www.pwccn.com

数据科学价值变现

Qingfeng Li

January 2018

pwc

目录

章节

1

掌握核心竞争力
数据科学及案列分析

数据科学及价值变现


• Linear Programming
• Non Linear / Integer Programming（非线性规划）
• Dynamic Programming
• Critical Path Analysis
• Simulation / Monte Carlo Analysis
• Forecasting
• Quality Control
• Inventory Control
• Queuing Theory
• Renewal Theory
运筹学
• Descriptive Statistics
• Probability
• Probability Distributions
• Estimation
• Correlation
• Regression
• Analysis of Variance
• Sample Surveys
• Design of Experiments
• Cluster Analysis
• Multivariate Analysis
• Time Series Analysis
（时间序列分析）
• Quality Control
• Econometrics
数理统计
• Neural Networks – MLP, RBF, Kohonen
• Decision Trees - CHAID, C 5.0, C&RT
• Link Analysis
• Association Rules
• Support Vector Machines（支持向量机）
• Genetic Algorithms
机器学习
价值变现

案列分享

背景信息：能源数据主要特性

能量来源主要是分布式可再生能源(如电、水），相比传统能源（气、煤），其不确定性和不可控性大；

能量使用侧用户负荷、运行模式等都会实时变化。比如工厂大型机组开工任务顺序不同，瞬时电压负荷不同，电流强度不同，电网上的能量消耗也非简单线性可估计。

以电能为例，无论是供电侧，还是用电侧，能量、物质和信息高度耦合（分布式、自组织式、按需紧、按需松）形成实体或非实体复杂网络；

从数据和信息层面从上而下可分为主干网、广域网、局域网三层，生产、存储、运输和终端消费又各有特点。每一层的工作环境和功能特性均不相同，这也造成每一层的动态特性尺度差异巨大。

能源的转化工作在由类型繁多、数量庞大的数据组成的高度信息化的环境中；

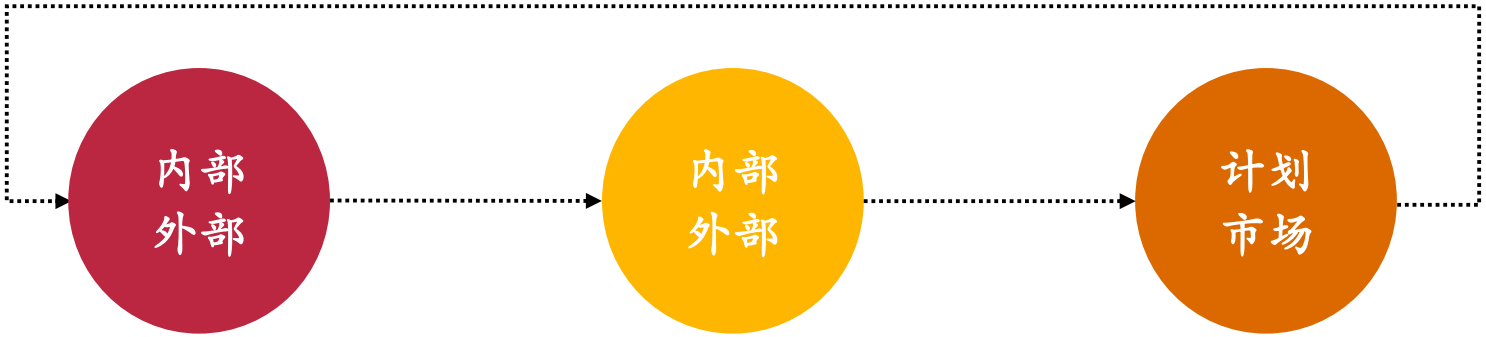
以电能为例，这些数据既包括电量相关数据，温度、压力、湿度等非直接用电量数据以及各类外部数据如气象数据、行业数据等。

案列分享（继续）

项目思路：大数据预测

输入：日电量数据+日气温+月平均气温+日期因子（历史真实数据）

输出：目标时间段内每天的日用电量（预测结果，同已有历史真实数据比对）



- 历史数据，训练数据
- 现在数据，验证模型
- 设定目标，预测结果

- 外部数据：气象数据、行业数据、政策数据等
- 内部数据：发电数据、燃料数据、检修计划等

- 市场电：辅助电量预测部门进行月度发电量预测
- 计划电：在国家相关政策驱动下，未来市场电比例上升后，实现对市场电用电量更准确的预测

数据科学实现步骤：

1 预测目标与变量假设

- 预测目标为发（用）电量
- 假设变量包括内部变量（x个）和外部变量（y个）等

2 数据收集与处理

- 异构体系下多种数据集成及捕获方式，丰富数据集（变量）
- 数据清理（异值剔除、缺失补齐、非结构化数据结构化）

3 变量选择处理

- 回归，选择效力因子强的变量作为关键变量（时间范围、天气）
- 相关性检验

4 模型建立、训练与验证

- 确定算法模型-时间序列预测法或回归（本次在成本范围下，选择前者）
- 不同数据集（变量-目标结果）下的模型验证及调优

案列分享（继续）

算法解释：回归分析与时间序列分析

回归分析（regression analysis）

- 有监督学习。其数据集是给定一个函数和它的一些坐标点，然后通过回归分析的算法，来估计原函数的模型。回归分析是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，在预测分析中运用十分广泛；
- 当数据之间存在多重共线性（自变量高度相关）时，宜使用岭回归（Ridge Regression）需要使用岭回归分析，岭回归通过给回归估计值添加一个偏差值，来降低标准误差可以很好的解决变量间的多重共线性问题，尤其是在处理工业界数据时表现优异；
- 在特定预测用例中，模型调优后，准确度非常高，但成本投入亦不低。

