

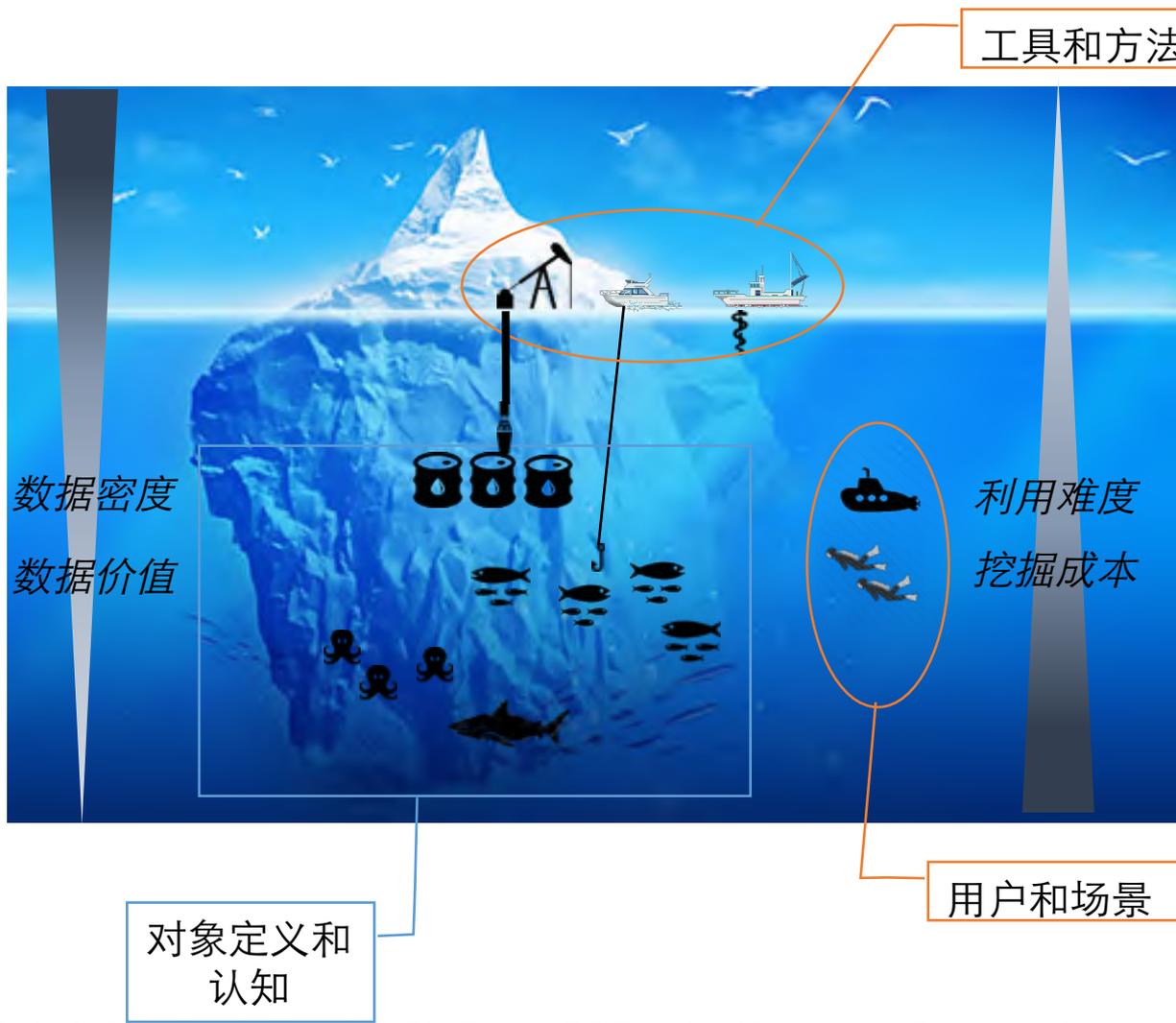
工业大数据的落地选择与分析实践

报告人：李一名

大数据落地的一些疑问

- 先建平台还是先做应用？
- 选择新的算法还是获取新的数据？
- 解决老问题还是发现新价值？

大数据冰山



如何进入良性循环？

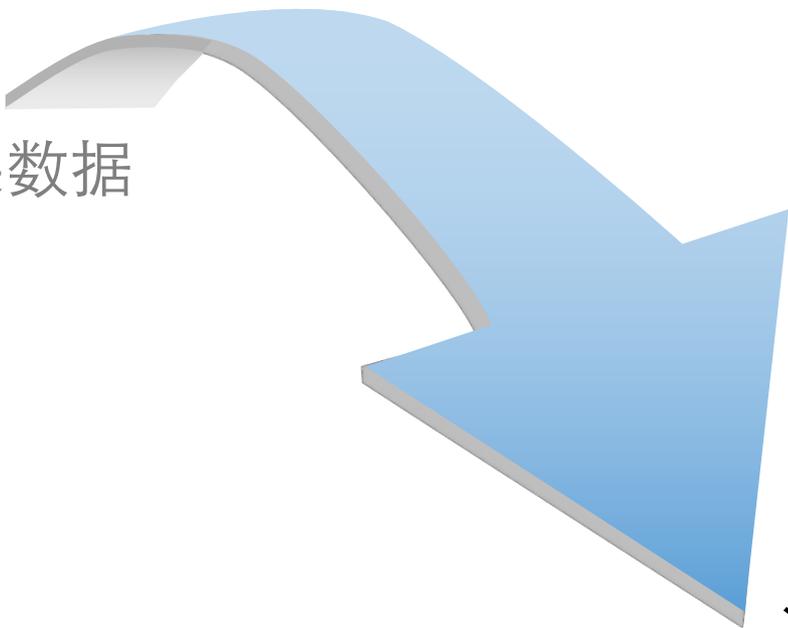
- 大数据必须附着于业务
- 大数据需专注于核心和盈利领域
- 基于大数据的自动化决策是未来主要方向

