UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

INTRODUCTION À

ET

Julien Nauroy - Direction Informatique

http://www.informatique-scientifique.u-psud.fr

université PARIS-SACLAY

Hadoop : Quel usage ?

- J'ai un ensemble de données assez grand
  - Disons quelques To
- J'ai des calculs à faire sur cet ensemble
  - Ex.: compter les occurrences de chaque mot
- Mon problème se découpe bien
  - « embarassingly parallel »
- Une seule machine ne suffit pas
  - Ex.: débit disque, contrainte de temps

Un problème classique ?

- Je souhaite traiter 8To de données

En local ?

En réseau ?

En local

- 8To = 1 disque, ouf ! (enfin presque)
  - Disque dur « classique », ~250€ HT
  - 30k IOPS en lecture séquentielle ~100mo/s
  - ~100mo/s => 20h de lecture continue

Un problème de temps

Sur le réseau

- Baie de 60 disques
  - Vitesse de lecture x60 => 20 min de transfert ?
  - NON : il faudrait 60Gbps de bande-passante
    - Contrôleur et réseau

Un problème de débit

Et les SSD ?

- 8To = 4 disques (par exemple)
  - 120k IOPS en lecture séquentielle ~400mo/s
  - 8To en ~1h15

Pas de problème de temps

Pas de problème de débit

Un problème de calcul
(à venir)

Hadoop : un paradigme adapté

- Prenons 15 machines
  - 4 disques durs par machine
  - Total 60 disques
- Découpons nos données en 15 parties
  - Chaque machine reçoit 1/15e
- Calculons les résultats partiels
- Agrégeons le résultat final

prélocalisation

paradigme spécifique

Hadoop – les fondamentaux

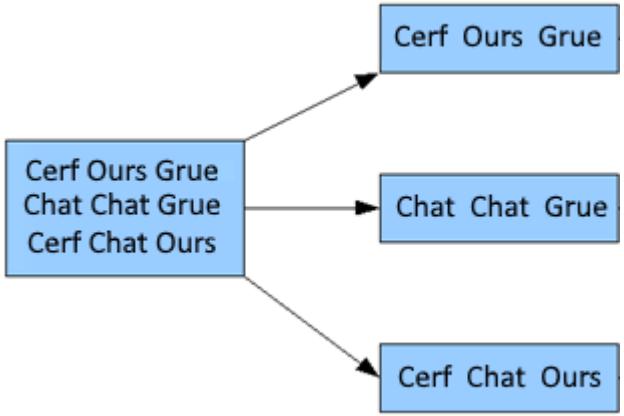
- Cœur de Hadoop
  - Un système de fichiers réparti
    - Hadoop Distributed File System (HDFS)
  - Un système de répartition des calculs
    - MapReduce
- Beaucoup d'autres outils
  - Stockage : Hbase, Parquet, …
  - Calcul : Hive (SQL), Spark, …

Le paradigme MapReduce

1. « **MAP** » : Lecture des données et production de **couples (clé, valeur)**
   - Les clés servent à organiser les données
   - Équivalent de « group by » SQL
2. « **REDUCE** » : Regroupement des clés et traitement des valeurs

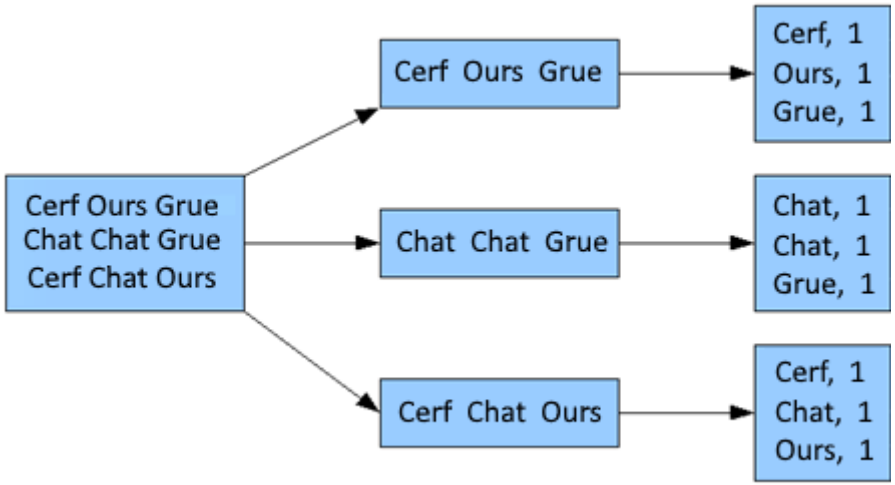
Un exemple : « wordcount »

- Préliminaire : Copie des fichiers
  - HDFS : système de fichiers réparti



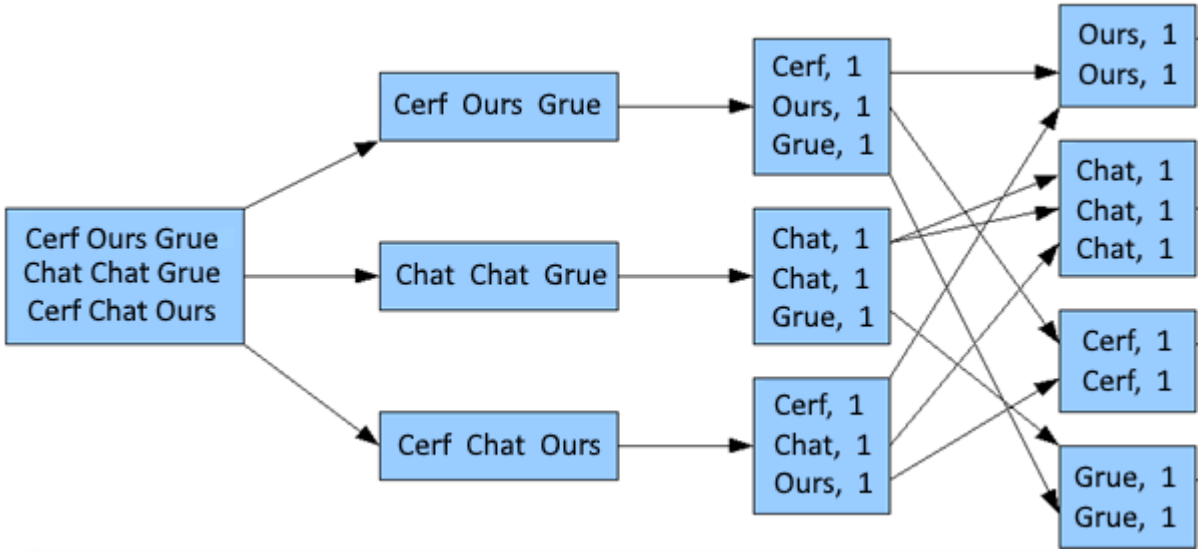
Un exemple : « wordcount »

- Etape 1 : **Map** – production de (clé, valeur)



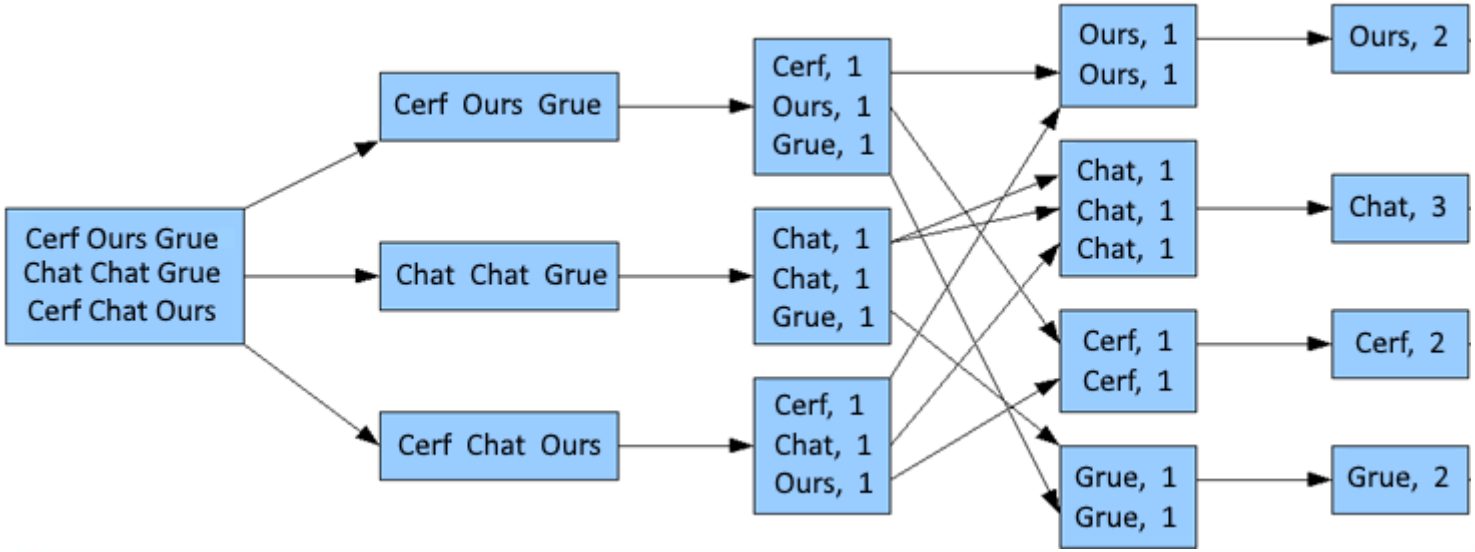
Un exemple : « wordcount »

