电商行业数据运营建设实践

兰亭集势 王庆恒

DTCC

2015中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2015

大数据技术探索和价值发现

提纲

- 一、对BI的思考
- 二、兰亭数据平台建设
- 三、大数据的高性能实现
- 四、高效的数据挖掘
- 五、BI的发展趋势

一、对BI的思考_定义

BI组成

- DW
- OLAP
- DM
- 查询报表
- 数据挖掘

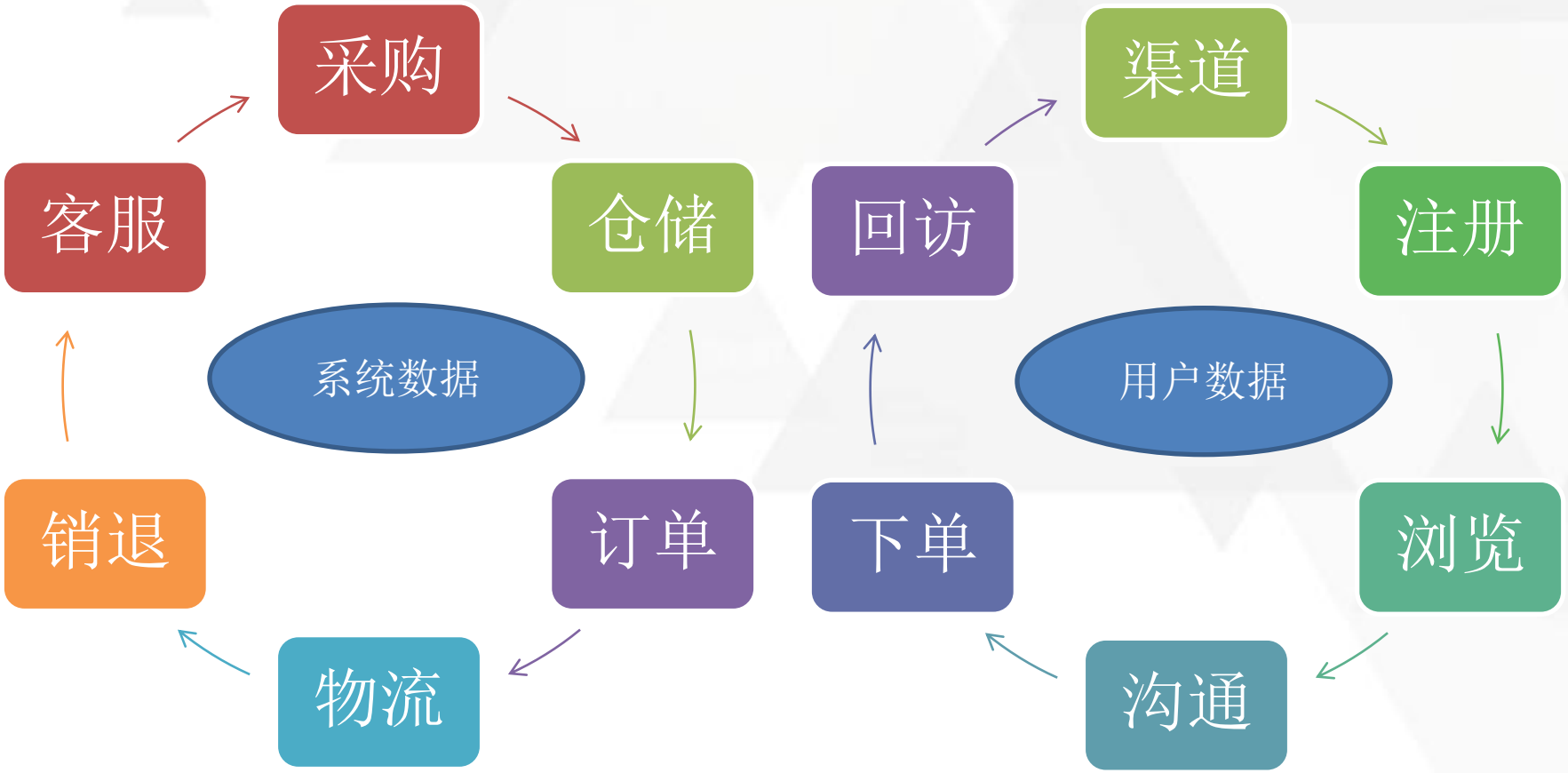
BI功能

- 信息搜集
- 管理
- 分析
- 处理

BI目的

- 知识(Knowledge)
- 洞察力(Insight)
- 辅助做出决策

一、对BI的思考_数据



一、对BI的思考_协作

一、对BI的思考_问题

响应较慢

- 开发速率跟不上互联网变化的节奏；
- 需求变化，又得根据流程重新开发；

灵活度不够

- 不能进行数据交互、自由组合；
