BDTC 2017 中国大数据技术大会
Big Data Technology 2017

SequoiaDB 巨杉数据库

金融级分布式架构专场

青云 QING CLOUD

Face++ 旷视

什么是“金融级”？

500亿

6亿

4/8/25

Gartner®

2017年数据库厂商推荐报告

Other Vendors to Consider for Operational DBMSs

Alibaba Cloud aliyun.com

传统交易型数据库云

SequoiaDB 巨杉数据库

分布式多模数据库

GBASE®

传统关系型数据库

VOLTDB　mongoDB.　redis　cloudera　HORTONWORKS　Couchbase　MariaDB

BDTC 2017 中国大数据技术大会
Big Data Technology 2017

SequoiaDB 巨杉数据库

金融级分布式数据库发展

- 广州巨杉软件开发有限公司－简称：巨杉软件
  - 成立于2011年，专注于新一代企业大数据平台研发
- 核心产品：SequoiaDB（巨杉数据库）
  - 中国第一款新一代分布式数据库
  - 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
- 与Cloudera, Databricks建立战略合作
  - 获得CDH、Spark产品认证发行权，并嵌入自有大数据产品中

- 2017 中国首次进入Gartner推荐厂商之一
- 唯一入选“2016硅谷大数据生态地形图”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”
- 2017 猎云最佳企业服务商


数据孤岛问题
性能问题海量/高并发/弹性
两地三中心容灾
数据集约化精细化
大数据操作/分析
攻
守
降低TCO（总体拥有成本）
可靠
可控
可扩

从金融级用户角度来看

从技术产品角度来看

物理隔离、逻辑隔离、
权限、用户、加密

两地三中心、同城双活、一致性

HTAP、实时处理能力、大数据

读写分离、数据域管理

原厂、原生态分布式管理

多模、多引擎

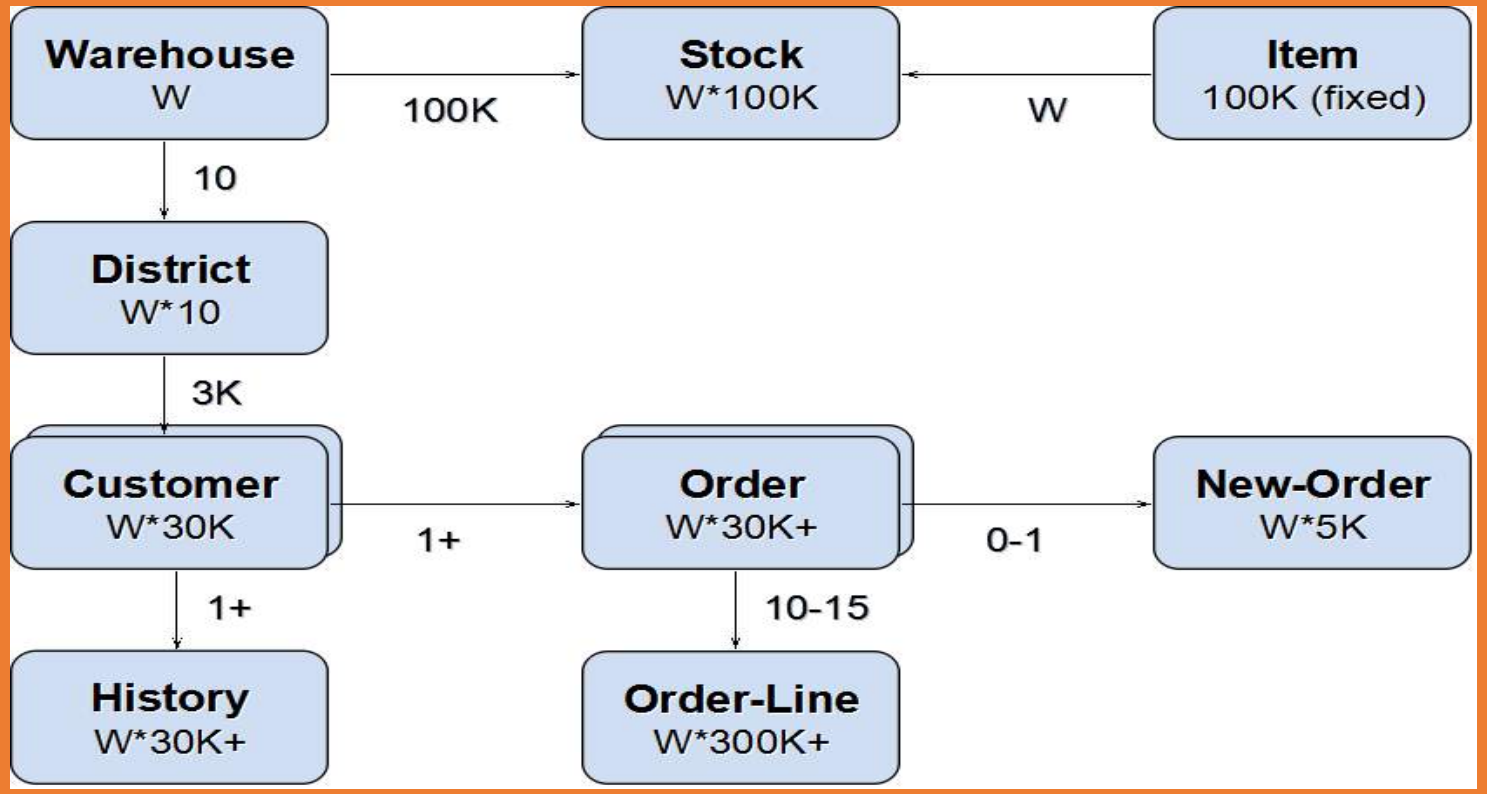
分布式架构

数据的应用范畴

- **交易型数据库**
  - Transactional DB
  - 传统OLTP业务
- **分析型数据库**
  - Analytical DB
  - 分析报表型业务
- **联机型数据库**
  - Operational DB
  - 在线高并发非交易型业务

