淘宝实时计算平台实践

离哲@淘宝网
@flyinweb

提纲

1.简介

2.现状

3.历史变迁

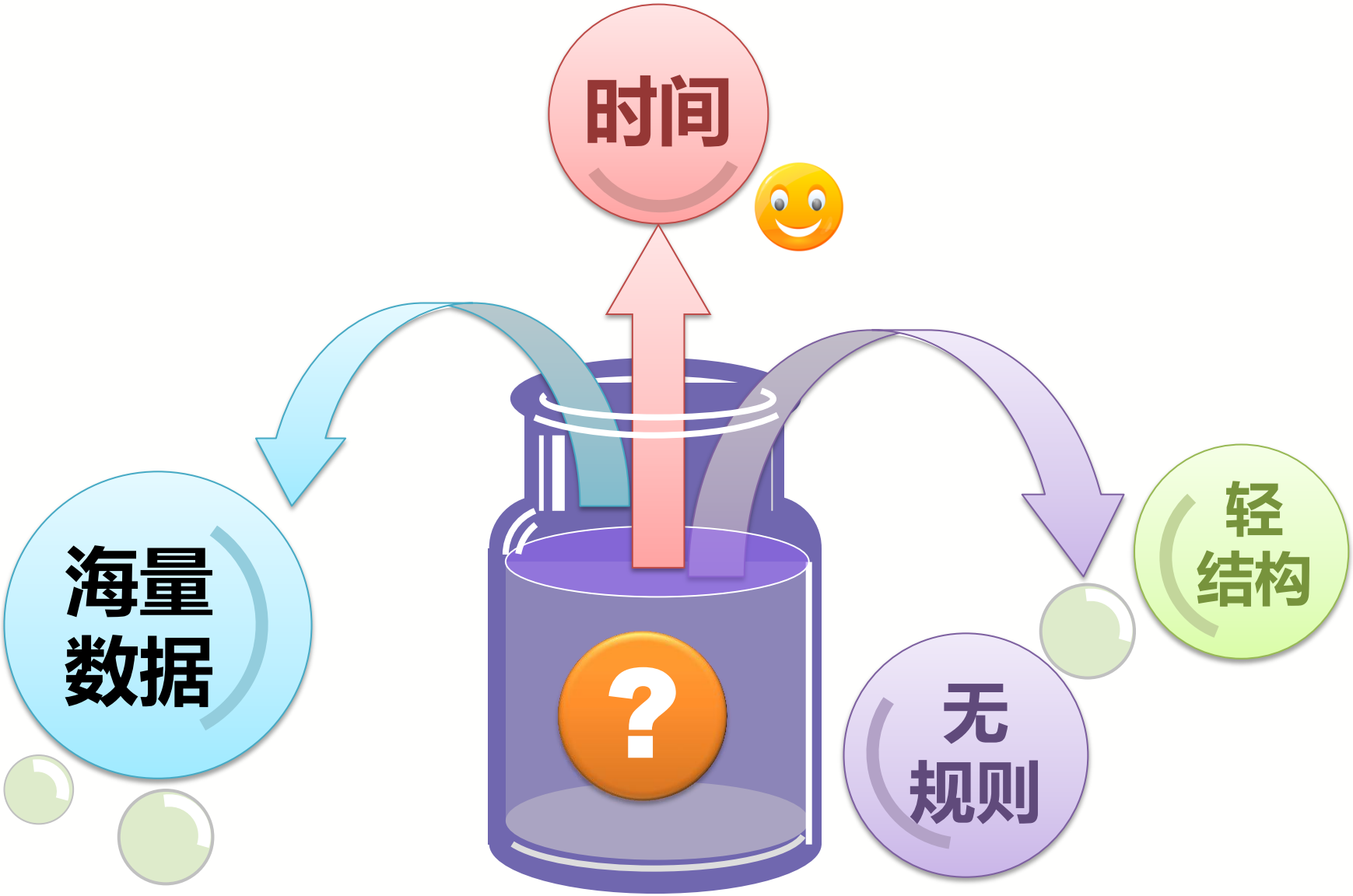
4.架构总览

5.关键特性

6.应用案例

7.未来展望

1.1 背景



1.2 简介

实时计算定义：

针对**历史**数据进行**即时**数据的获取和计算

相关：

RTOLAP(Realtime OLAP)

Grid Computing

In-memory database

METAMARKETS

GIGASPACES

ORACLE TIMESTEN

mobject precision data management

2.1 现状 @ 淘宝

一些数字：

✓ **已接入：**

搜索成交信息 > 5 亿 / 天

活跃用户信息 > 1 亿

➢ **即将接入：**

用户类目偏好信息 > 10 亿

用户品牌偏好信息 > 20 亿

>>>>>>>>>>>>>>>>>

淘宝指数

Taobao DataCube 数据魔方

cool.taobao.com

jianghu.taobao.com

性能：

QPS > 300

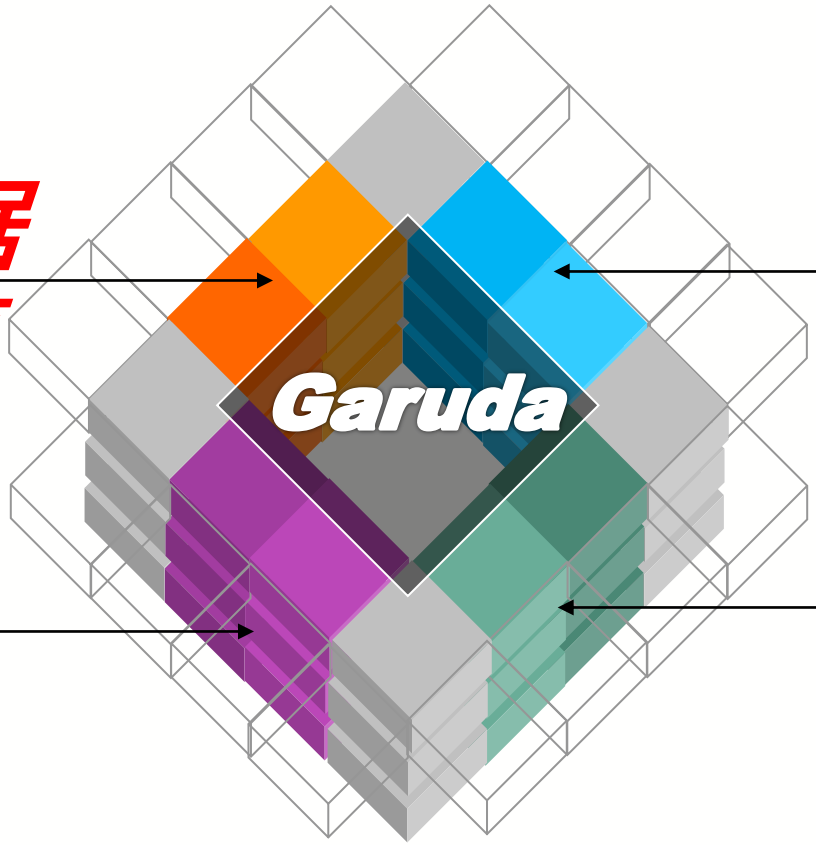
AVG Query Scan Row > 300 万

AVG Query Compute Column > 50

2.2 需求 @ 淘宝

- 海量数据
- 无法预算

- 高并发
- 高可用



- SQL
- Schema Free

- 低延时
- 计算精确

分布式/全索引/内存/数据库

3.1 历史版本一：

Redis集群

冗余ID 列表

分片统计结果

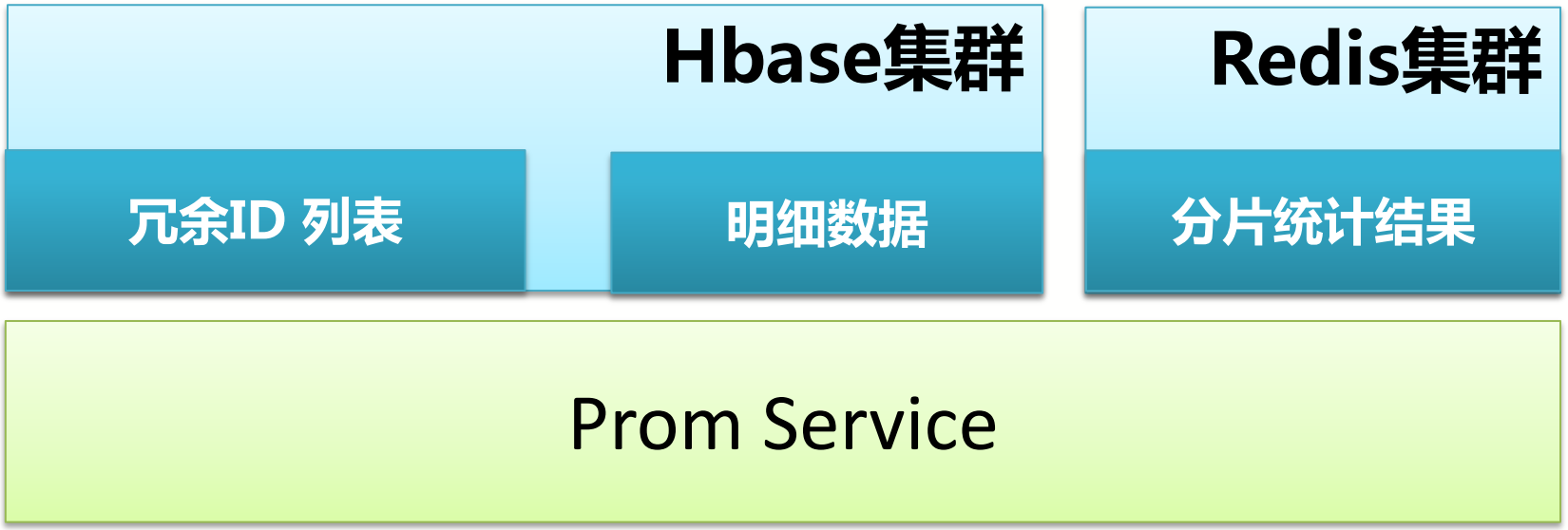
Tokyo Cabinet集群

明细数据

Prom Service

1.冗余度
2.明细数据慢
3.规则变复杂

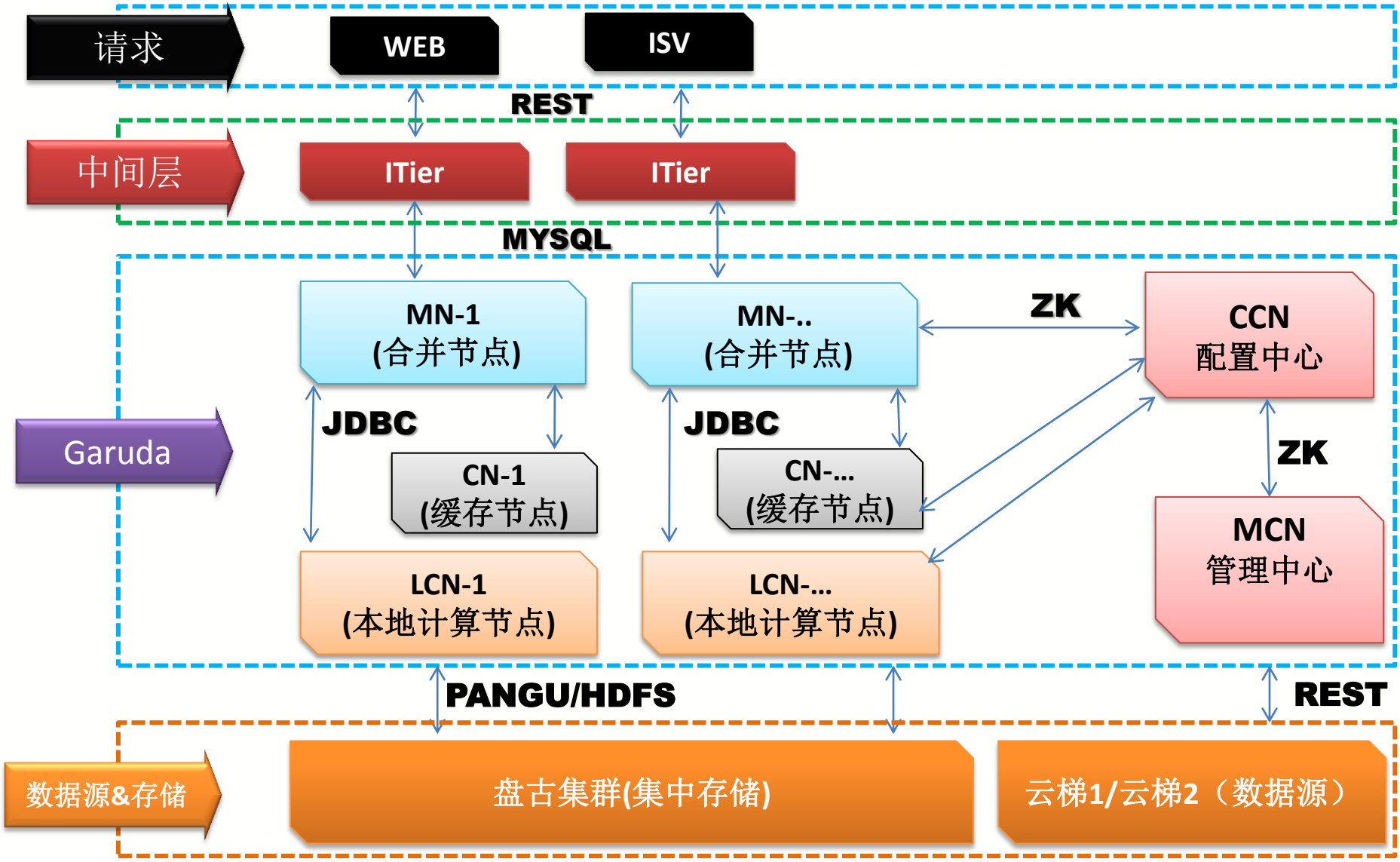
3.2 历史版本二：



1.倒排索引
2.本地计算(Hbase)
3.自动扩容(Hbase)

1.冗余度
2.网络带宽
3.定制性

4.1 架构总览


请求
WEB
ISV
REST
中间层
ITier
ITier
MYSQL
MN-1
(合并节点)
MN-..
(合并节点)
ZK
CCN
配置中心
JDBC
JDBC
Garuda
CN-1
(缓存节点)
CN-...
(缓存节点)
ZK
MCN
管理中心
LCN-1
(本地计算节点)
LCN-...
(本地计算节点)
PANGU/HDFS
REST
数据源&存储
盘古集群(集中存储)
云梯1/云梯2（数据源）

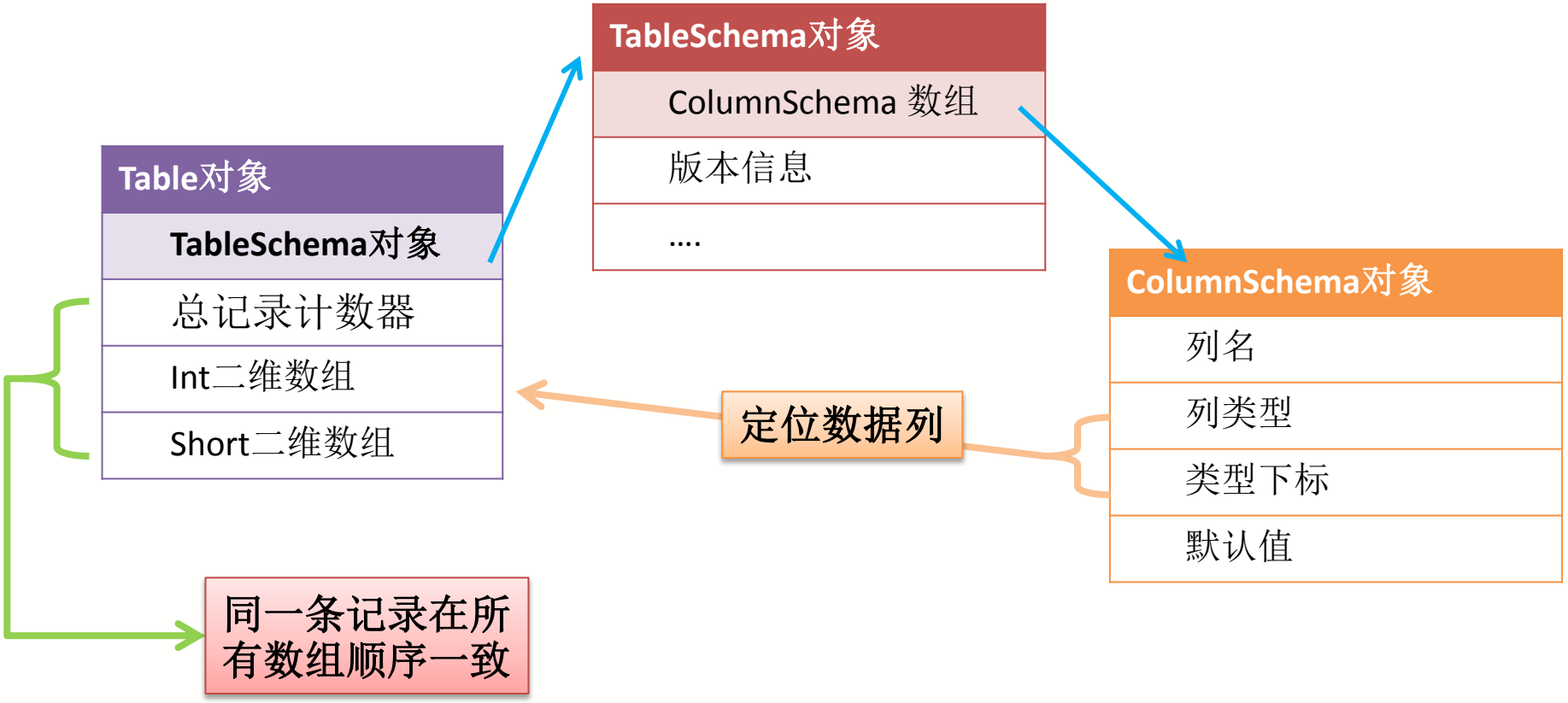
5 关键特性

Fixed/Free Schema（列存储）
Partition/TableGroup
全索引
本地计算
大表Join
缓存
资源管理调度
可用性

5.1 Schema

全部导入/局部导入

列存储



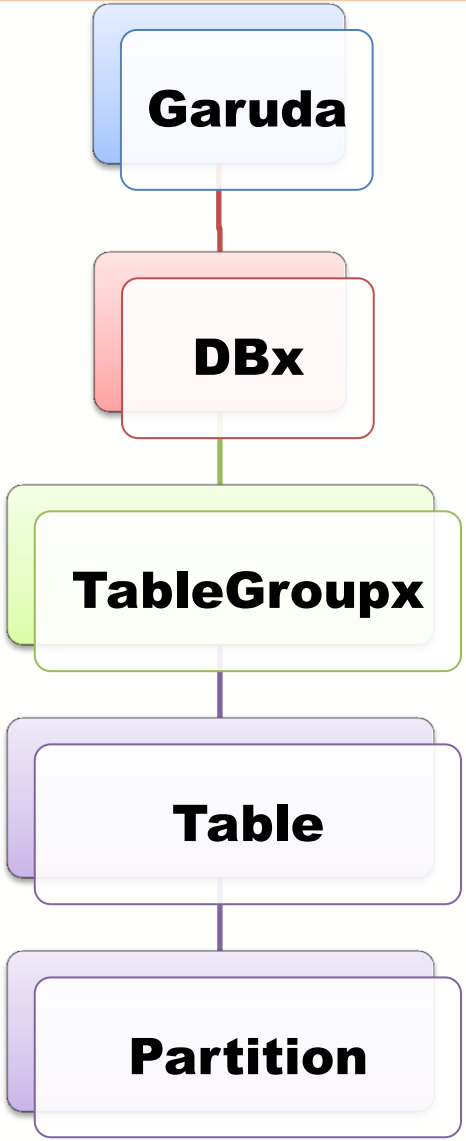
5.2 Partition/TableGroup

Partition

- Interval
- Range
- Hash

TableGroup

- Join
- PartitionGroup



5.3 全索引

- **计算列/索引列(倒置)**
  - 计算列 @ memory
  - 索引列 @ disk
- **索引**
  - Hash
  - B+Tree
  - Skiplist
  - **Bitmap**
- **倒排**
- **压缩**
  - String ?
  - PForDelta(7%)

Index array(abstract)	
tree<T,int[]>	SSD
skiplist<T,int[]>	SSD
hashmap<T,int[]>	SSD
unique<T,int>	memory

5.3 全索引

数据结构	数据集大小(亿条)	每次参与运算数据量(条)	线程数	每请求耗时(ms)	总耗时(ms)	每秒处理记录数
Array	5	200,000	100	10	177	112994350
	5	200,000	1000	5	772	259067358
Hashmap	5	200,000	100	653	1143	17497813
	5	200,000	1000	533	10838	18453589
SkipList	5	200,000	100	28959	41853	477863
	5	200,000	1000	47439	361795	552799
B+ tree	5	200,000	100	3922	6112	3272251
	5	200,000	1000	4261	58458	3421260

特别说明：此为单台 16core E5620 @2.40GHZ,24GB内存的测试结果。

5.4 本地计算

Master：

- ✓ SQL解析
- ✓ 路由分发
- ✓ 结果缓存合并

Localnode

- ✓ SQL解析
- ✓ 索引查找
- ✓ 计算

带宽？

5.4 本地计算

特殊：

- ✓ 跨TableGroup
  大表+小表（batch）
- ✓ 虚拟列
  group by partition_key
- ✓ AVG
  sum/count
- ✓ Limit
  全局limit * 阈值
- ✓ Order by rand() limit n
  rand(cardinality*limit+2)
- ✓ Distinct
  全局 Bitset+Bloom Filter

5.5 大表Join

- **TableGroup：**
  分区Join
- **附属表（支持M:N）：**
  存储：主表内存位置+自身内存位置
  加载：主表增加虚拟列
- **附加索引（支持M:N）：**
  存储：主表内存位置
  只能用来定位和count

SQLPlus | 本地节点SQLPlus | taobaoindex__trade

column信息

id	name	comment	type	primaryKey
0	thedate	日期	DATE	false
1	cat_id	叶子类目id	INT	false
2	order_id	交易ID	Long	true
3	auction_id	商品ID	Long	false
4	alipay_trade_...	交易商品件数	INT	false

