如何成为数据驱动发展的企业

程良 | Pivotal 中华区大数据技术总监

Pivotal

企业数据管理与使用现状

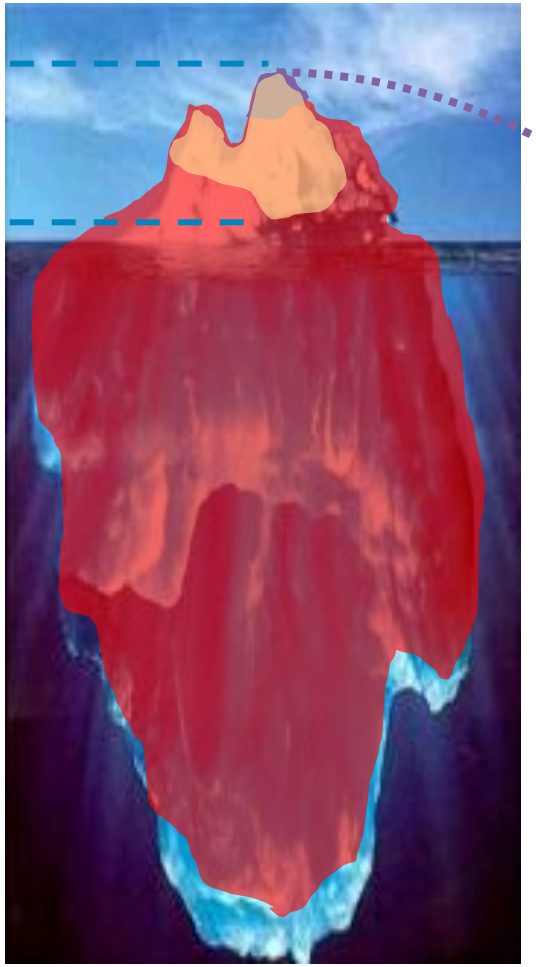
Pivotal | IT大咖说 知识分享平台

准备用于分析的数据

企业管理的数据

据统计，企业日常生产运营产生的数据80%都会被存储

但是准备用于分析的数据仅占3%



最终用户日常分析的数据仅有0.5%

分析后将产生的业务洞察运用到生产中的比例更是低于0.5%

数据在最终用户面前显露的现状

企业数据管理与使用面临的挑战

Pivotal | IT大咖说 知识分享平台

数据分析

- 更强大、丰富的业务洞察能力
  1) 以报表、多维分析和仪表盘为主，对业务现状进行分析，辅助业务决策
  2) 基于多形态的历史数据，通过挖掘算法，对业务发展趋势进行预测，辅助业务决策

