Experiments on integrating ROOT and Spark

E. Tejedor, D. Piparo, P. Mató

ROOT meeting

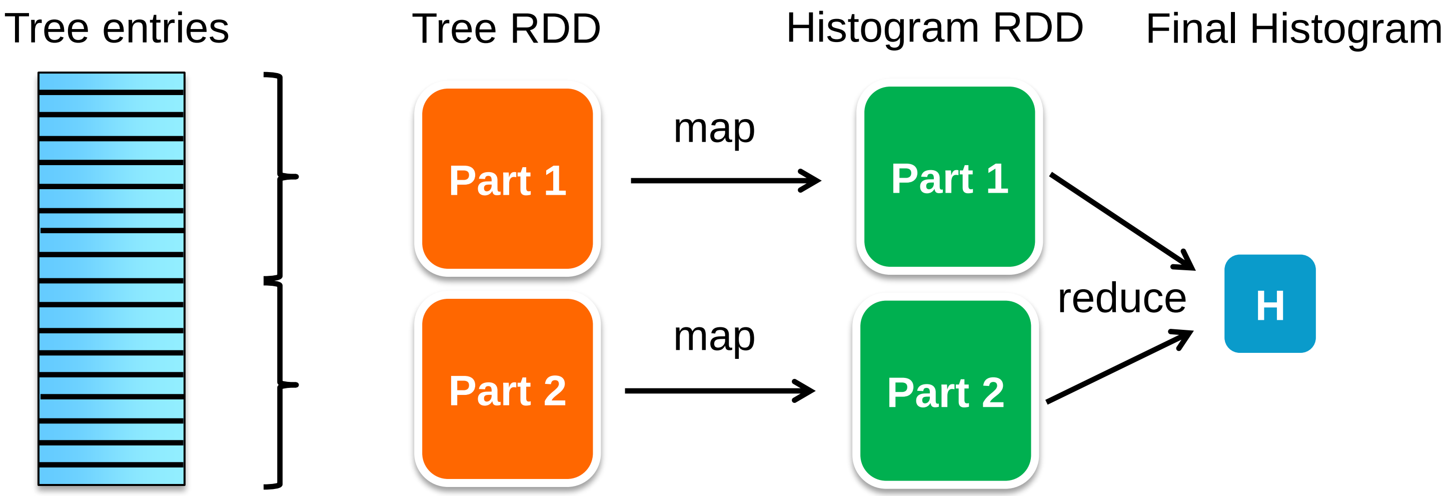
07/03/2016

Motivation, Objective

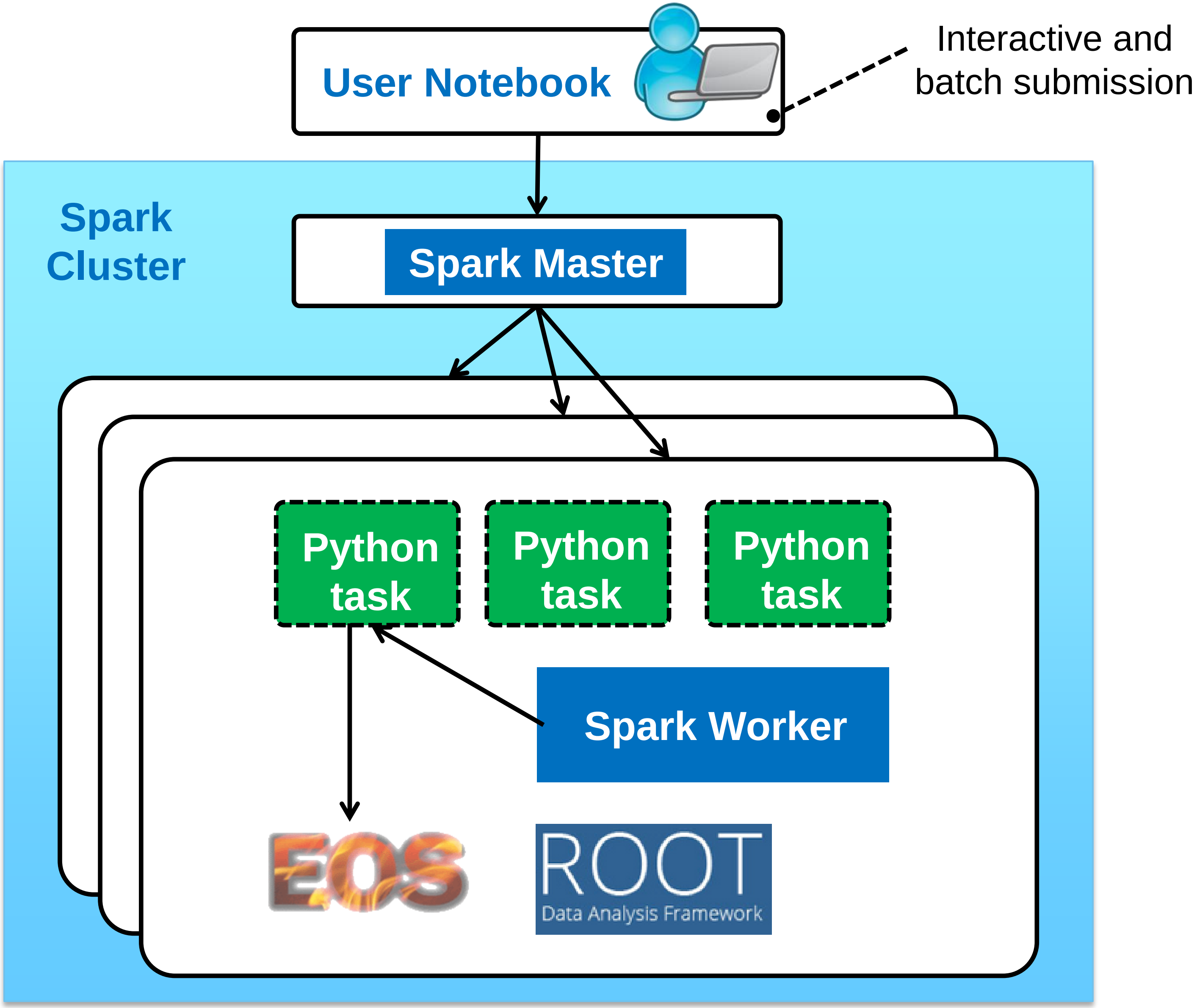
- Meeting with IT-ST in Analytics WG
  - Learnt how IT-ST uses Spark to do analysis on non-physics data (e.g. system monitoring logs), from notebooks
  - Started to explore how to use Spark with physics data, with ROOT
  - Potentially a good complement to DMaaS: distributed execution, not constraint to the container
- Objective: reuse existing technologies as much as possible
  - PySpark to leverage PyROOT
  - Storage: explore EOS as an alternative to HDFS
    - Avoid problem of splitting ROOT binary files
    - Avoid problem of data ingestion

Data Model

- Spark is based on the following main concepts:
  - RDD: distributed collection of items
  - Actions and transformations applied on RDDs (map, reduce, filter, etc.)
- Model ROOT data as RDDs
  - Tree: collection of entries, logically split – input of map
  - Histograms: output of map, input/output of reduce



Execution in a Cluster


User Notebook
Interactive and batch submission
Spark Cluster
Spark Master
Python task
Python task
Python task
Spark Worker
EOS
ROOT
Data Analysis Framework