



Classification des données cliniques : Solution Big Data

Khedidja Seridi

Salim Rahali

Introduction

Le diabète

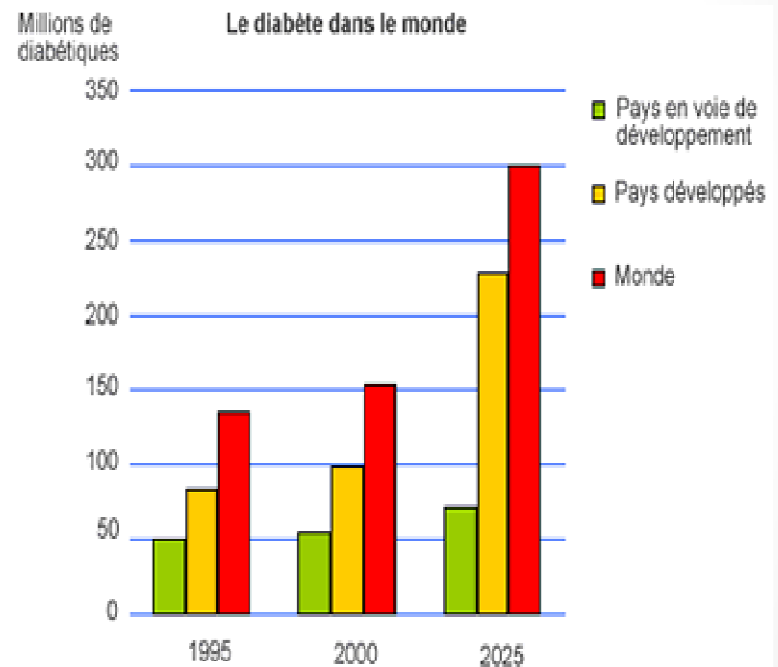
Maladie chronique qui présente des risques importants sur la santé des patients en terme de mortalité et invalidité.



Introduction

Nombre de personnes atteintes de diabète 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 :

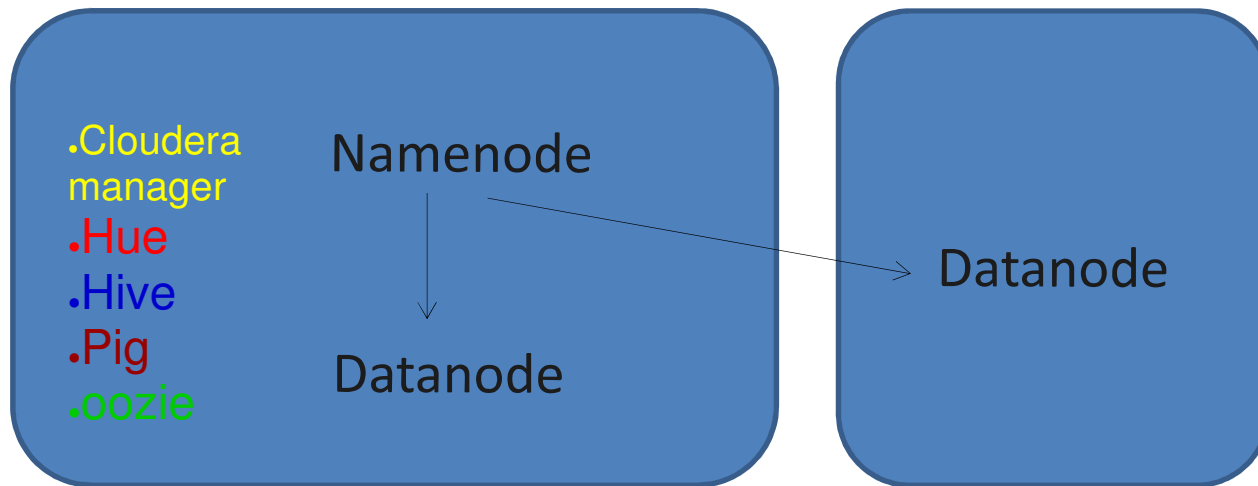
Diabetic_Data.csv: La race, le genre, l'âge, .. durée d'hospitalisation, nb de visite aux urgences l'année précédant, l'hospitalisation.

