光华管理学院 Guanghua School of Management

因思想 而光华

现代预测平台与选择
Forecasting Platforms and Selection

李丰

北京大学光华管理学院

https://feng.li/forecasting-with-ai

从模型到平台

From time series model to forecasting platform

从模型到平台

- 传统预测：单机 + 单模型（ARIMA, ETS）
- 现代预测：云端平台 + 自动化 + 并行计算
- “预测”已从算法问题 → 变为系统工程问题

预测平台的总体架构

- 数据层（Data Lake / Warehouse）
- 特征层（Feature Store）
- 模型层（Forecast Engine / AutoML）
- 服务层（API / Dashboard）
- 决策层（ERP / BI / SCM）

硬件体系概览

- 预测的规模化依赖计算与存储基础设施。
- 云计算平台： AWS, Azure, GCP
- 分布式计算框架： Spark, Ray, Dask
- GPU/TPU 加速： NVIDIA, Google
- 容器化与调度： Docker, Kubernetes

数据层硬件与软件

- 数据来源：ERP、CRM、POS、IoT
- 存储方式：数据湖（S3, GCS, HDFS）
- 计算引擎：SparkSQL, Presto, BigQuery
- 特征提取：Feast, Featuretools

计算层：分布式与弹性算力

- 批处理：Spark、Ray
- 流处理：Kafka、Flink
- 弹性扩展： Auto Scaling, Serverless
- GPU/TPU 加速的深度预测（TFT, Chronos）

软件栈概览

层	功能	示例工具
数据采集	ETL / Streaming	Kafka, Airflow, AWS Glue
分析计算	DataFrame/SQL	Spark, BigQuery
特征工程	自动化特征提取	PyCaret, Feast
模型训练	AutoML / Forecasting	Prophet, NeuralForecast, Chronos
部署与监控	API / MLOps	MLflow, FastAPI, Vertex AI

预测引擎的多样化

- 统计模型：ARIMA, ETS
- 机器学习模型：XGBoost, LightGBM
- 深度模型：TFT, N-BEATS, PatchTST
- 基础模型（Foundation Models）：Chronos, TimesFM

服务与部署层

- 部署方式：容器（Docker）+ 编排（Kubernetes）
- 持续集成与监控（CI/CD + MLOps）
- 预测服务接口：API、Dashboard、Webhook

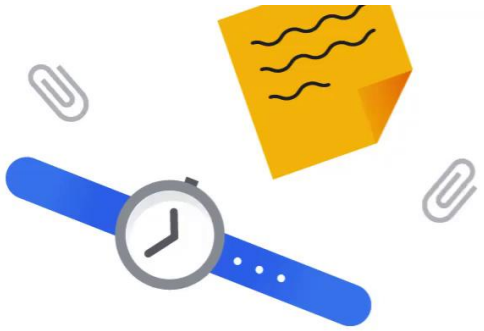
企业案例对比：系统架构视角

企业	系统名称	硬件/计算	软件栈	应用领域
Amazon	Forecast + SageMaker	EC2 GPU, S3	AutoML, API	零售/供应链
Google	Vertex AI Forecast	TPU, BigQuery	TFT, TimesFM	云服务
Uber	Michelangelo	Spark + Ray	MLflow, K8s	运营与需求
Meta	Prophet + PyTorch Infra	GPU + CPU	轻量模型部署	广告预测
Walmart	Blue Yonder + SAP IBP	Azure 云	SAP + LSTM	库存管理

案例： AWS Forecast 商业模式

- 用户可以使用APIs 在 Amazon Forecast 控制台导入时间序列数据集、训练预测变量和生成预测。
- Amazon Forecast 可自动执行大部分时间序列预测过程，使用者能够专注准备数据集和解释预测。
  - 提供了最佳的机器学习算法组合，自动执行复杂的机器学习任务。
  - 提供了从常用统计方法到复杂神经网络的各种训练算法。
  - 提供了多种自动处理数据集中缺失值的填充方法。
  - 可以自动合并内置数据集，改进模型。
- 使用 Amazon Forecast，您只需按实际用量付费。无最低费用，无预先承诺。
- Amazon Forecast 的成本取决于其所生成预测的数量、数据存储和训练时长。

