具有机器学习功能的MySQL HeatWave云服务

公益讲座11：00准时开始，请大家先浏览云技术微信公众号技术文章资料会在各群同步发布，已入群客户请勿重复入群！

20-19

数据库和云讲座群

甲骨文云技术公众号

具有机器学习功能的MySQL HeatWave

徐轶韬

Principal Solution Engineer
MySQL Global Business Unit

为何在云中使用MySQL？

安全性
保持数据库的安全性变得更为困难

快速部署与自动化
有助于改善系统管理与合规性

基础服务外包
有助于减少运营开销

节省成本
节省硬件成本和软件许可证成本

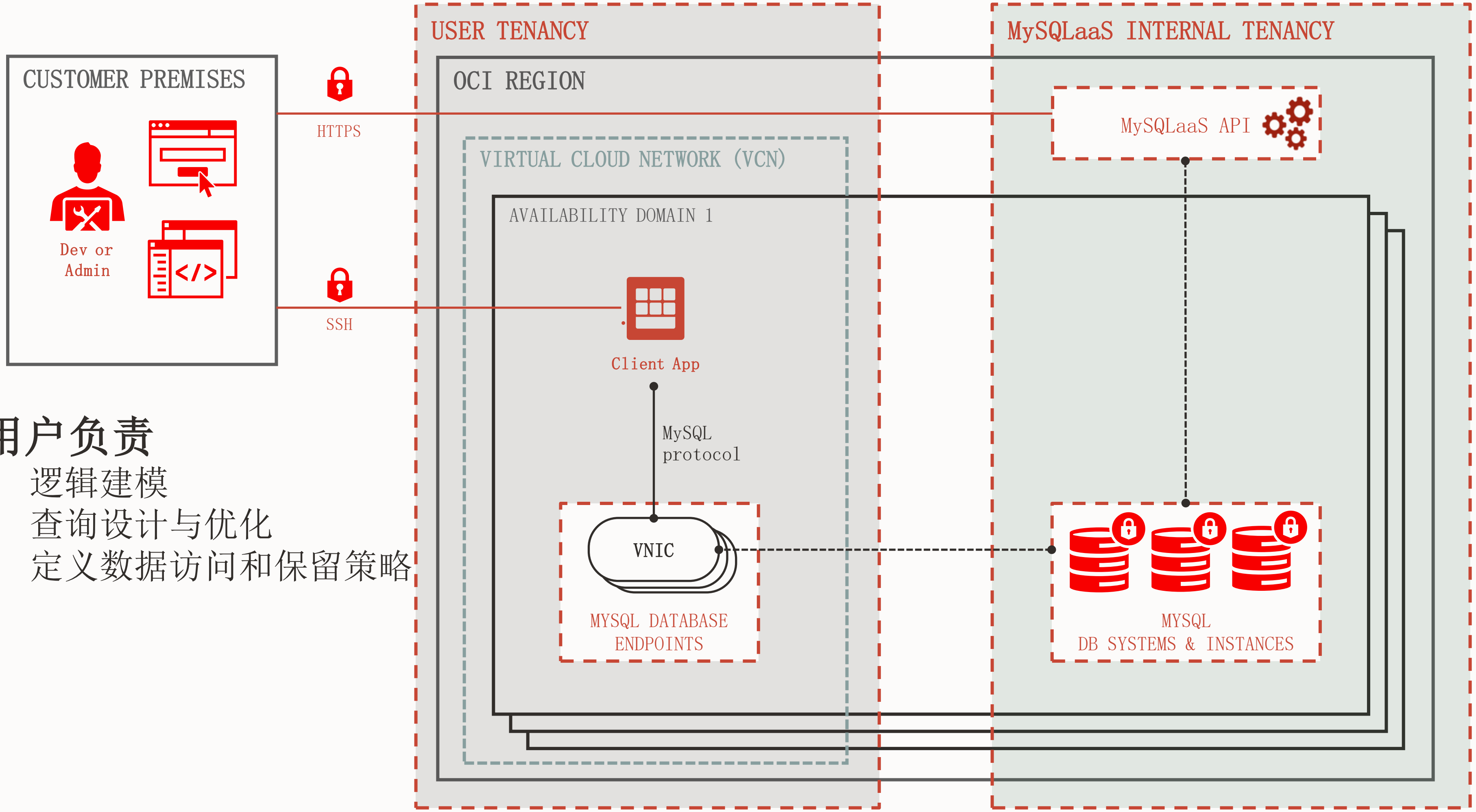
为何选择 MySQL Database Service (MDS)

使用MySQL企业版

1. 减少数据泄露风险
2. 满足合规要求
3. 获取最新的安全更新
4. 与本地部署的MySQL100%兼容
5. 来自MySQL专家的技术支持

更能节省成本

MDS使用MySQL企业版提供云服务



Oracle 负责

- 备份与恢复
- 具有自动故障转移功能的高可用性
- 水平与垂直扩展
- 数据库和操作系统补丁
- 监视和日志处理
- MySQL企业版中的高级功能

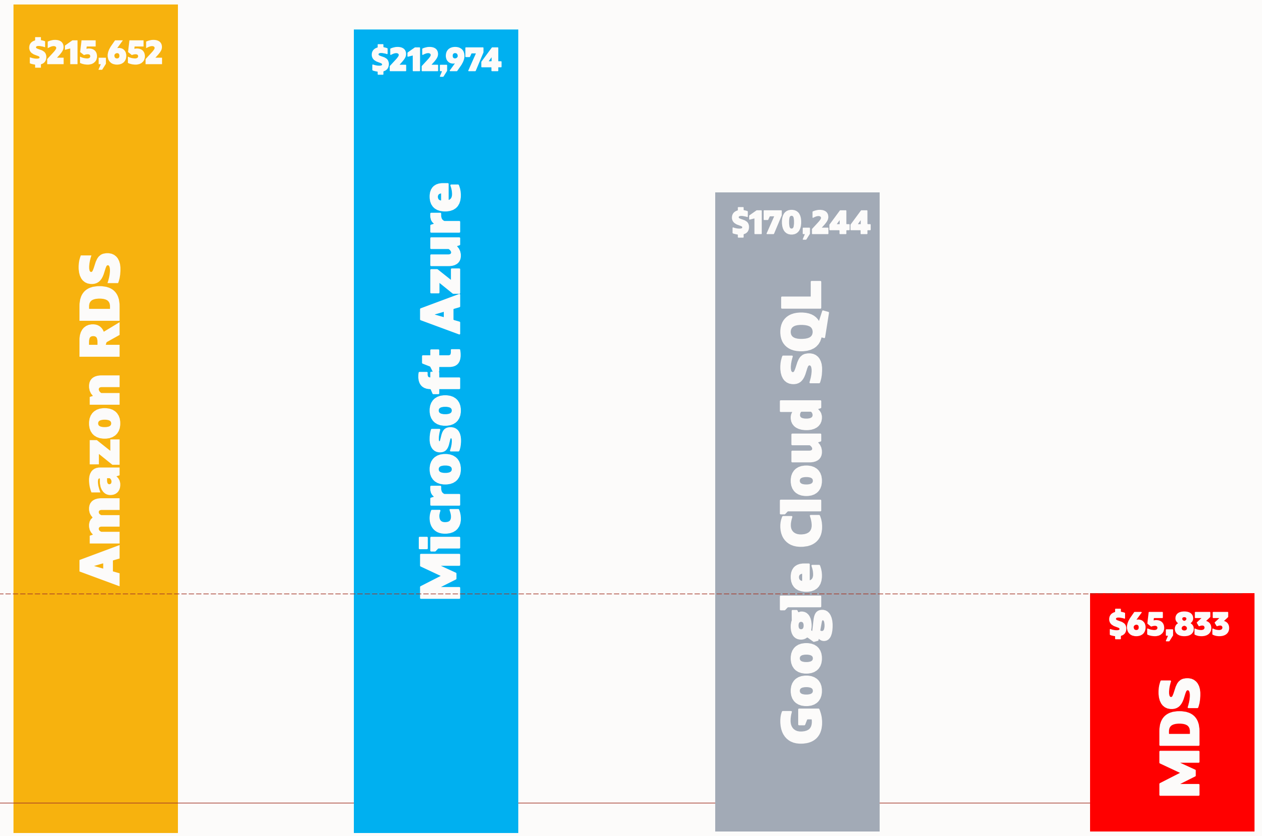
用户负责

- 逻辑建模
- 查询设计与优化
- 定义数据访问和保留策略

70% 与 Amazon RDS相比节省成本

MySQL Database Service: 1 Year TCO
100 OCPUs, 1TB Storage

MySQL Database Service: Standard E3 AMD 16GB/Core, all regions have the same price.
Amazon RDS: Intel R5 16GB/Core, AWS US East.
Azure: Memory Optimized Intel 20GB/Core, MS Azure US-East.
Google: High Memory N1 Standard Intel 13GB/Core, GCP Northern Virginia.
Configuration: 100 OCPUs, 1 TB Storage.



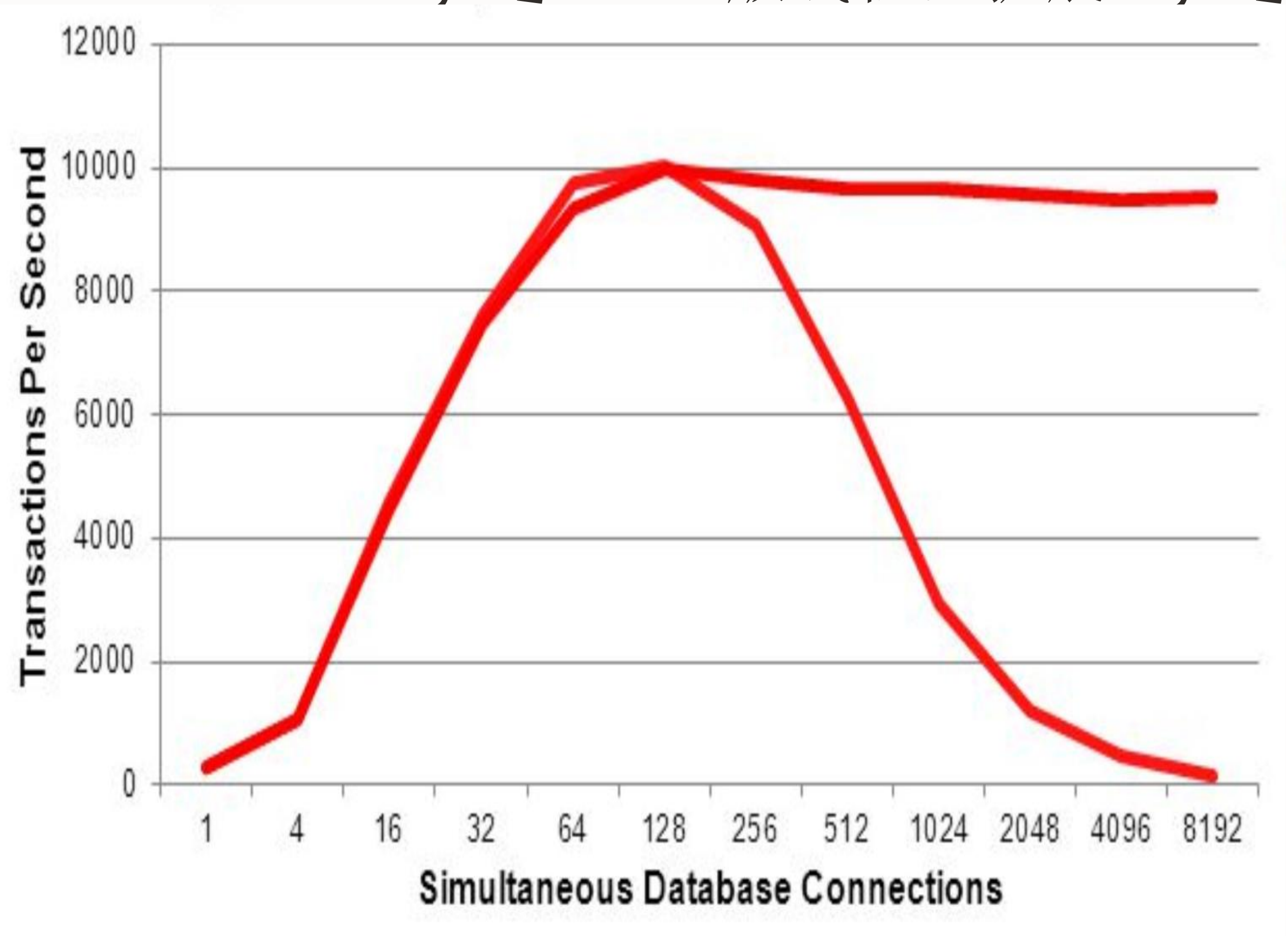
±30% Cost

inbound traffic is FREE and outbound traffic is FREE for the first 10 TB per month

MDS性能稳定

20x

MySQL 企业版线程池扩展MySQL



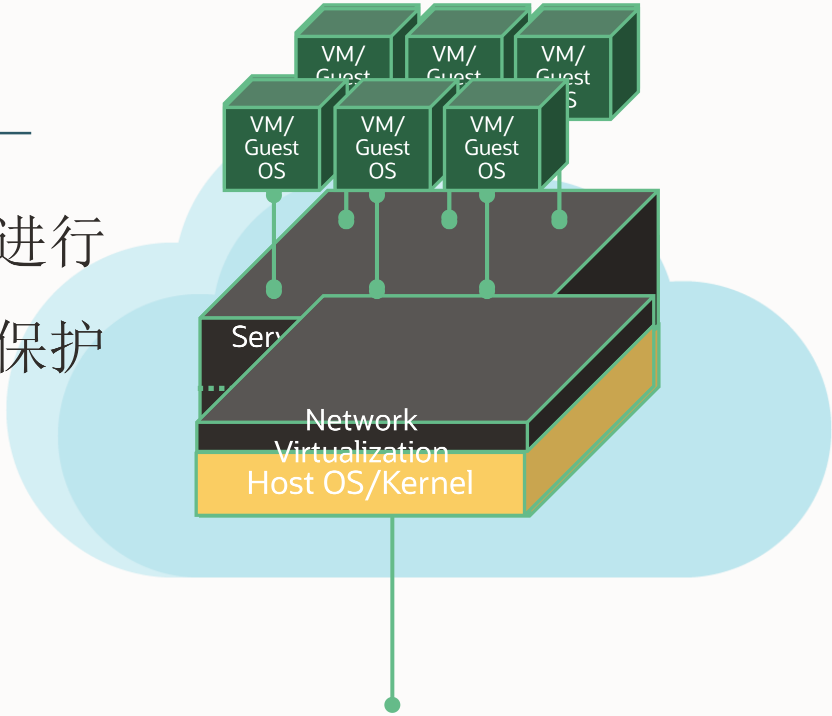
MySQL Database Service (MDS)
使用MySQL企业版线程池

