DAMS

中国数据智能管理峰会

DATA & AI MANAGEMENT SUMMIT

万亿数据库核心存储引擎实现与应用

演讲人：母延年

演讲人简介-母延年

录信数软

职业生涯的前十年 任职于 新浪、阿里、腾讯

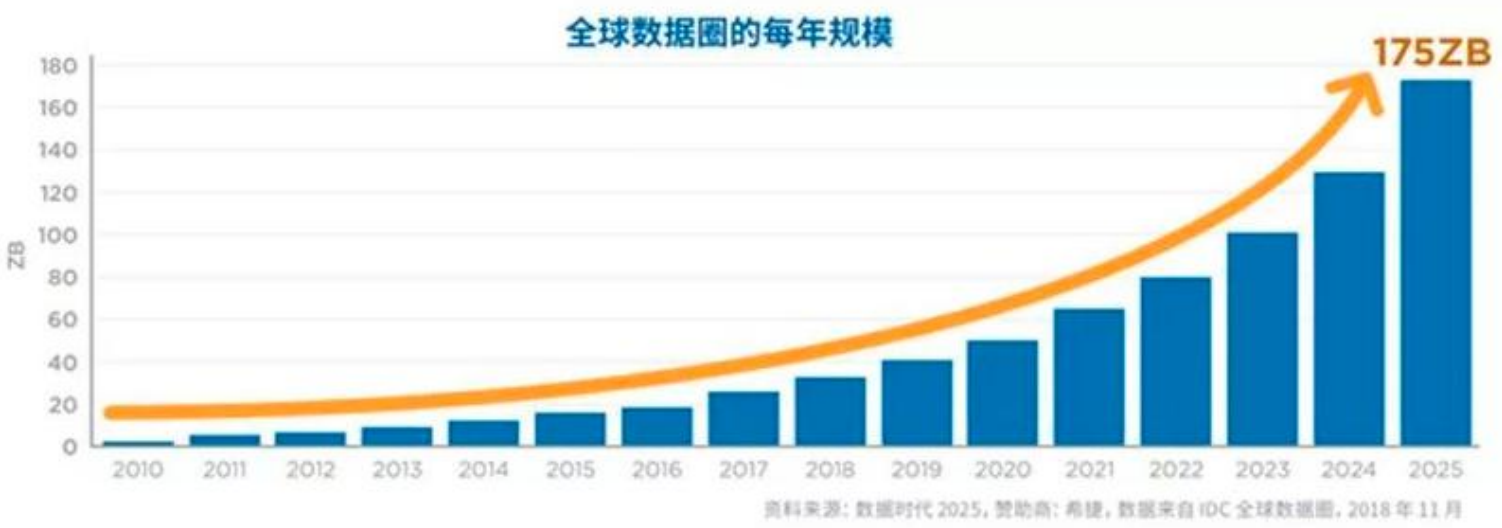
- 精通hadoop、lucene、storm等多种主流软件源码
- 阿里负责海狗（higo）、黄金策等多个项目的研发，开源2个项目，Mdrill、Jstorm
- 腾讯负责腾讯数据平台hermes项目开发，单日数据规模支撑3600亿+

最近五年 南京录信创始人&CTO

录信是lucene的中文发音，作为粉丝，以此命名公司

- 有200多套在生产系统使用超1年
- 其中1个系统数据规模过10万亿（1000+节点）
- 其中10+个系统数据规模过万亿（300+节点）
- 其中100+个以上系统数据规模过千亿
- 获12项技术发明专利