具体实践

- 平台：从数据仓库到大数据
- 应用：
 - 预测不可测结果，实现优化调整
 - 提升识别精度，实现智能自动判定
- 产品：解决技术的最后一公里

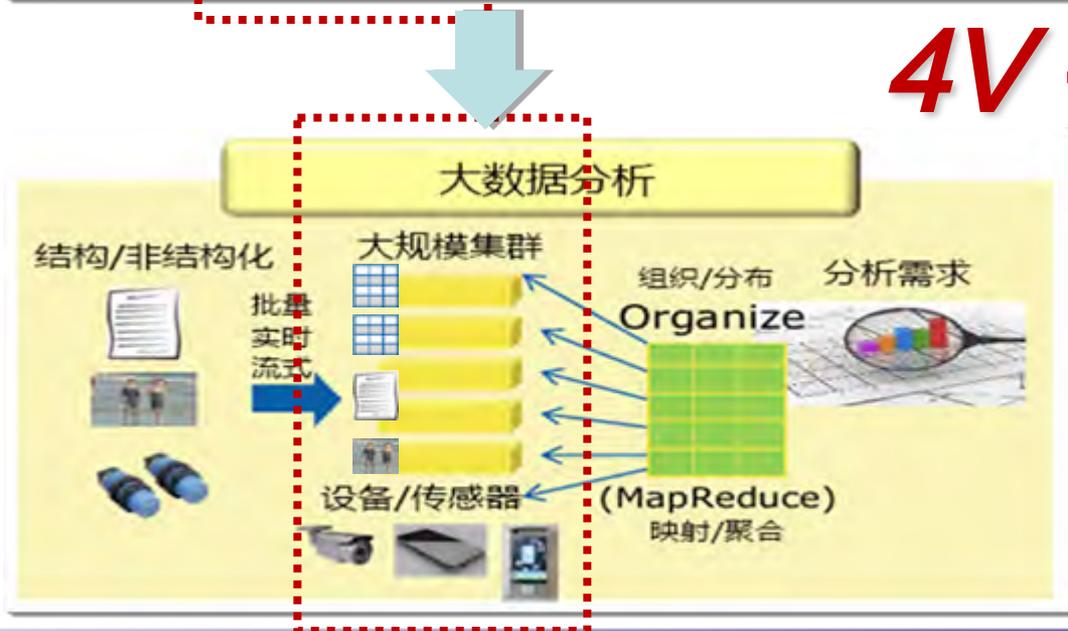
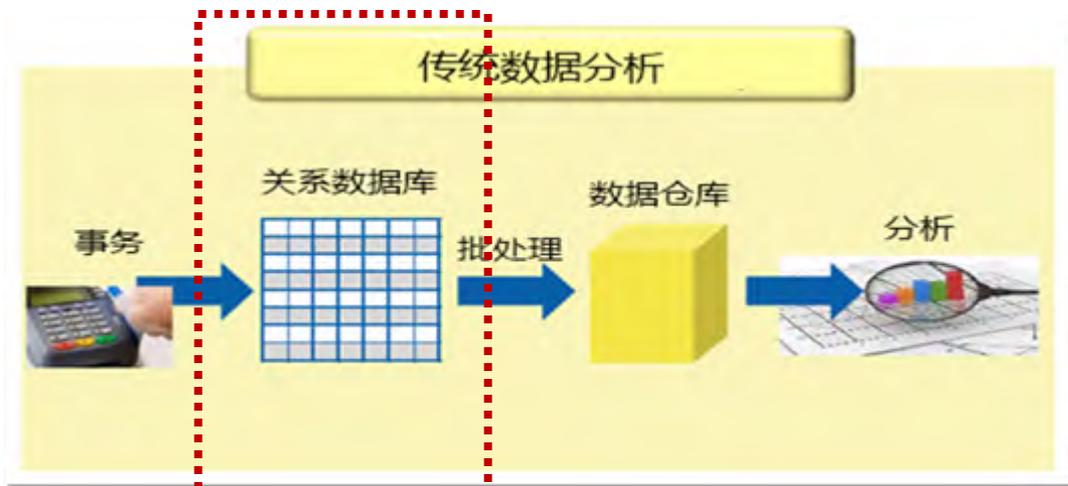


结果数据



过程数据

数据仓库的分析对象都是管理的结果数据、业务的状态数据，对这些数据的统计分析在一定程度上有效的帮助管理者把握企业运营态势，使决策有据可依。然而，环境、业务、人员都处于不断的变化之中，数据呈现的状态结果并不可靠，也不全面，决策者并不具有真正敏锐的洞察力。充分利用各领域的过程数据，挖掘潜在规律，突破固有认识，将推动商业、管理、技术跨越式创新。



4V →

更大的数据量
(Volume)

TB(10^{12} B) → PB(10^{15} B)

更及时地数据更新
(Velocity)

天 → 时/分

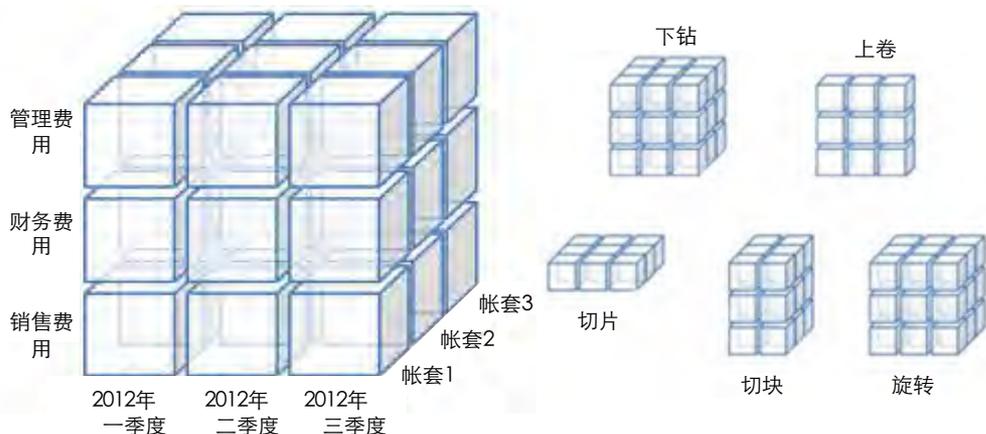
更丰富的数据类型
(Variety)

结构化(传统数据) → 结构化、半结构化、非结构化

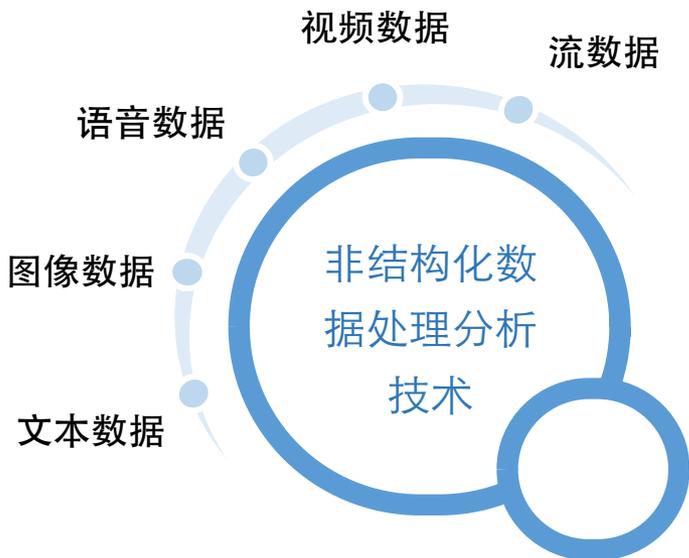
更经济的成本
(Value)

