数据集成为Hadoop保驾护航

久经验证的降低数据管理成本
的创新之路

INFORMATICA®

大数据是关于...

创新

“到2015年，那些将高价值、多样和最新的信息类型及来源集成到统一连贯的信息管理基础设施的组织，其财务表现将较业内同行优越**20%以上**。”，

Neil Chandler, Gartner

成本

“当前部署的85%数据仓库项目，都不能适当扩展规模以满足新的信息数量和复杂性要求”，“

Mark Beyer, Gartner

实施久经验证
的创新之路

大

v数据

回报率

=

数据价值

数据成本

随着数据呈指数级增长，
降低大数据成本

您如何权衡创新&成本？

您打算如何利用大数据来开发创新产品和服务？

欺诈侦查，
风险 & 投资组合分析
投资推荐

基于位置的服务

实时数据审计
医疗保险交易
合规性
国家安全

互联车辆

预测维护维修

治疗效果预测
患者监护
个性化医疗
合规性

主动客户沟通交流

药物识别
基因测序
合规性

忠诚度计划
游戏遥测

数据量持续增长，您如何降低 & 控制成本？

源数据

分析系统

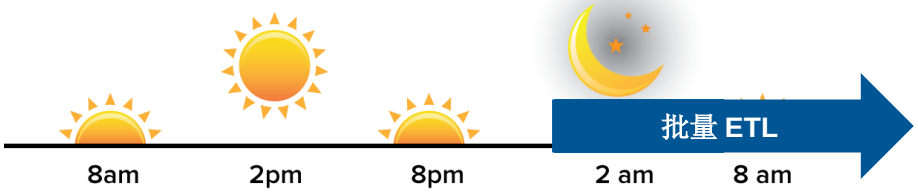
交易，
OLTP, OLAP

文档和电子邮件

社交媒体和网络日志

科学机器设备

数据库和数据仓库迅速
力不从心



批量窗口已到极限，
SLA处于危险之中

原始数据或不经常使用的数据
耗费能力

企业数据仓库

