成就直达号的大数据引擎技术专场

百度直达号 http://zhida.baidu.com/

百度开放服务平台 http://developer.baidu.com/

百度开放云 http://bce.baidu.com/

百度OLAP系统实践

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

IT168 ChinaUnix ITPUB

目录

- OLAP背景介绍
- Palo整体架构
- Palo关键技术
- Palo对外开放

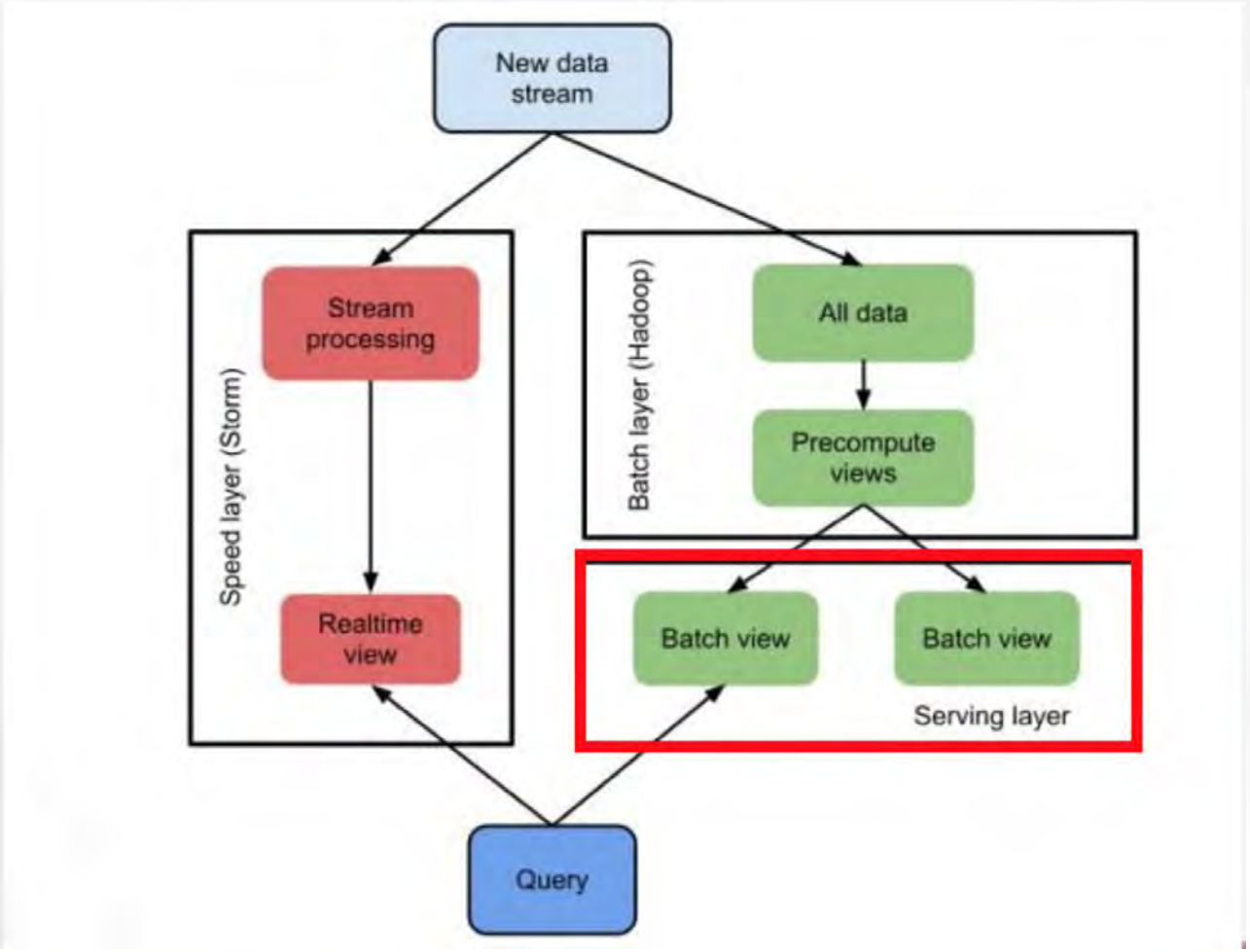
什么是OLAP

- Online Analytical Processing
  - Analytical Processing vs. Transactional Processing
  - Online vs. Offline (Interactive vs. Batch)

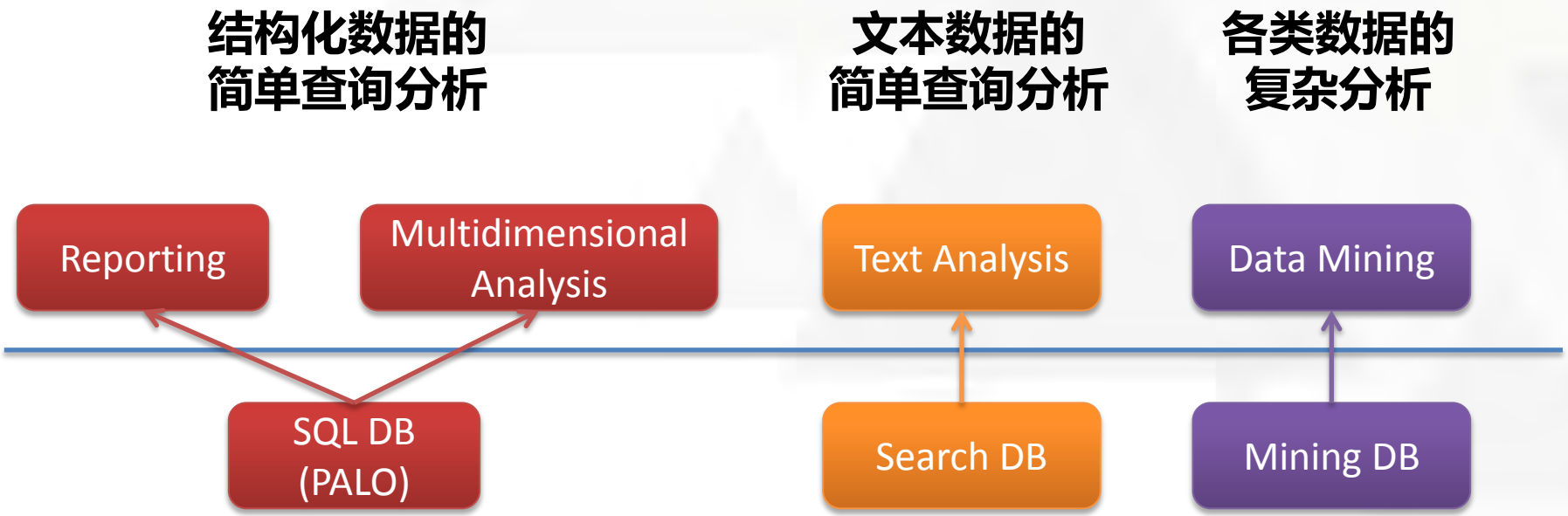
OLTP vs. OLAP

	OLTP	OLAP
面向应用	日常交易处理	明细查询，分析决策
访问模式	简单小事务，操作少量数据	复杂聚合查询，可以过大量数据
数据	当前最新数据	历史数据
数据规模	GB	TB ~ PB
数据更新	实时更新	批量更新
数据组织	满足3NF	反范式，星型模型

