



MongoDB在大规模数据获取和准实时分析中的应用



Proudly Powered by MongoDB

张锦杰
深圳市一面网络技术有限公司

电商
社交网络
新闻媒体
企业内部数据
数据采集

01

数据处理

数据清洗
自动化ETL

02

自然语言处理
图像识别

机器学习

03

分析建模

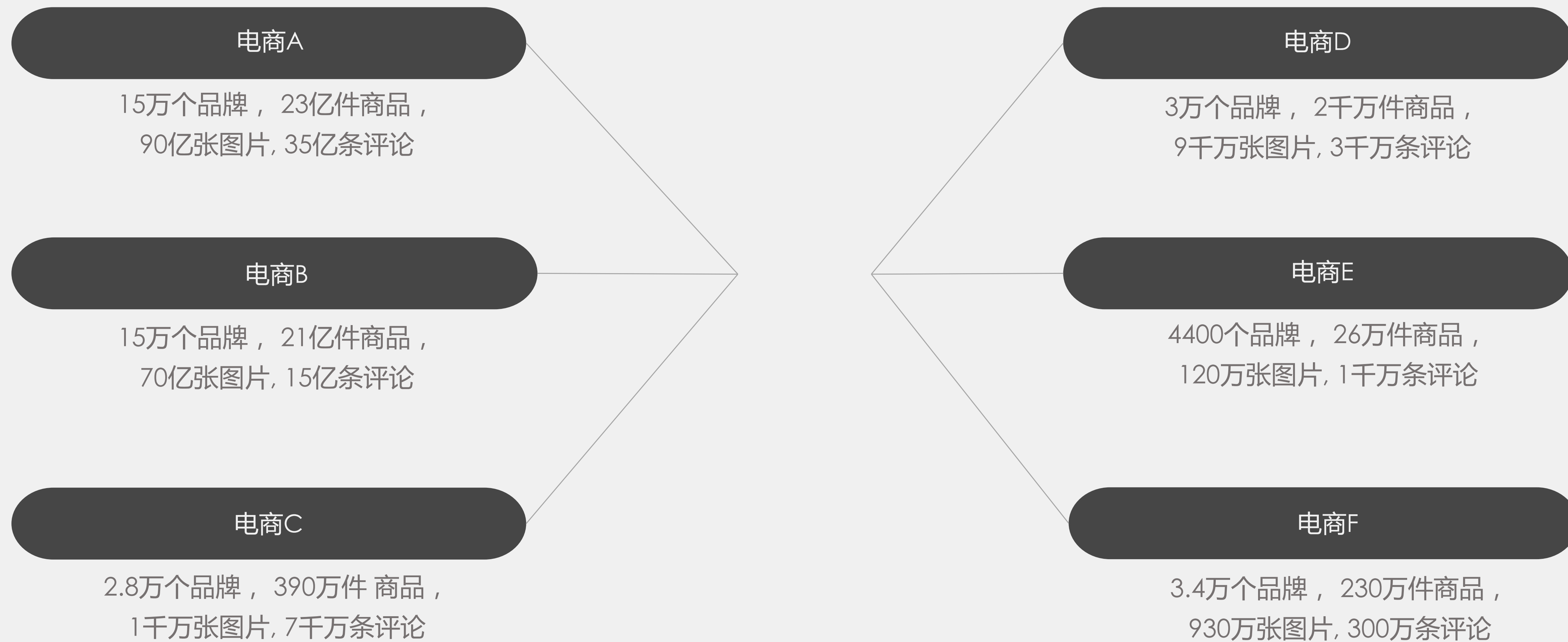
关联规则
聚类、分类
预测分析

04

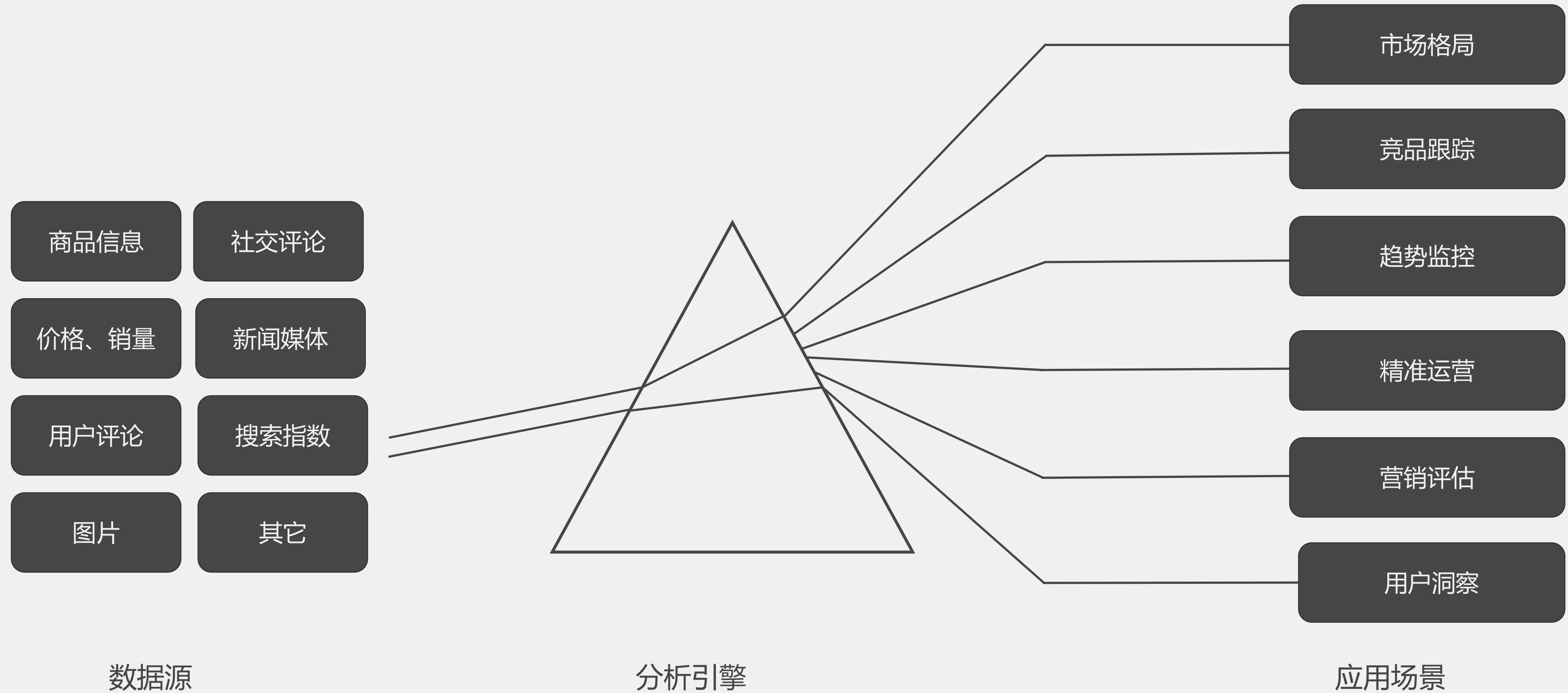
可视化
前端自助分析

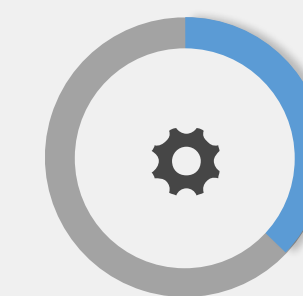
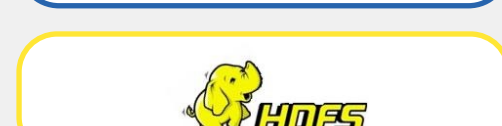
智能

