Ecole BIG-DATA et NoSQL

Qserv

base de données distribuée pour LSST

Fabrice Jammes (LPC Clermont)
Emmanuel Medernach (LPC Clermont)

LSST

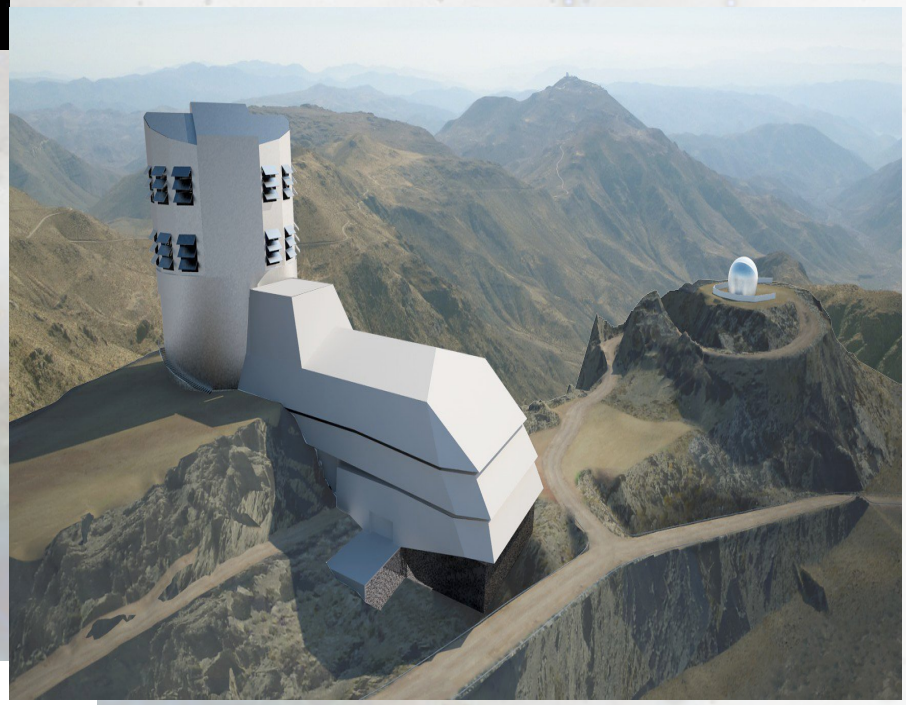
Téléscope « grand champ » destiné à la recherche de l'énergie noire et à l'étude des supernovae.

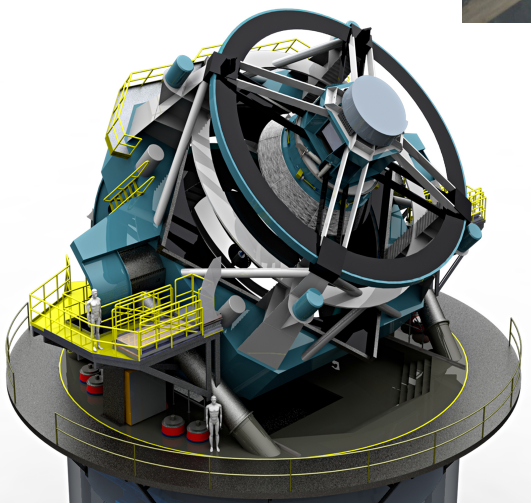
Ce téléscope, doté d'une CCD de 1m2, réalise un film de l'univers.

60 PB d'images => 5 PB de données de type catalogue.

LSST en une slide

- Expérience de cosmologie de 4ème génération :
  - Télescope de 8,4 m
  - Cerro Pachon (Chili)
  - Astronomie très grand champ : caméra 9,6□, Résolution 0,7"
  - A partir de 2020

