

« TECHNO DU BIG-DATA »

Stéphane Derrode, Dpt MI - Stephane.derrode@ec-lyon.fr



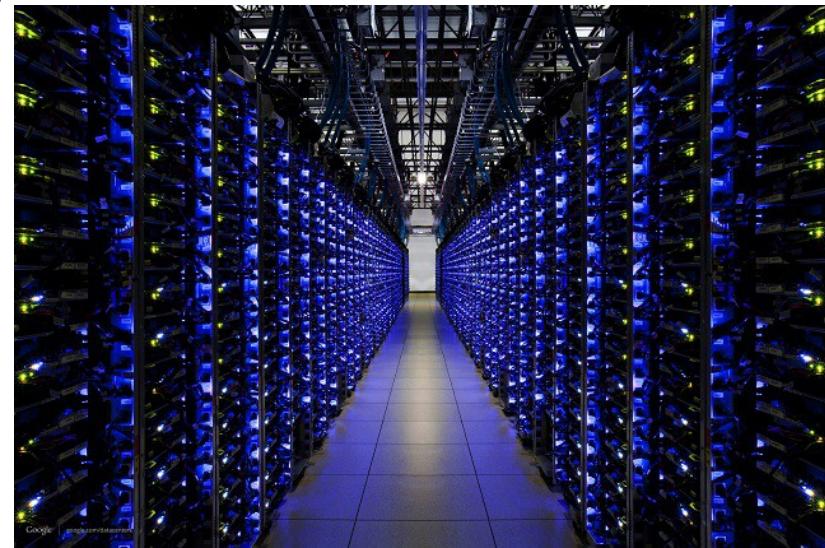
Sommaire

1. Big Data et parallélisme
2. Modèle MapReduce
3. HDFS
4. MapReduce sur cluster Hadoop
5. Générateurs et Itérateurs en Python
6. Ecosystème Hadoop

BIG DATA ET PARALLÉLISME

« Big Data »

- Exemples:
 - Google, 2008: 20 PB / jour, 180 GB / job
 - Web index : 50 Milliard de pages, 15 PB
 - Large Hadron Collider (LHC)@CERN : 15PB / année
- Capacité d'un (gros) serveur
 - RAM : 256 GB
 - DD : 24 TB
 - Vitesse de transfert du DD : 100 MB/s
- Solution : parallélisme
 - 1 serveur : 8 DD, lire le web en **230 jours**
 - Hadoop cluster @ Yahoo : 4000 serveurs, lire le web en // = **1h20**



Google data center

Tolérance aux erreurs

- Le problème du parallélisme
 - 1 serveurs bug tous les quelques mois
 - 1000 serveurs -> temps moyen avant bug < 1 jour
- Un « gros » job peut prendre plusieurs jours
 - Une panne matériel : c'est donc la normalité!
 - Parallélisme : impossible de relancer partiellement en cas de panne
 - Point de contrôle, réPLICATION : difficile à implémenter correctement
- Plateformes Big data : tout le monde doit pouvoir écrire des programmes
 - Encapsule le parallélisme
 - Encapsule la tolérances aux pannes
 - Codé une fois par des experts, profitables à tous (non experts)

MODÈLE MAP REDUCE

Inspiré de la programmation fonctionnelle

Deux fonctions très simples inspirées de la programmation fonctionnelle:

- **Transformation : map**

- $\text{map}(f, [x_1, \dots, x_n]) = [f(x_1), \dots, f(x_n)]$
- Exemple : $\text{map}(2^*, [1, 2, 3]) = [(2^*, 1), (2^*, 2), (2^*, 3)] = [2, 4, 6]$

- **Agrégation : reduce**

- $\text{reduce}(f, [x_1, \dots, x_n]) = f(x_1, f(x_2, \dots, f(x_{n-1}, x_n)))$
- Exemple : $\text{reduce}(+, [2, 4, 6]) = (+2 (+4 6)) = 12$

Ces fonctions sont génériques car elles prennent en paramètre une fonction: le développeur fournit les fonctions.

- $\text{map}(\text{toUpperCase}, ['hello', 'data']) = ['HELLO', 'DATA']$
- $\text{reduce}(\text{max}, [3, 45, 27]) = 45$

Les données sont toujours représentées par des paires (clé, valeur)

Une clé peut être de n'importe quel type

- ('Hello', 17)
 - 'Hello' est la clé (text)
 - 17 est la valeur (int)

Lorsque les données ne sont pas des couples (clé, valeur)

- Un texte est représenté par (numéro de ligne, contenu de la ligne)

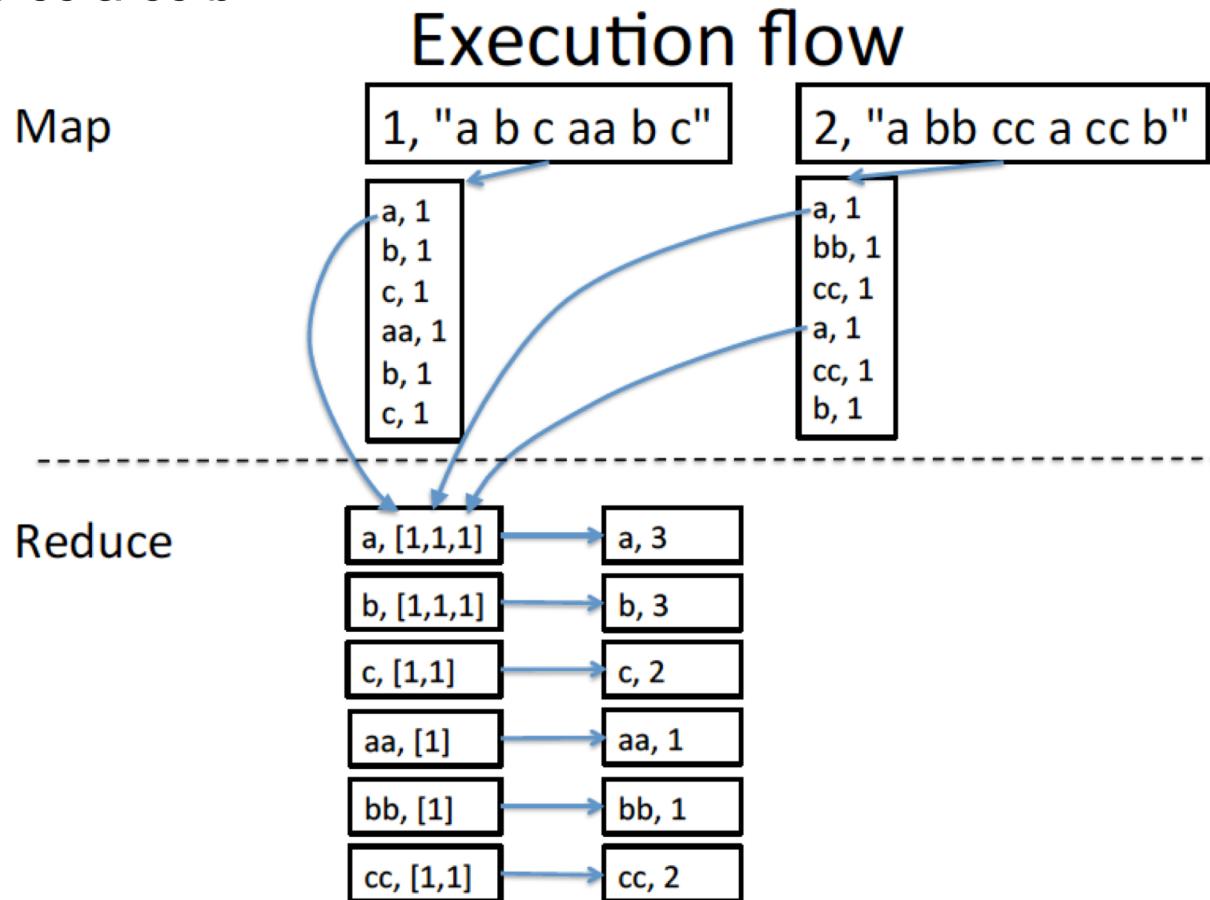
Map-Reduce appliqué à des couples (clé, valeur)