数据管理

更广泛、多样的数据管理能力（采集、处理、存储等）

1) 半结构化、非结构化等形态多样
2) 准实时、实时等时效性多样

数据运营

- 更敏捷、弹性的智能化运营能力
  1) 嵌入业务流程中，自动触发、智能化决策
  2) 敏捷开发
  3) 弹性伸缩

数据驱动型企业的定义

在不断增强面向过去的描述型分析基础上，建立数据科学家团队，基于企业数据平台上开展挖掘预测，发现业务洞察

Apply analytic algorithms
Discover insights
业务洞察的深入化

基于发现的业务洞察，建立敏捷、弹性伸缩的分析应用，与业务系统紧密互动，实现智能化决策，数据化运营

Deploy analytic apps at scale
业务运营的智能化

遵循开放标准，构建企业级数据平台。提供多样化的数据处理整合能力，满足多时效的数据服务需求

Store any type and size of data
数据管理的多样化

特征一：数据管理的多样化

Pivotal | IT大咖说 知识分享平台

半/非结构化数据

非结构化数据形态多样，需要进行结构化处理，再与业务数据整合，才能辅助业务决策，发挥业务价值

- 上市公司报告
- 外部财经
- 论坛贴吧
- 社交媒体
- 访问日志
- 位置信息
- 生活轨迹
- ……

数据形态

结构化数据

数据具有结构化程度高、时效性低的特性，需要按照分析主题整合，以分析挖掘等方式辅助业务决策

- 核心系统
- 国结系统
- 信贷系统
- 个贷系统
- 网银系统
- 总账系统
- 信用卡系统
- 客服系统
- ……

动态海量数据产生于业务运营中，具有数据量大、变化频率高特性，需要实时监控处理，与业务系统紧密互动

- 用户点击流
- 用户访问日志
- 用户行动轨迹
- 股票实时行情
- 金融高频交易
- ……

