



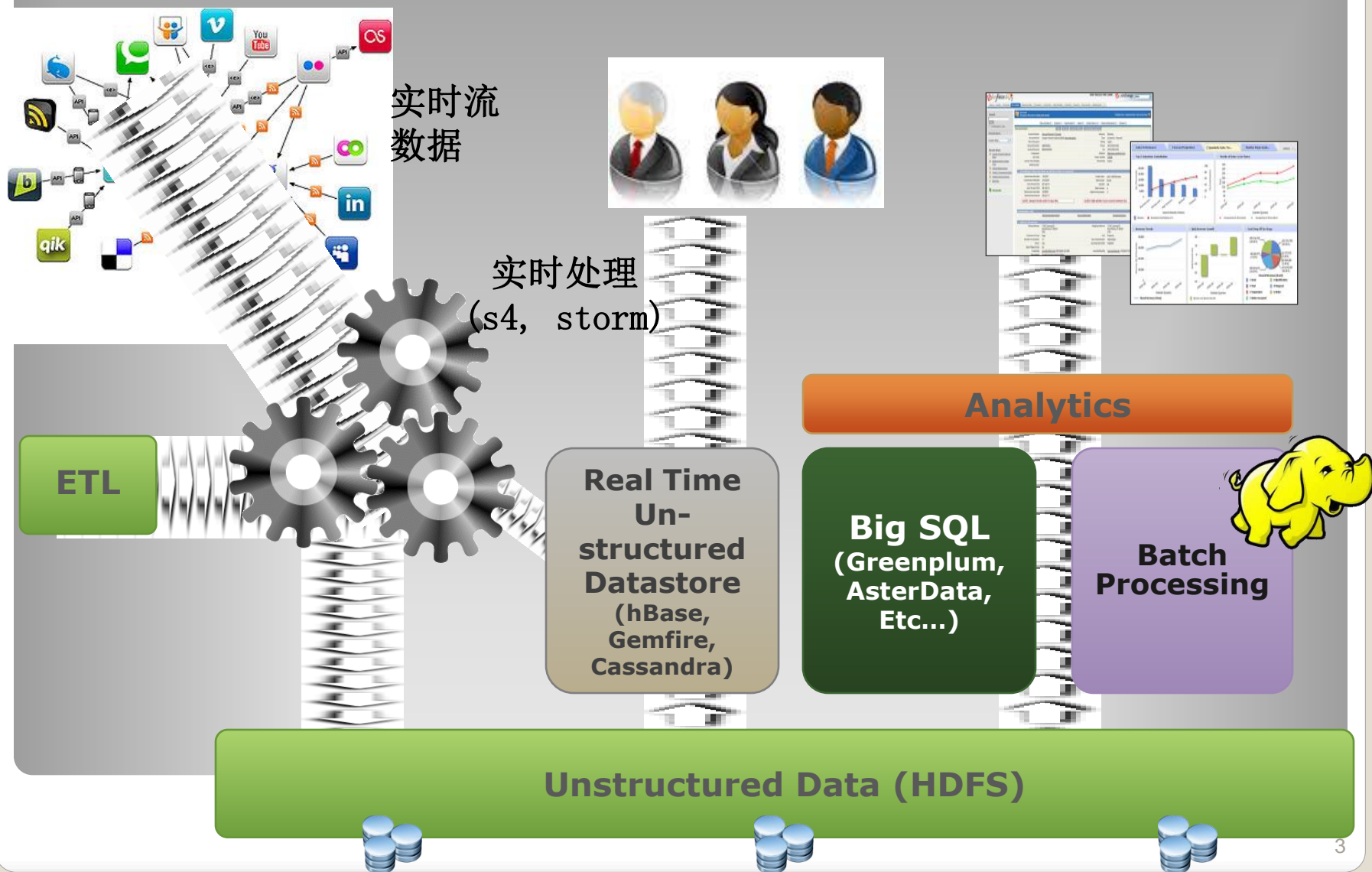
海量并行 (MPP) 内存数据仓库

—— 实现探讨

谢剑锋
Jason

 柏睿数据科技（北京）有限公司
 联想服务首席技术顾问
惠普实验室(总部)特邀研究员

统一的大数据系统的整体视图



通过虚拟化来统一大数据计算平台

■ 目标

- 简单、快速、即需地监控数据集群
- 允许混合负载
- 利用虚拟机来提供隔离（如：多租户）
- 通过虚拟拓扑来优化数据处理性能
- 通过虚拟拓扑来优化平台稳定性