- **Map**, f est appliquée sur chaque couple **indépendamment**
 $f(\text{clé}, \text{valeur}) \rightarrow \text{list}(\text{clé}, \text{valeur})$
- **Reduce**, f est appliquée sur **toute les valeurs** de même clé
 $f(\text{clé}, \text{liste(valeur)}) \rightarrow \text{list}(\text{clé}, \text{valeur})$

Les types des clés et des valeurs n'ont pas besoin d'être les mêmes en entrée et en sortie.

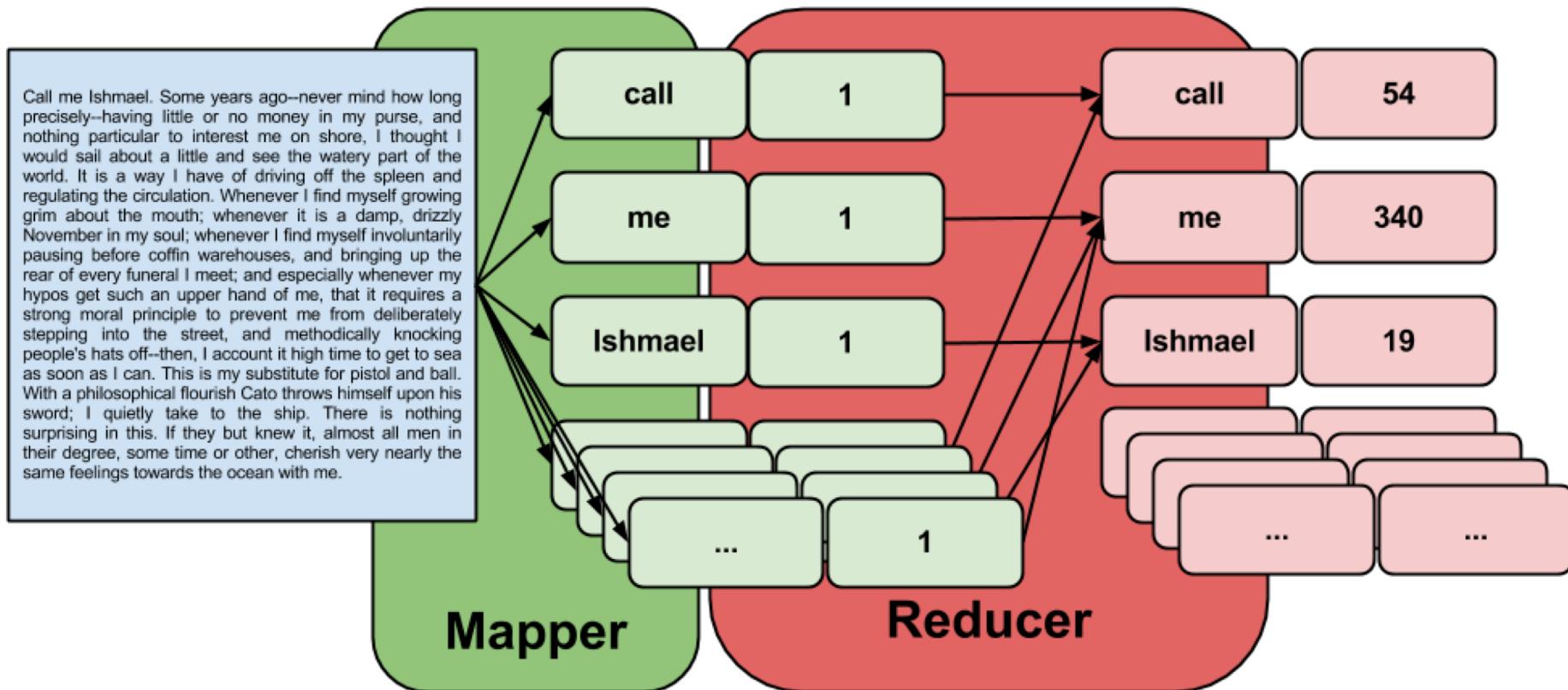
Exemple : comptage de la fréquence d'un mot

- Données d'entrée: un fichier A de 2 lignes
 - 1, 'a b c aa b c'
 - 2, 'a bb cc a cc b'



Map-Reduce famous Word Count

Le ‘Hello world’ du map-reduce



Hadoop map-reduce : natif en Java mais connecteur Python

Word Count in python (map)

```
#!/usr/bin/env python3
# fichier wc_mapper.py

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line=line.strip()
    # split the line into words
    words=line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        # tab-delimited; the trivial word count is 1
        print(word, '\t1')
```

Word Count in python (reduce)

```
#!/usr/bin/env python3
# fichier wc_reducer.py
import sys

current_word=None
current_count=0
word=None
for line in sys.stdin:
    line=line.strip()
    word, count=line.split('\t', 1)
    try:
        count=int(count)
    except ValueError:
        continue

    if current_word==word:
        current_count+=count
    else:
        if current_word:
            print(current_word, '\t', current_count)
        current_count=count
        current_word=word

if current_word==word:
    print(current_word, '\t', current_count)
```

