第七届

数据技术嘉年华

Data Technology Carnival

云·数据·智能 - 数聚价值智胜未来

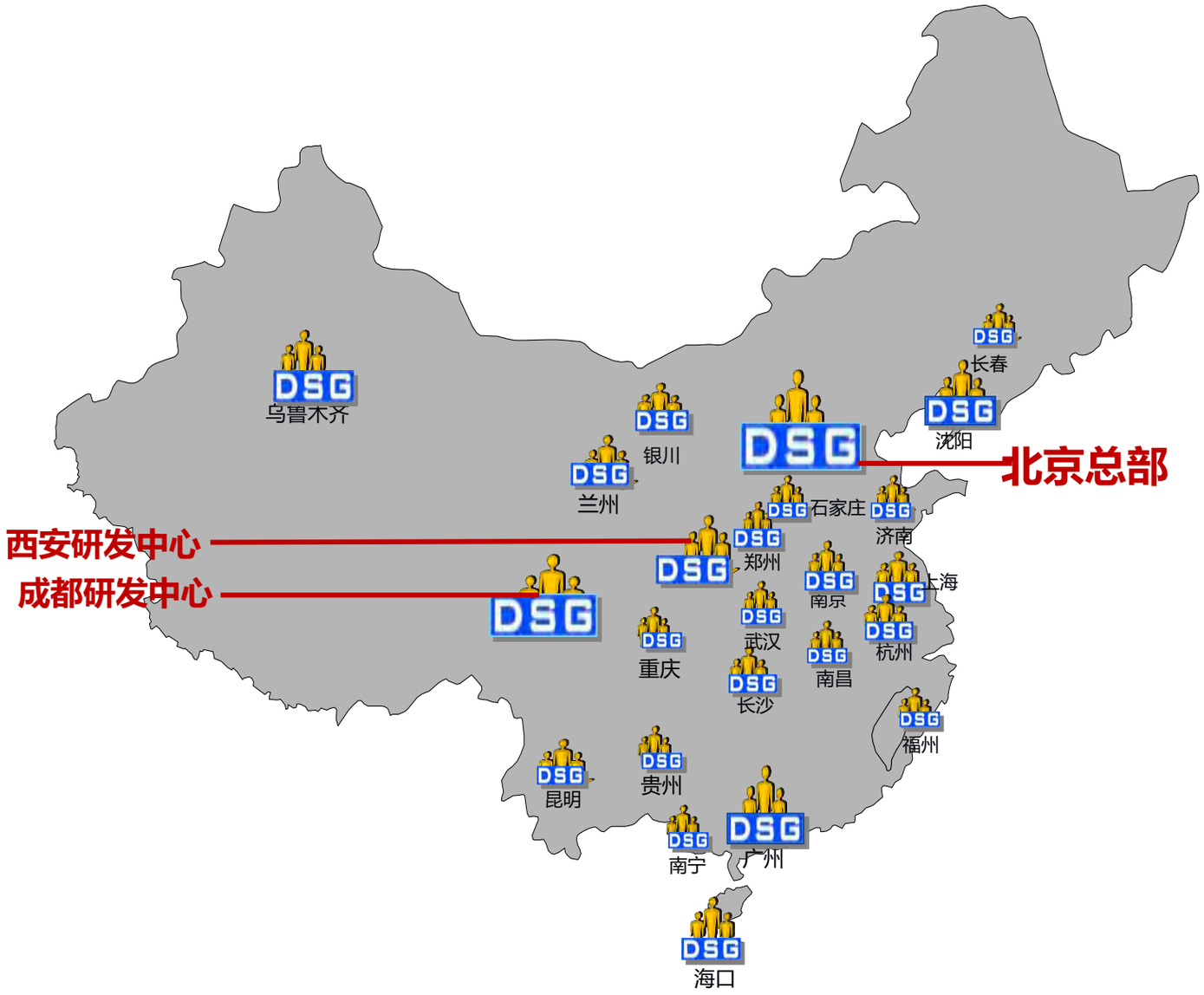
DSG
information management expert

大数据时代的DSG复制云

DSG公司创始人、总裁：韩宏坤

目录 / CONTENTS

DSG的现状：



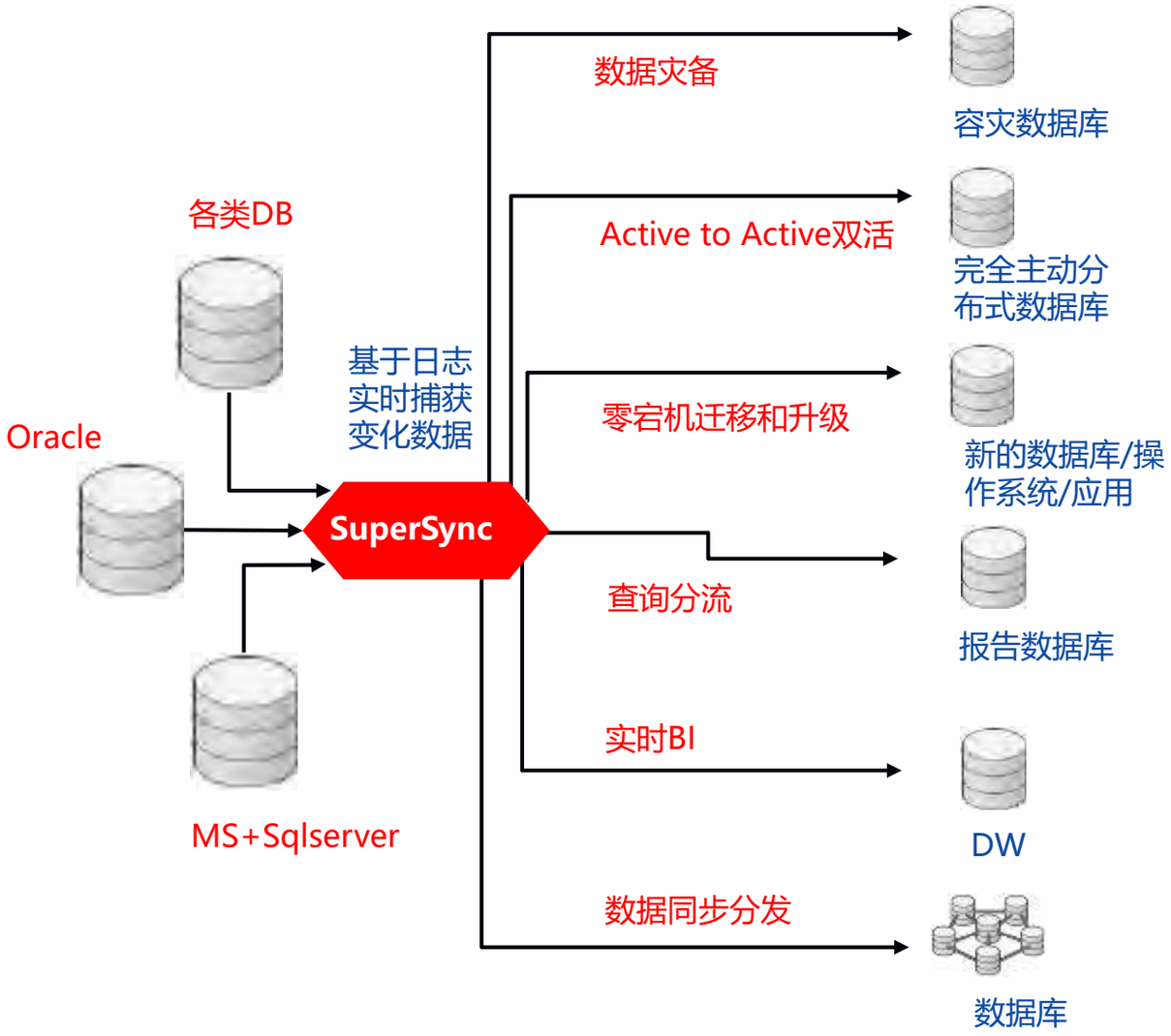
- **3**： 3家公司成立
- **240**： 员工人数约
- **1**： 总部北京
- **3**： 北京、成都、西安研发中心
- **25**： 25个省市设有办事处或者技术支持中心
- **800**：拥有电信、金融、政府等800多个大型客户
- **1.5亿**：2017年10月，资产为1.6亿

DSG的历史：最强的Oracle备份

产品名称		灾备级别	功能特点	厂家
DBP 容备云平台		高效备份、应用测试和数据迁移	• 支持BDMP BS、BDMP DC一体机所含有的所有软件功能 • 支持大型Oracle数据库统一灾备、测试迁移一体化： ➢ 高性能数据库首次全同步及准实时同步 ➢ 容灾端在线测试及报表查询 ➢ 数据库同平台迁移	DSG
一体机	BDMP-DC	数据中心级全数据云灾备管理一体机	• 支持BDMP BS 一体机功能 • 支持Oracle高性能备份 ➢ 智能全备 ➢ 非归档备份 ➢ 直接表恢复 ➢ 备份数据验证 ➢ 端到端数据流压缩与存储	DSG
	BDMP-BS	企业级备份容灾一体机	• 硬件一体机 • 传统数据库备份 • VMwareESXi、HyperV虚拟机备份 • CDP实时备份 • P-V、V-V实时容灾	DSG

DSG的历史：最强的Oracle复制，800多用户

DSG information management expert®



功能优势

- 全球独家支持：Rowid复制
- 支持源/目标安装、或第三方静默安装
- 支持UDT和IOT
- 高速首次全同步
- 数据一致性检查、高效修复
- 双活容灾
- 数据库异构迁移
- 统一监控管理

性能优势

- 日志并发分析、多次使用，提高分析效率、实现最高150GB/小时日志分析、实时增量日志可处理3Tbit/天以上
- 首次同步技术可实现200G-800G/小时
- 分钟级容灾切换、容灾切换时间小于5分钟
- 系统干扰小于5%
- 能分钟内高速处理数千DDL操作

DSG的现在：


DSG
产品体系
备份容灾事业部
BDMP BS 、BDMP DC、DBP
新概念备份、云备份、云容灾
、单产品监控、统一监控
复制容灾共享
SuperSync应用级双活容灾
Enhanced ETL、数据脱敏、大数据交换平台
统一监控
DSG服务和运维
NDMP容灾导航管理+数据迁移
数据库运维 + DB monitor+数据库审计
智能运维
大数据和应用开发
大数据应用开发平台
行业大数据方案
兆源数通
基于智能识别的安防
不良信息识别+商业信息识别
爬虫+深度舆情……

