Haier

海尔大数据技术构建和运营实践

海尔电器 大数据技术总监 黎洋

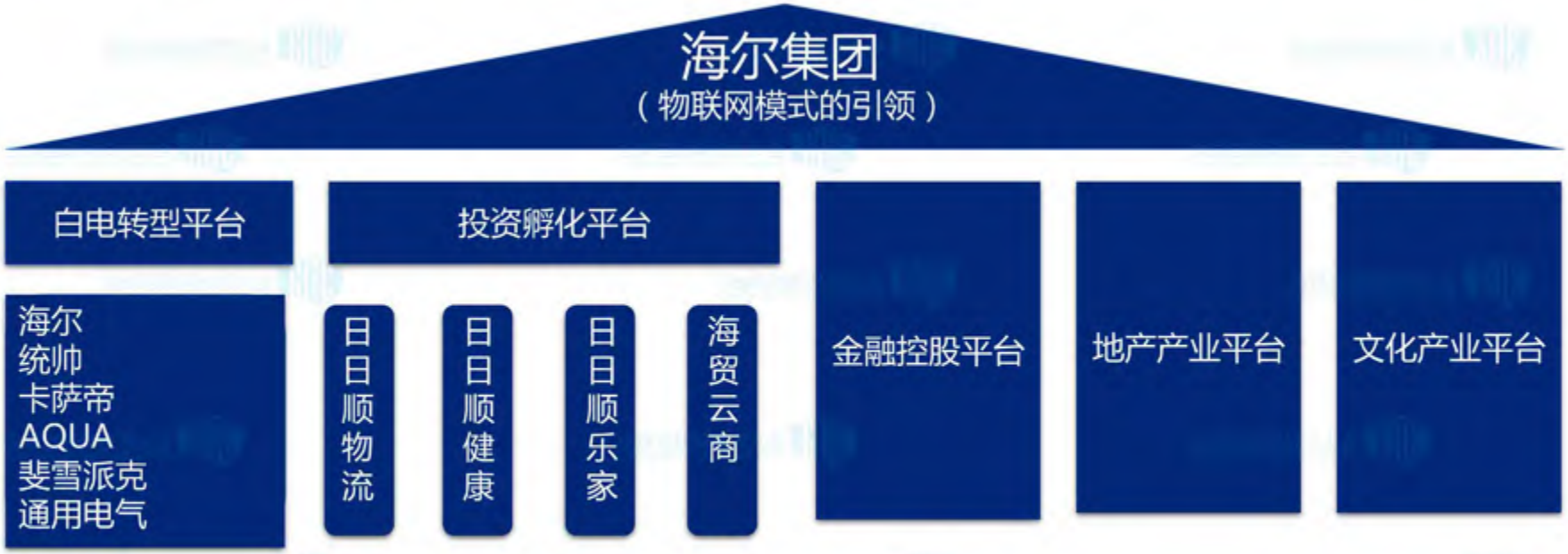
个人简介

Haier

- 曾在HP、 Standard Chartered 从事DBA，主数据管理，数据仓库、数据分析方向
- 2014.12 加入Haier，负责大数据平台搭建，数据治理，搭建运营指标体系

