Spark and Jupyter

IT - Analytics Working Group - Luca Menichetti

Analytics Working Group

- Cross-group activity (IT) which main goal is to perform analyses using monitoring data coming from different CERN IT services.
- Members of the AWG have the commitment to share their data with others using the Hadoop cluster (HDFS) as common repository.
  - This is implicit if their applications are already running inside Hadoop.
- An analysis task can therefore be performed inside the cluster using any Hadoop compliant technology
  - such as MapReduce, Spark, Pig, HBase, Hive or Impala

Analysis with Spark

- Why? Spark is the most popular Big Data project also because, including it in the analysis process, it is possible to gather together many steps required by the analysis workflow.
- To submit a job is not straightforward:

```
spark-submit –executors-num 10 –mem-executor 4G
   –executor-cores 4 –class MyMainClass myCode.jar args
```

  - Every change to the code requires to compile and to submit the jar again
  - Impossible to execute only the modified parts
- Developing in Spark within a web notebook will improve the overall analysis process and the learning curve.

Spark notebooks – current status

- Requirement: to fully exploit Hadoop and Spark, the notebook kernel has to run in a machine that is a client of the Hadoop cluster.
  - Read/Write HDFS directly (avoid data migration)
  - Spark application can scale up/down within the cluster
- Current solution: IPython, Jypiter and Apache Zeppelin (installed manually) running in Openstack nodes with puppetized Hadoop client installation.
  - Jupiter runs PySpark (Python API for Spark) as a kernel.
  - Zeppelin (multi-interpreter, interactive, web notebook solution) provides built-in Spark integration.

Limitations and issues

- Each user has to create it’s own virtual machine and install the notebook platform she/he would like to use
- No possibility to properly share notebooks (currently gitlab)
- Running in “yarn mode” (distributed job in the cluster)
  - Not fully supported
  - Still not clear how to manage the distributed application
- Each Hadoop client needs a Kerberos token used by Jupiter/Zeppelin instance to access HDFS and run jobs (users cannot share the instance).

Interests and ideas

- We are supporting users to run Spark with notebooks, however we will not provide this as a service. Work in progress to create a puppet module to install Jupyter/Zeppelin. We hope in a future this can be integrated in a proper service to provide ready-to-use Spark notebooks.
- To solve the sharing/portability problem (to avoid dependencies issue, both missing library or version).
- Features:
  - Additional interpreters for Spark components (SparkSQL)
  - Notebook versioning
  - Schedulers
  - Notebook web visualization

CERN