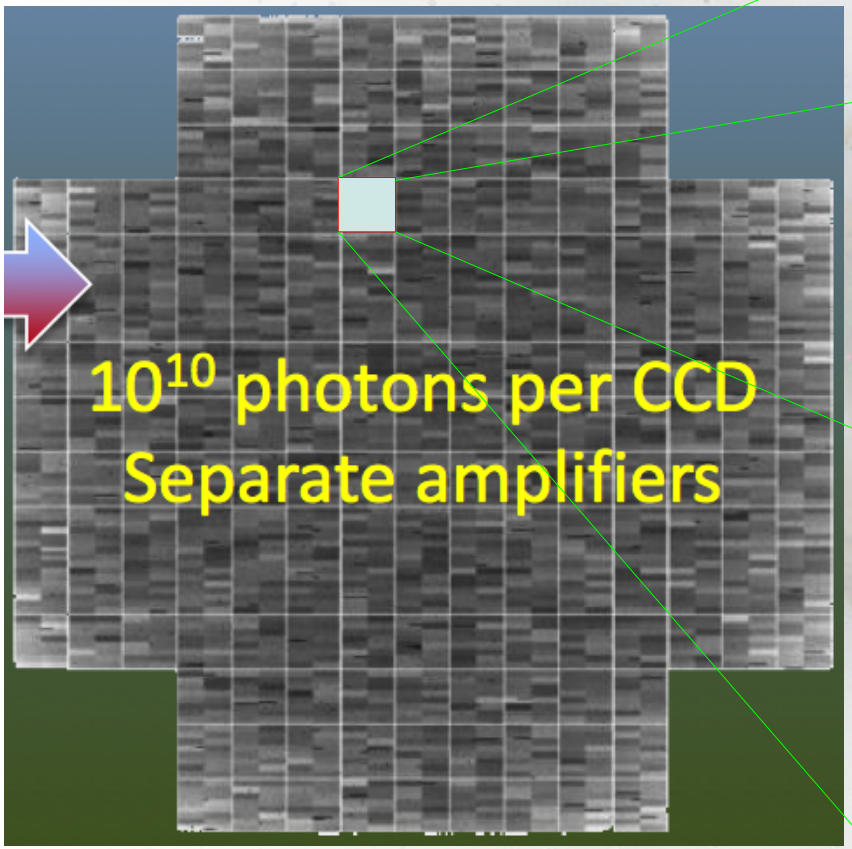


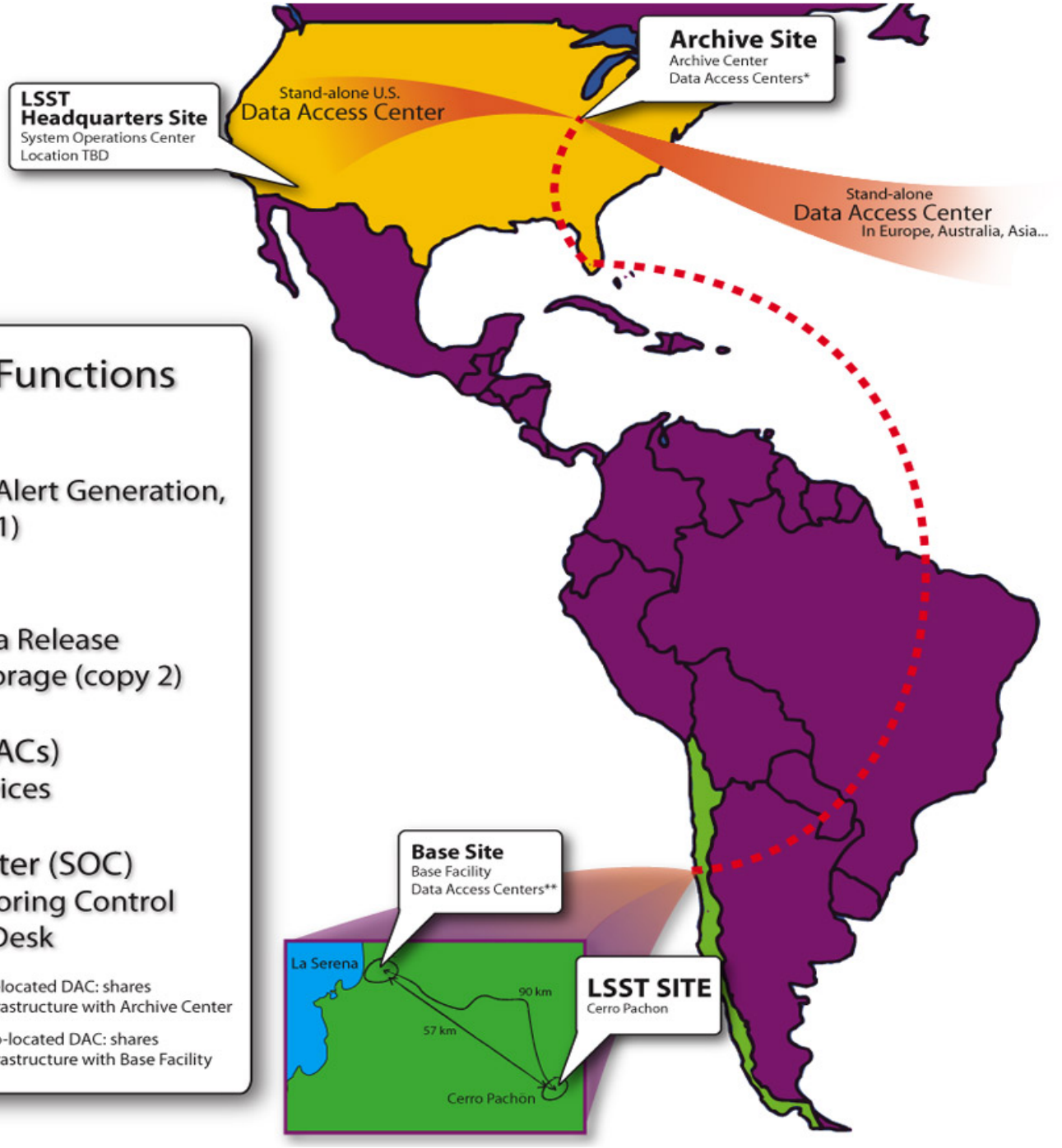


- Tout le ciel visible en 6 bandes (ugrizy) (20 000□)
- Poses de 15 s, 1 visite / 3 jours
- 10 ans, 60 Pbytes de données

Simulation des données :

~ 1/1 000 000 000 des données LSST !





