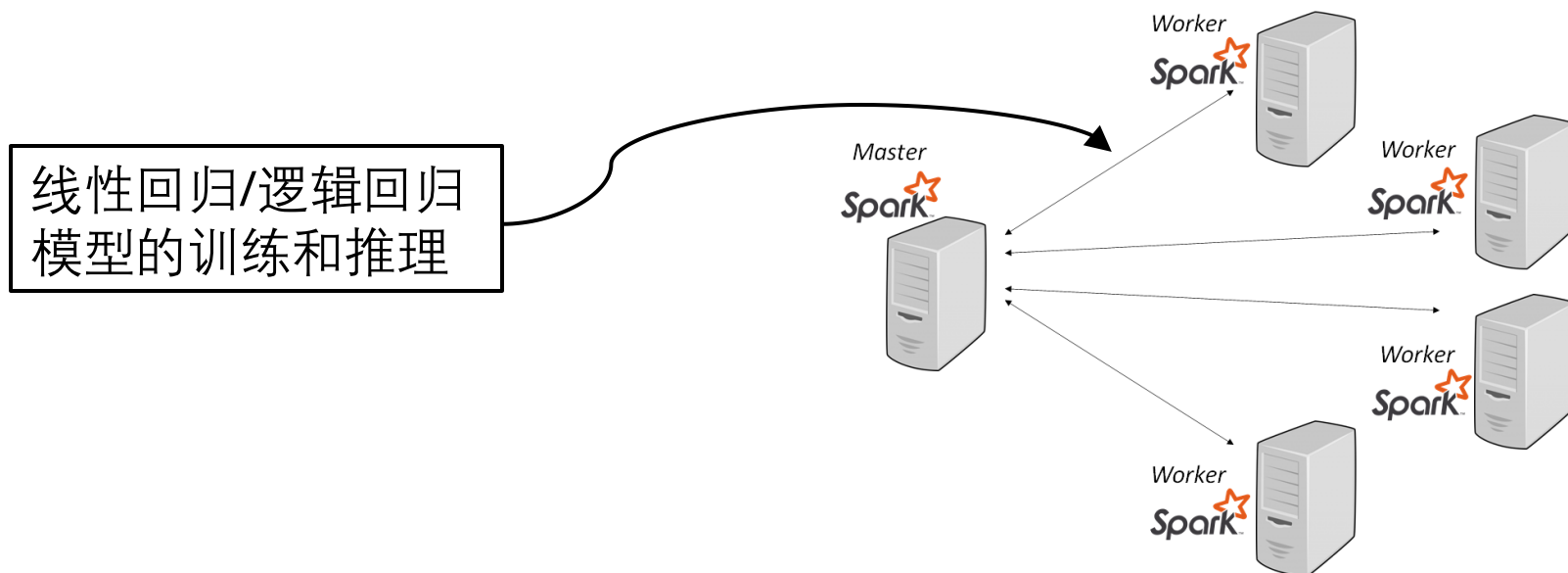


提交任务到Spark集群

实验任务

- 上节课我们已经完成了线性回归和逻辑回归在单个spark节点上训练和推理。
- 本节实验课请任选1个任务（线性回归或者逻辑回归），将其的代码提交到之前创建的集群中运行。



注意事项1

- 如果程序依赖除了代码以外的外部数据（如 CSV、TXT 文件），那么所有节点都必须能访问这些数据。通常可以通过分布式存储服务（如 HDFS）来确保数据在各节点间可见。所以首先，我们需要搭建一个分布式的存储系统。
-

