Recherche et développement autour de nouvelles technologies pour la manipulation de grandes masses de données

Présentation de l'entreprise

- Observatoire Astronomique de Strasbourg

- Chargé d'histoire : inauguré en 1881, riche patrimoine d'instruments et d'ancien ouvrages
- Unité mixte de recherche entre Université et CNRS
- Structuré en **trois équipes de recherche** et deux services d'observation de l'Institut National des Sciences de l'Univers (INSU)

Présentation de l'entreprise

- L'observatoire a pour mission de contribuer aux progrès de la connaissance par :
  - L'acquisition des données d'observation
  - L'élaboration des outils théoriques nécessaires
- Egalement chargé :
  - d'assurer la formation des étudiants et du personnel de recherche
  - d'assurer la diffusion des connaissances
  - de prendre part à des activités de coopération internationale

Présentation de l'entreprise

- Les 3 équipes de recherche :
  - Galaxies : étude de la formation et de l'évolution des galaxies ainsi que de la dynamique des étoiles et de la matière noire
  - Hautes Energies : s'intéresse aux sources galactiques et extragalactiques émettrices en rayons X
  - Centre de Données astronomiques de Strasbourg: équipe de recherche et service d'observation

Présentation de l'entreprise - CDS

- Créé en 1972 – Actuellement 33 personnes
- Héberge la base données de référence mondiale pour l'identification d'objets astronomiques
- Ses missions :
  - Rassembler des informations sous forme informatisée
  - Mettre à jour ces données en le comparant
  - Distribuer ces données à la communauté internationale
  - Mener des recherches en utilisant ces données

Présentation de l'entreprise - CDS

- 3 services principaux:
  - Simbad :
    - Base de données de référence pour la bibliographie des objets astronomiques situés hors système solaire.
    - 7 millions d'objets cross-matchés parmi 18 millions d'identifiants
  - VizieR :
    - Base de données qui regroupe plus de 11 000 catalogues d'objets célestes constitués de données relevées pendant des missions
  - Aladin :
    - Atlas interactif du ciel permettant de visualiser des images astronomiques

Présentation du stage

Problématique:

- Evolution des volumes de données, R&D autour de l'hébergement et de l'accès à ces données

Objectif:

- Tester des nouvelles technologies dans ce contexte

Présentation du stage

La première partie du stage:

- Mettre en oeuvre Hadoop et tester son utilisation notamment pour le service VizieR
  - Vérifier son comportement lors d'un passage à l'échelle
  - Test d'un framework libre d'Hadoop

Présentation du stage

Deuxième partie du stage:

- Développement d'un « parseur » VOTable (standard de l'Observatoire Virtuel » coté client en JavaScript.

Étape 1: Début sur Apache Hadoop

- Apprendre Hadoop

- Se plonger dans le domaine du «Big Data»

Étape 1: Début sur Apache Hadoop

Quelques contraintes:

- Aucune expérience personnelle sur Hadoop
- Peu de connaissance d'Hadoop au sein de l'Observatoire

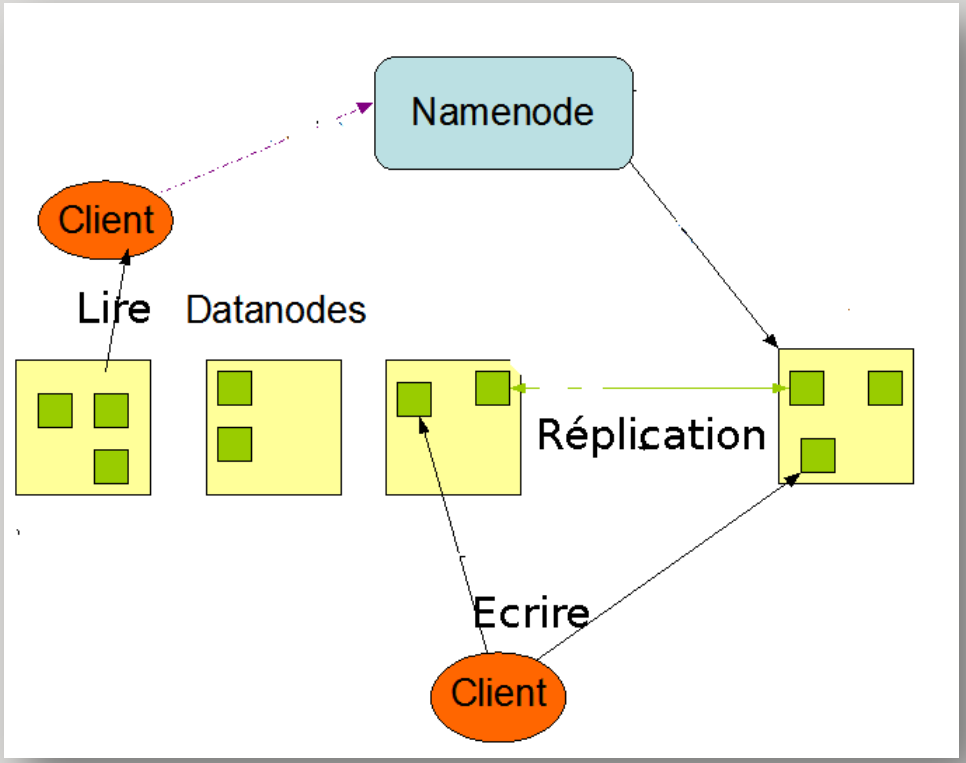
Étape 1: Début sur Apache Hadoop

Composants principales d'Hadoop:

- HDFS (Système de fichiers distribués)
- Implémentation de l'algorithme MapReduce

Étape 1: Début sur Apache Hadoop

- Un apperçu général de son architecture



La mise en place d'Hadoop

Objectif:

- Installer Hadoop sur une seule machine
- Créer un cluster Hadoop à partir de machines virtuelles (avec VirtualBox)

Problème rencontré:

- Documentation peu engageante

Test d'un framework Hadoop

- Adaptation à l'environement d' 

- Apprentissage des langages:
  - HiveQL (Apache Hive)
  - Pig latin (Apache Pig)
- Test de données dans ces langages.

Test d'un framework Hadoop

Apache Hive et Apache Pig:

- Il facilite la création des algorithmes MapReduce 
- Il support plusieurs types de fichiers. 
- La rapidité est assez faible 

Test d'un framework Hadoop

Une requête sur Apache Hive:

mode	q_mode	cl	sdss9	m_sdss9	im	raj2000	dej2000	obsdate	q	umag	e_umag
2		6	J170956.26-014812.6		Im	257.484441	-01.803517	2005.4300	3	25.848	0.861
1		3	J170956.41-014902.3		Im	257.485045	-01.817315	2005.4329	3	24.278	1.233
1		3	J170956.57-014926.7		Im	257.485724	-01.824095	2005.4329	3	28.793	0.931
2		6	J170958.73-014835.8		Im	257.494729	-01.809952	2005.4300	3	25.093	1.404
1		6	J170959.66-015625.6		Im	257.498588	-01.940466	2005.4329	3	25.497	0.877
1		6	J170959.96-014709.7		Im	257.499872	-01.786027	2005.4329	3	26.743	0.315
1		3	J171001.43-015620.5		Im	257.505995	-01.939048	2005.4329	3	25.294	1.339
1		6	J171001.97-015612.5		Im	257.508220	-01.936816	2005.4329	3	25.355	0.944
2		3	J171003.25-014634.5		Im	257.513546	-01.776257	2005.4301	3	24.098	1.714
2		3	J171003.91-014828.6		Im	257.516291	-01.807967	2005.4300	3	24.397	1.660
1		3	J171004.24-015455.6		Im	257.517698	-01.915462	2005.4329	3	21.129	1.505
2		6	J171004.28-015455.9		Im	257.517838	-01.915538	2005.4300	3	23.034	0.659
1		6	J171004.56-015516.4		Im	257.518998	-01.921248	2005.4329	3	25.783	0.743
1		6	J171004.68-015232.3		Im	257.519517	-01.875654	2005.4329	3	26.027	0.591
1		6	J171005.89-015336.7		Im	257.524565	-01.893535	2005.4329	3	21.828	0.209

Conclusion sur Hadoop

Avantages: 

- Traiter de vastes quantités de données.
- L'analyse à grande échelle.

Principal inconvénient: 

- La rapidité de l'implémentation de MapReduce est assez faible.

Début avec Elasticsearch

- Indexation de données
- Recherche du texte en temps réel

Début avec Elasticsearch

Utilisation:

- Indexation de données des catalogues
- De requêtes et de réponses de données en format JSON.

Début avec Elasticsearch

Avantages: 

- La réponse aux requêtes est super rapide.

inconvénients: 

- La documentation est très basique
- Il faut apprendre son propre langage

Elasticsearch sous Hortonworks

Objectif:

- Tester Hadoop en utilisant Elasticsearch :

Résultat:

- Une meilleure performance lors de l'écriture du texte
- Aucun changement par rapport à la lecture de données

Étape 2: Développement en Javascript

Objectif:

- Interpréter un fichier sous le format VOTable
- Pouvoir exploiter les données appartenant au fichier

Quelques contraintes:

- mémoire disponible gérée par le navigateur.
- chaque navigateur possède ses propres limitations et performances.

Étape 2: Développement en Javascript

Par exemple:

Conclusion

- Un stage qui m'a apporté des connaissances innovantes.
- Une expérience professionnelle.
- j'ai réussi à répondre aux besoins demandés pendant ce stage.
- Une introduction dans le domaine « Big Data ».

MERCI POUR VOTRE ATTENTION

Avez-vous des questions ?