数据集市

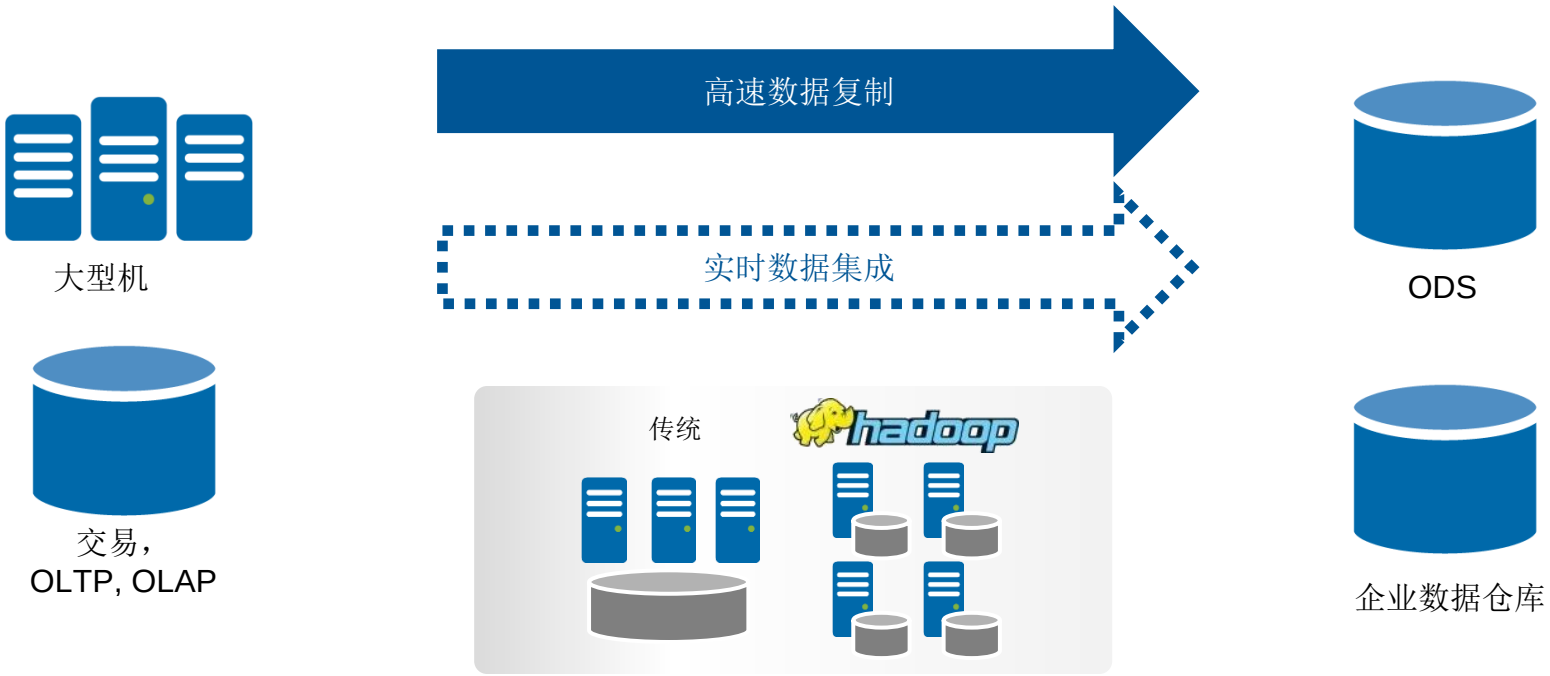
ODS

降低数据管理成本

- 将原始数据临时存储在低成本的商用硬件上
- 将 **ETL/ELT** 处理转移到低成本的商用硬件上

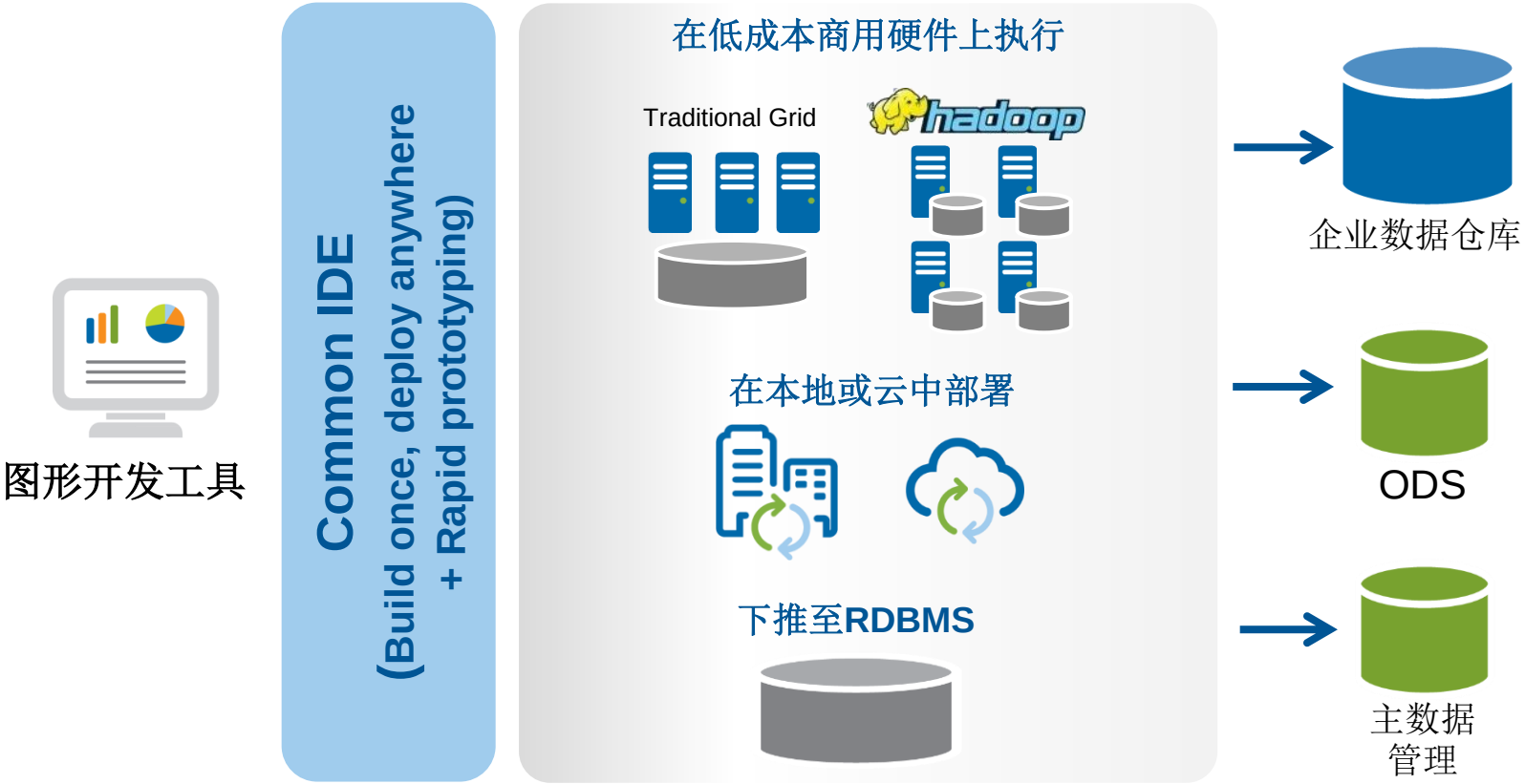
降低数据管理成本

- 借助实时数据集成，平滑实现ETL处理
- 借助高速数据复制，从源系统中卸载处理



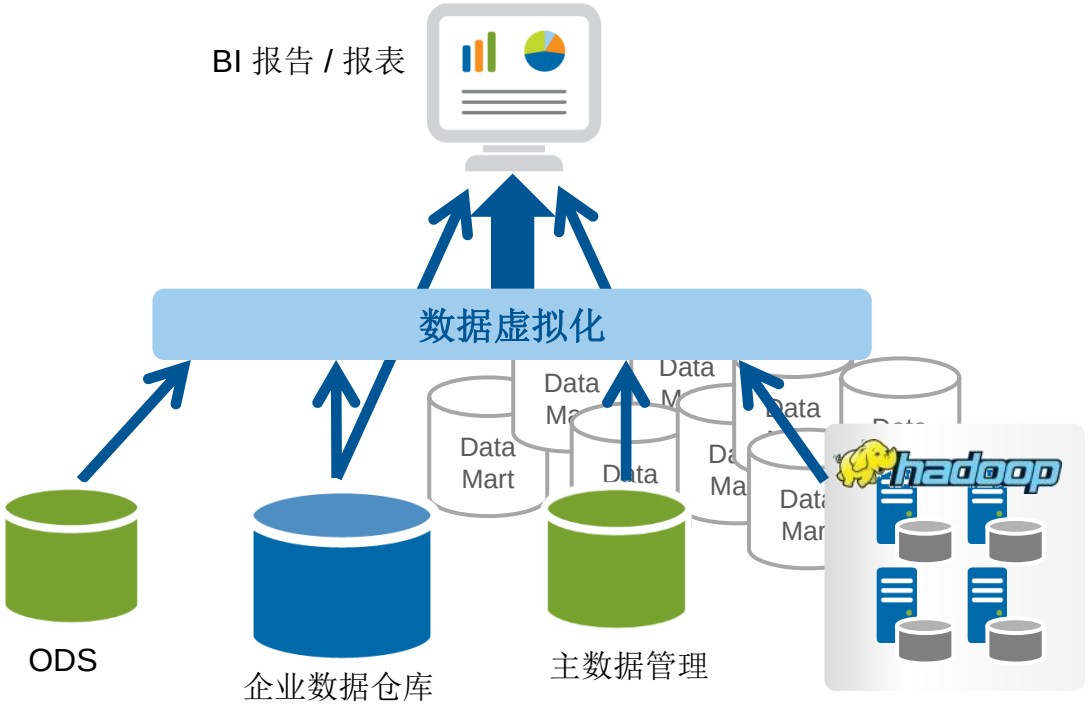
降低数据管理成本

- 借助通用的IDE，将生产效率提升两倍。开发人员通过一次开发，即可实现随地部署。



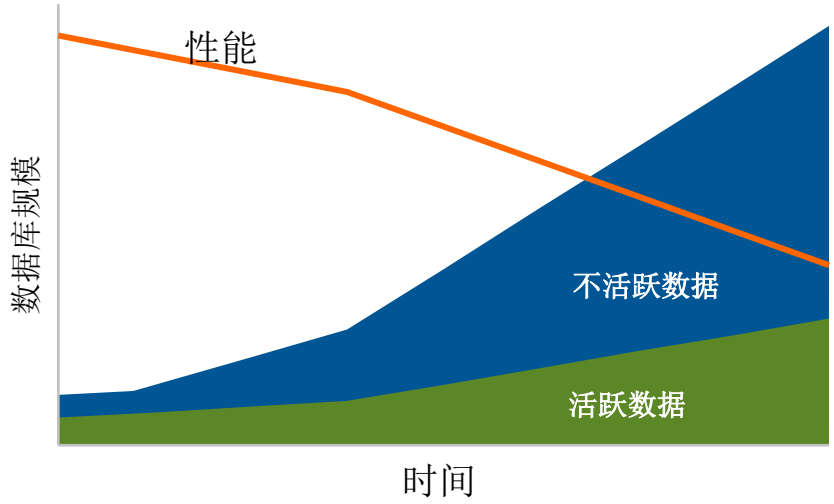
降低数据管理成本

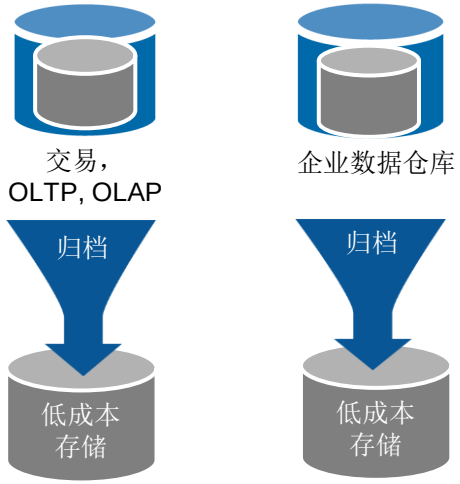
- 消除数据副本，通过数据虚拟化提升数据仓库能力



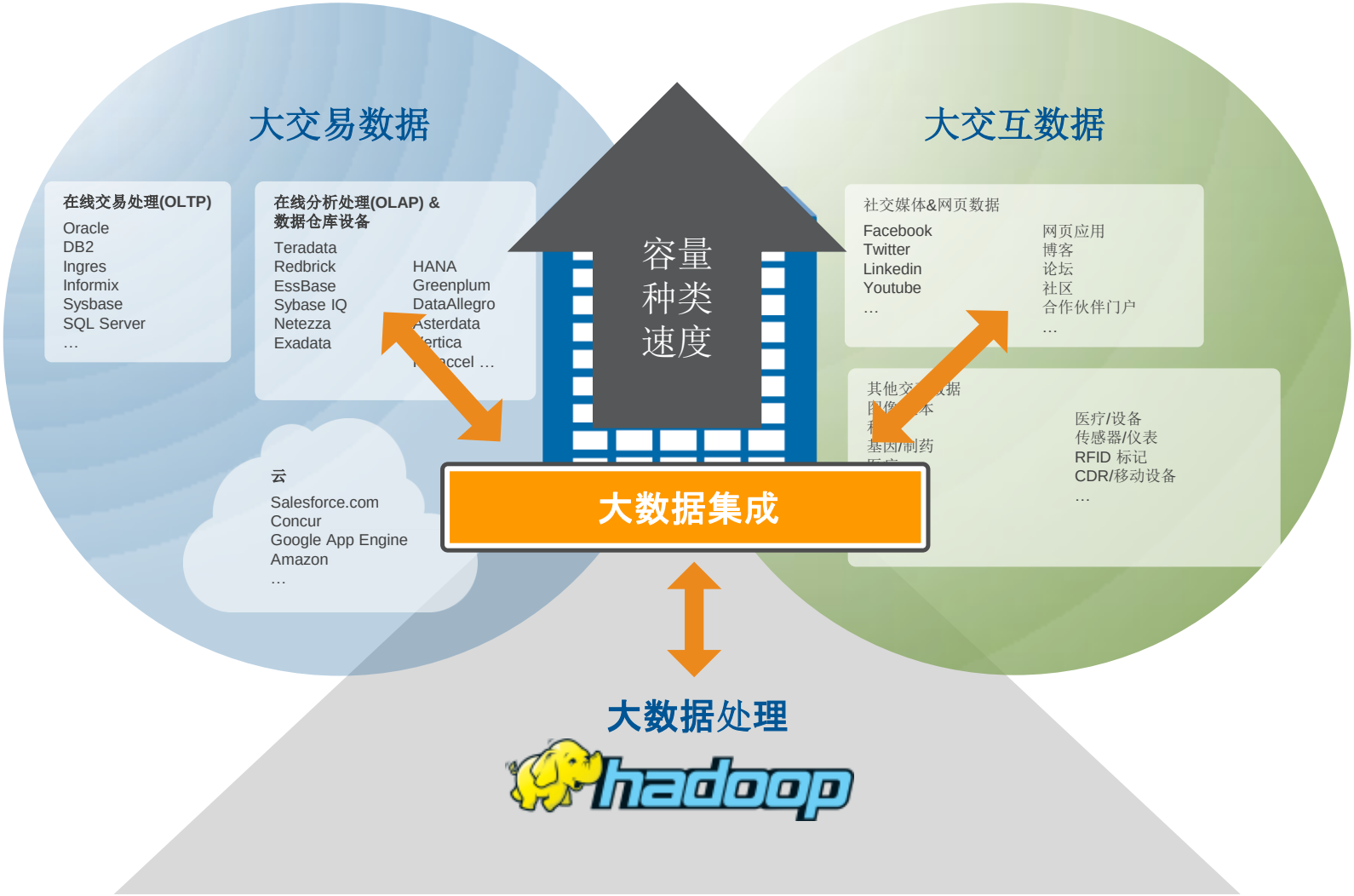
降低数据管理成本

- 识别休眠数据
- 将不活跃数据归档至低成本存储



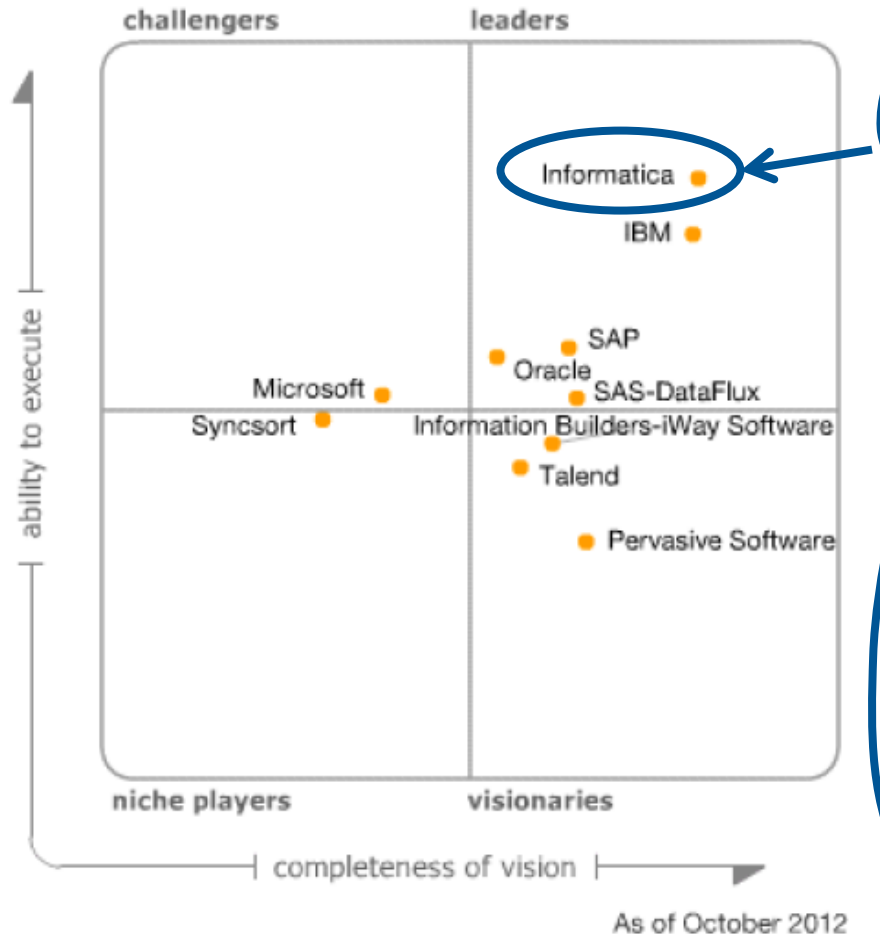


您如何实现大数据的最大回报？



… 以及将大数据项目风险降至最低

Informatica, 数据集成领域的绝对领导者



Source: Gartner (October 2012)