DSG
information management expert®

目录 / CONTENTS

数据库数据与大数据交换技术趋势

DSG information management expert®

1 人工汇聚数据

通过人工数据查询，将数据个性化发给不同用户

效率低、成本高、干扰系统、时效差

2 应用软件定时采集汇聚发送

增加自动化，减少了人工

开发成本高、对数据库影响较大、不灵活

跨应用、跨用户之间难以协调、成本高

3 数据库采集前置一体机

典型技术采用ODBC / JDBC技术

性能低、高干扰源端数据库、实时增量差、

不支持DDL，异构数据库、异构平台难

不支持实时转换，需要借助第三方ETL数据转换

4 数据库数据实时转换共享

代表产品DSG的Enhanced ETL

高性能、1-2%的低干扰，单数据库间实现一小时数百GB、支持实时增量

支持DDL、异构数据库、异构硬件平台

支持实时转换

5 实时大数据交换平台

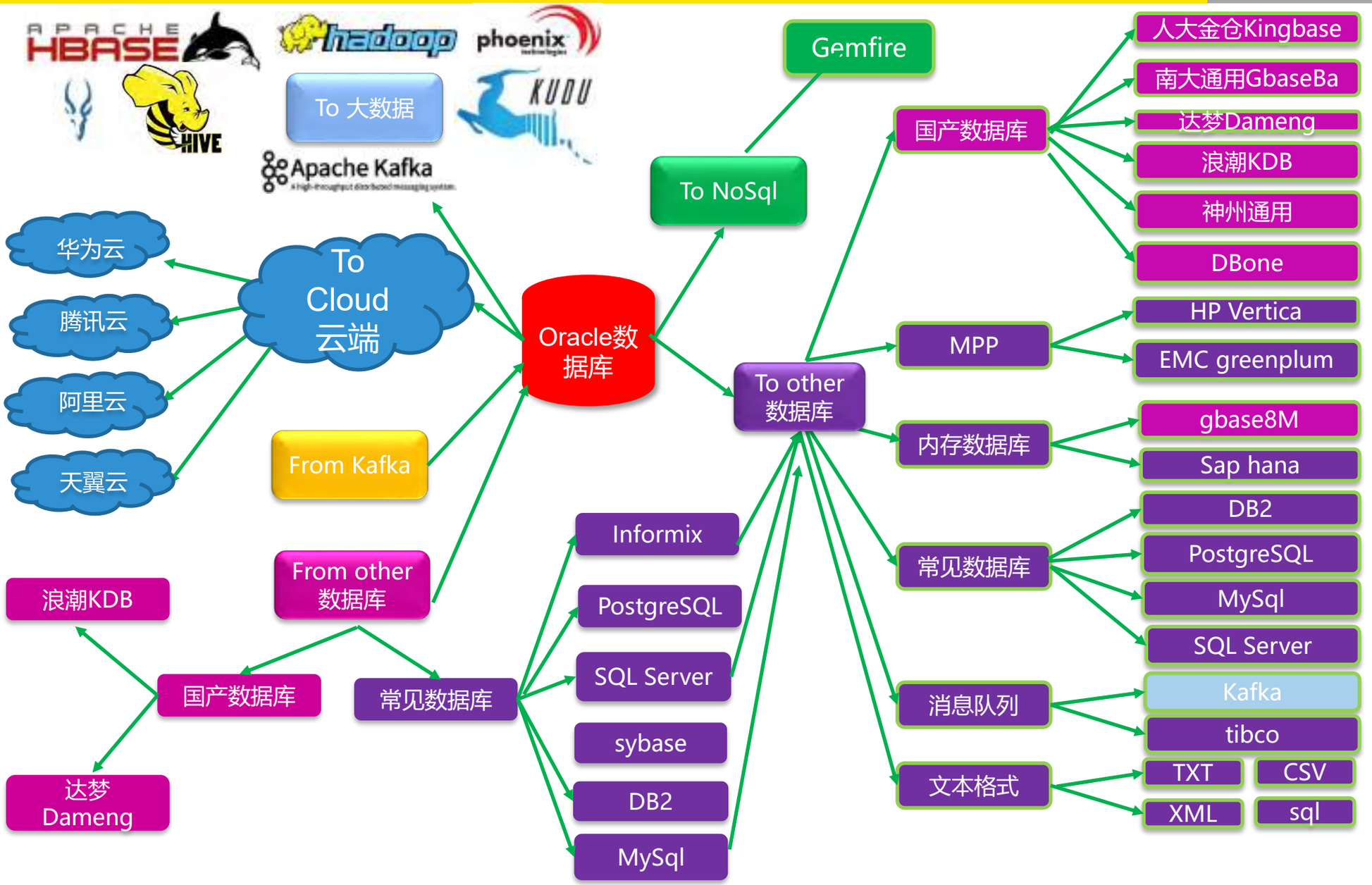
将数据库实时交换共享技术，与各类大数据技术集成开发，实现新一代大数据交换PaaS

支持实时采集、变换

支持可视化拖拉拽

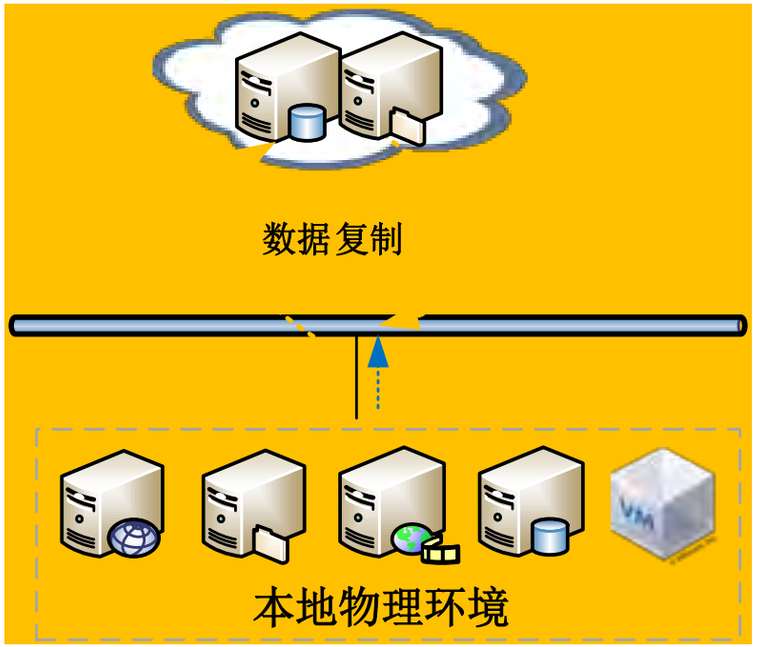
支持多租户虚拟管理等

DSG已经实现的：数据实时交换与共享支持系统框架



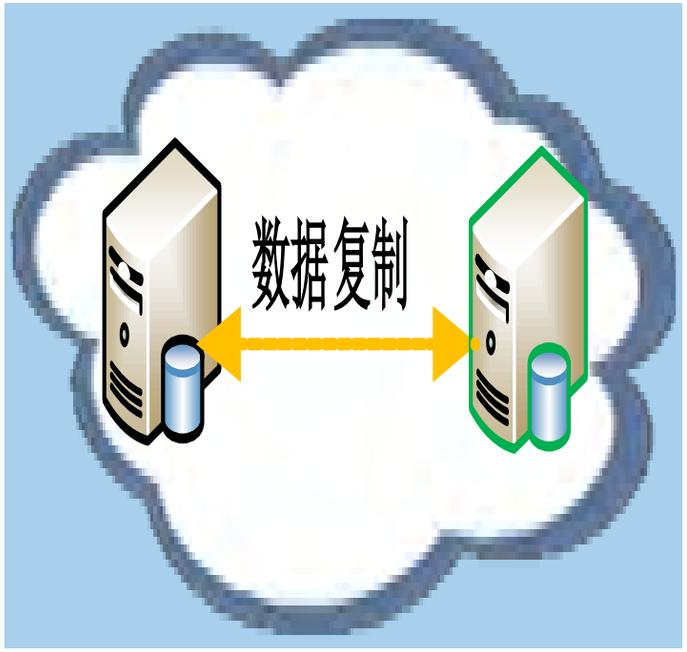
DSG:多种云和云间复制场景

云上云下间数据灾备



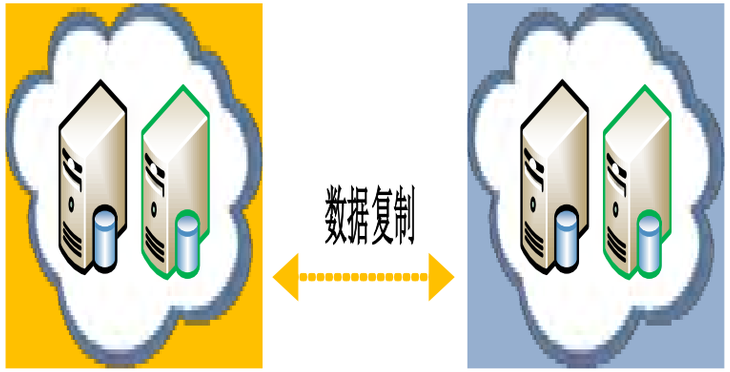
物理机灾备至天翼云
物理机灾备至政务云

同种云内数据灾备



天翼云灾备至天翼云
政务灾备至政务云

异种云之间灾备



天翼云灾备至政务云
政务云灾备至天翼云

衍生的三大新方向：

全球竞争力：DSG的数据库复制技术具有全球竞争力，与全球竞争对手 OGG、Quest、IBM CDC 、Informatic等知名厂家相比有显著的技术优势：高性能、多种DDL支持、多数据库和大数据之间环境支持。

国产数据库支持：DSG的技术是国内自主知识产权，对国产数据库的支持也最多，并远优于国外对手。

国内高端用户普遍使用：近800家，如建行、太平洋保险、中国移动、联通、电信、政府公安、证券等。

- **应用级双活容灾：**实现Unix、Linux、windows 环境下的应用级双活容灾、支持异构硬件环境、跨数据库、容灾端数据库可用

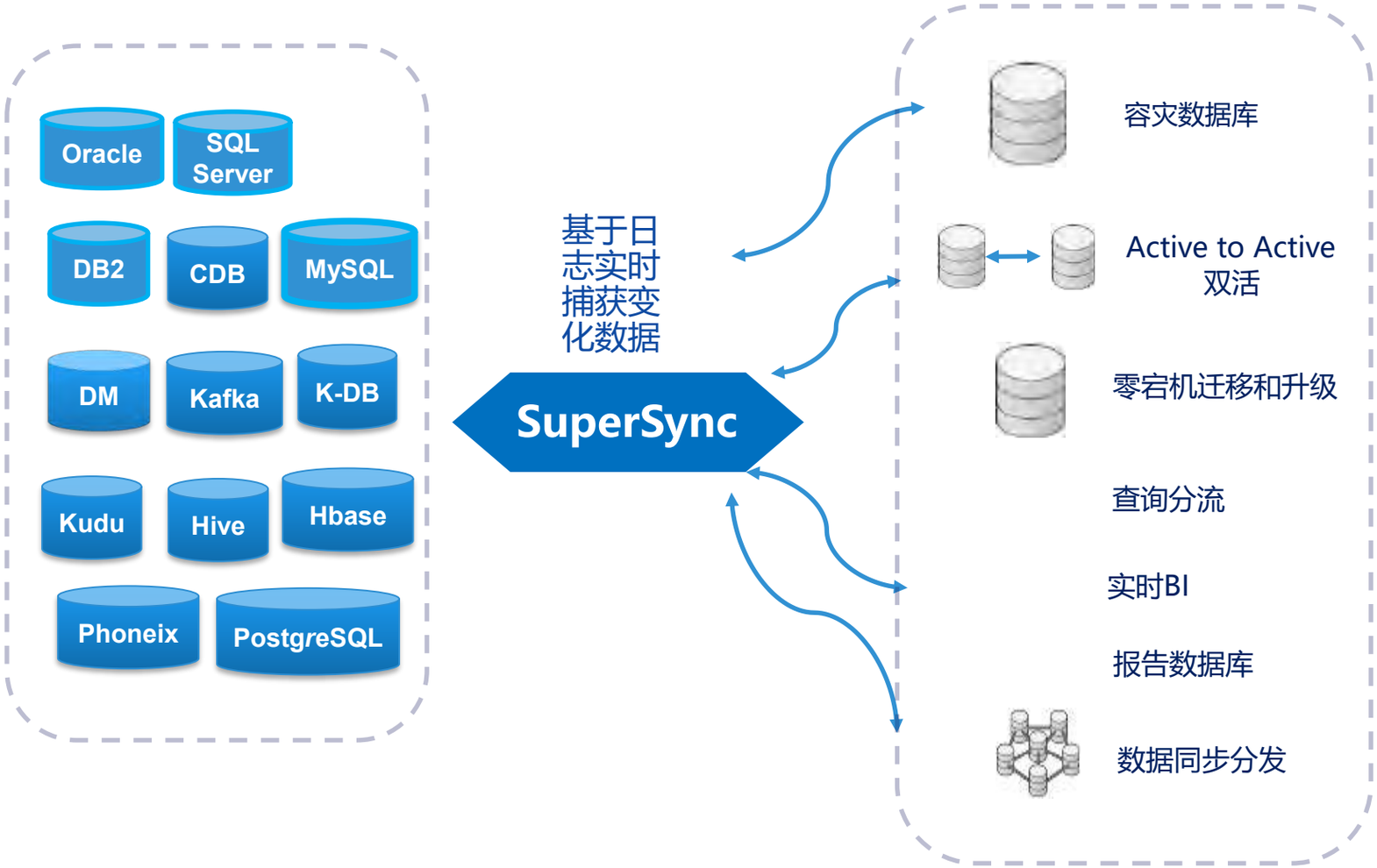
- **数据库实时共享交换：**实现Unix、Linux、windows 环境下异构数据库之间数据实时复制、实时转换、实时交换和共享

