Module: 165

Utiliser des bases de données NoSQL

ETML Informatique

ETML Informatique

Base de données NoSQL

Objectifs

1.4 Connaître diverses structures d'indexation

ETML Informatique

Base de données NoSQL

Les différentes structures d'indexations

Base de données
Les indexes

Le but d'une base de données est de fournir efficacement un résultat à chaque requête.

Pour cela plusieurs indexes permettent d'accéder directement à l'information recherchée.

Base de données
Les indexes

Découvrons premièrement certains concepts permettant l'optimisation de l'utilisation de volume important de données grâce aux indexes.

Base de données
Les arbres

L'index le plus utilisé est l'arbre B (ou BTree), il adopte une structure arborescente pour chercher toute valeur indexée. Les feuilles de cet arbre contiennent des liens vers les pages contenant la valeur.

Base de données
Le hachage

Le but d'une table de hachage est de placer les données par paquets, et à chaque donnée correspond un seul paquet de destination. Pour placer cette donnée, une **fonction de hachage** unique détermine le paquet.

Base de données
La distribution

Les bases de données distribuées existent depuis de nombreuses années ; le but de la distribution est de soulager le serveur central en répartissant les données sur plusieurs serveurs. Ce serveur central s'occupe ainsi de répartir la charge (données et requêtes), de fusionner le résultat et de gérer la cohérence des données.

Base de données
L'élasticité

C'est la capacité du système à s'adapter automatiquement en fonction du nombre de serveurs qu'il dispose et de la quantité de données à répartir.

Base de données
Le sharding

Le sharding est une technique permettant de distribuer des chunks (morceaux de fichiers) sur un ensemble de serveurs, avec la capacité de gérer l'élasticité (serveurs/données) et la tolérance aux pannes.

Base de données
Famille de distribution

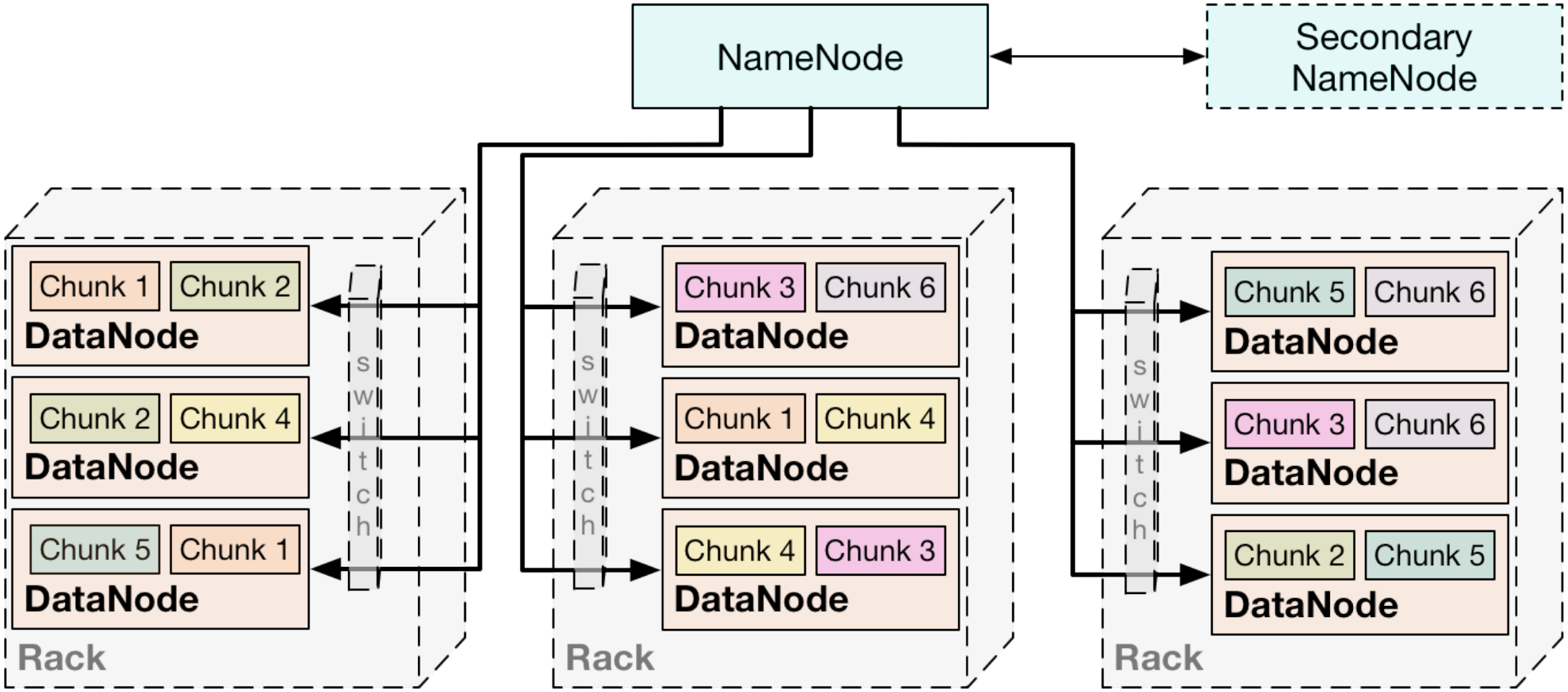
Trois familles de distribution pour le NoSQL existent :

- HDFS (basé sur la distribution),
- le clustered index (basé sur le BTree)
- le consistent hashing (basé sur les tables de hachage).

Base de données
HDFS

HDFS (Hadoop Distributed File System) est une technique de distribution de fichiers volumineux. Chaque fichier sera découpé en "chunk" de 64Mo pour être distribué sur le réseau. Chaque serveur de ce réseau est un **datanode** contenant plusieurs **chunks**. La répartition de ces chunks est définie par le serveur central, le **namenode**.

Base de données
HDFS

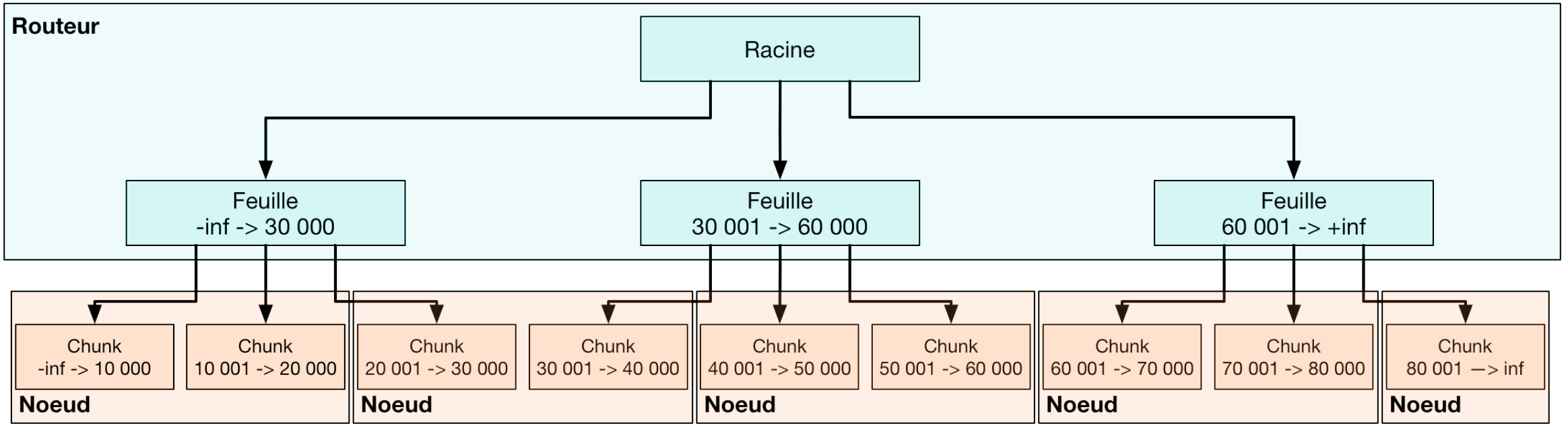


Base de données
Arbre distribué

La seconde famille de distribution de données repose sur le traditionnel BTree non-dense (ou **clustered index**) : il s'agit de l'arbre dont les données sont triées.

Un serveur central s'occupe de l'arborescence de cet arbre, et les feuilles (les données) sont prises en charge par les nœuds du cluster.

Base de données
Arbre distribué

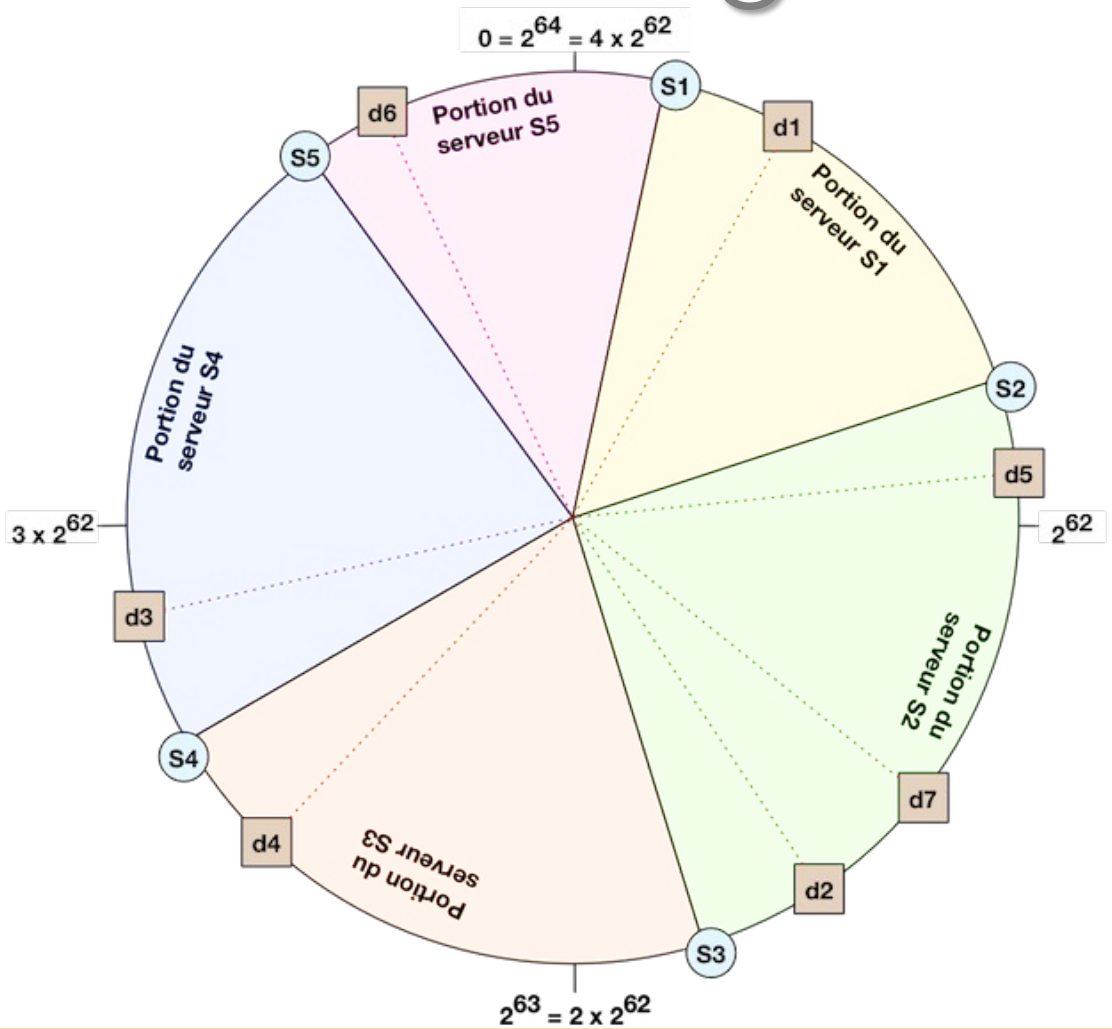


Base de données
Table de hachage distribué (DHT)

Avec da dernière famille, basée sur les tables de hachage, on distribue l'intégralité des informations dont on dispose : à la fois la donnée et la table de hachage.

C'est ce que l'on appelle le **Consistent Hashing**. La particularité de cette technique est que chaque nœud est à la fois client et serveur.

Base de données
Table de hachage distribué (DHT)



Chunk S1					
resp	d1				
réplicas	d5	d2	d4	d7	
Chunk S2					
resp	d5	d2	d7		
réplicas	d4	d3			
Chunk S3					
resp	d4				
réplicas	d3	d6			
Chunk S4					
resp	d3				
réplicas	d6	d1			
Chunk S5					
resp	d6				
réplicas	d1	d5	d2	d7	

ETML Informatique