

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

- 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

海尔集团

(物联网模式的引领)

白电转型平台

海尔
统帅
卡萨帝
AQUA
斐雪派克
通用电气

投资孵化平台

日日顺
物流

日日顺
健康

日日顺
乐家

海贸
云商

金融控股平台

地产产业平台

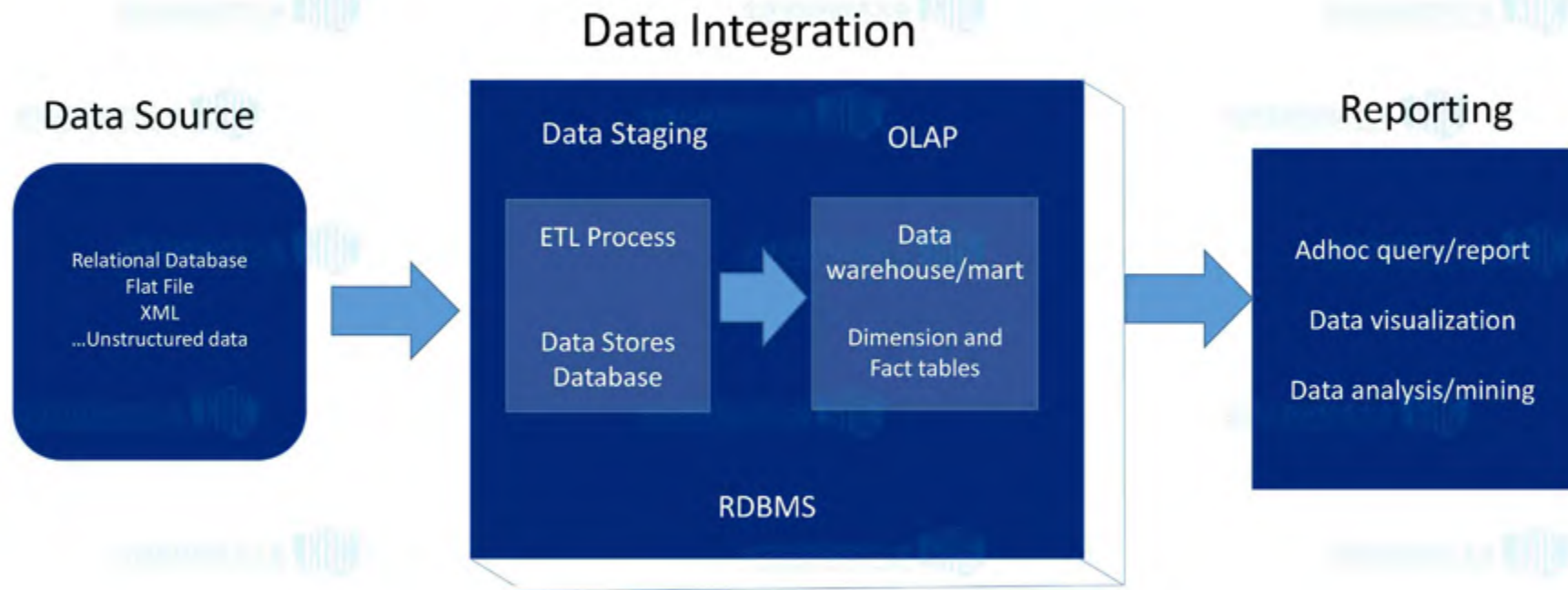
文化产业平台

- 大数据解决方案实践

- 运营案例分享

大数据解决方案实践

传统数据仓库解决方案



报表工具选型

REPORTING



Data Analysis

Data visualization

Adhoc Query/report

Adhoc Report方案



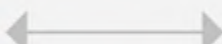
Application Express



Web Browser



Oracle Database
with APEX



Local Data
Source

Web Services



Remote Data
Sources and
Services

Database Link



Enterprise Data
Sources and
Services



Application Express

Haier

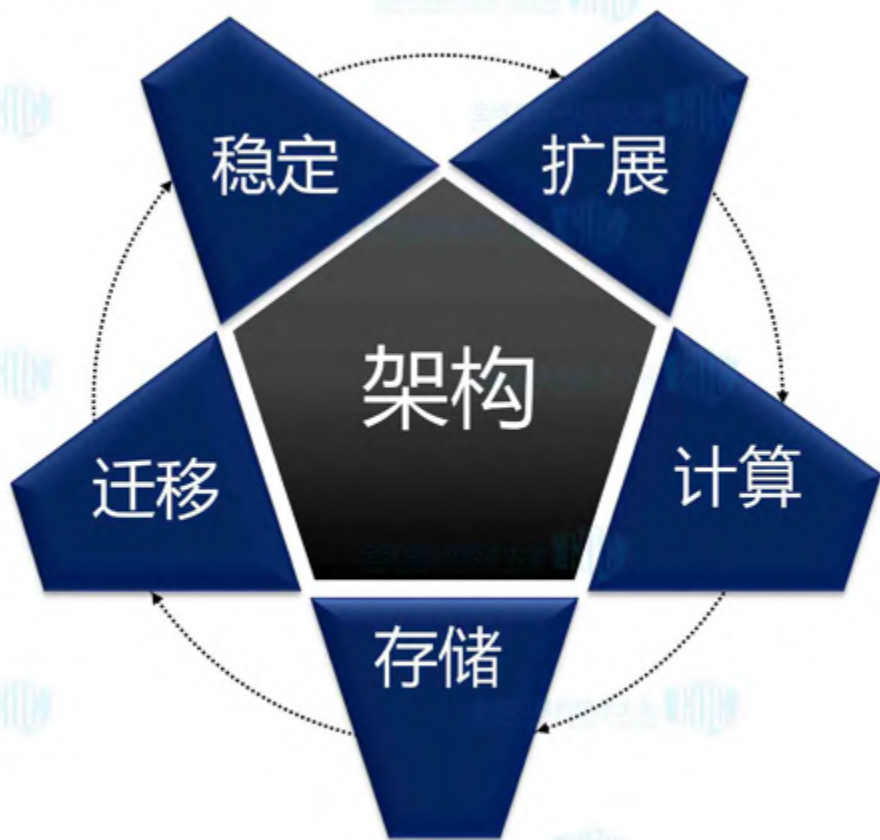
- SQL and PL/SQL
- Declarative framework , platform as a service
- Agile development
- Browser-based IDE
- Short learning curve
- No additional cost

The image displays three overlapping screenshots of the Oracle Application Express (APEX) interface. The top-left screenshot shows the 'Work by Developer' workspace with a project tree on the left and configuration panels on the right. The bottom-left screenshot shows the 'Dashboard' with a donut chart for 'Check Top Objects' (Applications: 44, Pages: 2,998, Packages/Applications: 1, Webpages: 8) and a line chart for 'Page Views' showing usage over time. The right-side screenshot shows a log viewer for 'v\$session' with columns for ID, Username, Session, Message, Level, and Check, listing various session events.

