

BIG  
DATA

# 数据治税

大数据构建智慧税务

刘诚忠

明略数据税务事业部总经理

# QCon

2016.10.20~22

上海·宝华万豪酒店

## 全球软件开发大会 2016

### [上海站]



购票热线: 010-64738142

会务咨询: [qcon@cn.infoq.com](mailto:qcon@cn.infoq.com)

赞助咨询: [sponsor@cn.infoq.com](mailto:sponsor@cn.infoq.com)

议题提交: [speakers@cn.infoq.com](mailto:speakers@cn.infoq.com)

在线咨询(QQ): 1173834688

团 · 购 · 享 · 受 · 更 · 多 · 优 · 惠

# 7折

优惠(截至06月21日)  
现在报名, 立省2040元/张

# 目录

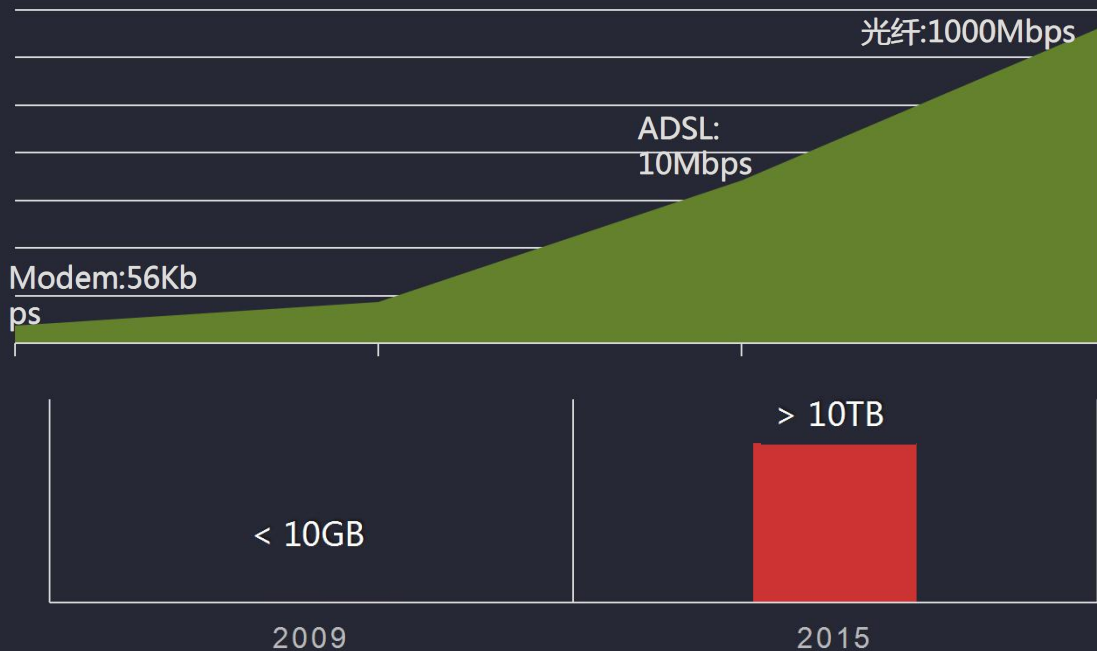
---

**01** 税务大数据背景

**02** 实践案例分析

**03** 行业观察

Change



数据爆炸

# Status of Taxation



税务系统信息

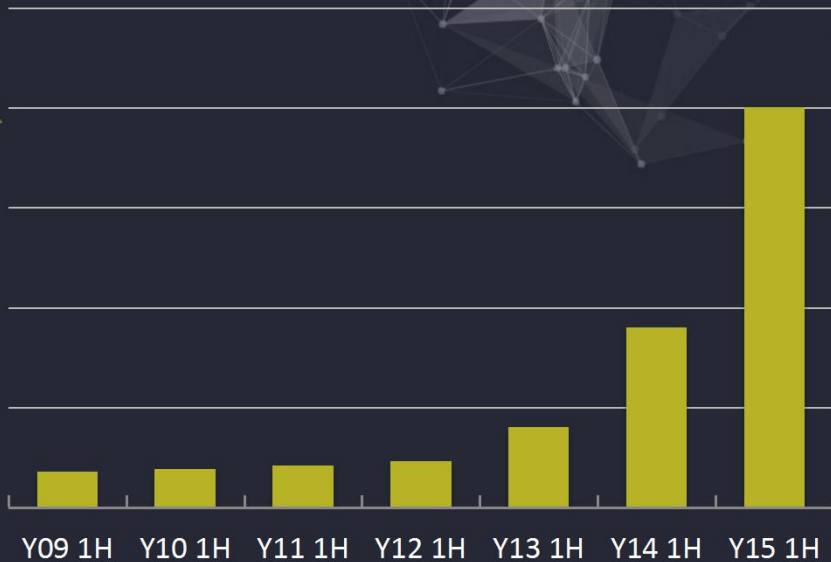


互联网数据



三方涉税信息

■ 自有系统



种类多

专业性强

关联性强

国税局主题数据

征管  
数据

增值税  
发票

车辆购  
置税

出口  
退税



# 2016

海淀破获特大虚开发票案

上百个空壳公司

涉案金额30多亿

流失税额达5亿余元

# 大型涉税案件 多方数据关联分析

4月份过后的国税局办税大厅