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

CentOs 6.2

12 GB, 2 proc, 80 GB

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

The screenshot displays the Hue web interface for the Metastore Manager. The top navigation bar includes 'HUE', a home icon, and menu items for 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. The 'Data Browsers' menu is open, showing 'Metastore Tables', 'HBase', and 'Sqoop Transfer'. The 'Metastore Tables' option is highlighted in yellow. The main content area shows the 'Metastore Manager' for the 'default' database. Under the 'ACTIONS' section, 'Create a new table from a file' is highlighted in yellow. The wizard is currently at 'Step 1: Choose File', with 'Step 2: Choose Delimiter' and 'Step 3: Define Columns' visible as subsequent steps. The wizard title is 'Name Your Table and Choose A File'. It contains three input fields: 'Table Name' (containing 'table_name'), 'Description' (containing 'Optional'), and 'Input File' (containing '/user/user_name/data_dir'). Below these fields is a checkbox labeled 'Import data from file' which is checked. A yellow warning box at the bottom states: 'Warning: The selected file is going to be moved during the import.'

quickstart.cloudera:8888/beeswax/create/import_wizard/default#

Développement

ra Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Home Query Editors Data Browsers Workflows Search Security File

Metastore Manager

Create a new table from a file

Create a new table manually

Choose a Delimiter

Beeswax has determined that this file is delimited by **commas**.

Delimiter: Comma (,)

Preview

Table preview

	col_4	col_5	col_6	col_7	c
	gender	age	weight	admission_type_id	c
2278392	8222157	Caucasian	Female	[0-10]	? 6 2
149190	55629189	Caucasian	Female	[10-20]	? 1 1
64410	86047875	AfricanAmerican	Female	[20-30]	? 1 1
500364	82442376	Caucasian	Male	[30-40]	? 1 1
16680	42519267	Caucasian	Male	[40-50]	? 1 1
35754	82637451	Caucasian	Male	[50-60]	? 2 1

Développement

HUE Home Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloud

Metastore Manager

default

ACTIONS

- Create a new table from a file
- Create a new table manually

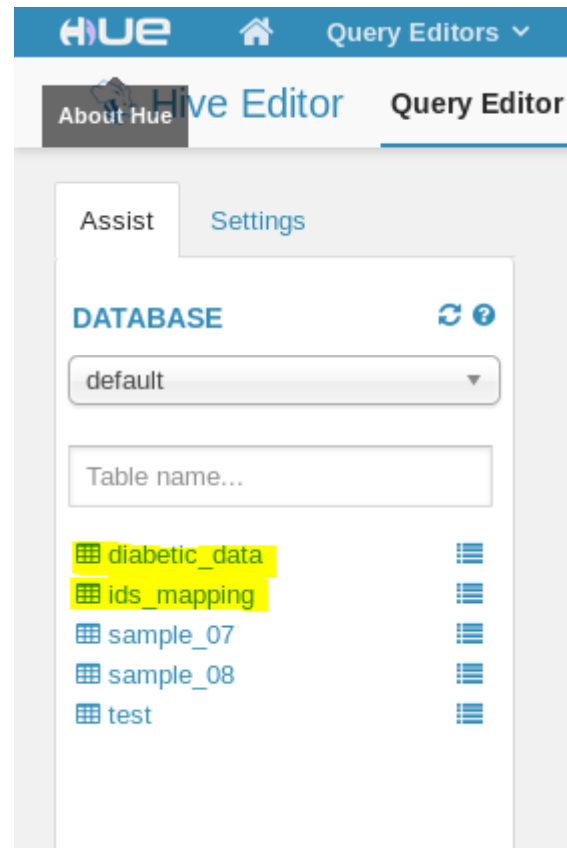
Step 1: Choose File Step 2: Choose Delimiter **Step 3: Define Columns**

Define your columns

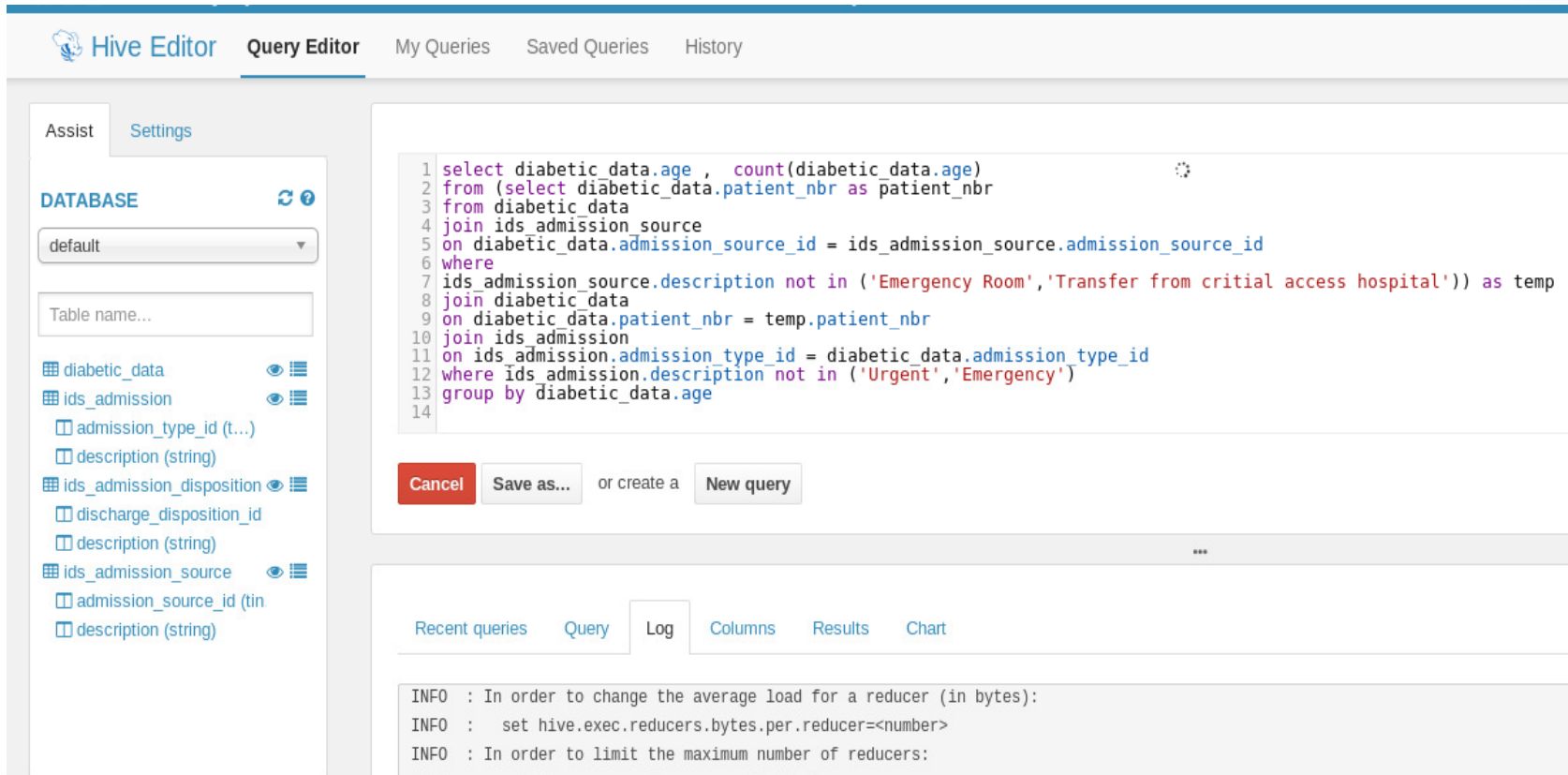
Use first row as column names Bulk edit column names

Column name	Column Type	Sample Row #1	Sample Row #2
encounter_id	int	2278392	149190
patient_nbr	int	8222157	55629189
race	string	Caucasian	Caucasian
gender	string	Female	Female
age	string	[0-10)	[10-20)

Développement



Developpement



The screenshot displays the Hive Editor interface. At the top, there is a navigation bar with 'Hive Editor' and 'Query Editor' (selected), along with links for 'My Queries', 'Saved Queries', and 'History'. On the left side, there is a sidebar with 'Assist' and 'Settings' tabs. Under 'Assist', there is a 'DATABASE' section with a dropdown menu set to 'default' and a 'Table name...' input field. Below this, a list of tables is shown with expand/collapse icons and visibility toggles: 'diabetic_data', 'ids_admission' (expanded to show 'admission_type_id (t...)' and 'description (string)'), 'ids_admission_disposition' (expanded to show 'discharge_disposition_id' and 'description (string)'), and 'ids_admission_source' (expanded to show 'admission_source_id (tin)' and 'description (string)').

The main area is a SQL query editor containing the following code:

```
1 select diabetic_data.age , count(diabetic_data.age)
2 from (select diabetic_data.patient_nbr as patient_nbr
3 from diabetic_data
4 join ids_admission_source
5 on diabetic_data.admission_source_id = ids_admission_source.admission_source_id
6 where
7 ids_admission_source.description not in ('Emergency Room','Transfer from critical access hospital')) as temp
8 join diabetic_data
9 on diabetic_data.patient_nbr = temp.patient_nbr
10 join ids_admission
11 on ids_admission.admission_type_id = diabetic_data.admission_type_id
12 where ids_admission.description not in ('Urgent','Emergency')
13 group by diabetic_data.age
14
```

Below the query editor, there are buttons for 'Cancel', 'Save as...', 'or create a', and 'New query'. At the bottom of the interface, there is a navigation bar with tabs for 'Recent queries', 'Query' (selected), 'Log', 'Columns', 'Results', and 'Chart'. The 'Log' tab is active, showing the following output:

```
INFO : In order to change the average load for a reducer (in bytes):
INFO : set hive.exec.reducers.bytes.per.reducer=<number>
INFO : In order to limit the maximum number of reducers:
```

Developpement

The screenshot displays the Cloudera Hue Job Browser interface. At the top, there is a navigation bar with various application icons and a search bar. Below this, the 'Job Browser' section is active, showing a search filter for 'cloudera' and a search text input field. A legend indicates job statuses: Succeeded (green), Running (orange), Failed (red), and Killed (black). The main area contains a table of job logs with the following columns: Logs, ID, Name, Status, User, Maps, Reduces, Queue, Priority, Duration, and Submitted. All four jobs listed are in a 'SUCCEEDED' state.

Logs	ID	Name	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1457463325565_0004	select diabetic_data....ata.time_in_hospital(Stage-2)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	35s	03/08/16 15:02:10
	1457463325565_0003	select diabetic_data....ata.time_in_hospital(Stage-2)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	46s	03/08/16 13:05:59
	1457463325565_0002	select diabetic_data....Urgent','Emergency')(Stage-3)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	38s	03/08/16 12:54:28
	1457463325565_0001	select distinct race from test(Stage-1)	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	43s	03/08/16 12:19:41

Showing 1 to 4 of 4 entries

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 critical 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:

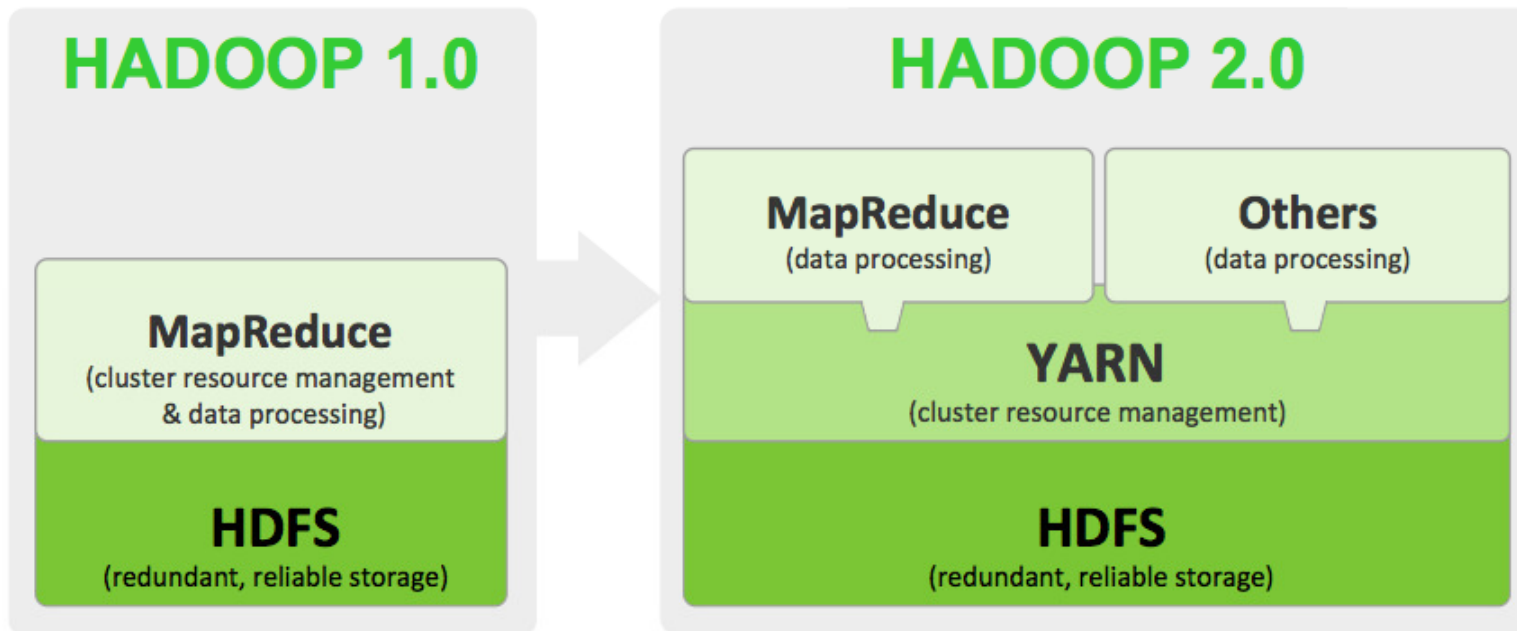
◆	◆ diabetic_data.age	◆ count_admission_before_emerg
0	[0-10)	38
1	[10-20)	306
2	[20-30)	1345
3	[30-40)	2400
4	[40-50)	5775
5	[50-60)	10179
6	[60-70)	13531
7	[70-80)	15672
8	[80-90)	8754
9	[90-100)	1062

