第5讲 上手实践PSI

查询未被使用的端口

Alice与Bob初始化，同时运行

生成spu

创建数据集

求交

查询未被使用的端口

```python
import socket
from contextlib import closing
from typing import cast

def unused_tcp_port() -> int:
    with closing(socket.socket(socket.AF_INET, socket.SOCK_STREAM)) as sock:
        sock.bind(("", 0))
        sock.setsockopt(socket.SOL_SOCKET, socket.SO_REUSEADDR, 1)
        return cast(int, sock.getsockname()[1])

print(unused_tcp_port())
```

The execution took in 333ms, finished at 11:05 晚上.

bob 的输出

```
42107
```

alice 的输出

```
55893
```

Alice与Bob初始化，同时运行

```python
import secretflow as sf

cluster_config = {
    'parties': {
        'alice':{
            'address': '172.16.0.69:55241',
            'listen_addr':'0.0.0.0:55241'
        },
        'bob':{
            'address': '172.16.0.70:45607',
            'listen_addr':'0.0.0.0:45607'
        },
    },
    'self_party':'alice'
}

sf.shutdown()
sf.init(address='local',cluster_config=cluster_config)
```

The execution took in 9s 494ms, finished at 11:06 晚上.

```
2024-11-29 15:05:53.824 INFO api.py:342 [alice] -- [Anonymous_job] Shutdowning rayfed intendedly...
2024-11-29 15:05:53.825 INFO api.py:356 [alice] -- [Anonymous_job] Wait for data sending.
2024-11-29 15:05:53.829 INFO message_queue.py:72 [alice] -- [Anonymous_job] Notify message polling thread[DataSendingQueueThread] to exit.
2024-11-29 15:05:53.929 INFO message_queue.py:102 [alice] -- [Anonymous_job] The message polling thread[DataSendingQueueThread] was exited.
2024-11-29 15:05:53.930 INFO message_queue.py:72 [alice] -- [Anonymous_job] Notify message polling thread[ErrorSendingQueueThread] to exit.
2024-11-29 15:05:53.976 INFO message_queue.py:102 [alice] -- [Anonymous_job] The message polling thread[ErrorSendingQueueThread] was exited.
2024-11-29 15:05:53.977 INFO barriers.py:469 [alice] -- [Anonymous_job] Stop sender proxy actor...
2024-11-29 15:05:53.980 INFO barriers.py:473 [alice] -- [Anonymous_job] Sender proxy actor stoped.
2024-11-29 15:05:53.980 INFO api.py:384 [alice] -- [Anonymous_job] Shutdowned rayfed.
/usr/local/lib/python3.10/subprocess.py:1796: RuntimeWarning: os.fork() was called. os.fork() is incompatible with multithreaded code, and JAX is multithreaded, so this will likely lead to a deadlock.
  self.pid = _posixsubprocess.fork_exec(
2024-11-29 15:05:57,630 WARNING services.py:1996 -- WARNING: The object store is using /tmp instead of /dev/shm because /dev/shm has only 67108864 bytes available. This will harm performance! You may be able to free up space by deleting files in /dev/shm. If you are inside a Docker container, you can increase /dev/shm size by passing '--shm-size=0.96gb' to 'docker run' (or add it to the run_options list in a Ray cluster config). Make sure to set this to more than 30% of available RAM.
2024-11-29 15:05:57,778 INFO worker.py:1724 -- Started a local Ray instance.
2024-11-29 15:05:59.402 INFO api.py:233 [alice] -- [Anonymous_job] Started rayfed with {'CLUSTER_ADDRESSES': {'alice': '0.0.0.0:55241', 'bob': '172.16.0.70:45607'}, 'CURRENT_PARTY_NAME': 'alice', 'TLS_CONFIG': {}}
2024-11-29 15:06:00.372 INFO barriers.py:284 [alice] -- [Anonymous_job] Succeeded to create receiver proxy actor.
(ReceiverProxyActor pid=3055) 2024-11-29 15:06:00.369 INFO grpc_proxy.py:359 [alice] -- [Anonymous_job] ReceiverProxy binding port 55241, options: ((('grpc.enable_retries', 1), ('grpc.so_reuseport', 0), ('grpc.max_send_message_length', 524288000), ('grpc.max_receive_message_length', 524288000), ('grpc.service_config', '{"methodConfig": [{"name": [{"service": "GrpcService"}], "retryPolicy": {"maxAttempts": 5, "initialBackoff": "5s", "maxBackoff": "30s", "backoffMultiplier": 2, "retryableStatusCodes": ["UNAVAILABLE"]}}]}')))...
(ReceiverProxyActor pid=3055) 2024-11-29 15:06:00.371 INFO grpc_proxy.py:379 [alice] -- [Anonymous_job] Successfully start Grpc service without credentials.
2024-11-29 15:06:01.246 INFO barriers.py:333 [alice] -- [Anonymous_job] SenderProxyActor has successfully created.
2024-11-29 15:06:01.247 INFO barriers.py:520 [alice] -- [Anonymous_job] Try ping ['bob'] at 0 attemp, up to 3600 attemps.
```

```python
import secretflow as sf

cluster_config = {
    'parties': {
        'alice':{
            'address': '172.16.0.69:55241',
            'listen_addr':'0.0.0.0:55241'
        },
        'bob':{
            'address': '172.16.0.70:45607',
            'listen_addr':'0.0.0.0:45607'
        },
    },
    'self_party':'bob'
}

sf.shutdown()
sf.init(address='local',cluster_config=cluster_config)
```

