CHAOUL Laurence, PANEM Chantal
CNES, Toulouse

PROCESSING GAIA'S BILLION STARS IN CNES, A BIG DATA STORY

gaia

EWASS 2017, June 29th 2017

The DPCC Key Numbers

Gaia DPCC = Gaia Data Processing Center in CNES

Up to 8 chains to be processed in parallel

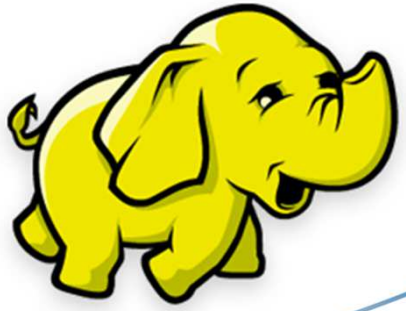
- Spectroscopic chains
- Objects processing chains (Multiple Stars, Solar System Objects, Galaxy, Quasars)
- Astrophysical Parameters Determination chain
- Various kinds of complex algorithms (object by object or global processings)

Number of objects to process

- 1 billion stars, 80 billions observations
- Increasing volume all along the mission, up to 4PB

Architecture built with a NoSql solution : Hadoop

- High Level of parallelization
- HDFS: distributed data storage



The first main lessons learned

- It works !
- High Performing Parallel Computing
- Able to handle tables with tens of billions records
- Very scalable

- Hadoop configuration very tricky
- No indexing capabilities : difficult to request the content of the tables => problem for validating the results
- The scientific algorithms shall be designed to be run in parallel

Proven architecture for Gaia DR2 processings and on the right track for Gaia DR3 !

More info ?

http://www.cosmos.esa.int/web/gaia/

http://smsc.cnes.fr/GAIA/

laurence.chaoul@cnes.fr