- Etape 2 : Regroupement des clés
  - « Shuffle & Sort » (automatique)



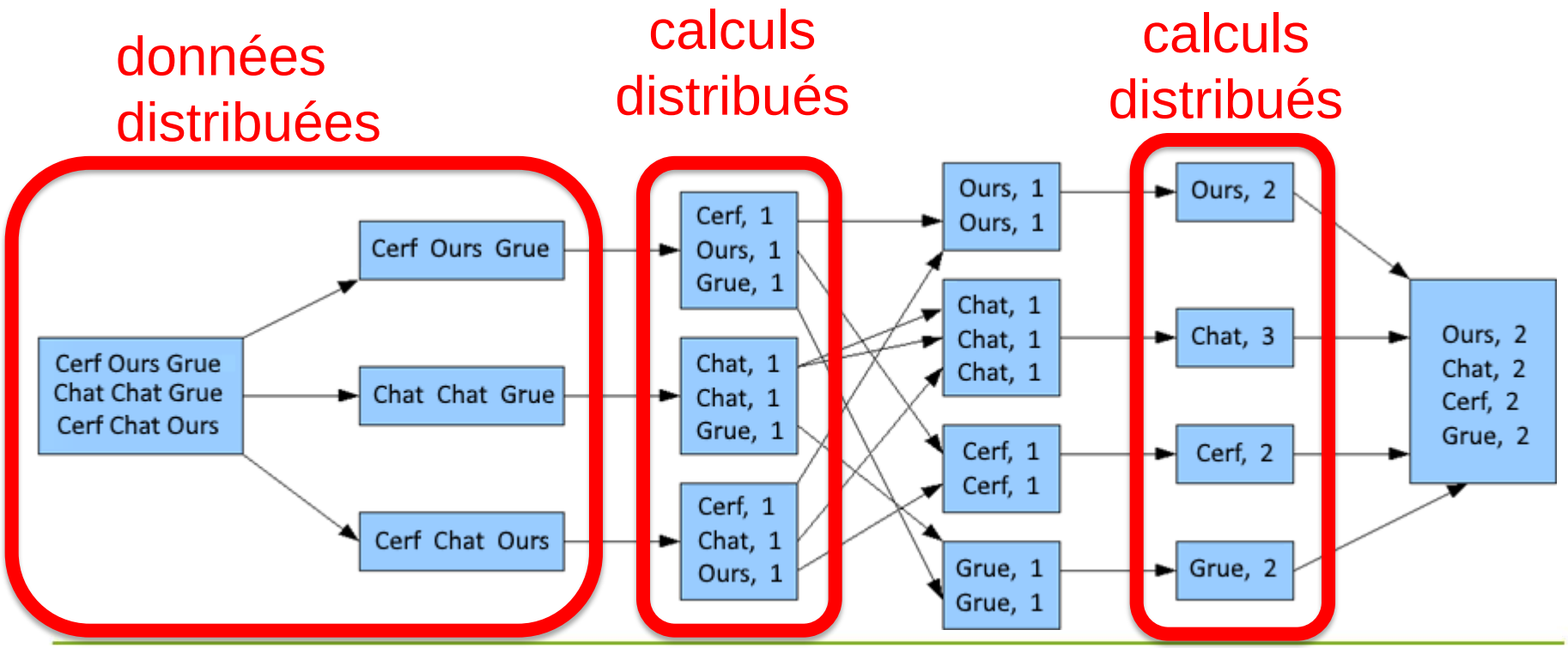
Un exemple : « wordcount »

- Etape 3 : **Reduce** – somme des valeurs



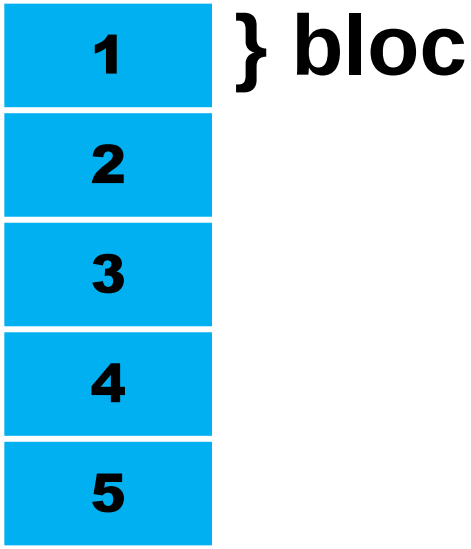
Un exemple : « wordcount »