# 营改增

一般纳税人 **400万户**

年开票量 **几十亿份**

分各行业，各区域进行 **快速** 统计

# 数据时代的税务方法论



01

数据是直接的核查信息

02

数据可以作为交叉比对的  
辅助手段

03

多方数据流动和联合形  
成数据生态



# 《“互联网+税务”行动计划》

## 19.互联网+涉税大数据

将手工录入等传统渠道采集的数据和通过互联网、物联网等新兴感知技术采集的数据以及第三方共享的信息，有机整合形成**税收大数据**。运用**大数据技术**，开发和利用好**大数据**这一基础性战略资源，支撑纳税服务、税收征管、政策效应分析、税收经济分析等工作，优化纳税服务，提高税收征管水平，拓展税收服务国家治理的新领域。

在互联网上收集、筛选、捕捉涉税数据和公开信息，通过分析挖掘，为纳税人提供更精准的涉税服务，为税源管理、风险管理、涉税稽查、调查取证等工作提供信息支持。通过舆情监控，对纳税人需求和关注及时了解，及时采取应对措施，提高税收工作的针对性和有效性。

挑战

> > 30s

传统架构的增长曲线

理想增长曲线

MB

GB

10GB

100GB

TB

10TB

100TB

PB

营改增

税收  
特点

数据  
爆炸

有一小部分企业最先面临并成功解决了的问题  
—— 用**大数据**的方式

**互联网企业**



单选题：当需要拖动更多的载重时

- A. 找一匹更强壮的马
- B. 找十匹马

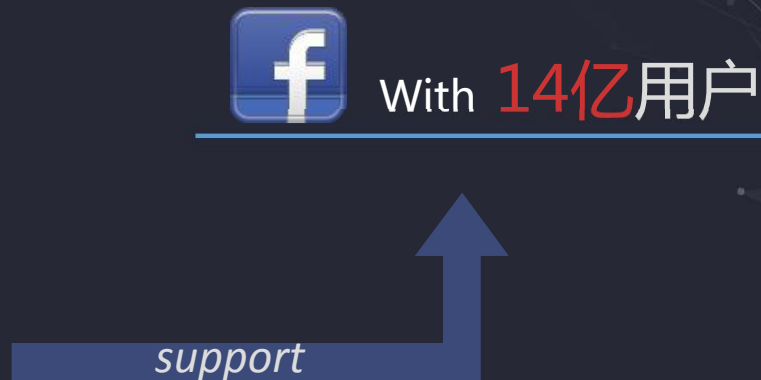
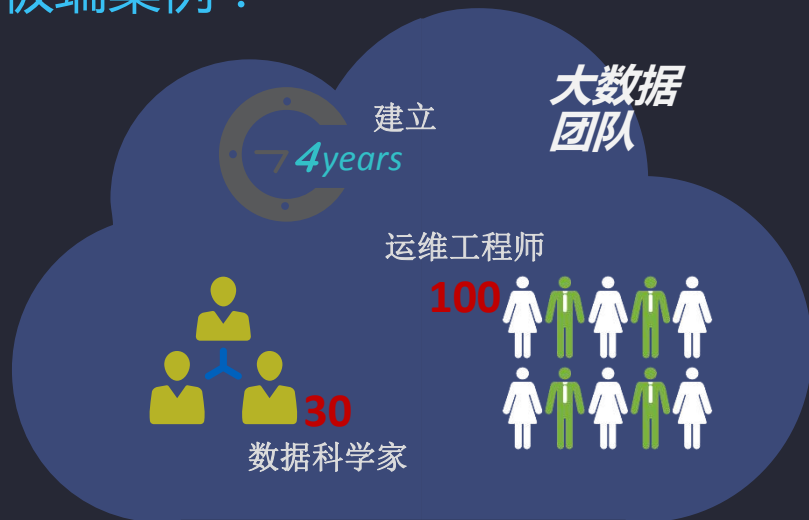
背后的思路：