时代背景——数据爆炸式增长



据IDC发布《数据时代2025》的报告显示，全球每年产生的数据将从2018年的33ZB增长到175ZB，2025年全球每天产生的数据量将达到491EB。

行业痛点

一、大数据里的产品种类很多，但每个种类内都很单一。

1.绝大部分系统采用单一的”暴力扫描”,性能低下。

2.少量系统有索引但功能受限只能KV,或全文检索,做不了复杂的统计。

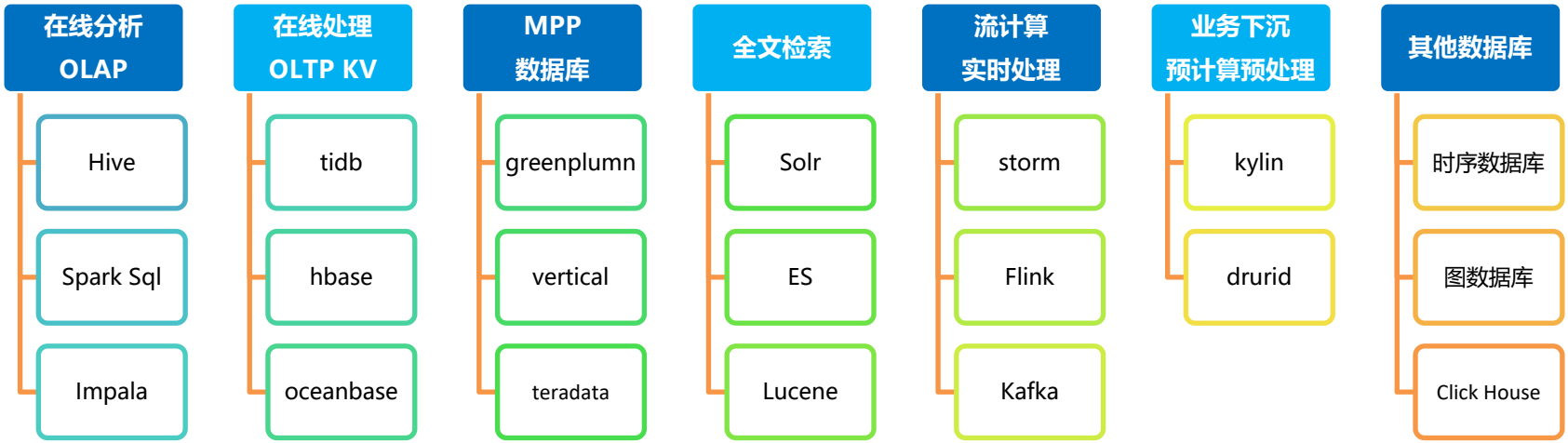
3.还有一部分系统采用预计算处理,不灵活也不能查看原始明细数据。

二、为了应对这些不完善,需要混合使用多种系统

1.数据存储的份数太多，浪费存储资源。

2.多个系统之间数据互通很难。

3.每个系统接口都不一样，学习与维护成本很高。

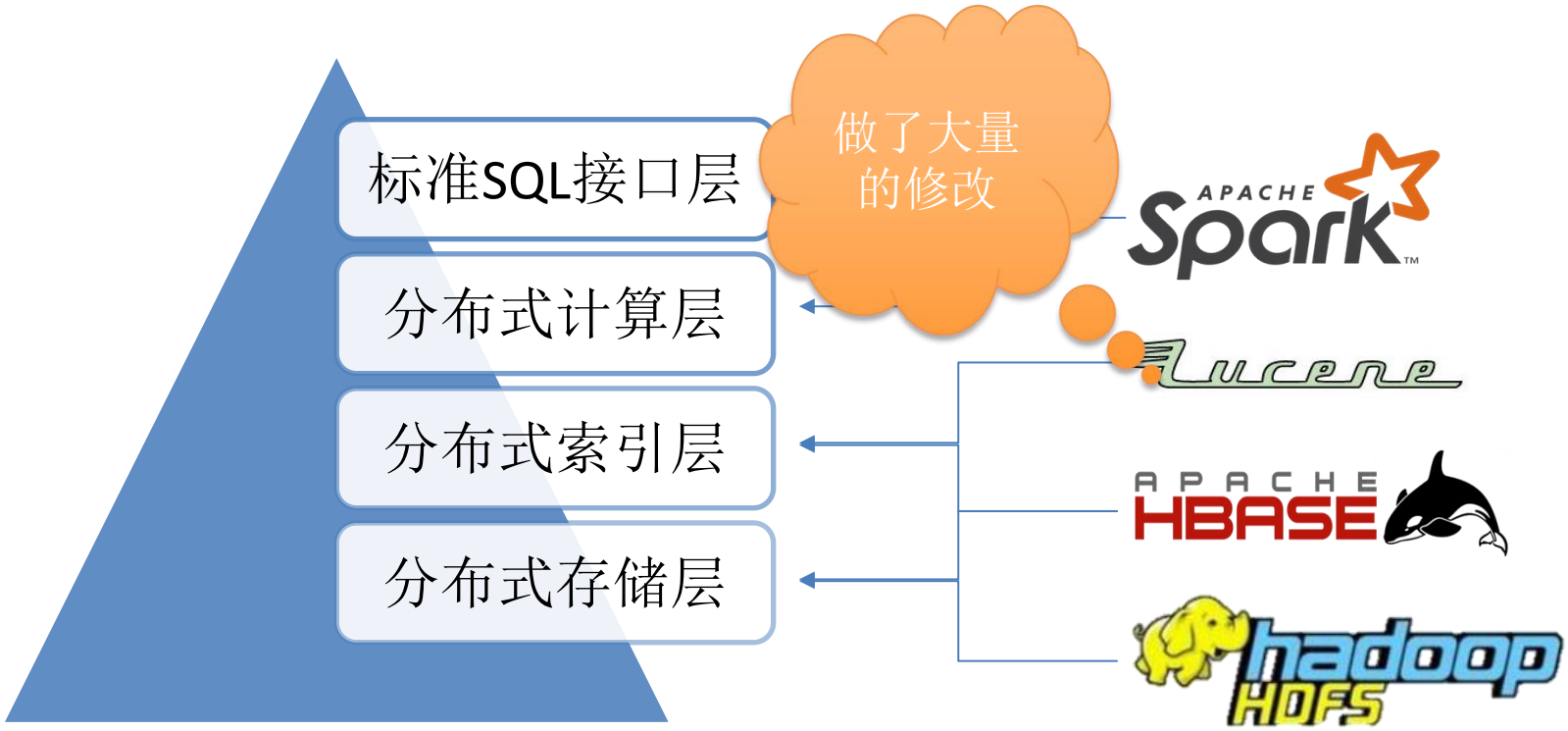


能不能将他们整合起来
一份数据、 一套系统、 一种接口

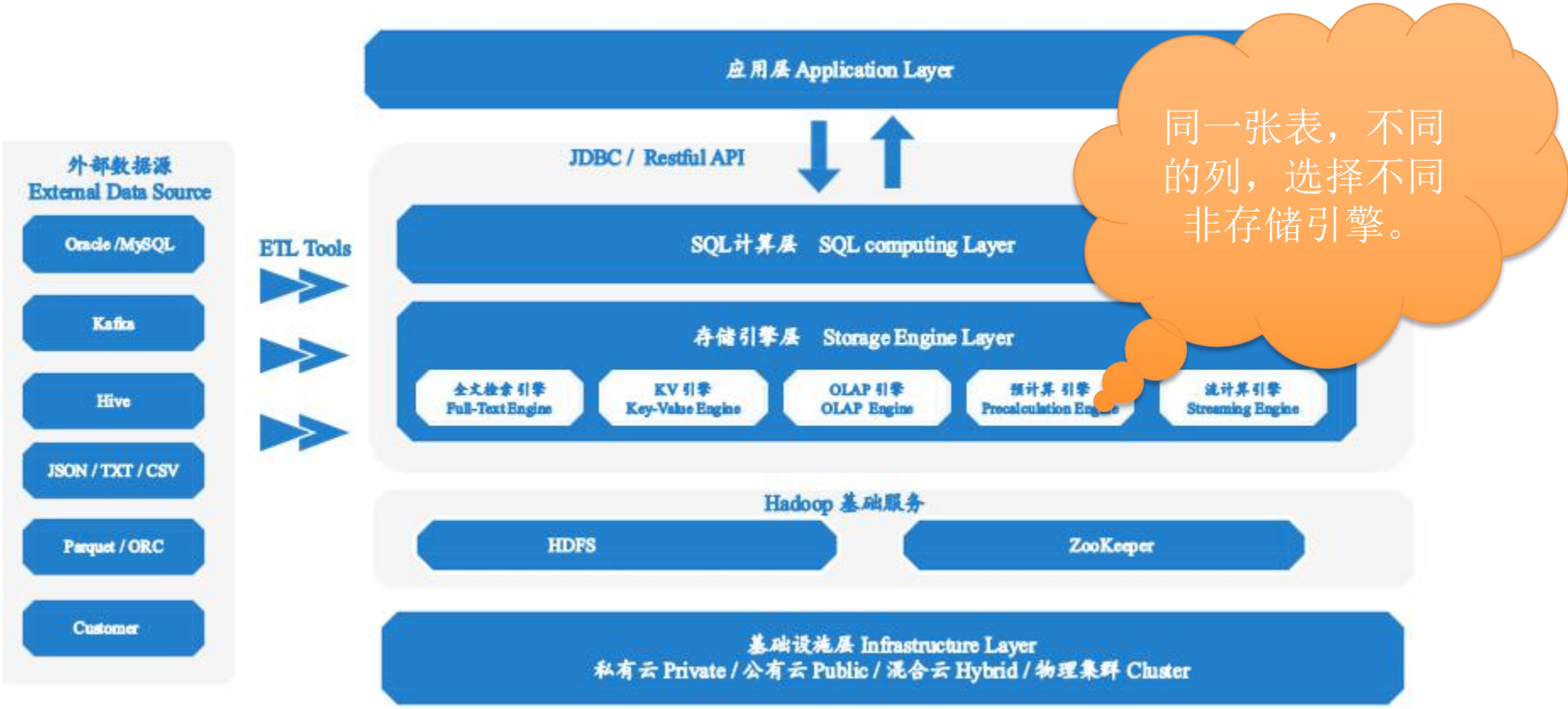
1:统一的标准SQL接口

2:十万亿规模,上千个节点

解决思路


标准SQL接口层
分布式计算层
分布式索引层
分布式存储层
做了大量的修改
APACHE Spark
Lucene
APACHE HBASE
hadoop HDFS

系统架构


外部数据源
External Data Source
Oracle /MySQL
Kafka
Hive
JSON / TXT / CSV
Parquet / ORC
Customer
ETL Tools
应用层 Application Layer
JDBC / Restful API
SQL计算层 SQL computing Layer
存储引擎层 Storage Engine Layer
全文检索引擎
Full-Text Engine
KV 引擎
Key-Value Engine
OLAP 引擎
OLAP Engine
预计算 引擎
Precalculation Engine
流计算引擎
Streaming Engine
Hadoop 基础服务
HDFS
ZooKeeper
基础设施层 Infrastructure Layer
私有云 Private / 公有云 Public / 混合云 Hybrid / 物理集群 Cluster
同一张表，不同的列，选择不同非存储引擎。