- **大数据实时共享交换平台：**实现Unix、Linux、windows 环境与大数据、互联网数据之间实时复制、转换和交换，是PaaS平台，支持多租户等

1、DSG SuperSync数据同步产品

- 标准化的单一技术解决多种需求
- 保证业务的连续性
- 同时满足实时数据共享和灵活的数据结构变换

- 低影响、高效率
- 跨平台、低成本，投资回报率高
- 快速部署、一体化的监控管理。



SuperSync 的优势

DSG information management expert®

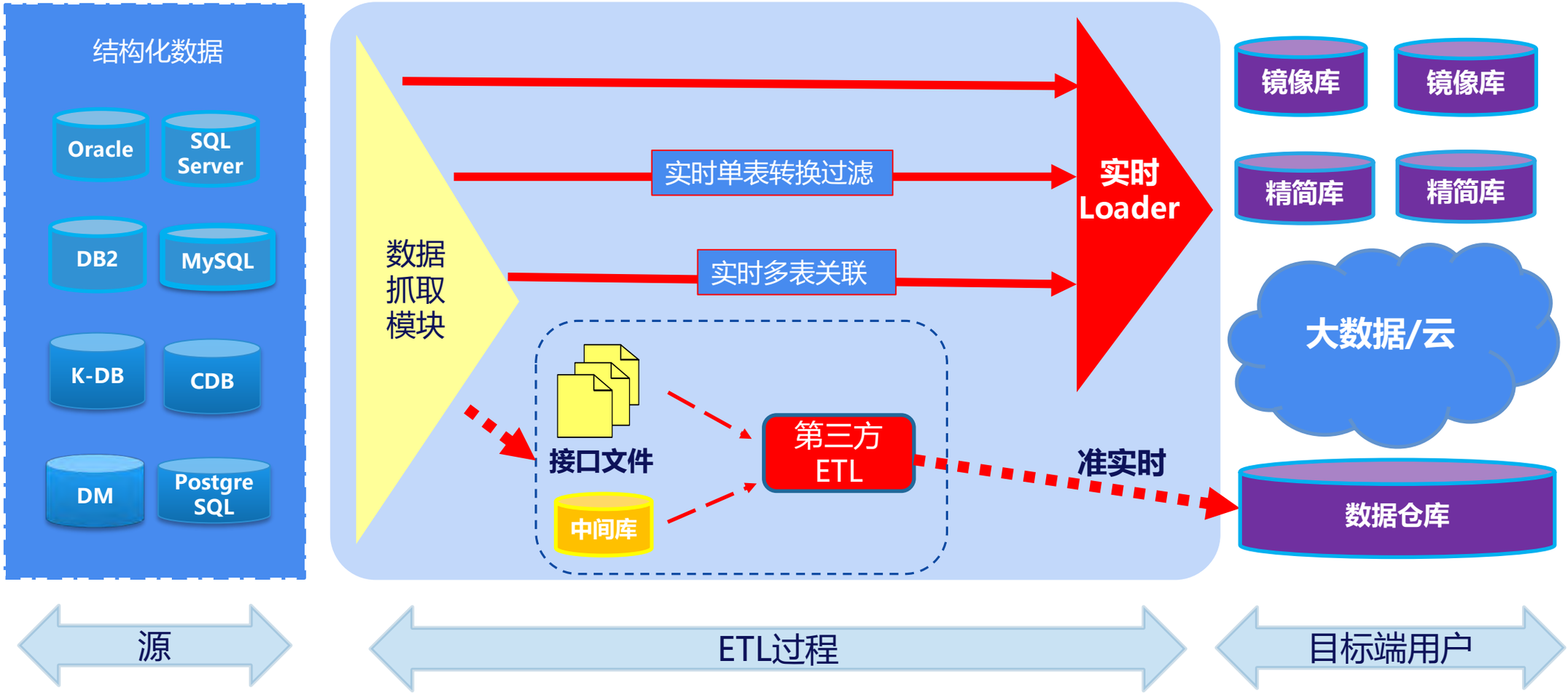
功能优势

- 全球独家支持：Rowid、PK/UK等复制
- 全球独家支持源/目标安装、或第三方静默安装
- 支持UDT和IOT
- 高速首次全同步
- 数据一致性检查、高效修复
- 双活容灾
- 数据库异构迁移
- 统一监控管理

性能优势

- 日志并发分析、多次使用，提高分析效率、实现最高200GB/小时日志分析、实时增量日志可处理3Tbit/天以上
- 首次同步技术可实现200G-800G/小时
- 分钟级容灾切换、容灾切换时间小于5分钟
- 系统干扰小于5%
- 能分钟内高速处理数千DDL操作

2、DSG Enhanced ETL数据库实时交换与共享平台



Enhanced ETL 的功能

功能	说明
数据实时抽取	• 通过源系统端的Agent进程对数据库Log日志进行实时分析，获取交易指令 • 将交易指令和交易数据经过格式转化生成数据格式；过滤转化为与生产应用相吻合的指令 • 实时传输到目标端系统
数据实时转换	• 复制指定的数据、表、列 • 支持数据集中，即多个相同结构的数据库中将数据整合到一个库中 • 同类的数据项集合放到一个表中 • 支持数据分发
实时存储和增量变化通知	• Agent将识别到的实时增量数据发送到中间数据库 • 在此中间库中维护一张和生产系统对应的数据表； • 对数据进行整合、过滤和判断后通知订阅方
支持ETL实现准实时数据抽取	• 支持增量抽取间隔到每几秒钟、几分钟、10分钟生成一个接口文件 • 支持从镜像库中获取数据
QETL	• 支持多表关联同步 • 只复制到多表关联结果集到目标端 • 支持复杂的sql模式 • 支持多种同步维护模式 • 保持分析日志模式而非sql查询模式

Enhanced ETL 的优势

DSG information management expert®

实时性

采用数据跟踪和push技术，安装在数据源端的数据变化跟踪程序Agent实时跟踪数据变化，将变化实时发送到数据需求端，数据可在秒级实现共享

可配置性

通过参数配置来定义需要共享的数据表(table)、共享的数据项（字段）、共享的条件（满足条件的记录）

低干扰性

通过数据库自身信息获取源系统上的改变并传送给目的系统，降低对生产系统性能影响

无需全量数据处理，只抽取增量数据，减少数据库存储压力

灵活性

提供选择表、选择字段和选择记录的复制

提供数据转换，如字段名映射、数据类型转换、数据运算等

支持中间库、定制文件到ETL的准实时转换

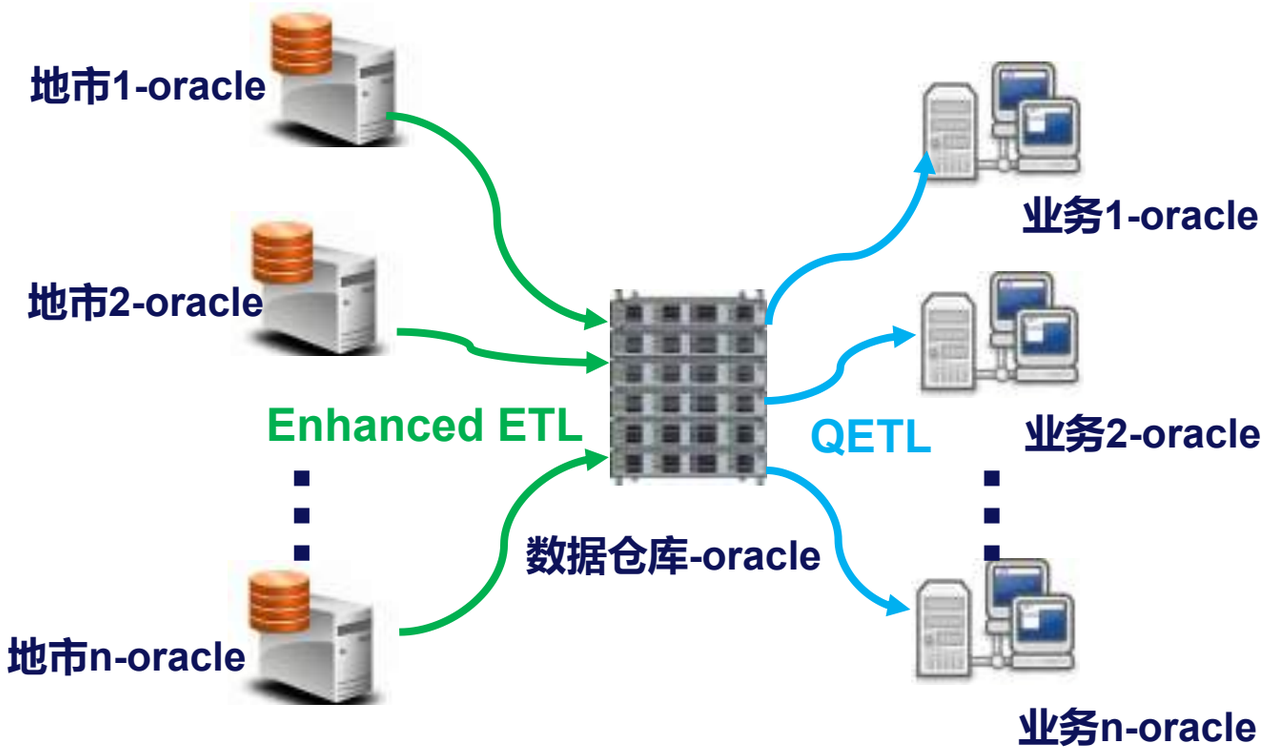
多样性

支持多种复制系统拓扑结构

Enhanced ETL 的典型应用

多表关联实时高性能复制场景

传统同步一般会采用链接、视图、触发器等方案解决，占用较大资源，还不能满足实时要求。传统ETL一般将全量数据、部分数据或者组合推送至目标端，由于推送的数据包含大量不需要的数据，不高效；特别涉及多个表的数据量大，而所需数据量又较小的情况下，非常适合只推送SQL语句查询结果集，以实表的情况存放在目标端实时使用。