海尔集团介绍

Haier



大纲

Haier

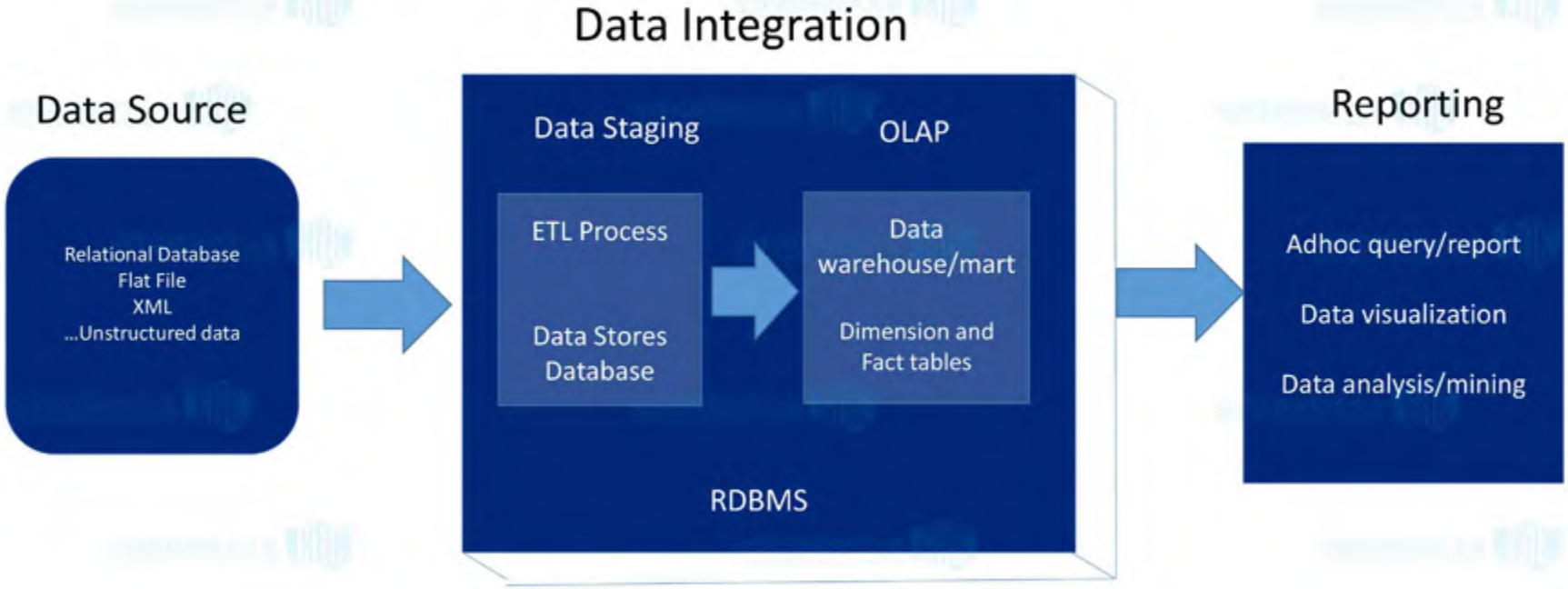
- 大数据解决方案实践
- 运营案例分享

Haier

大数据解决方案实践

传统数据仓库解决方案

Haier



报表工具选型

Haier

REPORTING

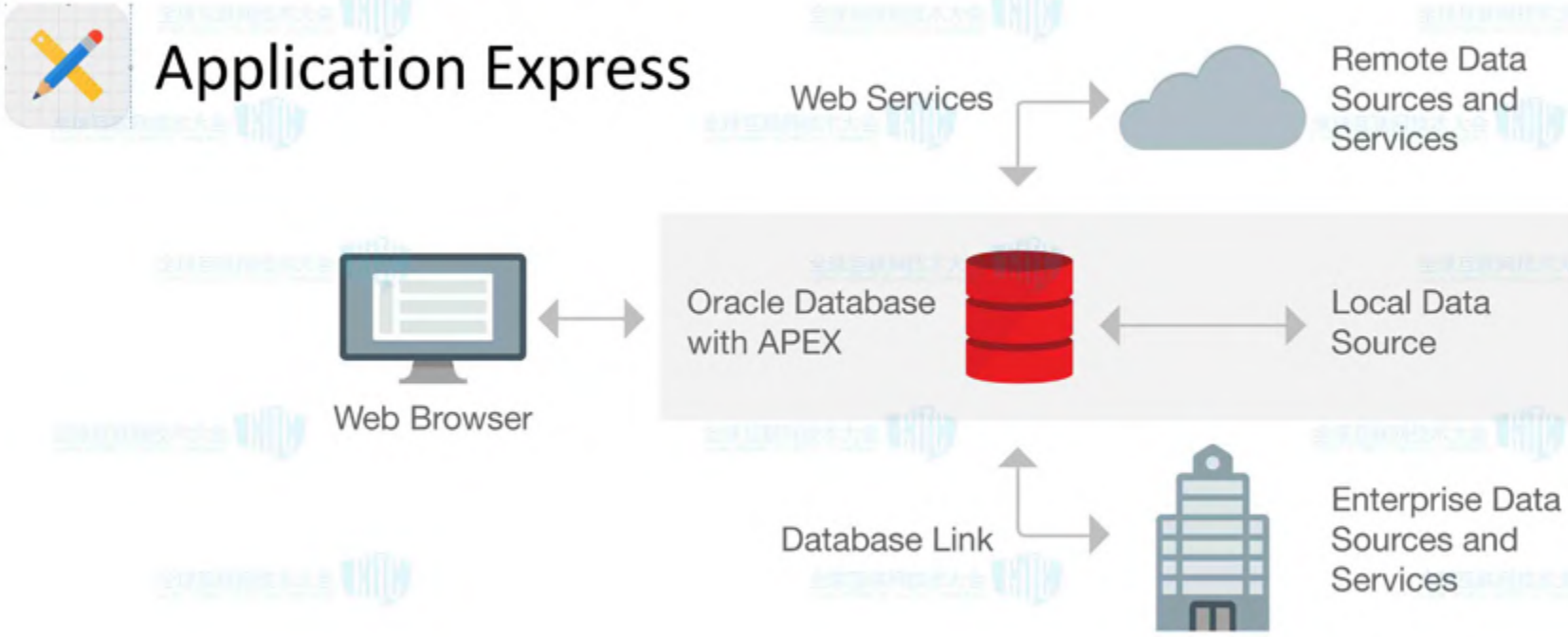
Data Analysis

Data visualization

Adhoc Query/report

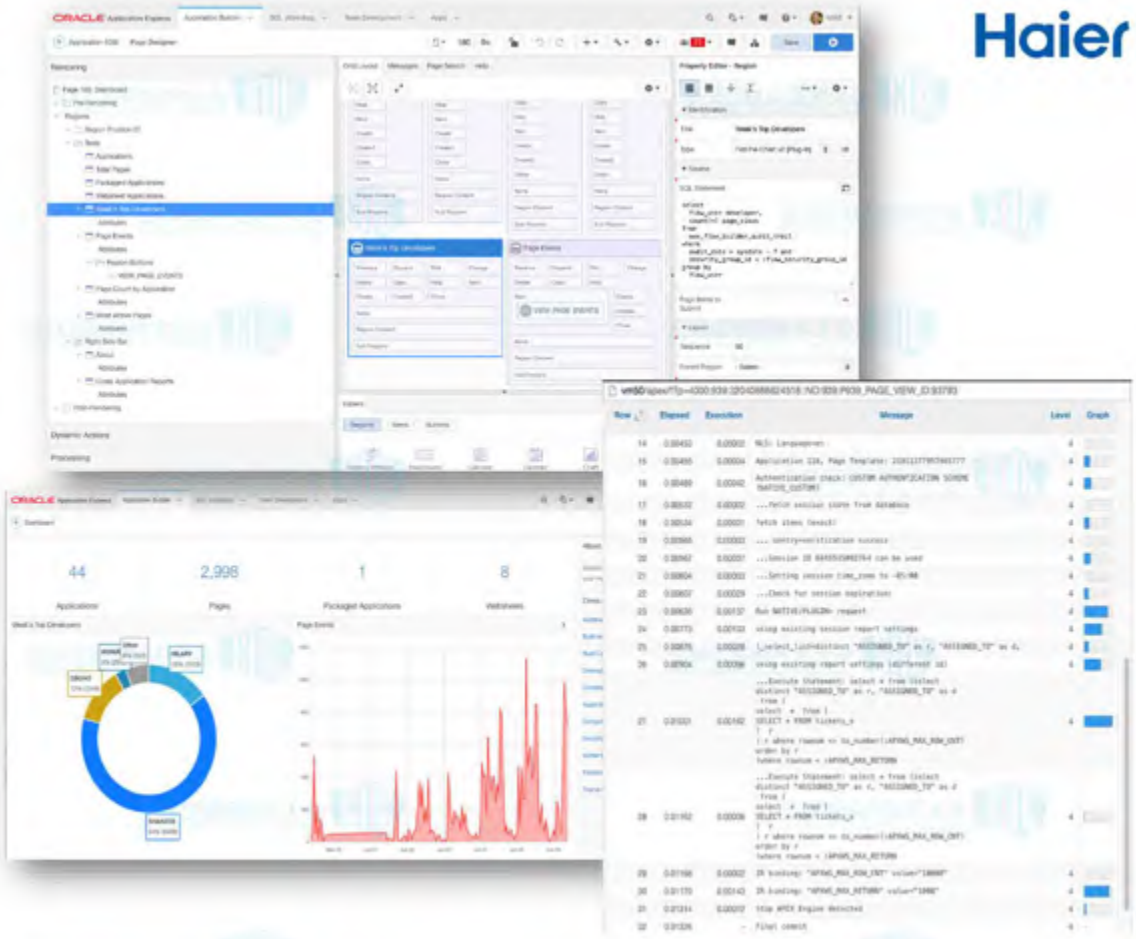
Adhoc Report方案

Haier



Application Express

- SQL and PL/SQL
- Declarative framework , platform as a service
- Agile development
- Browser-based IDE
- Short learning curve
- No additional cost



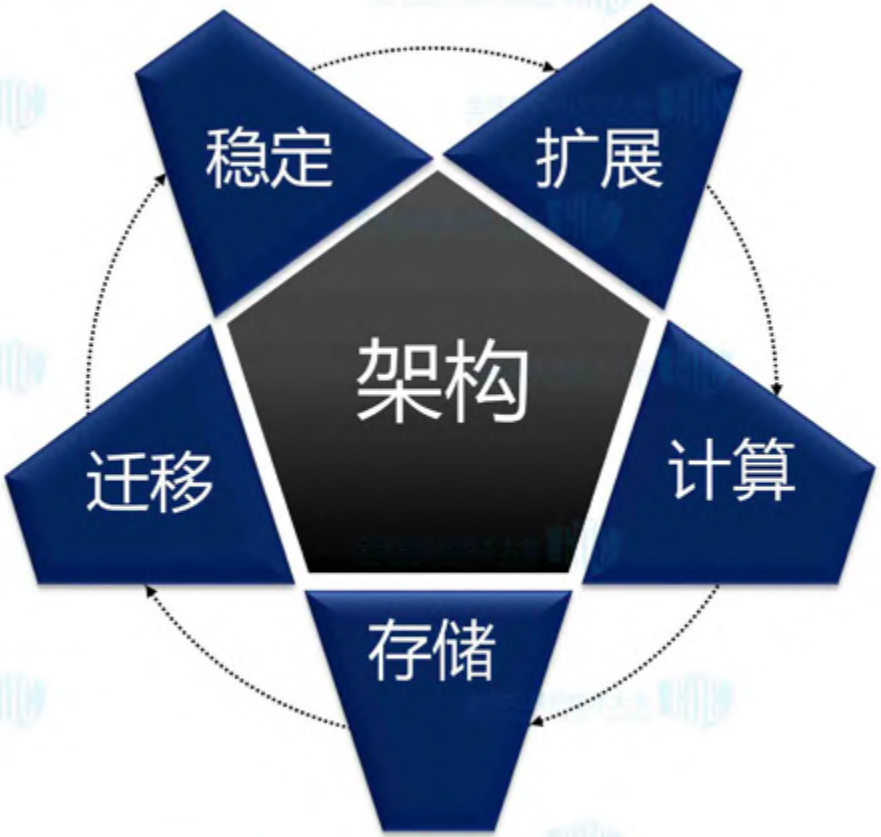
数据累积、维度爆炸带来的问题

Haier

- Performance
- Cost
- Scalability

架构升级考虑

Haier



系统整合

Haier

1, File system layout of HDFS

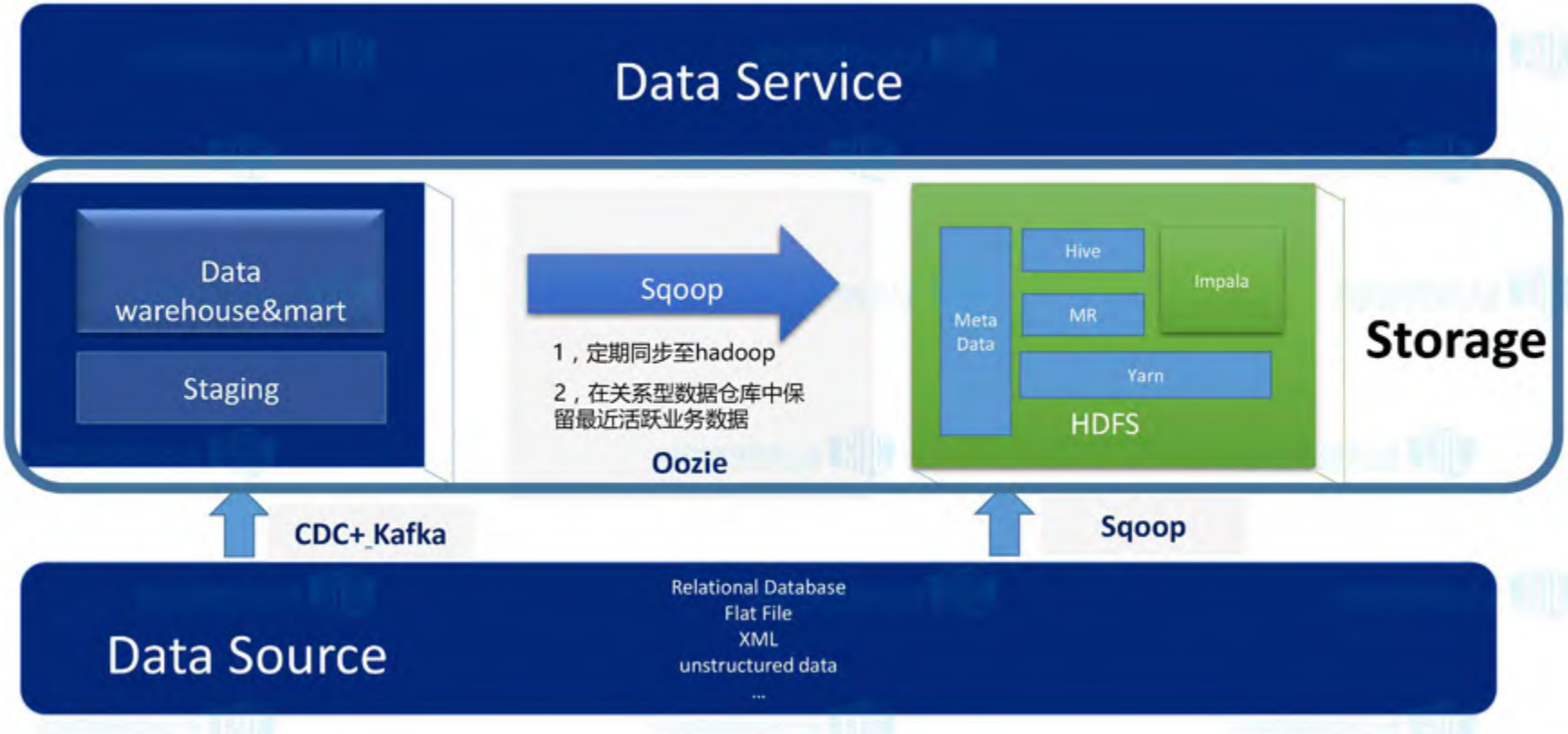
2, Batches or real-time ingestion

3, File format

4, Partition

引入Impala, 实时查询引擎

Haier



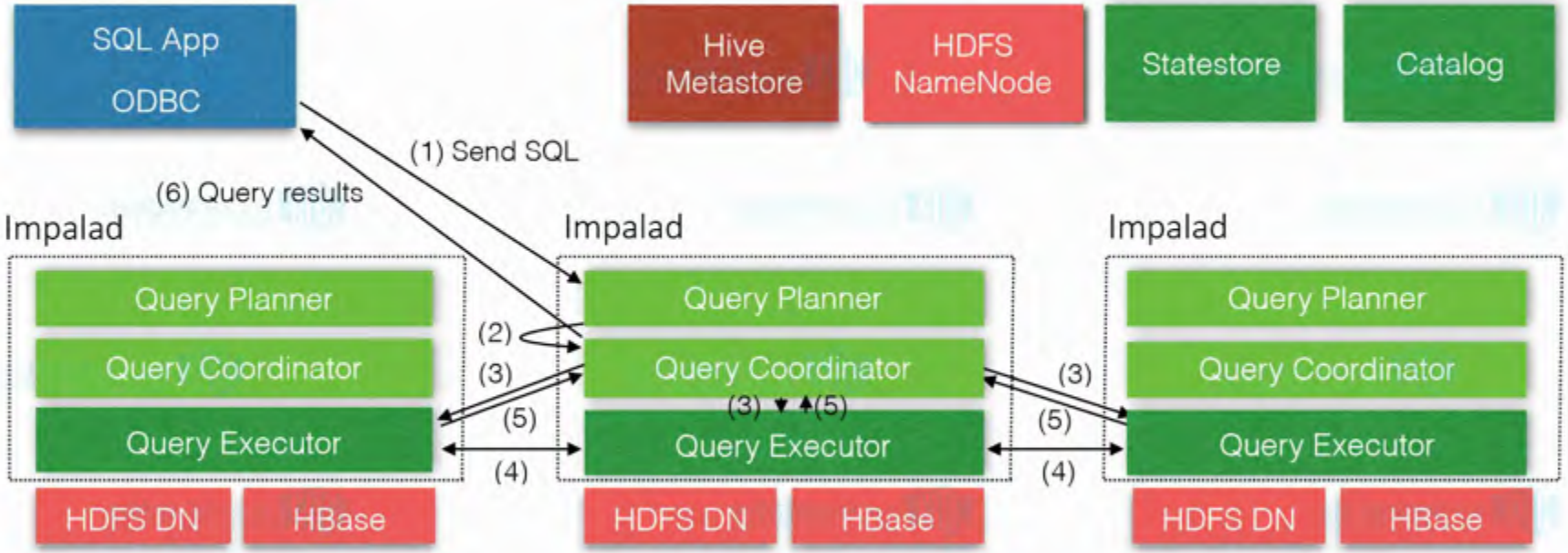
Why Impala?

Haier

- Parallel SQL query engine
- Utilizes standard Hadoop components
  (HDFS, Hbase, Metastore, Yarn)
- High-performance, fully open-source
- Industry-standard interfaces
- (odbc/ jdbc, Kerberos and LDAP, ANSI SQL)
- Predicate pushdown

Impala 架构

Haier



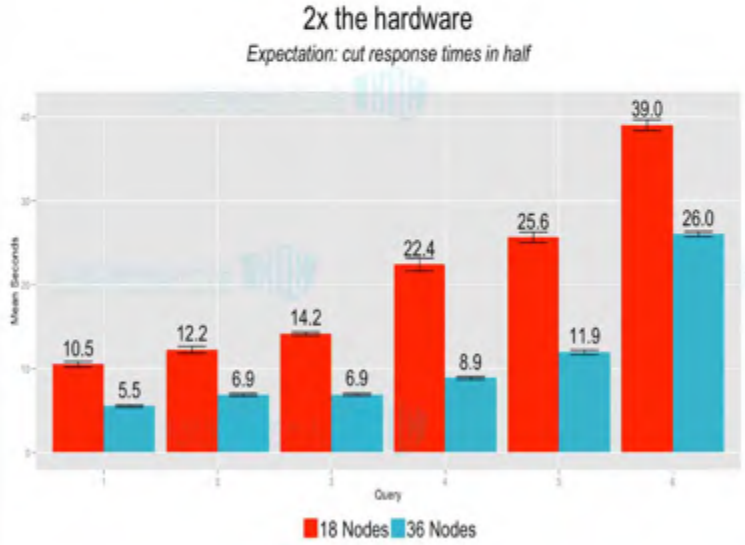
Impala性能测试

Haier

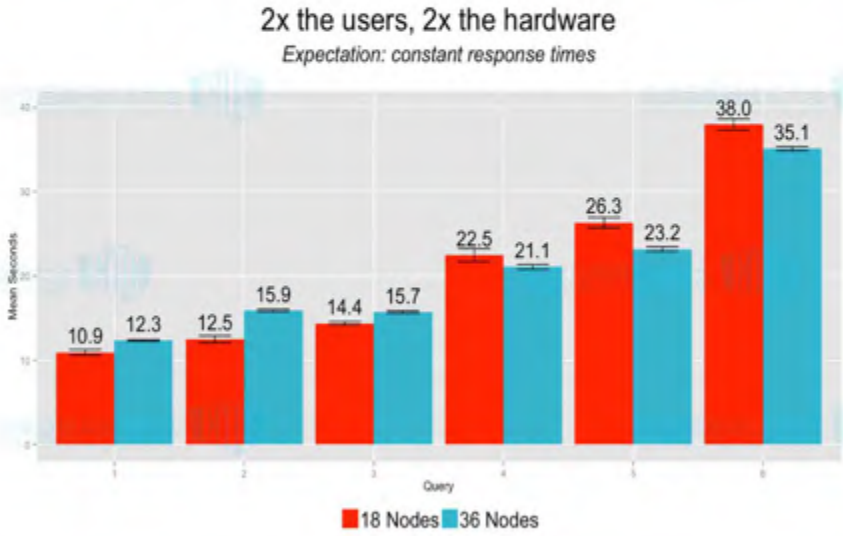
TEST SETUP: 15 TB TPC-DS data set

18 nodes vs 36 nodes

Latency



Concurrency



Impala SQL性能考虑

Haier

I/O

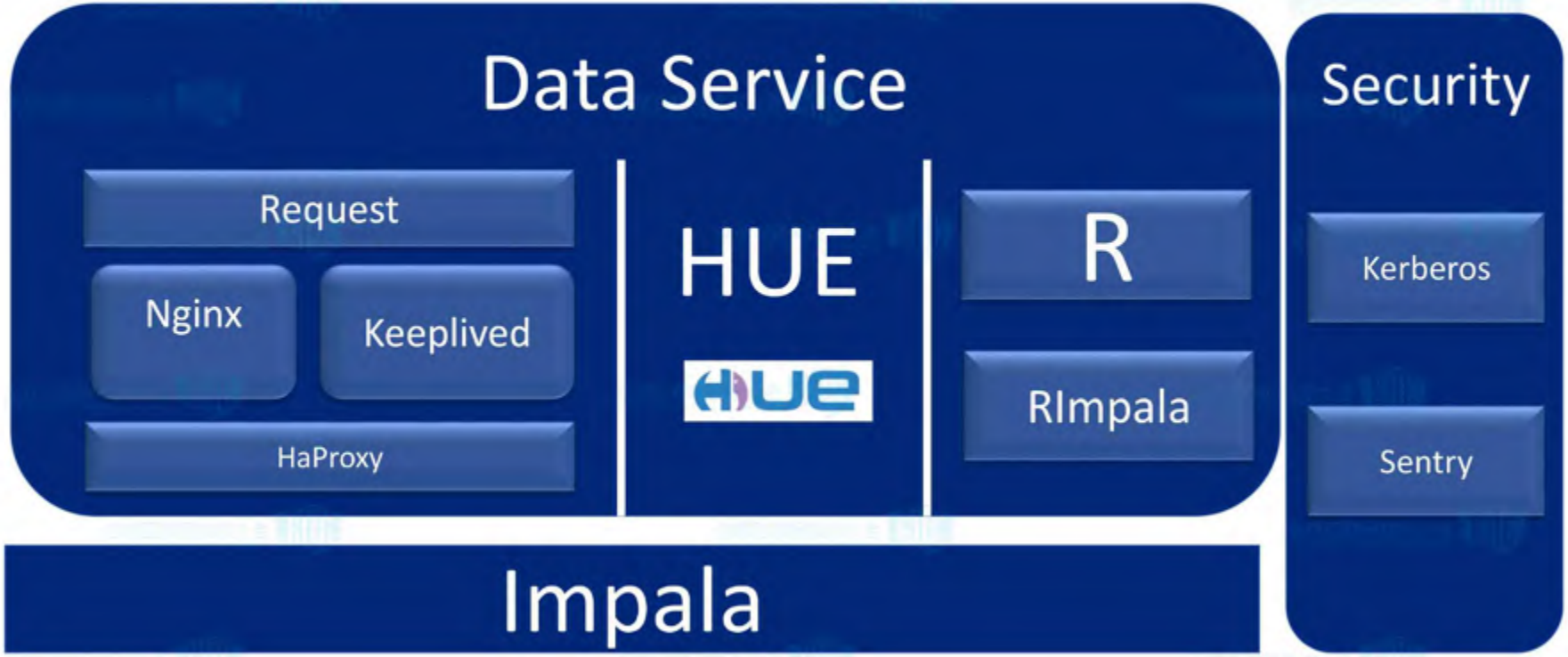
- File format
- Small files
  - data ingestion
  - Partition granularity
- Parquet block size
- Data transfer
- Data skew

Execution Plan

- Statistics
  - Table
  - Column
- Runtime filter
- Join
  - Join order
  - Table size
  - Broadcast/partitioned joins
  - Without stats.

Data Service Interface

Haier



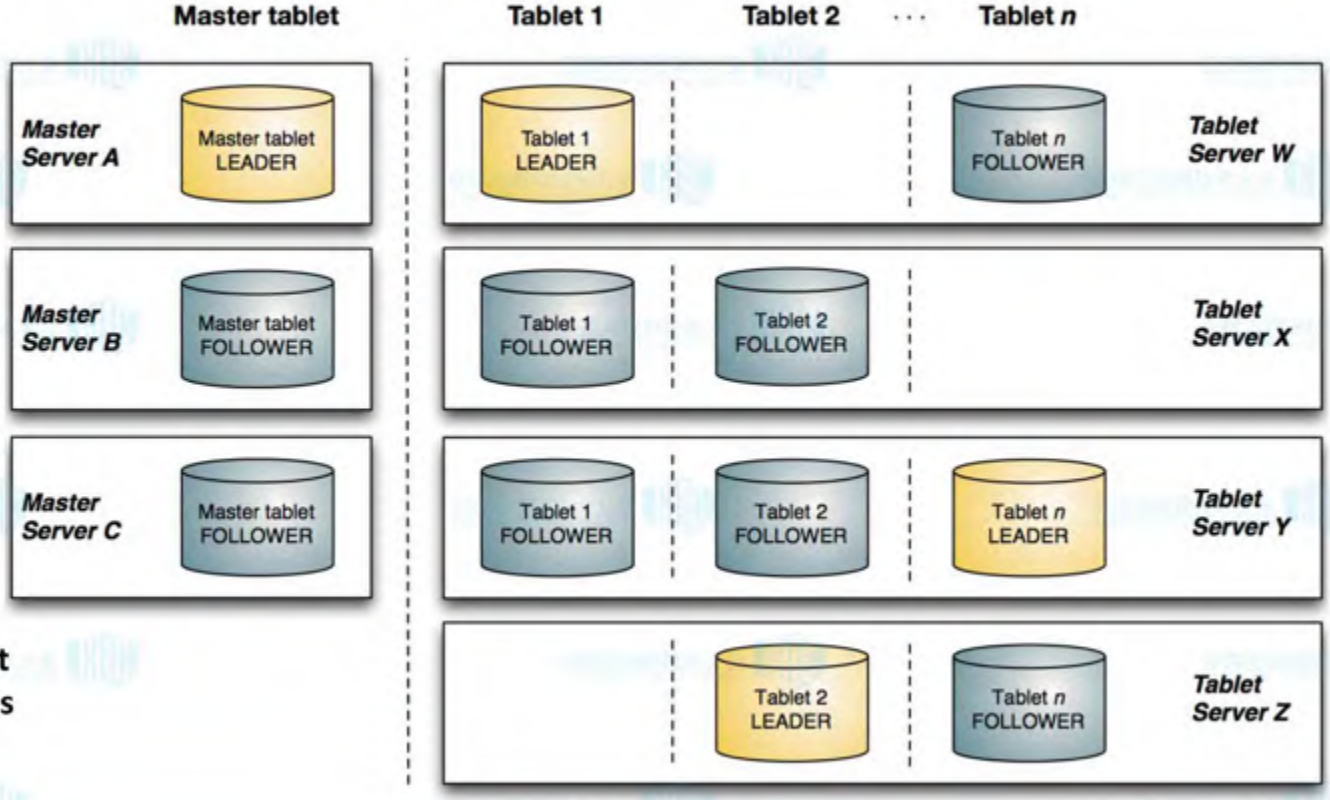
后续计划

Haier

- 数据仓库全部迁移，UDF开发
- 解决频繁数据加载更新问题

Kudu network architecture

Haier



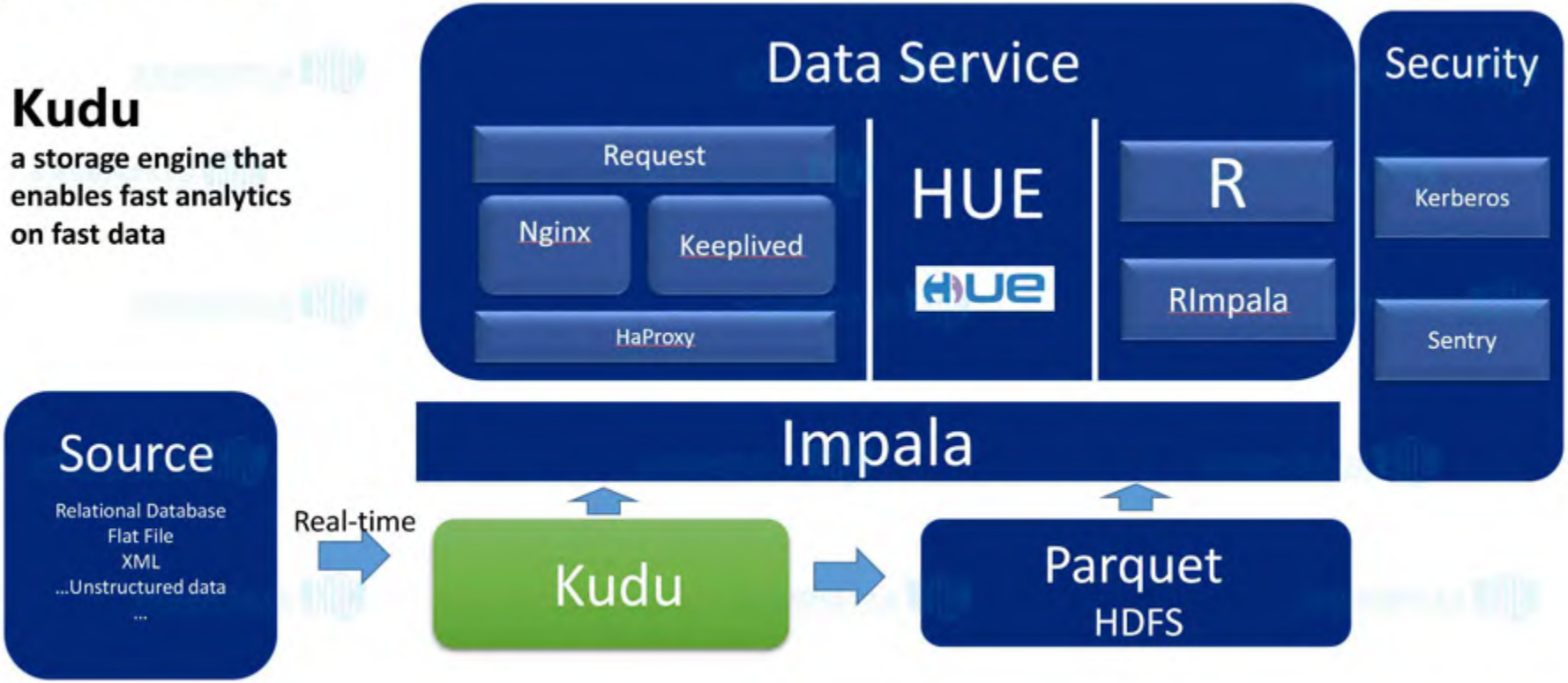
a storage engine that enables fast analytics on fast data

探索Kudu

Haier

Kudu

a storage engine that enables fast analytics on fast data

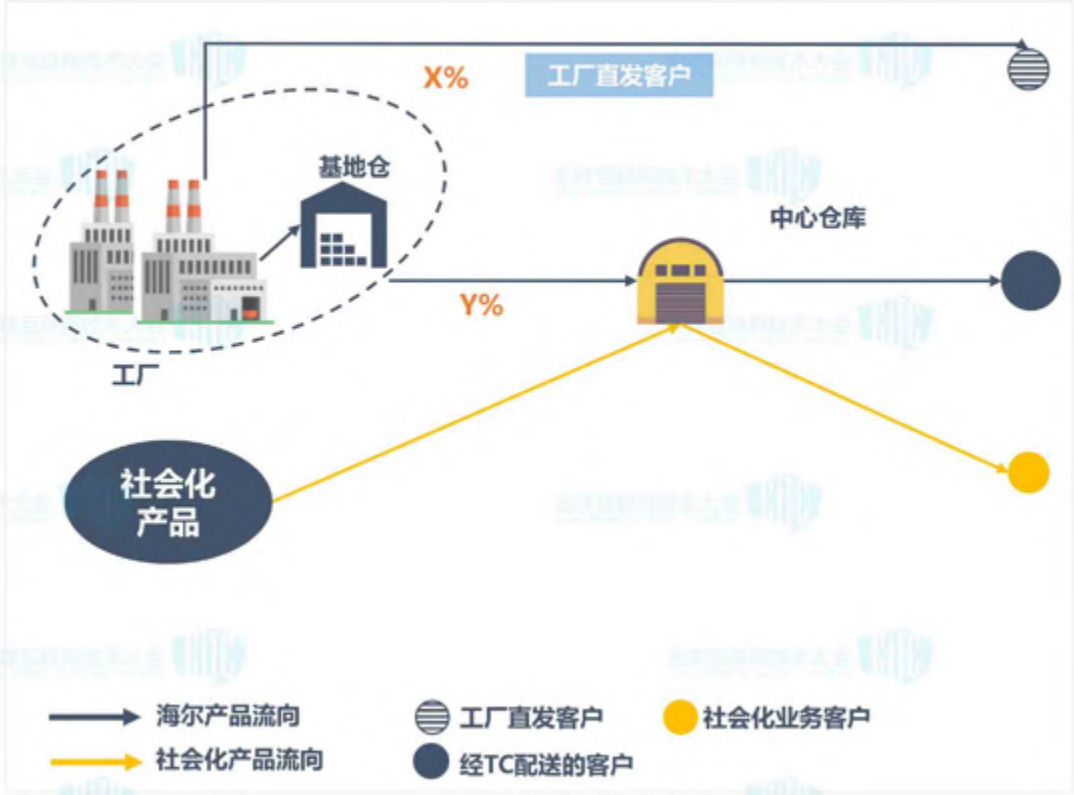


Haier

运营案例分享

海尔主要业务介绍

物流业务


X%
工厂直发客户
基地仓
中心仓库
Y%
工厂
社会化
产品
海尔产品流向
社会化产品流向
工厂直发客户
经TC配送的客户
社会化业务客户

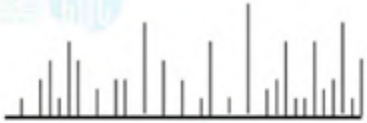
Activity Based Classification

Haier

- **需求分析主要维度**

1.间歇性（Intermittency）：平均的需求间隔
2.变异系数（Dispersion）：需求的离散程度
3.波动性（Variability）：需求的标准差

Activity Based Classification

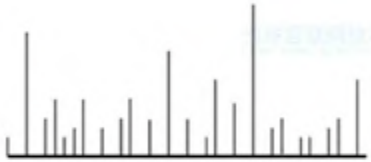


Non-intermittent

持续

Smooth (fast)

