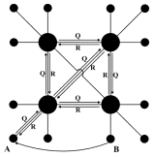
DATAGURU 炼数成金

Leader-us系列



大型分布式系统案例实战 第4周

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

说说内存计算

- Hazelcast & GridGain

说说内存计算

服务器内存究竟能多大、能多贵、能多快

2014年初，英特尔发布最新一代面向关键业务及实时计算的至强E7 v2

16根组成1T的价格是76800

16T-100T内存

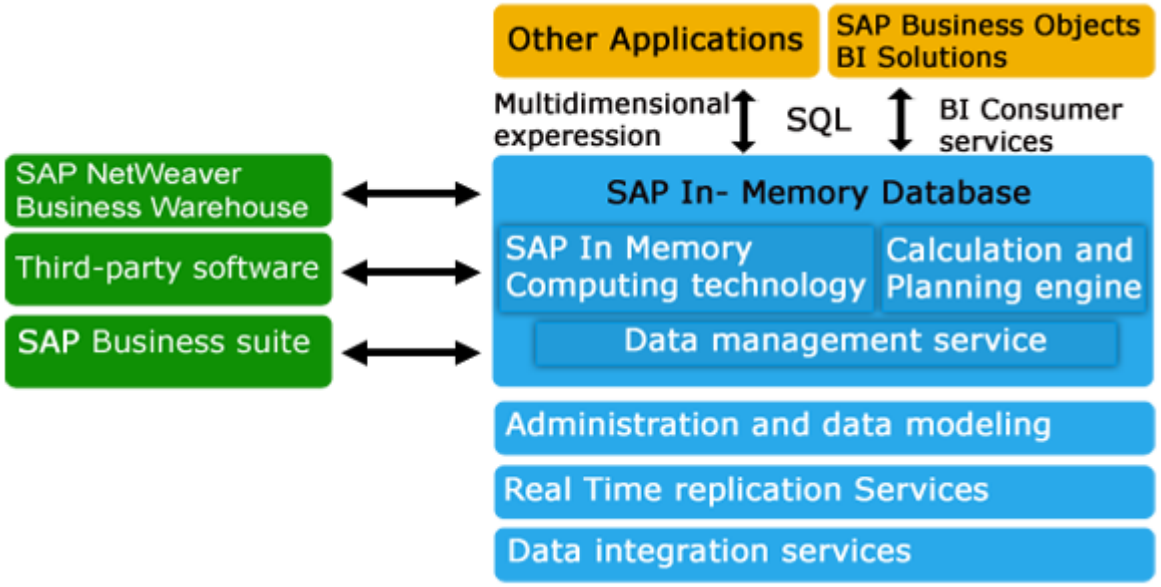
预计未来5到10年，关系数据库将彻底消失，届时所有的SAP产品都将使用HANA——2011年

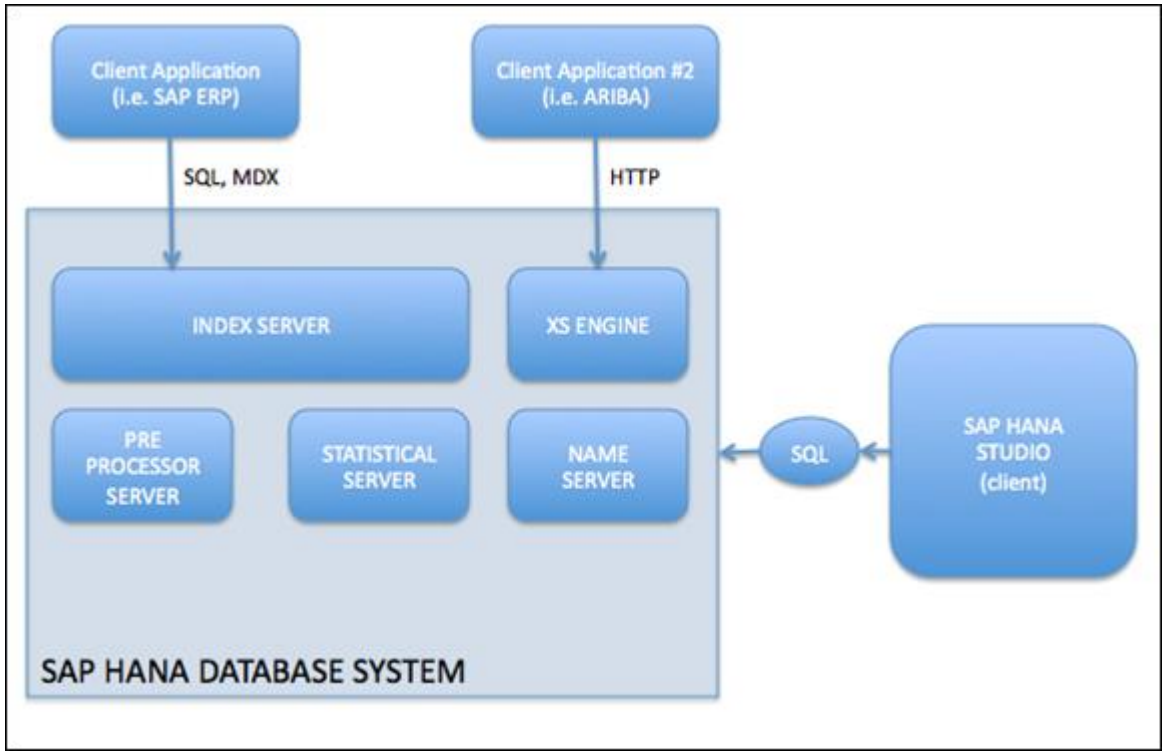
内存计算的模式将能帮助企业分析数据的速度提升10万倍（相比传统关系数据库）。这也就意味着以前需要数小时的分析现在几秒钟内就可以完成。”

