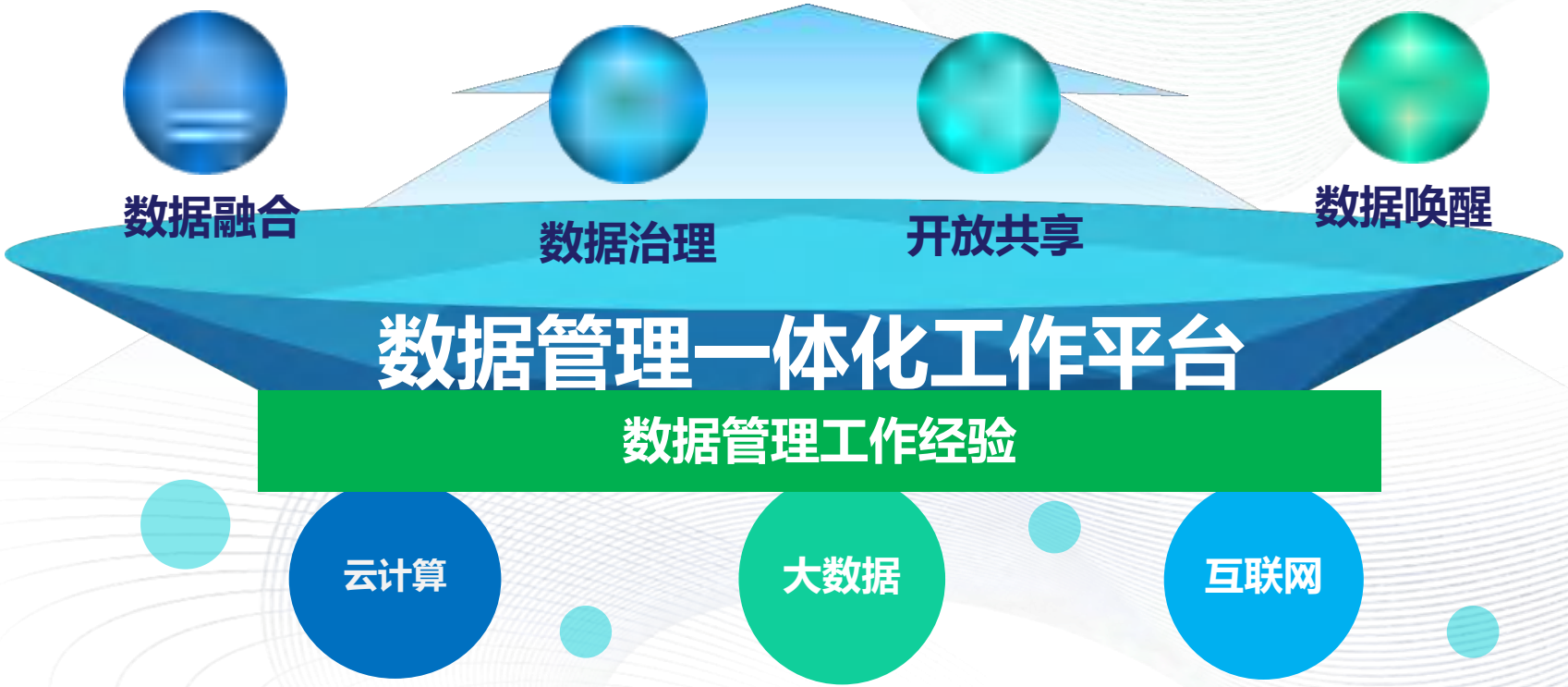
数据治理大数据平台

资源规划与建设

目录

01 项目概述-整体介绍

全面覆盖　全量提供　深度挖掘



01 项目概述-建设目标

聚合数据

- 对多来源、多种类涉税数据进行全面采集、整合、规范，形成全面可用的数据资产

统一治理

- 建立统一的数据标准，实现数据的统一采集、统一管理、统一应用

开放服务

- 以数据服务超市的形式将数据服务化、可视化，实现数据资产的内部统一、跨部门共享和对外开放

智能应用

- 通过应用创新中心提供灵活、智能的数据应用，实现海量涉税数据的挖掘应用

建设原则

01 标准先行

标准是数据管理工作的基础，严格执行总局已有的标准规范，并在此基础上，按需完善XX省地税相关标准。

02 融合提升

与各业务系统科学集成，有益补充各种数据源，建设全局性的数据管理系统，全面提升数据应用能力。

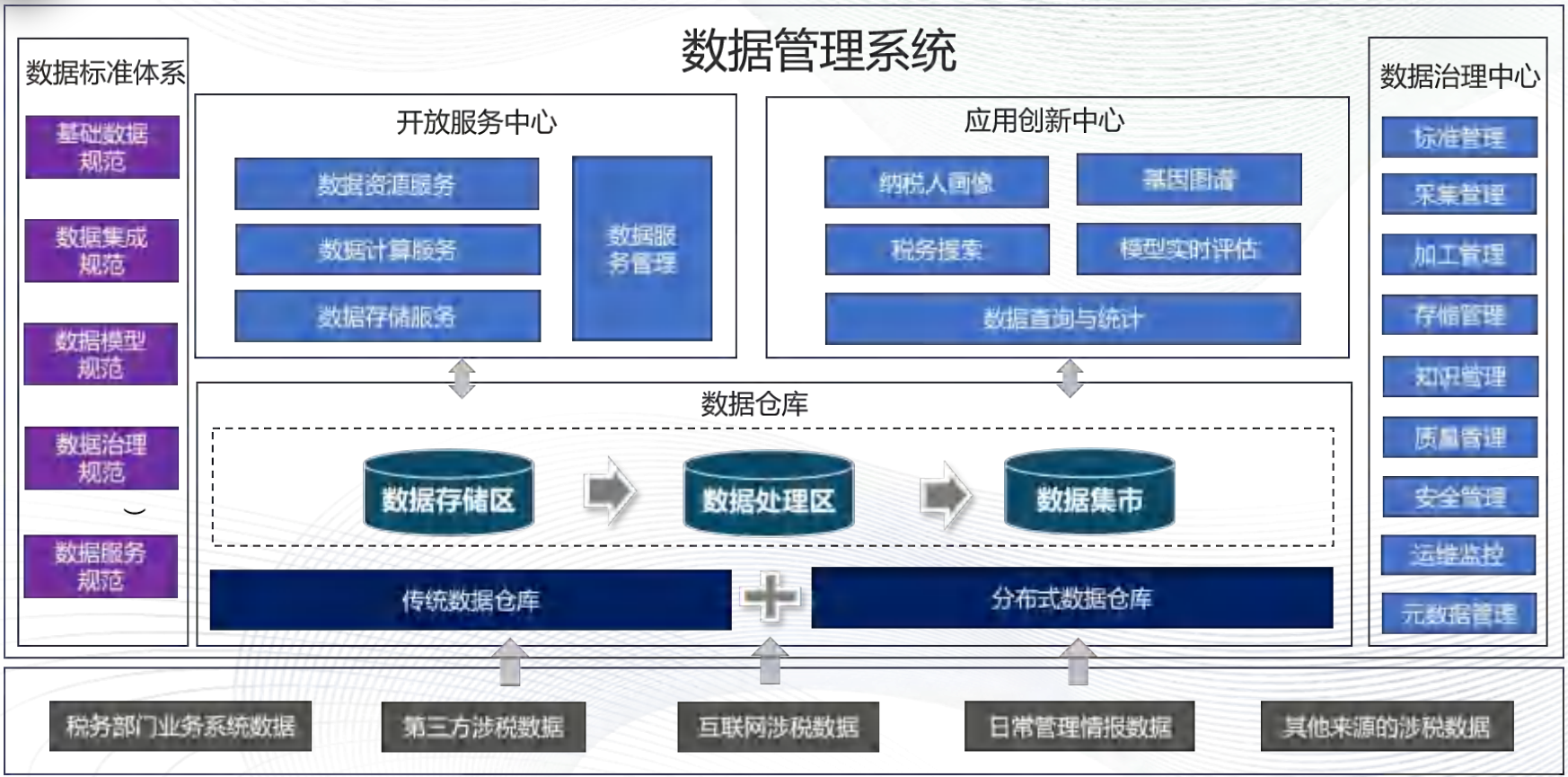
03 开放创新

打破数据管理条块分割的限制，实现数据资产内部统一、跨部门共享和外部开放的生态环境。

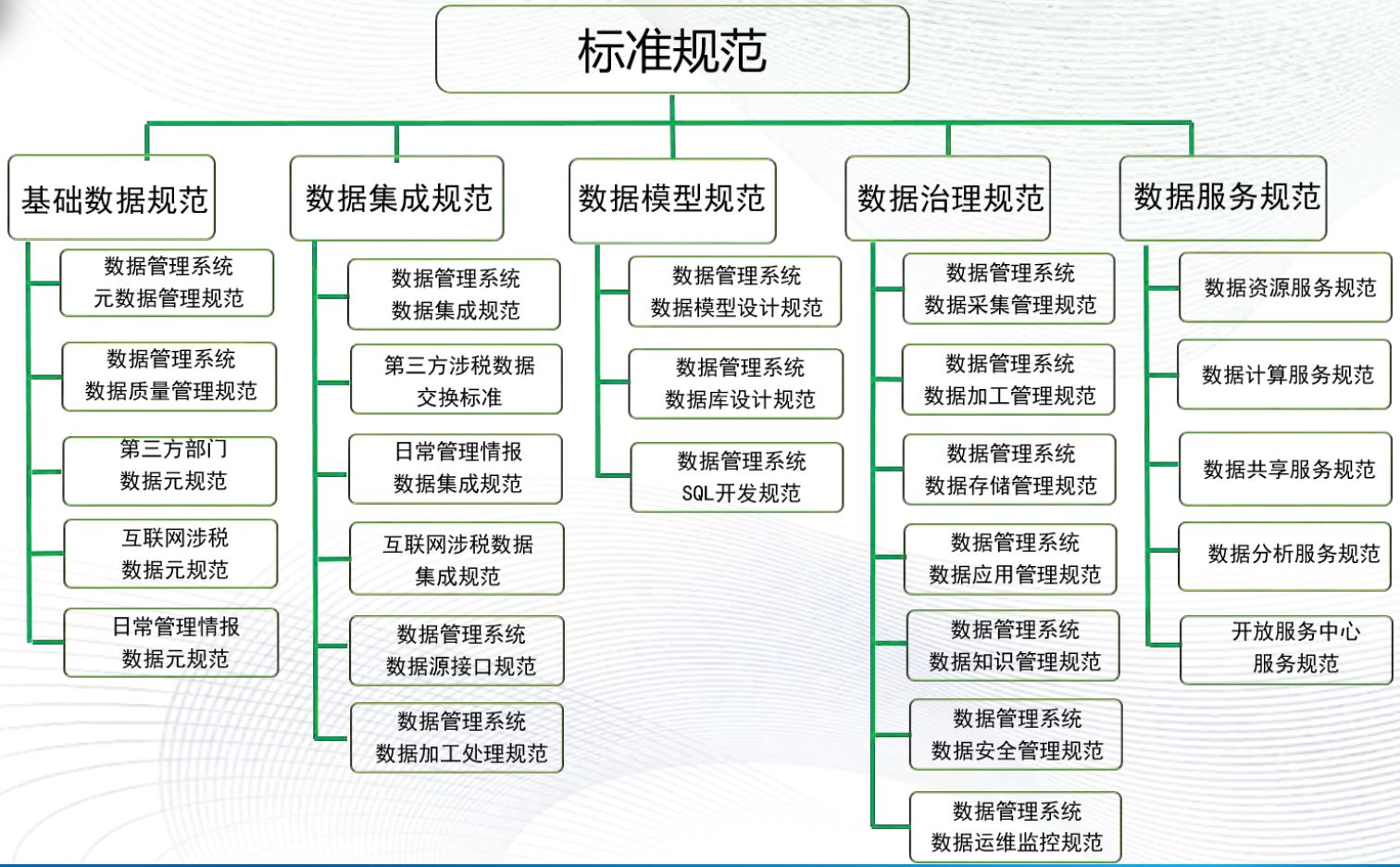
04 循序渐进

数据管理系统内涵丰富，建设工作不可一蹴而就，应采用“迭代”模式，循序渐进、有条不紊地进行系统的搭建工作。

01 建设内容-整体架构

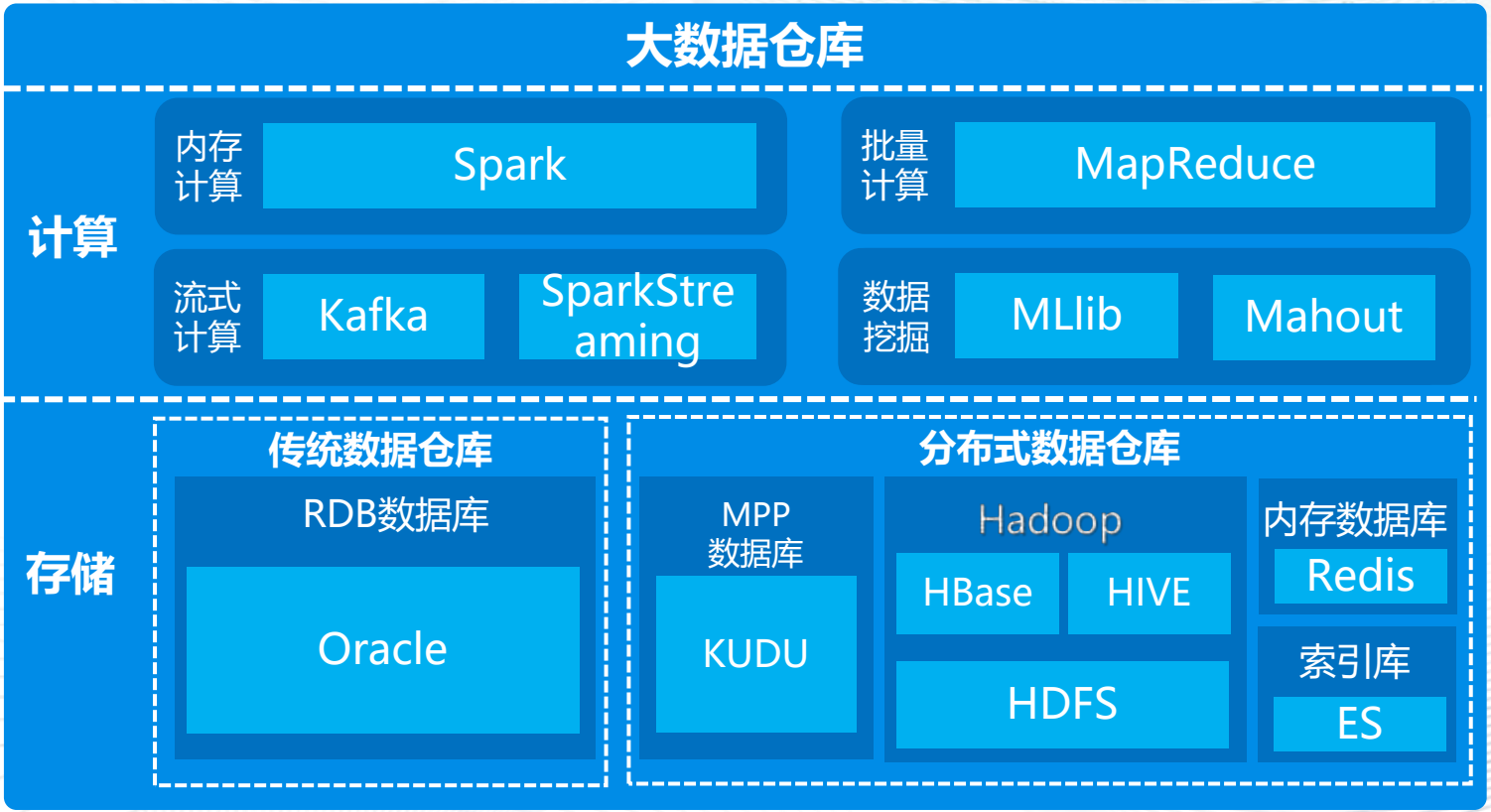

数据管理系统
数据标准体系
基础数据规范
