
PetaSky
Emmanuel Gangler - LPC

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*

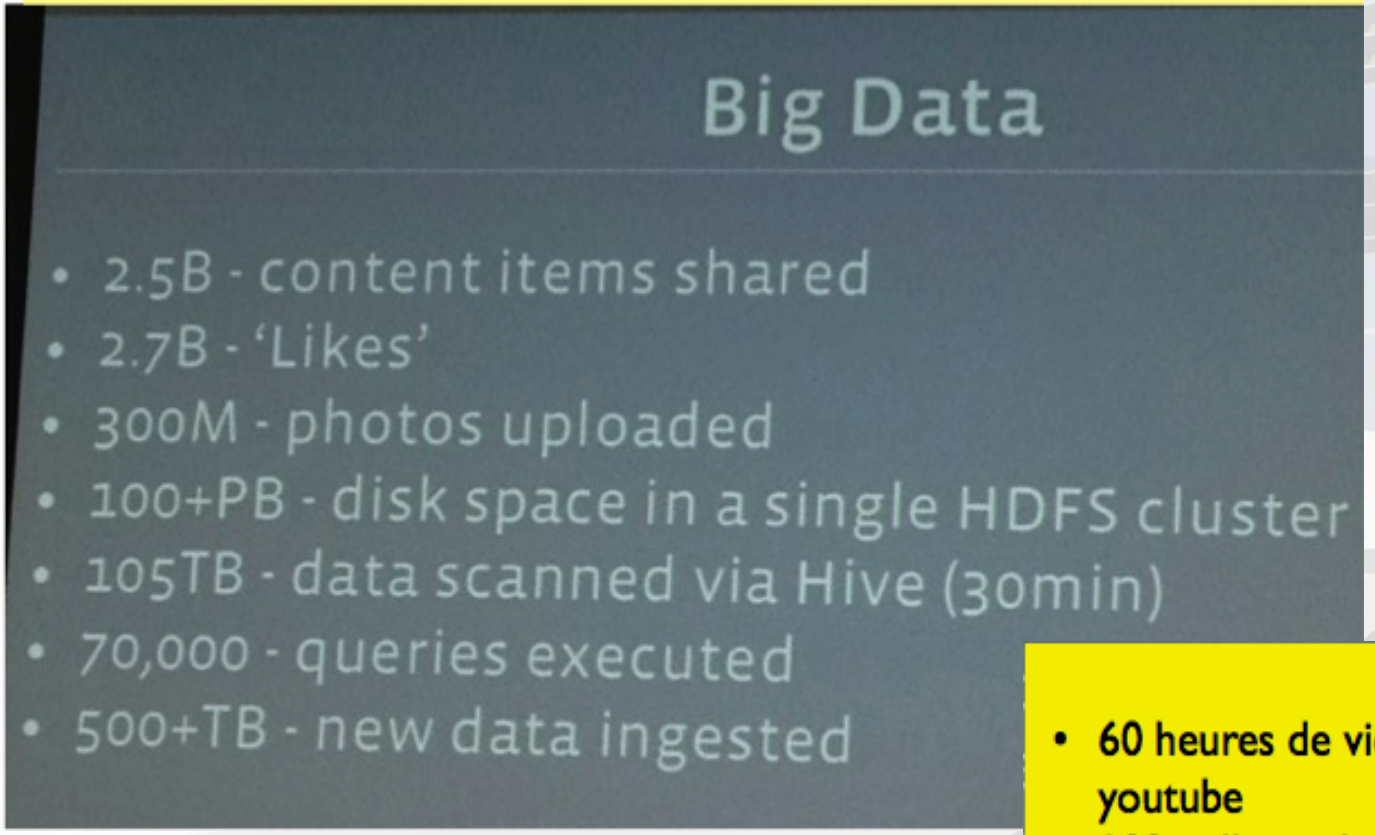
- Consortium pluridiscuplinaire :
  - IN2P3 : LPC, APC, LAL, CC
  - INS2I : LIMOS (Clermont-Ferrand), LIRIS (Lyon)

Position du projet

- Gestion de données astrophysiques de type LSST
  - Volume (~100 Pb en fin de projet, ~6 Pb de catalogues)
  - Complexité relationnelle (Objets, Sources, ...)
  - Hétérogénéité des formats (Images, catalogues)
- Problématiques
  - Passage à l'échelle et intégration de données (Axe 1)
  - Visualisation (Axe 2)
  - Analyse et fouille de données(Axe 3)

How big is big ?

Bigdata chez Facebook (Août 2012)



- Lire 1 TB : ~3H
- Objectif de Google : 1TB/s

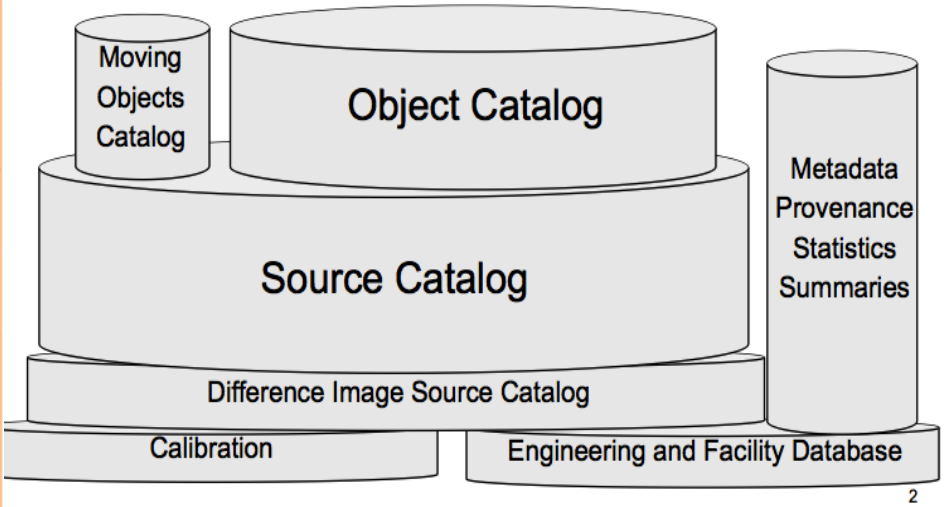
Google

- 60 heures de vidéo/minutes chargées sur youtube
- 100 millions de Go dans google search index
- 425 millions d'utilisateurs de gmail
- Google search crawler utilise 850 TB
- Google Analytics utilise 220 TB
- Google Earth utilise 70.5 TB

Axe 1 : stockage et gestion de données

(Mohand Said ; Emmanuel Gangler)

Table	Taille	#enregistrements	#colonnes (arité)
Object	109 TB	38 B	470
Moving Object	5 GB	6 M	100
Source	3.6 PB	5 T	125
Forced Source	1.1 PB	32 T	7
Difference Image Source	71 TB	200 B	65
CCD Exposure	0.6 TB	17 B	45



+ ~80 tables

- Quel modèle logique : relationnel, multidimensionnel, graphe ?
- Quel modèle physique : ligne, colonne, arbre, multidimensionnel ?
- Quelle architecture matérielle : partitionnement, multi-threading

Axe 1 : état des lieux

- Plateformes matérielles :
  - Au LPC : 5 machines, 1 maître + 4 nœuds, 10 TB/noeud
  - Au CC : 250 machines, 100 GB/noeud
- Logiciel :
  - Qserv (baseline LSST / MySQL) installé, data en cours de déploiement (100GB PT1.1 puis 2T W13)
    - Objectif = déploiement au CC pour FDR ; benchmarking
  - Benchmarks de Hive et HadoopDB (LIRIS)
    - Objectif = exploration des solutions alternatives
  - En discussion : comparaison avec solutions commerciales (Orange, Amazon)

Axe 2 : visualisation

(Florent Dupont, Guy Barrand)

- Visualisation d'images (et de données) de très grande taille
- Axes explorés :
  - Représentation des données ;
  - exploitation de la temporalité ;
  - interface aux catalogues.
- Exemple : découpage multi-résolution

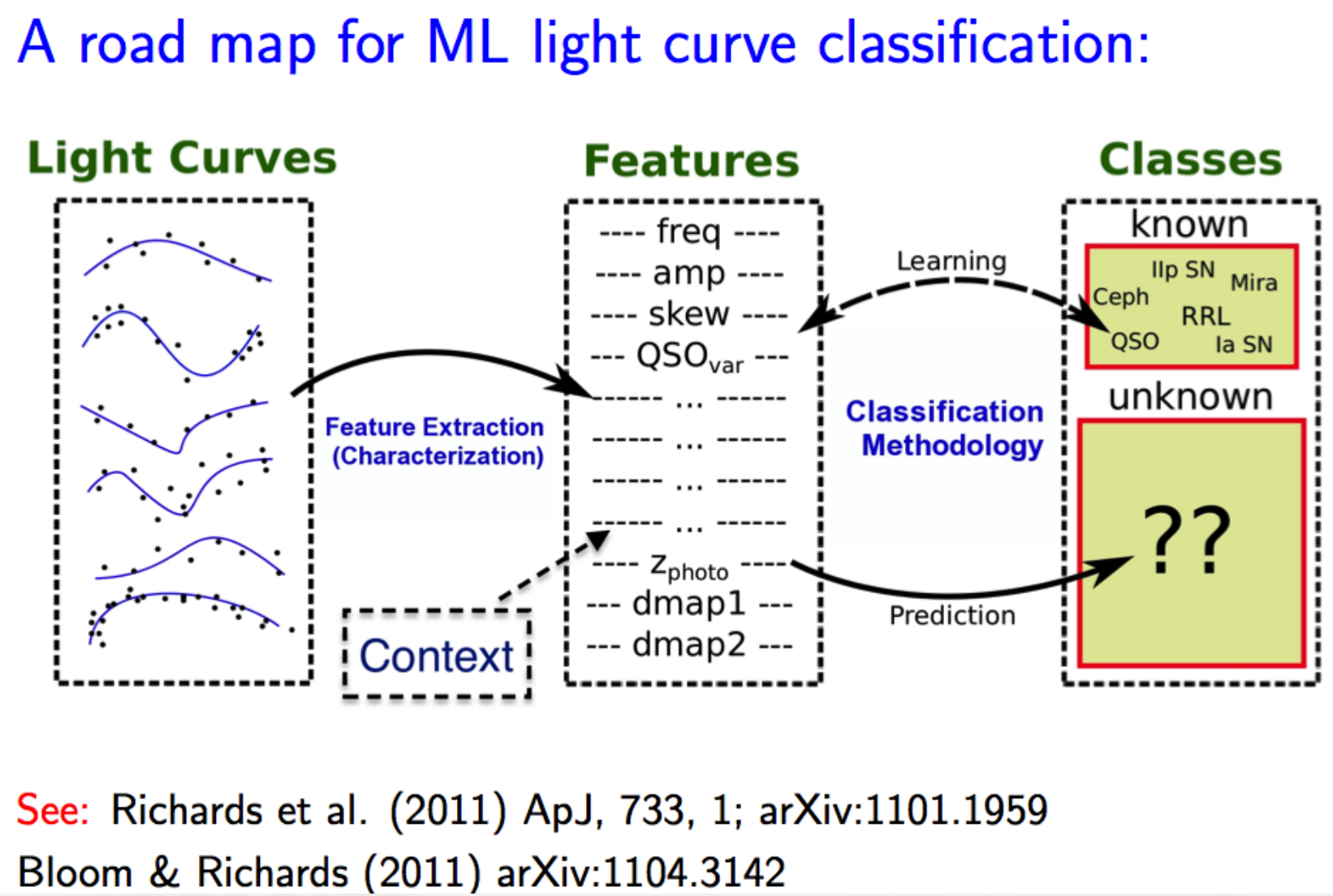
  http://liris.cnrs.fr/petasky/images/
- Plan de travail :
  - Solutions web
  - Mur d'images

Axe 3 : fouille de données

(Engelbert Mephu, Eric Aubourg)

- Objectifs
  - Classification des objets astrophysiques
    - Supervisée et non-supervisée
    - Caractérisation a priori ou non
  - Emergence de propriétés à partir des données
- Verrous
  - Passage à l'échelle
  - Caractérisation couplée à l'apprentissage
  - Incertitudes
  - Accès aux données (interface avec Axe 1)
- Feuille de route pour 2013
  - Séparation étoile-galaxie
  - Photo-z

Axe 3 : fouille de données



Succès du démarrage de PetaSky

- Projet sélectionné et financé (70 k€ 2012)
  - Plateforme de développement
  - Missions
- Apport de nouvelles approches
- Visibilité au CNRS