HUAWEI

迈向智能世界白皮书2024

云计算

云上跃迁，AI重塑千行万业

构建万物互联的智能世界

目录

01

趋势一：

构建AI原生基础设施，提供弹性、高效、多元的算力

02

趋势二：

构建以知识为中心的数据底座，为大模型高效供数

03

趋势三：

构建多模态多尺寸模型，实现场景与模型的最佳匹配

04

趋势四：

开放价值场景，使AI Agent成为企业AI落地的新方向

HUAWEI

01

趋势一：

构建AI原生基础设施，提供弹性、高效、多元的算力

HUAWEI

趋势一：构建AI原生基础设施，提供弹性、高效、多元的算力

联算比、容算比、存算比配比失衡，带来性能扩展、成本、效率全面挑战，AI基础设施架构需进行系统性变革

联算比 = 联接带宽/算力

Transformer	运算量	↑ 16.8倍 /2年
	内部连接带宽	↑ 1.4 倍/2年

容算比 = 内存容量/算力

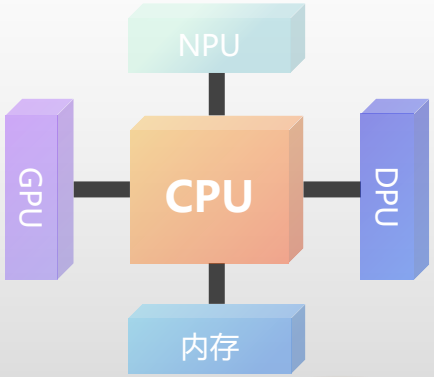
大模型	参数量	↑ 11.2 倍/2年
	内存容量	↑ 2 倍/2年

存算比 = 内存带宽/算力

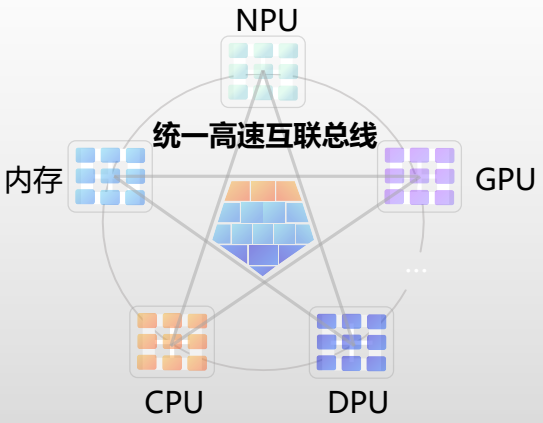
过去20年	硬件计算能力	↑ 60,000 倍
	内存带宽	↑ 100 倍

AI算力基础设施从单体算力到矩阵算力

基于主从架构的单体算力



基于全对等池化的矩阵算力

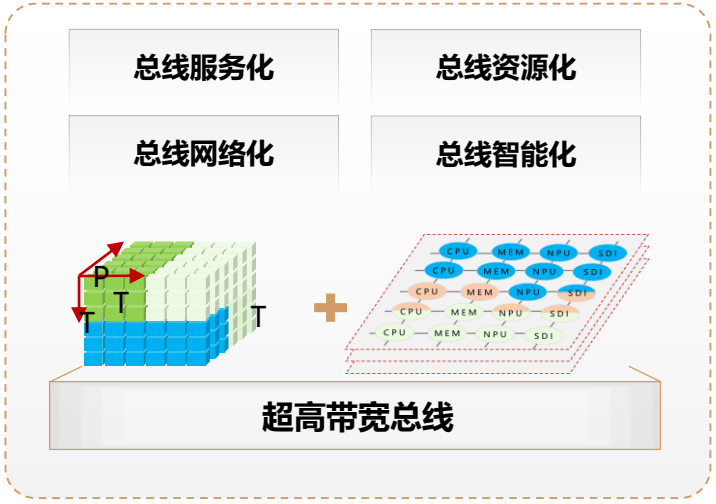


HUAWEI

行动建议：建设多样算力全对等、池化、可组合的新型AI算力基础设施

资源统一池化 / 突破算力边界

总线对等互联 / 突破网络边界



算力灵活组合 / 突破使用边界

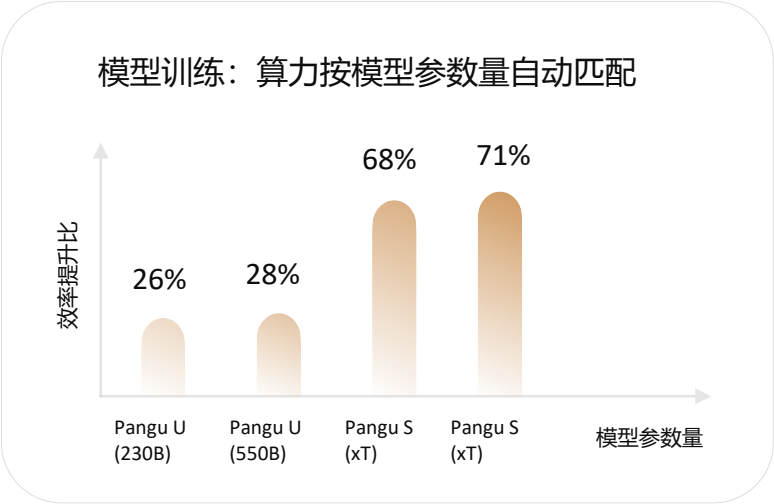
资源统一池化：
突破单服务器的能力边界，实现CPU、NPU、GPU、内存等多样资源统一池化。

