大数据的开源工具 第一部分

Pekka Barck，Teradata天睿国际架构咨询总监

2016年5月5日

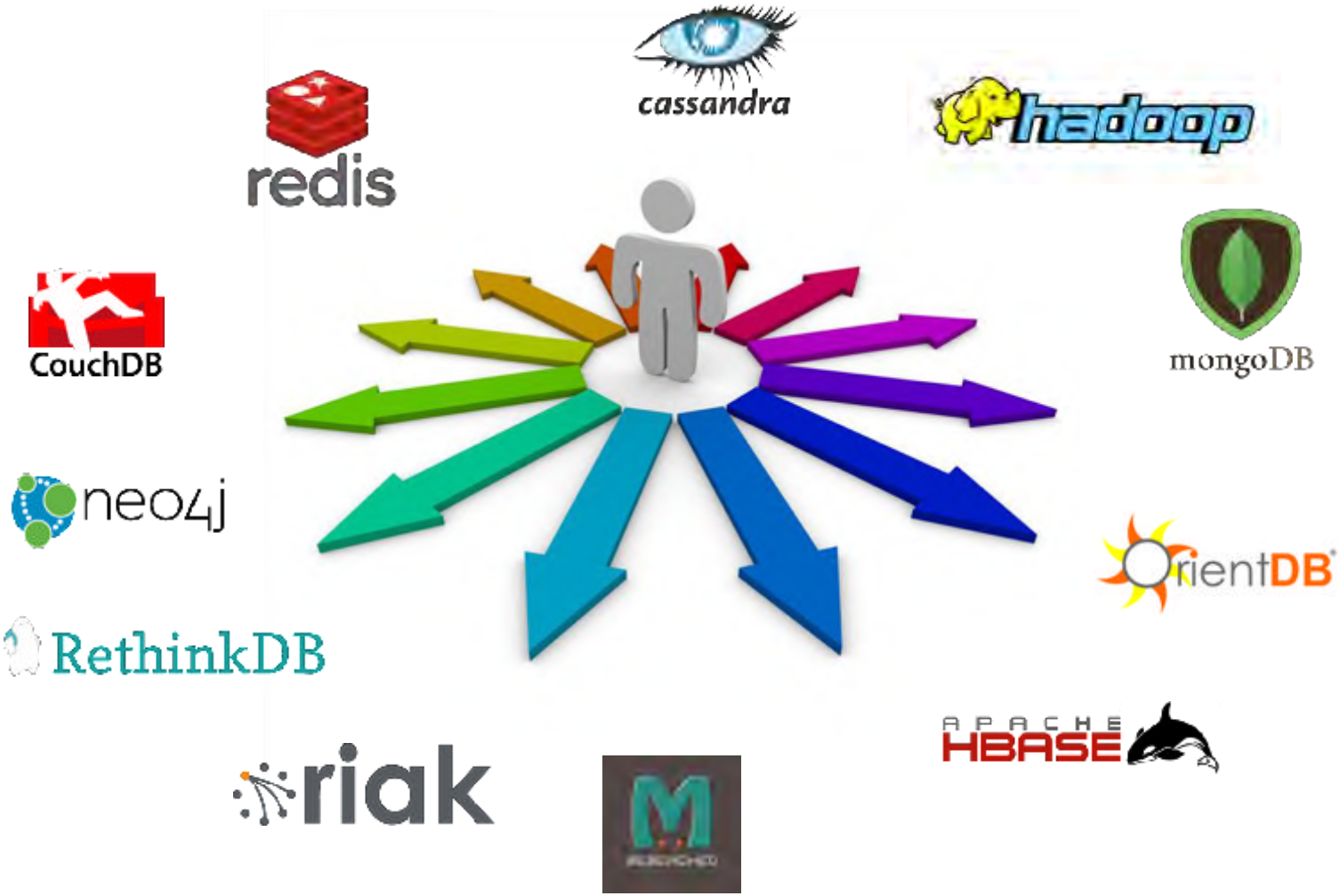
目录

- 本次演示从两个关键角度探讨开源产品：
  - 第一部分：平台
  - 第二部分：集成和摄取

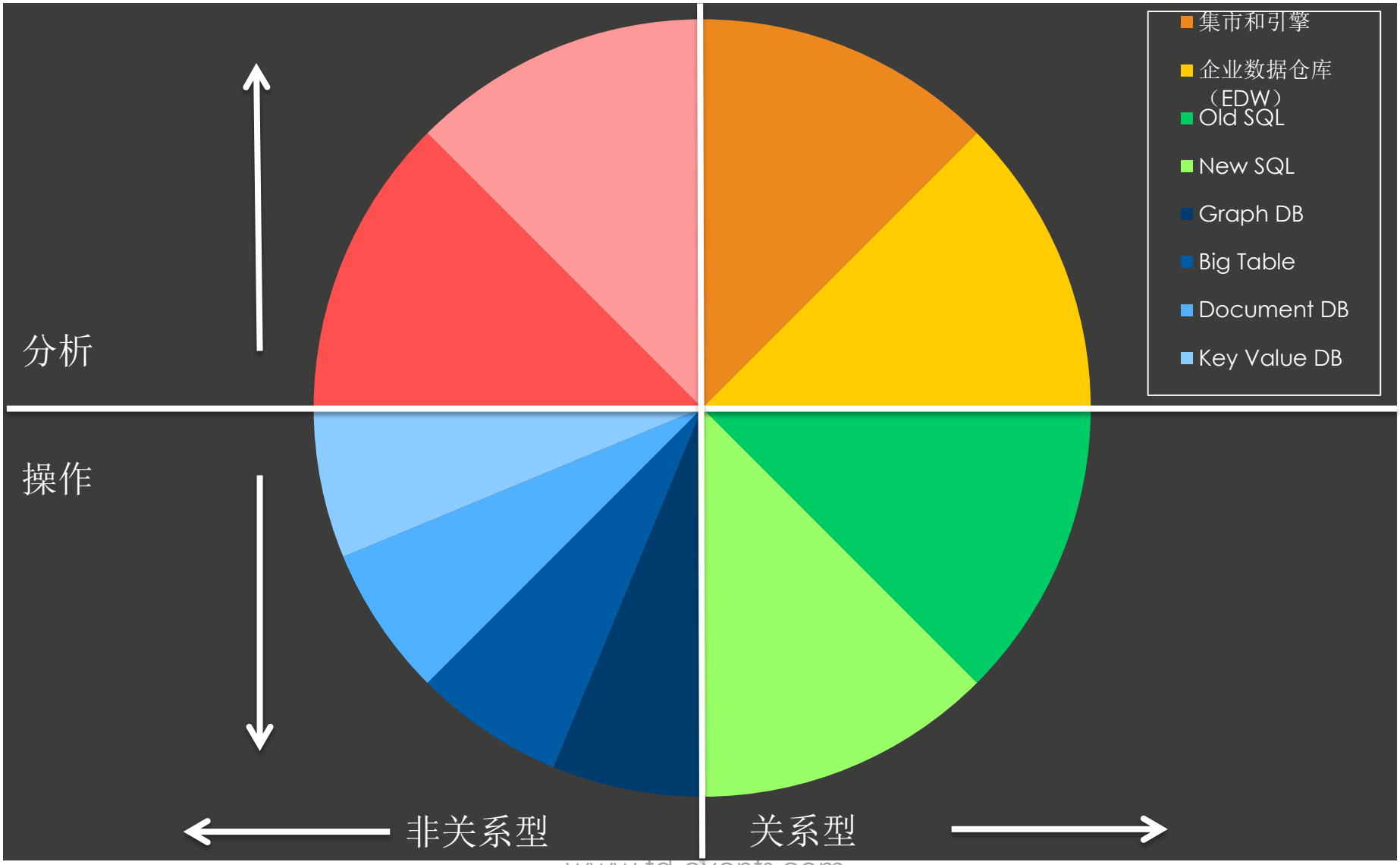
第一部分：数据平台

第一部分：平台——何去何从？

多个世界



存储维度：物理示例



你知道吗？

类似于Teradata的关系数据库的最大弱点是什么？

复杂的关系

TERADATA.

但开源也有他们的问题：