分布式索引的存储->HDFS

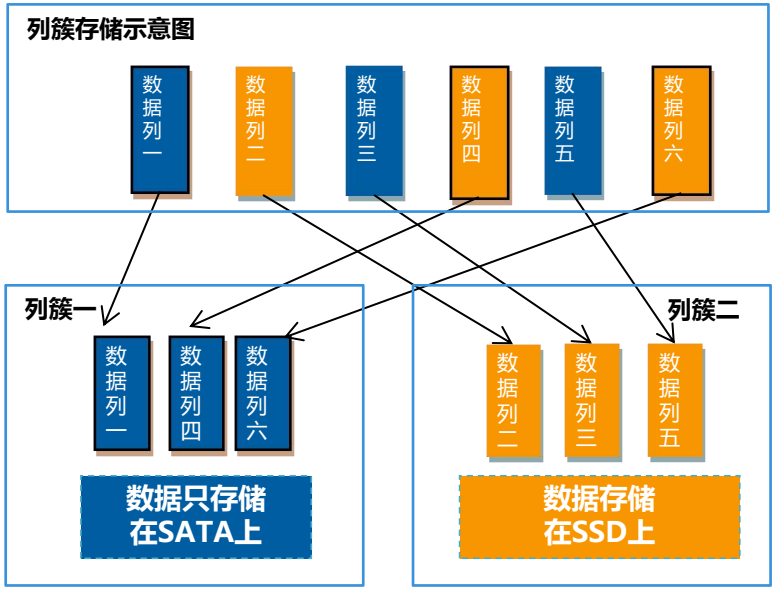
1：相对于本地文件系统的优点！
2：面临的主要问题，如何解决?
3：单库跨多个联邦存储。

全文检索->ES场景

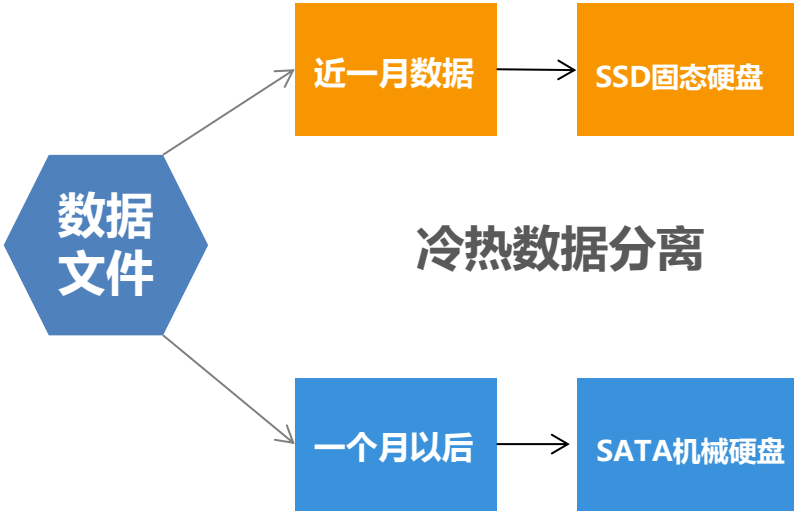

录信数软
Luxin DB software
汽车
全文检索
地理位置检索
统计分析
指数分析
交易时间:
开始日期
至
结束日期
银行名称:
请输入银行名称
商品名称:
请输入商品名称
交易金额最小值:
请输入最小交易金额
交易金额最大值:
请输入最大交易金额
经纬度:
请输入经度
请输入纬度
半径:
请输入半径
查询条数:
50
筛选
交易时间
2019 54829
2020 53522
性别
男 65205
女 43146
银行名称
青岛银行 734
宁夏银行 732
绍兴银行 729
常州农村 720
中国工商 718
交易类型
网上购物 29615
刷卡消费 29355
个人信息
银行信息
商品信息
交易信息
姓名: 於喆震
性别: 男
年龄: 27
身份证: 411424199207014075
联系方式: 13403951373
银行名称: 九江银行
账户类型: 借记卡
银行卡号: 6537823529497390
商户编号: 20190119867507
商户名称: 春珠汽车协会
商品名称: 左岸、爱登堡
商品类型: 服装
交易编号: 201903072140048865
交易时间: 2019/03/07 21:40:04
交易地点: 111.311756,23.441529
交易类型: 网上购物
交易金额: 2196
交易结果: 成功
万亿数据，秒级响应
["广西","115.643616,34.534885","借记卡","服装","8675","4504","4114","2019","商丘市","6537823529497390","预付费","13403951373","成功","07","201903072140048865","2196","20190307214004","春珠汽车协会,","20190119867507","左岸、爱登堡","河南","男","411424199207014075","111.311756,23.441529","湖州市商业银行","03","於喆震","115.643616,34.534885","网上购物","20190119","6114688126147426","梧州市","21","111.311756,23.441529","27","九江银行"]
姓名: 盛紫淑
银行名称: 中信银行
商户编号: 20190906552185
交易编号: 202002201857252559

列簇+异构

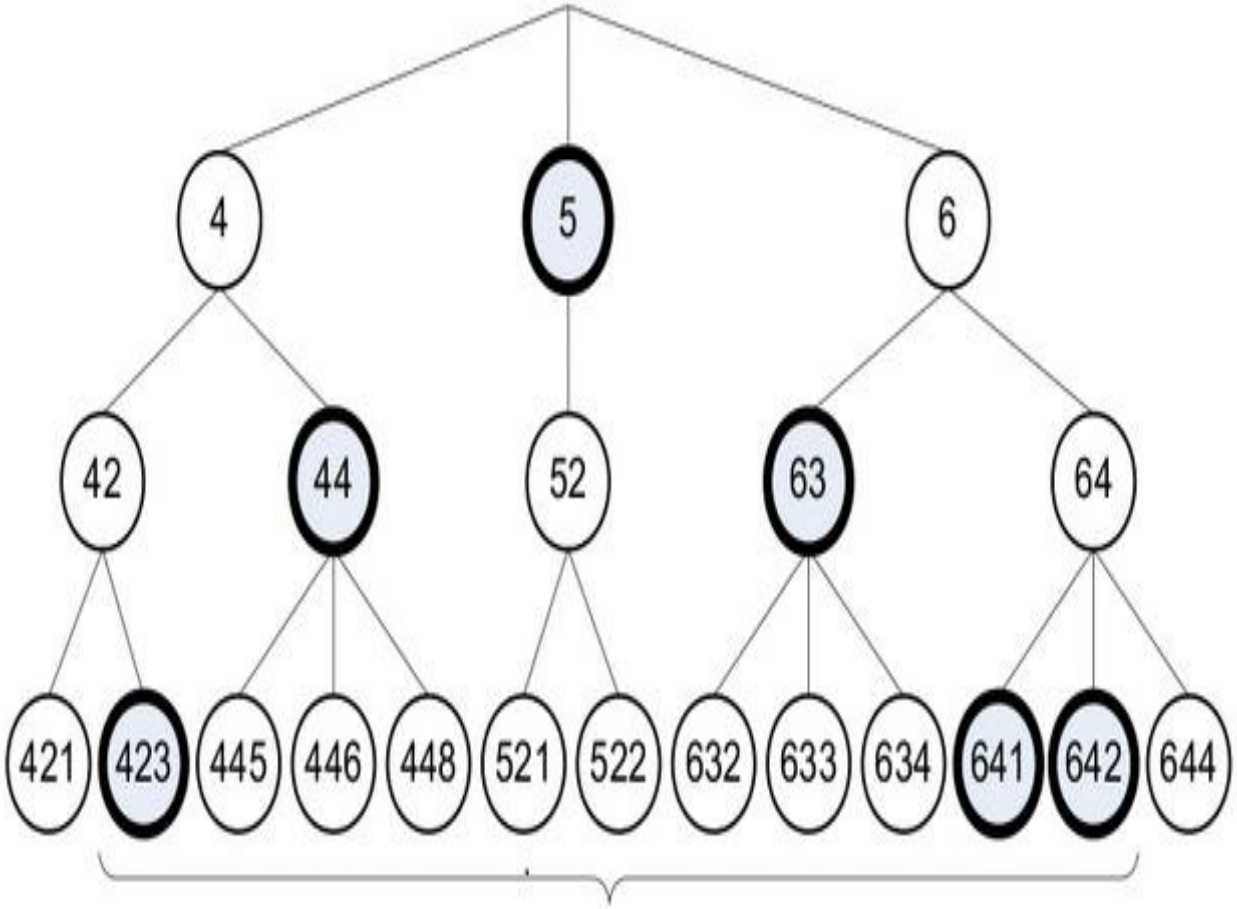
列簇存储



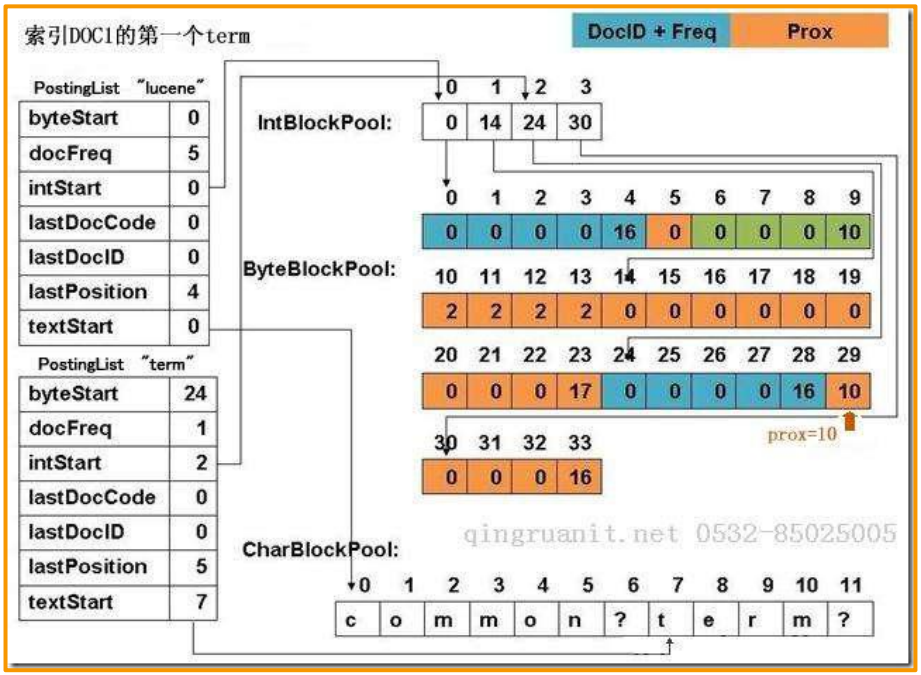
异构存储

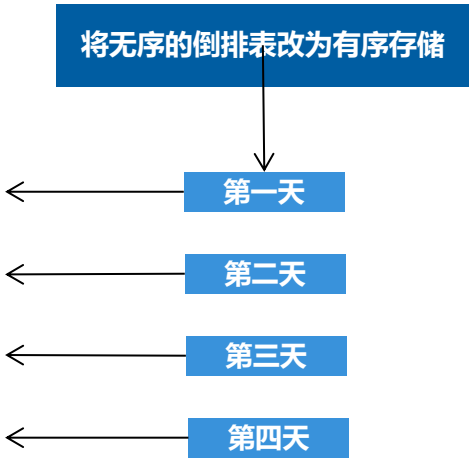


多层次索引-前缀与排序


4
5
6
42
44
52
63
64
421
423
445
446
448
521
522
632
633
634
641
642
644

倒排表数据按时序存储


索引DOC1的第一个term
DocID + Freq
Prox
PostingList "lucene"
byteStart 0
docFreq 5
intStart 0
lastDocCode 0
lastDocID 0
lastPosition 4
textStart 0
PostingList "term"
byteStart 24
docFreq 1
intStart 2
lastDocCode 0
lastDocID 0
lastPosition 5
textStart 7
IntBlockPool:
0 1 2 3
0 14 24 30
ByteBlockPool:
0 1 2 3 4 5 6 7 8 9
0 0 0 0 16 0 0 0 0 10
10 11 12 13 14 15 16 17 18 19
2 2 2 2 0 0 0 0 0 0
20 21 22 23 24 25 26 27 28 29
0 0 0 17 0 0 0 0 16 10
prox=10
30 31 32 33
0 0 0 16
CharBlockPool:
0 1 2 3 4 5 6 7 8 9 10 11
c o m m o n ? t e r m ?
qingruanit.net 0532-85025005


将无序的倒排表改为有序存储
第一天
第二天
第三天
第四天

计算框架->基于索引的Spark

将Spark底层的数据存储部分改成了基于分布式文件系统的索引，给Spark底层数据加了一层索引

查询的时候借助索引，避免了对数据的暴力扫描，查询与分析性能直接提升了100倍以上

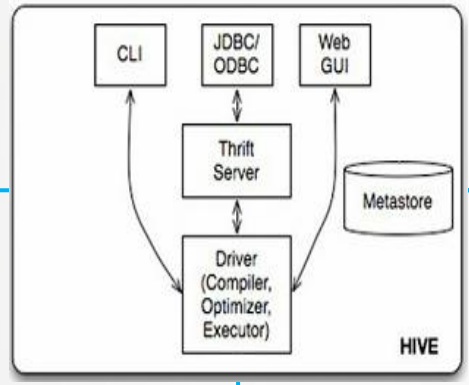
修正了大量的开源Spark的BUG，趟平开源Spark在生产系统中出现的各种问题

与spark融合后，查询与统计分析功能更强大

录信数软 Luxin DB software

- ✓ 支持SELECT语法：
  group by、order by、case when、sum、max、min、avg、count、
  join[left join[right join]]、in、not in、like、not like、with子句、union/union all、嵌套子查询等常见语法。
  支持1000+个条件组合联合匹配查询。

- ✓ 支持丰富的函数操作：
  支持数学函数(sin、cos、round、floor等)调用。
  支持字符函数(substring、concat等)调用。
  支持分析函数(row_number、rank、lag、lead等)调用。