3-5年内如何升级？

# 大道至简 分而治之

## 大数据提供了可扩展的数据应对机制

极端案例：



# 目录

---

**01** 税务大数据背景

**02** 实践案例分析

**03** 行业观察

我们关注

如何用**合理的成本**  
应对**海量增长**的数据

如何充分挖掘出  
数据价值来**提升业务**

达成前提：**数据的业务含义和关联**



# 大数据平台实施步骤

- ✓ 客服访谈
- ✓ 需求调研
- ✓ 整体规划

咨询规划

- ✓ 大数据基础平台
- ✓ 数据存储和计算能力

基础能力建设

- ✓ 数据汇集系统
- ✓ 各类数据大集中操作

数据汇集

- ✓ 数据治理体系规范
- ✓ 提高数据质量

数据治理

- ✓ 新型的数据应用
- ✓ 数据价值的变现

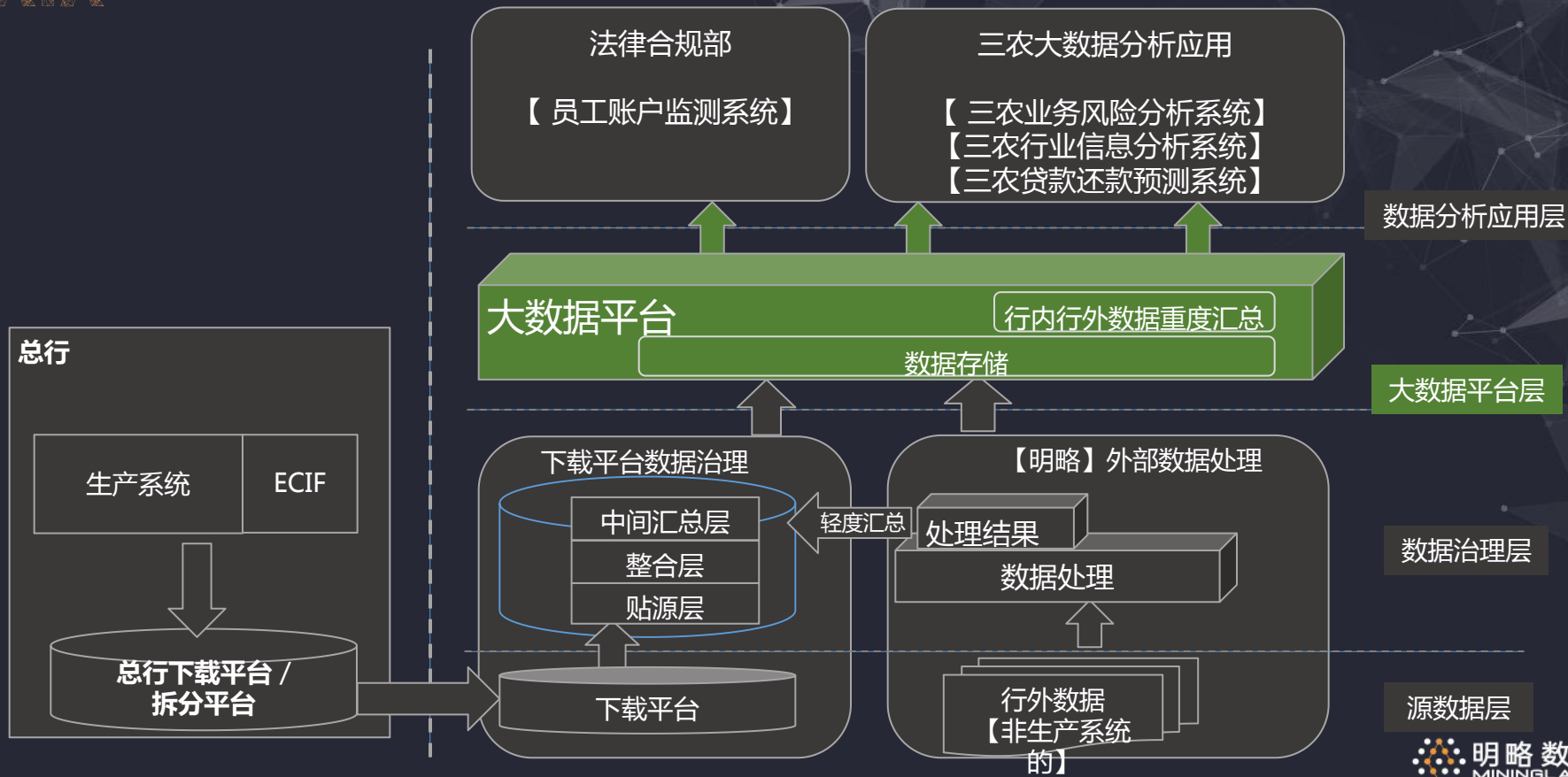
数据应用

- ✓ 数据运营
- ✓ 数据交互
- ✓ 数据合作
- ✓ 数据开发

数据运营



# 某行整体大数据架构





# 这些目前都不会发生

精细化的税源管理，大数据风控

纳税人画像分类，纳税遵从

基础环境  
数据资产管理

01

数尽其用  
数据分析挖掘

02

价值衍生  
数据运营

03

目前大多数政府部门  
还在第一阶段

开源产品成熟度  
远远不够

数据清洗占到成本70%以上

数据校验，转换，数值合并



统计

详情

- ☒ 省地税自有数据
- ☒ 三方数据
  - ☒ 保监局信息交换
    - ☒ 四川省财产保险公司车险赔付支出表
    - ☒ 四川省财产保险公司原保险收入及赔付支出表
    - ☒ 四川省省级财产保险企业名单
  - ☒ 财政厅信息交换
    - ☒ 财政拨款建设项目的拨款计划和拨款明细