流动快稳定



Erratic

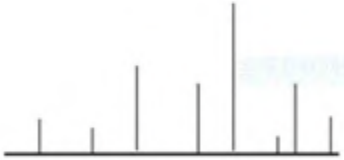
流动快不稳定

Intermittent

间歇

Slow

流动慢



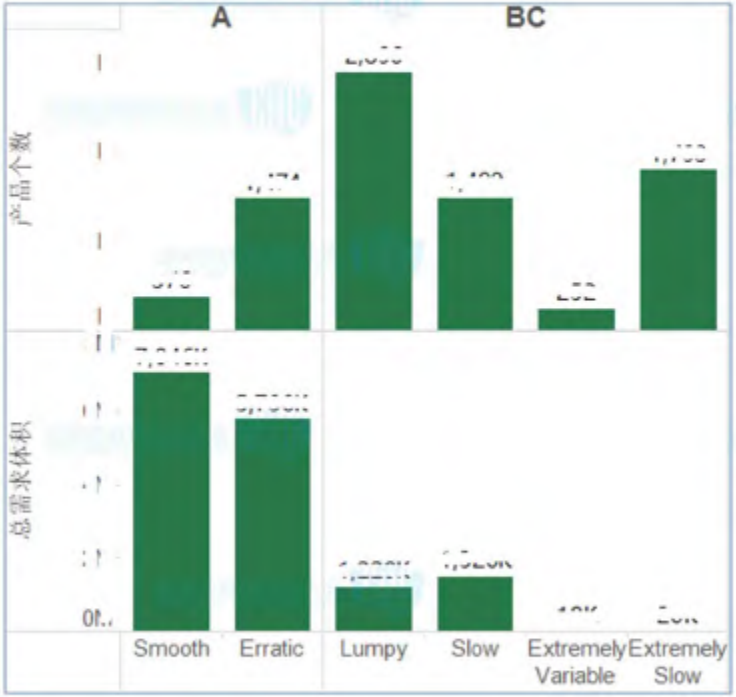
Lumpy

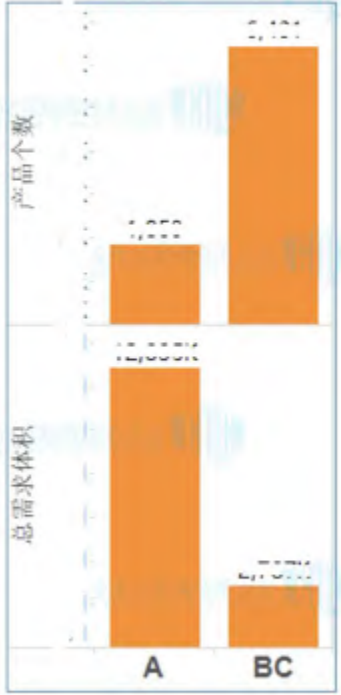
流动慢不稳定

Activity Based Classification

Demand Class	说明	产品分类
Extremely Slow	流动极端缓慢	BC
Smooth	流动快稳定	A
Erratic	流动快不稳定	A
Slow-Low Variable	流动慢-低波动	BC
Slow-Highly Variable	流动慢-高波动	BC
Lumpy	流动慢-不稳定	BC
Extremely Small	平均单量极小	BC
Extremely Variable	极端波动	BC

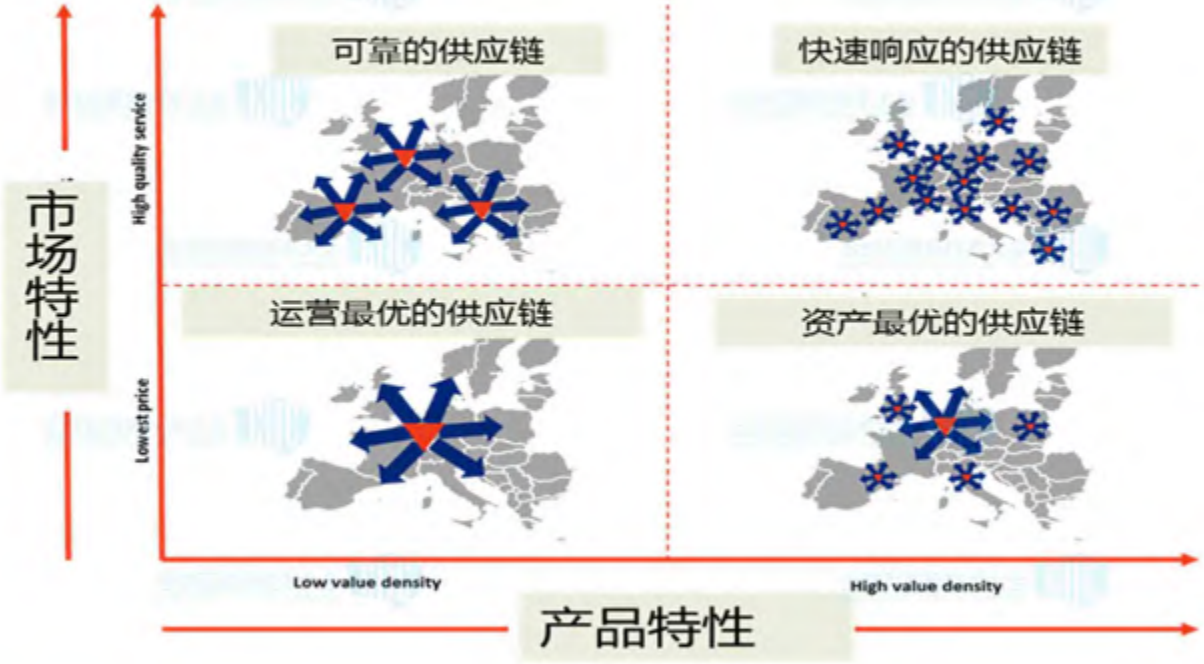
产品分类结果分析





1. A类产品产品数占22.3%；A类产品需求量占82%
2. 仅针对产品的需求特征进行分类，未对产品价值做分类
3. 针对不同产品分类，可以推荐不同的库存管理规则

优化方向



- 仓库选址
- 最优路径
- 配送成本

思考

Haier

- 明确价值，建立关联，小处着手
- 解决数据孤岛问题 VS 数据安全问题