VERTICA

Vertica统一分析仓库
最闪亮的那一颗皇冠上的明珠

刘定强 Vertica亚太首席技术专家/Field Chief Technologist

MICRO FOCUS

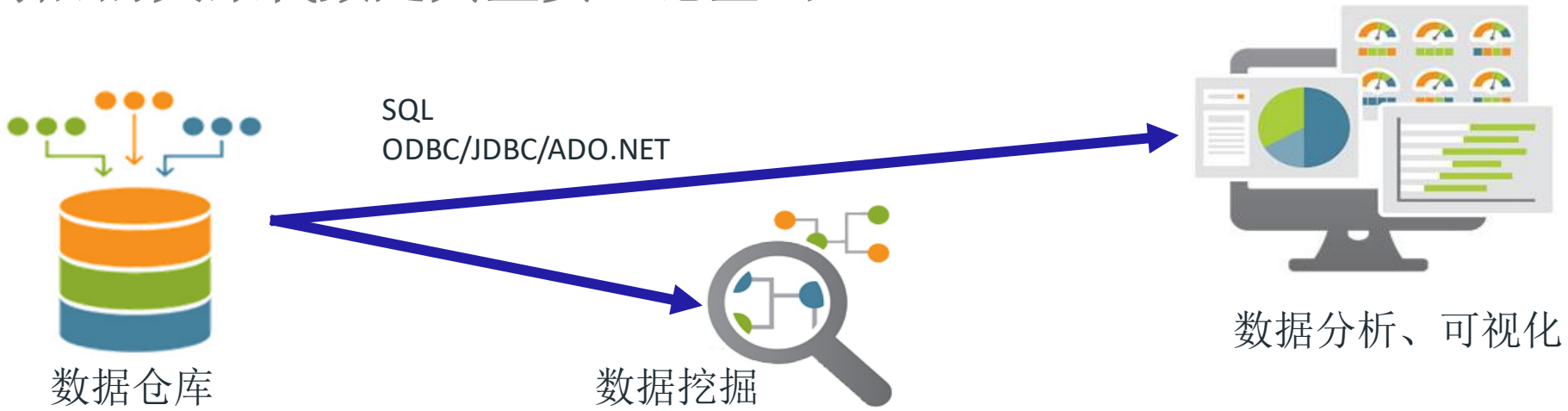
都说数据是金矿，数据库就是淘金的利器 和金库

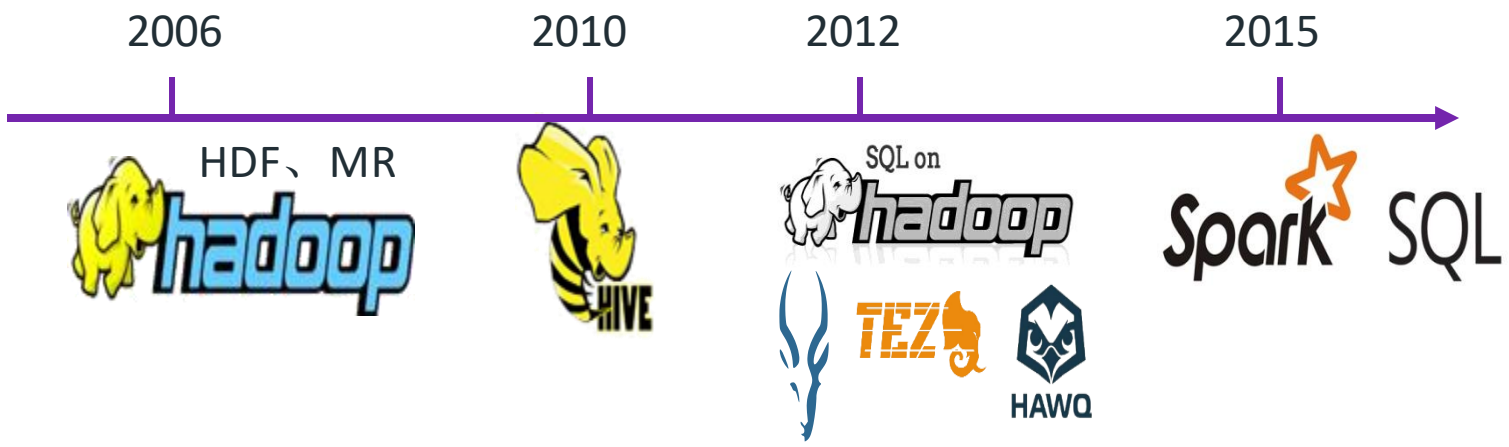
数百亿的市场规模

仍然在快速增长，尤其是大数据平台

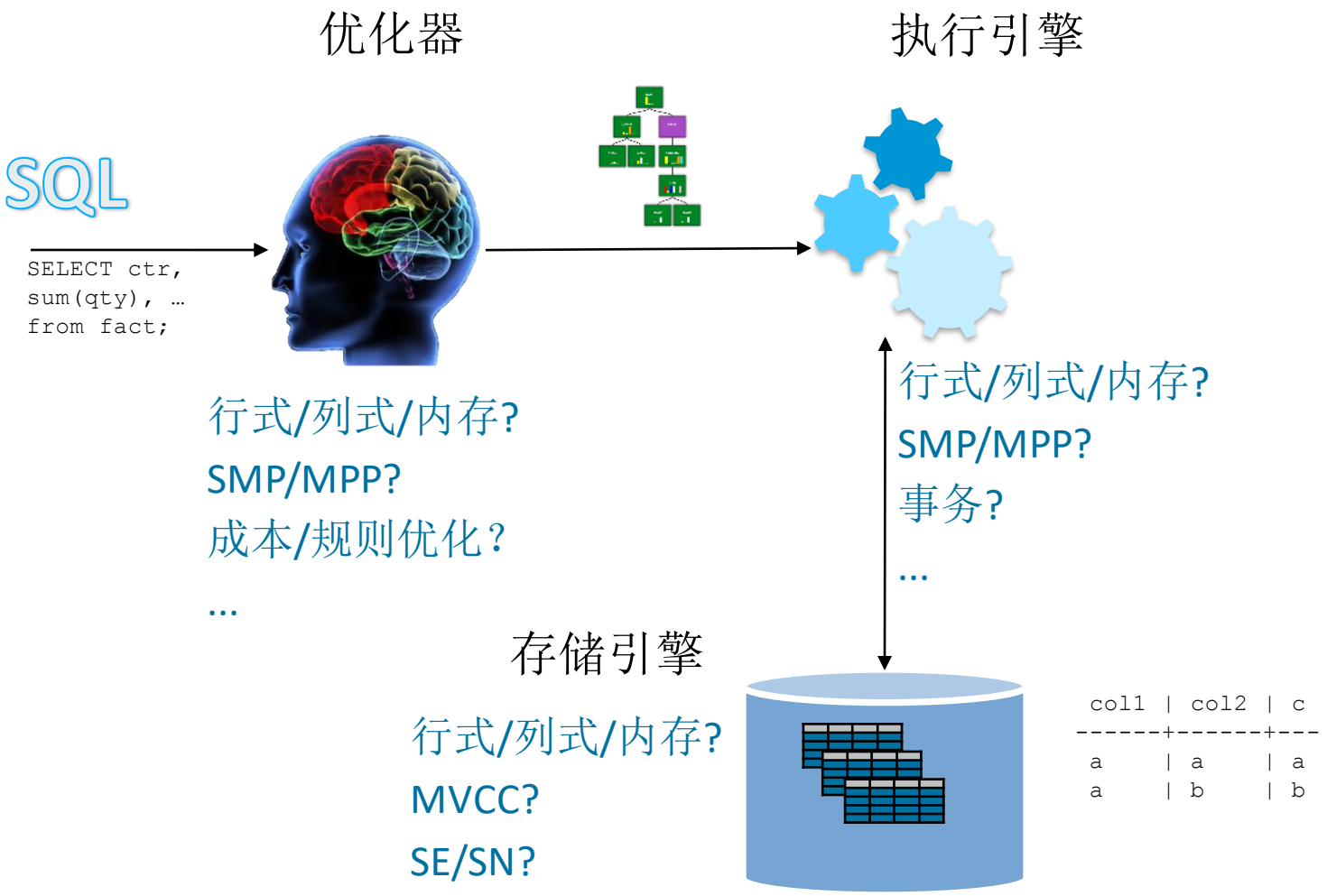
数据库有很多种，但SQL数据库仍是大多数人的最爱

背后的关系代数是其坚实理论基础





50来年数据库技术的发展，沉淀下来的稳定架构



数据库的核心部件：

优化器是核心

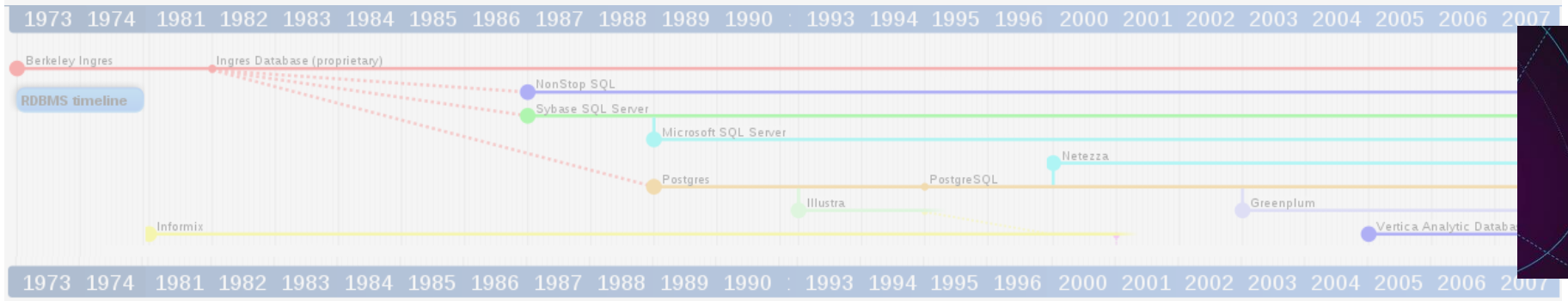
执行引擎是能力的关键

存储引擎是性能的基础

定位不同的数据库，会权衡不同的技术

One size does Not fit all !

Mike Stonebraker 亲手创建的数据库



70-80s 80-90s 00-10s

Ingres → Postgres → Vertica

Sybase
Microsoft SQL Server
NonStop SQL

Greenplum
Teradata Aster
Netezza
ParAccel
RedShift

Vertica
Vertica SQL on Hadoop

Mike Stonebraker
2014年度图领奖获得者

创造了数据库系统一系列奠基性基本概念和实际技术

一代比一代更好的数据库技术

MPP架构数据库的缘起

6. CONCLUSIONS

In scalable, tunable, nearly delightful data bases, SN systems will have no apparent disadvantages compared to the other alternatives. Hence the SN architecture adequately addresses the common case. Since SN is a nearly free side effect of a distributed data base system, it remains for the advocates of other architectures to demonstrate that there are enough non-tunable or non-scalable or non delightful problems to justify the extra implementation complexity of their solutions.

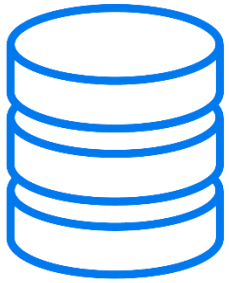
–The Case For Shared Nothing
Stonebraker, ‘85

VERTICA

什么是 Vertica？

Vertica 是一款高级分析平台，专为应对当前数据驱动型世界的规模和复杂性而精心打造。这款平台可将高性能 MPP 查询引擎的强大功能与高级分析和机器学习完美结合。