高端设备 → 低端并行

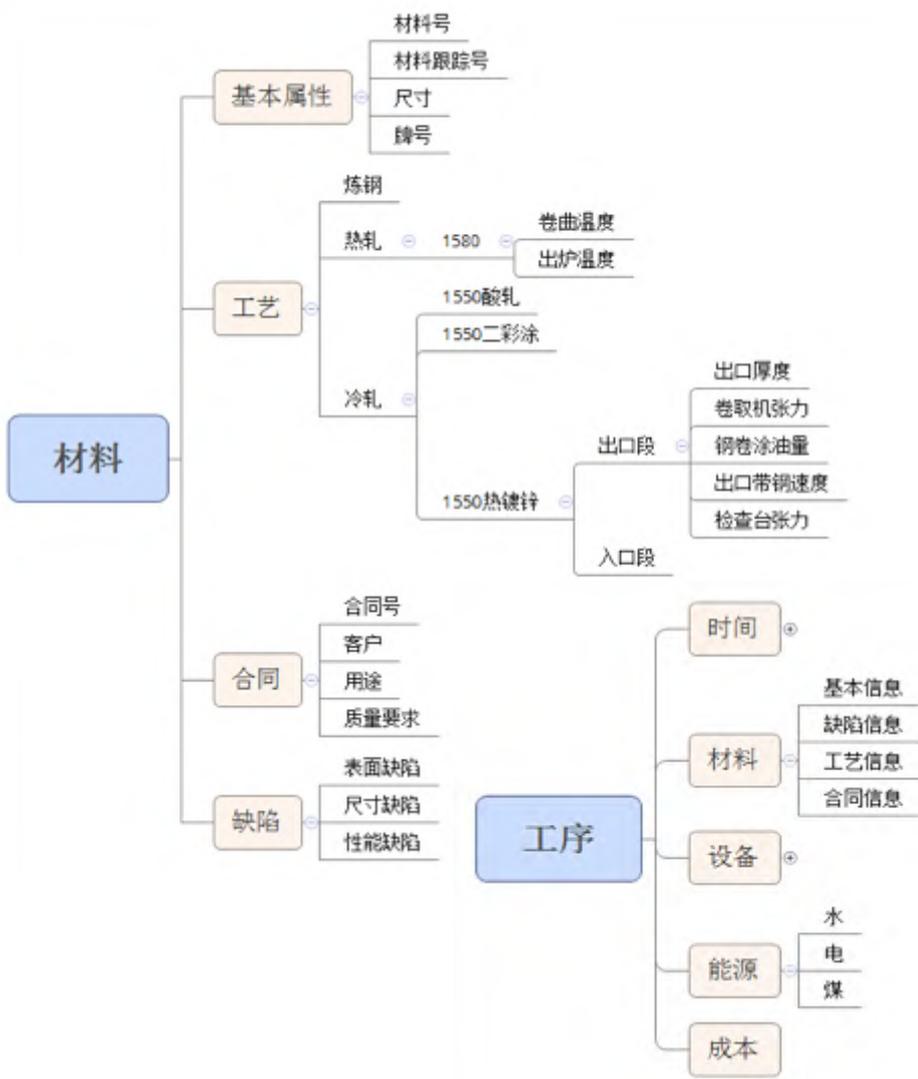
结构化数据OLAP分析



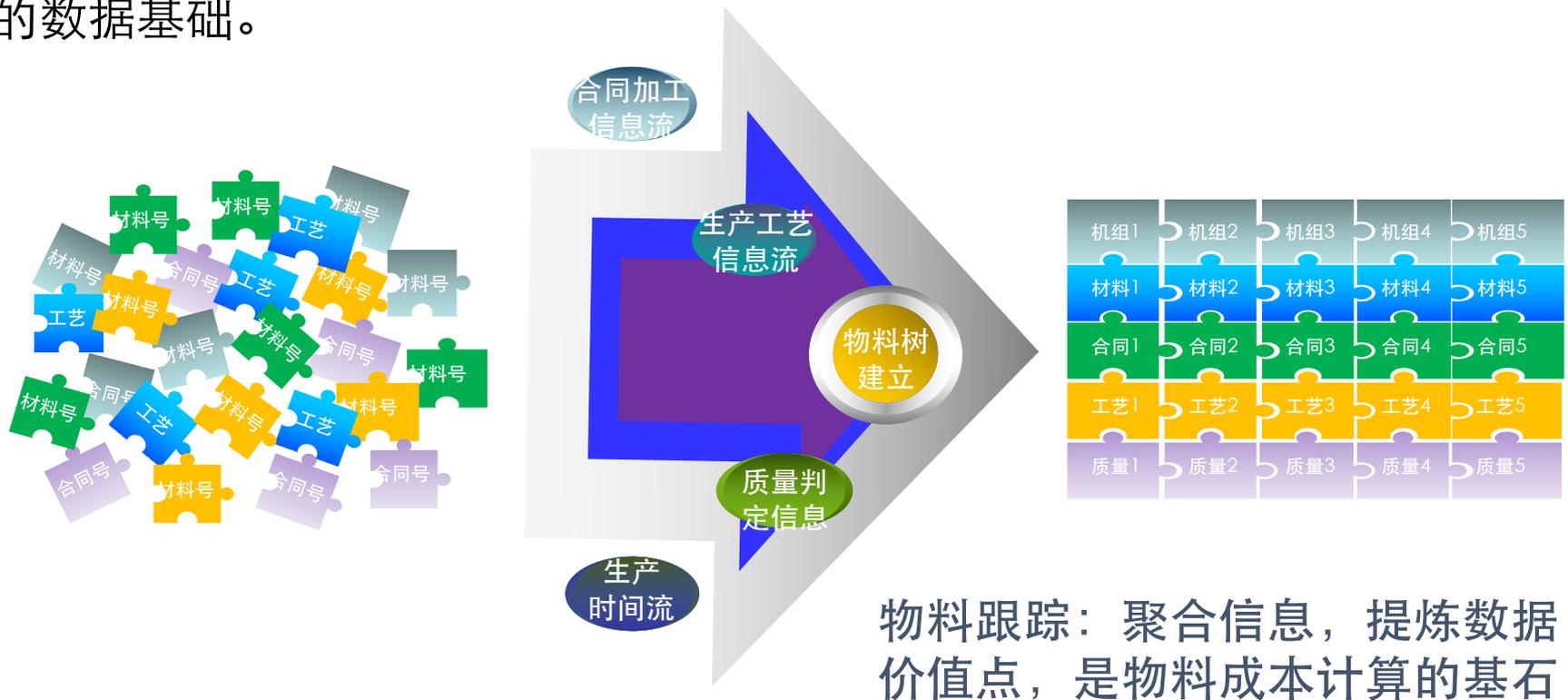
统计分析与数据挖掘



面向工业对象的大数据数据模型构建

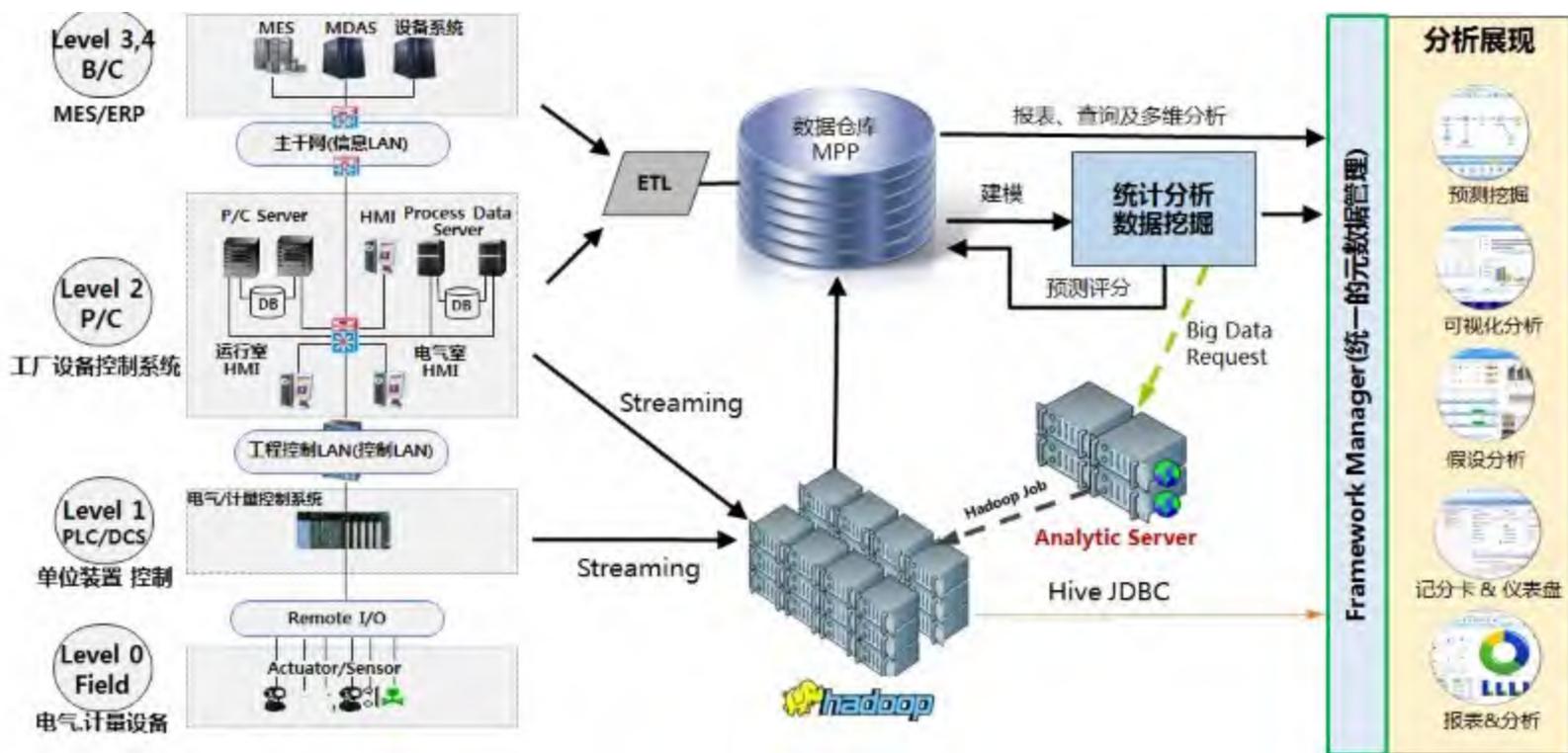


设计“亲属码”，编制物料“家谱”，在此基础上，融合每个工序的工艺信息、合同生产信息和质量判定信息等，将各个机组“孤立”信息，按照实物生产过程进行串接，并通过动态增量更新技术，构建物料信息流，形成整个项目的数据库基础。