时间序列分析（time series）

- 无监督学习。侧重研究数据序列的互相依赖关系，实际上是对离散指标的随机过程的统计分析。其基本思想是根据系统的有限长度的运行记录，建立能够比较精确地反映序列中所包含的动态依存关系的数学模型，并借以对系统的未来进行预报；
- 原理一是承认事物发展的延续性。应用过去数据，就能推测事物的发展趋势。二是考虑到事物发展的随机性。任何事物发展都可能受偶然因素影响，为此要进一步利用移动平均法、指数平滑法、模型拟和法等进一步优化模型；
- 特点：简单易行，便于掌握，但准确性差，一般只适用于短期预测。

由单一算法向多算法融合的延伸

有监督学习（分类，回归），半监督学习（分类，回归），无监督学习（聚类）
半监督聚类（有标签数据的标签不是确定的，类似于：肯定不是x，很可能是y）

案列分享（继续）

测试用例，确定模型：基于2014.10.01~2016.04.23（571天）的数据建立模型，预测2016.04.24~2016.04.30（一周7天）的日总用电量

输入

输入变量的数量越多，预测的结果就可能更精确，包含客户提供及数据顾问补充后的变量集：

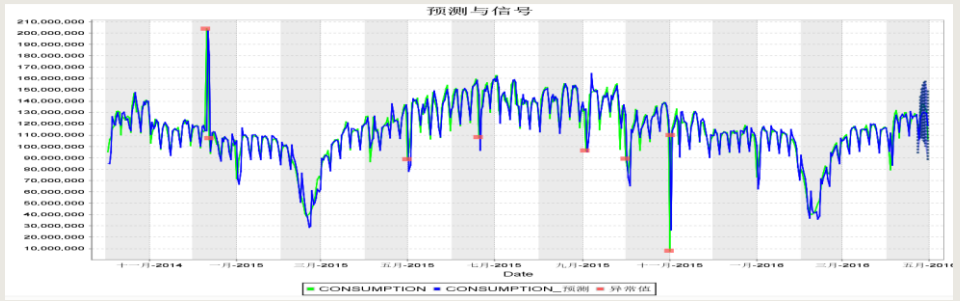
Column	Description
SundayMonthInd	Indicates if the date is a Sunday with the weekday's occurrence count in the month so far. 0 otherwise.
MondayMonthInd	Indicates if the date is a Monday with the weekday's occurrence count in the month so far. 0 otherwise.
TuesdayMonthInd	Indicates if the date is a Tuesday with the weekday's occurrence count in the month so far. 0 otherwise.
WednesdayMonthInd	Indicates if the date is a Wednesday with the weekday's occurrence count in the month so far. 0 otherwise.
ThursdayMonthInd	Indicates if the date is a Thursday with the weekday's occurrence count in the month so far. 0 otherwise.
FridayMonthInd	Indicates if the date is a Friday with the weekday's occurrence count in the month so far. 0 otherwise.
SaturdayMonthInd	Indicates if the date is a Saturday with the weekday's occurrence count in the month so far. 0 otherwise.
LastSunday	1 if last Sunday of the month. 0 otherwise.
LastMonday	1 if last Monday of the month. 0 otherwise.
LastTuesday	1 if last Tuesday of the month. 0 otherwise.
LastWednesday	1 if last Wednesday of the month. 0 otherwise.
LastThursday	1 if last Thursday of the month. 0 otherwise.
LastFriday	1 if last Friday of the month. 0 otherwise.
LastSaturday	1 if last Saturday of the month. 0 otherwise.
PenultimateSunday	1 if penultimate Sunday of the month. 0 otherwise.
PenultimateMonday	1 if penultimate Monday of the month. 0 otherwise.
PenultimateTuesday	1 if penultimate Tuesday of the month. 0 otherwise.
PenultimateWednesday	1 if penultimate Wednesday of the month. 0 otherwise.
PenultimateThursday	1 if penultimate Thursday of the month. 0 otherwise.
PenultimateFriday	1 if penultimate Friday of the month. 0 otherwise.
PenultimateSaturday	1 if penultimate Saturday of the month. 0 otherwise.
Workingday	1 if working day (Saturday, Sunday, Bank Holiday). 0 otherwise.
BeforeHoliday	1 if before holiday. 0 otherwise.
Holiday	1 if holiday (Saturday, Sunday, Bank Holiday). 0 otherwise.
ContributionToWorkingMonth	If working day: 1 divided by number of month's working days. 0 otherwise.
ContributionToMonth	1 divided by number of month's days.
MonthWorkingDayInd	Indicates if working day with the work day's occurrence count in the month so far. 0 otherwise.
ReverseMonthWorkingDayInd	Indicates if working day by counting down the work day's occurrence count in the month. 0 otherwise.
Last5WDinMonthInd	Indicates the month's last 5 working days by counting them up from 1 to 5. 0 otherwise.
Last5WDinMonth	1 if one the month's last 5 working days. 0 otherwise.
Last4WDinMonthInd	Indicates the month's last 4 working days by counting them up from 1 to 4. 0 otherwise.
Last4WDinMonth	1 if one the month's last 5 working days. 0 otherwise.
DayofTheWeek	Day of the week