数据集成规范
数据模型规范
数据治理规范
数据服务规范
开放服务中心
数据资源服务
数据计算服务
数据存储服务
数据服务管理
应用创新中心
纳税人画像
基因图谱
税务搜索
模型实时评估
数据查询与统计
数据治理中心
标准管理
采集管理
加工管理
存储管理
知识管理
质量管理
安全管理
运维监控
元数据管理
数据仓库
数据存储区
数据处理区
数据集市
传统数据仓库
分布式数据仓库
税务部门业务系统数据
第三方涉税数据
互联网涉税数据
日常管理情报数据
其他来源的涉税数据

01 建设内容-一套标准

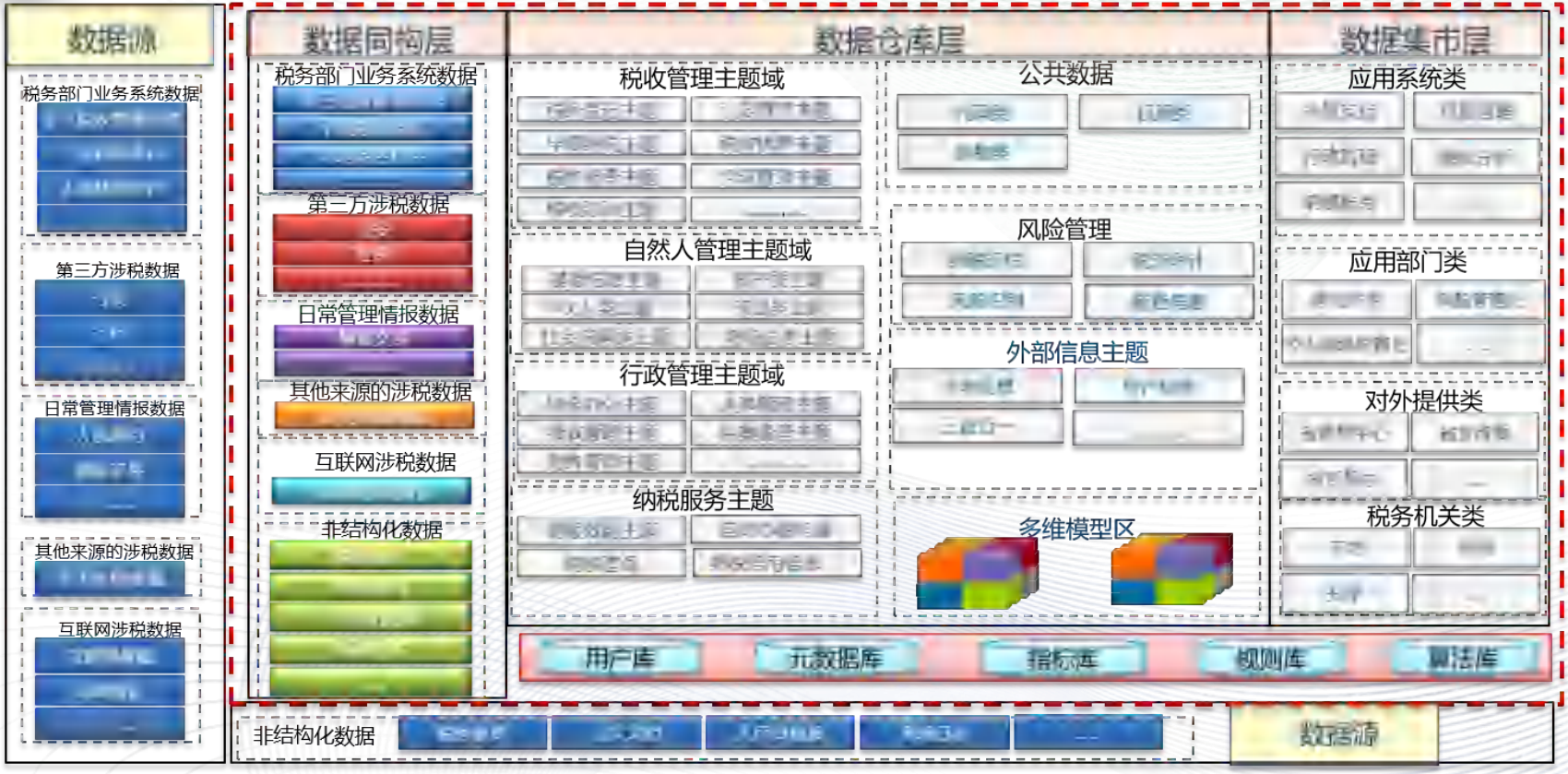

标准规范
基础数据规范
数据管理系统 元数据管理规范
数据管理系统 数据质量管理规范
第三方部门 数据元规范
互联网涉税 数据元规范
日常管理情报 数据元规范
数据集成规范
数据管理系统 数据集成规范
第三方涉税数据 交换标准
日常管理情报 数据集成规范
互联网涉税数据 集成规范
数据管理系统 数据源接口规范
数据管理系统 数据加工处理规范
数据模型规范
数据管理系统 数据模型设计规范
数据管理系统 数据库设计规范
数据管理系统 SQL开发规范
数据治理规范
数据管理系统 数据采集管理规范
数据管理系统 数据加工管理规范
数据管理系统 数据存储管理规范
数据管理系统 数据应用管理规范
数据管理系统 数据知识管理规范
数据管理系统 数据安全管理规范
数据管理系统 数据运维监控规范
数据服务规范
数据资源服务规范
数据计算服务规范
数据共享服务规范
数据分析服务规范
开放服务中心 服务规范