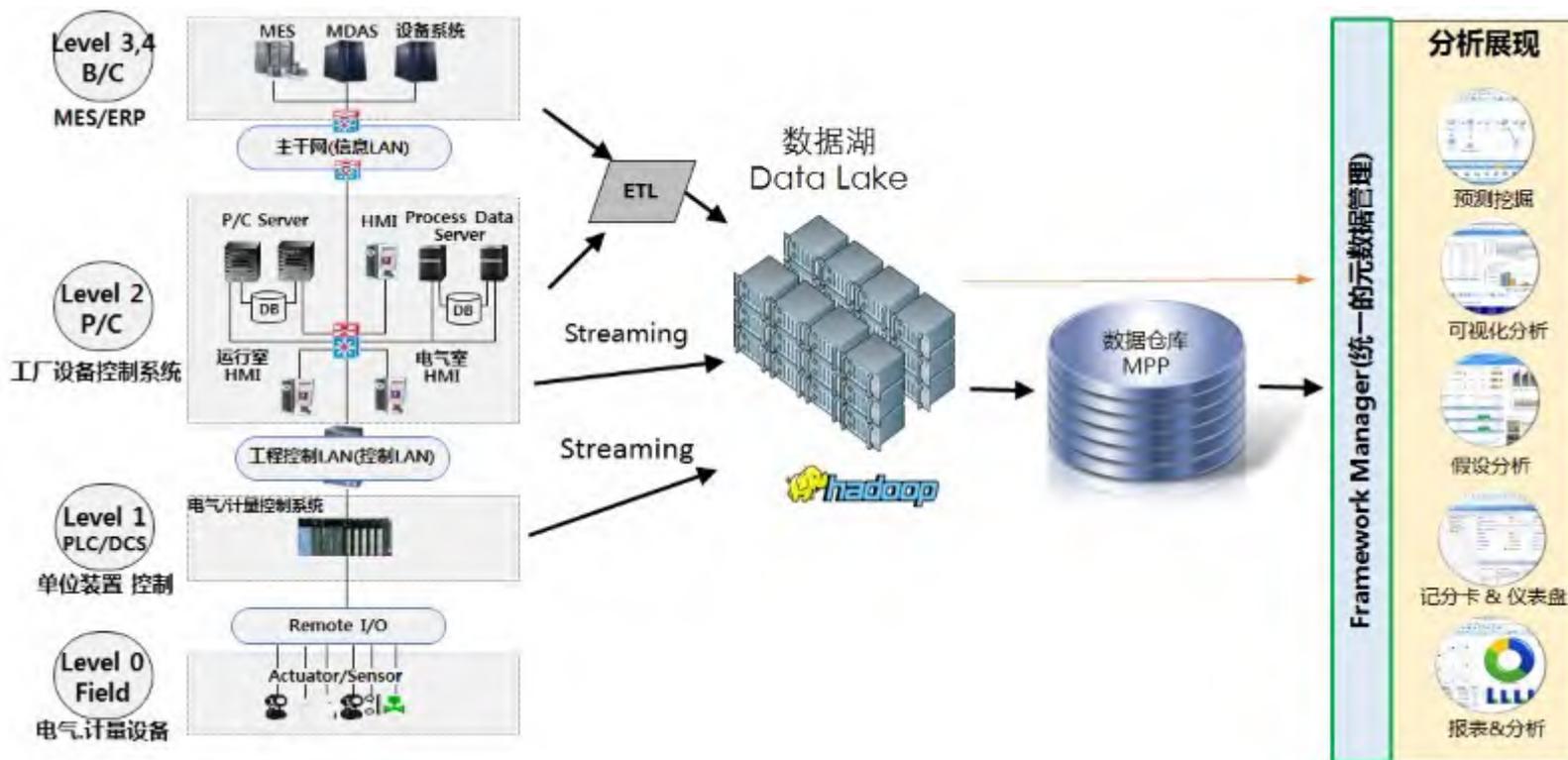


数据量大，物流产线交错，生产情况复杂多样，如回退（变号、不变号）多次，材料新增、删除、判废、修改等逻辑处理异常复杂。

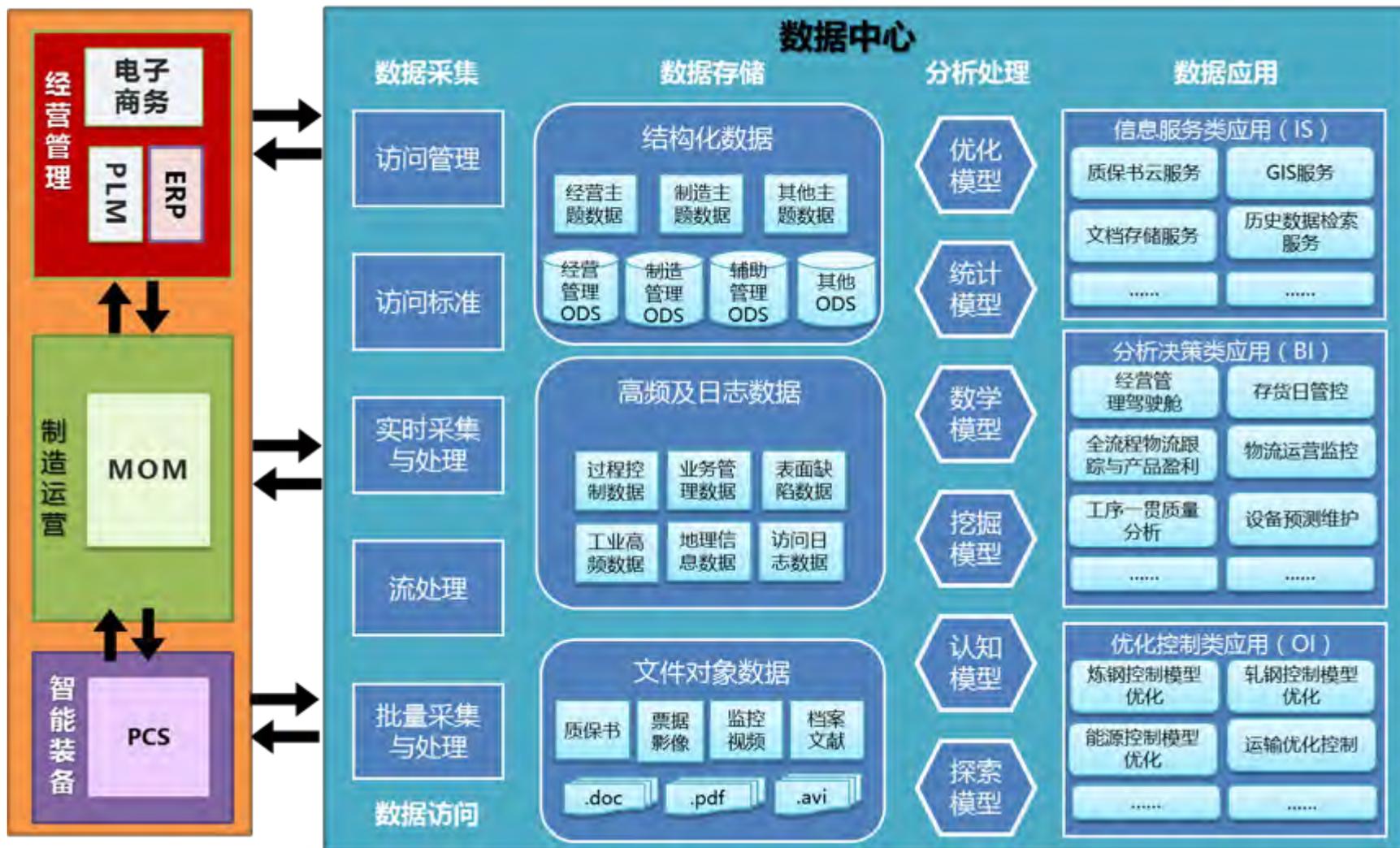
模式一：数据仓库和大数据平台的混合架构



模式二：从数据湖到数据仓库的层次架构



企业大数据中心应用规划

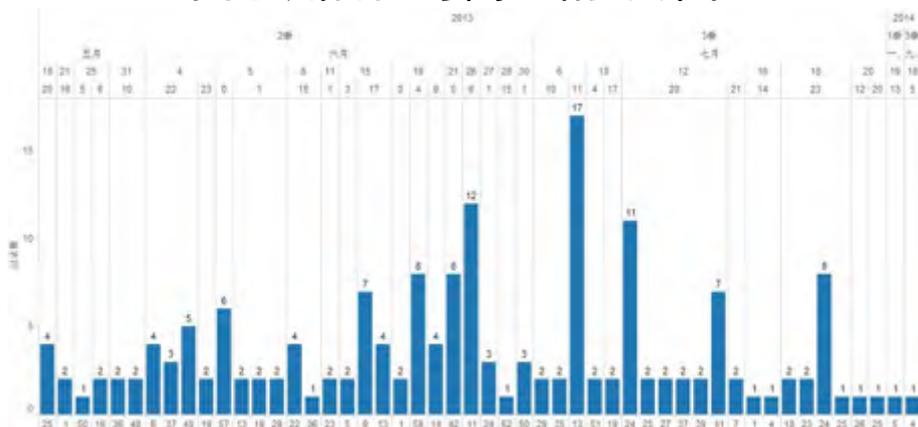


质量缺陷大数据分析逻辑：

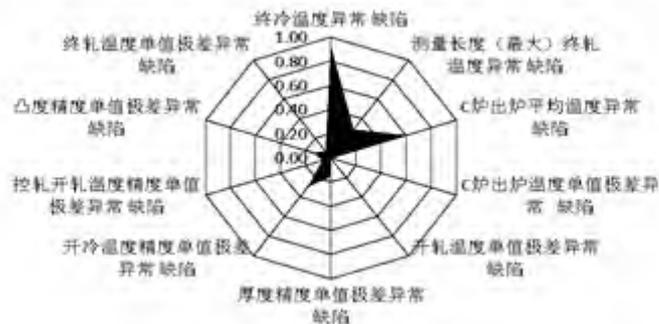
产品质量缺陷发生在特定的时空中，这个时空中的各类因素一定在某个微观程度上与无缺陷生产时空因素存在显著差异。

- ✓ 不同缺陷在时间上的聚集性
- ✓ 不同缺陷在不同因素上的分布倾向
- ✓ 不同缺陷在不同成份含量上的分布倾向
- ✓ 不同缺陷在成份含量上的共同特征
- ✓ 过程异常与缺陷的相关性

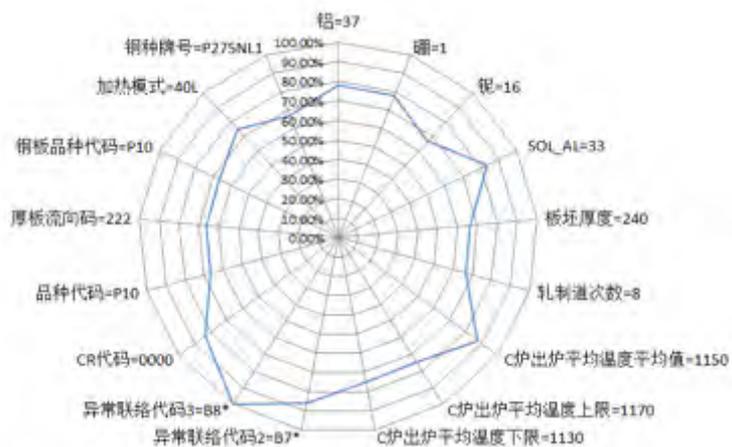
不同缺陷在时间上的聚集性



过程异常与缺陷的相关性



