

Oracle机器学习及人工智能

公益讲座11: 00分准时开始，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！



20-18

数据库和云讲座群



甲骨文云技术公众号

Oracle机器学习及人工智能

王炜

Oracle China Big Data Team

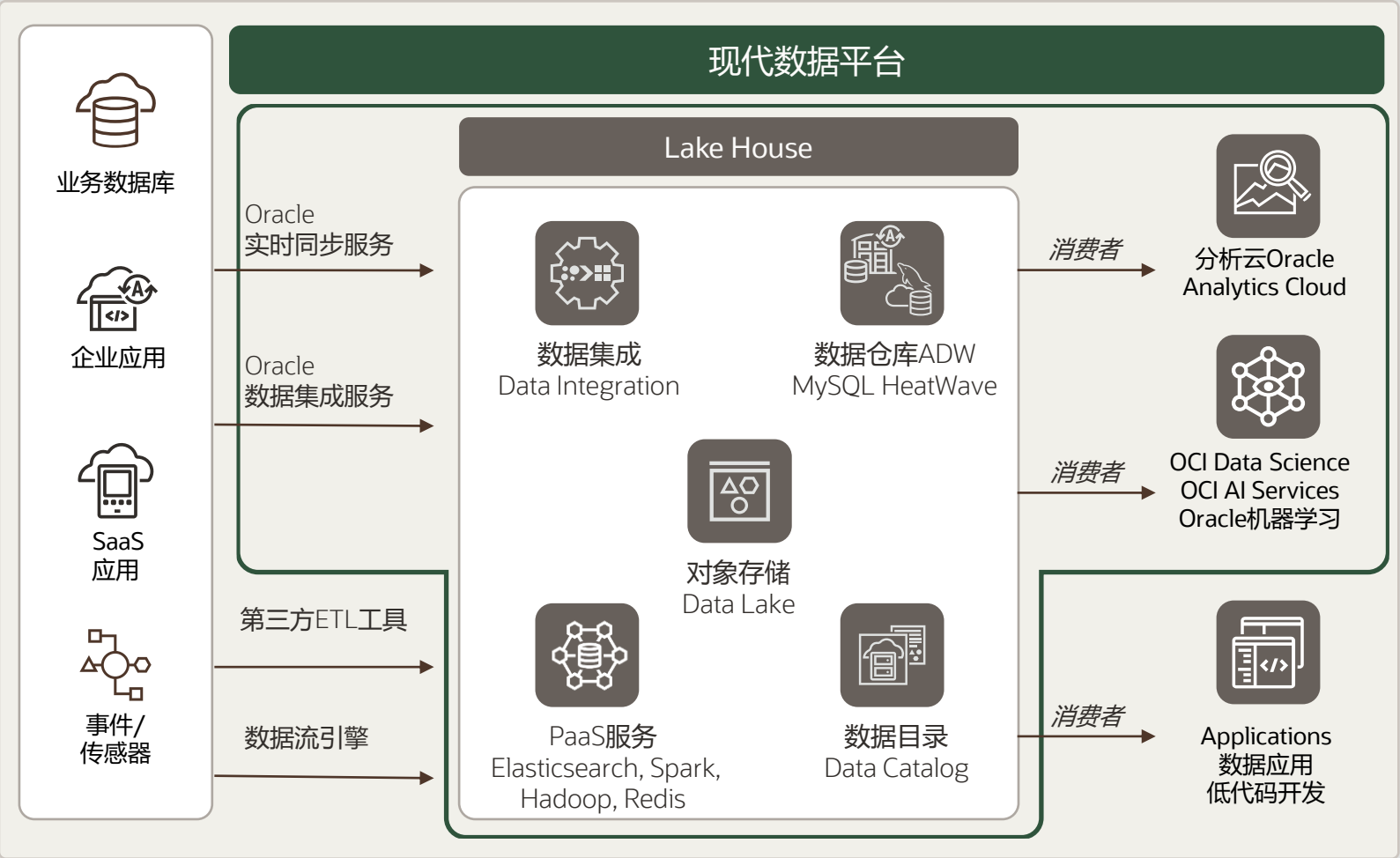
Safe Harbor

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Statements in this presentation relating to Oracle's future plans, expectations, beliefs, intentions and prospects are "forward-looking statements" and are subject to material risks and uncertainties. A detailed discussion of these factors and other risks that affect our business is contained in Oracle's Securities and Exchange Commission (SEC) filings, including our most recent reports on Form 10-K and Form 10-Q under the heading "Risk Factors." These filings are available on the SEC's website or on Oracle's website at <http://www.oracle.com/investor>. All information in this presentation is current as of September 2019 and Oracle undertakes no duty to update any statement in light of new information or future events.

Oracle融合数据平台助力企业实现云中敏捷数据分析

支撑企业实现从数据应用部署到数据价值的快速洞察



Autonomous Data Warehouse: 具有高性能存储和分析功能的自治数据管理平台

MySQL Heatwave: MySQL的高性能分析

Object Storage Data Lake: 低成本存储

Managed Open-source Services: Oracle 管理的平台云服务 (Spark、Hadoop、Elasticsearch、Redis)

OCI Data Integration: 为数据科学和分析轻松提取、转换和加载 (ETL) 数据。设计数据湖和数据仓库之间的无代码数据流

OCI Data Catalog: 维护数据湖和数据仓库用于数据发现的数据资产清单

OCI Analytics Cloud: 集成了经典KPI报表, 敏捷可视化和增强智能的分析工具

OCI Data Science: 预配置基于Python的机器学习协同开发和运行环境



Oracle 机器学习及AI整体解决方案

涵盖云服务、应用和全数据资产的统一 AI/ML 平台

Applications



OCI Digital Assistant



OCI Language



OCI Speech



OCI Vision



OCI Anomaly Detection



OCI Forecasting

AI Services



OCI Data Science



Oracle Database Machine Learning



OCI Data Labeling

Machine Learning Services

Data

Oracle数据科学云服务 (Data Science Service)



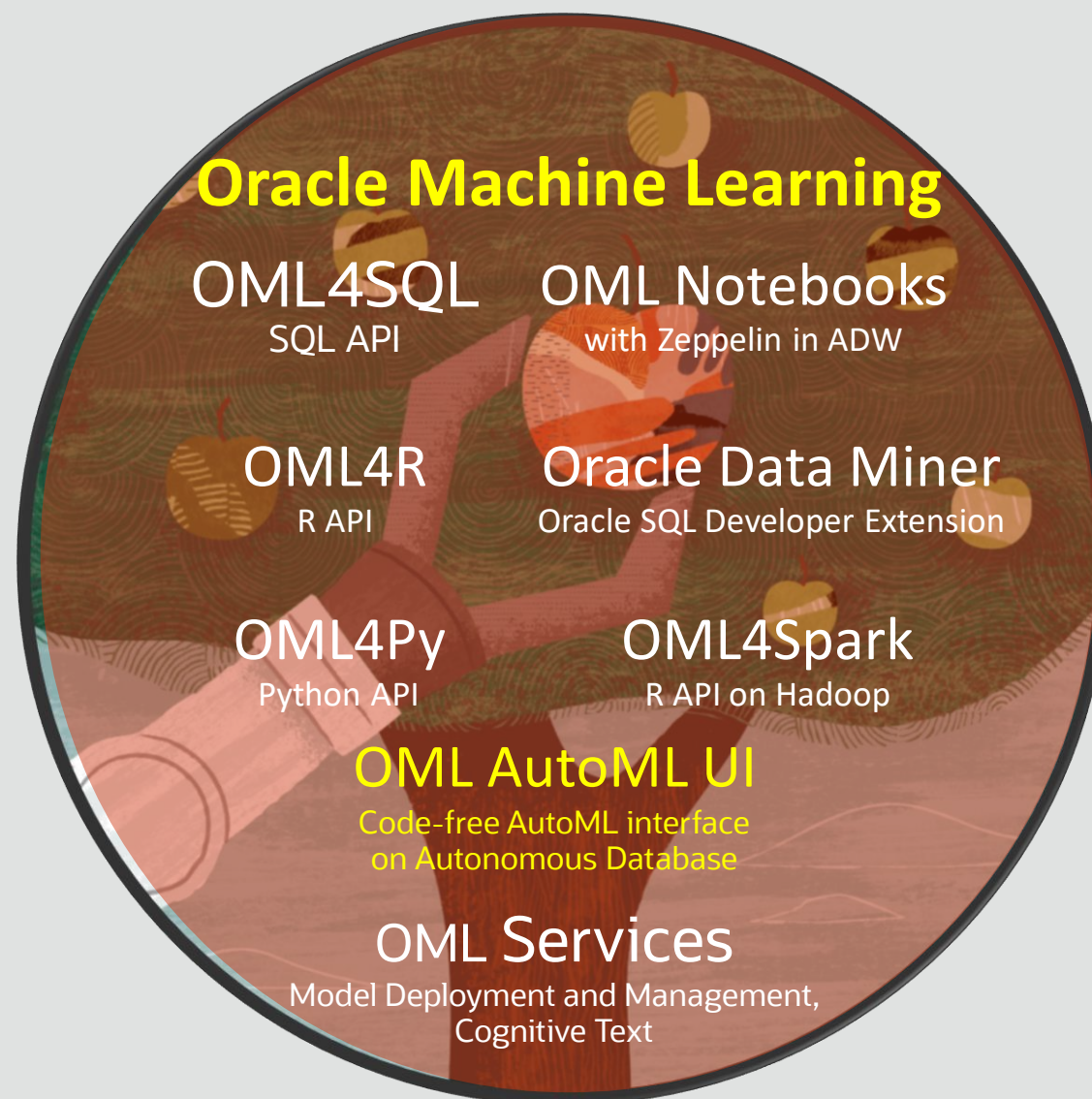
应用场景

- 客户需要高级分析和机器学习从现有数据资产中提取更多业务价值
- 客户拥有专业数据科学家，希望能兼容现有主流开源平台技术和应用资源

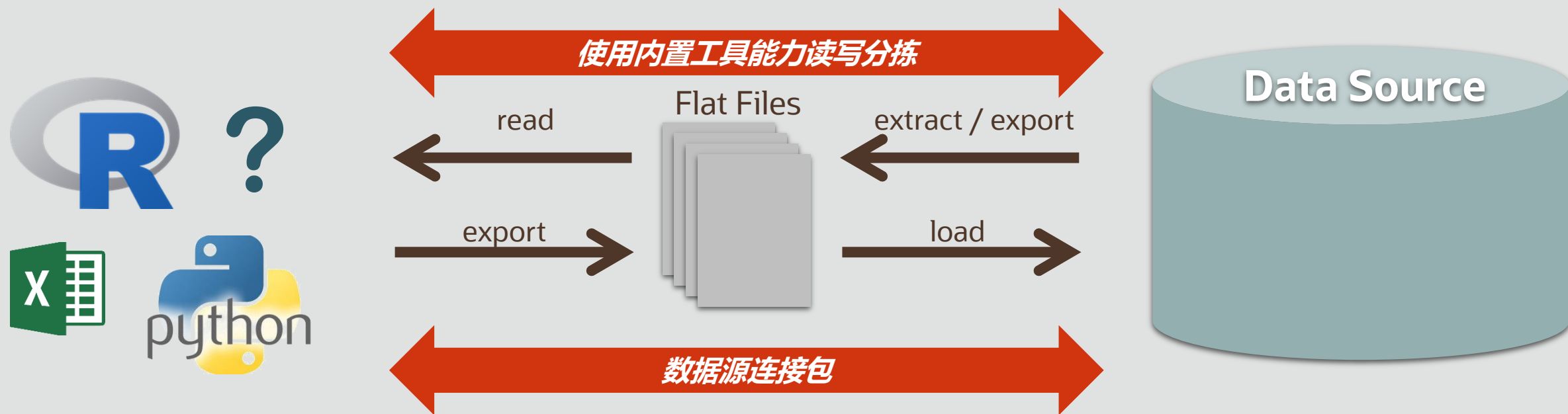
价值体现

- 功能强大：Oracle 云基础设施数据科学平台专为满足现代数据科学家的需求而设计，融合了最新的开源机器学习工具包与 Oracle 的专有技术。
- 可扩展：凭借 Oracle 云硬件的高速性能和可扩展性，数据科学家可以扩展其数据科学资源以应对组织中的大数据挑战。
- 协作：实现细粒度访问控制 and 安全性，集中管理数据科学资产，从而让数据科学家团队可以在一个协作式工作区协同作业。

Oracle数据库机器学习平台：面向数据的机器学习



传统分析和数据源交互



Deployment

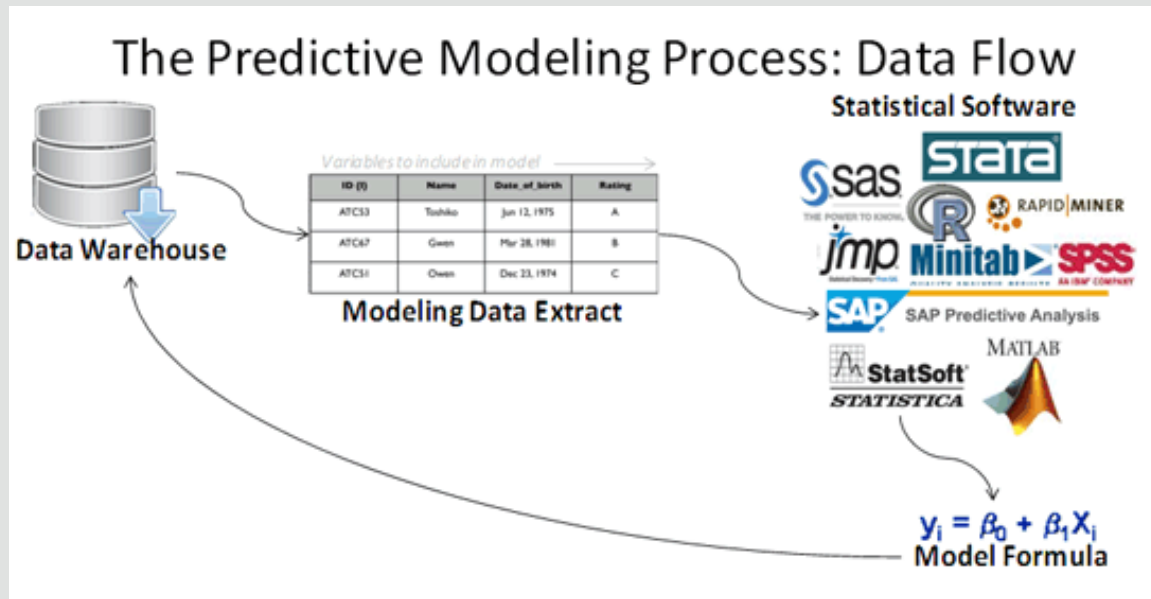
Ad hoc
cron job

- 访问延迟
- 范式转变：R/Python → 数据访问语言 → R/Python
- 内存限制 = 数据大小、内存处理能力
- 单线程
- 备份、恢复、安全问题
- 临时生产部署

传统 vs. Oracle机器学习/预测分析

传统 — “移动数据”

ORACLE®: “无需移动数据!”



Oracle Machine Learning Notebooks

自治数据库作为数据科学平台

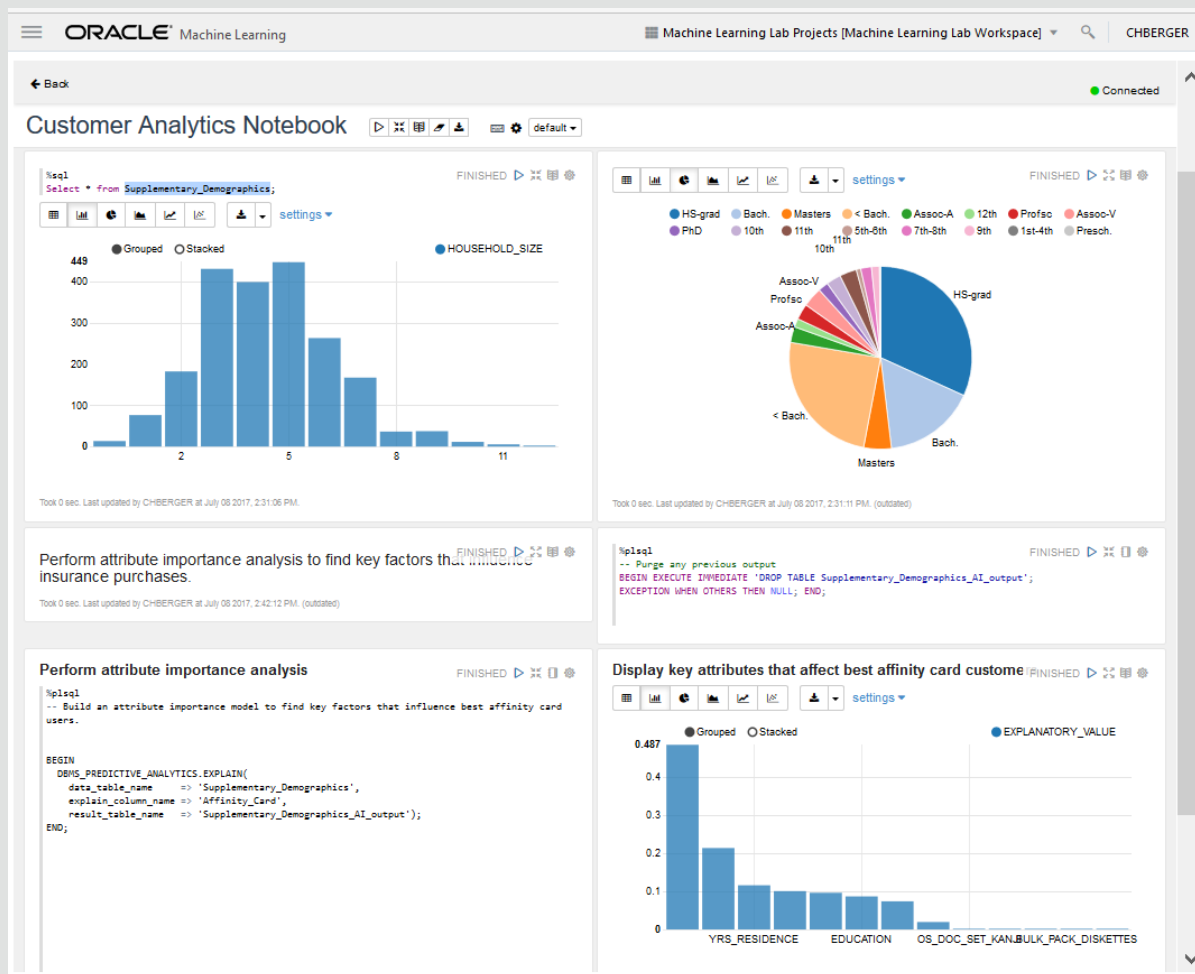


交互式Web页面

- 基于Apache Zeppelin
- 支持数据科学家、数据分析师、
- 应用程序开发人员、懂 SQL 的 DBA 以及Python开发人员
- 轻松共享笔记本和模板
- 权限、版本控制和执行
- 调度

包括在自治数据库中

- 自动预配、管理、备份
- 数据库算法和分析功能
- 探索和准备、构建和评估模型，
- 对数据进行评分，部署解决方案
- 即将支持 R 进行增强



Oracle Machine Learning for SQL

使 SQL 用户能够立即访问 Oracle 数据库和 Oracle 自治数据库中包含的 机器学习

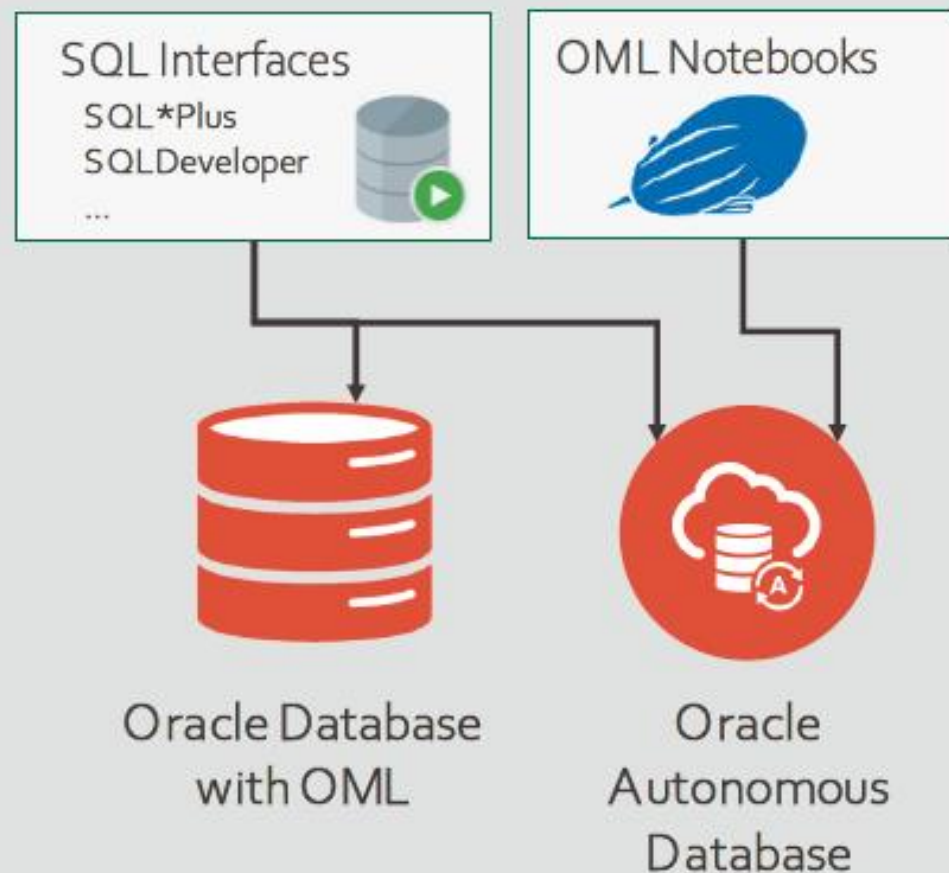
数据库中并行的，分布式算法

- 不需要将数据提取到单独的 ML 引擎
- 快速且可扩展
- 批量和实时评分
- 解释性预测详细信息

机器学习模型部署在全球领先数据库对象中

- 通过权限访问控制
- 审核用户操作
- 跨数据库导出/导入模型

跨Oracle堆栈利用机器学习

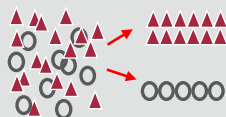


Oracle机器学习算法



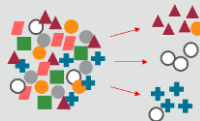
分类

朴素贝叶斯
逻辑回归 (GLM)
决策树
随机森林
神经网络
支持向量机
显式语义分析
XGBoost



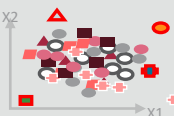
聚类

分层 K 均值
分层 O-Cluster
期望最大化 (EM)



异常检测

单类支持向量机 (One-Class SVM)
MSET-SPRT

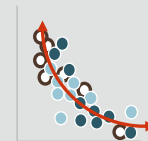


时间序列

预测 - 指数平滑
包括流行的模型,
例如具有趋势、季节性、不规则
性、缺失数据的 Holt-Winters

回归

线性模型
广义线性模型
支持向量机 (SVM)
逐步线性回归
神经网络
LASSO



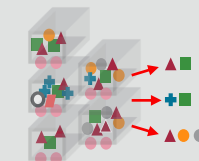
属性重要度

最小描述长度
PCA
非监督式 Pair-wise KL Div
CUR decomposition for row & A



关联规则

A priori/ 购物篮



预测查询

Predict, cluster, detect, features

SQL 分析

SQL Windows
SQL Patterns
SQL Aggregates



特征提取

PCA
非负矩阵分解
奇异值分解 (SVD)
显式语义分析 (ESA)

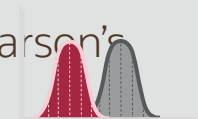
文本挖掘

算法支持文本标记化和主题提取
文档相似性的显式语义分析 (ESA)



统计函数

基本统计: min, max,
median, stdev, t-test, F-test, Pearson's
Chi-Sq, ANOVA, etc.



R 包

通过嵌入式执行的第三方 R 包
Spark MLlib 算法集成



模型部署

SQL—1st Class Objects
Oracle RESTful API (ORDS)
OML Microservices (Apps)



Oracle Data Miner 用户界面

创建分析工作流 – 数据科学家的生产力工具 – 人人都是数据科学家



- 用于 Oracle 数据库的SQL Developer扩展
- 自动执行典型的数据科学步骤
- 易于使用的拖放操作
- 快速定义和共享分析工作流
- 广泛的算法和数据转换
- 生成 SQL 代码，立即部署

The screenshot displays the Oracle Data Miner interface within the SQL Developer environment. The main workspace shows a workflow diagram with nodes such as 'Scatter Box plots etc. 1', 'Explore Data', 'CUST_INSUR_LTV1', 'Filter Columns', 'Multiple Classification Models', 'Most Likely Customers', and 'LIKELY_BUY_INSURANCE_CUSTMRS 1'. The 'Query Builder' pane at the bottom shows a SQL query for creating a model and selecting data. The 'Script Output' pane displays the results of the query, including a table with columns 'POLICYNUMBER', 'PERCENT_FRAUD', and 'RNK'. The 'Components' pane on the right lists various data science tasks and models available in the tool.

POLICYNUMBER	PERCENT_FRAUD	RNK
1	654	61.87
2	11069	57.37
3	7435	55.47

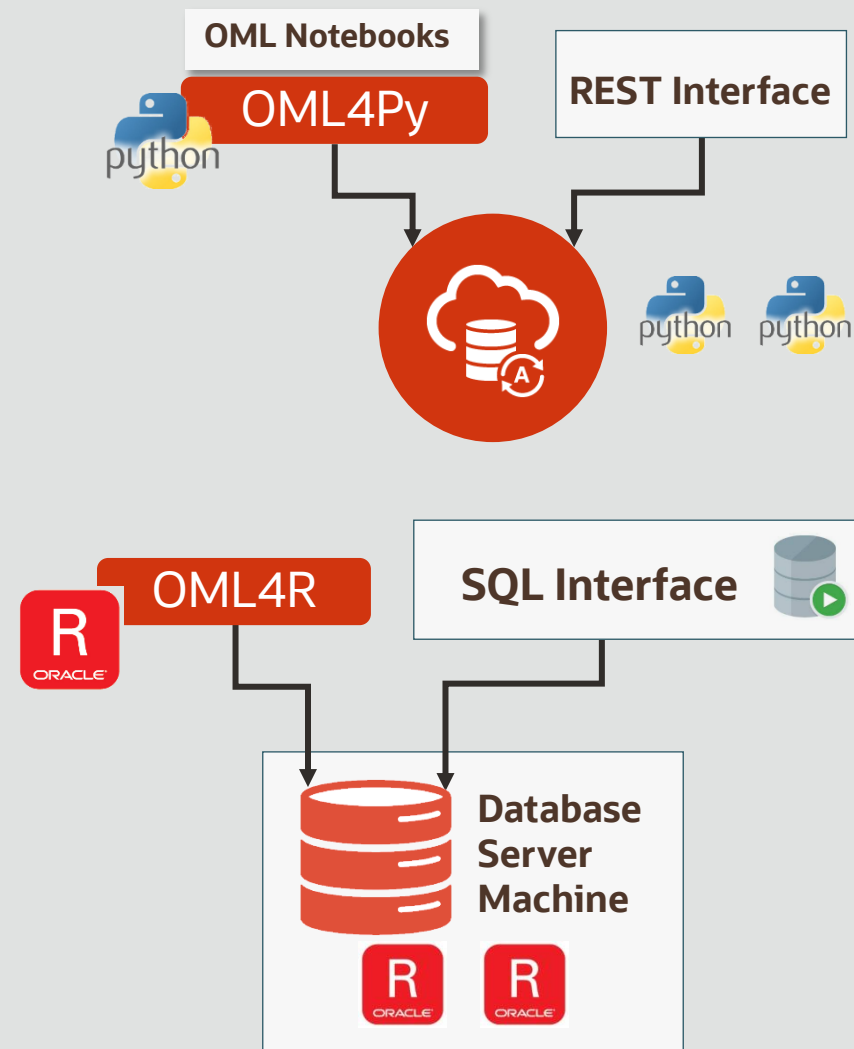


Oracle Machine Learning for R and Python

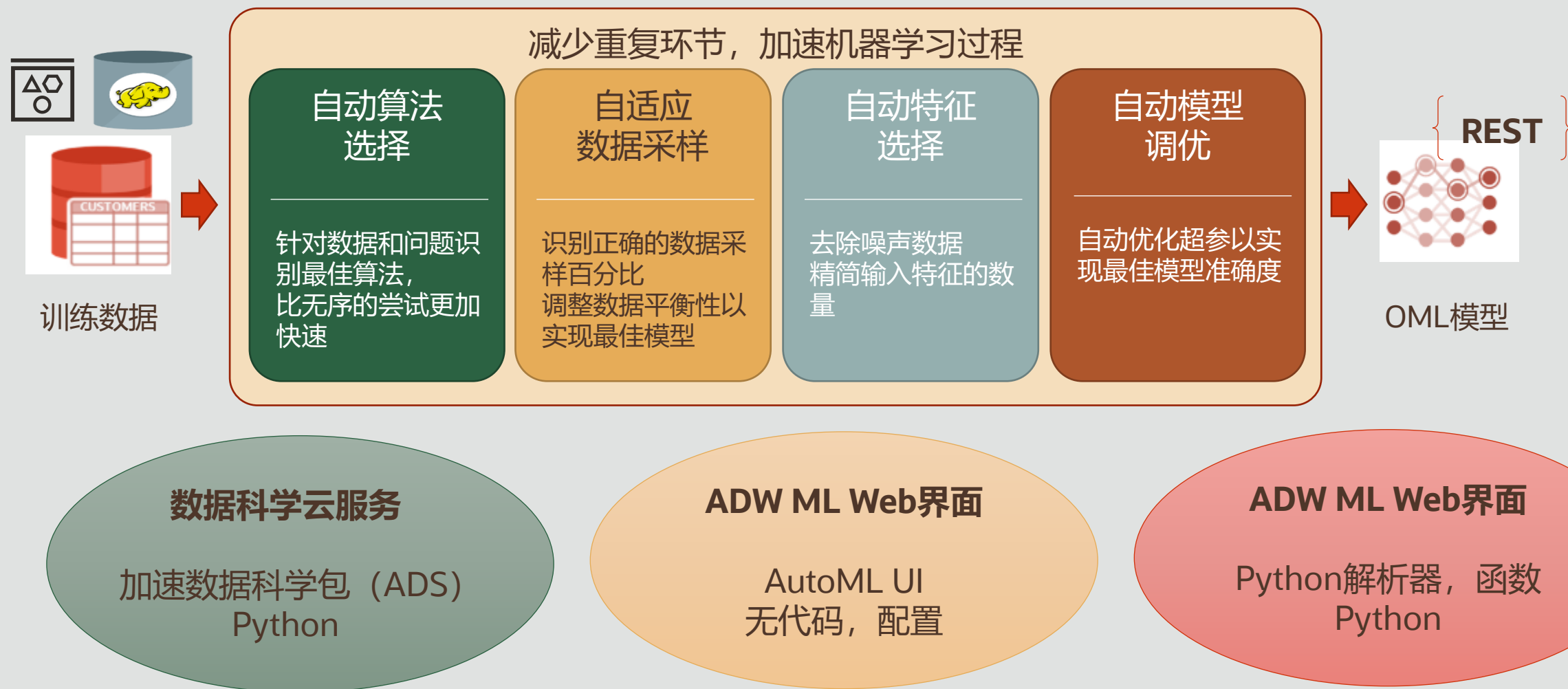
Oracle 数据库 – R / Python

Oracle Autonomous Database – Python

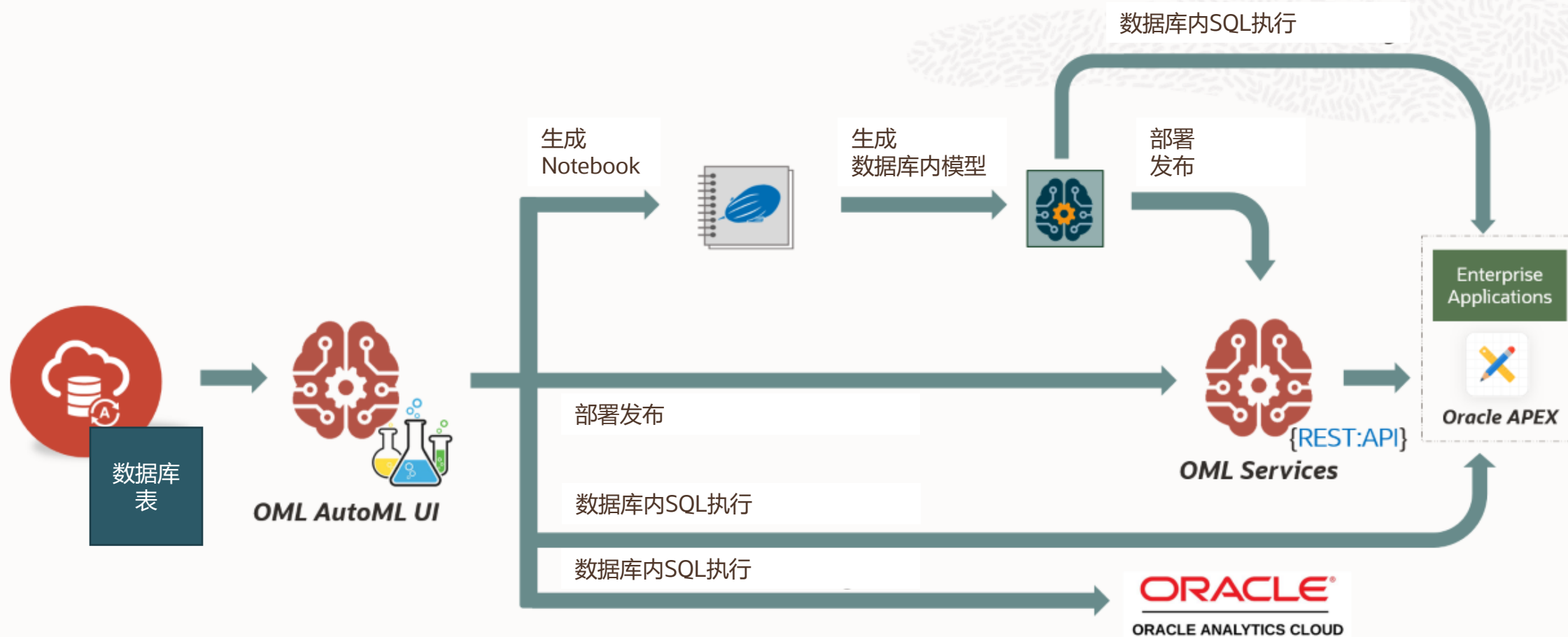
- Oracle 数据库作为 HPC 环境
- 利用代理对象，使数据保留在数据库中
- 数据库内并行化和分布式机器学习算法
- 在 Oracle 数据库中管理和调用 R 或 Python 脚本
- 使用开源包增强功能
- 通过 SQL 或 REST 将结果集成到应用程序和仪表板中
- OML4Py 自动化机器学习 AutoML
 - 自动特征选择、模型选择、超参数调整
 - 为数据科学家提供开源环境



Oracle 自动化机器学习 (AutoML) 能力



OML AutoML UI可与更多平台协作



Oracle ML Value to Customers

- • 使用 ML 支持数据科学家和分析师、开发人员和 DBA/IT
- 数据库内并行化和分布式机器学习算法，消除成本高昂的数据移动和延迟
- 快速、可扩展的数据探索、数据准备和 ML 算法
- 支持 30 多个算法：回归、分类、时间序列、聚类、特征提取、异常检测
- R 和 Python 集成支持数据科学家
- 易用的 ML 模型和 R/Python 脚本部署
- 自动执行关键 ML 流程步骤
- 利用现有的备份、恢复和安全机制和协议
- 从大多数企业数据存在的地方启动 - 将算法引入数据！

Oracle企业级数据探索加速实现业务价值



创建连接
选择连接类型

Oracle 应用产品

Oracle 数据库

AI

Action Ingres

AR

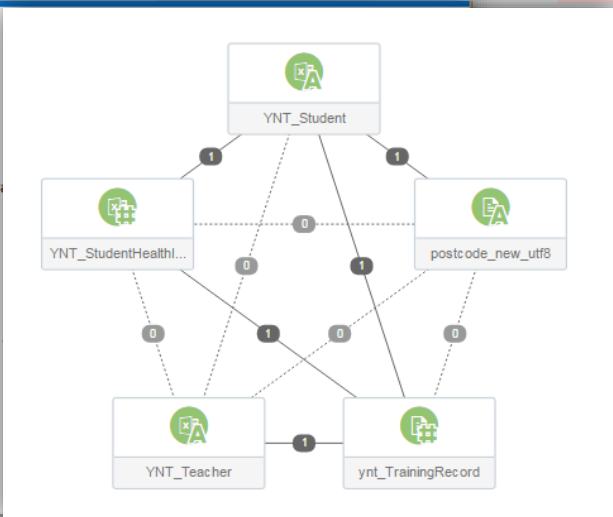
Amazon Redshift

AM

Action Matrix

AD

Apache Drill



Heart Disease Likelihood Prediction - 数据流

数据流步骤: 添加数据, 联接, 应用模型, 保存数据

model - Heart disease likelihood Training - CART

	Absent	Less Likely	Likely	Others	总计	精确率	召回率	F1
Absent	22	0	0	15	37 (62%)	76%	58%	0.66
Less Likely	4	0	0	4	8 (13%)	0%	0%	0.00
Likely	2	0	0	6	8 (13%)	0%	0%	0.00
Others	1	0	0	6	7 (12%)	19%	83%	0.31
总计	29 (48%)	0 (0%)	0 (0%)	31 (52%)	60 (100%)			

置信度: 0 100%
非置信度: 0 100%

Value: PredictedValue
ConfidencePercentage: PredictionConfidencePercentage
Group: PredictionGroup

Age	Sex	Chest Pain Type	Resting BP
50	FEMALE	Asymptomatic	110

混合数据

支持多种数据源, 包括用户自己上传Excel。生成数据集可按权限浏览和使用

轻建模

提供数据流对数据集进行处理加工, 通过关键字连接多个数据集, 构建数据统一视图

敏捷可视化

提供丰富的图表, 拖曳即可生成, 不需要编写代码, 支持移动端

相关性探索

支持关键字搜索, 视图内或全局条件过滤, 多图形联动下钻等

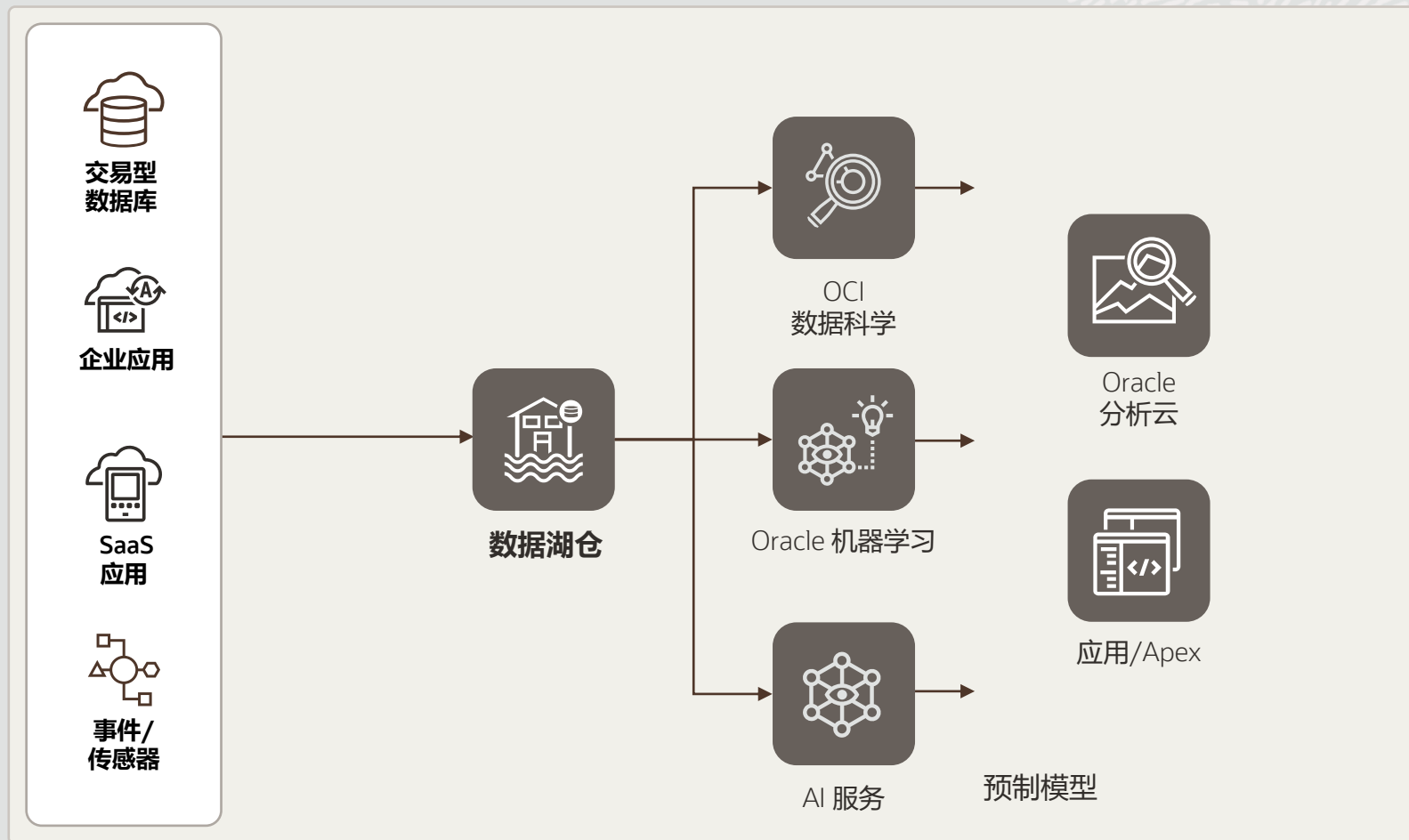
内置智能算法

开箱即用属性相关性分析和聚类, 趋势, 异常等算法, 也可通过数据流利用机器学习生成模型

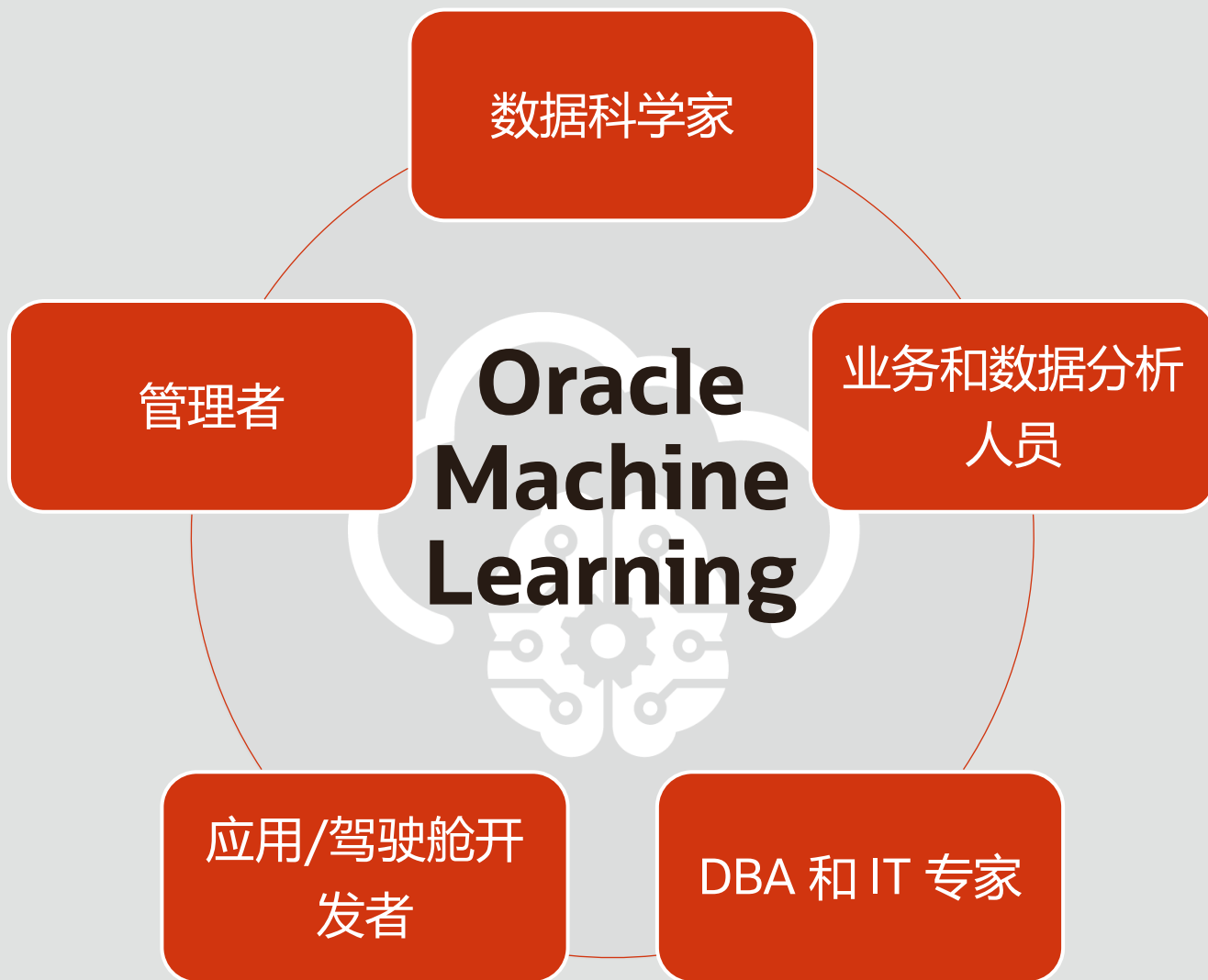
扩展和共享

支持图形扩展, 通过R和Python扩展算法, 支持和大数据云服务对接实现海量数据分析

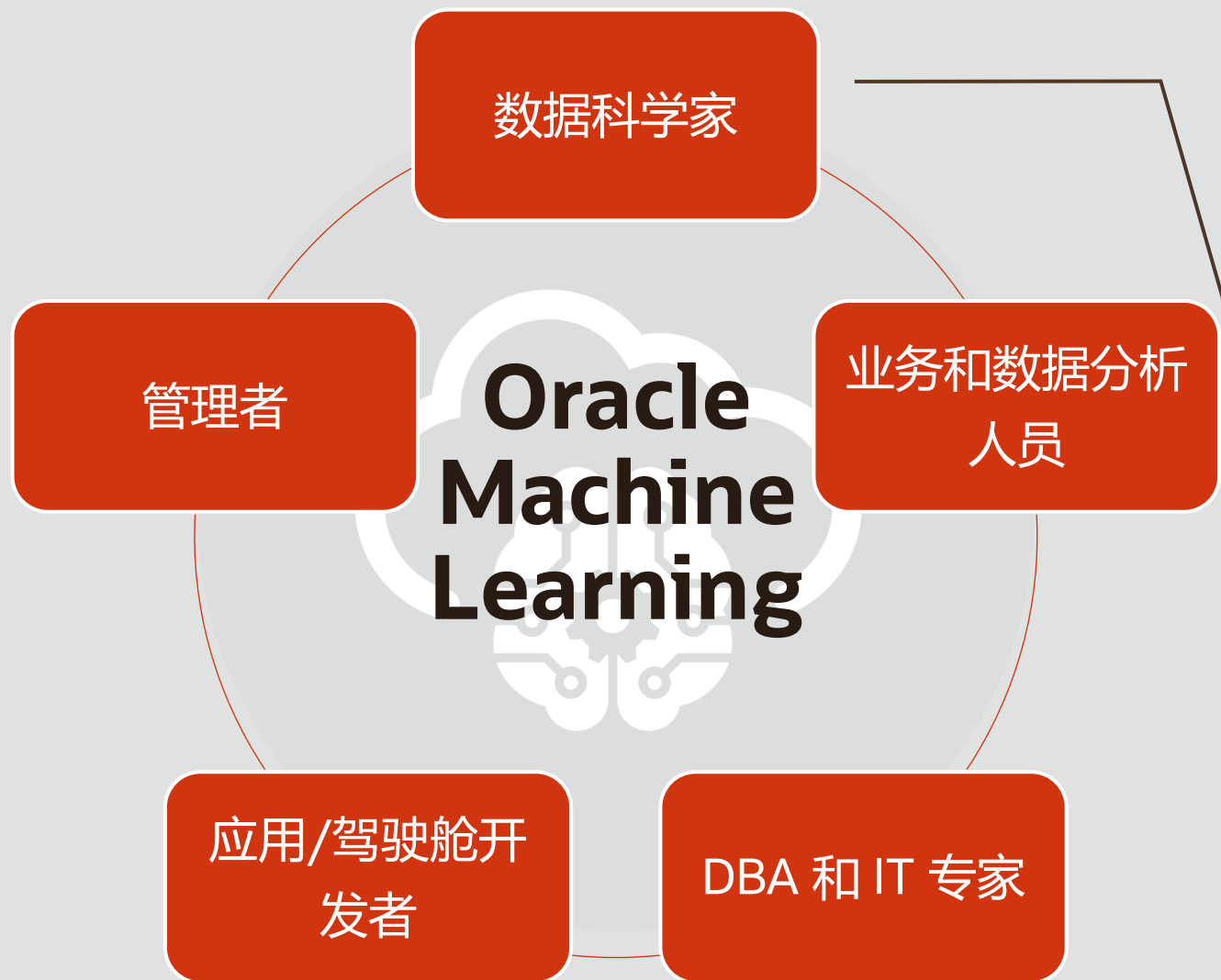
基于机器学习的增强决策



OML 支持企业不同角色用户

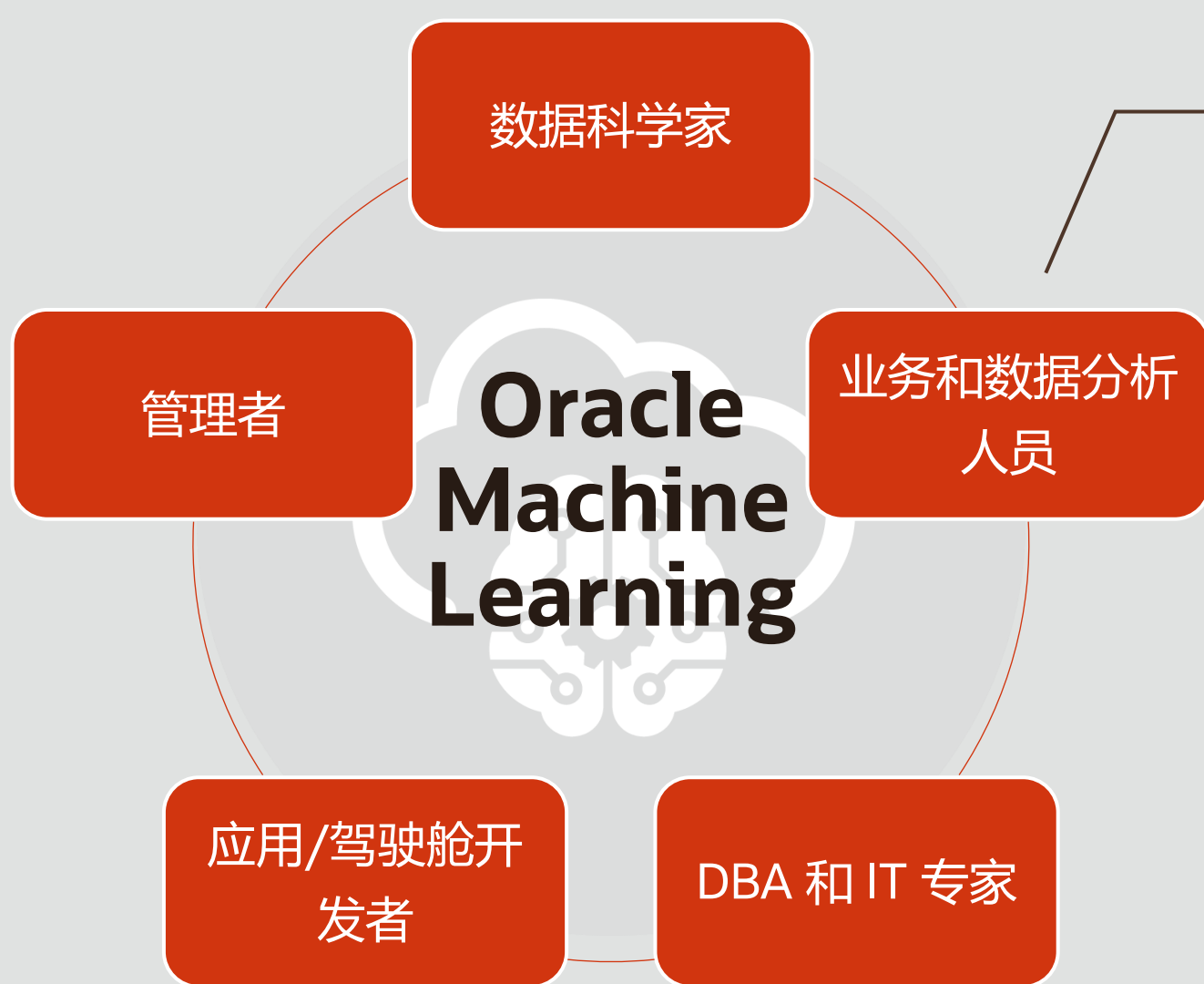


数据科学家



- 热门数据科学语言：Python、R、SQL
- 使用第三方包进行增强
- 可扩展性和性能
- 自动化提高的生产率
- 加强企业协作
- 整合和分析整个企业的数据

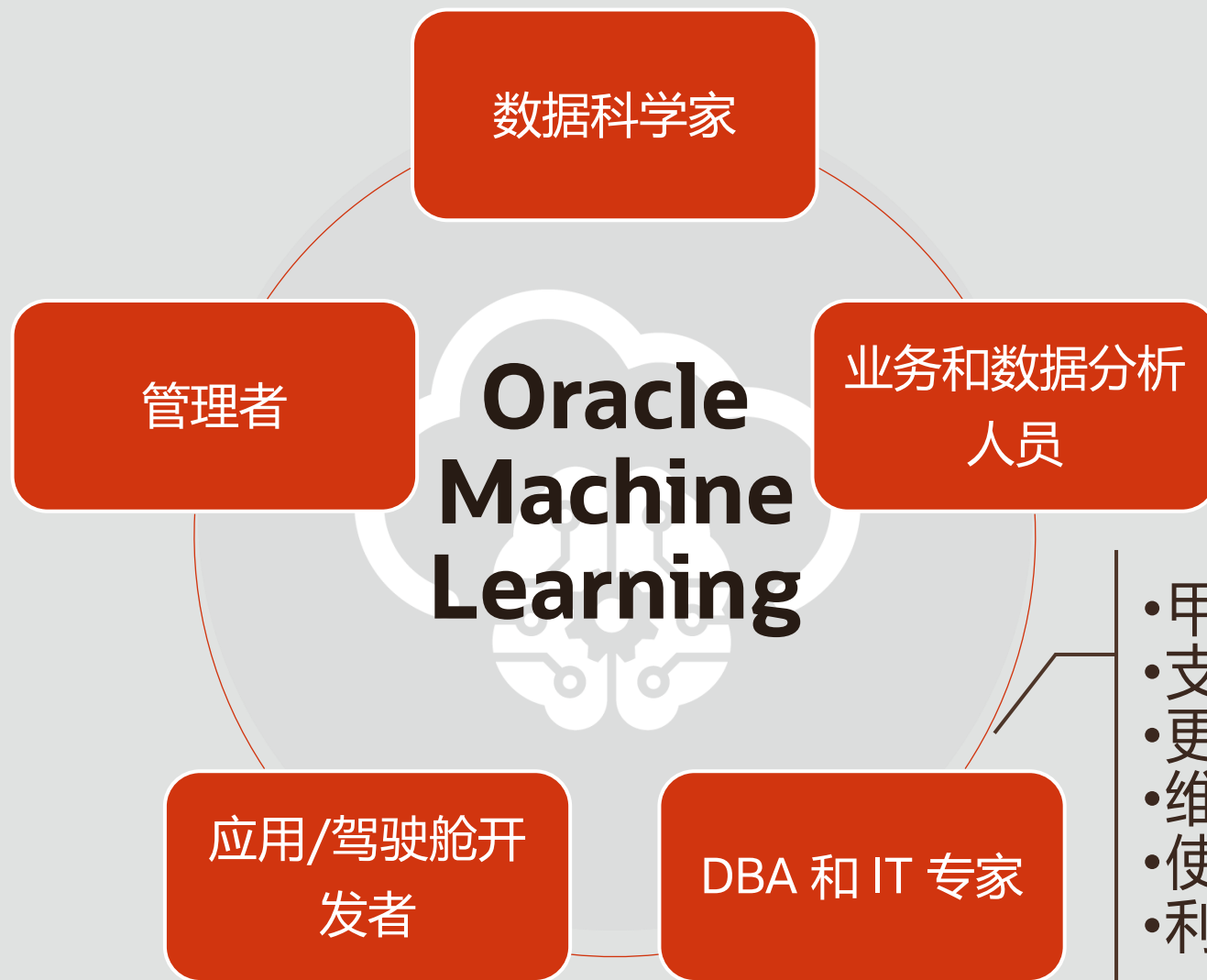
业务和数据分析人员



- 使用 ML 展开分析工具集
- 非 ML 专家可以使用 AutoML 快速启用
- 利用领域知识获得更好的结果
- 与数据科学家和 IT 协作



DBA 和 IT 专家

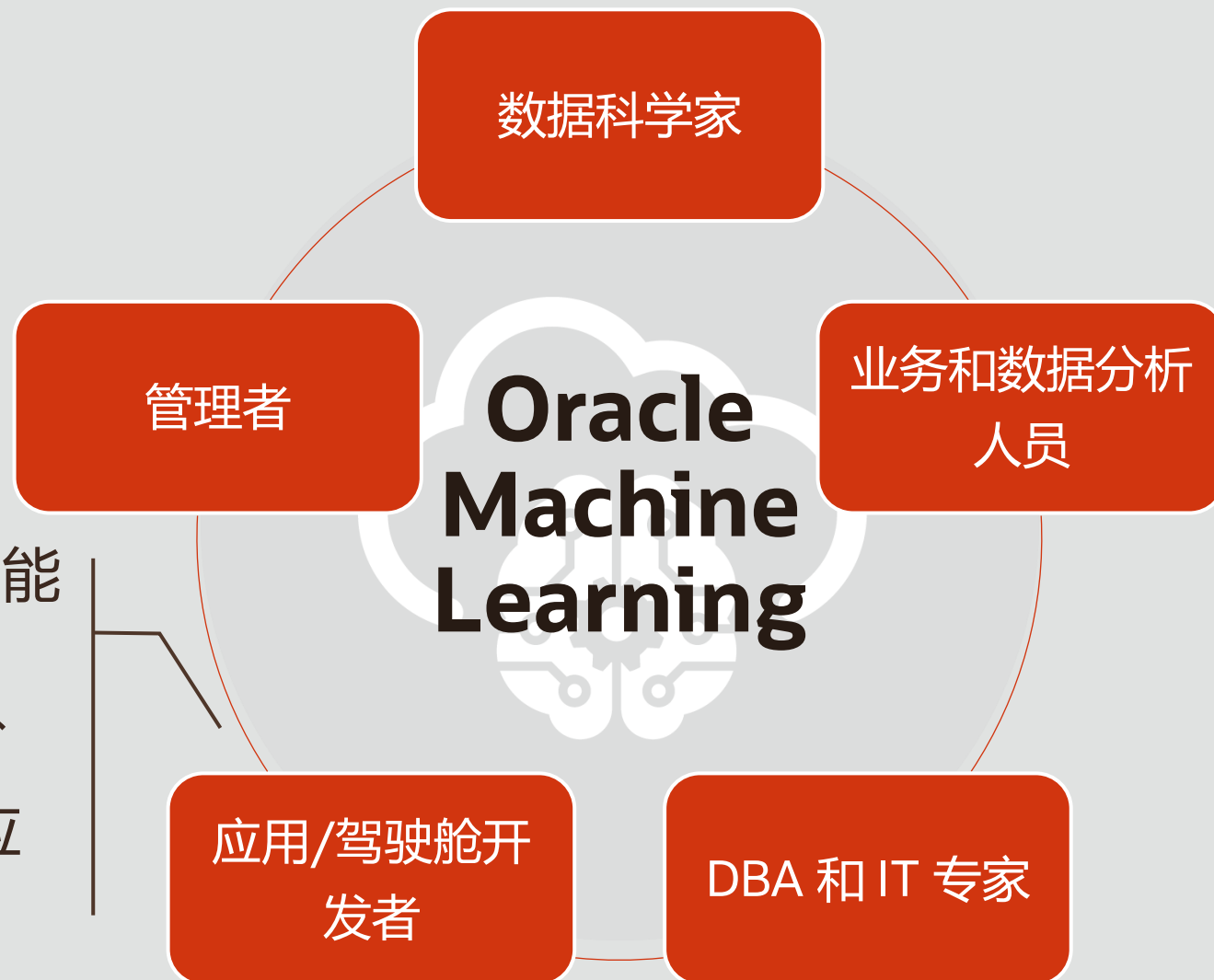


- 甲骨文投资带来更大的价值
- 支持可扩展性和性能
- 更简单、更精简的基础架构
- 维护数据安全、备份、恢复
- 使用 SQL, 扩展到 Python 和 R
- 利用数据库和大数据源

应用和驾驶舱开发人员

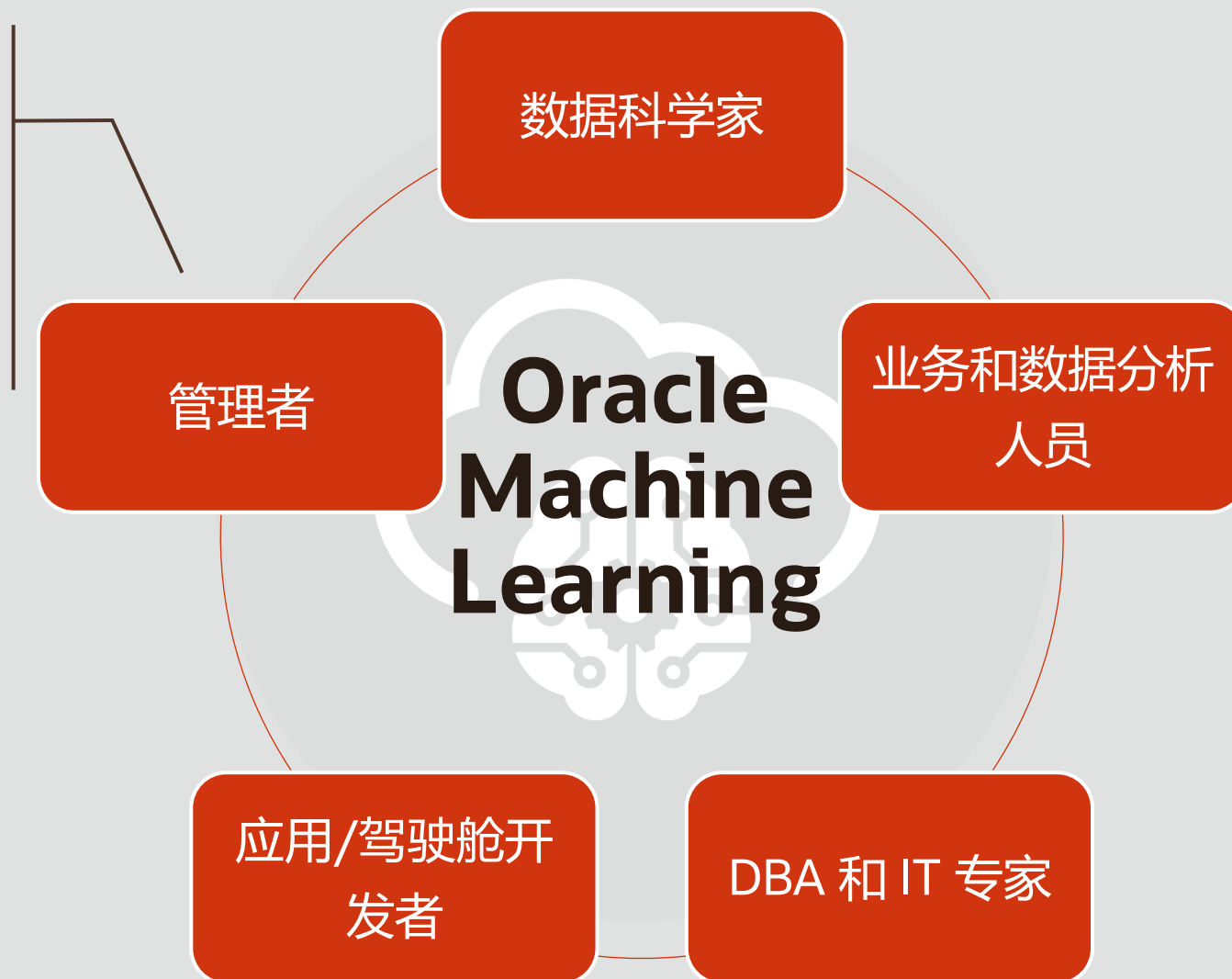


- 通过 Oracle 堆栈集成更快地实现智能解决方案
- 轻松使用数据科学家的 R、Python、SQL 脚本和快速部署解决方案
- 使用 SQL、REST 和 SODA API 在应用程序和仪表板中嵌入 ML



管理者

- 受益于世界一流的数据管理技术和支持
- 在整个企业中实现 ML 民主化, 以实现更好的数据驱动决策
- 更快地部署解决方案以实现 ROI



适合机器学习应用的业务场景

电信	消费品	金融服务	公共部门
客户分析 <ul style="list-style-type: none">客户保留（流失）客户细分、客户特征分析和行为分析产品组合分析（捆绑）客户交互优化系统异常检测客户特征优化	客户分析 <ul style="list-style-type: none">客户细分和特征分析市场组合分析客户交互优化购物篮和行为分析客户流失预警客户特征优化	客户和企业分析 <ul style="list-style-type: none">收入预测购物篮分析信用风险/策略分析（账龄分析）保险欺诈防范客户行外资产预测差异化定价	支出和公众分析 <ul style="list-style-type: none">支出预测欺诈识别安全/智能分析客户分析经济指标预测舆情分析
制造	医疗保健和生命科学	媒体	旅游服务
生产分析 <ul style="list-style-type: none">需求预测质量控制预测性资产维护收入优化成本费用分析采购分类人员流失分析	患者和企业分析 <ul style="list-style-type: none">患者疗效分析药物开发转化研究（个性化医疗）收入和需求预测分配和补货优化	广告分析 <ul style="list-style-type: none">广告优化营销优惠优化收入预测广告价格优化实时优先级分配	来宾分析 <ul style="list-style-type: none">细分、获取、保留、交互优化资产绩效分析娱乐场楼层布局优化收入预测和优化HR 欺诈防范分析

其它应用：自动驾驶，IT系统自动维护、图像/人脸识别、文字识别、NLP、语音识别等

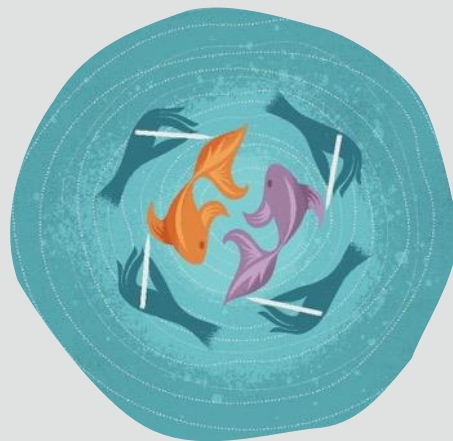
Oracle 机器学习 (Machine Learning)



自动化

- 以更少的投入，更快地获得更好的结果
- 公民智能——甚至是非专家人员

提升效率



扩展性

- 提供并行，分布式算法来处理海量数据
- 算法靠近数据，避免数据移动

更快实现业务目标



快速部署，面向数据

- 更快部署和更新数据科学解决方案
- 提供集成的机器学习平台

更多关注在创新

Thank You

