TERAPIX

Pipeline Terapix

Point de vue

Yannick

Préliminaires

- On ne fait pas un pipeline pour faire un pipeline mais pour produire et distribuer des données utilisables scientifiquement
- L'objectif ultime est la science que les produits de sortie du pipeline permettent de faire. Le pipeline et ses produits de sortie sont des aides aux chercheurs et il doivent répondre et s'adapter à leurs besoins/requetes. Sinon, les données ne seront pas utilisées, même si le pipeline est génial.
- La science qui doit être envisagée n'est pas nécessairement une science de routine, mais une science allant jusqu'aux limites
- Un produit Legacy se trouver en partie en contradiction à ce point de départ.

Les data d'entrée

- Contraintes: diversité des modes et des données
  - Data CFHTLS « Legacy » de nature différente: Deep/Wide/VeryWide
  - Data CFHTLS « optimisées » (SNLS, stacks cosmic shear, stack deep sub-arc-second)… science « aux limites »
  - Data PI : grande diversité; hétérogénéité des modes, des champs, de la qualité (certains programmes PI sont exécutés par le CFHT que si les conditions sont mauvaises…). Certains ne sont de fait pas intégrables dans un pipeline totalement automatique et l'intervention humaine est indispensable.
  - Data WirCam PI (pour le moment)
  - Data non Megacam (CFHT12K, UH8k)
  - Data non CFHT (WIFI, FORS, Omegacam, VISTA, SUBARU…)
  - Data multi-sites et/ou multi-bandes

Oublier le pipeline idéal

- Les data sont hétérogènes (conditions atmosphériques, pb instrument, modes d'observation très diverses, souvent imprévisibles ou inadaptés, contenu des méta-données hétérogènes ou incomplètes)
- Les logiciels de processing ont des bugs ont ne sont pas prévus à l'avance pour tous les modes
- Le matériel n'est pas stable à 100% et des pannes machines, électriques, réseau sont inévitables
- Les gens qui définissent les spec. et les besoins ne sont pas géniaux (et on le sait à l'avance)
- La production des données se fait avec des dates butoirs définies par des extérieurs (SAC, SG, Board, utilisateurs) qu'on DOIT prendre en compte
- Les ressources sont limitées

Pour répondre à ces contraintes

- Fléxibilité de fonctionnement
- Modularité des tâches/blocs avec des étapes qui permettent de revenir sur ses choix et proposer des alternatives
- Intervention autorisée des chercheurs pour optimiser ses choix de sélection des data et du processing
- Pas d'irréversibilité qui impliquerait qu'une fois lancé on doit attendre la fin ou tout stopper
- Le point clef est l'interface permettant des sélection, la description et l'accès aux données à chaque étape clef

Le choix

- Un web service?
- Mais est-ce gérable? (espace disque, CPU, réseau)
- Ne répond pas à l'aspect « release » officielles du CFHTLS. Besoin de critères « officiels »
- Le dernier point signifie dans tous les cas un mode pipeline automatique. MAIS il doit être flexible pour tenir compte des inattendus

Le choix

- Le concept de 3 STEP conservé:
  - STEP1: Analyse images individuelles, cartes de poids (QFITS-in pages, notes, weightmaps, catalogues)
  - STEP2: Calibrations, stacks monochromatiques (QFITS-out, Q-assessments plots, images+weigthmaps, catalogues)
  - STEP3: Productions panchromatiques: images chi2 ou Schi2, merged catalogues, Q-assessments plots
  -
- Des outils d'analyses statistiques globales manquent (e.g. fonction corrélation)
- Les outils de suivi des surveys + analyse statistiques globales sont insuffisants, ou pas prêts

Quelques suggestions

- L'interface avec la DB: trop rigide, on ne doit pas avoir à tout relancer ou attendre que tout soit fini à chaque fois d'un crash se produit. Une BD intermédiaire?
- Une option de sélection manuelle, une fois que les sélections de Laurent sont faites.
- Une étape Scamp découplée de Swarp? Il faudrait avoir des .head qui soient valides quels que soit le Swarp fait après. Idée: Un .head PEUT IL ÊTRE construit par défaut (pour les releases) uniquement à partir du champ central et de ses huit zones Megacam les plus proches voisines, ce .head est construit en même temps pour tous les champs ayant même RA,DEC+/- 7' ? Actuellement la distribution d'une calibration astrométrique n'a pas de caractère intrinsèque à l'image associée
- L'interface pour l'utilisateur n'est pas de la poudre aux yeux, même si elle est une apparence, c'est un point crucial qui détermine la popularité de l'accès
- Réfléchir à un parc de machines avec un fonctionnement plus stable. Machines/OS moins « à la pointe » mais dont la robustesse est validée?

Echelle de temps

- T0004 lancé le 15 septembre
- STEP1 terminé, data délivré au CADC le 15 octobre
- STEP2 terminé pour D, W, VW le 20 novembre
- STEP3 + Release candidate Deep+Wide le 10 décembre
- T0004 terminé le 21 décembre.
- Développement new pipeline: une fois T0003 terminé. Goal: 15 avril.