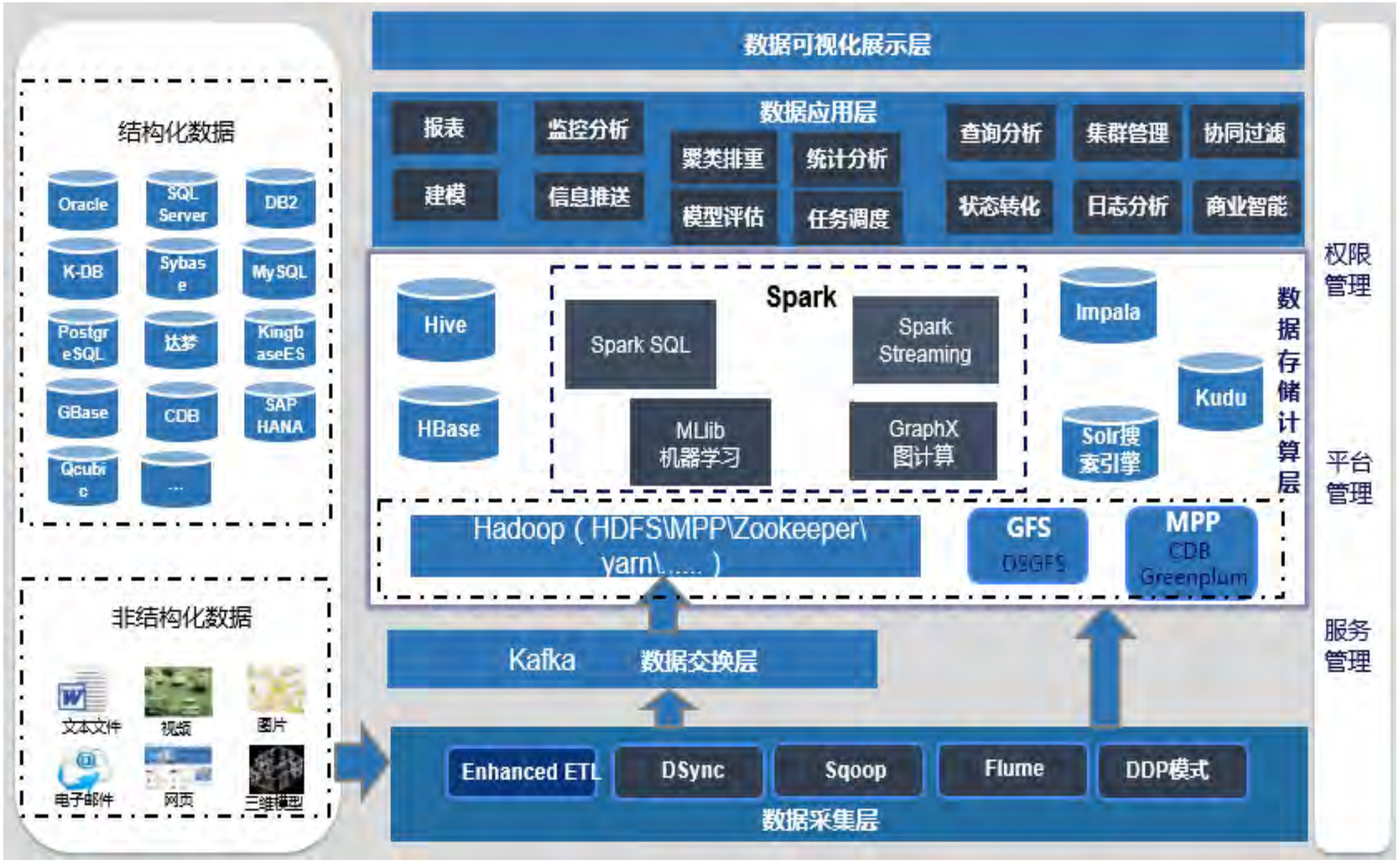
案例介绍

湖北某券商客户使用Enhanced ETL 将各地市的数据集中复制到数据仓库，然后使用QETL对数据实时导出并进行组合运算，将转换后的数据集实时同步到各业务库。

案例功能特点

- 源端为组合视图，目标端直接同步的结果集为表中；
- 目标数据涉及多个用户下的多张表，QETL只关注客户所需数据，最小化同步范围，节省系统资源；
- 只跟踪业务所需字段，避免了敏感字段数据的向下传递；
- 某行业某块业务所需的数据，业务上仅需要通过一条SQL从多张分别拥有百万和几亿条的数据表中提取所需结果，该结果返回仅有几十条数据，传统的推送方式要么达不到实时要求，要么占用较大资源，现通过QETL 实时的推送所需的几十条数据的结果集，简化了同步方式、大大提高了同步效率，深受客户好评，并将大量推广应用。

3、DataONE大数据应用开发统一平台（开放平台）



DataONE总体架构含：

1. 平台的统一流程调度管理
2. 数据交换分发
3. 数据海量存储
4. 数据查询
5. 数据容灾
6. 数据ETL转换
7. 数据分析
8. 数据挖掘
9. 数据应用与展示层

目录 / CONTENTS

发展史

起步期（2010-2014年）

关注平台是否可以稳定运行，多用于某类特定场景（详单查询、上网日志分析、银行交易查询与分析）；

平台开放期（2014-2016年）

平台和数据开放、多应用提供商、多用户访问，关注平台安全、数据安全，目前国内企业级市场多数在此阶段；

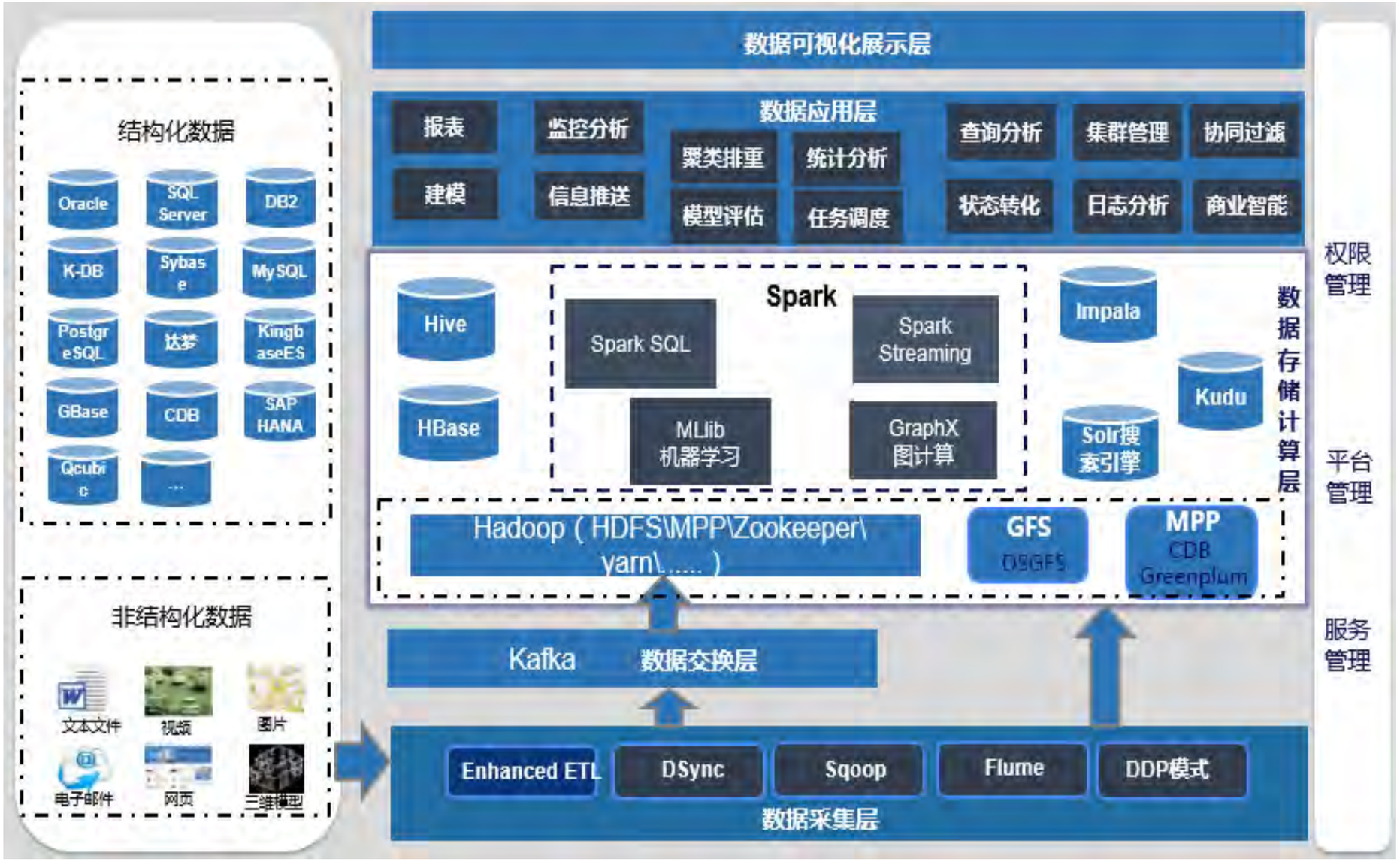
应用扩展期（2014-2016年）

提升数据的分析/数据ETL处理能力(统一的SQL处理能力和机器学习能力的支持)，进一步降低TCO、稳定性和大集群运维能力，目前的优秀技术实施处于这一阶段；

第4代Hadoop技术兴起（2016-2018年）

DSG公司从2012年起关注和研究Hadoop为代表的大数据技术，我们认为随着以往IOE架构被重新定义

DataONE大数据应用开发统一平台



DataONE总体架构含：

1. 平台的统一流程调度管理
2. 数据交换分发
3. 数据海量存储
4. 数据查询
5. 数据容灾
6. 数据ETL转换
7. 数据分析
8. 数据挖掘
9. 数据应用与展示层

DataONE的定位

DataONE功能特点

DataOne大数据应用开发平台提供了完整功能管理和流程管理大数据综合处理平台；

1. 用户可在DataOne平台上通过托拉拽实现以下功能：

- **大数据采集**
- **数据转换加工**
- **数据治理，建模**
- **分析**
- **容灾**
- **查询**
- **数据挖掘**
- **图形化展示**
- **流程管理和实时实时数据监控**

…..等一系列复杂的流程处理操作。

2. DataOne是PaaS开放平台:提供 了良好的接口，支持其他第三方平台的算法功能接入DataOne；在统一平台中，完成大数据的处理工作。

- 面向结构化和非结构实时、批量、PB级数据处理
- 实现百亿级数据的秒级查询
- 开放架构，已经支持近百种统计、分析、和挖掘算法，还可以加入第三方算法
- 集成主流图形化展示框架，多种图形化可选
- 全球领先的实时数据采集与数据转换和交换分发平台
- 开放PaaS平台，兼容用户现有Hadoop大数据系统，支持基于华为、CDH、星环等目前主流Hadoop平台，支持Hadoop开源1.6-2.8、Solr、Hbase、Hive、Spark1.1-2.1、Kudu、Flume、Sqoop、Kafka等大数据组件产品
- 该平台支持用户个性化二次开发大数据应用，降低用户投资成本，提高数据处理、使用效率

DataONE组件与部署

大数据应用开发平台	大数据平台数据调度组件
	统一权限认证组件
	大数据采集组件
	ETL组件
	数据存储组件
	数据容灾组件
	数据查询组件
	数据挖掘组件
	数据统计分析组件
	图形化展示组件