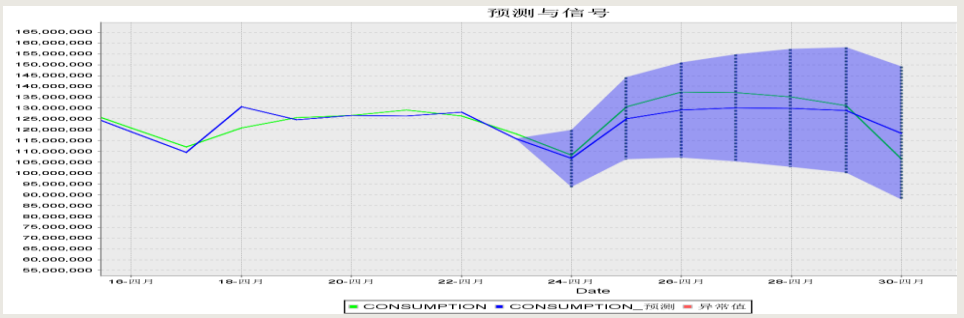
平均误差为4.75%，若剔除最后一天(2016/4/30为串休），*平均误差为3.71%*。

预测结果十分理想。

输出

时间序列模型（拟合度>98%和置信区间>95%）





日期	实际用电量	预测用电量	误差
2016/4/24	108,268,697.71	106,683,000	1.46%
2016/4/25	130,716,325.37	125,121,000	4.28%
2016/4/26	137,298,449.47	129,031,000	6.02%
2016/4/27	137,039,961.40	130,043,000	5.11%
2016/4/28	134,989,243.73	129,926,000	3.75%
2016/4/29	131,034,816.55	128,903,000	1.63%
2016/4/30	106,646,210.94	118,411,000	11.03%

案列分享（继续）

验证用例，优化模型：提供2014年10月1日~2016年7月31日期间内的每日用电量数据及天气数据，预测2016年8月1~7日（一周7天）用电量

输入

初步拟建模型两类种，一种包含天气因素，另一种排除天气因素;

包含天气因素模型的输入变量：气温+时间因子;

排除天气因素模型的输入变量：时间因子。

与实际值对比，一周总计误差达到了4.83%（排除天气6.92%）

2016/8/2（星期二）的误差较大（>40%)。其原因与2016年8月2号当天台风“妮妲”的到来有很大的关系

排除2016/8/2，一周总计的误差0.35%（排除天气2.21%）

输出

时间序列模型下的验证结果

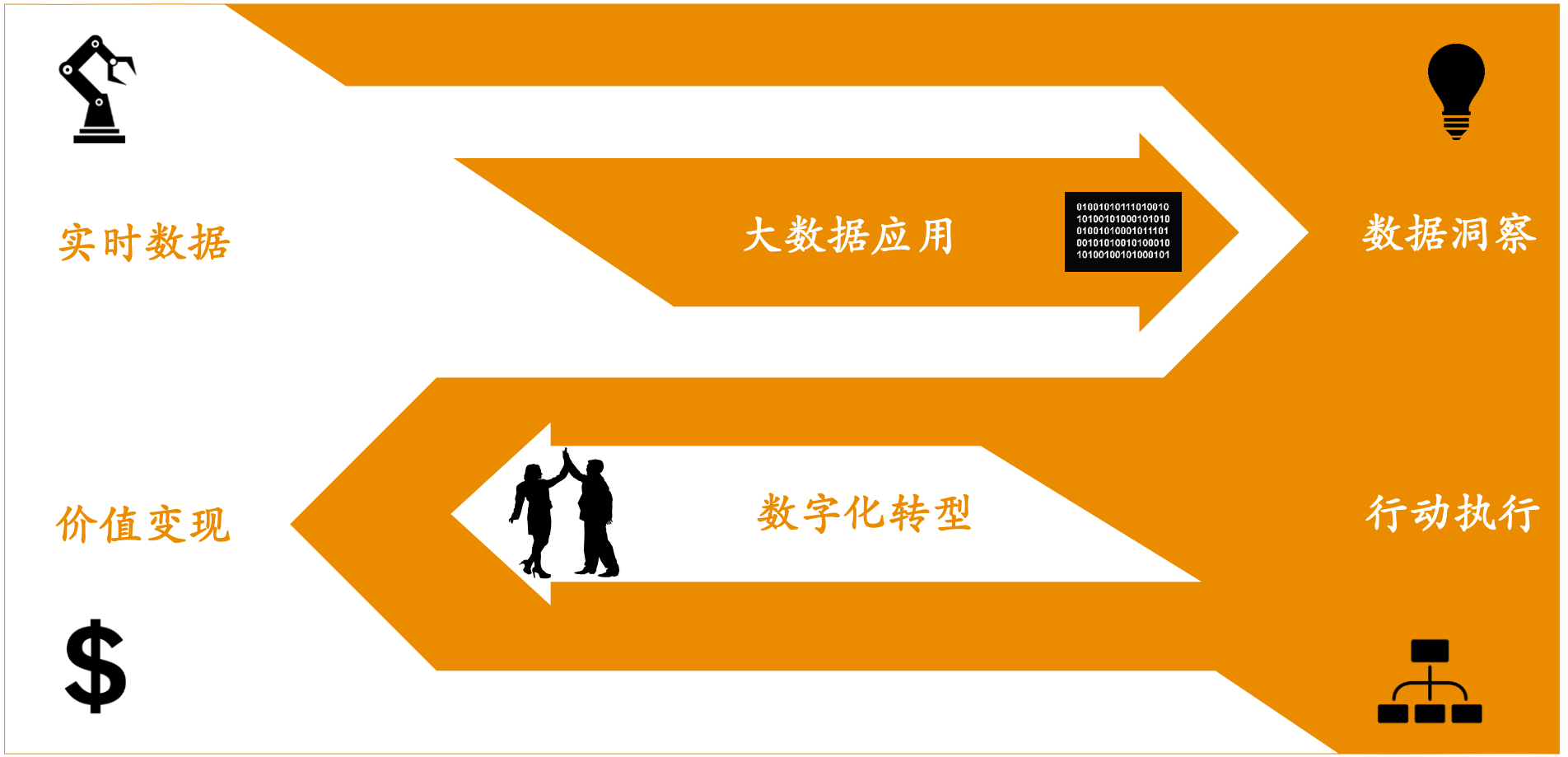
日期	包含天气因素	排除天气因素	实际值	误差（包含天气）	误差（排除天气）
一周总计	1007839000	102791600	96136064	4.83	6.92
一周总计（排除2016/8/2）	860091000	876020000	857112788	0.35	2.21