01

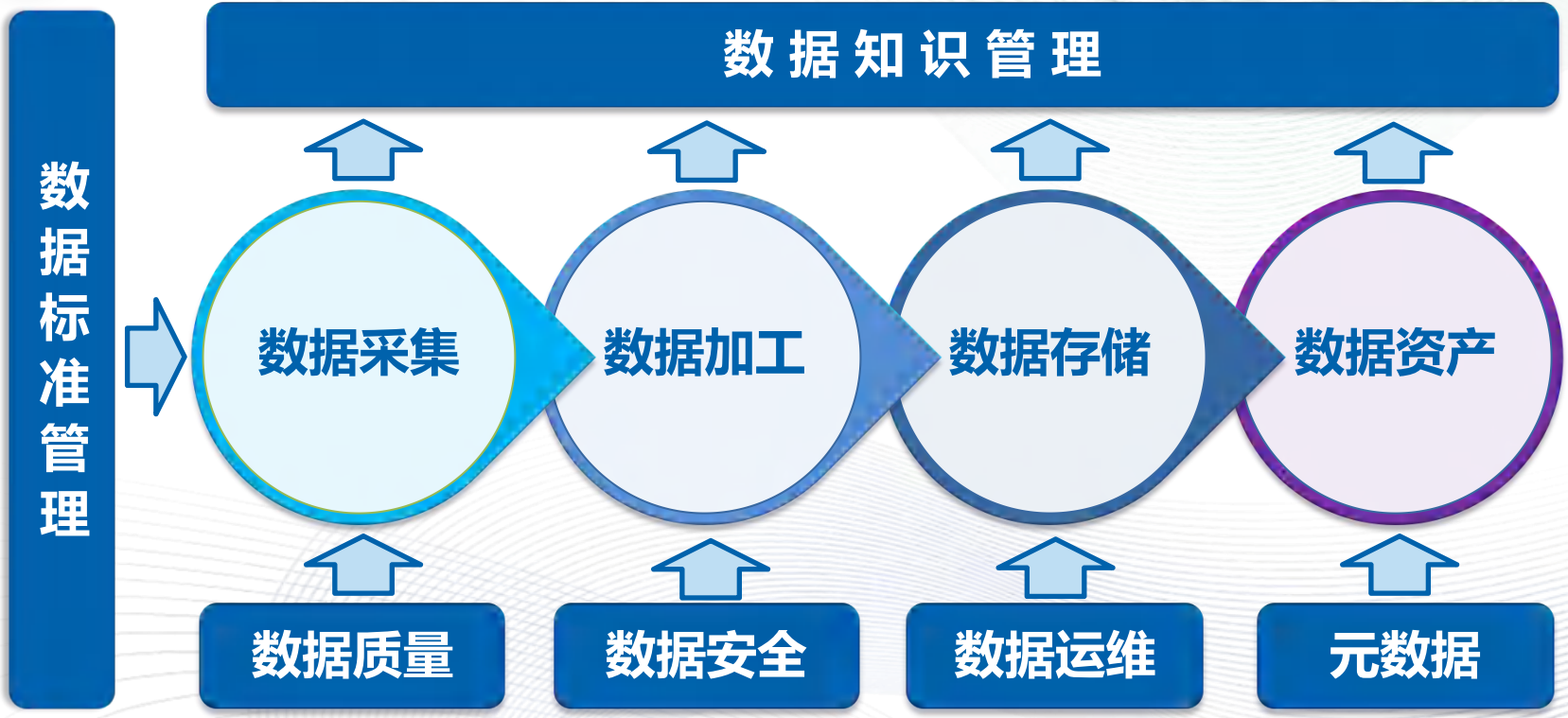
建设内容-仓库技术架构



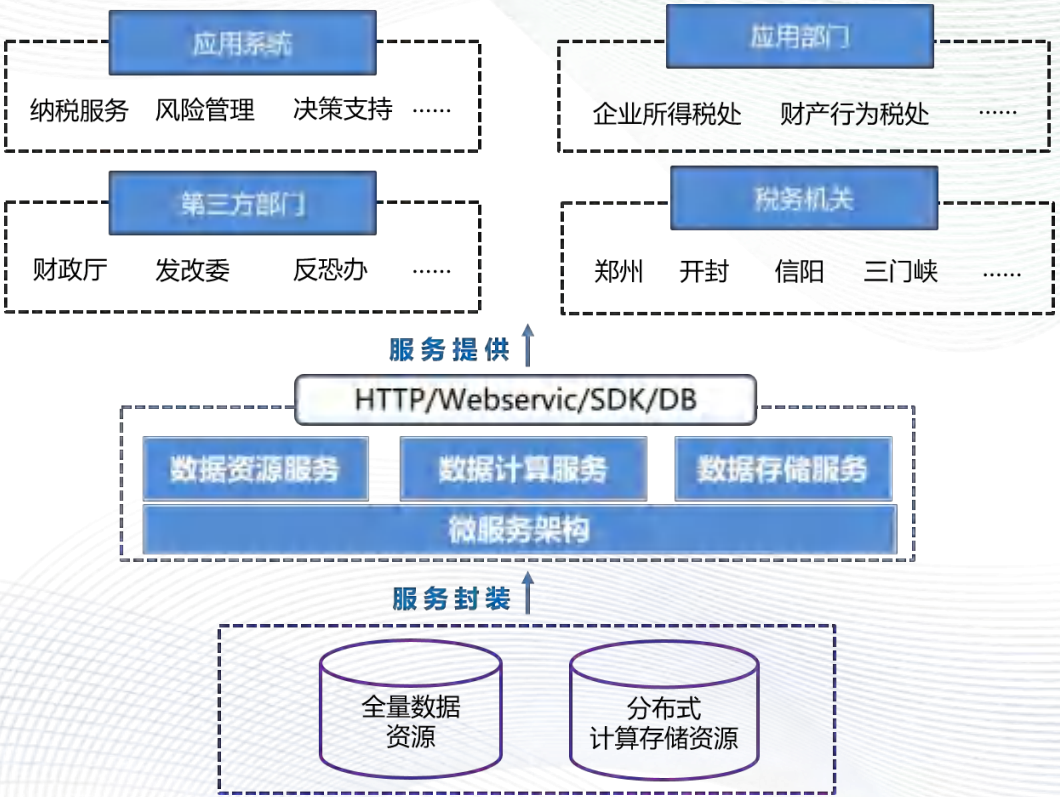
01 建设内容-仓库逻辑架构



建设内容-数据治理中心

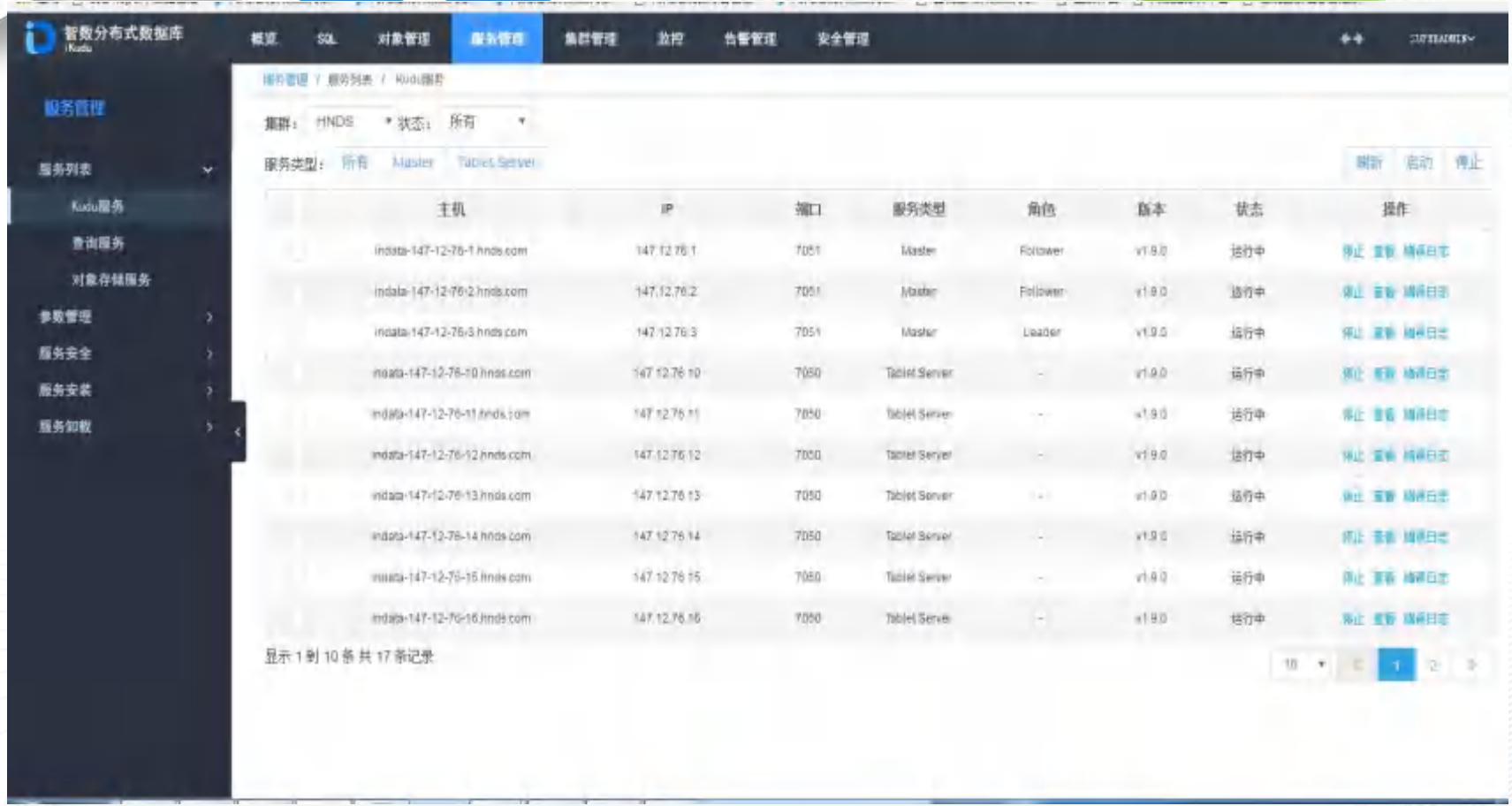


01 建设内容-开放服务中心

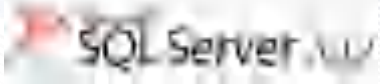


01 建设内容-应用创新中心

01 系统特点-系统架构先进



01 系统特点-系统架构先进

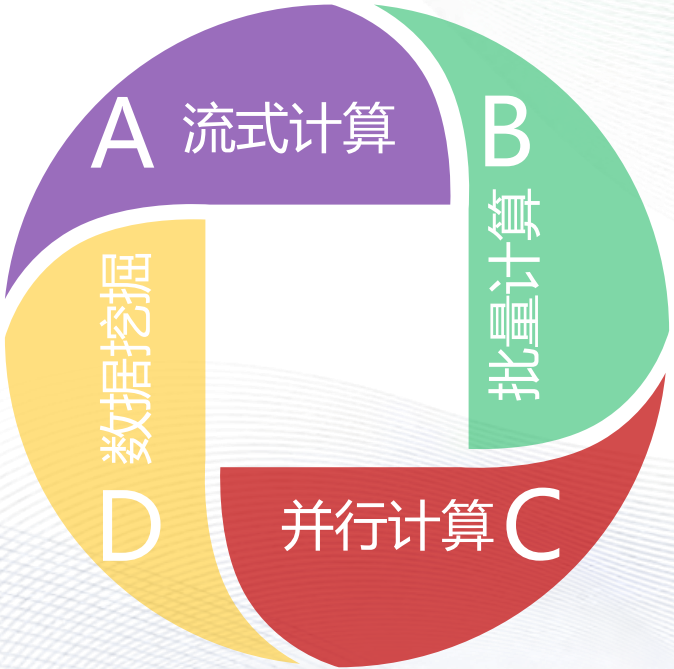


01 系统特点-运算能力强大

