Apache Flink

TensorFlow与Apache Flink 的结合与实践

陈戊超 阿里巴巴 / 技术专家

Apache Flink China Meetup 上海
- 2019年09月07日

Apache Flink

Contents

PART 01

Background

Background

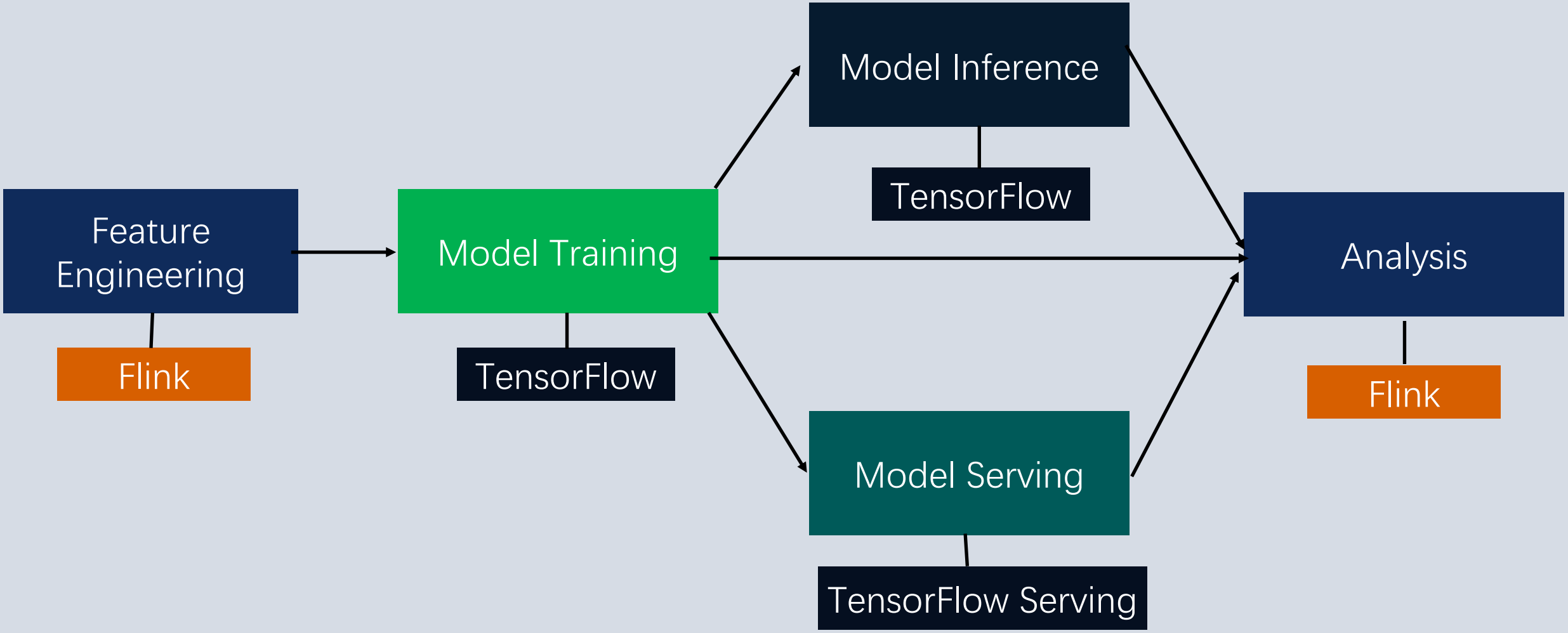
Apache Flink

TensorFlow is an open source software library for numerical computation using data flow graphs and is the most popular **AI computing framework**.

Flink is a framework and distributed processing engine for stateful computations over unbounded and bounded data streams。**Flink is widely used in data processing and Feature Engineering**

Machine learning workflow

Apache Flink



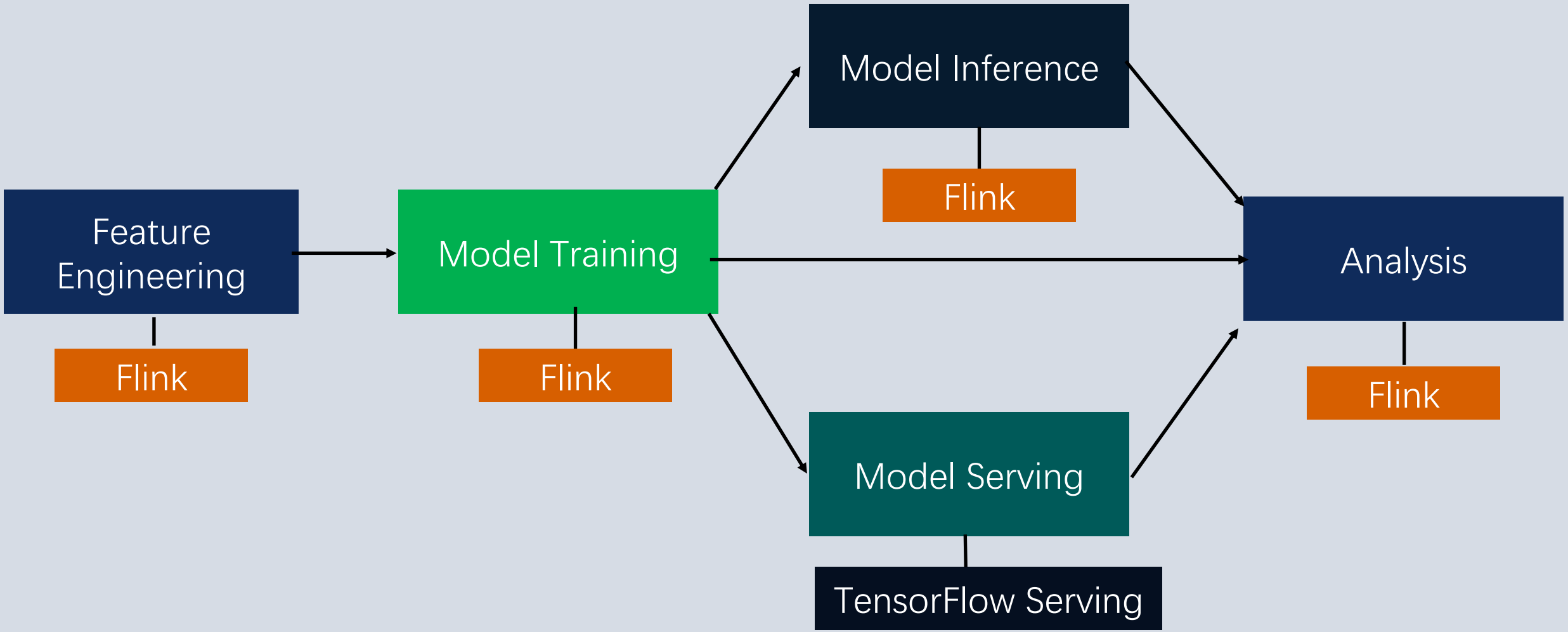
Problems

Apache Flink

- *Users do feature engineering, model training and model prediction with **two framework**.*
- *Distributed programs often run in clusters but **it's not friendly** to use TensorFlow for distributed training to determine IP and port first.*
- *TensorFlow Distributed Running Can't **Failover Automatically**。*

Goal

Apache Flink

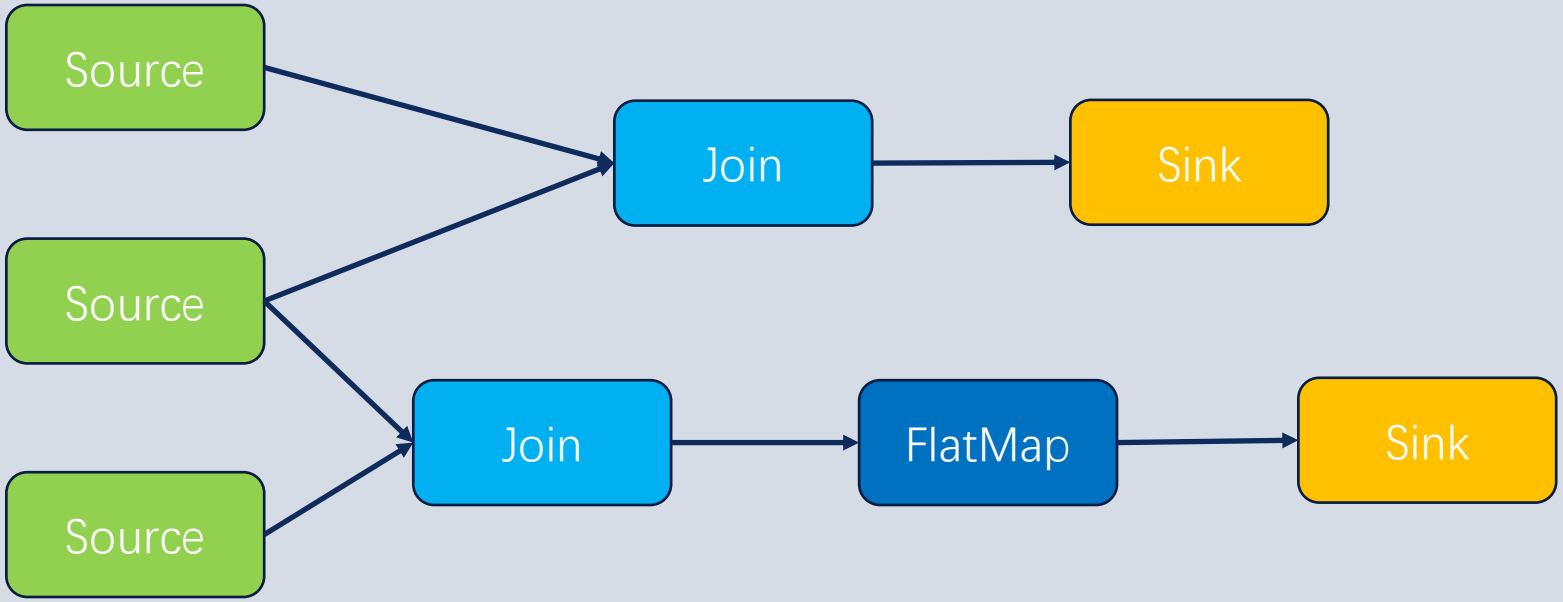


PART 02

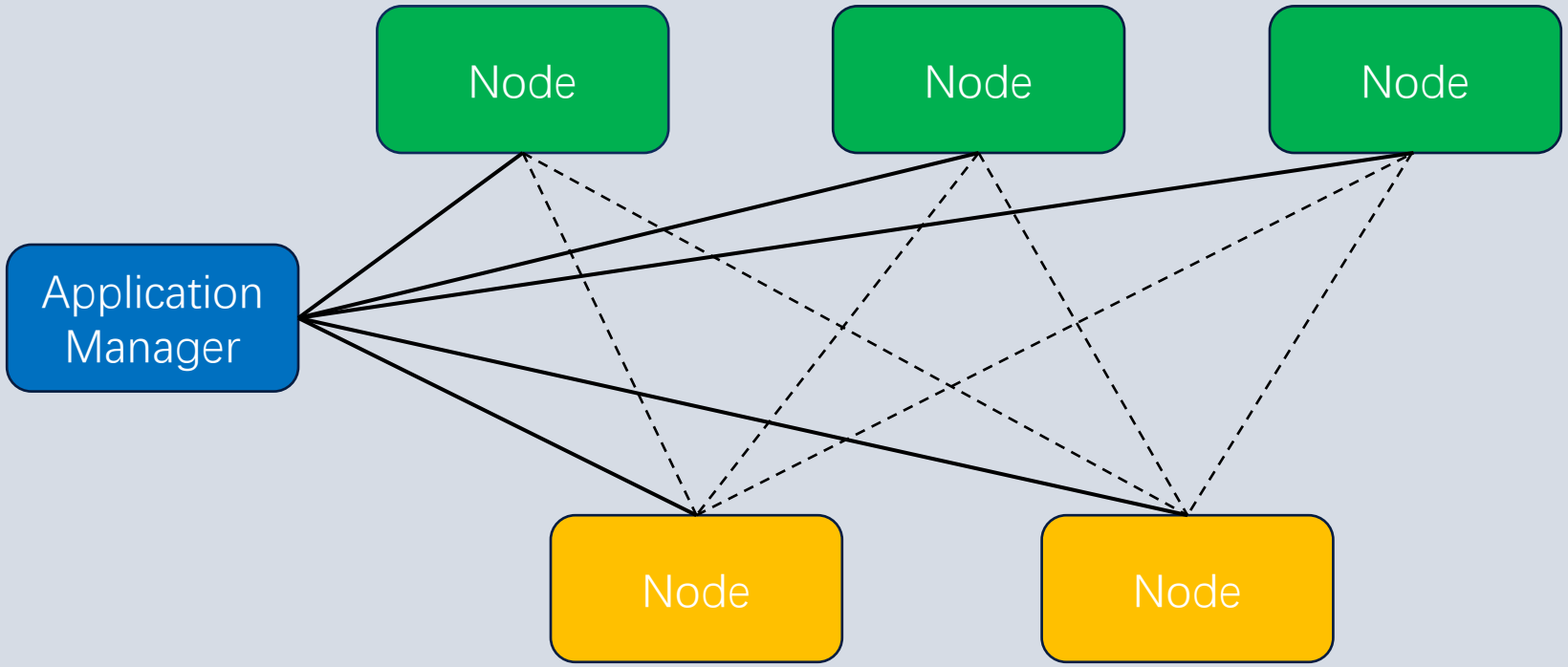
Machine Learning On Flink

Flink

Apache Flink

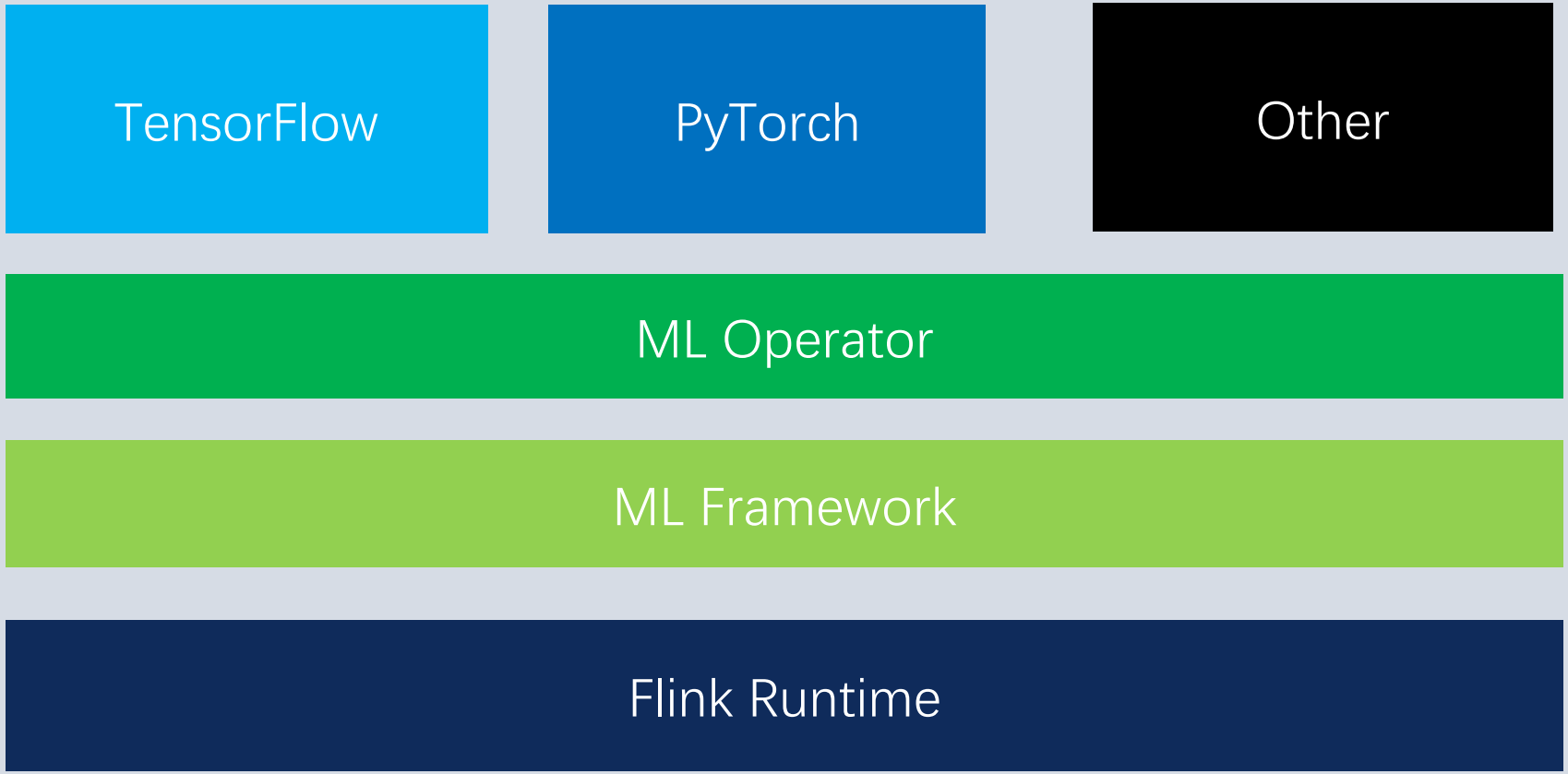


Machine learning cluster



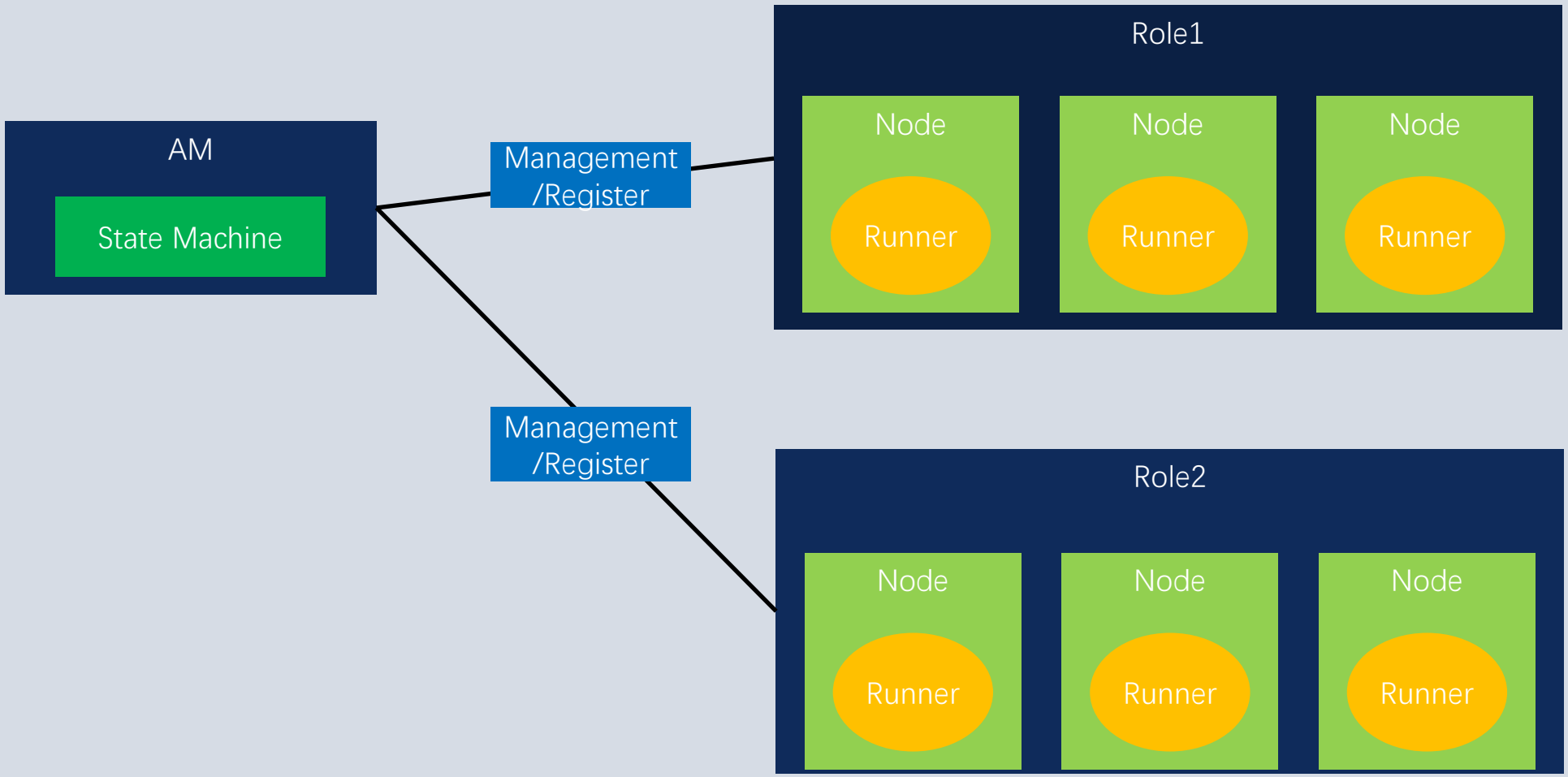
Machine Learning On Flink

Apache Flink



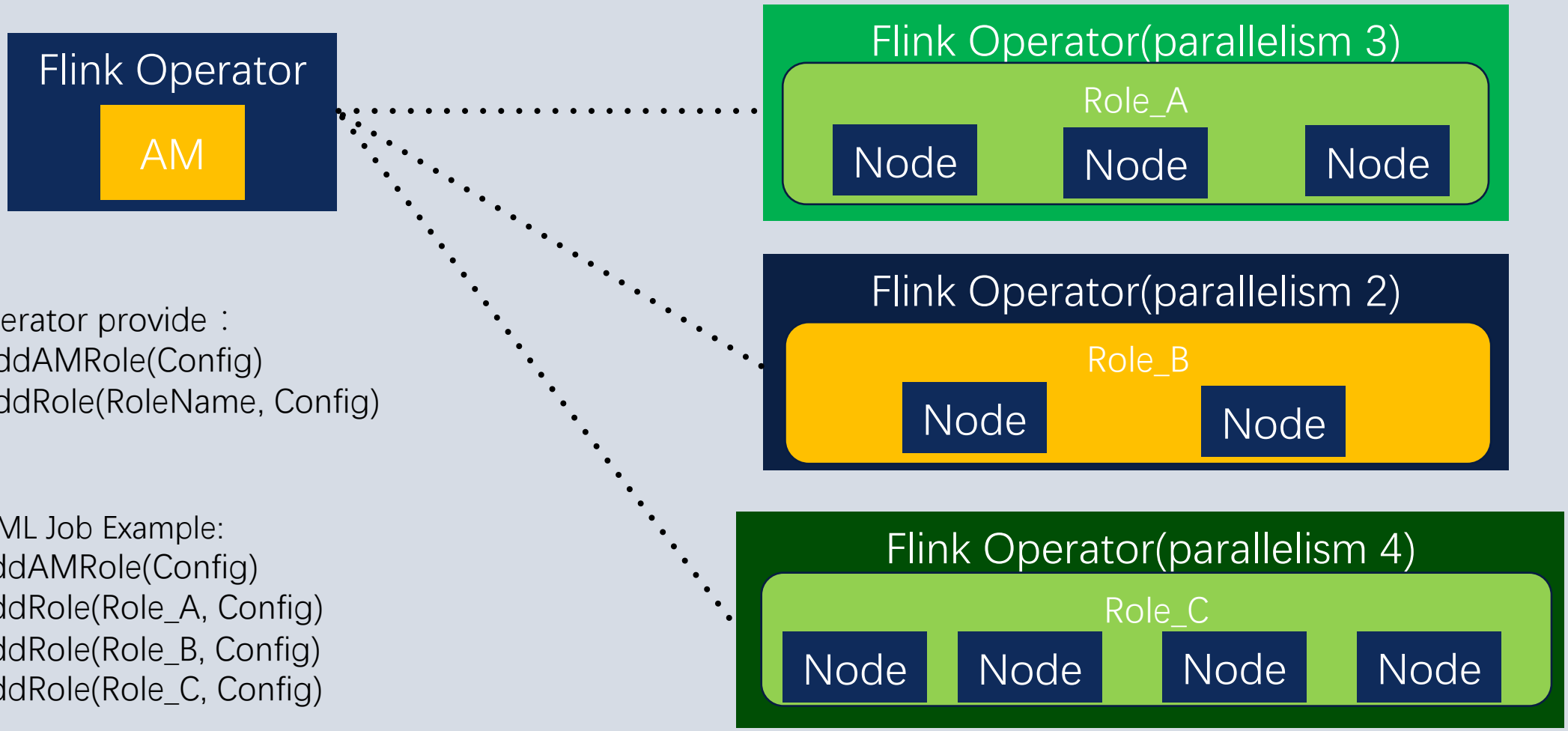
ML Framework

Apache Flink



ML Operator

Apache Flink



ML Operator provide：