上节课实验用到的数据在hdfs分布式存储系统内的地址为：

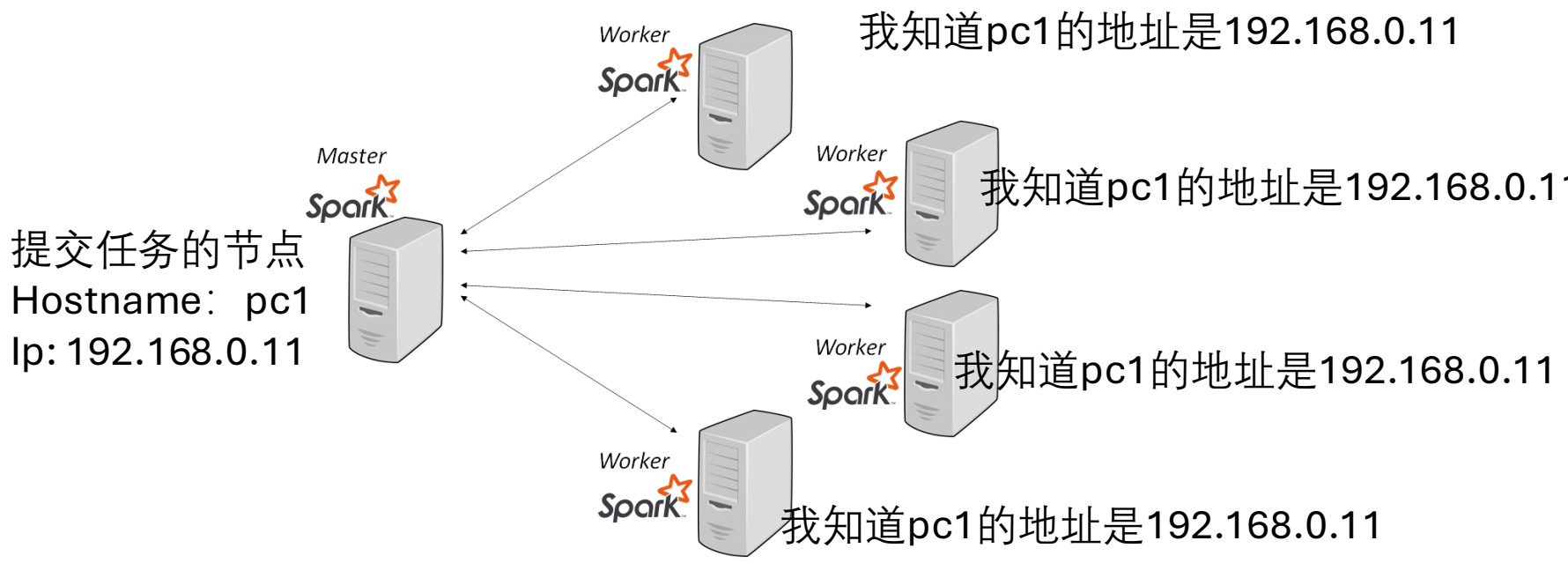
线性回归： `hdfs://nodes.pxymfang.com:9000/user/heyunan/cruise_ship_info.csv`

逻辑回归： `hdfs:// nodes.pxymfang.com :9000/user/heyunan/customer_churn.csv`

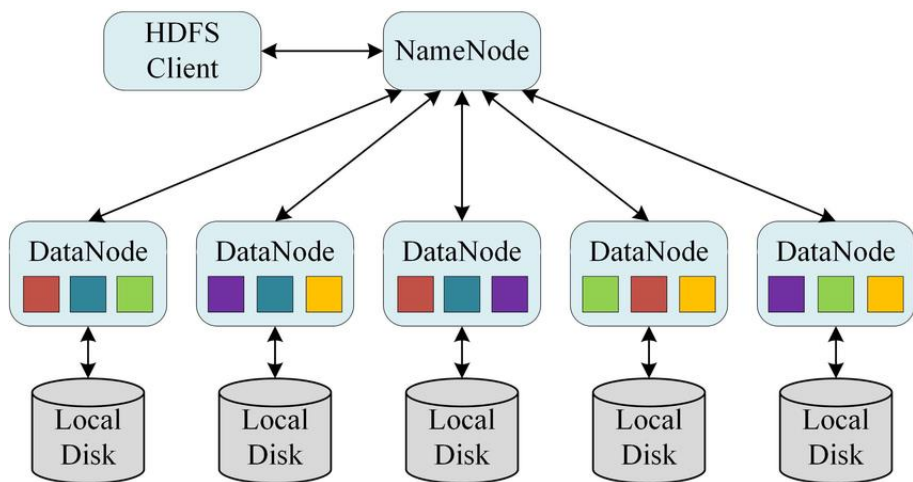
`hdfs:// nodes.pxymfang.com :9000/user/heyunan/new_customers.csv`