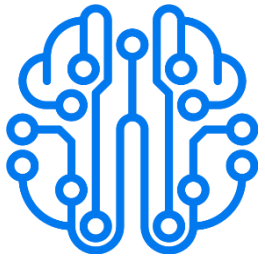
SQL 数据库



将数据**加载和存储**到专为**超快速分析**打造的数据仓库之中

+

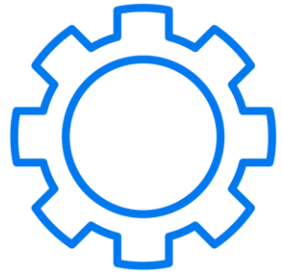
分析和机器学习



创建、**训练**和**部署**大规模的高级分析与**机器学习**模型

+

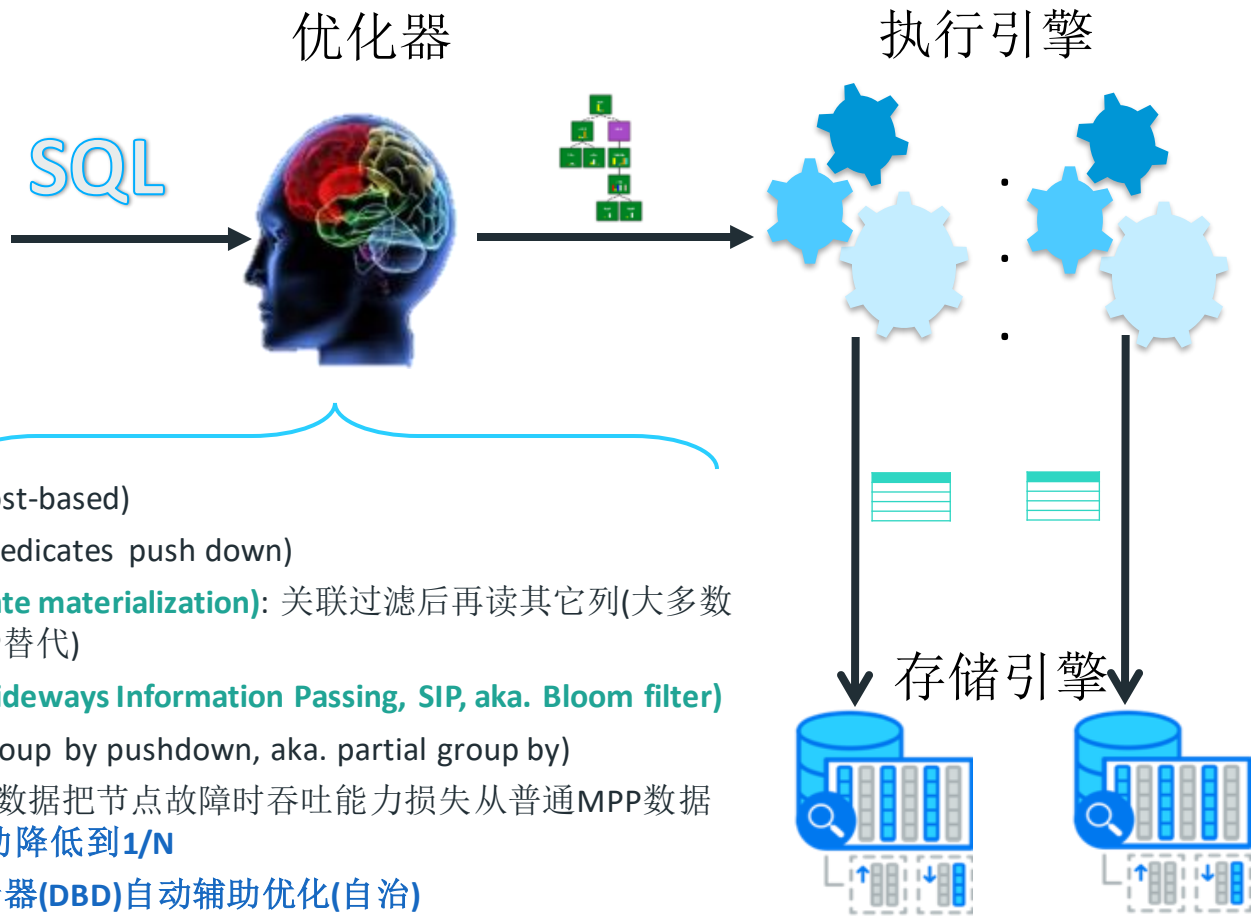
查询引擎



提出复杂的分析问题并快速**获得解答**，无论数据位于何处

Vertica核心的技术

最大化利用现代化的列式存储和计算、并行计算优化技术



- 成本优化(cost-based)
- 谓词下推(predicates push down)
- 延迟物化(late materialization): 关联过滤后再读其它列(大多数场景已被SIP替代)
- 关联下推(Sideways Information Passing, SIP, aka. Bloom filter)
- 分组下推(group by pushdown, aka. partial group by)
- 利用多副本数据把节点故障时吞吐能力损失从普通MPP数据库的1/2自动降低到1/N
- 数据库设计器(DBD)自动辅助优化(自治)
- 行级权限(row access policy)，列级权限(column access policy)
- ……

- 拉模式(pull mode)、向量化(vectorization)、即时编译(JIT)
- 同时支持无主节点并行(nodes parallelism)，多线程流水线并行(inter-operator/pipelined parallelism)、操作符并行(intra-operator parallelism)
- 直接处理列式编码数据
- 存储有序关联/分组(pre-sorted merge join/grpby)，分区Hash关联/分组(partitioned hash join/grpby)
- 资源池(resource pool)静态、动态隔离负载，子集群(sub clusters)横向隔离负载
- 容错组(fault groups)、分层网络(terrace routing)降低网络开销，提升可扩展性
- ……

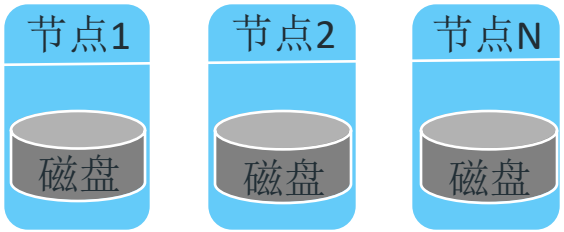
- 列式有序 (DSM, LSM-tree with WOS/ROS/DV container + position index，相当于索引组织表)一致哈希分布存储(分段或分片订阅)，自动管理K-safety容忍任意K个节点故障，多版本并发控制(MVCC)，支持一表多projections(相当于物化视图或索引)
- 无模式表(flex table, aka. schema on read)
- 智能裁剪(filter by position index, aka. zone maps)，按列取值范围裁剪文件和块，灵活的表分区表达式
- 实时聚合(live aggregate)，自动宽表(flattened tables)
- ……