Démonstration : en local (sur votre machine)

```
>> echo "foo foo quux labs quux" | ./mapper.py
>> echo "foo foo quux labs quux" | ./mapper.py | sort
>> echo "foo foo quux labs quux" | ./mapper.py | \
    sort | ./reducer.py

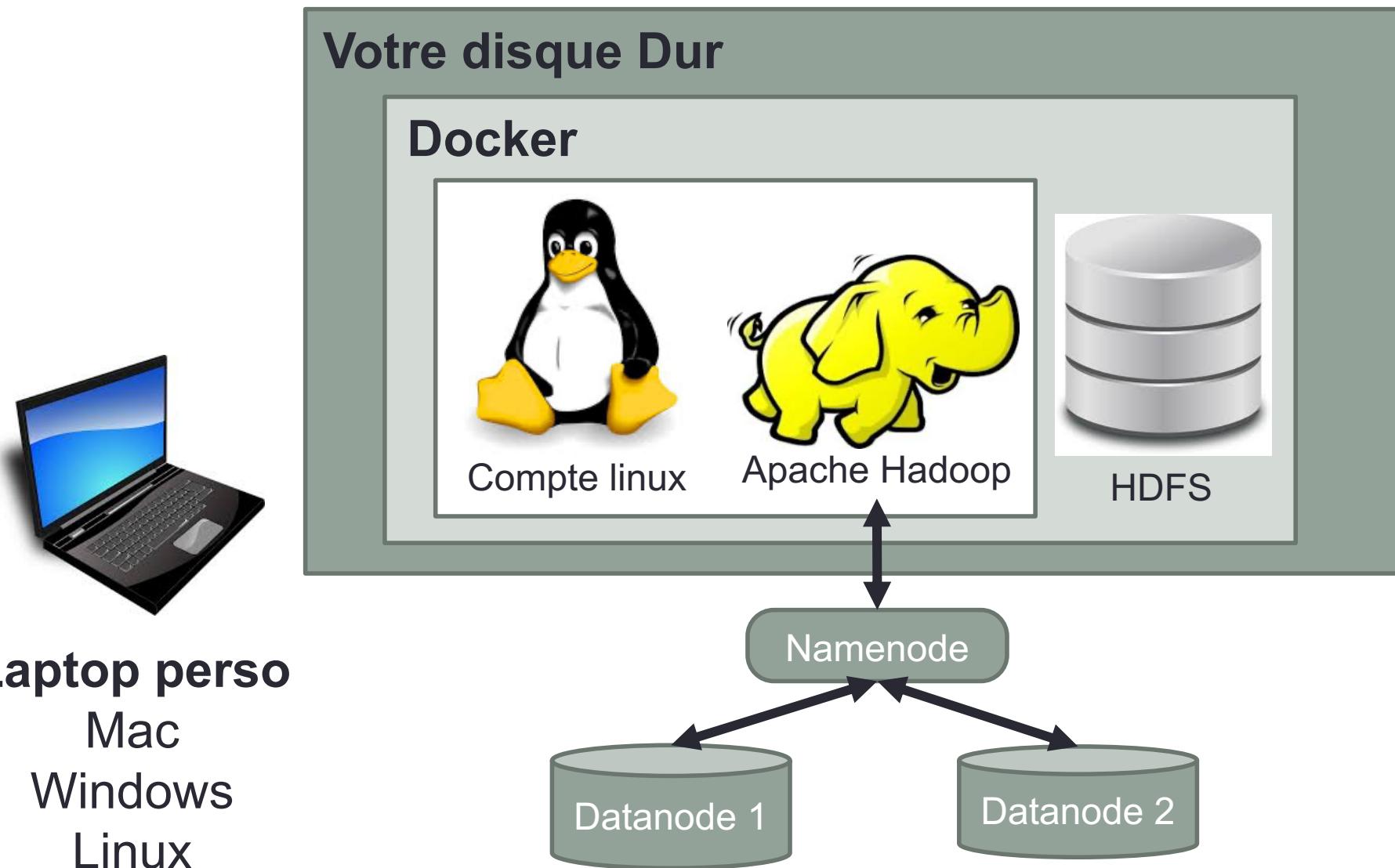
>> wget http://www.textfiles.com/etext/FICTION/dracula
>> more dracula

>> cat dracula | ./mapper.py
>> head -n 20 dracula | ./mapper.py | sort | ./reducer.py
```

HDFS

Hadoop Distributed File System

TP Hadoop : contexte technique



TP Hadoop : préparation

Enoncé du TP :

https://gitlab.ec-lyon.fr/sderrode/s9_mod21_bigdata_tp

Avant votre séance de TP : préparation requise

- Entrez dans le dossier *TP_Hadoop*
- Lisez *readme.md*
 - Installez *git* si requis
 - Commencez la partie 2 du TP : *Install_Docker_Hadoop.md*
 - Installer *Docker* en suivant le lien
 - Exécutez la commande suivante dans un Terminal:
`docker pull liliiasfaxi/spark-hadoop:hv-2.7.2`

Terminal : console permettant de taper de commandes. Natif sous Linux et sous Mac OS X. Sous Windows 10, utilisez "Windows powershell".

Système de fichiers distribués (HDFS)

- Objectifs:
 - Tolérant aux erreurs (redondance)
 - Performant (accès parallèle)
- Fichiers volumineux
 - Lecture et écriture séquentielle
- Traitement de données « au plus près »
Les données sont stockées sur les machines qui les traitent
 - Pour un meilleur usage des machines
 - Pour éviter les transferts réseaux (lag)
- Les données sont organisées en fichiers et répertoires
 - Mime les systèmes de gestion de fichiers standards
 - Les fichiers sont découpés en blocks (64MB) et éparpillés sur les serveurs avec réPLICATION (3 fois par défaut)
 - Si possible, traite les données sur les machines où elles sont stockées.

Architecture « maître / esclave »

- Un maître : le « NameNode »
 - Gère les noms de fichiers, les droits d'accès...
 - Stocke les metadata associées aux fichiers
 - Garde tout en mémoire RAM (maximum : 60M objets and 16 GO)
 - Supervise les opérations sur les fichiers et les blocks
 - Supervise la santé du système (échecs, crash), et équilibre les charges
- Des milliers d'esclaves : les « DataNode »
 - Stocke les données (blocks).
 - Les données ne transitent jamais par le « NameNode ».
 - Réalise les opérations de lecture et d'écriture.
 - Réalise les copies (replications) ordonnées par le « NameNode »
 - Vérifie régulièrement la santé du « NameNode »
 - Rapporte au « NameNode » si des blocks sont corrompus (checksum)

