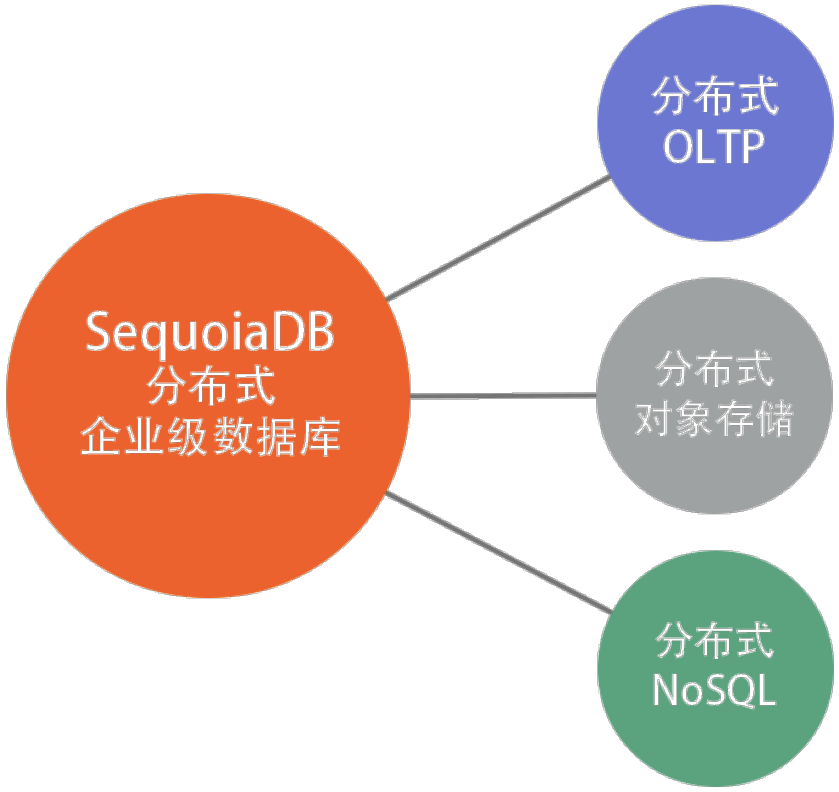
HTAP数据库技术与实践

@Yann

SequoiaDB

SequoiaDB 巨杉数据库

- SequoiaDB（巨杉数据库）
  - 中国第一款自助研发企业级分布式数据库
  - 完全自主研发，数据库引擎没有基于任何开源数据库源代码
  - 核心研发团队来自IBM北美DB2研发团队
- 中国第一款商业开源数据库产品
  - www.github.com/sequoiadb/sequoiadb
  - www.sequoiadb.com



行业认可与奖项

唯一入选 “2016硅谷大数据生态象限图”的中国公司
连续两年获得美国创新媒体《红鲱鱼》的“全球创新企业100强”
连续三年获评为美国科技媒体《快公司》 “中国50大创新公司”