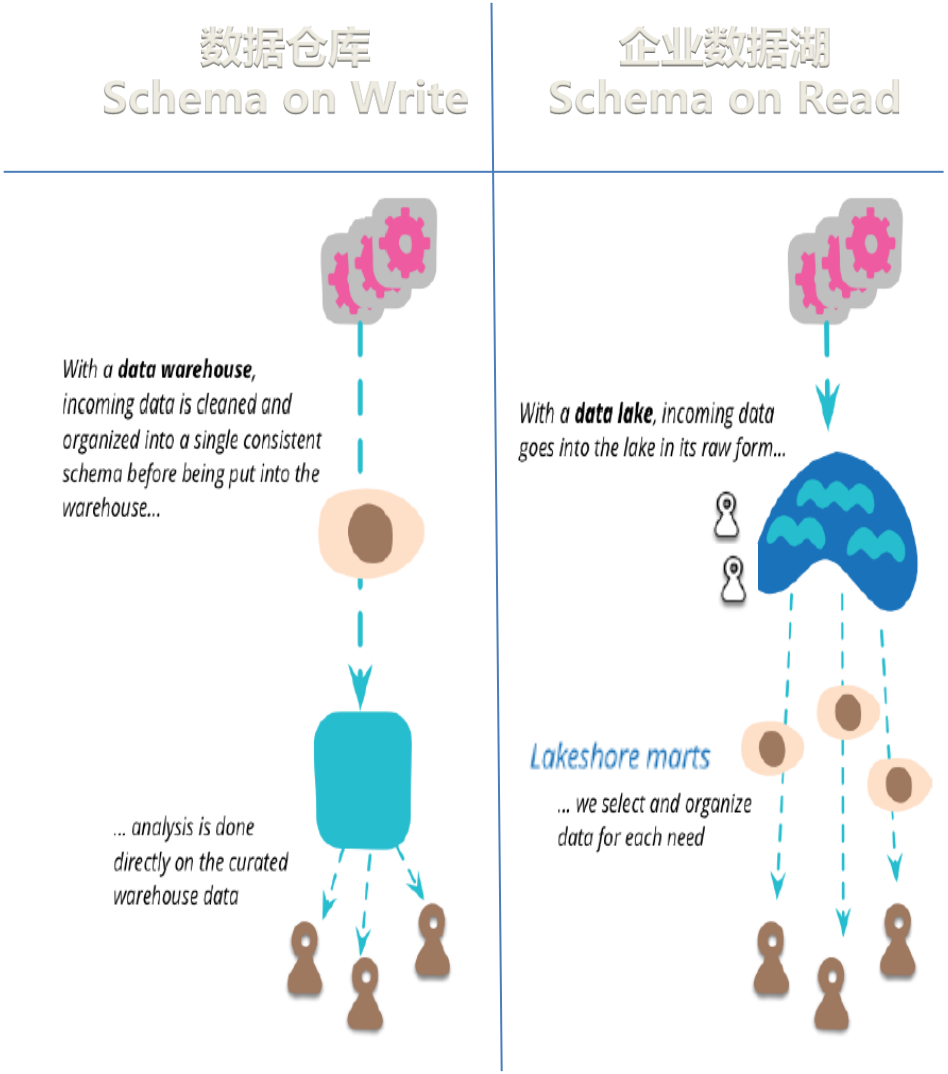
静态海量数据

产生频率

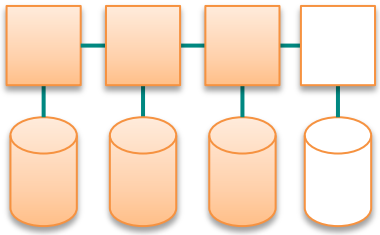
动态海量数据

数据湖实现多样化的数据管理

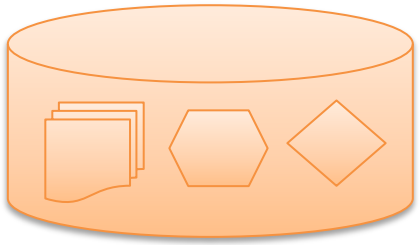
- 企业数据湖最早是由Pentoho CTO James Dixon于2010年提出的
- James认为：“如果把传统层次化的数据平台看做加工后的纯净水的话，数据湖则是未经处理和包装的原生态水库，不同源头的水体源源不断流入数据湖，带来各种分析、探索可能性”
- 近两年随着大数据的流行，数据湖被一些企业采用，搭建企业级基础数据平台



数据湖的技术路线

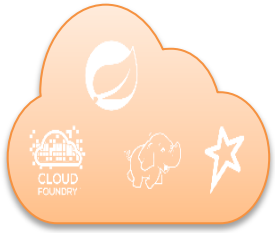


海量数据存储能力



管理各种类型、形态数据

弹性伸缩的计算能力



满足云友好、开源、分布式技术路线

特征二：业务洞察的深入化

- **随着大数据时代的到来，企业数据分析也不同于传统的BI分析：**
  1. **数据量急剧增大，数据形态更加多样、内容更加丰富**
  2. **分析方式不再是简单的查询报表，而是以数据挖掘、机器学习等预测类分析为主**

看到过去的描述型分析

- 整合企业内部的结构化数据
- 以报表、多维分析和仪表盘为主，反应当前业务现状，并对业务现状进行分析，辅助业务决策
- 典型应用：业务统计分析等

面向未来的预测型分析

- 整合企业内外部结构化、非结构化数据
- 通过数据挖掘，对业务发展趋势进行预测，辅助业务精准决策
- 典型应用：数据科学实验、精准营销
- 预测型分析出现了一类新的角色——数据科学家，这类人对业务和数据都非常了解，经常会有针对业务经营的想法或思路，他们需要海量历史数据，进行数据科学实验，验证自己的想法