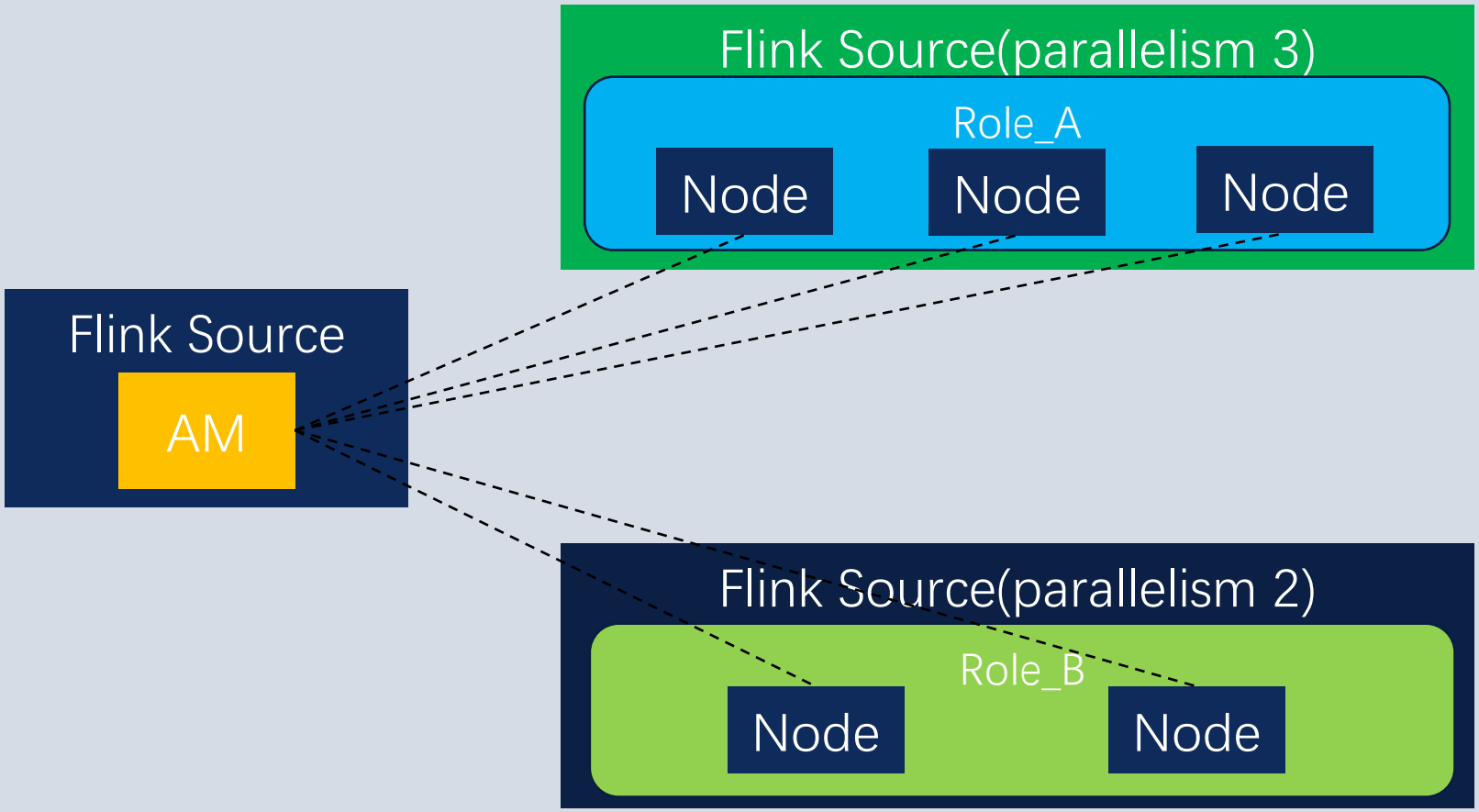
1. addAMRole(Config)
2. addRole(RoleName, Config)

Create ML Job Example:

1. addAMRole(Config)
2. addRole(Role_A, Config)
3. addRole(Role_B, Config)
4. addRole(Role_C, Config)

Batch Mode

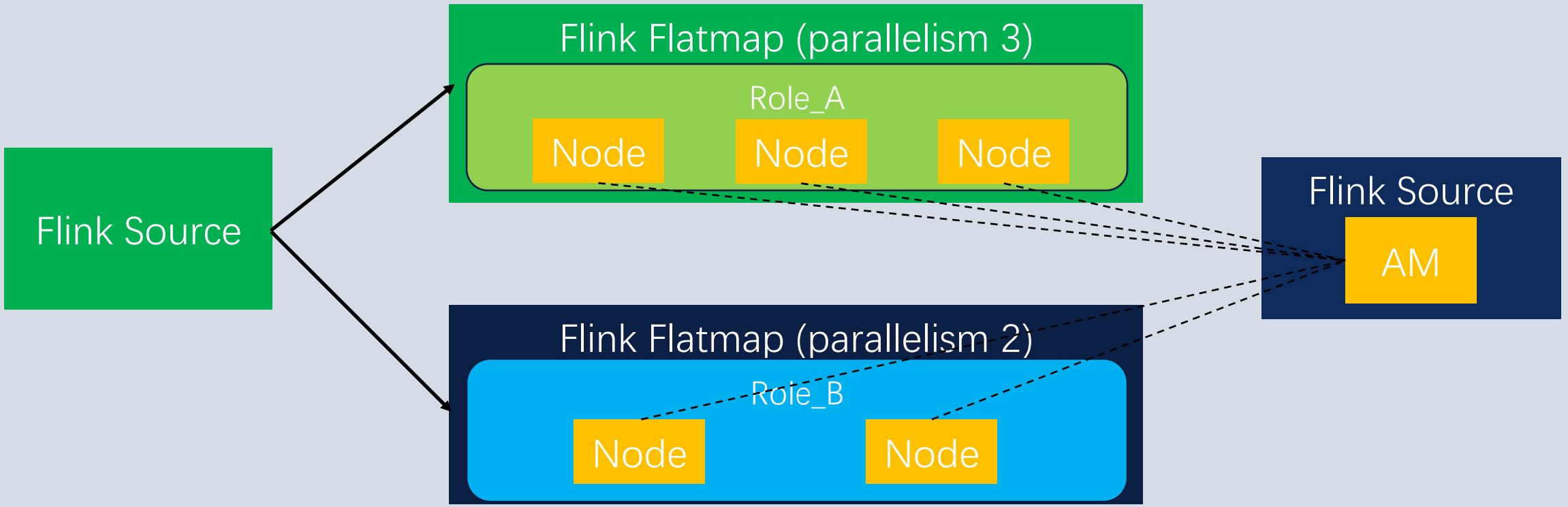
Apache Flink



Do not read data from flink， plan role as flink source operator

Stream Mode

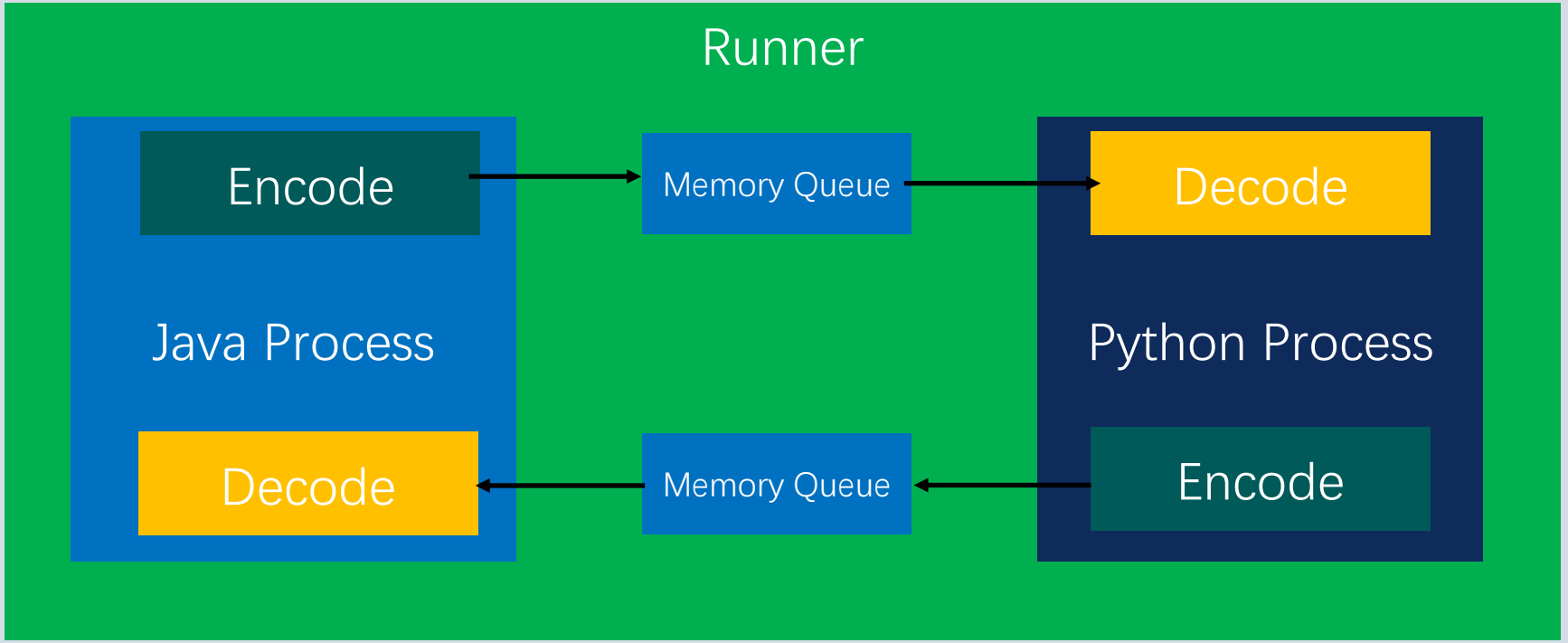
Apache Flink



Read data from flink ， plan role as flink flatmap operator

Data Exchange

Apache Flink



1. Encode transfer user define object to byte[]
2. Decode transfer byte[] to user define object

Encode and Decode is extensible

Summary

Apache Flink

1. Cluster creation and state management

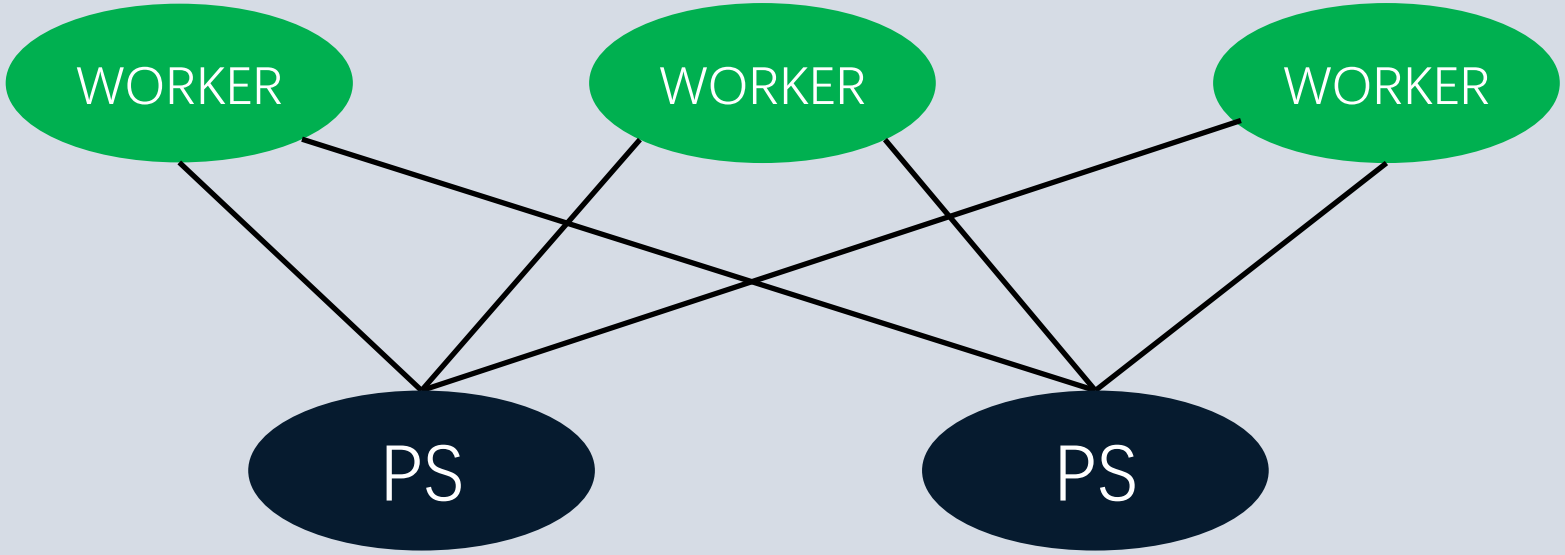
2. Java and python data exchange

PART 03

TensorFlow

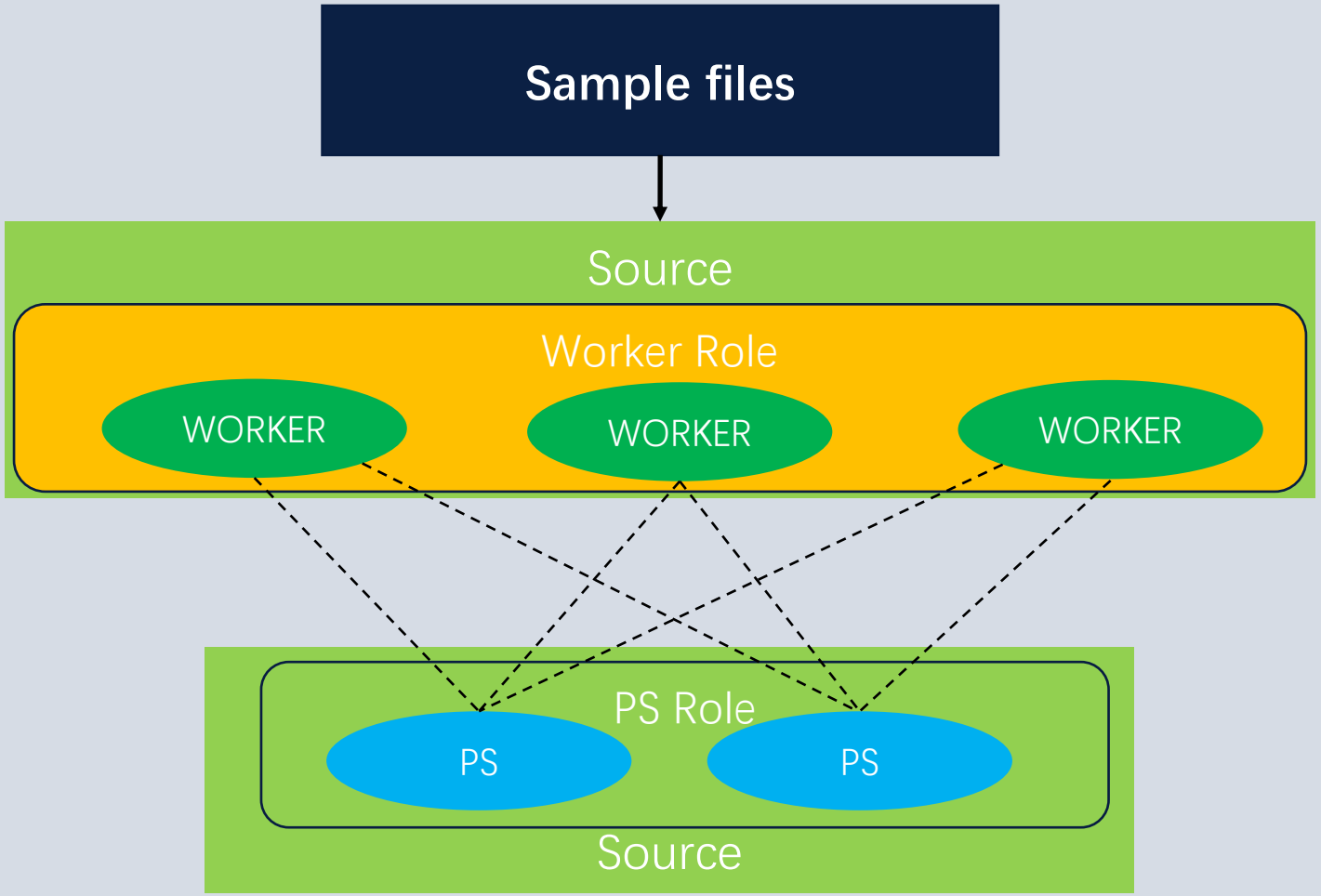
TensorFlow Training

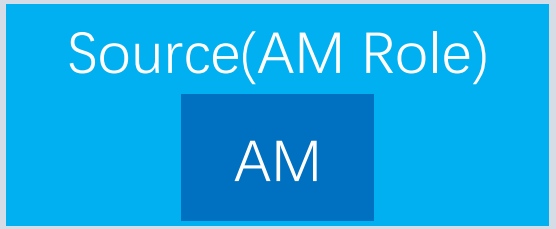
Apache Flink



TensorFlow Batch Training

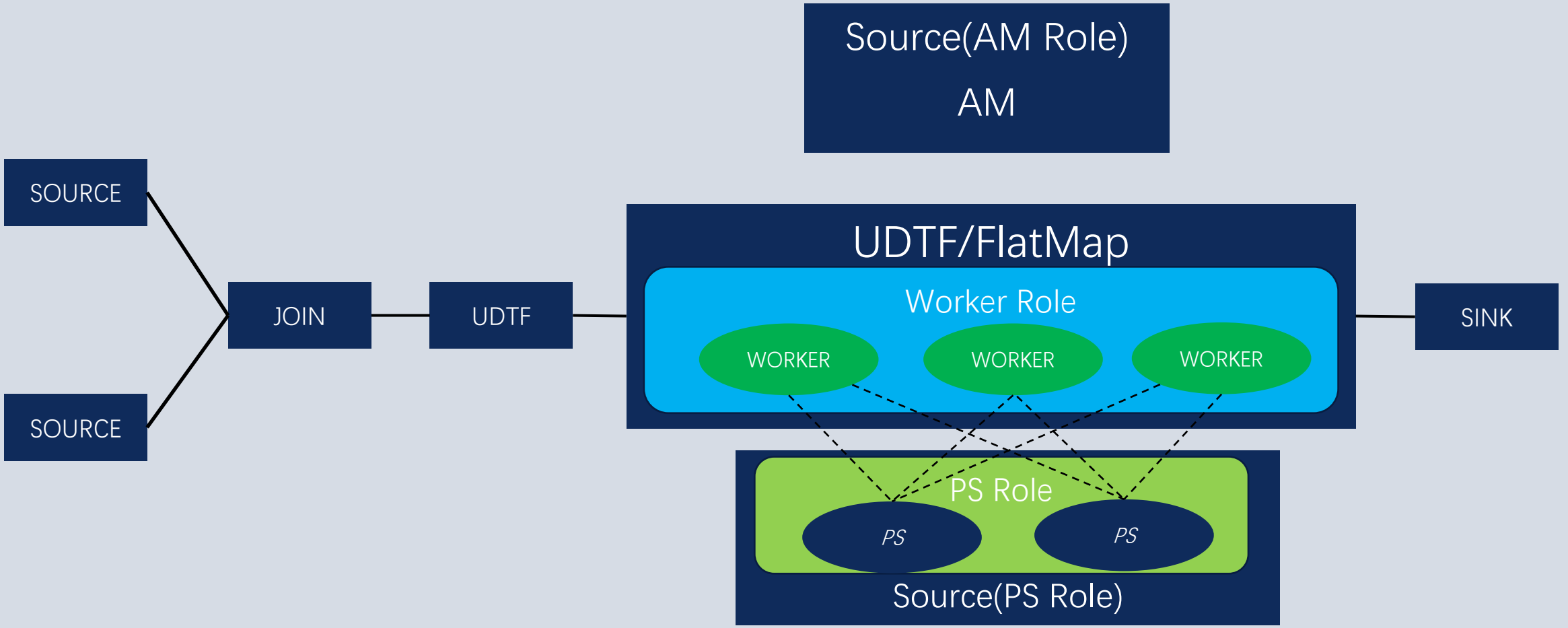
Apache Flink





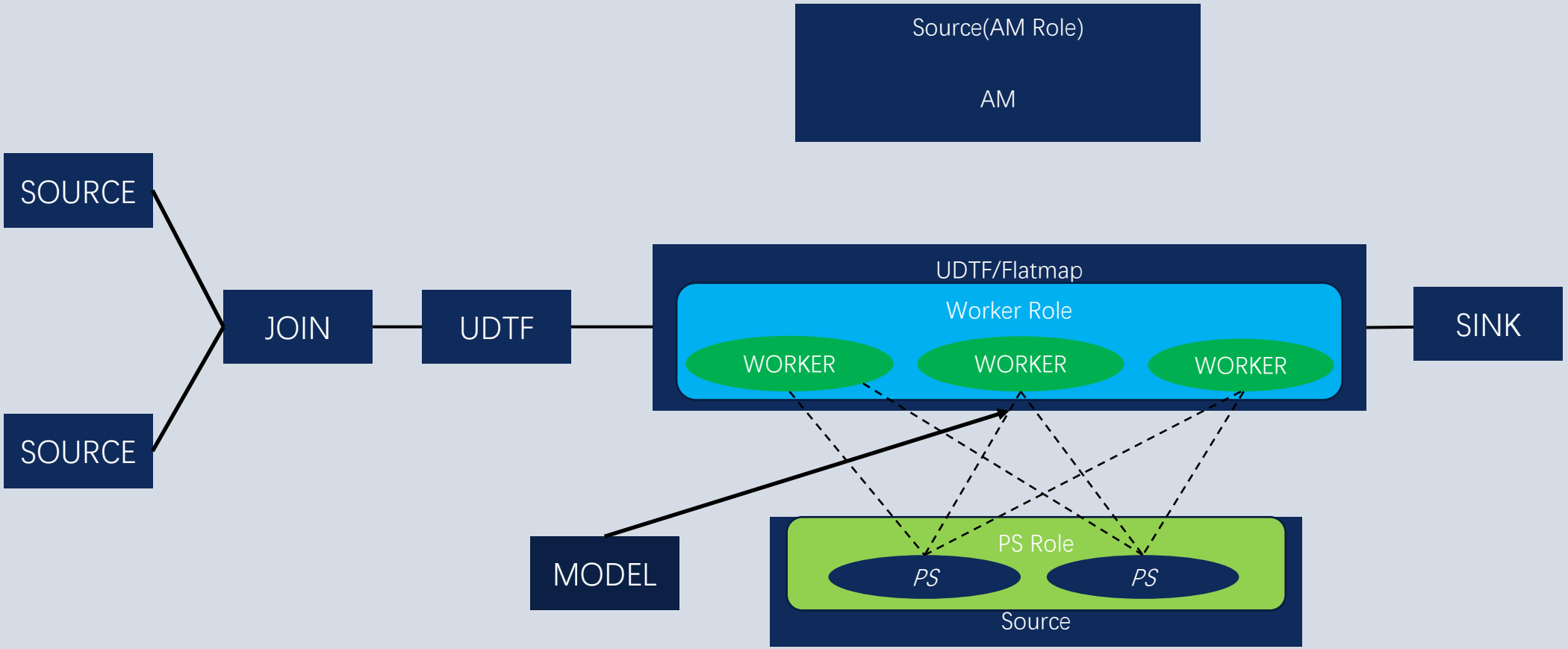
TensorFlow Stream Training

Apache Flink



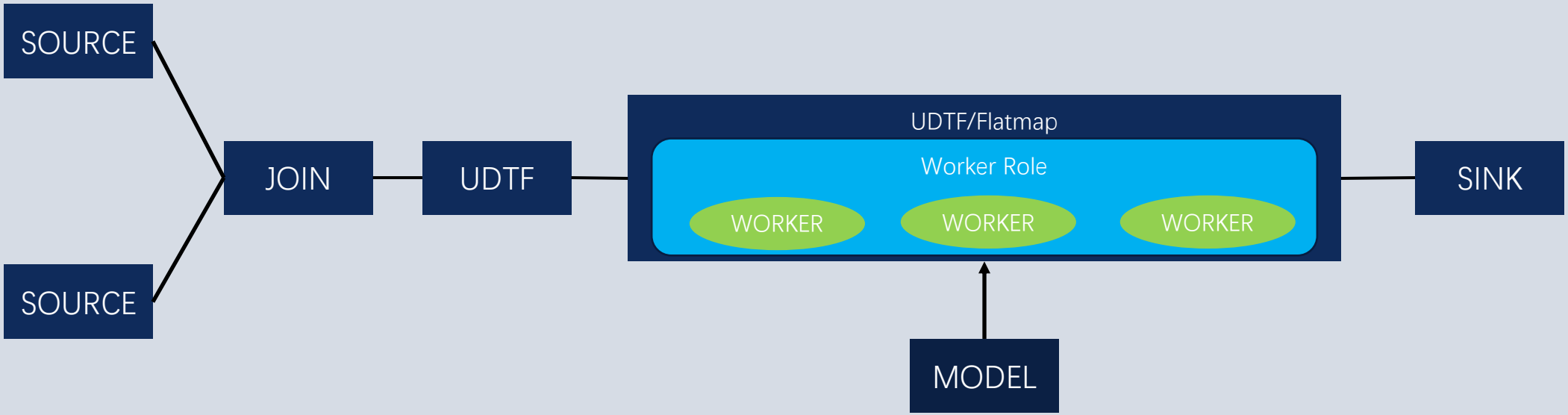
TensorFlow Stream Inference

Apache Flink



TensorFlow Stream Inference