数据累积、维度爆炸带来的问题

- Performance
- Cost
- Scalability

架构升级考虑



系统整合

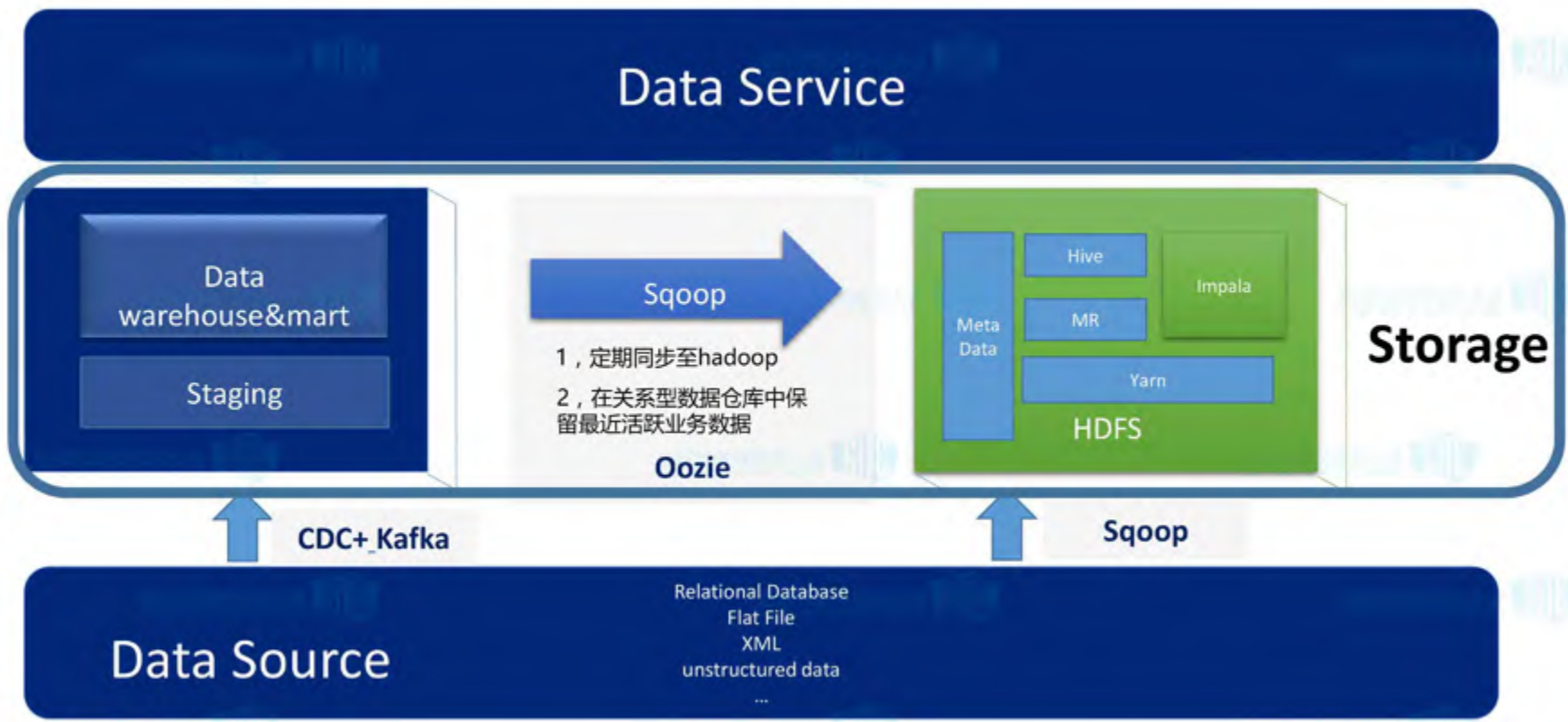
1, File system layout of HDFS

2, Batches or real-time ingestion

3, File format

4, Partition

引入Impala, 实时查询引擎

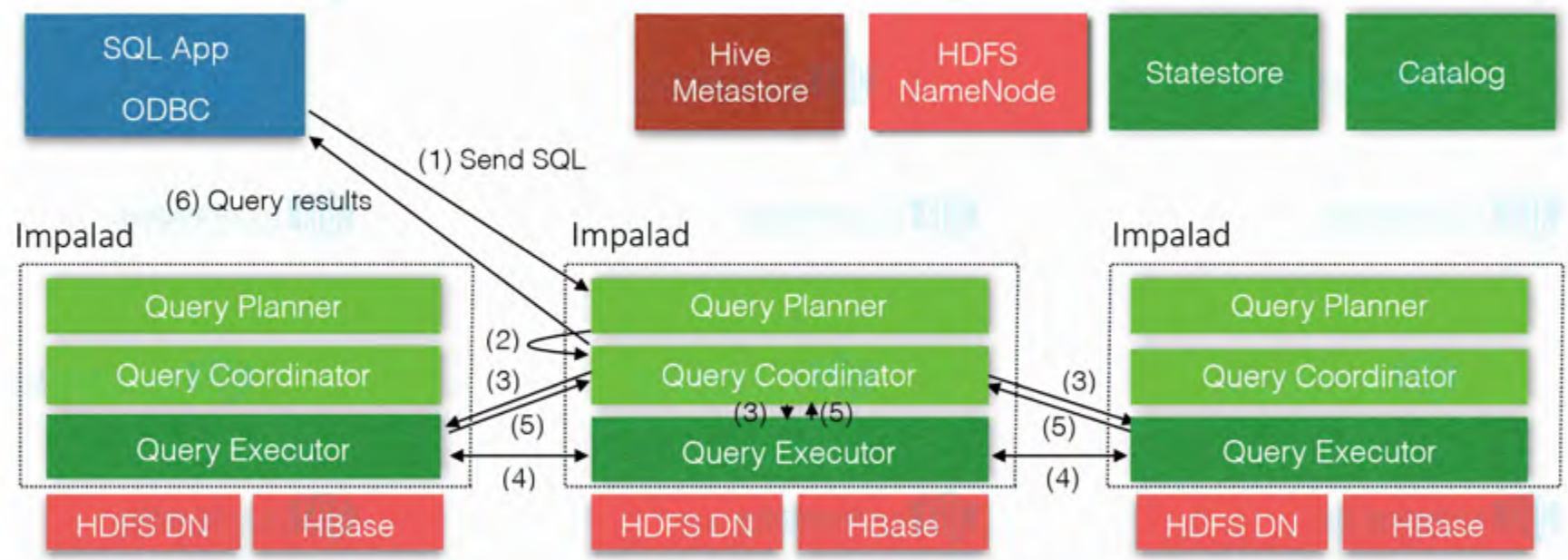




Why Impala?