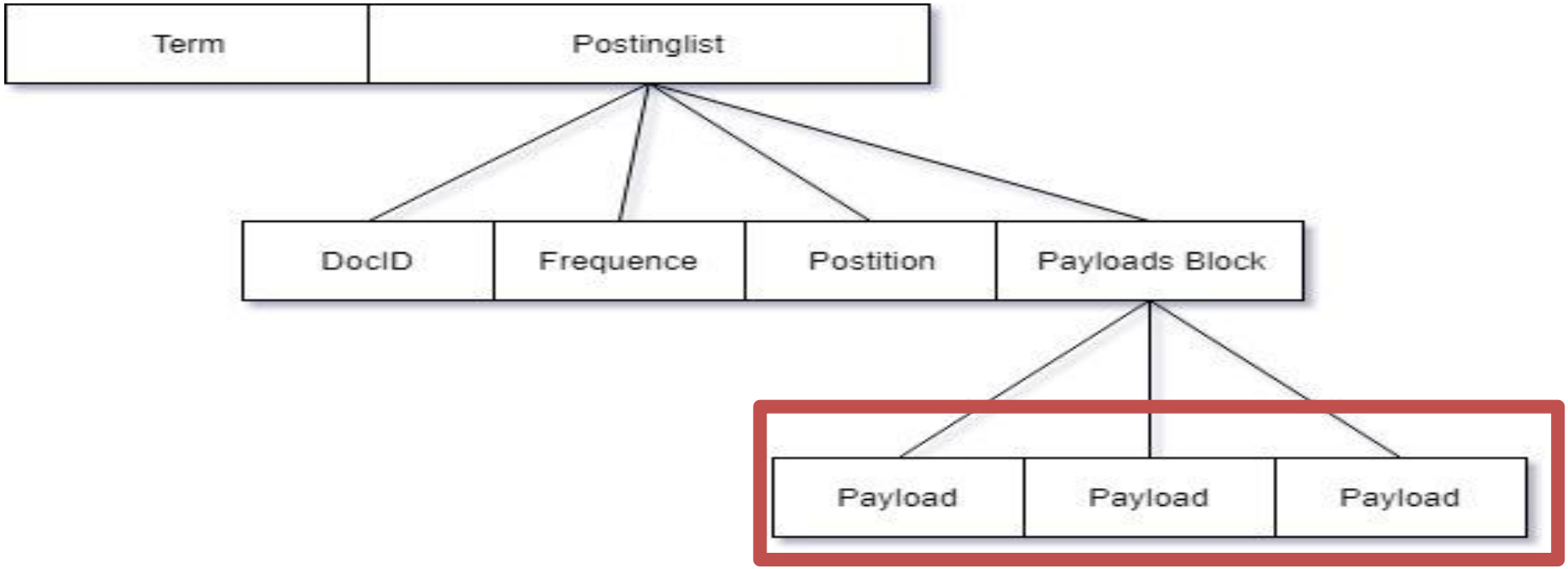


- ✓ 支持DDL语法：
  create table
  drop table
- ✓ 支持DML语法：
  支持数据插入insert语法
  支持数据删除delete语法。
- ✓ 支持数据分区清理truncate table语法。
- ✓ **支持数据的大规模导出操作。**

支持用户自定义(UDF)函数调用。
支持丰富的数据类型(string、long、int、double、char、like、text、geopoint等)，可以针对不同的业务场景进行表数据类型设计；同时也可以支持自定义分词数据类型。

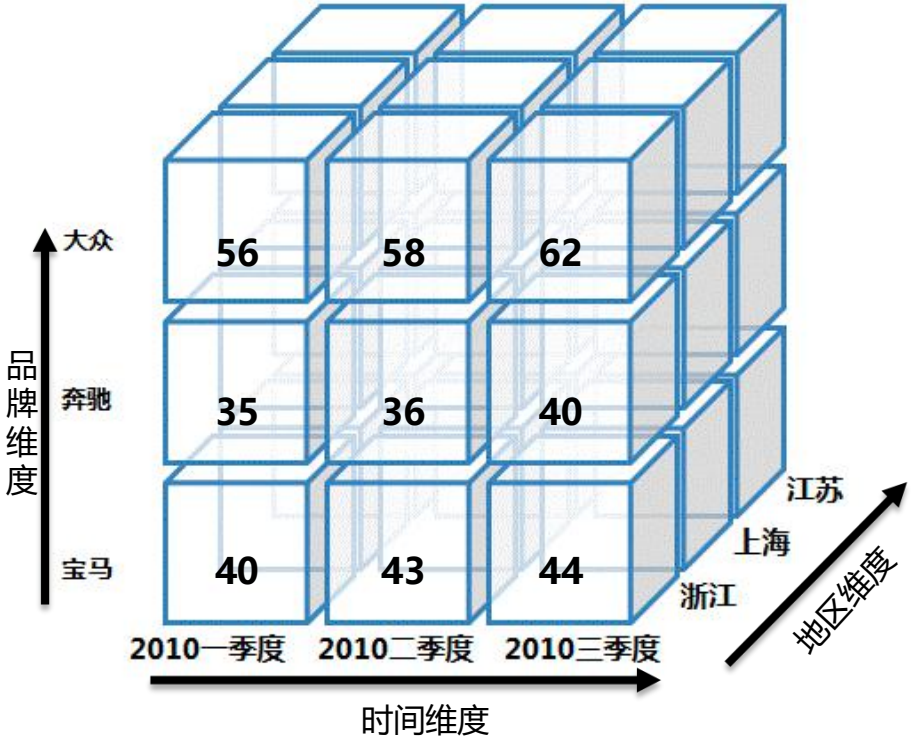
- 对手机号进行1*的检索。
- 对有数据倾斜的列，进行多表关联。
- Limit 10000000000
- Partition by 一个不均衡的列
- 过载控制