05



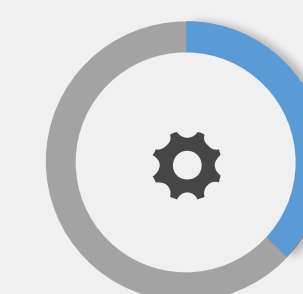
棱镜





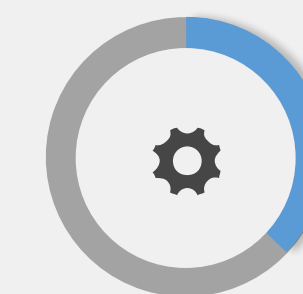
数据应用

基于WEB的交互分析
跨平台、响应式布局
事件通知及预警
分析建模



数据处理

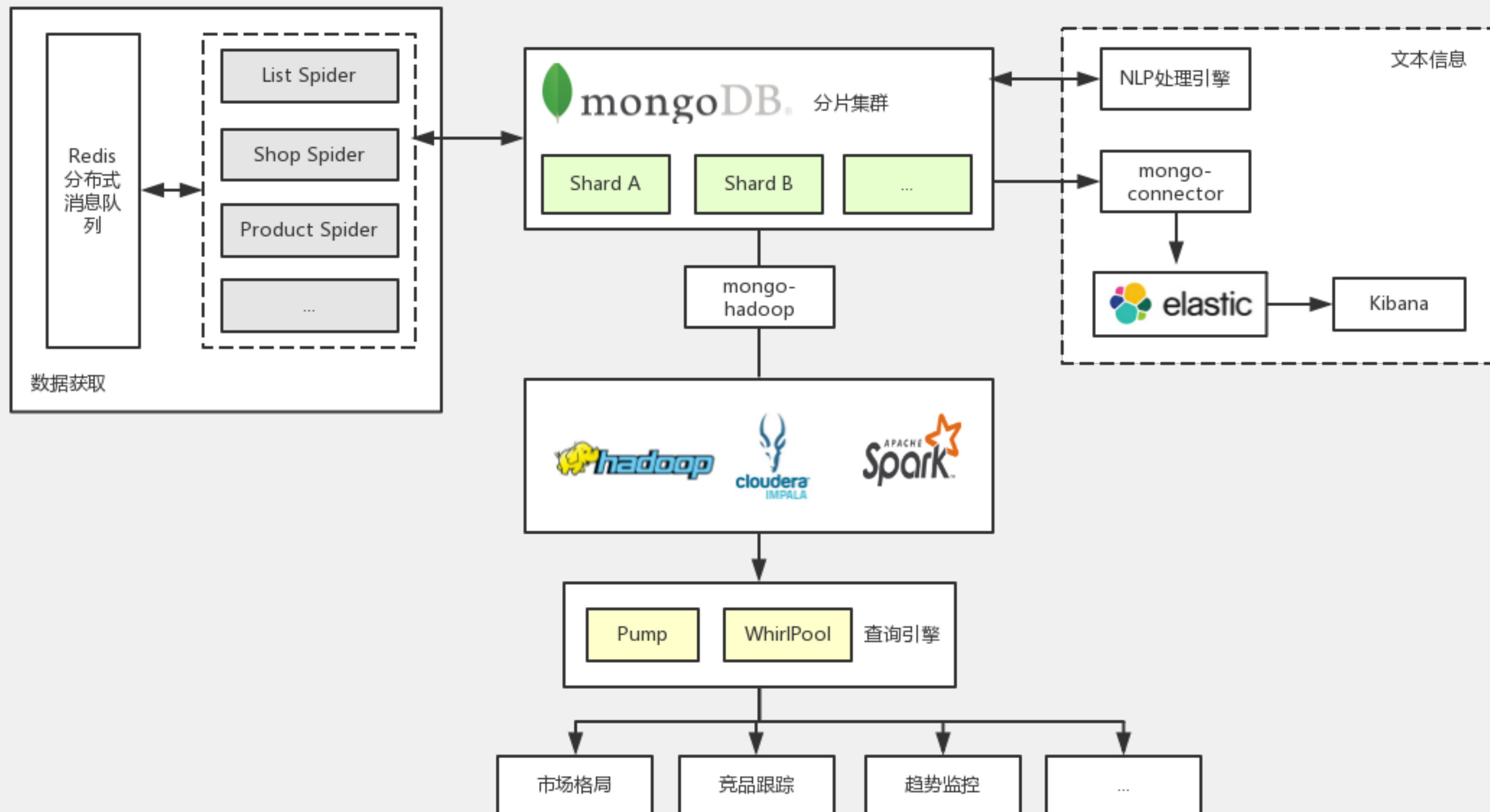
准实时查询
数据清洗、自动化ETL
机器学习



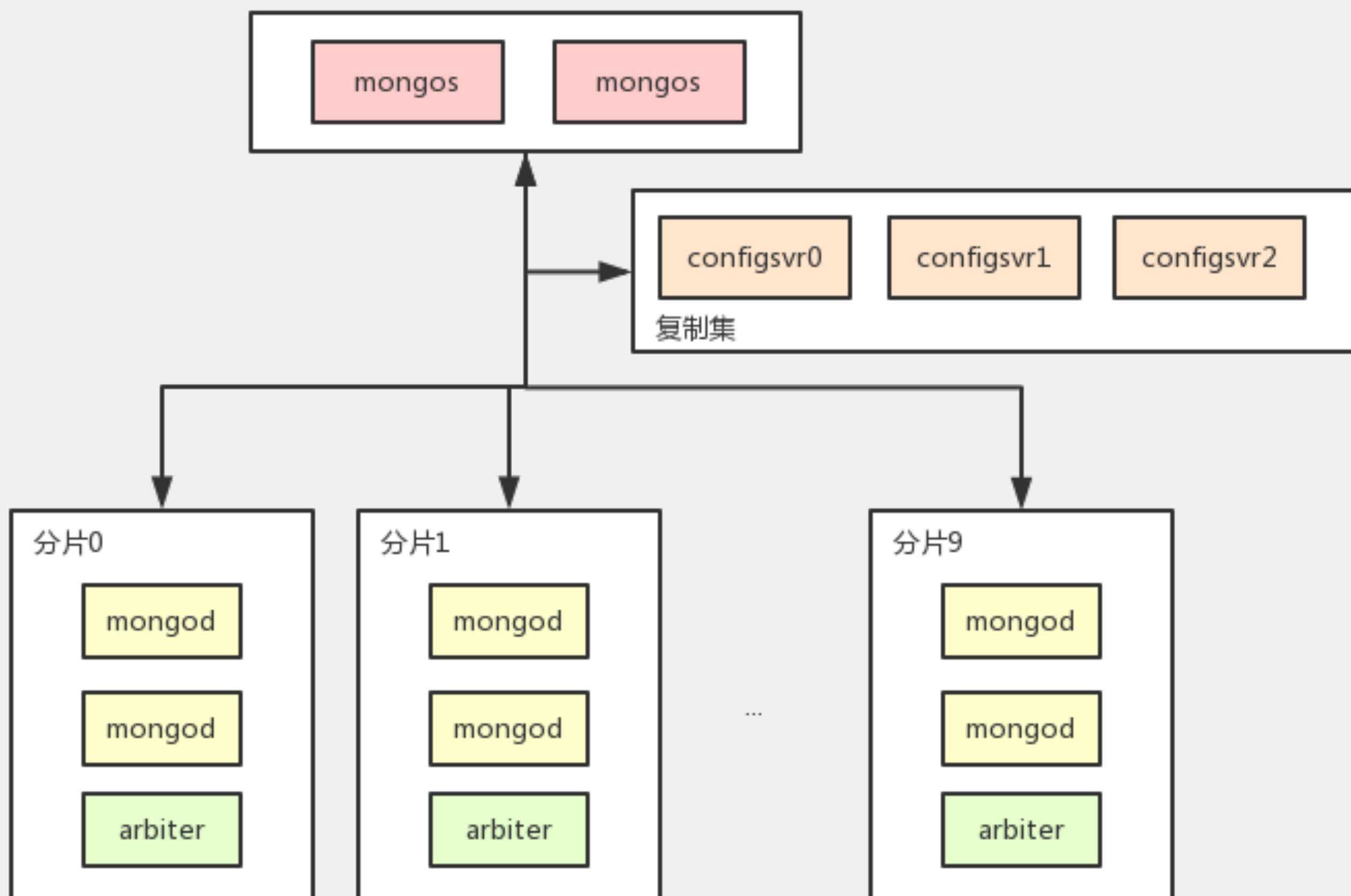
数据存储

分布式
大规模数据存储

大规模数据获取（电商数据）



MongoDB分片集群



分片集群特点

- 10个分片，每分片2个数据节点、1个arbiter
- 交叉分布，分布在10台物理节点
- 2个mongos节点冗余备份
- 3个configsvr0形成复制集，存储集群元数据

MongoDB分片集群

Shard Key选择：

- 特异性 (Cardinality)

尽量每条记录有不同的值

- 并行写入能力 (Write Scaling)

每条记录的shard key具有一定的随机性

- 查询隔离 (Query Isolation)

每次查询尽量到一个分片或少量几个分片去查询

组合分区键

Product_ID + Date

集群性能

- 100TB数据存储
- 每小时3千万条记录写入

Pump查询引擎

- 支持多种数据源，提供统一访问接口

MySQL、Impala、MongoDB、ElasticSearch

- JSON形式的HTTP API调用

- 提供数据查询缓存

- 提供客户端SDK，简化上层调用

```
{
  "ds_type": "sqla",
  "ds_dsn": "impala://ym-044:21050/tm_data_dress",
  "table": "shop",
  "metrics": [
    {
      "field": "*",
      "aggregation": "count",
      "alias": "total"
    },
    {
      "field": "shop_id",
      "aggregation": "count(distinct)",
      "alias": "different_shop_id_count"
    }
  ],
  "filters": [
    {
      "field": "created_at",
      "operator": "range",
      "value": ["2015-12-01", "2016-02-01"]
    }
  ],
  "groupby": ["week(created_at)"]
}
```

查询格式

```
ds = DataSource('mysql', 'mysql://root@localhost:3306/misc', 'shop')
q = Query(ds)

q1 = q \
    .count('*', 'count_*) \
    .cumsum('shop_id') \
    .count_distinct('shop_id', 'count_shop_id') \
    .metric('id', 'count', 'count_id') \
    .flt_range('created_at', ["2015-07-01", "2016-07-01"]) \
    .groupby(('created_at', 'week')) \
    .orderby('week(created_at)')
print(pretty(q1.to_dict()))
```

客户端SDK调用

WhirlPool查询引擎

- 基于Spark提供多维交叉分析

利用RDD进行复杂的聚合运算

- 提供类似于ElasticSearch的DSL查询语言

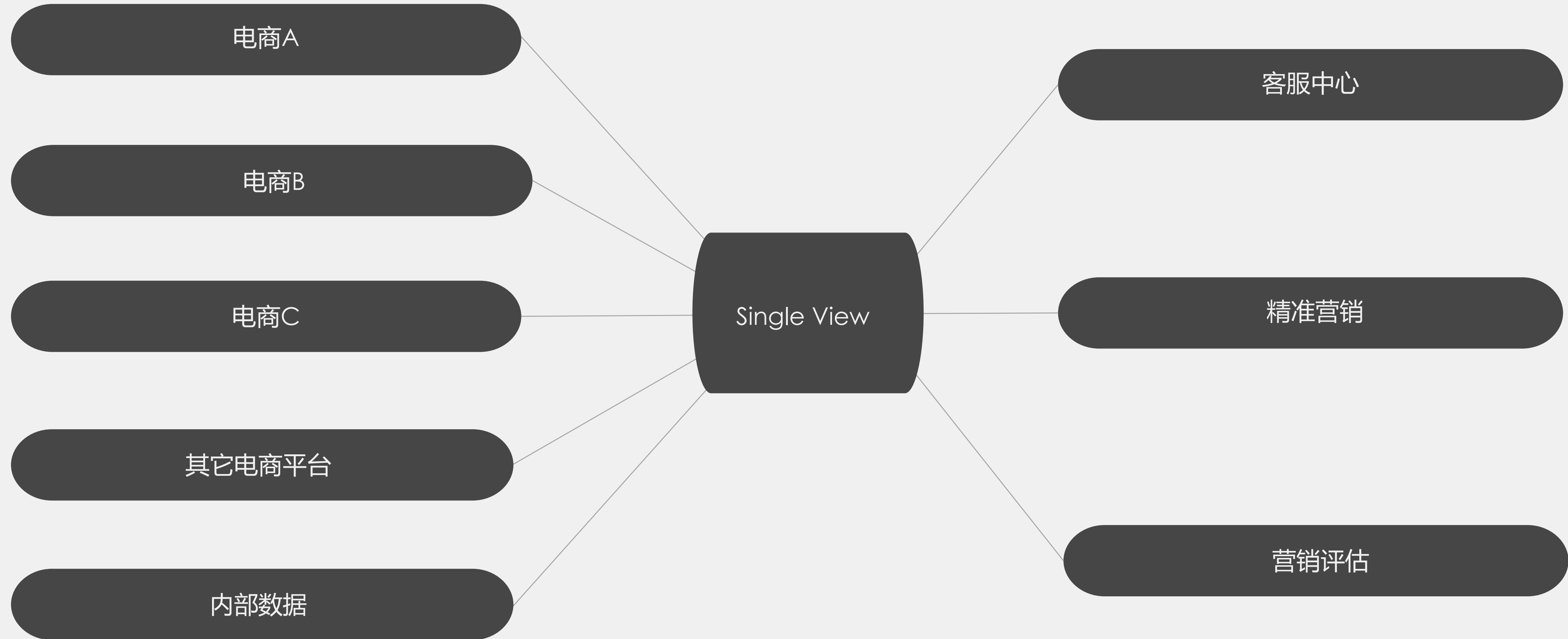
- HTTP API接口调用

- 查询语言更加人性化

- 提供数据查询缓存

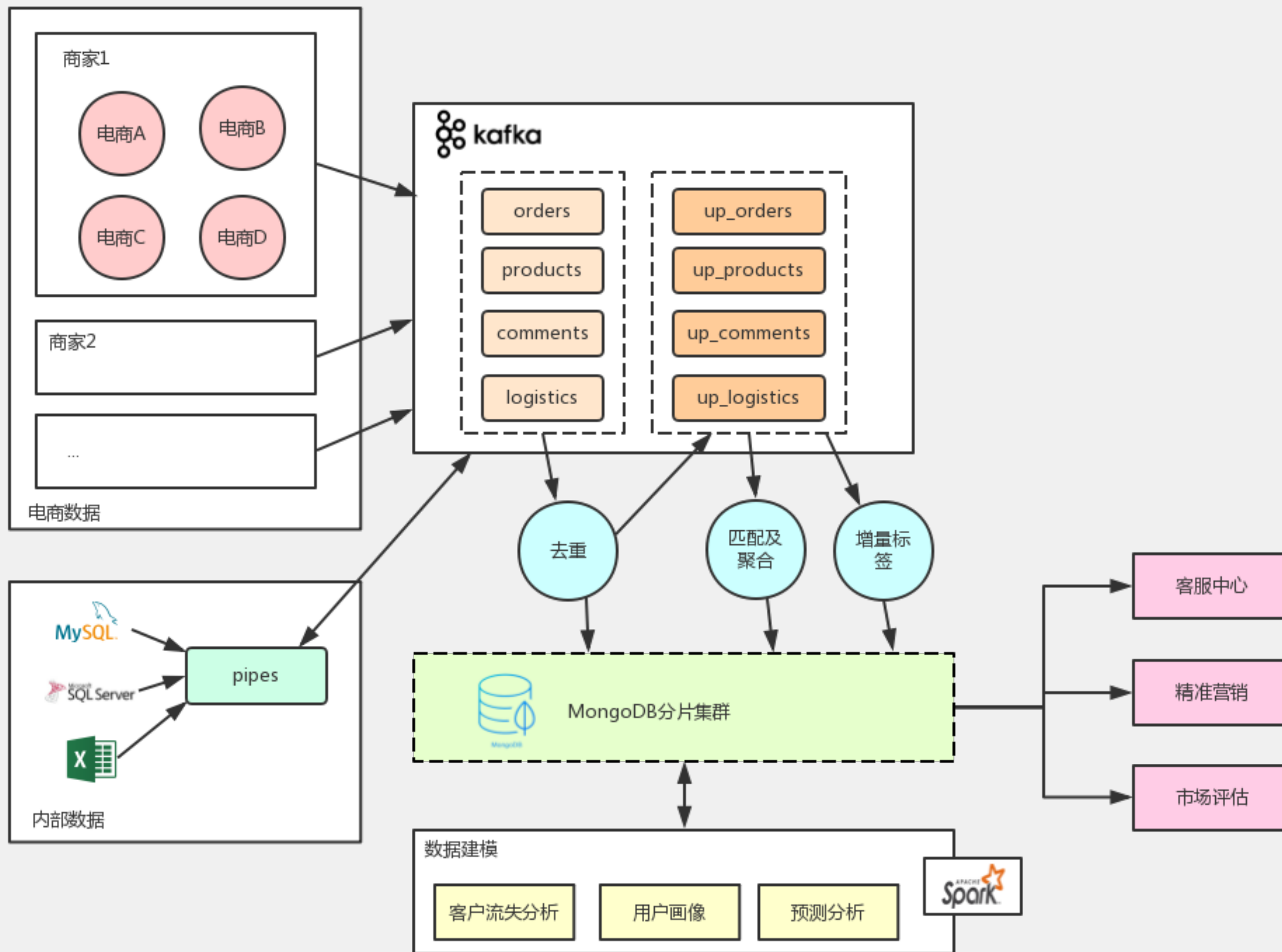
统计每个月每个品类商品在不同价格区间中的销量

```
{
  "aggs": {
    "date_histogram": {
      "field": "date",
      "interval": "month",
      "aggs": {
        "terms": {
          "field": "category",
          "aggs": {
            "histogram": {
              "field": "price",
              "interval": 10,
              "aggs": {
                "sum": {
                  "field": "sales"
                }
              }
            }
          }
        }
      }
    }
  }
}
```

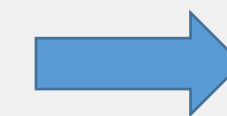
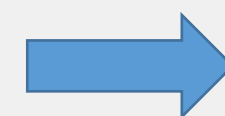
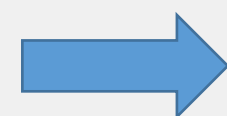