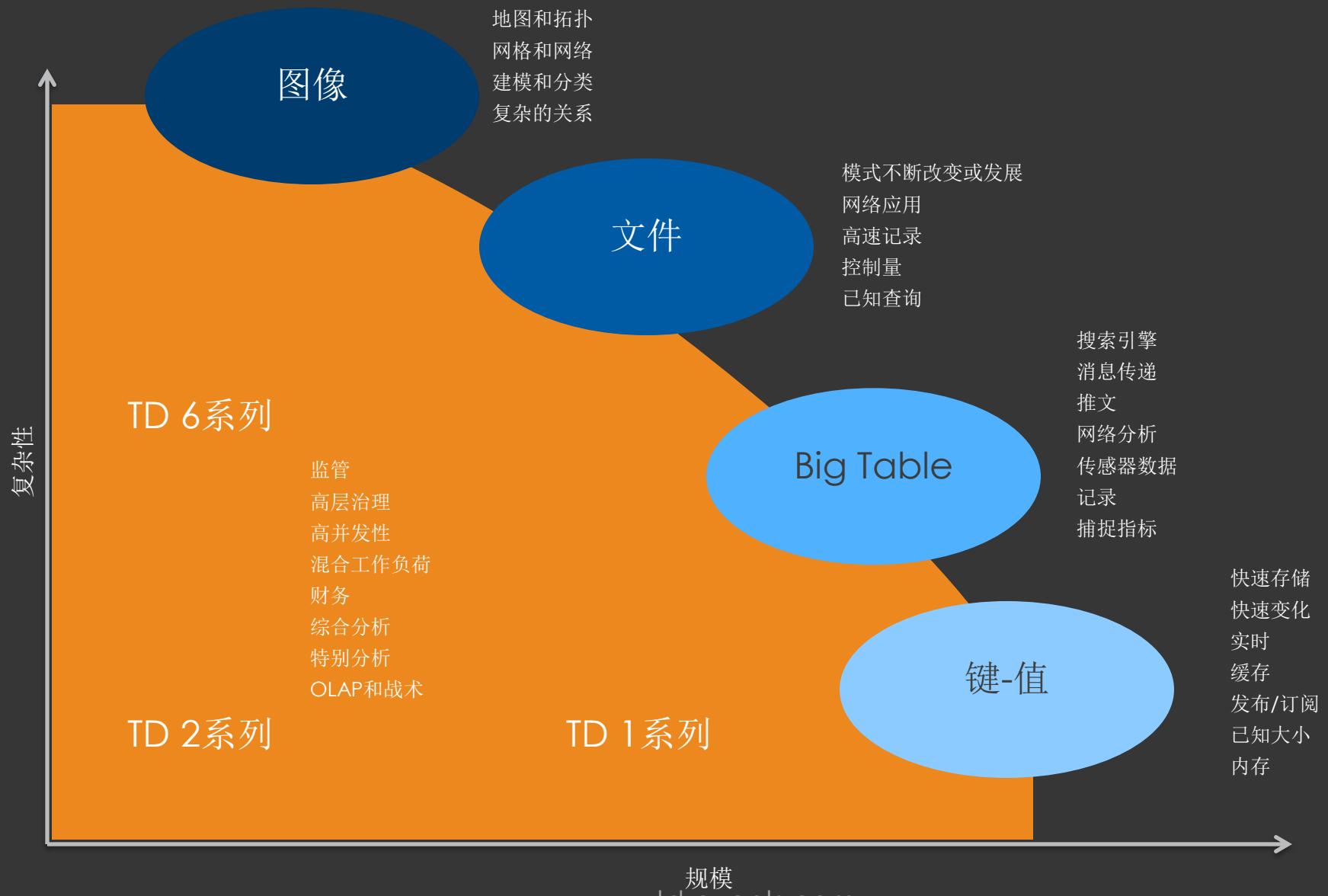
- 与开源NoSQL引擎相同，Teradata按线性方式扩展。
- 今年，“开源”NoSQL的市值为20亿美元。
- 在大多数情况下，开源NoSQL系统具有强大的分析性能。但是仍然存在下列问题：
  - 关联分析差
  - 无聚合
  - 很少或无索引
  - 无内置语言功能
  - 易用性差
  - 缺乏优化程序和工作负荷管理
  - 成熟度和企业可能性有限
  - 并发性水平低
- 使用带有传统分析数据库（如IBM、Oracle和Teradata）的XML和JSON，也可实现Schema-on-read。
  - 在没有schema的情况下，每个人应用数据的深度会参差不齐

结论

- 我想说什么呢？永远坚持使用Teradata？
- 不

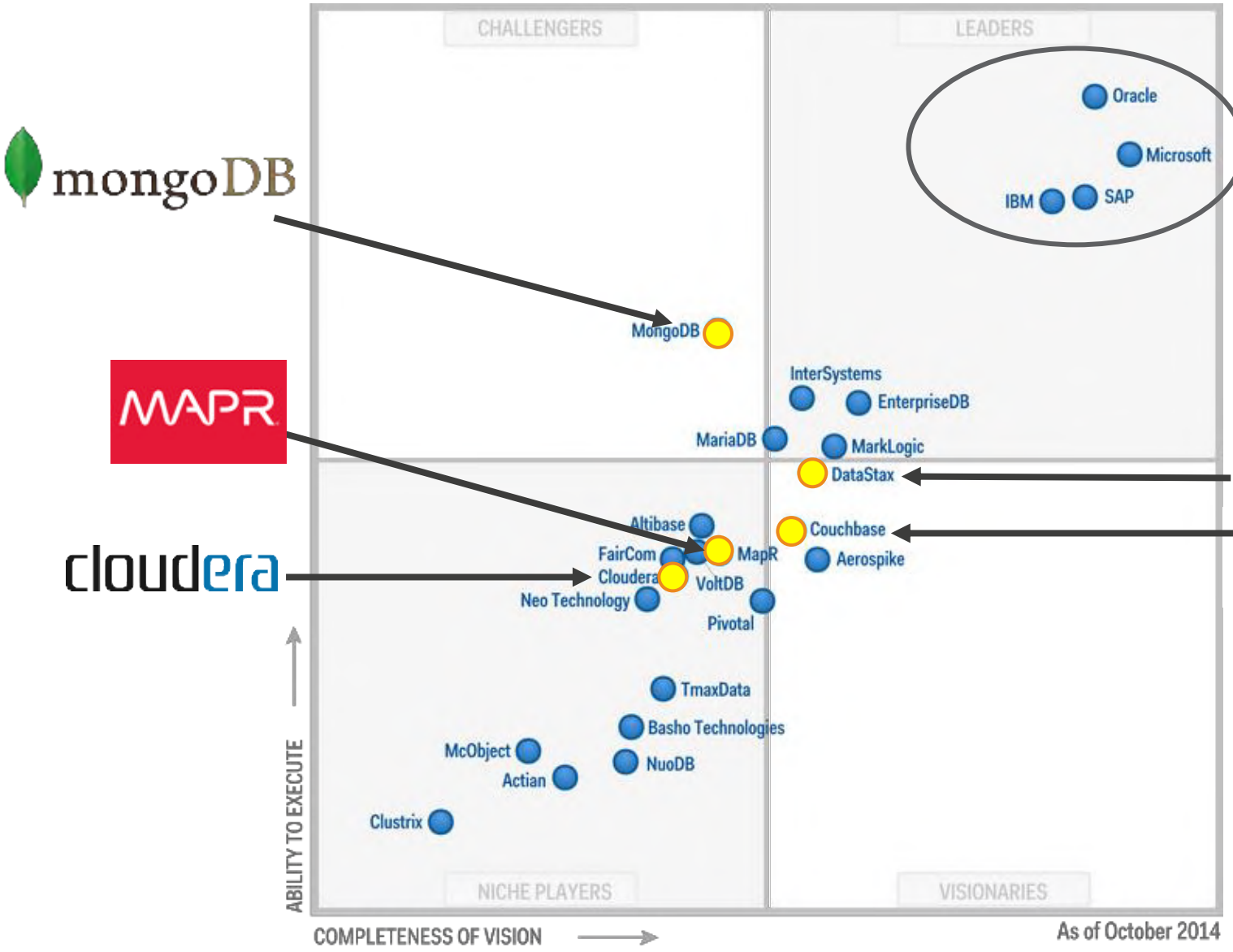
- 开源无疑是正确的方向，但应当务实前行。
- 总拥有成本。产生的实际效益。
  - 收入的影响
  - 节省的成本
  - 避免的成本
- 受益的时间。
- 长期效益。
- SQL和NOSQL的复合生态系统架构。