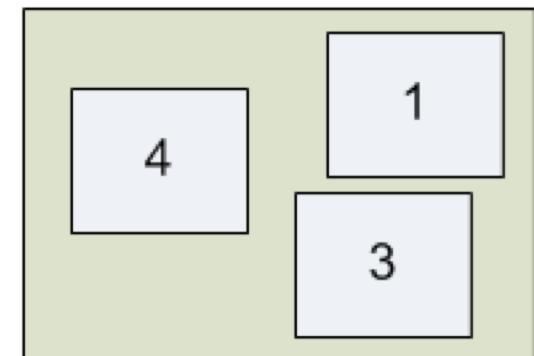
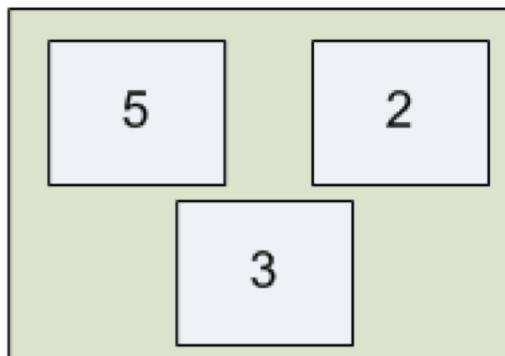
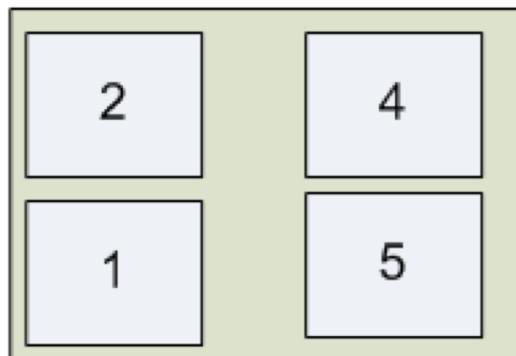
Architecture « maître / esclave », « master / slave »

1

NameNode:
Stores metadata only

METADATA:
`/user/aaron/foo → 1, 2, 4`
`/user/aaron/bar → 3, 5`

DataNodes: Store blocks from files



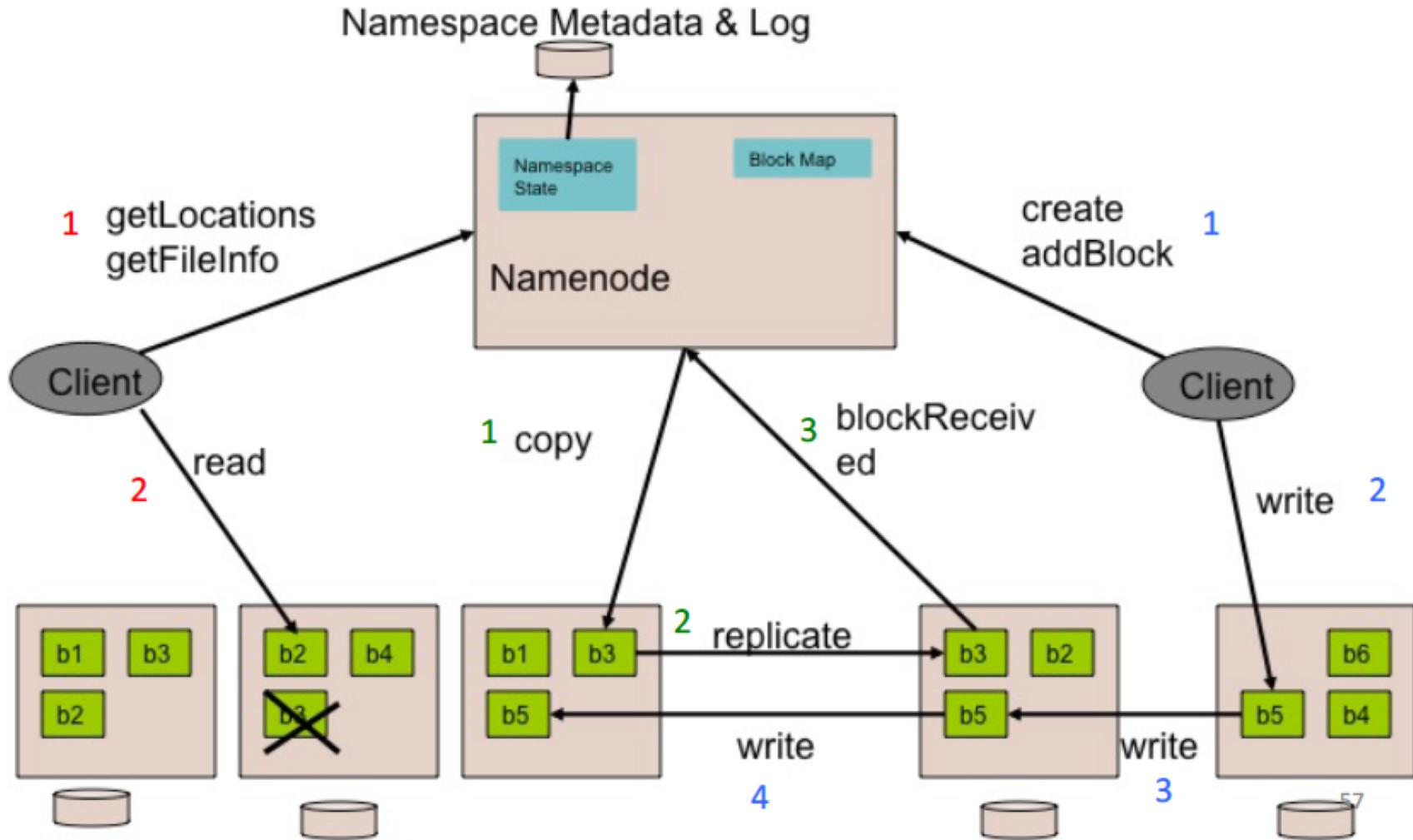
Ecrire un fichier sur HDFS

1. Le client envoie une demande au *NameNode* pour créer un nouveau fichier.
2. Le *NameNode* vérifie les autorisations du client et les conflits tels que existence du fichier.
3. Le *NameNode* choisit les *DataNodes* pour stocker les fichiers et leurs réplicas. Il constitue une pipeline de *DataNodes*.
4. Les blocks sont alloués sur ces *DataNodes*.
5. Le flux de donné est envoyé depuis le client vers le premier *DataNode* du pipeline.
6. Chaque *DataNode* transfert les données reçues au *DataNode* suivant dans le pipeline.

Lire un fichier sur HDFS

1. Le client envoie une demande au *NameNode* pour lire un fichier
2. Le *NameNode* vérifie que le fichier existe et construit une liste de *DataNodes* contenant les premiers blocks.
3. Pour chaque block, le *NameNode* envoie l'adresse des *DataNodes* les hébergeant (la liste est ordonnée en fonction de la proximité au client).
4. Le client se connecte au *Datanode* le plus près contenant le premier block du fichier
5. La fin de lecture du block
 - Ferme la connexion au *Datanode*.
 - Ouvre la connexion au *Datanode* contenant le block suivant.
6. Lorsque tous les blocks sont lus, le client questionne le *NameNode* sur les blocks suivants.