- Etape 4 : écriture du résultat
  - automatique



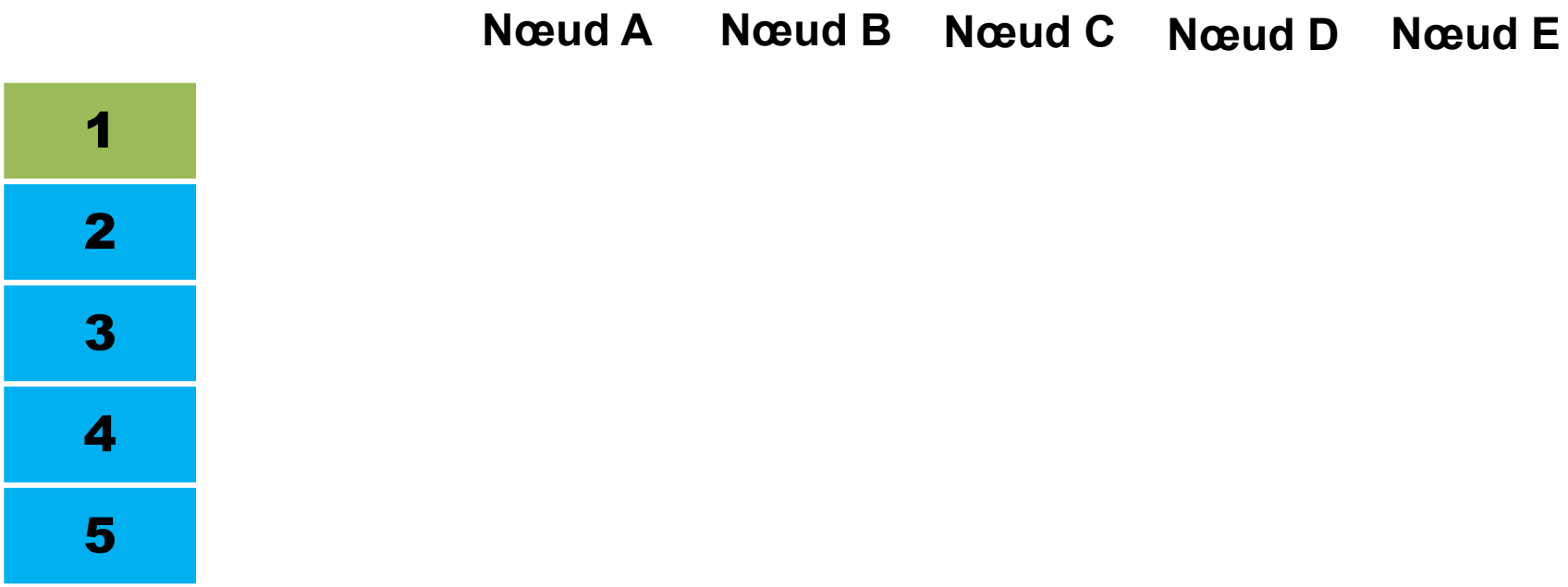
Fonctionnement de HDFS

1. Découpage des fichiers en blocs
   - Taille indicative : 128Mo



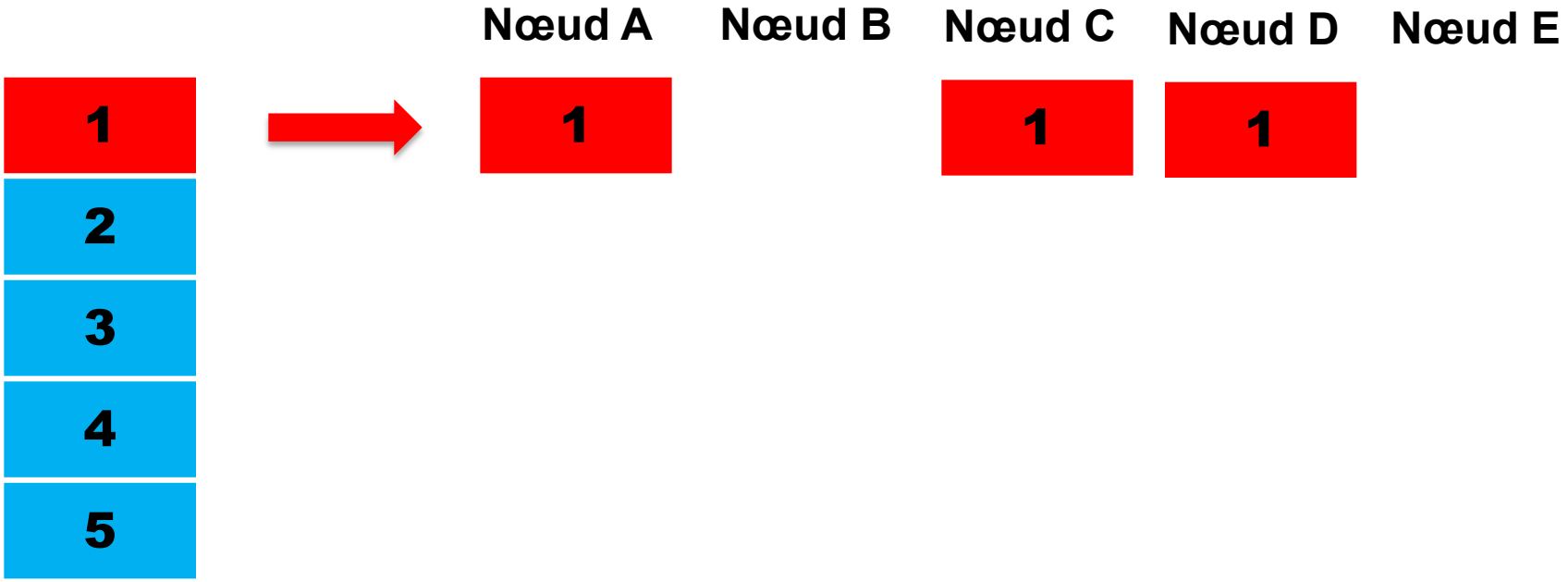
Fonctionnement de HDFS

2. Copie des blocs sur plusieurs nœuds



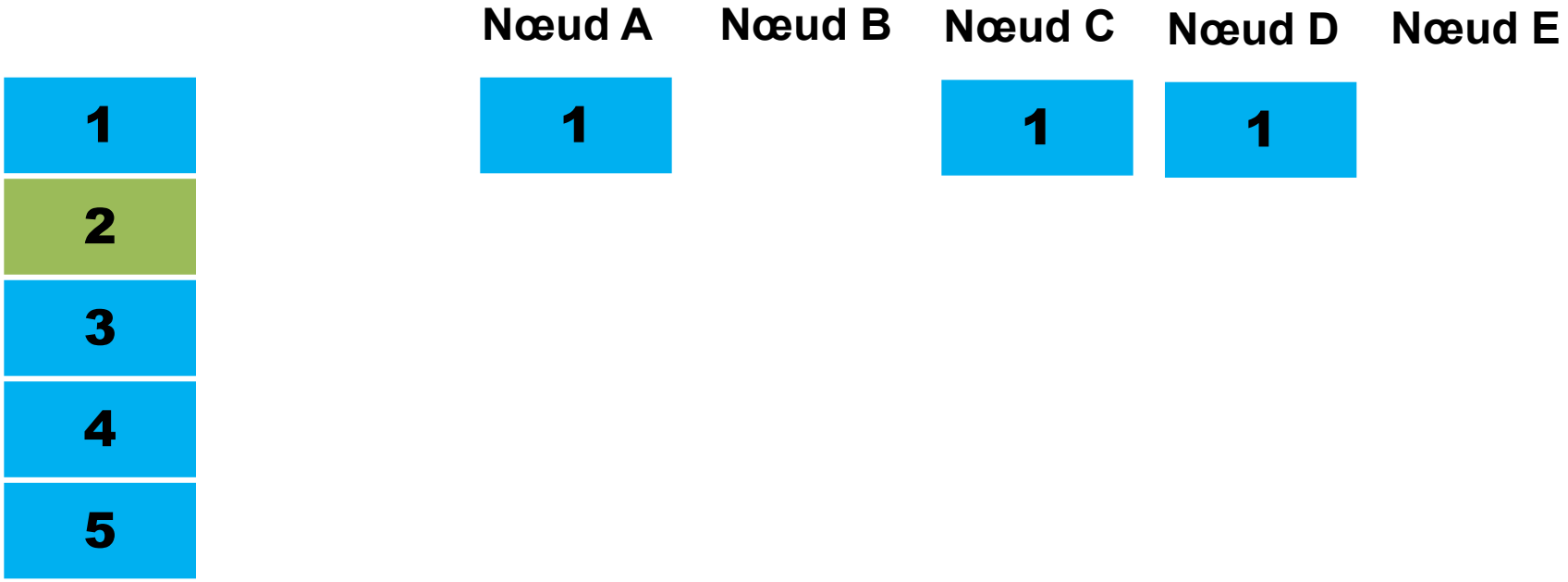
Fonctionnement de HDFS

2. Copie des blocs sur plusieurs nœuds
   - En général, 3 nœuds



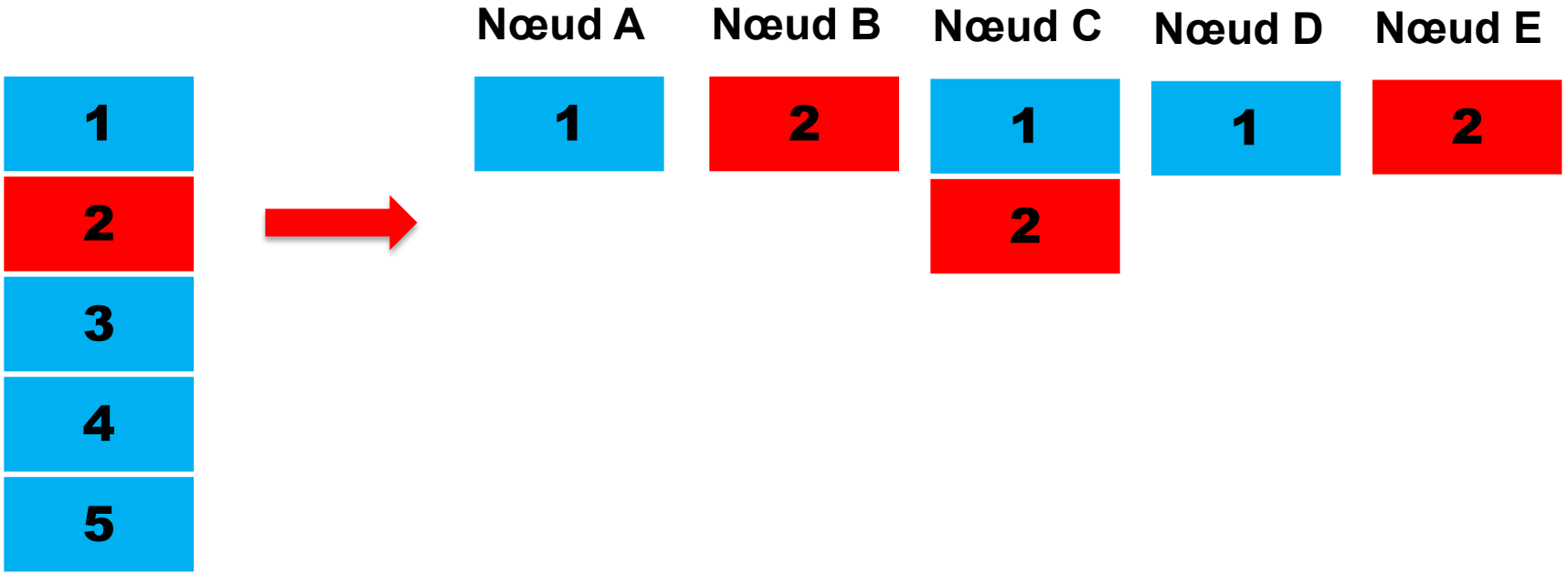
Fonctionnement de HDFS

2. Copie des blocs sur plusieurs nœuds
   - En général, 3 nœuds



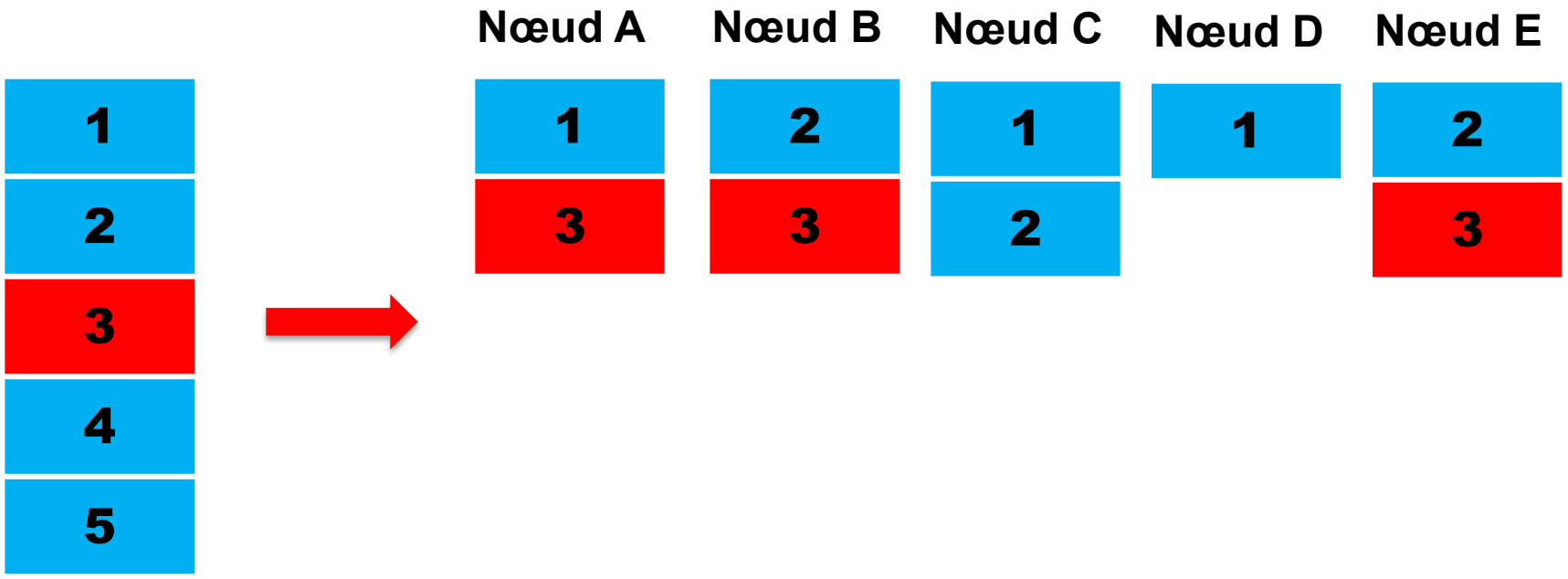
Fonctionnement de HDFS

2. Copie des blocs sur plusieurs nœuds
   - En général, 3 nœuds



Fonctionnement de HDFS

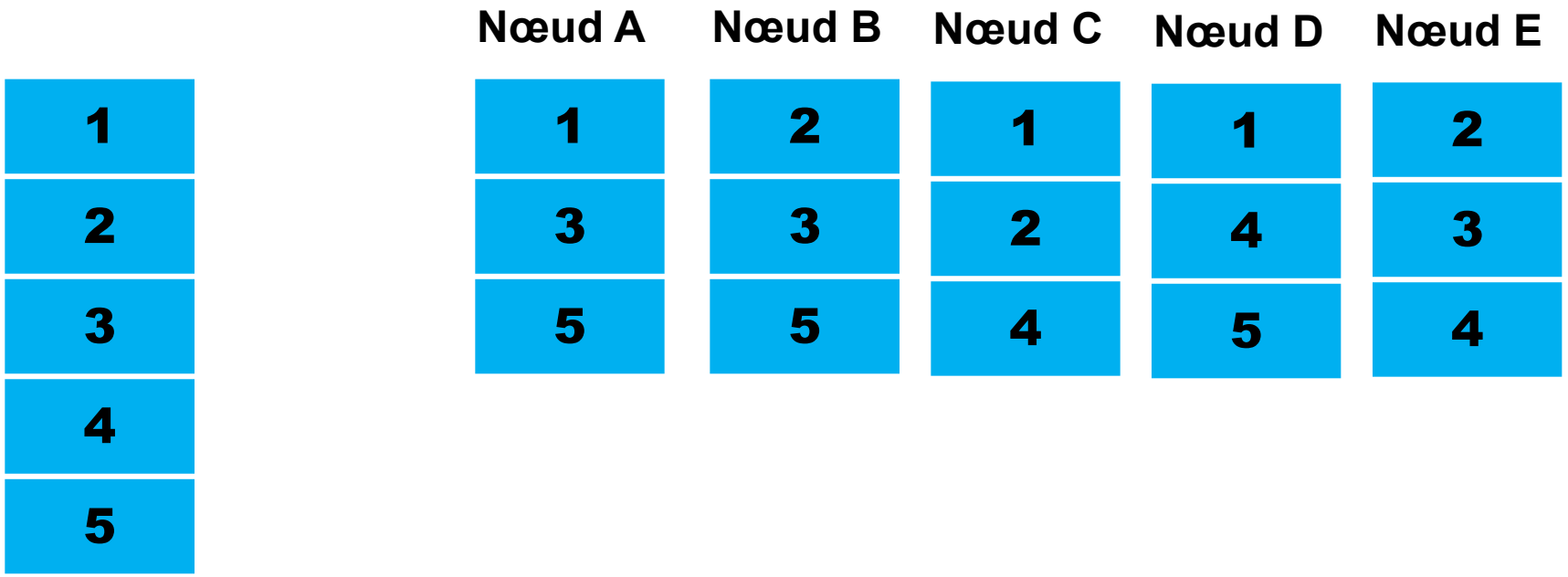
2. Copie des blocs sur plusieurs nœuds
   - En général, 3 nœuds



