UNIVERSITÉ DE TECHNOLOGIE DE BELFORT-MONTBÉLIARD

Big Data en Astronomie

Étude et implémentation de Hadoop et Spark, application au Cross-Match du CDS

Noémie Wali – Département Informatique

ST40 – Stage assistant ingénieur

Automne 2015

Tuteurs en entreprise
André Schaaff
François-Xavier Pineau

Enseignant suiveur UTBM
Nicolas Gaud

utbm

Observatoire astronomique de Strasbourg

Plan

- Présentation de l'Observatoire
- Sujet du stage
- Travail réalisé
- Résultats obtenus
- Conclusion

Présentation de l'Observatoire

Observatoire astronomique de Strasbourg

Crédits : topic-topos.com

- Fondé en 1881
- 3 équipes : Hautes énergies, CDS, Galaxies
- 80 personnes
- Objectifs : recherche, enseignement, services d'observations et diffusion des connaissances

Présentation de l'Observatoire

Centre de Données astronomiques de Strasbourg

- Créé en 1972
- Equipe de 30 personnes : documentalistes, informaticiens et astronomes

Objectifs :

- Collecte, enrichissement et distribution des données
- Recherche

- Membre fondateur de l'International Virtual Observatory Alliance (IVOA)

Présentation de l'Observatoire

Principaux services du CDS

Collection de données sous forme de catalogues
14 000 catalogues

Atlas interactif du ciel (images astronomiques numérisées) – 200 To de données

Base de données d'objets astronomiques

- 8 000 000 d'objets
- 314 136 références bibliographiques
- 12 839 269 citations d'objets dans les articles

Aujourd'hui : 1 000 000 de requêtes / jour sur l'ensemble des services

Sujet du stage

Étude et implémentation de Hadoop et Spark, application au XMatch du CDS

- Contexte : Big Data en astronomie

- Missions :
  - Technologies Hadoop et Spark d'Apache
  - XMatch du CDS (corrélation croisée)
  - Évaluation de l'architecture

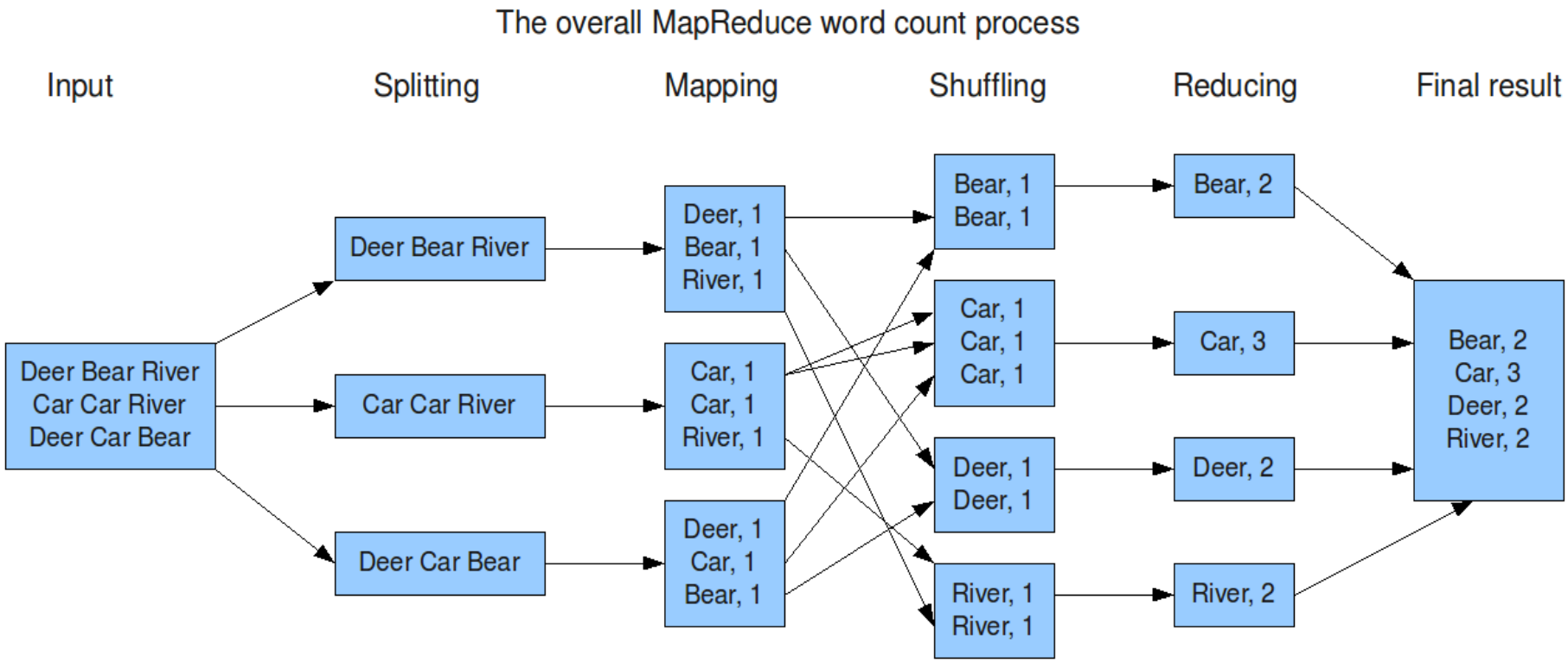
Travail réalisé

Apache Hadoop

- Framework libre de calcul distribué
- Hadoop propose notamment :
  - Système de fichiers distribué HDFS (*Hadoop Distributed File System*)
  - Architecture MapReduce

Travail réalisé

Fonctionnement du MapReduce - exemple



Crédits : Grégory PAUL

Travail réalisé

Spark

- Framework de calcul distribué
- Utilisation de la mémoire
- RDD = Resilient Distributed Dataset
  - Collection distribuée de données
- Interfaces pour Scala, Java, Python et R

Travail réalisé

Implémentation :

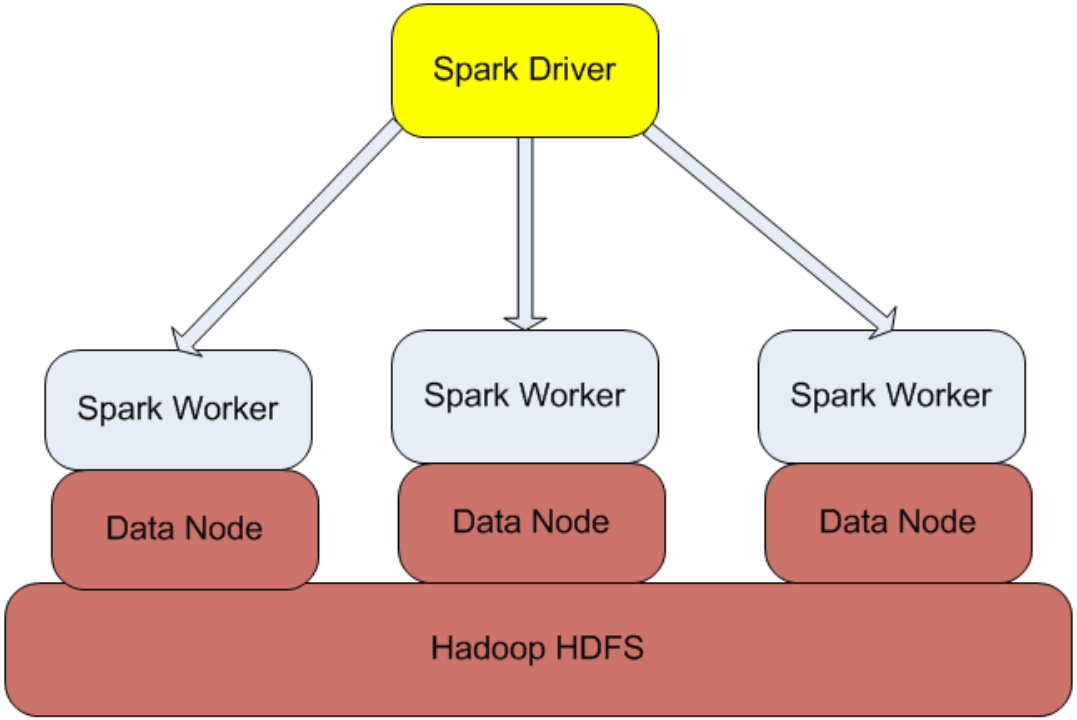
- Spark
- HDFS

Principales raisons :