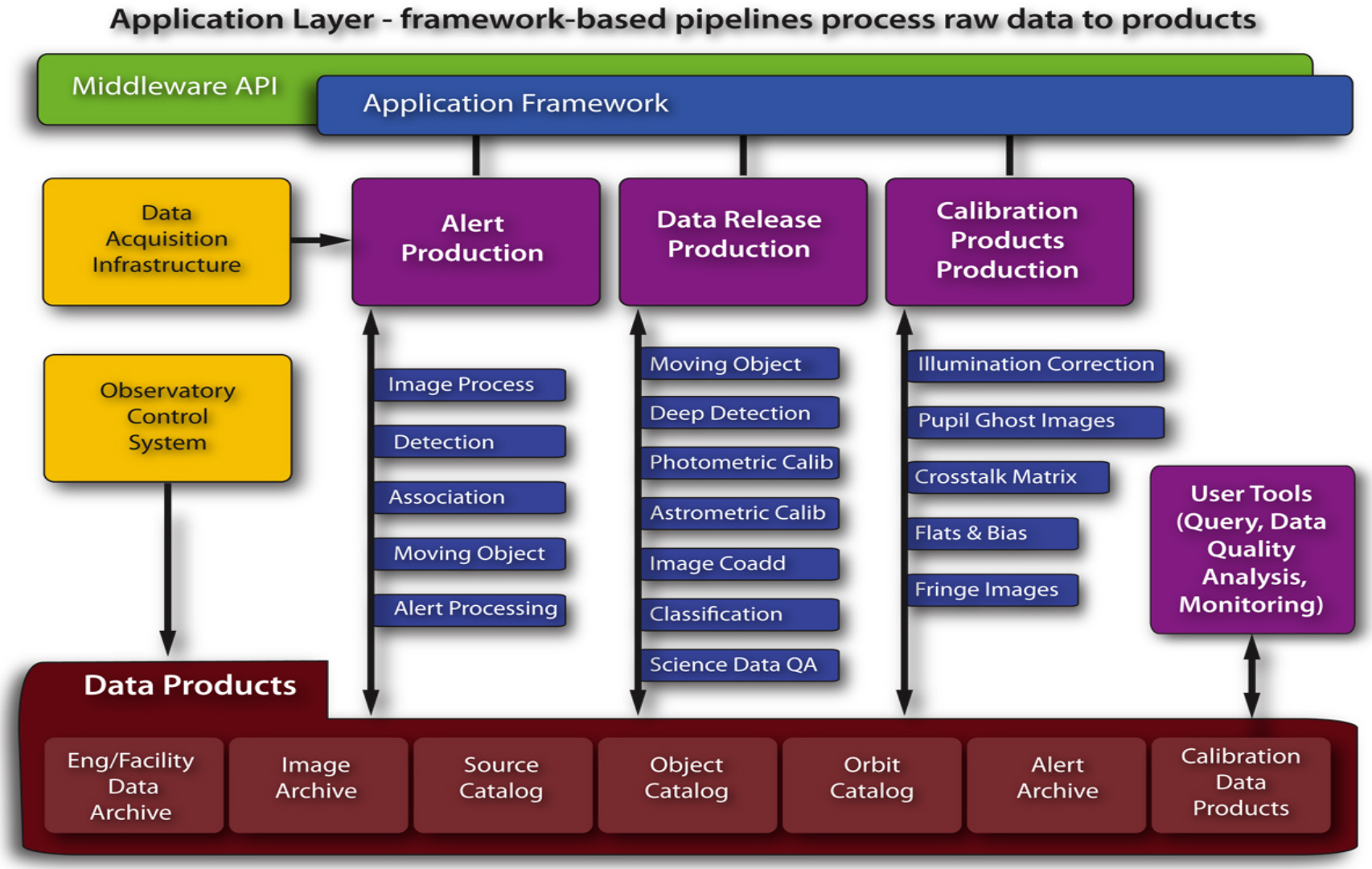
Site Roles and their Functions

- **Base Facility**
  Real-time Processing and Alert Generation, Long-term storage (copy 1)
- **Archive Center**
  Nightly Reprocessing, Data Release Processing, Long-term Storage (copy 2)
- **Data Access Centers (DACs)**
  Data Access and User Services
- **System Operations Center (SOC)**
  System Supervisory Monitoring Control & End User Support/Help Desk

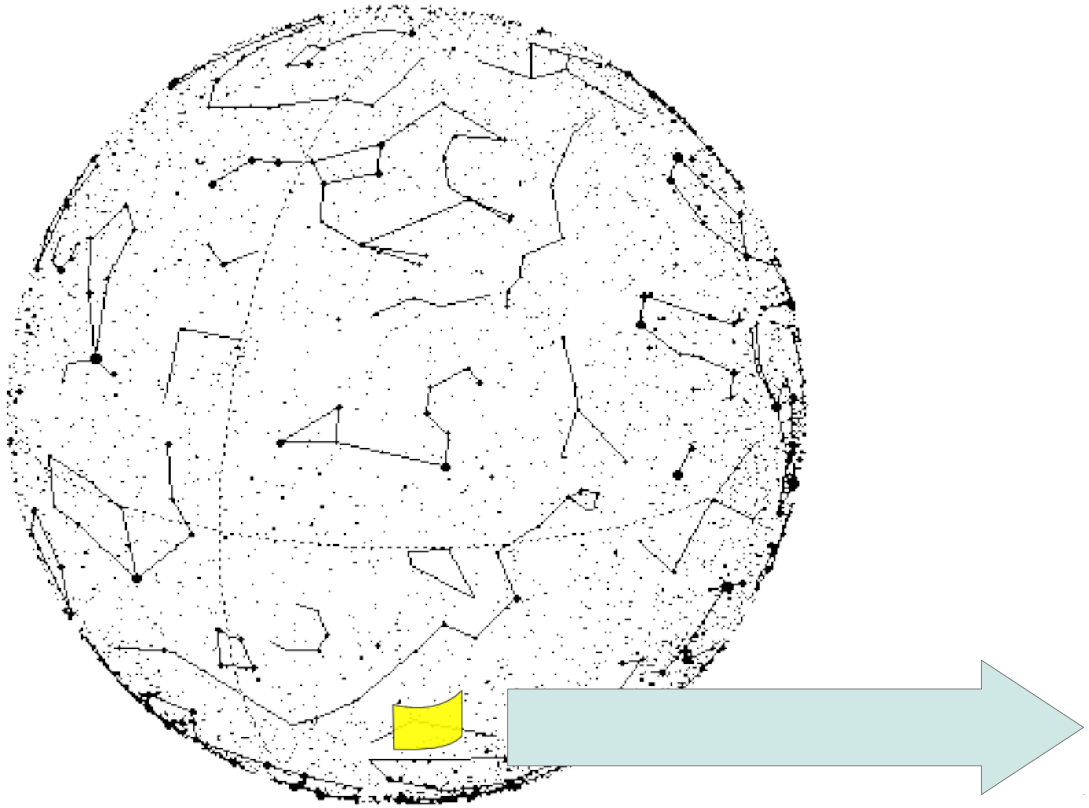
* Co-located DAC: shares infrastructure with Archive Center

** Co-located DAC: shares infrastructure with Base Facility

Couche applicative



Qserv: objectifs



Object
Source
...

Qserv : exigences fonctionnelles

SGBD relationnel, car les chercheurs connaissent bien le SQL.