- 不能二次分析；

无法支持大数据量

- TB,PB级数据无法响应

一、对BI的思考_解决思路

快速响应

- 快速响应：实时数据仓库和准实时数据仓库；
- 需求变化：最快响应；

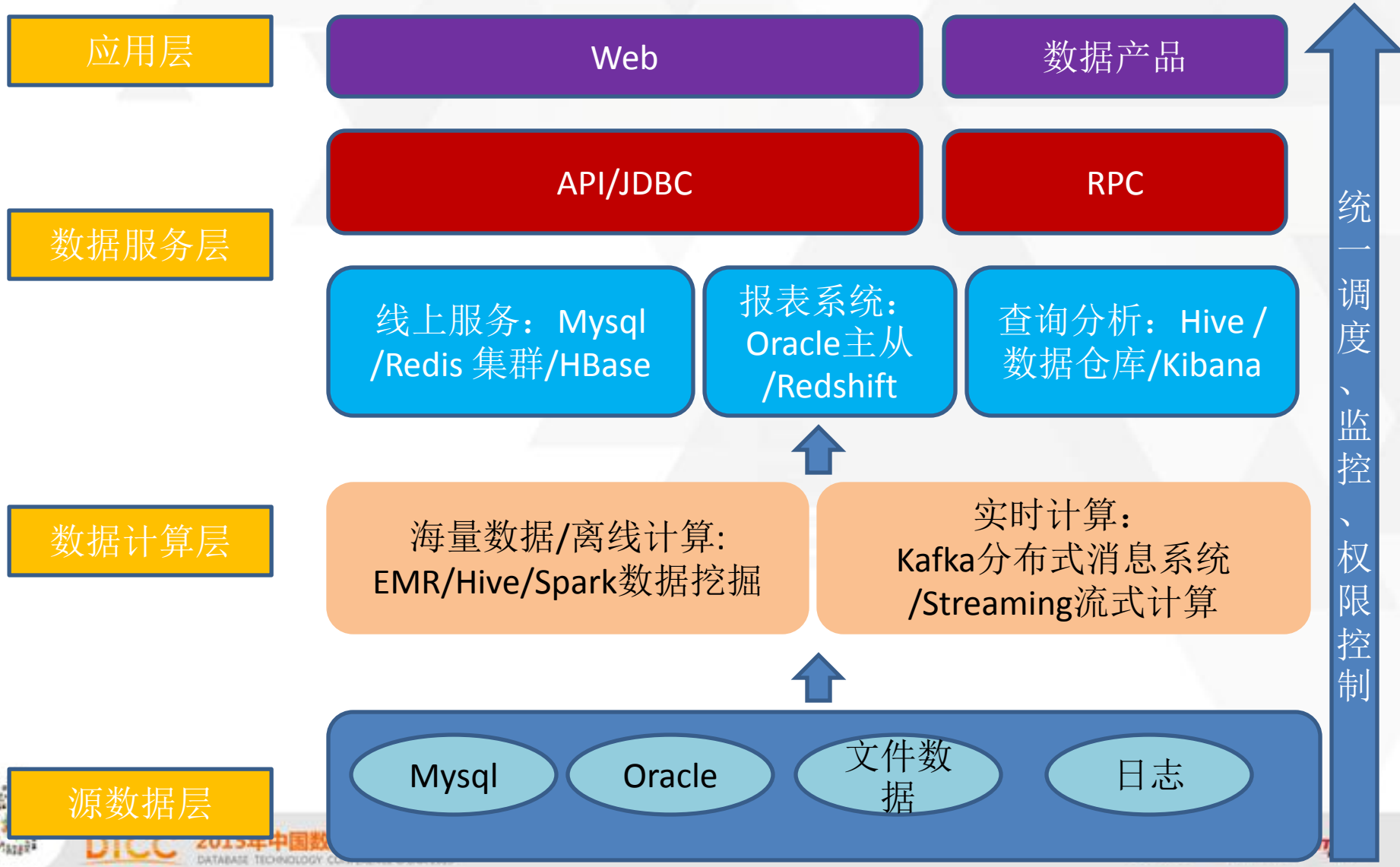
灵活多变

- 人与系统交互、维度自由组合；
- 多样的展现方式

基于云计算的大数据响应

- 海量数据处理：海量业务数据在线分析、云端部署

二、兰亭数据平台建设_架构



二、兰亭数据平台建设

1、数据

- 模型分层：ODS, DW, DM, RPT
- 范围：流量、销售、运营三大数据模型及数据集市
- 推荐：个性化推荐数据、商品推荐数据
- 排序：根据能效值排序的数据

2、架构

- 调度：基于Quartz开发的调度系统；
- 监控：判断程序执行返回状态，记录到日志表中，发送短信和邮件报警；
- 权限控制：数据仓库访问权限，申请、审批和授权；
- 线上支持：MySQL，HBase通过API接口支持线上服务

