

PROJET CASI

BIG DATA AND TWEET STREAMING

Équipe DataTeam
Christophe Cluizel et Thibaud Dauce

PRÉSENTATION DU PROJET



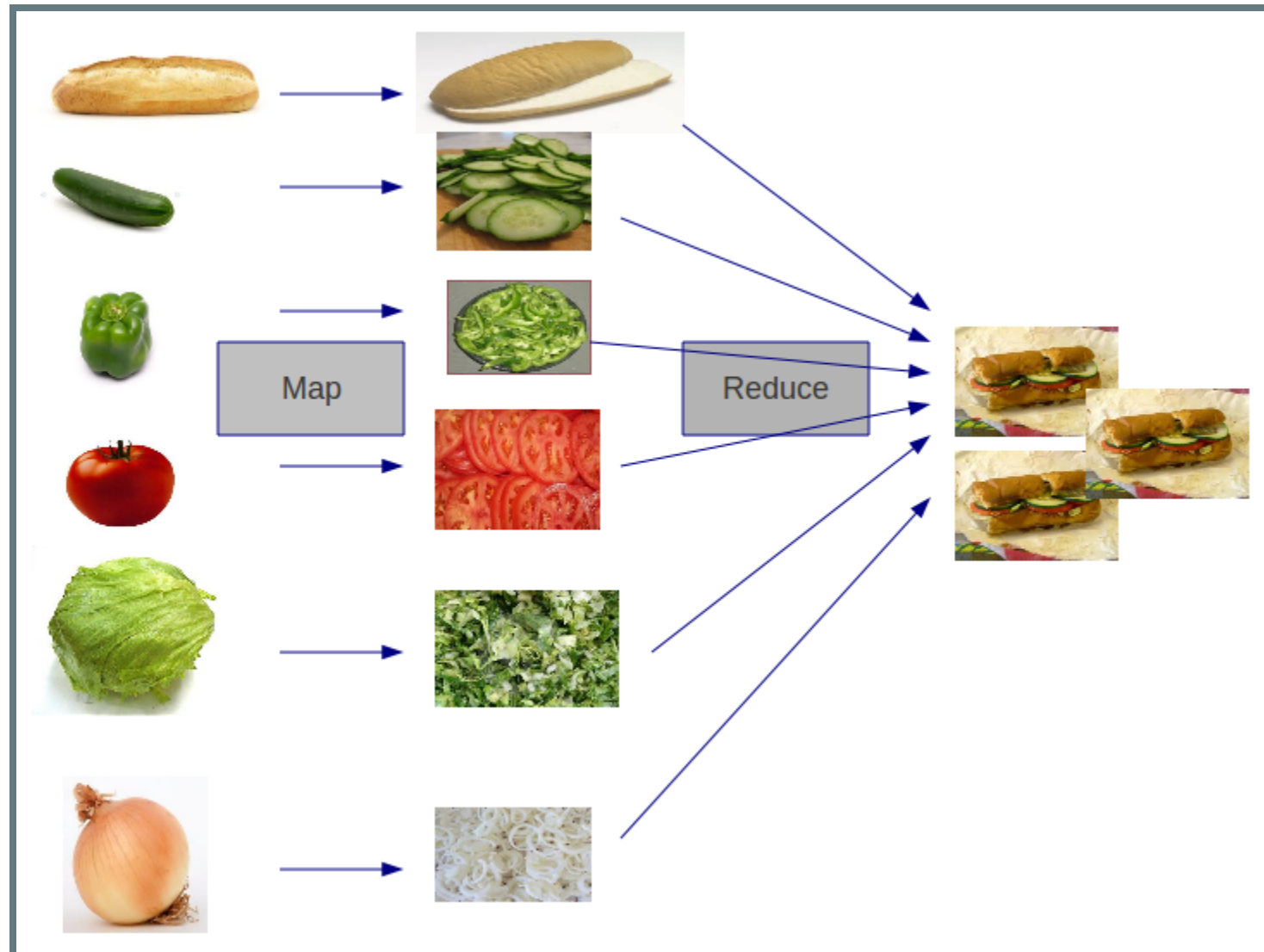
DE NOMBREUX ACTEURS...

- Google
- Facebook
- Twitter
- Amazon

... ET DE NOMBREUX PROJETS OPEN-SOURCE

- Hadoop Distributed File System
- Apache Hadoop
- