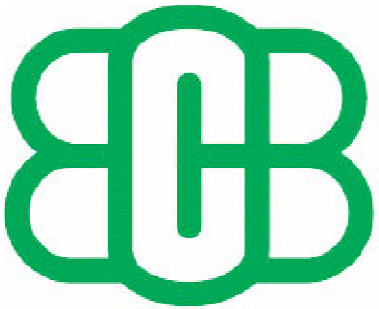


Classification des données cliniques : Solution Big Data

Khedidja Seridi

Salim Rahali

Introduction

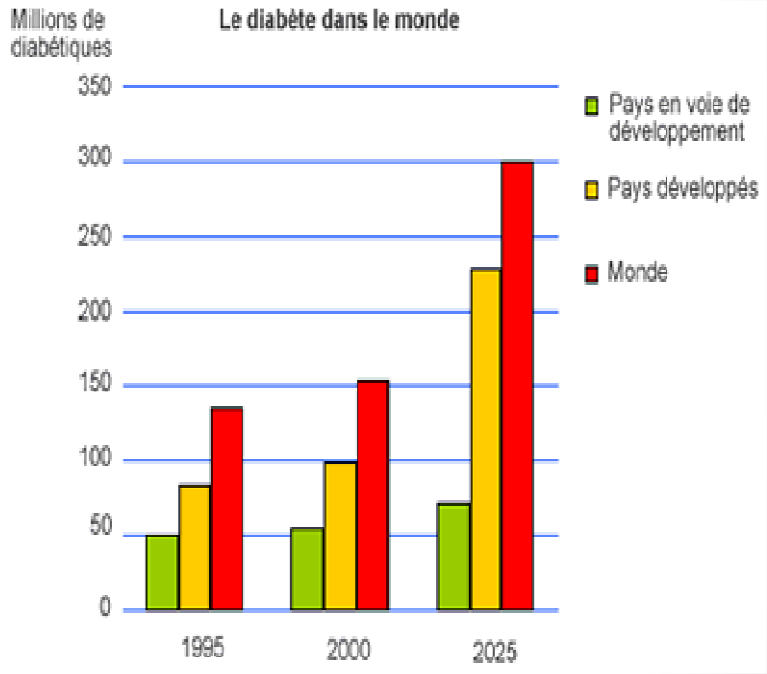
Le diabète

Maladie chronique qui présente des risques importants sur la santé des patients en terme de mortalité et invalidité.

Introduction

Nombre de personnes atteintes de diabètes est en augmentation permanente.

Impact important sur les budgets consacrés aux prises en charge médicales.



Besoin

Une classification des durées d'hospitalisation en fonction du nombre d'admission aux urgences.

Ministère de la santé publique

Le dataset :

Représente 10 ans (1999-2008) de soins cliniques dans 130 hôpitaux américains.
d'instances (ligne ou individu): 100000
d'attributs (colonnes ou critères) : 55

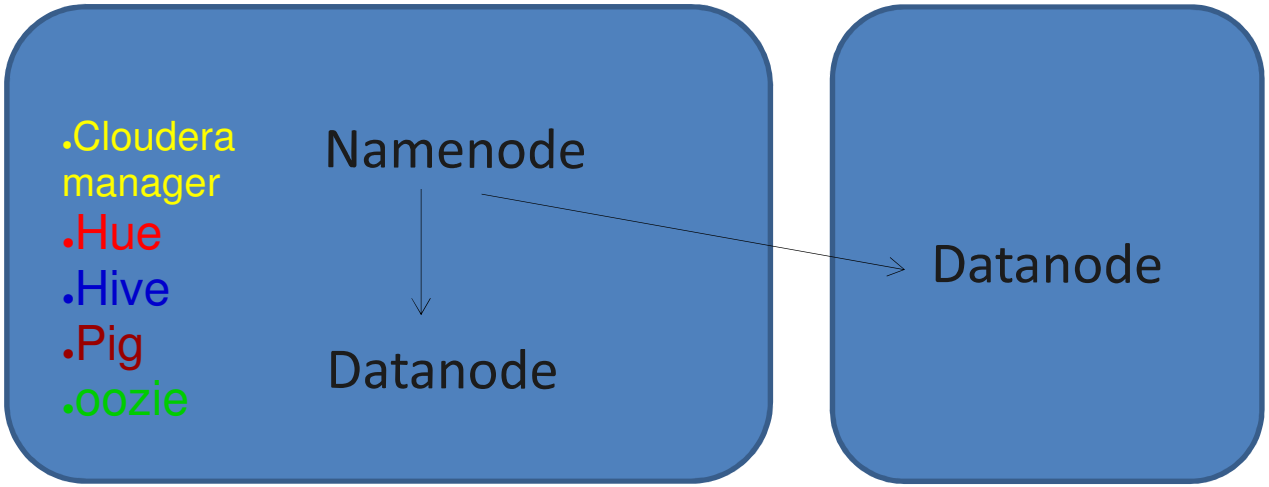
Les attributs :

Diabetic_Data.csv: La race, le genre, l'âge, .. durée d hospitalisation, nb de visite aux urgences l'année précédant, l'hospitalisation.