Technical Stack

Single View

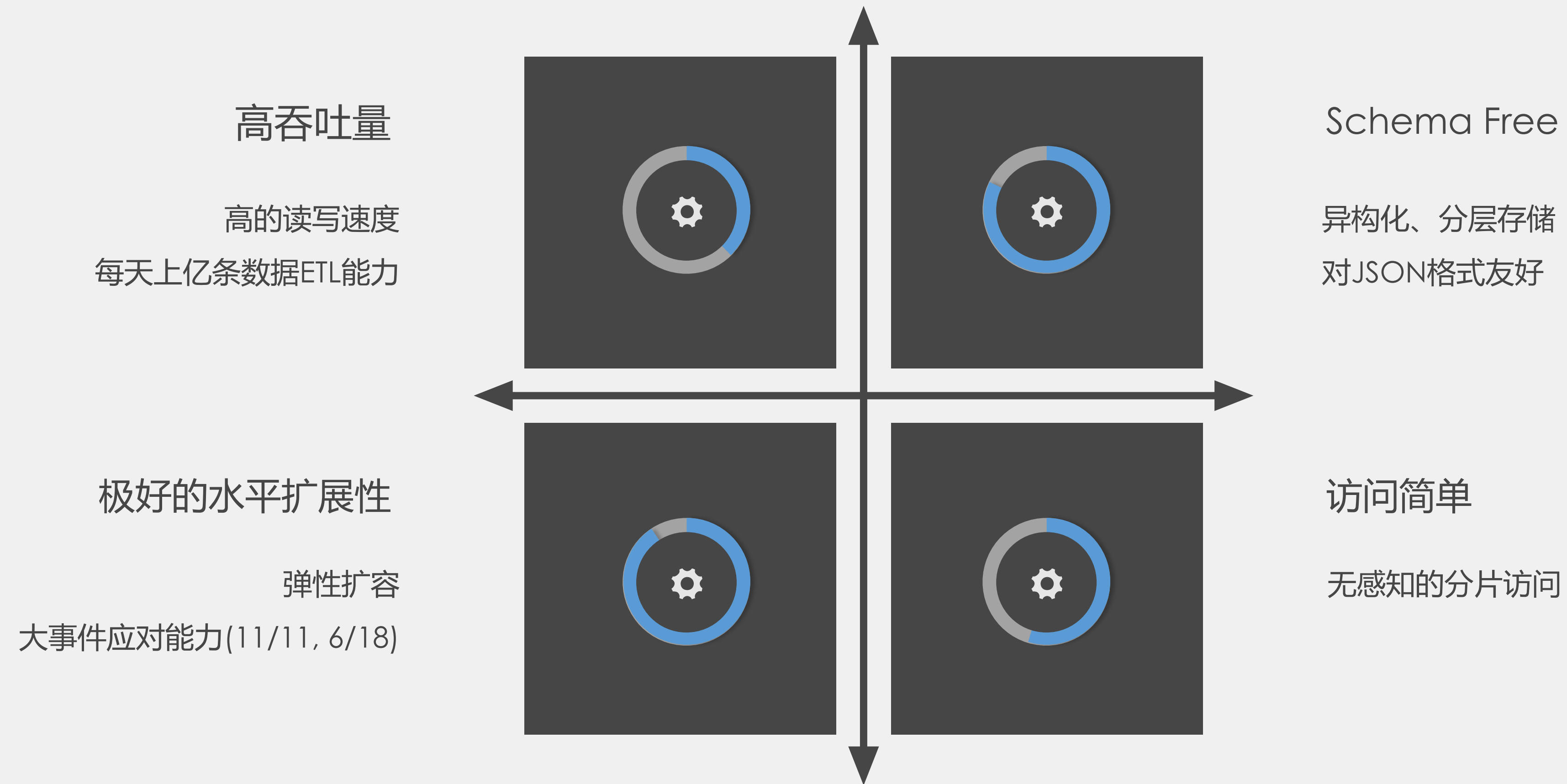


数据仓库进化史

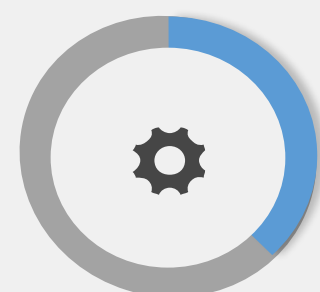


	优点	缺点
MongoDB	<ol style="list-style-type: none">1.schema-free适合与爬虫应用场景；2.提供完整的索引功能，包括secondary, compound索引；3.自动分库，极强的水平扩展能力；4.社区活跃，资料齐全；	<ol style="list-style-type: none">1.不支持类型转换函数；2.统计功能相对较弱，而且书写复杂，不易理解，比如实现count(distinct)功能；
Cassandra	<ol style="list-style-type: none">1.去中心化，具有极强的水平扩展能力；2.混合keyvalue和列式存储结构，需要极强的写入能力；3.提供primary index和secondary index索引功能，但索引可能对性能有较大影响；	<ol style="list-style-type: none">1.面向query的数据建模方式，上层query的更改对数据库有很大影响；2.collection类型不支持嵌套；3.集群运维不方便，添加节点、替换节点和删除节点需要遵从不同的操作流程和注意事项；4.缺乏图形化的监控工具；5.不能做aggregation统计；6.文档资料不完善，社区不够活跃；

MongoDB的优势

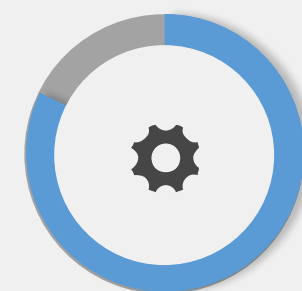


MongoDB的劣势



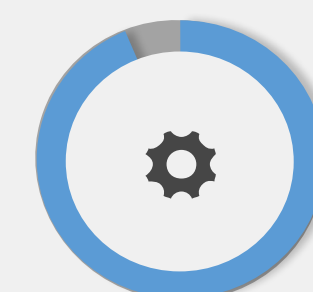
复杂查询

Hive & Impala



文本挖掘

ElasticSearch

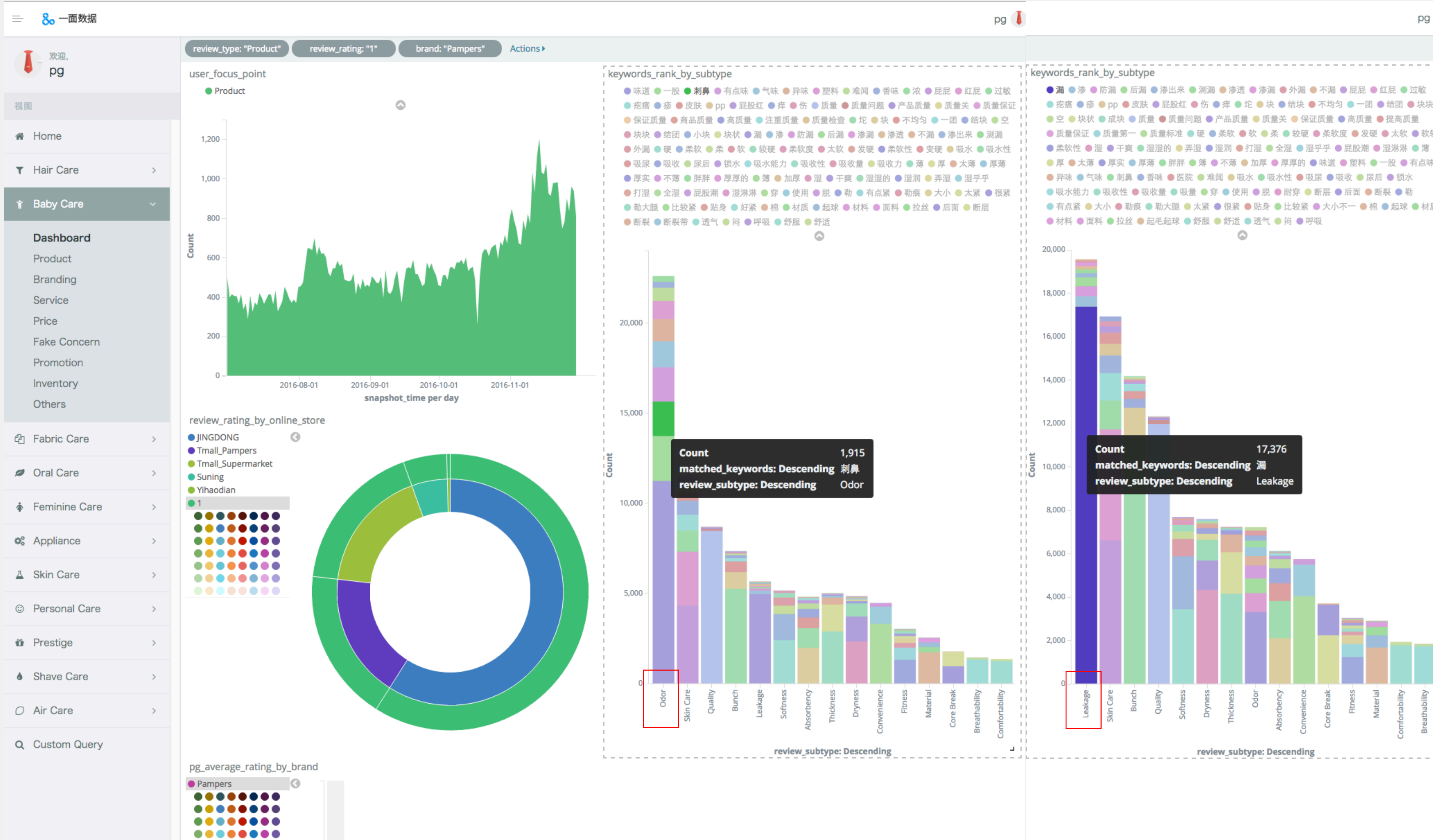


分析建模

Spark

Quantitative Rating & Reviews

通过用户评论改善产品质量



A品牌(某产品)

6个电商平台

120万条用户评论

20%负面评论来自于气味

B品牌(竞品)