日期	星期	包含天气因素	排除天气因素	实际值	误差（包含天气）	误差（排除天气）
2016/8/1	星期一	14640600	148271000	14426083	1.49	2.78
2016/8/2	星期二	14774800	151896000	10424785	41.73	45.71
2016/8/3	星期三	14552400	152694000	13438429	8.29	13.62
2016/8/4	星期四	15476100	152316000	14528431	6.52	4.84
2016/8/5	星期五	14661400	150810000	15061798	2.66	0.13
2016/8/6	星期六	13791900	141196000	14587590	5.45	-3.21
2016/8/7	星期日	12886700	130733000	13668965	5.72	4.36

价值变现
对于客户，对于我们

2

数字时代下的启思



案列分享（继续）

相对完整的项目实现步骤：

1 预测目标与变量假设

- 预测目标为发（用）电量
- 假设变量包括内部变量（x个）和外部变量（y个）等

2 数据收集与处理

- 异构体系下多种数据集成及捕获方式，丰富数据集（变量）
- 数据清理（异值剔除、缺失补齐、非结构化数据结构化）

3 变量选择处理

- 回归，选择效力因子强的变量作为关键变量（时间范围、天气）
- 相关性检验

4 模型建立、训练与验证

- 确定算法模型-时间序列预测法（或回归）
- 不同数据集（变量-目标结果）下的模型验证及调优

5 数据质量提升

- 加大训练集数据体量，获得更准确的机器学习结果
- 加入政策数据等非结构化数据

6 算法质量提升

- 后续将在使用中不断挖掘新的预测模型，实现由单一算法向 多算法融合的延伸
- 逐步深入研究自选变量，代入系 统以提高模型准确度

7 封装需求

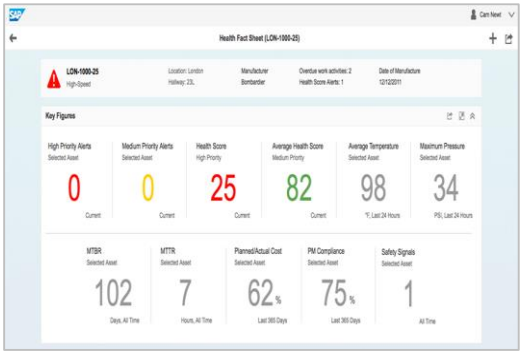
- 电力企业需要建立，部署多个行业的预测模型，时间序列可以自动批量建模
- 基于属性，加载模型、并分段，开启项目的任务，自动地创建多个模型
- 长远计划，模型工厂可以设置计划重复作业

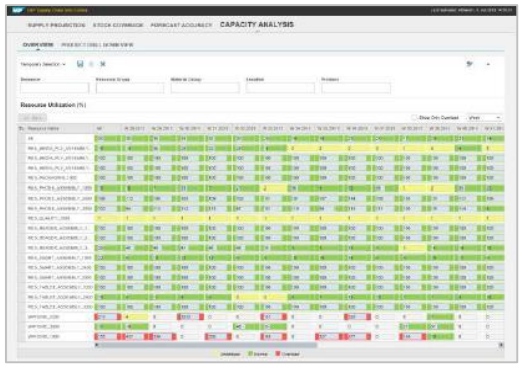
8 业务模式创新

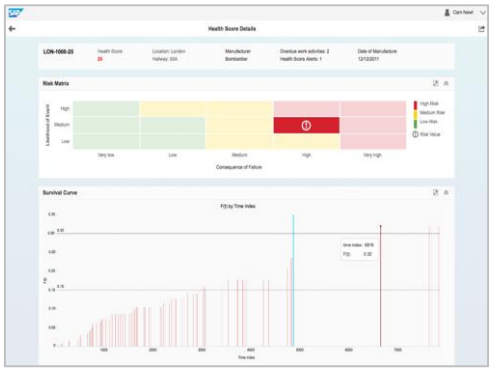
- 供电侧：结合国家新政策、计划生产及售电、智能电网
- 用电测：能源品质分析、能源计划优化、厂内微网管理