部署方式可以分为：

- 单一集群模式
- 统一集群模式
- 混合集群模式

图 多集群混合架构

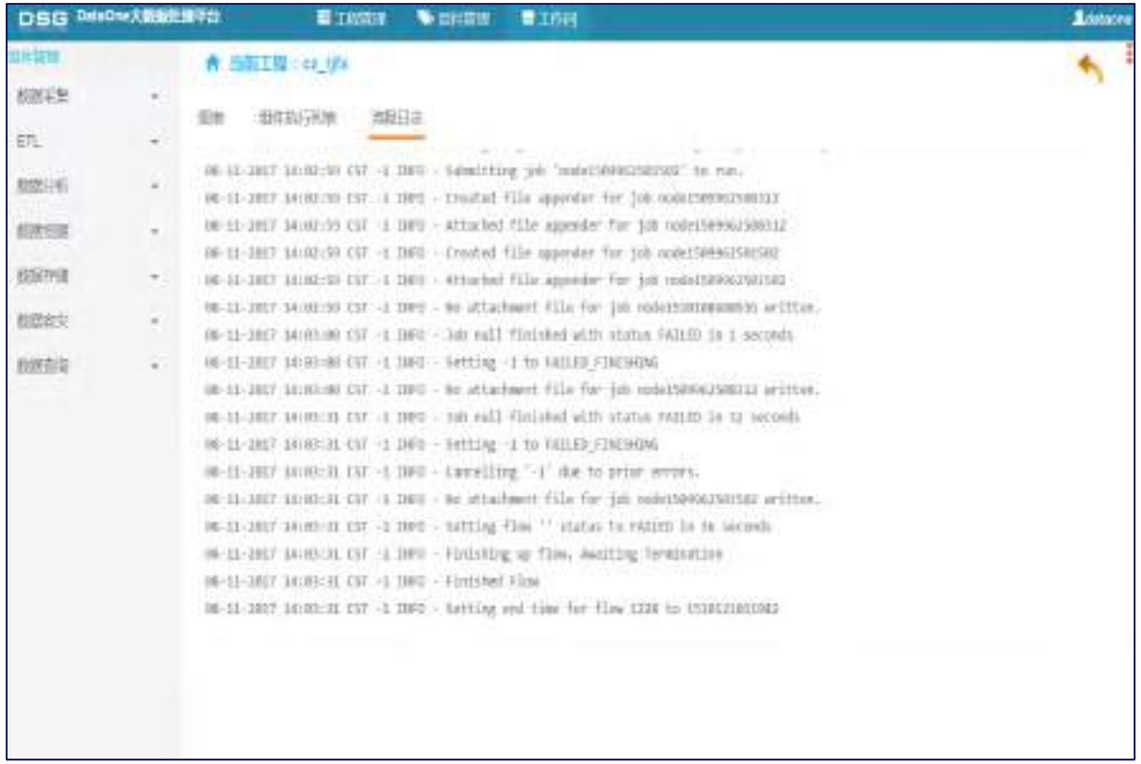
DataONE组件与可视化操作特点

特点

- 功能组件支持拖拉拽操作，直接对组件配置参数信息，通过连接线设置任务之间的关系；
- 组件可单独执行，也可以定时执行；
- 跟踪查看整个流程的执行情况和每个功能组件的执行进度；
- 组件可复用，并且支持第三方平台以REST标准接口接入；模块化的和可插拔的功能组件机制，具有良好的扩展和兼容性；

DataONE组件特点

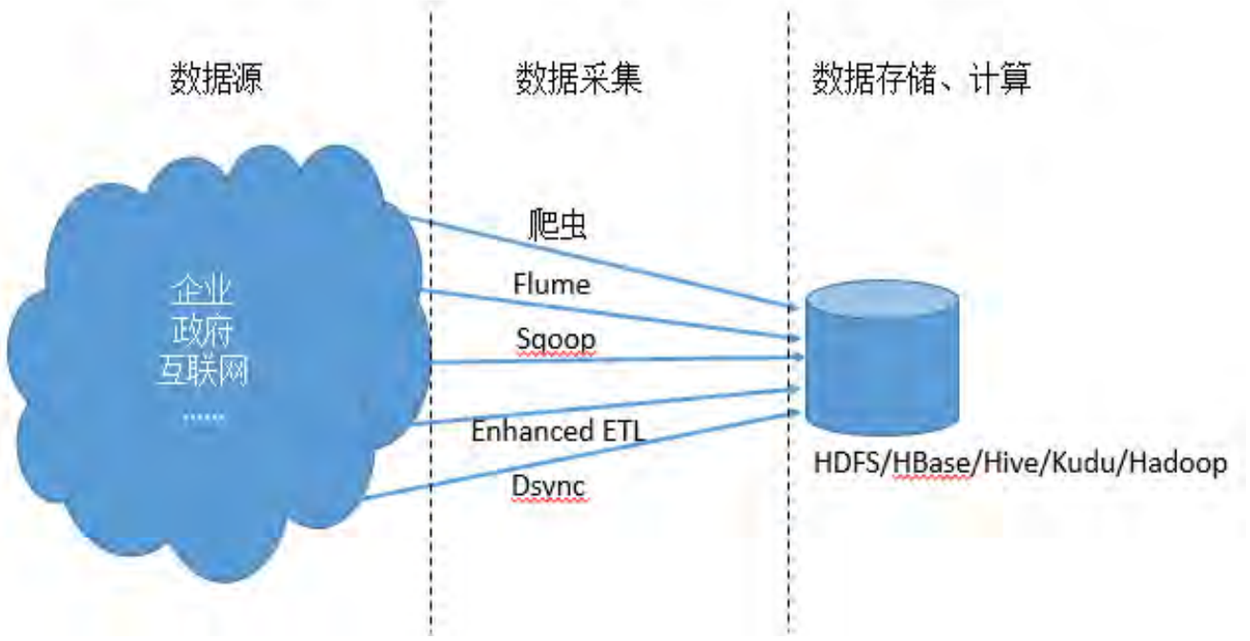
- 查看每个组件在流程执行中的进度、结果信息



- 查看每个组件在流程执行中的日志信息
- 查看整个工作流程的执行日志信息

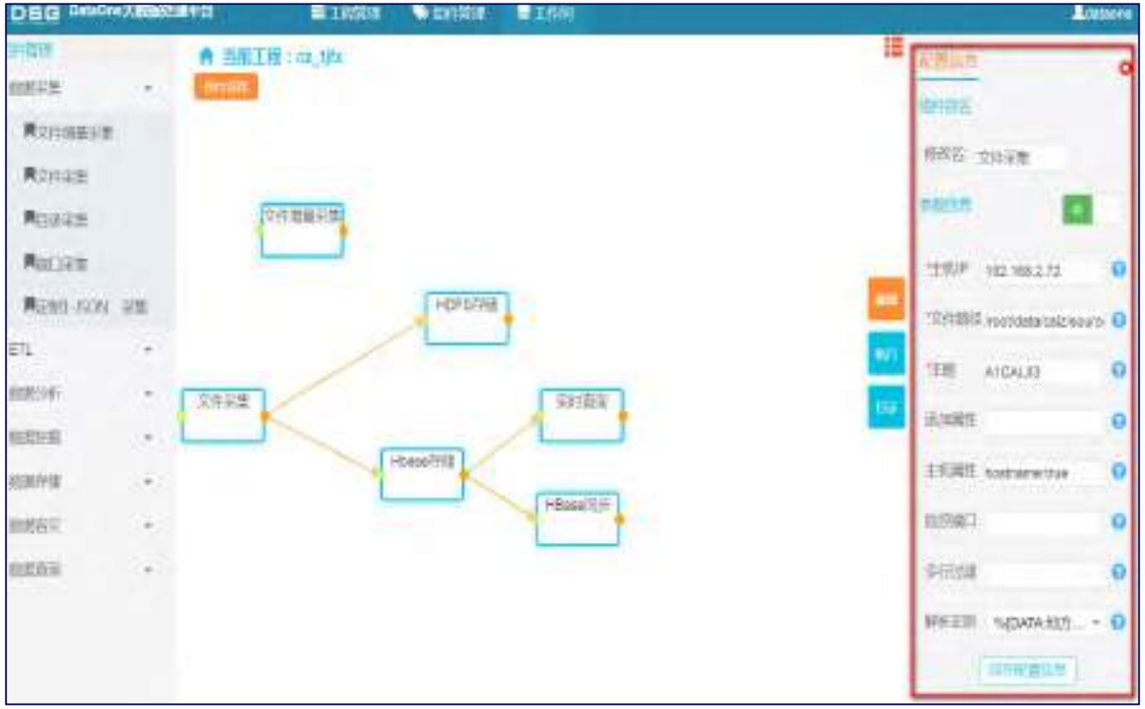
大数据采集与数据转换治理

基于全球领先的实时数据库复制技术

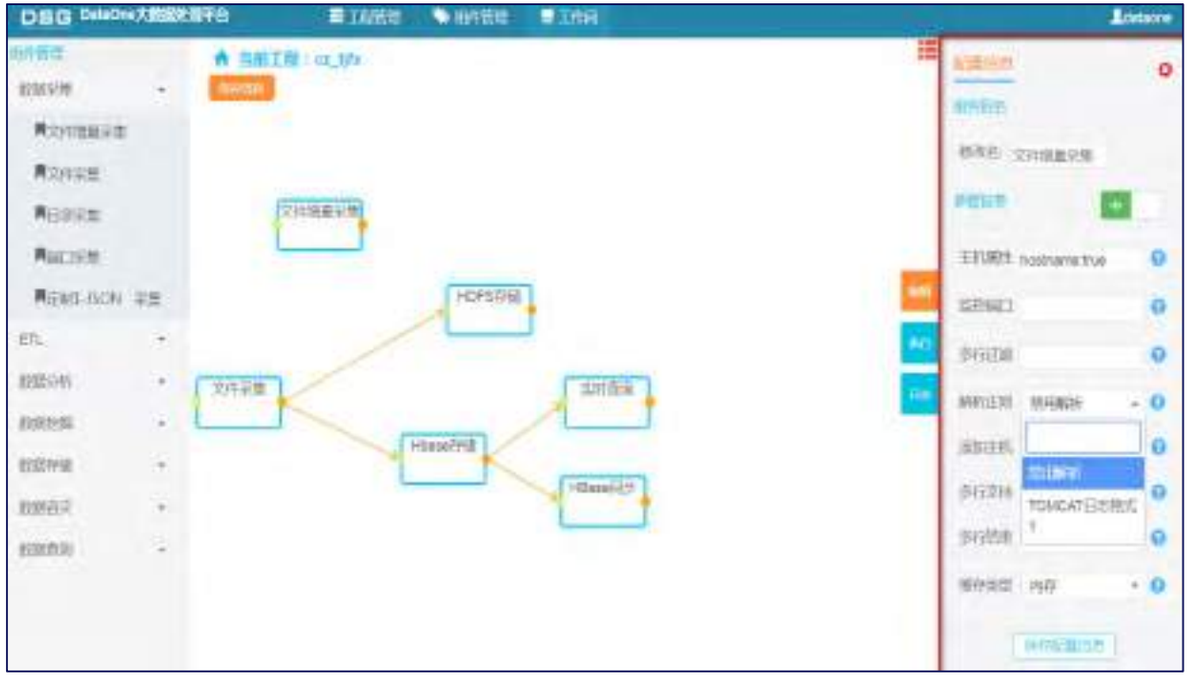


采集方式	说明
爬虫	• 从一个或多个指定初始网页的地址URL开始，读取URL列表，建立初始待爬队列； • 开始顺序爬信息，并不断在从当前正在爬取的网页上抽取新的URL放入带爬取队列；直到爬完为止。
Flume/ Logstash	• 分布式、可靠、高可用的海量日志采集、聚合和传输的系统，支持在系统中定制各类数据发送方，用于收集数据；同时，提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力
Sqoop	• 即 SQL to Hadoop ，在传统型数据库（如 Oracle、MySQL、DB2、PostgreSQL等）与Hadoop之间进行数据迁移的工具，充分利用MapReduce并行特点以批处理的方式加快数据传输。支持关系型数据库和Hive、HDFS，HBase之间数据的相互导入。
Enhanced ETL	• DSG公司一款革命性的数据共享产品，在异构数据库/文件系统之间实时、准实时交换数据的统一管理平台，实现跨平台、跨数据库、跨系统之间的批量数据同步以及实时增量数据同步；实现数据库到大数据平台之间的数据实时交换

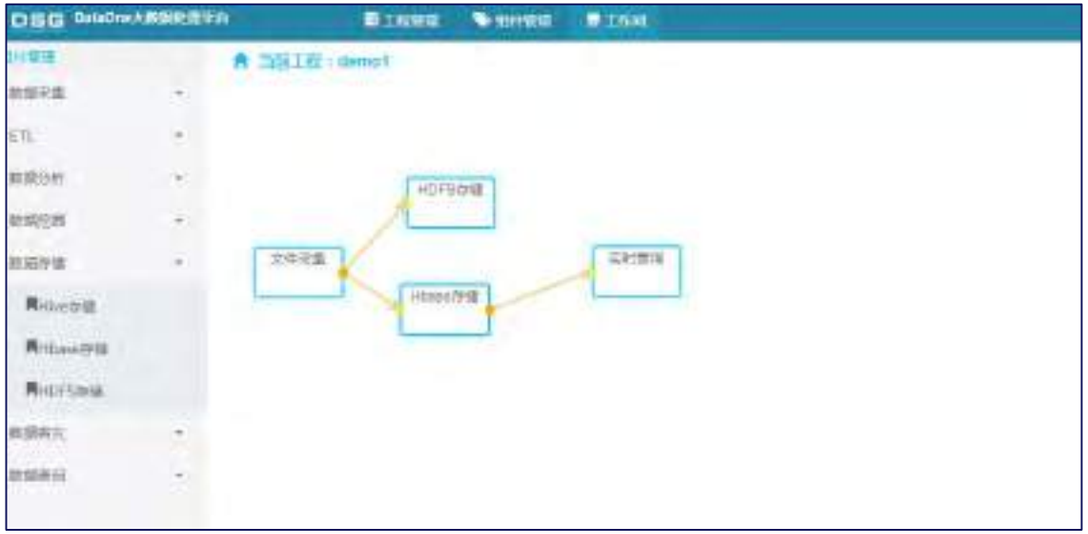
DataONE组件---数据采集



- 按数据类型分为结构化数据采集和非结构化数据采集
- 按数据量分为全量采集和增量采集
- 按采集方式分为文件采集，目录采集，端口采集

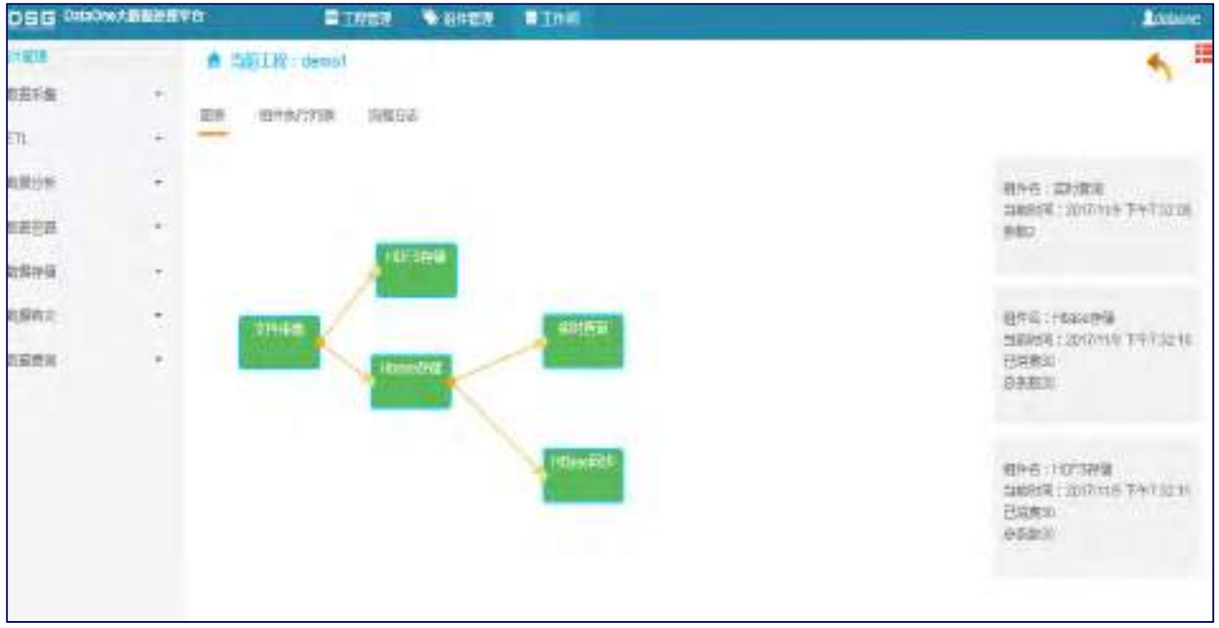


DataONE组件---数据存储



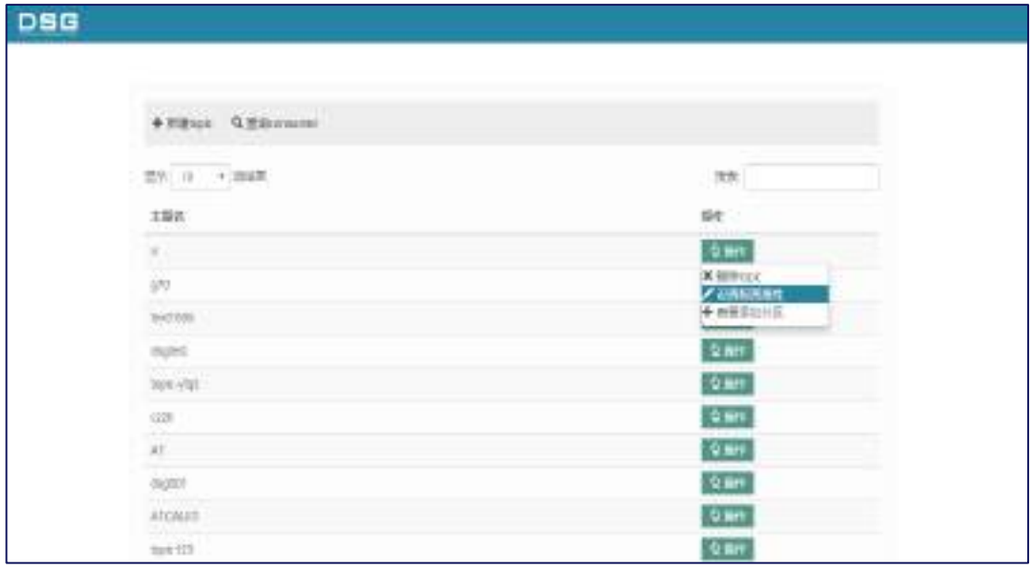
- 支持采集并缓存在Kafka和存入HIVE，HBASE、HDFS
- 支持在数据存入HBase后做实时数据查询(使用实时查询组件)
- 支持HBASE表实时同步到其他集群(HBase实时容灾组件)
- 支持在数据存入HIVE后做DSG报表自助分析(数据分析组件)
- 支持在数据存入HDFS后做数据挖掘，建立模型，使用模型进行预测(数据挖掘组件)。

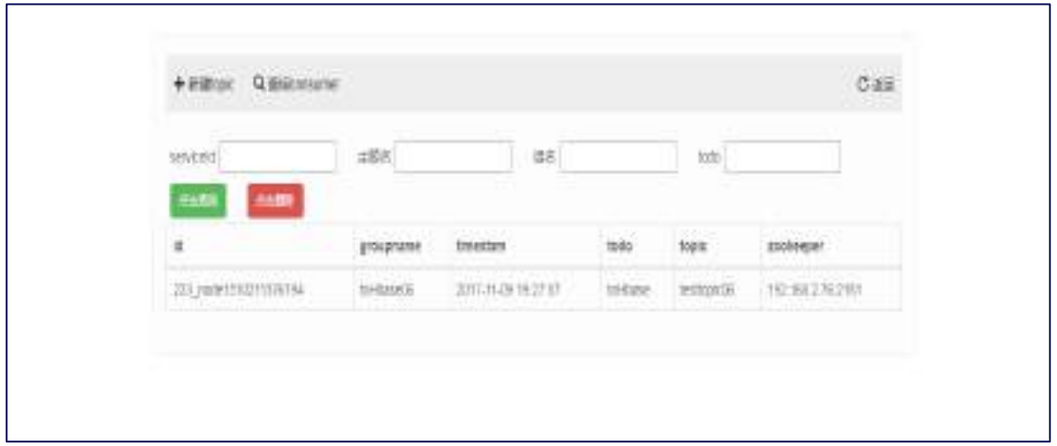
支持对Kafka中的数据进行监控，查看已采集并缓存在Kafka中的数据量和已存入的Hive的数据量