https://dev.lsstcorp.org/trac/wiki/db/queries

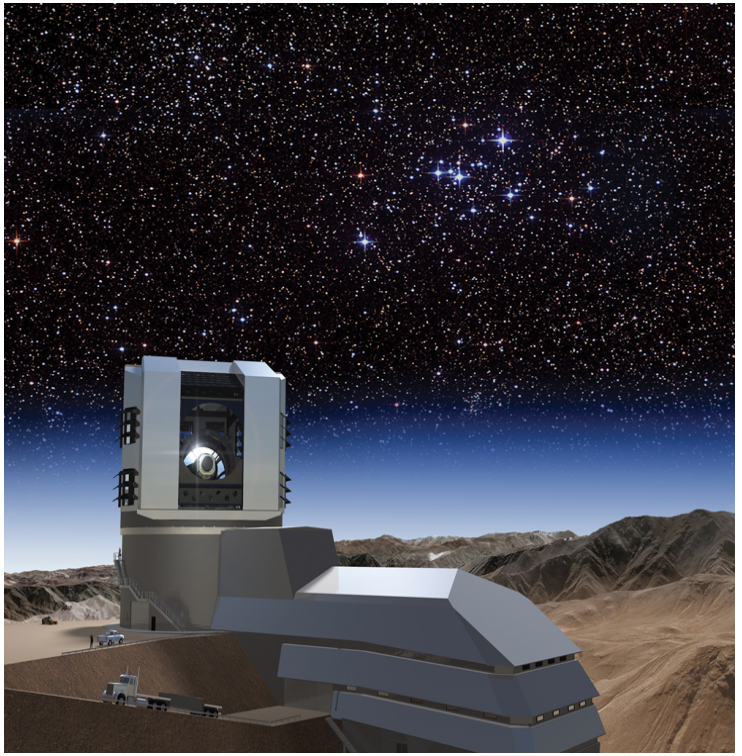
Optimisé pour pouvoir comparer des observations voisines.

La syntaxe SQL n'est pas entièrement supportée et les requêtes sont classées de 'simple' à 'impossible'.

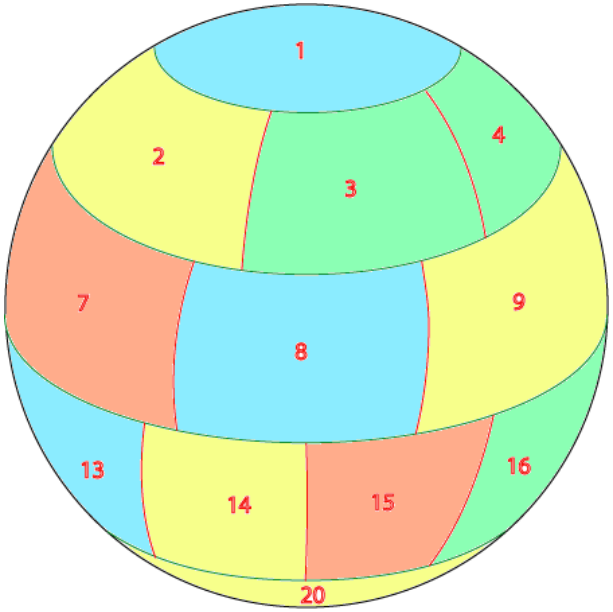
https://dev.lsstcorp.org/trac/wiki/db/queries/difficulty

