OpenStack/Hadoop/Spark
LSST

Journées MAESTRO
5-6 juillet 2016

Christian Arnault (LAL)
Réza Ansari (LAL)
Axel Chevarin (LAL)
Johann Cohen Tanoudji (LUPM)
R&D Spark

**Is the Spark technology appropriate for us?**
- Many experiences in the Big Data domain
  - But do they fit LSST needs?

**Spark:**
- From Apache Foundation
- (Hadoop/HDFS) + MapReduce + Spark + DataSets + *many components*
- ... components are Spark capable:
  - For various Data Formats
    - Text
    - Columnar data
    - SQL, NoSQL
    - Records
    - JSON
    - User defined types described by a schema
- Schedule and optimise distribution of:
  - Data blocks vs. CPU vs. Disk vs: Memory
- Map Reduce algorithms through functionnal programming
  - (Java), Scala, Python, R
R&D Project

• Set up R&D in the context of Université Paris Sud
  - Candidate for an « Equipements de Recherche Mutualisés » (ERM)
  - OpenStack clouds (VirtualData + Paris-Sud)

• Study several use cases
  - « Queries » on large number of FITS files
    • 2D WCS queries
    • Various selection based on FITS header keys
  - Production of large catalogs of simulated galaxies
  - Working on large images
  - ...

• Understand the scaling factors
  - Number of workers
  - Distribution of CPU, disk, memory
  - Relation with DTBs, with data formats

• Other may join this study
  - Johann, ...
The Spark technology

- **the Tungsten projet**
  - Memory Management and Binary Processing
  - Cache-aware computation
  - Code generation -> GPUs

- **Drill**
  - ANSI SQL
  - + Schema-free SQL Query Engine for Hadoop, NoSQL and Cloud Storage

- **Petasky**
  - Some measures have been done using Spark/Drill (to be announced)

- **R&D in the context of Strasbourg CDS**
  - Cross correlation of catalogs

The Spark world is rapidly evolving..
The Spark technology

• The data format
  - Spark expects data formatted around the (Key, Value) pair schema
  - However, many mechanisms permit to let Spark understand specific data formats
    • Through Serialization
    • Using JSON description,
    • JSON, CSV, SQL, NoSQL

• Several Spark packages can leverage some specialized data formats:
  - Hive (for SQL queries)
  - MongoDB => FITS headers
  - Parquet (DataSets, DataFrames) (cf HDF5)
  - Avro (binary serialization) => FITS headers
Status

- One installation is available at CCIN2P3:
  - Include: MondoDB + Spark + HDFS (installed by Osman)
    - 10^9 SNLS12 FITS headers has been uploaded into MongoDB
    - => 2D WCS queries
- One installation in the VirtualData cloud (LAL+UPSUD)
  - Up to 6 workers (HDFS, OpenStack, Spark + Avro + SciPy + Matplotlib)
  - Production of simulated galaxies (with Johann help)
    - => experiencing histogramming
  - Getting experience with M.R. algorithms
    - Working with Spark experts from Orsay University
    - A Spark Expert group is currently starting (Loops context)
  - Using the Spark statistics tools to measure the performances
    - CPU, Disk, IO, Memory
- No real measurements yet ...
  - Still ongoing...
Some references

- Hadoop/Spark and HDF5
  https://www.hdfgroup.org/pubs/papers/Big_HDF_FAQs.pdf

- Map Reduce in astro

- Spark v2.0

- Drill
  https://www.mapr.com/blog/apache-spark-vs-apache-drill
  https://drill.apache.org/

All these are topics for discussion

Next version coming soon