- Flexibilité
- API Java

Travail réalisé

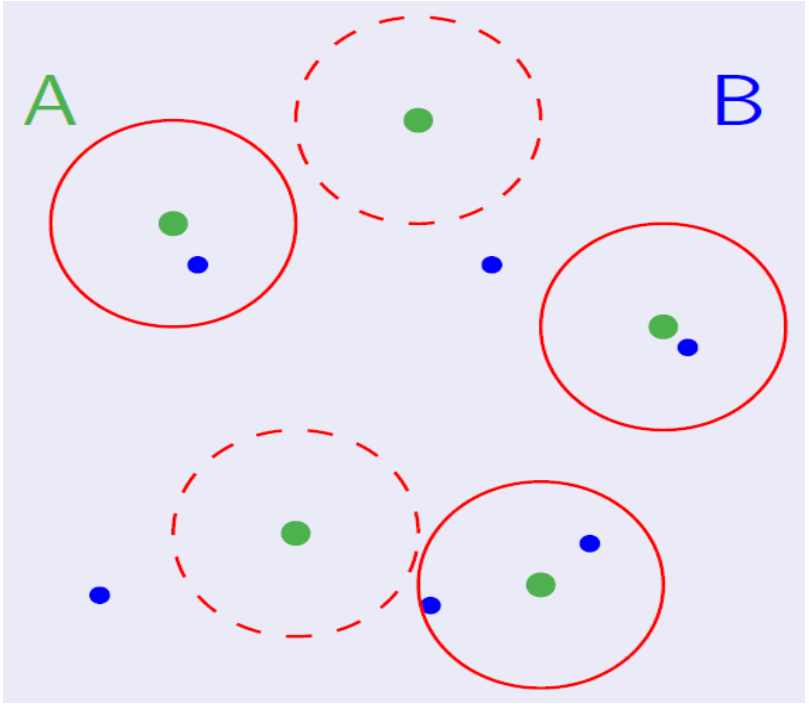
Installation et configuration de Spark et HDFS sur les machines



Crédits : BigHadoop

Travail réalisé

XMatch (corrélation croisée) :
Jointure floue entre 2 tables de plusieurs centaines de millions de données

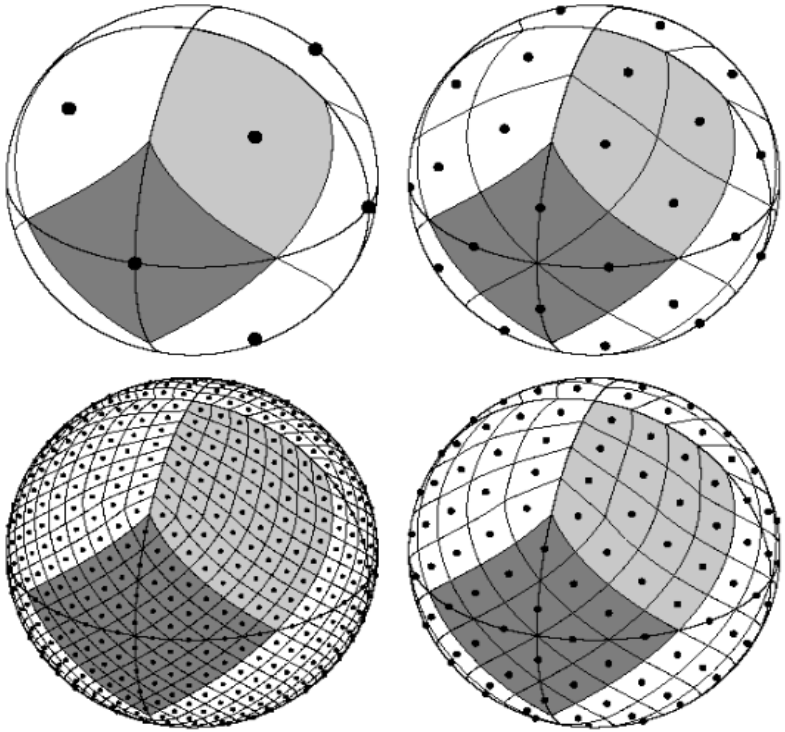


Travail réalisé

Une implémentation du XMatch en MapReduce

Couples (clé = n° de pixel, valeur)