- 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



Impala性能测试

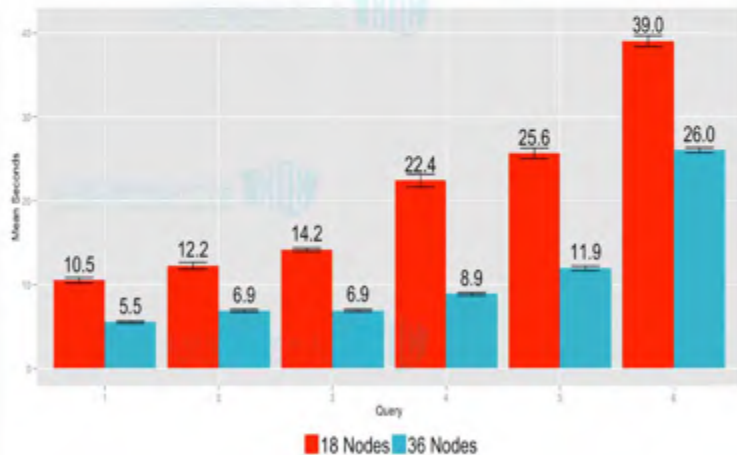
TEST SETUP: 15 TB TPC-DS data set

18 nodes vs 36 nodes

Latency

2x the hardware

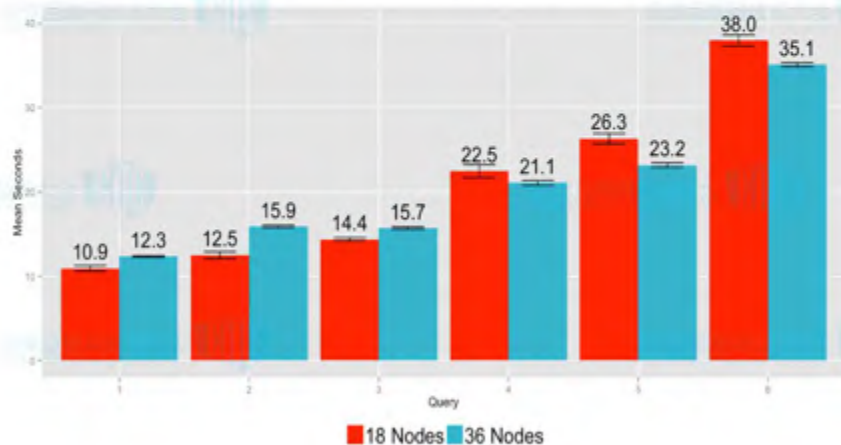
Expectation: cut response times in half



Concurrency

2x the users, 2x the hardware

Expectation: constant response times



Impala SQL性能考虑

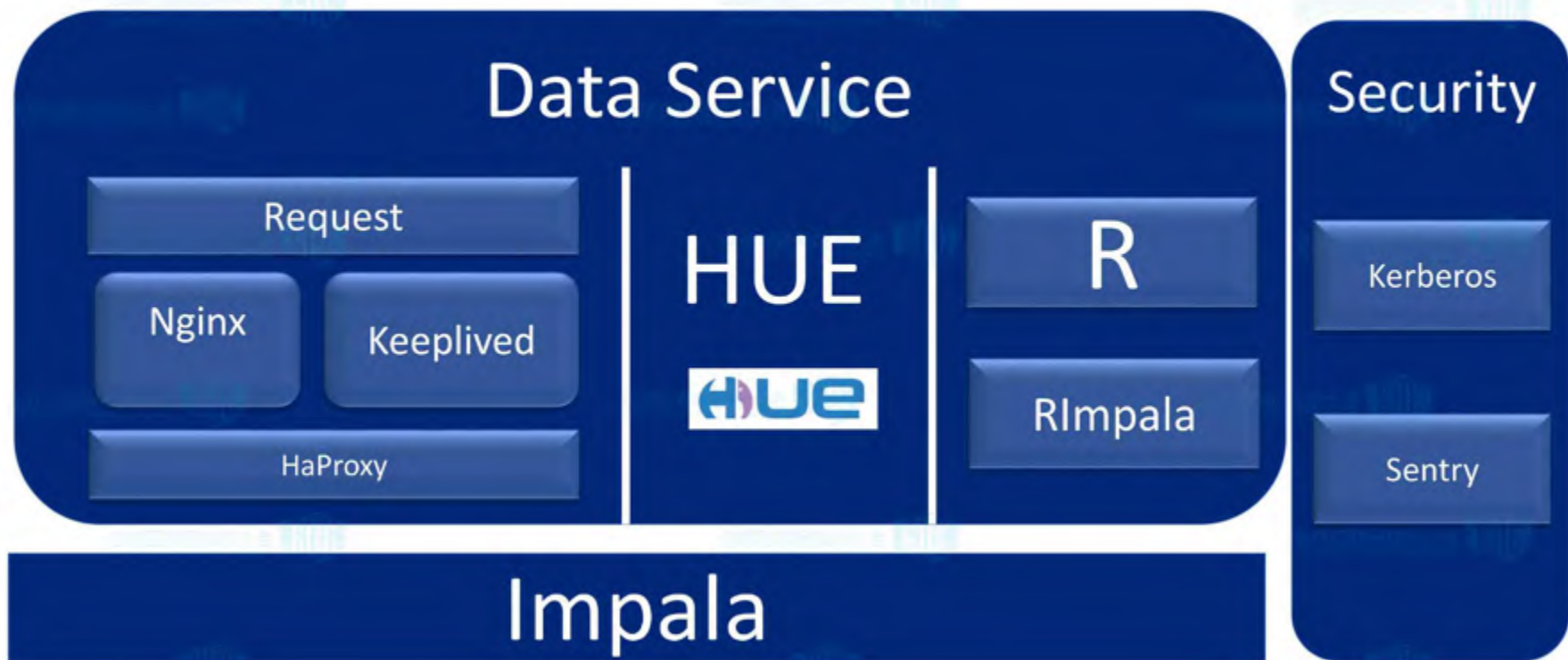
I/O

- 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.