二、兰亭数据平台建设

2、架构

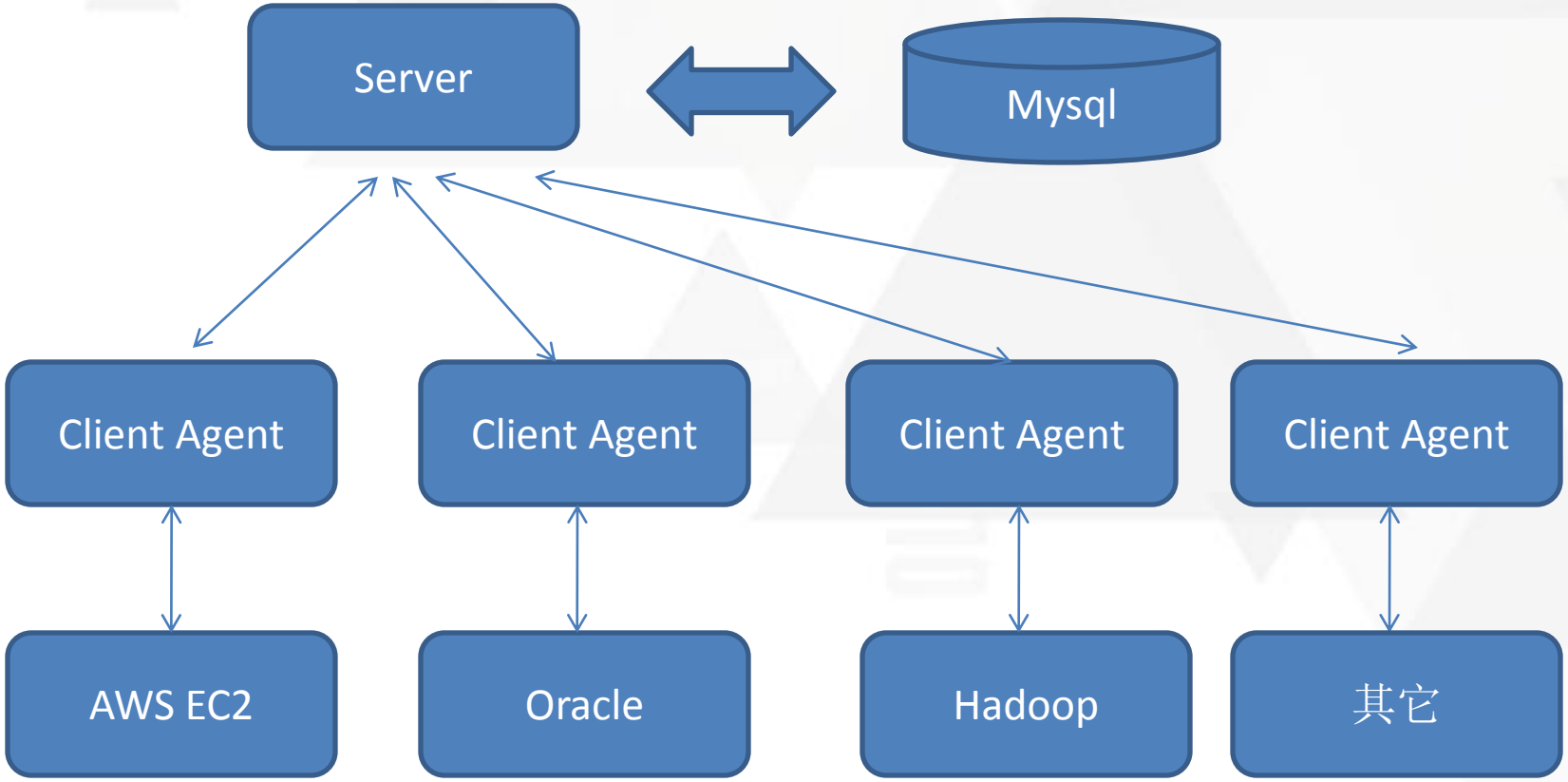
- 数据收集：DataExp ,shell 脚本，Kafka分布式消息系统，快速低成本收集日志等；
- 抽取与转换：存储和计算基于EMR/Hive/Spark/Oracle
- 存储：Oracle数据仓库存储量级较小的数据，Redshift存储量级大的数据
- 云存储：Amazon S3
- Redis集群：支持实时推荐等

3、应用

- 日常查询：临时查询；
- 报表服务：报表以web展示，或者邮件形式发送给用户；
- 自助服务：支持用户自定义查询，并保存为报表；
- 网站运营效果评估：响应时间、各指标用户数等；
- 流量分析：流量、点击率、转化率、跳出率等
- 推荐系统
- 搜索引擎

二、兰亭数据平台建设

- 统一调度系统：架构



二、兰亭数据平台建设

- 调度系统：运维、管理、监控、依赖

Job Configration

Basic | Alarm | Job Depends

10 records per page Search:

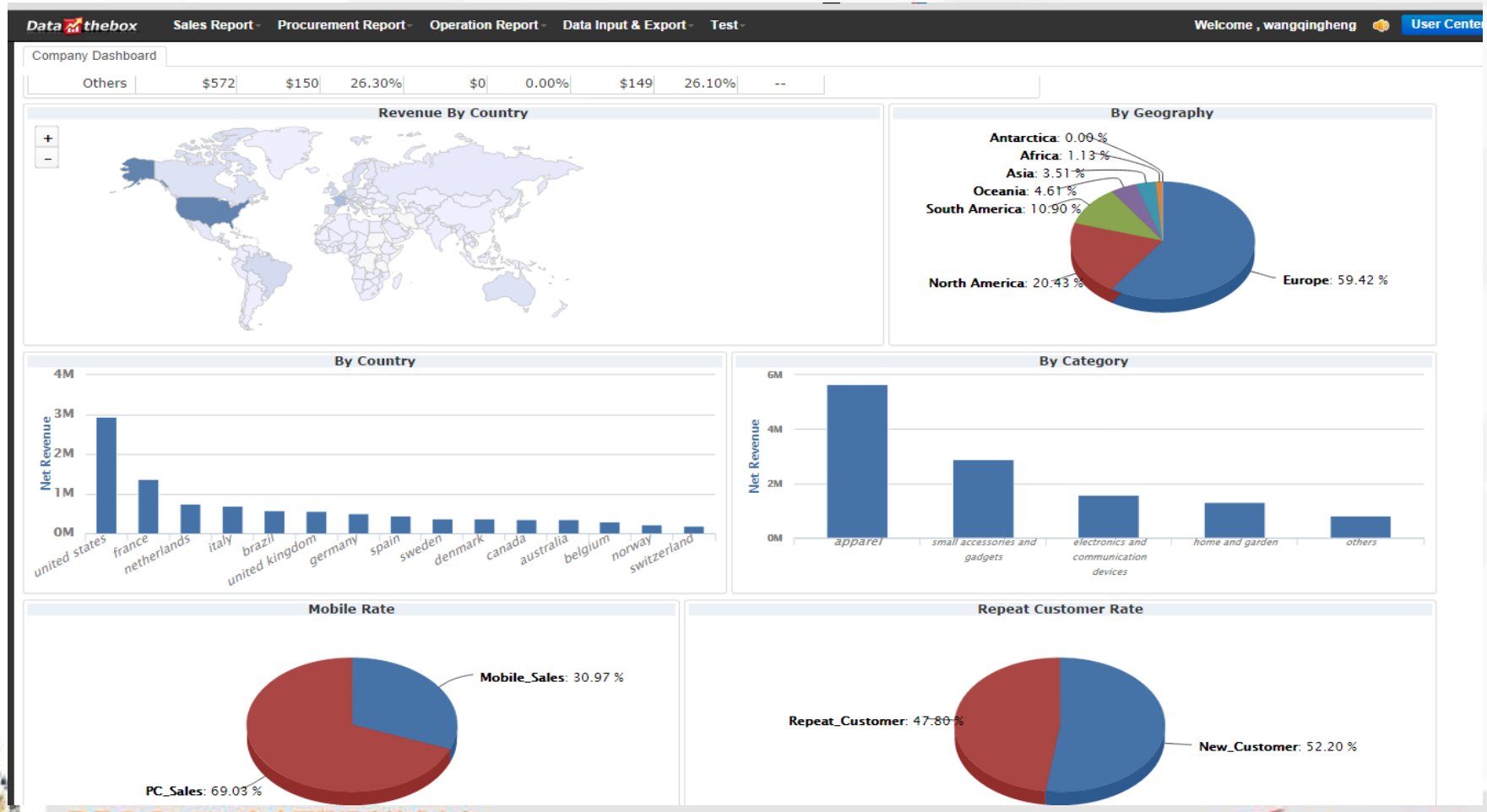
Job Id	Job Name	Job Type	Cron Expression	Script
16	clickstream_base_s3_litb_daily	定时	0 50 0 * * ?	/bin/sh /data1/bin/hive/bin/clickstream_base_s3/run_clickstream_base_s3_daily.sh litb ${BasicDate:-1:[yyyy-MM-dd]} ${BasicDate:-1:[yyyy-MM-dd]}
47	load_dw_v3_orders_daily	定时	0 10 4 * * ?	/bin/sh /data1/bin/hive/oracle/dw_v3_orders/run_load_dw_v3_orders_daily.sh ${BasicDate:-1:[yyyy-MM-dd]} ${BasicDate:-1:[yyyy-MM-dd]}
251	cid_products_refresh_mini	依赖		python /data1/bin/hive/bin/cid_products/run_cid_products_refresh_mini.py ${BasicDate:-1:[yyyy-MM-dd]}
228	load_dw_product_sales_cc_daily	定时	0 40 2 * * ?	/bin/sh /data1/bin/hive/oracle/dw_product_sales_cc/run_load_dw_product_sales_cc_daily.sh ${BasicDate:-1:[yyyy-MM-dd]} ${BasicDate:-1:[yyyy-MM-dd]}
198	cptree_update_drop_daily	依赖		sh /data1/bin/traffic_analyst/bin/run_cptree_update_daily.sh drop
176	clickstream_mobile_base_s3_litb_m	定时	0 5 0 * * ?	sh /home/nitb-bak/mobile_traffic/dailytraffic/run_clickstream_mobile_base_s3_history.sh litb_m ${BasicDate:-1:[yyyy-MM-dd]} ${BasicDate:-1:[yyyy-MM-dd]}
157	BadIPListJob_mini_api	依赖		sh /data1/bin/traffic_analyst/bin/run_apitraffic_daily.sh mini_api ${BasicDate:-1:[yyyy-MM-dd]} badip
138	ContentVisitJob_mini	依赖		sh /data1/bin/traffic_analyst/bin/run_pctraffic_daily.sh mini ${BasicDate:-1:[yyyy-MM-dd]} contentvisit
121	BadIPListJob_mini	依赖		sh /data1/bin/traffic_analyst/bin/run_pctraffic_daily.sh mini ${BasicDate:-1:[yyyy-MM-dd]} badip ${Master}
100	Load_ppv_litb	依赖		ssh hadoop1 hadoop jar /mnt/program/traffic_analyzer/traffic-analyzer-0.1.jar com.litb.bi.trafficanalyzer.pc.daily.WAStorer litb ppv ${BasicDate:-1:[yyyy-MM-dd]}