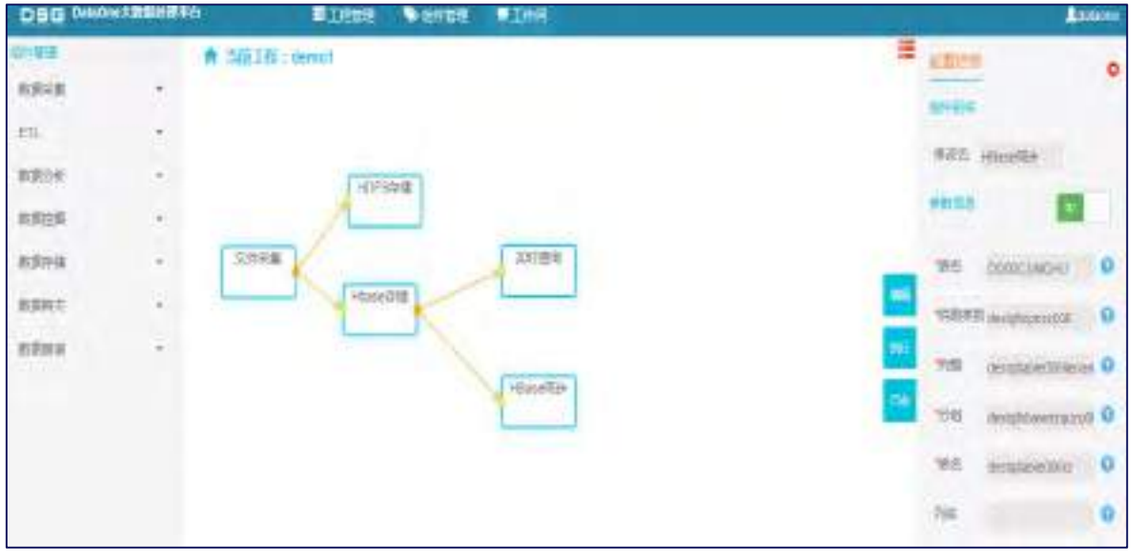
DataONE组件---数据存储

支持对Kafka进行管理。如：删除topic,设置分区属性，查看添加分区，管理consumer等





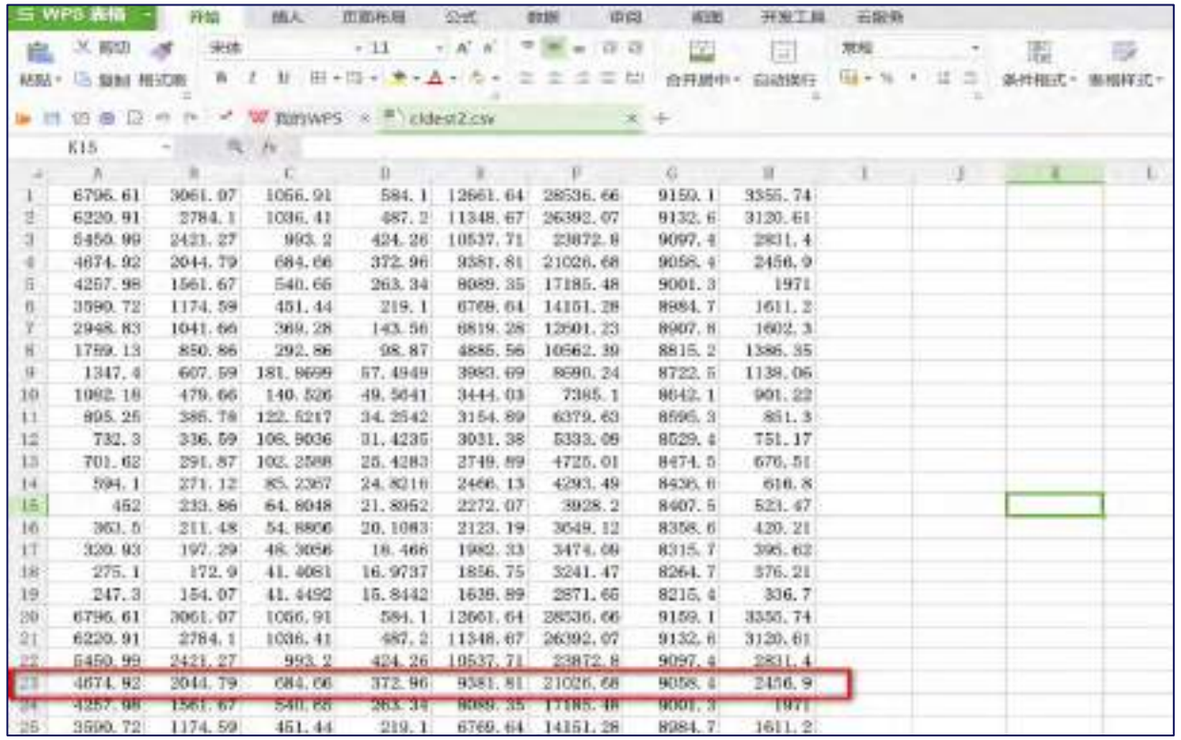
支持HBase表实时（和离线）复制到另一套HBase集群中



DataONE组件---数据查询

特点

- 支持离线数据查询
- 支持在线实时数据查询



	A	B	C	D	E	F	G	H
1	6796.61	3061.07	1056.91	584.1	12661.64	28536.66	9159.1	3355.74
2	6220.91	2784.1	1036.41	487.2	11348.67	26392.07	9132.6	3120.61
3	5450.99	2421.27	993.2	424.26	10537.71	23872.8	9097.4	2831.4
4	4674.92	2044.79	684.66	372.96	9381.81	21026.68	9058.4	2456.9
5	4257.98	1561.67	540.65	263.34	8089.35	17185.48	9001.3	1971
6	3590.72	1174.59	451.44	219.1	6769.64	14151.28	8984.7	1611.2
7	2948.83	1041.66	369.28	143.56	6819.28	12601.23	8907.8	1602.3
8	1759.13	850.86	292.86	98.87	4885.56	10562.39	8815.2	1386.35
9	1347.4	607.59	181.9699	57.4949	3983.69	8690.24	8722.5	1138.06
10	1092.18	479.66	140.526	49.5641	3444.03	7385.1	8642.1	901.22
11	895.25	386.78	122.5217	34.2542	3154.89	6379.63	8595.3	851.3
12	732.3	336.59	108.9036	31.4235	3031.38	5333.09	8529.4	751.17
13	701.62	291.87	102.2588	25.4283	2749.89	4725.01	8474.5	676.51
14	594.1	271.12	85.2367	24.8216	2466.13	4293.49	8436.6	616.8
15	452	233.86	64.8048	21.8952	2272.07	3928.2	8407.5	521.47
16	363.5	211.48	54.8806	20.1083	2123.19	3649.12	8358.6	420.21
17	320.93	197.29	48.3056	18.466	1982.33	3474.09	8315.7	395.62
18	275.1	172.9	41.4081	16.9737	1856.75	3241.47	8264.7	376.21
19	247.3	154.07	41.4492	15.8442	1638.89	2871.66	8215.4	336.7
20	6796.61	3061.07	1056.91	584.1	12661.64	28536.66	9159.1	3355.74
21	6220.91	2784.1	1036.41	487.2	11348.67	26392.07	9132.6	3120.61
22	5450.99	2421.27	993.2	424.26	10537.71	23872.8	9097.4	2831.4
23	4674.92	2044.79	684.66	372.96	9381.81	21026.68	9058.4	2456.9
24	4257.98	1561.67	540.65	263.34	8089.35	17185.48	9001.3	1971
25	3590.72	1174.59	451.44	219.1	6769.64	14151.28	8984.7	1611.2

- 实时查询首先通过DataOne的采集组件，实时将数据存入DataOne平台后，查询组件会自动建立索引，将数据从HBase中的数据实时缓存进Solr中，完成在线实时检索。
- 适用于海量数据的实时在线查询或者监控用。特点是条件过滤查询效率高，速度快，使用了数据同步，所以对原业务影响小。可以实现百亿级秒级查询

DataONE组件---数据ETL处理及举例

月份	地区	房地产投资	住宅	result
8	北京	2449.31	1143.2	3592.51
8	天津	1335.99	882.14	2218.13
8	河北	2677.18	1948.54	4625.72
8	辽宁	2975.13	2187.72	5162.85
8	上海	2085.67	1059.2	3144.87
8	江苏	5493.04	4082.32	9575.36

« ‹ 1 / 26 › » [1 - 6 / 155]

属性构造

属性构造是对两列数值类型的数据进行加减乘除的操作，并得到新的一列结果，可删除原有的列；

重新编码

1.对输入列进行重新编码
2.输出经过编码操作后附加编码列的数据记录

重新编码后的处理结果：

ber_no	gender	work_country	age	work_country_indexed
993	男	CN	31	0
065	男	CN	42	0
106	男	CN	40	0
189	男	US	46	1
546	男	CN	48	0
972	男	CN	64	0

« ‹ 1 / 167 › » [1 - 6 / 1000]

DataONE组件---数据ETL举例

月份	地区	房地产投资	住宅	新增列
8	北京	2449.31	1143.2	0
8	天津	1335.99	882.14	1
8	河北	2677.18	1948.54	2
8	辽宁	2975.13	2187.72	3
8	上海	2085.67	1059.2	4
8	江苏	5493.04	4082.32	5

« ‹ 1 / 26 › » [1 - 6 / 155]

自增序列

1. 增加新列包含自增序列、增加常量列、增加字符串列和增加日期列，即在数据上增加一列数据

2. 对该数据增加一列自增序列，从0开始自动编号

空值填充

1. 空值填补是对列中的空值通过人工输入和附近统计值的方式进行填补的，只针对数值类型的数据
2. 对age列的数据进行空值填补，通过人工输入的方式为50值替换
3、值替换是将数据列进行空值替换或者非数值替换为某值，可以同时选择多列

DataONE组件---数据ETL举例

类型替换

1. 类型转换是将数据中列的数据类型进行转换；
2. 有的数据是数值，但在数据库中的类型为string或者其他不能计算的类型，需要将该列数据转换为数值类型的；
3. 对work_country列的varchare类型强制转换为整型，所以输出的结果全为null.

数据过滤

1. 根据输入的列信息和过滤条件得到满足条件的数据
2. 此例过滤age字段区间（35，45）内的数据记录

批处理去重

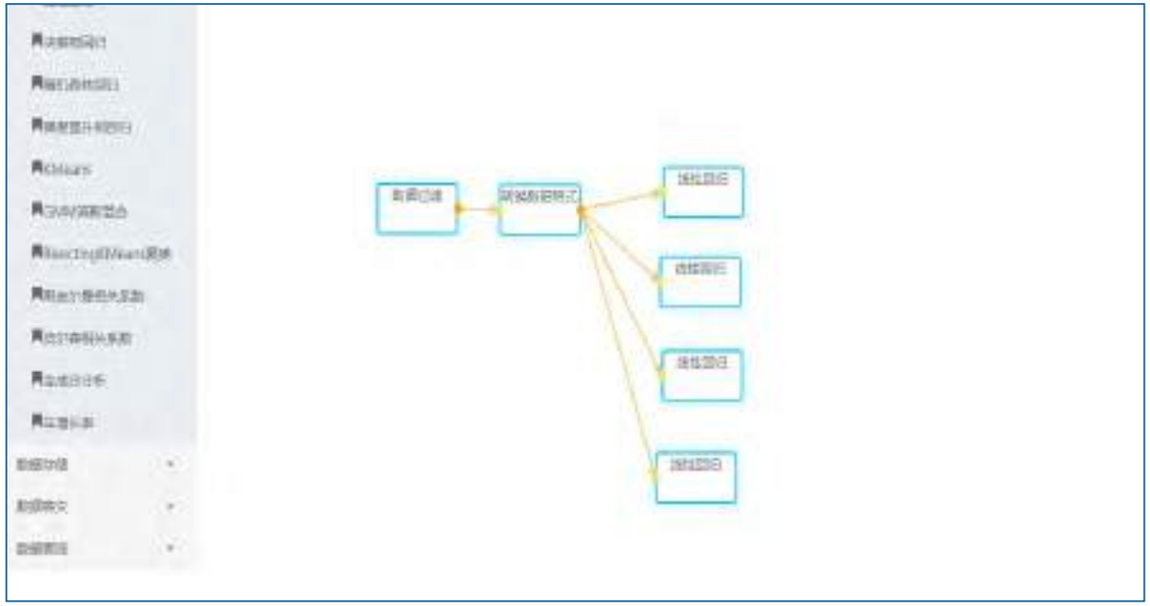
1. 数据去重组件可以针对一列或者多列数据进行去重，针对该列重复的行只会保留一行这样的数据；
2. 对work_country列数据去重后，得到的结果中不存在重复的国家.