Bussmann所在的团队在2010年10月份的时候，通过概念证明SAP数据分析速度有望提升14000倍，以前需要花费5小时分析处理完数据，现在仅需要1秒钟则可以迅速完成。

说说内存计算

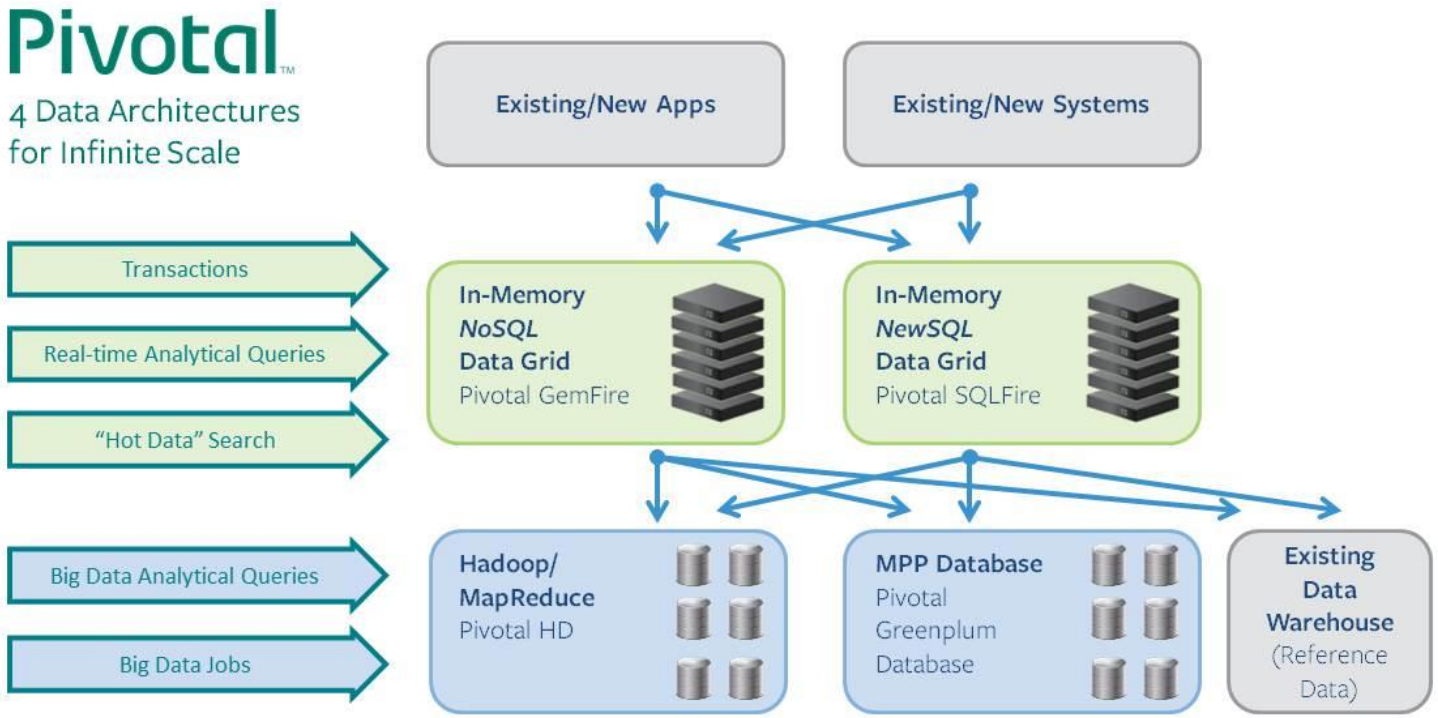
SAP HANA超级内存数据库





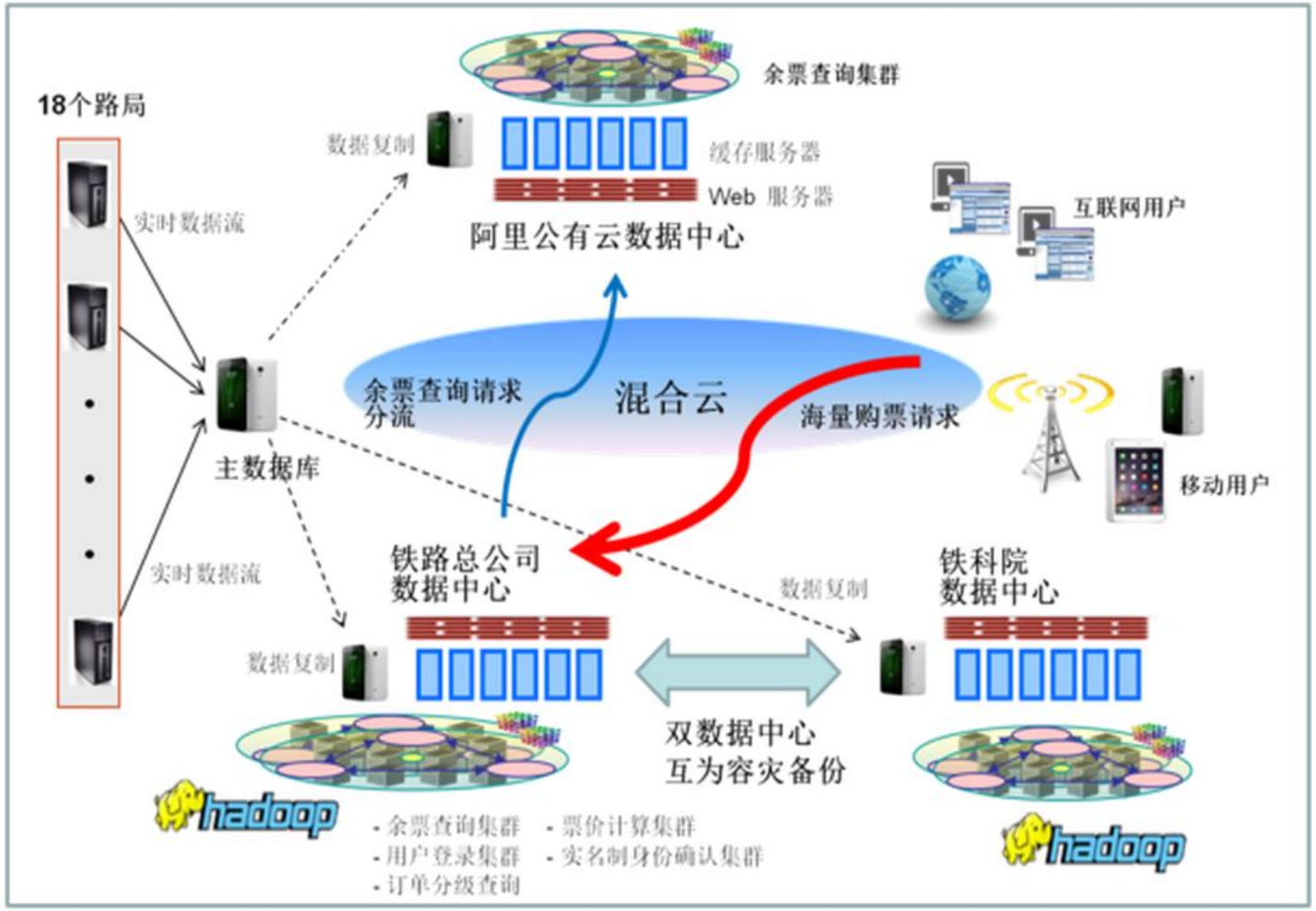
说说内存计算

分布式内存整合计算的典型代表Pivotal GemFire



说说内存计算

Gemfire+12306让中国人回家过年更方便？



部署数百个Pivotal Gemfire节点

在2015年春运购票高峰之前，考虑到超大并发会造成网络流量大以及阻塞的问题，今年特别在阿里云建立一个数据中心，由阿里云提供“虚拟机”的租赁服务，将基于Gemfire实现余票查询功能的系统以及Web服务部署在这些虚拟机上，以分流“余票查询”请求

技术改造之后，在只采用10几台X86服务器实现了以前数十台小型机的余票计算和查询能力，单次查询的最长时间从之前的15秒左右下降到0.2秒以下，缩短了75倍以上。2012年春运的极端高流量并发情况下，系统几近瘫痪。而在改造之后，支持每秒上万次的并发查询，高峰期间达到2.6万个查询/秒吞吐量，整个系统效率显著提高。

旧系统每秒只能支持300-400个查询/秒的吞吐量