预计算索引->kylin+流计算

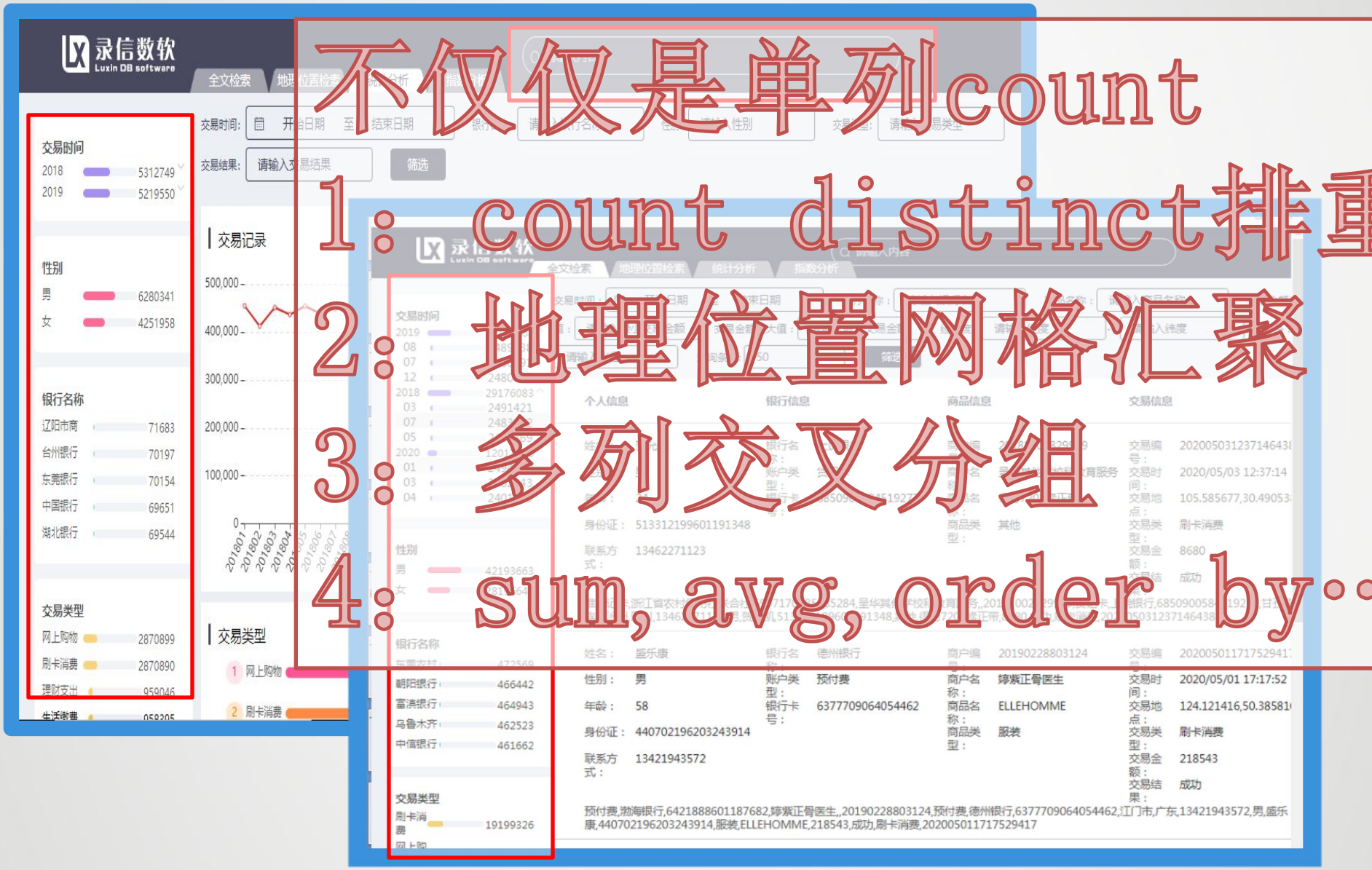


多维统计-> clickhouse场景

1：95%以上的值为null值。
2：碎片化数据多维统计分析。



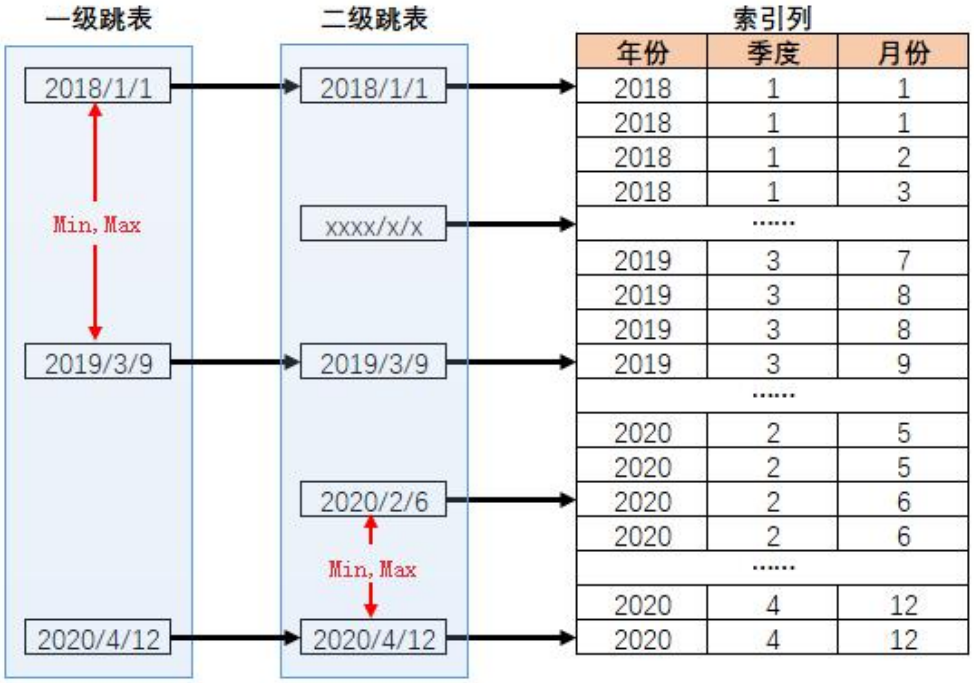
千人千面统计分析->vertical的projection