Showing 1 to 10 of 190 entries ← Previous | 1 | 2 | 3 | 4 | 5 | Next →

Previous Finish

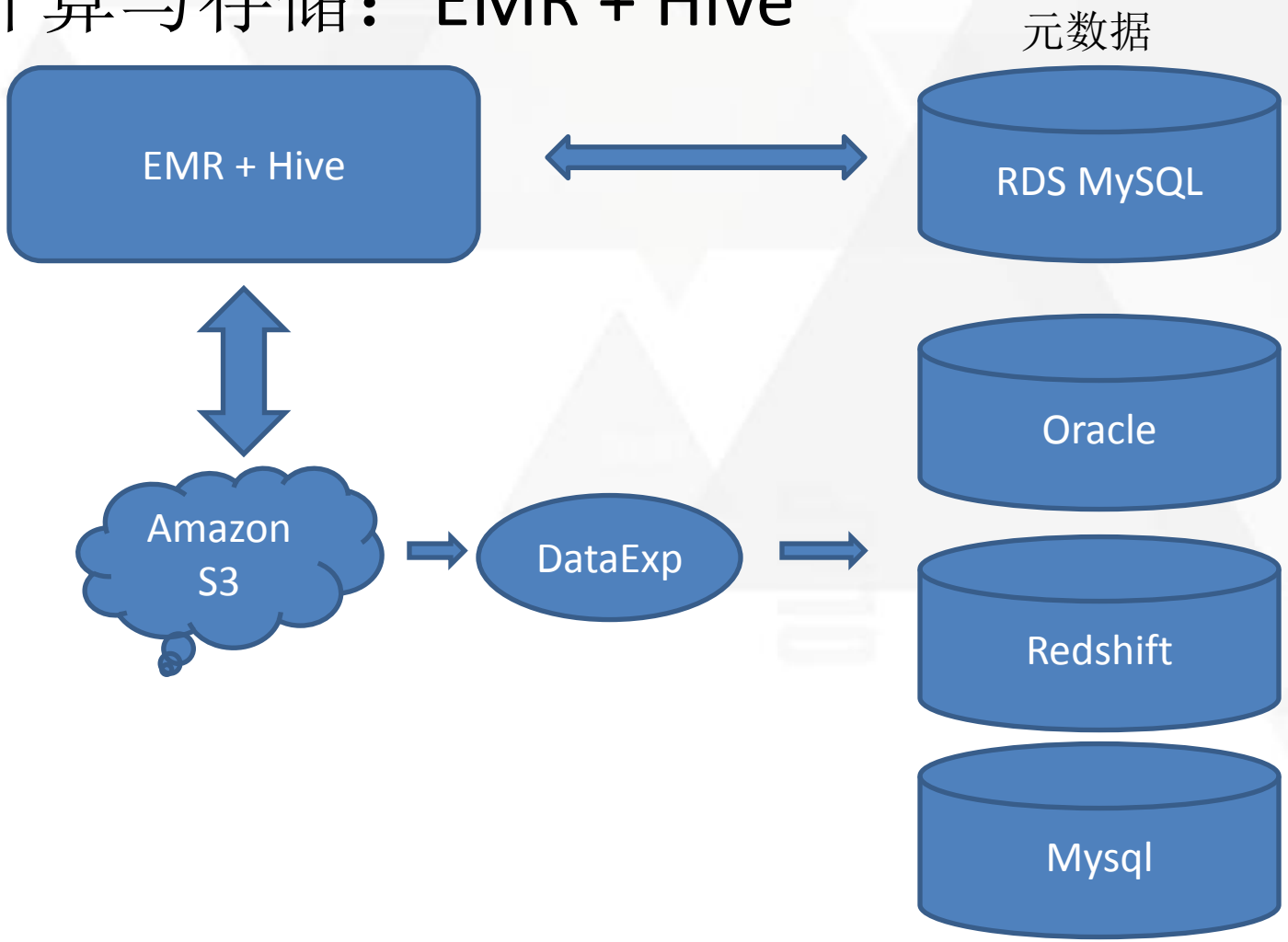
二、兰亭数据平台建设

- 报表系统：维度分析、动态展现、权限



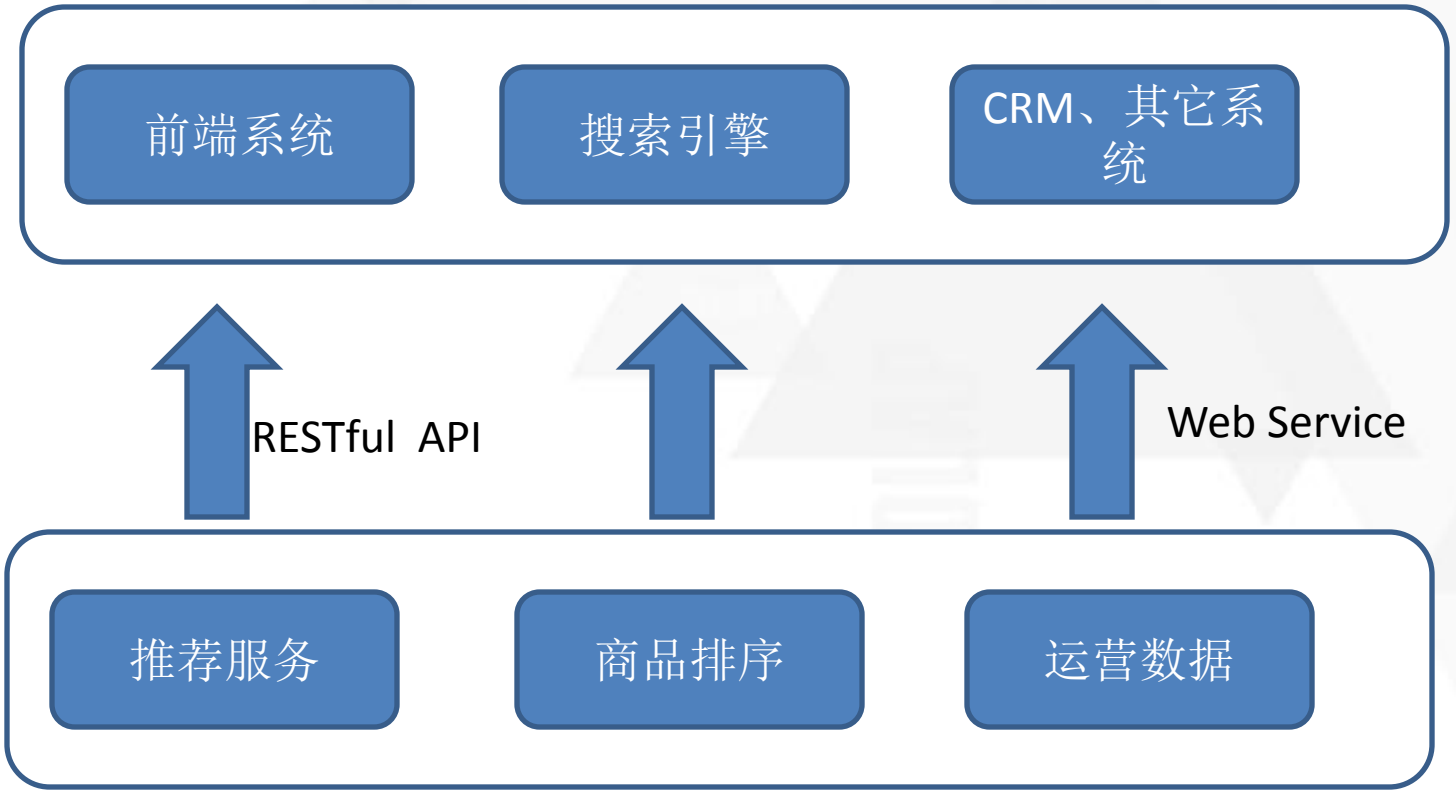
二、兰亭数据平台建设

- 计算与存储：EMR + Hive



二、兰亭数据平台建设

- 应用