数据量和复杂性



TERADATA.

2014年10月*操作数据库管理系统*的魔力象限



Oracle
Microsoft
SAP
IBM

mongoDB

MAPR

cloudera

DATASTAX

Couchbase

TERADATA

数据库排名：网络数据库

供应商	4/2016	5/2015	5/2014
Oracle	1	1	1
MySQL	2	2	2
MS SQL Server	3	3	3
MongoDB	4	4	5
Cassandra	8	8	9
Redis	9	10	13
Teradata	13	13	12
HBase	15	15	15
Hive	16	17	18
Splunk	18	18	20
SAP HANA	19	21	23
Couchbase	24	24	28

- 数据库排名结合了：
  - 网络上提到的次数
  - 谷歌搜索显示的趋势
  - Stack Overflow and DBA Stack Exchange 提到次数
  - 工作机会数量
  - LinkedIn资料数量
- 无法衡量
  - 设施、收入
  - 无LinkedIn人员

http://db-engines.com/en/ranking

Big Table

- 2006:BigTable是一个稀疏的、分布式的、持久化存储的多维度排序映射。Bigtable的设计目的是快速且可靠地处理PB级别的数据，并且能够部署到上千台机器上
- 分布式NoSQL数据库
- 不同行的列完全不一样：灵活模式。
- 高可用性
  - 可以回放：以记录写入操作开始。
  - 数据复制而防止集群故障。
- 大数据量的可扩展性
  - 线性可扩展性。
  - 模块化可扩展性。扩展时无需重新启动集群。
  - 大型数据集需要保存很长一段时间。
- 有利于非易失性数据
- 要求已知访问模式：访问模式相似的列位于同一列族。

Hbase一览

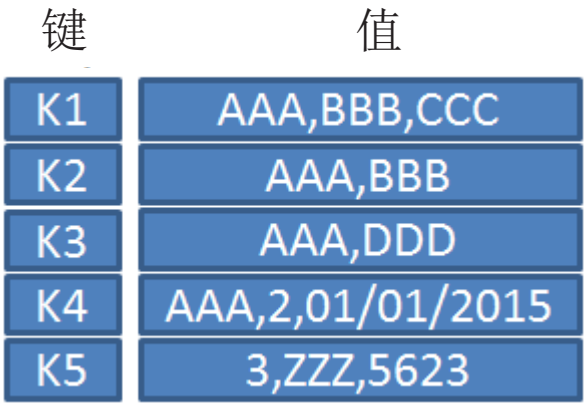
	详情
描述	主要用于存储、处理和更新大型数据集
主要特性	部署在HDFS上，分区到表，进一步分割为列族。
用例	• 日志分析或捕捉指标 • 支持搜索引擎或消息传递
优势	• 成批作业和延迟降低 • 与Hadoop集成 • 良好的读写一致性 • 非正则化数据 • 读取：给出快速搜索响应。 • 有利于进行基于距离的扫描。 • 比Cassandra更简单的一致性模型。
弱势	• 不可用于事务型应用或关系分析。 • 联合运算难以执行 • 聚合、上卷和分析交叉行速度缓慢。 • 全表扫描速度缓慢。 • 仅有一个表索引。 • 不包含任何查询优化程序。 • 结合MapReduce工作时，出现不可预测的延迟。 • 仅允许一个默认表排序 • HMaster是性能和可用性瓶颈。
竞争对手	• Cassandra、MongoDB、Couchbase、Oracle、Redis、VoltDB
客户	• 谷歌、雅虎、Facebook、eBay、Pinetrest…

DataStax Cassandra一览

	详情
描述	Apache Cassandra具有强大的集群，无模式，良好的管理工具。Hype：实时分析
主要特性	• 实时超快速写入，慢读，操作数据库 • 强大内存缓存功能（Memcache OSS） • Apache Spark内存中引擎，Spark streaming微量整批 • 多数据中心可扩展性
用例	• 个性化、电子邮件和社交媒体营销、物联网、欺诈检测
优势	• 高于平均客户满意度 • 多数据中心HA集群 • 内存中事务、搜索功能 • 几乎无意外停机 • 充满活力的开源社区 • 可调一致性。 • 随机读写速度快。 • 支持单行查询。最适用于实时快速查找。
弱势	• Cassandra查询语言——另一种较难的语言 • 产品成熟度——文档质量差、升级回归缺陷 • 在POC测试期间表现不佳（选择错误的应用程序） • 无参照完整性，无地理空间索引 • 不利于严格监管数据或严格授权规则 • 基于距离的行扫描有限。
竞争对手	• MongoDB、Couchbase、Hbase、Oracle、Redis、VoltDB
客户	• 450多家客户：Netflix、康卡斯特、Adobe、eBay、GNIP、Travelocity、Priceline、西夫韦、思科、Twitter、瑞银（UBS）、诺福克南方、腾讯