MySQL 社区版不具有线程池功能

安全性

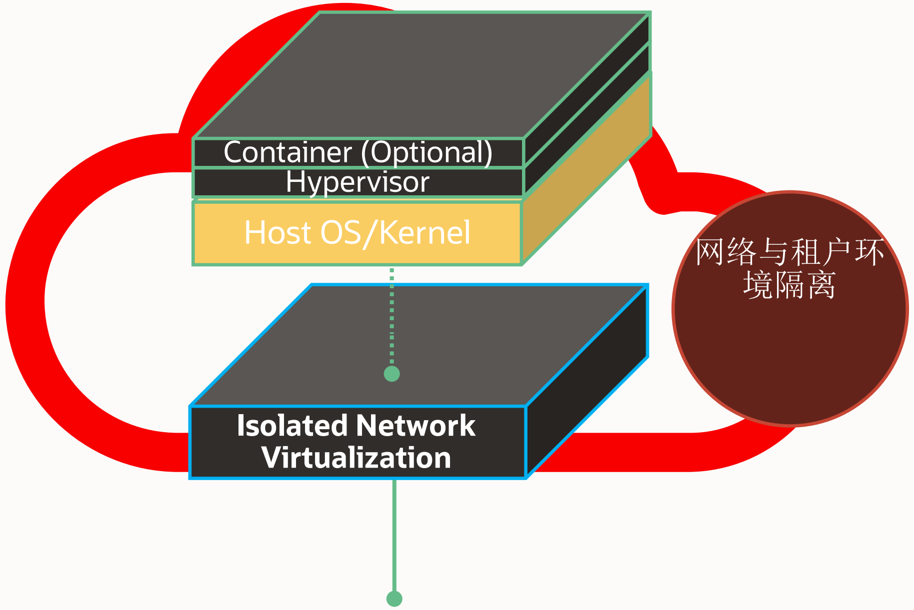
架构设计为对租户进行最大程度的隔离和保护

一代云:
最普遍的做法



租户之间

二代云:
Oracle Cloud Infrastructure-Wide



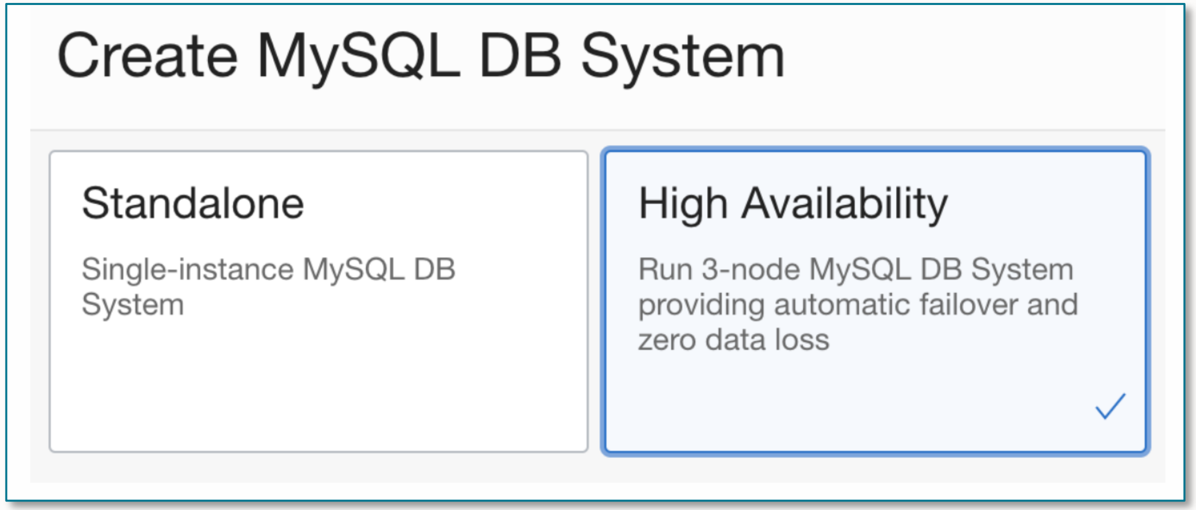
租户之间

MySQL Database Service: 高可用性

具有自动故障转移和零数据丢失的容错系统

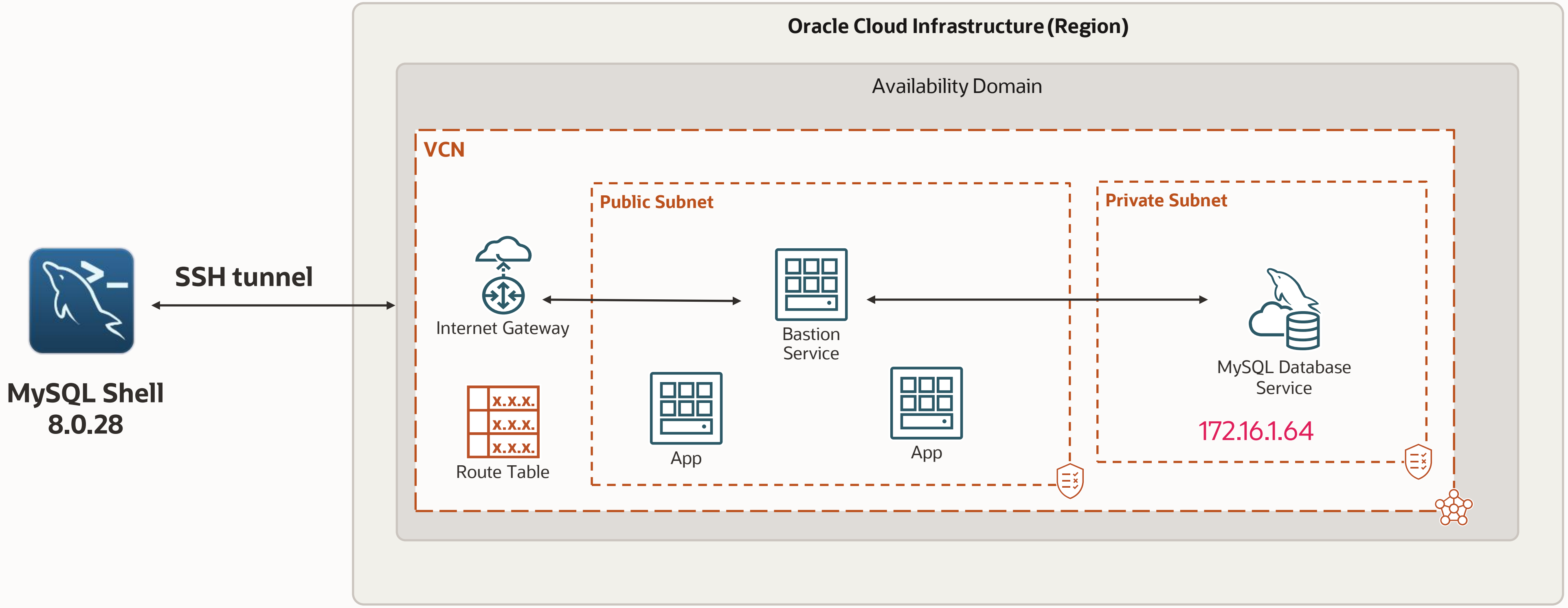
- 一键高可用性
- 自动故障转移
- 提高运行时间
- 数据零丢失

- Recover Time Objective (RTO): 分
- Recovery Point Objective (RPO): 0



通过堡垒主机连接

使用 MySQL Shell 8.0.28

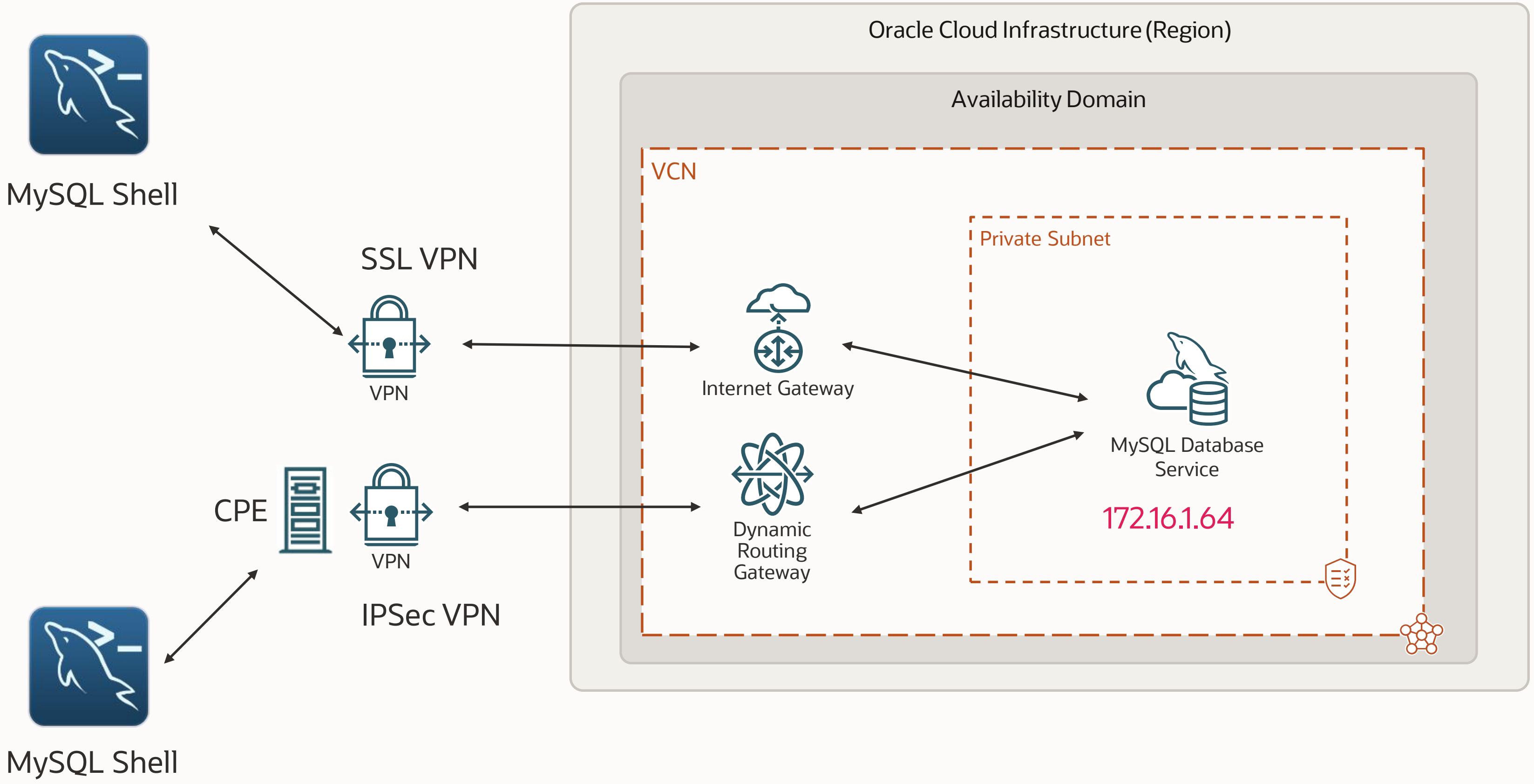


e.g. mysqlsh root@172.16.1.64:3306 --password=Welcome1! --ssh=opc@140.83.38.174 --ssh-identity-file=~/.ssh/id_rsa

通过VPN (SSL, IPsec)连接 MDS

使用 MySQL Shell 8.0.28

e.g. mysqlsh root@172.16.1.64:3306 --password=Welcome1!

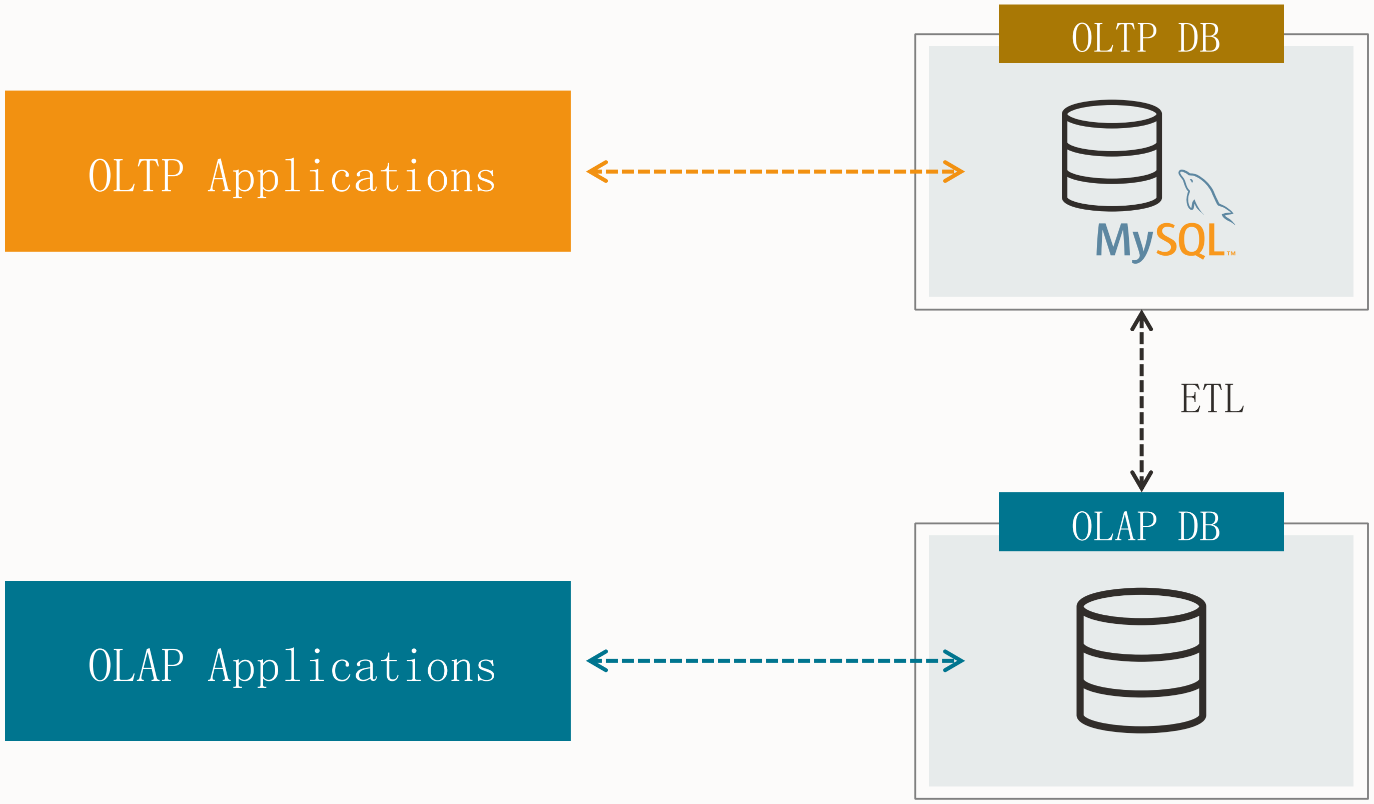


e.g. mysqlsh root@172.16.1.64:3306 --password=Welcome1!

挑战#1: 事务处理与分析处理的系统需要分开

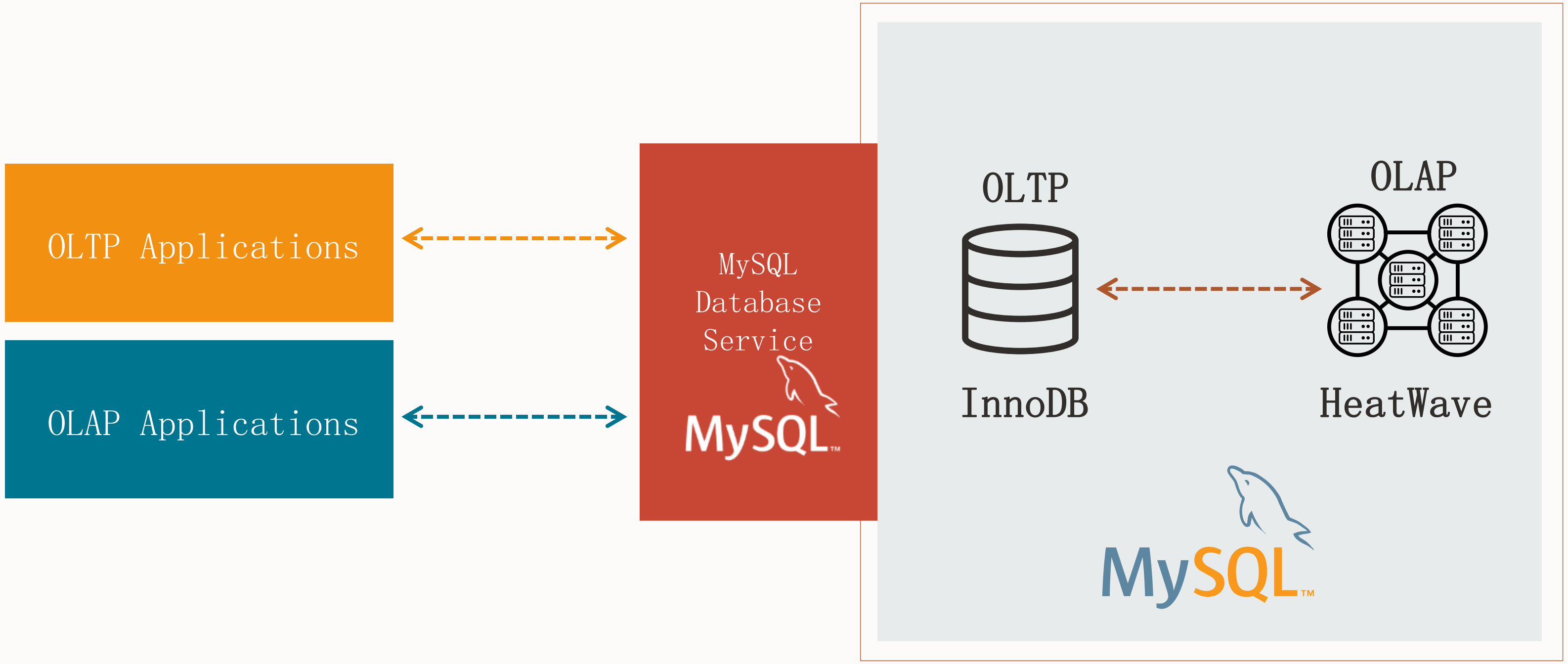
MySQL 用户需要分别使用OLTP 和OLAP系统

传统的行格式没有针对OLAP工作负载进行优化



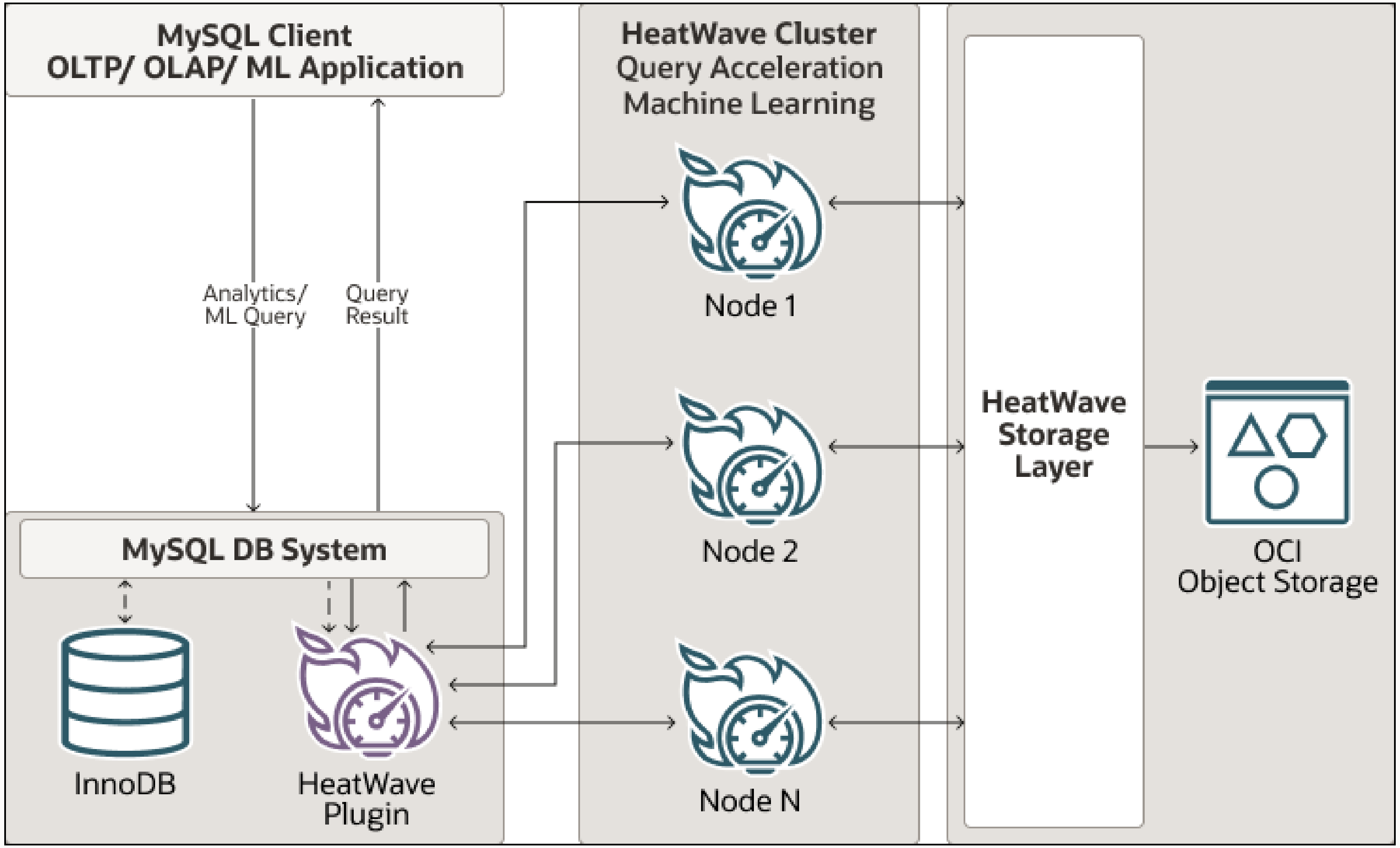
Oracle云中提供的MySQL HeatWave 数据库服务

无需重写查询和ETL、加速OLTP和OLAP工作



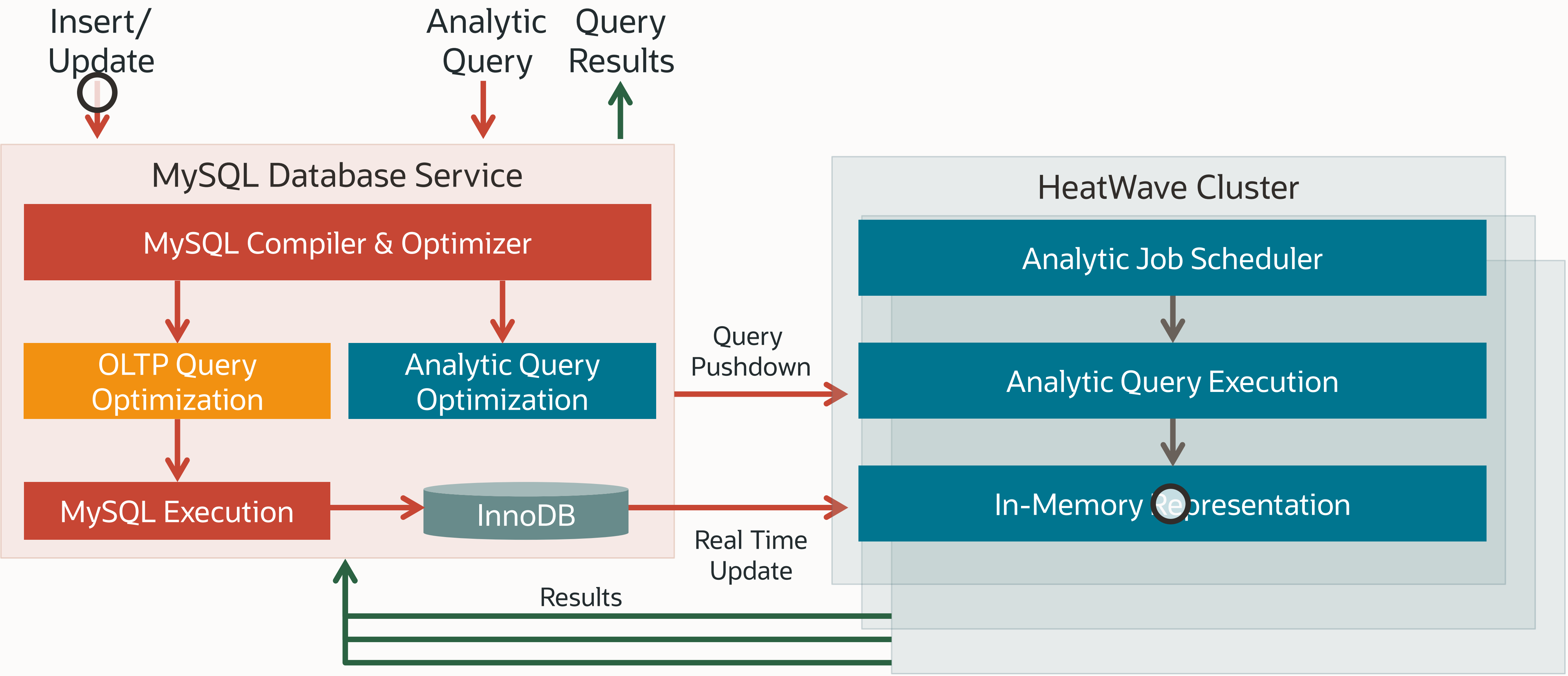
1. **一个MySQL**数据库，用于OLTP和OLAP应用程序
2. 数据分区存储在**内存**
3. 已有的应用程序**无需任何更改**