主动客户沟通交流

基于位置的服务

欺诈侦查

国家安全

预测维护维修

治疗效果预测

投资推荐

药物识别

基因测序

互联车辆

风险&投资组合分析

医疗费用

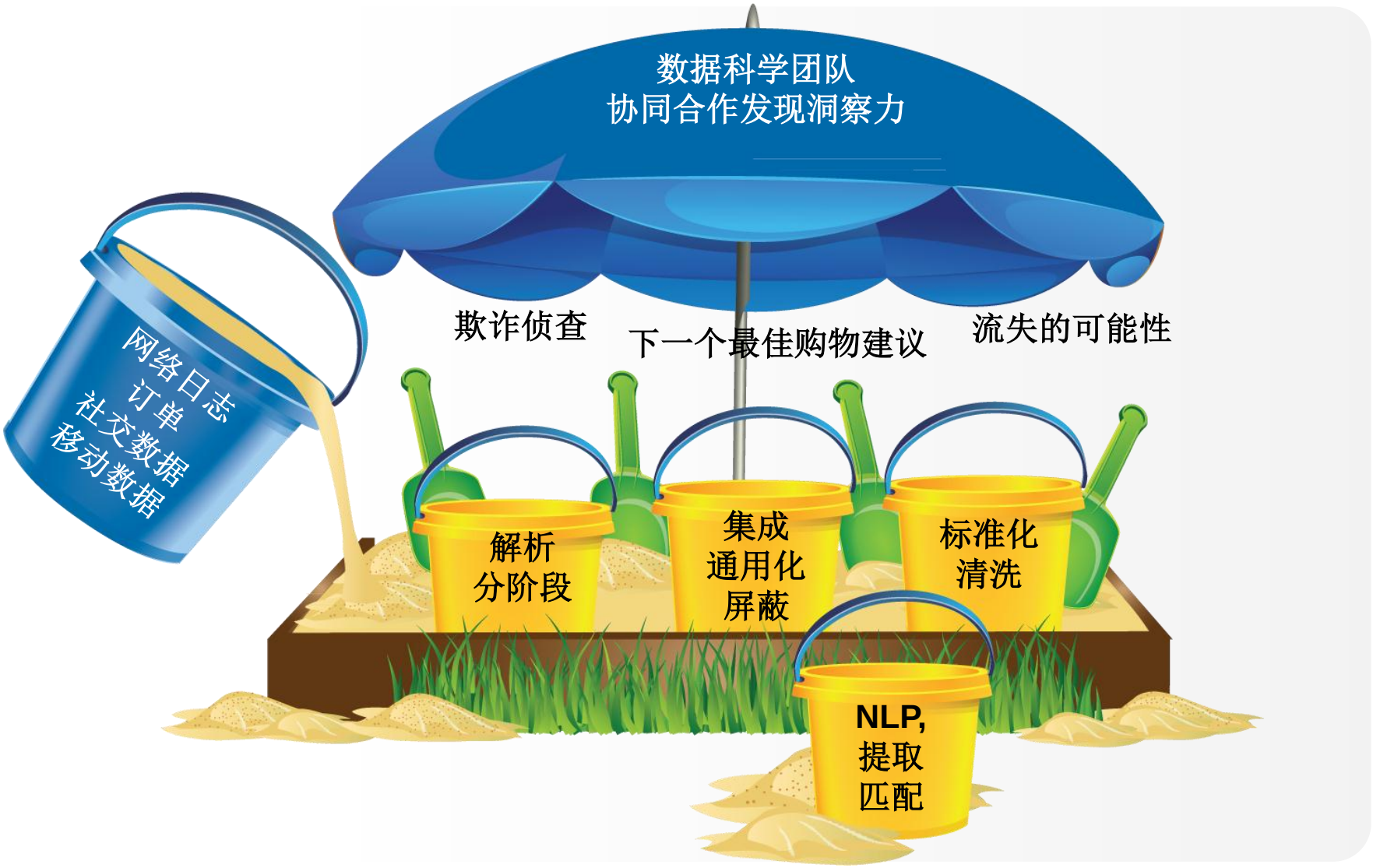
忠诚度计划

合规性

医疗保险交易

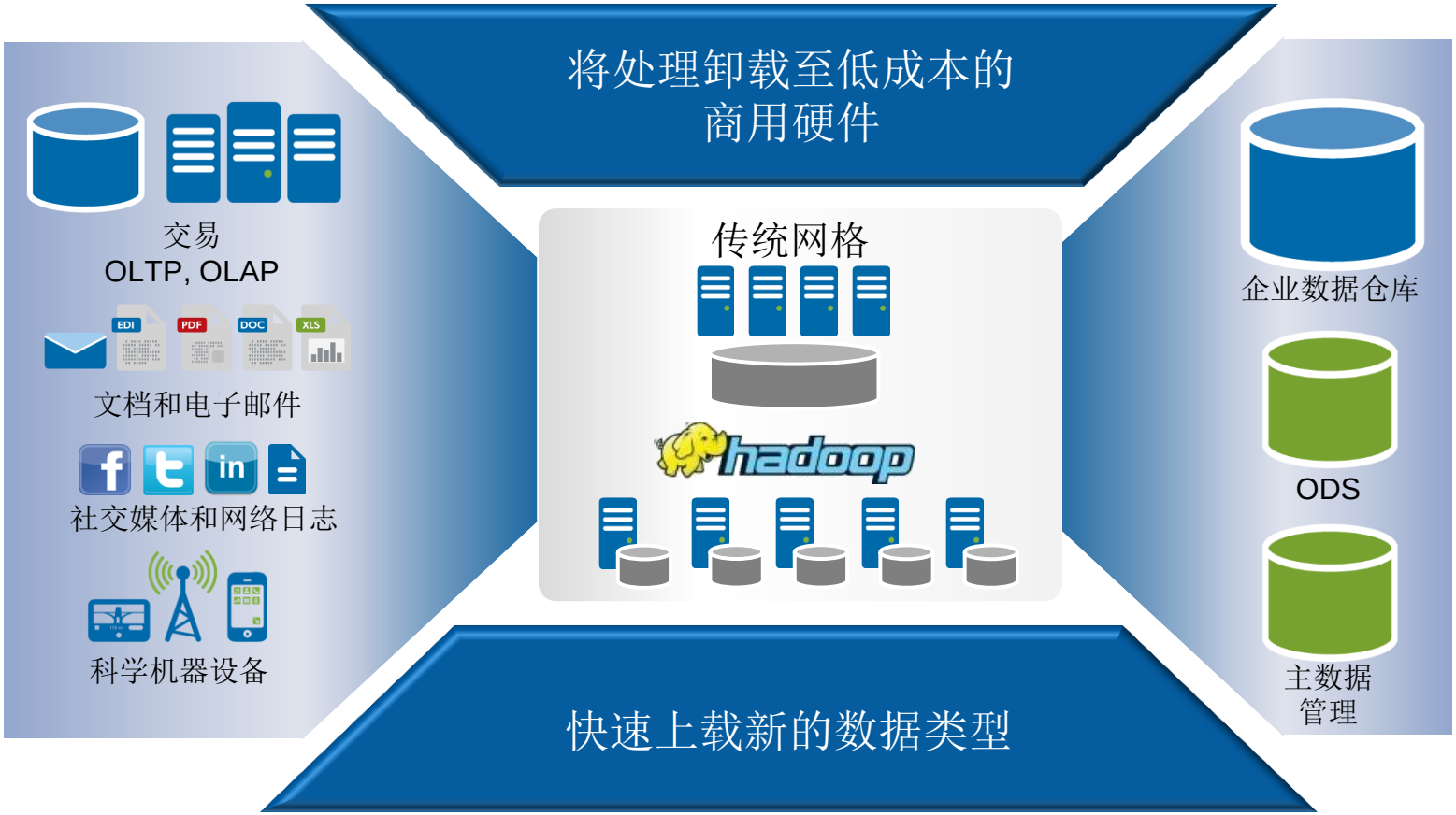
实施久经验证的创新之路

通过快速原型法和合作获得更快的洞察力



PowerCenter 大数据版

降低大数据项目成本

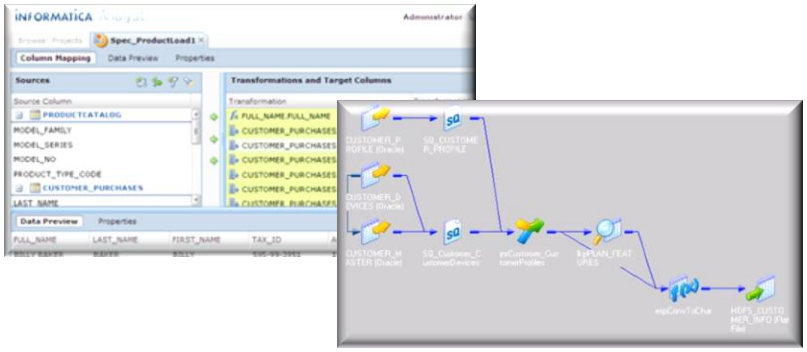


PowerCenter 大数据版

提高生产率，降低风险

分析师 &
数据科学家

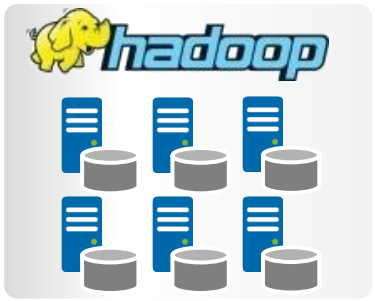
生产率提高近3倍



设计一次
随地部署

开发人员

传统网格



PowerCenter 大数据版

大数据之旅安全畅通

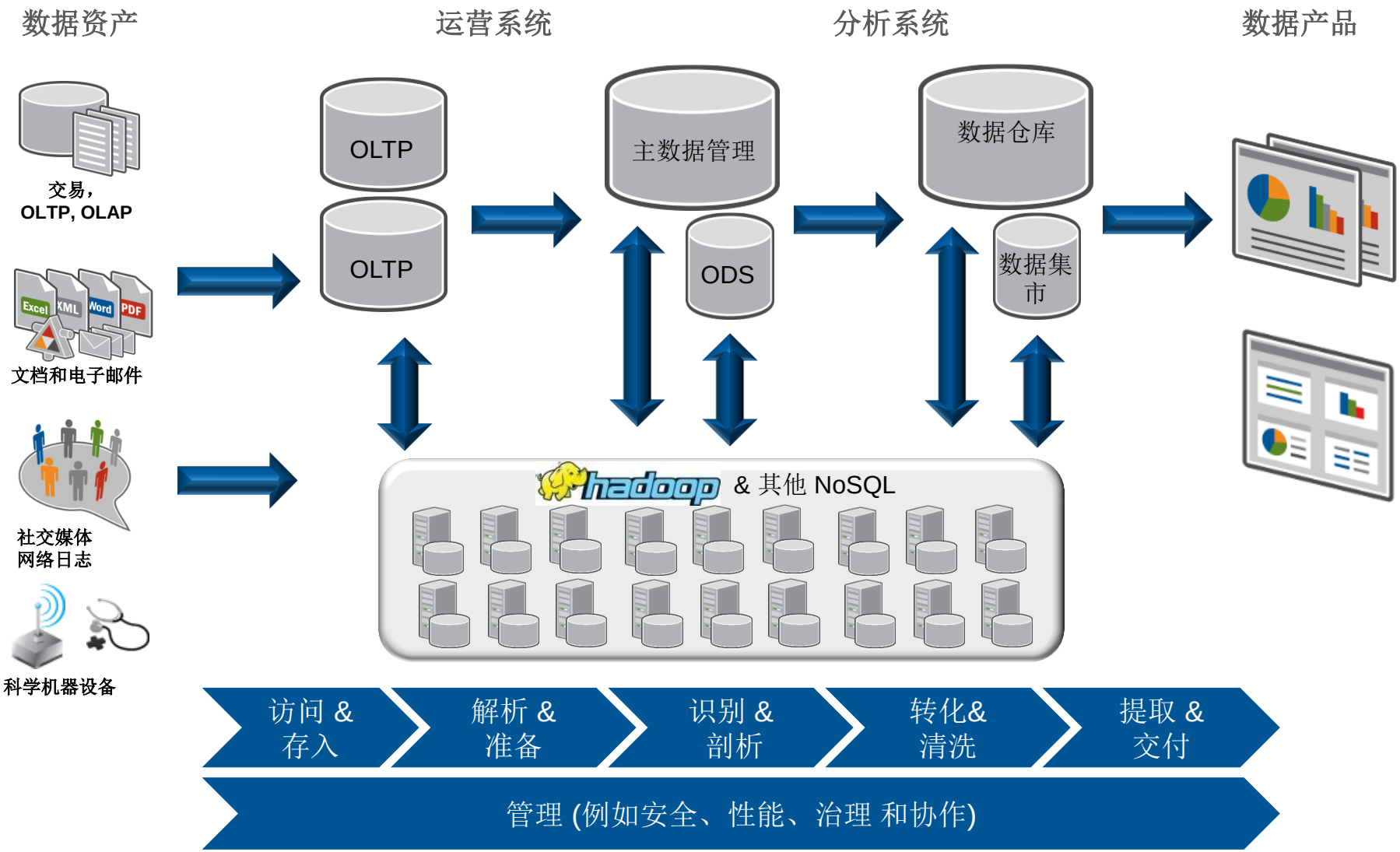
HADOOP核心：MAPREDUCE

分布式计算框架

在多个节点并行处理大量工作，并整合结果。

来源：Cloudera

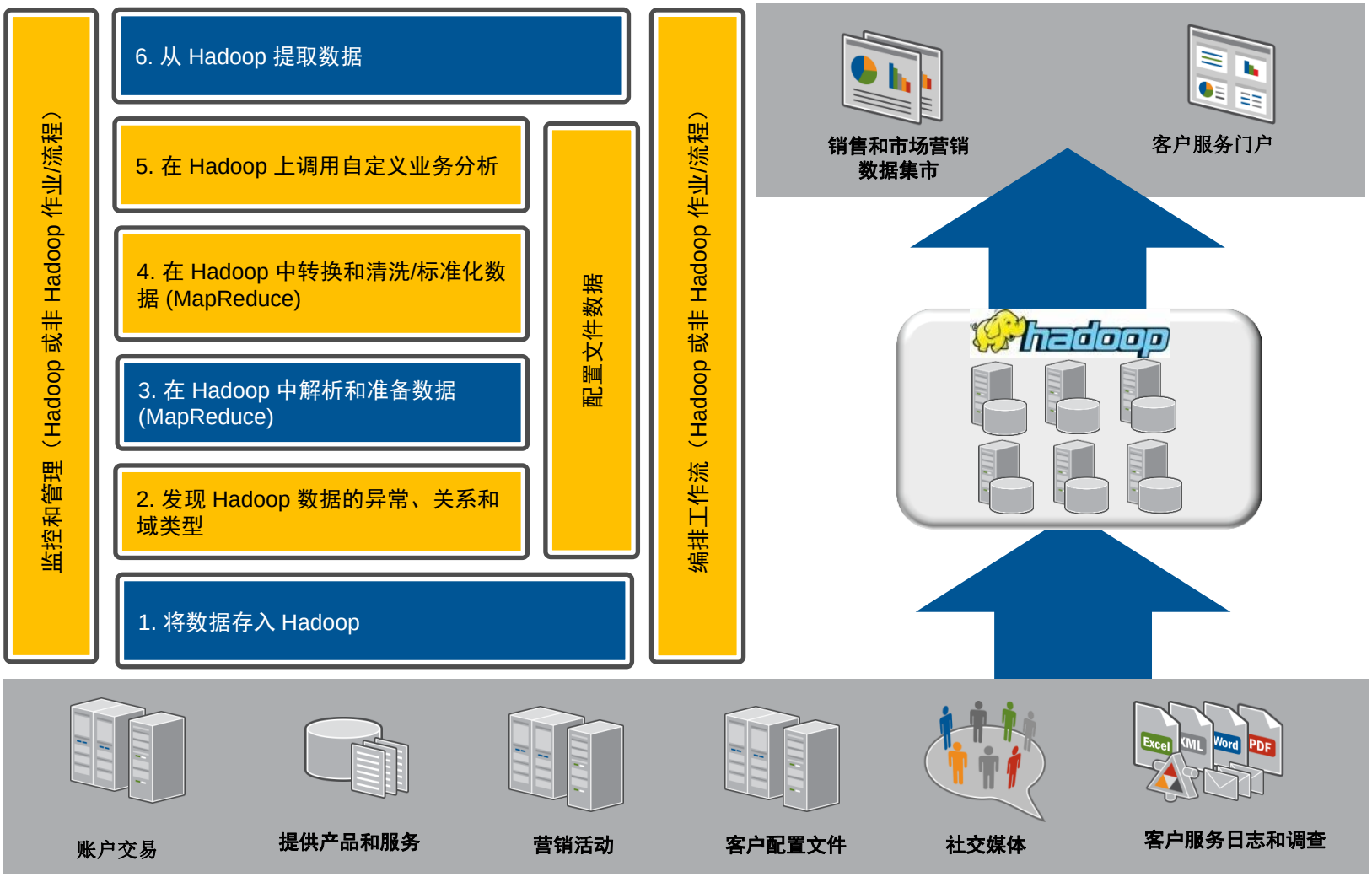
最大化大数据投资回报
Hadoop 补充现有基础设施



通过Informatica 释放Hadoop的强大功能

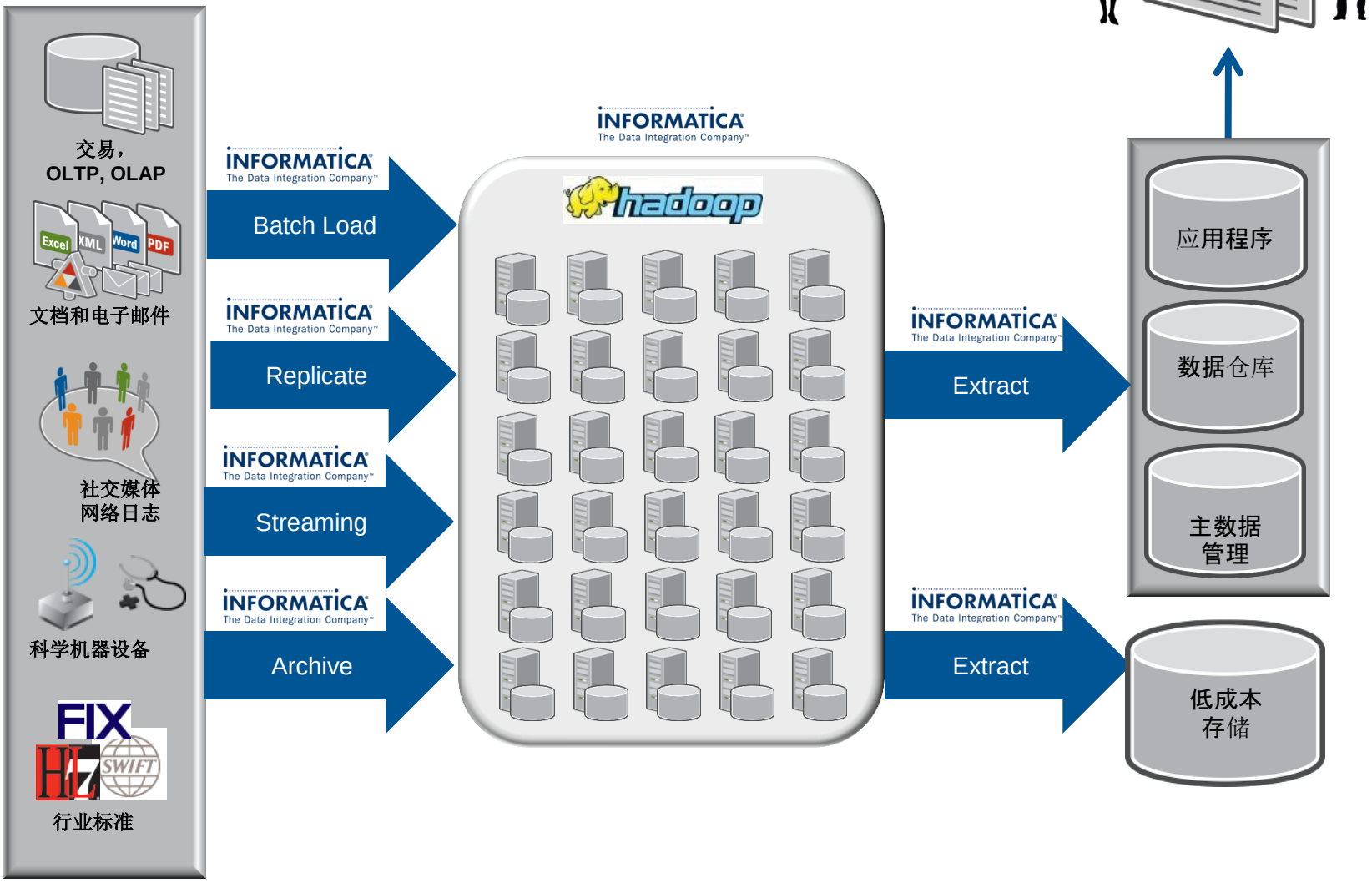
立即可用

2012年12月



数据存入和抽取

每小时移动数十TB的交易数据、交互数据和流数据



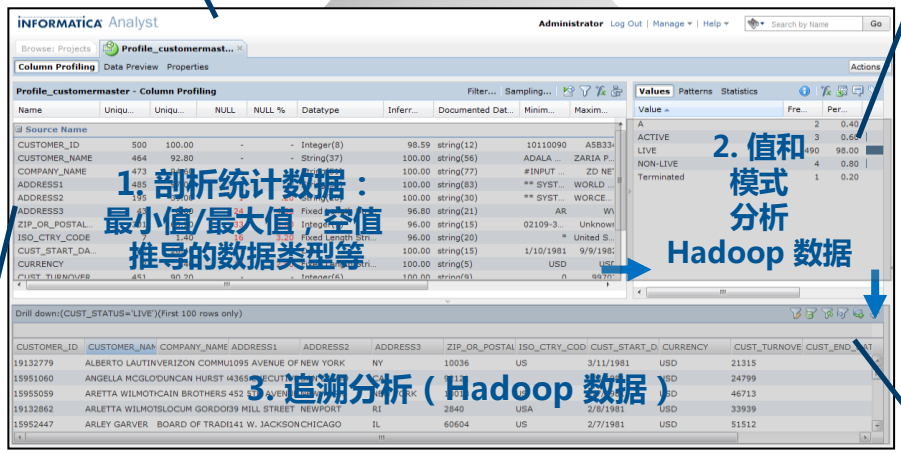
Hadoop 数据剖析结果

Hadoop 数据剖析结果 ——通过浏览器 接触企业中的任何人员

值和模式频率与不一致的/脏数据或意外模式隔离

CUSTOMER_ID example

Statistic	Value
Maximum Length	8
Minimum Length	6
Bottom (5)	10110090
	10110091
	10110092
	10110122
	10110124
Top (5)	A5B334
	A44563
	A23456
	19134136
	19134134



国家代码示例

Value	Fre...	Per...
NULL	16	3.20
United States	2	0.40
USA	8	1.60
US	464	92.80
U.S.A.	6	1.20
U.S.	3	0.60
*	1	0.20

Value	Fre...	Per...
NULL	33	6.60
Unknown	4	0.80
N/A	2	0.40
999999999	1	0.20
999999	2	0.40
98101	1	0.20
98006	1	0.20
98005	1	0.20
97210-3676	1	0.20
95135	1	0.20
94903	1	0.20

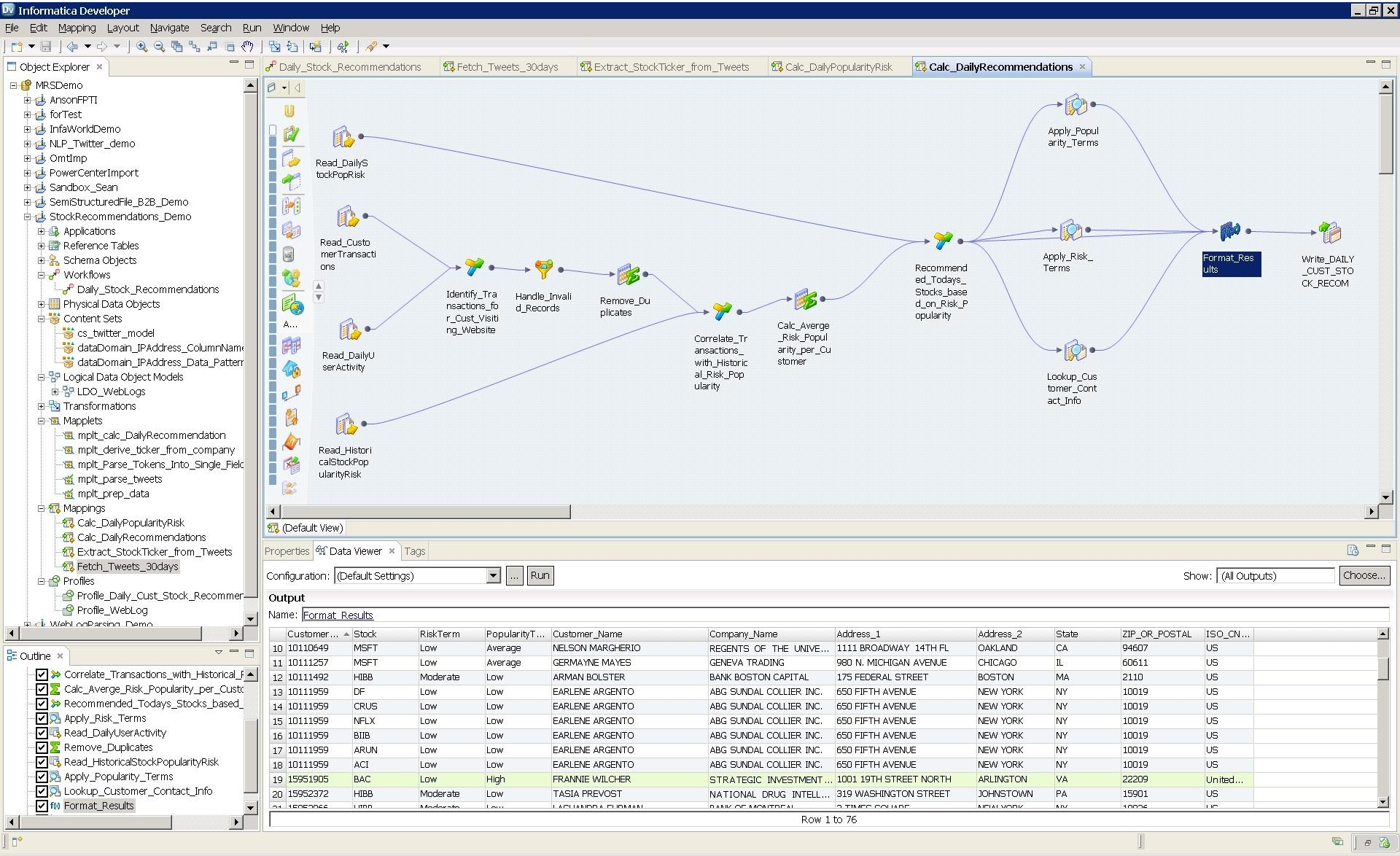
邮政编码示例

标识数据中的异常和反常现象的统计数据

IN(GORDY SPAROW G

CUSTOMER_ID	CUSTOMER_NAME	COMPANY_NAME	ADDRESS1	ADDRESS2	ADDRESS3	ZIP_OR_POSTAL	ISO_CTRY_C
10110239	GORDIE SPARROW	MCGREGOR GROUP	1740 BROADWAY	NEW YORK	NY	10019	US
10116657	GORDON SPARROW	MCcGREGOR GRP	1740 BRDWY	NY	NY	10019	US
10178890	GORDY SPAROW	UNKNOWN	BROADWAY	NEW YORK	NY	10019	USA

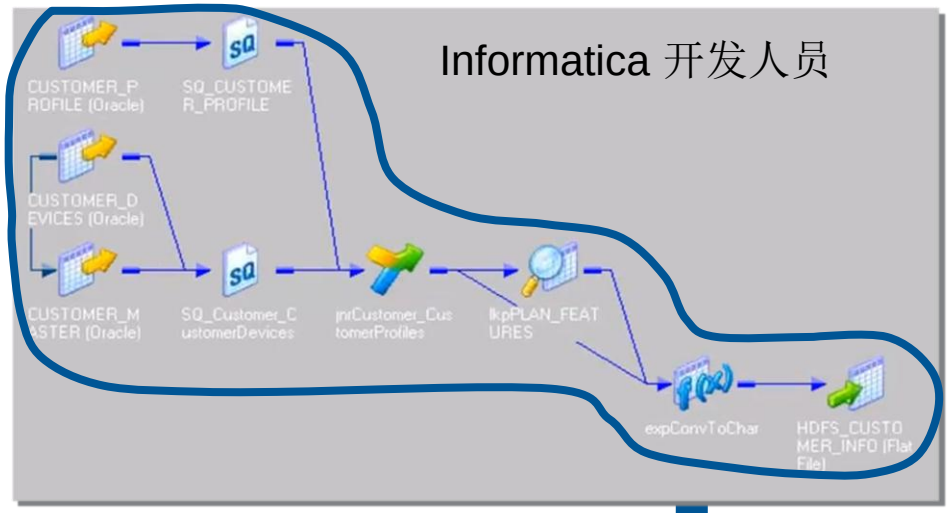
追溯实际数据值来检验整个数据集中的结果，包括可能的重复



	Customer...	Stock	RiskTerm	PopularityT...	Customer_Name	Company_Name	Address_1	Address_2	State	ZIP_OR_POSTAL	ISO_CN...
10	10110649	MSFT	Low	Average	NELSON MARGHERIO	REGENTS OF THE UNIVE...	1111 BROADWAY 14TH FL	OAKLAND	CA	94607	US
11	10111257	MSFT	Low	Average	GERMAYNE MAYES	GENEVA TRADING	980 N. MICHIGAN AVENUE	CHICAGO	IL	60611	US
12	10111492	HIBB	Moderate	Low	ARMAN BOLSTER	BANK BOSTON CAPITAL	175 FEDERAL STREET	BOSTON	MA	2110	US
13	10111959	DF	Low	Low	EARLENE ARGENTO	ABG SUNDAL COLLIER INC.	650 FIFTH AVENUE	NEW YORK	NY	10019	US
14	10111959	CRUS	Low	Low	EARLENE ARGENTO	ABG SUNDAL COLLIER INC.	650 FIFTH AVENUE	NEW YORK	NY	10019	US
15	10111959	NFLX	Low	Low	EARLENE ARGENTO	ABG SUNDAL COLLIER INC.	650 FIFTH AVENUE	NEW YORK	NY	10019	US
16	10111959	BIIB	Low	Low	EARLENE ARGENTO	ABG SUNDAL COLLIER INC.	650 FIFTH AVENUE	NEW YORK	NY	10019	US
17	10111959	ARUN	Low	Low	EARLENE ARGENTO	ABG SUNDAL COLLIER INC.	650 FIFTH AVENUE	NEW YORK	NY	10019	US
18	10111959	ACI	Low	Low	EARLENE ARGENTO	ABG SUNDAL COLLIER INC.	650 FIFTH AVENUE	NEW YORK	NY	10019	US
19	15951905	BAC	Low	High	FRANNIE WILCHER	STRATEGIC INVESTMENT...	1001 19TH STREET NORTH	ARLINGTON	VA	22209	United...
20	15952372	HIBB	Moderate	Low	TASIA PREVOST	NATIONAL DRUG INTELL...	319 WASHINGTON STREET	JOHNSTOWN	PA	15901	US

Informatica Hadoop 路线图

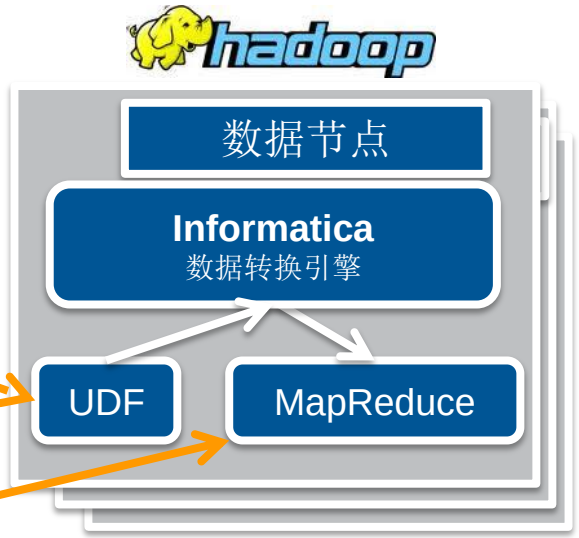
Hadoop MapReduce 处理



1. Informatica 映射转换成优化的 Hive HQL和用户自定义功能
2. 优化的 HQL 转换为 MapReduce
3. 在 Hadoop 上执行 MapReduce 用户自定义功能

```
SELECT
    T1.ORDERKEY1 AS ORDERKEY2, T1.li_count, orders.O_CUSTKEY AS CUSTKEY, customer.C_NAME,
    customer.C_NATIONKEY, nation.N_NAME, nation.N_REGIONKEY
    FROM
                    (
                    SELECT TRANSFORM (L_Orderkey.id) USING CustomInfaTx
                    FROM lineitem
                    GROUP BY L_ORDERKEY
                    ) T1
                    JOIN orders ON (customer.C_ORDERKEY = orders.O_ORDERKEY)
                    JOIN customer ON (orders.O_CUSTKEY = customer.C_CUSTKEY)
                    JOIN nation ON (customer.C_NATIONKEY = nation.N_NATIONKEY)
                    WHERE nation.N_NAME = 'UNITED STATES'
                    ) T2
    INSERT OVERWRITE TABLE TARGET1 SELECT *
    INSERT OVERWRITE TABLE TARGET2 SELECT CUSTKEY, count(ORDERKEY2) GROUP BY
    CUSTKEY;
```

Hive HQL



Entire mapping logic (all transformations) can be executed on Hadoop

Informatica HParser
处理各种各样的大数据

最广范围的大数据

平面文件和文档

定位

名称 = 价值

^/>限定<¥^

XML

行业标准

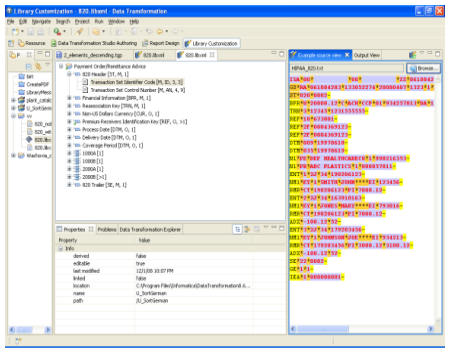
交互数据

社交

科学设备/传感器

生产力

- 直观解析环境
- 预定义转换



任何 DI/BI 体系架构

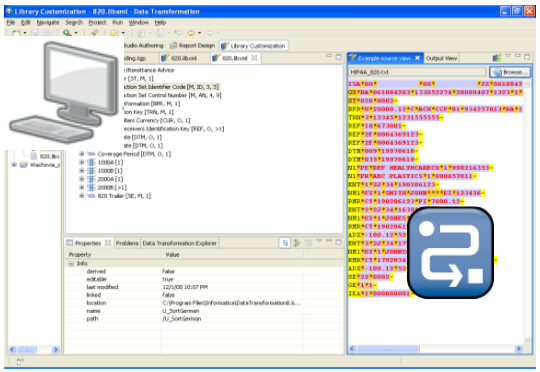
HIVE

PIG

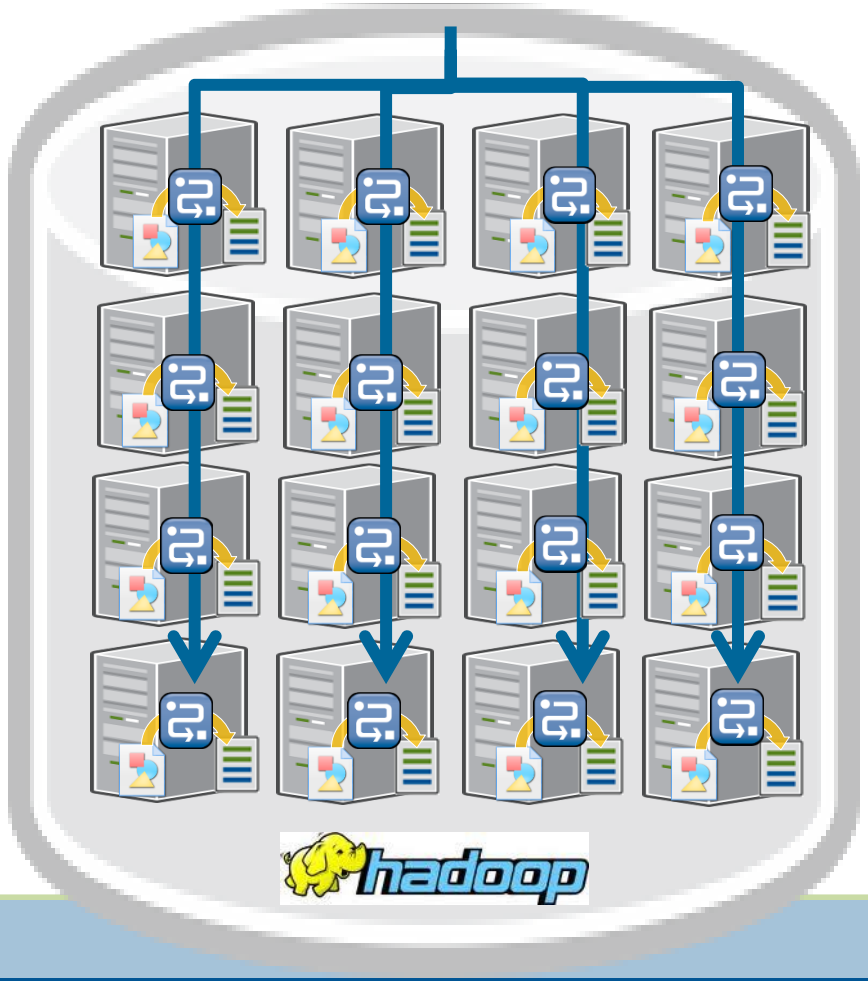
EDW

MDM

在Hadoop上解析和准备数据
工作原理如何?