面向未来的预测型分析过程

提出业务假想

数据科学家根据自己对业务的理解，有了一些新的业务想法或假设

这些想法或假设存在不确定性，需要验证其业务价值

探查企业数据

数据科学实验团队进行数据探查，确定数据实验所需数据可用性

针对搜集到的数据进行处理，主要包括：大数据的结构化处理、文本挖掘

训练挖掘模型

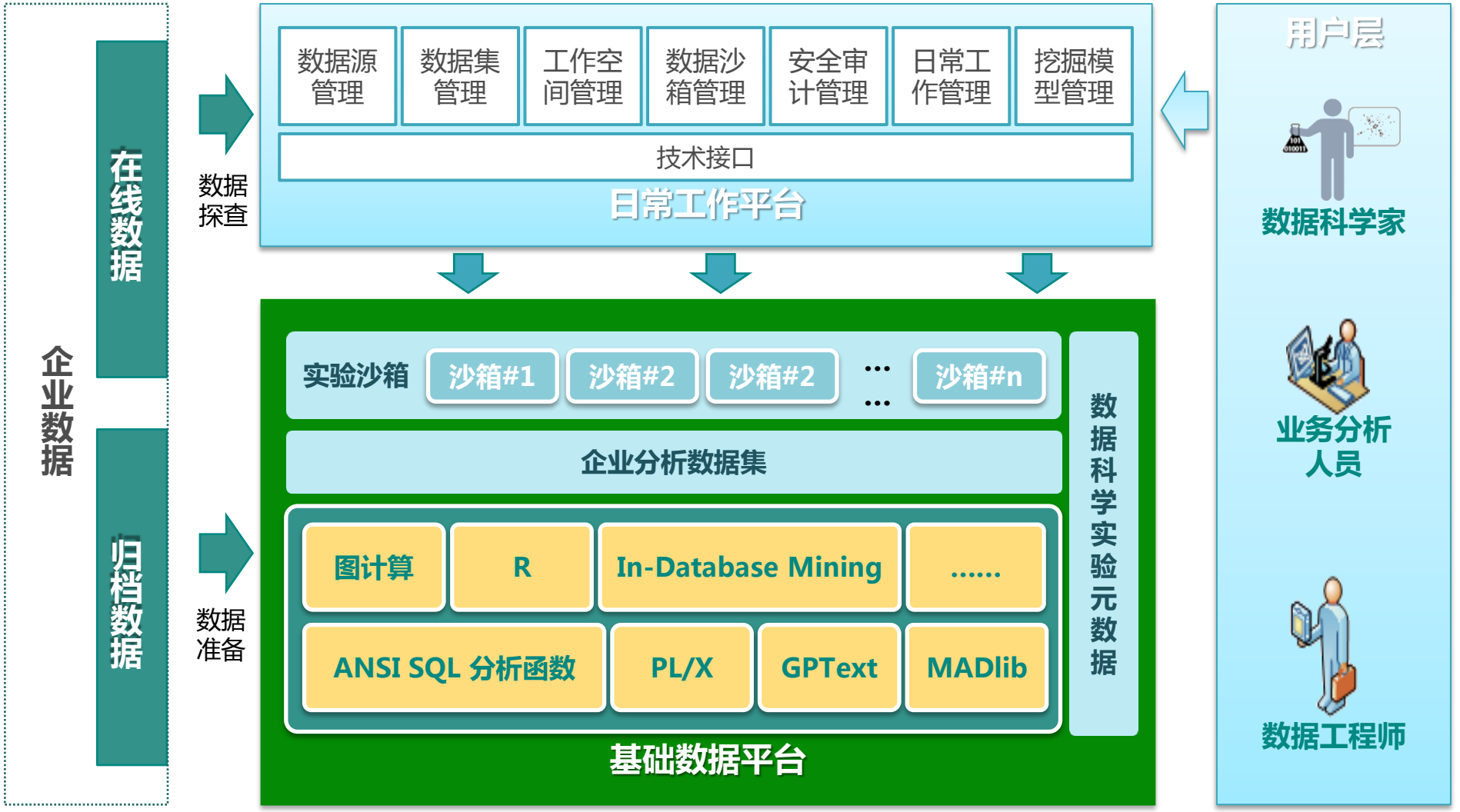
数据科学实验团队提取特征值，训练模型，最终完成模型设计

Predictive Analytics

产生业务洞察

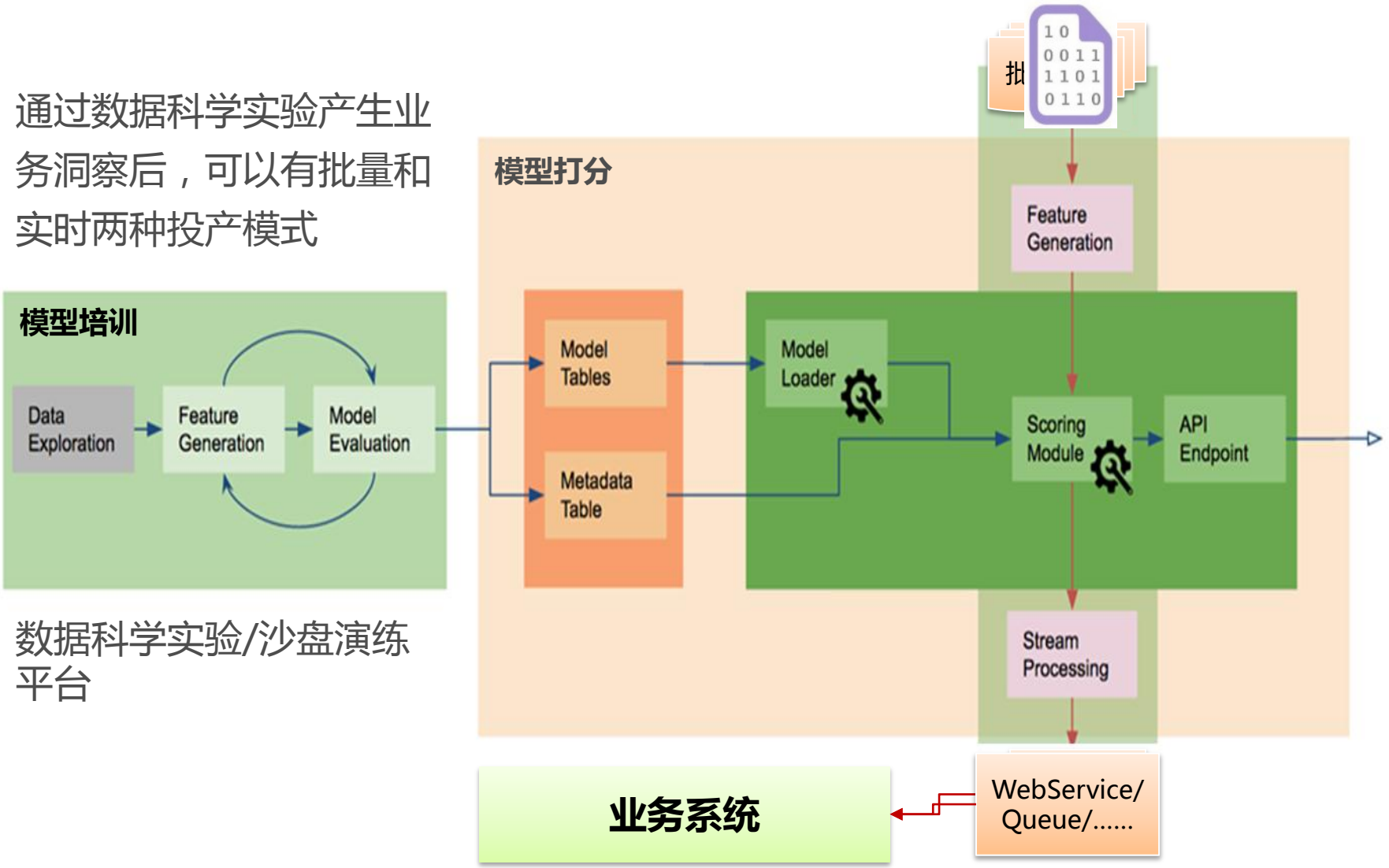
根据挖掘的结果，论证新的业务想法或假设是否具备业务价值，是否运用到生产中

支撑预测型分析的平台逻辑架构



预测型分析的成果投产

通过数据科学实验产生业务洞察后，可以有批量和实时两种投产模式



数据科学实验/沙盘演练平台