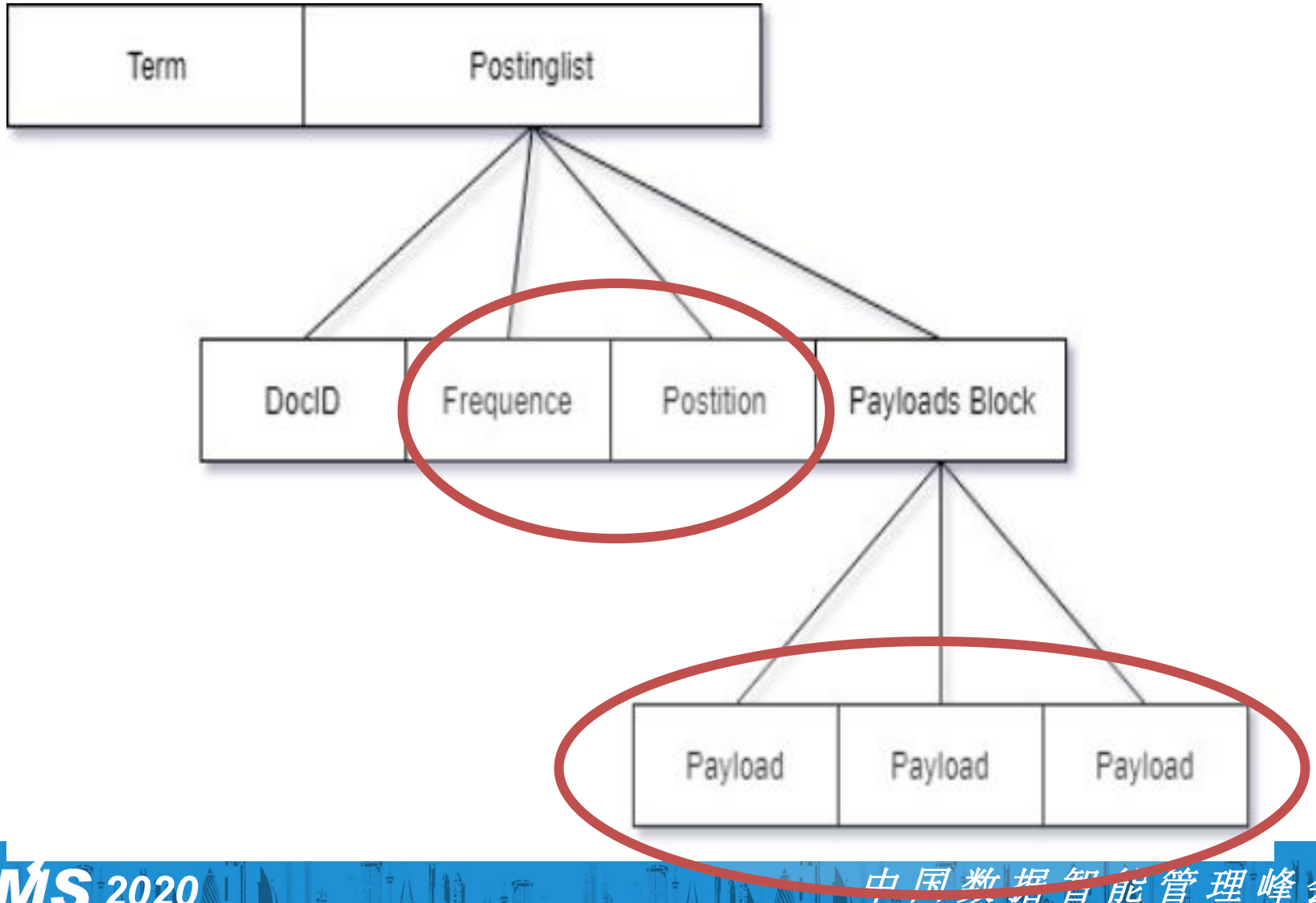
多维统计方案——多列联合索引

多维统计

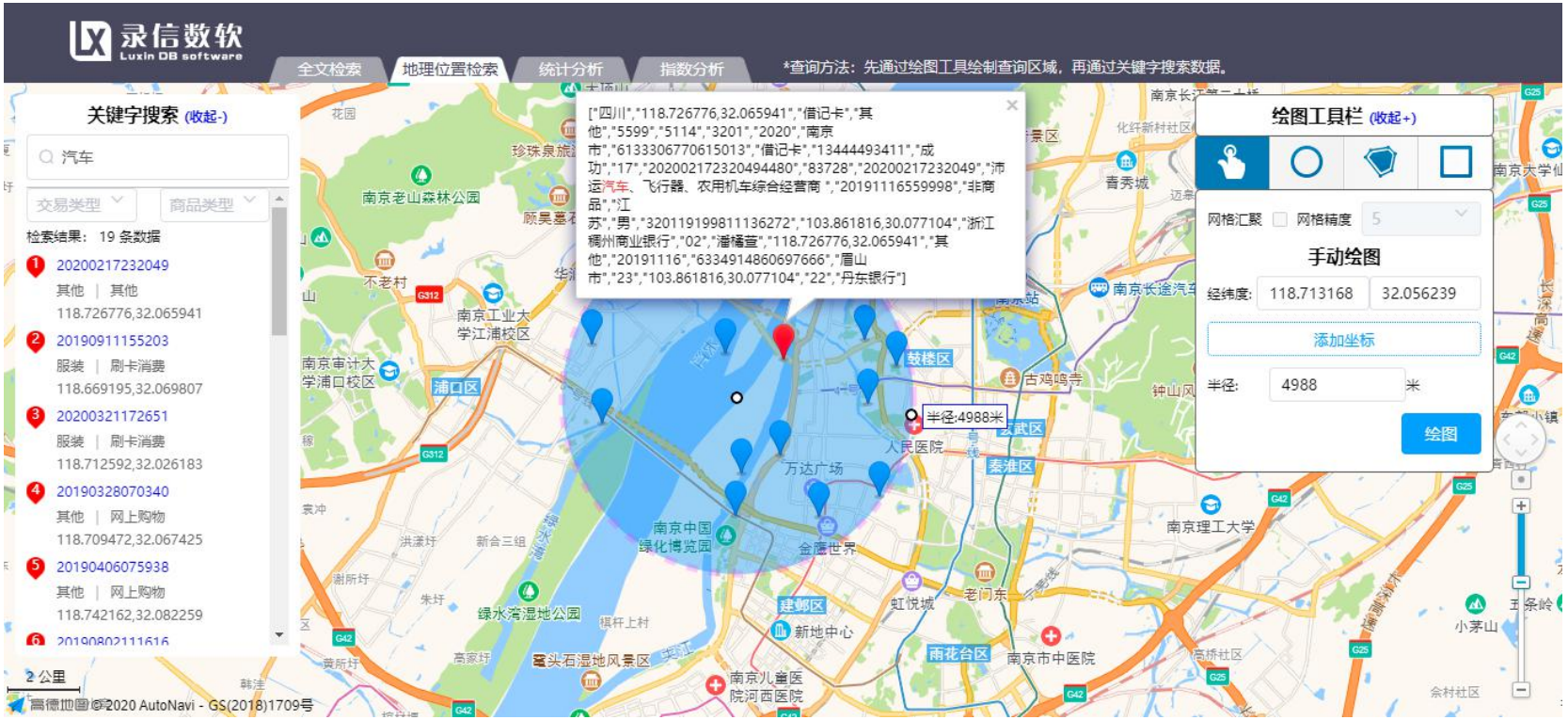
1、每列之间采用列存储。
2、干预数据的排序分布，让列存储的压缩更有效。
3、依据查询构造顺序读取。
4、多个列之间有层次关系。
5、结合分块存储。