交易型性能测试

- 测评方式
  - TPC-C测试结果
- 读写性能
  - 随机读操作性能
  - 随机写操作性能
  - 随机更新性能
- 事务相关
  - 提交回滚开销
- 一致性相关
  - 悲观/乐观锁
  - 主外键能力
  - 触发器能力
- 可靠性
  - ACID
  - 高可用性
  - 准实时灾备功能
  - 备份恢复

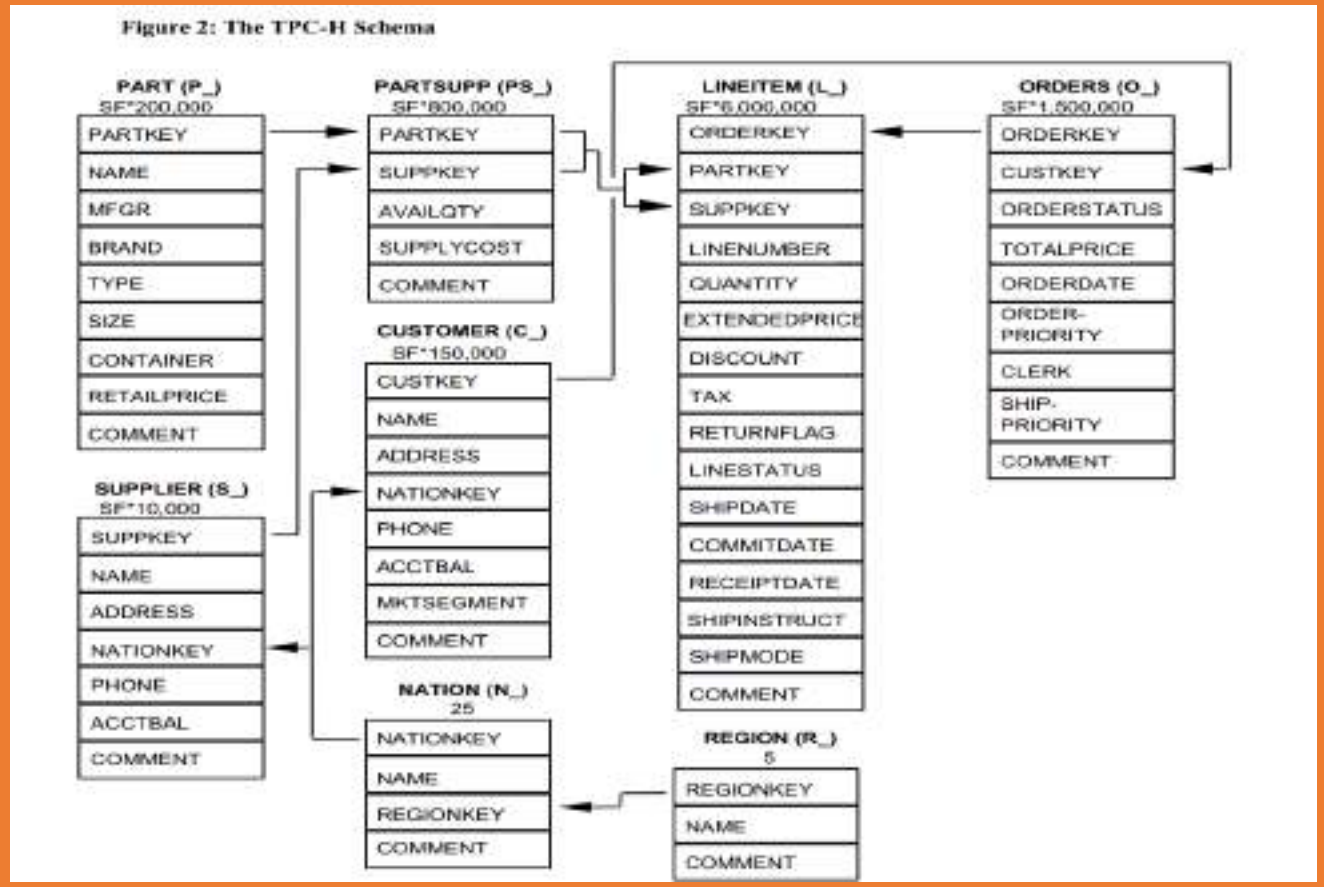


分析型性能测试

测评方式
- TPC-H测试结果
- TPC-DS测试结果

读写性能
- 批量读操作性能
- 批量装载操作性能
- 大表关联性能
- 大量数据聚集性能

批处理加工
- 存储过程能力



联机型性能测试

- 测评方式
  - YCSB测试结果
- 读写性能
  - 随机读性能
  - 随机写性能
  - 批量写性能
  - 随机更新性能
  - 并发能力
  - 可扩展性
- 可靠性
  - 高可用性
  - 准实时灾备功能
  - 备份恢复
  - 强一致性与最终一致性

Workload	Operations	Record selection	Application example
A—Update heavy	Read: 50% Update: 50%	Zipfian	Session store recording recent actions in a user session
B—Read heavy	Read: 95% Update: 5%	Zipfian	Photo tagging; add a tag is an update, but most operations are to read tags
C—Read only	Read: 100%	Zipfian	User profile cache, where profiles are constructed elsewhere (e.g., Hadoop)
D—Read latest	Read: 95% Insert: 5%	Latest	User status updates; people want to read the latest statuses
E—Short ranges	Scan: 95% Insert: 5%	Zipfian/Uniform*	Threaded conversations, where each scan is for the posts in a given thread (assumed to be clustered by thread id)

*Workload E uses the Zipfian distribution to choose the first key in the range, and the Uniform distribution to choose the number of records to scan.

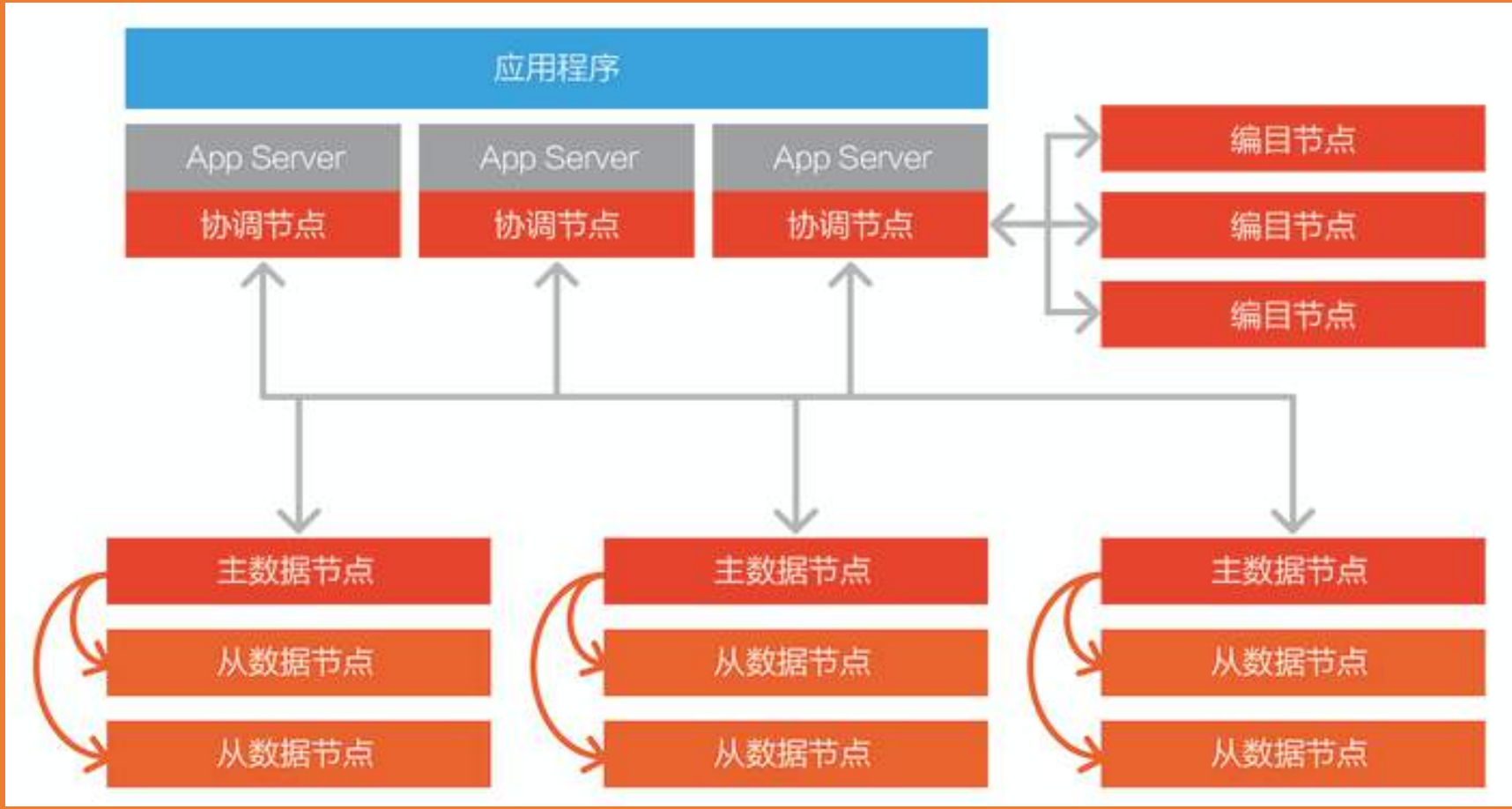
Table 2: Workloads in the core package

分布式的架构

计算分布

+

存储分布

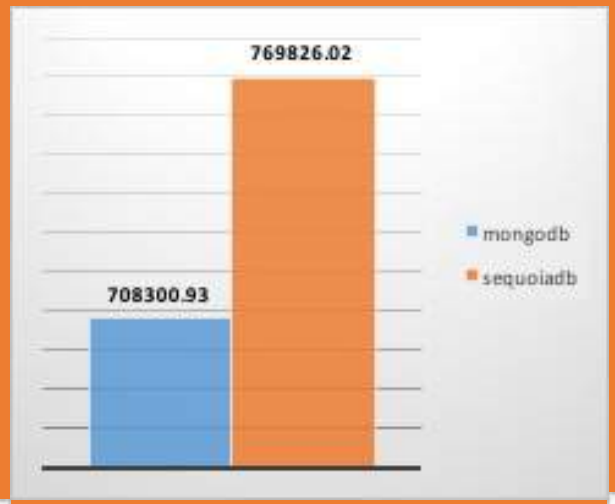


性能测试


YCSB测试客户端
10Gbit
Sequoia DB/ MongoDB集群
服务器1
服务器2
服务器3
服务器4
服务器5
服务器6
协调节点
编目节点
数据节点
数据组9
数据组11

性能测试

100%插入



100%读

50%读
50%插入

50%读
50%更新

性能案例

平均每日超过2亿条记录写入

高峰时段，同时有超过百亿级别的数据需要被检索、调用

系统保存3年内所有交易和持有数据

峰值并发量超过10000

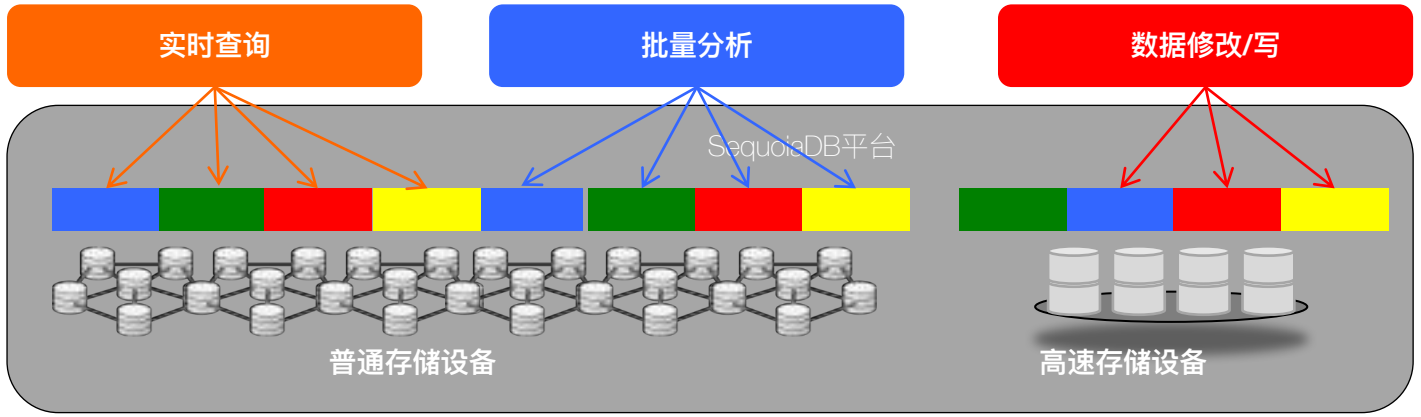
高峰时段，查询返回时间小于100ms

实际测试性能10倍于原有MySQL

操作涉及3张数据表的关联，总量超过3000亿条数据

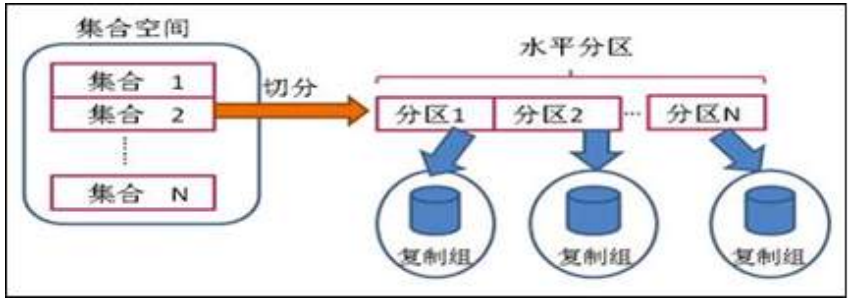
分布式架构优化：读写分离机制

- 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

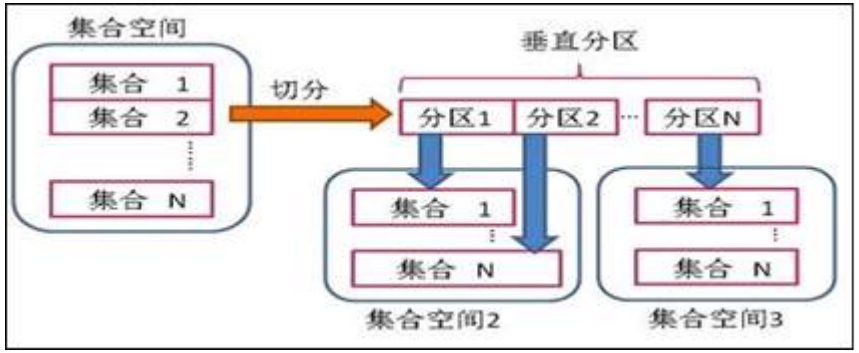


分布式架构优化：数据多维分区

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



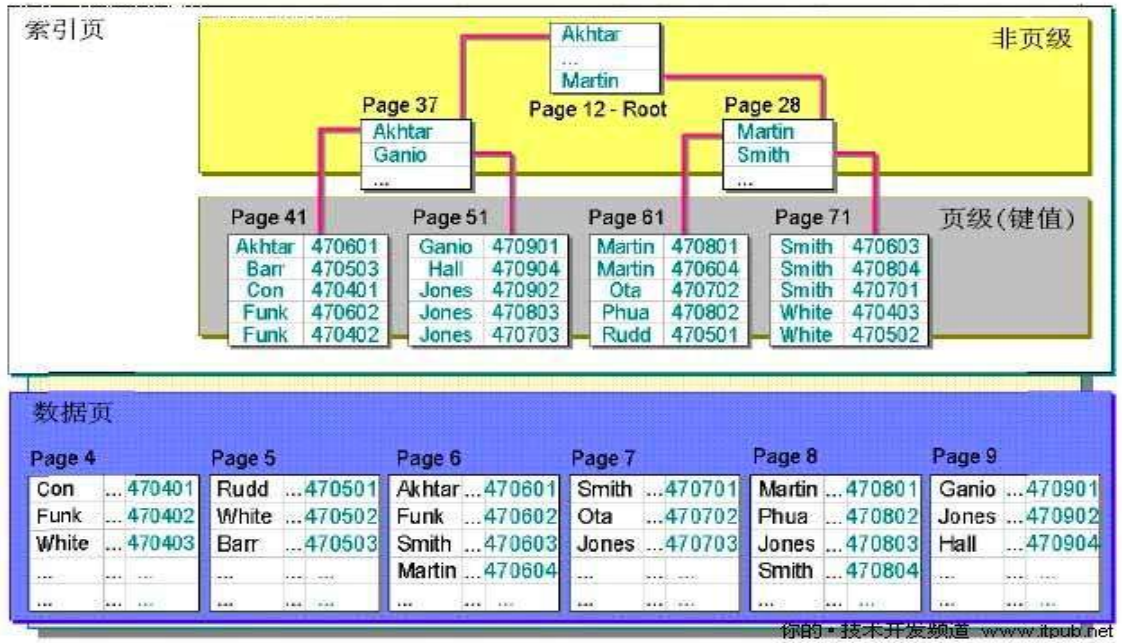
分别适合流水数据与快照数据



优势： 容量和性能可线性扩展

引擎内部优化：B树多维度索引

支持多字段索引
支持复合索引
支持唯一索引
B树索引与数据保持强一致
支持全文检索索引
全文检索索引与数据保持最终一致



从技术产品角度来看

物理隔离、逻辑隔离、权限、用户、加密

两地三中心、同城双活、一致性

HTAP、实时处理能力、大数据

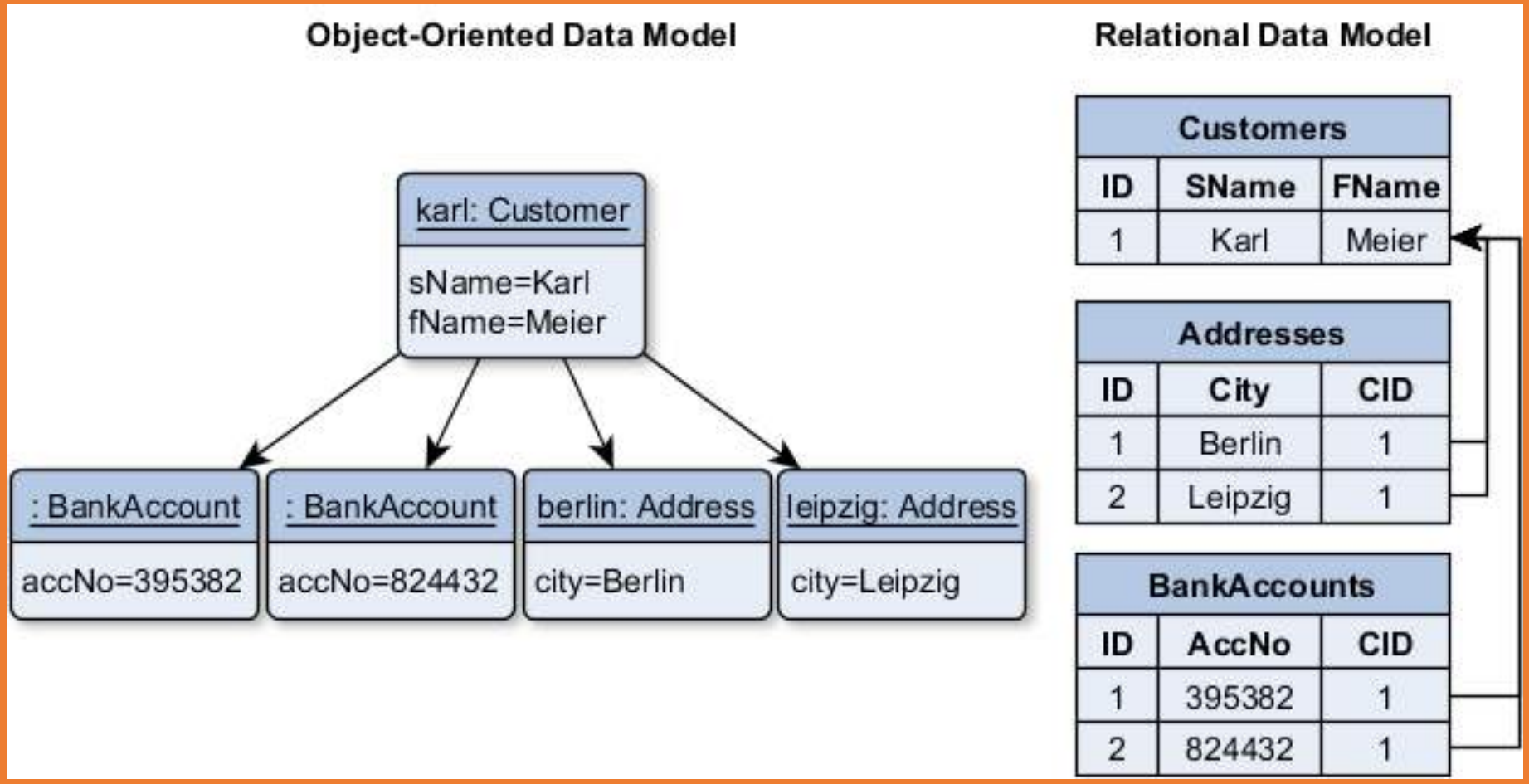
读写分离、数据域管理

原厂、原生态分布式管理

多模、多引擎

分布式架构

ORM 模型


Object-Oriented Data Model
karl: Customer
sName=Karl
fName=Meier
: BankAccount
accNo=395382
: BankAccount
accNo=824432
berlin: Address
city=Berlin
leipzig: Address
city=Leipzig
Relational Data Model
Customers
ID
SName
FName
1
Karl
Meier
Addresses
ID
City
CID
1
Berlin
1
2
Leipzig
1
BankAccounts
ID
AccNo
CID
1
395382
1
2
824432
1