Protocol de lecture/écriture/réplication sur HDFS



HDFS : Hadoop distrib. file system

- Se logger sur son compte
- Accès à l'espace de stockage HDFS

>> *hadoop fs -COMMANDER PARAMETRES*

>> *hadoop fs -ls data*

>> *hadoop fs -mkdir data/mots*

>> *hadoop fs -put livre.txt data/mots/liste.txt*

>> *hadoop fs -get ...*

>> *hadoop fs -rm -r -f sortie*

- Commandes

<https://www.edureka.co/blog/hdfs-commands-hadoop-shell-command>

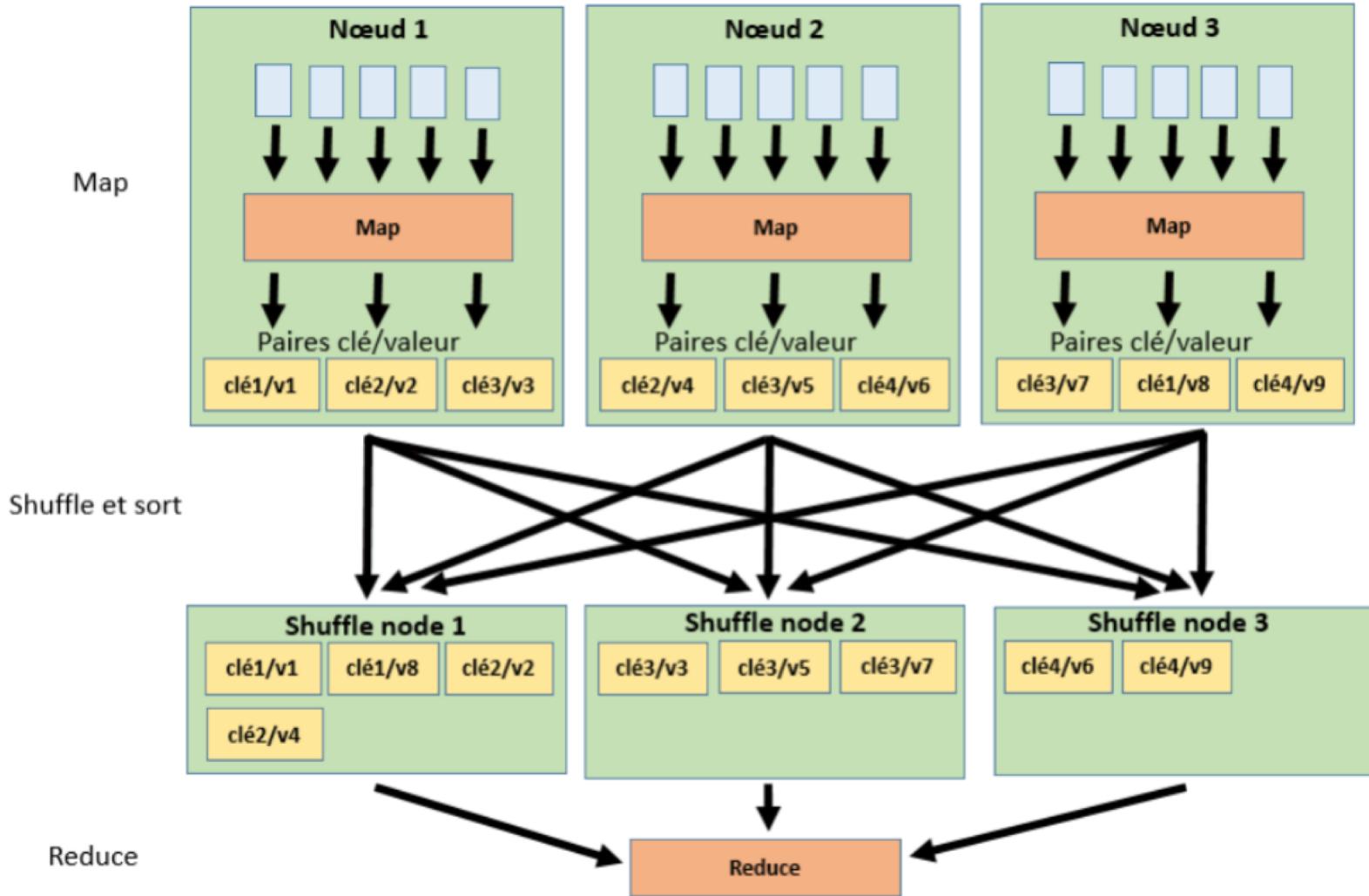
ou

<https://cdiese.fr/commandes-shell-courantes-pour-hdfs/>

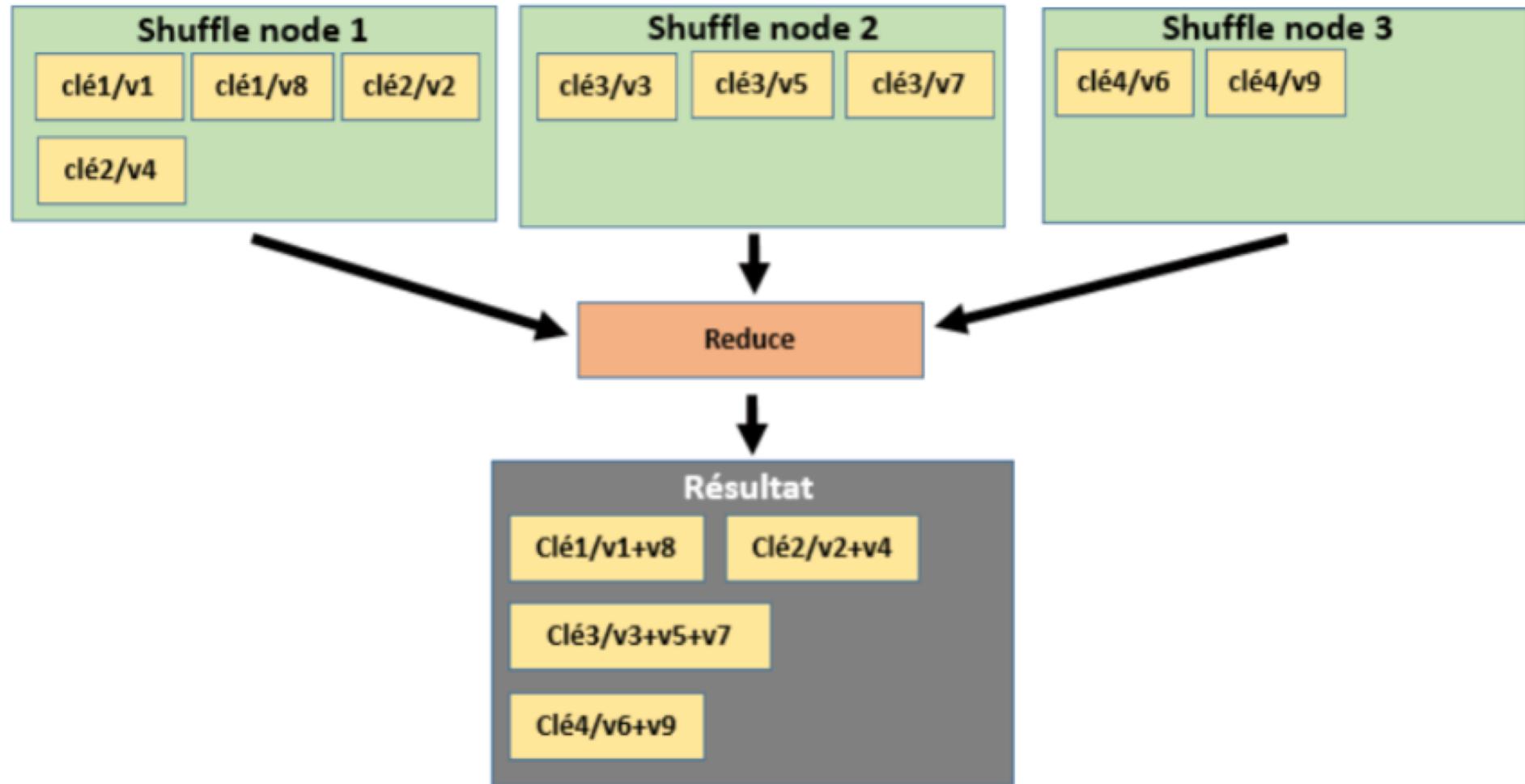
MAP-REDUCE

Map Reduce sur cluster Hadoop

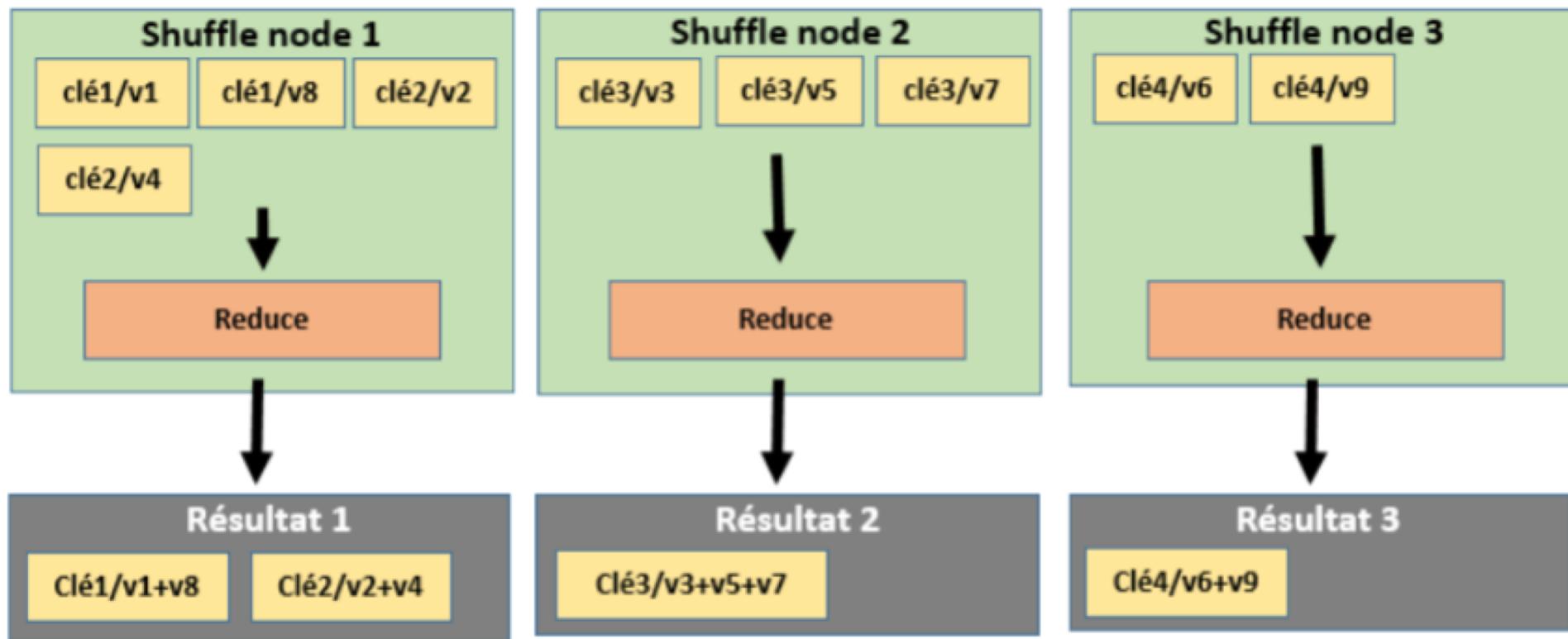
Map Reduce : shuffling and sorting



Map Reduce : reducing



Map Reduce : multiple reducing



WordCount sur un cluster HDFS

```
>> export STREAMINGJAR='.../hadoop-streaming-2.7.2.jar'  
  
>> hadoop jar $STREAMINGJAR -files wc_mapper.py,wc_reducer.py -mapper  
wc_mapper.py -reducer wc_reducer.py -input livres/dracula -output sortie
```

hadoop-streaming.jar : librairie Java utilisée lors de l'exécution d'opérations MapReduce.

Elle établit un lien entre Hadoop et le code externe MapReduce que vous fournissez.

-files : indique à Hadoop que les fichiers spécifiés sont nécessaires pour effectuer cette tâche MapReduce, et qu'ils doivent être copiés sur tous les nœuds de travail.

-mapper : indique à Hadoop quel fichier doit être utilisé comme *mappeur*.

-reducer : indique à Hadoop quel fichier doit être utilisé comme *reduceur*.

-input : indique le nom du fichier d'entrée sur lequel s'applique la tache (si vous spécifiez un répertoire, le Job s'appliquera à tous les fichiers du répertoire).

-output : le répertoire sur lequel la sortie sera écrite. Ce répertoire sera créé par la tâche. Il doit être détruit entre deux exécutions, sous peine d'un message d'erreur lors de l'exécution (>> hdfs dfs -rm -r -f sortie)

Nombre de taches en parallèle :

- D mapred.reduce.tasks=3

- D mapred.map.tasks=5

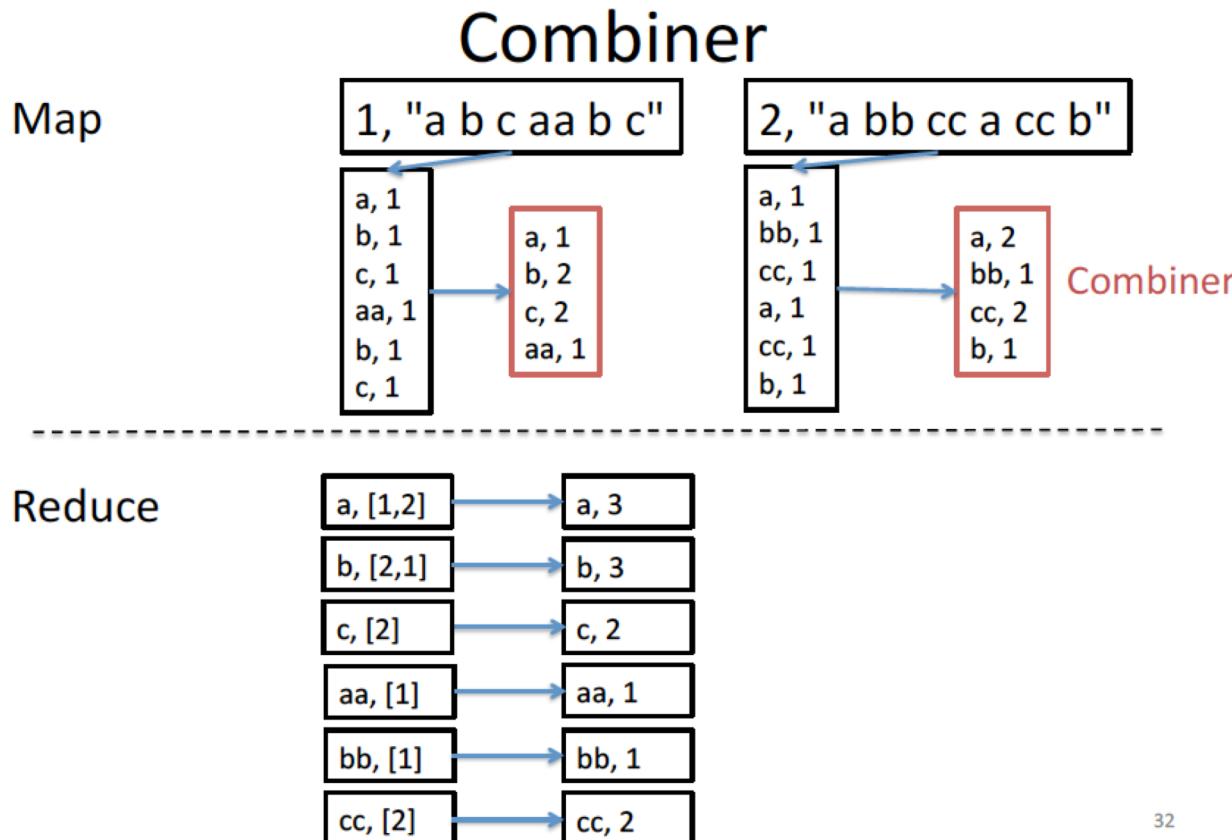
Hadoop n'honore pas nécessairement mapred.map.tasks (cela est considéré comme une éventuelle aide pour choisir automatiquement le nombre adéquat).

- D mapred.reduce.tasks=0 pour un job map-only

Le combiner

- Potentiel problème du mapper : de nombreuses paires (clé, valeur) à la sortie:
 - Envoyés au reducer à travers le réseau (cf shuffling)
 - Etape très couteuse en terme de temps d'exécution
- Ajout d'une opération : le combiner
 - Peut être vu comme un mini-reducer, qui travaille au niveau de chaque mapper, pour commencer l'agrégation.
 - C'est une étape Hadoop optionnelle : le résultat ne doit pas dépendre d'elle.

Le combiner



32

```
>> hadoop jar $STREAMINGJAR -files wc_mapper.py,wc_reducer.py -mapper  
wc_mapper.py -combiner wc_reducer.py -reducer wc_reducer.py -input  
livres/dracula -output sortie
```

Si la fonction « reduce » est à la fois commutative et associative, alors on peut utiliser le reducer comme combiner !

GÉNÉRATEUR ET ITÉRATEUR EN PYTHON

Map-Reduce improved!

Iterators en python

Un **itérateur** est une sorte de curseur qui a pour mission de se déplacer dans une séquence d'objets. L'itérateur permet de parcourir chaque objet d'une séquence sans se préoccuper de la structure sous-jacente.

Une liste et une liste en compréhension sont des itérateurs

```
# liste : iterator
liste=[1,2,3,4,5,6,7,8,9,10]
for x in liste:
    print(x)
```

```
# liste en comprehension : iterator
a_list=[1,9,8,4]
A=[elem*2 for elem in a_list]
print(A)
```

Iterators en python

```
class Fib:
    def __init__(self, max):
        self.max=max

    def __iter__(self):
        self.a=0
        self.b=1
        return self

    def __next__(self):
        fib=self.a
        if fib>self.max:
            raise StopIteration
        self.a, self.b = self.b, self.a+self.b
        return fib

if __name__=="__main__":
    fib=Fib(100)
    for n in fib:
        print(n, end=' ')
print(list(Fib(200)))
```

Generators en python

Un **générateur** permet de simplifier la création d'itérateurs.

Le mot clé **yield** est un peu similaire au **return** des fonctions sauf qu'il ne signifie pas la fin de l'exécution de la fonction mais une mise en pause et à la prochaine itération la fonction recherchera le prochain **yield**.

```
def generateur1():
    yield "a"
    yield "b"
    yield "c"

g1=generateur1()
for v in g1:
    print(v)

def generateur2(n):
    for i in range(n):
        if i==5:
            print("Ceci est le 5eme tour")
        yield i+1

g2=generateur2(10)
for v in g2:
    print(v)
```

WordCount amélioré avec itérateurs et générateurs

```
import sys

def read_input(file):
    for line in file:
        # split the line into words
        yield line.split()

def main(separator='\t'):
    # input comes from STDIN (standard input)
    data=read_input(sys.stdin)
    for words in data:
        # tab-delimited; the trivial word count is 1
        for word in words:
            print(word, separator, '1')

if __name__=="__main__":
    main()
```

Amélioration avec itérateurs et générateurs

```
# fichier wc_mapper_improved
from itertools import groupby
from operator import itemgetter
import sys

def read_mapper_output(file, separator='\t'):
    for line in file:
        yield line.rstrip().split(separator, 1)

def main(separator='\t'):
    data=read_mapper_output(sys.stdin, separator=separator)

    # groupby groups multiple word-count pairs by word,
    # and creates an iterator that returns consecutive keys and their group:
    # current_word - string containing a word (the key)
    # group - iterator yielding all ["current_word", "count"] items
    for current_word, group in groupby(data, itemgetter(0)):
        try:
            total_count=sum(int(count) for current_word, count in group)
            print(current_word, separator, total_count)
        except ValueError:
            # count was not a number, so silently discard this item
            pass

if __name__=="__main__":
    main()
```

ECOSYSTÈME HADOOP

Qui utilise Hadoop?