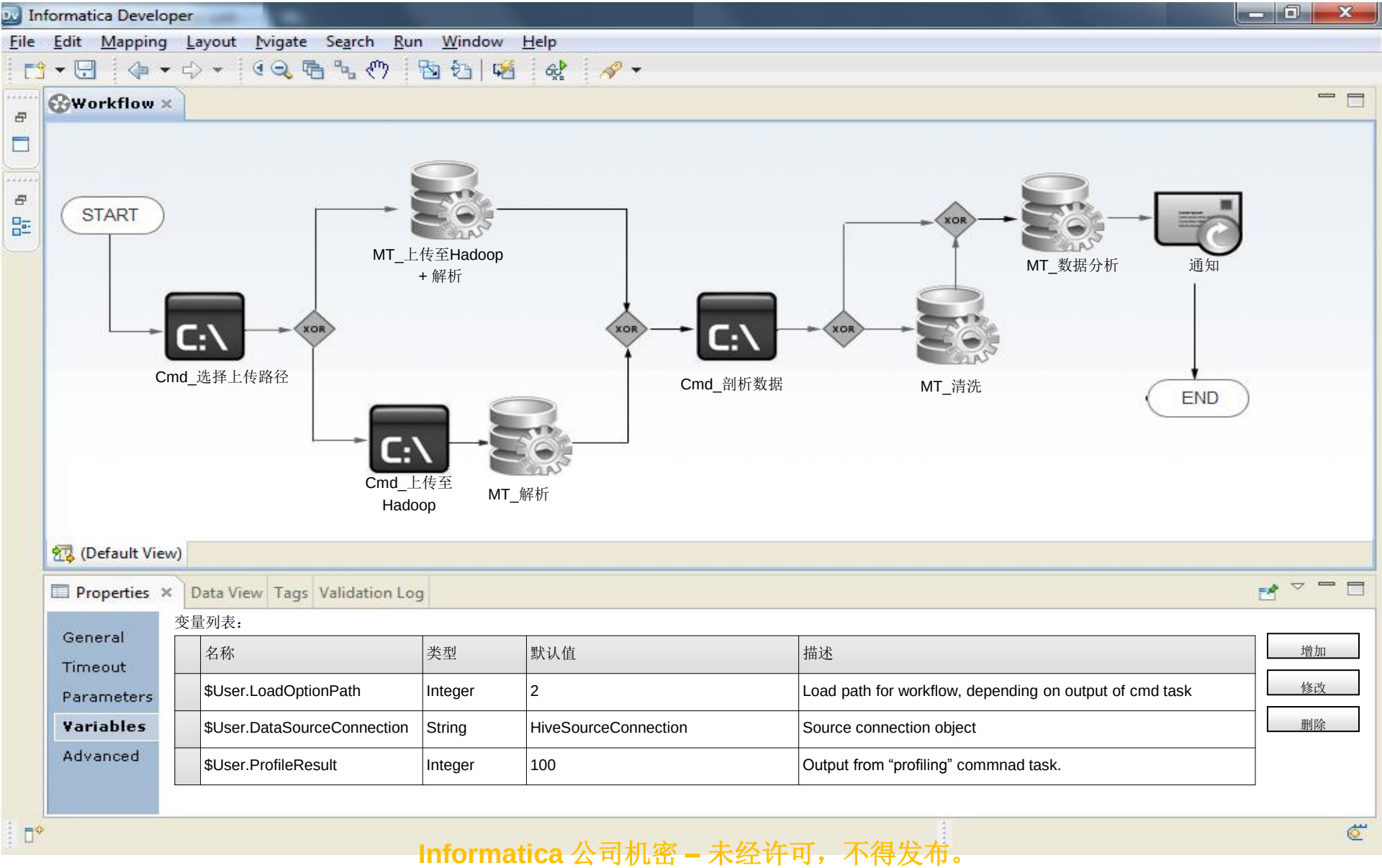
```
hadoop … dt-hadoop.jar
… My_Parser /input/*/input*.txt
```



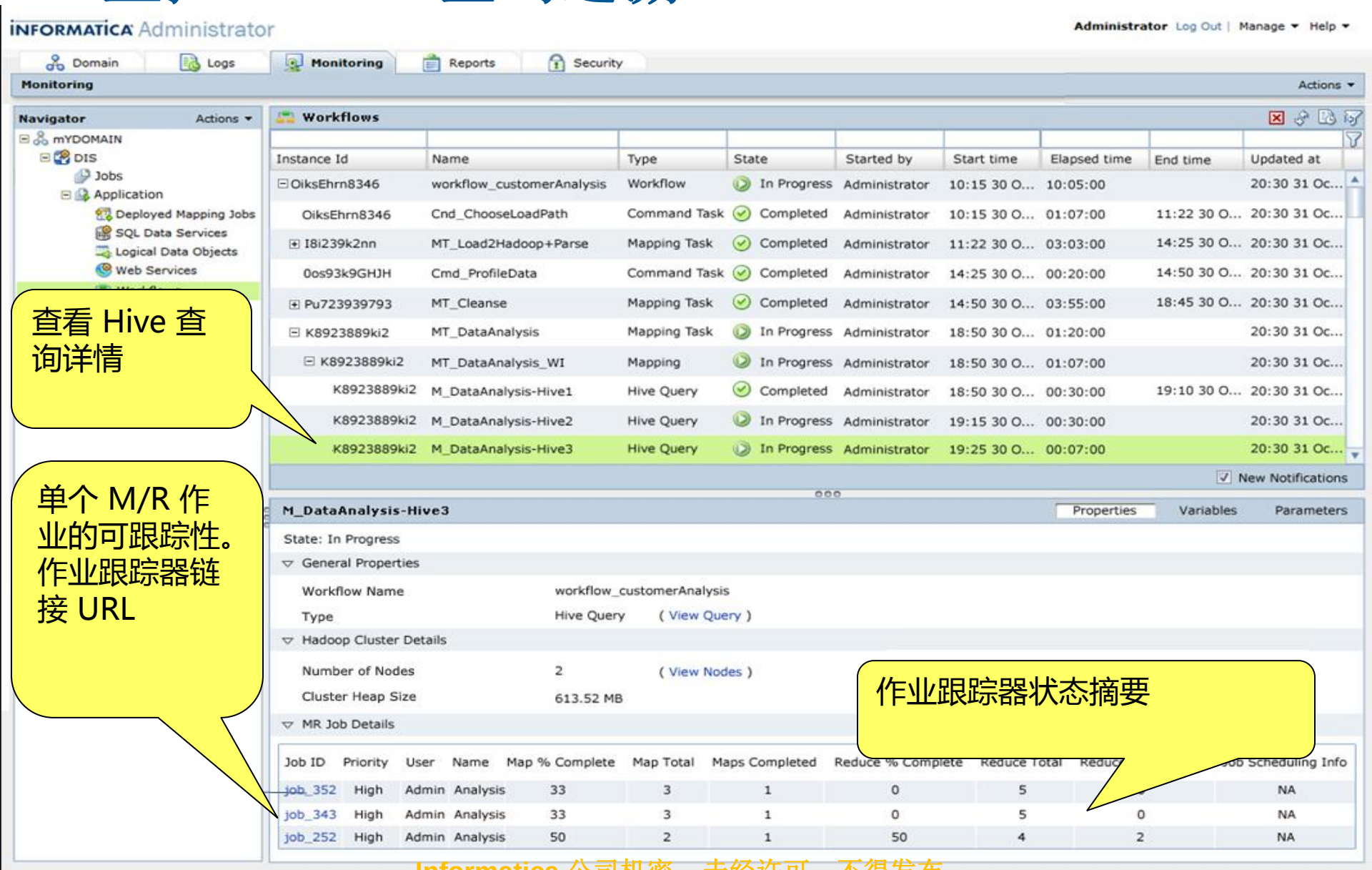
1. 在 HParser 可视化工作室中定义解析器
2. 在 Hadoop 分布式文件系统 (HDFS) 上部署解析器
3. 运行 HParser 提取数据，并在 Hadoop 产生表格格式