Spark Streaming

A

采用内存计算和流式计算技术，满足可视化实时展示、事中风险监控、重点企业数据比对等业务的实时性处理要求。



Hvie

B

采用分布式大数据计算引擎，支持数据聚合、汇总、比对等各种常见的数据分析场景，支持分钟级的离线批量数据处理。

Spark+MLlib

D

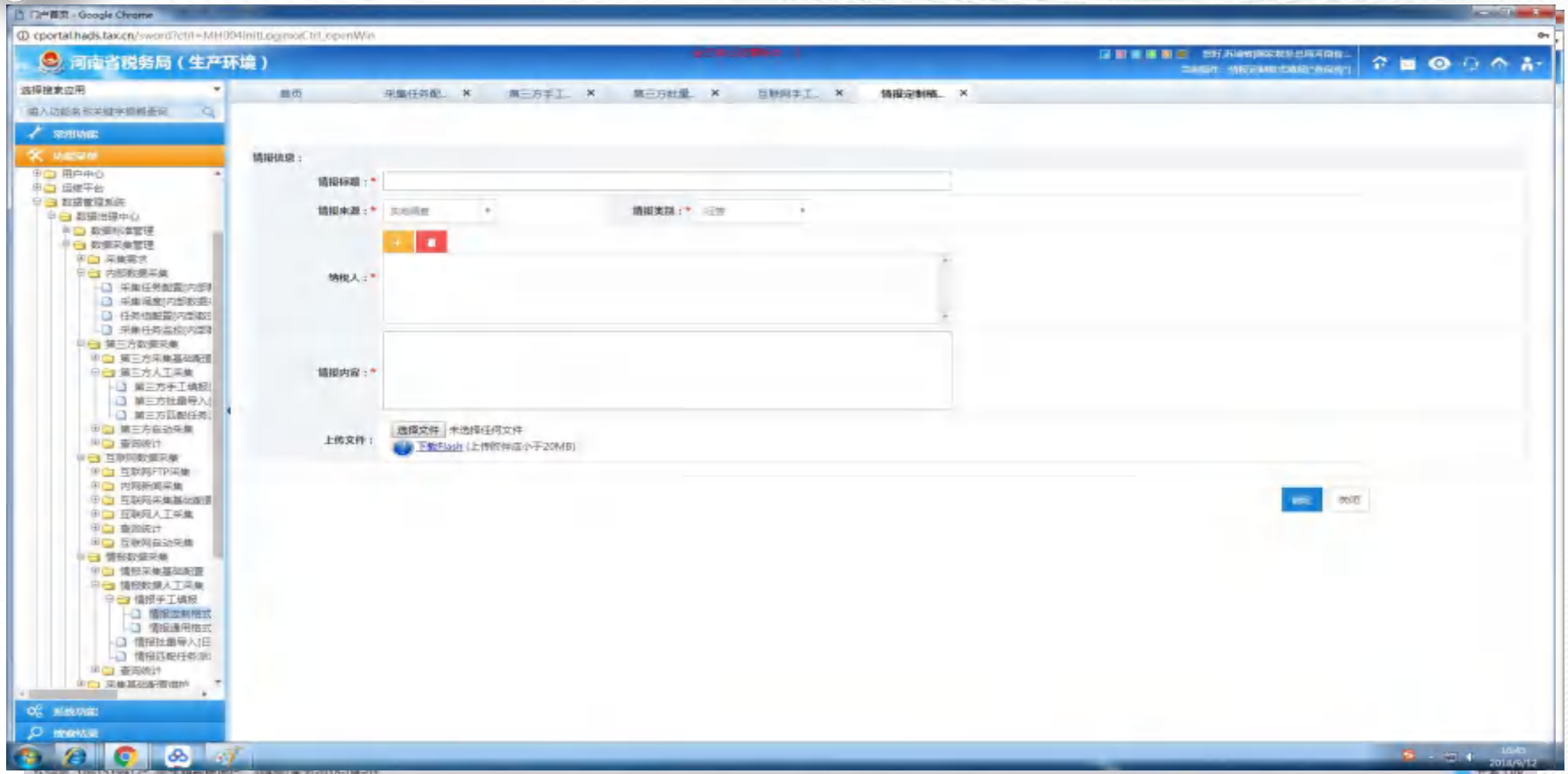
采用spark计算引擎和MLlib算法库，挖掘数据中隐含的知识，实现数据的深度利用。

MPP

C

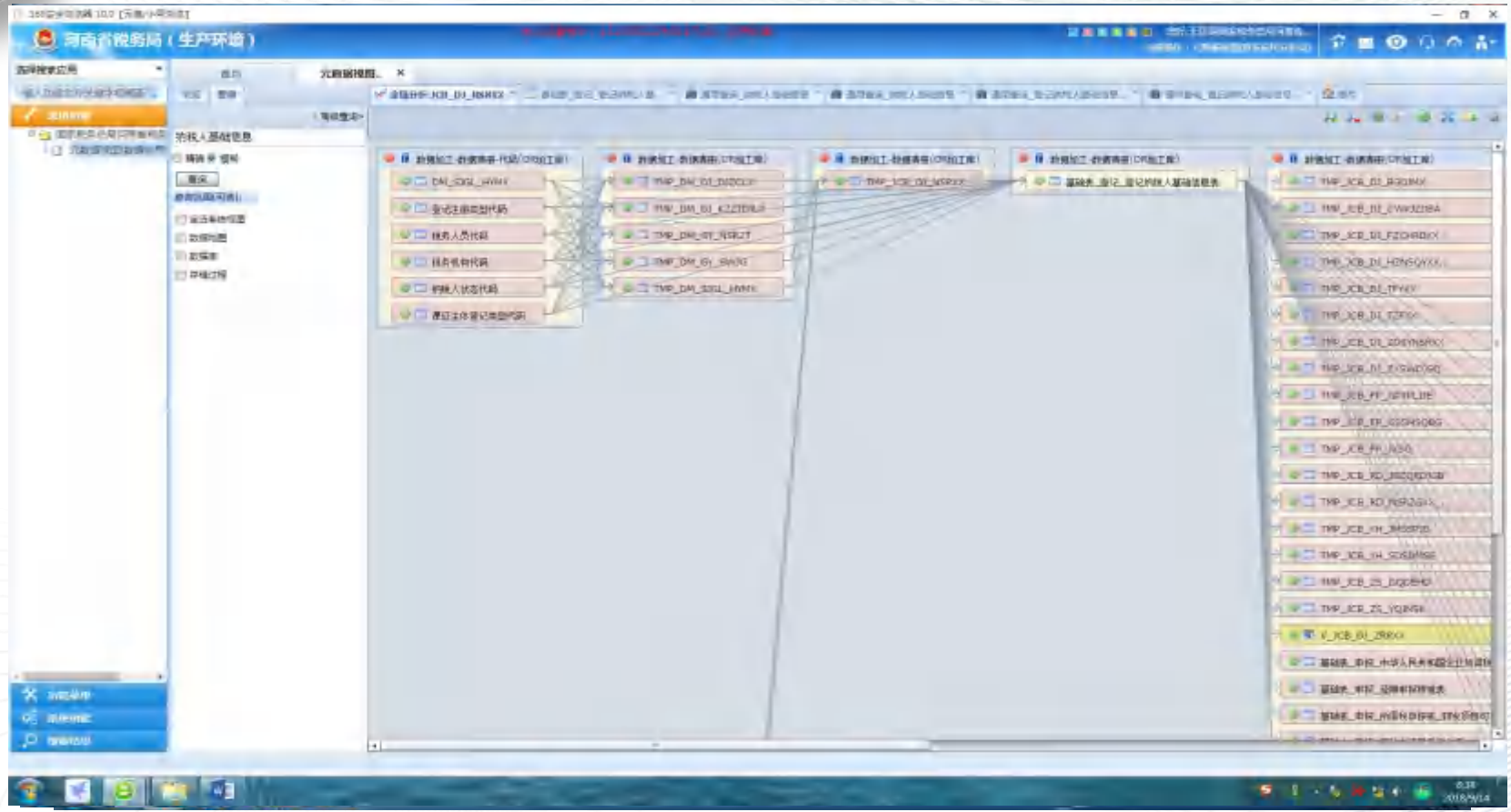
采用大规模并行处理技术，支持多服务器、多处理器、多进程并行处理，支持秒级的交互式数据计算和即席查询。