Apache Flink



TensorFlow Example

Apache Flink

```
import tensorflow as tf
cluster = tf.train.ClusterSpec({
    "worker": [
        "A_IP:2222",
        "B_IP:1234",
        "C_IP:2222"
    ],
    "ps": [
        "D_IP:2222",
    ]})
isps = False
if isps:
    server = tf.train.Server(cluster, job_name='ps', task_index=0)
    server.join()
else:
    server = tf.train.Server(cluster, job_name='worker', task_index=0)
    with tf.device(tf.train.replica_device_setter(worker_device='/job:worker/task:0', cluster=cluster)):
        w = tf.get_variable('w', (2, 2), tf.float32, initializer=tf.constant_initializer(2))
        b = tf.get_variable('b', (2, 2), tf.float32, initializer=tf.constant_initializer(5))
        addwb = w + b
        mutwb = w * b
        divwb = w / b
```

TensorFlow Example

Apache Flink

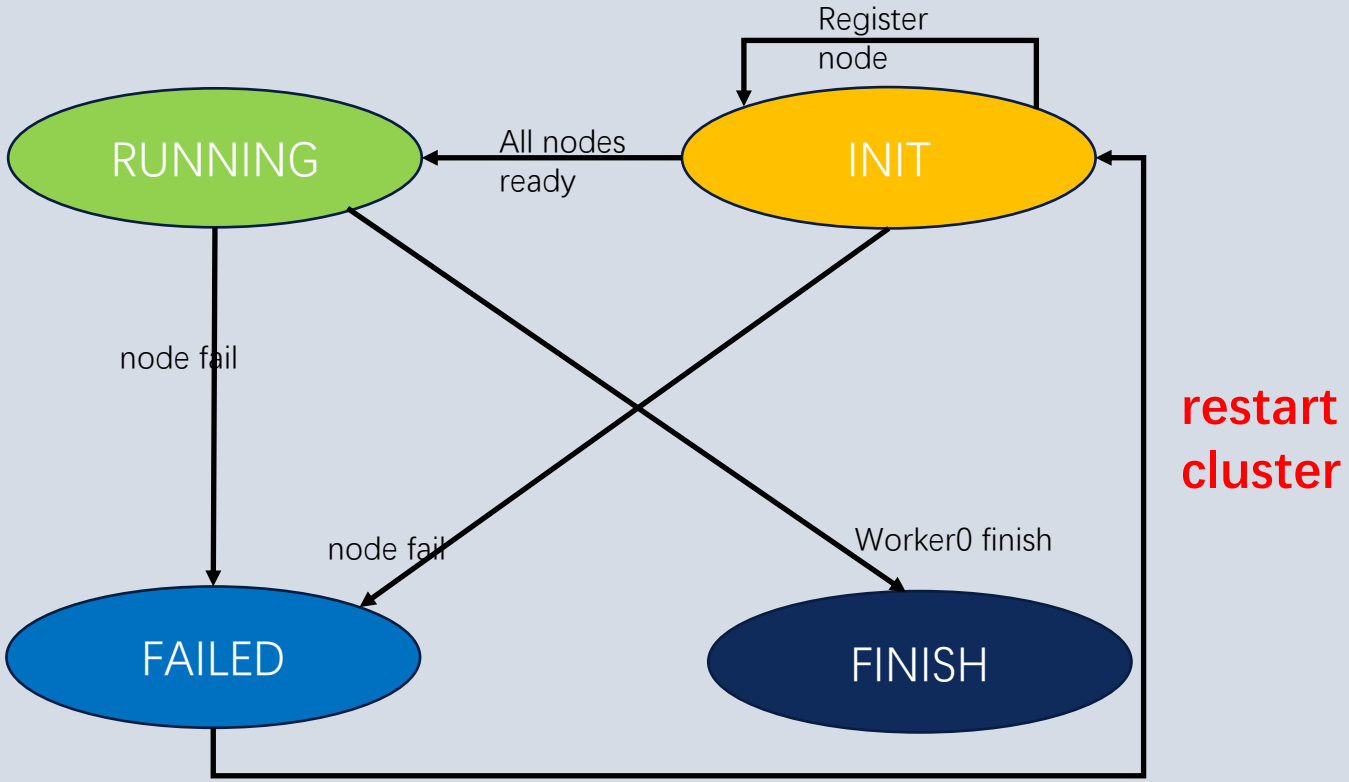
```
import tensorflow as tf
from flink_ml_tensorflow.tensorflow_context import TFContext

def run_main(context):
    tf_context = TFContext(context)
    cluster = tf_context .get_cluster()
    job_name = tf_context .get_job_name()
    task_index = tf_context .get_task_index()
    if 'ps' == job_name:
        server = tf.train.Server(cluster, job_name=job_name, task_index=task_index)
        server.join()
    else:
        server = tf.train.Server(cluster, job_name=job_name, task_index=task_index)
        with tf.device(tf.train.replica_device_setter(worker_device='/job:worker/task:0', cluster=cluster)):
            w = tf.get_variable('w', (2, 2), tf.float32, initializer=tf.constant_initializer(2))
            b = tf.get_variable('b', (2, 2), tf.float32, initializer=tf.constant_initializer(5))
            addwb = w + b
            mutwb = w * b
            divwb = w / b

if __name__ == "__main__":
    stream_env = StreamExecutionEnvironment.get_execution_environment()
    train(3, 1, run_main, properties=None, stream_env=stream_env, input_ds=None, output_row_type=None)
```

TensorFlow Failover

Apache Flink



TensorFlow Applications

Apache Flink

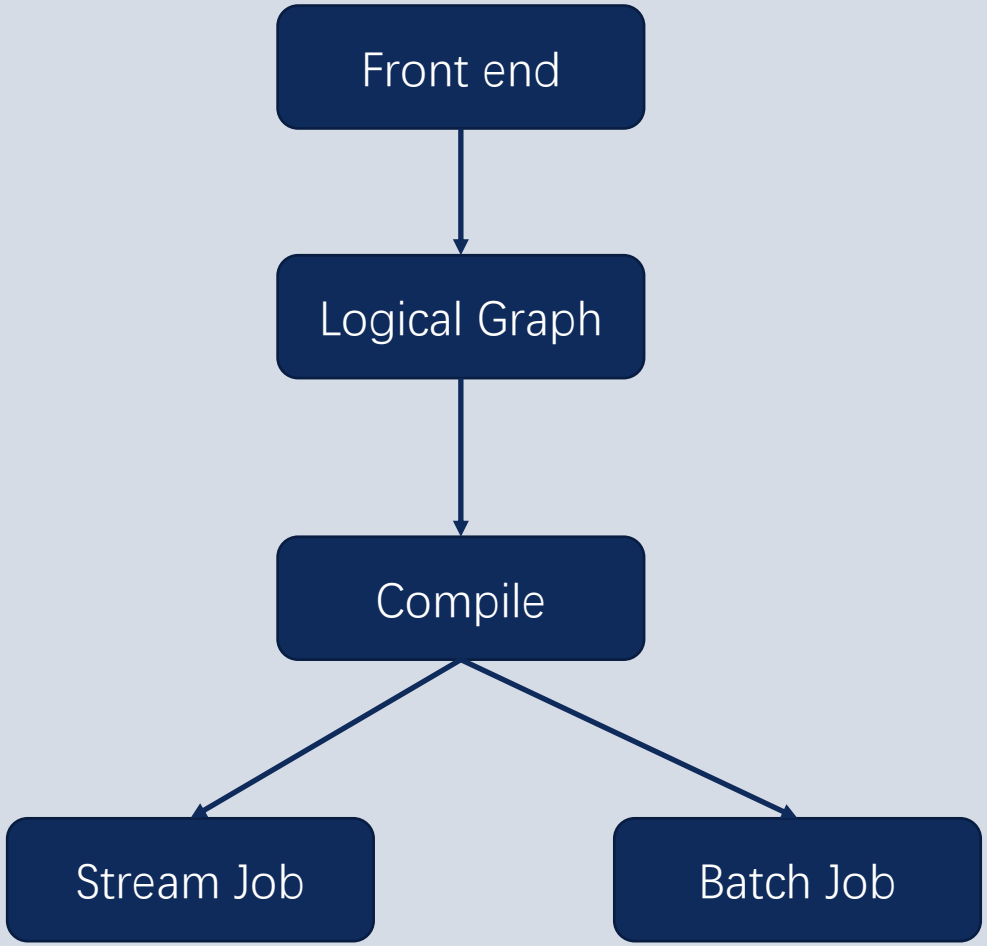
1. Search Ranking
2. Recommender System

Apache Flink

PART 04

Future

F u t u r e



Apache Flink

https://github.com/alibaba/flink-ai-extended

FLINK FORWARD

Organized by 阿里云

Created by ververica

Flink Forward Asia

全球最大的 Apache Flink 官方会议

预计 **2000+** 参会人员， 2019年11月28-30日 @北京国家会议中心

国内外一线厂商悉数参与

阿里巴巴、腾讯、字节跳动、intel、 DellEMC 、Uber、美团点评、Ververica ...

大会官网，查看更多

FLINK FORWARD

Organized by 阿里云

Created by ververica

Flink Forward Asia

The largest Apache Flink official conference all over the world

Expected 2000+ participants

28th November 2019 - China National Convention Center @Beijing

Numbers of Top domestic and foreign companies

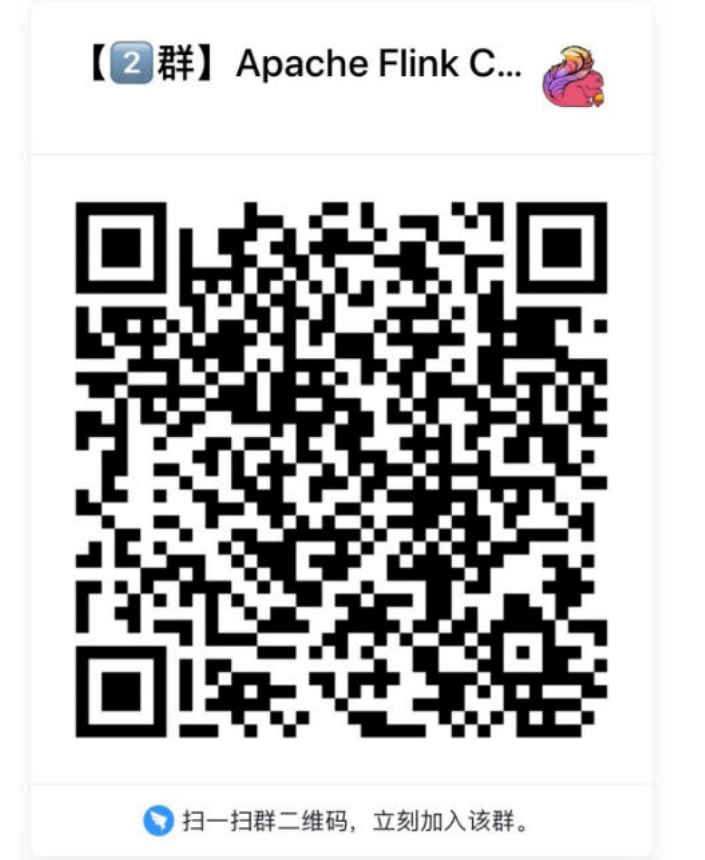
Alibaba, Tencent, Bytedance, Intel, DellEMC, Uber, Meituan Dianping, Ververica...

Official website

Apache Flink

Apache Flink 社区微信公众号「 Ververica」

Meetup动态 / Release 发布信息 / Flink 应用实践



Apache Flink

THANKS

Apache Flink China Meetup

· SHANGHAI