Google

ebay



Microsoft

YAHOO!

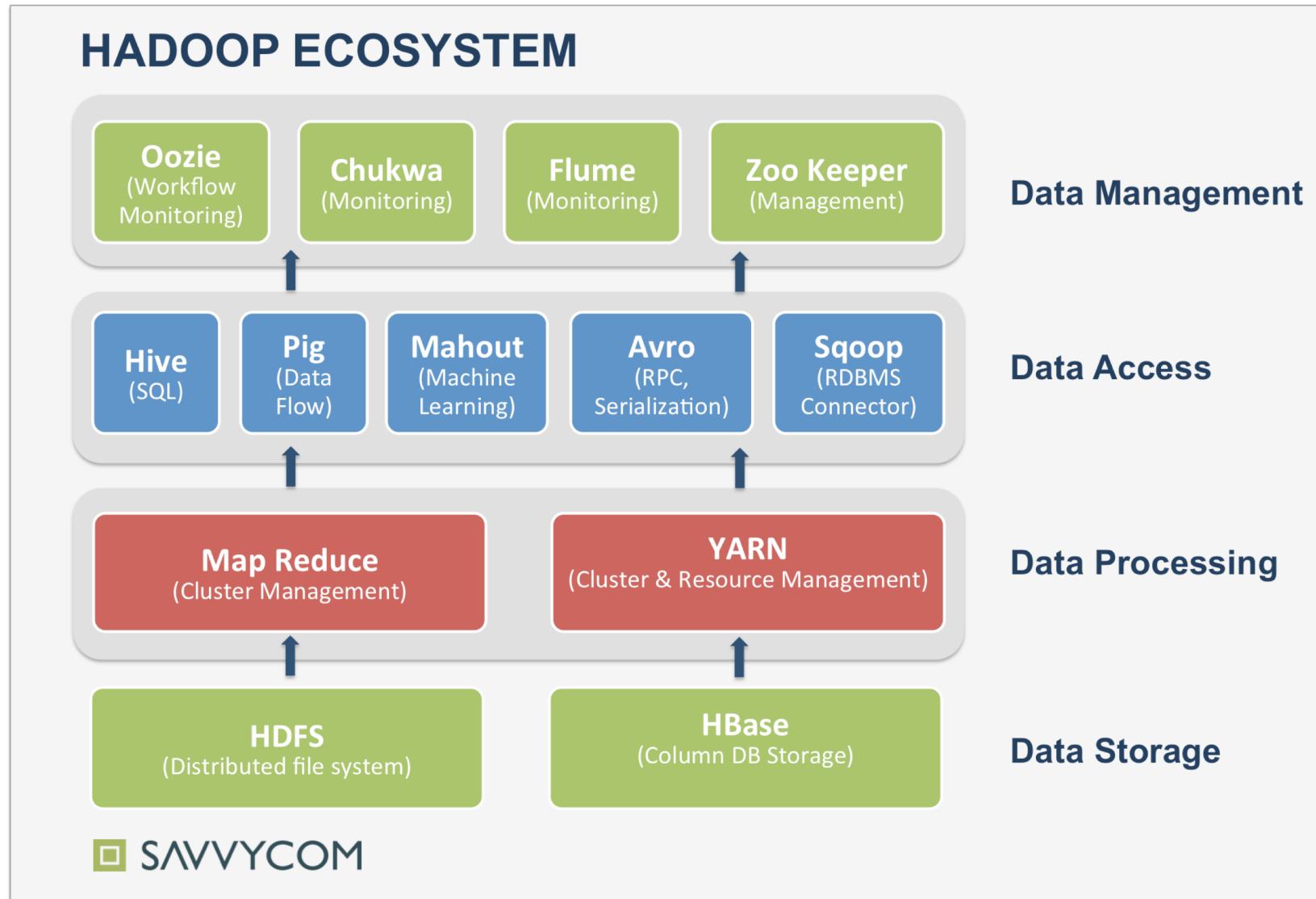
twitter



Massachusetts
Institute of
Technology

amazon.com

Ecosystème Hadoop



Description complète : <https://hadoopecosystemtable.github.io>

Quelques outils 1/3

HIVE et **PIG** sont deux outils qui permettent d'interagir avec les données contenues dans HDFS en exécutant des jobs MapReduce de façon transparente.

Hive est un outil développé par Facebook qui permet d'utiliser HADOOP par le biais de requêtes HiveQL. Ce langage est proche du SQL. Il permet de construire un modèle de données relationnel basé sur les données contenues dans HDFS.

Ce modèle de données est stocké dans un *métastore* qui gère la définition des tables ainsi que les métadonnées.

Pig est un outil développé par Yahoo qui permet d'exécuter des jobs mapreduce en utilisant un langage de scripting (PigLatin). Les scripts PigLatin permettent de travailler sur des données contenues dans HDFS. Ces données sont traitées par MapReduce qui retourne le résultat de ses calculs à Pig.



Quelques outils 2/3

Flume: Il s'agit d'un outil permettant d'injecter de gros volumes de données en temps réel dans HDFS. Cet outil est capable de « streamer » des données depuis n'importe quelle source pour les ajouter dans HADOOP. Flume est extensible : il intègre des données en provenance de sources variées.



HBASE: C'est une base de données non relationnelle distribuée. Elle a la particularité d'utiliser les données directement dans HDFS sans passer par des tâches MapReduce. HBASE présente donc des caractéristiques assez similaires à celles de HDFS (capacité à gérer des volumes de données de plusieurs Po, forte tolérance de panne...). HBASE est bien adaptée pour gérer des données parsemées comme par exemple une table de plusieurs milliers de colonnes avec une majorité de cellules vides.



Quelques outils 3/3

Oozie est un outil d'ordonnancement de jobs HADOOP. En effet, de nombreux traitements ne peuvent être réalisés par un seul job. Il faut alors créer et gérer une chaîne de traitement. C'est ce que permet de faire Oozie. Il est également capable de gérer des dépendances temporelles, d'exécuter une tâche plusieurs fois de suite ou encore de faire remonter d'éventuelles erreurs.



HUE est un outil permettant d'obtenir une interface graphique de HADOOP. Cette interface graphique comprend un navigateur de fichiers permettant d'accéder à HDFS, un navigateur de job MapReduce, un navigateur HBASE ainsi qu'un éditeur de requête pour Hive et Pig.