注意事项2

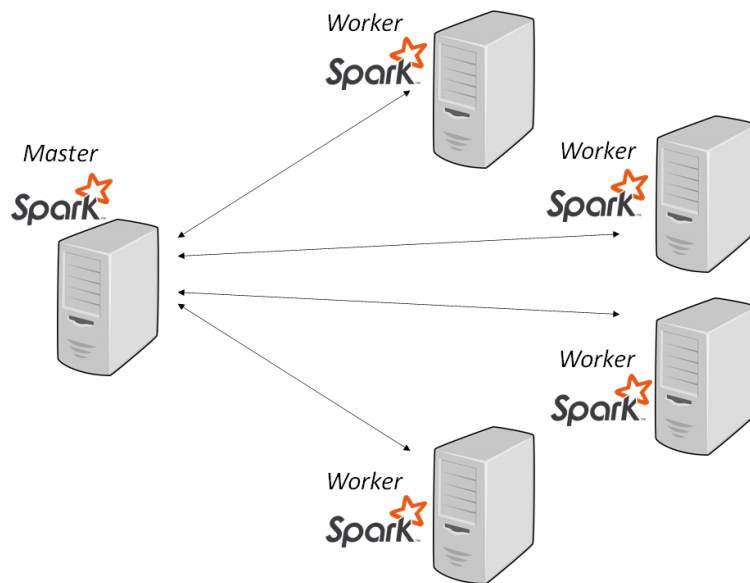
- 在集群任务中，不论你从哪个节点提交作业，**要求其它节点都能够正确解析提交作业节点的 hostname**。电脑的hostname可以在CMD或者PowerShell中输入“hostname”查询。其映射信息（映射关系指的是hostname与ip地址的映射关系）存储在C:\Windows\System32\drivers\etc\hosts 中。



分布式系统



数据在这里
(分布式存储)



运算在这里
(分布式运算)

提示

建议按照以下顺序完成实验任务

- 先尝试在单节点运行程序（与上节课不同的是使用的是在分布式存储系统中的数据），这一步是解决可以顺利从分布式存储系统中导入数据的问题。
- 然后将程序提交到集群。这一步是解决集群中节点可以相互通信，共同完成训练和推理任务的问题。

最终使用的命令类似于这样：

`spark-submit`

参数1指定master节点：`--master spark://192.168.0.177:7077`


参数2告诉 Spark，在访问 HDFS 时，底层的 Hadoop 文件系统客户端应该使用 DataNode 上报告的主机名，而不是默认的内部 IP 地址：`--conf`

`"spark.hadoop dfs.client.use.datanode.hostname=true"`

要处理的目标程序：`linear_regression.py`或者`logistic_regression.py`

提示

如果很长时间（比如运行时间超过了30秒）都没有结果，你就要打开Spark Master 的 Web UI（控制台）默认监听在 8080 端口（`http://<master-host>:8080`）

 3.5.3

Application: WordCount

ID: app-20250328104147-0017

Name: WordCount

User: heyunan

Cores: Unlimited (32 granted)

Executor Limit: Unlimited (1 granted)

Executor Memory - Default Resource Profile: 1024.0 MiB

Executor Resources - Default Resource Profile:

Submit Date: 2025/03/28 10:41:47

State: FINISHED

▼ Executor Summary (1)

ExecutorID	Worker	Cores	Memory	Resource Profile Id	Resources	State	Logs
------------	--------	-------	--------	---------------------	-----------	-------	------

▼ Removed Executors (1)

ExecutorID	Worker	Cores	Memory	Resource Profile Id	Resources	State	Logs
0	worker-20250317140003-192.168.0.236-38601	32	1024	0		KILLED	stdout stderr

标准输出
标准错误输出
可以在这里查找原因