RED HERRING GLOBAL WINNER 100

“全球创新企业Top100”
——《红鲱鱼》
美国最具影响力商业媒体

FAST COMPANY 50 MOST INNOVATIVE COMPANIES

“中国创新企业50强”
——《快公司》
美国著名创新媒体

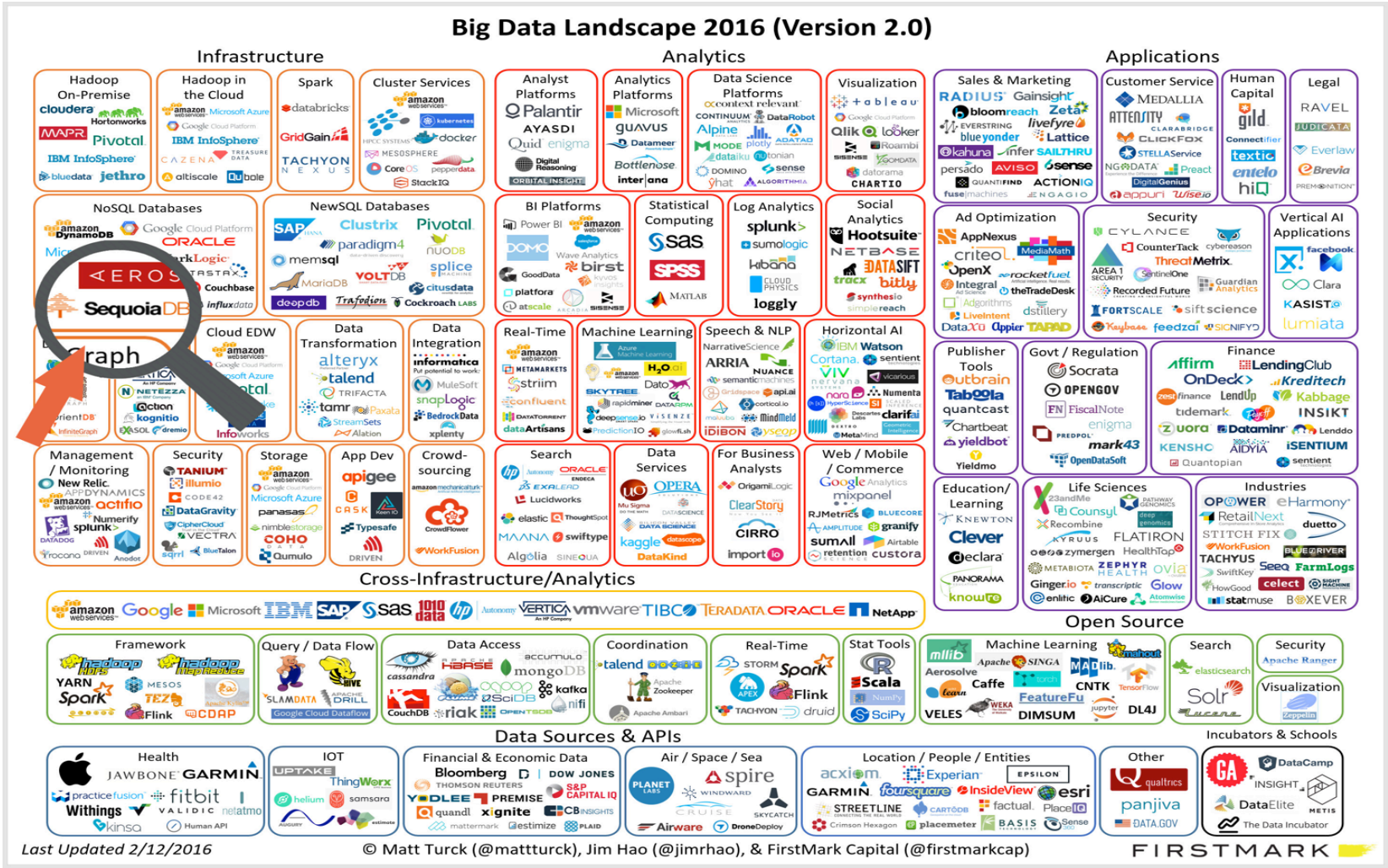
中国开源软件推进联盟（COPU）颁发的“2015年度优秀开源项目”奖
中国电子信息产业发展研究院评选的“2015中国金服务.数据库领域最佳产品”奖

联合创始人兼CTO – 王涛

- 入选“快公司”评选的“2014中国商业年度最具创意人物100”榜单
- 2014年获得国内知名IT门户 CSDN评选的， “TOP 50 最具价值CTO”奖项

2017年6月 参加Spark峰会
发布技术演讲

SPARK SUMMIT 2017
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 5-7 | MOSCONE CENTER | SAN FRANCISCO


Big Data Landscape 2016 (Version 2.0)
Infrastructure
Analytics
Applications
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
Sales & Marketing
Customer Service
Human Capital
Legal
NoSQL Databases
NewSQL Databases
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Ad Optimization
Security
Vertical AI Applications
Cloud EDW
Data Transformation
Data Integration
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Publisher Tools
Govt / Regulation
Finance
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
Education / Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other
Incubators & Schools
Last Updated 2/12/2016
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
FIRSTMARK

HTAP能力

HTAP这种事务和分析并存的数据库能力需求旺盛，但是这越来越依赖于一体化基于数据库分析能力（In-DBMS Analytics），特别是对于实时以及过程处理中的数据分析要求。这种分析能力和目前大数据的数据分析和挖掘不同，强调更多的实时性和交互操作能力。HTAP因此也划分成“决策点HTAP”和“加工过程HTAP”。

分析的场景

事后分析　　事前 / 事中分析

实时分析　　流处理

MapReduce ? 