■ 充分利用虚拟化

- 可伸缩的扩展性能
- 依靠高可靠性来保护关键服务，如：Hadoop的Name Node及Job Tracker
- 资源控制和共享：重用低利用率的内存及CPU
- 对负载进行优先级控制：在混合环境中优化及限制资源的使用

统一的分析云将被极大的简化



SQL集群



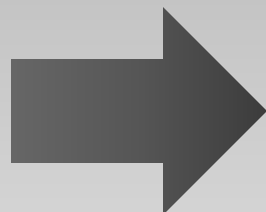
NoSQL集群



Hadoop 集群

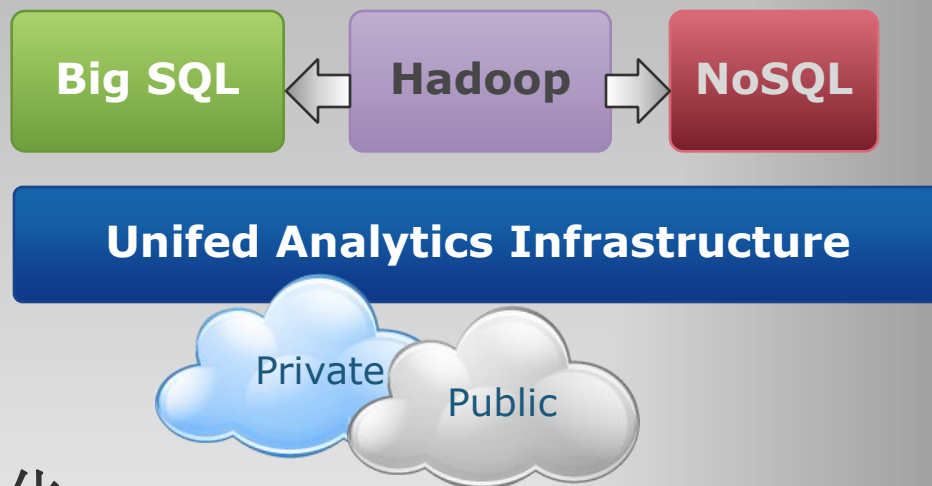


Decision Support 集群



■ 简化