MySQL HeatWave 数据库服务架构



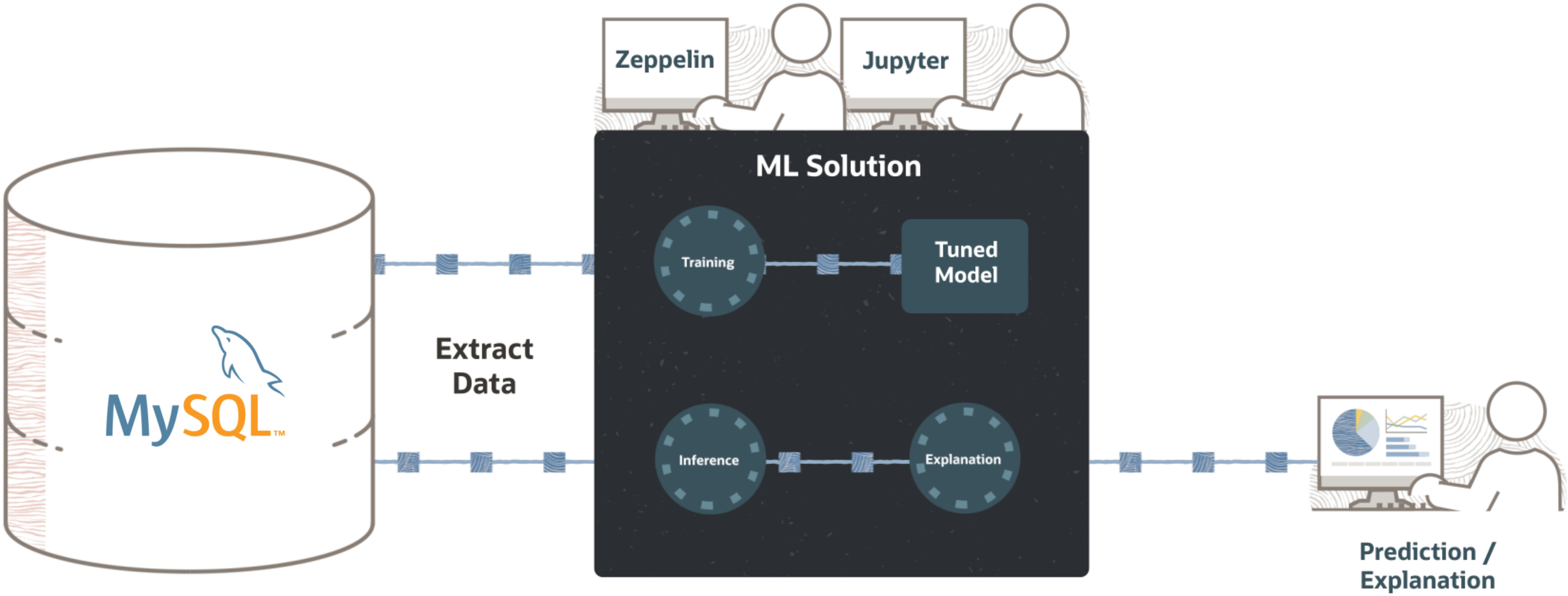
HeatWave: 实时分析

插入MySQL的数据可以在一秒内进行分析查询



挑战 #2: 用于机器学习的工具和服务需要分开

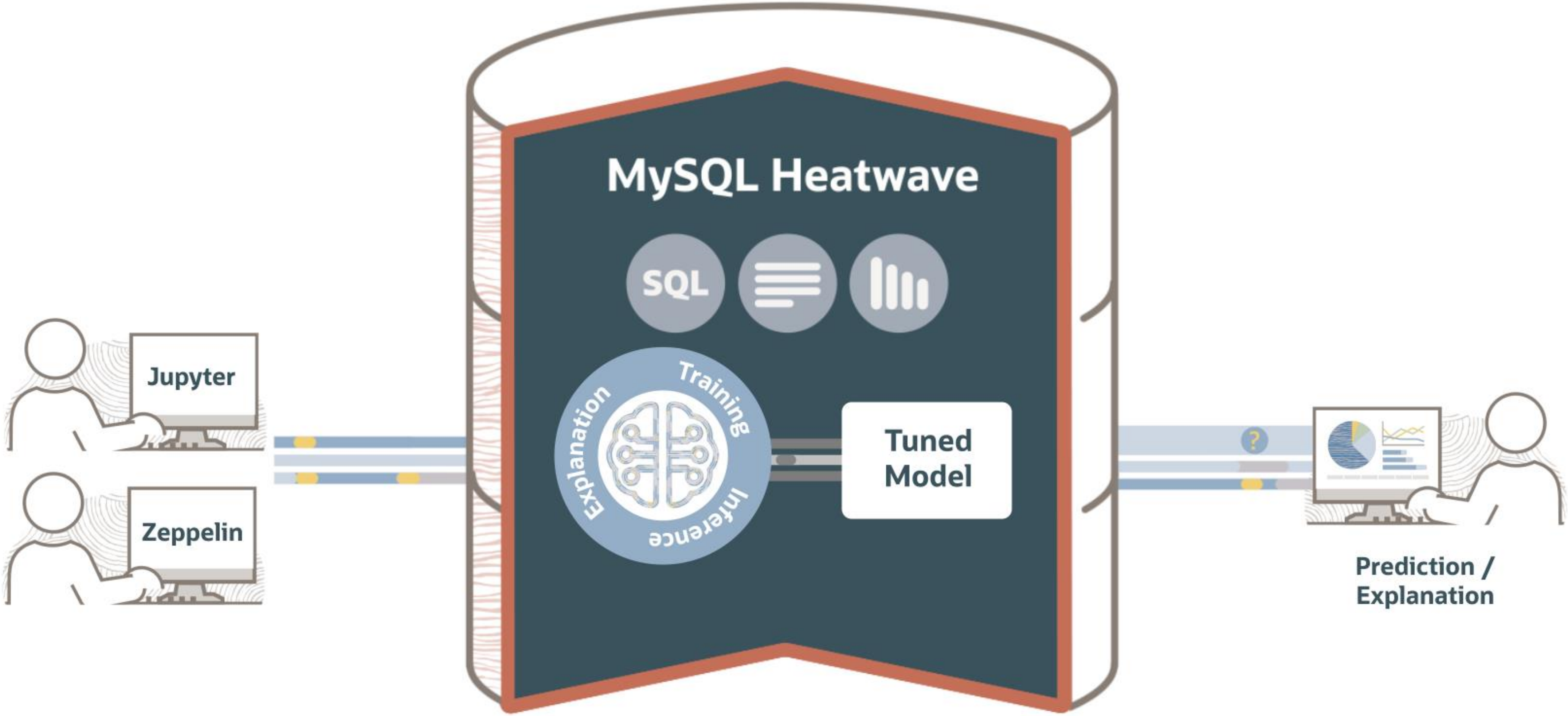
需要使用ETL分离数据进行训练与预测



- 复杂,耗时
- 增加成本和风险
- 需要学习新的工具/语言

当使用其他数据库时，情况会更糟……

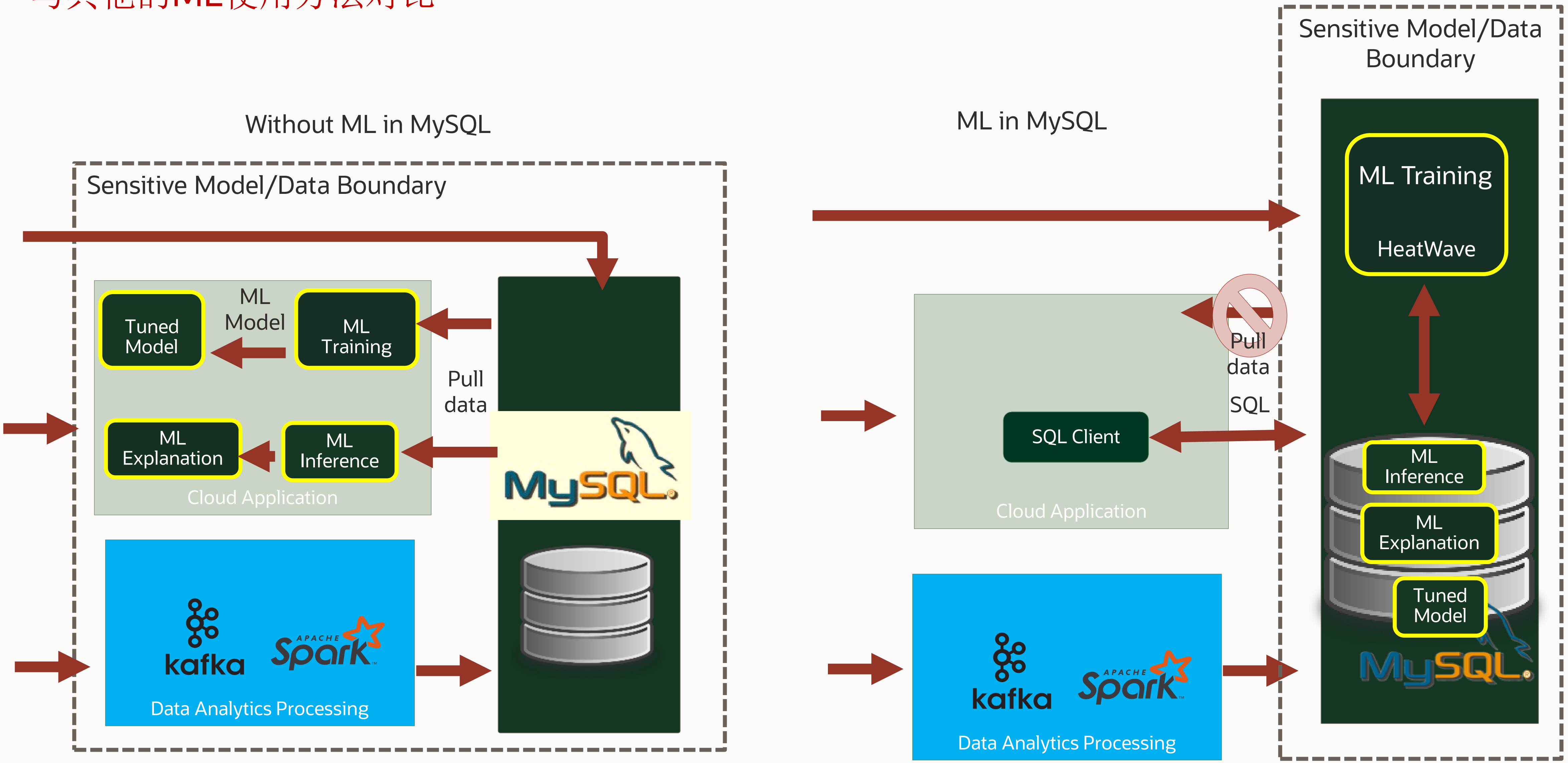
使用MySQL HeatWave的机器学习



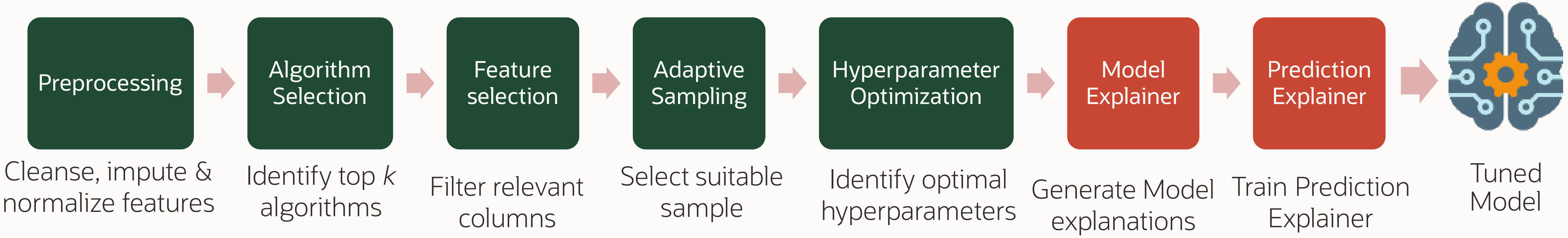
加速ML计划，增加安全性，降低成本

Heatwave ML

与其他的ML使用方法对比



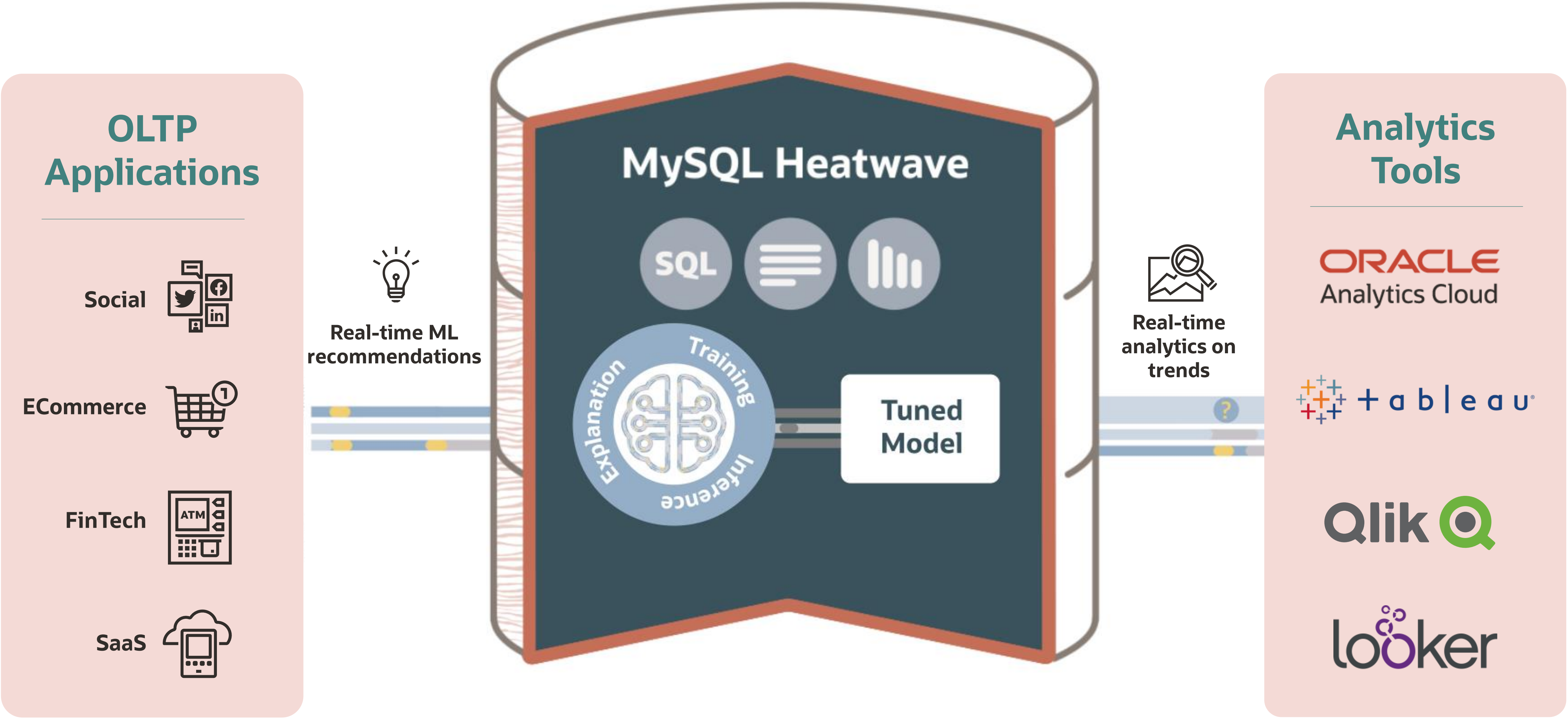
Oracle AutoML
自动生成一个准确和快速的模型



AutoML有多个自动化处理步骤，用于精确预测

- 非迭代管道设计
  - 利用算法和数据集的元数据学习
- 智能选择，减少每个阶段的尝试
  - 根据输入数据集特征选择算法/参数，准确率更高，运行时间更短
  - 通过使用元数据学习实现
- 可扩展设计
  - 所有阶段都利用了节点间和节点内的并行性，减少了运行时长

Machine learning in action with MySQL HeatWave



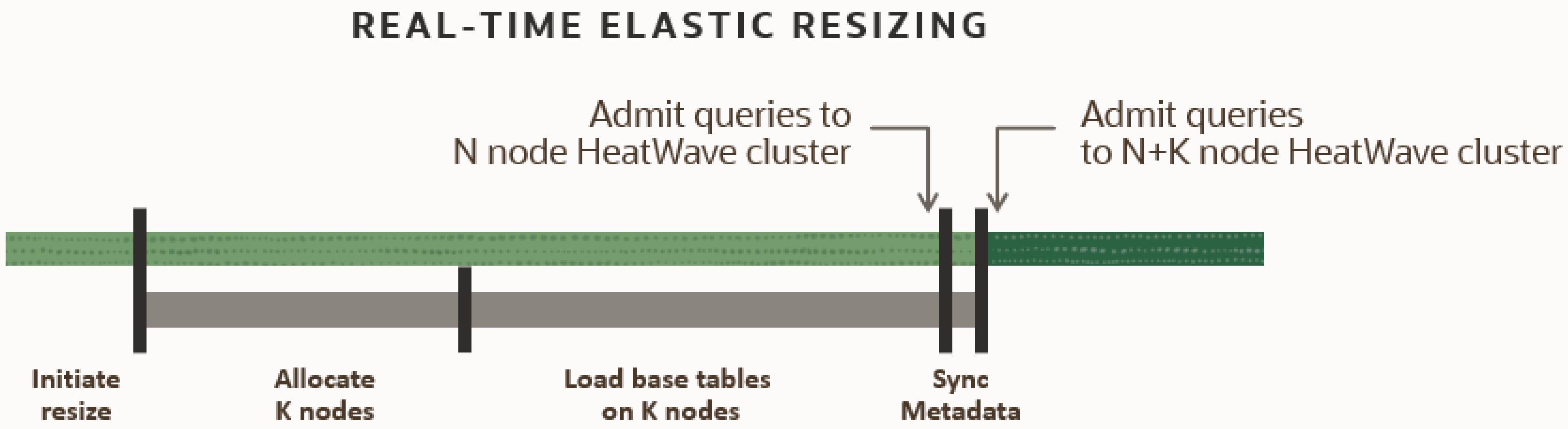
HeatWave ML vs Redshift ML: 基准测试

- ✓ 产生更准确的结果
- ✓ 训练模型平均速度快25倍
- ✓ 1% 的成本
- ✓ 添加更多模型时可以扩展

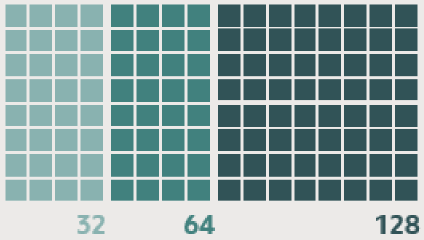
基准测试详细内容: https://www.oracle.com/mysql/heatwave/performance/

零宕机时间OLTP, OLAP, 及 ML

通过增加或减少任意数量的节点、调整HeatWave集群的大小



Snowflake
Only scale to
their sizes



AWS Redshift
10 –15 minute query
blackout and manual
balancing required

挑战 #3: 人工管理工作耗时严重

人工管理任务消耗资源

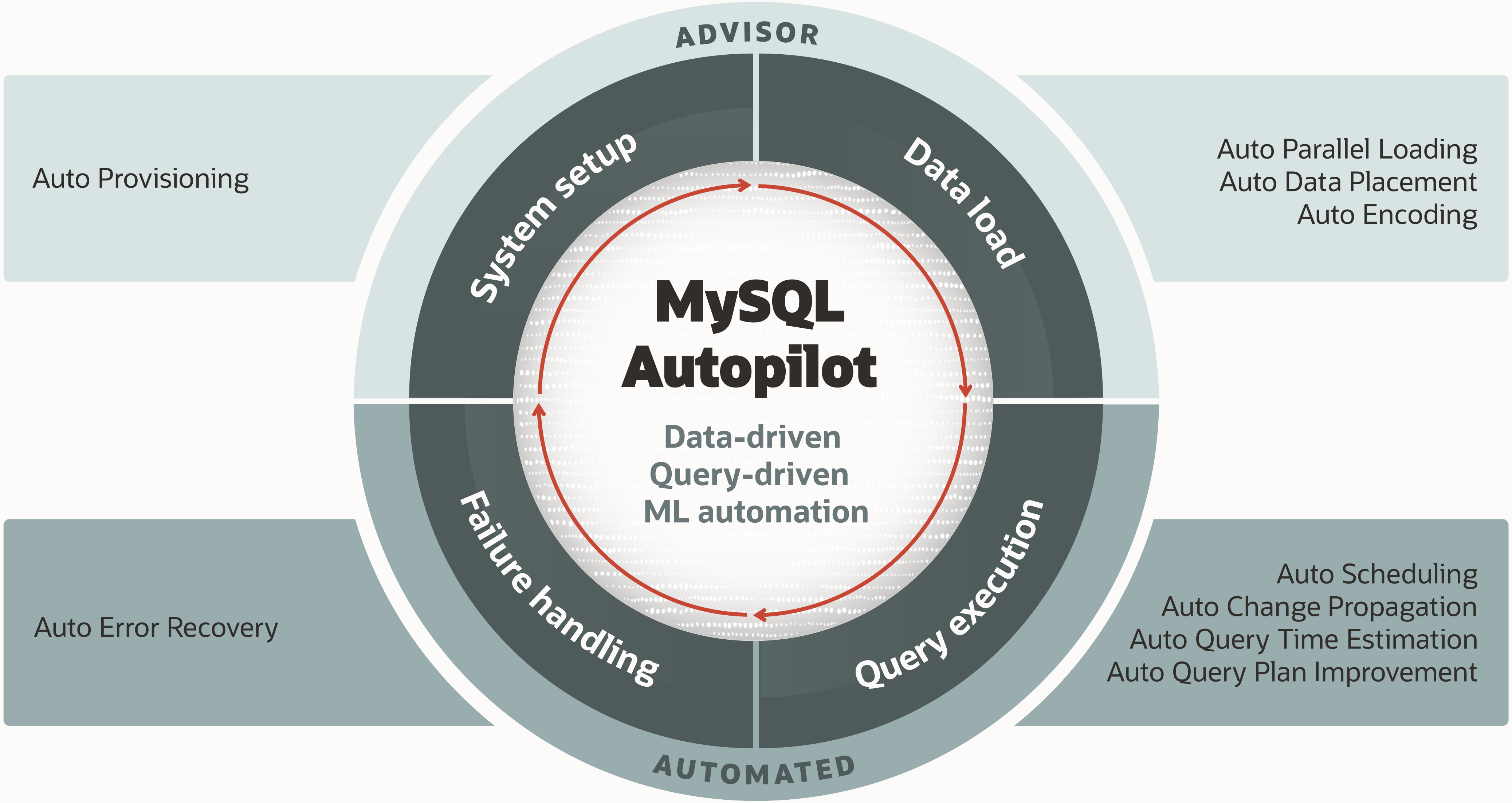
- **本地部署：**
  - 数据库管理：配置、备份、HA、补丁、安全等
  - 操作系统管理：安装、打补丁、升级……
  - 基础设施管理：服务器的购买和维护，存储
  - 数据中心管理：空间、电源、冷却、灾难恢复等
- **云服务：**
  - 配置：调整数据库的大小
  - 数据加载：优化加载时间，内存使用，编码，数据放置
  - 查询执行：性能调优，查询的优先级排序
  - 失败处理：处理错误恢复的操作

基于Oracle Cloud Infrastructure 的 MySQL Database Service

开发，管理，支持100%由Oracle提供

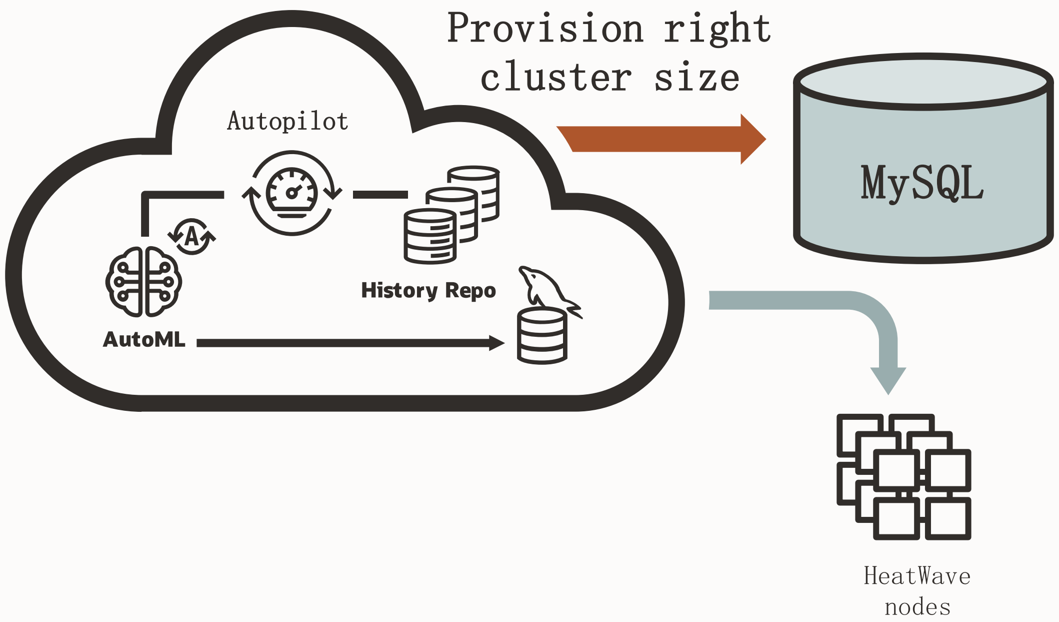
	Automation	MySQL On-Premises	MySQL Database Service
Database	Scaling	✗	✓
	Backup	✗	✓
	Security Patch & Upgrade	✗	✓
	Provision & Configure	✗	✓
OS	OS Security Patch & Upgrade	✗	✓
	OS Installation	✗	✓
Server	Hardware Purchase & Maintenance	✗	✓
Storage	Storage Purchase & Maintenance	✗	✓
Data Center	Rack & Space	✗	✓
	Power, HVAC, Networking	✗	✓

MySQL HeatWave的基于机器学习的安全自动化



自动部署

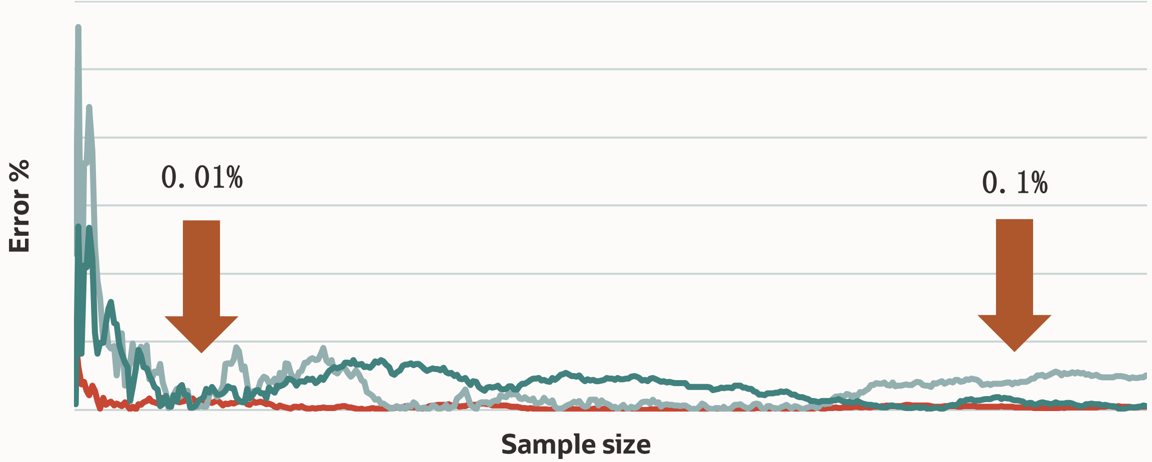
机器学习预测内存使用以估计集群大小



自动部署结果

精度高，效率高，全自动

Adaptive sampling

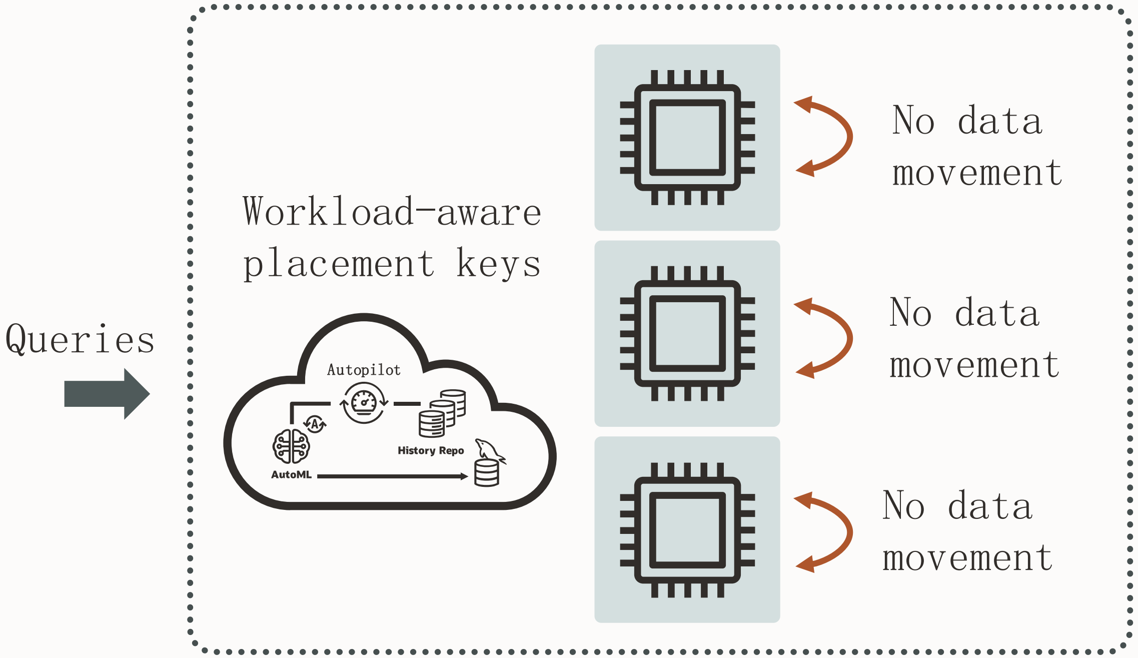


< 0.1% of data scanned for prediction

Datasets	TPCH 1024G	TPCDS 1024G	Cust A	Cust B
Accuracy in memory prediction	98.4%	96.9%	98.3%	96.9%

自动数据放置

- 系统根据查询预测并显示内存中分区数据的最佳列
- 运行时预测改进



自动部署结果

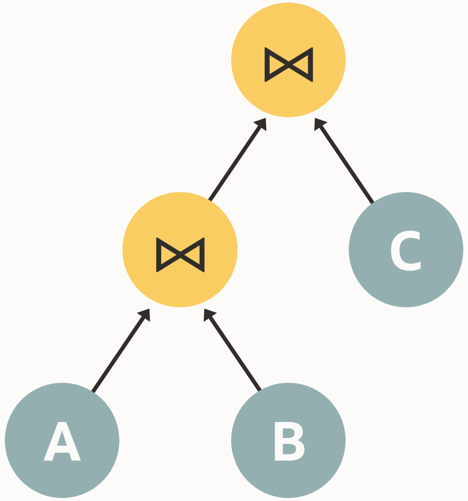
随着模型不断学习，性能提高

Dataset	Time with primary key placement	Predicted improvement	Actual improvement
TPCH 1024	332 sec	26%	**37%**
TPCH 4096	373 sec	20%	**25%**

自动改进查询计划

优化器根据前面执行的查询，学习和改进查询计划

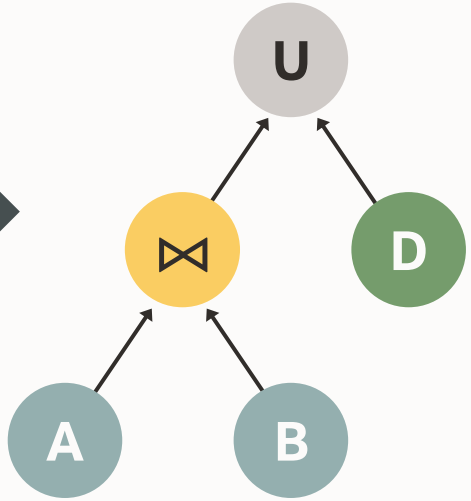
Query #1
A ⋈ B ⋈ C



Node	Statistics
A	70
B	150
A ⋈ B	1000
C	…
A ⋈ B ⋈ C	…

Runtime stats

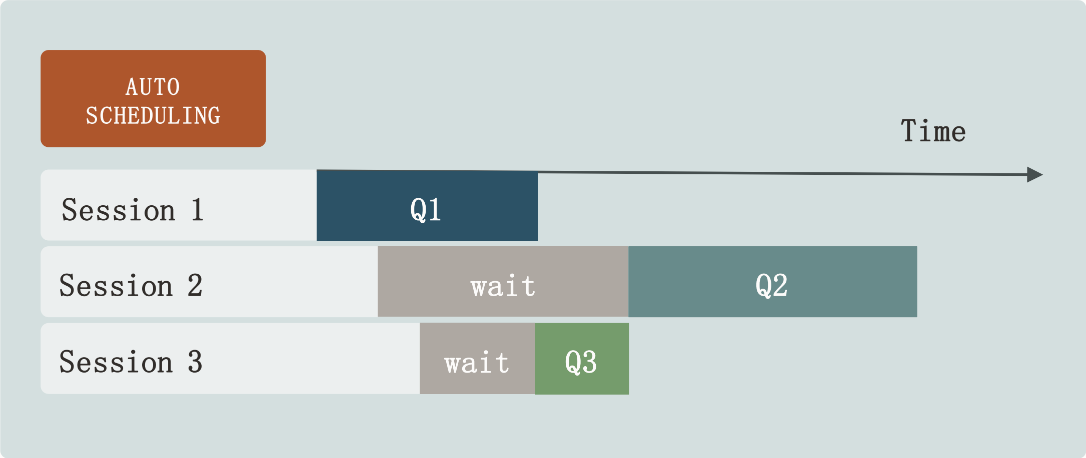
Query #2
A ⋈ B U D



MySQL Autopilot improves TPCH, TPCDS 24TB performance by **40%**

自动调度

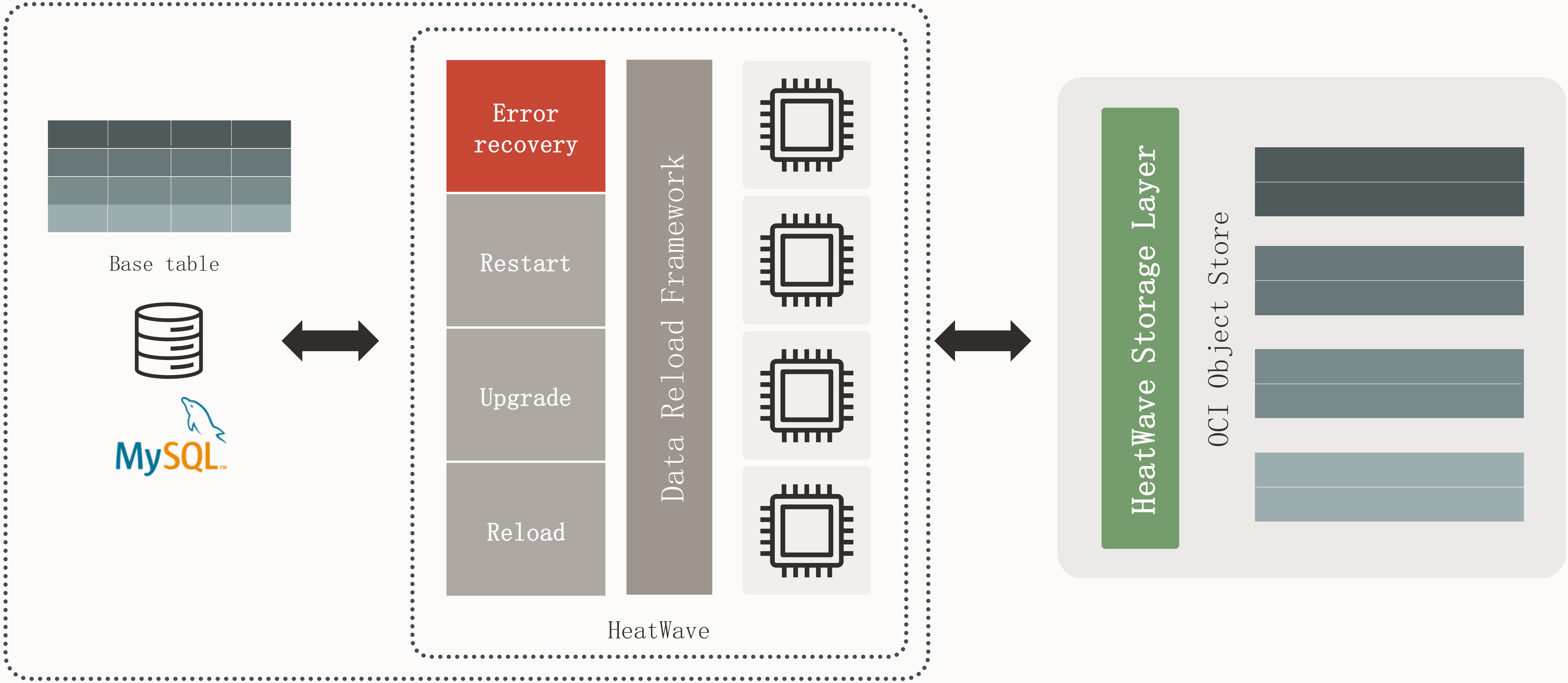
预测执行时间并对较短的查询进行优先级排序，以减少总体等待时间



减少混合工作的等待时间
(OLTP + OLAP)

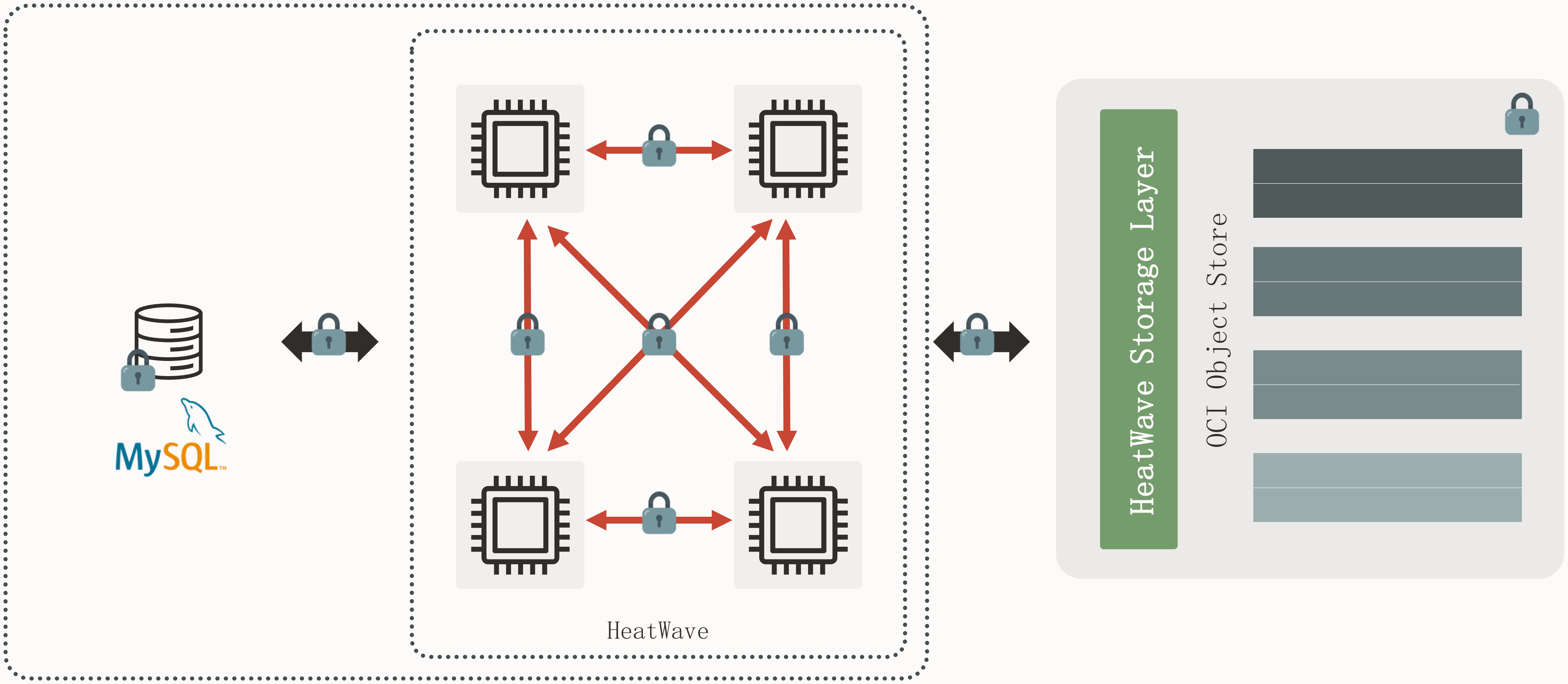
MySQL HeatWave 扩展数据管理

重新加载任意数量的数据的时间固定



MySQL HeatWave的安全强化

数据处于加密状态



高级别安全性

用于实现额外安全措施的内置服务器功能

- 使用密钥生成和数字签名的非对称加密：使用公钥和私钥增加对机密数据的保护，并实现数字签名来确认签署文档的人的身份。
- 数据屏蔽和脱敏:帮助保护私人数据免受外部攻击和恶意员工的攻击。
- 数据库防火墙：防止特定于数据库的攻击，如SQL注入。