JSON Object Mapping

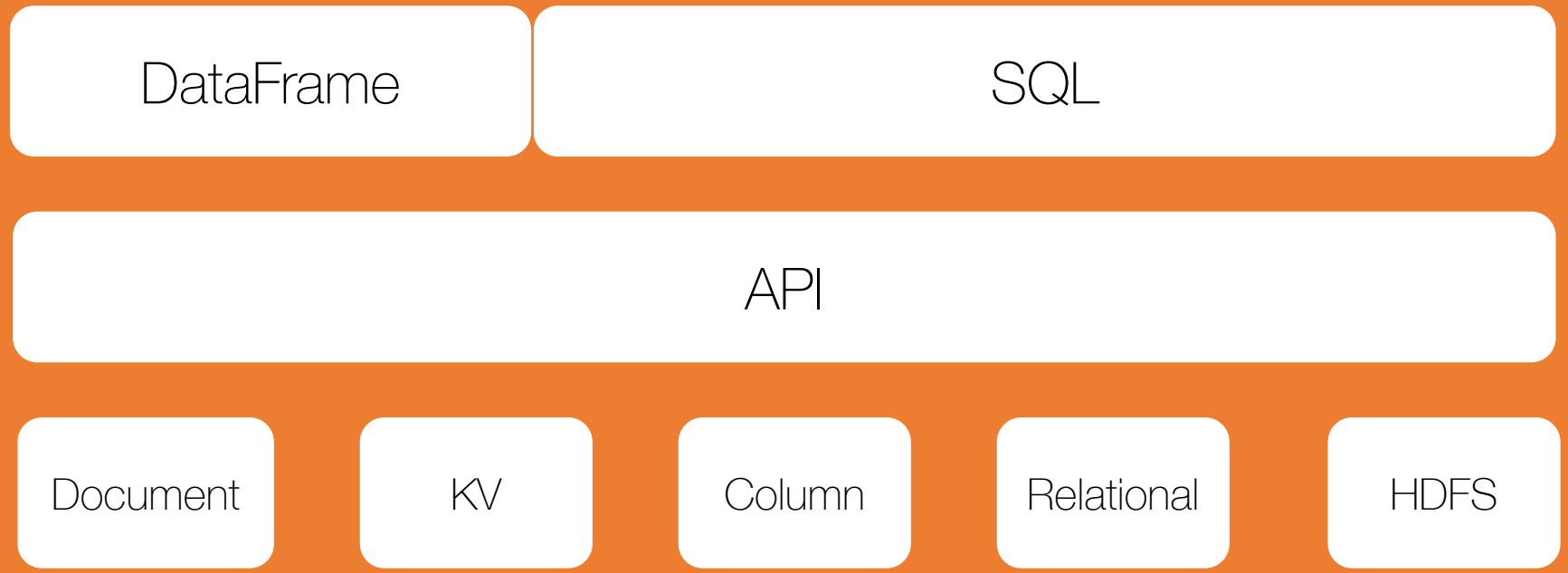
```
private void JSONSerilaize()
{
    // Serializaion
    Employee empObj = new Employee();
    empObj.ID = 1;
    empObj.Name = "Manas";
    empObj.Address = "India";

    // Convert Employee object to JOSN string format
    string jsonData = JsonConvert.SerializeObject(empObj);
                      jsonData "{\"ID\":1,\"Name\":\"Manas\",\"Address\":\"India\"}"
    Response.Write(jsonData);
}
```

流水、Snapshot、Table 和 SQL

1	036010	99999	1930	05	27	54	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	3.5
2	037770	99999	1930	05	02	50.7	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	3.5
3	037770	99999	1930	12	04	40	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	3.5
4	039730	99999	1930	03	17	41	4	33.3	4	997.3	4	9999.9	0	999.9	0	2.2
5	030910	99999	1930	09	10	56	4	9999.9	0	1012.9	4	9999.9	0	999.9	0	8.0
6	030050	99999	1930	08	29	54.8	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	7.0
7	031590	99999	1930	05	05	45.5	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	6.0
8	036010	99999	1930	04	28	53.7	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	6.0
9	038560	99999	1930	05	21	53.2	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	3.5
10	038560	99999	1930	06	29	58.5	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	7.0
11	039730	99999	1930	02	23	40.2	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	8.0
12	039730	99999	1930	03	26	47.5	4	45.8	4	1024.1	4	9999.9	0	999.9	0	8.0
13	039730	99999	1930	05	04	51	4	9999.9	0	1016.5	4	9999.9	0	999.9	0	4.0
14	039730	99999	1930	08	31	57	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	8.9
15	030910	99999	1930	01	27	33	4	9999.9	0	1009.5	4	9999.9	0	999.9	0	8.9
16	033790	99999	1930	11	01	46.8	4	45.7	4	1007.2	4	9999.9	0	999.9	0	8.0
17	036010	99999	1930	08	12	58.3	4	9999.9	0	9999.9	0	9999.9	0	999.9	0	8.0
18	030050	99999	1930	04	21	41	4	34.5	4	1010	4	9999.9	0	999.9	0	12.0

90% 的开发、程序员、数据科学家都习惯于用SQL

查询和分析：SQL + 多模



从技术产品角度来看

物理隔离、逻辑隔离、权限、用户、加密

两地三中心、同城双活、一致性

HTAP、实时处理能力、大数据

读写分离、数据域管理

原厂、原生态分布式管理

多模、多引擎

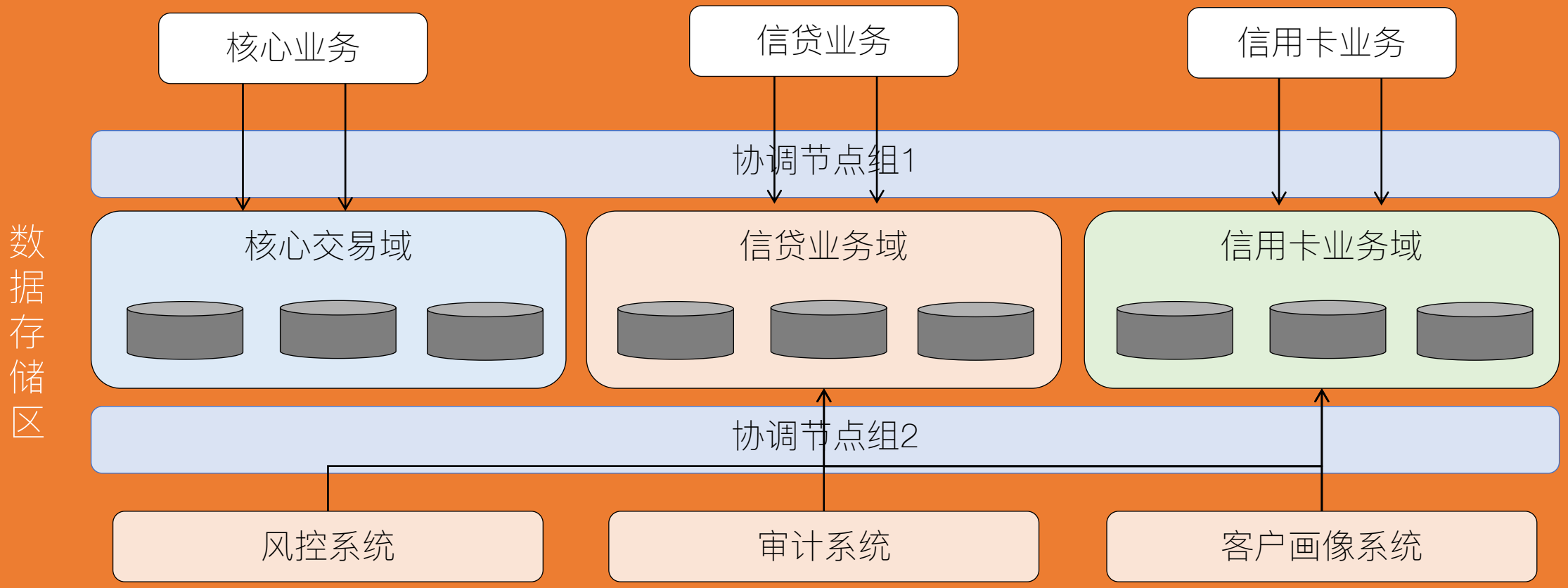
分布式架构

"Chief data officers (CDOs) and other senior data and analytics leaders are therefore looking to data management to simplify and provide a cohesive data management ecosystem across what was once siloed data."