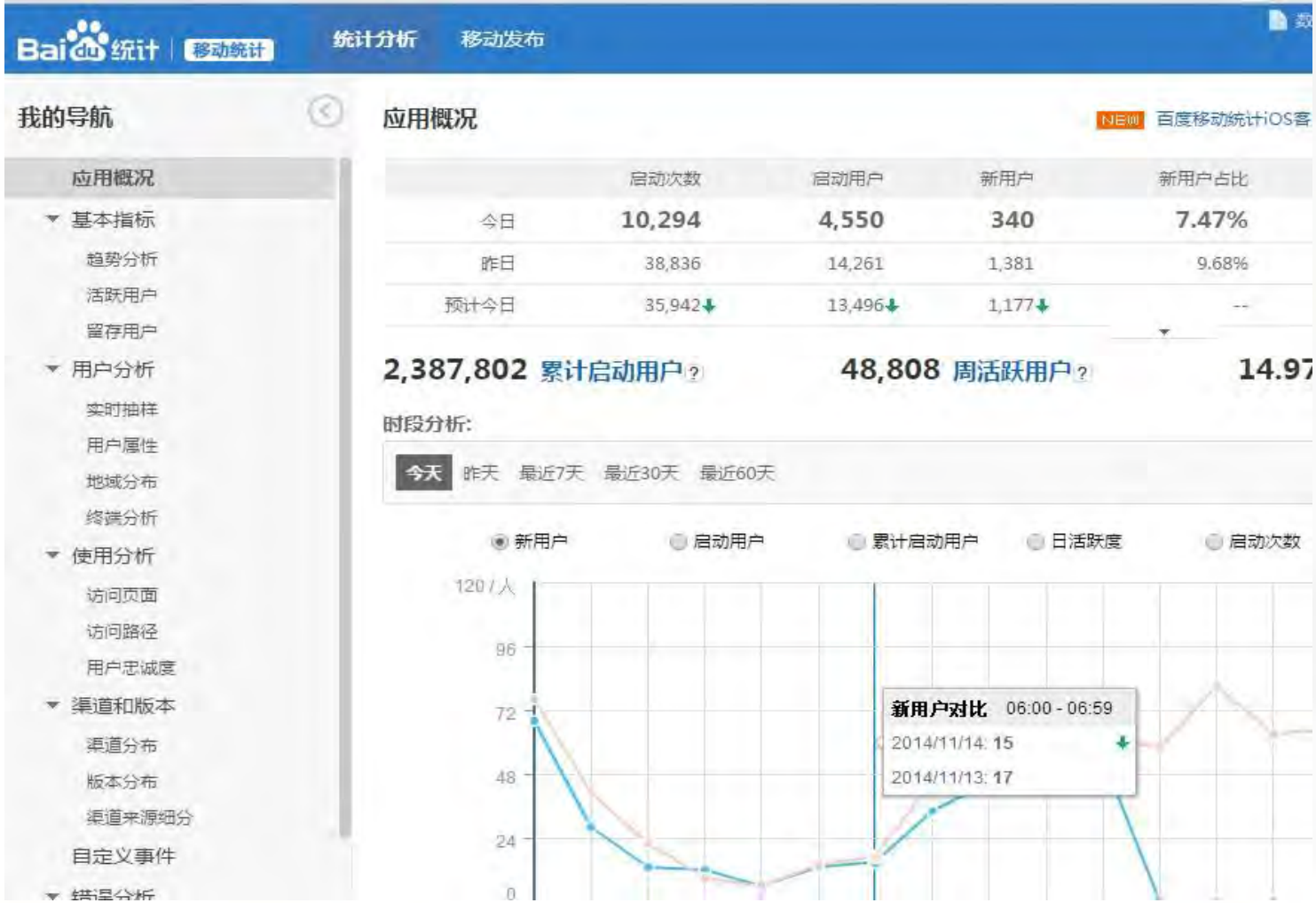
Online vs. Offline



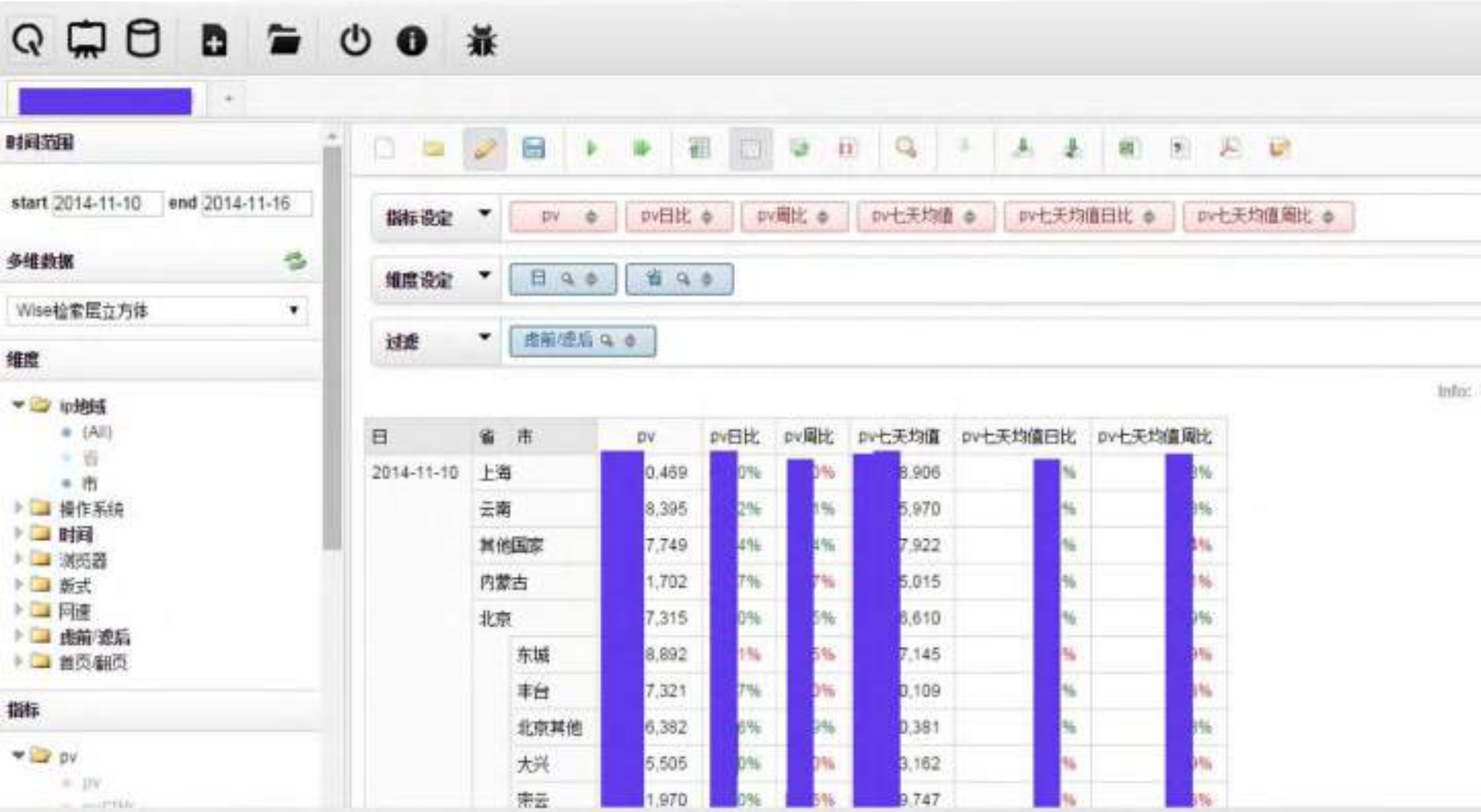
OLAP - Interactive Data Analysis



OLAP应用-在线报表


Baidu 统计
移动统计
统计分析
移动发布
我的导航
应用概况
基本指标
趋势分析
活跃用户
留存用户
用户分析
实时抽样
用户属性
地域分布
终端分析
使用分析
访问页面
访问路径
用户忠诚度
渠道和版本
渠道分布
版本分布
渠道来源细分
自定义事件
应用概况
NEW 百度移动统计iOS客
启动次数 启动用户 新用户 新用户占比
今日 10,294 4,550 340 7.47%
昨日 38,836 14,261 1,381 9.68%
预计今日 35,942 13,496 1,177 --
2,387,802 累计启动用户
48,808 周活跃用户
14.97
时段分析:
今天 昨天 最近7天 最近30天 最近60天
新用户 启动用户 累计启动用户 日活跃度 启动次数
120 / 人
96
72
48
24
0
新用户对比 06:00 - 06:59
2014/11/14: 15
2014/11/13: 17

OLAP应用-多维分析


时间范围
start 2014-11-10
end 2014-11-16
多维数据
Wise检索层立方体
维度
ip地域
(All)
省
市
操作系统
时间
浏览器
版式
网速
虚前/虚后
首页/翻页
指标
pv
指标设定
pv
pv日比
pv周比
pv七天均值
pv七天均值日比
pv七天均值周比
维度设定
日
省
过滤
虚前/虚后
日
省 市
pv
pv日比
pv周比
pv七天均值
pv七天均值日比
pv七天均值周比
2014-11-10
上海
0,469
云南
8,395
其他国家
7,749
内蒙古
1,702
北京
7,315
东城
8,892
丰台
7,321
北京其他
6,382
大兴
5,505
密云
1,970

商业产品

产品	简介	技术特点	收购情况
Netezza	2000年在美国成立 Netezza TwinFin	✓ 软硬一体机 ✓ 采用FPGA数据过滤代替索引	2010年9月20日，IBM出资17.8亿美元收购
Greenplum	2003年在美国成立 Greenplum Database	✓ 行存 + 列存 ✓ Shared-Nothing集群	2010年7月6日，EMC出资3亿美元收购
Vertica	2005年在美国成立 Vertica Analytic Database	✓ 列存 ✓ Shared-Nothing集群	2011年2月，HP出资3.5亿美元收购
Aster Data	2005年在美国成立 nCluster	✓ SQL-MapReduce ✓ Shared-Nothing集群	2011年7月6日，Teradata出资2.63亿美元收购
ParAccel	2005年在美国成立 PADB	✓ 列存 + 自适应压缩 ✓ Shared-Nothing集群	2013年Actian出资1.5亿美元收购，Redshift宣称使用ParAccel

Vendor and Appliance	Memory (GB)	Total Cores	Compression	User Storage (TB, Compressed)	List Price
EMC Greenplum Data Computing Appliance	768	48	4 to 1	144	$2,000,000
IBM PureData System for Analytics N1001-010	n/a	112	4 to 1	128	$1,599,000
Microsoft SQL Server 2012 Parallel Data Warehouse[1]	2,304	144	5 to 1	340	$1,569,970
Oracle Exadata Database Machine X3-2	2,048	128	10 to 1	450	$13,580,000
Teradata Data Warehouse Appliance 2690	768	96	4 to 1	146	$1,168,000

tive Analysis of Web-Scale Datasets (Google 2010)
Implementation On The MapReduce Framework *(Google 2011)*
cessing a Trillion Cells per Mouse Click (Google 2012)

Hortonworks

The Stinger Initiative: Making Apache Hive 100 Times Faster

cloudera

Cloudera Impala: Real-Time Queries in Apache Hadoop, For Real

Google bigquery

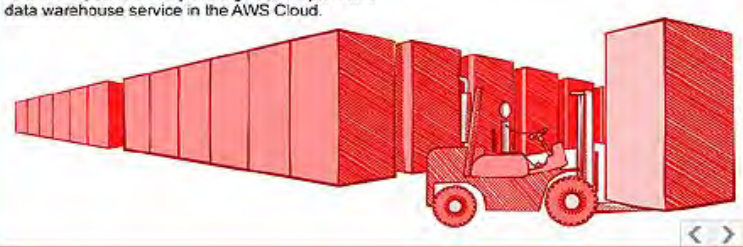
Compose Query

Recent Queries

Apache Drill Distributed system for interactive analysis.

Introducing **Amazon Redshift**

A fast and powerful, fully managed petabyte-scale data warehouse service in the AWS Cloud.



MemSQL, The Real-Time Analytics Platform.

Spark SQL

Download Libraries Documentation Examples Community FAQ

Spark SQL is Spark's module for working with structured data.

Mesa: Geo-Replicated, Near Real-Time, Scalable Data Warehousing

Ashish Gupta, Fan Yang, Jason Govig, Adam Kirsch, Kelvin Chan
Kevin Lai, Shuo Wu, Sandeep Govind Dhoot, Abhilash Rajesh Kumar, Ankur Agiwal
Sanjay Bhansali, Mingsheng Hong, Jamie Cameron, Masood Siddiqi, David Jones
Jeff Shute, Andrey Gubarev, Shivakumar Venkataraman, Divyakant Agrawal
Google, Inc.

ABSTRACT

Mesa is a highly scalable analytic data warehousing system that stores critical measurement data related to Google's

ness critical nature of this data result in unique technical and operational challenges for processing, storing, and querying. The requirements for such a data store are:

目录

- OLAP背景介绍
- Palo整体架构
- Palo关键技术
- Palo对外开放

Palo

- ***A MPP-based Interactive Data Analysis SQL DB***
- A Google Mesa Clone, is simpler and better than Mesa
- **面向百TB~PB级别，结构化数据，毫秒/秒级分析**
- **百度大数据部OLAP团队研发**
- **第三代产品 Doris -> OlapStroageEngine -> Palo**
- **PALO意为“玩转OLAP”**
- **80+产品线使用，400+机器，单一业务最大百TB，15年预计部署1000台**

Palo设计原则和定位

- **简单可依赖**

低成本

1/10 ~1/100 Cost

线性扩展

100~200节点 / 1000 TB

支持云化部署

高可用

99.9999 % Uptime

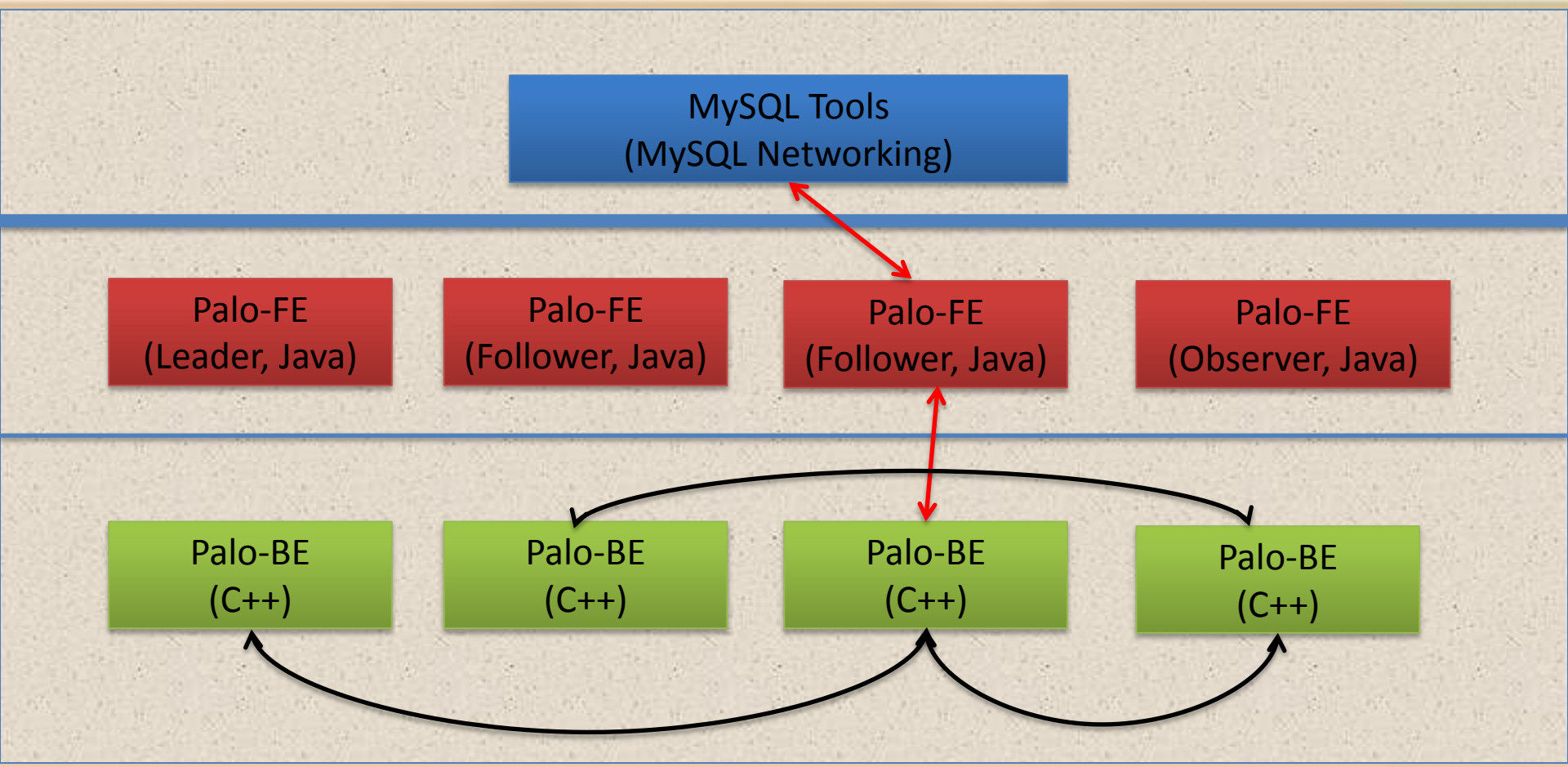
高查询性能

10W QPS/ 100GB/s

高加载性能

10 TB / Hour

整体架构

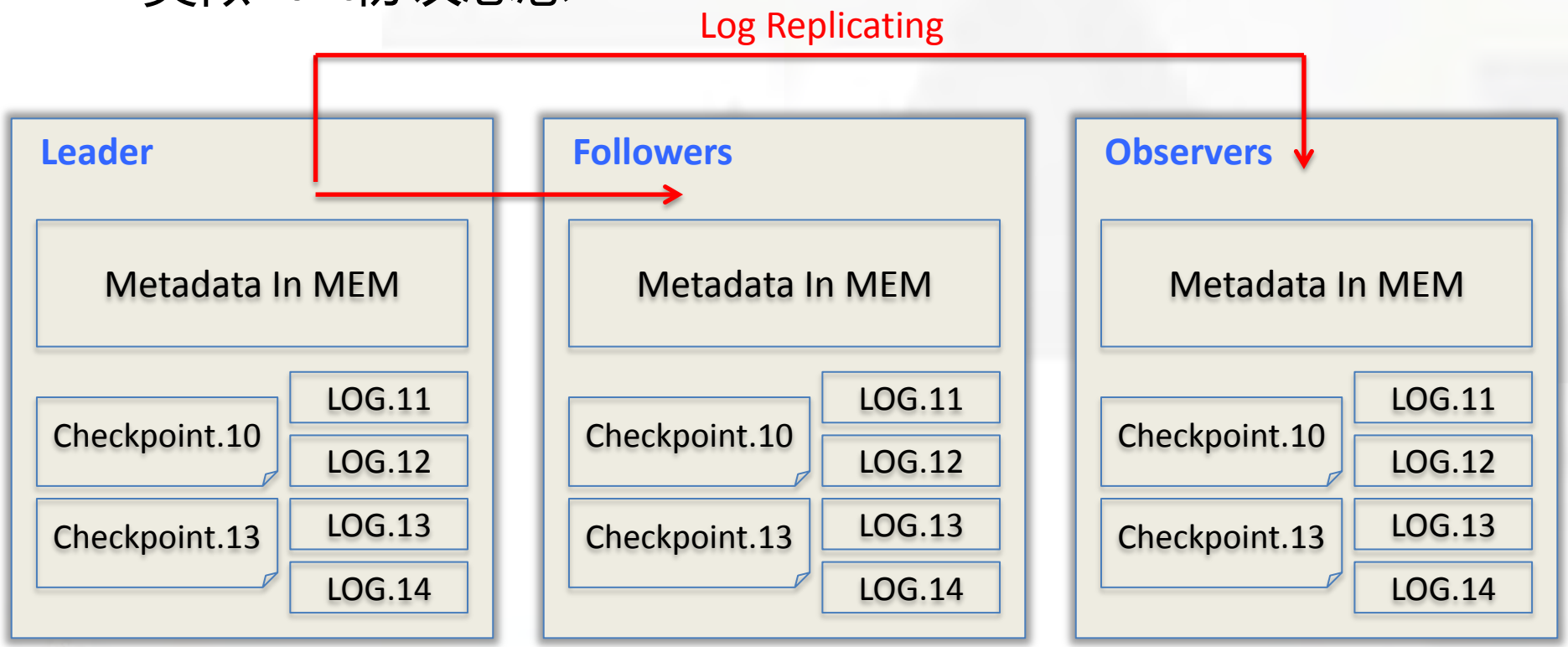

MySQL Tools
(MySQL Networking)
Palo-FE
(Leader, Java)
Palo-FE
(Follower, Java)
Palo-FE
(Follower, Java)
Palo-FE
(Observer, Java)
Palo-BE
(C++)
Palo-BE
(C++)
Palo-BE
(C++)
Palo-BE
(C++)

目录

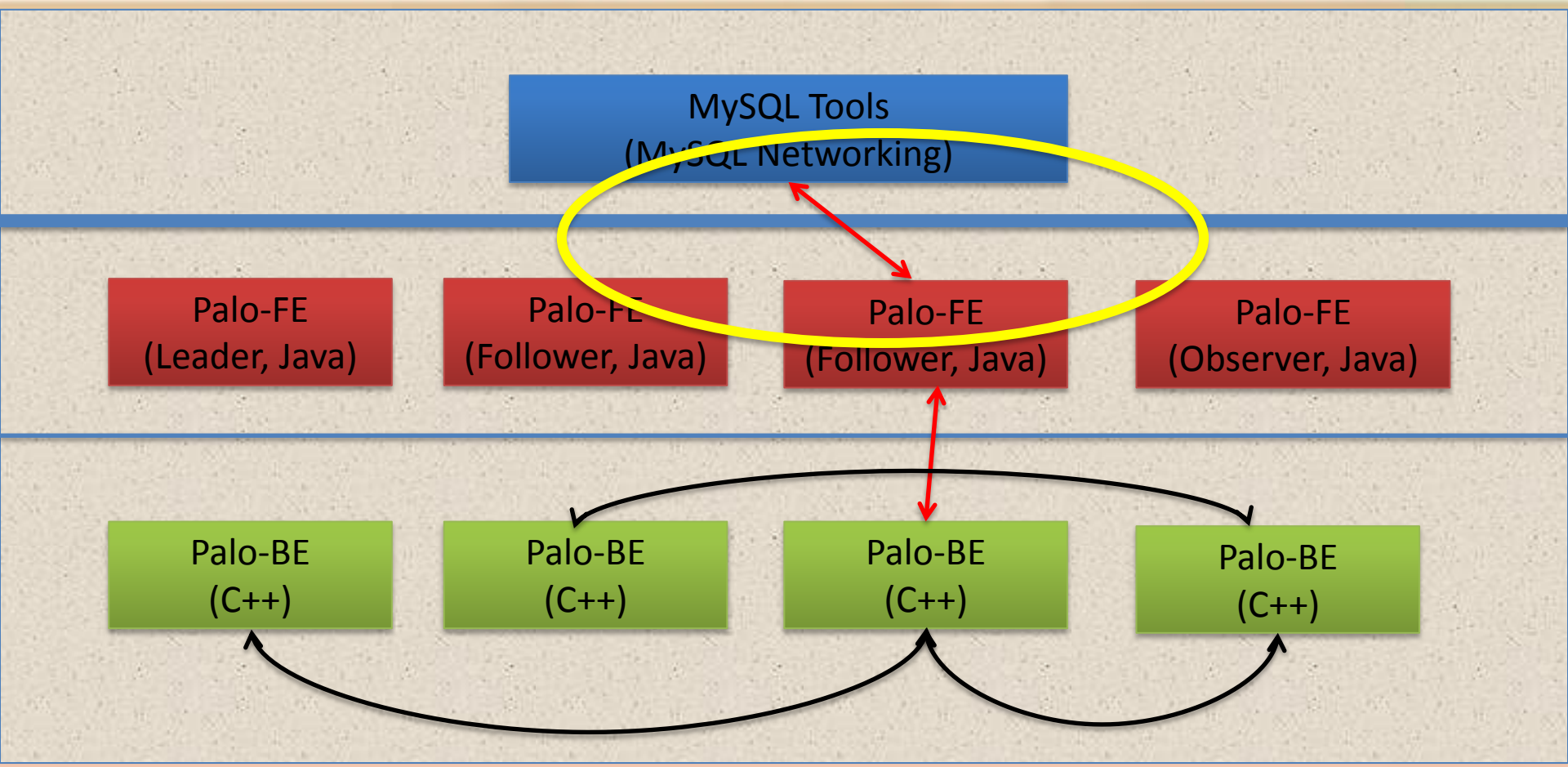
- OLAP背景介绍
- Palo整体架构
- Palo关键技术
- Palo对外开放

Frontend Metadata Management

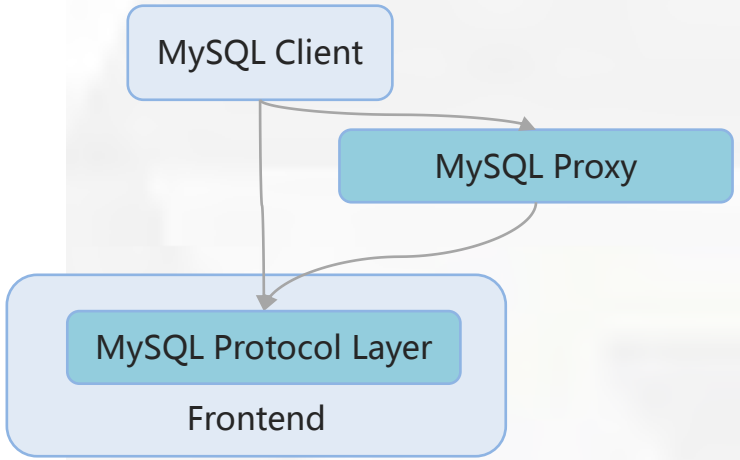
- State Machine + Replicated Log
- 类似Raft协议思想



