CCTC 2017 中国云计算技术大会
Cloud Computing Technology Conference 2017
CSDN

INFINIDATA:基于Spark的统一数据管理与探索平台

熊永平

北京邮电大学网络技术研究院

提纲

- 设计背景
- INFINIDATA大数据平台
- INFINIDATA可视化探索
- 应用实践

数据应用的5个发展阶段

数据应用5阶段演进模型

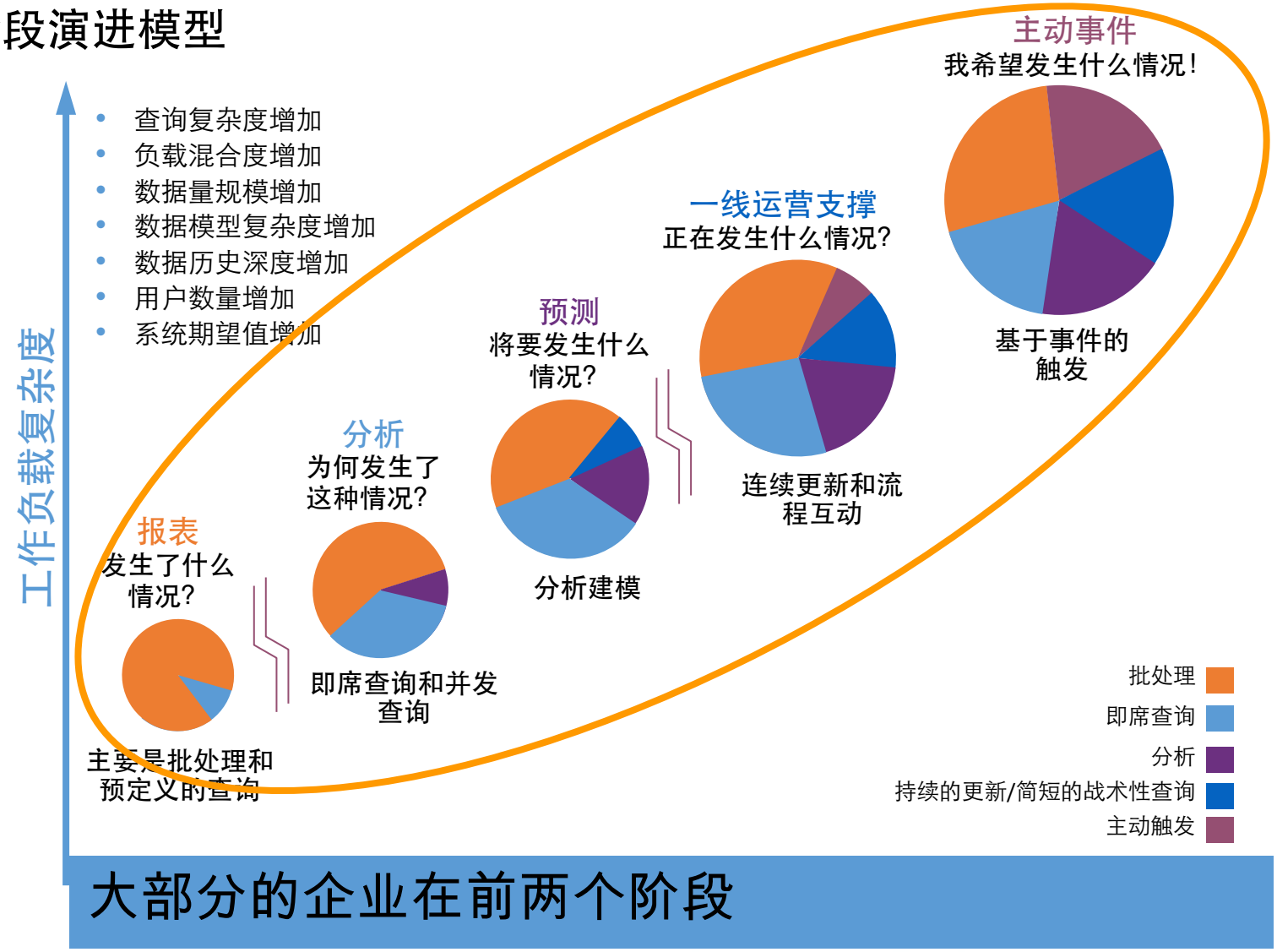
第五阶段

第四阶段

第三阶段

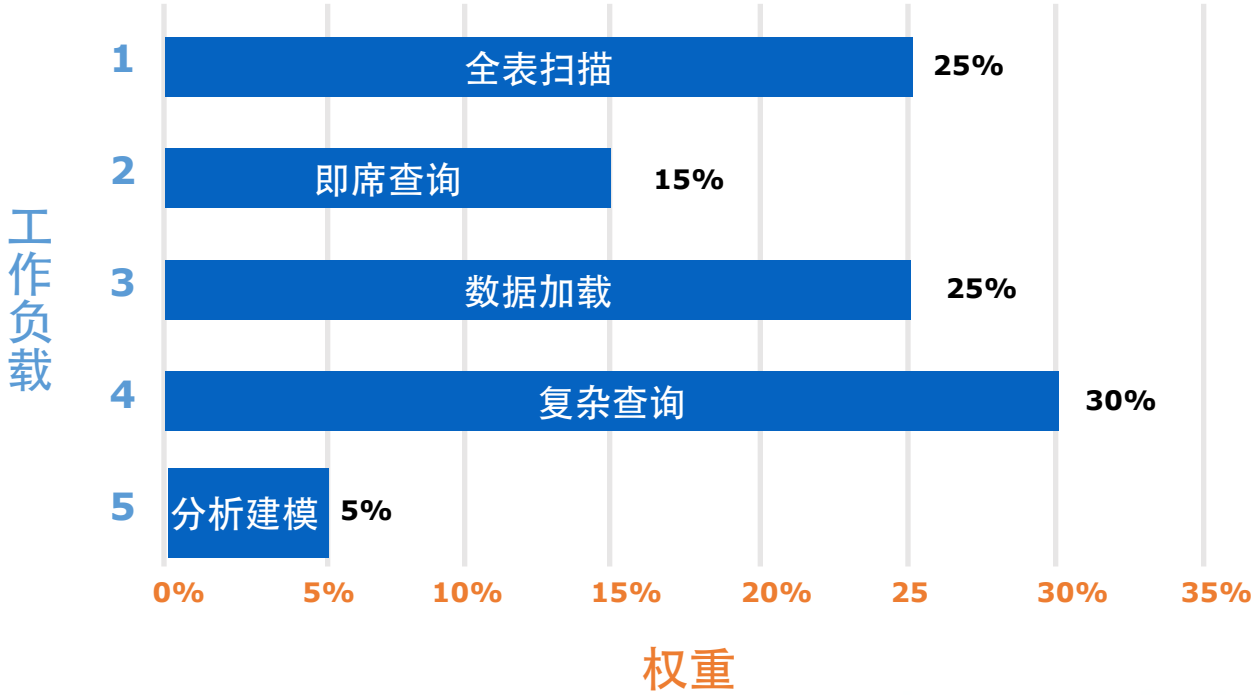
第二阶段

第一阶段



大部分的企业在前两个阶段

- 典型负载
  - 即席查询SQL：报表、简单查询、汇总
  - 复杂检索：多字段检索、模糊检索、全文检索
  - 全表扫描：离线DAG计算任务、ETL处理流程、预测等
  - 交互式探索：自助交互式建模