01 系统特点-采集方式丰富多彩



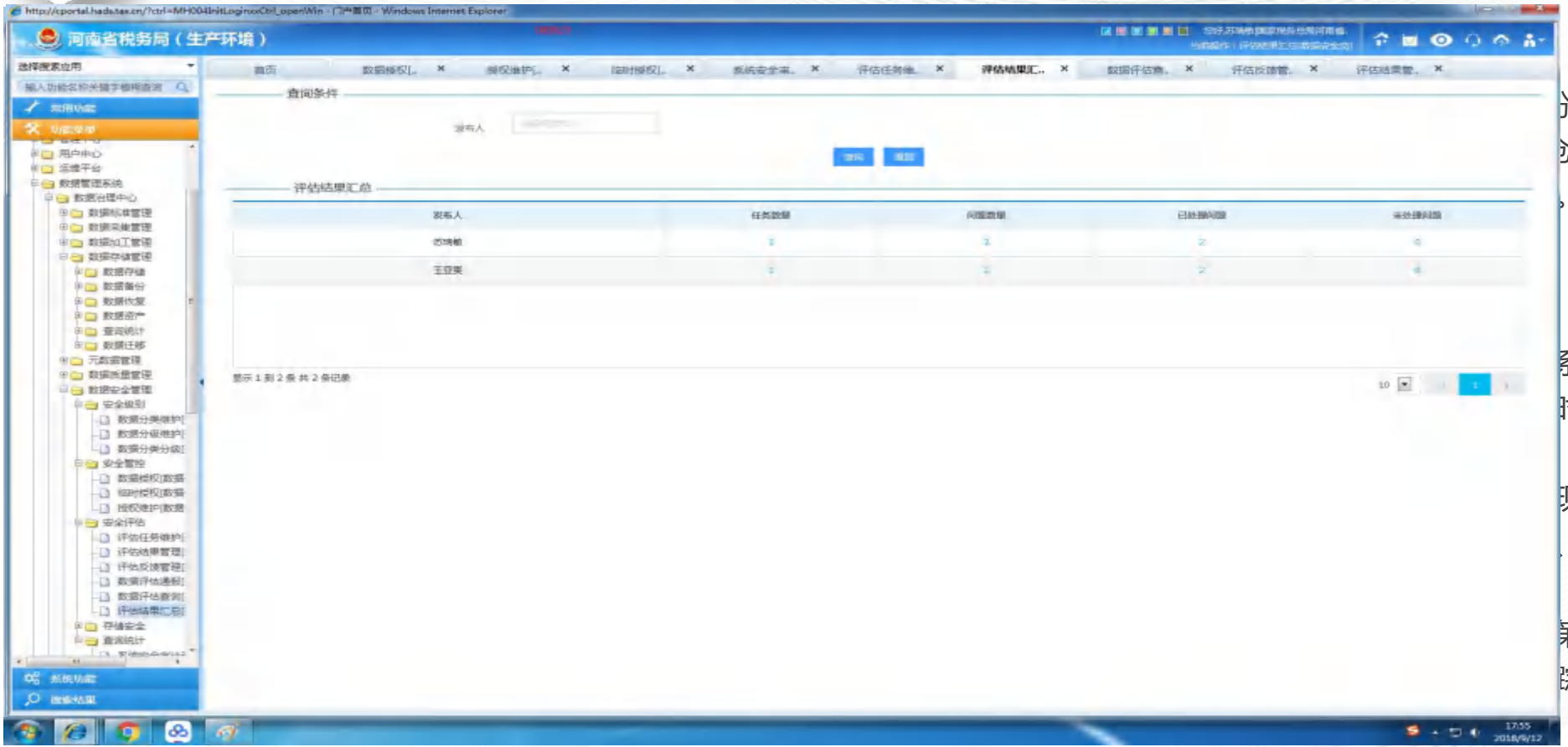
01 系统特点-数据质量闭环管理

01 系统特点-数据资产直观清晰

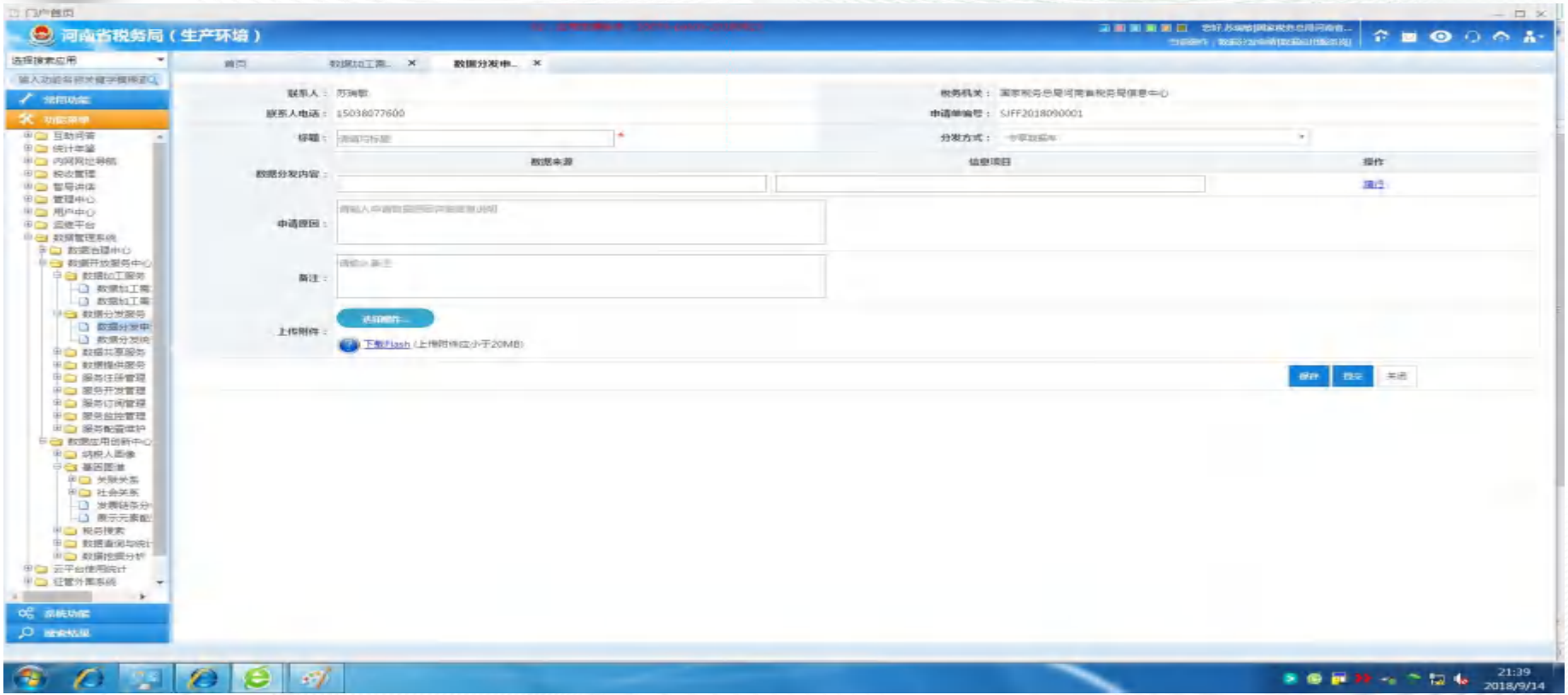


01 系统特点-数据服务便捷高效

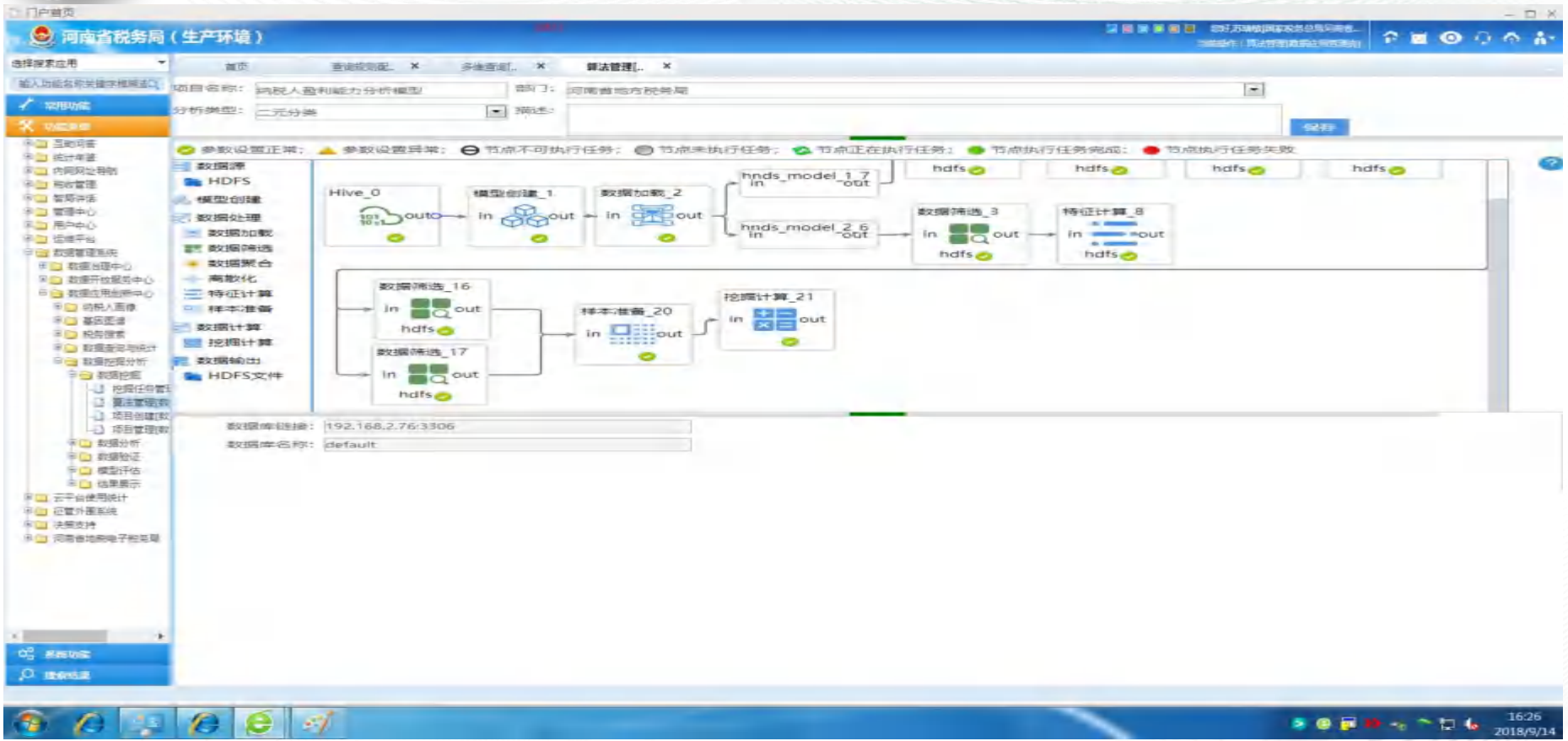
01 系统特点-数据安全严密管控



01 系统特点-个性化需求响应快捷



01 系统特点-数据应用创新多元



01 系统特点-即时掌握数据状态

数据资产

挖掘分析

数据质量

数据标准

纳税人画像

税务搜索

查询统计

基因图谱

元数据

01 平台建设成果

数据仓库	集中统一税务数据资源平台
数据治理	全局参与数据管理工作平台
开放共享	全面开放数据服务平台
丰富工具	提供多种数据分析工具

02 建设背景

1. 新形势下税收工作的要求
2. 数字税务建设的现实需要
3. 新技术新理念的成功运用
4. 总局大数据云平台的功能延伸

02 建设目标

全面集成各类数据

实现数据标准化

实现业务能力提升

数据管理系统

数据中台

云计算

大数据

微服务

02 系统定位

涵盖了大量跨省数据资源，为开展全国性纳税人关联分析提供了数据支持

口径统一、体系完备的数据查询统计功能，满足了日常数据查询、统计和报表

3、着眼于基层亟待解决的数据应用难题，不贪大求全，对决策支持系统1包已经具备的查询统计菜单功能不再重复建设。

1、着眼于数据管理全流程，侧重于数据治理、数据服务和数据挖掘分析，支撑各级税务机关创新应用。

2、着眼于大数据处理技术在税务领域的运用，侧重于对全量历史数据的聚类分析、趋势分析。

02 建设思路

全省统一部署的税务大数据平台

02 建设原则

开放性

平台面向不同应用系统和功能需求，采用开放式设计，提供丰富的数据访问及应用集成接口

统一性

平台提供统一的界面风格，操作方式应符合用户使用习惯

扩展性

平台的设计和建设要充分考虑网络和硬件的扩展需要。

安全性

平台建设要保证软、硬件安全、可靠地运行，要有容灾、容错预案等。

维护性

平台的数据资源、组件工具、开发运行都应提供方便、灵活、直观的维护手段，方便进行维护和管理。

建设策略

01 夯实基础

一是以数据全面整合、统筹规划、目录先行、标准服务为核心策略；二是以数据应用服务为核心。

02 丰富应用

丰富数据应用，解决广大税务干部在数据分析应用方面的难点、痛点问题。

03 加强管控

围绕统一规划、统一管控、统一开发、统一应用的建设策略进行项目建设。

04 加强创新

1、通过数据分析、数据挖掘、数据可视化等工具，打造创新应用能力支撑体系；

2、通过机制创新、人才培养、协作共享等，形成创新生态，激发全局创新活力。

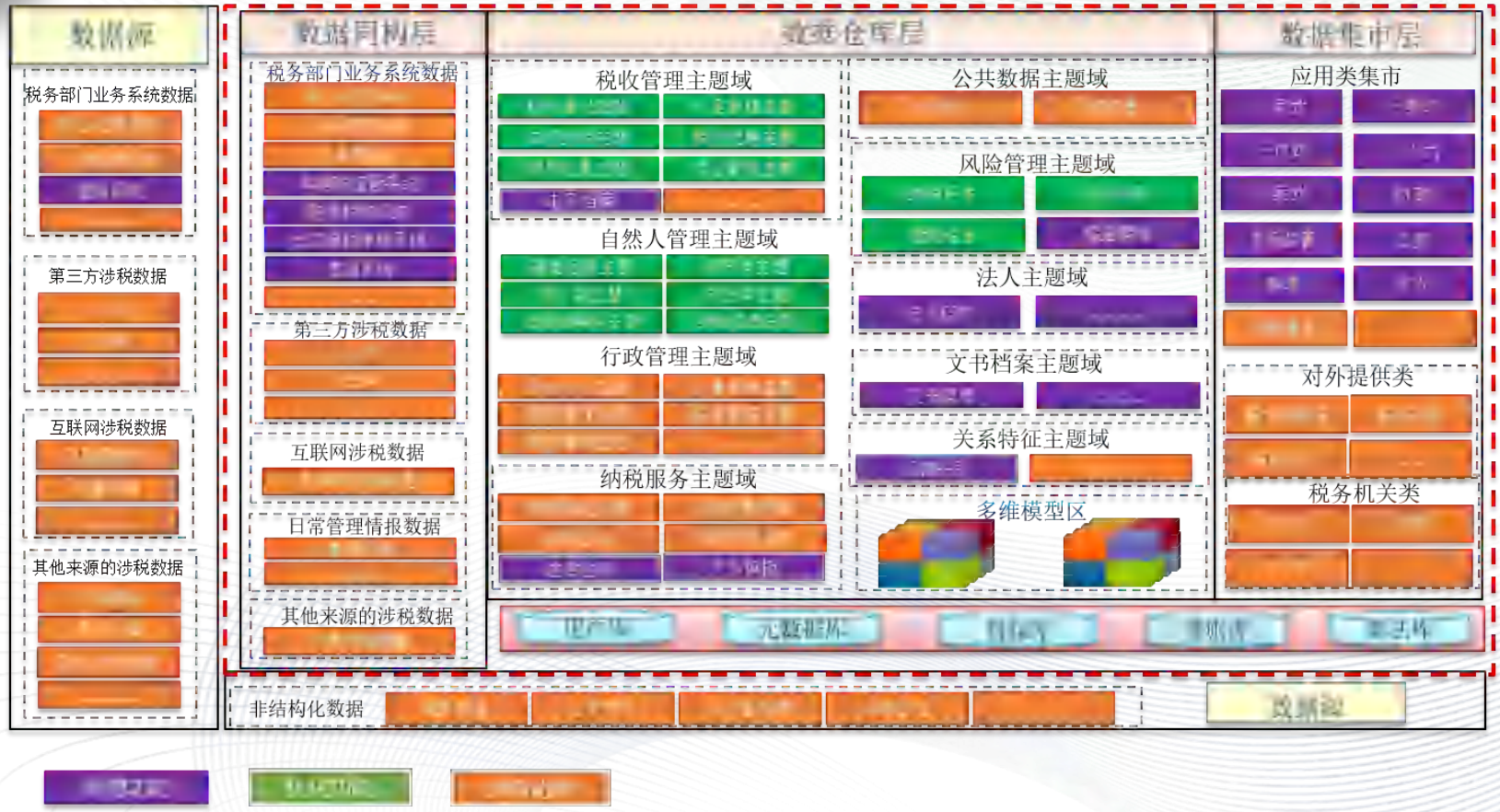