说说内存计算

内存计算为什么这么快

当前计算架构的瓶颈在存储，处理器的速度按照摩尔定律翻番增长，而磁盘存储的速度增长很缓慢，由此造成巨大高达10万倍的差距。

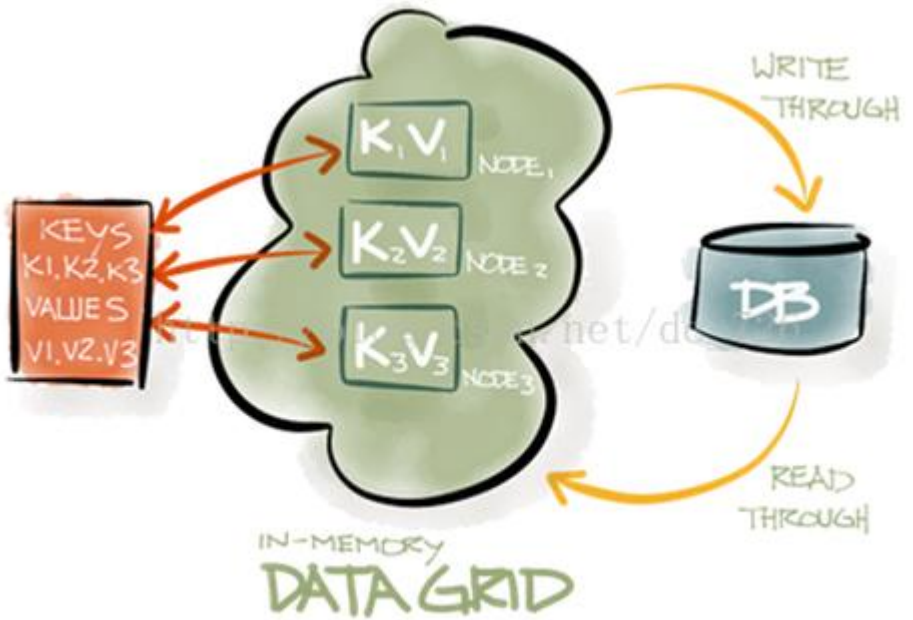
IMDG

In Memory Data Grid（“内存数据网格”）

- 首先自然是网格式分布式存储。
- 所有数据存于内存（RAM）。
- 存储服务器数量可随时增减。
- 数据模型是非关系模型，而是基于对象模型。
- 在网格内的某一台存储服务器的启动和关闭不会影响到网格内的其他服务器。

IMDG产品

比较		Hazelcast	GridGain
使用性	安装	Maven引入Jar包即可，无需安装软件	
	客户端	支持各种语言的客户端	
	框架集成	集成Hibernate、Web Session、Spring	
基本功能	分布式计算工具	分布式的**集合、并发包、消息队列、调度器**	
性能	性能配置	**内存索引、Near-Cache、数据亲和性**	
可靠性	数据备份	分区数据冗余备份	
	持久化	read-through，write-through/behind	
	事务	保证数据一致性	
扩展性	自动分区	支持**本地、分区、复制**三种方式	
	动态拓扑	动态添加删除结点，自动rebalance	



Hazelcast

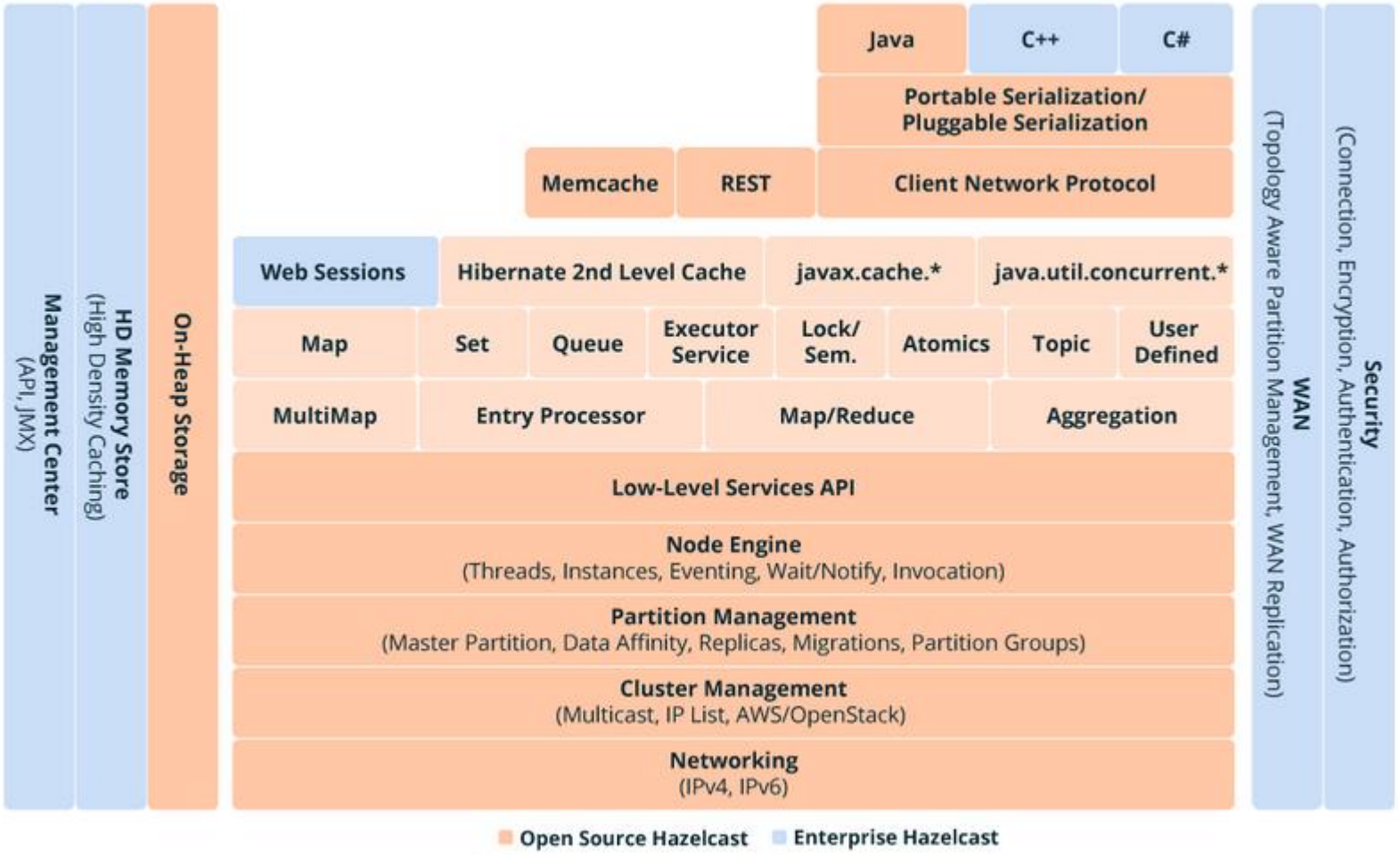
Hazelcast是什么

- 2008年开源项目，目标是一个简单的分布式Java Map
- 目前的定位是Java分布式内存网格

We have 92 contributors to Hazelcast of which 68 are external. While our competitors talk about building a community we have been doing it for six years.

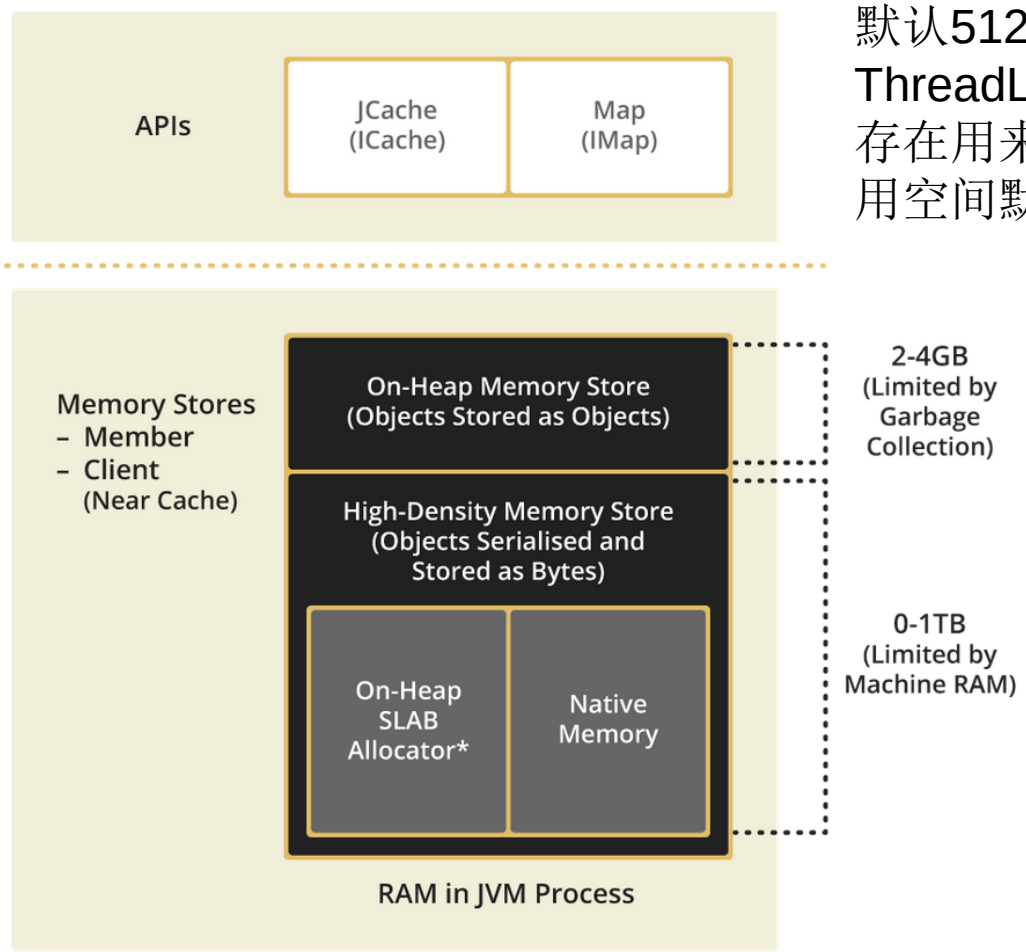
Hazelcast

Hazelcast架构

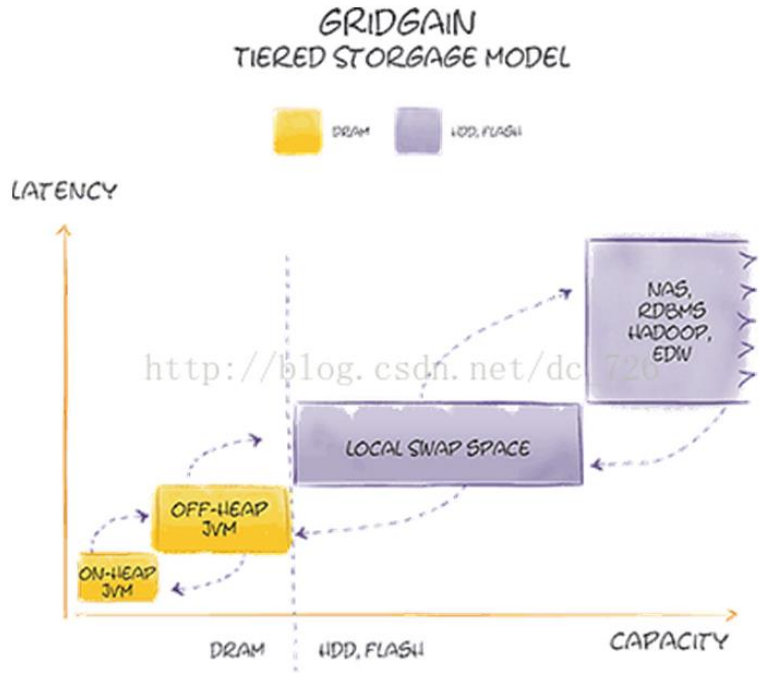


Hazelcast

Hazelcast企业版的最强特性之堆外缓存



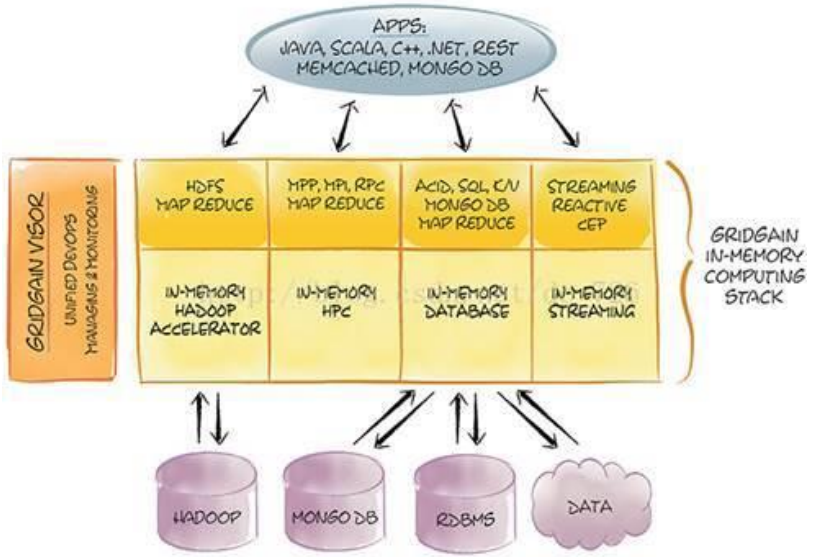
默认512M OS的内存池，分为默认4M大小的管理单元，默认采用ThreadLocal方式管理内存单元。
存在用来存放Index、偏移量等Metadata信息的内存单元，MetaData的占用空间默认是12%



GridGain

作为另一款主流的开源数据网格产品，GridGain是Hazelcast的强有力竞争者。同样提供了社区版和商业版，近日GridGain的开源版本已经进入Apache孵化器项目Ignite(一款开源的内存计算(In-Memory Computing)IMC中间件)，目前Apache正在迁移GridGain开源版本的代码到Ignite项目

- **GridGain在开源版本中就提供了堆外存储功能**，当堆和非堆内存都不足时，还可以开启SWAP，将数据溢出到磁盘
- GridGain使用**2PC(两阶段提交)协议**实现分布式事务，事务级别支持三种事务隔离级别
- GGFS(GridGain In-Memory File System)，类似Spark生态圈中的Tachyon，能够加速MapReduce任务的执行。
- 流式数据/事件处理，可以作为CEP事件处理器。

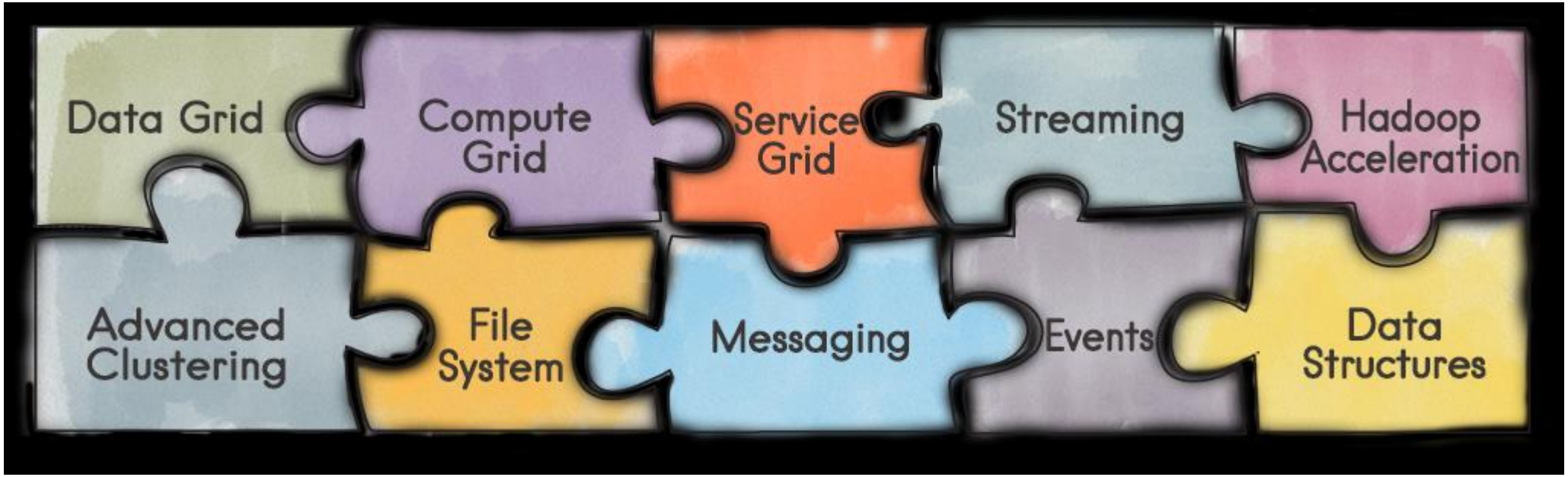


Apache Ignite项目凭借其业界领先的事务处理能力在新兴的混合型的OLTP/ OLAP用例方面更胜一筹。特别是针对Hadoop，Apache Ignite将为现有的Map/Reduce，Pig或Hive作业提供即插即用式的加速，避免了推倒重来的做法