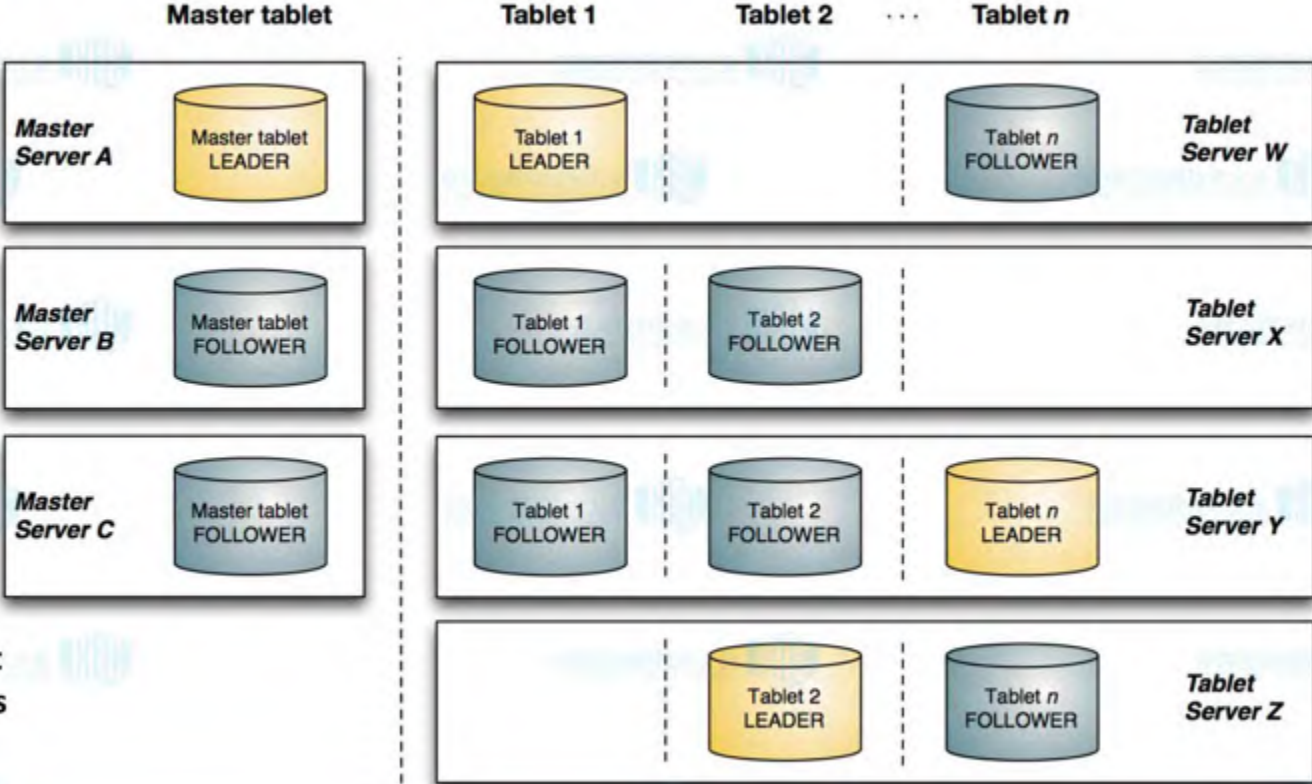
Data Service Interface



后续计划

- 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



a storage engine that enables fast analytics on fast data



探索Kudu

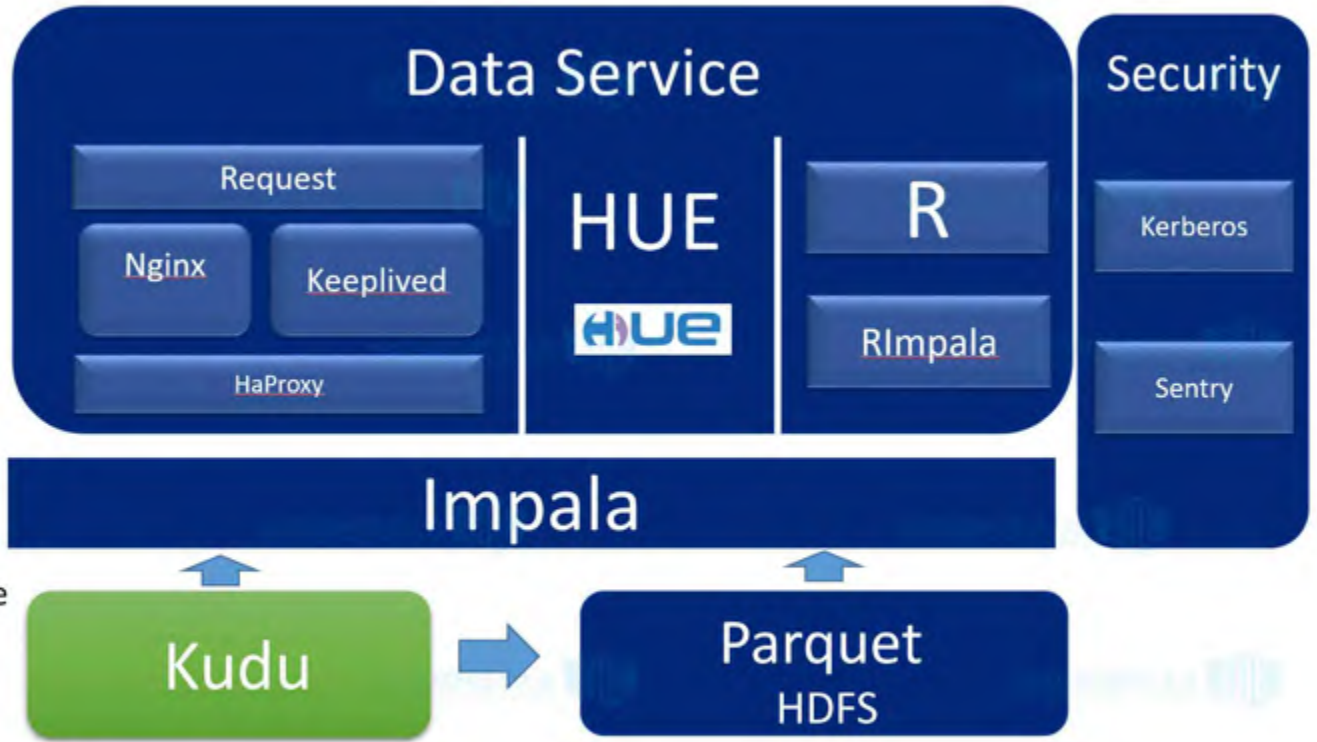
Kudu

a storage engine that enables fast analytics on fast data

Source

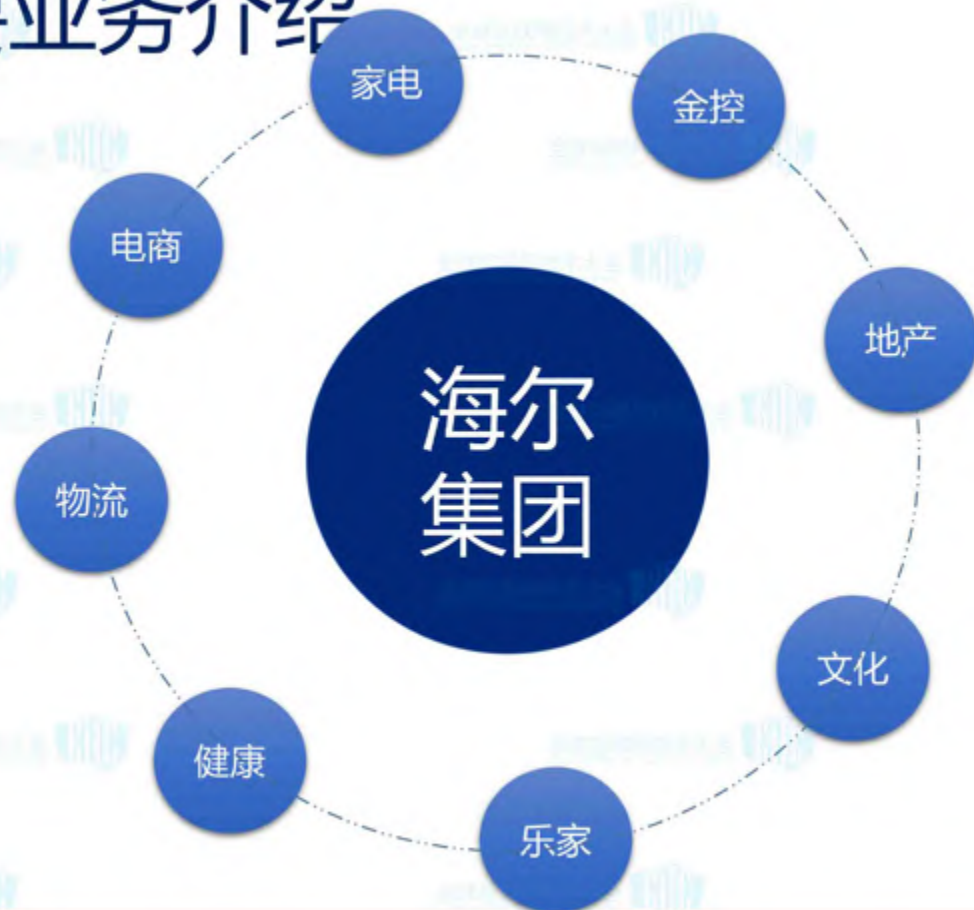
- Relational Database
- Flat File
- XML
- ...Unstructured data
- ...

Real-time

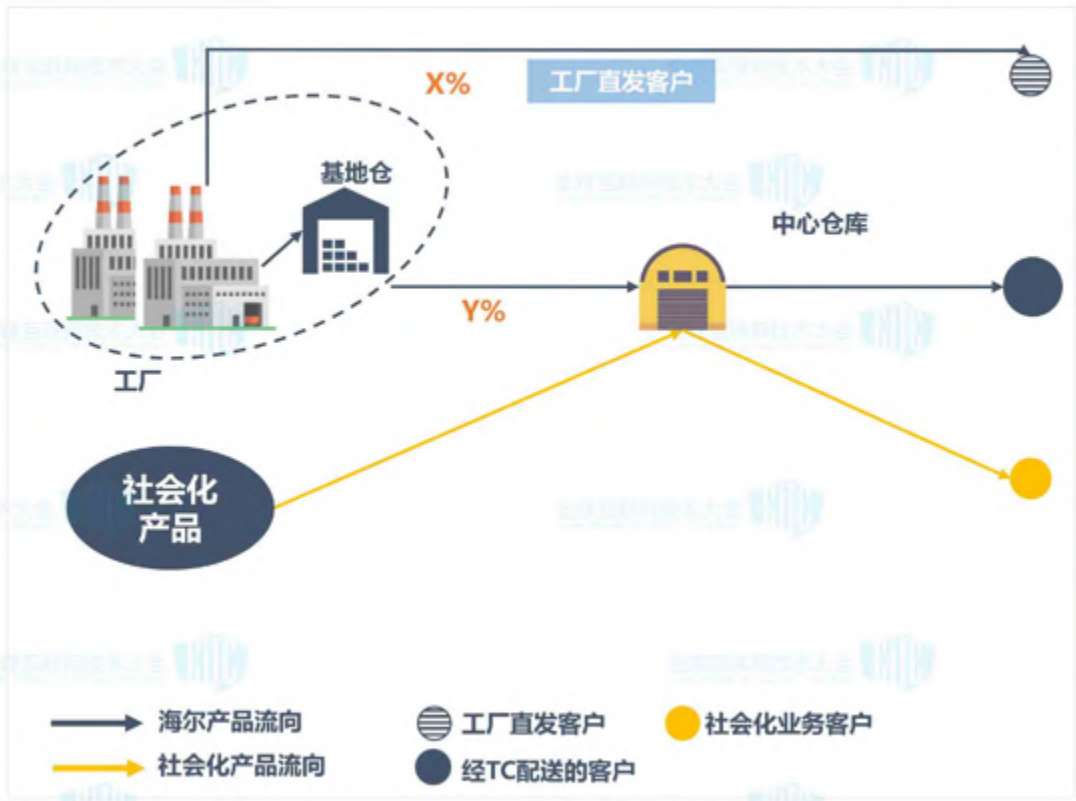


运营案例分享

海尔主要业务介绍



物流业务



Activity Based Classification

•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

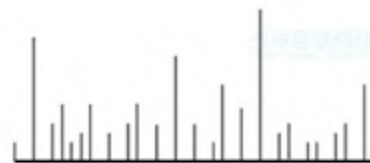
Activity Based Classification



Non-intermittent
持续



Smooth (fast)
流动快稳定



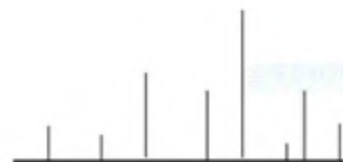
Erratic
流动快不稳定



Intermittent
间歇



Slow
流动慢



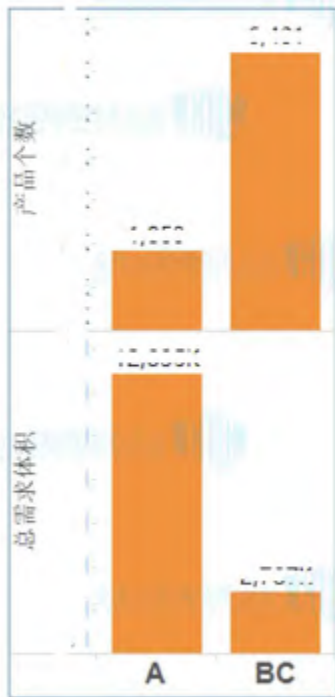