不同缺陷与特定因素间的相关性



不同缺陷在成份含量上的共同特征

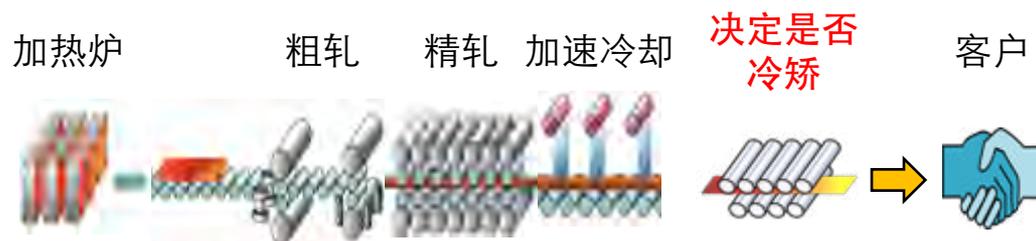
缺陷代码	缺陷名称	总条数	最大相同项集
006	粘膏	7	AS=2, CU=1
007	氧化铁皮剥落	12	CU=1
025	下表面边裂	21	CU=1
040	角部翘皮(上表传动侧)	15	CU=1
051	轧制异物压入	8	ZR=1
063	热矫压痕	58	AS=2
162	端部开裂(尾部)	9	CU=1,ZR=1
203	厚度超上限	18	AS=2, CU=1, NI=1
250	瓦楞浪	8	CU=1, NI=1, CR=3, MO=1
254	全宽翘曲	5	CU=1, NI=1, AS=2
406	上表折叠	5	ZR=1

- 目标：以未进行冷矫的子板为对象，建立预测模型，通过预测需要进行冷矫的厚板，降低冷矫率，同时减少因残余应力引起的客户质量异议。

- 分析对象

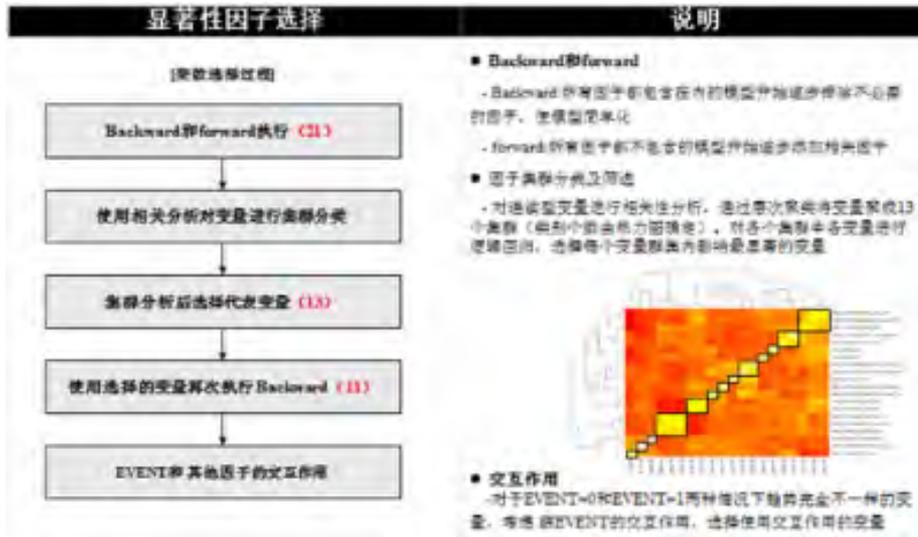
- 牌号：选择一个具体型号产品
- 时间段：2015.01 ~ 2015.07
- 数据：加热炉信息表/轧制信息表/热矫冷矫信息表+客户质量异议数据