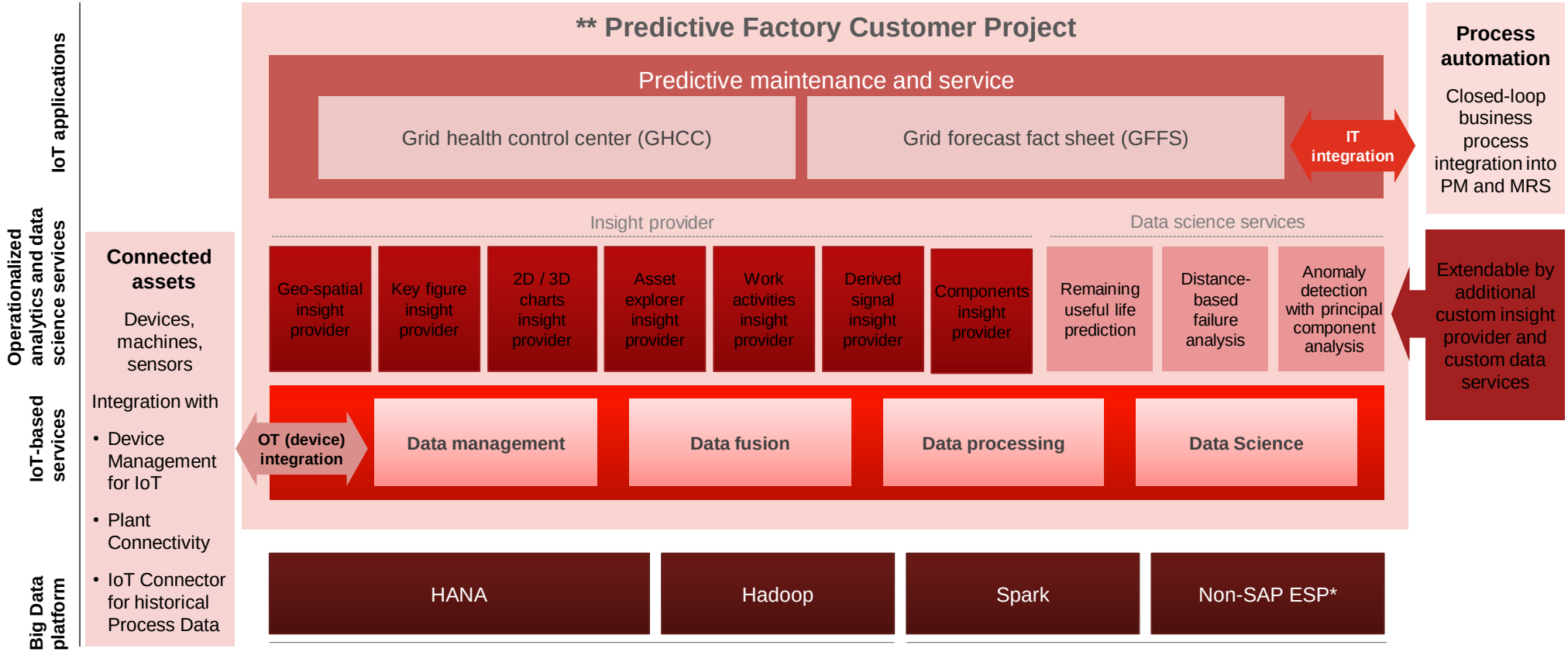
案列分享（继续）

系统功能解决方案：





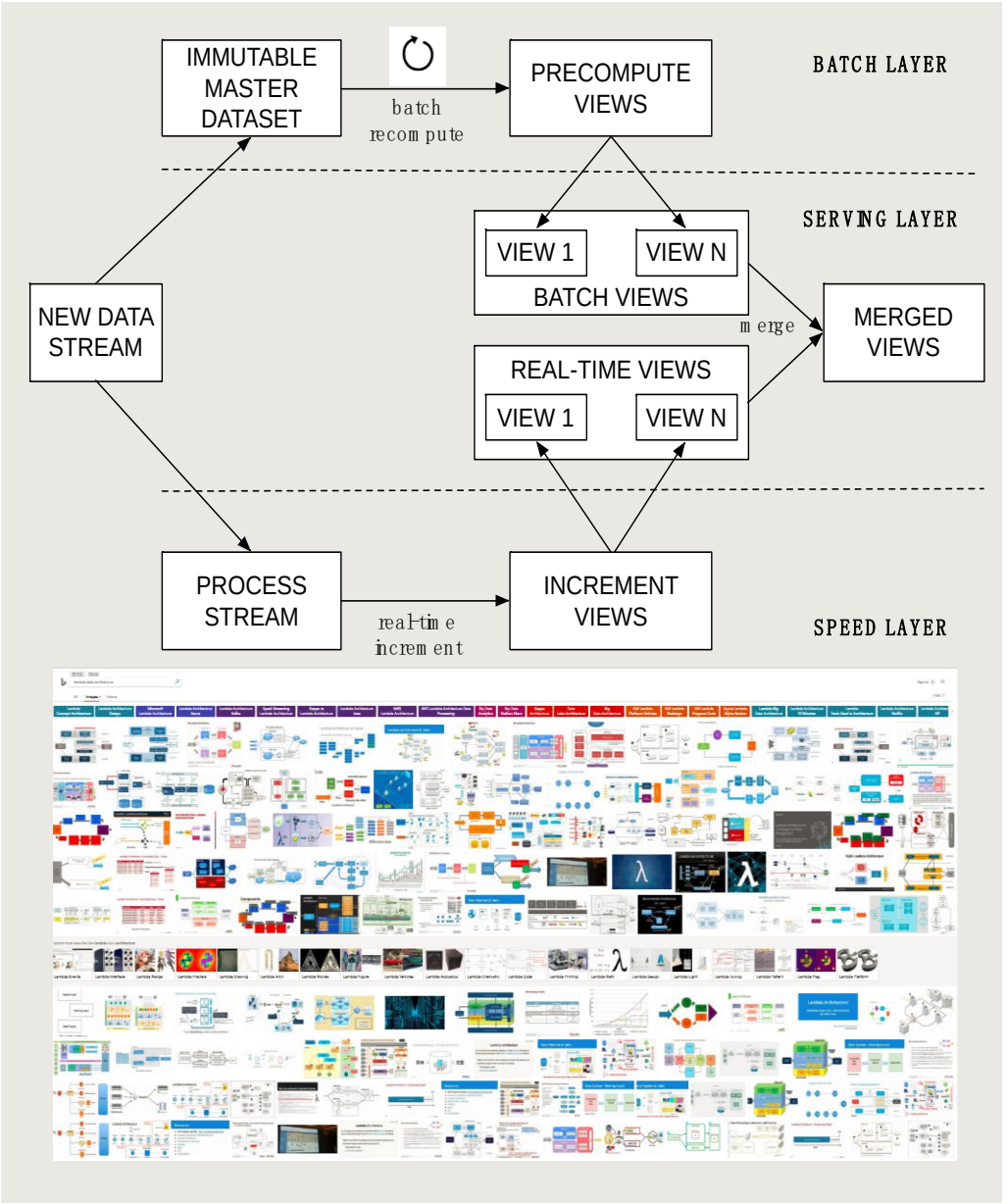




案列分享（继续）

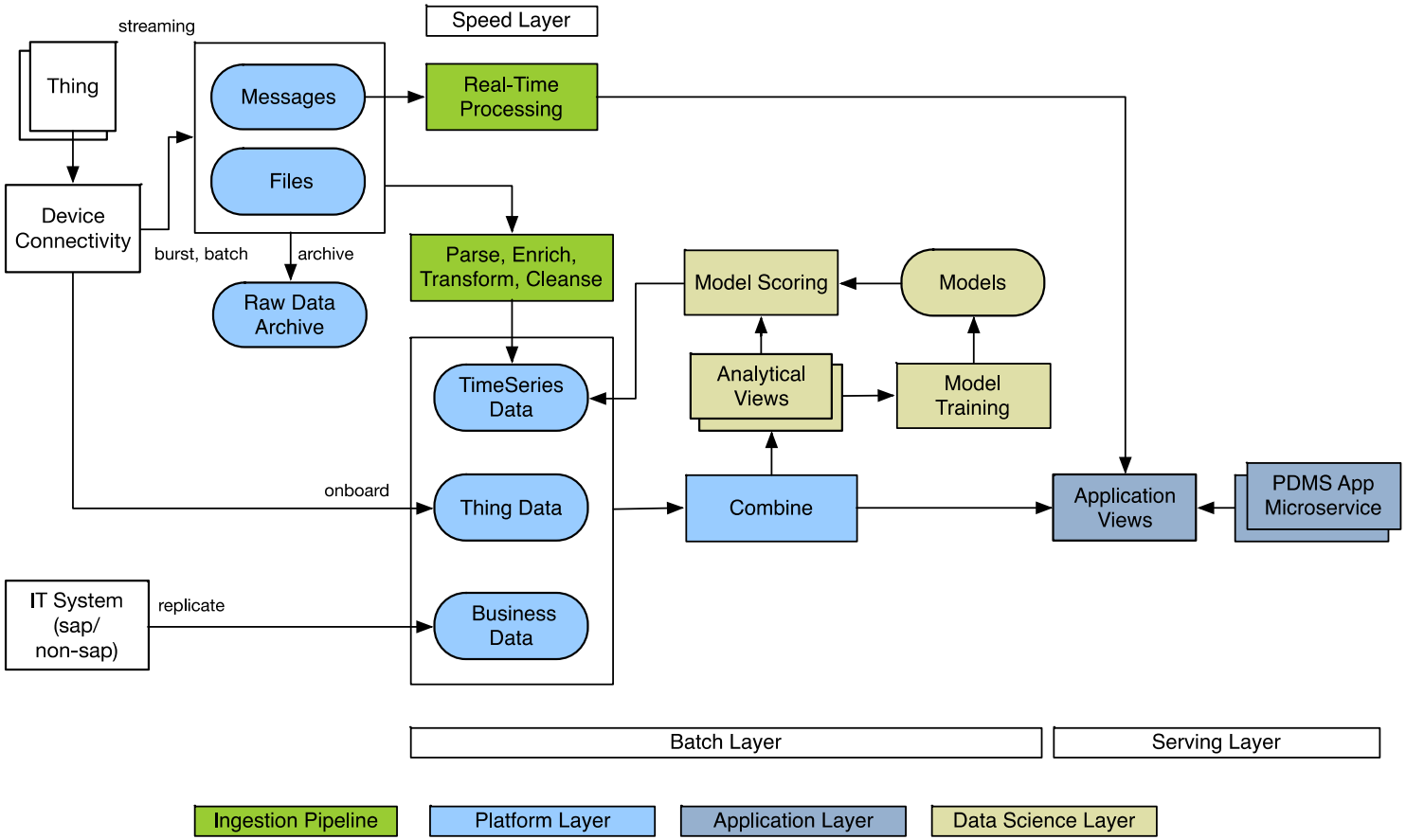
系统核心设计原理：Lambda大数据处理架体系（主流思想之一）

- 所有进入系统的数据，都会被分发到批处理层（batch layer）和快速处理层（speed layer）
- 批处理层（batch layer）有两个作用：管理master的数据（raw数据）：比如用HDFS来存储以及为数据转换为批处理视图做预处理
- 服务层（serving layer）用于加载和实现数据库中的批处理视图，以便用户能查询
- 快速处理层（speed layer）用于处理新数据和服务层更新造成的高延迟补偿
- 任何query的答案，都能通过合并批处理视图和实时视图的结果来获得