Difficultés rencontrés

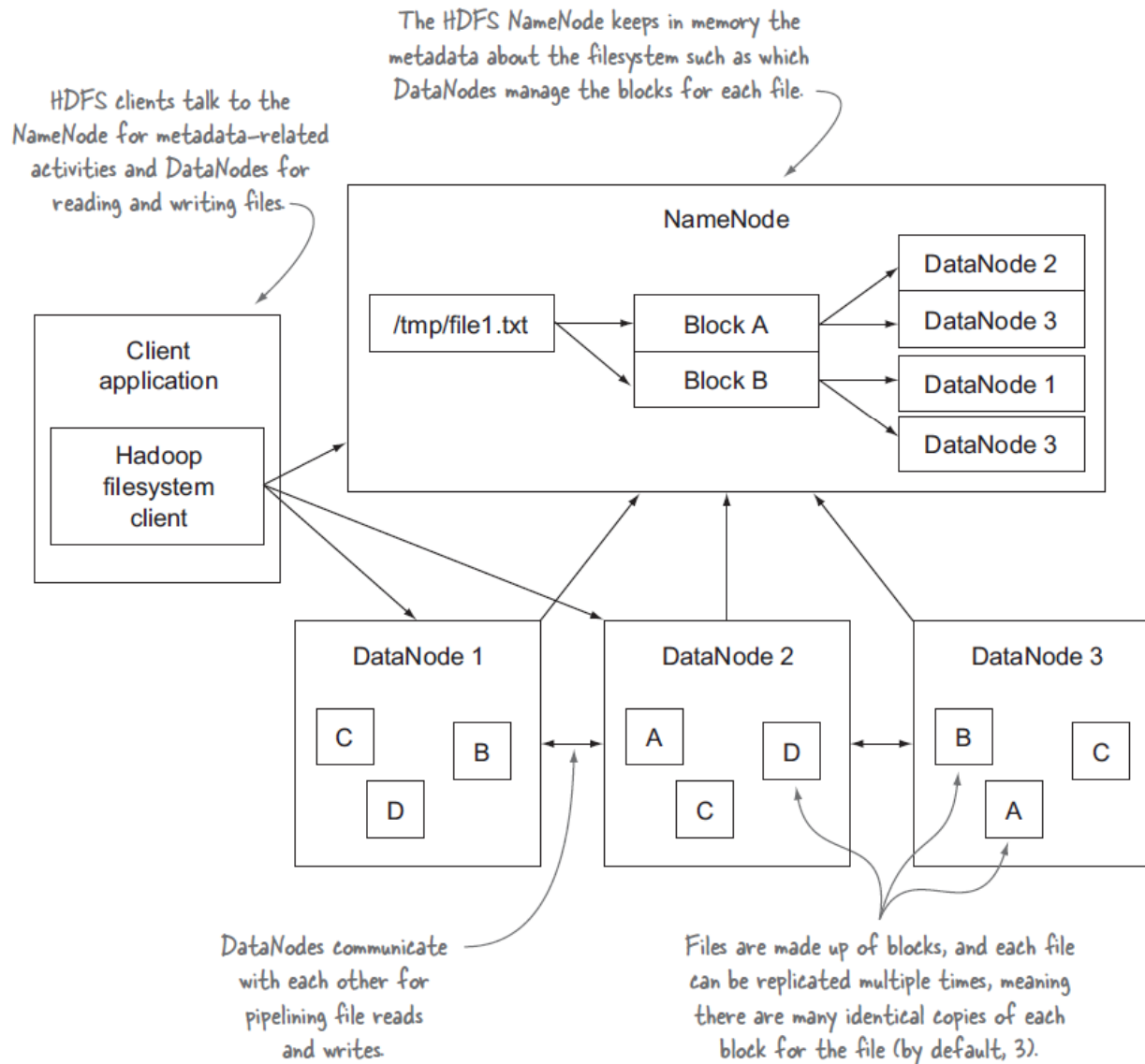
Installation et administration

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

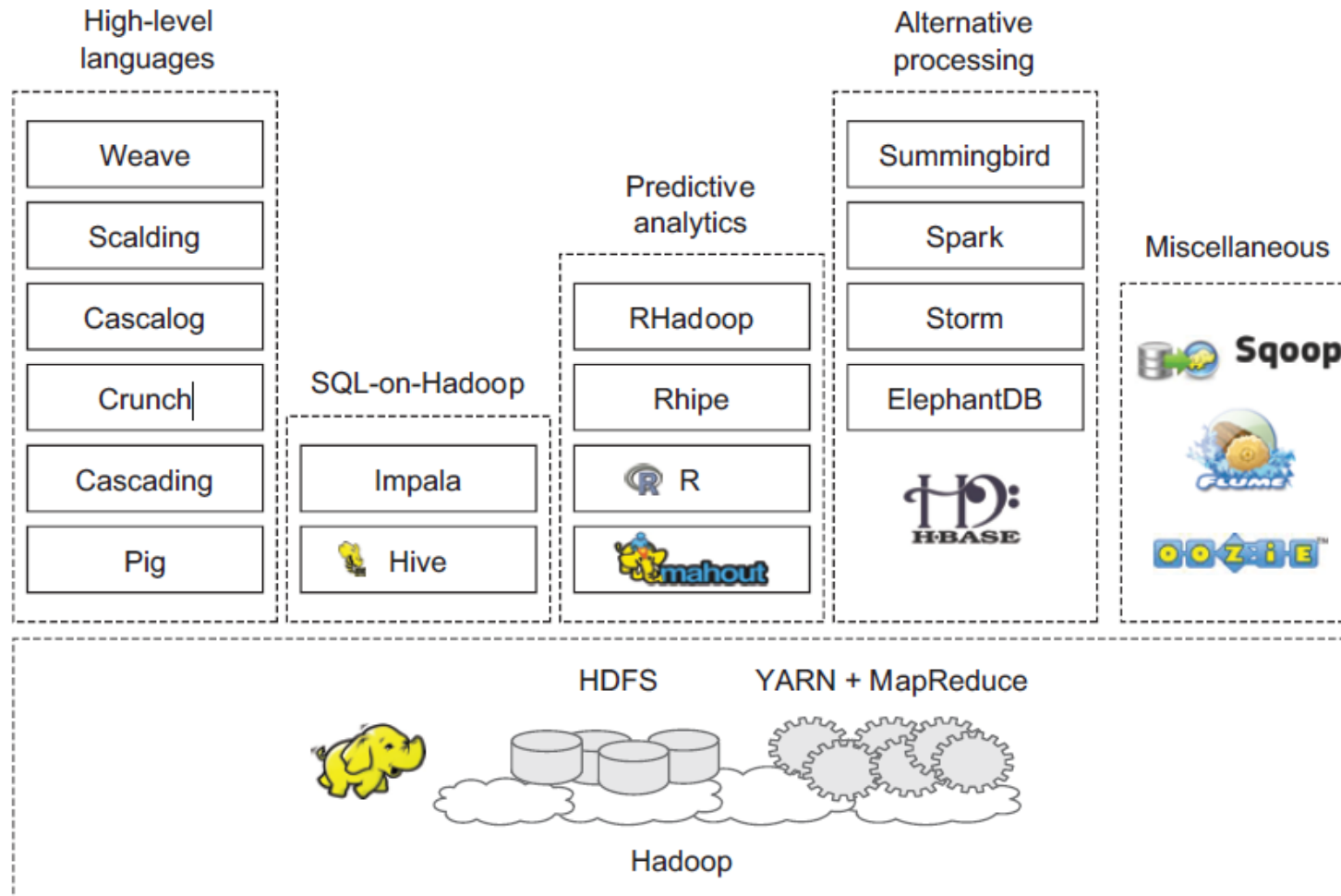
Compétences aquises



Compétences aquises



Compétences acquises



Conclusion