Apache Ignite成为未来的快速数据世界（Fast Data world），如同Hadoop是今天的大数据。

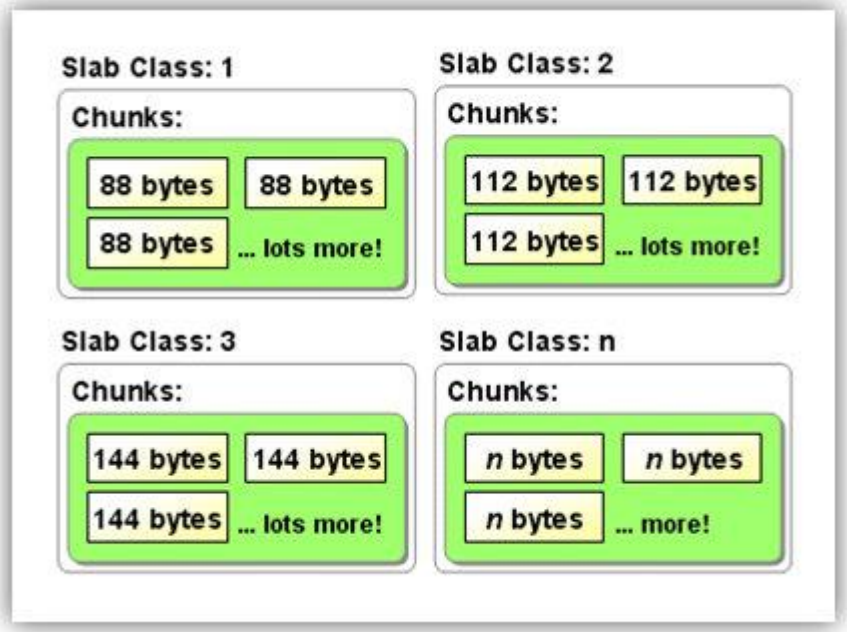
GridGain

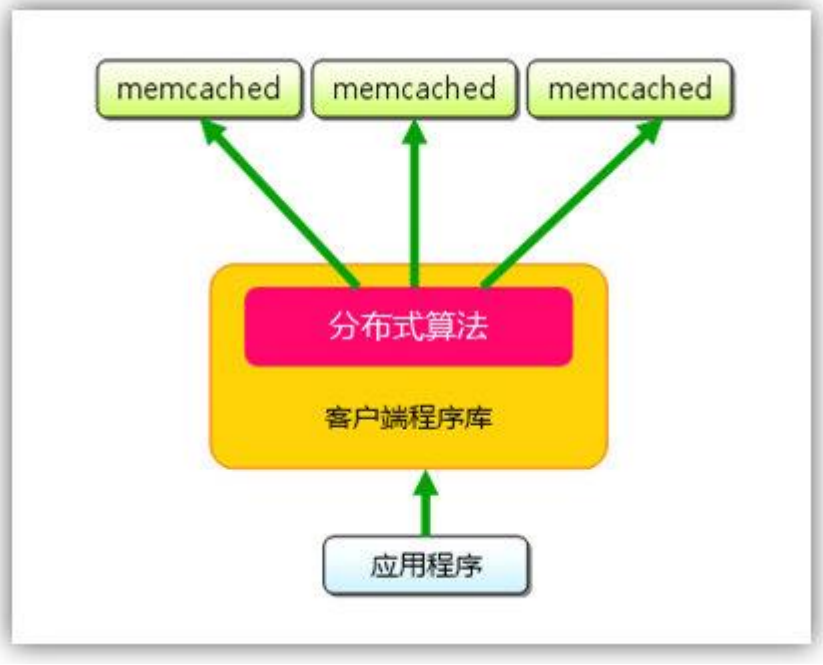
GridGain是一个融合了各种实时/内存计算技术的平台



Memcached

Web系统中使用最为广泛的分布式Key/value缓存中间件





Memcached最佳实践

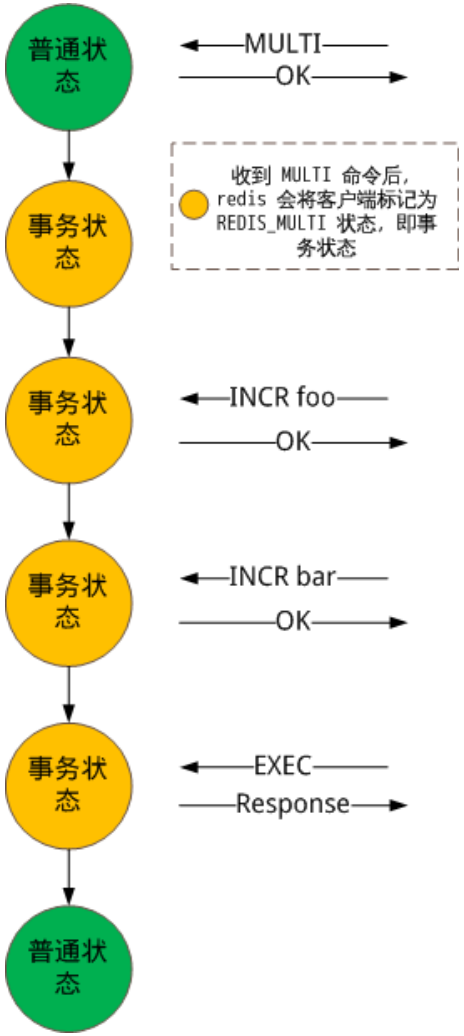
- 最适合存储小数据，并且存储的数据是大小一致的
- Memcached在很多时候都是作为数据库前端cache
- 虚机上不适合部署Memcached
- 确保Memcached的内存不会被Swap出去
- 不能便利所有数据，这将导致严重性能问题
- Local Cache+ Memcached这种分层Cache还是很有
- Memcached启动预热是一个好办法

Redis

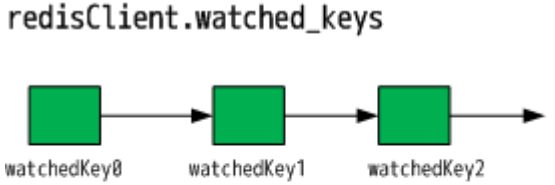
- 是Memcached的一个强有力的补充，同时也是一个有显著不同的新产品
- Redis扩展了key-Value的范围，Value可以是List，Set，Hashes，Sorted Set等数据结构，同时增加了新指令针对这些数据结构：如Set的union，List的pop等操作指令
- 比如Subscribe/publish命令，以支持发布/订阅模式这样的通知机制等等
- Redis通过Multi / Watch /Exec等命令可以支持事务的概念，原子性的执行一批命令。
- Redis 3.0开始实现Cluster方案，但没有采用一致性Hash

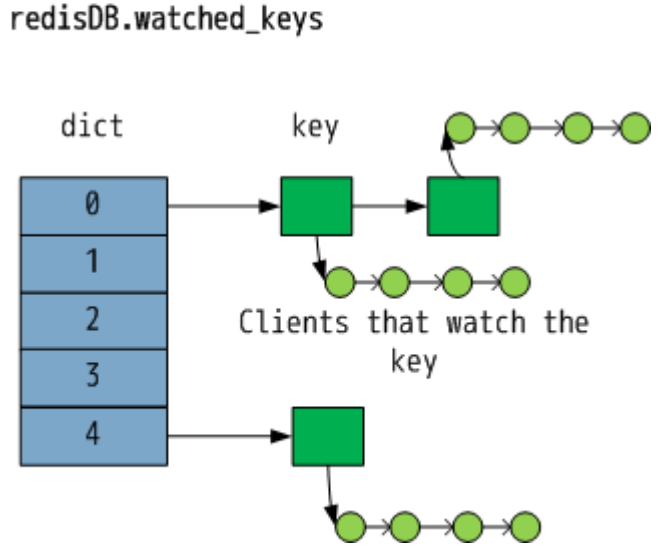
Redis

Redis事务问题



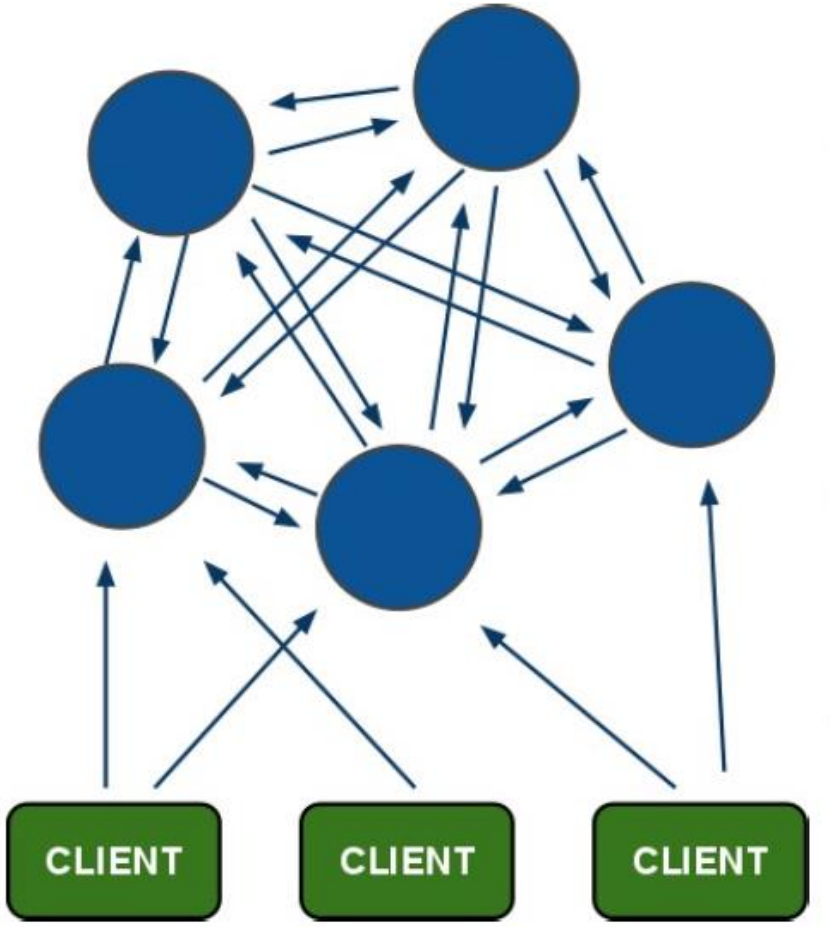
redis 数据集结构体 redisDB 和客户端结构体 redisClient 都会保存键值监视的相关数据

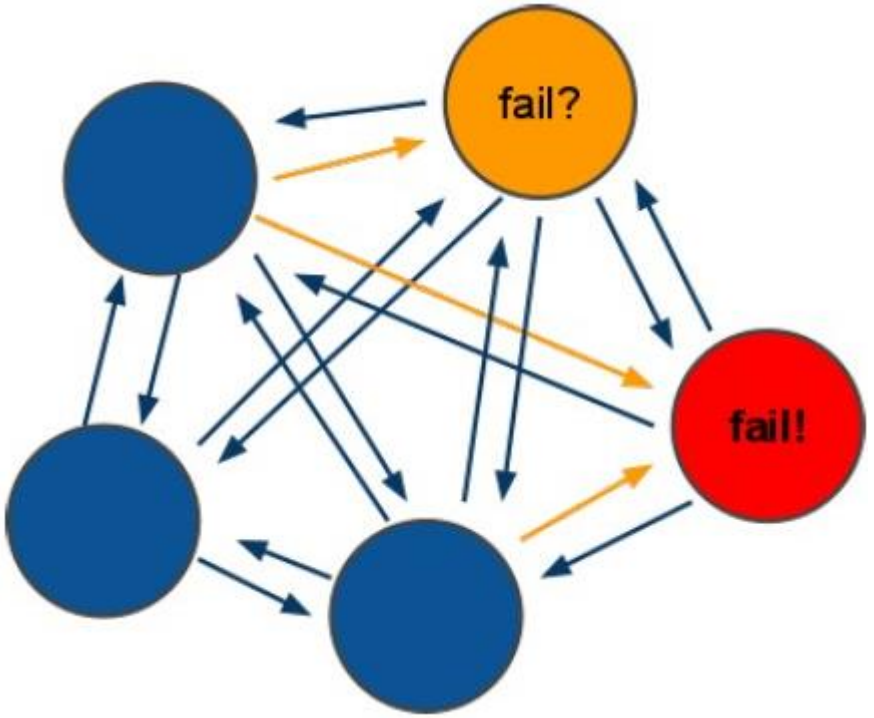




Redis

一个Redis集群包含16384个哈希槽(hash slot)，数据库中的每个键都属于这个16384个哈希槽的其中一个，集群使用公式CRC16(key) % 16384来计算键key属于哪个槽，其中CRC16(key)语句用于计算键key的CRC16校验和。





超过半数的Master节点之间的通讯故障后需要新选择Master

Redis vs Memcache

- 简单数据结构下Memcache的多线程架构有优势
- 仅仅从用做缓存的角度，Memcache还是无法被替代
- Redis具备向数据库考虑的能力，但这些方面并没有特别强的优势
- 不要求关系数据库质量级别的交易时，Redis可以取代一些特定场合的数据库操作，比如秒杀这样的系统

分布式系统存储之基于内存的两表Join演示

A表为Person { id、 name }
B表为Order{id,orderId,amount（金额）}
关联关系为order.orderid=person.id
求计算结果 select p.id,sum(o.amount),count(*) from person p ,order o where order.orderid=person.id order by sum limit 1000

Person、Order的数据分别存在CSV文件中，数量各自是1亿、10亿，数据随机制造，保证基本都有关联

采用Hazelcast或GridGain的API完成，需要找出这两个产品中最合适的API来完成Join、Group、Order等操作

Thanks

FAQ时间