- 厚板制造工艺流程图



显著因子间相关性分析

- 对导出的55个候补因子进行逐步逻辑回归，按照backward方法剔除变量和按照forward方法添加变量，得到最终进入模型的变量因子



使用交叉验证，选择最优模型，良品命中率为 92%，不良命中率为 77.8%:

$$Y = \frac{\exp(f(x))}{1 + \exp(f(x))} = \frac{1}{1 + \exp(-f(x))}$$

阈值:0.015（大于阈值则为不良）

$$f(x) = -7.063 + 0.02051 * X1 - 0.04188 * X2 - 18.83 * X3 + 0.001405 * X4 + 0.0303 * X5 + 0.003801 * X6 + 0.05597 * X7 - 7.178 * X8$$

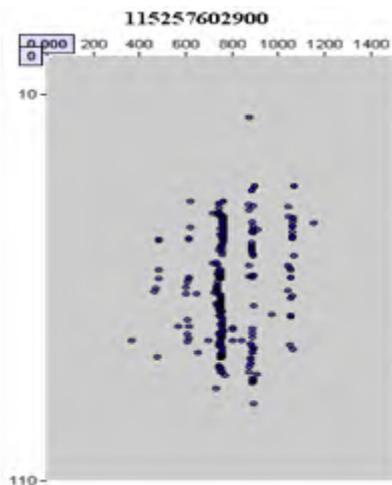
✓ 3个热轧机组保护渣的缺陷特征分析

- 时间：缺陷记录时间201501-201505
- 缺陷类型：HD钢质保护渣系夹渣
- 分析数据包含钢卷（个数）：162446、缺陷（个数）：4361904

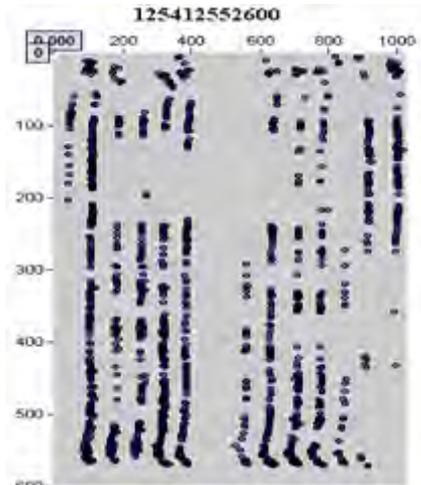
✓ 形成9个特征描述维度

- 面别倾向性、传动侧分布集中度、操作侧分布集中度、中心线位置分布集中、头部分布集中度、尾部分布集中度、长度方向分布集中度、宽度方向分布集中度、网格密度

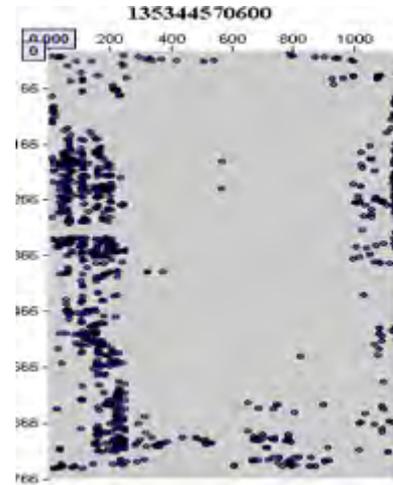
✓ 完成16万个热卷的细化特征描述，并可提供基于特征的搜索查询



具有上表面缺陷集中，中心线位置分布集中，长度方向分布集中，网格分布密度较小（平原）4个特征



具有传动侧分布集中，操作侧分布集中，长度方向分布集中，宽度方向分布集中，网格分布密度很高（高原）5个特征

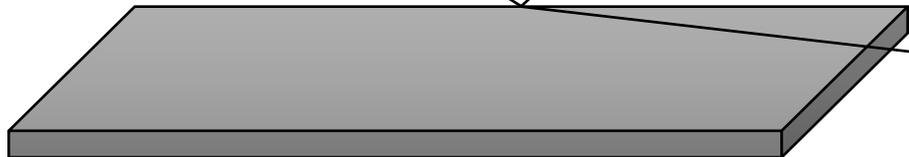
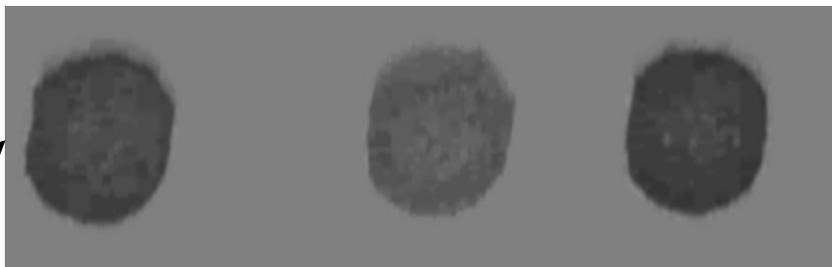