VERTICA

50年来关系数据库技术积累！

MICRO FOCUS

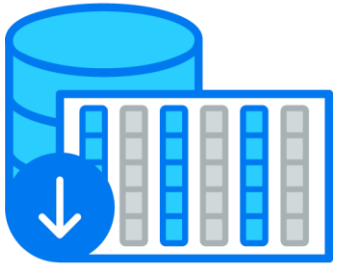
性能根植于 Vertica 的核心架构



大规模并行处理

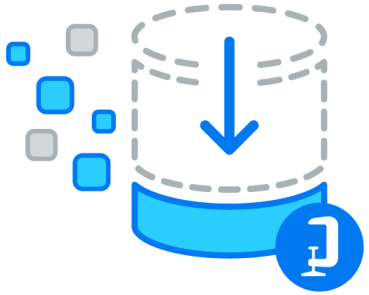
节点内、节点间并行，
无主节点、无单点故障，

线性扩展以支持更快的
性能或更多的用户



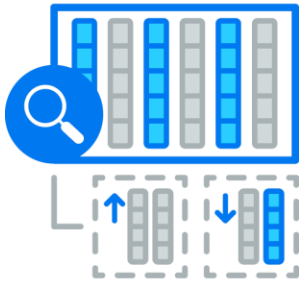
列式存储

只读取必要的数据，
降低 I/O 提升性能



高级压缩

减少高达90%的存储空间，进
一步降低I/O提升性能



性能优化的Projections

对查询优化的存储格式，
实时聚合、自动扁平化

自动化辅助设计优化

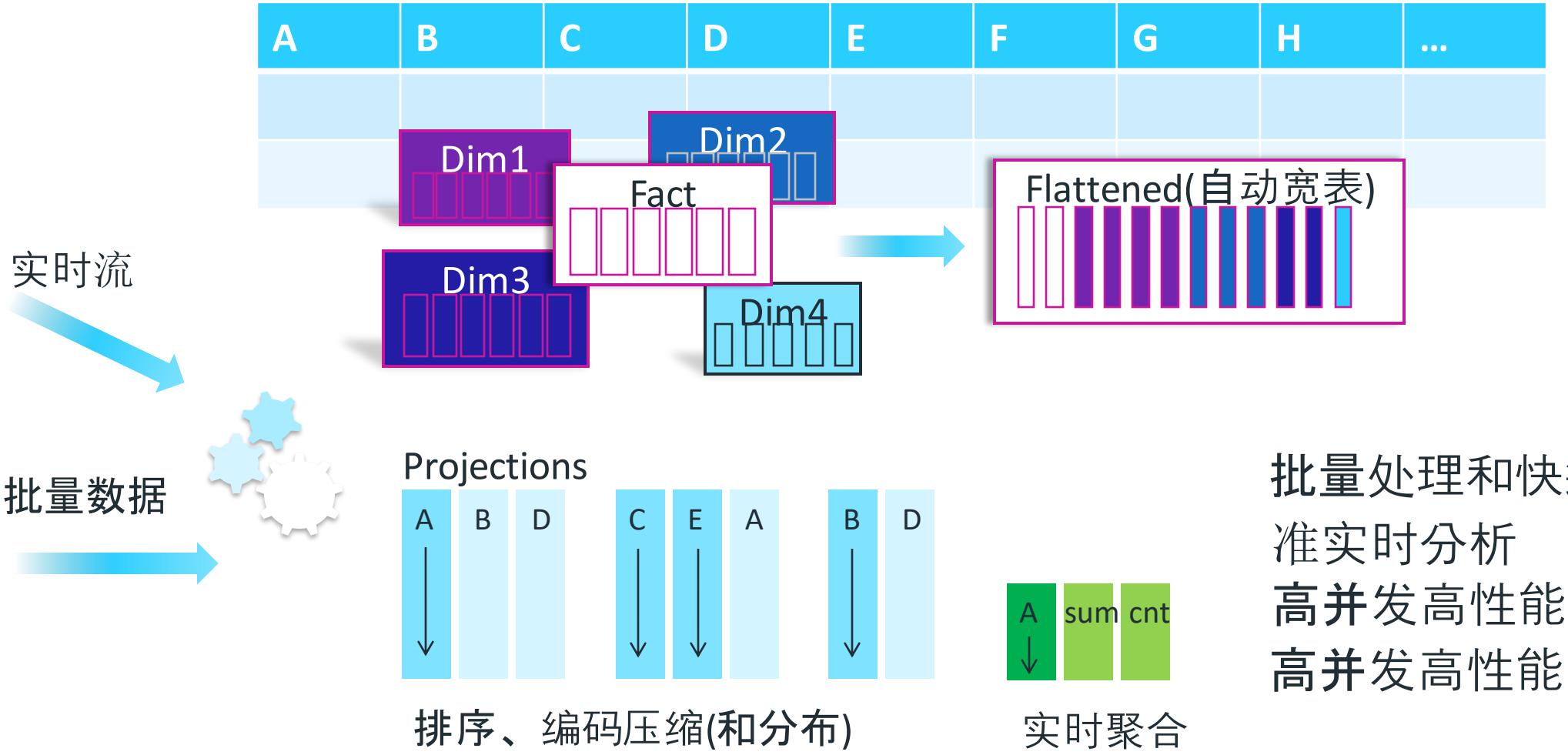
- ✓ 比传统数据库 **快 50x ~ 1000x**
- ✓ 比 Hadoop 等数据湖 **快 1 个数量级**
- ✓ 不牺牲**ACID**事务和**SQL**标准特性，**性价比更高**

VERTICA

MICRO FOCUS

批量和准实时分析融合

多种排序和分布方式存储、实时和批量数据加载、实时聚合、自动宽表



批量处理和快速分析

准实时分析

高并发高性能实时聚合

高并发高性能实时关联

云原生的计算存储分离架构(Eon Mode)

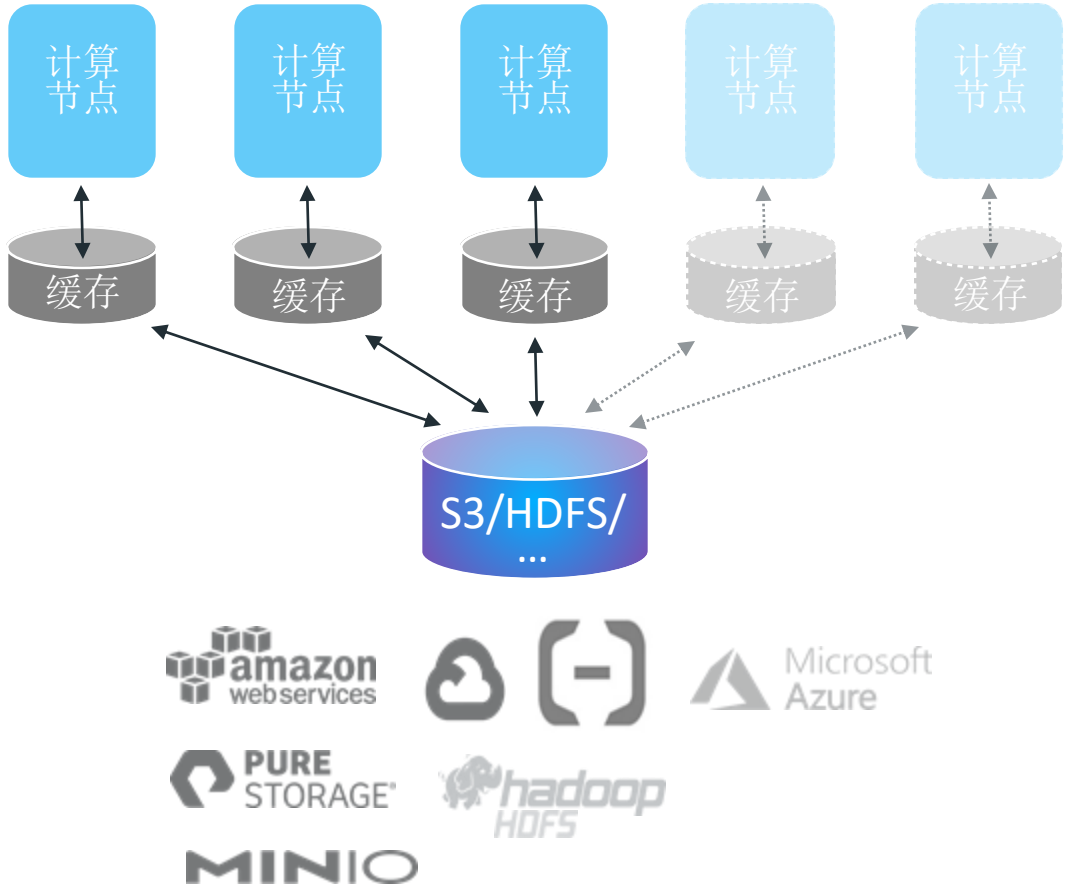
多云，混合，随处部署

分钟级快速线性扩展基础架构
弹性扩展应对工作负载变化、季节性或高峰负载时期

隔离分析工作负载，提升性能
子集群隔离工作负载并，提升并发/吞吐能力，多租户

简化数据库操作，数据容易共享
快速部署，极速恢复故障节点，出色的工作负载均衡

闲时休眠，用时开启
需要分析时可快速开启，不需要时可随时休眠计算节点



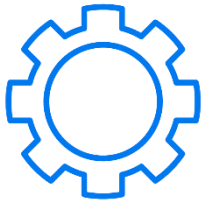
从根本解决MPP数据库扩容、故障恢复和高并发扩展三大难题

端到端的库内机器学习

Vertica 支持整个预测分析流程

SQL 数据库

分析和机器学习



查询引擎

数据分析

- 统计摘要
- 时间序列
- 创建会话
- 模式匹配
- 日期/时间代数
- 窗口分区
- 序列
- 等等...

数据准备

- 异常值检测
- 标准化
- 非平衡数据处理
- 采样
- 缺失值插补
- 等等...

建模

- 支持向量机
- 随机森林
- XGBoost
- 朴素贝叶斯
- 逻辑回归
- 线性回归
- 岭回归
- 交叉验证
- 等等...

评估

- 模型级别的统计数据
- ROC 表
- 错误率
- 提升表
- 混淆矩阵
- R 平方
- MSE
- 等等...

部署

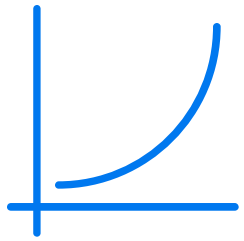
- 任意部署
- 数据库内评分
- 大规模并行处理
- 速度
- 可扩展性
- 安全性
- 等等...