    - ☒ 财政厅工资发放个人所得税扣缴信息
  - ☒ 发改委信息交换
    - ☒ 国民经济和社会发展规划主要指标
  - ☒ 工商局信息交换
    - ☒ 工商设立、变更、注销、吊销登记情况
  - ☒ 公安厅信息交换
    - ☒ 涉税信息交换\_网上网下时间统计
  - ☒ 国税信息交换
    - ☒ 查补税款信息
    - ☒ 车购税信息
    - ☒ 财务报表信息
    - ☒ 普通发票信息
    - ☒ 企业所得税信息
    - ☒ 税务登记信息



开始扫描

# BIG DATA 省国税局大数据性能提升

## 需求

按年度、按区县统计各个纳税人的纳税贡献度。单个纳税人的贡献度为：当年实际纳税金额总和/当年计税依据总和\*100%，算出贡献度后，按年度、区县聚合后进行排序

## 测试准备

- ✓ 数据集合是纳税人的申报明细记录，总共1000万条记录
- ✓ 测试环境是7台物理机器搭建的MDP集群。

## 测试结果

- ✓ 传统数据仓库 25分钟
- ✓ 大数据SQL引擎 9秒

基础环境  
数据资产管理

01

数尽其用  
数据分析挖掘

02

价值衍生  
数据运营

03



# 省级地税单位大数据平台落地后规划

## 数据管理能力

- ✓ 通过建设元数据管理系统和规范，清晰勾勒出地税税务数据资产的全貌，为后续基于数据的应用提供质量上的可靠保障
- ✓ 通过建设ETL可视化套件，将数据分析挖掘处理的关键过程去技术化，暴露在税务业务人员面前，使其能更深入理解和创新

## 数据应用能力

- ✓ 通过建立税源分析平台，并在其上构建种类丰富、形势多样的大数据应用，真正使税务数据资产变现。在政策决策辅助、税款追缴、提高纳税服务质量、降低纳税成本等方面发挥作用



