SPARK

pour la bioinformatique

Ludovic.Legrand@toulouse.inra.fr

Apache Spark

Framework de calcul distribué

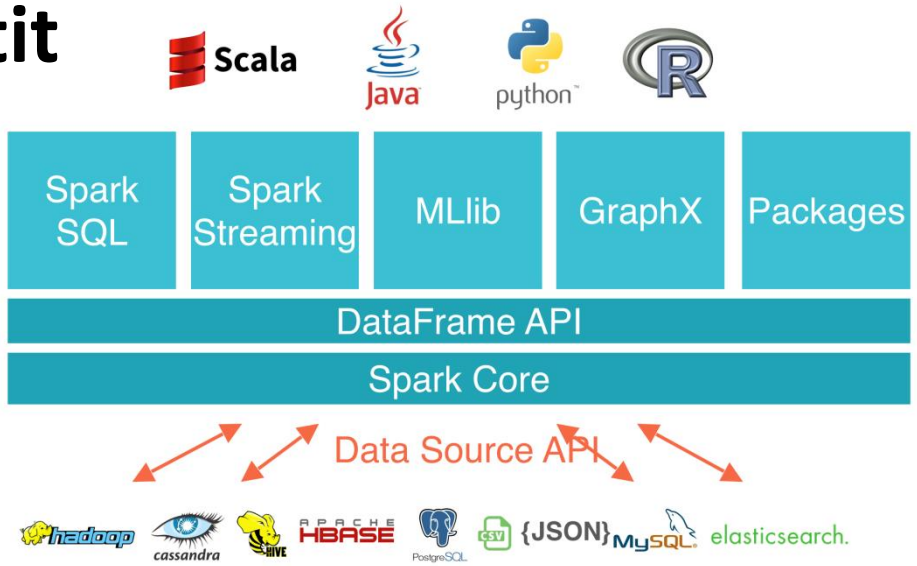
- né à Berkeley en 2009
- Top-level project Apache depuis 2013

Intégré à l'ecosystème Hadoop

- YARN/Mesos, HDFS, Hbase/Cassandra/Elastic …

Remplace Hadoop petit à petit

- Jusqu'à 100x rapide (RAM)
- API simple et ouverte
- Traitement interactif



Projet

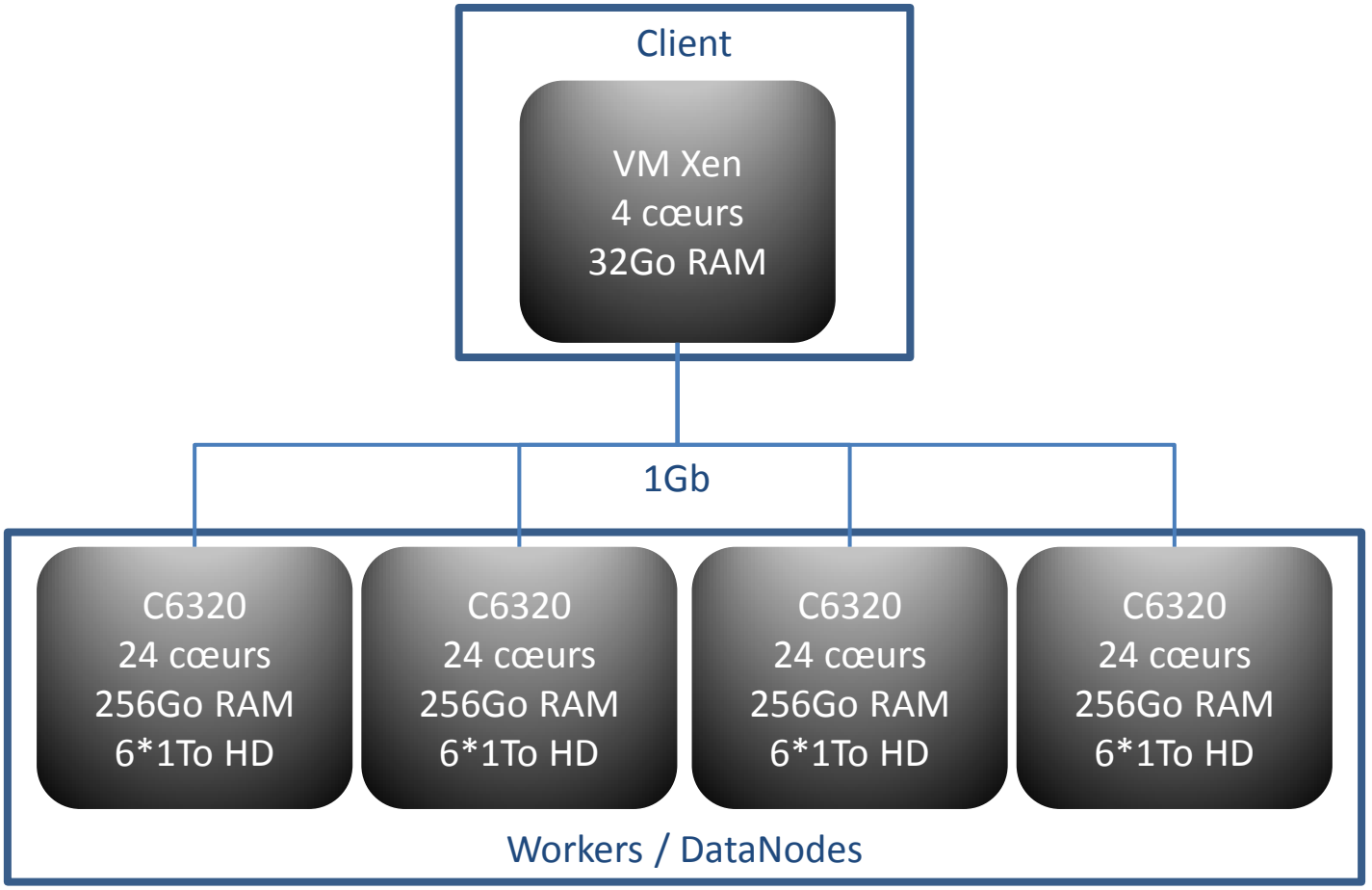
Evaluer Spark dans le cadre de la bioinformatique

- Evaluation d'outils déjà existants
- Développer des outils pour accélérer nos traitements
- Evaluer le paradigme MapReduce

Adéquation avec la bioinformatique

- Traitement de type pipeline
  - Ecriture/lecture sur disque important
  - Enchainement de traitement
- Manipulation de fichiers texte multiligne structurés

Infrastructure



Ressources cluster « utiles »

- 176 hyperthreads
- 650 Go RAM « overhead 350 Go »
- 6 To

Configuration

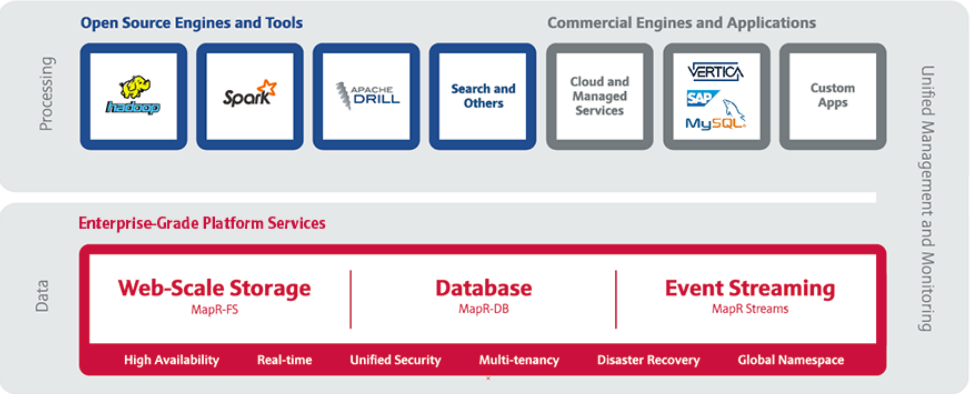
- Réplication 3x
- 20*1 To pour les données
- 4 * 1 To pour le système
- Carte RAID mode HBA

Distributions Hadoop

3 concurrents: MapR, Cloudera, Hortonworks

MapR (Open source edition)

- Intégration de l'ecosystème Hadoop
- Performant et user friendly
- MapR-FS
  - Système de fichiers compatible Hadoop
    - API HDFS, métadonnées distribuées
  - Compression des données
  - Système de fichiers POSIX
    - Export NFS
    - création/édition/délétion



Installation MapR

Installation

- Simple et automatique si sous Ubuntu/Centos/Redhat
- Dépôt logiciel (apt/yum)
- Interface web

Les problèmes rencontrés

- « Bidouiller » le script d'installation sous Debian
- Attention aux « locales » (en_US.UTF-8)
- Documentation dense mais manque d'exemples pratiques
- Difficulté à installer un client sur le frontal

Nos attentes

Accélérer l'analyse des données

- Suivre le débit de la production des données
- Mieux exploiter les données déjà produites

Infrastructure à faible coût et généraliste

- « Commodity hardware »
- Faible coût d'administration
- Mixer « Big Data » et cluster classique

Perspectives

Court terme

- Benchmark et prise en main des outils existants
  - Recherche de variants génomiques
  - Comptage de mots
- Prise en main de l'API Scala et développement d'outils *ad hoc*
  - Accélérer l'assignation taxonomique
- Evaluer les outils d'exploration
  - Spark SQL et Drill pour le SQL
  - GraphX pour les graphes

Moyen terme

- Evaluer des outils de Machine Learning (SparkML)
- Evaluer Mesos

Partenaires et financement

INRA pour le financement du matériel

Les CATIs

- BBRIC
- MIAGO
- CGI
- BIOS4BIOL