Vertica 数据库内机器学习的优势

数据库内机器学习改变了数据科学家和分析师与数据交互的方式

规模扩展

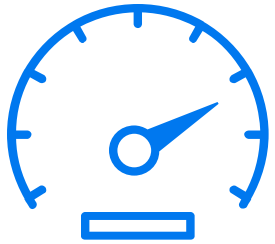
普及预测分析应用程序



让组织内的更多用户通过简单的 SQL 界面使用机器学习

速度

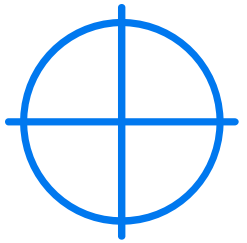
缩短机器学习项目的上市时间



利用 Vertica 的大规模并行处理能力，以业务所需的速度构建和训练模型

准确性

部署预测用例并保持竞争优势



以所有历史数据为基础运行机器学习模型，并非只基于部分缩小采样的数据集

使用 VerticaPy 将分析提升到新的水平

SQL Back-end

Much of the heavy computation is done by Vertica. Model storage & management in Vertica.

Open-Source

Users can contribute. No added cost on software. Constantly updated roadmap.

Python Front-end

Python objects which expose scikit/pandas-like functionnality

High Security

Data stays in Vertica (security, integrity, scalability).

Jupyter Notebooks

Jupyter/Python – popular tool of choice for data scientists & analysts

in-DataBase ML

Fast & Scalable, Vertica ML is unique in the ML space with its C++ implementation.

Rendering Capabilities

Dynamic & Responsive Charts using Matplotlib & Highcharts

HIGHCHARTS matplotlib

Integrations

Simplify models deployment using Vertica ML integrations.

TensorFlow PMML λ

GitHub

Issues are solved quickly. The API is quickly evolving. Unitary Tests are available.

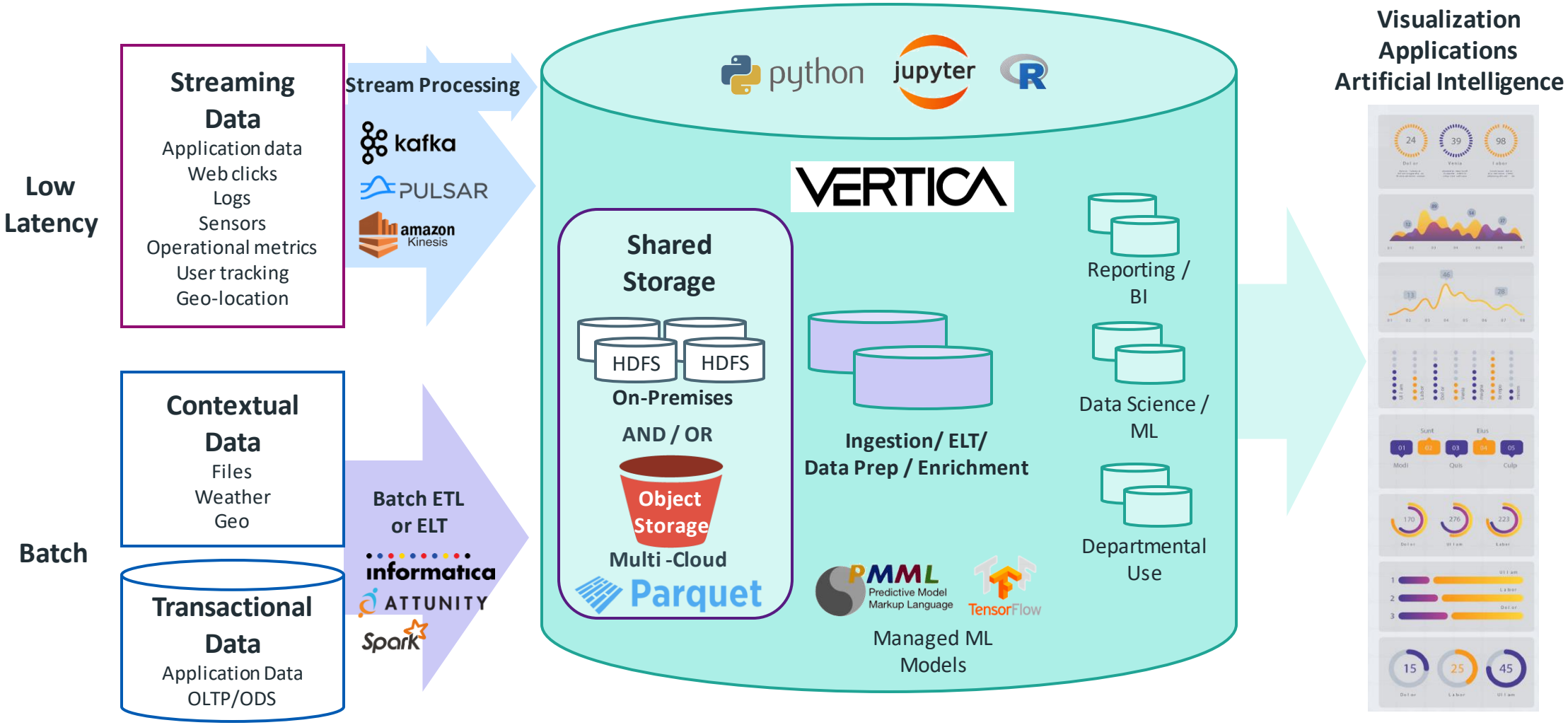
Data Science Lab

Possibility to follow the entire Data Science cycle without moving the data

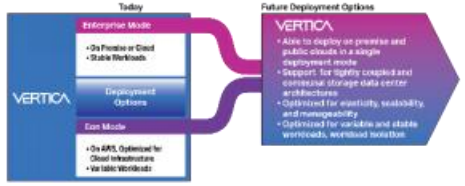
https://www.vertica.com/python/
https://github.com/vertica/VerticaPy

Vertica统一分析仓库

融合了仓库和数据湖，可采用计算与存储分离或无共享架构，并提供端到端的库内机器学习

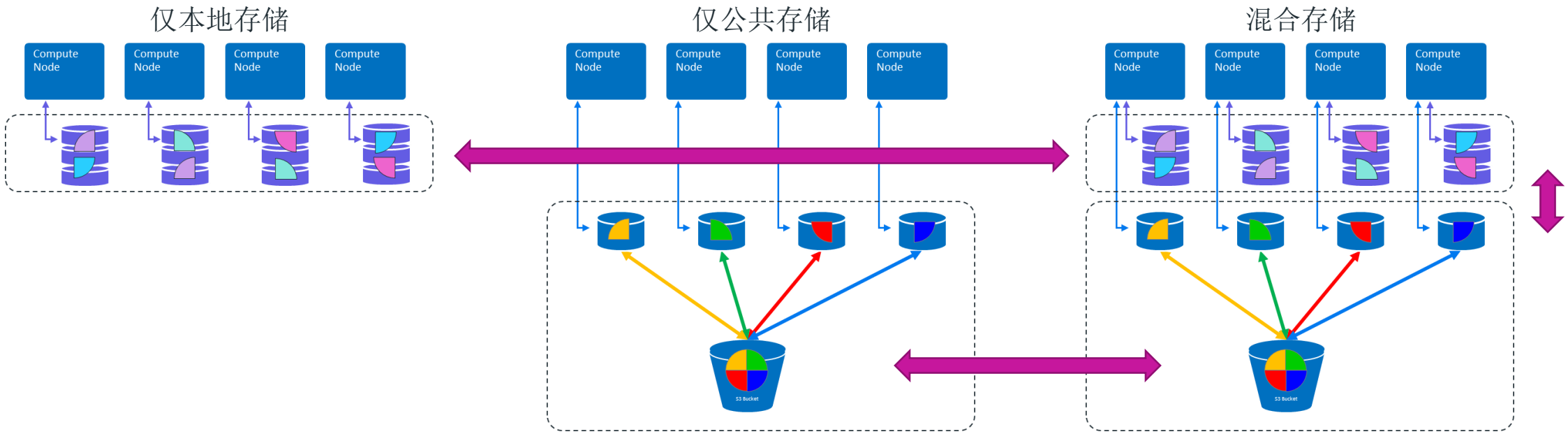


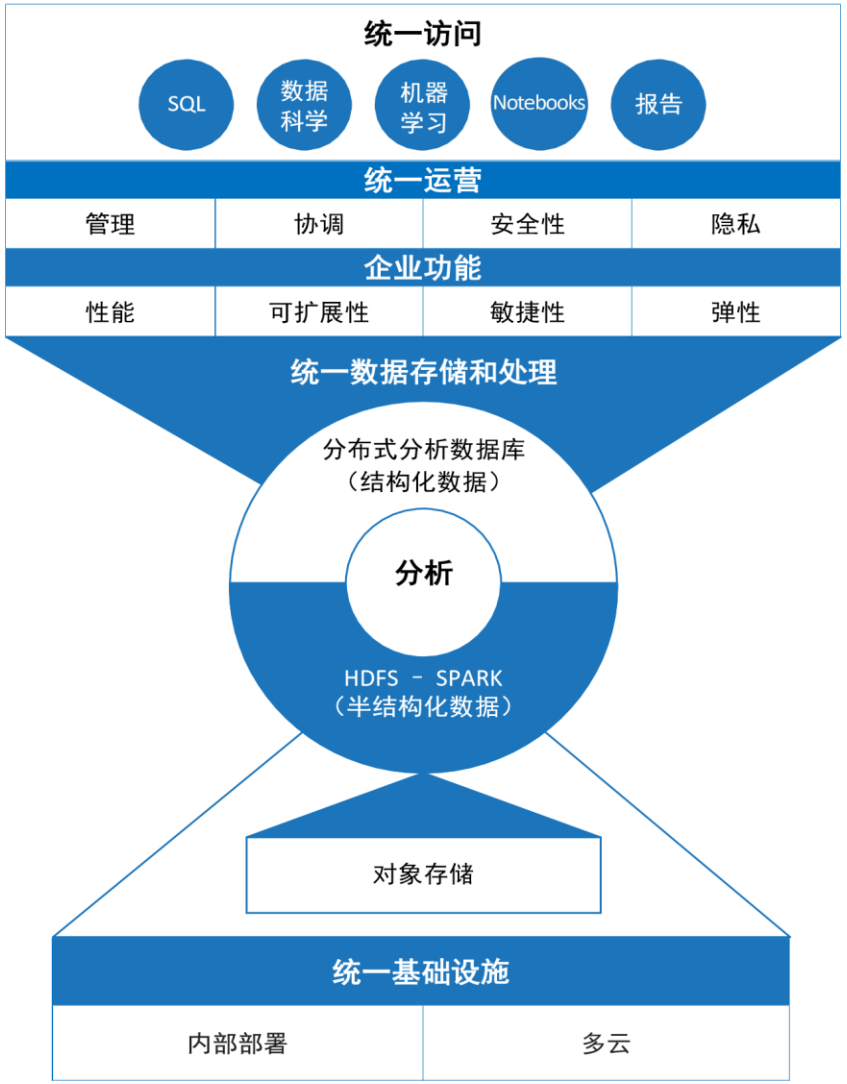
Eon模式和企业模式正在逐步融合



- **自由地将数据存**储在您想要的地方， 本地或公共存储
- **在存**储位置之间自由移动数据
- **用于混合和多云部署的更多子集群**选项
- **流**畅灵活的分片/**分段**选项

【未来】





超大规模数据，极速分析性能

基于标准SQL分析性能，对超大规模数据进行极速分析。

就地分析

无需数据移动可就地分析对象存储数据湖中开放格式数据，数据仓库和数据湖融为一体。

高级分析和预测性分析

时间序列、模式匹配、地理位置以及端到端的库内机器学习。

计算存储分离的云原生架构

冷热数据自动区分，热数据缓存计算节点保证性能，完整数据存储在对象存储确保持久化、多场景复用、弹性扩展和更低的总体成本。

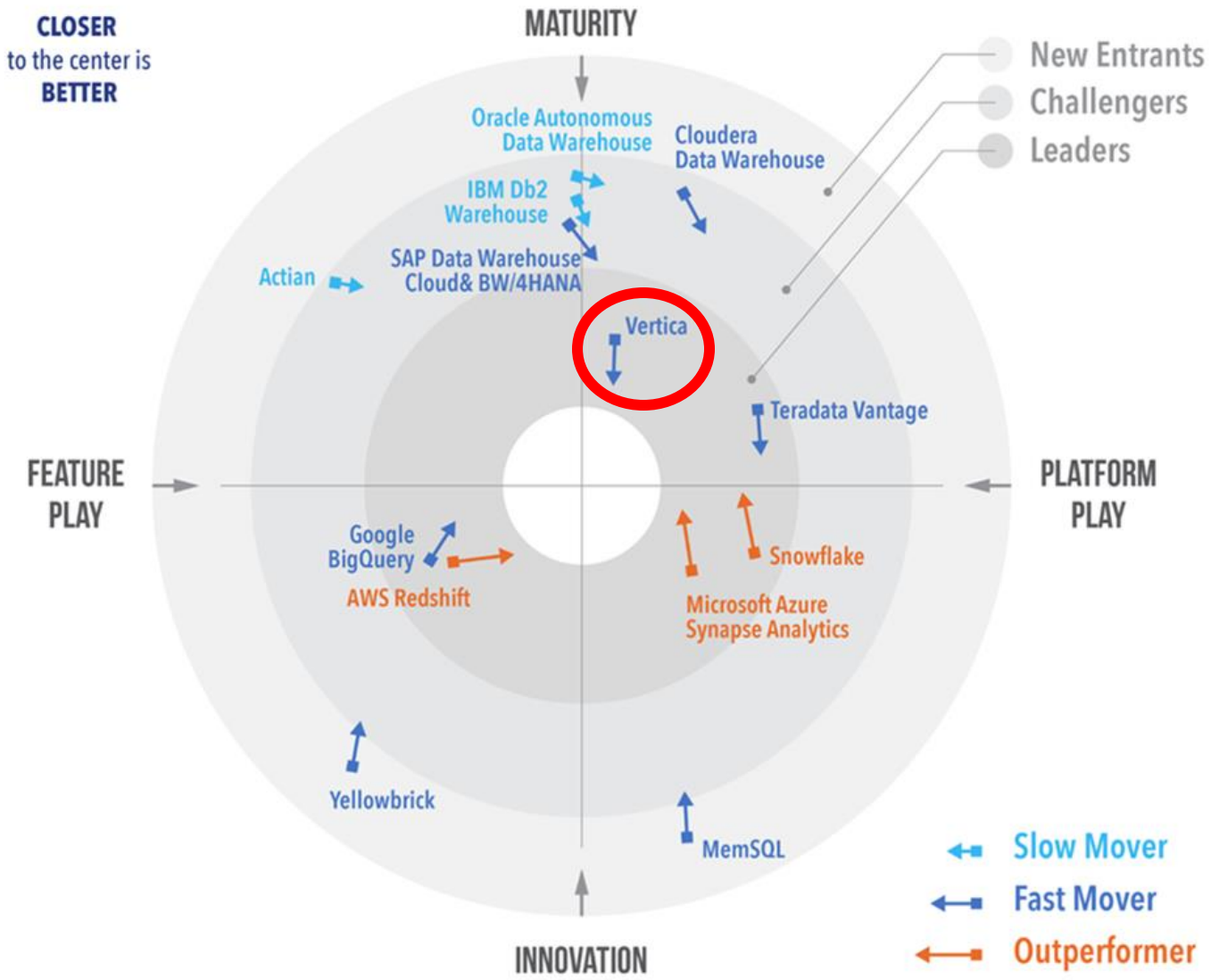
可企业内部和云端融合部署

既可部署在云端，也可以部署在企业内部。

Vertica在第三方数据仓报告中的领先位置

GIGAOM

"Vertica's platform offers strong performance; a wide variety of built-in analytics functions; excellent integration with data science workloads and external data lakes; and notable flexibility in its deployment options."



Source: GigaOm 2020

©GigaOm

VERTICA https://gigaom.com/report/gigaom-radar-for-evaluating-data-warehouse-platforms/

Vertica 助力全球顶级数据驱动企业

Vertica 为无数应用程序和服务提供支持，时刻推动数据驱动型世界的运转

TRANE	Cerner	Uber	GUESS	MTS	OPTIMAL
智能建筑	医疗保健/ EMR 分析	网约车	客户分析	网络优化	预测维护
AT	SUUNTO	THE CLIMATE CORPORATION	intuit	wayfair	CYBERBIT PROTECTING A NEW DIMENSION
路线优化	穿戴分析	智能农业	软件优化	点击流分析	安全性分析

Vertica行业应用场景– 概览

将分析**纳入竞争要素**的各行业示例

策略	医疗保健	零售业	电信业	物联网	营销/广告投放	IT 基础设施	金融服务
I. 客户体验管理	加速交付服务 降低费用 人口健康方案	增强购物体验	增加订购者	预测维护	广告定位	应用程序性能	投资管理和建模
II. 运维分析	医院效率	定价和库存管理	基础设施优化	供应链效率	根据目标市场营销提升客户利益	动态资源管理	实时比较分析
III. 保障和欺诈检测	医疗保健欺诈检测 付款人保护 提供商不法行为	信用卡欺诈损失预防	语音、视频和数据可靠性 SIM 卡欺诈	智能测量工具管理	垃圾邮件防御 广告投放	流量优化和授权访问控制	授权访问和欺诈检测

Vertica支撑了全球10大电信运营商中的7家

现代化的数据仓库

elisa Telia telenor

中国移动 China Mobile

中国电信 CHINA TELECOM

YOTA МегаФон MTS

kt SK telecom

Sprint verizon KDDI

多用途分析(含数据仓库)