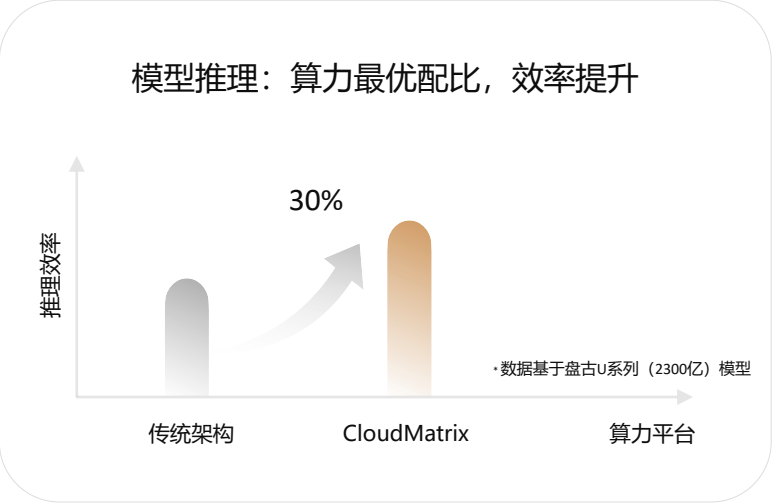
总线对等互联：
提供超高带宽Scale-up网络， 网络带宽对比传统网络提升一个数量级以上，达到 Tb级。

算力灵活组合：
通过调度系统匹配最优算力，覆盖百亿到万亿级模型训练推理场景，提供最优性能、最佳性价比的AI算力。

HUAWEI

行业实践：华为云数据中心系统性架构升级，模型训练及推理效率大幅提升





AI-Native云基础设施CloudMatrix

新一代的多元算力对等全互联架构

新型高速互联总线

HUAWEI

02

趋势二：

构建以知识为中心的数据底座，为大模型高效供数

HUAWEI

趋势二：构建以知识为中心的数据底座，为大模型高效供数

当前很多企业都建有数据仓库和数据湖用于企业商业智能，但面对AI大模型，当前数据平台还不能很好地为大模型使用，需要进行大量数据清洗、知识抽取，从而构建服务于AI模型的、以知识为中心的数据底座：

1、对数据湖的数据进行语义抽取，将数据与业务语义进行关联，转化为知识；

2、通过AI自动生成大模型所需要的问答对、图文对，能将业务知识高效供给大模型预训练和微调；

3、基于企业已积累的大规模数据，通过AI自动构建起企业的知识图谱。让业务人员可以通过企业大模型，用自然语言对话方式来快速实现各类任务。

HUAWEI

行动建议：三层架构为大模型高效供数

智能决策层

业务决策/数智应用

- 决策应用
- 自助可视分析应用
- 自助数智应用发布
- 数智应用市场

Other Apps

- 第三方开发应用

AI训练推理服务工具

- ModelArts
- Other AI/ML Platforms
- LLM/SLM

数智服务层

AI4Data Engine

- 智能搜索
- 数智助手
- 智能Workflow
- 智能洞察

Data4AI Engine (training/serving ready datasets)

- RAG GraphRAG
- 数据准备算子
- 数据集成算子
- 特征处理

DaaS Engine数据服务

- NL2SQL引擎
- AutoETL视图优化
- 查询优化
- 增强分析

MaaS Engine模型服务

- 模型服务
- 业务仿真
- 运维/运营优化
- AutoAI/ML

知识湖层

知识湖-知识服务层

- 动态本体
- 知识图谱
- 智能索引
- 知识管理&安全

知识湖-知识处理层

- 信息爬取与集成
- 知识定义
- 知识生成
- 知识融合

智能决策层

- **决策应用**：快速编排智能应用，应用构建效率提升10x↑
- **数智应用市场**：开发、发布原生数智应用，MarketPlace实现应用共享与交易；

数智服务层

- **DaaS Engine**：提供统一的智能数据服务能力，包括NL2SQL、AutoET、查询优化、数据见解等；10x↑
- **MaaS Engine**：提供统一的模型服务；
- **AI4Data Engine**：通过AI赋能数据治理、数据分析、数据处理；
- **Data4AI Engine**：提供高效数据准备能力，为AI模型训练提供高质量数据；

知识湖层

- **知识处理层**：实时挖掘数据特征、向量表征、基于规则与模型生成知识图谱、基于反馈自学习；效率相比人工10x↑
- **知识服务层**：知识统一检索服务、知识关系图谱服务、动态本体知识集成服务

HUAWEI

行业实践：助力某公司财经法务打造数智融合应用

业务痛点与解决方案

业务痛点	解决方案
业务决策逻辑来自专家头脑经验、IT系统无法感知与学习：	1、决策流开发：业务人员自助创建决策流水线，将专家经验沉淀到系统。
业务人员找数据难、理解数据难、IT人员投入大量人力数据开发、清洗；	1、定义业务对象-业务关系本体模型； 2、本体集成，自动发现业务系统相关数据，建立物理数据与业务本体映射关系，并且自动化数据集成； 3、通过AutoETL系统自动完成数据清洗转换
业务人员、IT人员彼此缺乏领域知识；	1、AI帮助下，法务人员自助创建决策流水线； 2、智能推荐仿真算法模型、流水线模板；
决策逻辑个性化，难以泛化、通用化，无法批量复制	1、执行决策流，执行what-if analysis 业务仿真； 2、自动从知识湖抽取知识，自动创建不同场景的变更流水线
缺乏业务影响分析	1、变更业务影响分析，支持NL问答 2、智能触发决策执行API，对接周边系统
缺乏业务监控、决策过程不可解释	1、决策链每次决策过程可视化、提供日志可解释 2、持续的业务运作优化

HUAWEI

03

趋势三：

构建多模态多尺寸模型，实现场景与模型的最佳匹配，解决复杂问题

HUAWEI

趋势三：构建多模态多尺寸模型，实现场景与模型的最佳匹配，解决复杂问题

生成式AI掀起了企业智能化升级的新一轮浪潮，大模型也开始走入千行万业，行业中的数据不仅存在文本和图片类型，更有大量的专业数值数据，例如钢铁行业的时序数据和工艺参数，气象行业的多源空间异构数据，医疗行业的医学影像和基因测序数据等等。多模态大模型可以将不同类型的数据通过预处理转化为统一的表示形式，然后结合多个模态的信息，进行联合建模和分析，提高模型的准确性和泛化能力，解决更为复杂的问题，在各行业各场景中有巨大潜力。

行业场景

金融	医疗	零售	制造
信贷风控	辅助诊断	营销文案	工艺优化
远程银行	药物研发	数字人	辅助设计
自动化客服	方剂优化	平台开发	排程排产
经营报告	文献生成	个性化营销	预测性维护