CDO和数据管理者都在寻求简化的数据管理方式以及一个统一连贯的数据管理生态，以消除独立的**数据孤岛**。

-- Gartner Hype Cycle for Data Management, 2017

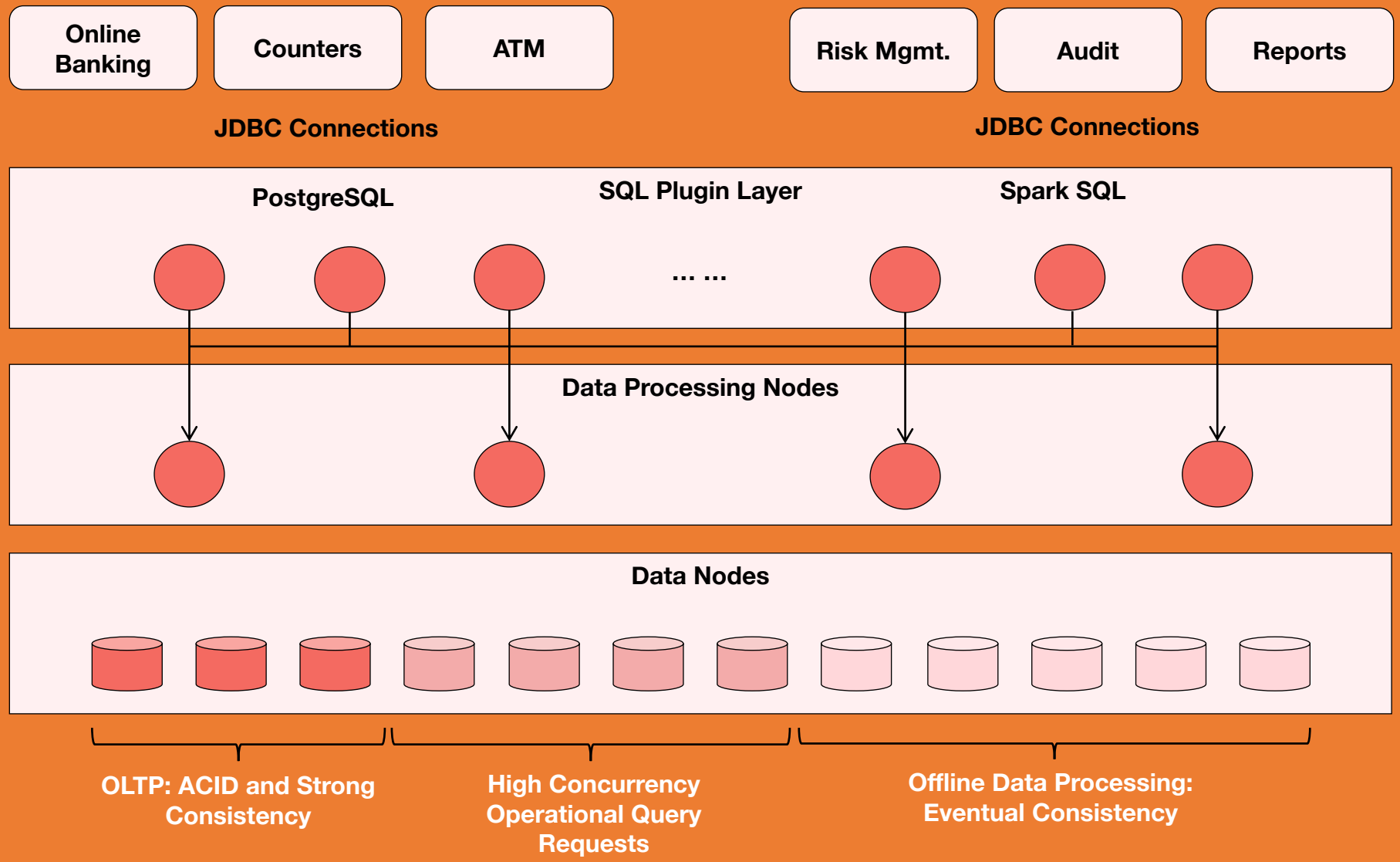
数据域逻辑与物理隔离


核心业务
信贷业务
信用卡业务
协调节点组1
核心交易域
信贷业务域
信用卡业务域
数据存储区
协调节点组2
风控系统
审计系统
客户画像系统

HTAP

读写分离

分域管理



总结

物理隔离、逻辑隔离、
权限、用户、加密

两地三中心、同城双活、一致性

HTAP、实时处理能力、大数据

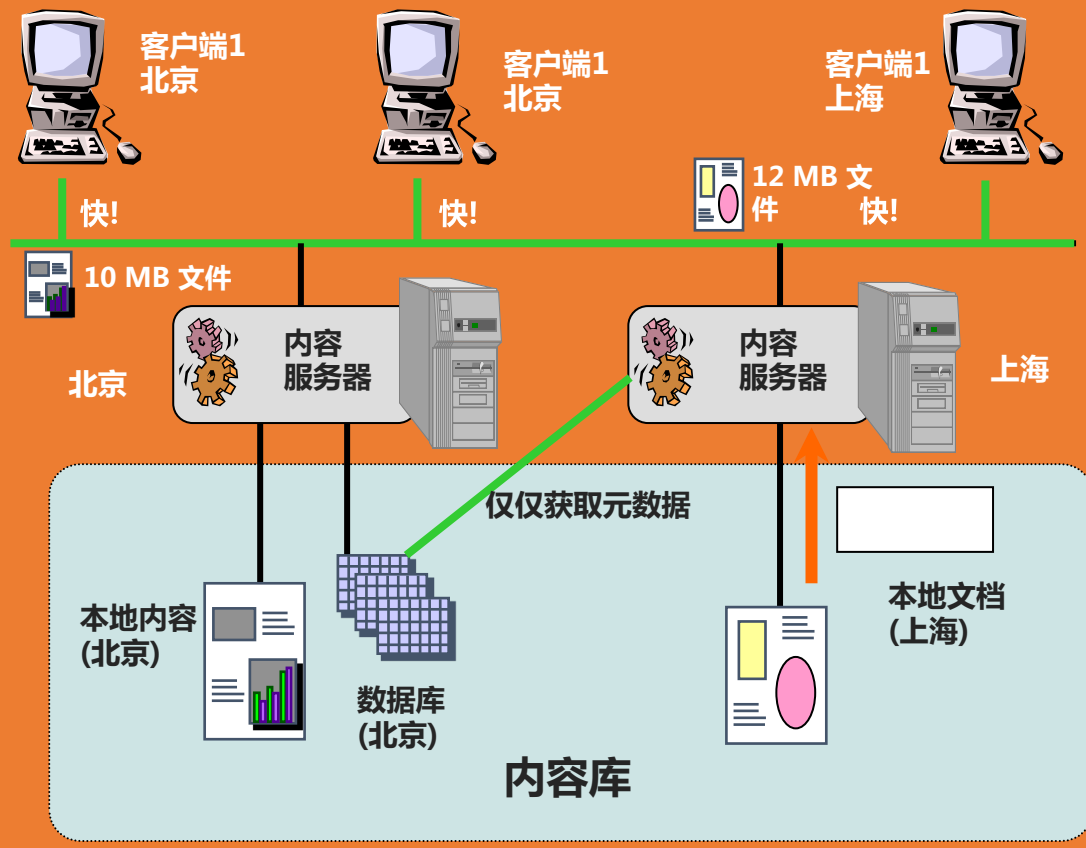
读写分离、数据域管理

原厂、原生态分布式管理

多模、多引擎

分布式架构

多数据中心分布式架构



特色：

- 统一管理，分布部署
- 所有的内容管理信息存放在总部，但是各地可拥有独立的本地内容文件
- 用户属于“总部 - 分公司”架构，且分公司具备相当数量的本地内容需要管理时的最佳选择；
- 部署有一定复杂度，也需要本地具备相关的IT人员和设备

总结

物理隔离、逻辑隔离、
权限、用户、加密

两地三中心、同城双活、一致性

HTAP、实时处理能力、大数据

读写分离、数据域管理

原厂、原生态分布式管理

多模、多引擎

分布式架构

隐私、安全是数据发展的重中之重

通用数据保护条例(GDPR)

违反GDPR条例要求的企业将面临高达全球年收益额4%或2000万欧元的罚款，以金额最高为准。

总结

物理隔离、逻辑隔离、
权限、用户、加密

两地三中心、同城双活、一致性

HTAP、实时处理能力、大数据

读写分离、数据域管理

原厂、原生态分布式管理

多模、多引擎

分布式架构