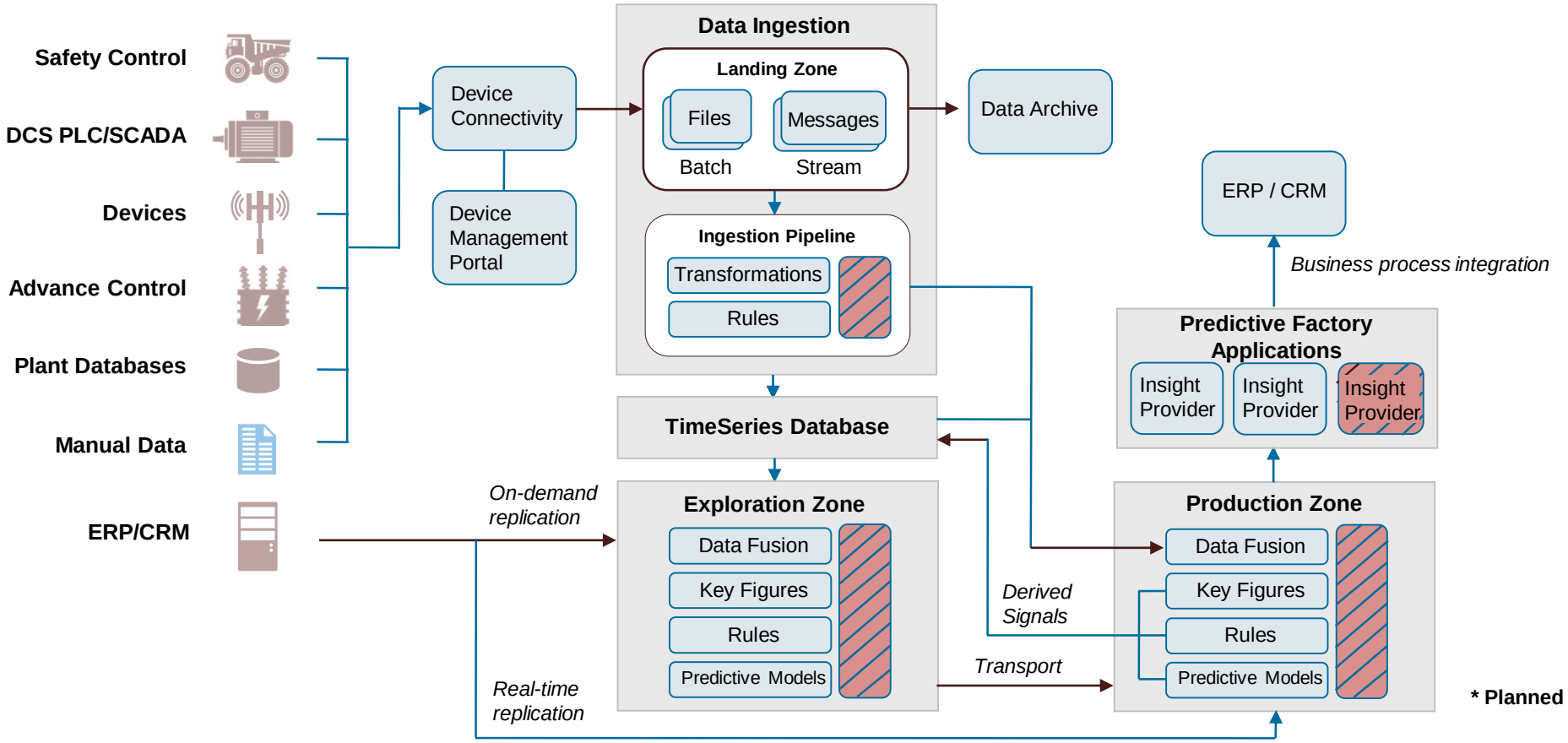
案列分享（继续）

系统技术单元架构：



案列分享（继续）

系统技术实现架构：

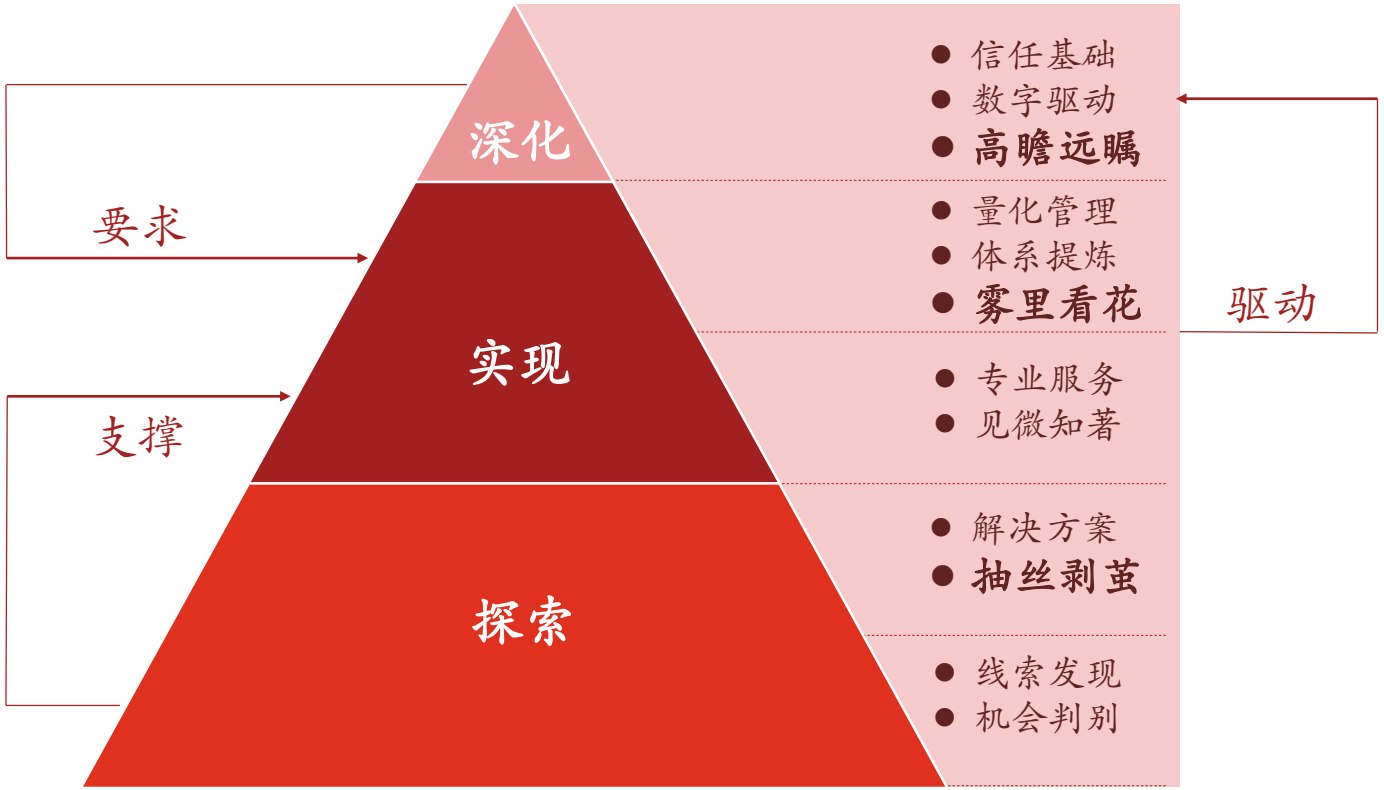


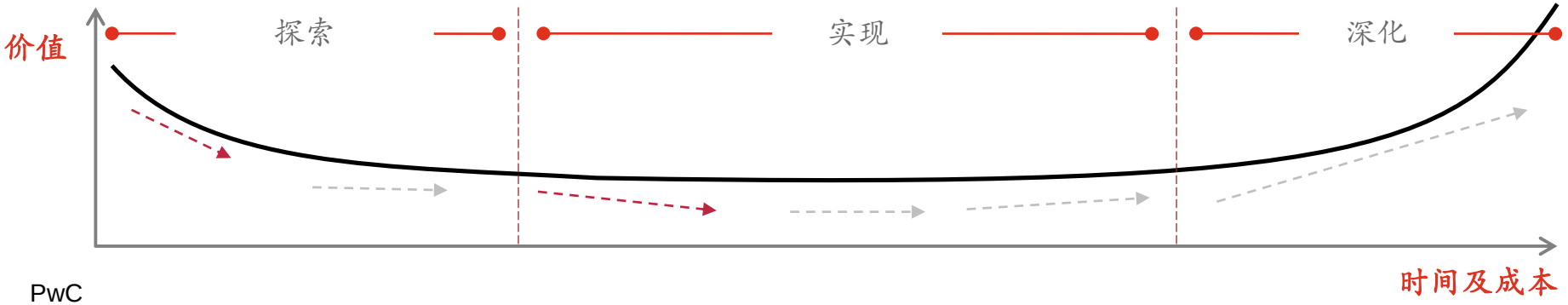
价值管理

3

案列分享（继续）

基于商业价值和时间维度，专注于从一系列“转化”活动中持续提炼价值





“海不则细流，故能成其大。山不拒细壤，方能就其高”

感谢聆听，欢迎探讨

李青峰 *Qingfeng Li*

PwC, Risk Assurance

Email: qingfeng.li@cn.pwc.com