模态分类：文本、图像、视频、语音、数值数据

关键挑战

数据质量

- 高质量多模态数据集有限：视觉等其他模态数据的收集、标注成本较文本高，导致多模态数据集少得多；
- **MaaS Engine：AI合成数据有望改善数据枯竭问题：**但如何构造可控的，高质量合成数据仍充满挑战；

算法能力

- 多模态大模型算法和工程难度更大：在表征、对齐、推理、生成、迁移、量化等环节均面临更多难点；
- 模态融合和和协同：如何有效地融合不同模态的信息，同时保持各模态的独特性和互补性，是技术难题；

算力资源

- 多模态大模型对算力的要求高于单模态：一般而言，同样信息量情况下，文本数据量<图片数据量<视频数据量，多模态大模型需要处理的信息量更大；

可信安全

- 可解释性和透明度：多模态大模型往往复杂且不透明，提高其可解释性对于信任和应用生产非常关键；
- 安全和治理：需要确保多模态模型的应用符合伦理标准，防止潜在的滥用和风险；

HUAWEI

行动建议：夯实多维度行业数据，优先在高价值场景中实践多模态

关键洞察	行动建议
多模态丰富了人类与AI的交互形态，可以激发更具创造力更具价值的生成式AI应用	1、明确企业中哪些场景结合多个数据模态时能比使用单模态AI创造更高的业务价值 2、思考如何使用具有创造力的生产方式颠覆传统的作业方式
多模态数据面临质量和专业性挑战，企业同时也面临算力挑战	1、充分理解和评估多模态模型的技术复杂性，合理整合多种模态的输入和输出数据 2、通过强大的数据工程和数据治理，确保数据质量，识别并降低数据隐私和安全方面的风险 3、寻找可靠且经济的算力资源
多模态的落地是分阶段的，多模态之间的协同也需要不断磨合	1、从高容错度、高价值的场景入手，基于已有的多模态模型试点，验证其技术可行性和业务价值 2、对多模态的不同技术路线持开放态度，随时使用最适合的技术路线解决业务挑战

HUAWEI

行业实践：多模态识别能力帮助动车检测

大家所乘坐的每一辆高铁，都需要回到日检车间进行基础安全检测，全国每天需要检测的动车超过3万节。这个工作，现在需要大量的巡检工人在凌晨时段去完成，耗时耗力。

华为云盘古高铁大模型对海量的高铁行业数据进行预训练与精调，大模型强大的检测能力能覆盖所有检测项点，故障识别准确率可达到90%以上。在此基础上，采用国内首创的二维图片+三维点云+激光光谱的多模态融合诊断技术，能精准识别超限、异物、变形、丢失、破损、断裂、漏油、松动等各种复杂故障。多模态融合诊断比起单模态，故障识别准确率可提升到98%以上。最后，将行业经验与大模型结合，通过“高铁故障数据生成算法”，构造了多种罕见的故障形态，丰富了样本集，并通过数据反哺，边用边学，使得故障识别准确率进一步提高到99%以上。

BRI 北铁科

HUAWEI

分割与检测一切的能力

大模型开箱精调，
故障识别准确率即可达到
90%+

多模态融合诊断能力

二维图片+三维点云+激光光谱等多模态融合诊断技术，故障识别准确率可达到
98%+

行业故障数据生成算法

对高铁场景罕见故障样本生成，通过数据反哺，边用边学，故障识别准确率进一步达到
99%+

HUAWEI

行业实践：多模态可控时空生成帮助自动驾驶训练

华为云盘古大模型升级采用了创新的STCG技术，结合场景视频生成、4D BEV视频生成、自动驾驶仿真库及路网信息等，能很好的理解物理规律，可以基于多模态的视频生成技术，生产和实际场景相一致的驾驶数据，让我们灵活的增加控制条件生成不同路况、不同光照、不同天气的训练视频数据，加速自动驾驶技术的快速成熟。