02 总体架构设计


金税三期标准规范
业务前台
数据应用中心
数据分析应用集市
数据挖掘应用专题
智能应用空间
自主创新应用空间
...
数据中台
数据服务中心
数据资源服务
数据计算服务
数据存储服务
数据应用服务
数据分析挖掘服务
数据仓库
数据同构层
数据仓库层
数据集市层
分布式数据仓库
传统关系数据库仓库
数据治理中心
数据标准管理
数据采集管理
数据加工管理
数据资产管理
元数据管理
数据质量管理
数据安全管理
数据运维监控
数据知识管理
数据源
税务部门业务系统数据
第三方涉税数据
互联网涉税数据
...
金税三期安全体系

02 应用架构设计


数据应用中心
数据分析应用集市
全景式数据应用集市
一局式监控分析
一员式监控分析
一户式监控分析
一人式监控分析
一案式监控分析
外部数据分析应用集市
财政数据应用
市场监管数据应用
公安数据应用
住建数据应用
水利数据应用
专题数据分析应用集市
税源分析
收入分析
数据挖掘应用专题
申报数据挖掘专题
发票数据挖掘专题
登记数据挖掘专题
自然人挖掘专题
智能应用空间
智能推送
智能预测
自主创新应用空间
自助挖掘
数据服务中心
服务目录
服务检索
服务发布
数据仓库
数据治理中心

02 数据架构设计


数据源
税务部门业务系统数据
第三方涉税数据
互联网涉税数据
其他来源的涉税数据
数据贴源层
税务部门业务系统数据
第三方涉税数据
互联网涉税数据
日常管理情报数据
其他来源的涉税数据
业务仓库层
税收管理主题域
自然人管理主题域
行政管理主题域
纳税服务主题域
公共数据主题域
风险管理主题域
法人主题域
文书档案主题域
关系特征主题域
多维模型区
数据集市层
应用类集市
对外提供类
税务机关类
非结构化数据

02 技术架构设计


应用组件
可视化
工作流
CMSP
单点登录
门户管理
权限管理
.....
数据服务
RESTFUL
WebService
EJB
服务网关
负载均衡
服务路由
数据计算
流式计算
kafka
Spark Streaming
批量计算
Spark
数据挖掘
Python
MLlib
数据存储
关系型数据库
Oracle
分布式
hadoop
KUDU
索引库
ES
内存数据库
redis
数据采集
nifi
OGG
互联网采集
数据来源
数据库
互联网
文件
音视频

02 技术架构设计

1、Hadoop 2.X升级至Hadoop 3.X，分布式存储、计算更高效。

2、升级后hive支持更新处理，能够更好的支持数据加工

3、引入流式数据同步工具NIFI，提高数据同步的效率和数据可靠性

4、升级搜索引擎升级到Elasticsearch，进一步提升搜索引擎效率和搜索命中率；

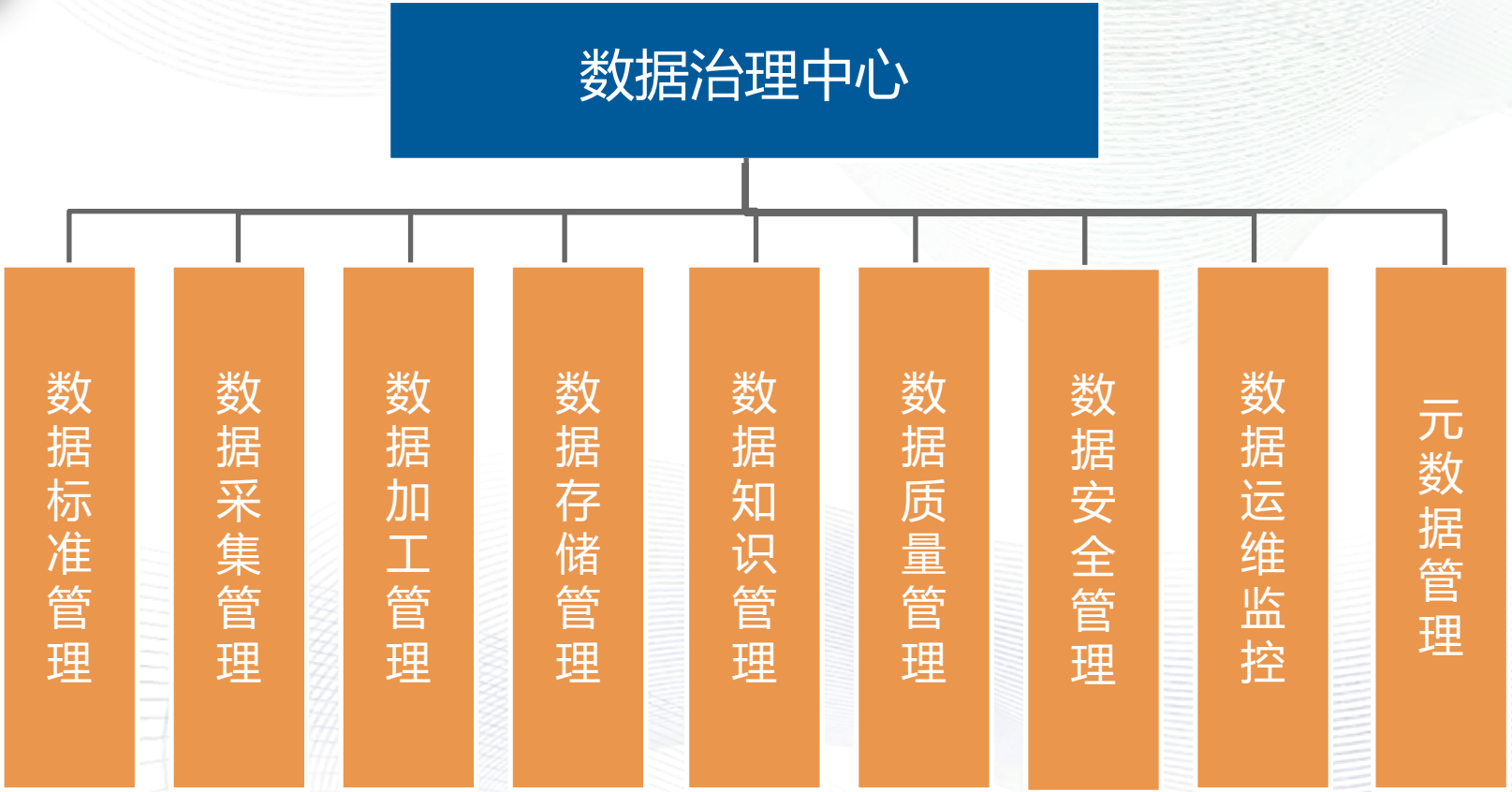
5、引入多租户功能，实现HDFS、Hive、HBase、Kafka、Spark等大数据服务的多租户管理功能；

6、升级MLlib、Mahout算法库版本，丰富算法内容和提升算法效率，为数据挖掘和智能应用提供有力支撑；；

02 数据治理中心

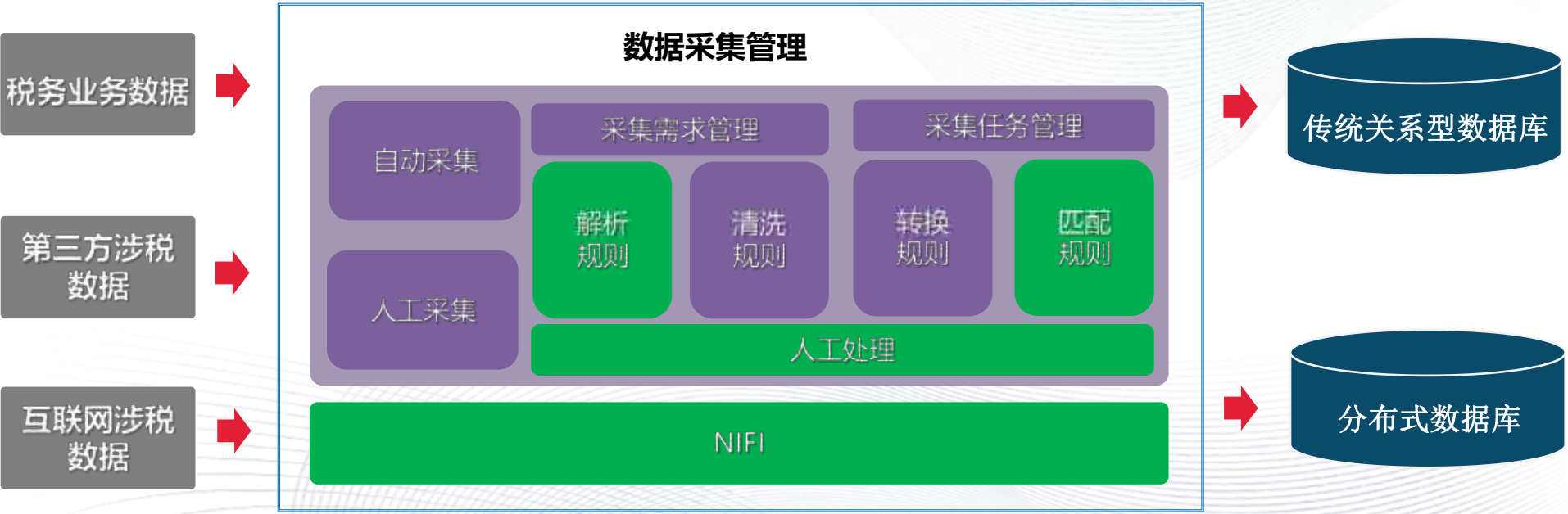
数据标准规范化	• 实现数据标准的规范化管理，构建“有标可依、依标可行、行而有效”三位一体的、可持续发展的数据标准体系
数据采集全面化	• 全面采集税务业务系统数据、第三方涉税数据、日常管理情报数据、互联网涉税数据和其他来源的涉税数据
数据管理流程化	• 实现流程化导向式的数据管理，使数据管理的角色和职责有明确划分，数据认责清晰，提升数据管理效率
数据资产可视化	• 实现对数据间流转、依赖关系的影响和血缘分析，使数据资产可视化
数据质量度量化	• 全方位管理全局的数据质量，实现可定义的数据质量检核和维度分析，以及任务化的问题跟踪处理

数据治理中心


数据治理中心
数据标准管理
数据采集管理
数据加工管理
数据存储管理
数据知识管理
数据质量管理
数据安全管理
数据运维监控
元数据管理

02 数据治理中心-数据标准

02 数据治理中心-数据采集



02 数据治理中心-数据加工

支持脚本在线运行、在线编写、版本控制、脚本检索、脚本共享等功能

对数据进行聚合运算、分组、计数、求最大值和最小值、求和、求均值等操作，并能够快速创建各类数据加工逻辑

02 数据治理中心-数据资产

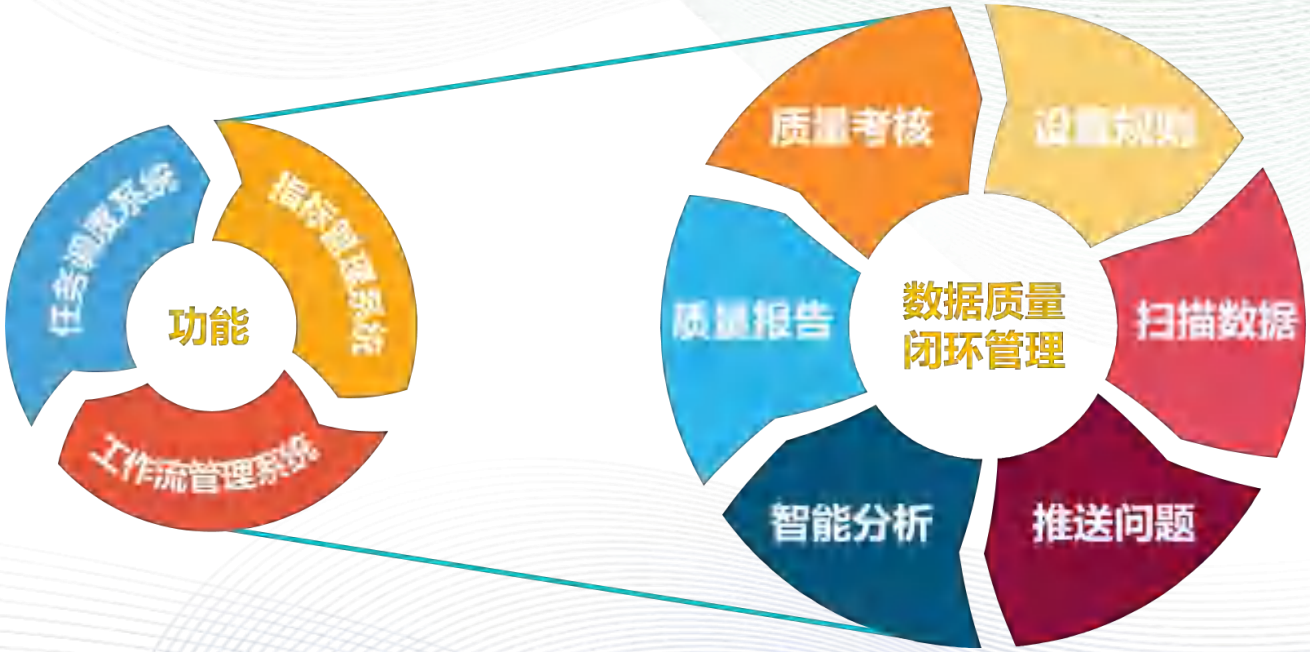
- 数据资产目录
- 数据资产配置
- 数据资产检索
- 数据资产报告
- 数据资产分析

- 表证单书
- 数据模型
- 数据服务
- 质量检测规则
- 标签
- 业务代码表
- 数据元

02 数据治理中心-元数据

1、元数据活跃度分析
2、一致性分析的功能模块

02 数据治理中心-数据质量



新增纳税人“一户式”数据质量分析、并通过网页、手机应用等方式推送至纳税人端提醒整改；问题数据总体监控多维度群体分析、多指标关联分析、纳税人维度智能分析、数据治理情况分析

02 数据治理中心-数据安全管理

01
数据加解密

02
数据脱敏管理

03
数据访问权限

04
日志审计

数据治理“一张图”

运维预警

组件维护

02

数据治理中心-知识管理

互助问答

风险案例共享

02 数据服务中心

各级税务机关

数据服务 | 专享库

查询统计应用

省市县审批

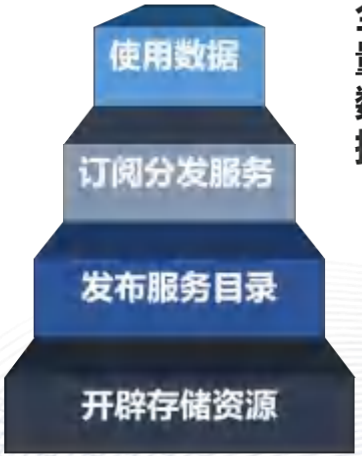
加工需求 | 查询统计

数据服务申请

场景A

数据加工服务

各省辖市局



场景D

数据分发服务

各级税务机关

场景B

数据共享服务

第三方部门

数据服务 | 数据文件

FTP | 开放数据库

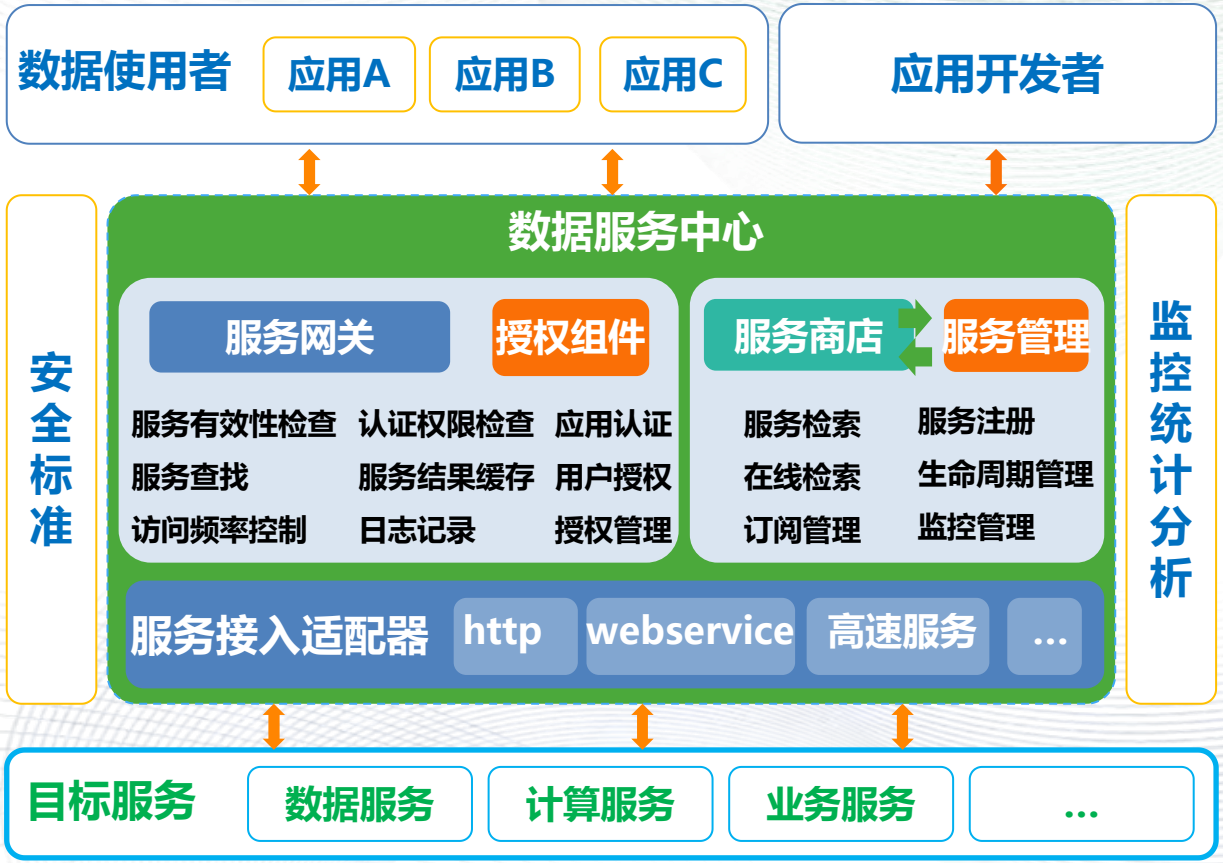
省市县审批

数据提供申请

场景C

数据提供服务

02 数据服务中心


数据使用者
应用A
应用B
应用C
应用开发者
数据服务中心
服务网关
授权组件
服务有效性检查
认证权限检查
应用认证
服务查找
服务结果缓存
用户授权
访问频率控制
日志记录
授权管理
服务商店
服务管理
服务检索
服务注册
在线检索
生命周期管理
订阅管理
监控管理
服务接入适配器
http
webservice
高速服务
...
安全标准
监控统计分析
目标服务
数据服务
计算服务
业务服务
...

02 数据服务中心


数据资源服务
数据计算服务
数据存储服务
数据分析服务
数据挖掘服务
数据应用服务
数据服务中心
全量数据资源
分布式计算存储资源

02 数据应用中心

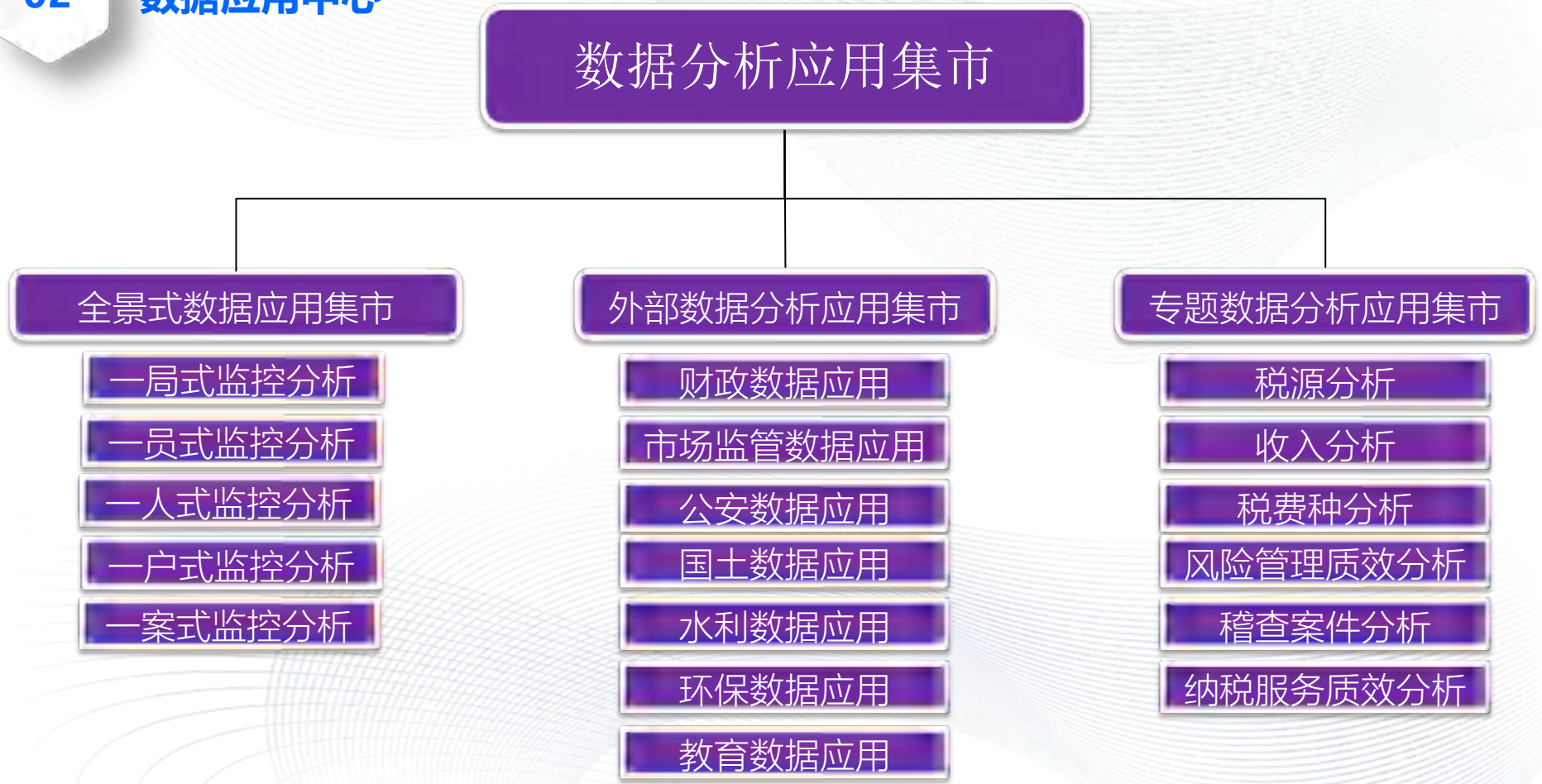
数据分析应用集市应涵盖通用查询、全景查询、历史数据查询和外部数据查询等功能，并按使用对象、管理对象等维度对集成的历史各阶段内外部、互联网数据进行归集、整理和分析，用于服务各级税务机关有效开展税收管理。

建设数据挖掘应用专题，深度挖掘数据潜在价值，辅助领导决策。通过提供数据分析和数据挖掘两方面应用，支持跨部门横向全局数据分析，也支持业务主题细分纵向深度挖掘，满足税收管理的需要。

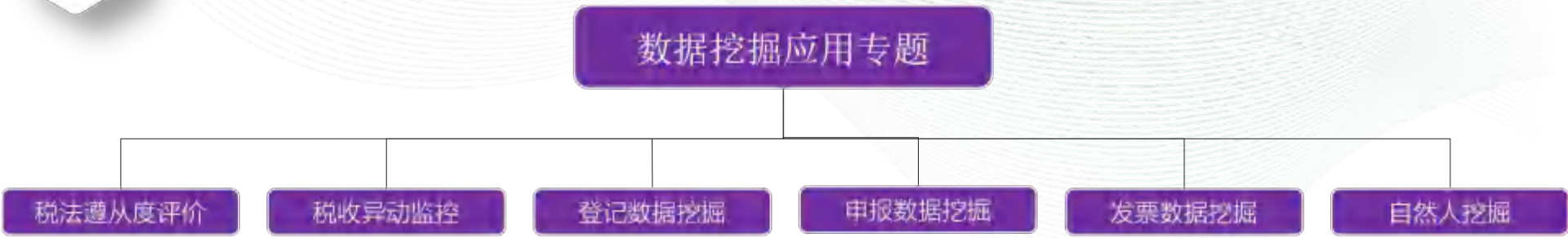
围绕税收管理工作热点、难点、痛点，构建智能化的信息处理模型、预测算法工具和搜索引擎，为税务人员开展个性化数据应用提供工作平台。

基于大数据仓库、数据服务中心以及大数据处理能力，快速开发和交付包括画像管理、关系图谱、票流分析、模型评估、数据探查、自主挖掘等在内的一大批创新应用，发挥涉税大数据价值和效益，各级税务人员个性化创新应用需求提供数据挖掘和分析工具。

02 数据应用中心


数据分析应用集市
全景式数据应用集市
一局式监控分析
一员式监控分析
一人式监控分析
一户式监控分析
一案式监控分析
外部数据分析应用集市
财政数据应用
市场监管数据应用
公安数据应用
国土数据应用
水利数据应用
环保数据应用
教育数据应用
专题数据分析应用集市
税源分析
收入分析
税费种分析
风险管理质效分析
稽查案件分析
纳税服务质效分析

02 数据应用中心



指标化纳税人涉税行为，深度挖掘纳税人行为数据，为纳税人税法遵从度评价提供参考。

从多个视角对税收异动变化情况进行全面监控，包括：税收总体走势异动监控、税收大幅下降企业数量趋势分析等内容。

深度挖掘纳税人登记信息，发现登记异常企业，包括：关键人员交叉关系、多级控股、注册地址异常、登记行为异常等。

深度挖掘各税（费）种申报数据、财务报表、第三方采集数据、互联网数据等相关数据，通过云计算和大数据智能技术发现申报逻辑异常企业。

深度挖掘发票开具、取得数据，发现发票使用异常企业，包括：进销项异常、循环开票、滞留票异常、库存商品销售异常、连续顶额开票且金额突增等。

聚合工资薪金、个体工商户、土地房产、车船信息、房产租赁所得、劳务报酬、股权红利、产权所得、偶然所得等相关数据，识别隐藏在数据中的高收入高净值的双高人群。

02 数据应用中心

智能应用空间

智能推送

根据用户工作岗位、订阅情况、使用习惯、关注焦点等信息，智能推送用户需要的或可能关心的内容，并依据使用情况自动调整。

智能预测

构建相关性分析、回归分析、时间序列分析等算法，建立科学的预测模型工具，为用户进行数据的趋势分析和预测提供辅助手段。

智能搜索

构建智能化搜索引擎，实现模糊匹配、智能纠错、语义搜索等智能化搜索方式。

02 数据应用中心

画像管理

自然人单户画像、自然人群体画像、风险特征类画像标签

关系图谱

利用大数据智能算法和数据分析能力，充分挖掘纳税人登记信息、业务交易信息、投资信息、控制关系，以图形化的方式直观展示。

票流分析

是基于企业间开具的增值税发票，分析企业间购销的上下游关系，进行图形化展示。

模型评估

新增模型评估功能模块，通过对标注好的数据，进行学习、训练得到一套能够进行分类的算法。构建模型工厂，模拟运行指标模型，进行可视化展示，辅助模型迭代优化，提升模型的精准度。

自助挖掘

构建自定义、可视化的自助分析工具，采用拖拽、关联等方式创建个性化的分析应用，加强全链条自助分析功能。

数据探查

新增数据探查模块，提供交互式的统计分析工具，方便快速的计算和展示数据各种维度的统计值，从而全方位的洞察数据的质量以及统计分布特性。

自定义菜单

根据不同用户的身份和使用需求，对经常使用、关注、关联的相关功能模块进行汇集，提供用户自定义菜单。