Fonctionnement de HDFS

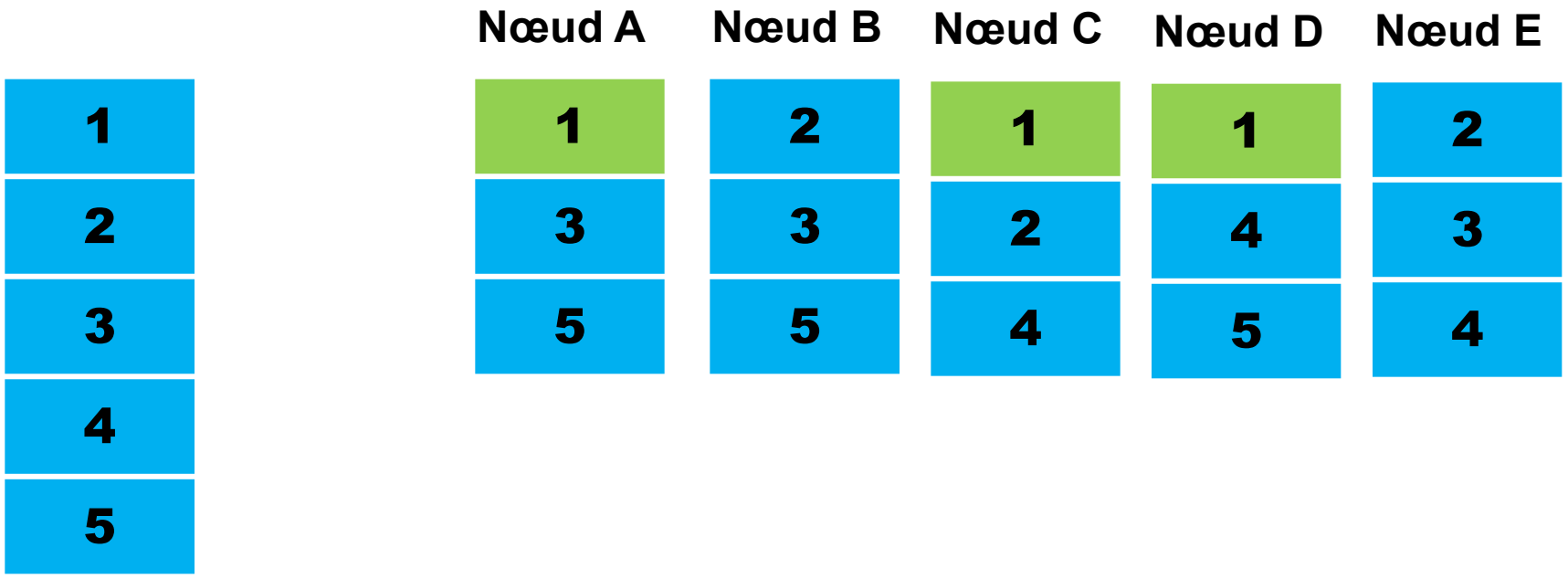
2. Copie des blocs sur plusieurs nœuds

- En général, 3 nœuds



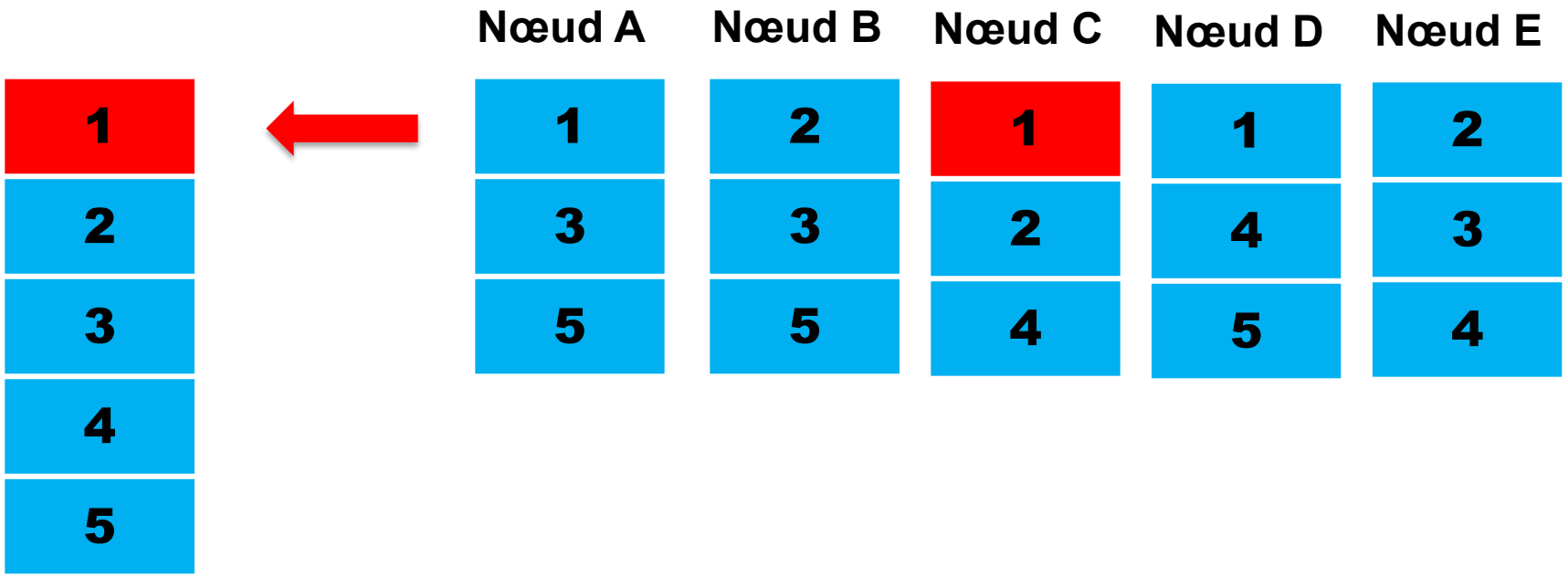
Fonctionnement de MapReduce

1. Sélection des nœuds portant les calculs
   - Les données doivent être sur le nœud



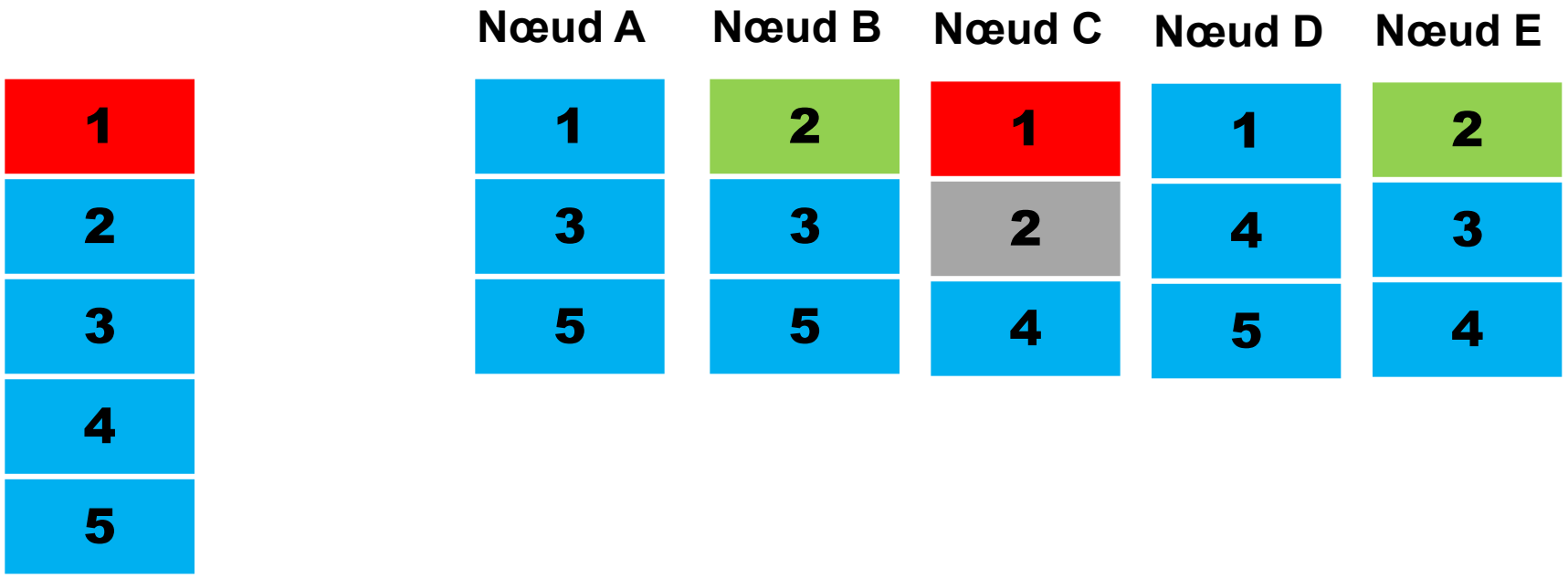
Fonctionnement de MapReduce

1. Sélection des nœuds portant les calculs
   - Les données doivent être sur le nœud



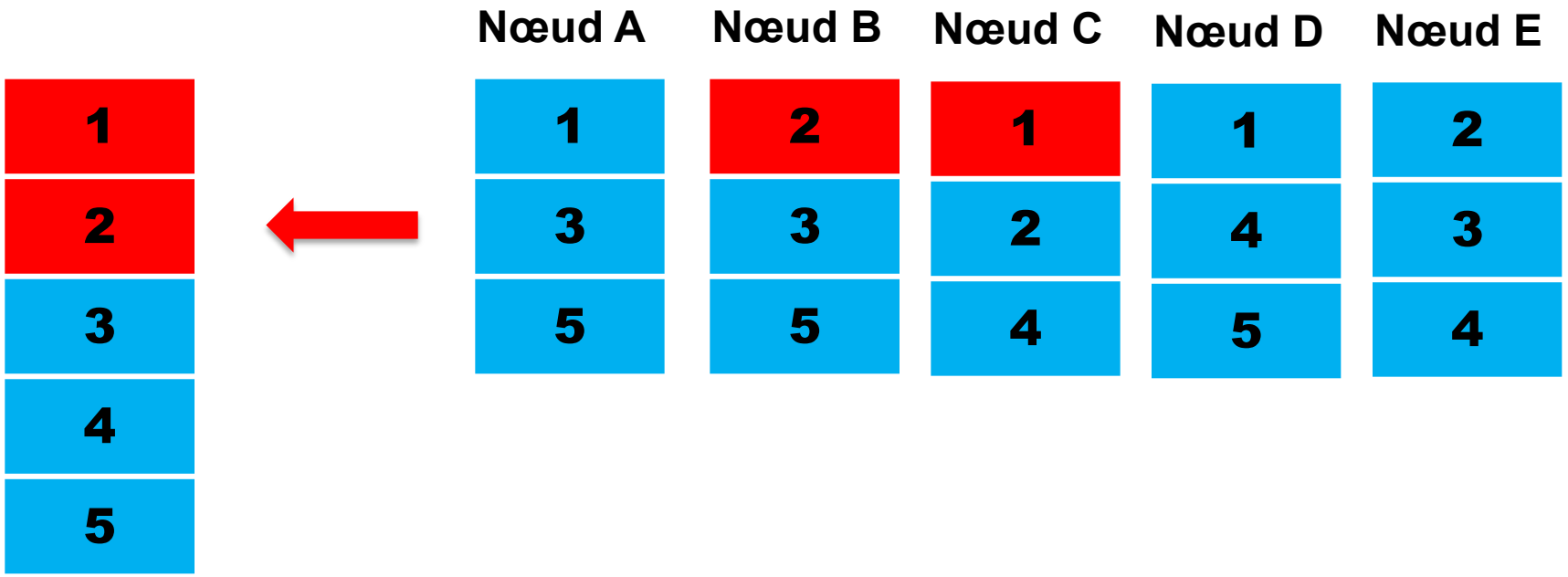
Fonctionnement de MapReduce

1. Sélection des nœuds portant les calculs
   - Le nœud ne doit pas être occupé



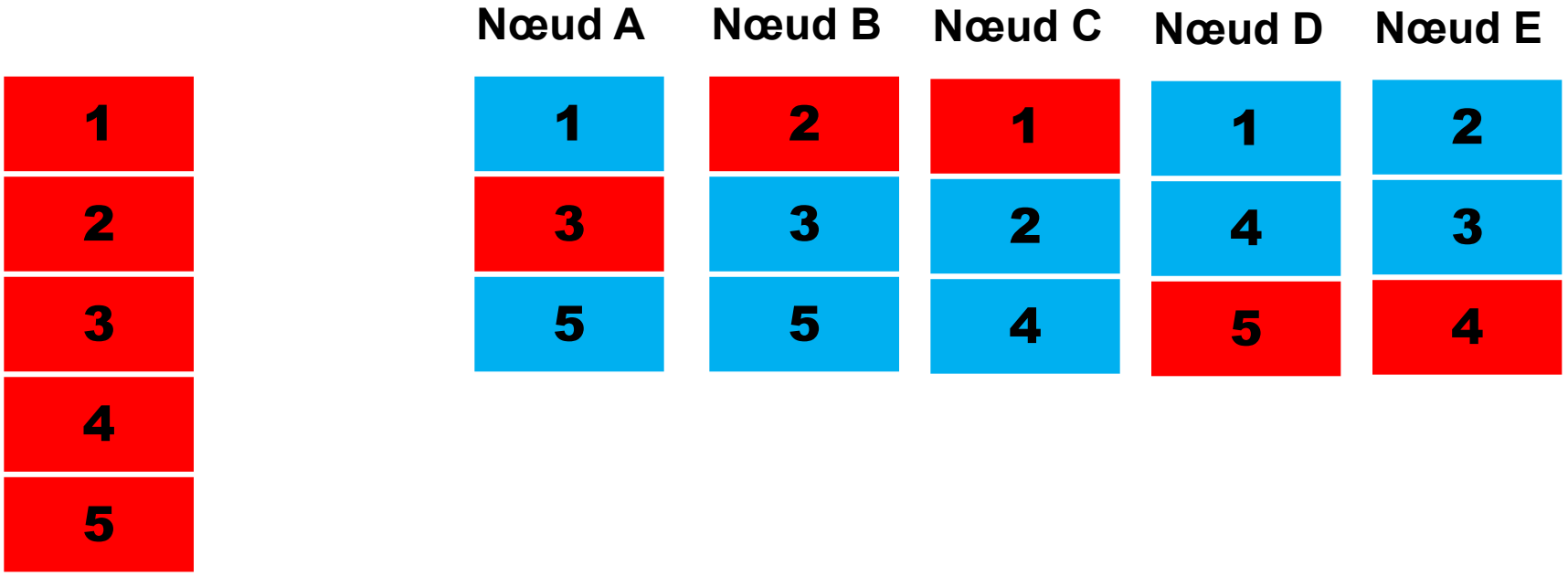
Fonctionnement de MapReduce

1. Sélection des nœuds portant les calculs
   - Les données doivent être sur le nœud



Fonctionnement de MapReduce

1. Sélection des nœuds portant les calculs
   - Tous les blocs doivent être traités



Forces de Hadoop

- Données de grande taille
  - Lorsque goulet d'étranglement = débit disque
    - En particulier débit disque >> débit réseau
  - Point fort : pré-localisation des données
  - Décompression à la volée (Gzip, LZO…)
- Passage à l'échelle
  - 3400 machines trient 100TB en moins de 3h
- Fiabilité
  - Redondance dans HDFS

Faiblesses de Hadoop

- HPC (systèmes parallèles synchronisés)
  - Pas de communication possible
- Temps réel
  - Autres projets, p.ex.: Storm
- Infrastructures de petite taille
  - < 5 nœuds = sous-optimal
- Workflows à beaucoup d'étapes
  - Ecritures intermédiaires sur disque

EXECUTION D'UN JOB MAPREDUCE

Représentation des données

```
public class NCDCData implements Serializable {