增加和删除交通参与者

生成没有车辆的街道

增加一辆行驶车辆

增加多辆行驶车辆

融合感知与决策数据 支持端到端模型训练

交通参与者跨视角形态稳定 符合物理规律

HUAWEI 9

04

趋势四：

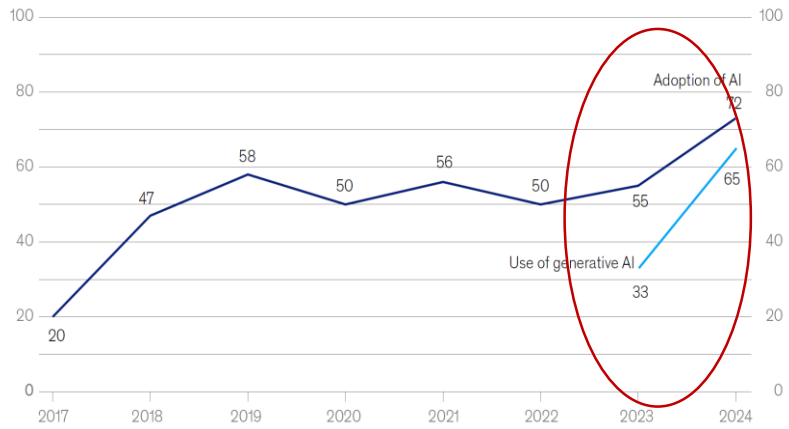
开放价值场景，使AI Agent成为企业AI落地的新方向

HUAWEI

趋势四：开放价值场景，使AI Agent成为企业AI落地的新方向

最新调查报告显示企业在最少一个业务功能采用AI的比例从2023年的55%大幅增大到2024年的72%

Organizations that have adopted AI in at least 1 business function，% of respondents