比Amazon Aurora, Redshift, 及 Snowflake更为安全。

为分布式云做好准备：可以部署在公有云和用户的数据中心

公有云

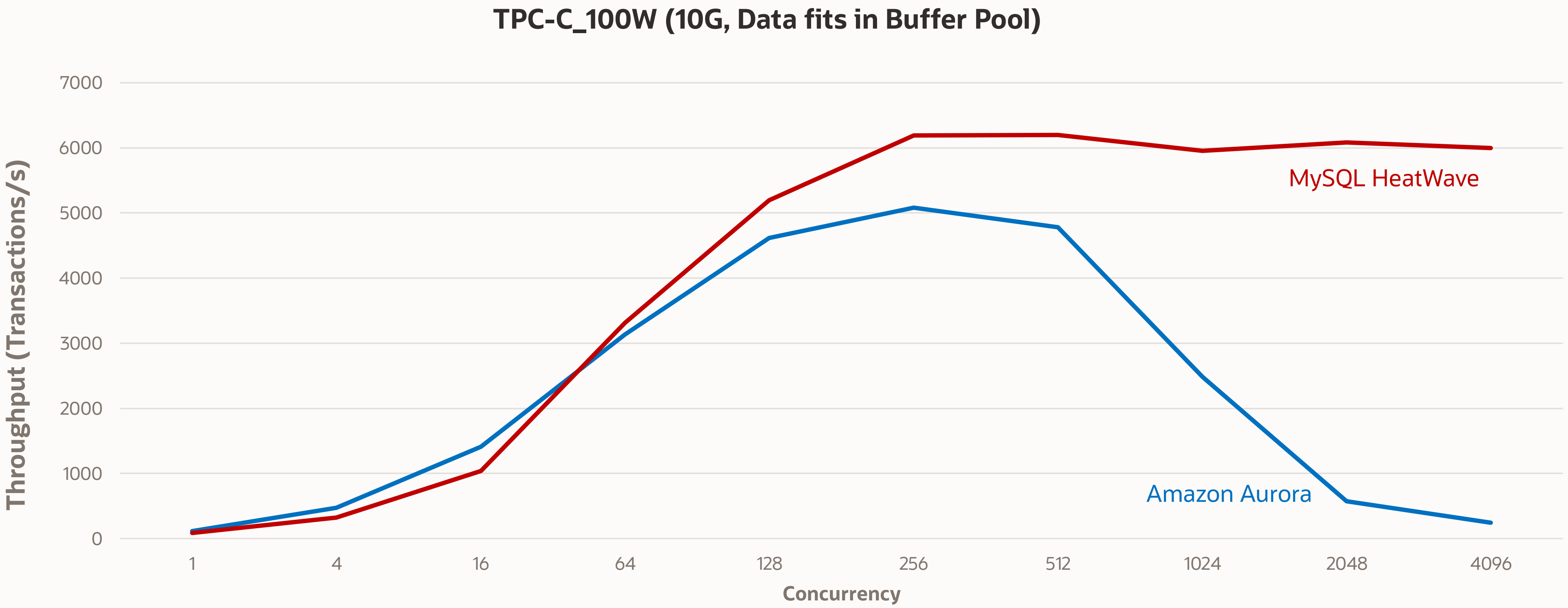
对每个云进行优化，以提供最佳的性价比

MySQL HeatWave on AWS

- MySQL HeatWave运行在AWS上，针对AWS基础设施进行了优化
- 数据不会离开AWS——节省流量成本，并避免合规性审批
- 访问MySQL HeatWave最低的延迟
- 与AWS生态系统紧密集成- S3, CloudWatch, PrivateLink

MySQL HeatWave为OLTP提供了比Aurora高10倍的吞吐量

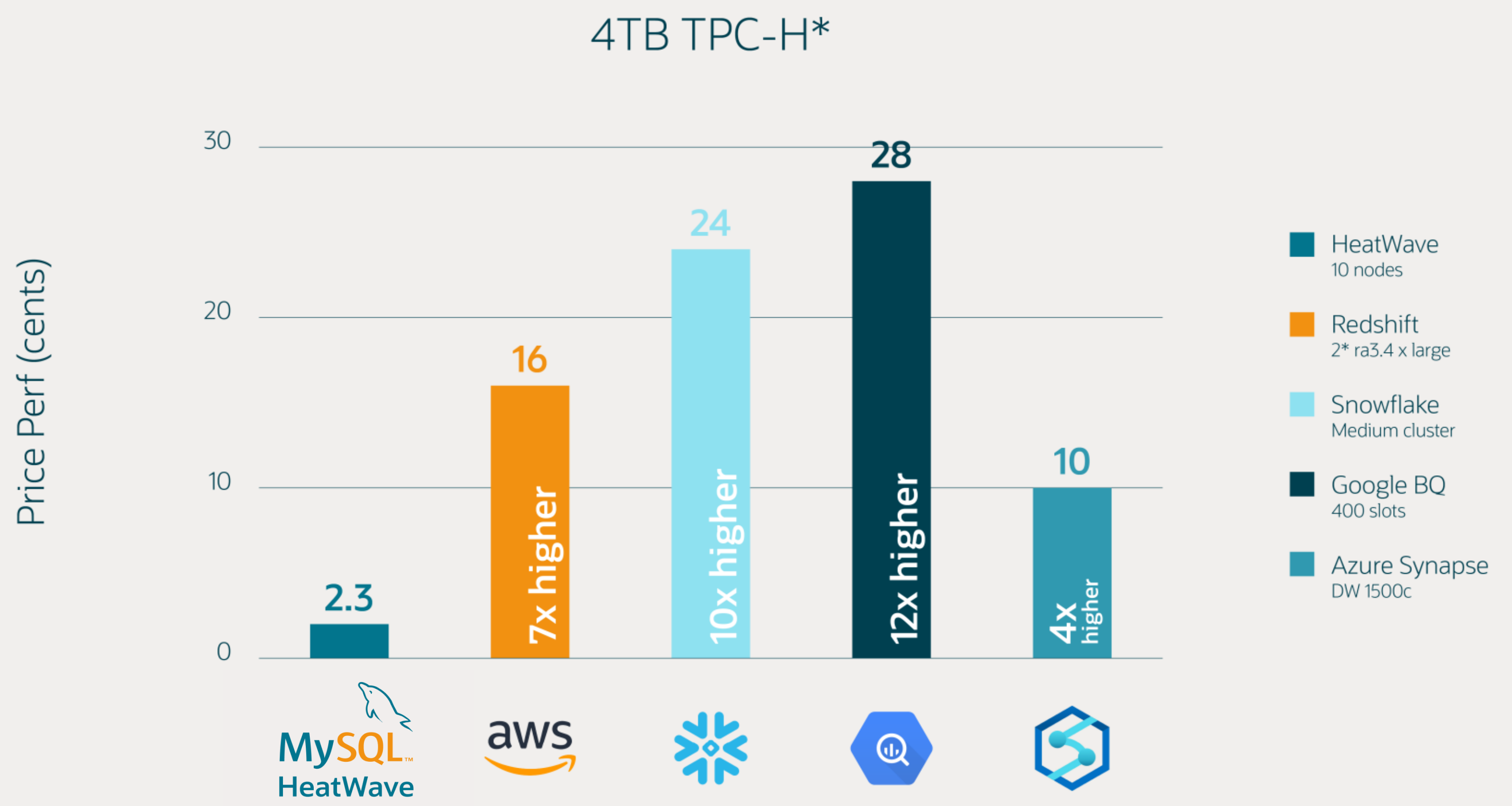
自动化线程池



*Benchmark queries are derived from the TPC-C benchmarks, but results are not comparable to published TPC-C benchmark results since these do not comply with the TPC-C specifications.

性价比对比

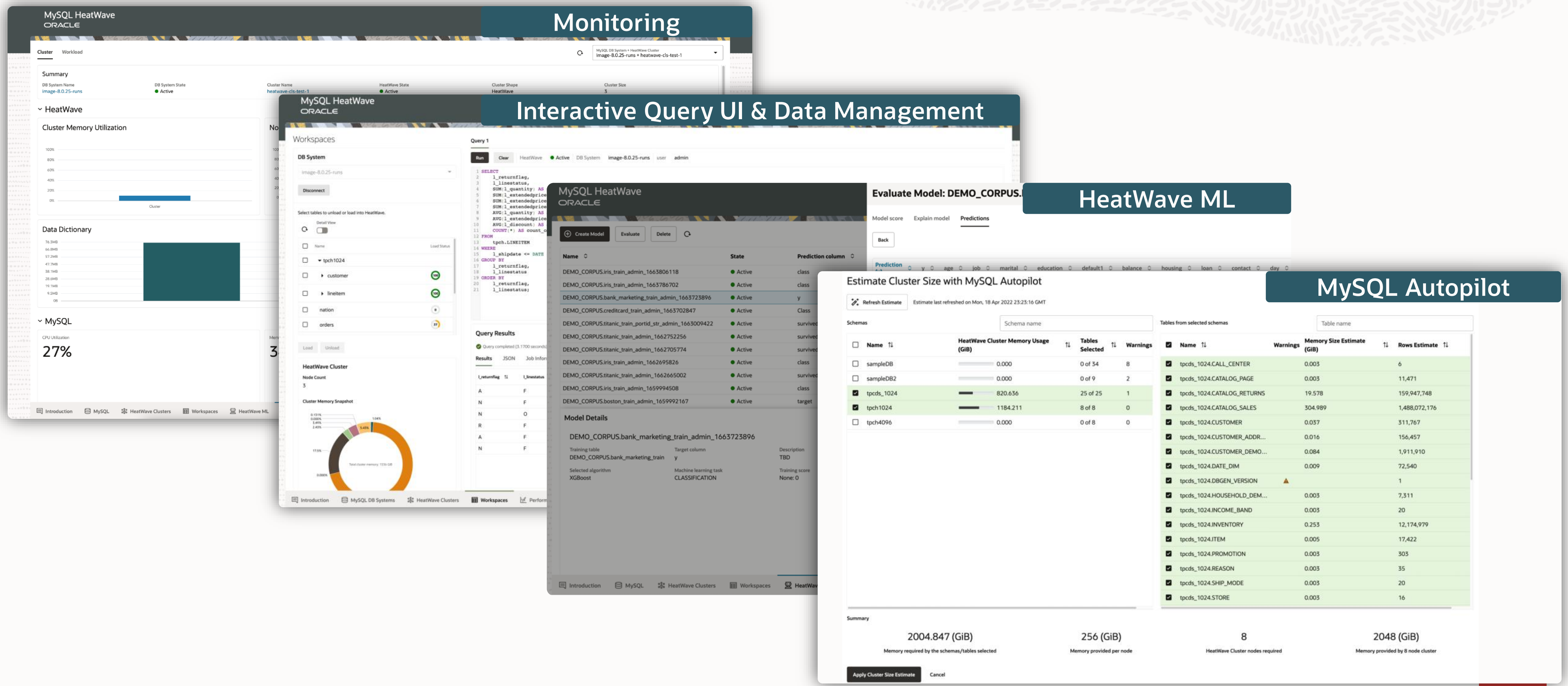
比Redshift好7倍、比Snowflake好10倍、比Big Query好12倍、比Synapse好4倍



Only compute costs are considered here
Pricing for Redshift is based on 1-year reserved instance, paid upfront. Snowflake price is based on standard edition
Price for Google BigQuery is based on monthly flat rate commitment. Azure Synapse is based on 1-year reserved pricing

*Benchmark queries are derived from the TPC-H benchmarks, but results are not comparable to published TPC-H benchmark results since these do not comply with the TPC-H specifications.

MySQL HeatWave 控制台



开发路线图OCI 及 AWS

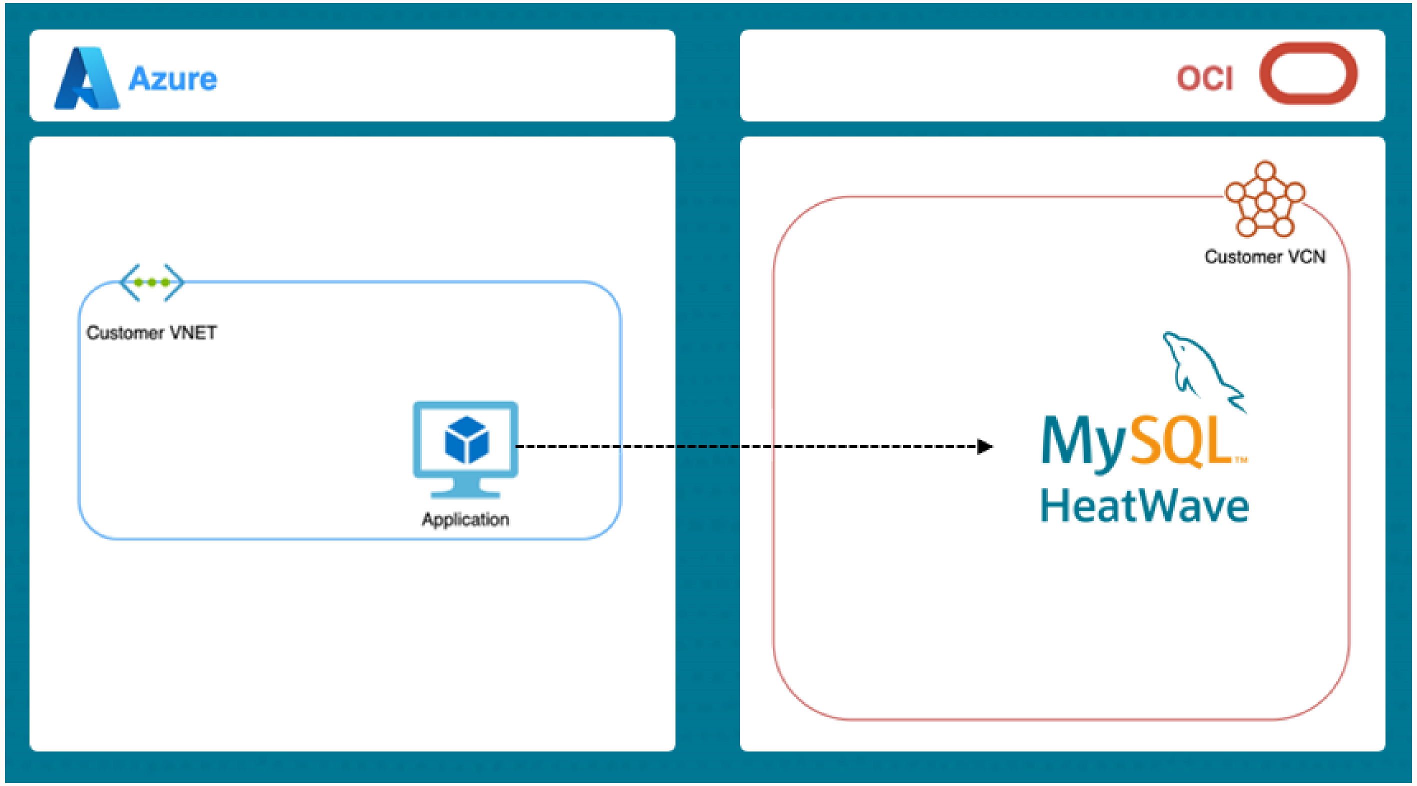
As of October 2022 - https://www.oracle.com/mysql/#roadmap

Key capabilities	Available on OCI	Available on AWS
Fully managed service	✓	✓
OLTP and OLAP in MySQL	✓	✓
Query acceleration for analytics and mixed workloads	✓	✓
Machine learning-powered automation (MySQL Autopilot for HeatWave and OLTP)*	✓	✓
Advanced security*	✓	✓
In-database machine learning (HeatWave ML)	✓	✓
Scale-out data management	✓	Coming soon
Interactive query and data management console	Coming soon	✓
Performance and workload monitoring from the console	Coming soon	✓
Adding HeatWave to any MySQL shape	Coming soon	✓

* Auto thread pooling and auto shape prediction in MySQL Autopilot as well as the MySQL HeatWave database firewall will be available soon on OCI.

MySQL HeatWave on Azure

通过Azure VNET连接 OCI的MySQL HeatWave

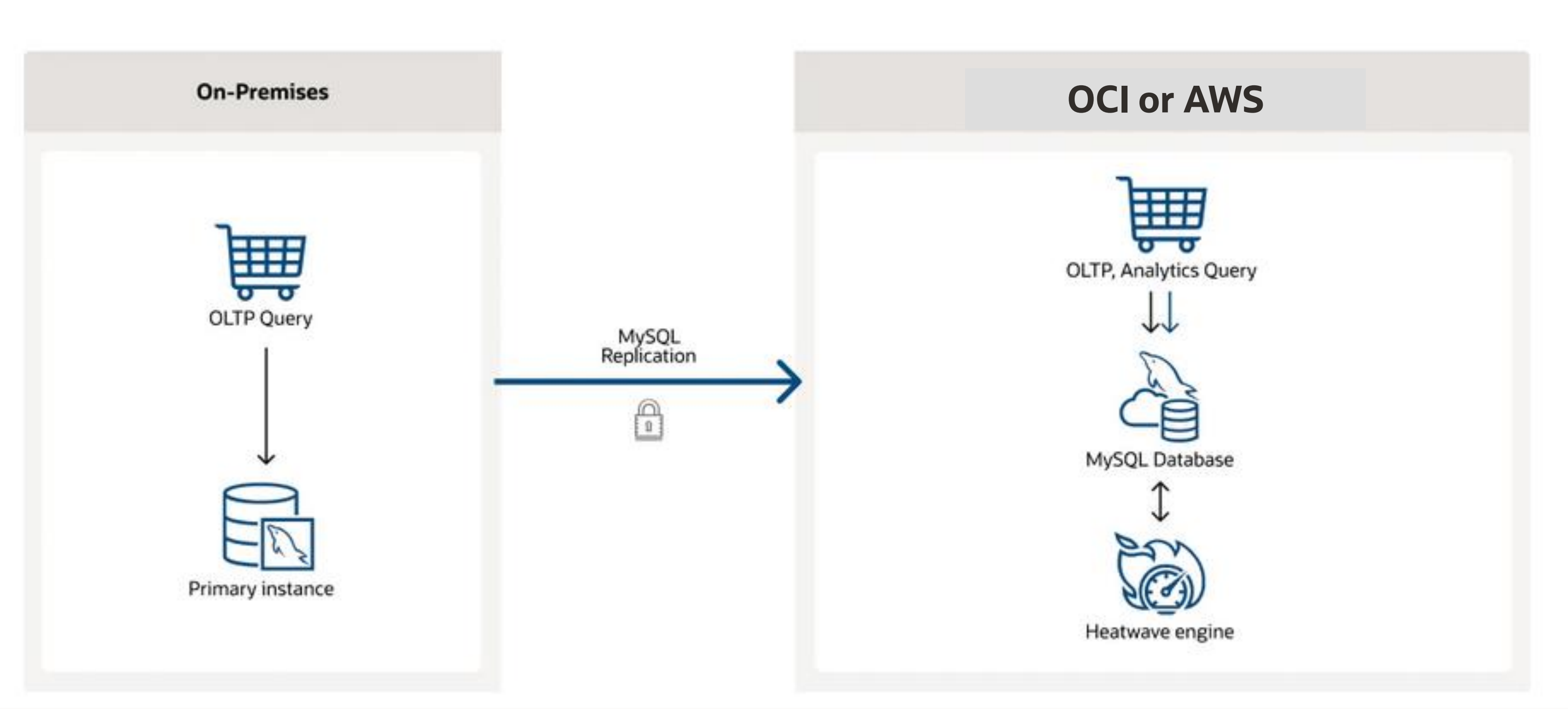


- azure原生用户体验
- 自动化集成身份、网络和监视
- 私有互连和网络延迟小于2毫秒
- 使用Microsoft Azure服务和MySQL HeatWave协作支持

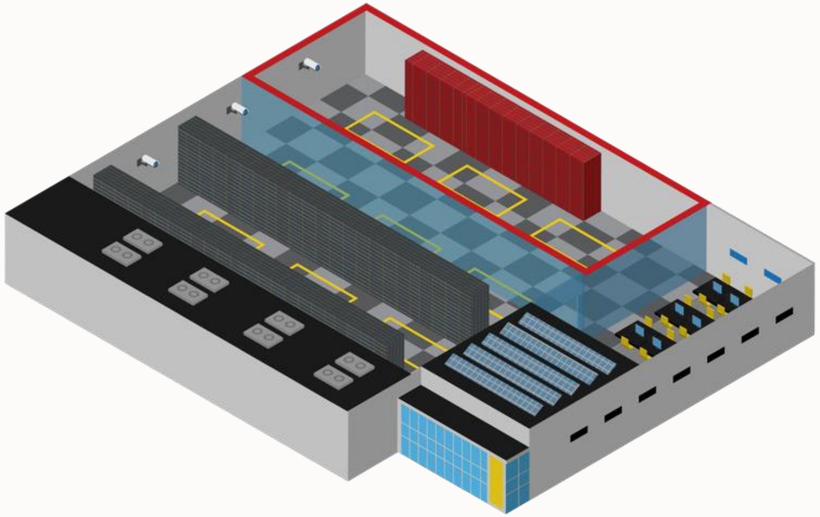
https://www.oracle.com/cloud/azure/oracle-database-for-azure

混合部署

本地部署OLTP，云端部署 OLAP



部署在用户的数据中心
Oracle Dedicated Region Cloud@Customer



独立的云区域

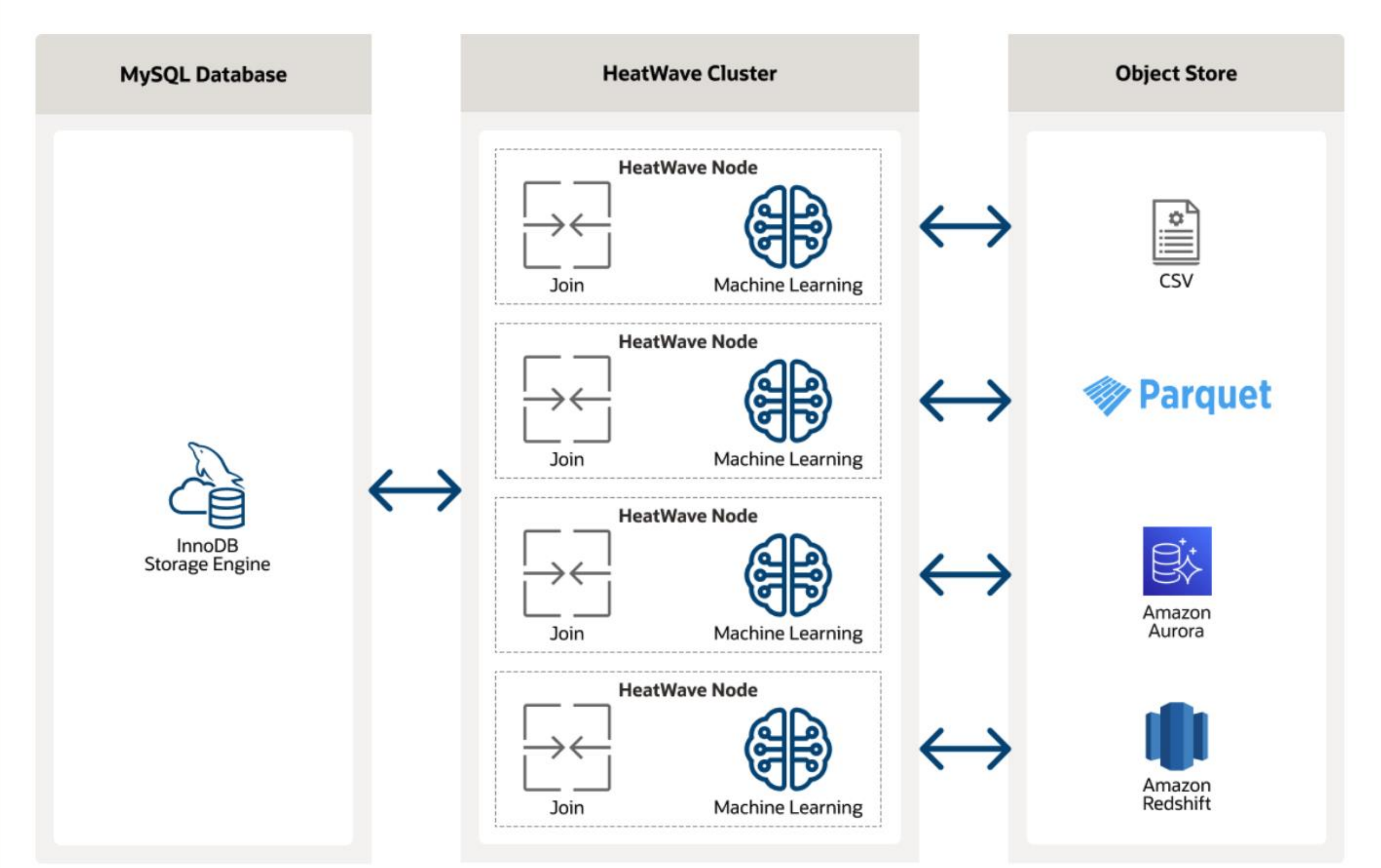
MySQL HeatWave和所有Oracle公有云服务在用户的数据中心提供

公有云的经济和安全性

满足数据驻留和延迟需求

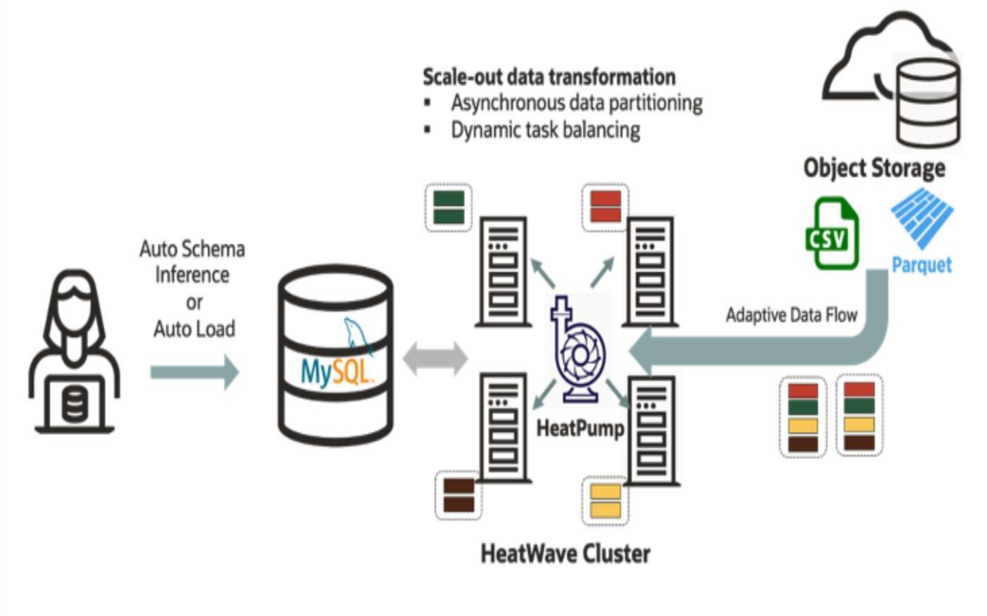
MySQL HeatWave Lakehouse (Beta)

MySQL HeatWave Lakehouse



MySQL HeatWave Lakehouse的优势

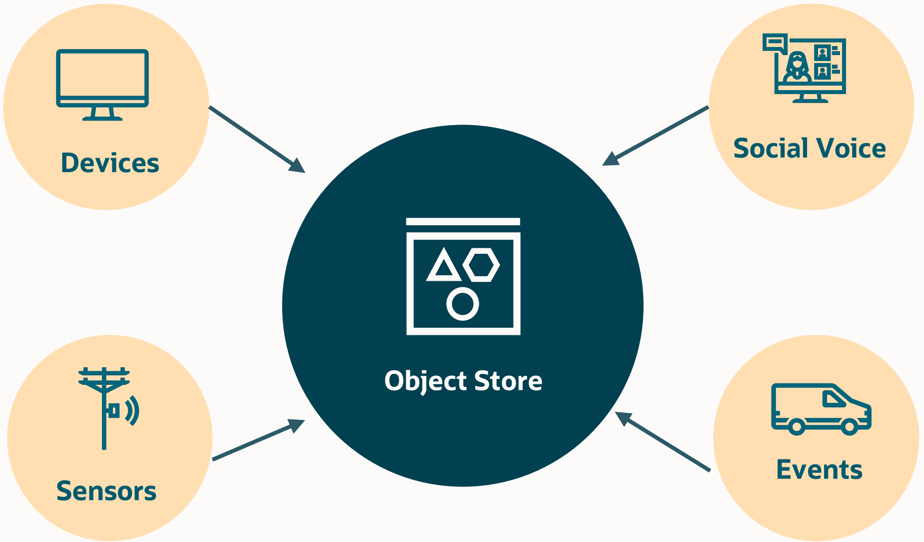
1. 横向扩展的体系结构，可以快速抽取、管理和执行查询，最多可处理400 TB的数据。
2. MySQL Autopilot自动化了常见的数据管理任务，包括半结构化数据的自动模式推断和自动加载。
3. 数据库和数据湖的统一查询引擎。
4. 保持与MySQL语法100%兼容。
5. 高可用的托管数据库服务，它可以在计算节点故障的情况下自动恢复加载到HeatWave集群中的数据。
6. 通过自动压缩相关列，高效地使用集群内存，提供高达2倍的压缩比——确保客户从所提供的HeatWave集群中获得最大收益。
7. 使用Pre-Authentica安全访问控制方法，完全控制对数据湖源的访问。



数据库外存储的数据显著增长

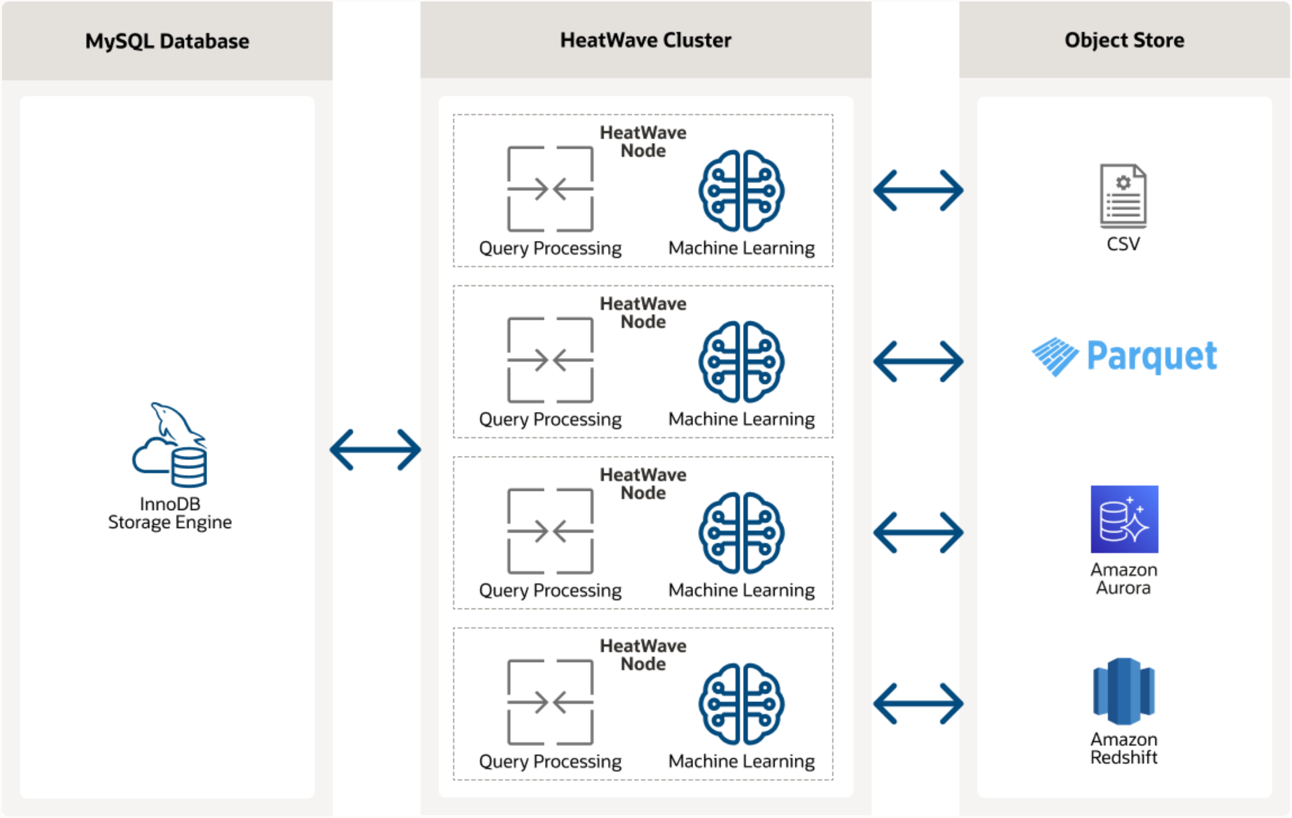
管理海量数据具有挑战性

- 2020年每人每秒产生1.7 MB数据。到2025年，该数字翻一番(TechJury)
- 2025年预计180 ZB的数据(IDC)
- 超过80%的数据都在文件中(CIO)
- 通常存储在对象存储的数据湖中
- 99.5%的收集数据未被使用(grow.com)



MySQL HeatWave Lakehouse (Beta)

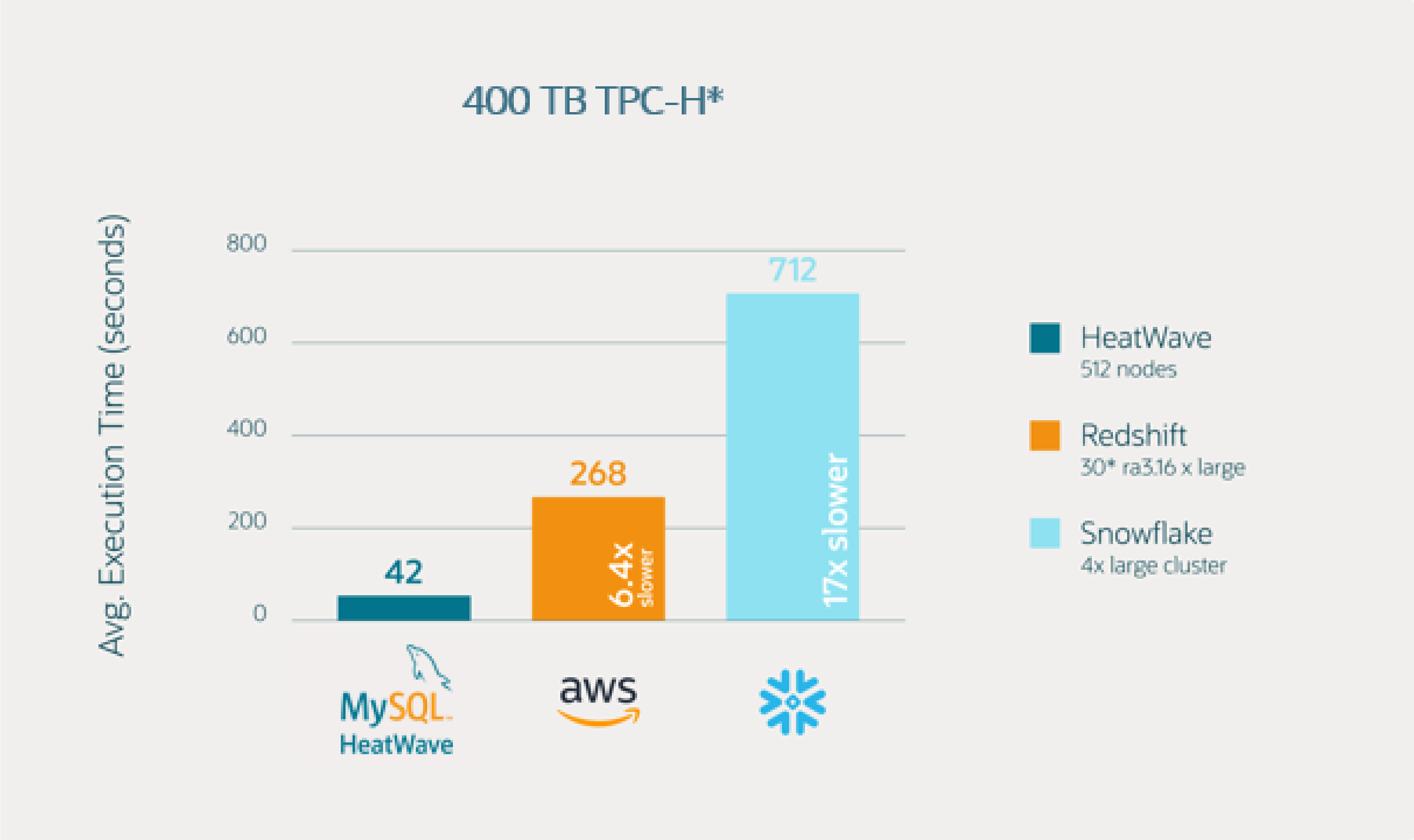
在对象存储中以各种文件格式处理和查询100TB的数据



- 使用标准MySQL命令跨MySQL、对象存储或两者查询数据
- 高达400TB的数据—HeatWave集群可扩展到512个节点
- 查询数据库和查询对象存储一样快速

MySQL HeatWave Lakehouse的查询性能

Snowflake的17倍,，Redshift的6倍

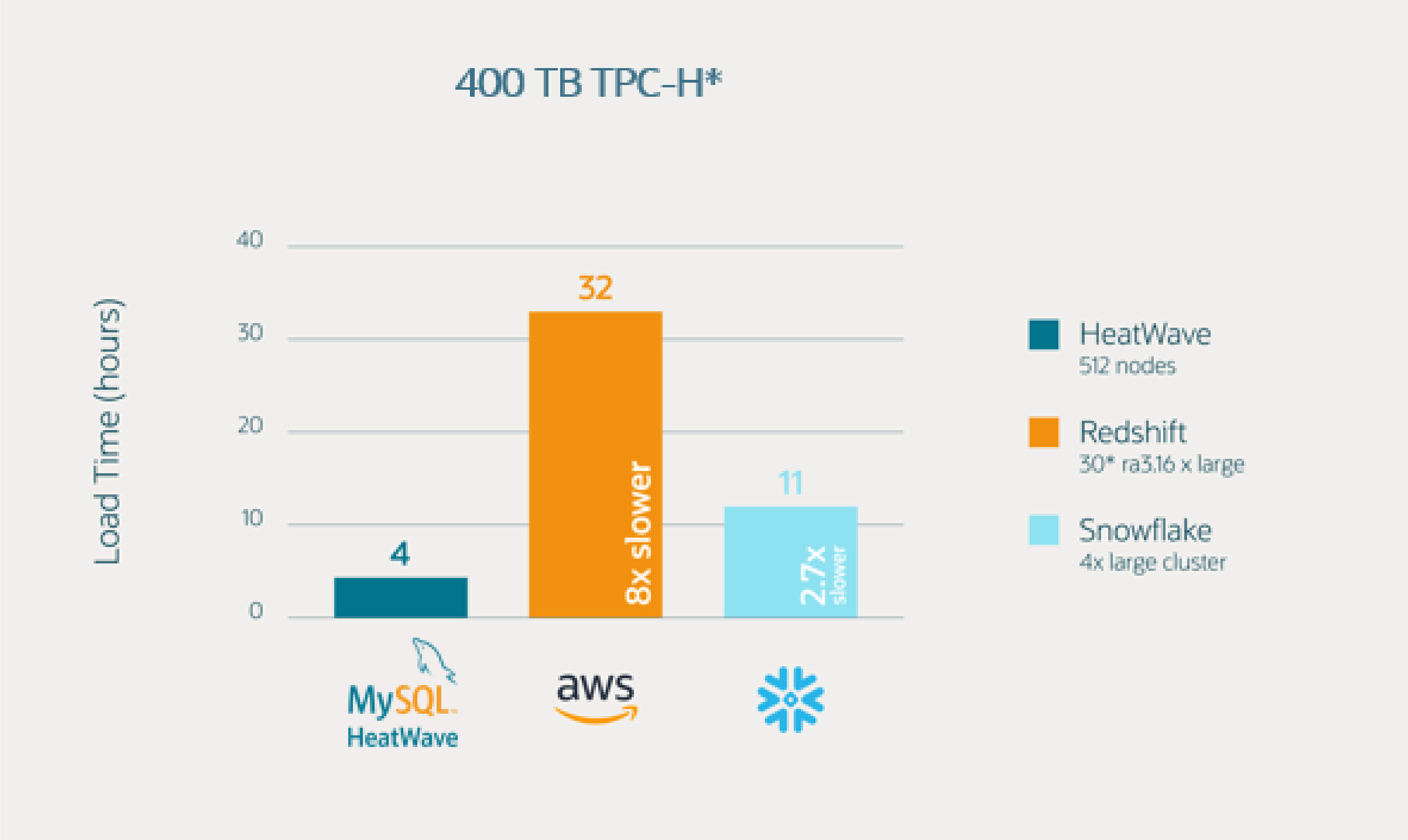


Annual provisioned cost for HeatWave $1,589,036; for Redshift $2,261,760; for Snowflake $2,242,560
Pricing for Redshift is based on 1-year reserved instance, paid upfront. Snowflake is based on standard edition

*Benchmark queries are derived from the TPC-H benchmarks, but results are not comparable to published TPC-H benchmark results since these do not comply with the TPC-H specifications.

MySQL HeatWave Lakehouse 的加载性能

Redshift的8倍，Snowflake的2.7倍



Annual provisioned cost for HeatWave $1,589,036; for Redshift $2,261,760; for Snowflake $2,242,560
Pricing for Redshift is based on 1-year reserved instance, paid upfront. Snowflake is based on standard edition

*Benchmark queries are derived from the TPC-H benchmarks, but results are not comparable to published TPC-H benchmark results since these do not comply with the TPC-H specifications.

MySQL Autopilot for MySQL HeatWave Lakehouse

机器学习驱动的自动化

自动部署

- 自适应采样原始文件和收集统计数据
- 估计要加载的数据的内存占用

自动推断模式

- 用于推断列数据类型的示例原始文件
- 生成DDL以创建表

自适应采样

- 自适应采样部分文件以收集统计信息
- 为各种Autopilot功能收集的统计

自动加载

- 预测加载时间
- 生成加载脚本

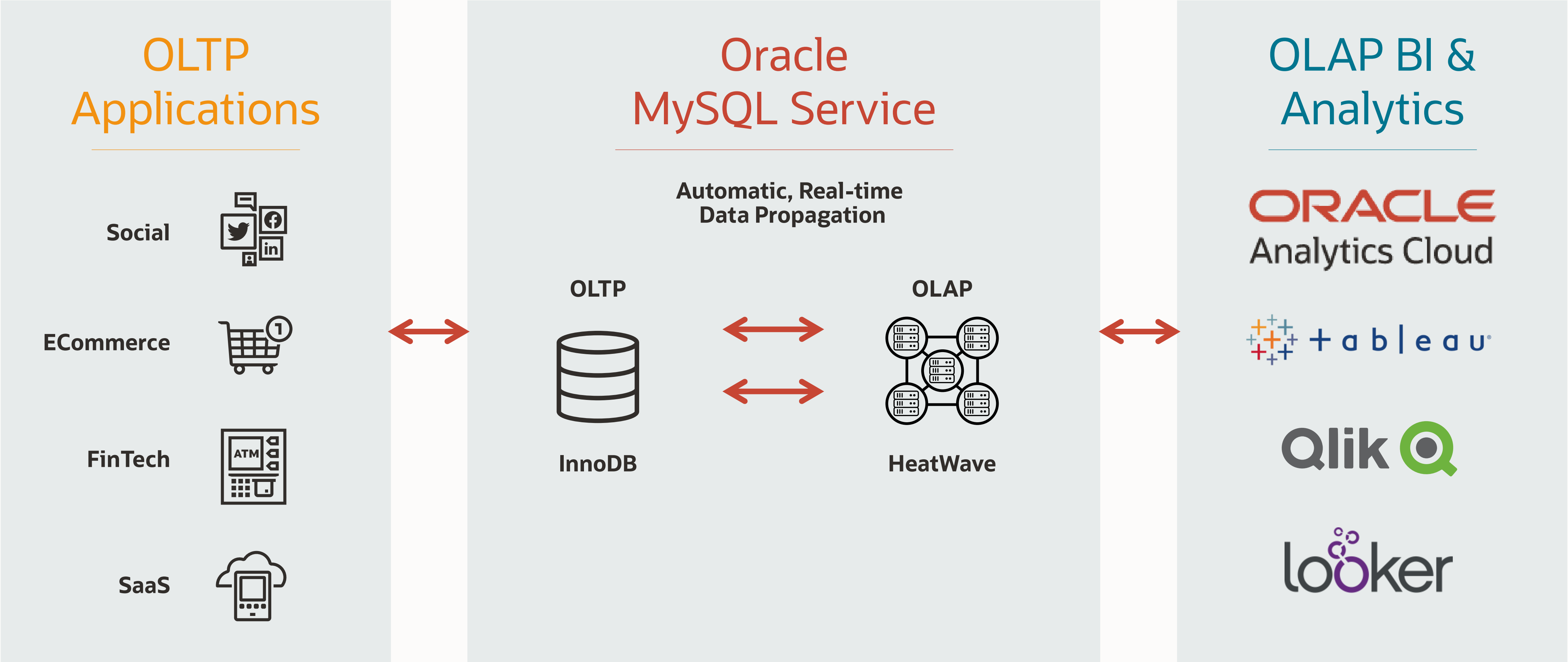
自适应数据流

- 系统适应对象存储的性能
- 提高系统性能和可靠性

自动改善查询计划

- 在运行查询时持续收集统计信息
- 增强未来执行计划

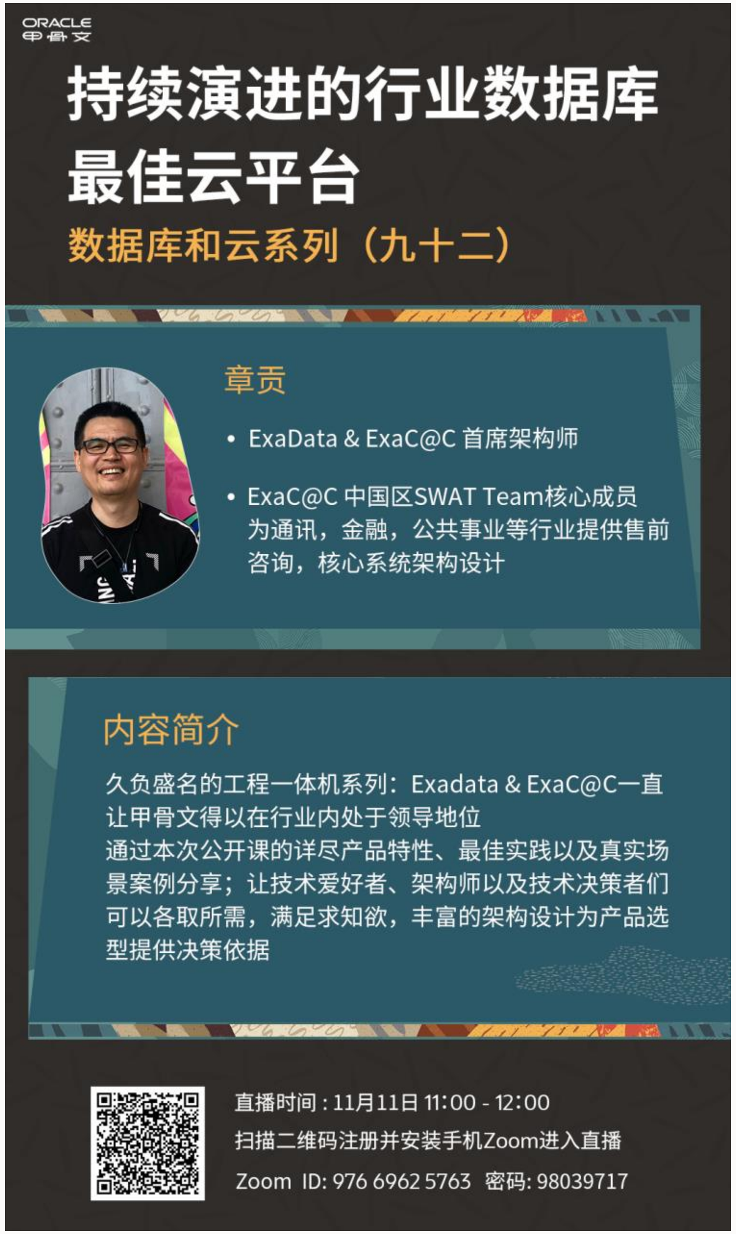
所有与MySQL兼容的分析应用程序无需任何更改即可运行



MySQL HeatWave 用户

MySQL™

oracle.com/heatwave



20-19

数据库和云讲座群

甲骨文云技术公众号

技术专家1V1深入交流