IDS_Mappings.csv: Contient les libellés et le mapping (Types d'urgences)

Source :
http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

Le cluster : Architecture(1)



Cloudera QuickStart

CentOs 6.2

12 GB, 2 proc, 80 GB

CentOs 6.2

8 GB, 1 proc, 80 GB

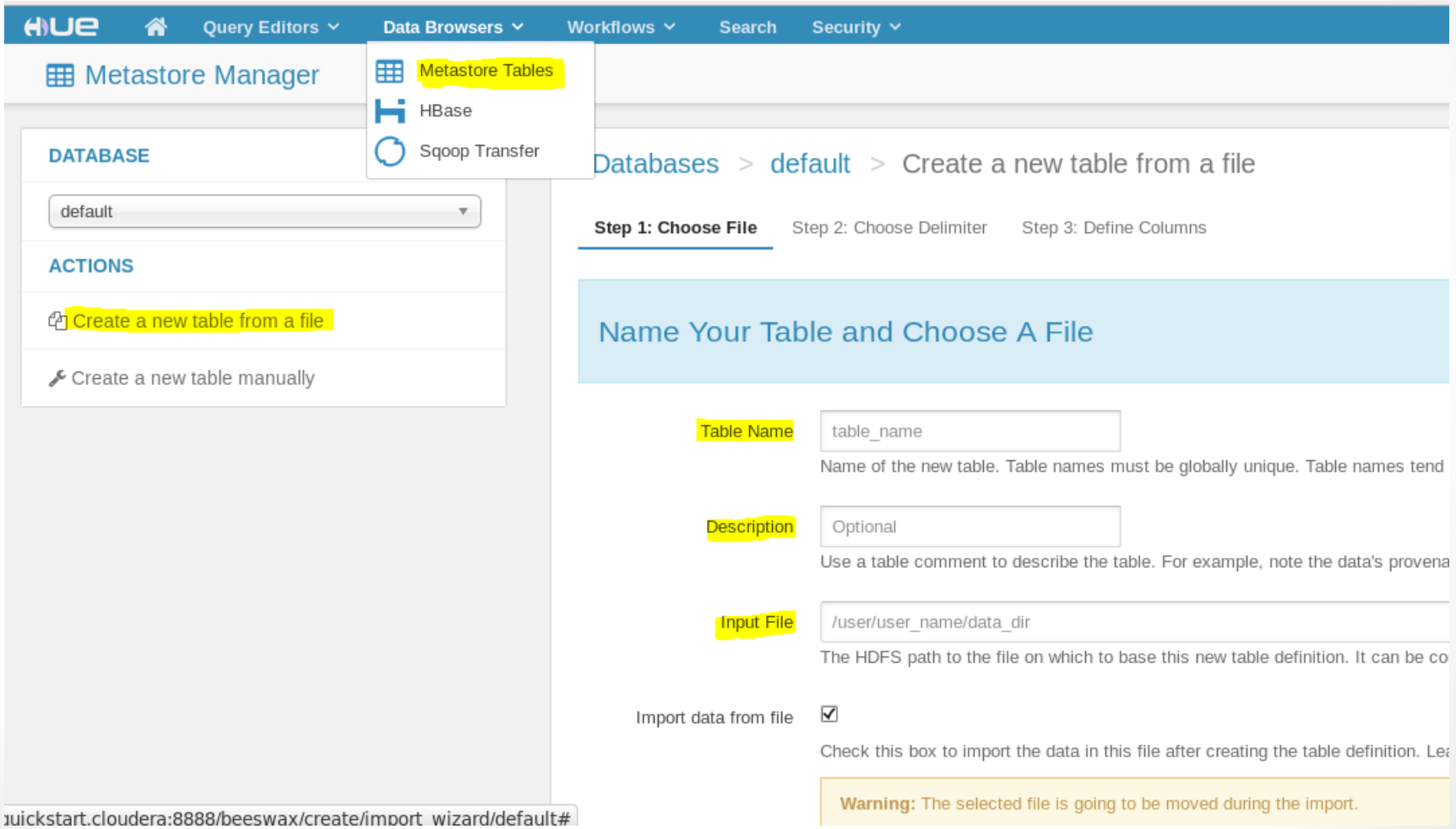
Le cluster : Architecture(2)