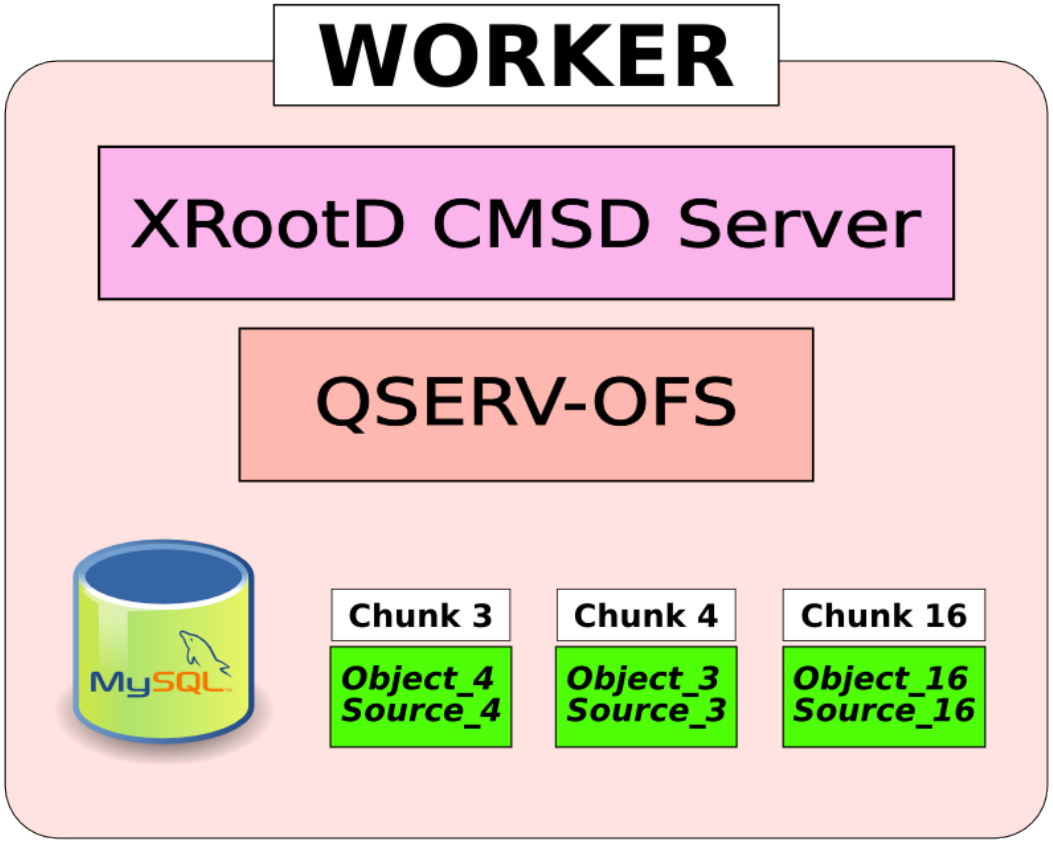
Qserv : exigences techniques

- Données distribuées en mode « Shared-nothing »
- Moteur d'exécution parallèle
- Disques locaux sur les esclaves
- Shared scans
- Open Source