The execution took in 9s 888ms, finished at 11:06 晚上.

```
2024-11-29 15:05:53.828 INFO api.py:342 [bob] -- [Anonymous_job] Shutdowning rayfed intendedly...
2024-11-29 15:05:53.828 INFO api.py:356 [bob] -- [Anonymous_job] Wait for data sending.
2024-11-29 15:05:53.830 INFO message_queue.py:72 [bob] -- [Anonymous_job] Notify message polling thread[DataSendingQueueThread] to exit.
2024-11-29 15:05:53.908 INFO message_queue.py:102 [bob] -- [Anonymous_job] The message polling thread[DataSendingQueueThread] was exited.
2024-11-29 15:05:53.909 INFO message_queue.py:72 [bob] -- [Anonymous_job] Notify message polling thread[ErrorSendingQueueThread] to exit.
2024-11-29 15:05:53.911 INFO message_queue.py:102 [bob] -- [Anonymous_job] The message polling thread[ErrorSendingQueueThread] was exited.
2024-11-29 15:05:53.912 INFO barriers.py:469 [bob] -- [Anonymous_job] Stop sender proxy actor...
2024-11-29 15:05:53.914 INFO barriers.py:473 [bob] -- [Anonymous_job] Sender proxy actor stoped.
2024-11-29 15:05:53.915 INFO api.py:384 [bob] -- [Anonymous_job] Shutdowned rayfed.
/usr/local/lib/python3.10/subprocess.py:1796: RuntimeWarning: os.fork() was called. os.fork() is incompatible with multithreaded code, and JAX is multithreaded, so this will likely lead to a deadlock.
  self.pid = _posixsubprocess.fork_exec(
2024-11-29 15:05:57,700 WARNING services.py:1996 -- WARNING: The object store is using /tmp instead of /dev/shm because /dev/shm has only 67108864 bytes available. This will harm performance! You may be able to free up space by deleting files in /dev/shm. If you are inside a Docker container, you can increase /dev/shm size by passing '--shm-size=1.01gb' to 'docker run' (or add it to the run_options list in a Ray cluster config). Make sure to set this to more than 30% of available RAM.
2024-11-29 15:05:57,789 INFO worker.py:1724 -- Started a local Ray instance.
2024-11-29 15:05:59.467 INFO api.py:233 [bob] -- [Anonymous_job] Started rayfed with {'CLUSTER_ADDRESSES': {'alice': '172.16.0.69:55241', 'bob': '0.0.0.0:45607'}, 'CURRENT_PARTY_NAME': 'bob', 'TLS_CONFIG': {}}
2024-11-29 15:06:00.420 INFO barriers.py:284 [bob] -- [Anonymous_job] Succeeded to create receiver proxy actor.
(ReceiverProxyActor pid=2918) 2024-11-29 15:06:00.417 INFO grpc_proxy.py:359 [bob] -- [Anonymous_job] ReceiverProxy binding port 45607, options: (('grpc.enable_retries', 1), ('grpc.so_reuseport', 0), ('grpc.max_send_message_length', 524288000), ('grpc.max_receive_message_length', 524288000), ('grpc.service_config', '{"methodConfig": [{"name": [{"service": "GrpcService"}], "retryPolicy": {"maxAttempts": 5, "initialBackoff": "5s", "maxBackoff": "30s", "backoffMultiplier": 2, "retryableStatusCodes": ["UNAVAILABLE"]}}]}'))...
(ReceiverProxyActor pid=2918) 2024-11-29 15:06:00.419 INFO grpc_proxy.py:379 [bob] -- [Anonymous_job] Successfully start Grpc service without credentials.
2024-11-29 15:06:01.314 INFO barriers.py:333 [bob] -- [Anonymous_job] SenderProxyActor has successfully created.
2024-11-29 15:06:01.315 INFO barriers.py:520 [bob] -- [Anonymous_job] Try ping ['alice'] at 0 attemp, up to 3600 attemps.
```

生成spu

```
import spu

cluster_def = {
    "nodes": [
        {
            "party": "alice",
            "address":"172.16.0.69:38979",
        },
        {
            "party":"bob",
            "address":"172.16.0.70:43115",
        },
    ],
    "runtime_config":{
        "protocol":spu.spu_pb2.SEMI2K,
        "field":spu.spu_pb2.FM128,
    },
}

spu = sf.SPU(
    cluster_def,
    link_desc={
        "connect_retry_times": 60,
        "connect_retry_interval_ms": 1000,
    },
)
```

The execution took in 101ms, finished at 11:06 晚上.

创建数据集

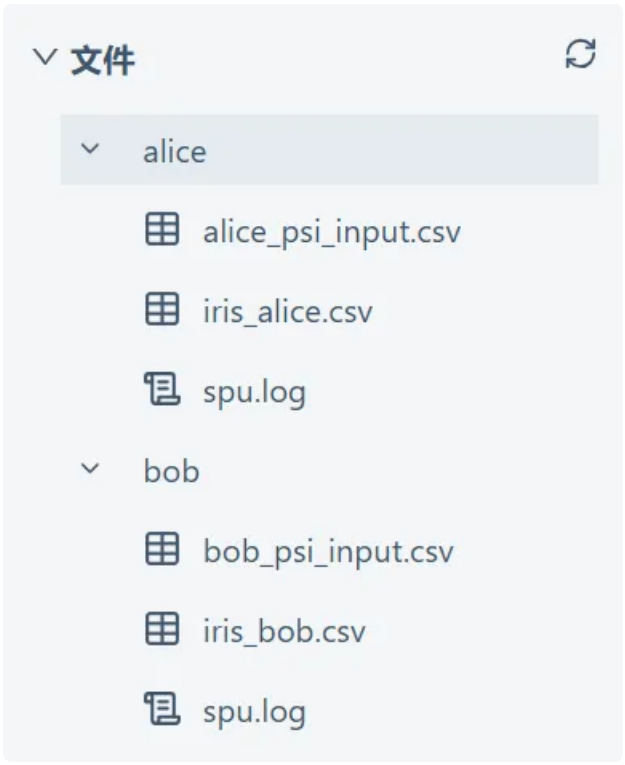
分别在alice和bob节点上运行

```python
import pandas as pd

alice_df = pd.DataFrame({
    "name": ["mike","tony","john","kate"],
    "age":[15,13,20,21]
})

import os
current_dir = os.getcwd()

alice_df.to_csv(f"{current_dir}/alice_psi_input.csv", index=False)
```

The execution took in 007ms, finished at 11:06 晚上.

```python
import pandas as pd

bob_df = pd.DataFrame({
    "name": ["alice","tony","john","kitty"],
    "sex":['F','M','M','F']
})

import os
current_dir = os.getcwd()

bob_df.to_csv(f"{current_dir}/bob_psi_input.csv", index=False)
```

The execution took in 007ms, finished at 11:06 晚上.



求交

```python
import os
current_dir = os.getcwd()

spu.psi(
    keys={"alice":["name"],"bob":["name"]},
    input_path={"alice":f"{current_dir}/alice_psi_input.csv", "bob":f"{current_dir}/bob_psi_input.csv"},
    output_path={"alice":f"{current_dir}/alice_psi_output.csv", "bob":f"{current_dir}/bob_psi_output.csv"},
    receiver="alice",
    broadcast_result=True,
    protocol="PROTOCOL_ECDH",
    ecdh_curve="CURVE_25519"
)


```

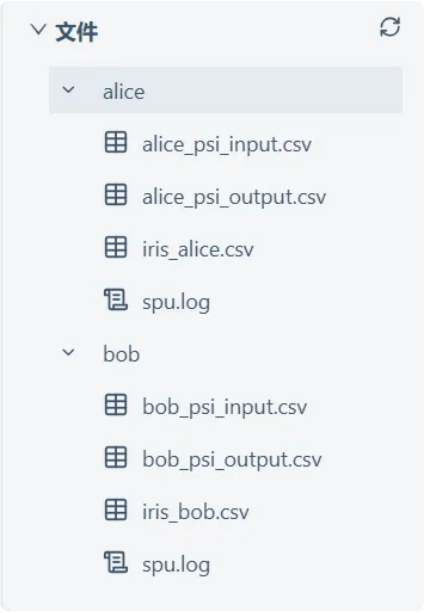
The execution took in 395ms, finished at 11:08 晚上.