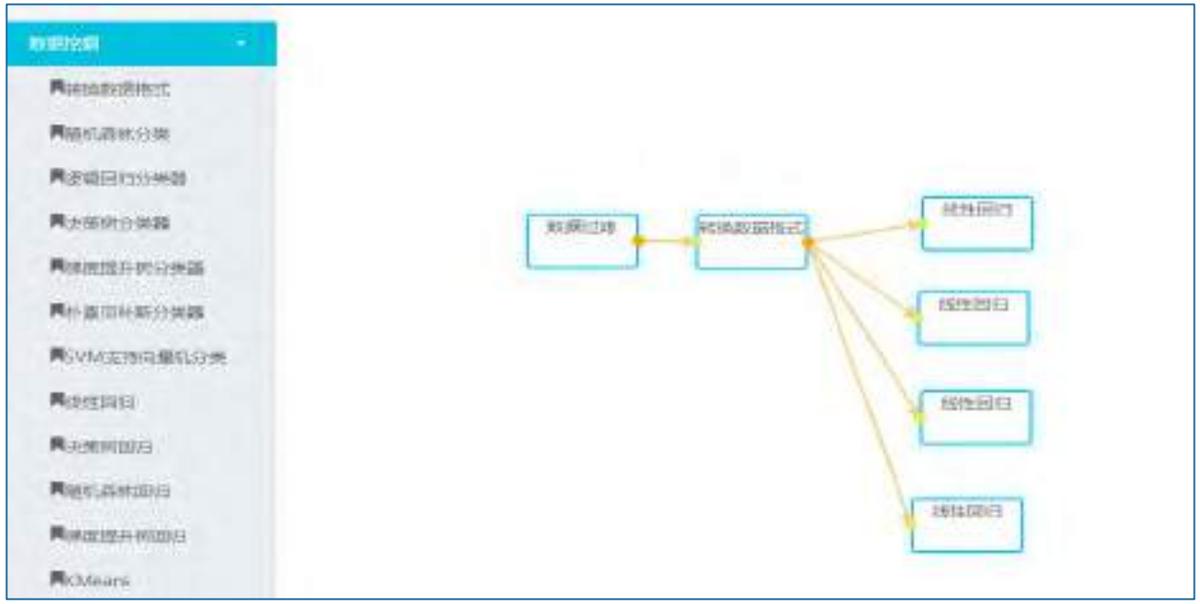
DataONE组件---数据挖掘

使用模型进行预测的标准流程是先做ETL，
接着格式转换，最后选择合适算法生成模型。

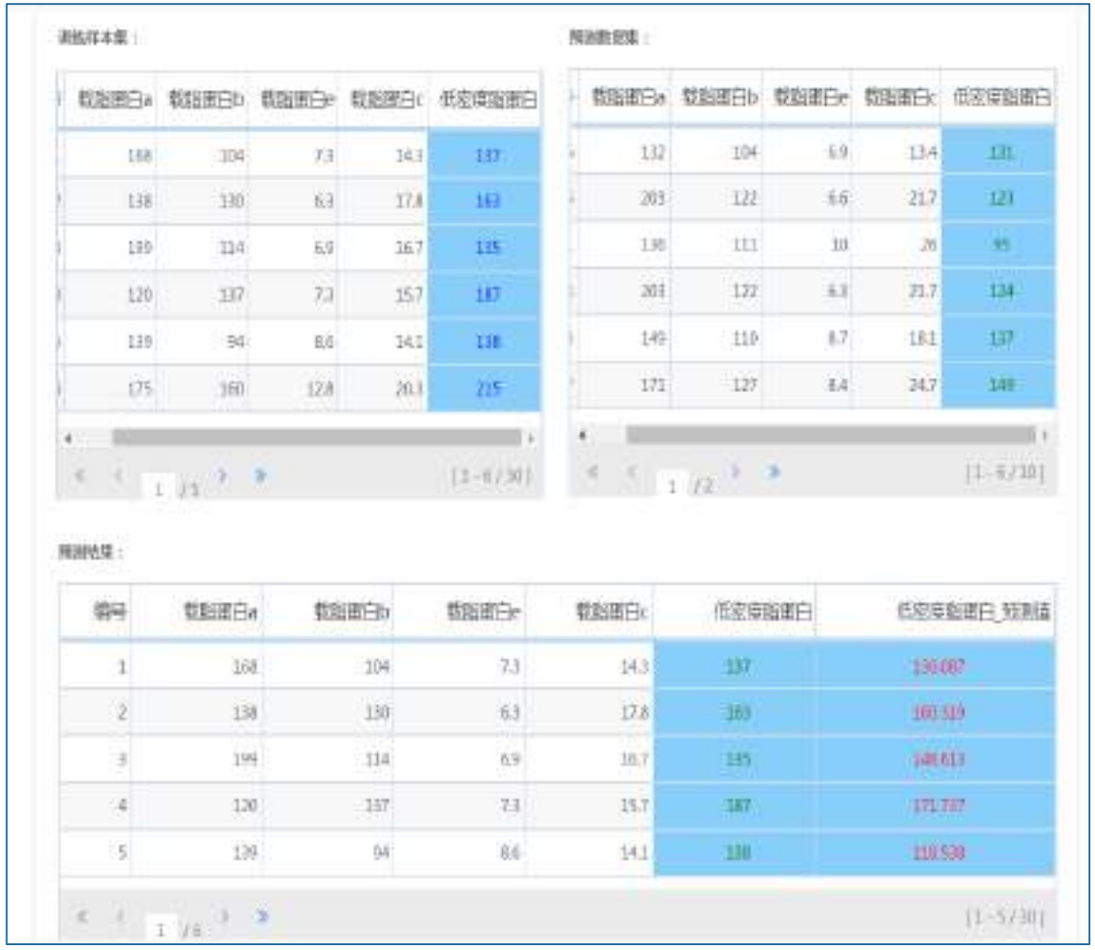
数据挖掘目前主要有几大类：

- 聚类分析
- 时序预测
- 回归分析
- 分类
- 神经网络





数据挖掘---线性回归模型举例



线性回归模型

回归分析(regression analysis) ,一个统计预测模型，用以描述和评估应变量与一个或多个自变量之间的关系。

分析：

1. 用来拟合一个变量与其他解释变量之间的线性关系。最终呈现回归方程以及模型检验的结果。用户可以增加一个线性回归预测节点对数据进行预测；
2. 本例的线性回归方程为：

7.36929480167497+0.130192489933246 6*载脂蛋白a+1.363810654337033 9*载脂蛋白b+(-2.219345743810897)*载脂蛋白e+(-0.4457326860357491)*载脂蛋白c

根据历史数据来对指定数据进行预测

数据挖掘--- KMeans聚类模型举例

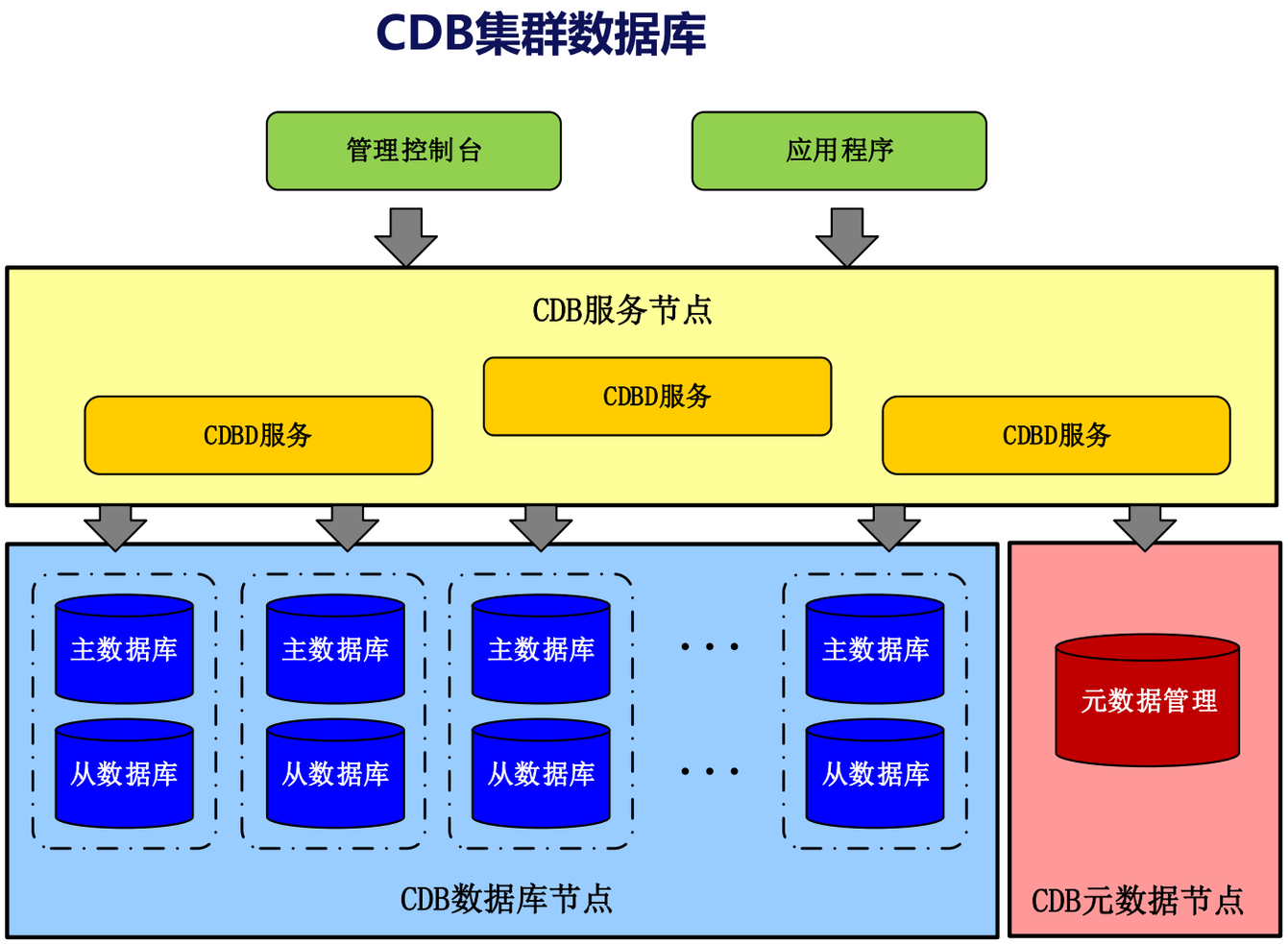
Bisecting k-means聚类算法，即二分k均值算法，它是k-means聚类算法的一个变体，主要是为了改进k-means算法随机选择初始质心的随机性造成聚类结果不确定性的问题，而Bisecting k-means算法受随机选择初始质心的影响比较小，是为了克服K-means算法收敛于局部最小值的问题而提出的

DSG自助报表分析组件是一个分布式搜索，提供Hadoop之上的SQL查询接口及多维分析（OLAP）能力以支持超大规模数据。

- 支持拖拽生成报表。

直接将MYSQL库中的数据导入到Hive中，使用DSG数据分析组件建模后，即可拖拽生成得到可视化报表，并且这些报表还可以切换展示效果。

大数据采集&应用—存储计算层



功能特点

- 动态添加、删除数据库
- 支持多种数据导入到CDB中
- 支持CDB数据导出成文件、SQL、xf1、xdt格式，方便加载到其他数据库中
- 支持标准编程接口，实现二次开发
- 数据备份
- 用户操作权限管理
- 图形化操作界面

性能与客户评价

- CDB是一款高可用、高性能、高安全性的集群数据库，可弹性扩展，无需改变企业现有设备，降低企业成本
- CDB的高并发分析能力，可用于企业经营分析系统中，同时还可以用于归档查询库，分担主业务压力