MySQL Networking Protocol

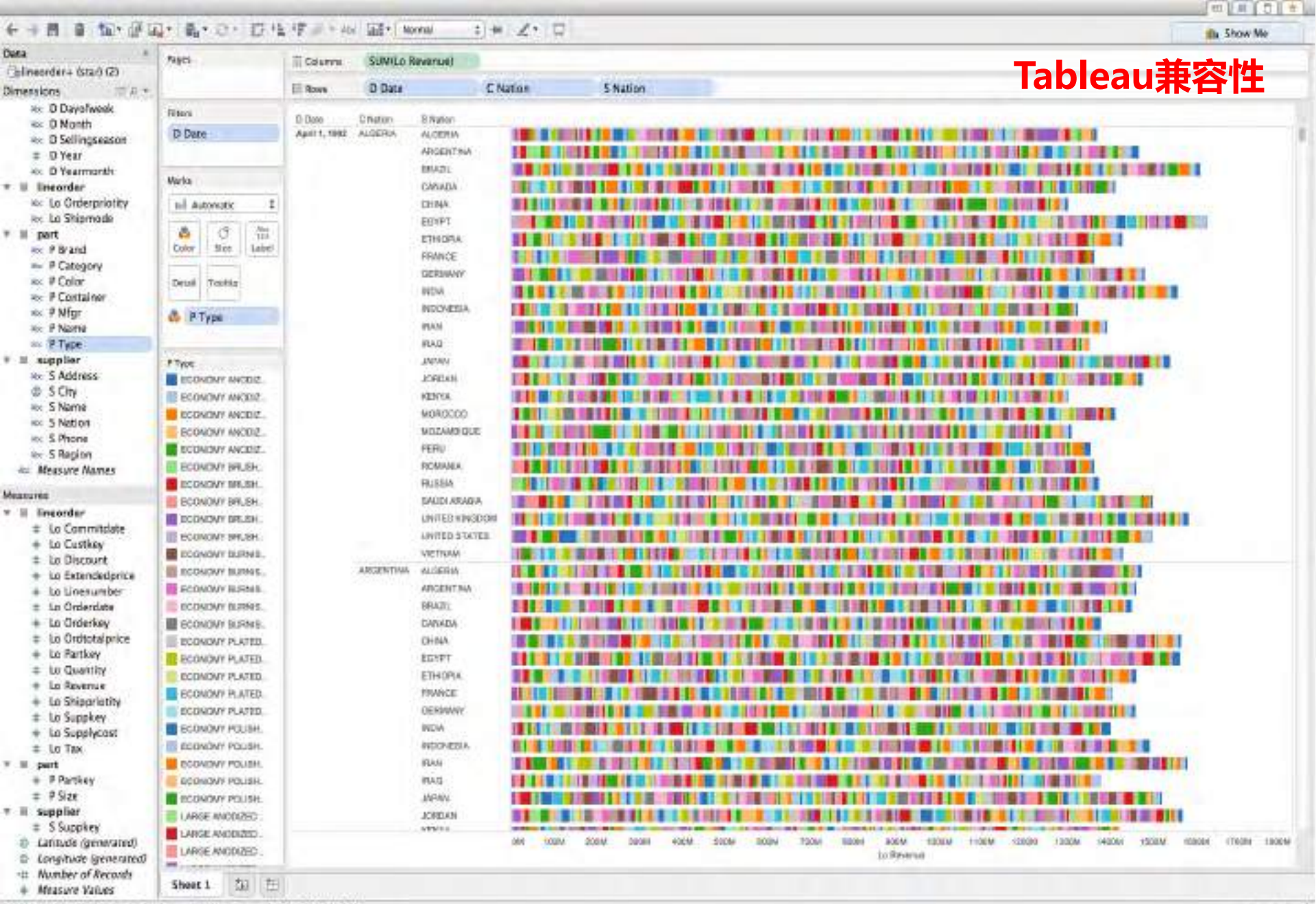


```
test@mry-laptop:~$
test@mry-laptop:~$ mysql -h tc-inf-devop01.tc.baidu.com -P 8276 -u maruyue
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 0
Server version: 4.1.2 (Powered by Palo 2.0 Beta)

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| demo               |
| fc                 |
| information_schema |
| lbs                |
| searchbox          |
| test               |
+--------------------+
6 rows in set (0.01 sec)

mysql> use test;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+-----------------+
| Tables_in_test  |
+-----------------+
| fc_cmatch_fact  |
| tbldim_pn       |
| tbldim_querytrade |
| tbldim_region   |
| tbldim_wbws     |
| tbldim_wns      |
| tbldim_wnt      |
+-----------------+
7 rows in set (0.01 sec)
```



- ✓ 轻量级客户端
- ✓ 与上层应用兼容容易
- ✓ 学习曲线平缓，方便用户上手使用
- ✓ 利用MySQL相关工具，比如MySQL Proxy

Tableau兼容性


Data
Pages
Filters
Marks
Columns
SUM(Lo Revenue)
Rows
O Date
C Nation
S Nation
Dimensions
Measures
Sheet 1

BIEE兼容性


ORACLE Business Intelligence
1.4
计分析
总览_
平台
(所有列值)
版式
(所有列值)
日期
介于
应用
重置
说明：用户里为各渠道用户里加和，并未做渠道间的去重处理。
说明标题
分析 - 编辑 - 导出
总览
查询条件
版式：(所有列值)
平台：(所有列值)
日期
9,154
155
23,358
7,305
658
476
485
401
9,812
038
6,831
,509
3,429
666
2,533
,981
6,659
752
8,394
,124
9,167
37
7,940
,264
9,149
70
9,474
869
2,117
99
4,806
,559
,011
248
7,238
372

R语言兼容性

```
➜ test_r  R

R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> library(RMySQL)
Loading required package: DBI
> con <- dbConnect(MySQL(), user="root", password="[redacted]", dbname="demo", host="tc-inf-devop01.tc.baidu.com", port=8276)
> dbListTables(con)
[1] "cumulative_detail_test" "fc_cmatch_fact"         "tbldim_pn"
[4] "tbldim_querytrade"      "tbldim_region"          "tbldim_wbws"
[7] "tbldim_wos"             "tbldim_wpt"             "ud_test"
> rs <- dbSendQuery(con, "select * from tbldim_region")
> d1 <- fetch(rs, n = 10)
> d1
   pid cid province     city
1    1   0     北京 北京其他
2    2   0     上海 上海其他
3    3   0     天津 天津其他
4    4   0     广东 广东其他
5    5   0     福建 福建其他
6    8   0     海南 海南其他
7    9   0     安徽 安徽其他
8   10   0     贵州 贵州其他
9   11   0     甘肃 甘肃其他
10  12   0     广西 广西其他
> 
```

存储模型

Keys

Values

Date	PublisherId	Country	Clicks	Cost
2013/12/31	100	US	10	32
2014/01/01	100	US	205	103
2014/01/01	200	UK	100	50

(a) Mesa table A

Date	AdvertiserId	Country	Clicks	Cost
2013/12/31	1	US	10	32
2014/01/01	1	US	5	3
2014/01/01	2	UK	100	50
2014/01/01	2	US	200	100

(b) Mesa table B

Base表

AdvertiserId	Country	Clicks	Cost
1	US	15	35
2	UK	100	50
2	US	200	100

(c) Mesa table C

Rollup表

Figure 1: Three related Mesa tables

引自Google Mesa Paper

Values聚合方式Sum, Replace

Date	PublisherId	Country	Clicks	Cost
2013/12/31	100	US	+10	+32
2014/01/01	100	US	+150	+80
2014/01/01	200	UK	+40	+20

(a) Update version 0 for Mesa table A

Date	AdvertiserId	Country	Clicks	Cost
2013/12/31	1	US	+10	+32
2014/01/01	2	UK	+40	+20
2014/01/01	2	US	+150	+80

(b) Update version 0 for Mesa table B

Delta更新

Date	PublisherId	Country	Clicks	Cost
2014/01/01	100	US	+55	+23
2014/01/01	200	UK	+60	+30

(c) Update version 1 for Mesa table A

Date	AdvertiserId	Country	Clicks	Cost
2013/01/01	1	US	+5	+3
2014/01/01	2	UK	+60	+30
2014/01/01	2	US	+50	+20

(d) Update version 1 for Mesa table B

Figure 2: Two Mesa updates

Compaction

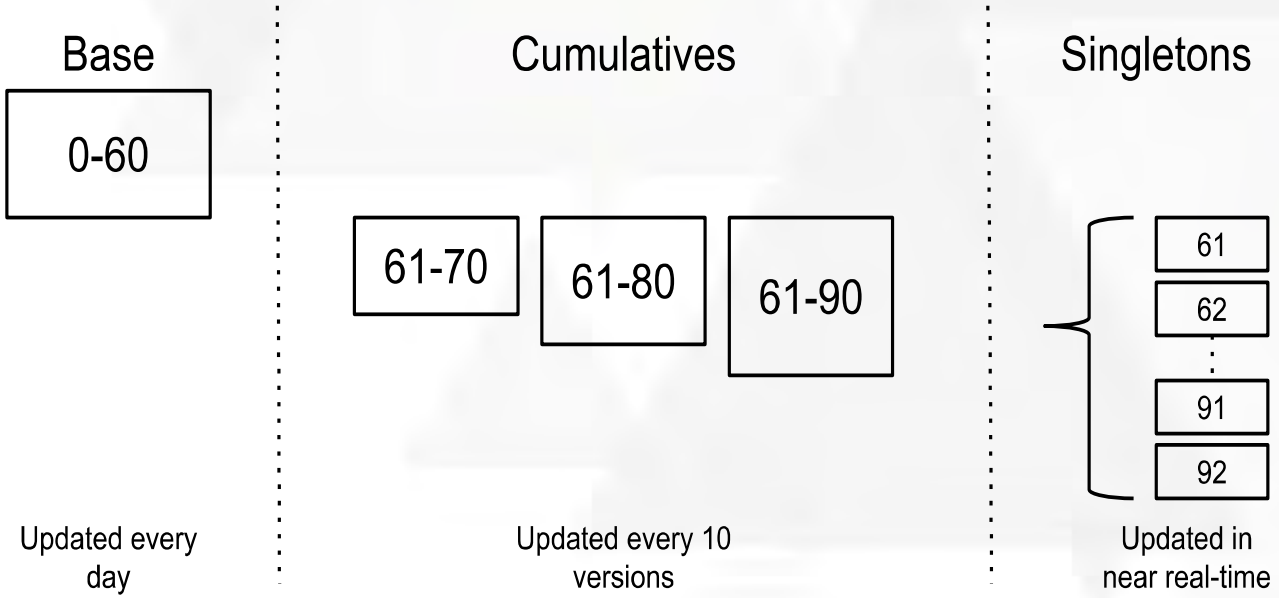
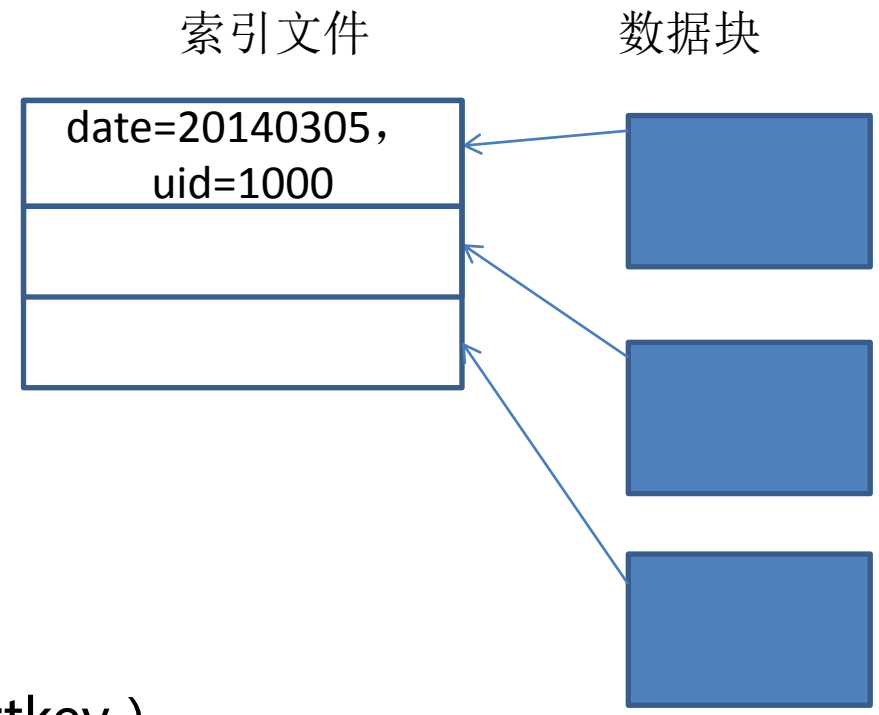


Figure 3: A two level delta compaction policy

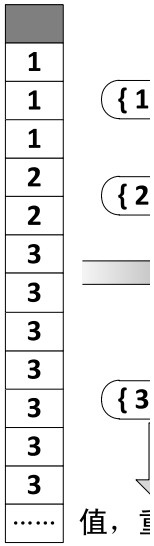
引自Google Mesa Paper

存储格式 – 行列存

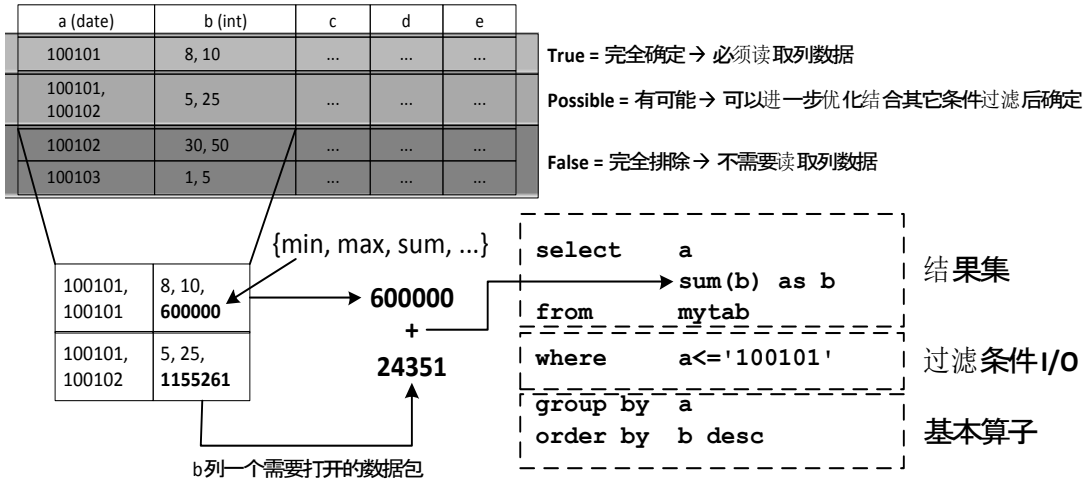
- 数据块存储
  - 每个块包含256行
  - 块内按列存储
  - 块整体压缩
- 稀疏索引
  - 索引驻内存
  - 每个块一个索引项
  - 部分Key作为索引（Shortkey）



存储格式 – 列存储

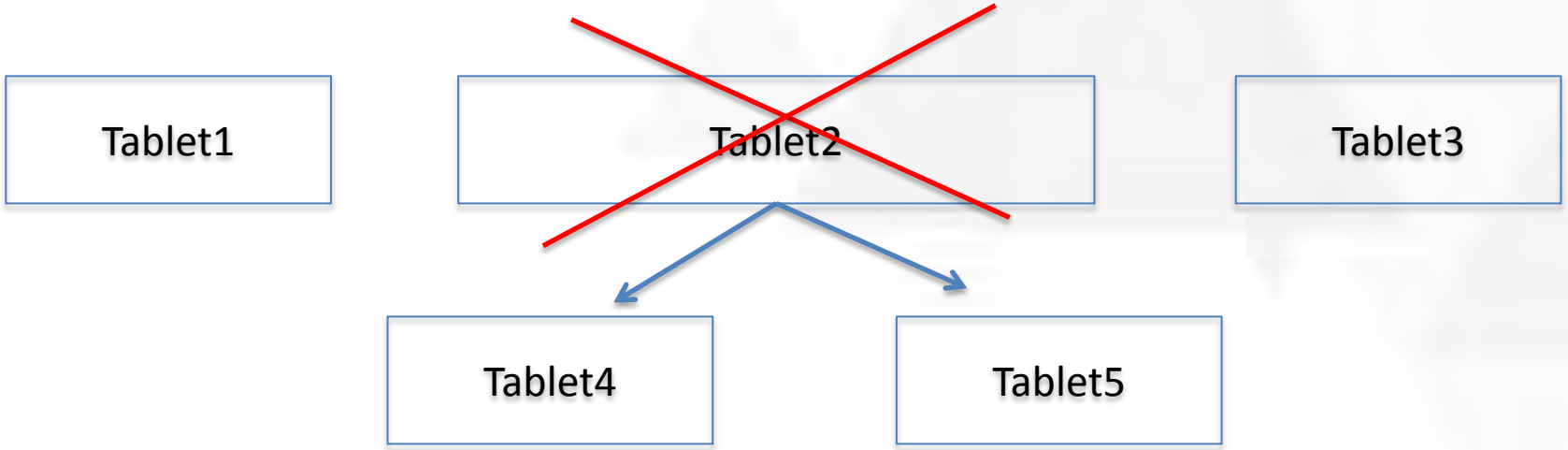


✓ Palo存储引擎利用原始过滤条件以及min、max和sum智能索引技术将数据集查询范围尽可能地缩小，可以大大减少I/O，提升查询性能



数据分区

- 单层分区
  - Hash Partition
  - Elastic Range Partition



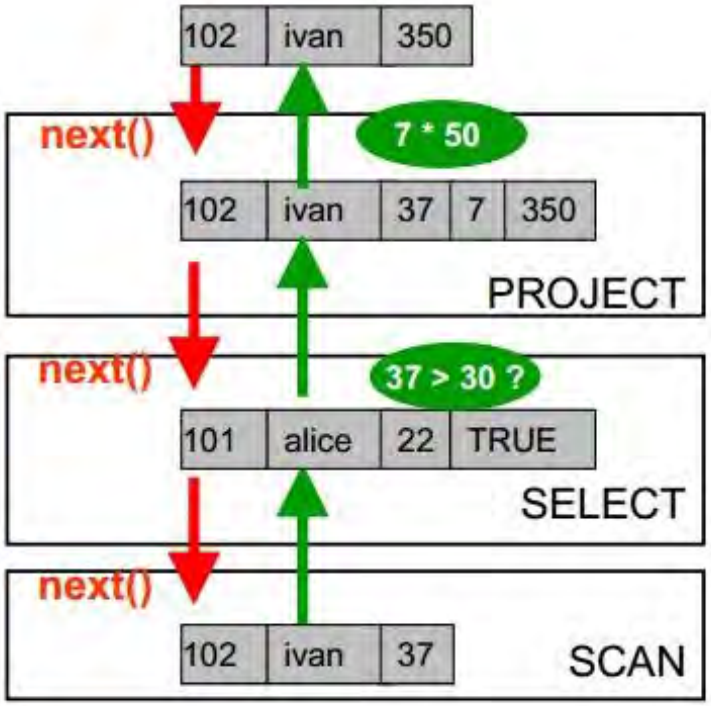
数据分区

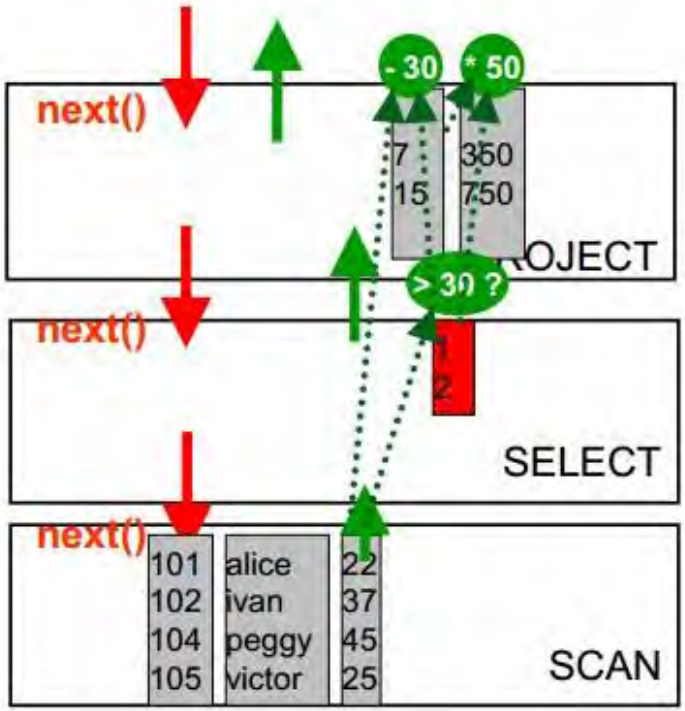
- 单层分区问题
  - 导入性能低，元数据更新频繁
  - 冷数据，会重复进行compaction，浪费资源
  - 冷热数据部方便进行异构存储介质优化（SSD/SATA）
  - 数据删除比较低效
- 复合分区

```
PARTITION BY LIST(date) (
        PARTITION 20150101 VALUES IN (“20150101”),
        PARTITION 20150102 VALUES IN (“20150102”)
)
SUBPARTITION BY HASH(user) PARTITIONS 16;
```

向量执行引擎

```
SELECT  id, name
        (age-30)*50 AS bonus
FROM    employee
WHERE   age > 30
```

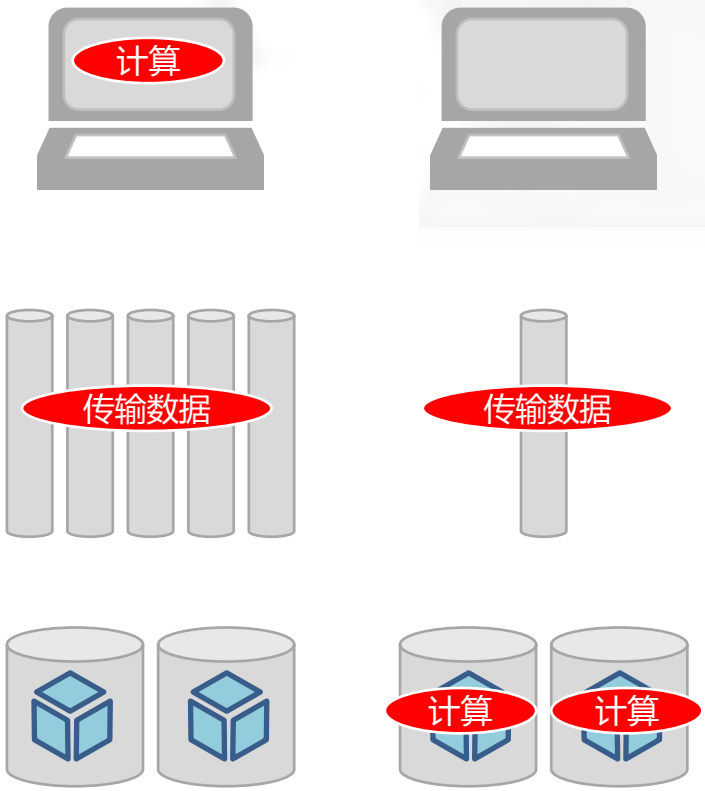




向量执行引擎

- 行式执行引擎的问题
  - 每行一次函数调用，打断CPU流水，不利于分支预测
  - 指令和数据的cache miss
  - 编译器不友好，不利于循环展开，SIMD
- 设计思路
  - 单条处理到批量处理
  - 行式处理到列式处理
- 效果
  - Star Scheme benchmark性能提升2~4倍

库内分析



time-stamp	userid
10:00:00	238909
00:58:24	7656
10:00:24	238909
02:30:33	7656
10:01:23	238909
10:02:40	238909

(a) Raw click data

time-stamp	userid	session
10:00:00	238909	0
10:00:24	238909	0
10:01:23	238909	0
10:02:40	238909	1
00:58:24	7656	0
02:30:33	7656	1

(b) Click data with session information

- UDF
- UDAF
- UDTF

```
SELECT ts, userid, session
FROM sessionize (
        ON clicks
        PARTITION BY userid
        ORDER BY ts
        TIMECOLUMN ('ts')
        TIMEOUT (60)
        );
```

Palo核心技术

- 存储引擎
  - 分布式存储引擎，单表容量可以到百TB~PB
  - 复合分区
  - 列存储，高效压缩和编码，智能索引
  - 小批量导入 + 批量原子提交，MVCC
  - 高效的分布式数据导入
  - 在线create rollup，schema change
  - 完善的分布式管理框架（自动副本均衡，副本修复）

Palo核心技术

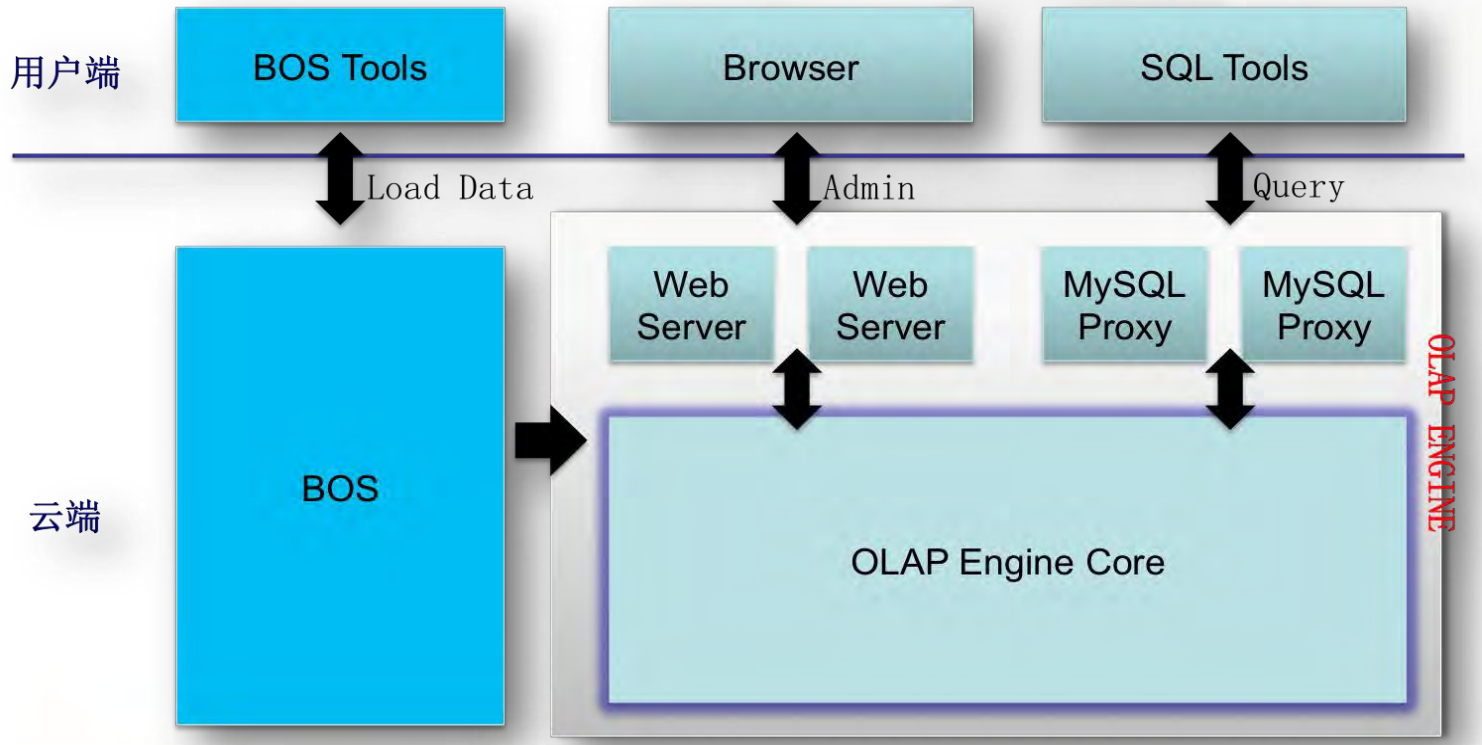
- 查询引擎
  - 实现Mysql网络协议，易用，和各种BI工具无缝对接
  - **Share-noting MPP架构，可扩展性好**
  - 大表分布式join (shuffle和broadcast)
  - 谓词下推，复杂谓词下推
  - **Rollup表智能选择**
  - Partition pruning
  - 向量执行引擎
  - 库内分析
  - 丰富的SQL表达：窗口函数
  - 多租户资源隔离

目录

- OLAP背景介绍
- Palo整体架构
- Palo关键技术
- Palo对外开放

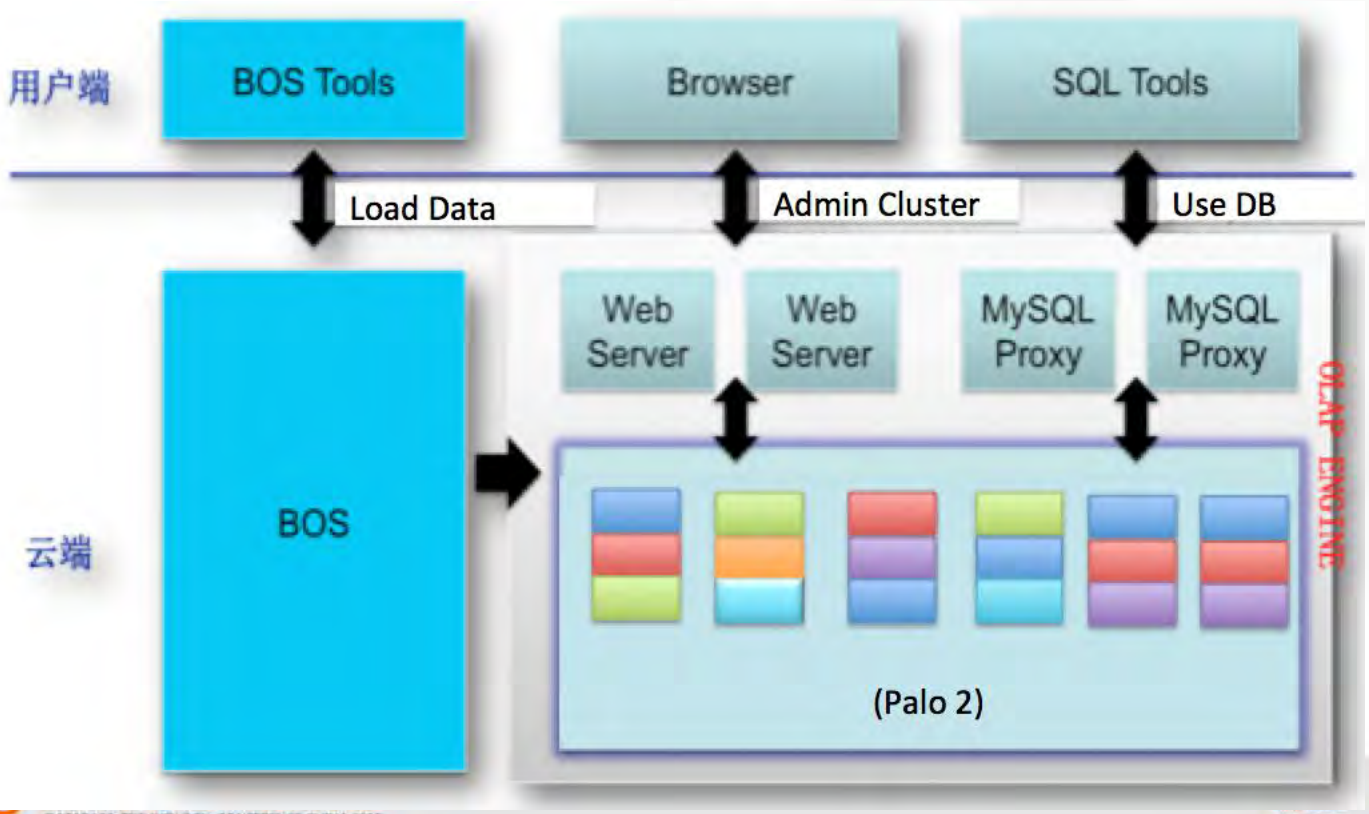
基于Palo的云分析 – OLAP Engine

- OLAP Engine Alpha 2014.9
- BigQuery 模式



基于Palo的云分析 – OLAP Engine

- OLAP Engine Beta 2015下半年，目前正在邀请用户测试
- Redshift模式



开源计划

- 2015下半年，10月左右，开源Palo
- 2016年开源云化框架


ChinaUnix
ITPUB
THANKS

成就直达号的大数据引擎技术专场

百度直达号 http://zhida.baidu.com/

百度开放服务平台 http://developer.baidu.com/

百度开放云 http://bce.baidu.com/