Fonction	Os	Ram	HDD
NamNode	Ubuntu 12.04	8Go	80Go
DataNode	Ubuntu 12.04	4Go	80Go
DataNode	Ubuntu 12.04	4Go	80Go

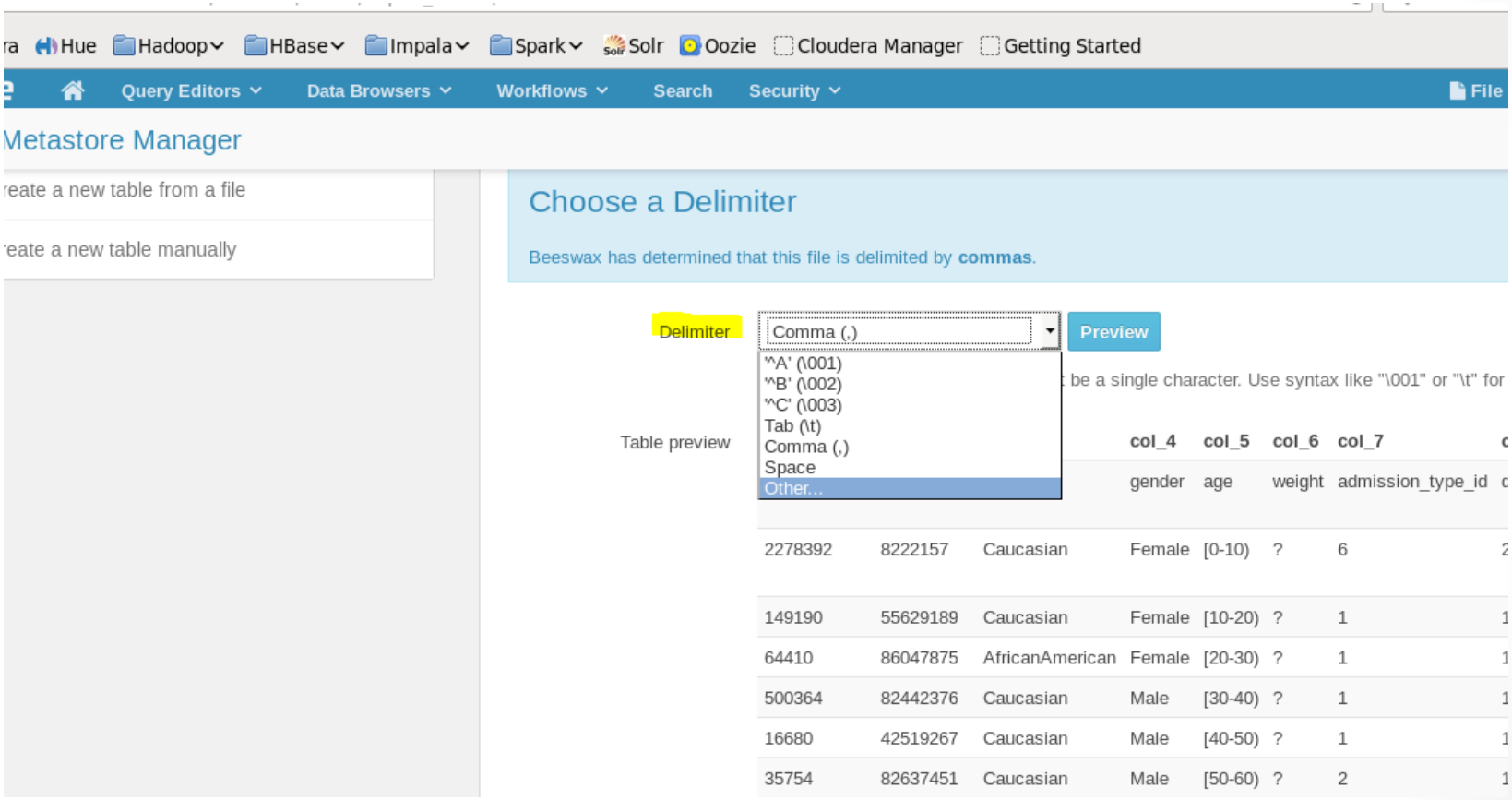
Développement

Pig	Hive
Used by Programmers and Researchers	Used by Analysts
Used for Programming	Used for Reporting
Procedural data-flow language	Declarative SQLish language
Works on the Client side of a Cluster	Works on the Server side of a Cluster
For Semi-Structured Data	For Structured Data

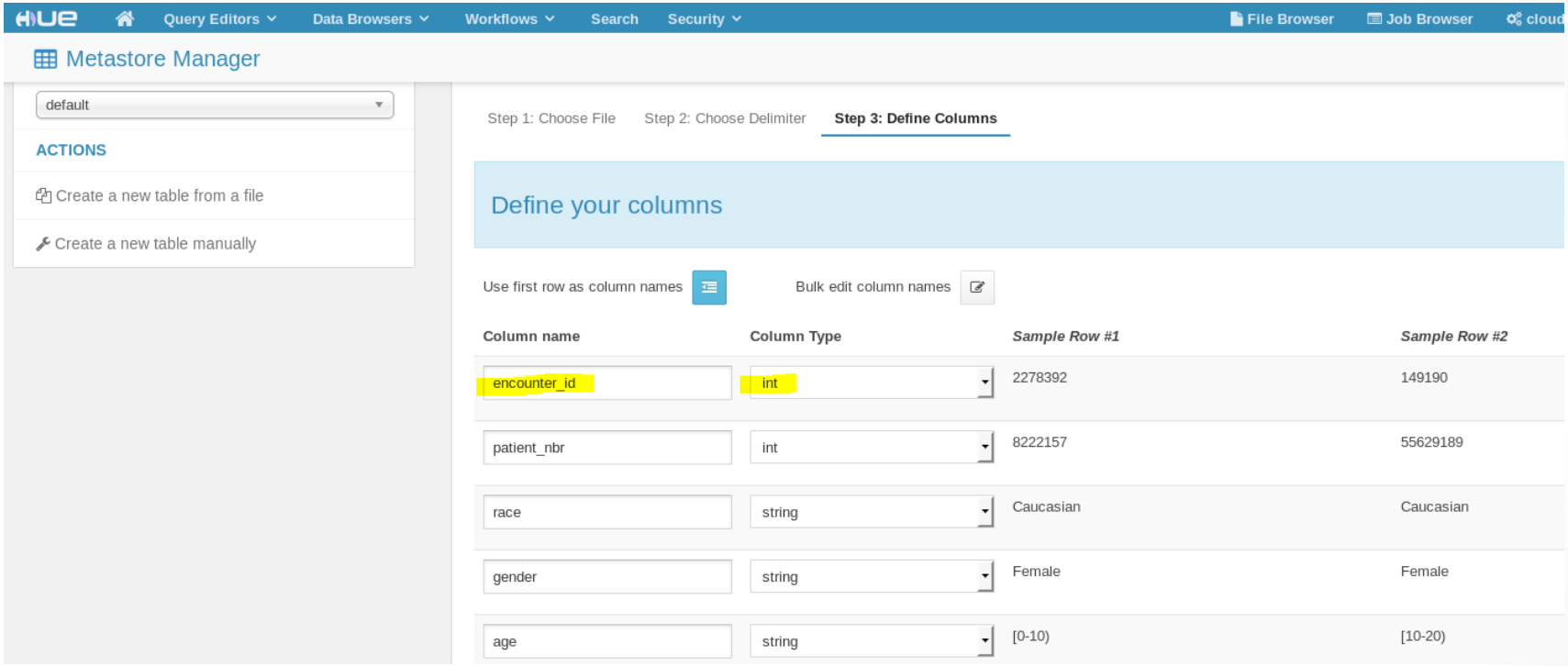
Développement



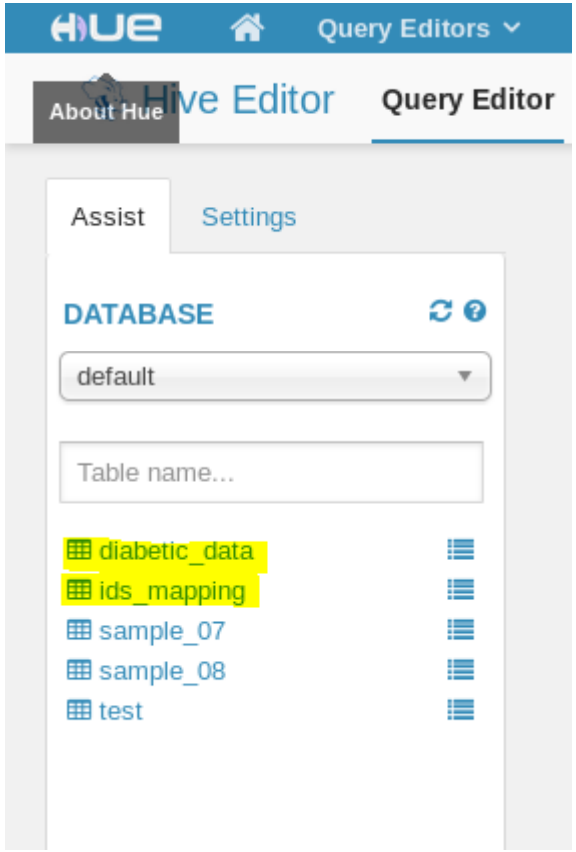
Développement



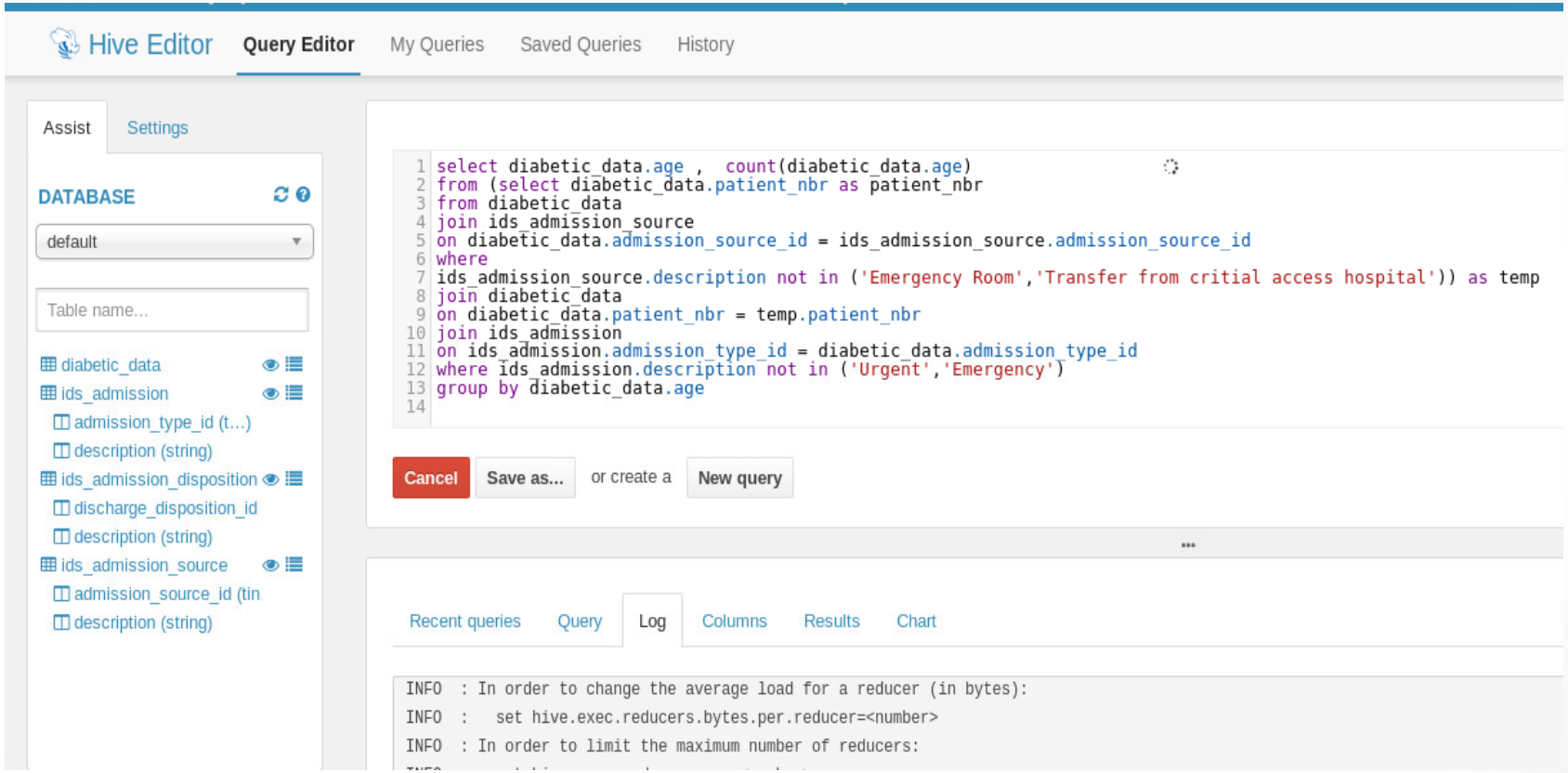
Développement



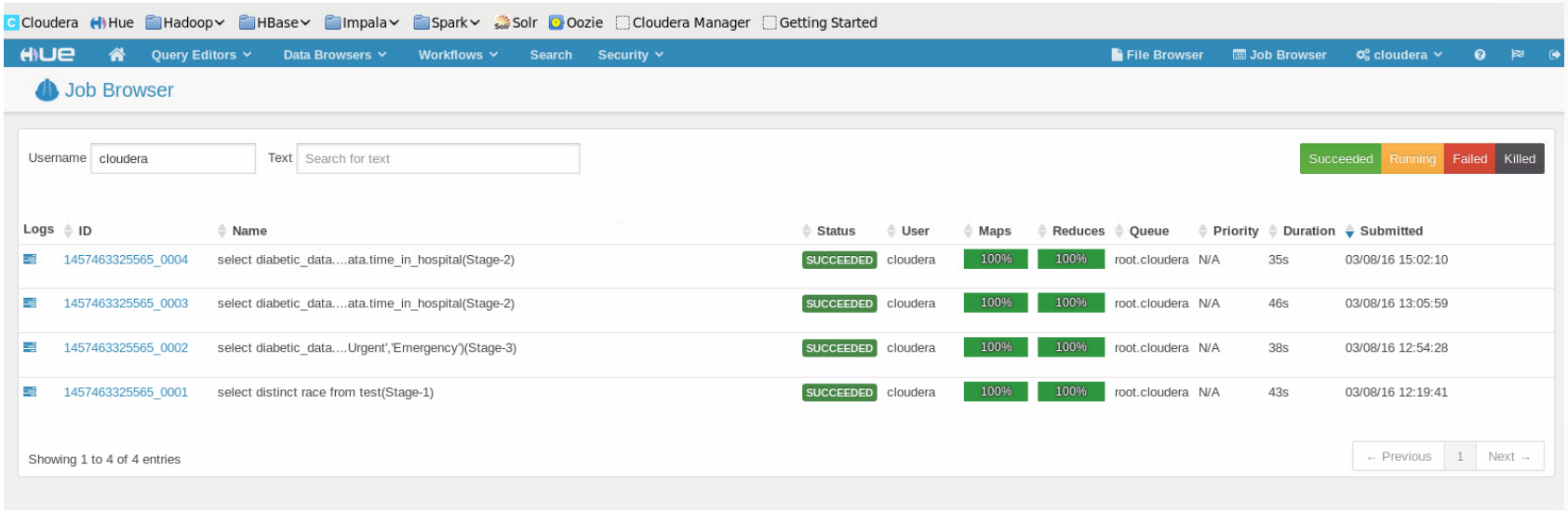
Développement



Developpement



Developpement



Résultats:

Besoin pour Projet:

Avoir le nombre de visites d'urgences avant hospitalisation par tranche d'âge.

Codes:

```
select diabetic_data.age ,  count(diabetic_data.age)