5%负面评论来自于气味

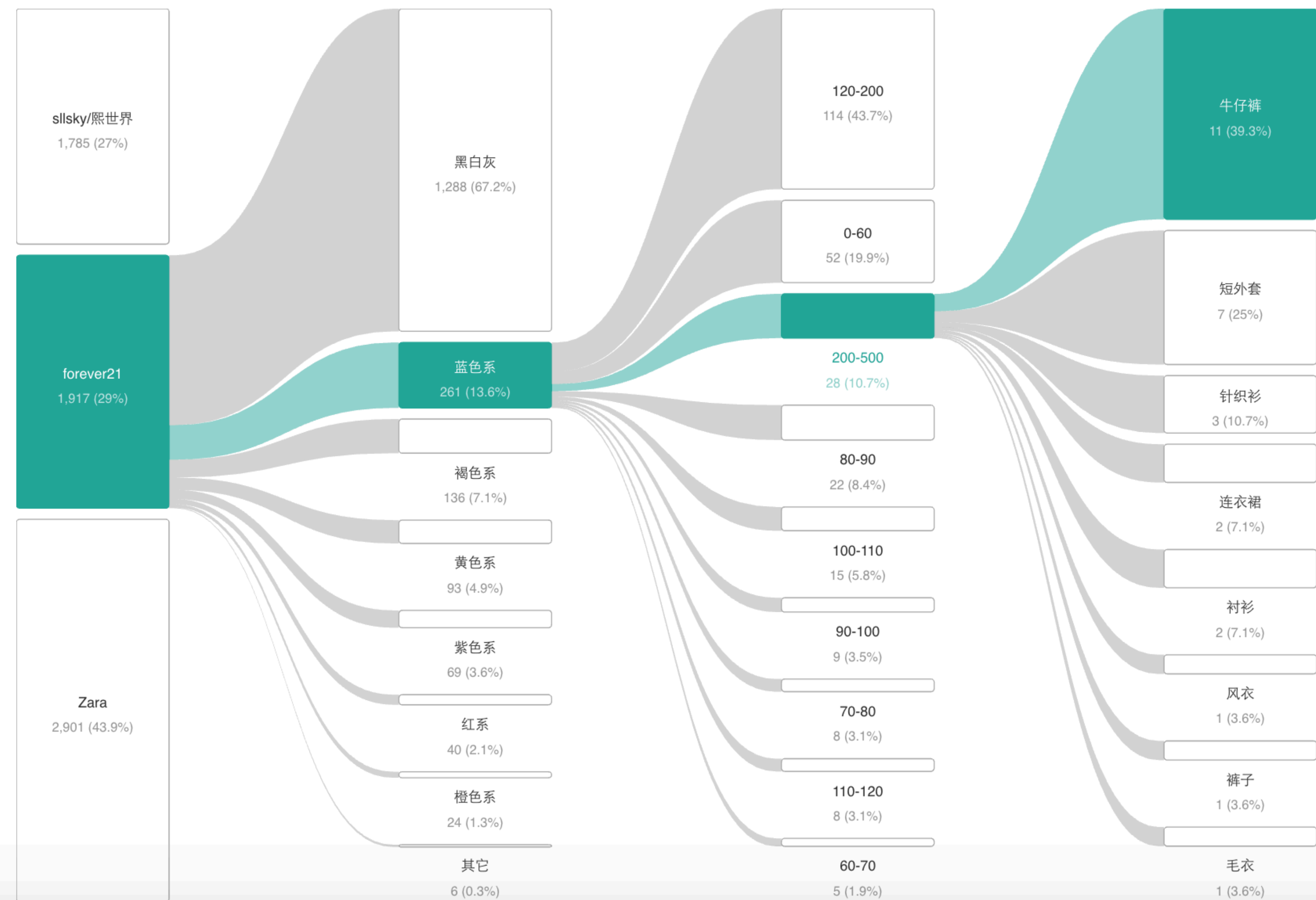
但是15%负面评论来自于侧漏

Market Landscape

竞争对手策略分析

销售额 销量 SPU数 **SKC数**

颜色 价格带 品类 提示：标签可选、可拖拽排序；SKC支持颜色分析 全选 (3)



