SWAN and Spark as a Service
POW 2019

IT-DB-SAS, 29th Oct 2019

Prasanth Kothuri, Piotr Mrowczynski

Highlights

- SWAN
  - Ability to offload Spark computations to Kubernetes Cluster from the Jupyter notebook
  - Provision of dedicated LCG software view for NXCals
  - Development of Jupyter extension to offload Spark computations to user managed kubernetes cluster
  - Early prototype of Porting SWAN to IT Container Service (Kubernetes)
  - Technical consultancy on Spark technology for AWAKE data reduction, TOTEM analysis and RDataFrame
  - Development and setup of testing pipelines for Spark notebooks for early detection of issues with configuration, LCG software and cluster setup

Highlights

- SWAN
  - Development of monitoring and alerting for notebook spawn errors and performance
  - Refactoring deployment of Hadoop and Spark configuration to CVMFS
  - Redesign of the SWAN ‘configure environment’ page
  - Publishing of example Spark notebooks for main usecases of Spark service
    - Distributed ROOT RDataFrame
    - AWAKE Analysis
    - Machine Learning with Apache Spark
  - Update SWAN to limit the spark ports to a range allowed on Hadoop clusters firewall
  - Propagation of user/local python packages from Spark driver to executors

Highlights

- SWAN
  - Improvements to usage of Spark in SWAN
    - Access to driver logs
    - Access to Spark history server
    - Access to Spark Application metrics
  - Refactoring of SparkConnector, SparkMonitor Jupyter extensions
  - Presentation of Spark integrations to SWAN at HEPIX
  - Organize and participate in SWAN User Forum

Highlights

- Spark as a Service
  - Successful delivery of ‘Introduction to Spark Data APIs training’
  - Migration of Spark on Kubernetes cluster to cpu-optimized hardware
  - Development of Helm charts to initialize Spark services in kubernetes cluster
  - Development of grafana dashboards to monitor health of the cluster
  - Improvements and fixes to hadoop-xrootd connector
  - Testing of Spark streaming on K8s and K8s jobs to submit Spark/YARN ETLs

Highlights

- Backup and Recovery
  - Development of recovery pipelines to validate backups
  - Published KB article on Backup and Recovery policy for Hadoop service

POW 2019 – SWAN and Spark as Service

SWAN

- Productionize and integrate SWAN service in IT-DB-SAS
  - Train IT-DB-SAS service managers on SWAN service
  - Build service operations guide
  - Establish ROTA and snow FE
  - Participate in the governance/development of SWAN service

- Migrate SWAN service to kubernetes
  - Secrets management
  - Logging, metrics and alerting
  - HA configuration (e.g. k8s master)
  - Different levels of QoS with e-groups and node labels

- Jupyterlab
  - Migrate all Jupyter notebook extensions to jupyterlab; total 10 extensions, at the minimum we are responsible for SparkConnector, SparkMonitor, HdfsBrowser and K8sSelection
  - Unification of ‘cloud containers’ with user managed clusters (k8sselection extension)

SWAN

- Address requirements from the Users
  - Better spec. instances and QoS for dedicated e-groups
  - Git integration

- Organize and participate in delivering Academic training on SWAN and Spark

- Notebook widgets and Cell magics
  - %%sql execute Spark SQL
  - %%info access spark application information
  - %%conf access spark application configuration
  - %%topandas return spark dataframe as pandas DF

- Schedule notebooks ?

Spark as a Service

- Spark as a Service
  - SPOT instances and autoscaling
  - Spark 3.0 and dynamic allocation SPARK-24432
  - Technical consultancy and documentation on RDataFrame interactive mode usecases
  - Technical consultancy and documentation on S3 usecases
  - Technical consultancy and documentation on spark-root batch mode usecases
  - Deliver Introduction to Spark APIs training

Backups and Hadoop-xrootd

- Backup and Recovery
  - Refactoring of backups
  - Resurrect web UI to consume backup information and request recoveries ?

- Hadoop-xrootd
  - Avoid dependency on platform (gcc version) for C++
  - Maven integration