Payloads压缩与按列存储-适合检索后的统计分析



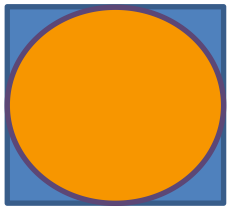
区域检索->数据预分布的变种



临近存储->数据预分布的变种

原生Lucene

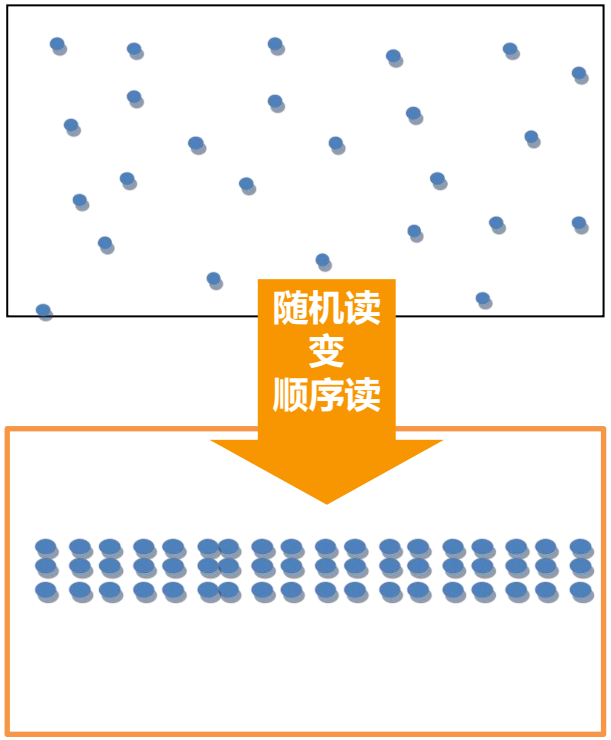
采用GeoHash选择正方形，再根据DocValues进行二次验证裁剪。



黄色部分
需要剪切验证

优化方案

通过地理位置临近数据**临近存储**的方式构造硬盘上的连续读取，大幅度的减少随机读取的次数，从而提升查询响应的速度。

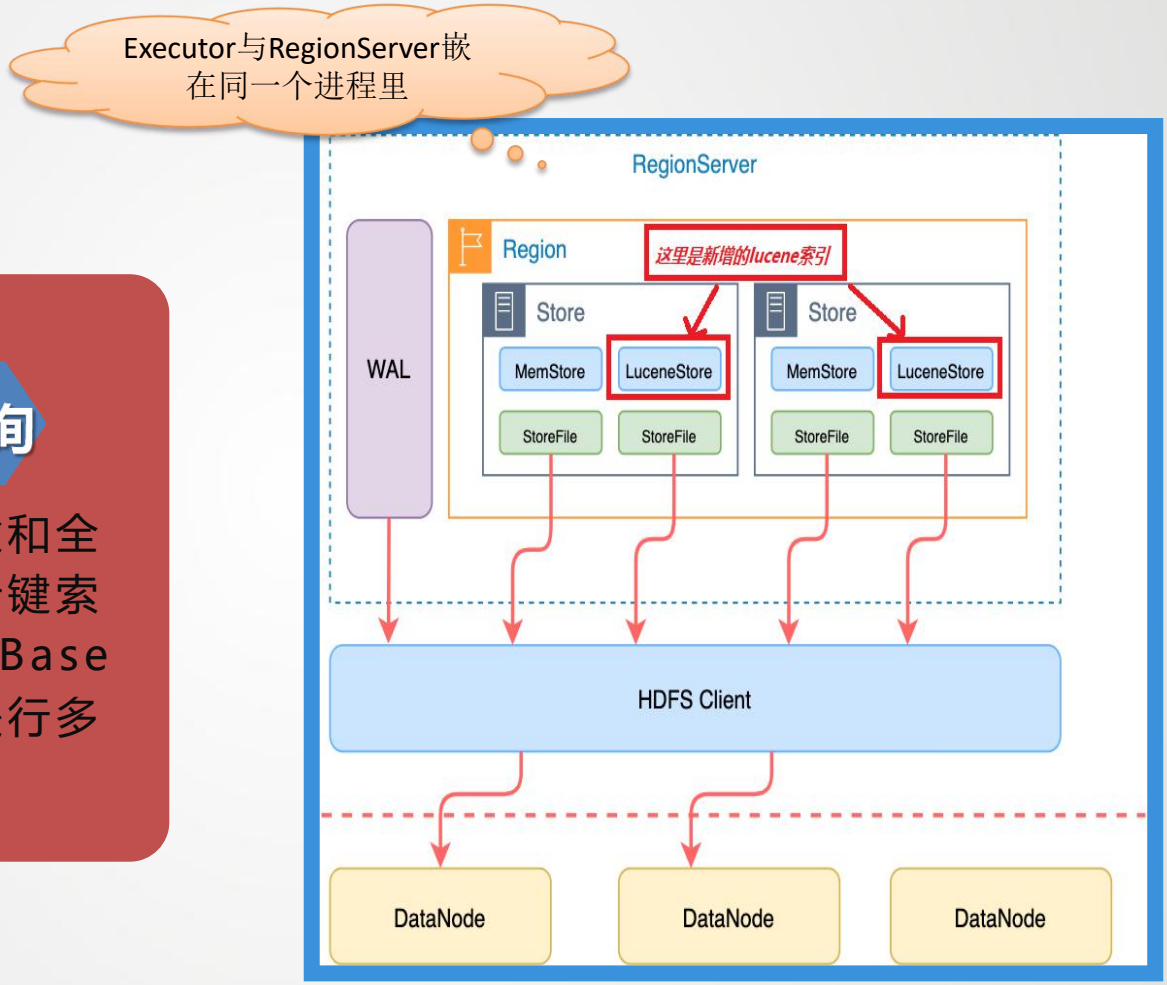


实现 Hbase 二级索引的方法

录信数软
Luxin DB software

HBase多维度查询

本身只提供基于行键和全表扫描的查询，而行键索引单一，需要采用HBase的二级索引方案来进行多条件的查询。



实现 Hbase 二级索引的关键技术

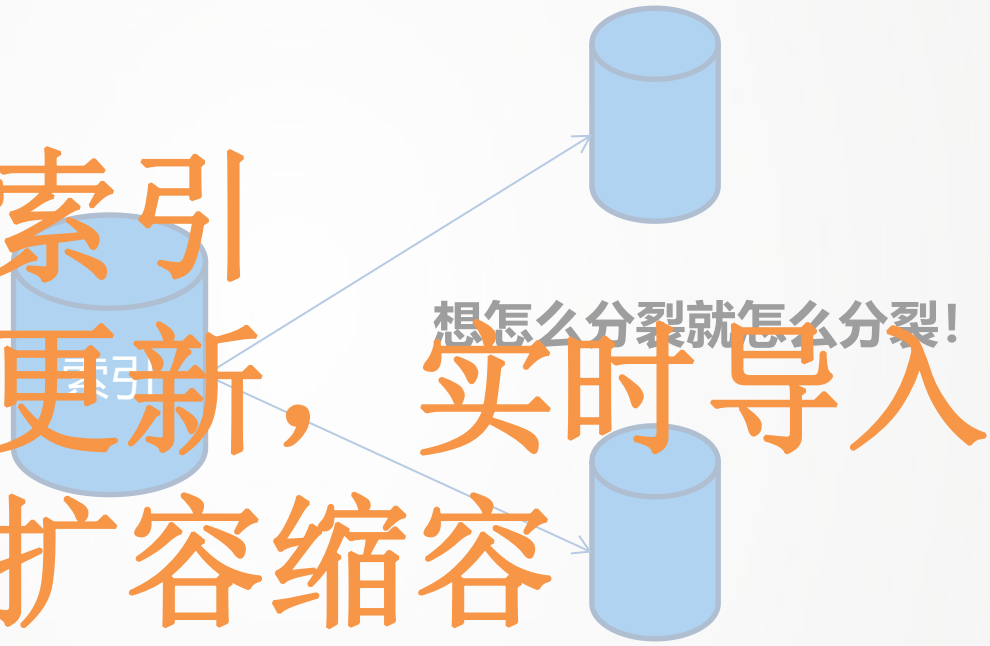
录信数软
Luxin DB software

特点

1：多维索引

2：实时更新，实时导入

3：随意扩容缩容



其他要实现的

1：任务调度，IO调度。
2：物化视图。
3：索引加载问题。

应用场景——公安军队

场景描述

公安部门汇集了全网全维度的海量数据，包含互联网数据、社会数据、通讯数据等，通过实时检索、关联碰撞，为各警种提供智能研判的关系网络（同行、同住同飞、同出入境等），大幅增加可用情报线索，提升侦破水平。

应用场景——汽车行业

场景描述

车联网迅速发展，数亿T-BOX采集的万亿级别数据经过分析可让经销商、主机厂、国家监管部门多方受益。尤其随着国六标准出台，国家要对汽车尾气排放进行监管，需要大量采集油耗和排放数据。

录信与中国汽车技术研究中心（中汽研）合作，针对国六标准的推行，对商用车和乘用车的尾气数据、油耗数据、加油站地理位置信息进行监控，并提供整体存储和检索分析解决方案，为一汽解放、江淮汽车等主机厂的创新型应用提供坚实基础

Q&A

南京录信软件技术有限公司

DAMS

中国数据智能管理峰会

DATA & AI MANAGEMENT SUMMIT

THANK YOU !