键-值数据库

- 键和值匹配。
  - 通过唯一键存储和访问数值。
- 无架构，无关系。
- 分布式数据库
- 有利于
  - 快速存储数据。
  - 快速修改数据。
  - 实时队列处理。
  - 缓存、存储供未来频繁使用。
  - 处理发布和订阅。
  - 处理数据状态。
- 限制因素
  - 数据量不可预知。
  - 数据集内存成本较高



键-值：Basho Riak一览

	详情
描述	分布式容错键值数据库基于Amazon Dynamo设计
主要特性	• 通过散列分布，形成一致的低延迟响应时间 • 自动化数据分布
用例	• 网络会话数据 • 网页内容管理 • 消息传递和聊天 • 全文搜索、索引、查询
优势	• 多数据中心集群+复制 • 无单点故障 • 不停机升级和维护 • 链接——单向轻量级数据关系 • LOB：JSON、图像、视频、电子邮件、UID、会话数据参考
弱势	• Riak无法加入对象——要求自定义编码 • 对Riak开源和实施的开发人员牵引较少 • 复合查询成本较高 • 功能缺失或较弱
竞争对手	Redis、MemcacheDB、MongoDB、Cassandra、MapR、HBase
客户	百思买、Healthx、Netflix、Rackspace（投资者）、波音、Comcast、特纳广播、气象频道、戴尔、Yammer、AT&T、AOL、Ask.com、赛门铁克、Github、vox Disqus。Basho拥有超过200家企业客户。

Redis概览

	详情
描述	Redis是一个极其高速的缓存中间件组件。它支持大规模网络负载，以及其他具有次秒级要求的应用。有几家门户网站使用Redis缓存所有最近的数据，以提供次秒级的用户体验。
主要特性	• 发布和订阅消息处理 • 通过快照和日志记录实现持久性 • HA主从复制（共3张） • 键可以是字符串、列表、集合或散列 • 范围和散列节点分区
用例	• 网络应用——广告推荐、粉丝列表、留言板跟踪 • 网络游戏 • 推文——向所有粉丝发送推文。 • 众包 • 移动应用
优势	• 简单键值服务器 • 简单查询系统 • 快速 • 适用于交易
弱势	• 数据集必须装入内存 • 集群
竞争对手	MemcacheDB、MongoDB、Cassandra、MapR、HBase

文档数据库

- 嵌套信息（递归和分层）
- 复杂的数据结构
- 启用Javascript。
  - 查询是Javascript表达式。
- 无模式数据库。

文档数据库：MongoDB

	详情
描述	MongoDB是一个开源文档数据库（JSON），扩展到100秒OLTP服务器，共用分区和复制
主要特性	• 自动分享横向可扩展性（无共享） • 二级索引，包括地理空间索引（B-树） • 广泛的语言支持
用例	• 内容管理、呼叫中心支持、产品目录 • 网络应用后端 • 半结构化内容 • 高速日志记录和实时分析
优势	• 应用程序代码开发较为简单 • 多模型存储- JSON、内存、WiredTiger快速写入 • 客户满意度高 • 合作伙伴生态系统丰富
弱势	• 高并发或高交易量 • 参照完整性需求 • 稳定性问题，特别是集群扩展 • ACID有限（写入到磁盘） • 解决查询的所有数据必须存储在BSON文档中
竞争对手	• CouchBase、CouchDB、Cassandra、Oracle、SQL Server、MySQL
客户	• 超过2500名付费用户：eBay、思科、Orange、ADP、西班牙电信、乐华梅兰、迪斯尼、Foursquare、Intuit、高盛；

文档数据库：Couchbase一览

	详情
描述	Couchbase = CouchDB + Membase（均为开源）。
主要特性	• 低延迟随机存取读写 • 双向复制 • MVCC - 写入操作不妨碍读取 • 升级期间不间断运行，任何节点都可以离线
用例	• 社交和移动游戏（Zynga）、移动数字钱包、互联网/移动广告目标
优势	• ACID一致性有限 • 键值或文档数据库 • 自动分区
弱势	• JSON + ISAM +内存缓存+复制 • 质量问题：故障或不可靠，部分发生在故障转移情况下 • 很难与其他数据库管理系统集成 • 查询是预定义的——更像视图
竞争对手	• MongoDB、DataStax、甲骨文
客户	• 450多家商业客户：乐购、LinkedIn、Orbitz、贝宝、美国在线、思科、DeutschePost/DHL、本田、salesforce.com、Vimeo、Zynga、尼尔森、NTT Docomo、Orange沃达丰、Sabre、韩国电信