具有上表面缺陷集中，传动侧分布集中，操作侧分布集中，长度方向分布集中，网格分布密度一般（丘陵）5个特征

划痕



油污



过酸洗



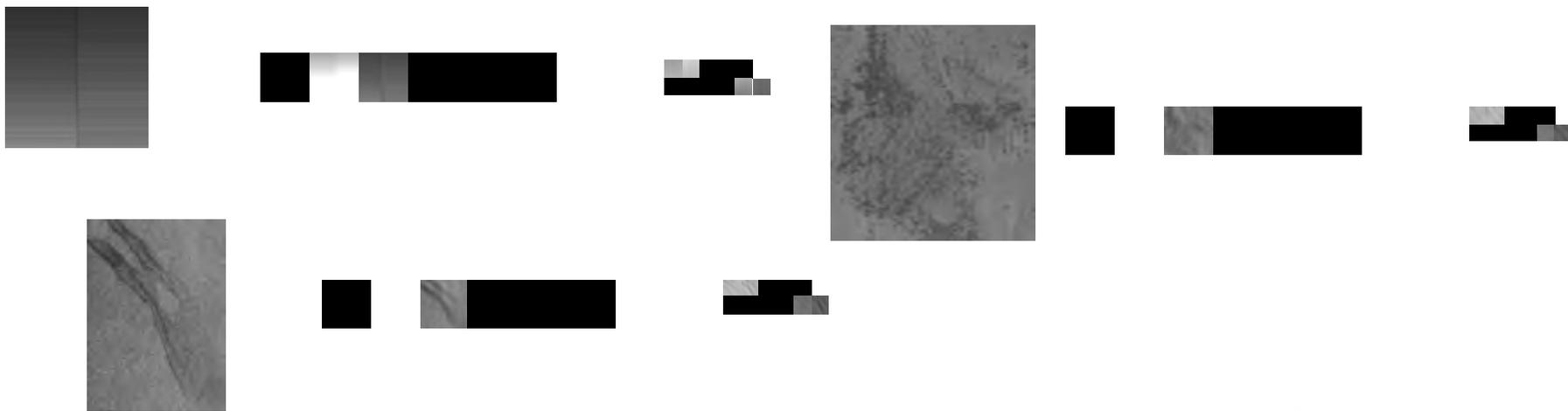
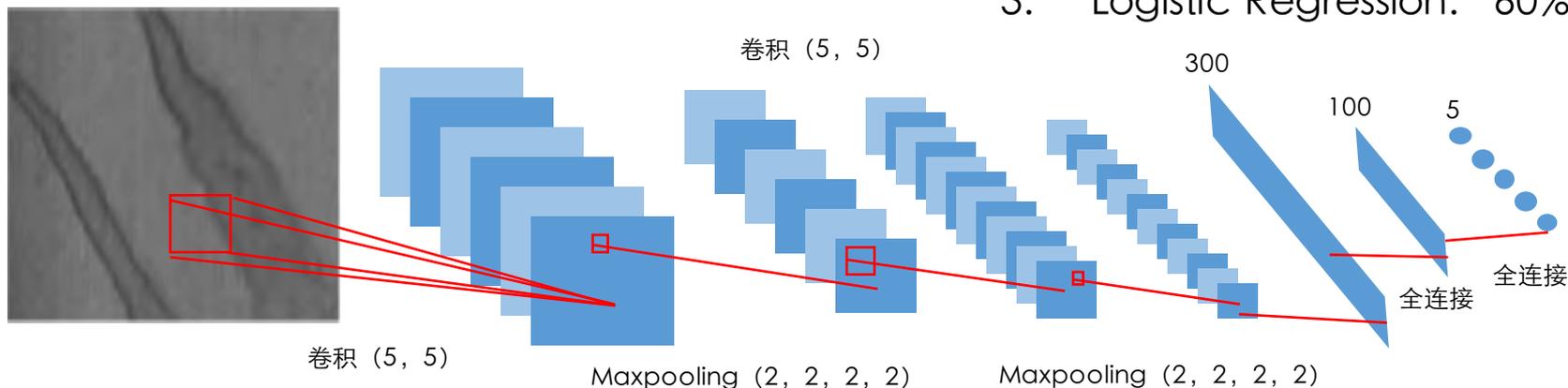
过酸洗



硌伤

CNN模型

- 测试集上准确率
- 1. CNN: 90.82%
- 2. MLP: 64%
- 3. Logistic Regression: 60%



1

大数据平台服务

2

大数据应用开发框架

3

大数据（数据及应用）管理平台

大数据平台服务



大数据应用开发框架

项目
应用
开发
框架

Java

C

.....



标准
服务
应用
开发
框架

统一数据访问框架

计算任务开发框架

分析算法开发框架

组态展现开发框架

报表应用开发框架

自助分析开发框架

页面组件开发框架

GIS应用开发框架

搜索服务开发框架

遵循服务应用开发框架的标准、规范，并在服务应用框架提供的开发环境中进行开发、管理和部署

大数据（数据及应用）管理平台

数据建模

数据安全

数据接入

数据溯源

数据质量

应用接入

运行监控

统一认证

日志分析

宝信大数据平台 xInsight



✓ 大数据开放数据引擎及云端开发

标准可靠
严格定义基础数据模型和元数据，确保数据质量、统一可靠

自助计算
提供类似Hadoop的自助平台，自主定义计算任务、自助计算资源分配、自助管理数据

数据标准
基础指标
基础维度
对象管理

数据管理
任务调度
数据授权

开放数据接口
支持JSON、XML、REST等接口
支持SQL、HQL、JDBC等接口

云端部署平台
支持公有云、私有云、混合云部署

跨平台访问标准
标准数据API
第三方应用接入

多租户管理
代码管理
应用组件
在线测试工具包
一键部署
运行监控

天气预报 上海

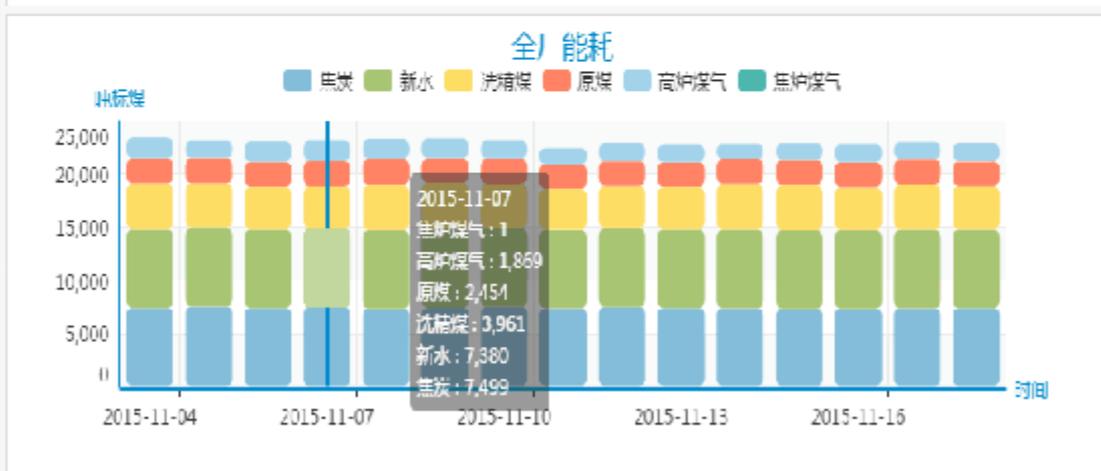
13-16°C
小雨
2015年11月15日紫外线最弱

昨日能耗

前日能耗(吨标煤)	昨日能耗(吨标煤)
12183707.00	12206660.00

昨日费用(元)

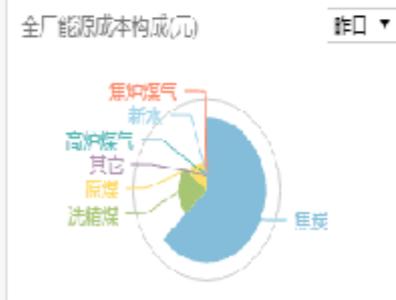
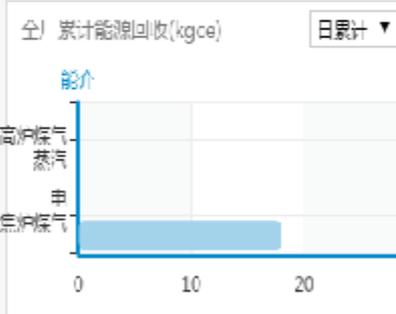
21365476.00
相对与前日: 0.19%



	吨钢综合能耗(kg)	吨钢总量(万吨标煤)	吨钢耗新水(m3)	吨钢能耗成本(元/吨)	吨产量(万吨)	绿色发电量(kwh)
今年	770.91	1117.11	4.33	1100.03	1440.03	1039.21
去年	775.82	1125.15	4.87	1123.05	1437.05	1056.62

分厂单耗 昨日

分厂名称	分厂能耗	先进值
轧材	196.30	10
原料	14.77	0.2
焦化	147.84	30
能源中心	0.00	2



The screenshot shows a web browser window with the address bar displaying `localhost:8080/toolkit/df/designer/designer.do?pagename=energy`. The browser tabs include "宝信大数据应用开发平台" and "Bootstrap 可视化布局器". The browser's address bar also shows search engines like Baidu and various bookmarks.

The main interface is titled "可视化布局设计器" (Visualization Layout Designer). It features a central workspace labeled "Container" where a mouse cursor is visible. To the left, there is a "布局设置" (Layout Settings) panel with a list of grid configurations and their corresponding "拖动" (Drag) buttons:

- 12 + 拖动
- 6 6 + 拖动
- 2 10 + 拖动
- 3 3 3 3 + 拖动
- 8 4 + 拖动
- 4 4 4 + 拖动

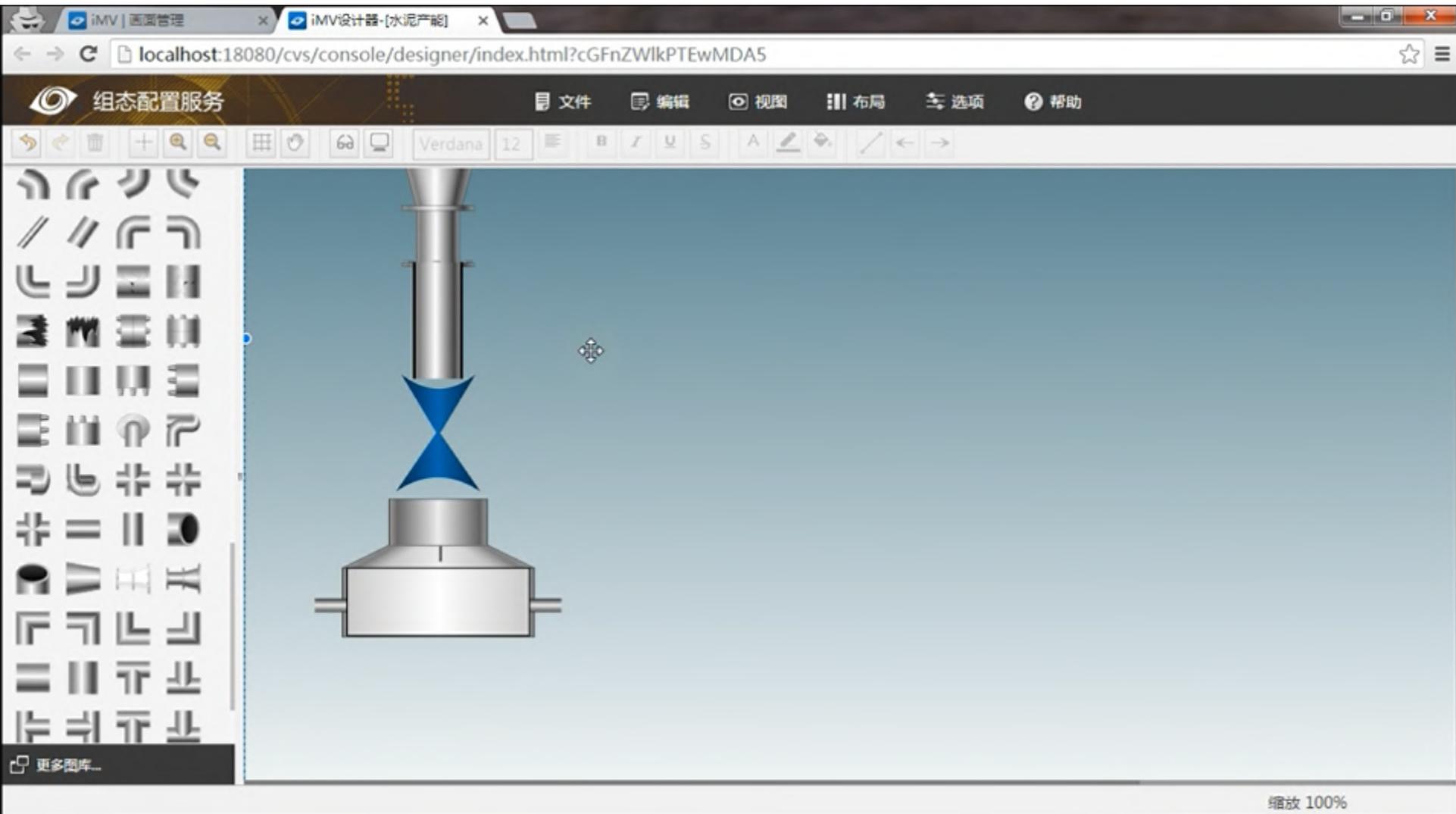
Below the layout settings, there is a vertical menu of visualization components:

- 信息化组件
- 柱状图
- 折线图
- 饼图
- 仪表盘
- 地图
- 雷达图
- 散点图

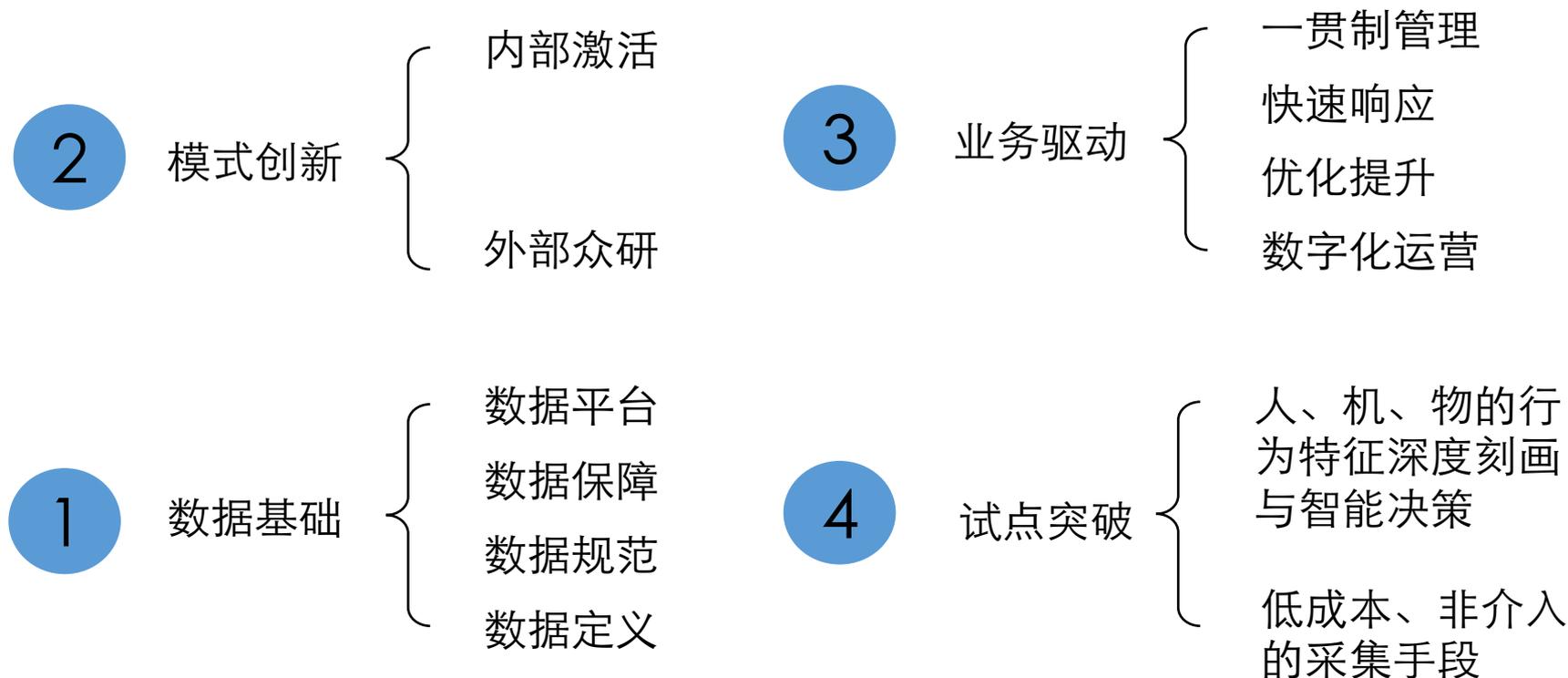
On the right side, there is a "属性框" (Properties Panel) with a sub-tab "后台数据源绑定框" (Backend Data Source Binding Box). A green button labeled "绑定事件" (Bind Event) is visible in this panel.

水泥产能





工业大数据应用落地的关键要素



感谢聆听