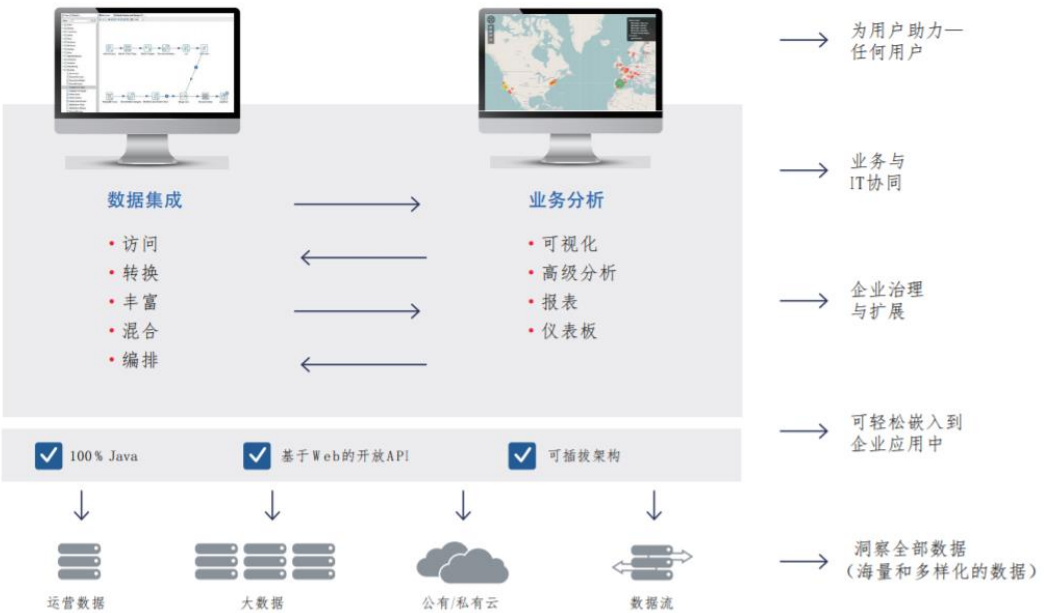
长虹数据集成软件CHDI

产品技术白皮书V8.2

长虹佳华是一家国企控股的香港上市公司（股票代码 08016.HK），是整合、优化全球资源的专业 IT 解决方案服务商与 IT 产品分销商，专业位置及信息服务终端产品生产商和服务商。

长虹佳华定位于新型的 IT 综合服务企业，以“做帮助成长、支持成功的好伙伴”为企业经营理念，以卓越的营销服务、专业的解决方案、自主知识产权专有设备、多元化产品，为全球 IT 领导企业和本土渠道合作伙伴及客户提供高效、专业的帮助与支持，帮助合作伙伴和客户成长、成功。

长虹大数据平台（CH Big Data Platform）是以工作流为核心的，强调面向解决方案而非工具组件的，基于 JAVA 平台的数据集成软件(Data Integrated)套件 DI，包含数据抽取，抽取转换，数据集成，数据挖掘，机器学习等端到端的数据集成和数据分析平台。CH Big Data Platform 是纯 JAVA 编写，可以在 Window、Linux、Unix、MacOS 上运行，数据集成高效稳定，数据分析简单易用。

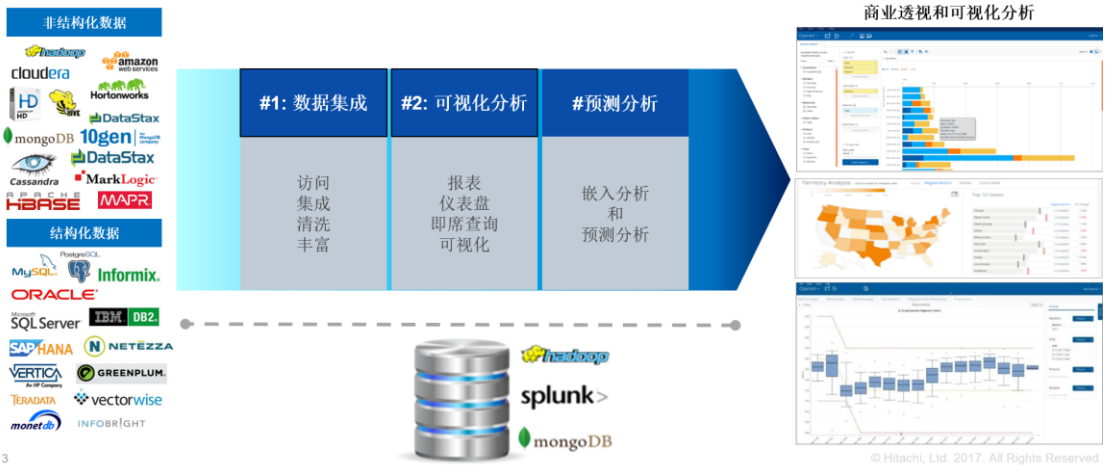


长虹大数据解决方案

1 数据集成分析平台解决方案

长虹佳华在大数据集成分析耕耘 10 多年，不但可以处理传统数据仓库所需要的结构化数据，而且处理融合非结构化数据和 Hadoop 数据湖，提供按需数据集市的自服务平台打通异构数据仓库的壁垒。

方案架构



模块说明

- 传统数据包含 ERP 企业资源计划系统，销售系统，客户管理系统，财务系统等文件，云服务，企业级应用数据。
- 通过 CHDI 的 ETL 功能加载到企业级数据仓库 EDW 中,通过 ETL 功能加载到不同业务系统的数据集市，（物理上可以与 EDW 在一个集群中），通过 CHBI 包含的

仪表盘设计器，嵌入式分析，自服务分析，仪表盘，操作型报表，移动端分析等功能提供给一线经理，分析人员，最终用户，管理人员和客户等用户使用。

- 大数据包含网络数据，位置信息，网页数据，社交媒体数据等非结构化数据，云服务社交媒体数据。
- 通过 CHDI 的 EL 功能加载到企业数据湖 Hadoop 或 NoSQL 存储中，通过 ETL 功能加载到不同业务系统的分析型数据集市，通过 CHBI 包含的仪表盘设计器，嵌入式分析，自服务分析，仪表盘，操作型报表，移动端分析等功能提供给一线经理，分析人员，最终用户，管理人员和客户等用户使用。

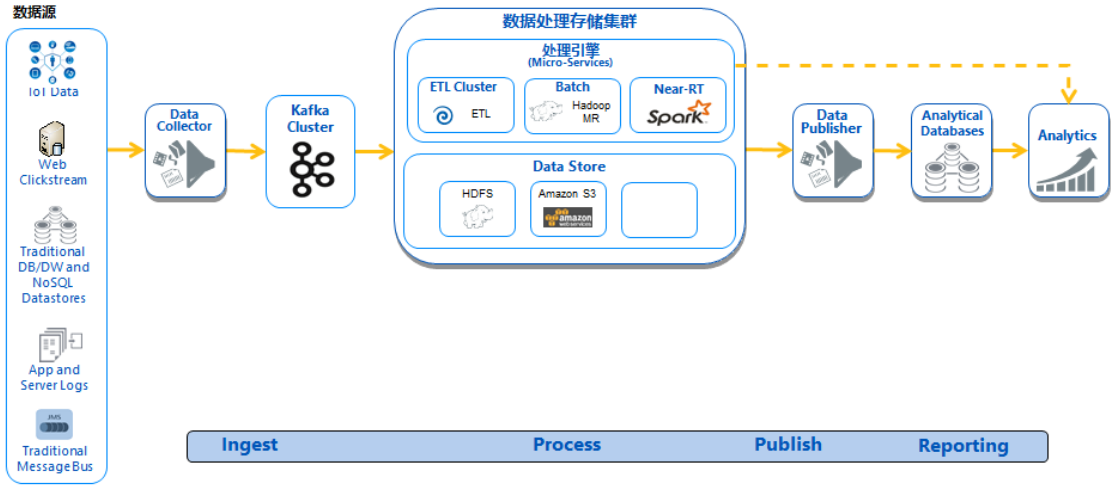
长虹数据集成分析平台的功能

- 长虹数据集成分析平台是一个端到端的大数据集成分析工具，可以支持不断融合不同数据形态，并且支持不断进化的大数据架构。
- CHDI 的计算下压功能，可以很好支持将同一次数据处理任务适应各种执行引擎，可以在 CHDI 服务器完成，也可以在 Spark 集群完成，特别适合海量数据复杂情况下数据移动和处理计算下压到 Spark 集群进行大规模并行处理
- 长虹数据集成分析平台的自助式数据准备功能，可以融合企业级数据仓库和 Hadoop 的数据交互，可以按需提取数据集市，只有在需要的时候才会进行数据准备服务，充分发挥业务人员的自主能动性。

2 实时流式处理分析平台解决方案

长虹数据集成分析平台不但含有 ETL 清洗转换业务流程等处理步骤，还有包含 Kafka Producer，Kafka Consumer，Get Data From Stream，Spark 等处理步骤支持流式处理的全流程。

方案架构



模块说明

- 数据源包含 IoT 数据工业互联网数据，网站点击流数据，传统数据库数据仓库和 NoSQL 数据库，APP 和服务器日志，传统消息队列
- 通过 CHDI 工具采集数据，通过 Kafka producer 进入 kafka 集群

- 数据存储和处理集群包含 ETL 服务器集群，Hadoop 批处理，Spark 实时处理集群。
- 通过 CHDI 将数据发布到分析型数据库
- 通过 CHBI 可以直接访问中央数据存储和处理集群中的 Hadoop 数据或 Spark 数据，或者访问分析型数据库数据。

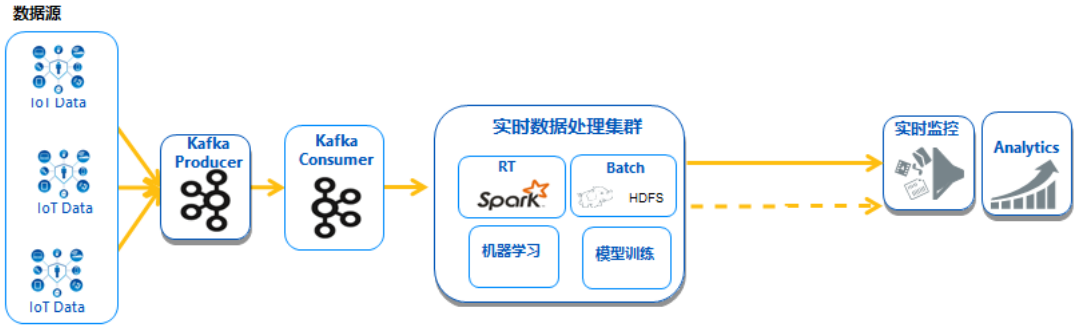
长虹数据集成分析平台的功能

- 长虹数据集成分析平台是一个端到端的实时数据集成软件，可以流式处理 IoT 工业数据，网站日志，数据库等各种数据源，并直接展示为业务分析。
- CHDI 有 Kakfa Producer，Kafka Consumer，Get Data From Stream 等流式处理步骤与 Kafka 良好的集成，同时支持 SparkSQL，支持 Scala/Java 调用 Spark 引擎。

3 机器学习解决方案

长虹数据集成分析平台实时处理能力能够支持机器学习和数据建模，同时 Weka，R，Python，Spark 机器学习算法的集成能够支持各类异常分析和问题分析，并且能够将机器学习模型直接集成到实时问题检测中。

架构



功能模块说明

- 数据源是 IoT 数据，工业互联网数据，各个生产线传感器的数据
- 通过 CHDI 工具采集数据，通过 Kafka producer 进入 kafka 集群
- 实时数据处理集群通过 Kafka Consumer 集成 IoT 数据，包含 Spark 实时处理，Hadoop 批处理，并且支持各种机器学习算法应用例如根本原因分析等分析模型。
- 通过 CHBI 可以直接实时展示业务运行指标,并且交互式展示分析模型应用的结果。

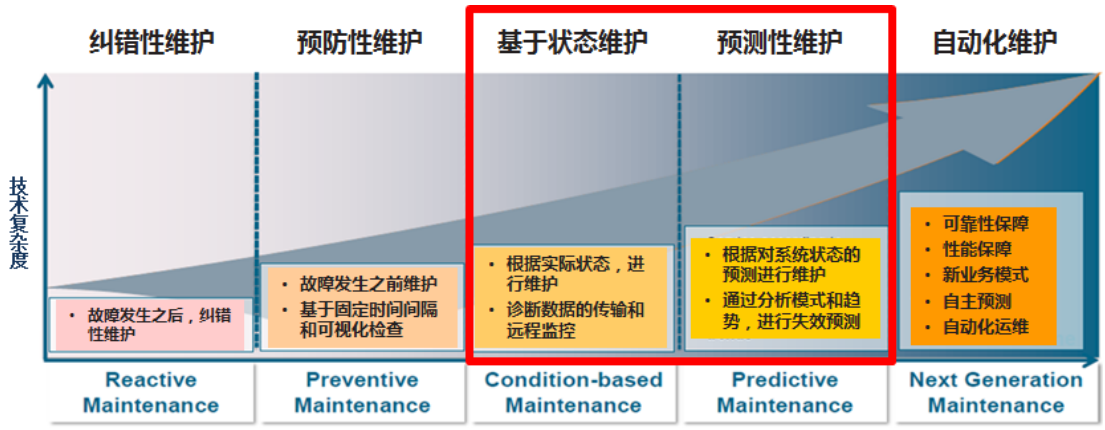
长虹数据集成分析平台的功能

- 长虹数据集成分析平台是一个端到端的实时数据集成软件，可以直接处理集成流式处理 IoT 工业数据，实时进行数据展示和业务分析。
- CHDI 有 Kakfa 和 MQTT 步骤与 Kafka 和工业互联网消息队列协议良好的集成，同时有 Weka，R，Python，Spark 等机器学习算法库进行良好的集成，进行模型

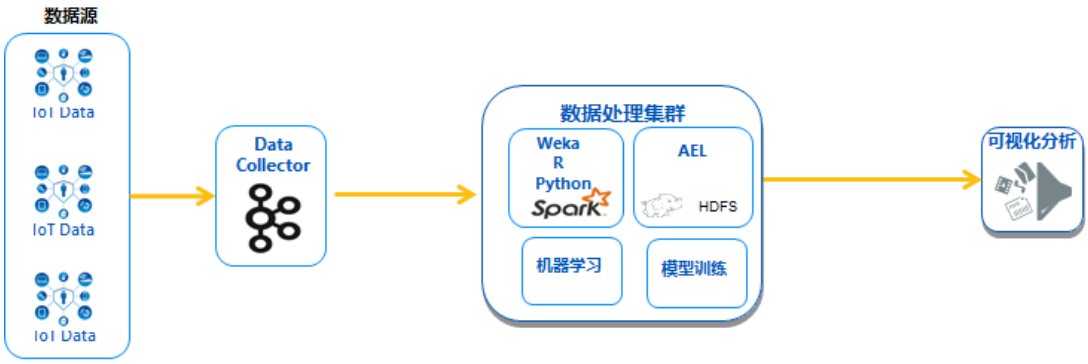
的训练部署和更新，将模型直接与异常检测的数据处理流程无缝集成。

4 预测分析解决方案

长虹数据集成分析平台直接支持 Weka，R，Python，Spark 机器学习算法的集成，并且能够将机器学习模型直接集成到数据处理流程中，能够支持从纠错性维护，预防性维护，基于状态维护，预测性维护以及自动化维护的全生命周期的设备维护的全流程分析中。



架构



功能模块说明

- 数据源是 IoT 数据，工业互联网数据，各个生产线传感器的数据
- 通过 CHDI 工具采集数据，通过 Kafka 或 MQTT 等集成工业互联网数据
- 通过 CHDI 中的 Weka，R，Python，Spark 等机器学习算法库进行数据特征提取，训练模型，模型部署，以及模型更新。
- 通过 CHDI 可以在数据处理过程中直接浏览数据进行分析和优化模型。

长虹数据集成分析平台的功能

- 长虹数据集成分析平台是一个端到端的实时数据集成软件，可以直接处理集成流式处理 IoT 工业数据，实时进行数据展示和业务分析。
- 长虹数据集成分析平台有 Kakfa 和 MQTT 步骤与 Kafka 和工业互联网消息队列协议良好的集成，同时有 Weka，R，Python，Spark 等机器学习算法库进行良好的集成，进行机器学习的全流程数据挖掘分析。

长虹数据集成分析平台是一个端到端的开放和可插拔的集成分析平台。CHDI 和 CHBI 的平台底层开源，因此具备良好的开放性，使客户能够最大资源利用到生态链中

的资源。CHDI 和 CHBI 服务不仅组件化，而且通过 Rest API 和 Web Services 与各种应用集成，实现各种嵌入式分析应用。长虹数据集成分析平台可以整合机器学习，客户不仅可以利用 CHDI 实现数据准备，而且可以嵌入各种预测模型（无论是基于 Weka、R、Python、Spark 等）。

长虹数据集成软件 CHDI

长虹数据集成软件（CHDI）是国内领先的数据集成平台，设计该产品的目的就是通过整合大数据和数据仓库，梳理数据业务流程，把各种数据放到大数据平台里，然后以一种指定的格式流出。长虹数据集成软件产品允许管理来自不同数据库的数据，通过提供一个图形化的用户环境来描述想做什么，而不是想怎么做。长虹数据集成软件产品中有两种脚本文件转换 Transformation 和任务 Job，转换 Transformation 完成针对数据的基础转换，任务 Job 则完成整个工作流的控制。

CHDI 因其可与任何数据类型连接的普适性，以及具有高性能 Spark 和 MapReduce 执行的能力，长虹佳华简化并加快了现有数据库与新数据源集成的过程。CHDI 的图形设计器包括：

- 直观的拖放式设计器可简化分析数据管道的创建。
- 丰富的预置组件库用于访问、准备和混合来自关系型数据源、大数据存储库、企业应用等的数据。
- 能够在任何数据的准备步骤中立即检查数据，并立即访问分析结果，包括

图表、可视化图形和报告。

- 强大的编排功能用于协调和组合转换结果，包括通知和告警。
- 集成的企业调度程序用于协调工作流和调试程序，以便测试和调整作业执行状态。

CHDI 加快大数据源的集成速度，并降低复杂性。长虹提供了：

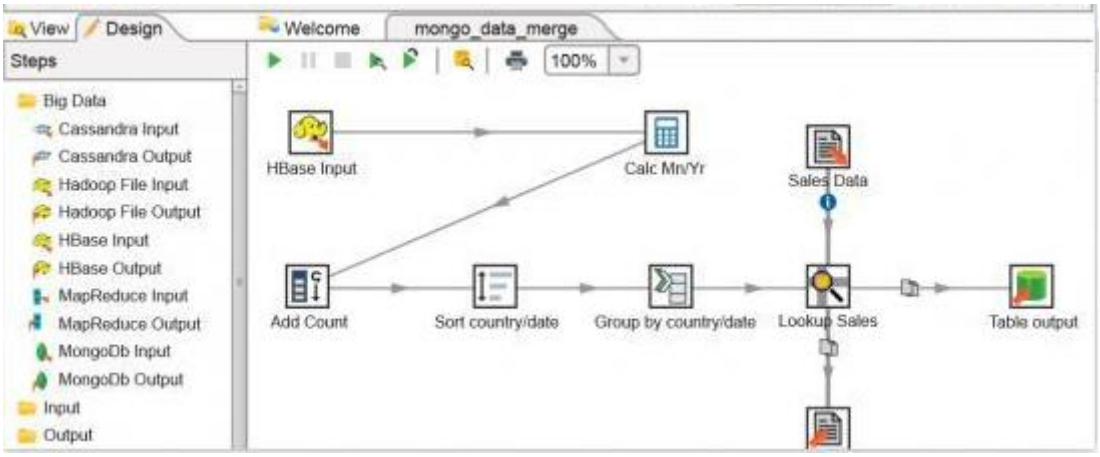
- 无代码的 Hadoop 数据转换设计提供了强大能力。与手动编码相比，工作效率提高 15 倍，并通过在集群内执行实现高性能。
- 以基于模板的方法，通过元数据注入特性集将数据源快速加载到 Hadoop 中。
- 在 Spark 和长虹的本机引擎等执行引擎之间无缝切换，以适应数据量和转换复杂性。

CHDI 为管理数据集成项目的运行提供了现成的能力，包括：

- 共享存储库，使数据分析师、开发人员和数据管理员能够互相协作。
- 内容管理、版本控制和锁定，轻松确定作业版本，以供回退到以前的版本。
- 控制用户和角色的安全权限以及与第三方安全系统的集成；能够设置创建、读取或执行作业和转换的权限。

长虹数据集成软件（CHDI）具有如下的独特功能，保障企业级数据集成平台的易用、易开发、易布署、易扩展和高可靠。

1、拖拽式开发的简单可视化设计器



- 图形提取-转换-加载(ETL)工具，以常规方式来加载和处理大数据源。
- 丰富的预建组件库能访问和转换来自广泛数据源的数据。
- 支持自定义编码并可由图形界面调用，分析图像和视频文件以创建有意义的元数据。
- 动态转换，使用变量决定映射域，验证和改进规则。
- 集成调试器用以检测和调试任务执行过程。

2、零编码要求的大数据集成

- 零编码要求的大数据集成完整的可视化开发工具消除了 SQL 编码或编写 MapReduce Java 函数。
- 通过本地支持的 Hadoop、NoSQL 和分析数据库可广泛的链接到任何类型数据或数据源。
- 并行处理引擎确保高效的性能和企业可扩展性。
- 支持提取和融合现有的多元数据，以生成高质量的实时分析数据。

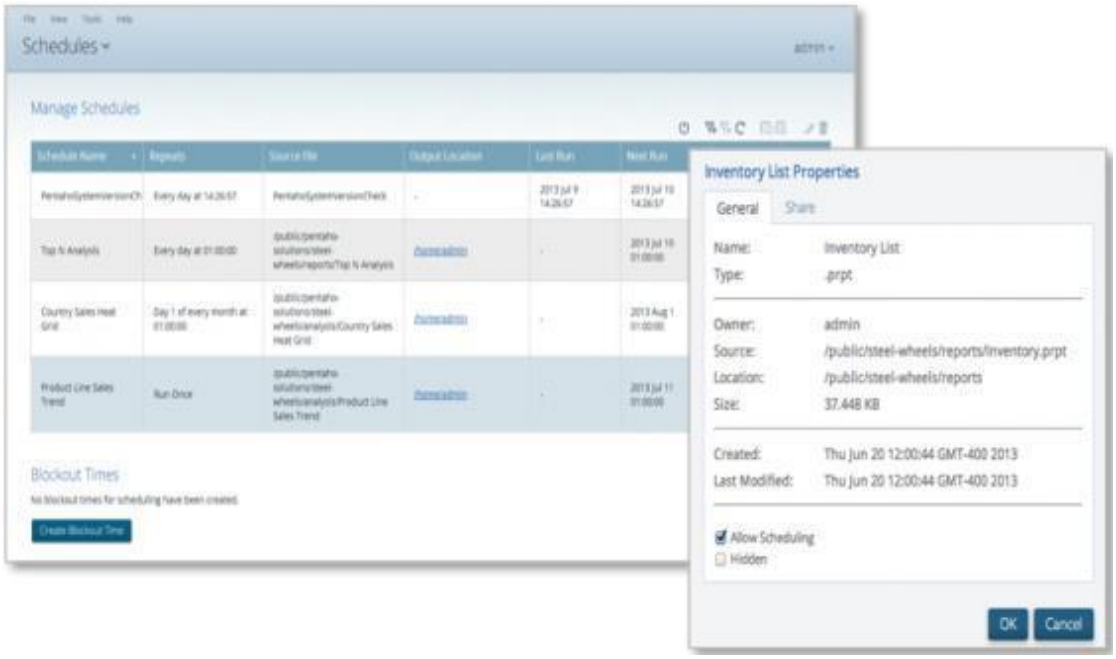
3、灵活支持所有大数据源

- 支持从 Cloudera, Hortonworks, MapR 到 国产华为 FusionInsight，星环 TDH 等最新的 Hadoop 系统。
- 包含针对 Cassandra、MongoDB 等 NoSQL 数据库的插件，也可以连接到 Amazon Redshift 和 Splunk 等专业的数据商店。
- 当使用新的版本和功能时，自适应大数据层为企业节省了大量的开发时间。
- 高度的灵活性，降低了大数据体系变化所带来的风险和孤立区。
- 反馈和分析增加的用户和机器数据的数量，包括网页内容、文档、社交媒体和日志文件。
- 通过灵活的集群分布，可以将 Hadoop 数据任务集成到全面的 IT/ETL/BI 解决方

案中。

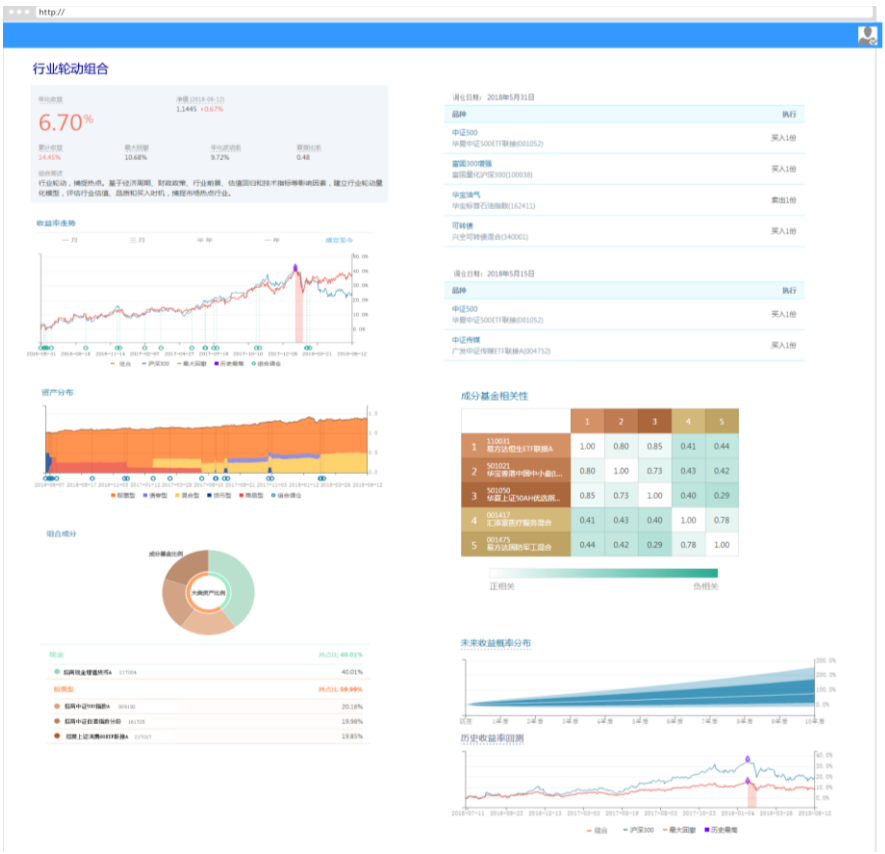
- 支持并行批量数据加载工具，以高效的加载数据。

4、强大的管理能力



- 管理用户和任务的安全权限。
- 从最近成功检查点上重启任务，并从当前失败中回滚作业执行。
- 集成了 LDAP 和 Active Directory 中现有的的安全术语。
- 设置用户的操作权限：读取、执行或创建。
- 进度数据集成过程实现了有序的流程管理。
- 监测和分析数据集成处理的性能。

5、嵌入式分析能力

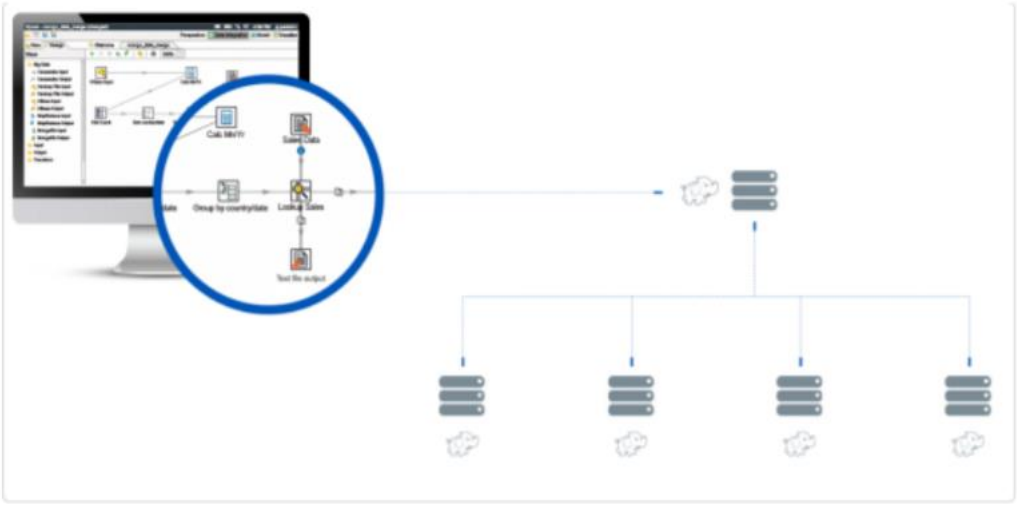


- 全面的数据访问、集成和挖掘平台（支持大数据）。
- 绚丽的可视化，交互式报表，即席分析和自定义面板。
- 实时反馈数据库与任何数据类型的未知连接。
- 使用 CSS 高度定制化、基于 Web 的用户界面来匹配应用程序的品牌、外观。
- 为简单的 SaaS 和云开发提供了多用户共享结构。
- 集成了安全、认证、单点登录等多项功能。

6、开放的架构和标准，支持广泛的扩展

- 服务器支持主流的操作系统，包括 Windows、Linux、MacOS 等。
- 现代化，100%的 Java 平台构建标准，像 REST 风格的 Web 服务接口，
- 方便集成到任何 Web 应用程序。
- 能与企业安全框架无缝集成，通过开放的 API 能扩展到第三方图表和图形。
- 能将复杂的分析方案轻松地嵌入到移动设备和平板电脑应用程序中。
- 产品路线图、源代码和组件都是可见的，以满足不断变化的客户需求。

7、自适应并行计算处理平台

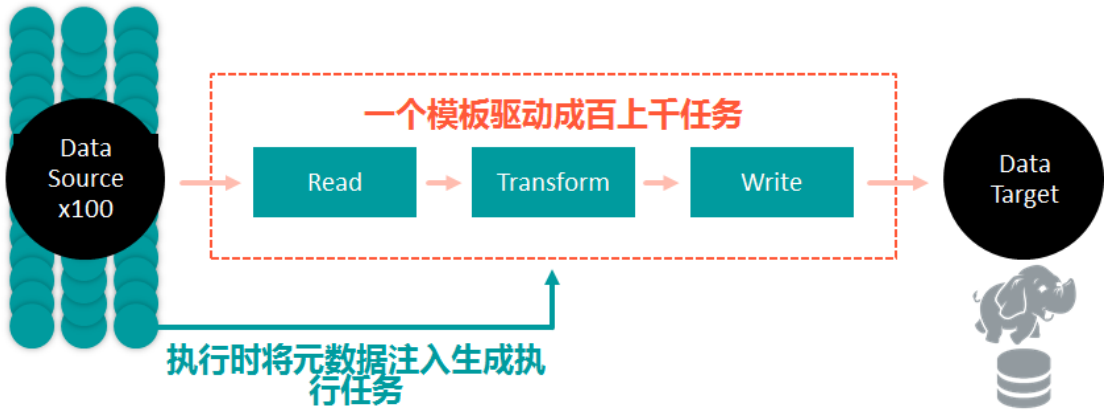


对于在 CHDI 的图形化拖拉开发的作业，可以一次编写多处执行。用户可以选择该作业的执行引擎，无论是 CHDI 的服务器引擎还是利用 Hadoop 平台的计算能力，实现大数据集成能力的平民化和 Hadoop 资源最大化利用，CHDI 具有不可比拟的优势：

开发、计算和存贮层分离。统一的图形化开发生成转换逻辑，开发过程全程可视化。大数据开发技术平民化。转换逻辑作业可以自定义执行引擎，可以选择 CHDI 服务器或 Spark 引擎，保证程序可适应未来技术发展；大数据技术不断更新的适应性。充分利用 Hadoop 存贮层的计算能力。

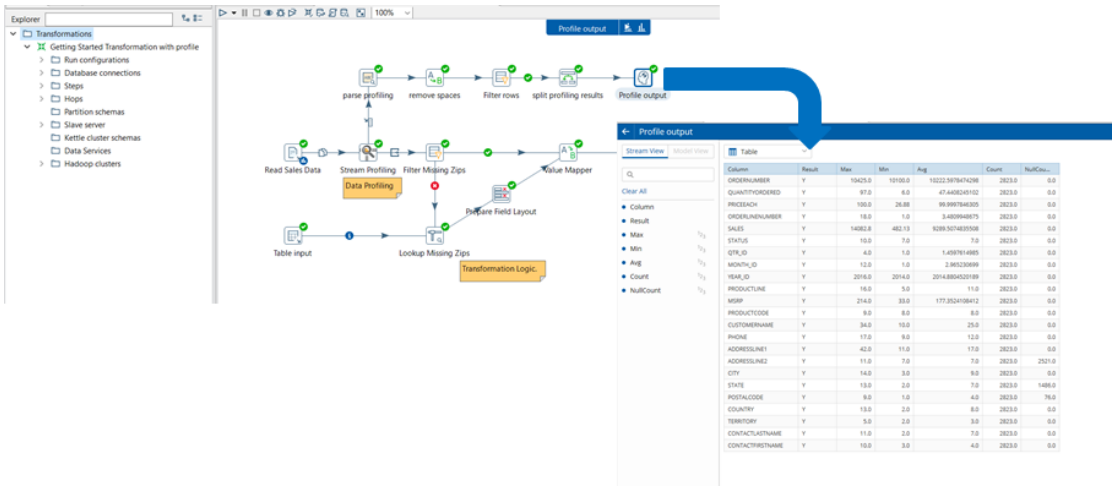
8、批量数据快速集成

CHDI 的元数据注入能够使用一个数据处理模板驱动成百上千的任务，循环执行任务快速开发进行快速数据集成。能够大大地减少开发时间，开发成本和项目风险。

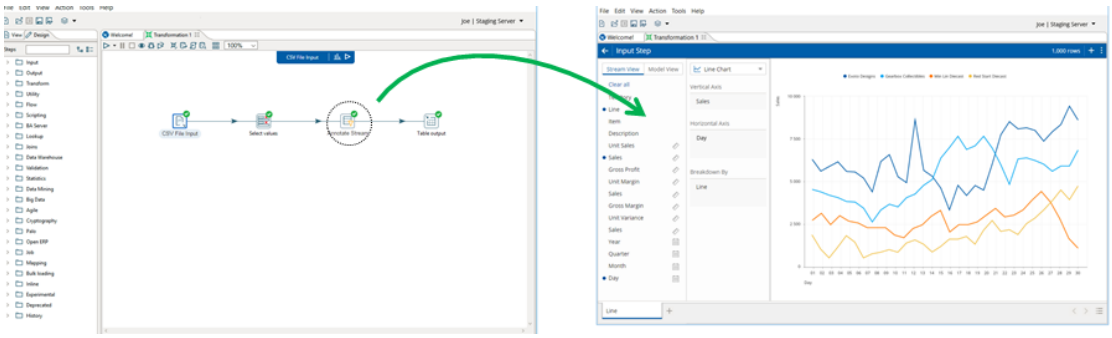


9、数据集成全程可视化

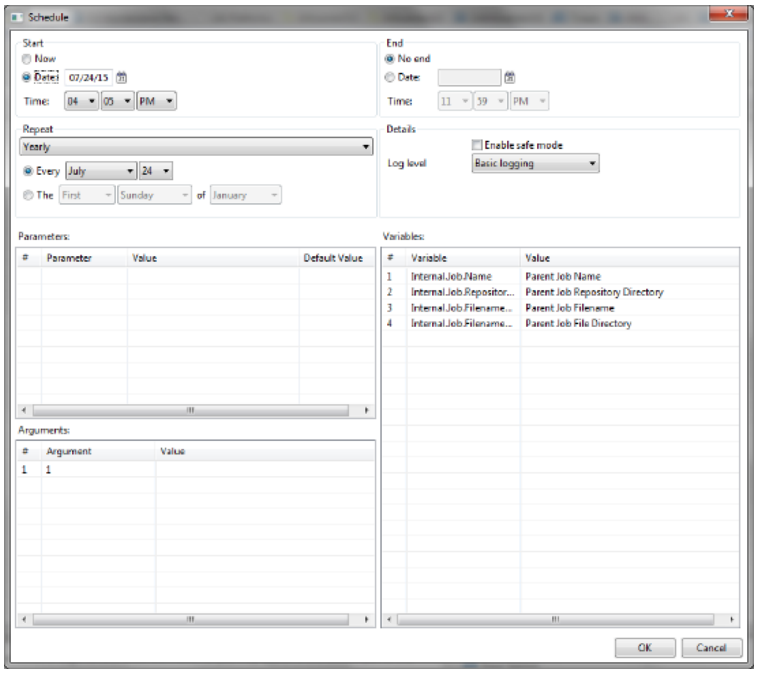
CHDI 在数据处理的任何一个过程中，直接可视化数据，挖掘数据价值



CHDI 无需调用其他系统，在数据处理流程中即可使用可视化的图表，图型，随机查询，分析数据。任意一点都可以检查数据和发布数据，更快地分析，挖掘数据价值。

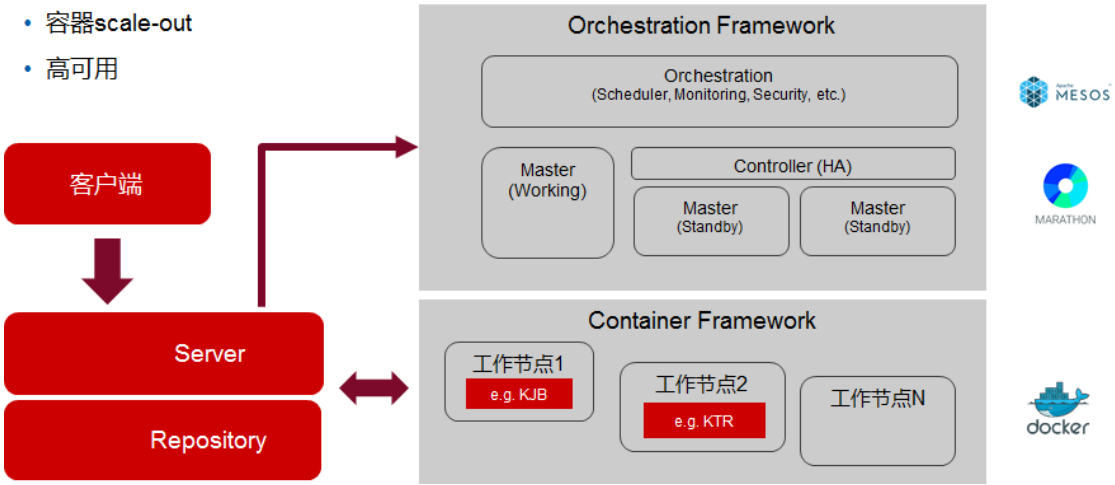


CHDI 支持调度可视化，连接上 CHDI Repository，打开 JOB 或 Transformation 来编辑 Schedule

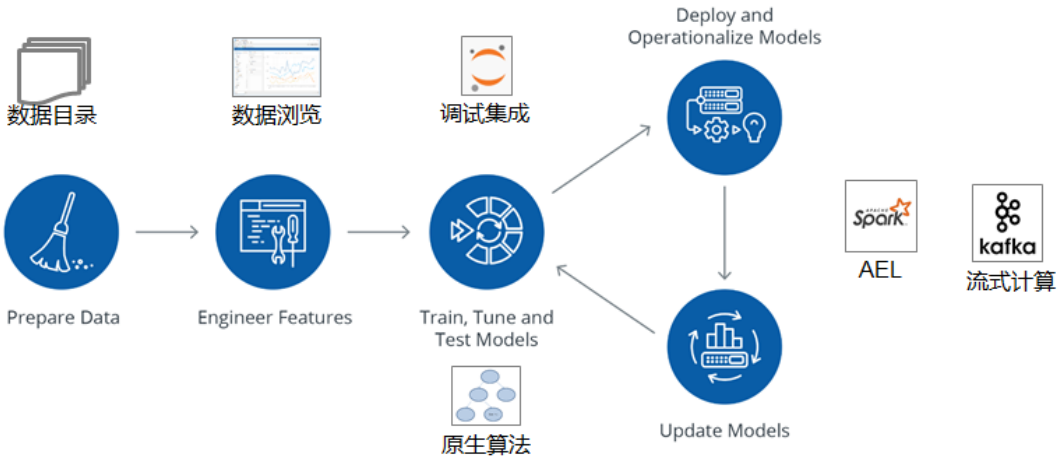


10、服务器集群及高可靠性 HA

CHDI 支持容器技术。CHDI 服务器和工作节点支持 HA。CHDI 服务器支持多个 Master 主备模式响应客户端请求，对于客户端的请求会分布于各个工作节点，降低高峰期客户端请求的处理压力。



11、机器学习及预测建模能力



大多数企业都在将模型投入工作的过程中挣扎，因为数据专业人员通常在孤岛中开展业务，并在为更新工作流建模而准备数据的过程中产生了瓶颈。CHDI 平台可实现协作，并消除以下四个关键方面的瓶颈：

1. 数据准备及特征工程：CHDI 可以轻松地将 ERP 和 CRM 等传统资源与传感器和社交媒体等大数据源结合在一起。CHDI 还在易用的拖放环境中加快了特性设计、数据自动打通、数据转换和数据验证流程等难度大且费用高昂的任务。
2. 训练、调整和测试模型：数据科学家经常采用试错方法而在模型中寻求复杂性、性能和准确性的适当平衡。通过针对 R 和 Python 等语言以及 Spark MLlib 和 Weka 等机器学习库的集成，CHDI 使数据科学家能够更快地无缝地训练、调整、构建和测试其模型。
3. 部署和运行模型：CHDI 允许数据专业人员轻松地将数据科学家开发的模型直接嵌入到运营工作流中。他们可以利用现有数据和特性设计成果，从而显著缩短部署时间。借助可嵌入的 API，企业还可以在现有应用中融入 CHDI 的全部能力。
4. 定期更新模型：Ventana Research 发现，只有不到三分之一 (31%) 的企业

使用自动化流程进行模型的更新。借助 CHDI，数据工程师和科学家可以使用新数据集重新训练现有模型，或使用 R、Python、Spark MLlib 和 Weka 的定制执行步骤进行特性的更新。预置工作流可以自动更新模型，并对现有模型进行存档。

12、平台架构说明

在基础资源方面，CHDI 服务器架构在通用 X86 PC 服务器硬件或共有私有云平台的 Windows、Linux 和 Mac 操作系统上，运行在 Jetty、Tomcat 等应用服务器和 J2EE、Spring 等中间件上，支持 Hadoop、数据库、API、文件等各种数据源的集成。

在 CHDI 平台方面，客户端设计和分析工具包含：

- CHDI 设计器是开发工具，用来开发和设计转换 Transform 和任务 Job，开发和管理调度等。
- CHDI 平台还增加了企业级服务所需要的安全特性，包含 AD/LDAP 的集成，Kerberos，Hadoop Sentry 和 Knox 的集成，支持认证和授权，支持多租户和用户组管理。
- 在扩展能力方面，CHDI 支持开放式标准架构和开放式 API，支持第三方扩展和嵌入式开发包括 Java、REST API、Web Service 等。
- CHDI 支持数据科学全流程，支持 Weka、R、Python、Spark 等算法嵌入至

CHDI 数据处理流程中。

北京

北京市丰台区南四环西路188号18区26号楼长虹科技大厦 邮编：100738 电话：010-58292000 传真：010-58292000

上海

上海市静安区北京西路1701号静安中华大厦602单元 邮编：200040 电话：021-62889117 传真：021-62889115

广州

广州市天河北路898号信源大厦3408室 邮编：510898 电话：020-38182838 传真：020-38182835

深圳

深圳市福田区华强北路群星广场B座2408室 邮编：518000 电话：0755-25327693 传真：0755-83534550