图形数据库

- 树状结构，节点和边缘通过关系互连
- 有利于
  - 社交网络连接
  - 公共交通连接
  - 地图
  - 网络拓扑
  - 复杂的关系
  - 建模
  - 分类

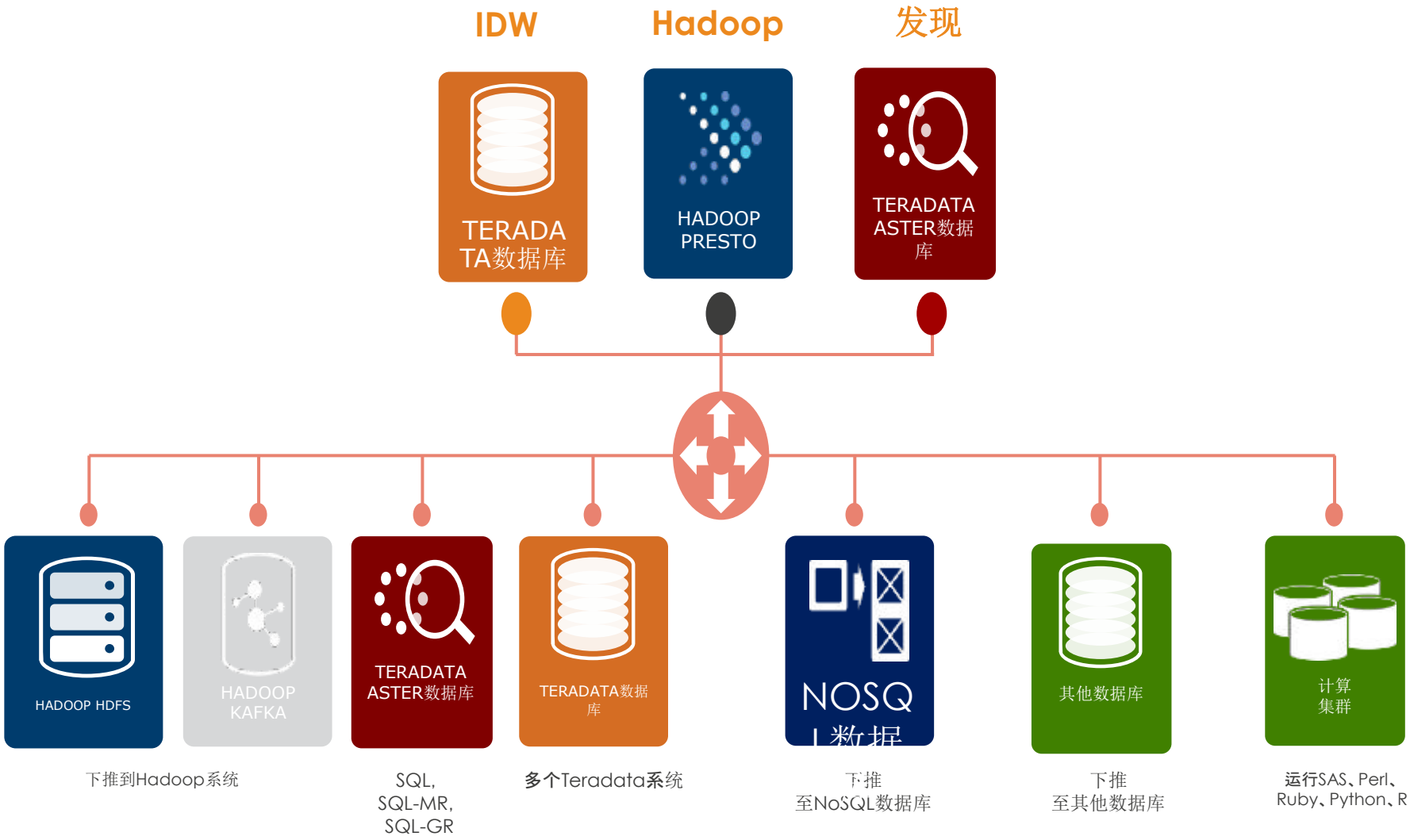
图形：Neo4J的更多信息

	详情
描述	Neo4J存储节点（键-值对）和边缘（节点之间的一个或多个指针）来表示关系。对于需要快速查询层次结构和关系的OLTP应用程序而言，倒退回到1970年代网络数据库的较快。
用例	实时建议引擎、社交网络、欺诈检测、网络配置、主数据管理层次结构
主要特性	• 本地图像存储和处理、模式自由 • ACID属性 • HA借助Hadoop Zookeeper、通过集群，只读缩放 • 设计用于OLTP • REST API，用于网络服务调用
竞争对手	Objectivity InfiniteGraph、OrientDB、GraphDB、YarcData、IBM DB2 NoSQL图形库、Oracle Spatial and Graph、Giraph GraphLab
优势	• 节点或关系索引——B-树、R-树 • Lucene文本搜索、排序、缓存、范围查询 • 查询速度比RDBMS快1000倍
弱势	• 存储的数据少；仅限节点/指针/属性标签 • 自带数字语言不是SQL或SPARQL • 预测分析有限；也只局限于图形
客户	Adobe、思科、德国电信、Tmobile、Intuit、必能宝、惠普、Gamesys、巴伐利亚；≈200家客户