SQL + Join 

- 基础：数据管理稳定性与可靠性；
- 性能：高并发、实时性的数据读写、查询能力；
- OLTP能力：事务处理能力；
- HTAP综合能力：同时应对事务处理与分析处理；
- 扩展性：海量数据的容量扩展和管理能力；
- 运维便利性：智能集中式的运维管理管理；
- 开发便利性：SQL的支持；
- 多模：结构化、半结构化、非结构化；

SequoiaDB

分布式数据库架构（早于Raft实现）

多模 Multi-Model

SQL | SparkSQL

分布式OLTP	分布式对象存储	分布式NoSQL
• 完整 ACID • MVCC	• 高吞吐量 • 高并发	• 性能 • 高并发

巨杉数据库 -- 分布式数据库HTAP实现

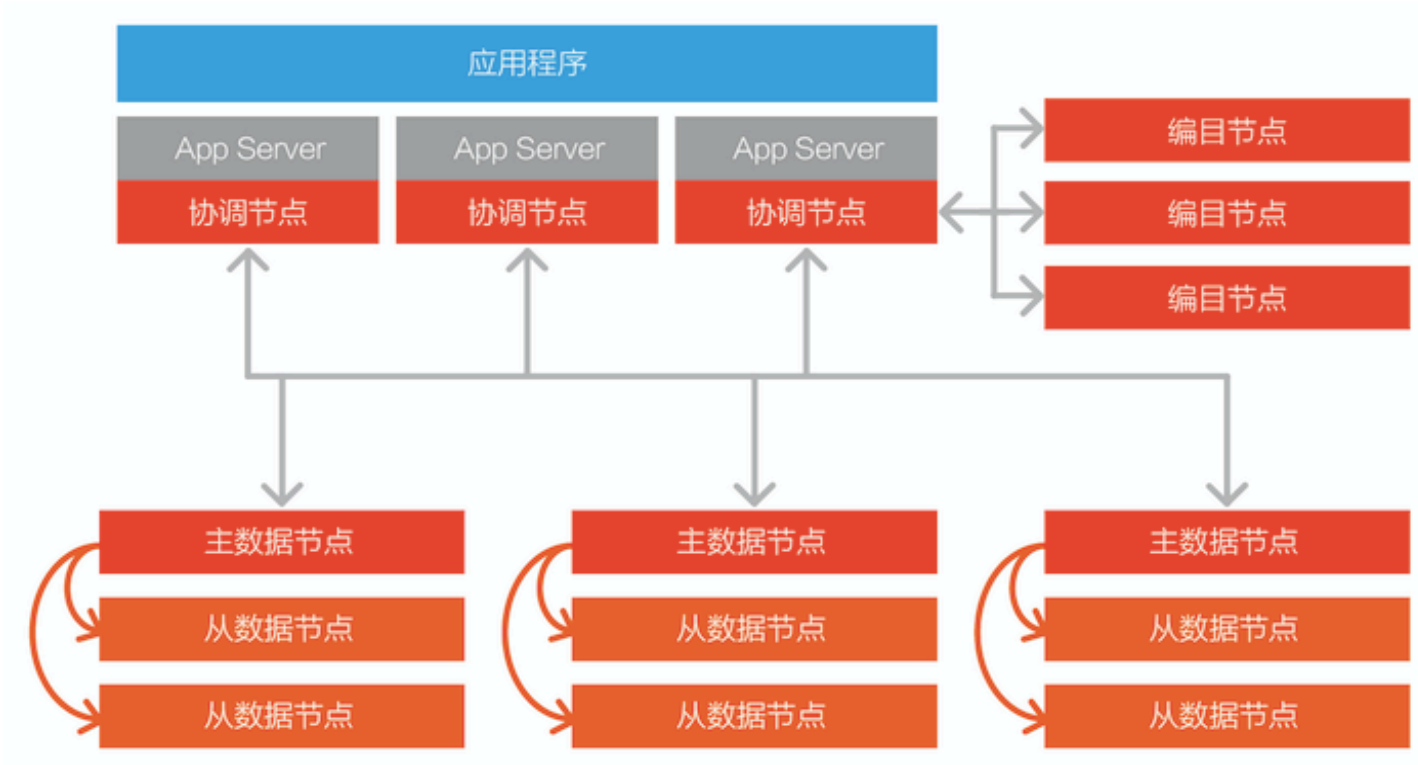
SequoiaDB

SequoiaDB分布式数据库架构

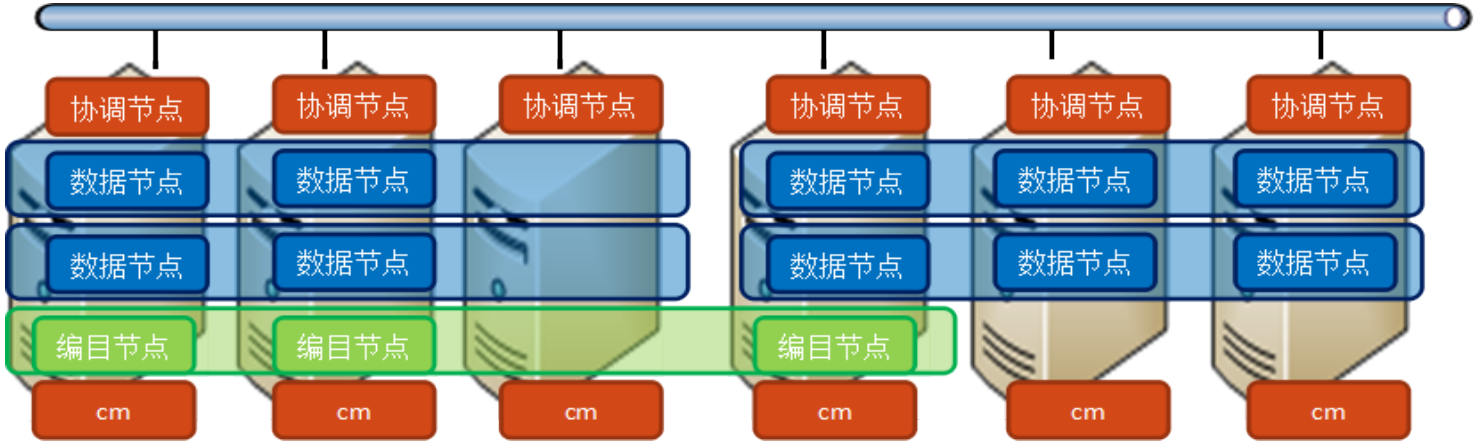