特征三：业务运营的智能化

信息变化的实时捕获

实时采集企业业务流程运转过程中产生的数据，如：股票交易、客户点击流等

业务规则的实时决策

针对采集到的实时数据流，运用业务规则，进行实时决策（如：合规监察、发现秒级营销机会等）

决策结果的实时执行

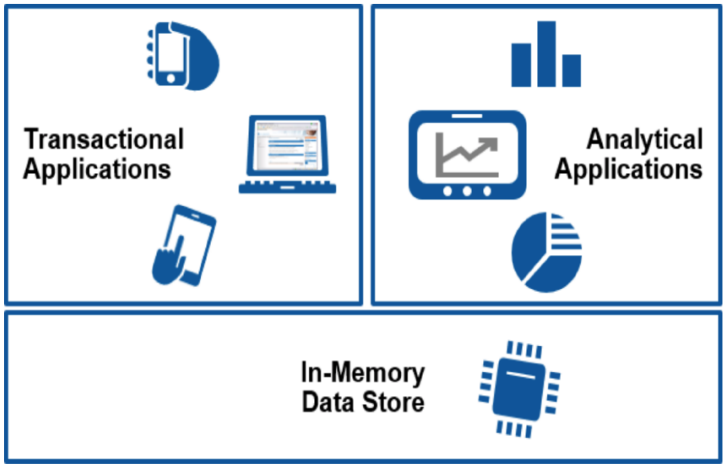
根据实时决策结果，驱动业务流程执行，如：秒级营销、风险预警

业务运营的智能化的技术实现

Hybrid Transactional Analytical Processing (HTAP) refers to the bringing together of transactional and analytical processing in the same system, on the same data. This enables real-time analytics