日志引擎：Splunk

	详情
描述	Splunk通过搜索，对非结构化数据执行日志文件分析。 它主要帮助CIO管理服务器、网络和安全
主要特性	• 轻松构建BI仪表盘 • 非结构化数据搜索和报告 • DB连接到Oracle、SQL Server、Teradata（2015）
用例	• 网络日志、服务器日志、应用程序日志、安全日志、移动日志
优势	• 易于使用非结构化数据仪表盘 • 非结构化服务器日志索引 • 日志搜索、仪表盘和统计功能 • 企业安全应用、HadoopOps、Vmware、移动设备
弱势	• 无数据集成或复杂的数据模型 • “仅是一个日志搜索引擎” • 仅非结构化文本分析；非预测分析
竞争对手	• 服务器日志：Sumo Logic、TIBCO LogLogic、Ipswitch、SolarWinds、Tripwire • SIEM：IBM Q1 Labs、HP ArcSight、McAfee。
客户	• 来自90个国家的7400家客户：巴克莱（Barclays）、思科、威瑞森、中国移动、奥托、Target、乐购、百思买、ING Direct、LinkedIn、法国Orange、澳洲电信、美国邮政、梅西百货、Ping、家得宝、JC Penney、洛斯、沃尔玛、德国电信、沃达丰、康卡斯特、美国银行、贝宝、百事可乐等。

配有Presto的Teradata QueryGrid™——未来状态



TERADATA

第二部分：数据摄取

新业务用例

网站活动

物联网/传感器

电子邮件合规

销售高峰

跟踪物流

跟踪社交媒体

客户满意度

预订代理

建议

安全数据

TERADATA.

Apache Spark起源

- 服务器集群的平行中间件
  - Spark.apache.org（2014）
- 由加州大学伯克利分校AMPlab开发
  - DataBricks提供费用支持
- 流动数据小批量
- 最活跃的、快速增长的Apache项目之一：来自超过200个公司的750个贡献者。
- Spark将取代MapReduce：比MapReduce快10到100倍。开源Apache项目
- 顶级用例
  - SQL-on-Hadoop
  - 机器学习

摄取：Spark

- 有利于迭代和交互式处理。
- 缩放并适用于分布式系统。
- 相同的平台可用于实时和批处理
- 完全满足大多数情况。有利于批量、流、机器学习和图像。
- 在几个月让内存分析平民化。
- 代码是beta质量。
- Spark消耗大量的内存，缓存使用有限。
- 不同集群之间需要进行通信时，很难建立复杂流程。
- 导致较多，自己要补丁和手动修改的系统
- 内存问题很难处理。要求对配置有一定了解。
- 竞争产品：Apache Storm、Apache Flink、DataTorrent RTS

Spark包括多个组件

Apache Storm起源

分布式和容错实时计算

- 起源于BackType/Twitter，2011年底开源
- 在Clojure中实施，结合一些Java
- 基于disruptor模式，实现低延迟复杂事件处理
- 开源：12个核心提交者，加~ 70个贡献者

摄取：Storm

- 分布式计算框架，主要针对流媒体实时分析。
- 每个节点可处理高达1亿字节/秒。Spark无法企及。
- 可以执行单个事件
- 内存占用低于Spark。
- 良好的容错性：如果一个worker失效，Storm将重新启动它。
- 针对流媒体的缩放比Spark更好
- 经常与Apache Kafka结合使用。
- Storm需要ZooKeeper（用于协调），这可能成为一个瓶颈。
- 竞争产品：Microsoft StreamInsight、TIBCO StreamBase、Amazon Kinesis、Software AG Apama。

Apache Kafka起源

- 起源于LinkedIn，2011年开源
- LinkedIn尝试管理：
  - 2000亿事件/天
  - 700万事件/秒（写）
  - 3500万事件/秒（读），通过复制
- 今天…，截至2014年，LinkedIn甚至不是最大的Kafka用户！
  - Netflix、Twitter、Spotify、Loggly、Mozilla、Airbnb、思科、Gnip、Squer、优步等……
- 用于下载各类事件，例如：
  - 指标：操作遥测数据
  - 跟踪：LinkedIn用户的所有功能
  - 排队/集成：LinkedIn应用之间

摄取：Kafka

- Kafka在大多数大数据流应用中扮演着重要的背景角色。
- 适用于大数据中的分布式发布-订阅消息传递的“标准”
- 消息吞吐量高。通过LinkedIn以每秒200万写入进行基准测试。允许经纪人支持成千上万的客户，同时保持分布式提交日志的容错性。
- 入站流智能分区，允许并行读写。
- 消息被保存。这促进倒带。有利于实时处理数据以及随后进行批量分析。
- 通过分布式提交日志实现耐久性。
- 使用数据流智能分区进行并行读写。
- 这些分区数据流复制到数量可配置的复制品中，进而被写入磁盘，防止数据丢失，并启动“回放”数据流历史的能力。
- 竞争产品：RabbItMQ、ActiveMQ、Apache Flume、Teradata Listener

摄取：NiFi起源

自动化系统之间的数据流

- 最初由美国国家安全局开发
- NiFi是一个Apache孵化项目
- 很多美国政府机构在使用。
- 如何与Spark和Storm区分？
  - NiFi启动双向和点对点，例如EAI应用程序。
- 适用于敏捷开发和测试
- 数据流可视化。
- 配有数据流模式。