Découpage HEALPix du ciel



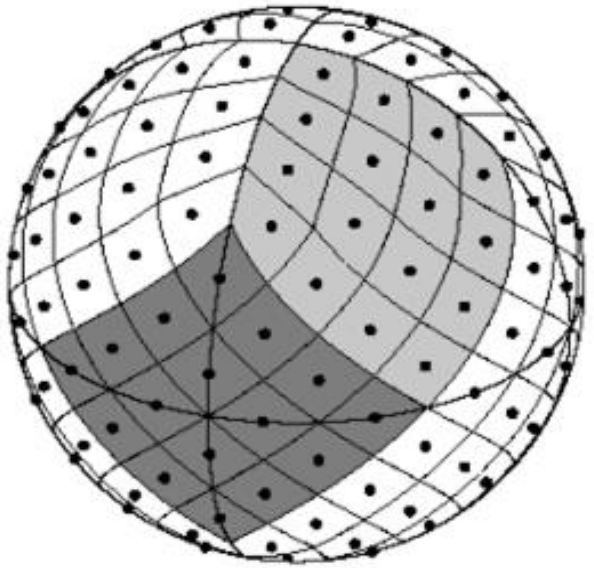
Crédits : HEALPix – arXiv:astro-ph/0409513

Travail réalisé

Effets de bord

-> Jointure floue

-> Duplication des sources dans les cellules voisines si besoin



Crédits : HEALPix – arXiv:astro-ph/0409513

Co-location des données

-> Enregistrement des clés communes à 2 RDDs sur les mêmes nœuds

Résultats obtenus

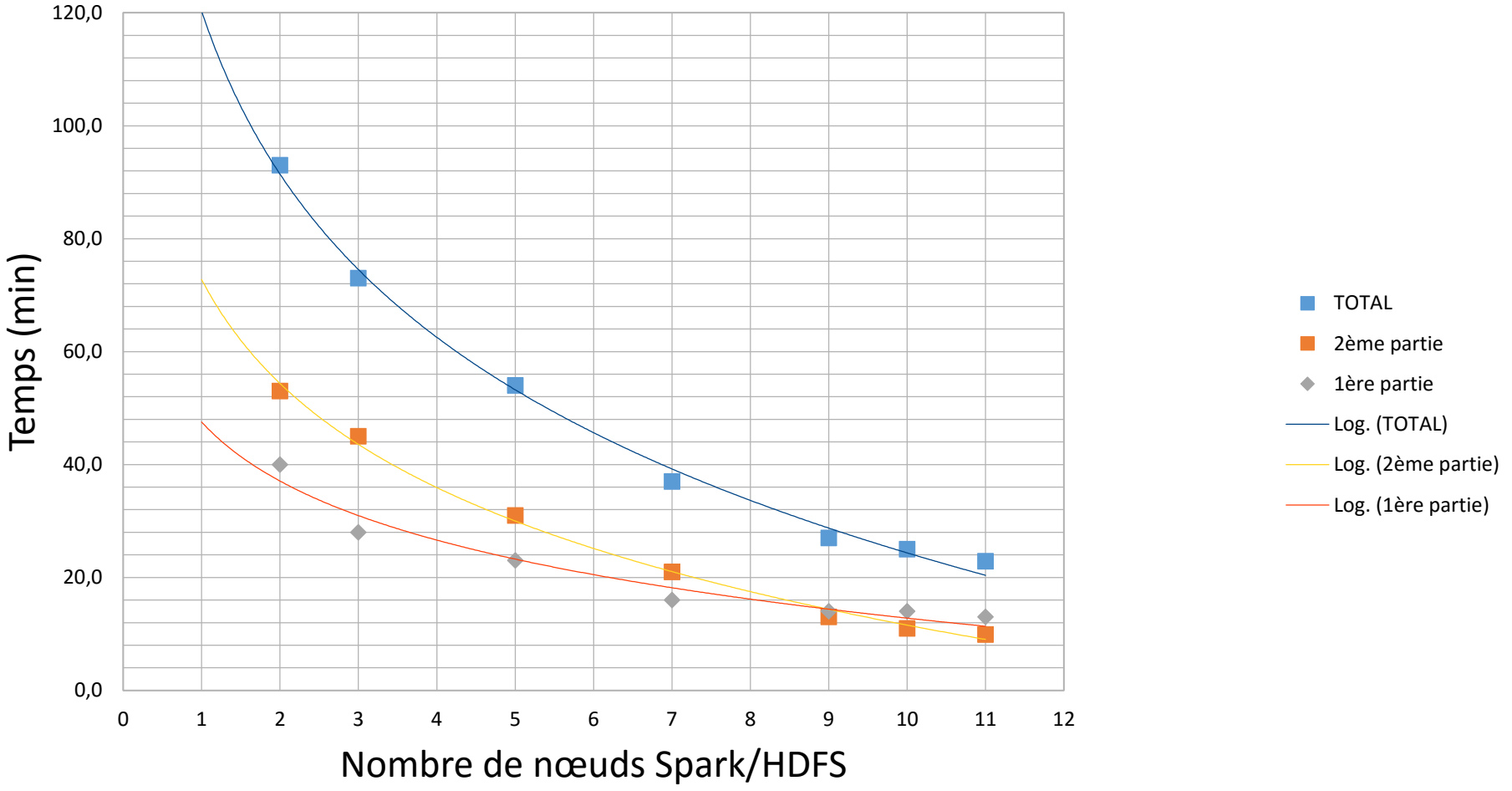
Données en entrée : fichiers de 54GB et 58GB ; 355 000 000 et 470 000 000 d'objets
Données en sortie : 49 208 820 d'éléments en sortie

RAM des serveurs OVH : 32Go

Cross-Match (duplication des sources faite dans la 2e partie ; avec toutes les données en sortie)											
Taille des blocs HDFS = 128MB pour les fichiers en entrée ; sdss7.csv et 2mass.csv répliqués 2x											
HashPartitioner	60 partitions										
Taille des blocs HDFS en sortie	32MB										
Nombre de nœuds Spark/HDFS	1	2	3	4	5	6	7	8	9	10	11
1ère partie : préparation des données		**40,0**	**28,0**		**23,0**		**16,0**		**14,0**	**14,0**	**13,0**