AT&T UNITED GROUP

中国电信 CHINA TELECOM

欺诈管理、合规

YOTA МегаФон maxcom

BİLGİ TEKNOLOJİLERİ VE İLETİŞİM KURUMU

OEM（探针、DPI、网优等）

ERICSSON NETSCOUT Anritsu

amdocs VIAVI EMPIRIX

RADCOM

sysmech sandvine Intelligent Broadband Networks YUVO

VERTICA

IOT 场景

NOKIA

高级IT运营

OTE vodafone

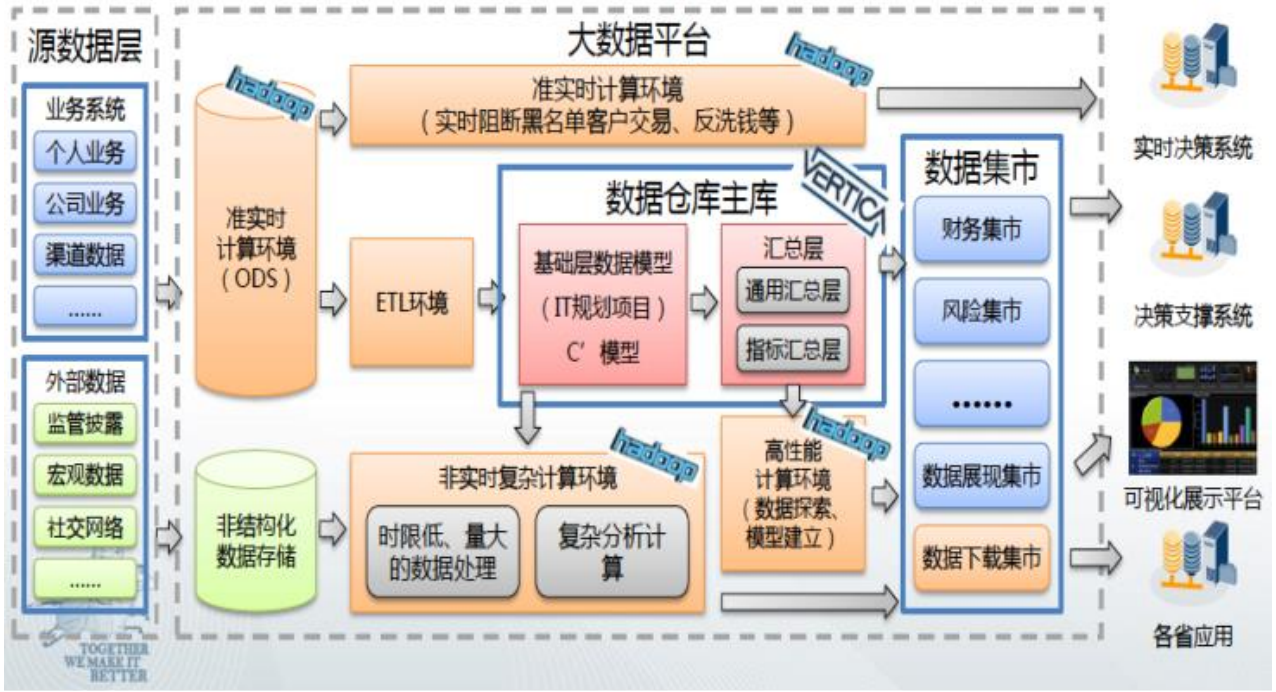
某国有大银行数据仓库

背景与业务需求：

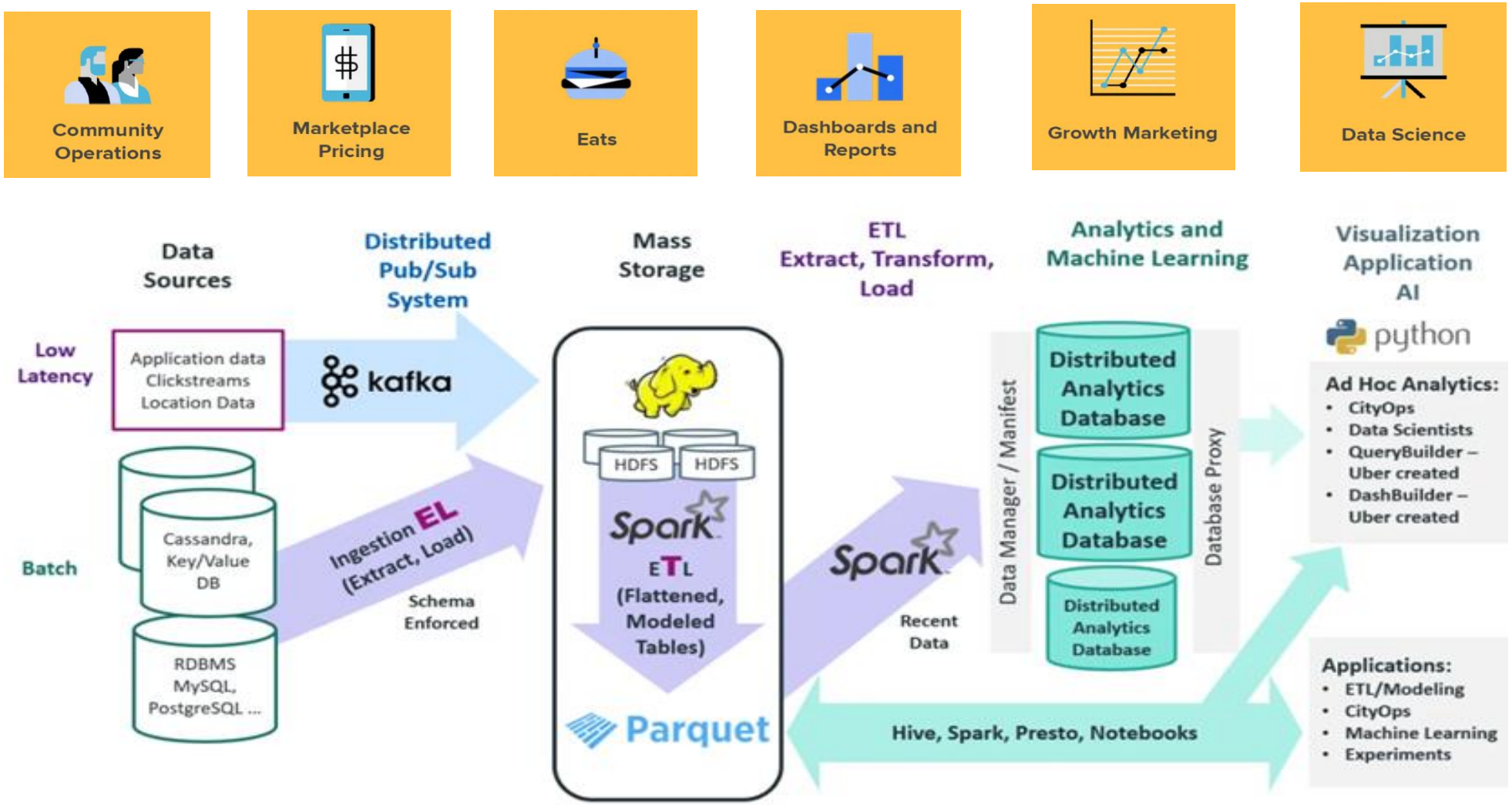
- 数据仓库和集市，原有平台（Teradata）高昂的升级和运维成本阻止了进一步扩容，无法满足业务增长的要求
- 大数据平台架构包含MPP DB、Hadoop、R、streaming等框架，需要更开放、灵活的数据库仓库基础平台

Vertica作为数据库仓库平台的价值:

- 列式计算带来的高性能，比原平台性能提升x10
- 高扩展能力，120节点2PB数据仓库
- 与Hadoop无缝集成，快速数据探索、集成和分享
- 无缝集成分布式R，支持模型快速演进



在Uber，数据参与每一个决策

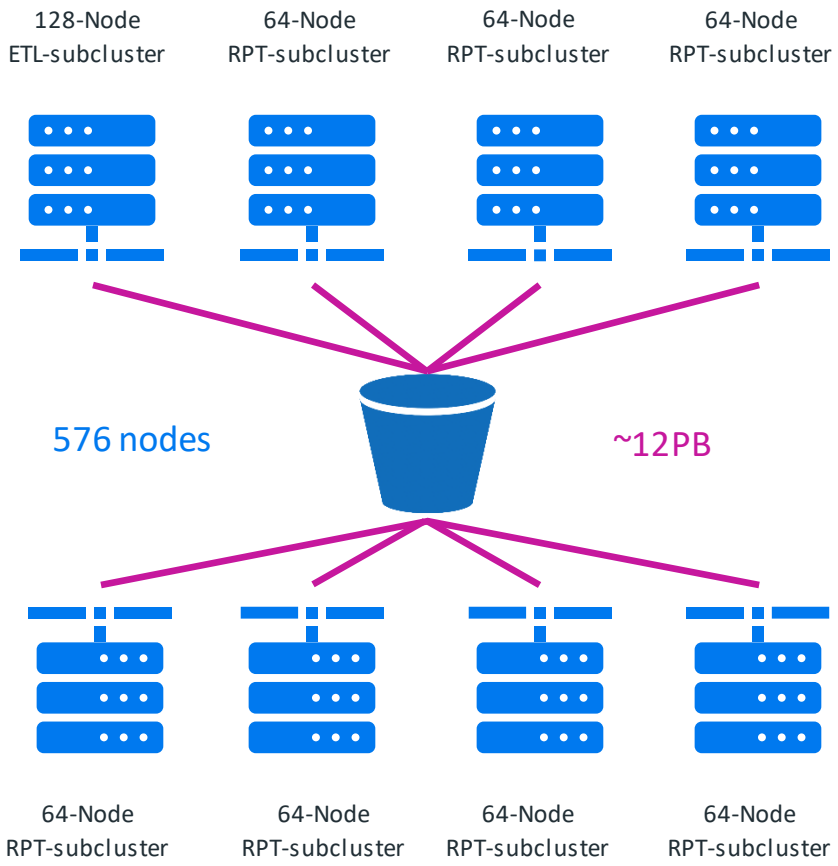


http://bigdatapage.com/can-presto-sql-on-hadoop-replace-your-data-warehouse/

UBER

提供数字广告行业成功所需的可伸缩性、稳定性和性能

按需及时扩展和收缩，高效的云经济模式



- 低成本云端部署，易于容量规划
- Vertica简化管理、自助服务能力
- Vertica每天处理40,000 报表
- 近12PBs数据
- 576节点按需伸缩

“我们支持实时出价领域的买家，帮助代理商和品牌以更有针对性和更明智的方式花费广告预算。但是支出只是挑战的一部分，透明度也是关键。我们不只是说我们在明智地支出他们的预算；我们向他们展示了从最初的印象到转化。 这就是我们为什么用Vertica。”

theTradeDesk

VERTICA

MICRO FOCUS

VERTICA

扫描二维码关注官方微信

谢谢！

MICRO FOCUS