计算分布

+

存储分布



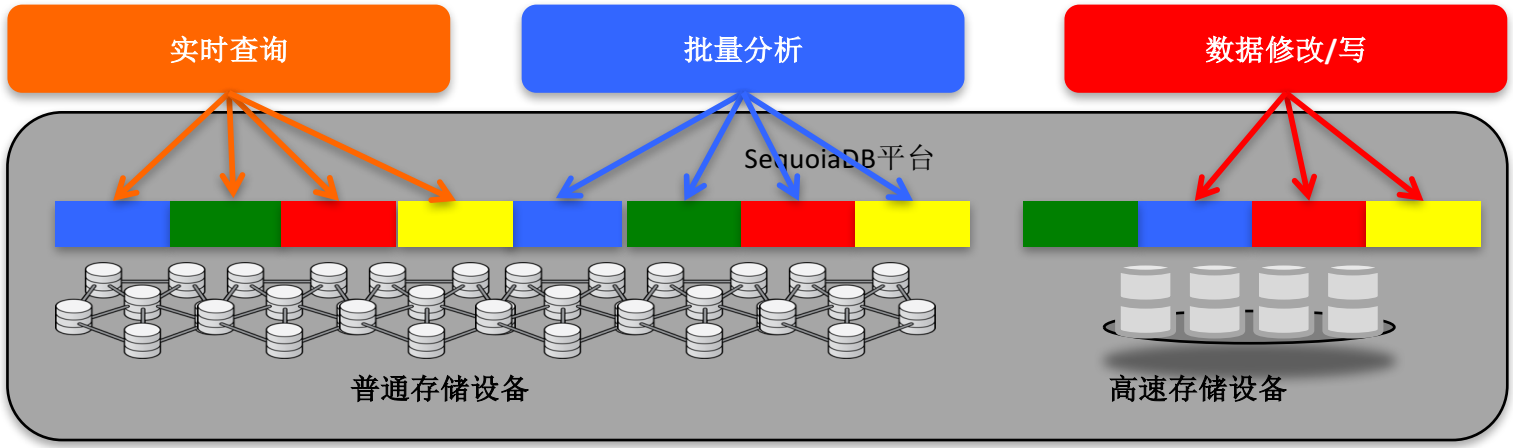
SequoiaDB物理架构



角色	功能
协调节点	胖客户层，从编目读取数据分布信息，从数据节点读取数据
编目节点	负责元数据信息存储，包括组信息、表切割信息
数据节点	负责数据表存储，提供查询、聚集、数据复制功能
CM节点	负责集群管理，包括watchdog, 节点增删启停

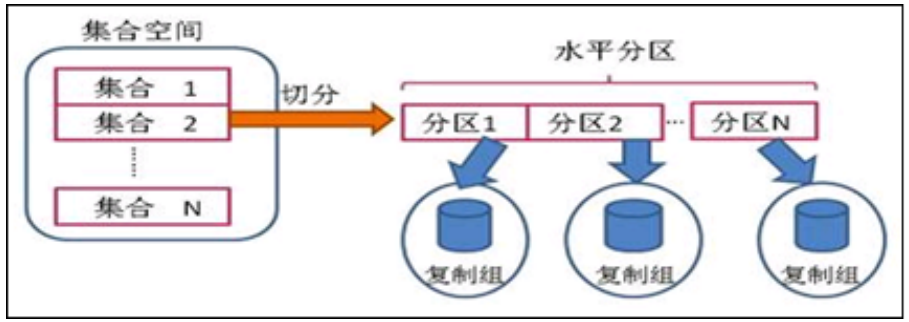
分布式架构优化：读写分离机制

- 数据在多个分布节点内自动复制，并实现写请求和读请求的自动分离，避免读请求对数据写入的影响。
- 此外，可进一步定制数据分布策略，保证不同类型业务可以运行在同一平台上，但同时又不会互相干扰，比如：
  - 冷/热数据区分离
  - 写交易的“强一致性”和“弱一致性”分离
  - 查询/批量分离

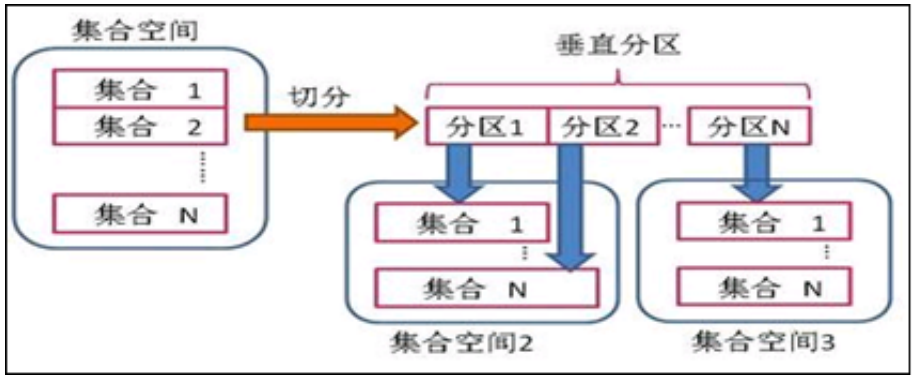


分布式架构优化：数据多维分区

SequoiaDB支持水平分区和垂直分区。水平分区尽可能选择唯一性较高的字段，垂直分区尽可能选择时间或区域这种相关性较高的字段。一个表可以同时为水平分区与垂直分区



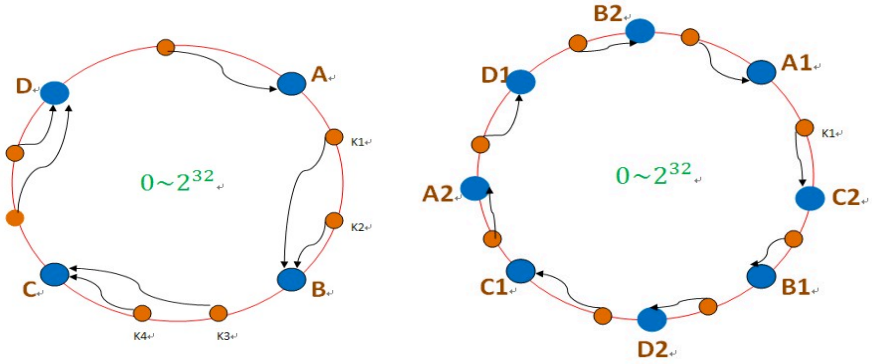
分别适合流水数据与快照数据



优势： 容量和性能可线性扩展

分布式架构优化：一致性散列机制

- SequoiaDB可以指定自定义分区键，不指定分区键的情况下使用记录ID作为分区键，可以保证数据随机散列分布；
- 特殊情况下，当出现热点数据时，热点数据所在分区可以使用split命令进行切分，进一步细化粒度减少热点；
- 切分过程当中为全在线操作，对业务无感知；
- SequoiaDB支持多维数据分区的机制，在大容量磁盘的配置中性能表现最佳。



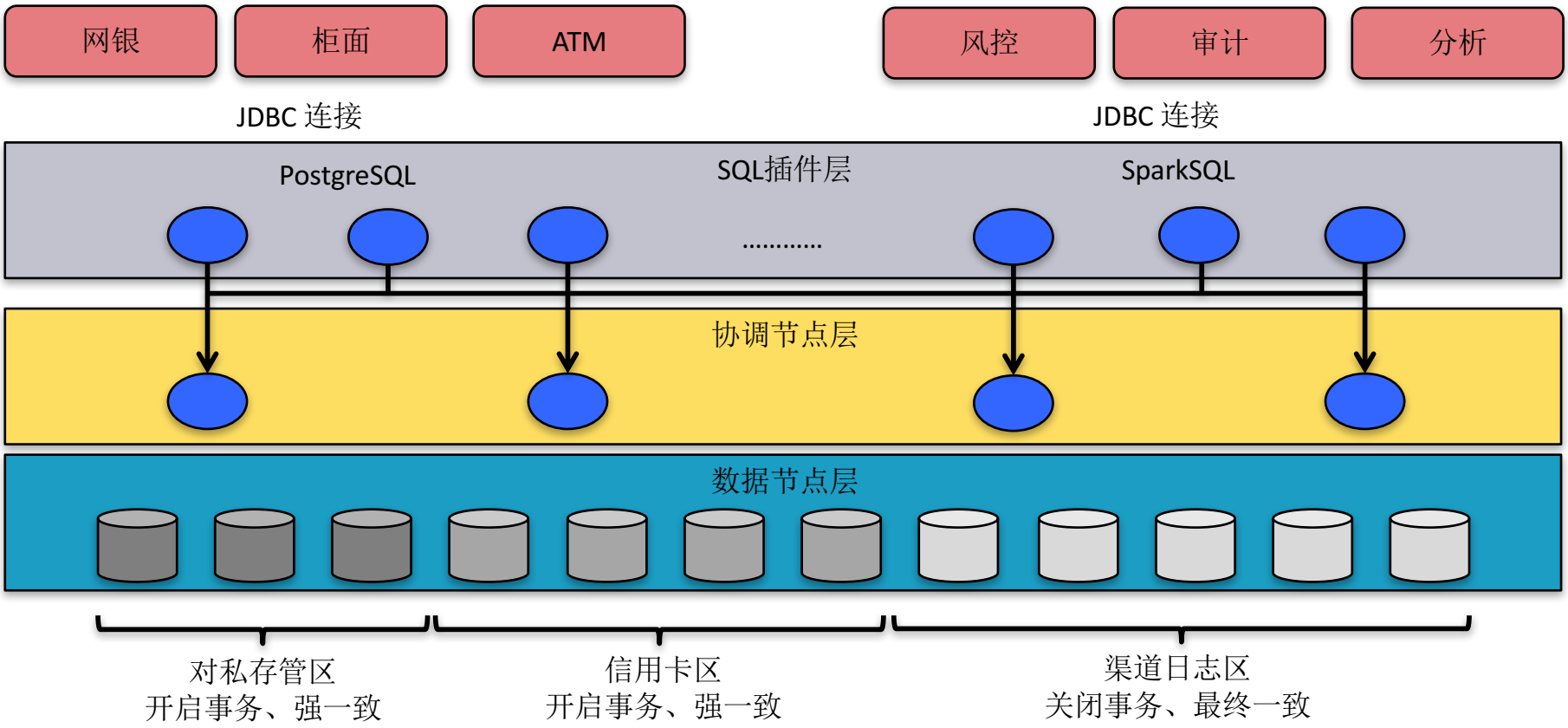
一致性散列：只需移动少量数据即可完成切分，不需全部数据导入导出。

分布式架构优化：可配置强弱一致性机制

一致性因素	读主节点	读备节点
W=N	强一致，无数据丢失，事务操作符合传统关系型数据库模型	强一致，无数据丢失，读操作受节点分布影响会有脏读时间
W!=N	强一致型，极端情况下可能发生数据丢失，事务操作符合传统关系型数据库模型	最终一致，可能发生读取不到写入的数据，极端情况下可能发生数据丢失

持久性因素	备节点响应策略	数据可靠性
物理同步	数据在备节点写入事务日志	全部节点宕机不会造成数据丢失，但会损失性能
逻辑同步	数据在备节点处理但未写入事务	主备节点同时宕机可能会造成数据丢失，数据查询强一致
半同步	数据在备节点成功接收但还未处理	备节点宕机可能会造成数据丢失，最终一致性
异步	数据不需要被备节点感知	主节点宕机可能会造成数据丢失

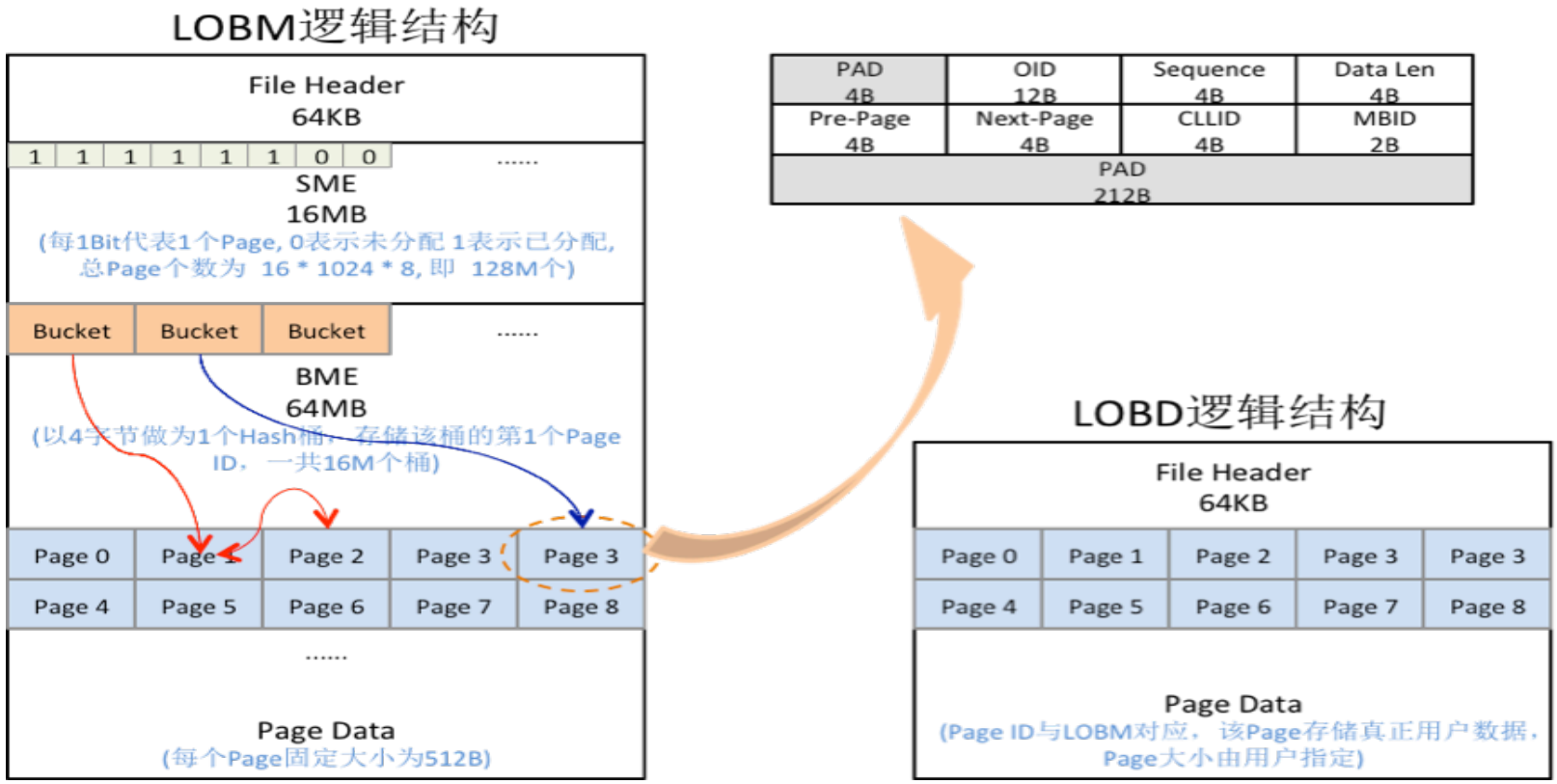
分布式架构优化：SQL与存储引擎隔离



分布式架构优化：分布式事务机制

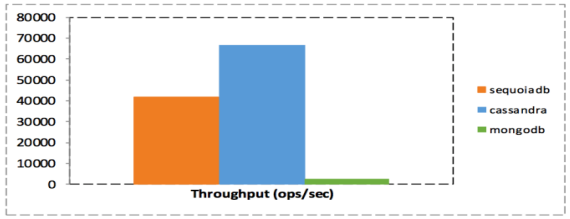
- 二段提交
- 协调节点首先发起预提交
- 当所有数据节点响应成功后，进行统一提交

引擎内部优化：非结构化块存储机制

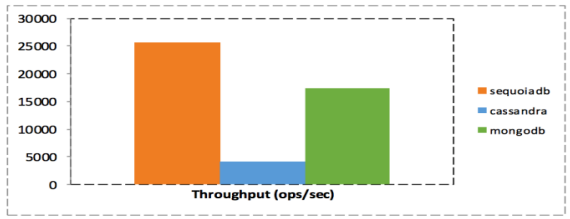


第三方性能对比（对比MongoDB、Cassandra）

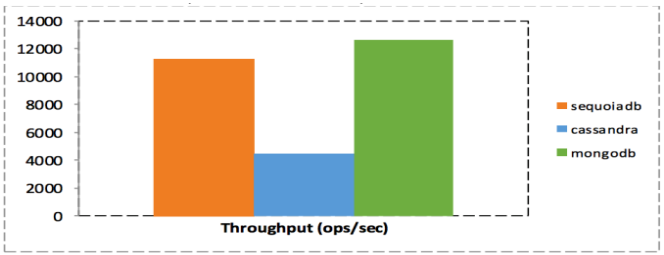
巨杉数据库的技术在业界领先，如今的SequoiaDB 2.6版本更是在各项企业级功能上超越了硅谷同类产品。同时，对比众多硅谷的同类产品，SequoiaDB巨杉数据库在各项性能指标都保持绝对领先。



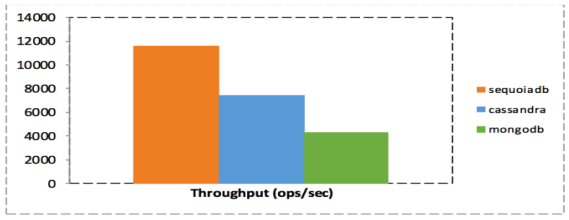
100%写入



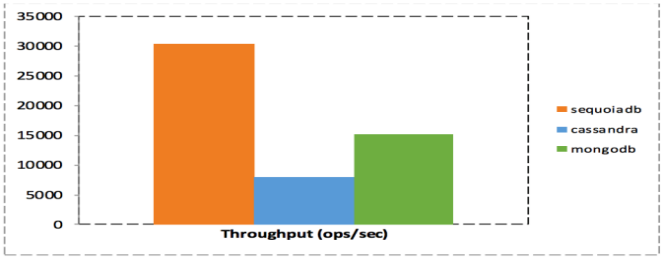
95%读
5%更新



100%读



95%读
5%写入



50%读
50%更新

http://www.bankmark.de/wp-content/uploads/2014/12/bankmark-20141201-WP-NoSQLBenchmark.pdf

第三方性能对比（中国标准测评中心）

- 100并发写入

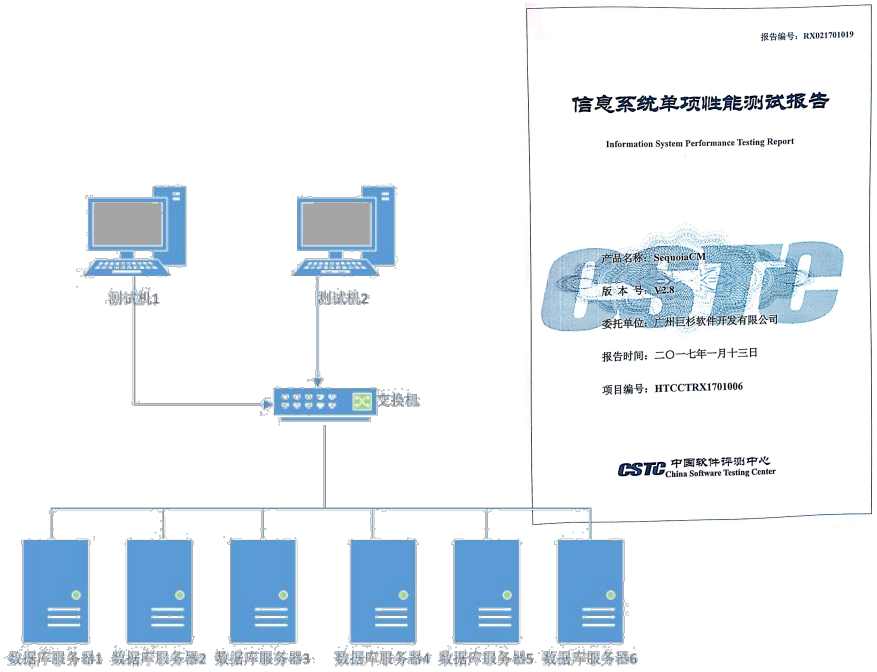
吞吐量（MB/秒）	50KB	200KB	1MB	10MB
整个集群（6节点）	353.3	1232.8	2223.3	2077.5
平均单物理节点	58.9	205.5	370.5	346.3
数据写入平均时延（ms）	14	16	44	466

- 100并发读取

吞吐量（MB/秒）	50KB	200KB	1MB	10MB
整个集群（6节点）	352.9	1225.1	2410.9	2653.3
平均单物理节点	58.8	204.2	401.8	442.2
数据读取平均时延（ms）	14	16	41	369

- 100并发（50个写入、50个读取）

吞吐量（MB/秒）	50KB	200KB	1MB	10MB
整个集群（6节点）	363.7	1266.0	1872.8	2323.8
平均单物理节点	60.6	211.0	312.1	387.3
数据存取平均时延（ms）	写：11 读：16	写：14 读：15	写：35 读：106	写：293 读：749



型号：DELL Power Edge R730xd
CPU：Intel Xeon E5-2650 V4 2.20GHz*2
内存：256GB
硬盘：SAS盘3.6TB*6 SSD盘1.2TB*6
网卡：BroadcomCorporation NetXtreme II BCM57800
1/10 Gigabit Ethernet

谢谢！