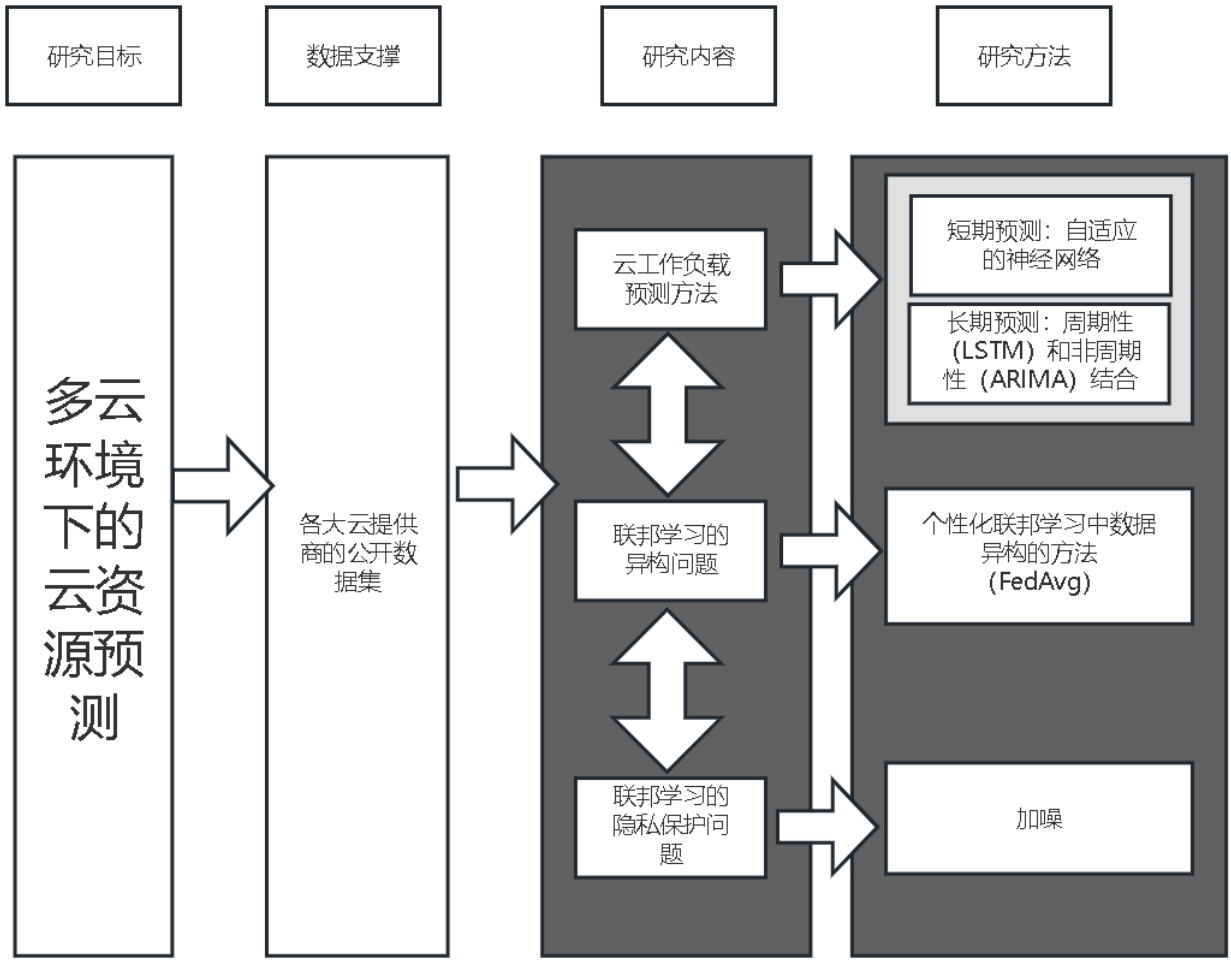
毕业论文构想版本库

主流：隐私保护和异构性的多云环境下的工作负载预测



单云环境下云资源预测的动机

云计算的背景

云工作负载的特性即挑战：高变换和高维度

真实云工作负载的形式特点（unknown a priori 不可知、dynamic fluctuation 动态波动）--> 缺乏清晰、稳定的趋势 or 季节性-->单一静态的预测不准确\时间序列模型可能不是最好的方法

云工作负载的动机

云资源调度 → auto-scaling → 云资源预测 → 云工作负载预测

云资源形式的复杂特点 / auto-scaling的reactive特质 → 云资源供应不足/过度问题 → 低成本效益、SLA违例

多云环境下云资源预测的动机

很难将视频精确地推荐给少数用户

Go to annotation“Due to the difference of user distribution in these two areas, the characteristics of data produced by them are quite different, and forming information islands between them. If each server makes video recommendation decisions merely based on its local data, it will bring information deviation, leading to nonprecise recommendation.”
由于这两个地区用户分布的差异，其产生的数据特征差异较大，形成了信息孤岛。如果每个服务器仅根据其本地数据进行视频推荐决策，会带来信息偏差，导致推荐不精准。

Go to annotation“Each cloud is in charge of collecting data from the areas it covers, leading to information islands among these clouds. In this case, the video recommendation results based on the local data performed by the single cloud will bring information deviations and inaccurate recommendation results for the minority of users.”
每个云负责从其覆盖的区域收集数据，导致这些云之间存在信息孤岛。在这种情况下，基于单一云端执行的本地数据的视频推荐结果会给少数用户带来信息偏差和不准确的推荐结果。

对于分布式性质的场景，集中式方法存在的缺点

Go to annotation“1) Single point of failure. The centralized server may fail due to attacks, which threatens the reliability of the system.
2) All data are required to send to the centralized server and process centrally, leading heavy cost and communication overhead for centralized system [13].
3) Considering the difference of data distribution, the centralized recommendation results cannot be representative of the preference of single local areas.”
1)单点故障。集中式服务器可能因遭受攻击而失效，威胁系统的可靠性。
2)所有数据都需要发送到集中式服务器并进行集中处理，导致集中式系统成本和通信开销较大[13];
3)考虑到数据分布的差异性，集中式推荐结果不能代表单个局部区域的偏好。

不同云提供商掌握的用户数据集具有不同的主要特征，但他们依然要为各自的少量用户提供服务（例如：A云提供商提供的服务主要用于机器学习，而B云提供商提供的服务主要用于部署web应用） ⟶ 在云际环境下创建预测模型，综合不同云提供商的用户数据训练更全面的模型。
因此全局数据和本地数据均是Non-IID分布的

⇓

目前现有的云工作负载预测在多云环境下的限制：

- 没有隐私保护，云提供商之间不愿意互相共享数据
- 没有考虑异构性，不同云提供商之间具有数据异构性和系统异构性

	数据异构	系统异构
联邦学习下的异构	non-iid 不同客户端数据分布\类型不一样	客户端的带宽、计算能力、内存、硬盘空间不一样
云际环境下的异构	用户特征不一样	服务器硬件条件不一样

⇓

采用联邦学习框架

数据集

云工作负载预测方法

数据预处理

将数据处理成时间序列

建模预测

短期学习：自适应的神经网络
长期学习：周期性（LSTM）与非周期性（ARIMA）结合

评估

联邦学习

在联邦学习的应用场景中，各个设备上的数据是由设备/用户独立产生的，不同设备/用户的非同源数据具有不同的分布特征，而每一个设备在进行本地学习的时候，所学习的训练数据是 Non-IID 的。因此研究提升 Non-IID 数据的学习效率，对于联邦学习具有重要意义。联邦学习允许用户在不需要集中存储数据的情况下，从本地存储的数据中共同获得共享模型的好

处。客户端的本地数据通常基于特定用户对移动设备的使用，因此任何特定用户的本地数据集都不能代表总体分布。主要研究数据Non-IID的异构性

异构性

设备/系统异构

数据/统计异构

FedAvg

[1602.05629] Communication-Efficient Learning of Deep Networks from Decentralized Data (arxiv.org)

处理异构性：

1. 在不同设备上的数据和标签分布都是IID时，FEDAVG已经被证明可以很好地逼近在集中收集的数据上训练的模型。但事实上，由于用户偏好和使用模式，每个设备上的数据是Non-IID的。
2. 每轮随机选择设备子集，随机选取的局部数据集可能无法反映全局视角下的真实数据分布，聚合这些发散的模型可以减慢收敛速度并大幅降低模型精度。
3. 允许设备进行E个epochs的本地训练，进行更多的本地计算。较大数量的局部历元可能会导致每个设备趋向于其局部最优，而局部最优可能与全局的目标相反，可能会损害收敛，甚至导致方法发散。E是个固定常数，而设备具有系统异构性，对于无法完成E个epochs的设备仅仅是将其删除。

减少联邦学习通信开销：

允许设备进行E个epochs的本地训练，通过平均来自客户端设备的模型权重，进一步减少通信轮数。而不是应用传统的梯度下降更新。

FAVOR

Optimizing Federated Learning on Non-IID Data with Reinforcement Learning | IEEE Conference Publication | IEEE Xplore

处理异构性：

每轮使用强化学习的代理选择一个设备子集。

减少联邦学习通信开销：

1. 客户端上传模型权重的更新给服务器。
2. 使用主层次分析压缩模型权重。

Federated Learning with Non-IID Data

[1806.00582v2] Federated Learning with Non-IID Data (arxiv.org)

处理异构性：

通过创建一个在所有边缘设备之间全局共享的数据子集来改进Non-IID数据的训练。但直接将全局数据分发到边缘客户端会带来很大的隐私泄露风险，这种方法需要在数据隐私保护和性

能改进之间进行权衡。此外，全局共享数据与用户本地数据的分布差异也会导致性能下降。且会带来额外的通信开销，依赖于大量的可用的公共数据。

FedProx

[1812.06127] Federated Optimization in Heterogeneous Networks (arxiv.org)
提出了FedProx，一个处理联邦网络中固有的系统和统计异构性的优化框架。FedProx允许在本地跨设备执行可变数量的工作，并且依赖于近端项来帮助稳定方法。
处理异构性：

1. 考虑到系统异构性，不对所有设备假设统一的局部epochs，聚合那些可能在FedAvg中被抛弃的设备的部分解。
2. 引入近端项，考虑全局和局部模型的不相似性，以限制局部更新的影响。

模型异构

FL+HC

Federated learning with hierarchical clustering of local updates to improve training on non-IID data | IEEE Conference Publication | IEEE Xplore
提出了另一种针对联邦学习的层次聚类框架。该方法适用于更广泛的非IID设置，并允许在每轮FL模型训练过程中对客户子集进行训练。分类算法的复杂度为$o(n^3)$（n为客户端的数量），但只用执行一次（除非添加新的设备），以降低计算和通信负载。

1. Transfer Learning
2. Model Distillation

隐私保护

1. 差分隐私添加噪声
2. 同态加密