商品

SPU/SKC数

类别 / 颜色 / 款式

价格/销售额/销量

趋势监控

时尚女鞋

流行趋势

销售额/销量/SKU数

分类特征

品类

风格

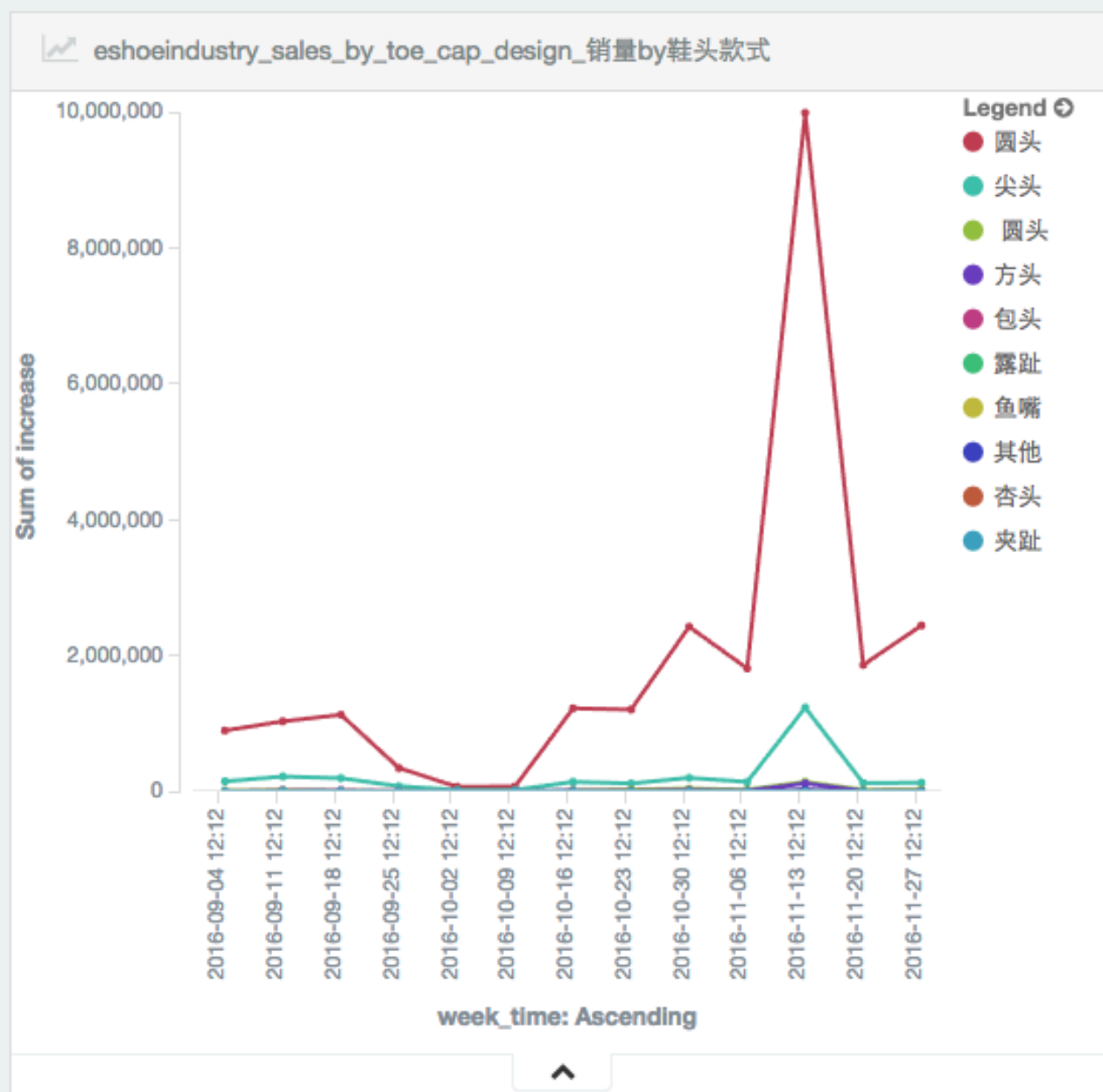
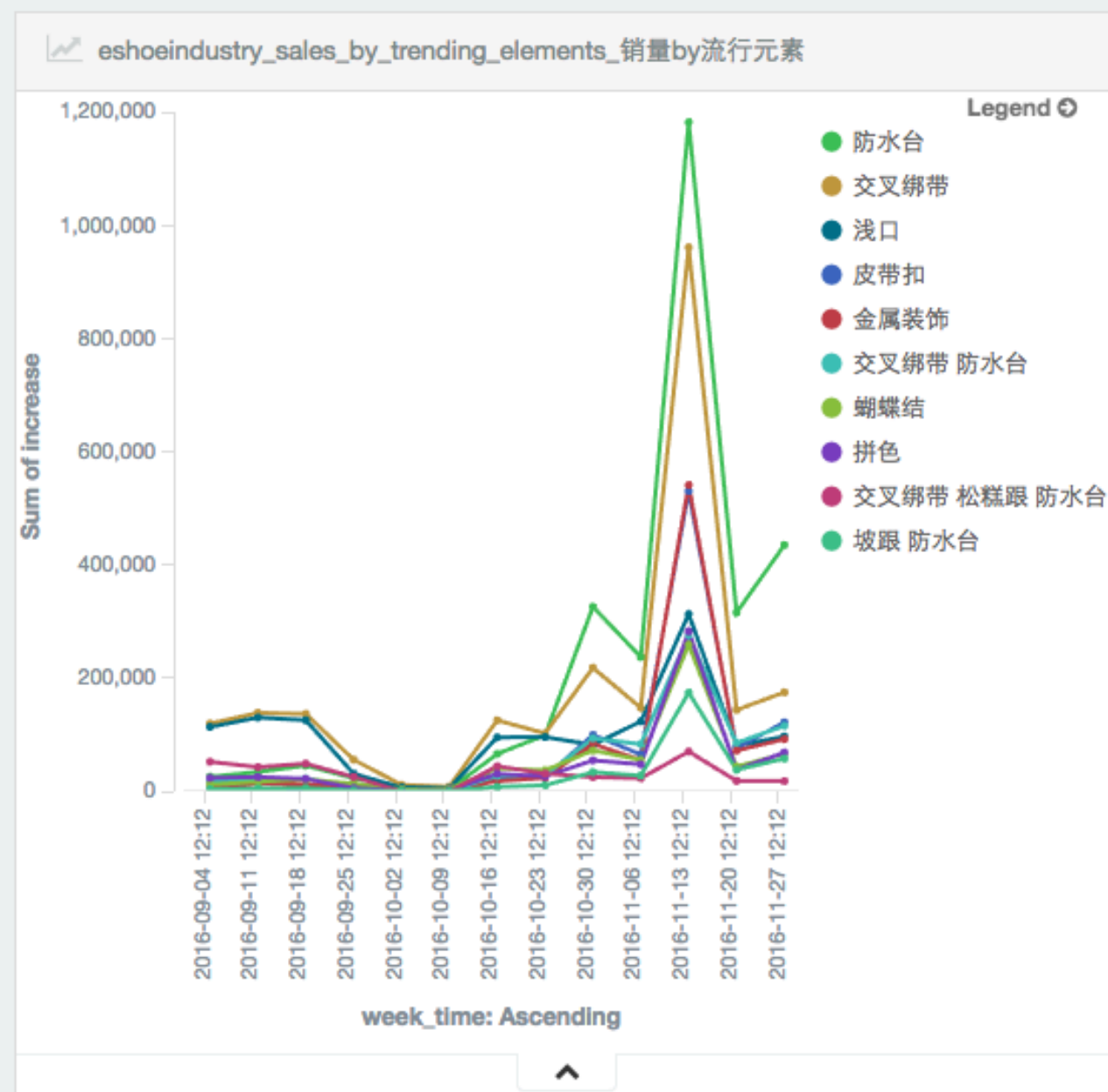