against transactional data as it is created. Gartner identifies in-memory computing technologies as necessary for supporting HTAP, as disk-based data storage does not provide the performance required.

Gartner

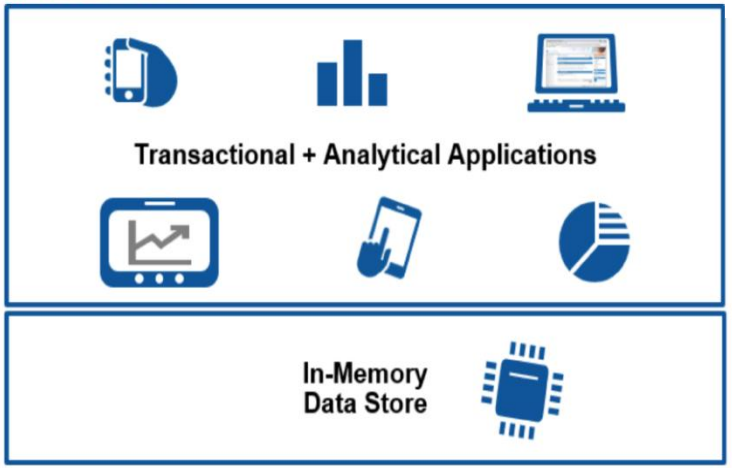
Figure 1. Point-of-Decision HTAP



HTAP = hybrid transaction/analytical processing

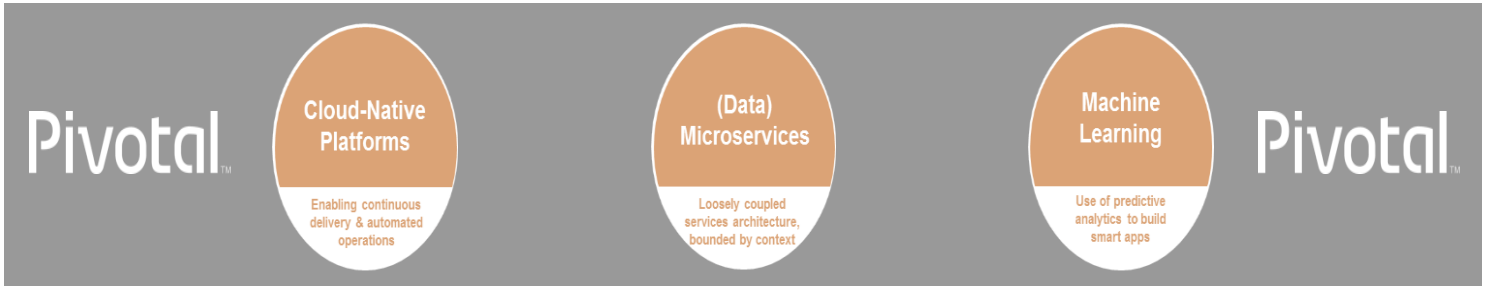
Source: Gartner (June 2016)

Figure 2. In-Process HTAP

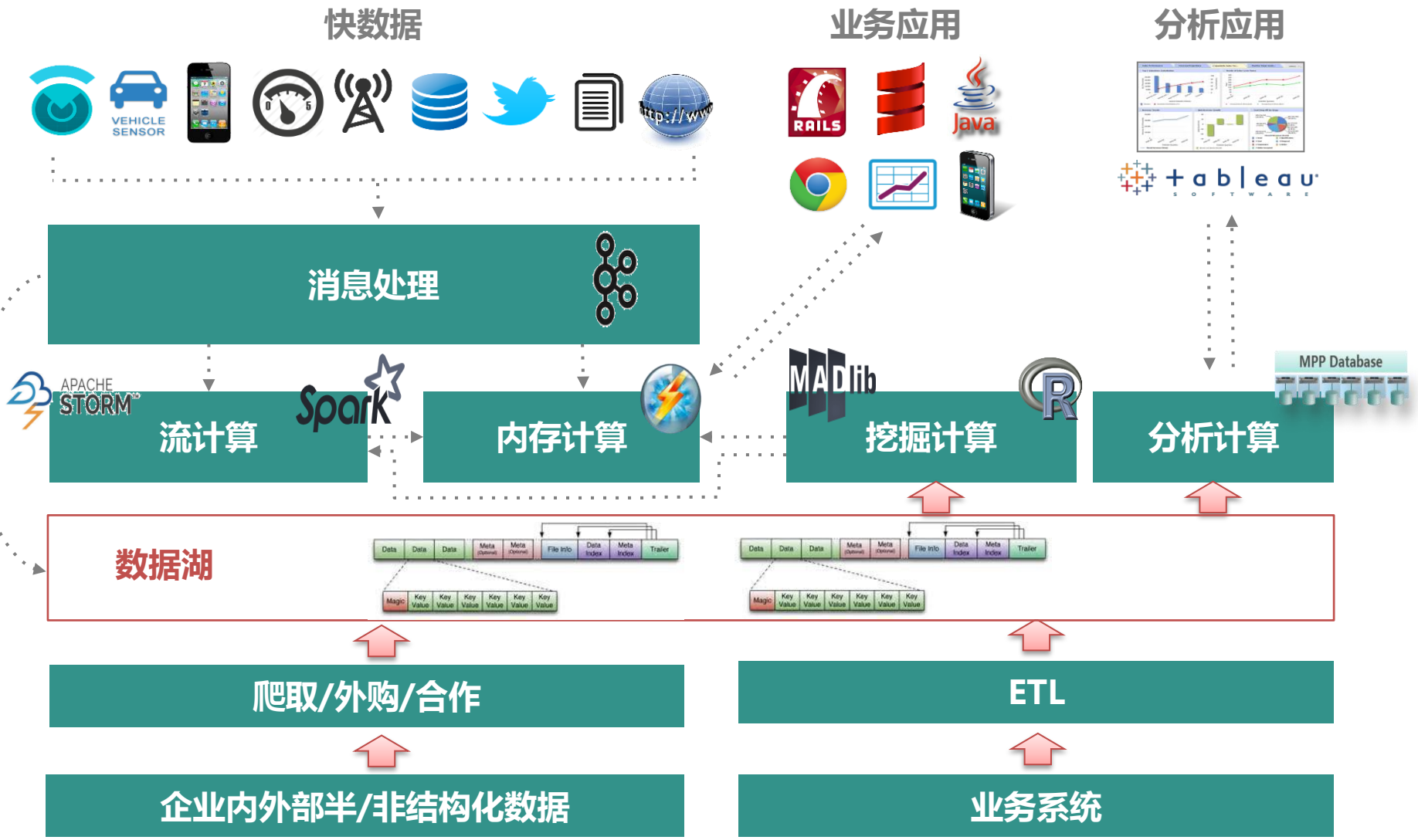


HTAP = hybrid transaction/analytical processing

Source: Gartner (June 2016)



Pivotal 建议的数据驱动型企业参考架构



总结……

BECOMING A DATA-DRIVEN COMPANY…

IS NOT Just about deploying Hadoop

OR How many Data Scientists you have

IT'S ALL ABOUT HOW YOU OPERATIONALISE YOUR INSIGHTS