        from (select diabetic_data.patient_nbr as patient_nbr
                from diabetic_data
                join ids_admission_source
                on diabetic_data.admission_source_id = ids_admission_source.admission_source_id
                where
                ids_admission_source.description not in ('Emergency Room','Transfer from critial access
                hospital')) as temp
        join diabetic_data
                on diabetic_data.patient_nbr = temp.patient_nbr
                join ids_admission
                on ids_admission.admission_type_id = diabetic_data.admission_type_id
        where ids_admission.description not in ('Urgent','Emergency')
        group by diabetic_data.age
```

Résultats:

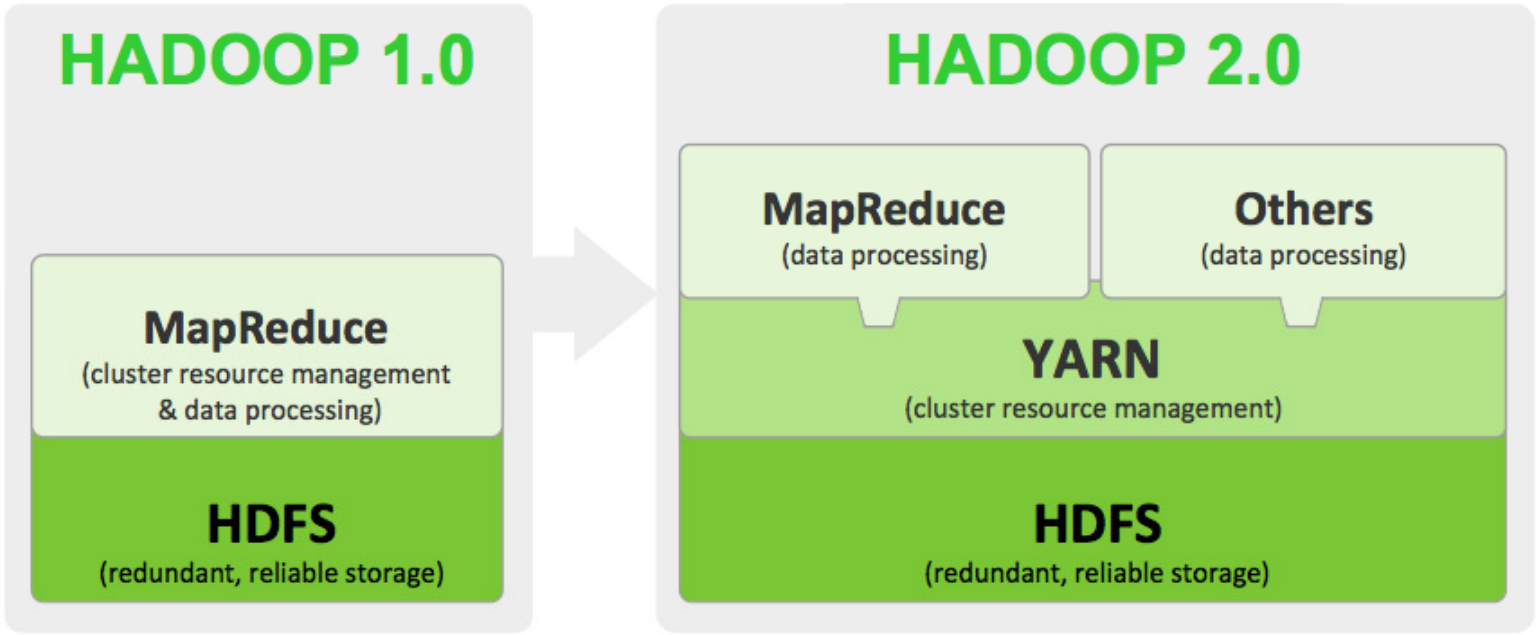
	diabetic_data.age	count_admission_before_emerg
0	[0-10)	38
1	[10-20)	306
2	[20-30)	1345
3	[30-40)	2400
4	[40-50)	5775
5	[50-60)	10179
6	[60-70)	13531
7	[70-80)	15672
8	[80-90)	8754
9	[90-100)	1062

Difficultés rencontrés

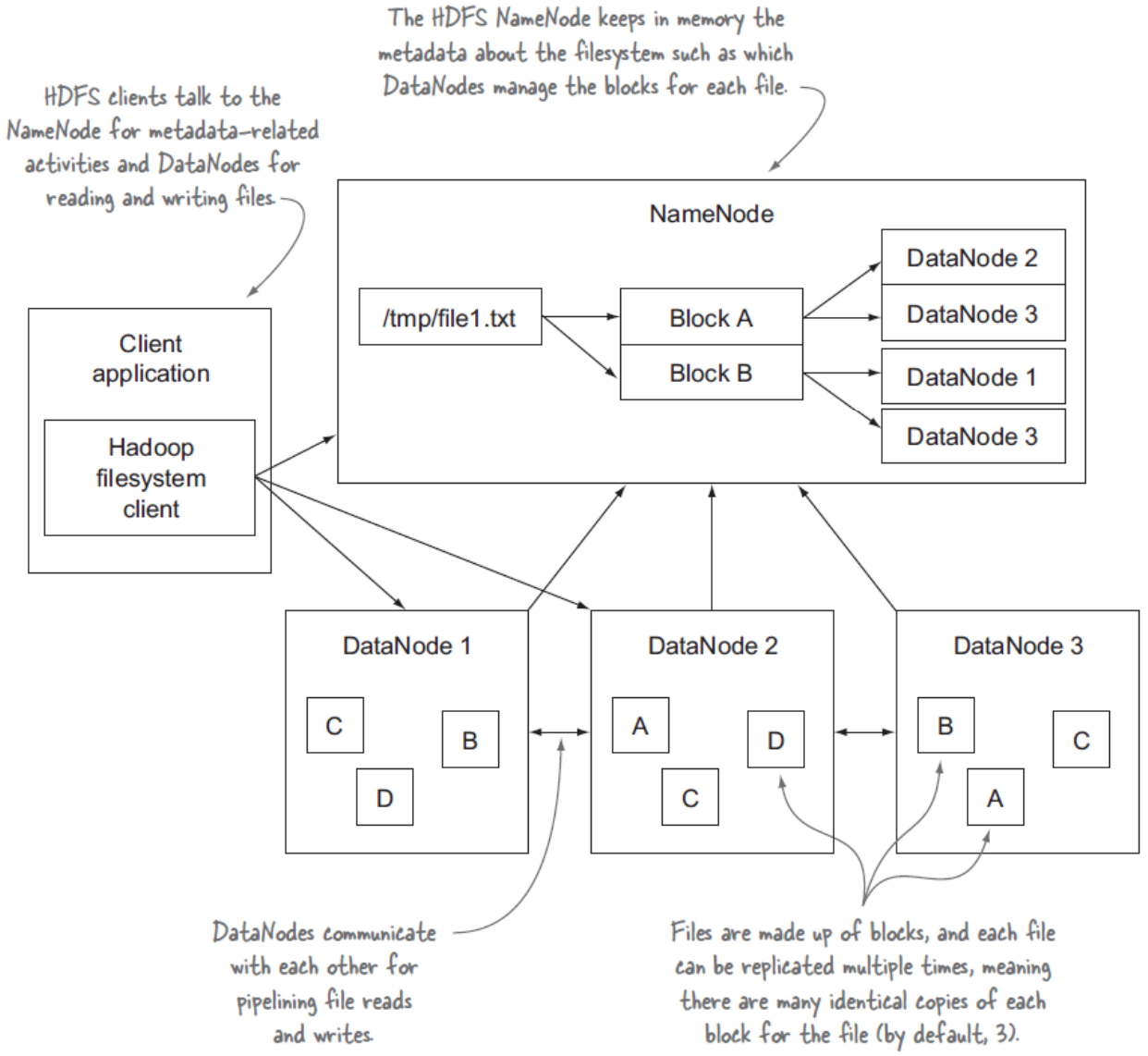
Installation et administration

- Monter en place un cluster stable.
- Ajouter les Data Nodes(Roles)