三、大数据的高性能实现

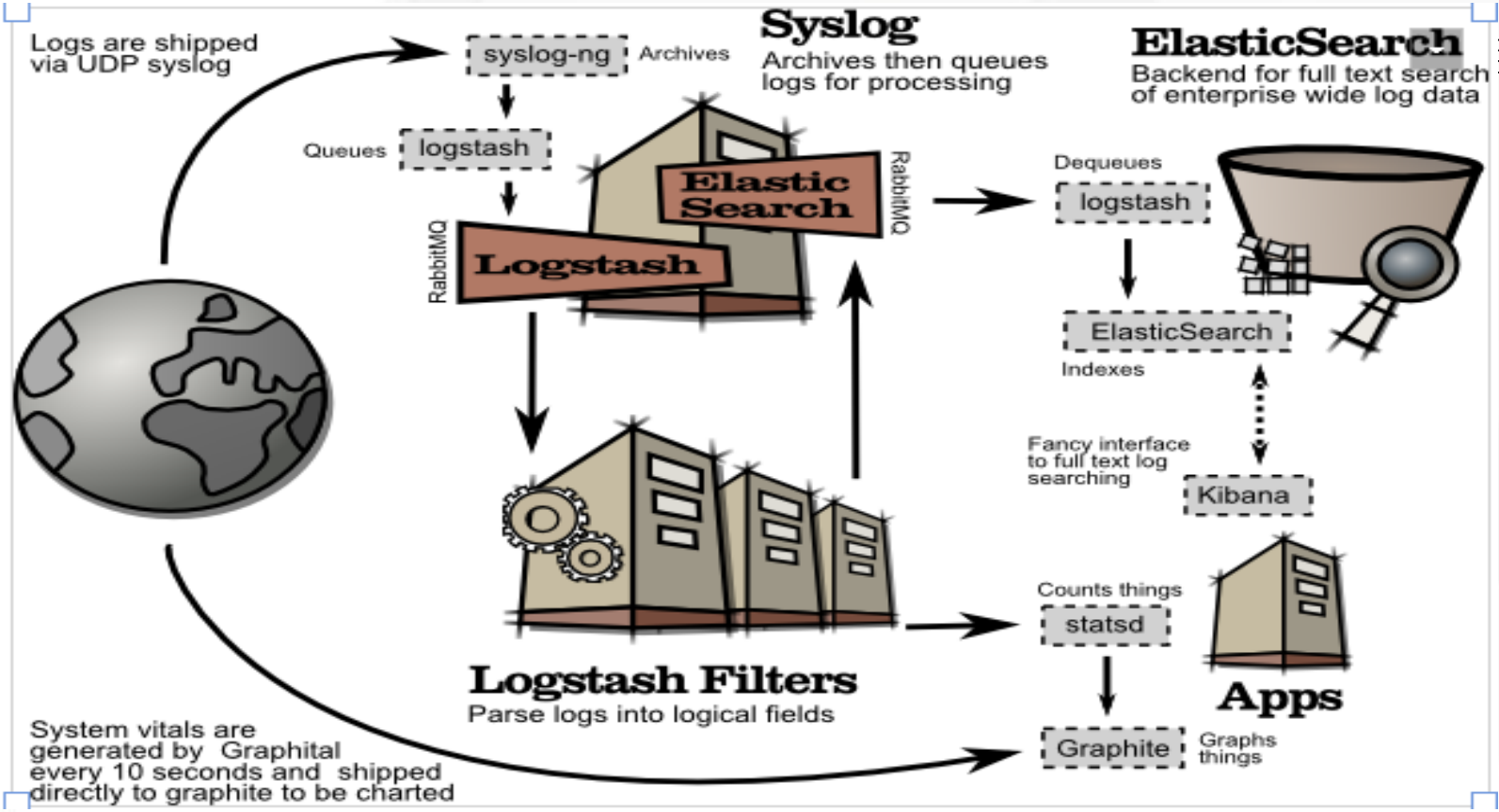
- 1、Oracle读写分离
  - 一拖二
  - 主库写，从库实时复制
  - 多从库用于查询、统计
  - 硬件加速：SSD盘加速

三、大数据的高性能实现

- 2、Redshift
  - Amazon Redshift 是一种快速、完全托管的 PB 级数据仓库解决方案;
  - 列式存储;
  - 数据压缩;
  - 区域映射;
  - MPP并行处理框架:在不停机的情况下实现扩展或收缩;