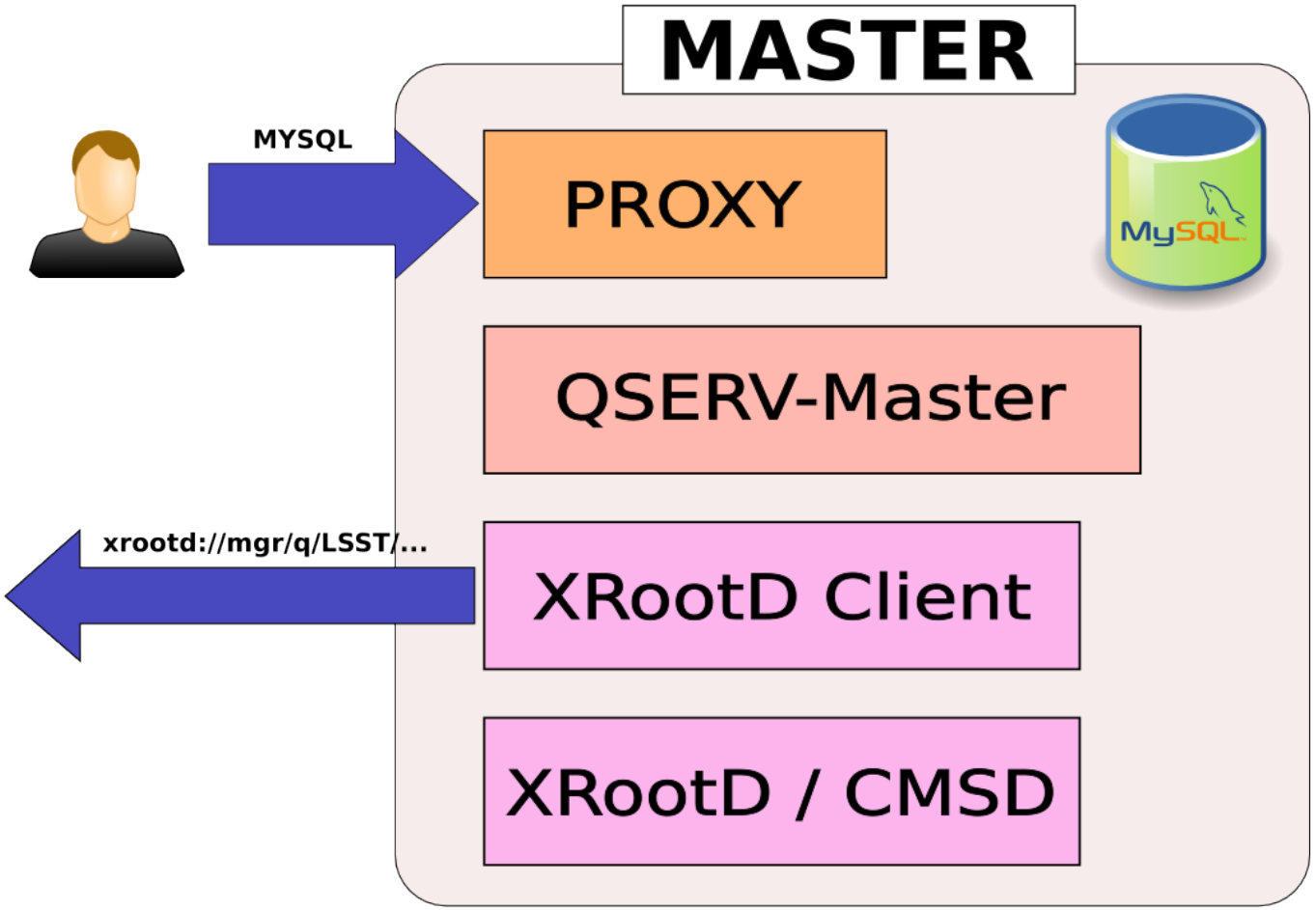


Worker Qserv

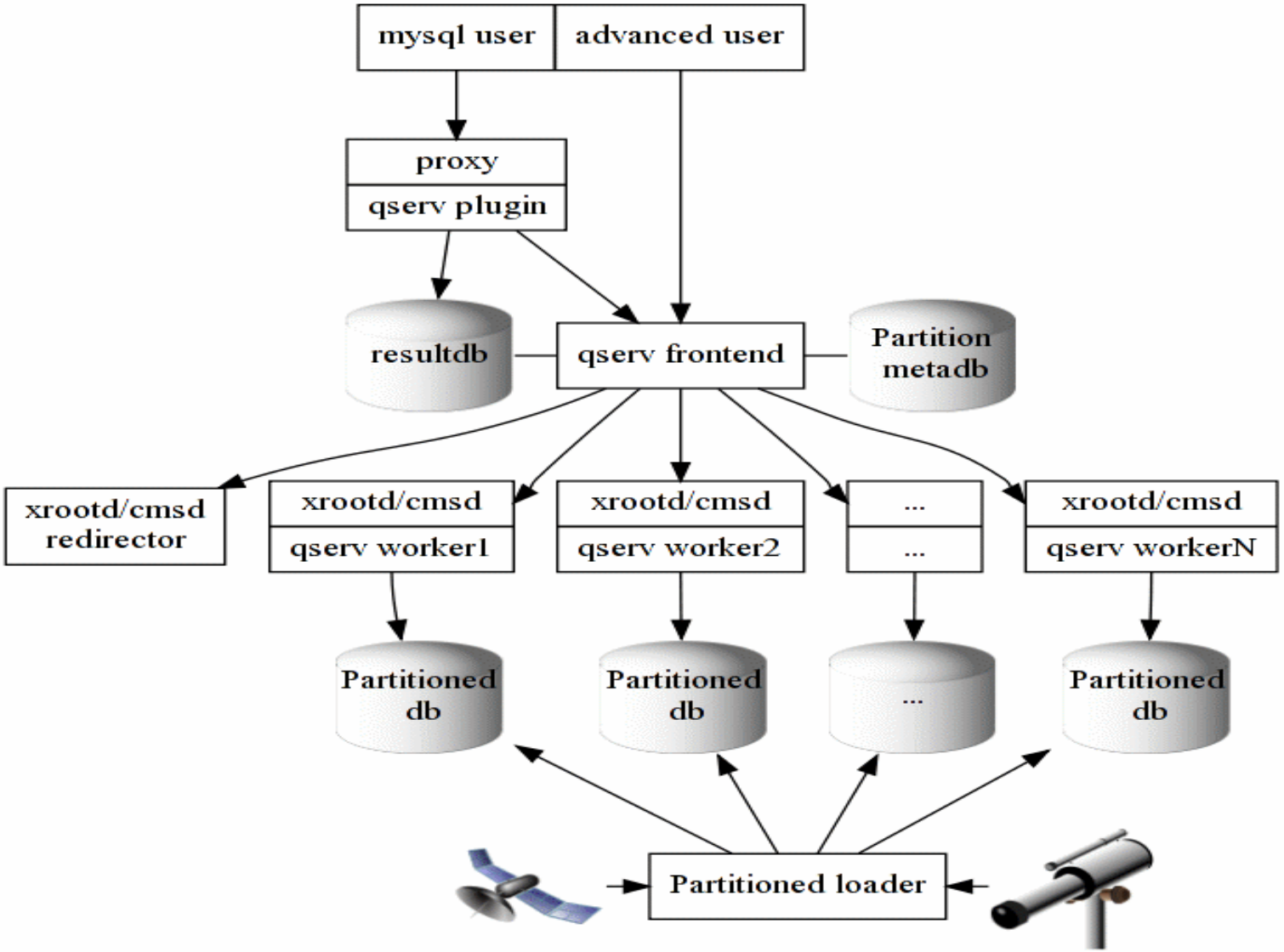




Master Qserv



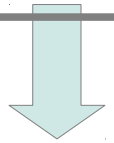
Au dessus de MySQL et xrootd



Réécriture de requêtes

MASTER

```
SELECT count(*)
FROM Object
WHERE qserv_areaspec_box(1,3,2,4)
   AND scisql_fluxToAbMag(zFlux_PS) BETWEEN 21 AND
21.5;
```



WORKER

```
SELECT count(*)
FROM Object_456
WHERE scisql_s2PtInBox(...)
   AND scisql_fluxToAbMag(zFlux_PS) BETWEEN 21 AND
21.5;
```

Avantages d'une solution ad-hoc

- Code bien adapté au besoin métier :
  - géométrie sphérique
  - recherche sur les proches voisins
- Maîtrise du code
- Contact étroit avec les développeurs Xrootd et MySQL (MariaDB)

Inconvénients

=> Les méthodes génériques implémentées dans les outils Map-Reduce sont à re-écrire :

- Partitionnement et chargement des données
- Supervision et gestion des pannes
- Répartition de charge et haute-disponibilité

En France : Petasky

Réponse à l'appel d'offre du défi de la MI du CNRS sur les BigData
2012 : Mastodons

*La disponibilité de très **grandes masses de données** et la **capacité de les traiter de manière efficace** est en train de modifier la manière dont nous faisons de la science.*

Consortium pluridiscuplinaire :

- **IN2P3 : LPC, APC, LAL, CC**
- **INS2I : LIMOS (Clermont-Ferrand), LIRIS (Lyon)**

=> Expertise + Financement (Infrastructure, missions)

Activités réalisées au LPC

Procédure d'installation automatique validée par SLAC

https://dev.lsstcorp.org/trac/wiki/db/Qserv/InstallAuto

Intégration continue (install, tests) au NCSA

Benchmarking (tuning MariaDB, jeux de données, requêtes SQL)

https://dev.lsstcorp.org/trac/wiki/mysqlLargeTablesAtIn2p3

Contribution à l'installation des 300 nœuds au CC-IN2P3 (3TB)

Passage à l'échelle

Machine de dév :

- 1 noeud, 20 MB (PT1.1, 1.2, W13)

Plate-forme de dév :

- 30 noeuds, 200GB

Plate-forme de benchmark :

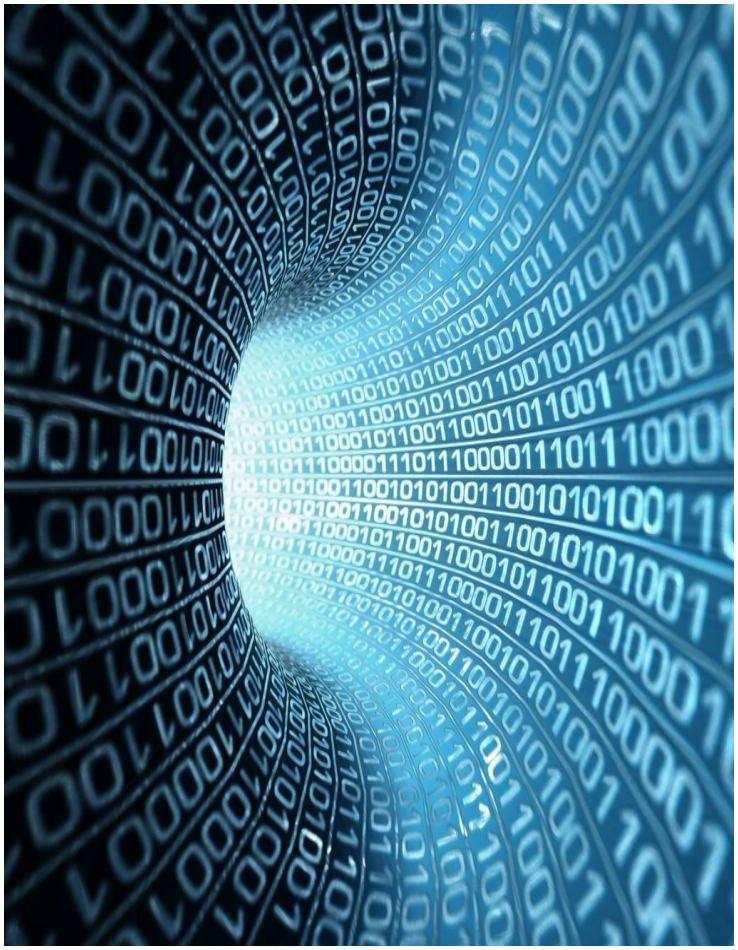
- 300 noeuds, 3TB

Lecons apprises

Difficulté de manipuler des données (via le réseau ou sur disque)

Adopter une démarche itérative :

- expérimenter sur 1 seul nœud
- travailler avec des petits jeux de données
- planifier une montée en charge très progressive (nœuds et données)



Questions/réponses

Remerciements à Emmanuel Gangler, Christian Arnaud, Dominique Boutigny et au service informatique du LPC.