- 单一的硬件基础架构
- 快速、简易的环境控制

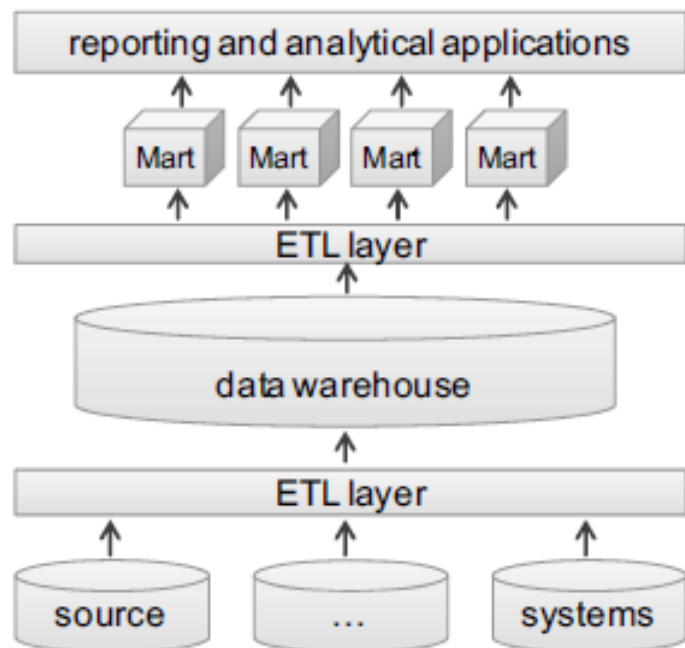


■ 优化

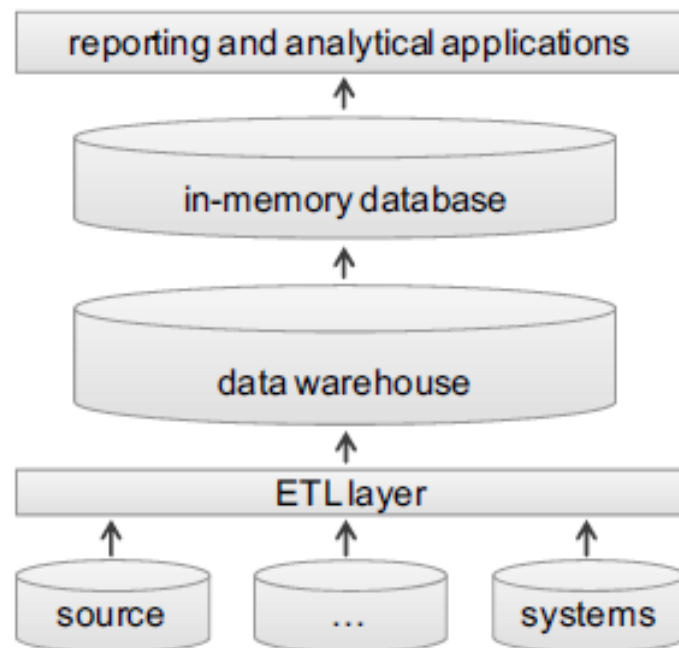
- 共享的资源 = 更高的利用率
- 可伸缩的资源 = 快速的即需资源访问

并行内存计算 及 持久化

Old landscape



New landscape

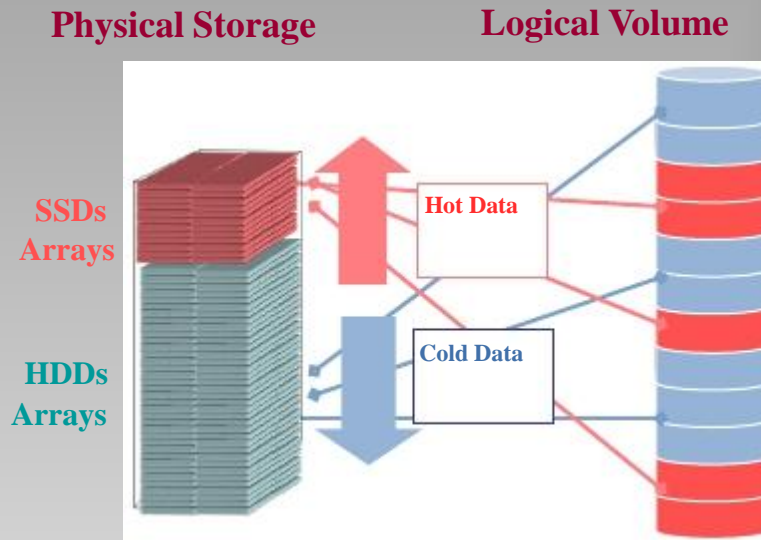
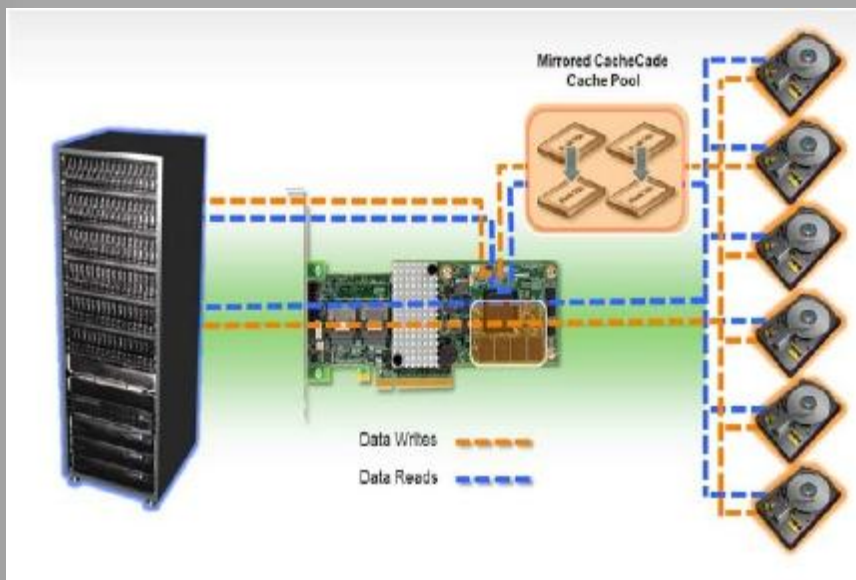


● 关键技术:

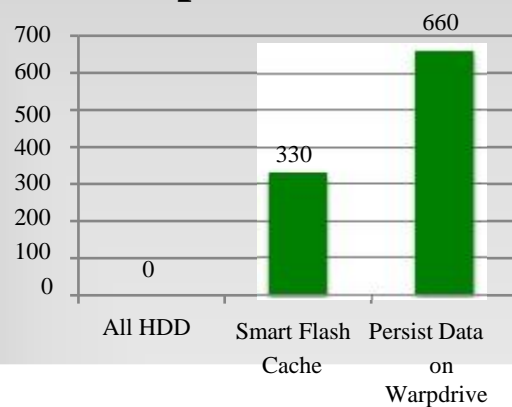
- Share-Nothing, MPP 海量并行架构
- 基于内存分区的数据集市
- 海量并行内存计算

- 虚拟化, 云
- 性能优化
- 固态内存

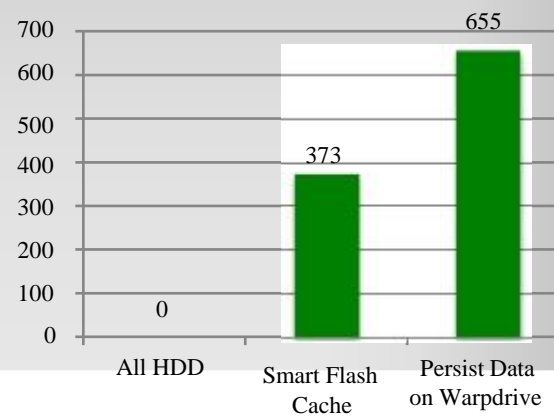
并行内存为大数据提供实时缓存



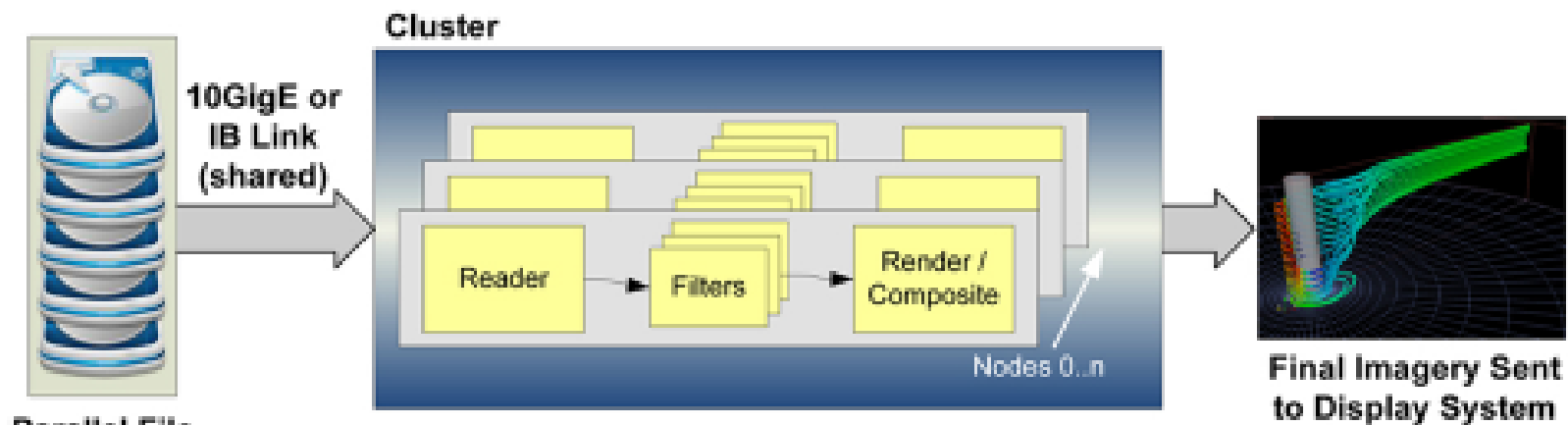
ResponseTime



TPS

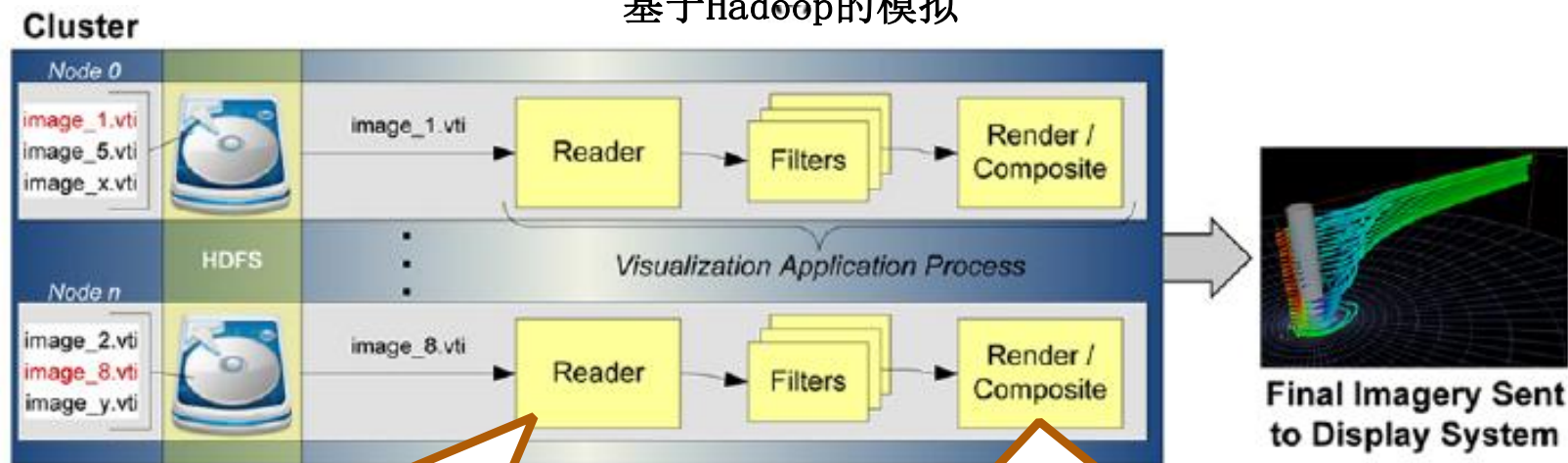


并行内存针对大数据的应用场景



Parallel File System













传统模拟对比
基于Hadoop的模拟



















MapReduce可以基于MPP内存

渲染组合可以基于内存,数据库同理

现有大数据处理平台的技术比较 - I

Capability	Cloudera CDH	EMC / GP UAP	MAPR	HortonWorks	Open Source	MPP In- memory with Hadoop
低延迟 任务调度	Impala only	No		No	No	
混合负载	No	No		No	No	
快速的 抢占式调度	No	No	No	No	No	
时间敏感 SLA保证	No	No		No	No	
使用计费及 分析插件	No	No	No	No	No	
可恢复的 Hadoop任务	No	No	No	No	No	No
POSIX 文件系统	No		NFS only	No	No	NFS or Gluster
企业级 文件系统功能	No			No	No	

现有大数据处理平台的技术比较 - II

Capability	Cloudera CDH	EMC / GP UAP	MAPR	HortonWork s	Open Source	MPP In- memory with Hadoop
SQL 的支持	 <i>Impala</i>	 <i>Pivotal</i>	 <i>Drill</i>	Via open source only	<i>Impala, Drill</i>	
大表 的支持	No	No		No	No	
外部数据的链接		GP DB built-in	No	No	No	
加速器	No	No		No	No	
完整的硬件及 软件的支持	Through HW partners		Through HW partners	No	No	
单一厂商支持	Through HW partners		No	No	No	
全功能的 私有云管理功能	No	No		No	No	

注: Hadoop 1.0 is based on the Hadoop 20.205 branch (it went 0.18 -> 0.19 -> 0.20 -> 0.20.2 -> 0.20.205 -> 1.0).

The project having matured to that point.

Hadoop 2.0 is from the Hadoop 0.23 branch, with major components re-written to enable support for features like High Availability, and MapReduce 2.0 (YARN), and to enable to scale out past 4,000 machines per cluster.