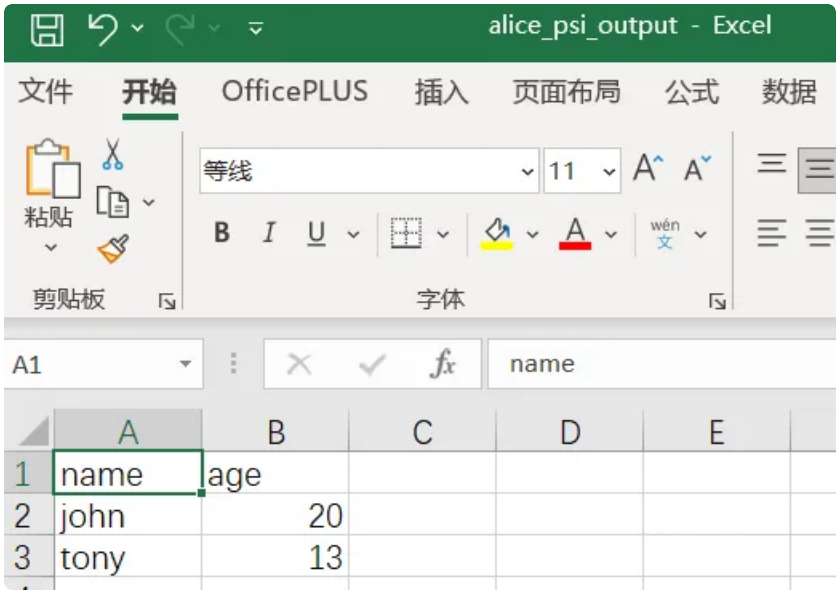
alice 的输出

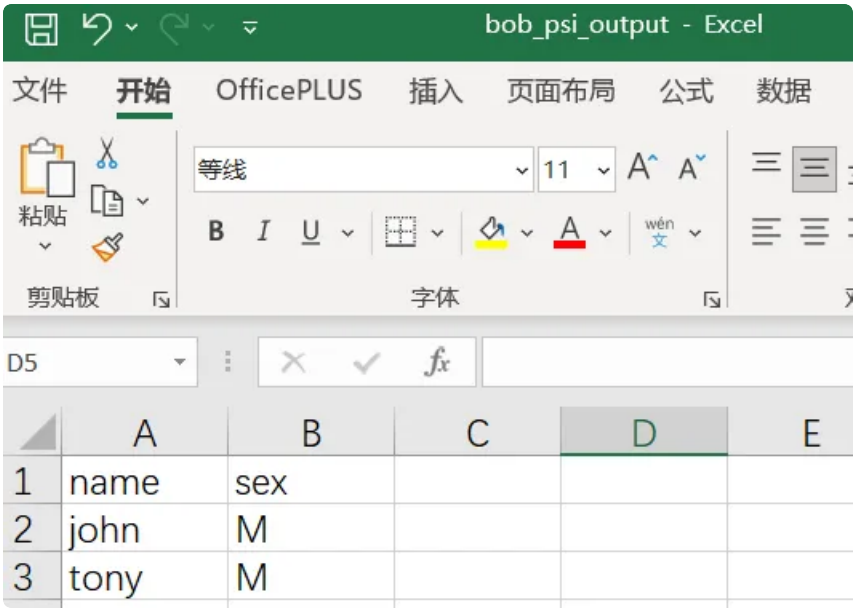
```
[{'party': 'alice', 'original_count': 4, 'intersection_count': 2},
 {'party': 'bob', 'original_count': 4, 'intersection_count': 2}]
```

bob 的输出

```
(SPURuntime(device_id=None, party=bob) pid=2989) [2024-11-29 15:09:13.433] [info] [launch.cc:119] PSI config: {"protocol_config":{"protocol":"PROTOCOL_ECDH","role":"ROLE_SENDER","broadcast_result":true,"ecdh_config":{"curve":"CURVE_25519"}},"input_config":
{"type":"IO_TYPE_FILE_CSV","path":"/home/secretnote/workspace/bob_psi_input.csv"},"output_config":{"type":"IO_TYPE_FILE_CSV","path":"/home/secretnote/workspace/bob_psi_output.csv"},"keys":["name"],"left_side":"ROLE_RECEIVER"}
(SPURuntime(device_id=None, party=bob) pid=2989) [2024-11-29 15:09:13.433] [info] [sender.cc:43] [EcdhPsiSender::Init] start
(SPURuntime(device_id=None, party=bob) pid=2989) [2024-11-29 15:09:13.433] [info] [interface.cc:78] [AbstractPsiParty::Init] start

(SPURuntime(device_id=None, party=bob) pid=2989) [460.395]       perfetto.cc:45899 Configured tracing session 5, #sources:1, duration:0 ms, #buffers:1, total buffer size:1024 KB, total sessions:1, uid:0 session name: ""

(SPURuntime(device_id=None, party=bob) pid=2989) [2024-11-29 15:09:13.764] [info] [interface.cc:136] [AbstractPsiParty::Init][Check csv pre-process] start
(SPURuntime(device_id=None, party=bob) pid=2989) [2024-11-29 15:09:13.765] [info] [csv_checker.cc:243] Executing script to get duplicates: LC_ALL=C tail -n +2 /tmp/71c5894e-ed9b-4bfe-bf20-dd61c9e924d8.psi_checked | LC_ALL=C sort --parallel=2 --buffer-size=1
--stable | LC_ALL=C uniq -d > /tmp/71c5894e-ed9b-4bfe-bf20-dd61c9e924d8.psi_checked_duplicates

[{'party': 'alice', 'original_count': 4, 'intersection_count': 2},
 {'party': 'bob', 'original_count': 4, 'intersection_count': 2}]
```


文件
alice
alice_psi_input.csv
alice_psi_output.csv
iris_alice.csv
spu.log
bob
bob_psi_input.csv
bob_psi_output.csv
iris_bob.csv
spu.log


alice_psi_output - Excel
文件 开始 OfficePLUS 插入 页面布局 公式 数据
粘贴
剪贴板
等线 11
字体
A1
name
A B C D E
1 name age
2 john 20
3 tony 13


bob_psi_output - Excel
文件 开始 OfficePLUS 插入 页面布局 公式 数据
粘贴
剪贴板
等线 11
字体
D5
A B C D E
1 name sex
2 john M
3 tony M