Google Vertex AI 预测模式



Agenda

Source: https://www.youtube.com/watch?v=rzVR4Ioja6g

理解预测平台架构

Understanding Forecasting Platform Architecture

预测平台时代的人才需求结构

- 从“建模人才”到“系统型人才”
  - 传统阶段：重“预测模型师”（Statistical Modeler）
  - 平台阶段：更重“跨界融合者”（Hybrid Talent）
- 能理解数据、算法、系统与业务的“四懂型”人才
- 企业未来最稀缺的不是算法专家，而是能把预测嵌入业务流程、连接“技术–管理–决策”的桥梁型人才。

角色类型	核心能力	典型工具 / 平台
数据工程师	数据采集、清洗、ETL、建数仓	Spark, Airflow, AWS Glue
机器学习工程师	模型开发、调参、部署	SageMaker, Vertex AI, PyTorch
MLOps 工程师	模型运维、自动化管线	Docker, Kubernetes, MLflow
业务分析师	业务理解、模型解释与落地	Power BI, Tableau, Excel
AI 产品经理	统筹模型—业务接口	ERP, CRM, Forecasting APIs

为什么管理者需要理解预测平台架构

- 管理决策正在“平台化”
  - 企业预测已从 分析师手动建模 → 云端智能平台自动建模
  - 管理者不再只关注“结果”，而要理解 预测从数据到决策的路径

- 理解架构 = 掌握智能决策的底层逻辑

架构层	管理价值	典型问题
数据层	判断数据质量与时效	数据是否可信？是否能支撑决策？
计算层	评估资源与成本	算力够吗？投资云资源是否值得？
模型层	选择预测策略	何时用传统模型？何时用AI模型？
部署层	落地与反馈	如何将预测嵌入ERP/供应链？

新趋势与发展方向

- 预测的云原生化（Cloud-native Forecasting）
- 时间序列基础模型（Chronos / TimesFM）
- 低延迟分布式推理
- 跨系统协同预测（ERP + Forecasting）
- MLOps 与 可解释预测

企业预测平台的取舍

API vs. Self-Hosted LLMs: Trade-offs for Forecasting Platforms

使用 API（托管方案）

- 优势：
- ✅ 即开即用，维护成本低
- ✅ 模型更新快（直接享受最新版本）
- ✅ 推理速度与稳定性高（云端优化）
- ✅ 支持多语言与多模态输入（文本、图、表）

- 适用场景：
  - 管理分析类任务
  - 预测结果总结、情境分析、商业洞察生成
  - 原型验证（PoC）与教学场景
- 局限：
  - 成本随调用量线性增长（Token 计费）
  - 数据需上传云端，存在隐私与合规风险
  - 模型逻辑不可定制

自建部署（私有或混合云）

- 优势：
  - ✅ 数据安全可控（敏感预测数据不出本地）
  - ✅ 可针对业务词汇/行业语料微调
  - ✅ 成本在高频调用下更低（长期固定算力）
  - ✅ 可与内部系统（ERP、Forecast API）深度集成
- 适用场景：
  - 金融、能源、制造等高隐私预测业务
  - 企业内部知识库问答与模型解释系统
  - 需要长期定制与内部优化的预测流程
- 局限：
  - 初始部署复杂，需要 GPU、MLOps 团队
  - 模型迭代慢，更新成本高
  - 性能随硬件限制（推理延迟、显存瓶颈）

建议

- “预测平台的智能化，不在于模型更大，而在于能否在成本、隐私、可控性之间找到平衡。”
- 先用上再说，切勿因噎废食。

维度	推荐方案
数据安全性要求高	✅ 自建部署
预算有限	✅ API 调用
需要快速上线	✅ API 调用
需要深度定制与本地知识整合	✅ 自建部署
长期预测自动化系统	✅ 混合方案（API + 私有部署）

讨论环节

- 企业选择预测平台时，应优先考虑硬件、软件还是算法？
- 开源与商业平台的取舍标准是什么？
- 在你了解的行业中，预测平台落地的最大障碍是什么？