分区信息

tableName	start	end	dataSize	n
trade__2012...	20120524	20120525	5759544278	4

附加索引信息

name	type	pkName	valueType
keyword	Hashmap	order_id	STRING

WHERE
keyword contains（‘iphone’，‘智能'）

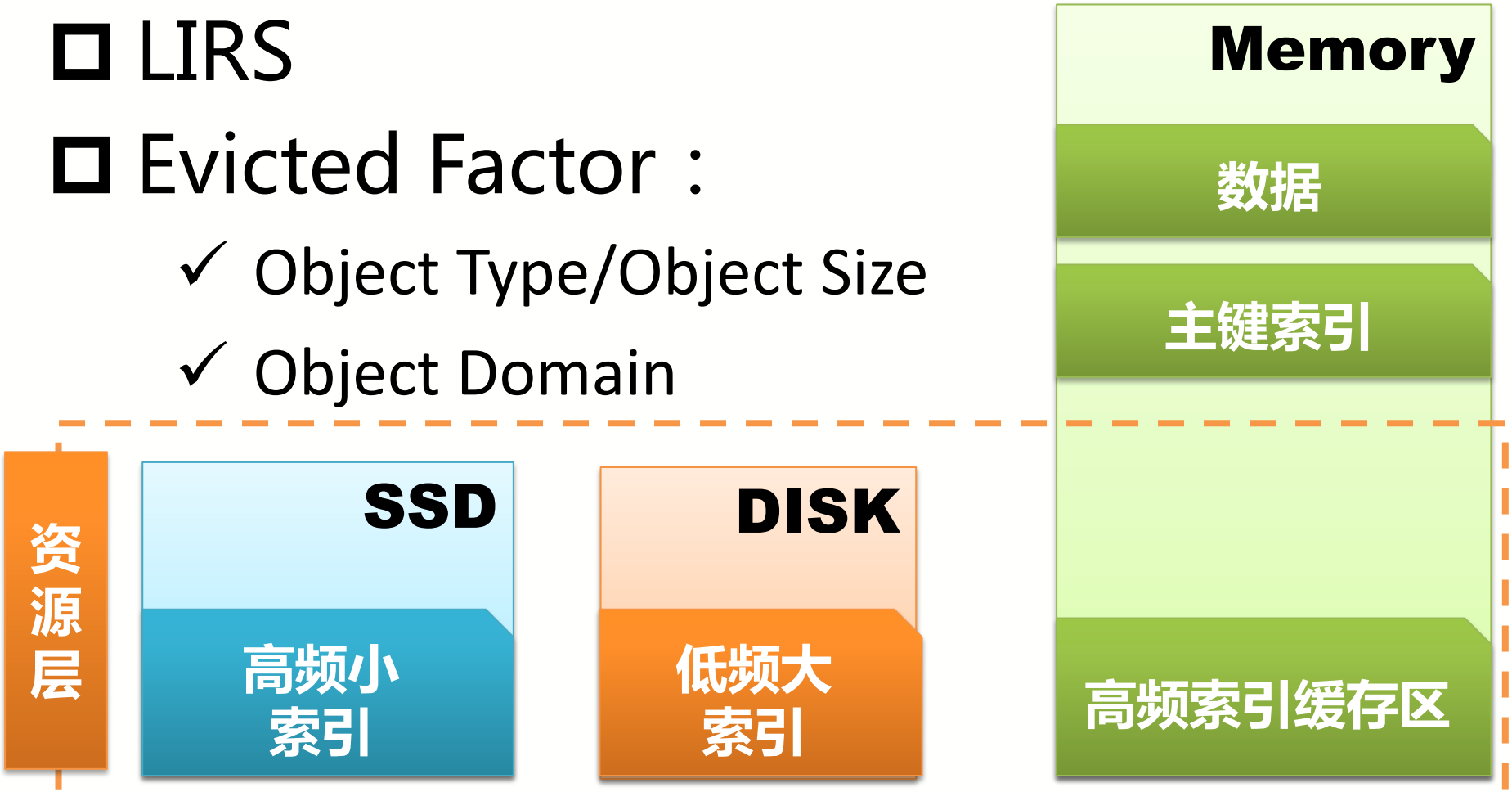
WHERE **cpv in（‘x1'，‘x2'）**

5.5 大表Join

5.6 缓存

本地节点缓存：

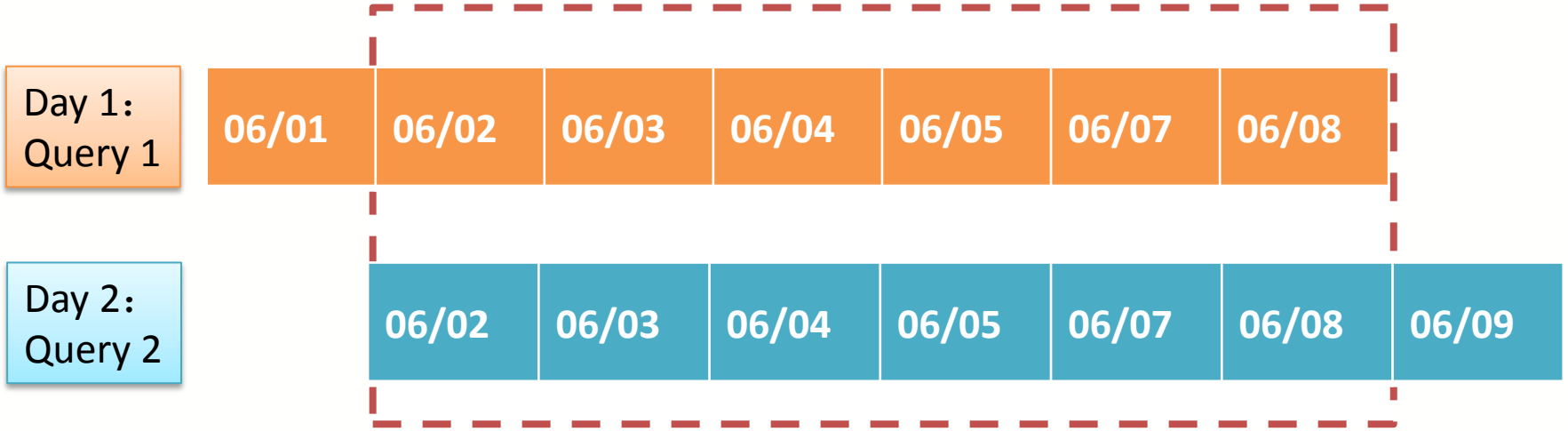
- LIRS
- Evicted Factor：
  - ✓ Object Type/Object Size
  - ✓ Object Domain



5.6 缓存

Master节点缓存：

- LIRS
- SQL cardinality
- Partition result



5.7 资源管理调度

动态规划算法

Monitor 服务器分布式锁（主/备）

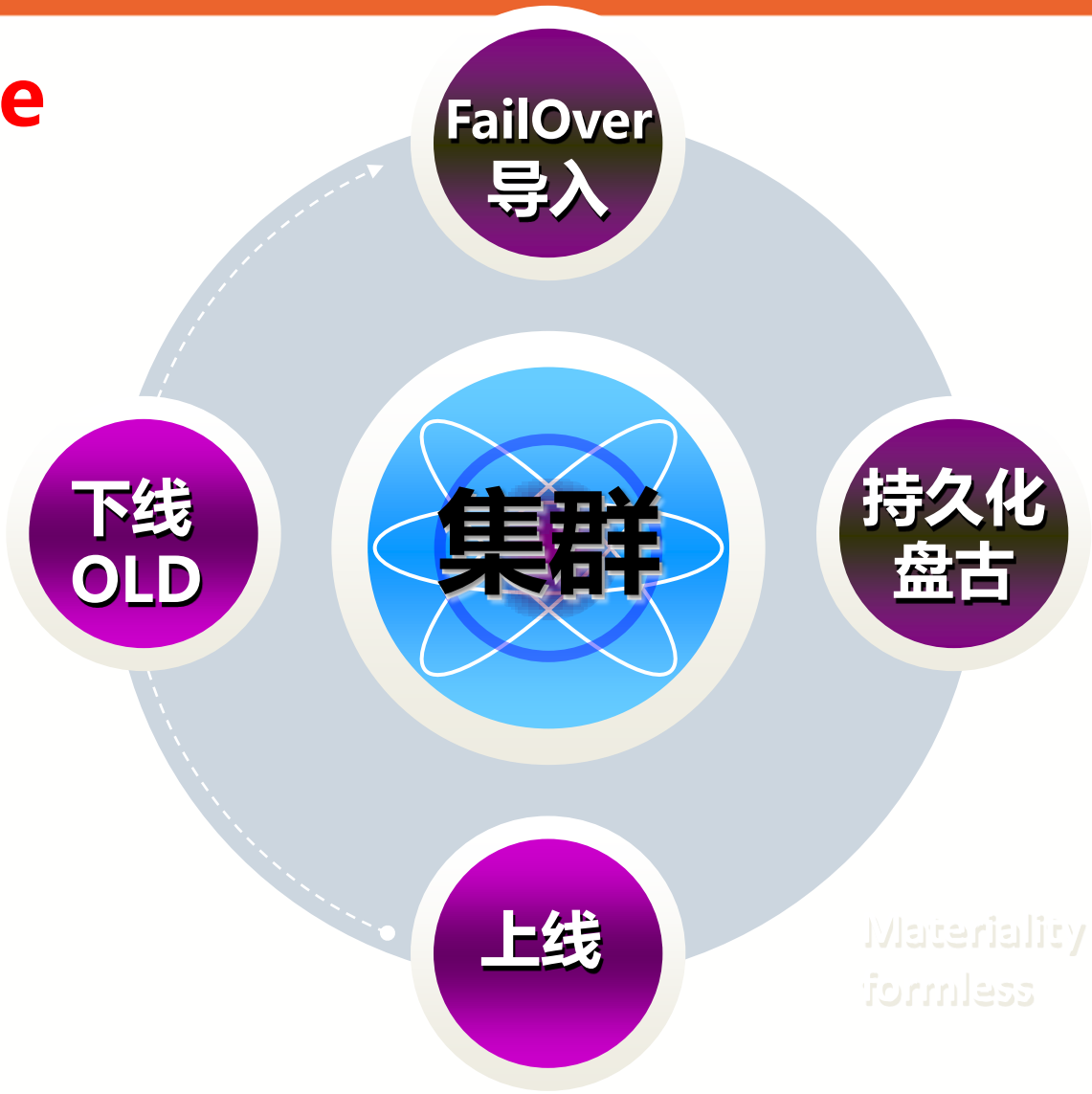
参数：

- 可用内存、可用磁盘（Buffer阈值）
- 每个表占用的内存、磁盘
- 最小可用实例数
- 最小Failover机器数
- 每个分区最小可用份数（在线上集群）
- 每个表最多保留分区数（Rotate）
- 超时设置(上线/下线/导入 超时)
- 表组信息
- 虚拟机组
- ....

5.8 可用性

- Failover Rotate
- 资源虚拟化(T4)

- Heartbeat
- 双机房
- 任务分布式锁
- 任务持久化
- 任务跟踪JobID
- 执行时间监控



6 应用案例

淘宝指数 + Taobao DataCube 数据魔方

- ✓ 交易信息+搜索信息 > 5亿条 * 7 天（即将接入30天）
- ✓ 附加索引 3个
- ✓ 表4张
- ✓ QPS > 300
- ✓ 平均响应时间<2s

6 应用案例

SNS：Jianghu.taobao.com

- ✓ 用户信息>100,000,000
- ✓ 表 1 张
- ✓ QPS > 600
- ✓ 平均响应时间 < 90ms

7 展望

实时数据源（MVCC）
迭代计算
资源隔离(DB/USER/表)
索引离线计算（hadoop）
存储结构优化
SSD优化

谢谢!

Q&A