mapToPair (sdss7.csv)		7,8			5,1		4,9		4,9	4,8	4,7
saveAsHadoopFile (sdss7.bin)		10,0			5,7		2,7		2,0	2,3	1,5
mapToPair (2mass.csv)		8,5			5,7		5,2		5,2	5,1	5,0
saveAsHadoopFile (2mass.bin)		13,0			6,5		3,6		1,9	1,6	1,4
2ème partie : jointure		**53,0**	**45,0**		**31,0**		**21,0**		**13,0**	**11,0**	**9,9**
mapToPair (sdss7.bin)					7,2		4,7		3,5	3,0	2,9
flatMapToPair (2mass.bin)					11,8		8,3		5,5	4,9	4,3
saveAsTextFile (crossMatch_D.txt)					12,0		7,6		3,4	2,4	2,3
TOTAL		**93,0**	**73,0**		**54,0**		**37,0**		**27,0**	**25,0**	**22,9**

Résultats obtenus

Temps de XMatch en fonction du nombre de nœuds



Conclusion

- Alternative au XMatch du CDS
- Résultats obtenus correspondent aux objectifs fixés
- Optimisation possible du code par la co-location des données
- Maîtrise des technologies, recherche et autonomie

Merci de votre attention

Noémie Wali - Département Informatique

ST40 – A15 – Big Data en astronomie

utbm

Observatoire astronomique de Strasbourg