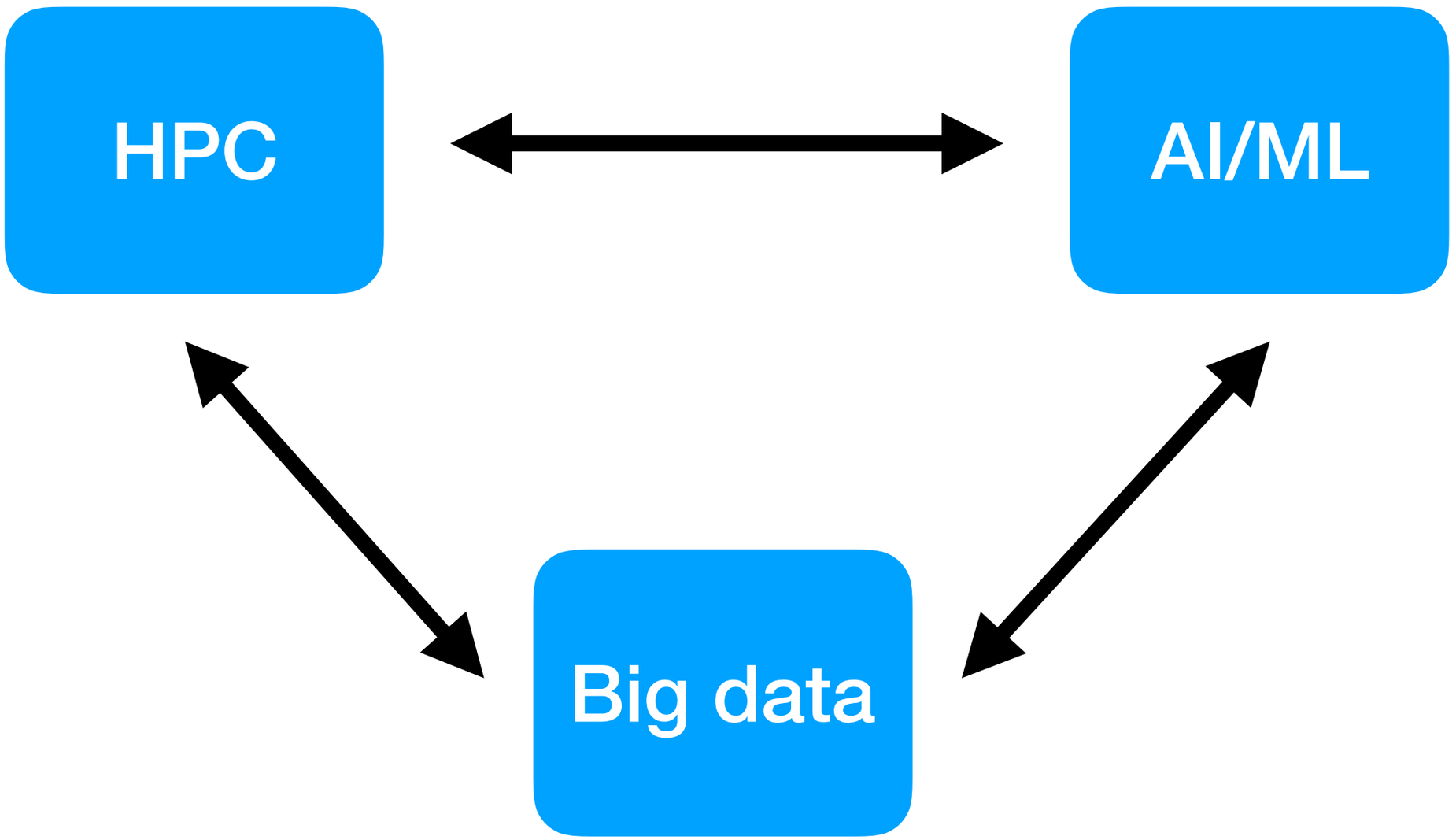
HPDA

Éric Aubourg • 20 décembre 2018

High performance data analytics



Contexte

- Traitement de données LSST (100 Po sur 10 ans)
- Images sur système de fichier distribué et catalogues/ métadonnées dans une base de données distribuée

- Projet : traitement d’images avec réseaux de neurones (bayesiens ?), MCMC.

Quels outils ?

- Tensorflow / Tensorflow probability : deep learning, réseaux bayesiens, MCMC… ; Keras (PyTorch ?)
- Apache Spark sur HDFS
- Apache Hive / Google BigQuery / AWS Athena & Redshift

Quels types de machines ?

- Rien de figé… Flexibilité !
- Machines avec GPU (NVidia Tesla V100) ou TPU, RAM>=60 Go
- Machines multiCPU/RAM++ (96 cœurs, 768 Go)
- Ferme HDFS (HDFS+GPU pour Spark+TensorFlow ?)
- Besoins très fluctuants.

Où ?

- Virtualisation, containers : Docker, Singularity ? Souplesse, approche DevOps, portabilité, indépendance par rapport aux configurations pré-installées…

- Utilisation du CC-IN2P3, du cloud Amazon (bourse de $27k l'an dernier), du cloud Google lors de hack days LSST.

    - Au CC pour le moment GPU moins puissants (K80, V100 courant 2019), moins de RAM. Migration AWS vers CC pas si simple…

    - Avec checkpoint sur les jobs, amazon tarif spot : V100 (8 cœurs, 60 Go) à $1/heure. Envisagé en attendant les V100 au CC en cas de besoin.

- Hack week LSST/DESC à l'automne 2018 : pic à 6 machines V100 sur AWS. Besoin de flexibilité.