混合工作流编排

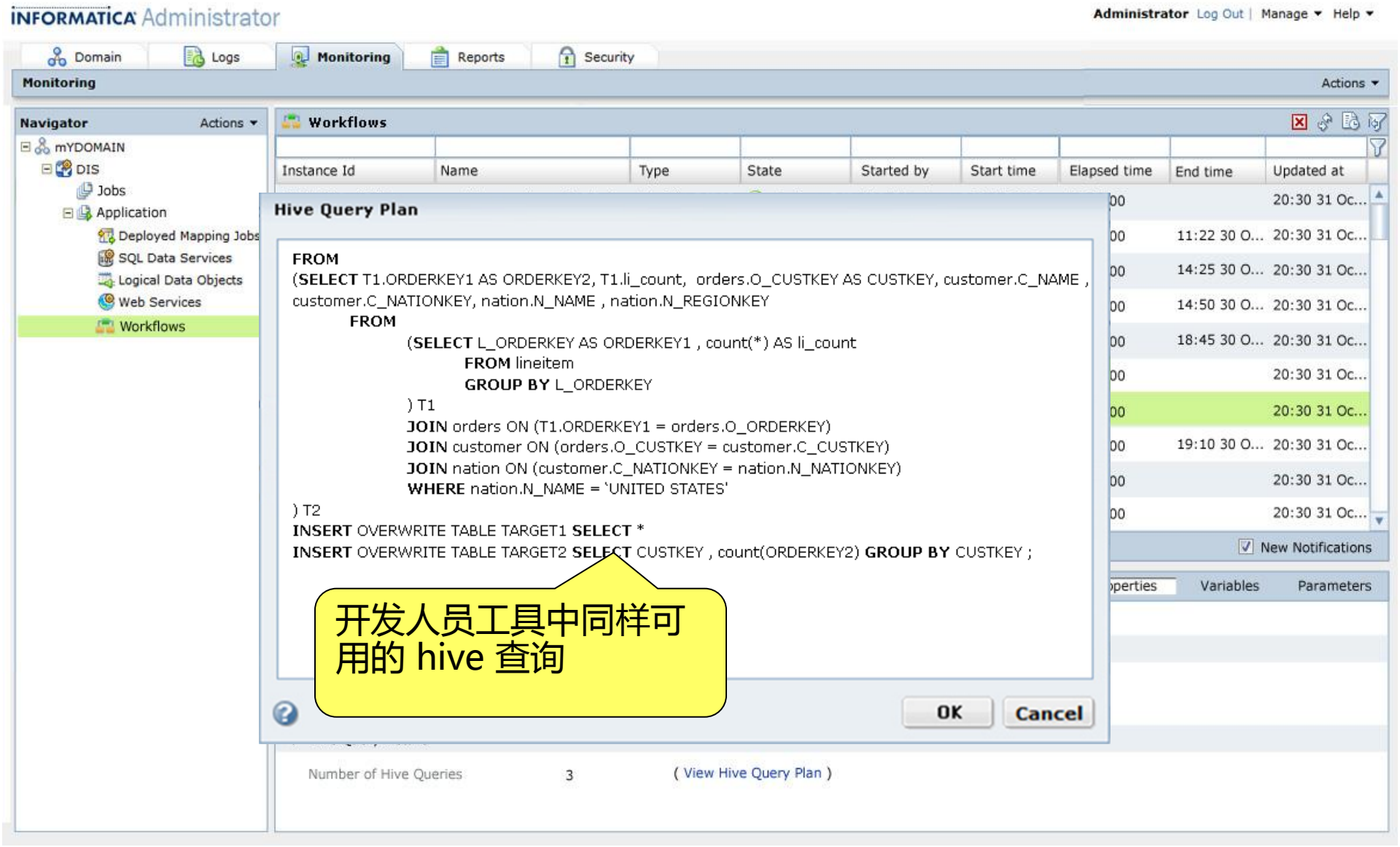
在Hadoop和本地环境中运行任务工作流同一



监控 – Hive 查询追溯 M/R

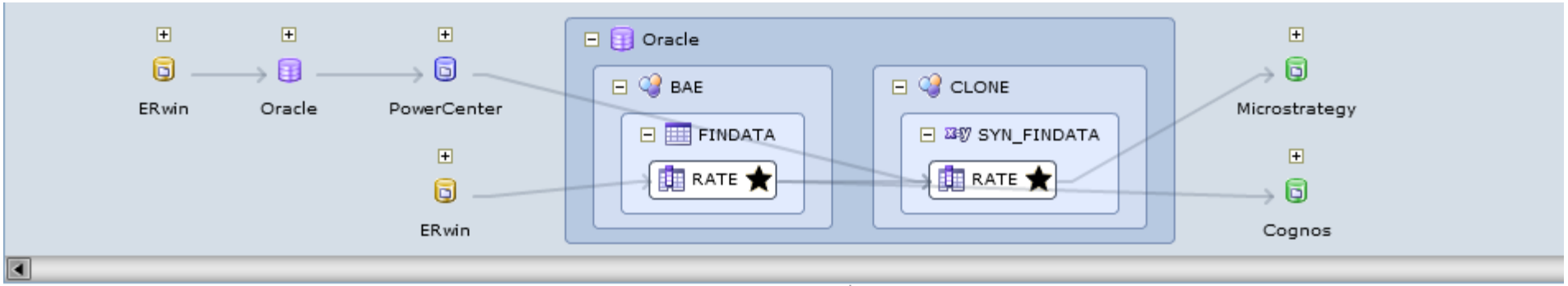


监控 – Hive 查询计划详情



数据沿袭和业务术语表

元数据管理路线图



RATE

Impact Summary - Downstream

Class	Name	Location
Report	ContractorsByDepartment	MM/Cognos/content/FinancialDemoData/Reports/ContractorsByDepartment
Microstrategy Report	Contractors by Department	MM/Microstrategy/Financial/Public Objects/Reports/Contractors by Department
Oracle Synonym	SYN_FINDATA	MM/Oracle/CLONE/Synonyms/SYN_FINDATA
Oracle View	V_FINDATA	MM/Oracle/BAE/Views/V_FINDATA

Impact Summary - Upstream

Class	Name	Location
Mapping	m_Fin_Fact_Table_xml	MM/PowerCenter/Metadata Manager - MM_Financial_dm/Mappings/m_Fin_Fact_Table_xml
Oracle Procedure	COMPLICATED_STUFF	MM/Oracle/BAE/Procedures/COMPLICATED_STUFF
Oracle Synonym	SYN_FINDATA	MM/Oracle/CLONE/Synonyms/SYN_FINDATA
Oracle Table	TRANSACTIONS	MM/Oracle/BAE/Tables/TRANSACTIONS
Table	FINDATA	MM/ERwin/Model_4/Tables/FINDATA
Table	TRANSACTIONS	MM/ERwin/Model_4/Tables/TRANSACTIONS

先进技术转化为常规IT部署

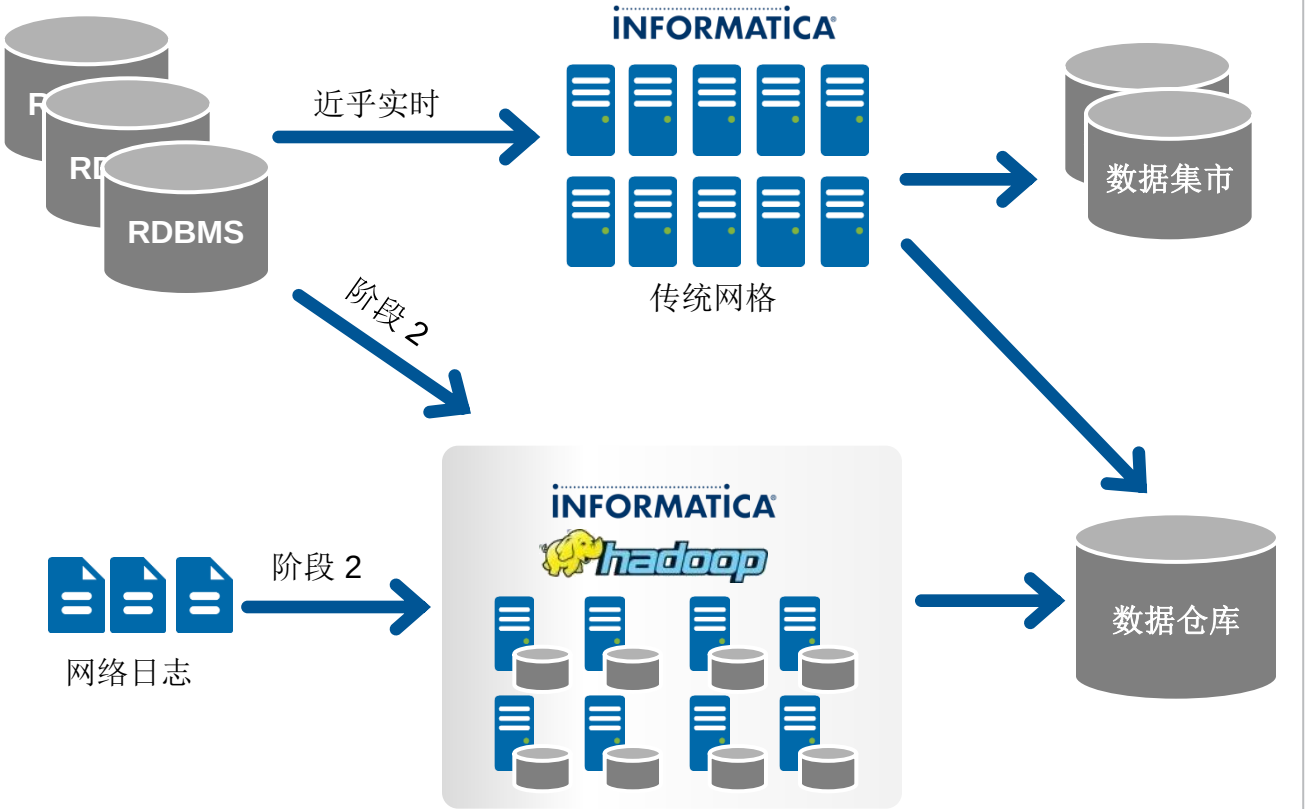
- **重复性**
  - 可预测、可重复的部署和方法
- **与快速的 Hadoop 变化隔离**
  - 经常推出新版本和项目
  - 避免对错误的技术下注
- **现有资产的重复使用**
  - 应用现有集成逻辑向 Hadoop 加载数据
  - 重新使用现有数据质量规则验证 Hadoop 数据
- **现有技能的重复使用**
  - 使 ETL 开发人员能够利用 Hadoop 的功能
- **治理**
  - 执行并验证数据安全性、数据质量和法规遵从政策
  - 可管理

扩展ETL 并控制成本

为大数据分析奠定基础

挑战：随着数据量和处理负荷的迅速增长，对更快的数据驱动型决策的需求不断增加

解决方案



结果

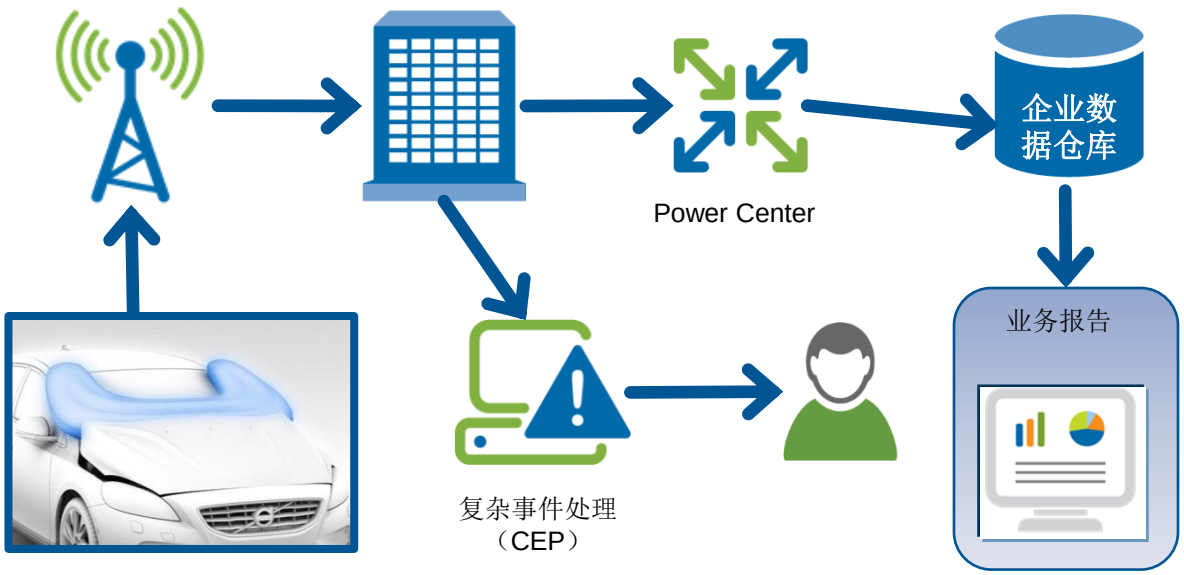
- 经济高效地拓展性能
- 降低硬件成本
- 通过在统一数据集成平台上的标准化，增加了灵活度

大型国际金融机构

互联车辆项目
开发创新产品和服务

挑战：为“互联车辆”计划，在年底前实现实时收集汽车数据

解决方案



- 持续收集所有车辆的所有信息
- 所有车辆在年底时，都将把数据传送到中央Teradata数据仓库
- 利用PowerCenter, CDC和CEP 实现实时数据集成

结果

- 助力实现互联车辆的目标：
  - 嵌入移动技术提升客户体验
  - 预测维修维护和提高燃料效率
  - 电话道路救援和自动调度服务

大型国际汽车制造商

PowerCenter 大数据版

降低大数据项目成本

加速创新产品和服务的上市速度

将采用新技术的风险降至最低

将 Hadoop扩展至整个企业

Informatica助您实现大数据的最大回报

www.informatica.com.cn