    public String USAFID;
    // Note : la temperature est exprimee
    // en dixiemes de degres
    public int airTemperature;
    public String airTemperatureQuality;
    // (...)

    public NCDCData(String s) {
        this.USAFID = s.substring(4, 10);
        // (...)
    }
}
```

Mapper

```
public class DemoMapper extends
        Mapper<Object, Text, Text, IntWritable> {

    public void map(Object key, Text value, Context context)
            throws IOException, InterruptedException {
        // line contient une ligne du fichier
        String line = value.toString();
        // On cree la structure NCDCData a partir de la ligne
        NCDCData data = new NCDCData(line);
        // Si la qualite de l'acquisition de temperature
        // est bonne, alors on cree le couple (cle, valeur)
        if (data.airTemperature != 9999) {
            context.write(new Text(data.USAFID),
                    new IntWritable(data.airTemperature));
        }
    }
}
```

Typage

Production (clé, valeur)

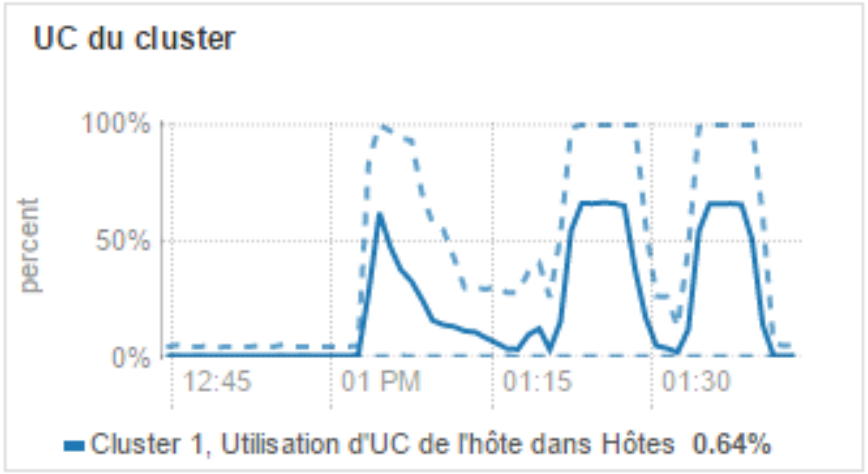
Reducer

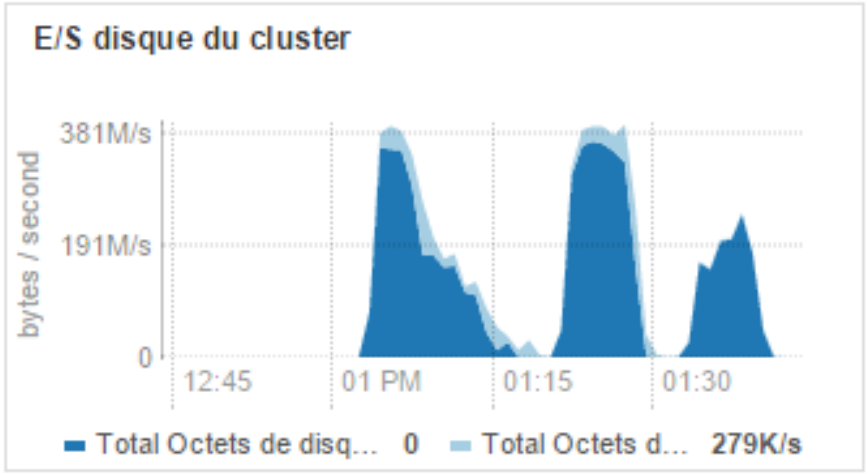
```
public class DemoReducer extends
             Reducer<Text, IntWritable, Text, IntWritable> {

    public void reduce(Text key, Iterable<IntWritable> values,
            Context context) throws /* ... */ {

        int max = Integer.MIN_VALUE;
        for (IntWritable val : values) {
            int intval = val.get();
            if (intval > max) max = intval;
        }
        IntWritable result = new IntWritable(max);
        context.write(key, result);
    }
}
```

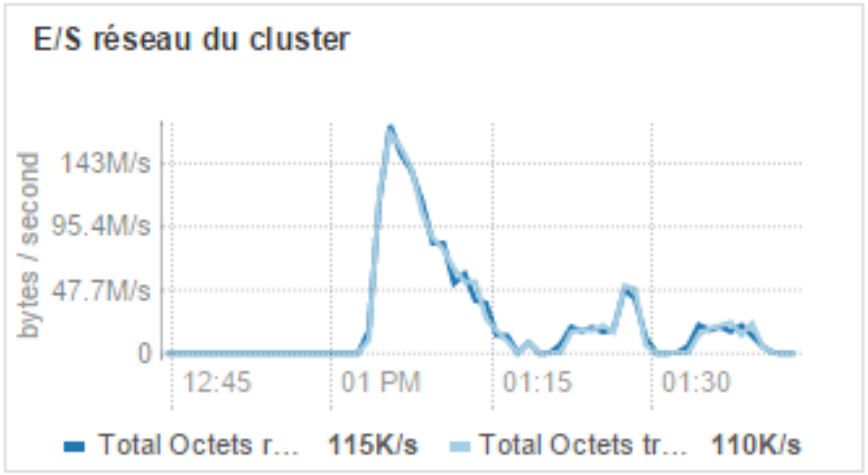
Typage

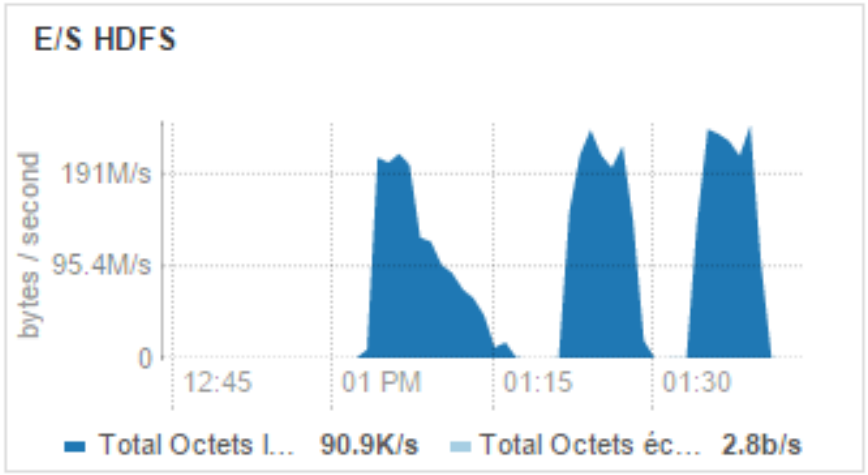
Production (clé, valeur)

