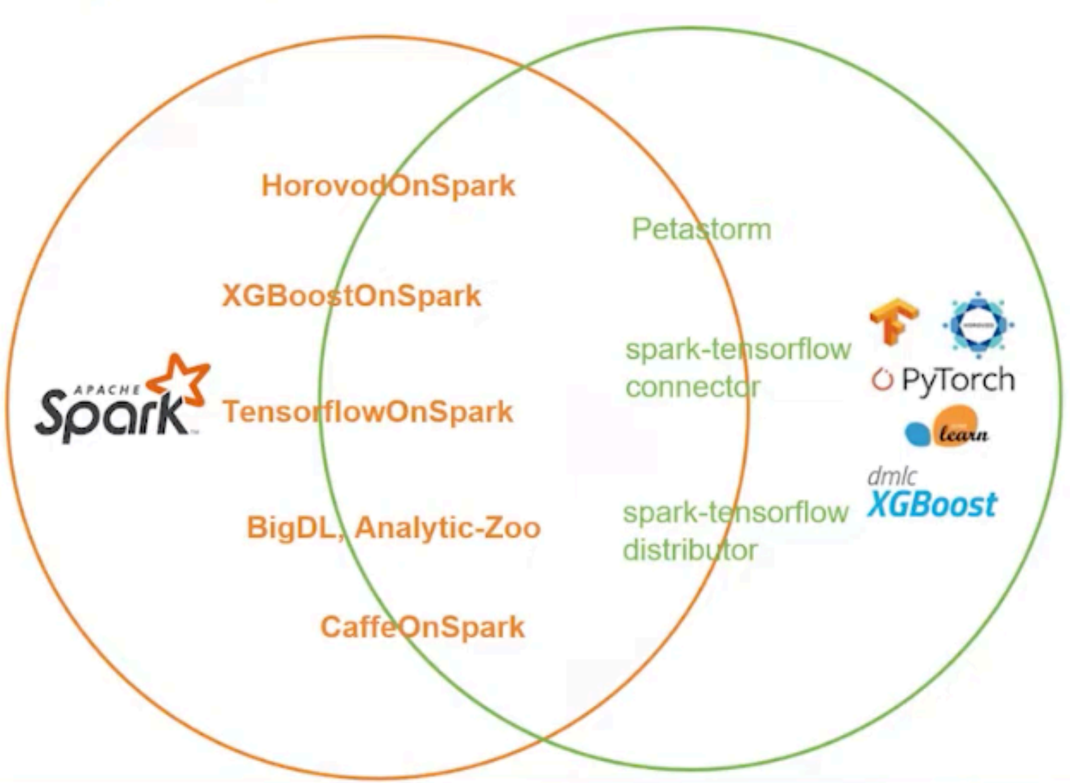
Integrating Big Data and AI Using Spark and Ray

Wanru Zhao

Spark and Ray

Big Data and AI

- Massive data is critical for better AI.
- Distributed training will be a norm.
- Many community efforts to integrate big data with AI.

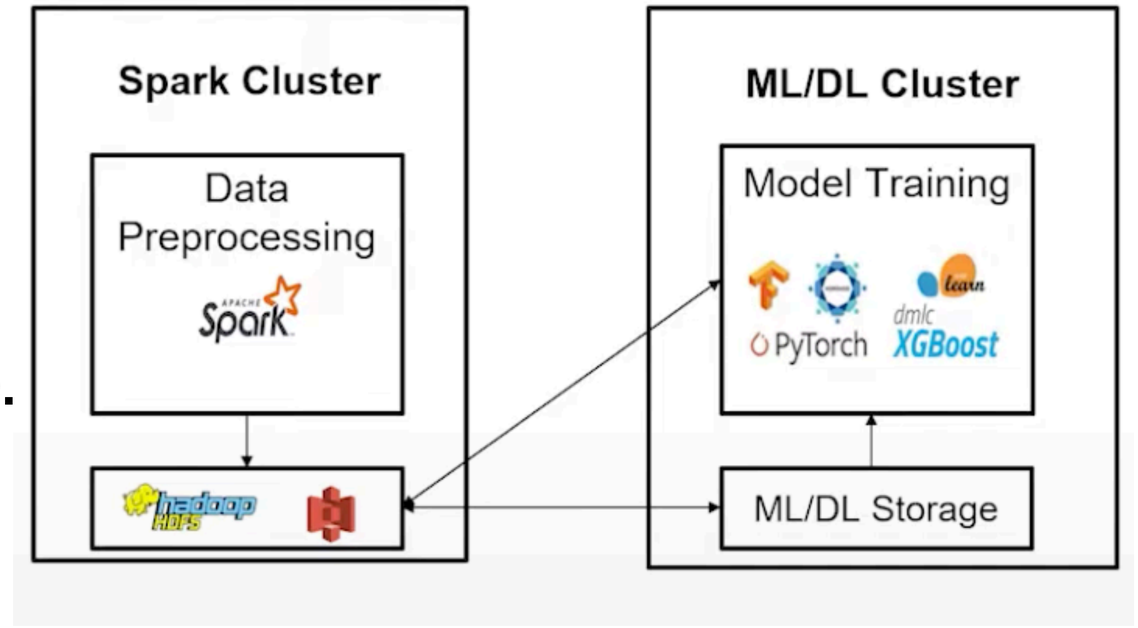


Big Data + AI

Separate Spark and AI Cluster

Challenges:

- Data movement between clusters.
- Overhead of managing two clusters.
- Segmented application and glue code.

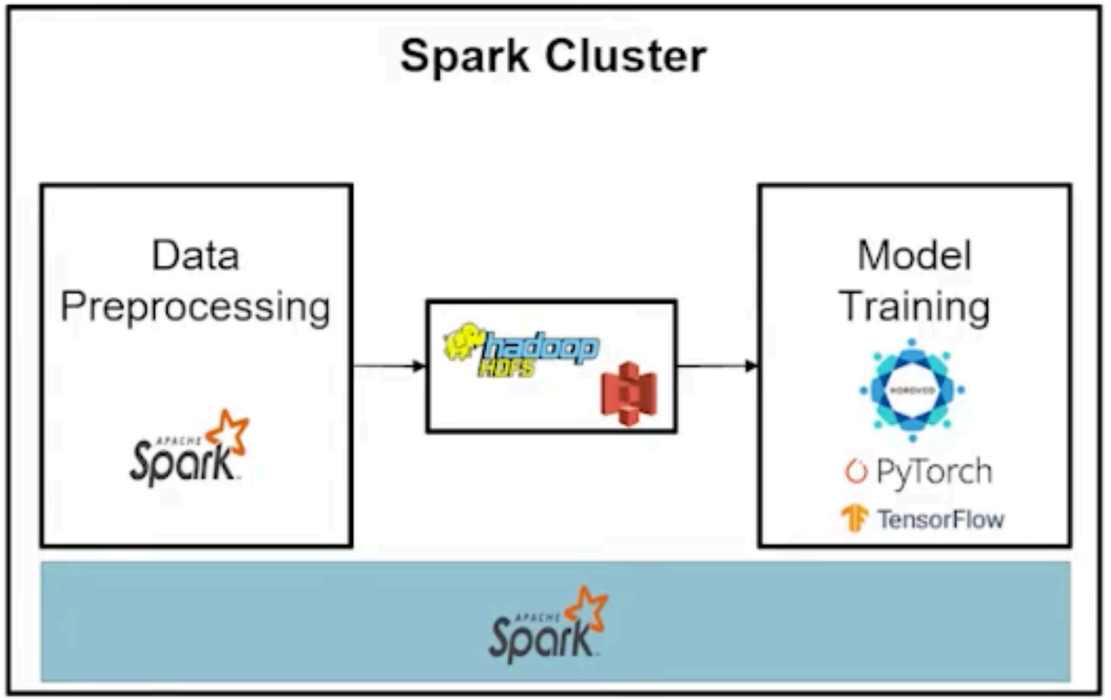


Big Data + AI

Running ML/DL Frameworks on Spark

Challenges:

- Specific to Spark and requires ML/DL frameworks supported on Spark.
- Data exchange between frameworks relies on distributed filesystems like HDFS or S3.



RayOnSpark
in Analytics Zoo

- Allow users to directly
  - Run Ray programs on existing Big Data clusters
  - Write Ray code inline with their Spark code
    - process the in-memory Spark RDDs or DataFrames

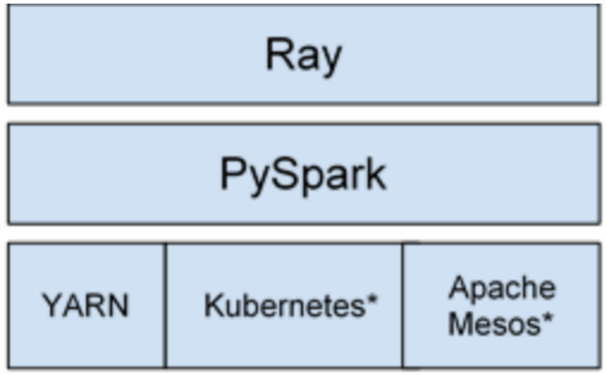


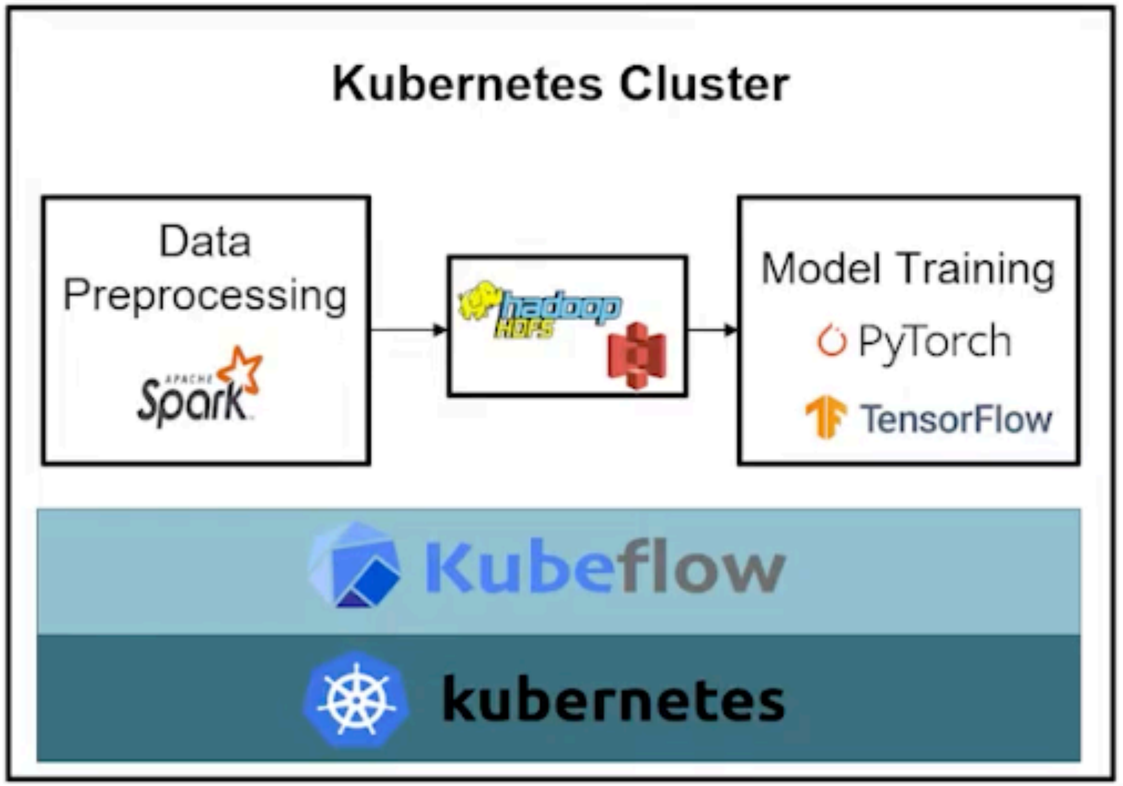
Figure1: Deploy Ray* on Apache Spark*

Big Data + AI

Running on Kubernetes

Challenges:

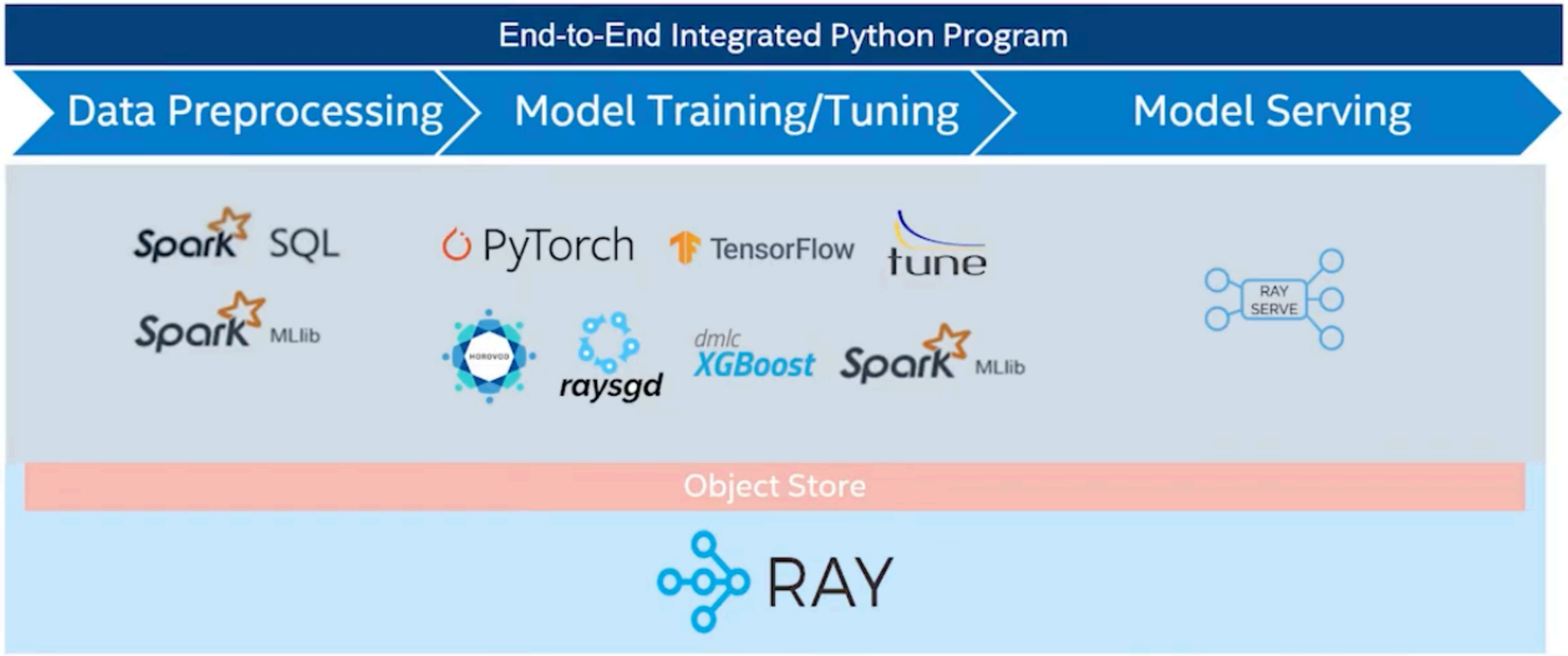
- The pipeline must be written in multiple programs and configuration files (v.s. a single python program).
- Data exchange between frameworks relies on distributed filesystems like HDFS or S3.



RayDP

Build End-to-End Pipeline using RayDP and Ray

- RayDP provides simple APIs for running Spark on Ray and integrating Spark with distributed ML/DL frameworks.



Goals

- Comparison between Ray and Spark on both Big Data and AI tasks.
- Comparison of different methods of integrating big data with AI.
- Examine the architecture of SparkOnRay and RayDP in detail.

Plan and Progress

- ✓ Read the papers and related work
- ✓ Go through the tutorial for Ray and Spark
- Go through the tutorial for RayDP
- Go through the tutorial for RayOnSpark and Analytics Zoo
- Run and evaluate RayOnSpark examples
- Run and evaluate RayDP examples:
  - Spark + XGBoost on Ray
  - Spark + Horovod on Ray
  - Spark + Horovod + RayTune on Ray
- Write down the results

References

- Analytics Zoo is an open source Big Data AI platform: https://github.com/intel-analytics/analytics-zoo
- https://medium.com/riselab/rayonspark-running-emerging-ai-applications-on-big-data-clusters-with-ray-and-analytics-zoo-923e0136ed6a
- https://github.com/oap-project/raydp
- RayDP: Build Large-scale End-to-end Data Analytics and AI Pipelines Using Spark and Ray

Thank you!

Suggestions?