现有解决方案

MPP数据仓库

- TeraData、Greenplum、SAP HANA等
- BI生态成熟、非SQL任务很难支持
- 巨贵

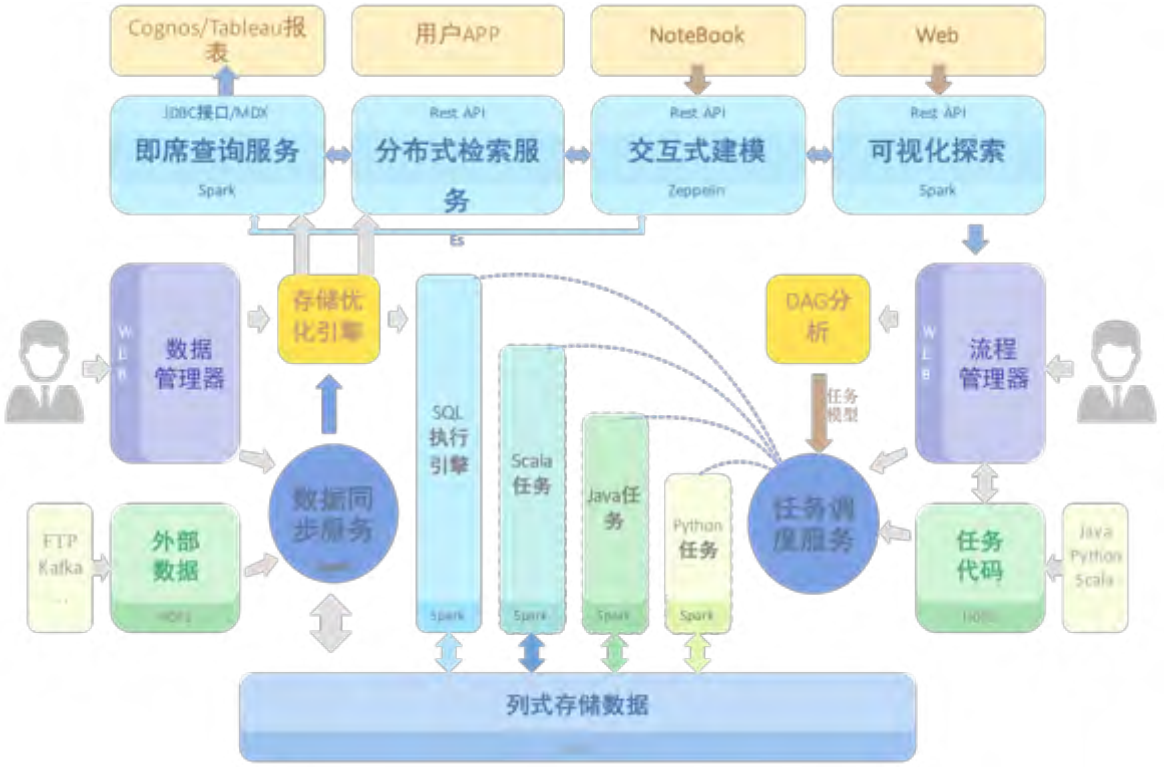
大数据平台

- HDP、CDH、星环等
- 技术先进、开源开放、坑多
- 暴露底层组件太多，运维和使用技术曲线陡峭

用户对大数据平台的期望

- 最好看起来像数据库一样，管理方便，使用简单
- 利用最新的大数据计算技术获得高性能和扩展能力
- 不需要掌握各种底层组件
- 兼容运行已有的数据库存储过程
- 统一管理各种数据处理任务
- 稳定可靠

2、INFINIDATA平台


Cognos/Tableau报表
用户APP
NoteBook
Web
JDBC接口/MDX
即席查询服务
Spark
Rest API
分布式检索服务
Rest API
交互式建模
Zeppelin
Rest API
可视化探索
Spark
Es
数据管理器
存储优化引擎
SQL执行引擎
Scala任务
Java任务
Python任务
DAG分析
任务模型
流程管理器
FTP
Kafka
外部数据
数据同步服务
任务调度服务
任务代码
Java
Python
Scala
列式存储数据

一体化的数据管理

- 特性
  - 全量数据和表结构Schema自动导入
  - 增删改等增量数据的智能同步
  - 同时支持原始表和衍生表
  - 支持对表数据和表结构Schema的变化轨迹溯源
  - 自主选择存储引擎和分区分桶优化
  - 数据变化自动触发相关的计算任务

易用强大的计算流程管理

- 特性
  - 借鉴关系代数思想，计算流程等价于表的函数变换
  - 计算流程统一管理，计算任务历史可追踪
  - 计算逻辑和中间结果可共享
  - 自动分析计算任务的依赖关系图进行全局调度优化
  - 支持PLSQL存储过程和非SQL（Scala或Python）的复杂计算任务

标准高效的数据查询服务

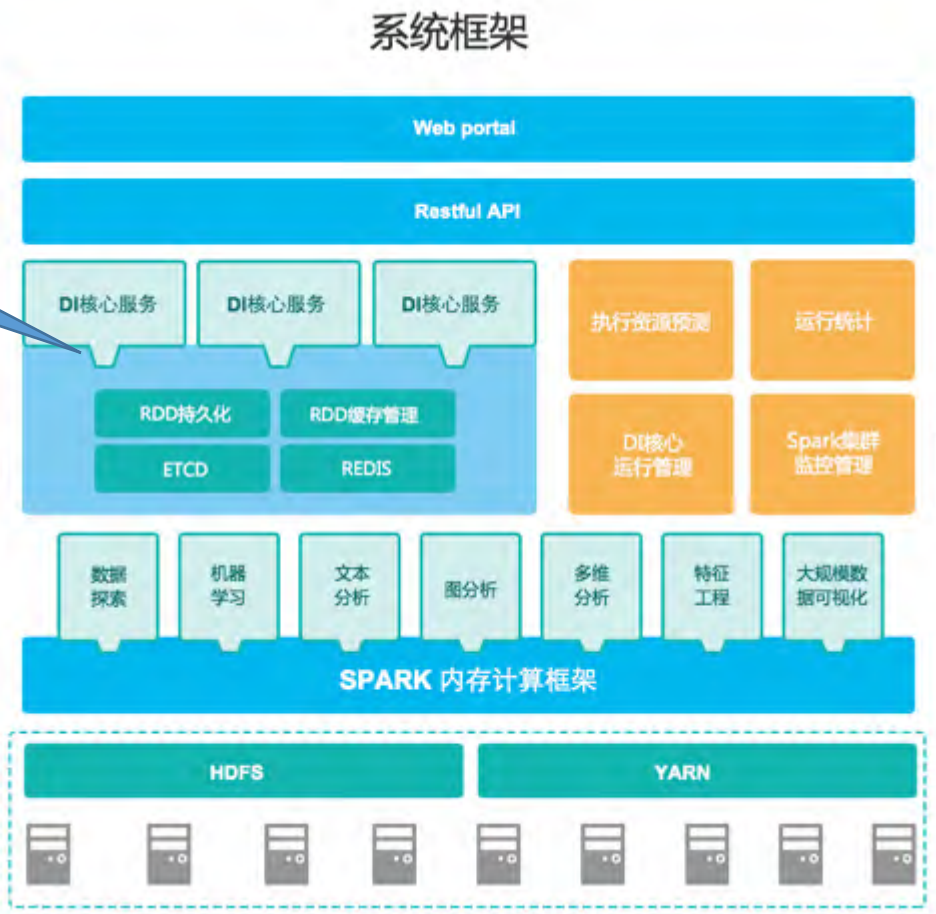
- 特性
  - 提供标准的JDBC访问接口，对常用的Cognos等报表服务提供driver
  - 提供MDX语言的建模和OLAP分析引擎服务
  - 提供标准Es接口提供数据检索服务
  - 动态分层功能，将访问最频繁的数据保存在内存中，同时将很少访问的数据移至磁盘
- 开源组件优化集成
  - 修改hive2.3、spark2.1等相关组件的bug和源代码约80处

3、INFINIDATA可视化探索

可视化探索DI

常驻内存服务

- 每个工程运行在一个单独的Spark环境
- Spark环境资源由YARN分配调度
- DI和Spark常驻内存，通过消息队列交互
- 利用RDD保存探索过程中的各种中间表



可视化探索功能

相关分析

- 行分析（聚类）
- 列分析（变量聚类）
- 值分析（频繁项）

离散矩阵分析

- 自相关分析
- 互相关分析

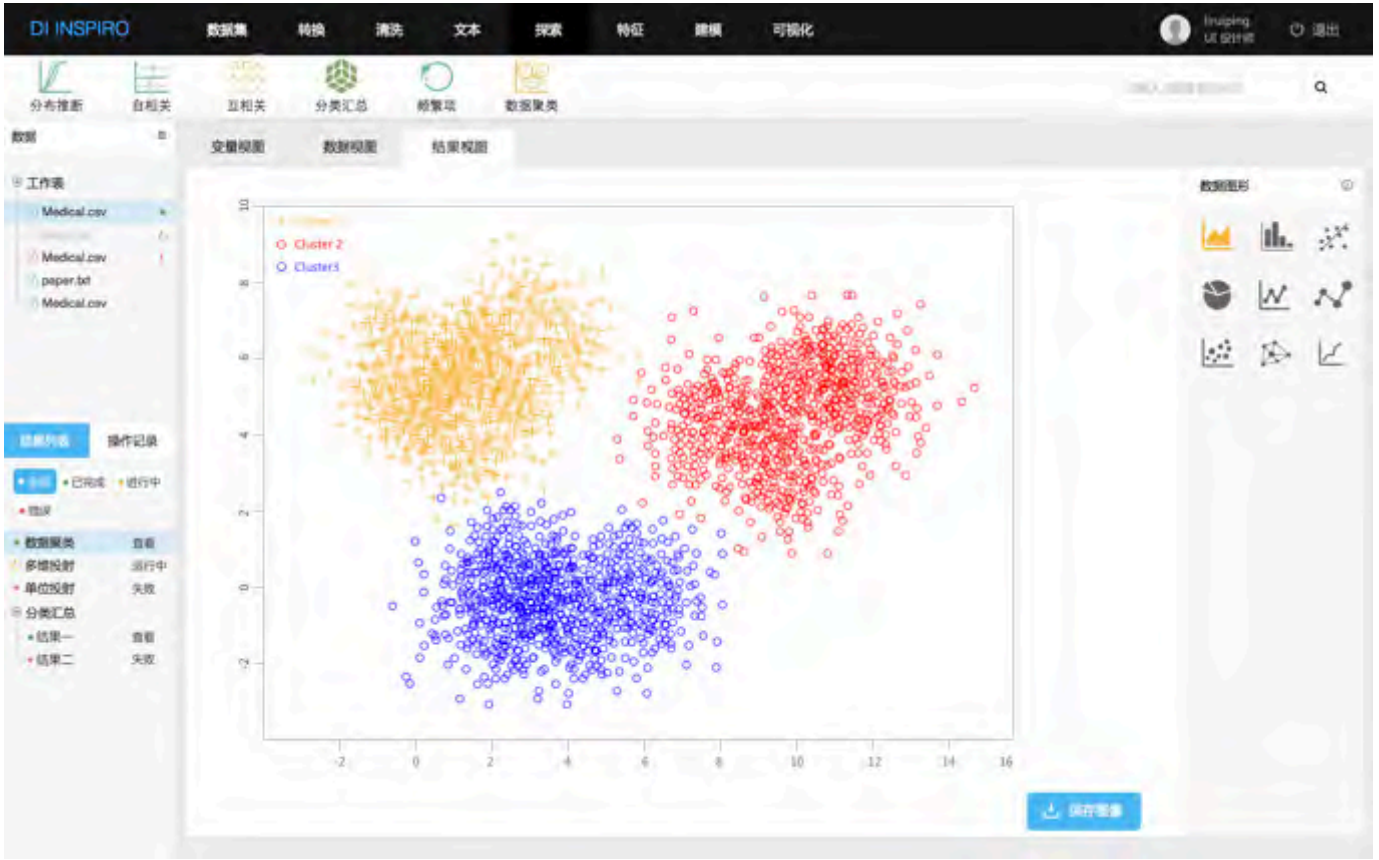
OLAP分析

- Mondrian建模
- 多维度分析

可视化分析

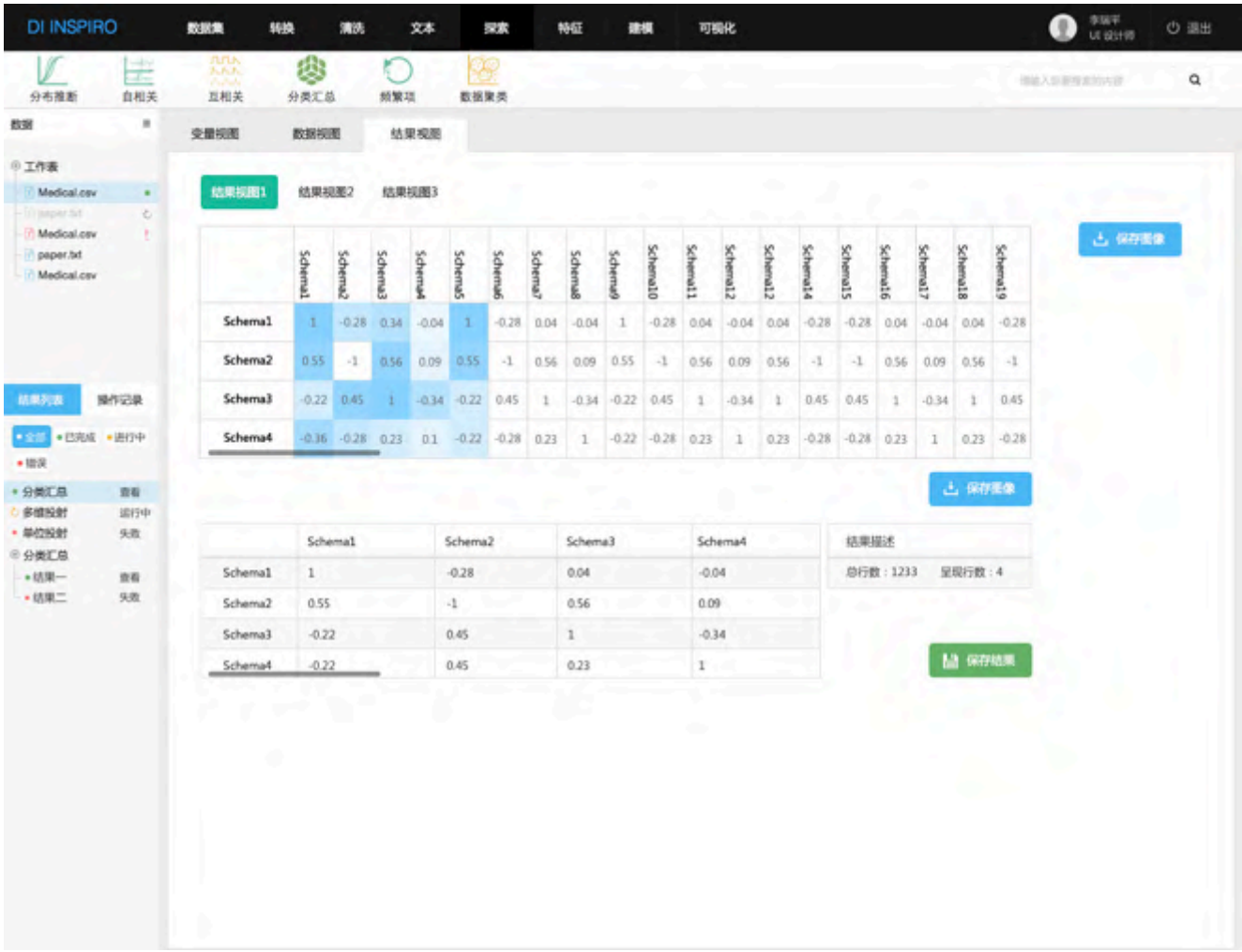
- 散点图
- 直方图
- 箱图
- 3维散点图

K-means聚类分析


DI INSPIRO
数据集
转换
清洗
文本
探索
特征
建模
可视化
分布推断
自相关
互相关
分类汇总
频繁项
数据聚类
变量视图
数据视图
结果视图
工作表
Medical.csv
Medical.csv
paper.txt
Medical.csv
操作记录
Cluster 2
Cluster1

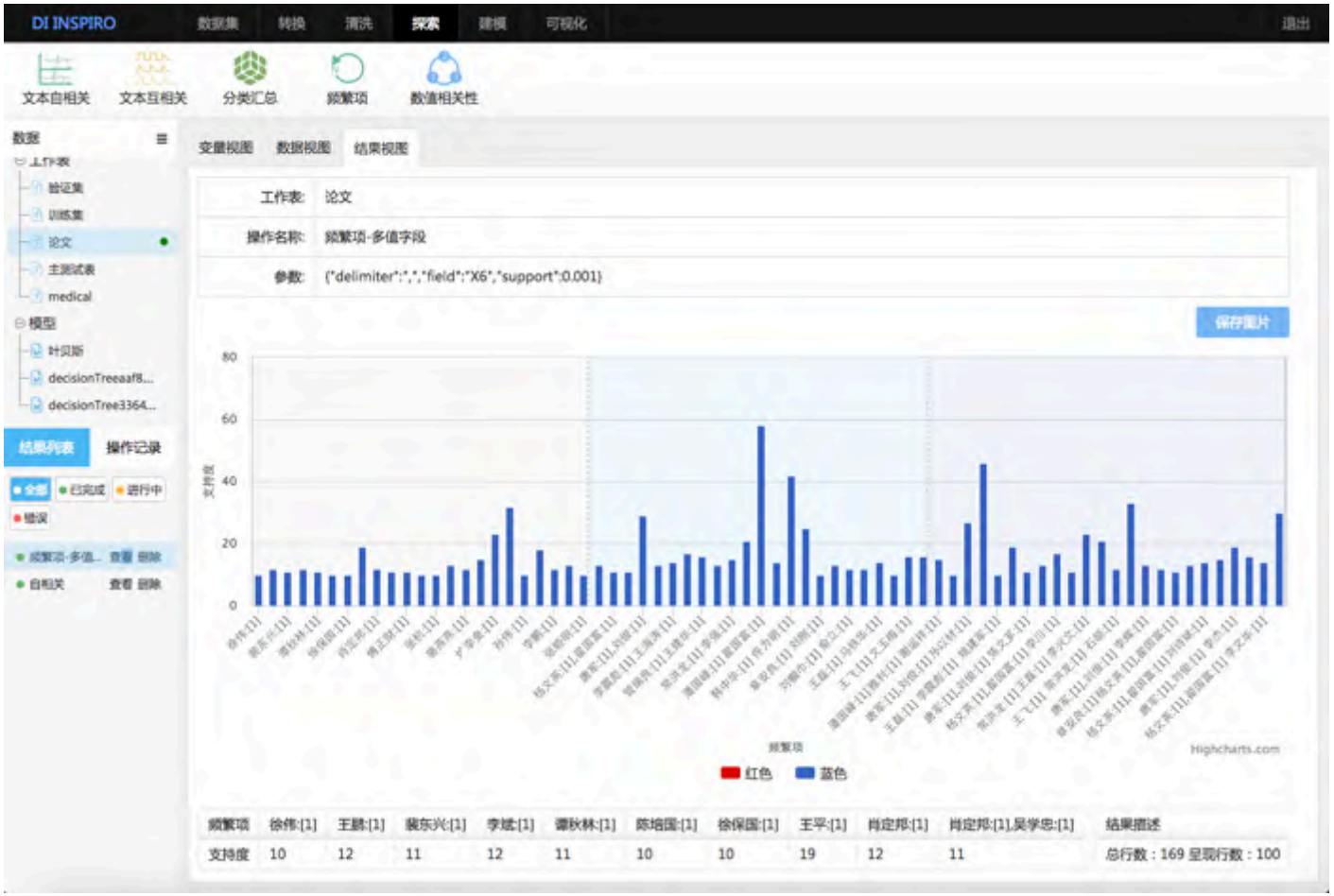
列相关-pearson相关性

- 热力图展现列之间相关性
- 发现基础变量和衍生变量



频繁项分析

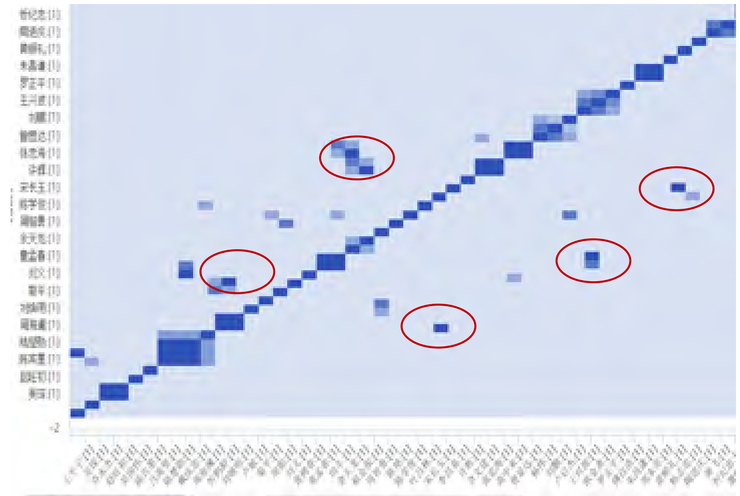
属性值关联性



离散矩阵分析

分析模型

- 图分析
  - 社会网络分析
  - 药品关联分析
  - 公共安全
    - （相同时间/机场）乘坐相同航班的同乘分析
    - （相同时间/地点）的紧密通话客户分析
  - 科技领域
    - 研发相类似技术领域的竞争对手分析
    - 论文合作关系



publication_time	title	abstract	author
2000_1期	[illegible]	-	-
2000_1期	[illegible]	[illegible]	[illegible]
2000_1期	[illegible]	[illegible]	[illegible]
2000_1期	[illegible]	[illegible]	[illegible]
2000_1期	[illegible]	[illegible]	[illegible]
2000_1期	[illegible]	[illegible]	[illegible]
2000_1期	[illegible]	[illegible]	[illegible]
2000_1期	[illegible]	[illegible]	[illegible]
2000_1期	[illegible]	[illegible]	[illegible]

自相关矩阵分析

原始数据表

	C1	C2	C3	C4	C5
R1	X_1	Y_1	A_1,A_2,A_3	B_1,B_2,B_3	Z_1
R2	X_2	Y_2	A_2,A_3	B_2,B_4	Z_2
R3	X_3	Y_3	A_1,A_4,A_5	B_2,B_3,B_6	Z_3
R4	X_4	Y_4	A_2,A_5	B_1,B_4	Z_4
R5	X_5	Y_5	A_3,A_4	B_1,B_5	Z_5

投影

矩阵变换

	R_1	R_2	R_3	R_4	R_5
A_1	1	0	1	0	0
A_2	1	1	0	1	0
A_3	1	1	0	0	1
A_4	0	0	1	0	1
A_5	0	0	1	1	0

共现相关性

	A_1	A_2	A_3	A_4	A_5
A_1		1	1	1	1
A_2	1		2	0	1
A_3	1	2		1	0
A_4	1	0	1		1
A_5	1	1	0	1	

COS相关性

	A_1	A_2	A_3	A_4	A_5
A_1	1	0.408	0.408	0.5	0.5
A_2	0.408	1	0.667	0	0.408
A_3	0.408	0.667	1	0.408	0
A_4	0.5	0	0.408	1	0.5
A_5	0.5	0.408	0	0.5	1

pearson相关性

	A_1	A_2	A_3	A_4	A_5
A_1	1	-0.167	-0.167	0.167	0.167
A_2	-0.167	1	0.167	-1	-0.167
A_3	-0.167	0.167	1	-0.167	-1
A_4	0.167	-1	-0.167	1	0.167
A_5	0.167	-0.167	-1	0.167	1

保险欺诈合谋分析

理赔数据表

案件编号	人员	车牌	地点	金额
344561	段建华，张华，许卫	湘A2BA32，湘AA1391，湘ZG00069	板仓南路	20000
344562	罗坚，肖蓉	湘J7ZH83，湘AL5S85	开元西路	50000
344563	王丽萍，刘双泉	湘A65N90，湘A1661K	寿昌路	100000
344564	彭发兵，周辉，苏英雄	湘A2ZB92，湘B2HL12，湘A2KA19	人民路	70000
344565	张斌，王丽萍，	湘AT8137，湘A65N90	湘江东路	10000

矩阵变换

	344567	344568	344569	344570	344571
谢前	0	1	1	0	1
敬春桥	0	0	0	1	0
罗坚	0	1	0	0	1
肖蓉	0	0	1	0	0
刘双泉	0	1	0	0	0

	谢前	敬春桥	罗坚	肖蓉	刘双泉
谢前	1	-0.612	0.667	0.408	0.408
敬春桥	-0.612	1	-0.408	-0.25	-0.25
罗坚	0.667	-0.408	1	-0.408	0.612
肖蓉	0.408	-0.25	-0.408	1	-0.25
刘双泉	0.408	-0.25	-1	-0.25	1

pearson相关性

	谢前	敬春桥	罗坚	肖蓉	刘双泉
谢前		0	2	1	1
敬春桥	0		0	0	0
罗坚	2	0		0	1
肖蓉	1	0	0		0
刘双泉	1	0	1	0	

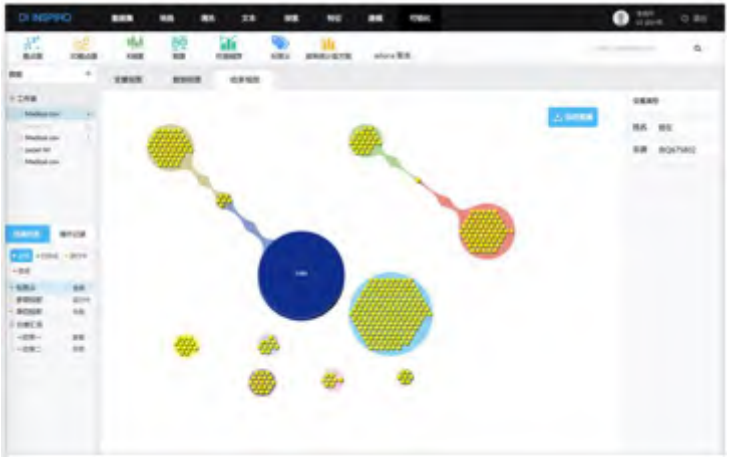
共现相关性

cos相关性

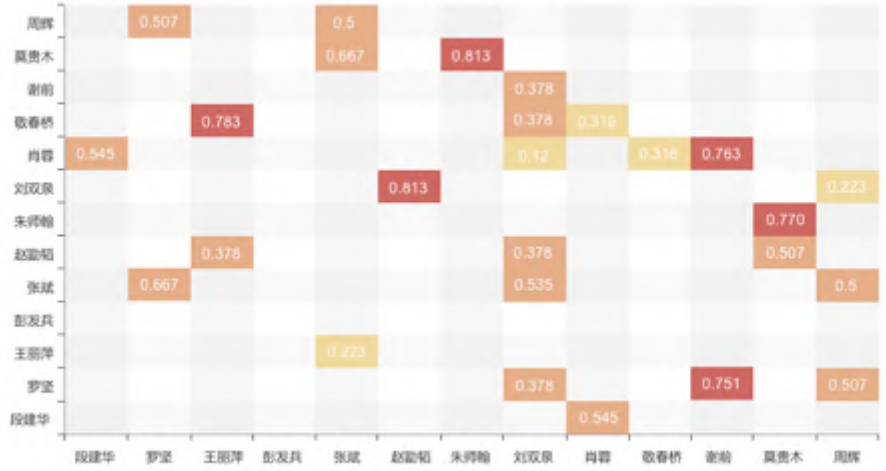
	谢前	敬春桥	罗坚	肖蓉	刘双泉
谢前	1	0	0.816	0.577	0.577
敬春桥	0	1	0	0	0
罗坚	0.816	0	1	0	0.707
肖蓉	0.577	0	0	1	0
刘双泉	0.577	0	0.707	0	1

自相关分析可视化

相关矩阵分析map



Aduna图合作关系可视化



互相关矩阵分析

原始数据表

	C1	C2	C3	C4	C5
R1	X_1	Y_1	A_1,A_2,A_3	B_1,B_2,B_3	Z_1
R2	X_2	Y_2	A_2,A_3	B_2,B_4	Z_2
R3	X_3	Y_3	A_1,A_4,A_5	B_2,B_3,B_6	Z_3
R4	X_4	Y_4	A_2,A_5	B_1,B_4	Z_4
R5	X_5	Y_5	A_3,A_4	B_1,B_5	Z_5

投影

矩阵变换

	B_1	B_2	B_3	B_4	B_5
A_1	1	2	2	0	0
A_2	1	2	1	2	0
A_3	1	2	1	1	1
A_4	1	1	1	0	1
A_5	1	1	1	1	0

共现互相关性

	A_1	A_2	A_3	A_4	A_5
A_1		3	3	3	3
A_2	3		5	0	0
A_3	3	5		2	0
A_4	3	0	2		3
A_5	3	0	0	3	

cos互相关性

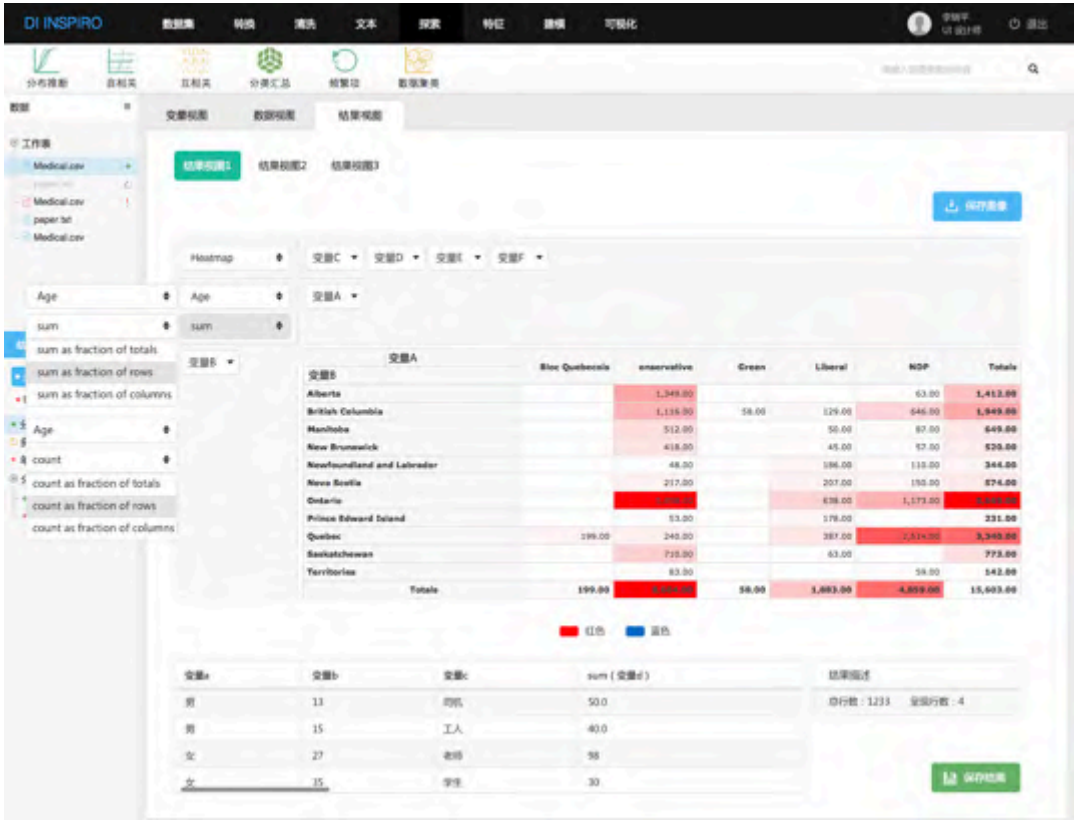
	A_1	A_2	A_3	A_4	A_5
A_1	1	0.7	0.783	0.849	0.849
A_2	0.7	1	0.894	0.566	0.849
A_3	0.783	0.894	1	0.791	0.791
A_4	0.849	0.566	0.791	1	0.8
A_5	0.849	0.849	0.791	0.8	1

pearson互相关性

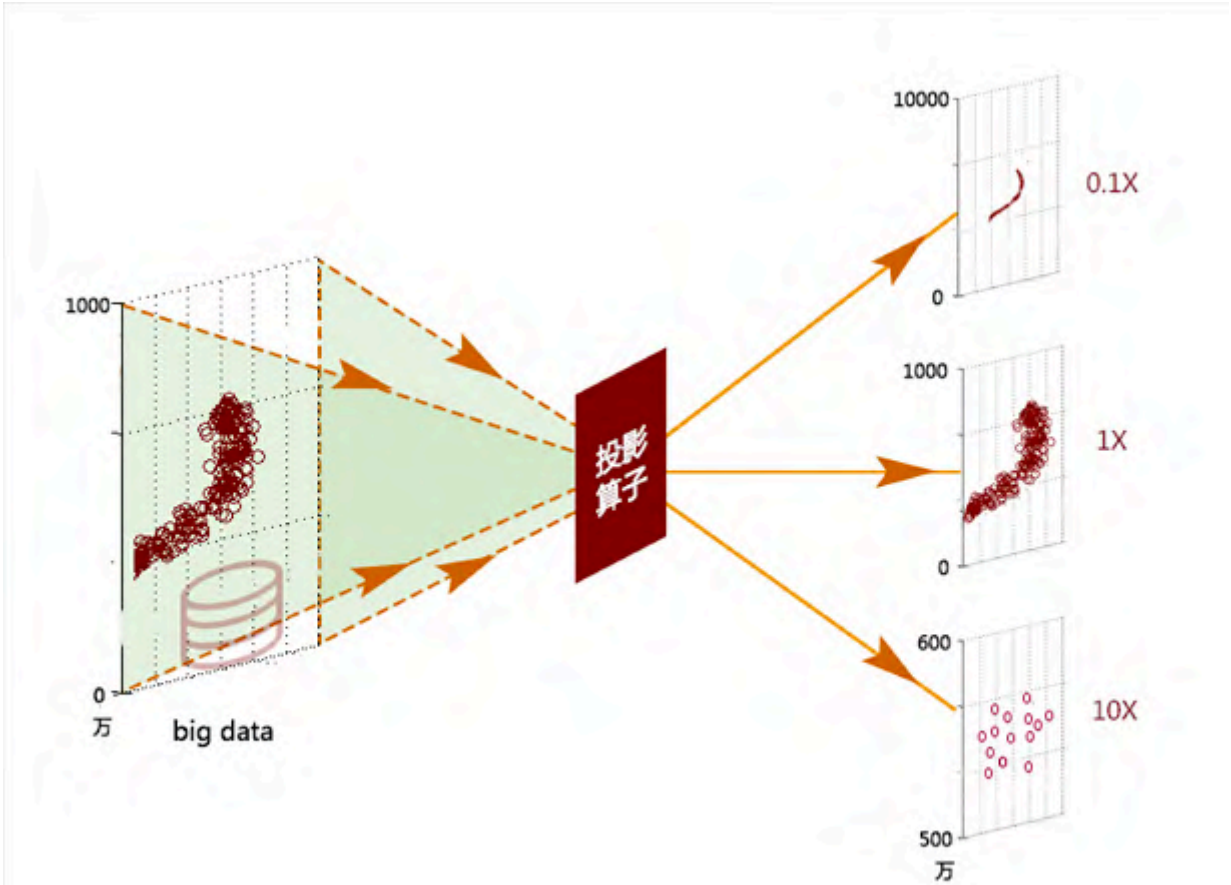
	A_1	A_2	A_3	A_4	A_5
A_1	1	0.25	0.354	0.548	0.548
A_2	0.25	1	0.707	-0.548	0.548
A_3	0.354	0.707	1	0	0
A_4	0.548	-0.548	0	1	-0.2
A_5	0.548	0.548	0	-0.2	1

多维度汇总分析

- 基于MDX的数据模型
  - Mondrian引擎

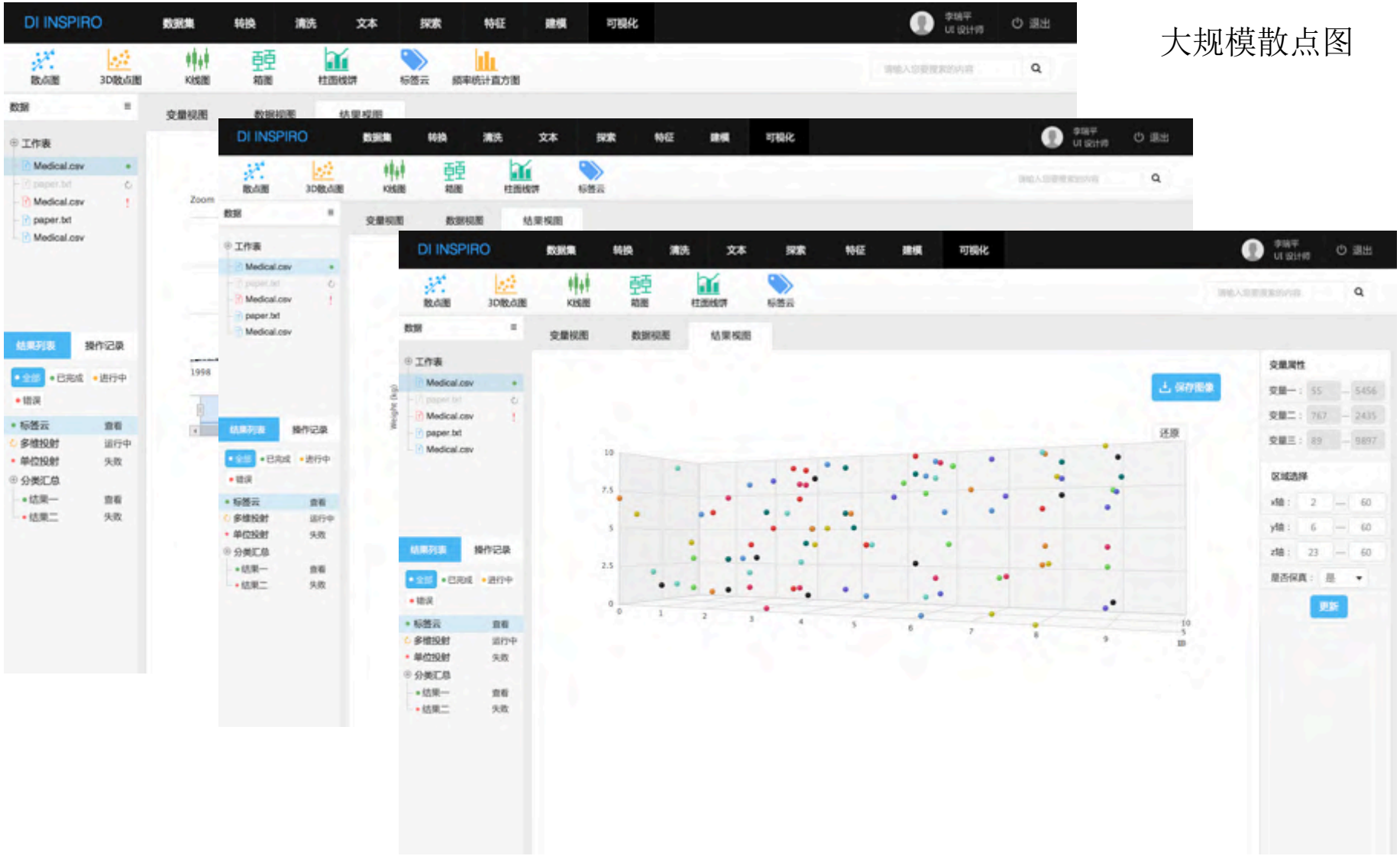


压缩投影的大数据可视化



- 基于保持数据概率分布不变的思想
- 将原始数据根据缩放级别和距离远近将原始数据映射成特定显示区域的矩阵

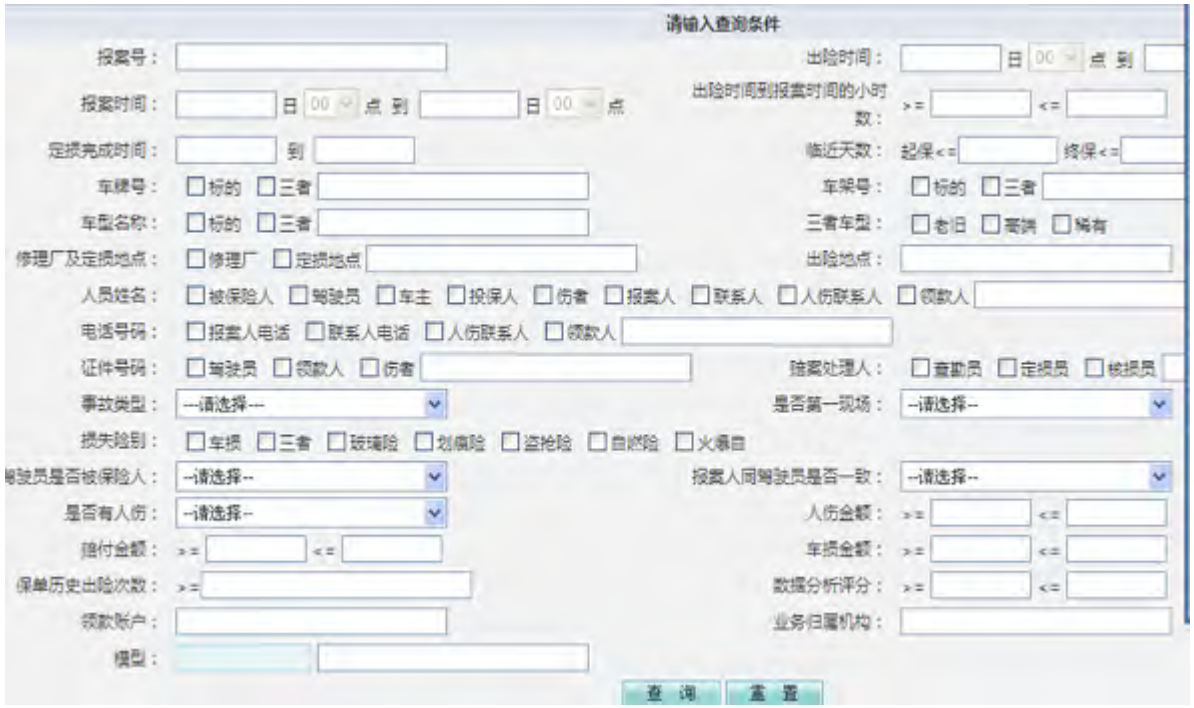
数据可视化



大规模散点图

4、应用实践

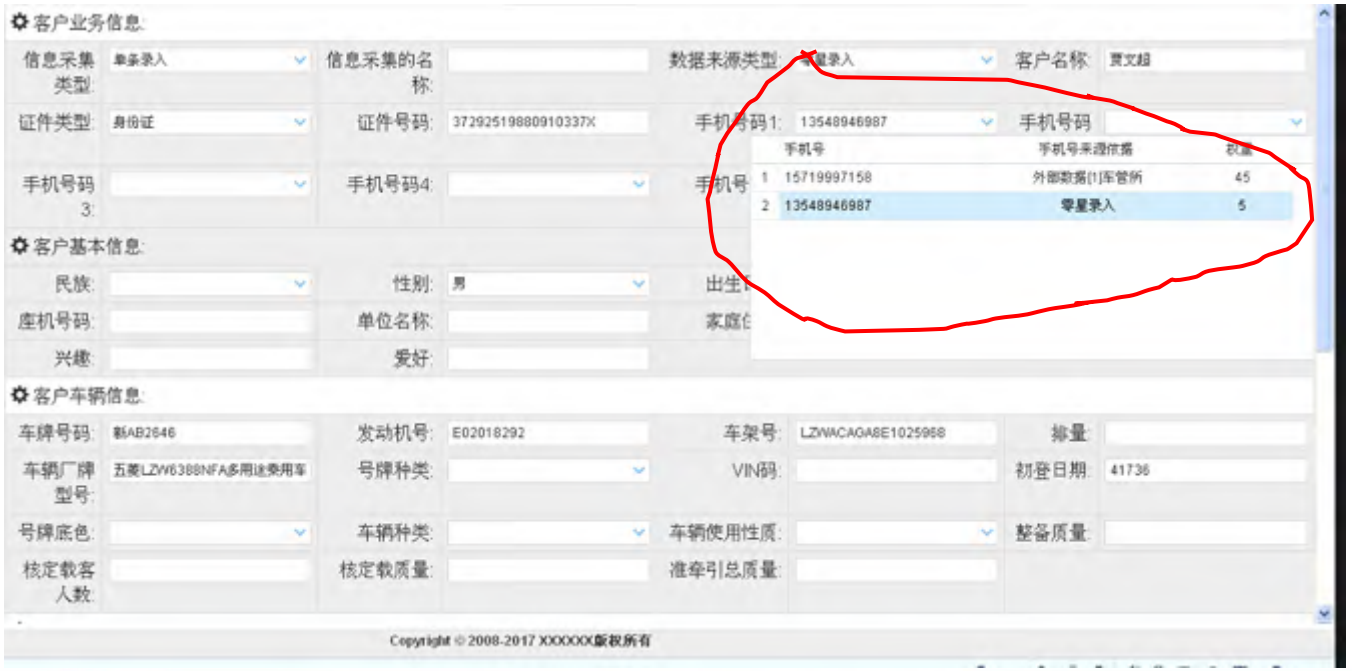
- 应用需求
  - 每天增量从22张原始业务表生成3张宽表
  - 原始表数据量7000万条，宽表2亿条记录
  - 对3张宽表近200个字段进行模糊组合检索



- 需求
  - 100多个维度、60多个指标，
  - 单表储数据量大小为6-30G
  - 修改Hive JDBC支持Cognos报表
  - DAG任务流程运行生成事实表和维度表

	2016年	2017年	年份
本年终保数量	25863	1032	26895
本年未结赔付件数	25285	1328	26613
本年未结赔款	1352067273.070000	24478550.950000	1376545824.020000
本年未核赔付件数	24932	1289	26221
本年已核赔付件数	186057	2939	188996
本年未核赔款	.000000	.000000	.000000
本年已核赔款	595834179.240000	16753960.080000	612588139.320000
本年非零赔款已核赔付件数	5547	430	5977
本年非零赔款已结赔付件数	151943	2702	154645
本年新车购置价	200439283754.810060	26548525007.000008	226987808761.810090
本年简单赔付率	78569	2706	81275
本年有效报案数量	221602	4556	226158
本年已释签单保费	24492138.164092	465256.912296	24957395.076388
本年有效报案数量/本年已释签单保费	873	351	1224
本年标准保费	6991574091.370002	891420936.090000	7882995027.460002
本年商业险标准保费	6991552660.020002	891420936.090000	7882973596.110002
本年NCD调整保费	628185007560.100100	63019723564.250015	691204731124.350100
本年NCD核保调整保费	604700562090.704470	60221147865.475906	664921709956.180300
本年NCD核保渠道调整保费	603032358009.099980	60042105420.350014	663074463429.450200

- 需求
  - 保险客户信息真实性较低，无法服务于精准化营销和客服资源的精准化投放
  - 每天导入来自车管所、电信公司、俱乐部、分布在承保理赔各环节的碎片化信息
  - DAG流程处理各来源数据并进行交叉核验，生成用户画像



子曰：人而无信，不知其可也
----《论语·为政》

如果没有加这个微信，不知道还可不可以做大数据?

谢谢！

欢迎交流！