Exemples d'exécution



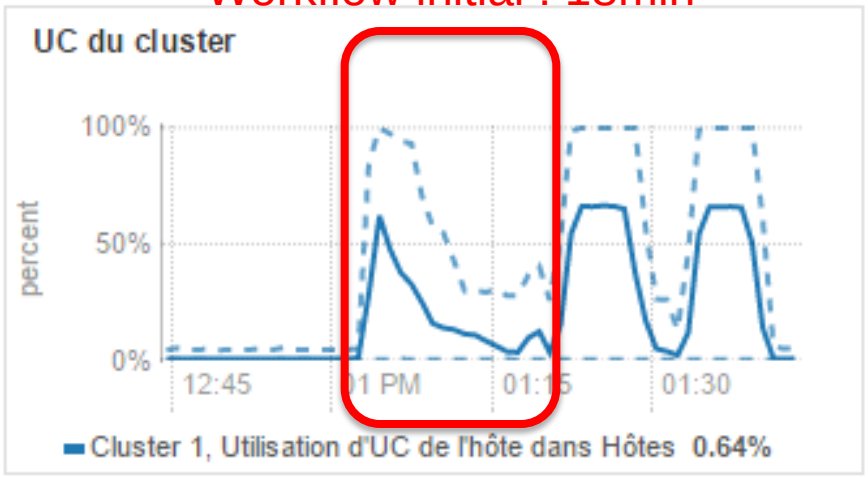


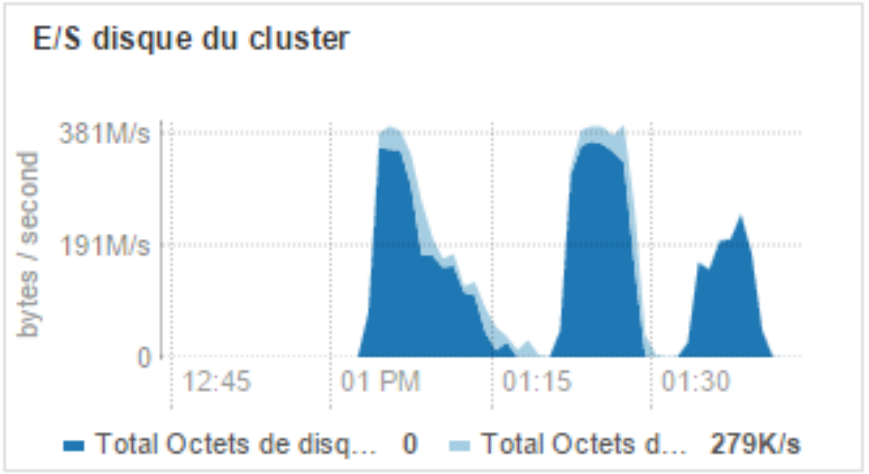


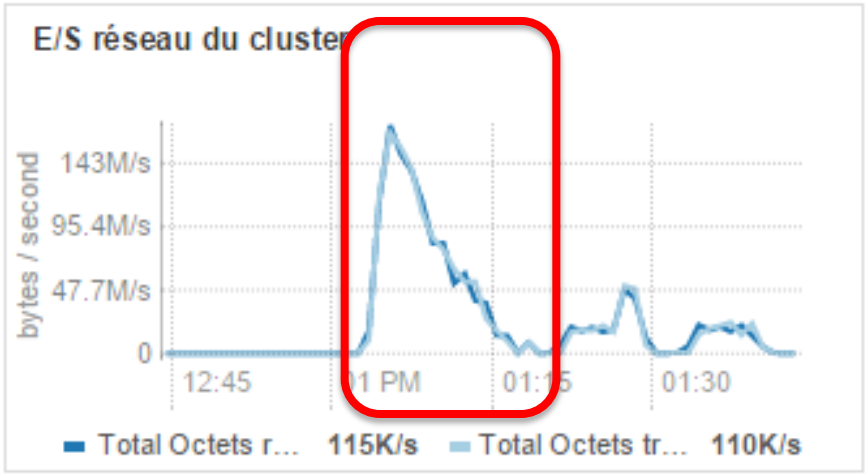


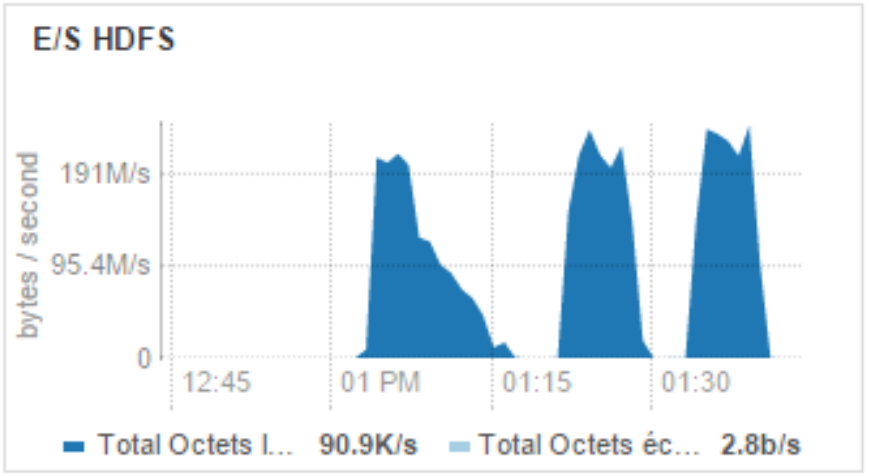
Exemples d'exécution

Workflow initial : 18min



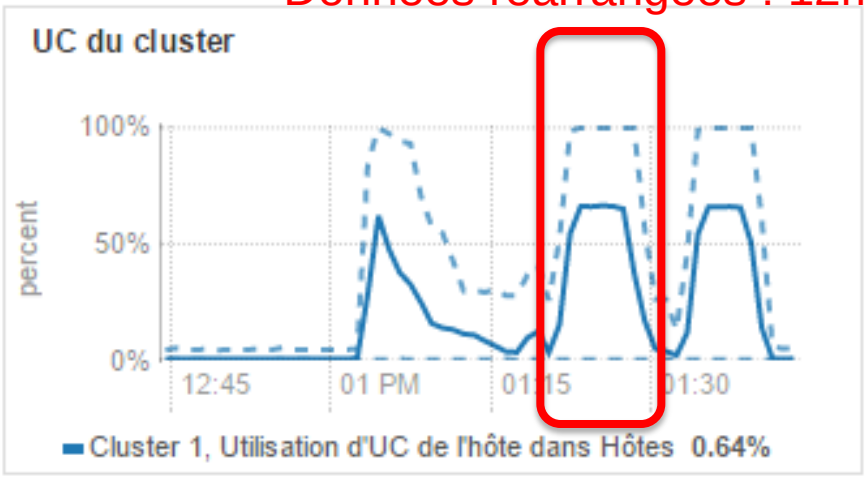


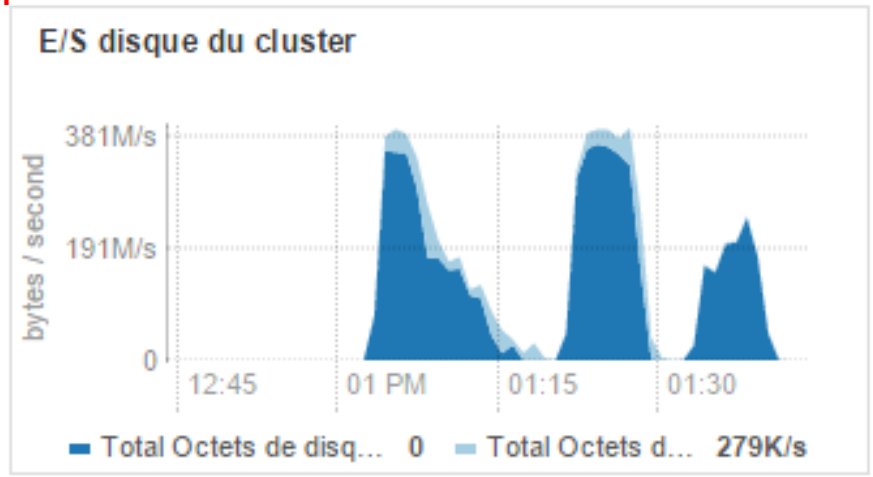


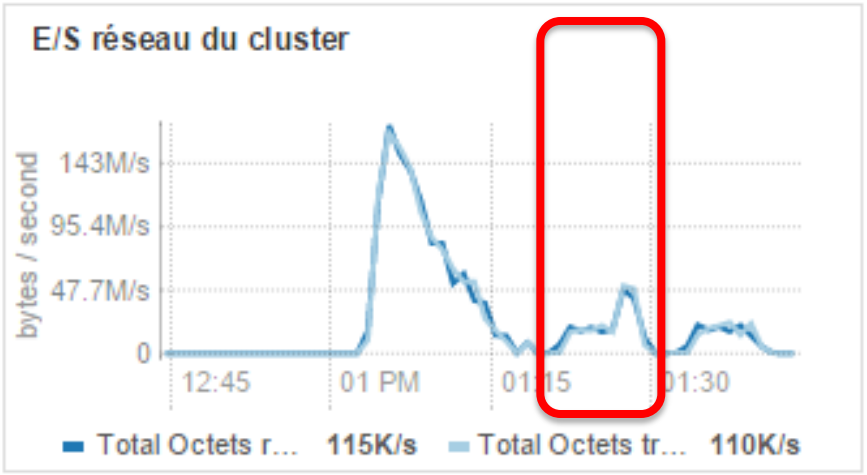


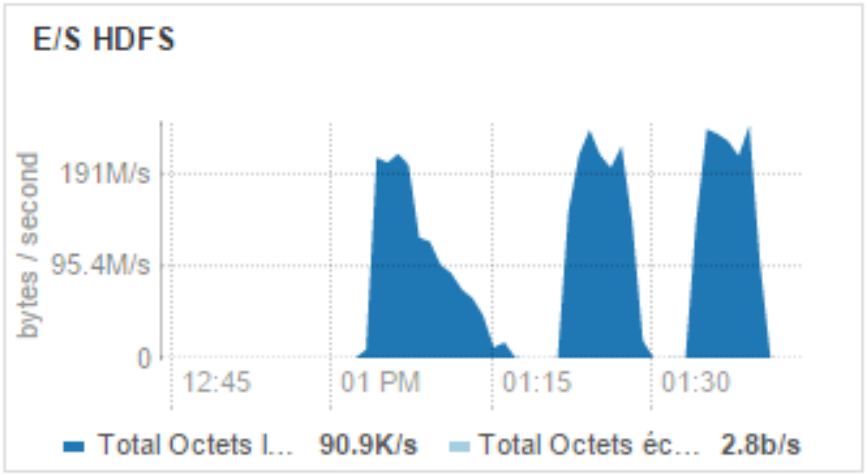
Exemples d'exécution

Données réarrangées : 12min



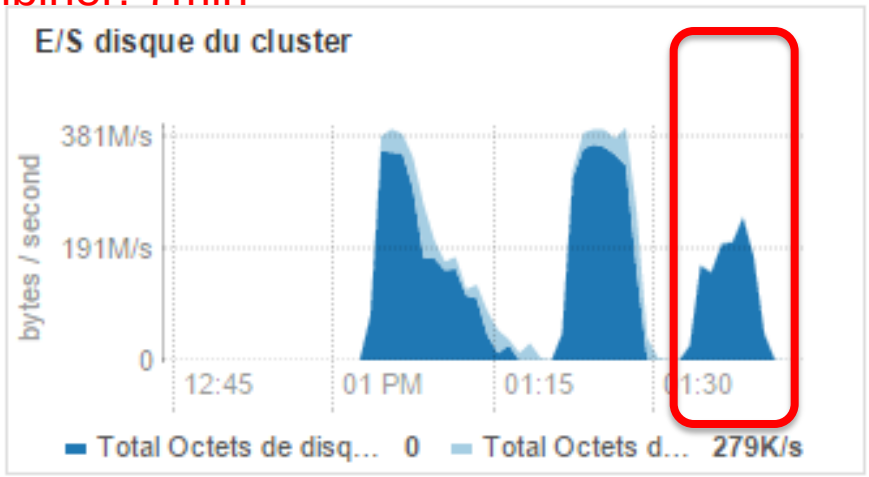


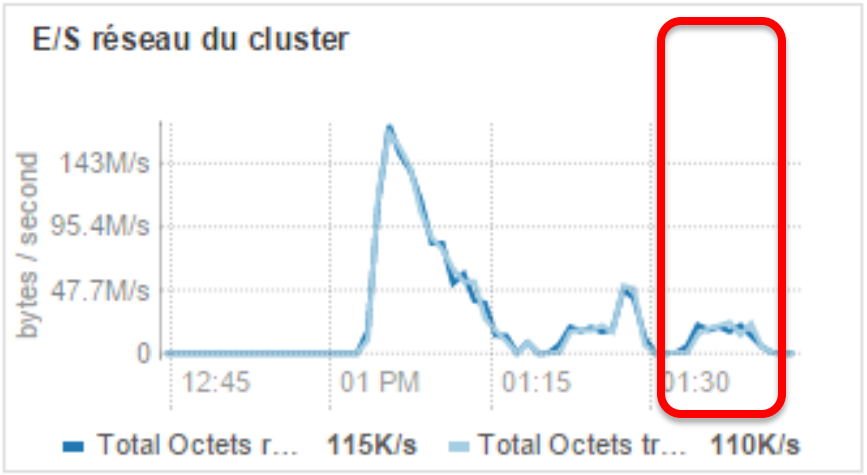


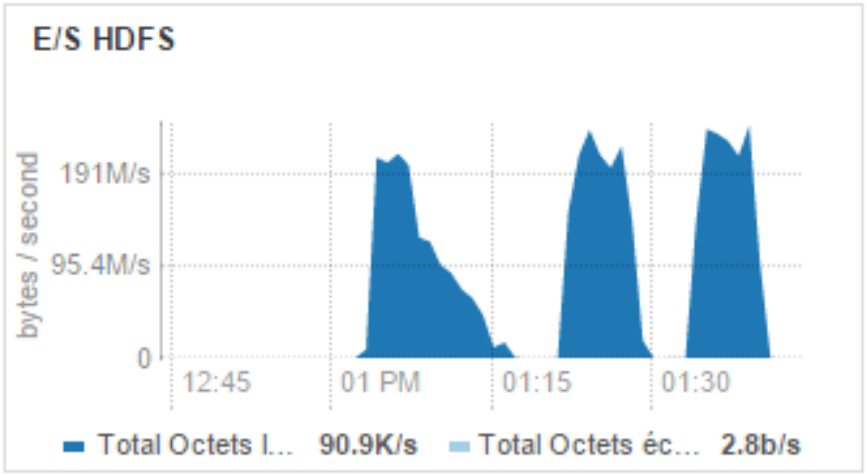


Exemples d'exécution

Données réarrangées + Combiner: 7min







INFRASTRUCTURE DE LA DI U-PSUD

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

université PARIS-SACLAY

Infrastructure

- 11 machines physiques
  - IBM Bladecenter
  - 8 cœurs, 16Go RAM, 150Go disque
- 10 machines virtuelles
- Total :
  - ~200 cœurs
  - ~3To stockage utile

Merci

- Questions ?