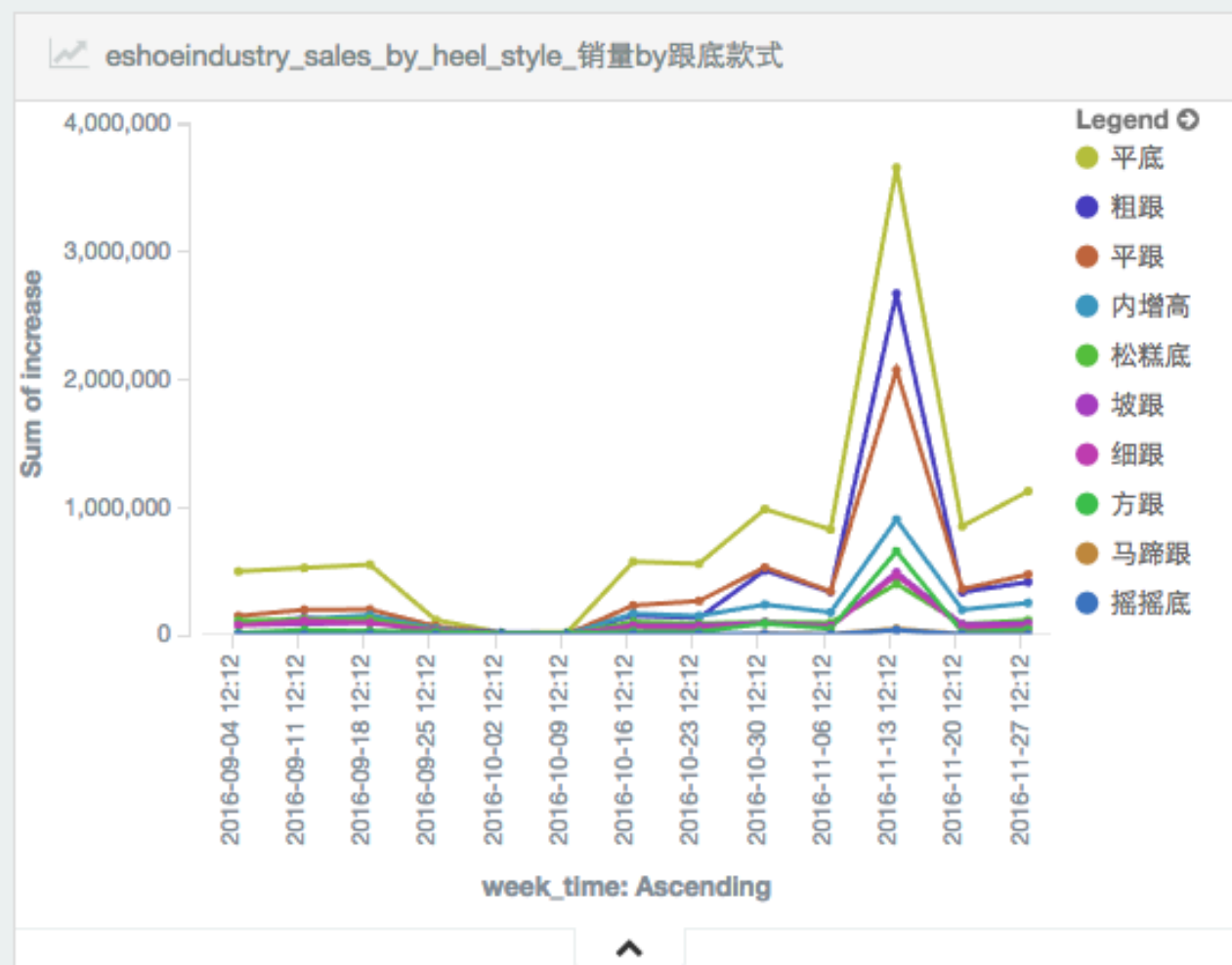
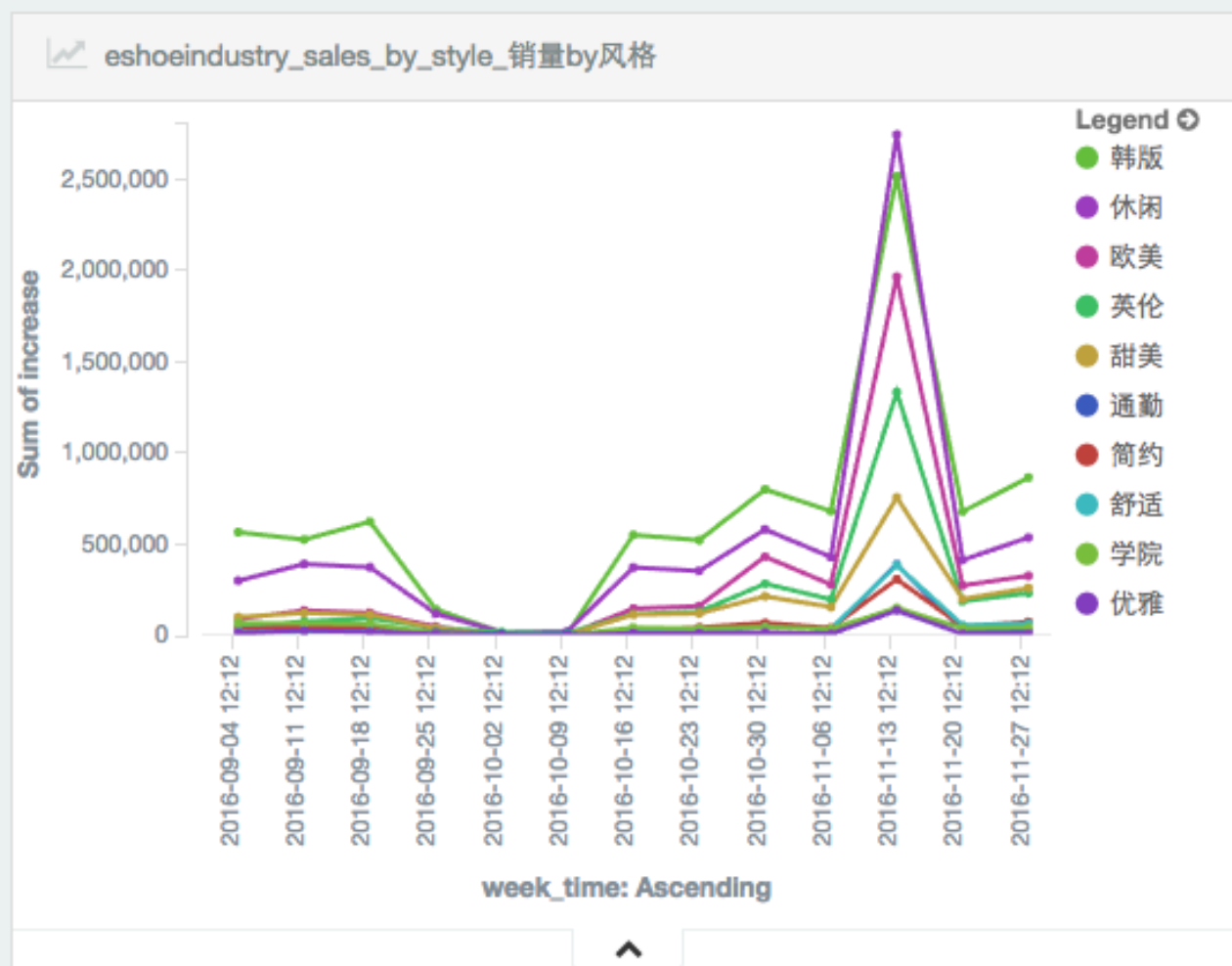
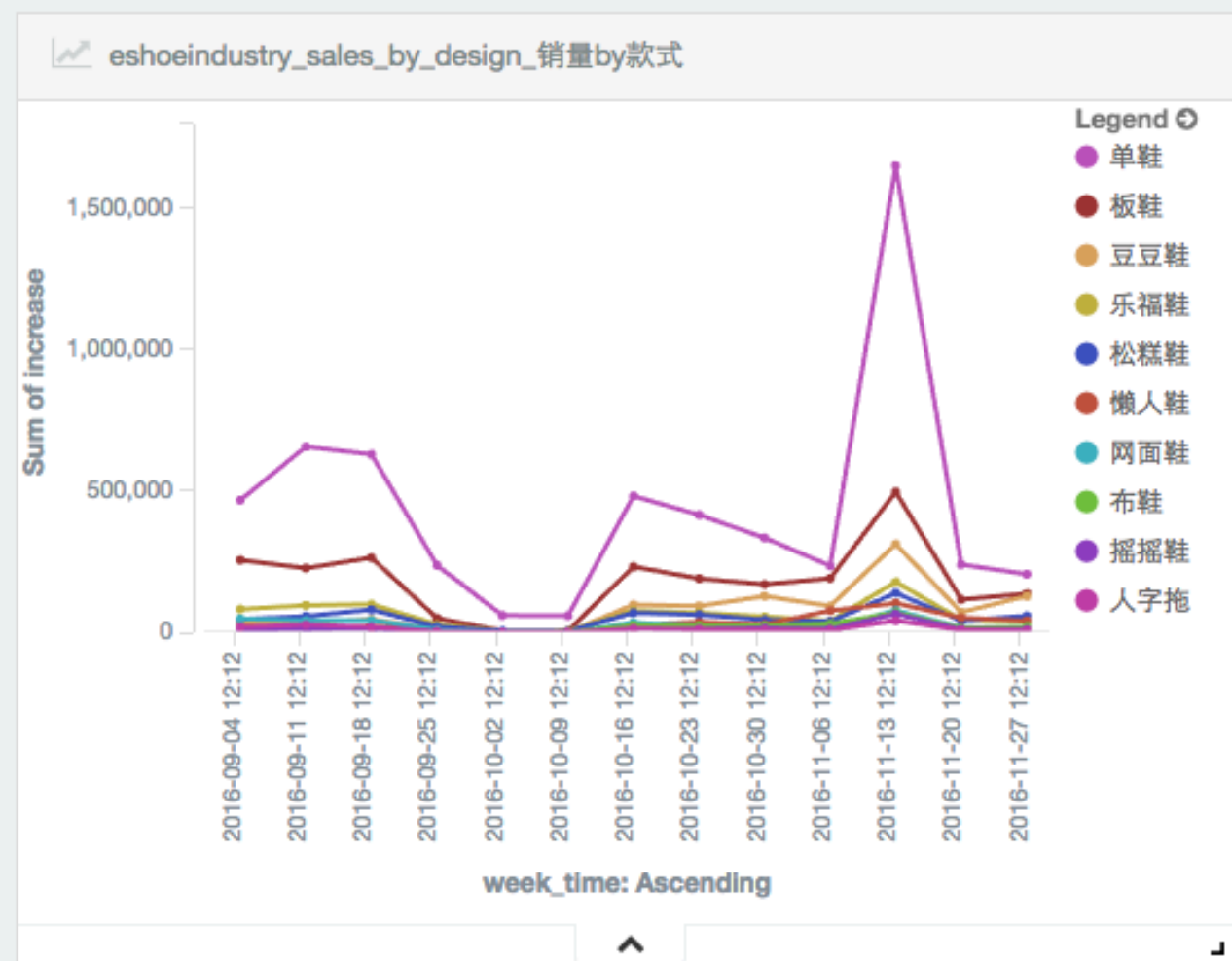
跟型

跟头

开口

流行元素

颜色





If you can't measure it,
you can't improve it.

EMAIL data@yimian.com.cn
<http://yimian.com.cn>

Phone +86 755 - 86503625