来源: McKinsey & Company

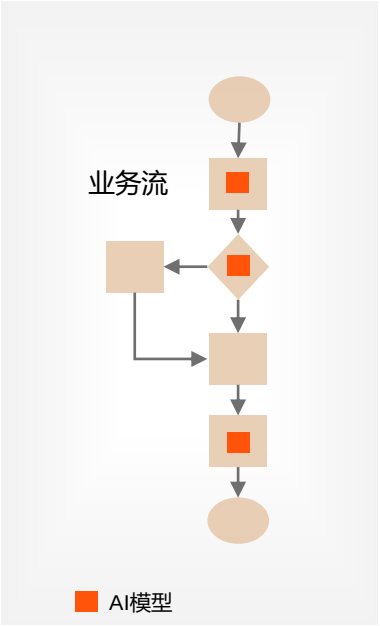
大模型基础能力持续提升以及AI Agent技术逐步成熟，使得

① 大模型越来越多地替代传统算法完成单点功能，比如CV、预测类

② 面向各行业各职能领域员工，基于角色的助手类应用成为新常态

③ 基于业务场景的AI Agent成为企业AI落地的新方向

① Embedded模式

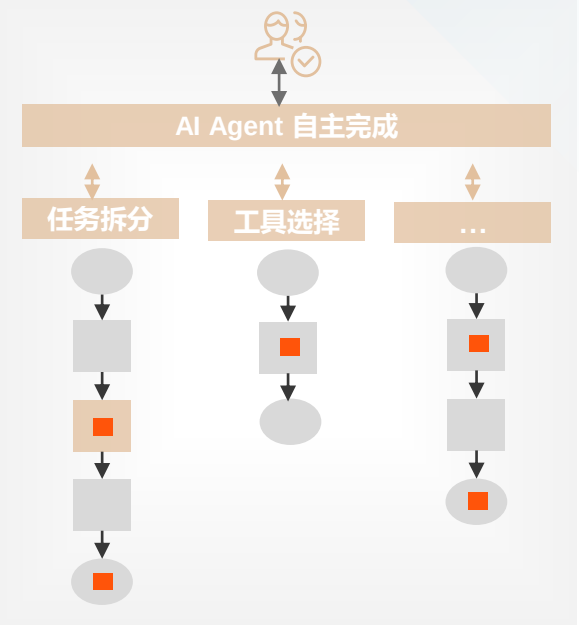


用大模型替代传统算法，完成单点功能

② Copilot模式

基于员工角色，用AI大模型构建助手应用或工具

③ AI Agent模式



人设定目标，由AI Agent自主完成业务场景闭环

HUAWEI

行动建议：通过Agent开发套件满足agent应用快速便捷开发需求

企业在开发和部署agent时可从以下几个方面进行设计：

1、明确业务应用和目标：在部署agent之前，企业应明确自身的业务需求，包括希望agent解决的问题、提升的效率、以及预期的业务价值；

2、选择合适的agent开发套件：根据企业的需求，选择支持灵活部署模式、易于集成现有系统、并提供强大AI能力的agent平台；

3、应用编排：结合agent开发套件提供的插件、知识库等功能，编排及配置agent应用；

4、采用分阶段的方法进行部署，从问答类AI Agent开始，逐步引入更复杂的应用，如客服助手、会议助手等，并在每个阶段进行性能评估和模型更新。

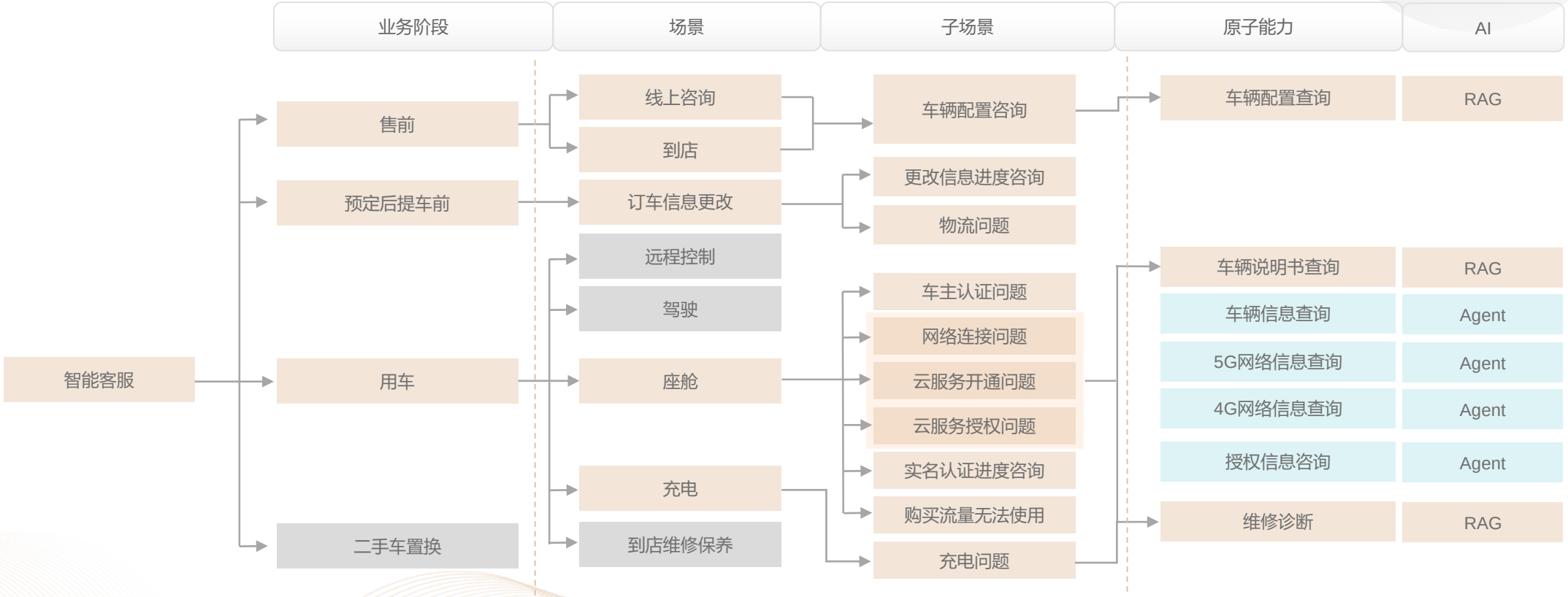
HUAWEI

行业实践：汽车行业智能客服Agent

通过Agent实现车主用户针对座舱、配置、维修诊断等相关领域的智能搜索、会话应答等能力，承担机器人客服角色。

1）座舱知识、车型配置知识的查询、应答。

2）配置类问题、维修诊断类问题，通过客服机器人自动应答。



Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

Copyright©2024 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI