CYPHY
赛　凡

云梦数据仓

B2100 Big Data Appliance

概述

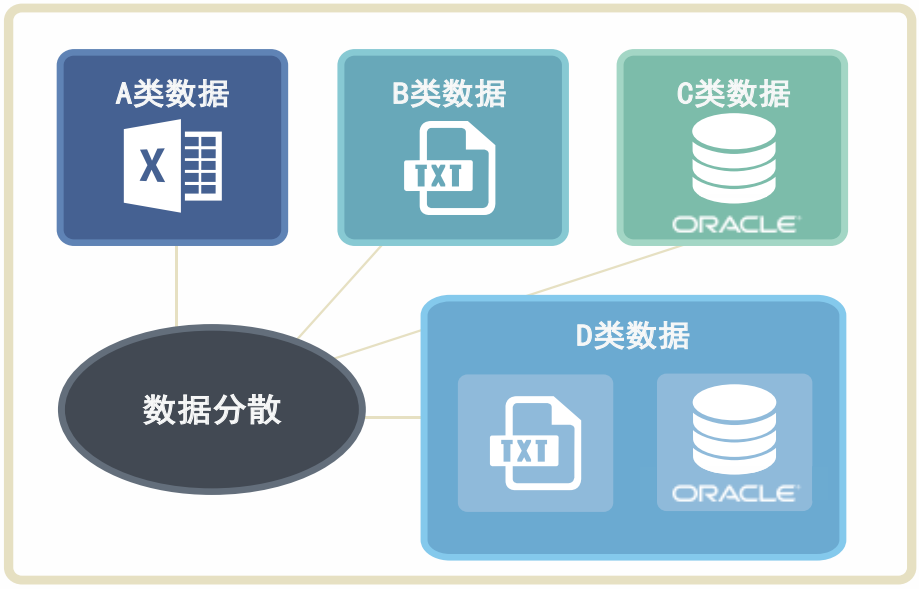
信息化建设经过近十年的高速发展，涌现出了各种各样的数据。每个单位、企业都积累了大量的数据，但早期由于缺乏统一的规划，导致现在的数据无论是存储方式还是数据标准都千差万别，这为数据分析工作带来了巨大的麻烦。我们必须要先把这些数据都清洗成统一的标准，才能在数据分析的工作中发挥这些数据潜在的价值。

云梦数据仓是一款软硬件一体化的数据清洗设备，一站式支持数据迁移、数据同步、数据交换和数据整合，可对结构化及非结构化数据进行清洗整理，全面解决因数据杂乱无章给您带来的困扰。

常见的数据问题

数据杂乱

在当前的企业或单位内部，由于历史原因，数据存储方式是多样化的。有些数据采用 Excel 格式文件存放，有些数据采用 TXT 格式存放，有些数据存放于关系型数据库中（Oracle、SQL 等），有些数据则更加复杂，部分存放于文件中，部分存放于数据库中。



数据质量低

在现实中即使是同一类数据，其数据标准也不一致。以上图 D 类数据为例，在 TXT 文件中可能有 32 个字段，而其在 Oracle 数据库中却只有 28 个字段。

即使是全部位于 TXT 文件中的 B 类数据，可能其质量也不高，例如对于主叫号码字段，有的带有 IP 拨号前缀（17909134071648O5），有的却不带，有的带区号，有的不带区号。对于日期字段，有的是 2009-09-08 方式存储，有的却是 2009-9-8 方式存储。

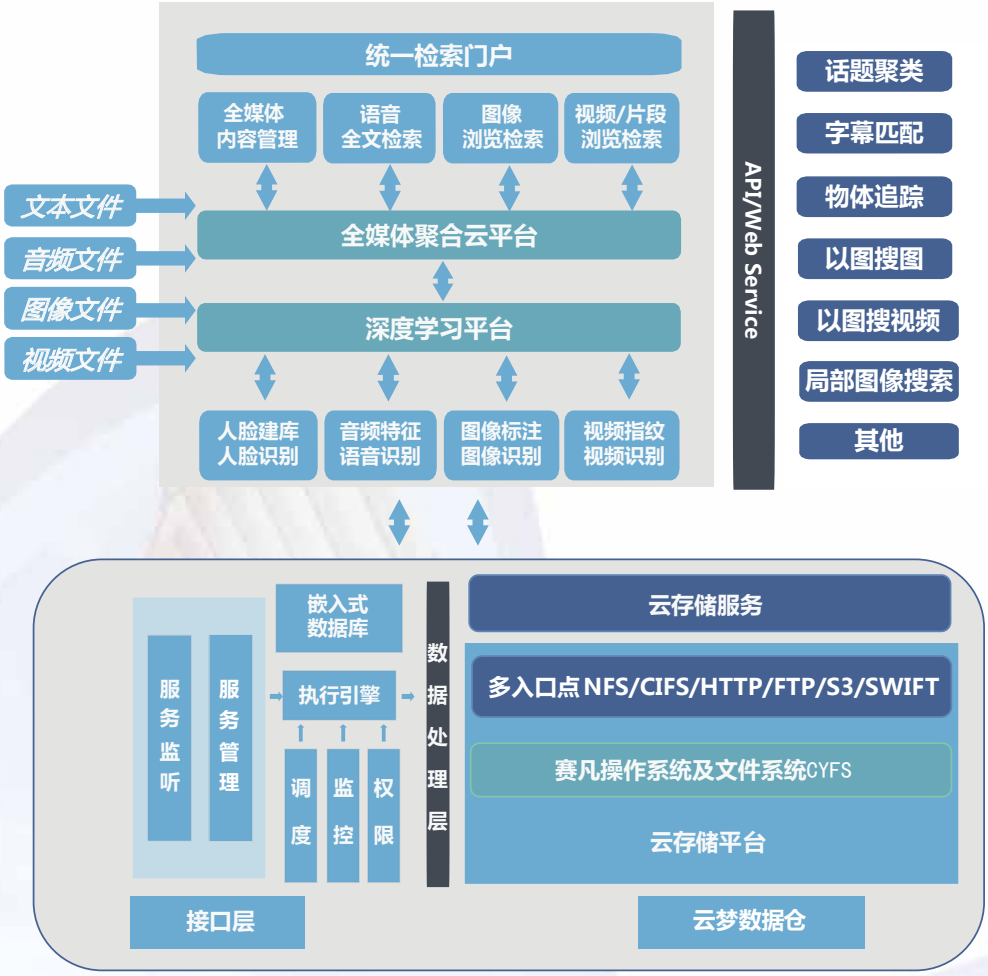
数据杂乱					
姓名	身份证	电话号码	出生日期	地址	...
張雲親	32108 2584698457456	13974558758	1989-09-23	南京市下关區四平二路 319 号 102 室	
劉楊	320107 1975 02073434	17909 13407164805	1975207	江蘇省仪征市新城鎮鋪薪村第十七組 30 號	
往艳芳	44068119820105024X	01058988666	1982/01/05	北京市昌平區 5 号楼 3 單元 401 號	
李金嵐	360425 197201142025	010-89874521	19720114	北京市朝阳区紅旗村 458 號	
张天赐	421051972011425025	13798665485	19720114	江西省九江市永修县军山毛纺 2 厂生活區	
徐洪艳	320382 45112545521	58466954	1987-7-1	江苏省苏州市四户镇竹园村 5 組 317 號	
博涵	650102199101	13569884574	1991-01-18	南京市玄武區南林一村 1 号	

效率低，通用性差

针对以上情况，有些企业和单位采用了一些专用的数据清洗工具，但仍存在两个问题：

1. 不具备通用性，这些专用的数据清洗工具处理的数据格式有限，且与特定的数据库产品绑定。
2. 整体工作效率不高，数据清洗的经验无法在工具内部进行分享，导致每个人都需要重新去设计数据清洗规则，且无法对既定模式的数据清洗规则进行定时执行，导致昨天做过的事情，今天又得重新做一次相似的工作。

云梦数据仓结构



技术优势

全面的数据源支持

支持各种主流数据库（Oracle、SQL Server、DB2、MySQL、Sybase、PostgreSQL 等）的全量和增量数据抽取和装载，还支持 TXT、CSV、Excel、XML 文件、消息服务器、LDAP 服务器、WebService 等数据的抽取和装载。

丰富的数据转换清洗规则

内置近 40 种数据清洗转换规则，并可动态扩充。对于简繁体、汉字拼音、乱码处理、字符集转换、中文数字等中文特有的问题都提供了内置的转换规则进行处理。

批量文件处理与文件同步

支持各种文件的批量读取，自动识别新增和修改的文件，并且能够在本地和远程服务器之间同步文件夹。

多重协议访问支持

支持通过 CIFS、NFS、FTP、WebDAV 对云梦数据仓同时访问。

高性能硬件平台

采用 2 颗 Intel 64bit Xeon 4 核高性能处理器；
对外接口采用 4x 10GbE 高速接口，单台设备性能可达 600MB/s。

高性能软件架构

采用基于流水线的多线程架构，并支持数据分区处理和并行装载，可以充分发挥硬件性能，数据处理能力，并且可以随着系统 CPU 和 I/O 性能的提升而同步提升。

安全性

一体化结构

19 英寸标准机架式 2U 专用硬件平台，嵌入式系统结构，安全可靠。

断点续传机制

支持自动重连与断点续传机制，当外部数据源出现故障且故障排除后可以自动重新进行数据同步。数据在不同的节点之间传输时通过消息确认、消息持久化和重传机制可以确保数据不丢失。

权限控制与加密传输

提供基于角色的权限控制机制，责权明晰。支持传输加密，不同交换节点之间传输的数据都是经过加密处理，防止信息泄露。

数据保护技术

底层 RAID 保护机制，支持文件级 RAID5、6、7，支持同一 RAID 最多在三块硬盘同时损坏时正常运行；多副本技术支持重要数据多重备份，支持 VTL 及 NDMP 备份；支持读写快照保护。

扩展性

插件机制

采用基于 OSGI 整体架构，可以方便快速的接入新类型的数据源；定制特殊的 / 复杂的业务逻辑转换组件或规则。

第三方支持

提供 API 接口，通过 API 第三方应用可以动态的创建、执行流程，获取监控和统计信息。

容量扩充

云梦数据仓自带 48TB 存储空间，最多可扩展至 792TB 存储空间。

应用场景

分散数据归一

使用云梦数据仓的数据清洗功能，可以轻易实现对结构化数据及非结构化数据进行集中清洗，并存储到数据库中。

云梦数据仓内嵌多种数据连接和装载通道，可支持绝大多数结构化（Oracle、SQL Server、MySQL、DB2、Access、Sybase、GreenPlum 等）及非结构化数据（TXT、Excel、CSV、XML 等），轻松实现对数据的内容识别及数据装载。

脏数据清洗

- 使用云梦数据仓可以轻松的实现数据清洗功能，可对结构化数据及非结构化数据进行标准化清洗，并形成统一的格式。
- 云梦数据仓内置 40 多种数据清洗规则，通过对不同清洗规则的组合使用，可以实现绝大部分的数据清洗功能，对于通过组合规则实现不了的清洗功能，还可以通过编写自定义函数的方式来实现特定格式的清洗。

清洗前

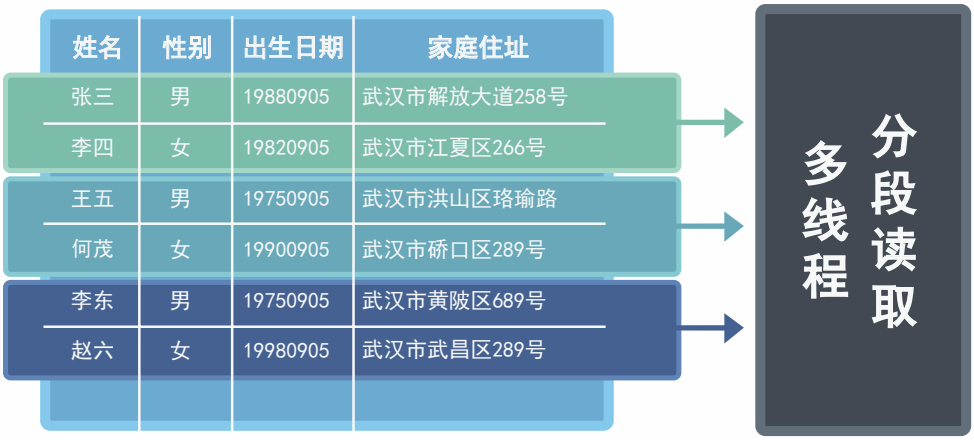
姓名	身份证	电话号码	出生日期	地址	...
張雲親	32108 2584698457456	13974558758	1989-09-23	南京市下关區四平二路 319 号 102 室	
劉楊	320107 1975 02073434	17909 13407164805	1975207	江蘇省仪征市新城鎮鋪薪村第十七組 30 號	
往艳芳	44068119820105 024 X	01058988666	1982/01/05	北京市昌平區 5 号楼 3 單元 401 號	
李金嵐	360425 197201142025	010-89874521	19720114	北京市朝阳区紅旗村 458 號	
张天赐	421051972011 425025	13798665485	19720114	江西省九江市永修县军山毛纺 2 厂生活區	
徐洪艳	320382 45112545521	58466954	1987-7-1	江苏省苏州市四户镇竹园村 5 組 317 號	
博涵	650102199101	13569884574	1991-01-18	南京市玄武區南林一村 1 号	

清洗后

姓名	身份证	联系电话	出生日期	地址	...
张云亲	321082584698457456	13974558758	1989-09-23	南京市下关区四平二路 319 号 102 室	
刘杨	320107197502073434	13407164805	1975-02-07	江苏省仪征市新城镇铺薪村第十七组 30 号	
往艳芳	44068119820105024X	01058988666	1982-01-05	北京市昌平区 5 号楼 3 单元 401 号	
李金岚	360425197201142025	01089874521	1972-01-14	北京市朝阳区红旗村 458 号	
张天赐	421051972011425025	13798665485	1972-01-14	江西省九江市永修县君山毛纺 2 厂生活区	
徐洪艳	320382451125455 21	02158466954	1987-07-01	江苏省苏州市四户镇竹园村 5 组 317 号	
博涵	650102199198547 85	13569884574	1991-01-18	南京市玄武区南林一村 1 号	

性能优化

云梦数据仓高性能一体化硬件平台，搭配创新性的文件分段读取、流水线作业处理、清洗规则并行三大核心技术，可使数据处理性能相比其他传统数据清洗工具，性能提高 10 倍以上。



定期清理

云梦数据仓具有任务调度功能，仅需将之前做好的数据清洗规则添加上任务调度，一旦检测到有新数据到来，即可触发数据清洗规则，自动完成数据的清洗和装载功能，无需做任何操作。同时还可以设置任务间隔运行的时间、触发条件。

模式分享

云梦数据仓具有数据清洗规则导出功能，可以将所设计好的清洗规则导出成文件，在其他设备上进行还原即可使用，无需重新设计规则，节省劳动。

技术规格

规格型号	云梦数据仓
尺寸	19 英寸标准机架式 2U12 盘位
接口	2x10GbE + 2x1GbE 对外接口
硬盘数量	热插拔硬盘 12 块起，支持 SSD 缓存加速
容量	10TB/30TB/160TB(SAS) 48TB/144TB/768TB(NL-SAS)
RAID 级别	支持 RAID0,1,5,6,7,10,50,60,70, 支持快速重建
数据源	支持 Oracle、SQL Server、DB2、MySQL、Sybase、PostgreSQL 等
转换清洗	内置近 40 种数据清洗转换规则，并可动态扩充
批量处理	支持各种文件的批量读取，自动识别新增和修改的文件，并且能够在本地和远程服务器之间同步文件夹
扩展	数据源适配、流程节点、转换规则都支持插件机制，支持自定义函数编写
接口	支持 API 接口，支持第三方应用
数据服务	支持快照、多副本备份、自动精简配置及 VTL、云盘（可选）
环境参数	1100W 冗余电源、冗余风扇 电压：100~240V 50~60Hz 工作温度：10℃ ~35℃ 非工作温度：-10℃ ~50℃ 工作湿度：20%~80% 非工作湿度：10%~95% 非冷凝

微信官方账号

如需获取更多资讯，请登陆 www.cyphytech.com

赛凡信息科技（厦门）有限公司（Cyphy Technology Co., Ltd.）成立于2011年，总部位于厦门软件园，在北京设立研发及营销中心，并在上海、广西、四川、台湾及香港等地设立办事处。

赛凡科技以业界领先的技术、产品和解决方案，为企业提供安全的云存储服务。自成立开始，赛凡科技始终坚持中国智造、自主可控、技术领先的经营理念，持续为政府、医疗、教育、公安安防、科研院所、金融、能源等多个行业用户应需定制贴近应用的产品和解决方案。

400 热线：400 879 8066
咨询信箱：service@cyphytech.com

2015 年 12 月创建，2016 年 3 月第一次修改