统一的基于分析的云

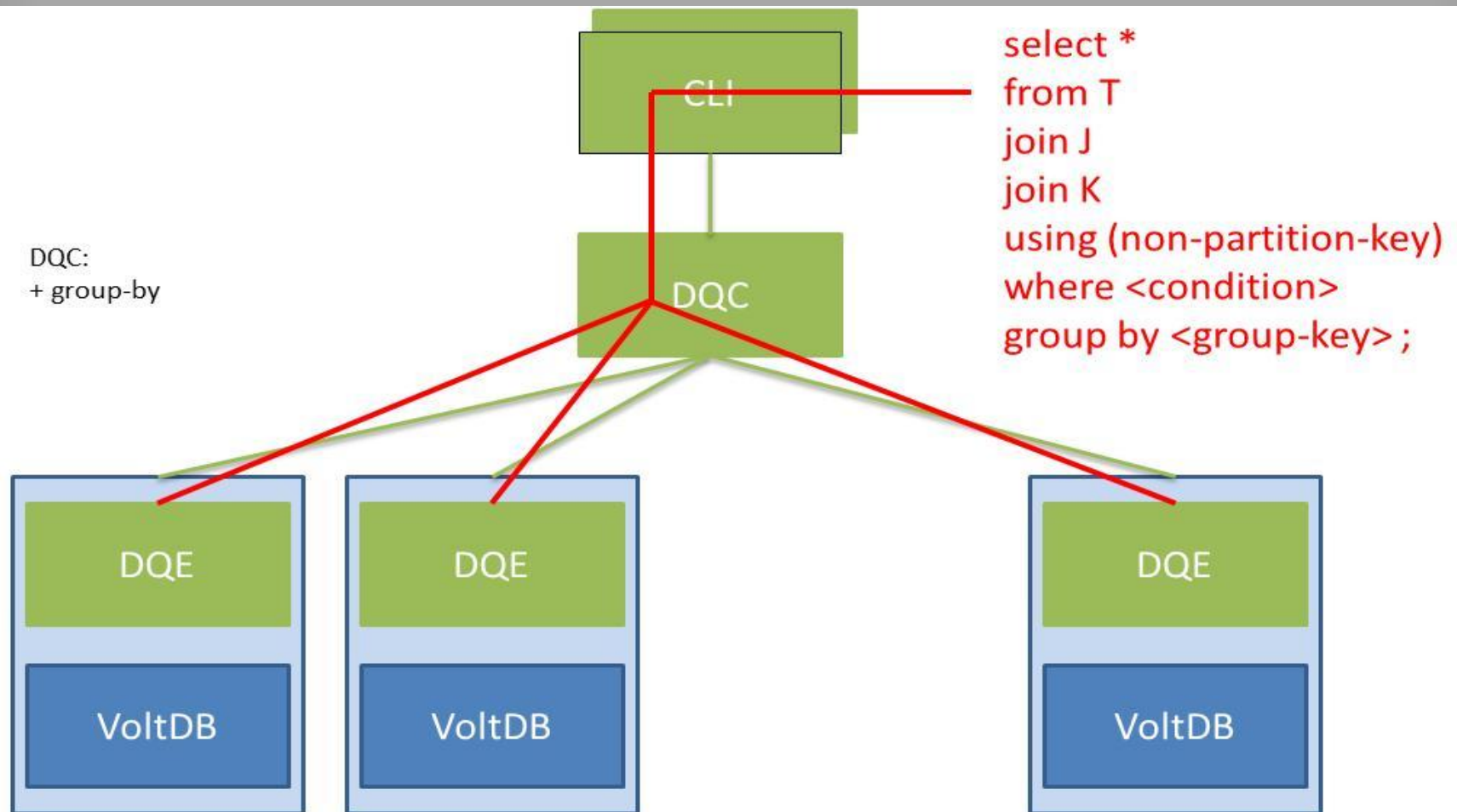
市场对大数据的要求 - 阶段I

统一的系统 - 解决大数据的存储

预先整合的系统，便于管理及使用

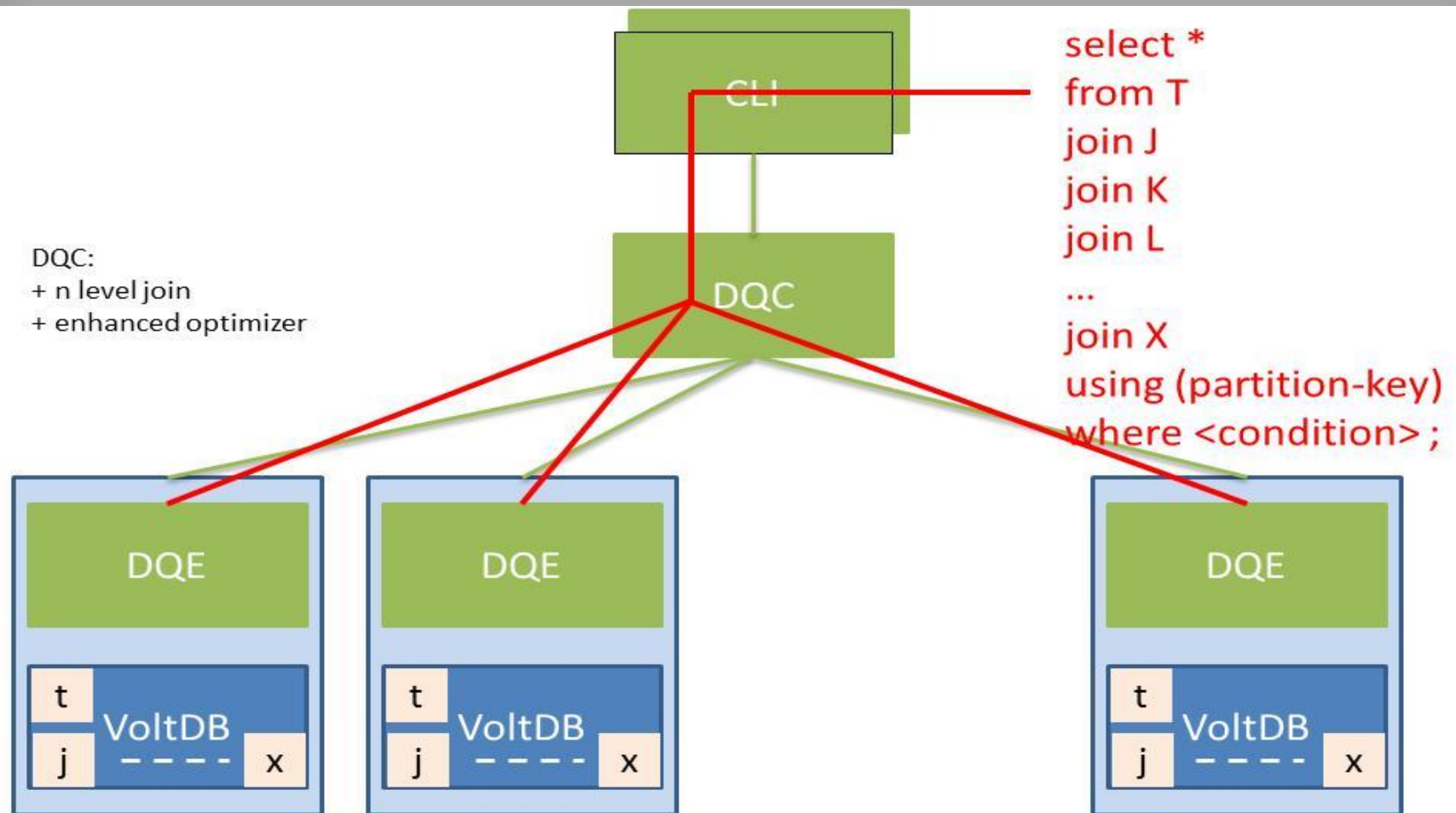
- 平台 - 大容量的索引 Pre-integrated using Hadoop Foundation,
- 整合的文本分析 - 面向非结构化的数据
- 可用性 - 易用的HDFS用户管理工具及查询工具
- 企业级功能 - 存储监控，任务调度工具，安全机制
- 支持 以基于搜索的，基于文本的XML数据模型
 - 文本存储于事务性存储库内
- Schema-Free:
 - 无需对数据库模式有深入的了解
 - 文字的索引与文本结构同时存在
- 充分利用标准的商业化硬件

实现细节I - Non-Partitioned 2-way Join + Group By



T & J & K are distributed. Aggregation of the group-by will need to be done on DQC

实现细节II - Partitioned n-way JOIN across n-nodes



All tables are distributed. Each part of the join can be fully resolved on each (due to using same partition key)

市场对大数据的要求 - 阶段II

■ 实时的流数据分析 - 解决流数据的分析

- 针对导入的数据执行实时的“流式”的分析查询
- 全速更新即时导入的数据
- 调度及执行上百个复杂查询
- 能够进行亿级维表和事实表JOIN, 同时无需对维表及事实表进行预处理
- 性能
 - 能以>GB/Sec的速率来进行流数据分析
 - 在使用高度规范的标准SQL查询时, 能有可预期的毫秒级反应速度
- 高可用性
 - 能达到无单点故障
 - 自动的故障恢复, 提供极端的高可用性
 - 自动的故障修复, 应当不会中断进程的操作即便是在节点故障的情况下

实现设想 - 围绕流式分析

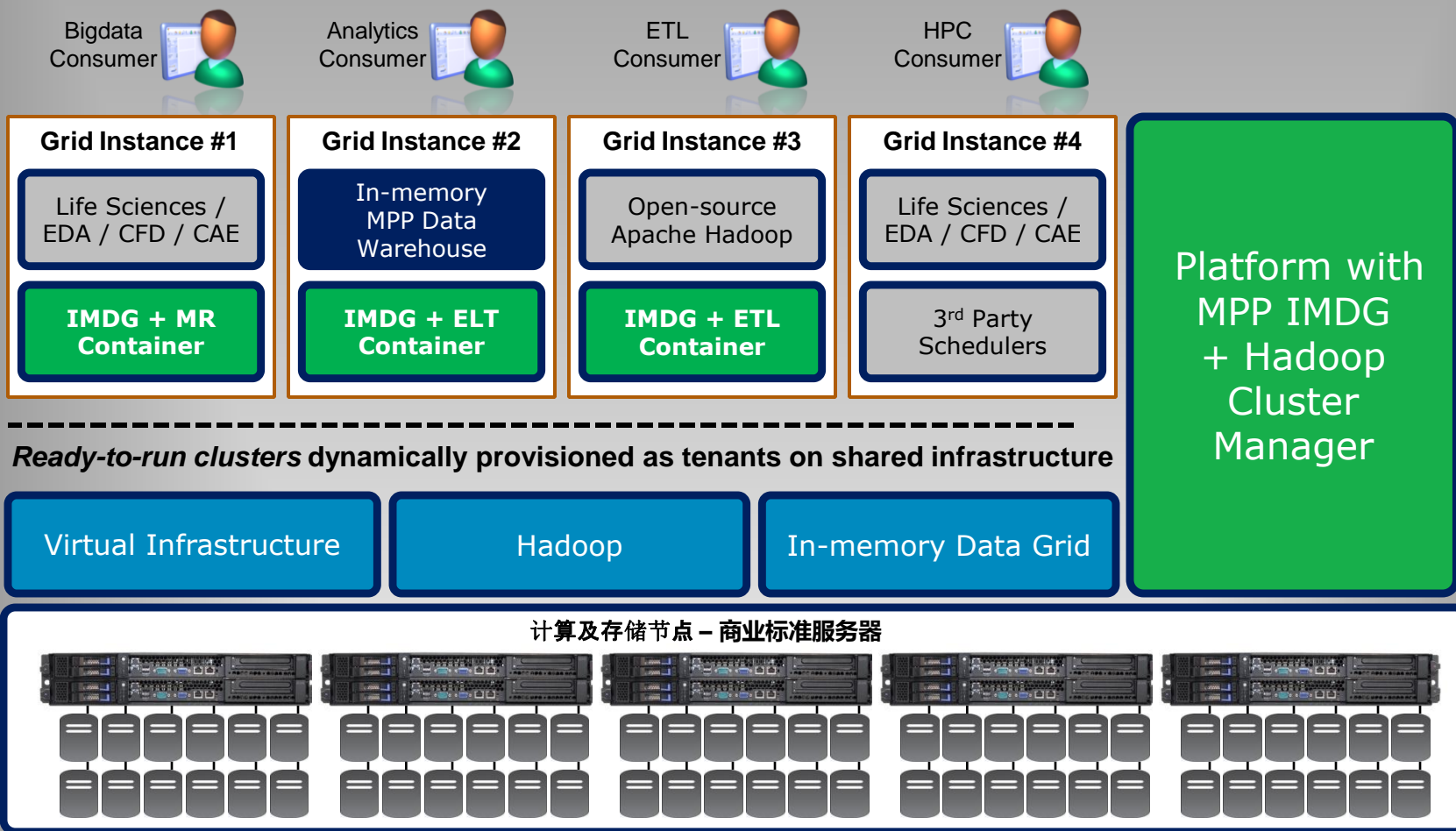
功能描述

- 1 SQL Windowing
- 2 Query concurrency up to 1000 concurrent queries
- 3 Scale-up testing to 1024 IMDB nodes
- 4 High Availability (ie recovery from DQS component failures)
- 5 Dynamic Cluster Management eg handling dynamic addition of IMDB nodes
- 6 Materialized Views (over JOINS)
- 7 Compression of network traffic within DQS
- 8 Workload Management - Master and Forwarding

市场对大数据的要求 - 阶段III

- 基于内存计算的分析应用部署：
解决即需的快速应用部署
 - 实时加载及大规模部署分析应用
 - 分析应用以虚拟机的形式存在
 - 大规模动态的调度内存节点为分析应用虚拟机服务
 - 分析应用全速响应外部App或传感器
 - App以桌面虚拟的方式来展现
 - 分析应用动态生成及执行大量复杂查询
 - 分析应用动态生成及调整复杂查询

类似于大型机的虚机及混合负载的分析平台



Q & A