对齐，云托管 VM、自有物理机、本地开发机(Mac/Ubuntu) 都能运行
- 私有输出零依赖，不强绑定任何平台、体系，平等支持各云平台和私有环境

8.1.4更稳定

CloudCanal 采用 E2E(end-to-end),基于租约的容灾,自动重试等机制让数据流动链路更加稳定

- E2E 模式将链路缩到最短，没有分布式流式计算引擎，没有消息中间件，只为可靠、实时传递数据到对端数据源(对端可搭配流式计算引擎、消息中间件等)
- 分布式场景下，基于租约的一致性动作是相对可靠的方式，CloudCanal 将容灾探测、调度(手动/自动)做得更加可靠、可控
- 自动重试机制让长距离数据传输同步更加少让人操心，99% 的问题得到更加快速的解决

8.1.5更全面

CloudCanal 的功能完整性和开发团队专业性让其更加全面。
尽力补齐数据源到数据源的核心能力，结构迁移/同步、数据初始化(全量)、数据同步和数据校验一个都不能少(部分能力补齐中)。
团队成员既有数据库内核开发出身，也有超大规模实时数据同步开发运维经验加持，更具备专业做云计算产品经验和能力。懂数据库，懂超大规模，懂云产品商业和服务模式。

9 产品兼容列表

CloudCanal 数据融合系统数据库组合清单		
序号	**产品模块**	**产品能力组成**
1	**在线数据构建**	1.MySQL->Redis 数据迁移、同步、校验
		2.MySQL->ElasticSearch 数据迁移、同步、结构迁移

	(ToC)	3.MySQL->RocketMQ 数据迁移、同步
		4.MySQL->RabbitMQ 数据迁移、同步
		5.MySQL->MySQL 数据迁移、同步、校验、订正、结构迁移
		6.MySQL->PostgreSQL 数据迁移、同步、校验、结构迁移
		7.MySQL->SQLServer 数据迁移、同步、校验、订正、结构迁移
		8.MySQL->Oracle 数据迁移、同步、校验、订正、结构迁移
		9.PostgreSQL->PostgreSQL 数据迁移、同步、校验、结构迁移
		10.PostgreSQL->MySQL 数据迁移、同步、结构迁移
		11.Oracle->Oracle 数据迁移、同步、结构迁移
		12.Oracle->PostgreSQL 数据迁移、同步、结构迁移
		13.Oracle->MySQL 数据迁移、同步、结构迁移
		14.SQLServer->SQLServer 数据迁移、同步、结构迁移
		15.RocketMQ->MySQL 数据迁移、同步
		16.RabbitMQ->Mysql 数据迁移、同步
2	实时数仓 / 小数仓构建 (面向后台或运营)	1.MySQL->Greenplum 数据迁移、同步、校验、结构迁移
		2.MySQL->ADB for MySQL 数据迁移、同步、结构迁移
		3.MySQL->TiDB 数据迁移、同步、结构迁移
		4.MySQL->Kudu 数据迁移、同步、结构迁移
		5.MySQL->ClickHouse 数据迁移、同步、校验、结构迁移
		6.MySQL->MongoDB 数据迁移、同步、结构迁移
		7.MySQL->StarRocks 数据迁移、同步、结构迁移
		8.MySQL->Doris 数据迁移、同步、结构迁移
		9.PostgreSQL->Kudu 数据迁移、同步、结构迁移
		10.Oracle->Kudu 数据迁移、同步、结构迁移
		11.Oracle->TiDB 数据迁移、同步、结构迁移
		12.Oracle->StarRocks 数据迁移、同步、结构迁移
		13.Oracle->OceanBase 数据迁移、同步、结构迁移
		14.Oracle->ClickHouse 数据迁移、同步、结构迁移
		15.Oracle->PostgreSQL 数据迁移、同步
		16.PostgreSQL->OceanBase 数据迁移、同步
		17.PostgreSQL->Doris 数据迁移、同步

		18.PostgreSQL->ClickHouse 数据迁移、同步
		19.StarRocks->Mysql 数据迁移、同步
		20.Greenplum->Starrocks 数据迁移、同步
		21.Greenplum->PostgreSQL 数据迁移、同步
		22.Greenplum->Greenplum 数据迁移、同步、结构迁移
3	**超高并发在线数据构建**	1.MySQL->PolarDB-X 数据迁移、同步、结构迁移
		2.MySQL->Oceanbase 数据迁移、同步、结构迁移
		3.SQLServer->Kafka 数据迁移、同步
		4.PolarDB-X->Kafka 数据迁移、同步
		5.Oceanbase->MySQL 数据迁移、同步
		6.Oceanbase->Oceanbase 数据迁移、同步
		7.Oceanbase->StarRocks 数据迁移、同步、结构迁移
		8.PostgreSQL->Kafka 数据迁移、同步
		9.MongoDB->TiDB 数据迁移、同步
		10.MongoDB->MongoDB 数据迁移、同步
4	**大数据构建**	1.MySQL->Kafka 数据迁移、同步
		2.MySQL->Hive 定时数据迁移
		3.Kafka->Kafka 数据迁移、同步
		4.PostgreSQL->StarRocks 数据迁移、同步
		5.Kafka->ElasticSearch 数据迁移、同步
		6.Kafka->MongoDB 数据迁移、同步
		7.Kafka->Oracle 数据迁移、同步
		8.Kafka->StarRocks 数据迁移、同步
		9.Kafka->ClickHouse 数据迁移、同步
		10.Oracle->Kafka 数据迁移、同步、结构迁移
5	**云数据生态构建(专项适配)**	1.Aliyun RDS for MySQL 支持 (数据连通能力等同 MySQL)
		2.Aliyun RDS for Pg 支持(数据连通能力等同 PostgreSQL)
		3.Aliyun RDS for Redis 支持(数据连通能力等同 Redis)
		4.Aliyun RDS for MongoDB 支持(数据连通能力等同 MongoDB)
		5.Aliyun PolarDB for MySQL
		6.Aliyun ADB for MySQL
		7.Aliyun ADB for Pg(数据连通能力等同 Greenplum)

		8.Aliyun ADB for ClickHouse(数据连通能力等同 ClickHouse)
		9.Aliyun ElasticSearch(数据连通能力等同 ElasticSearch)
		10.Aliyun Kafka(数据连通能力等同 Kafka)
		11.Aliyun RocketMQ(数据连通能力等同 RocketMQ)
6	**备注**	随着软件支持数据库组合种类增多，此表会做相应更新； 其他暂未支持的链路或能力，如果客户有要求，可联系我方按照人力投入进行报价评估；