首页



图析



地图



档案



协作



Admin



纳税人A



借贷



控股



诉讼



合同



诉讼



借贷



金融机构



关联的税人



执法机构A



纳税人B



执法机构



金融机构



统计

查询

显示

选中

删除

反选删除

挖掘

关系

预测

评估

决策

导航器



统计

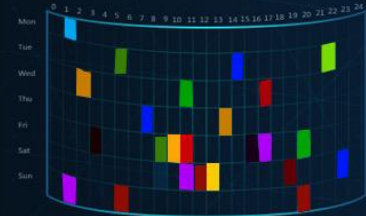
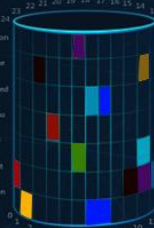
详情

时间轴

划分单位

西巷街道

属性

☐ 所有对象☐ 选中的对象

税种信息



营业税



个人所得税



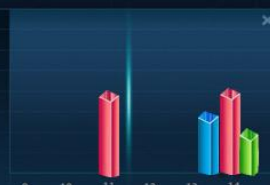
附加税



环保税

8  
7  
6  
5  
4  
3  
2  
1

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31



优化纳税服务

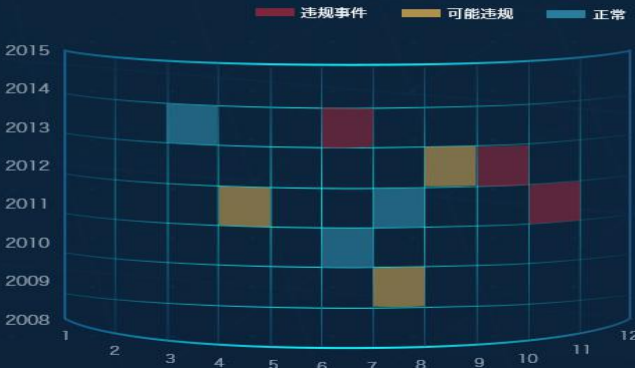
- 办税次数总和
- 评价转换值
- 税款欠缴次数
- 缴纳时间距离开始时间小于5天的次数
- 纳税人从申报开始到最终拿到完税证明的时间



社会经营关系



涉税事件追踪



税源画像



基本信息:  
企业名称: xxx 企业税号: XX123456 法人: 杨某  
地址: 北京市昌平区北苑路北中煤集团8号楼  
注册地址: 北京市昌平区东小口中煤建设集团大厦



基础环境  
数据资产管理

01

数尽其用  
数据分析挖掘

02

价值衍生  
数据运营

03

# BIG DATA 数据运营的价值体现

应用：提升服务能力

宏观

简化：让数据主控权回归业务人员

微观

基础：数据处理能力极大提升

领导视图 风控	行业报告 跨部门
稽查 税源分析	纳服 增值服务

对内

对外

# 目录

---

**01** 税务大数据背景

**02** 实践案例分析

**03** 行业观察

客户真实需求  
突破传统数仓

TAX

大量发票，三方，互联网数据进入视野

以纳税人为核心进行快速组织和检索

深化风控和预测模型

数据汇集  
关联分析

感知 & 学习

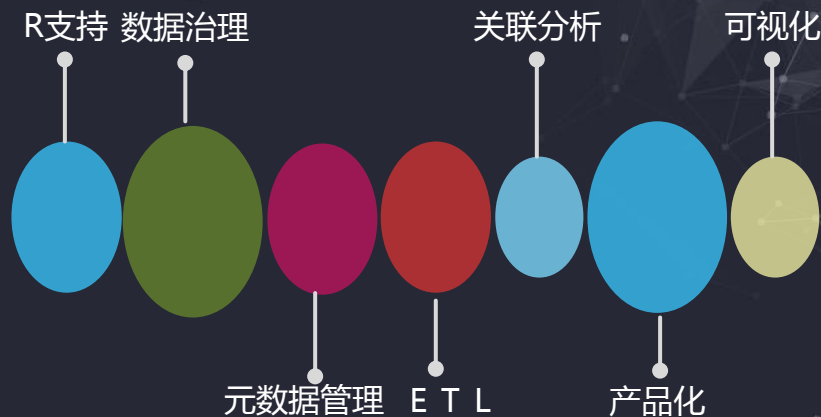


# 理想与现实的差别

## 憧憬



## 现实



真实需求：高效直观的统计分析



我们需要的大数据公司？





# 经验丰富的数据猎手

互联网级别的数据经验

统计和工程结合，核心参与开源

咨询到落地的全栈服务

相对成熟的商业化产品



欢迎一起迈入2016 ——

大数据元年

Thanks !