三、大数据的高性能实现

- 3、Logstash、 ElasticSearch、 Kibana



四、高效的数据挖掘

1、 Hadoop在数据挖掘中的问题

- MP模式 vs复杂的机器学习算法
- 多次迭代问题
- 中间数据的处理
- 开发周期长

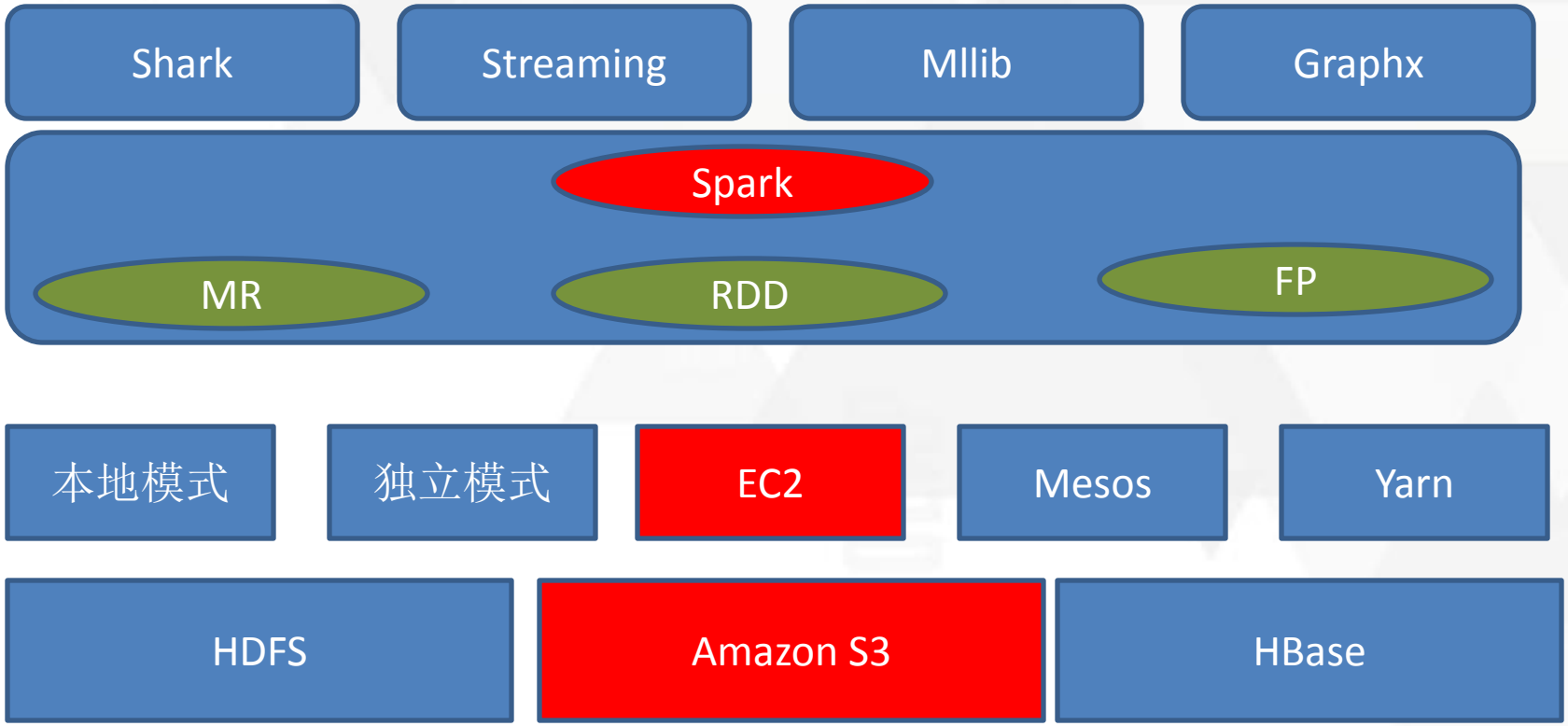
四、高效的数据挖掘

2、Spark

- RDD（**Resilient Distributed Dataset**）
  - ✓内存计算:计算的中间结果保存在内存中，不需要读写HDFS;
  - ✓快速迭代;
  - ✓DAG
- Scala
  - ✓函数式编程（FP）;
  - ✓Actor模型：基于线程和基于事件的Actor;
  - ✓并发能力;

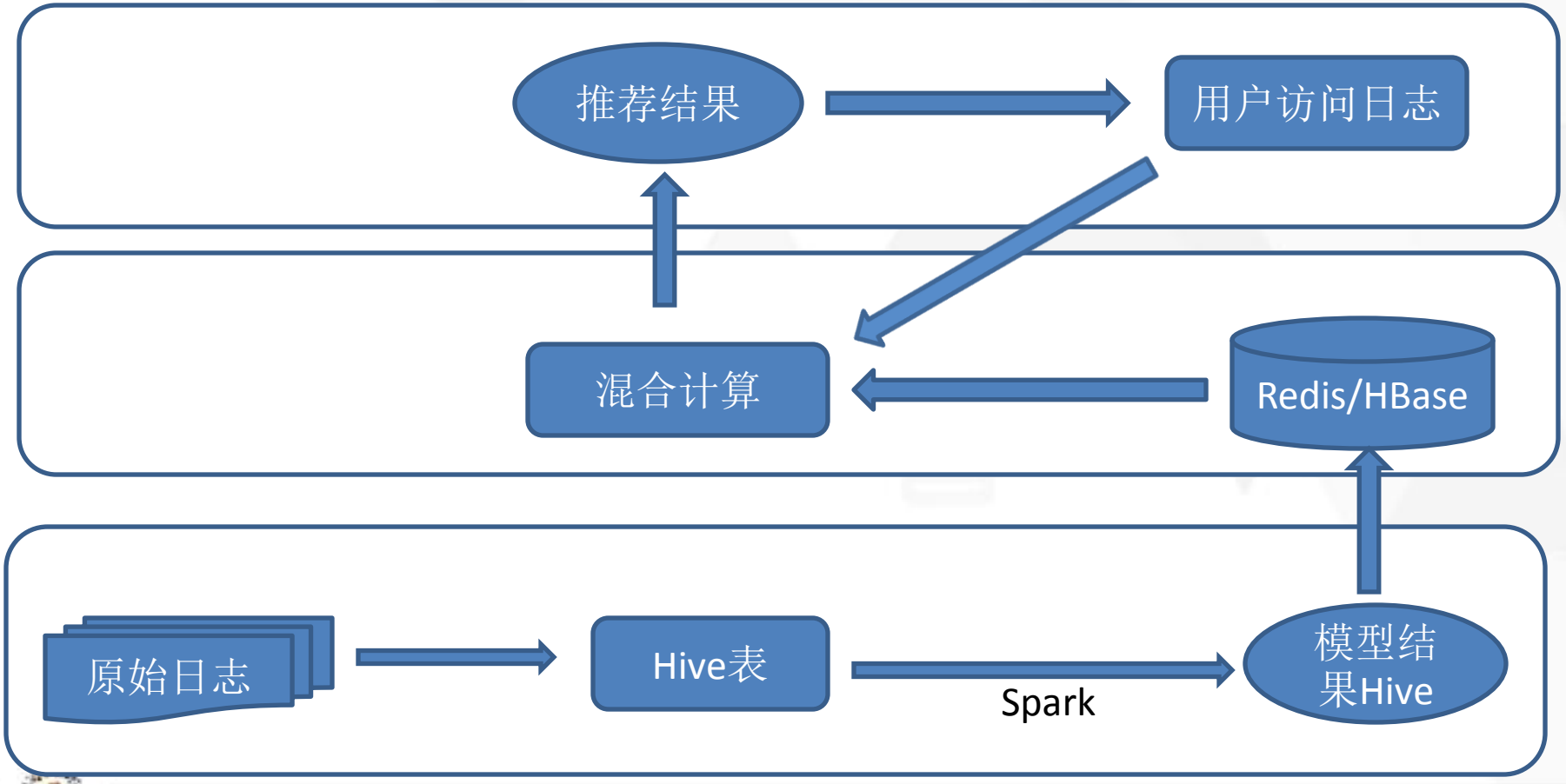
四、高效的数据挖掘

- 3、Spark架构



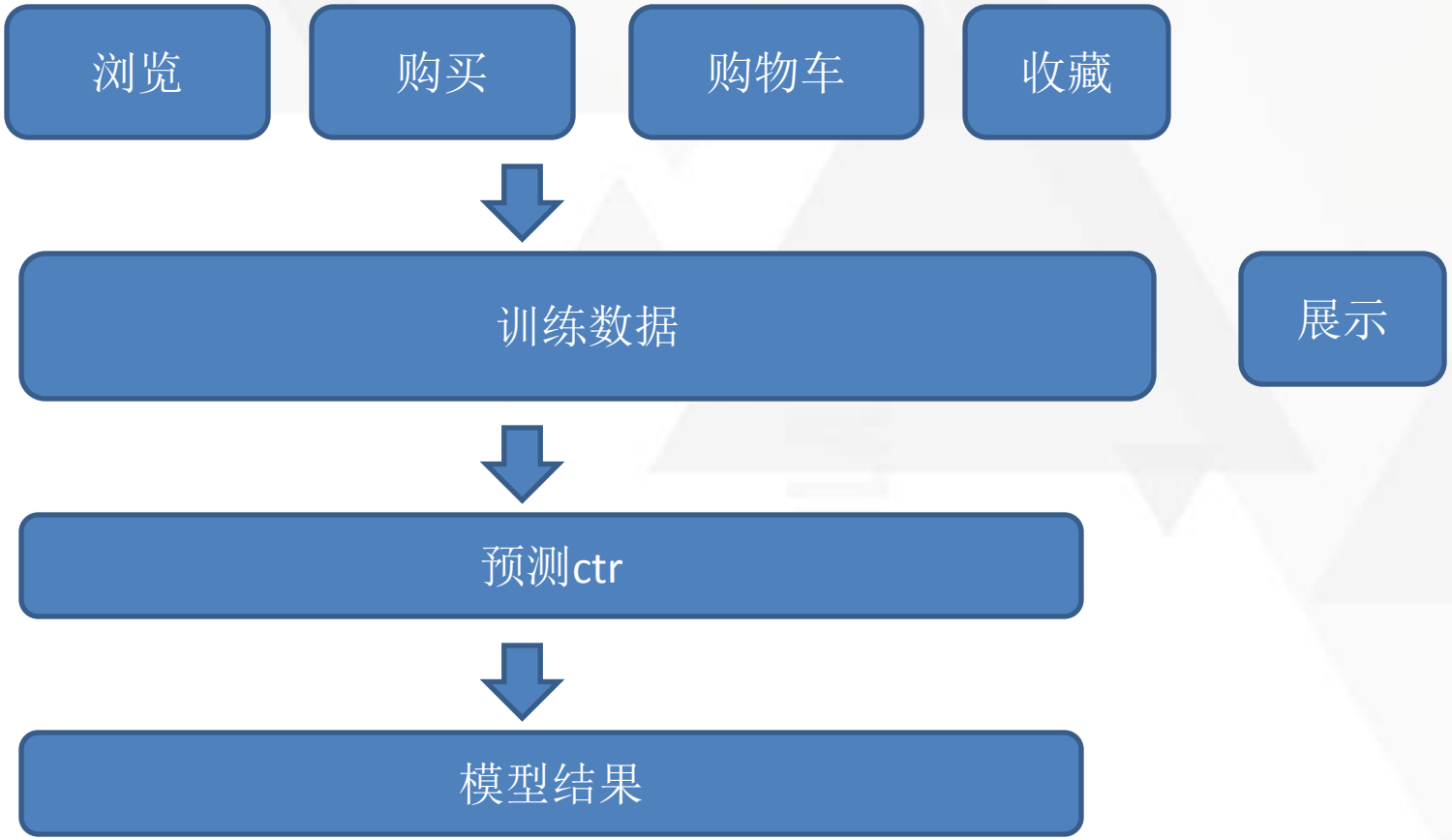
四、高效的数据挖掘

4、兰亭推荐系统架构



四、高效的数据挖掘

5、基于用户行为的商品推荐



五、BI的发展趋势

- 更加灵活的可定制商业智能；
- 更易用更人性化；
- 多样的展示方式；
- 移动BI；
- 云计算与云部署；
- 海量数据处理；

联系我们

- 姓名：王庆恒
- 邮箱：wangqingheng@lightinthebox.com


ChinaUnix
ITPUB
THANKS