- Garder les services et les hôtes fonctionnels.

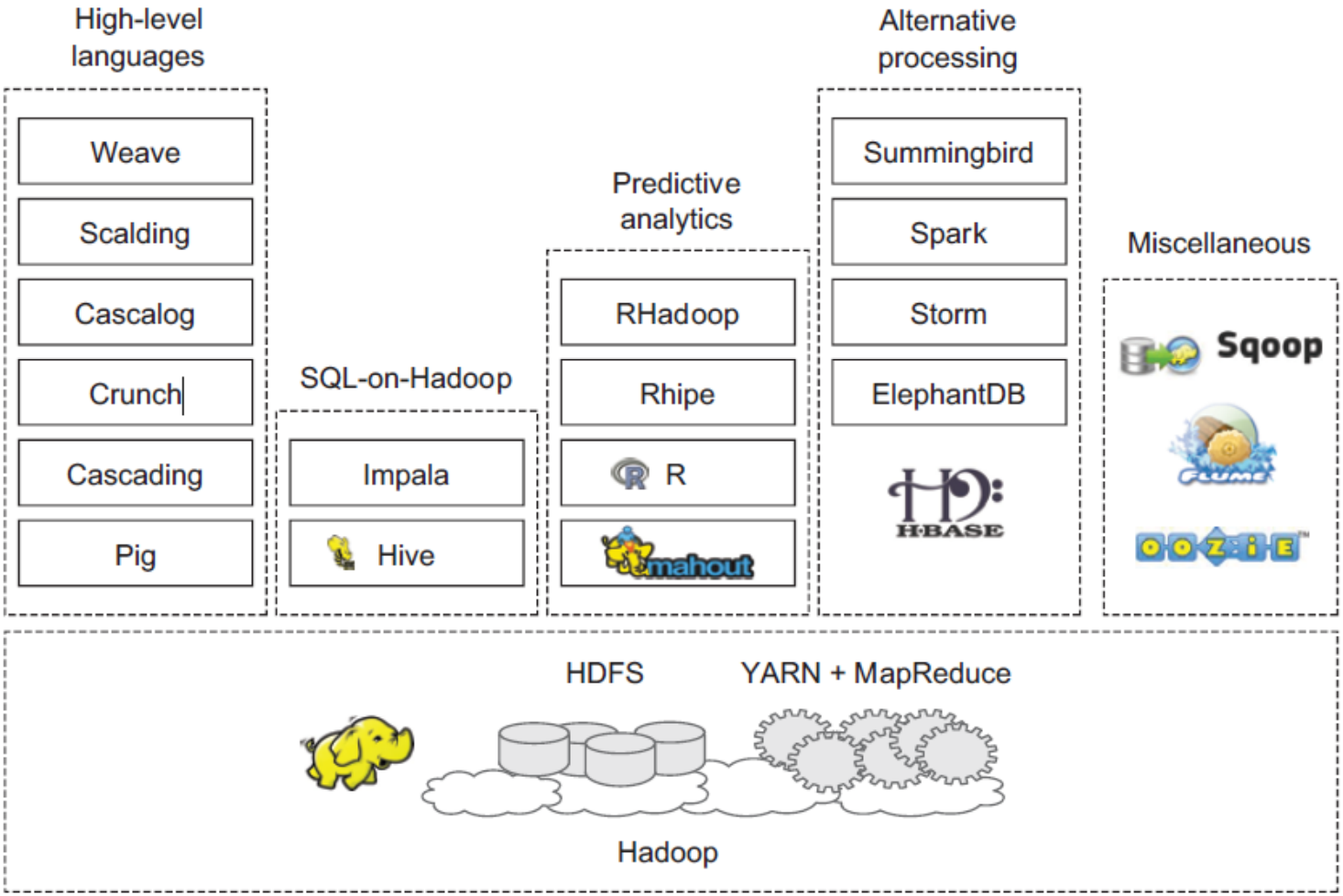
Compétences aquises



Compétences aquises



Compétences aquises



Conclusion