Apache NiFi

- 通过集群扩展
- 良好的细粒度流特定配置
  - 容错性。支持所有排队数据的缓冲。持续预写式日志。
  - 很少延迟和吞吐量
  - 优先级排序。启动队列重要数据流的优先级排序。
  - 转换。可以执行简单的转换。
  - 提供数据沿袭，可追踪数据来源。
  - 透明度：NiFi记录和优化流的元数据。
- 实时调试和数据流变化
- 有利于更好地保护数据中心之外,如传感器或其他数据源的数据。
  - 双向SSL
  - 可插拔授权
  - SSH、HTTPS、加密内容

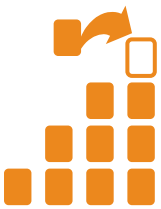
Teradata Listener与自己动手（Do-it-Yourself）

建造完成，随时可以运行

- 预先集成软件
- 强大的功能和平稳的升级

单一供应商支持

- 一个电话
- 世界级支持



可扩展性和可用性

- 设计用于集群增长
- 内置故障转移

企业平台

- 多个用例
- 关键任务

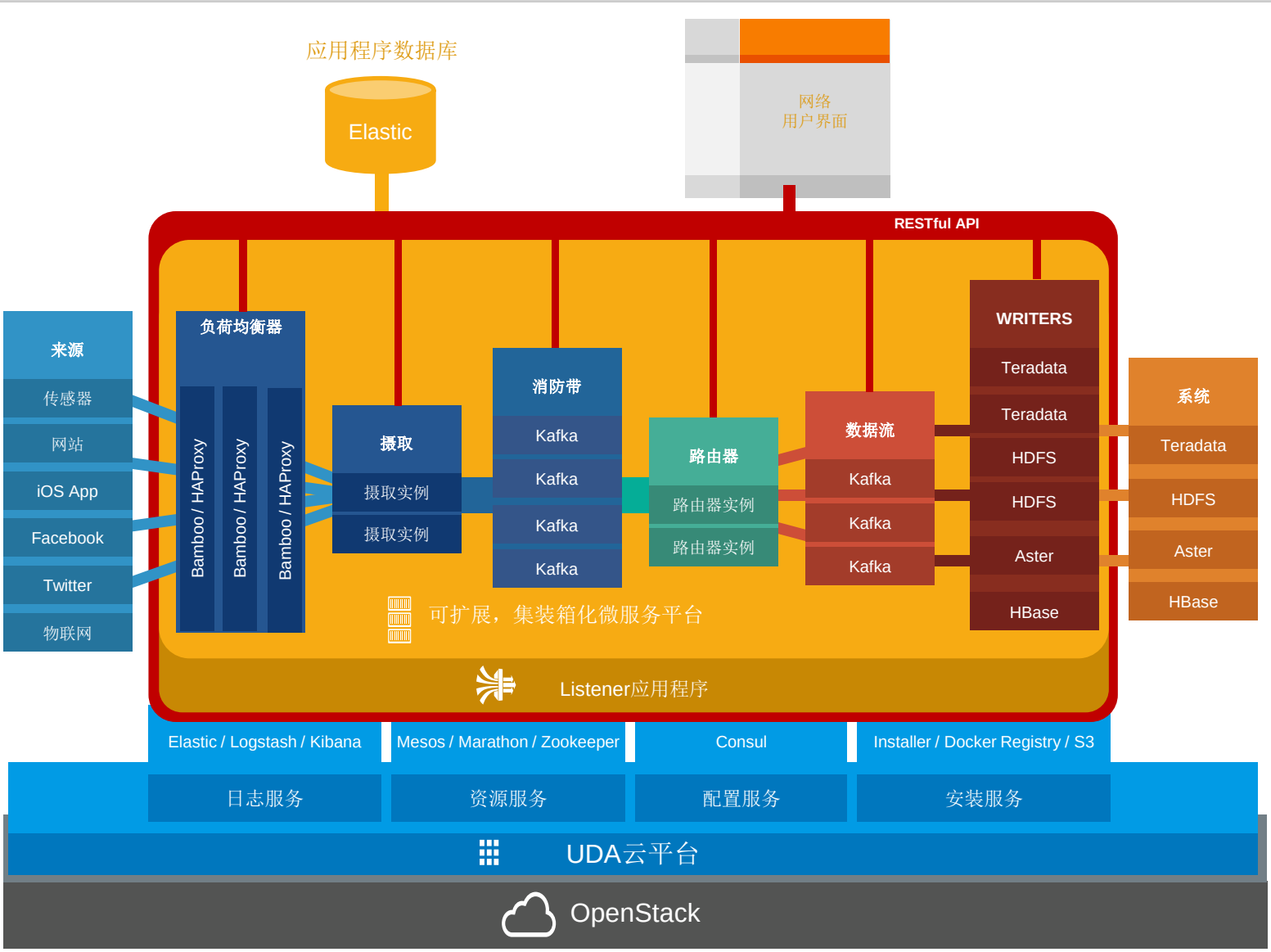
预构建UDA集成

- 简单连接
- Teradata、Aster、Hadoop

自助服务与治理

- 生产力
- 业务控制

Listener逻辑架构



Listener架构

- Teradata Listener由两组组件构成
- 构成Teradata Listener应用程序的服务集合

- 为未来UDA解决方案提供公共平台的全套服务