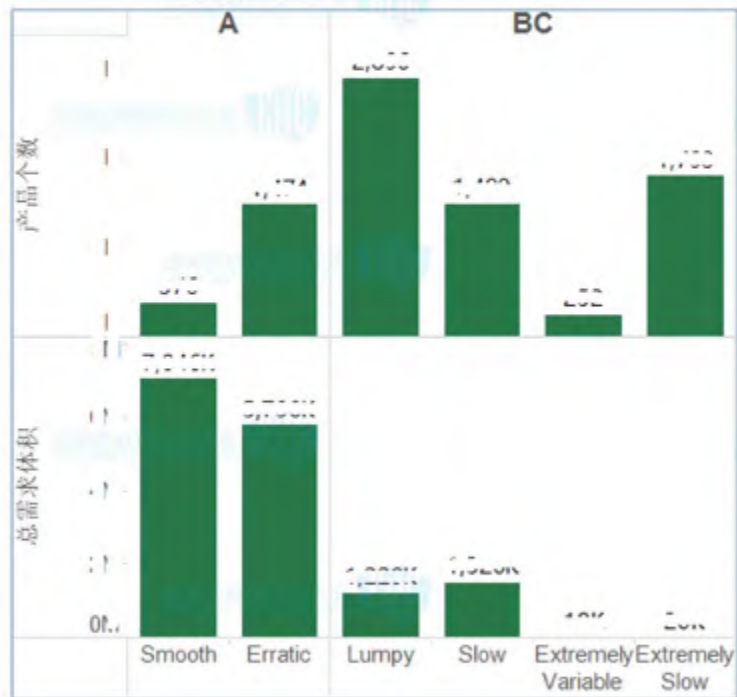
Lumpy
流动慢不稳定

Activity Based Classification



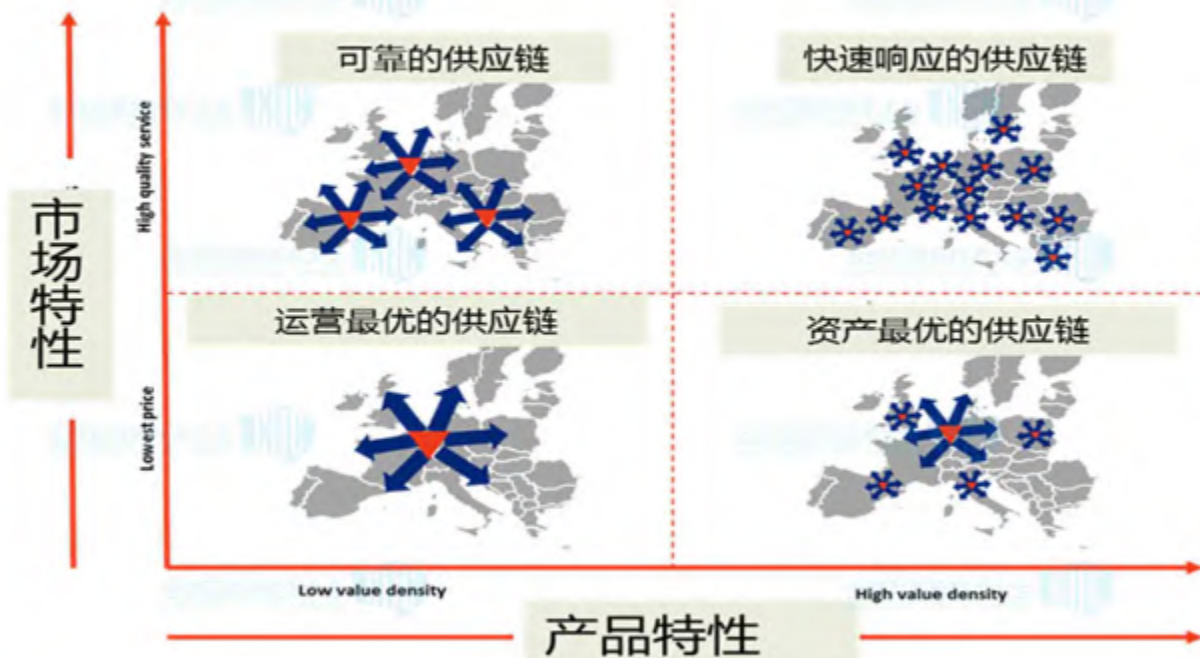
Demand Class	说明	产品分类
Extremely Slow	流动极端缓慢	BC
Smooth	流动快稳定	A
Erratic	流动快不稳定	A
Slow-Low Variable	流动慢-低波动	BC
Slow-Highly Variable	流动慢-高波动	BC
Lumpy	流动慢-不稳定	BC
Extremely Small	平均单量极小	BC
Extremely Variable	极端波动	BC

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

- 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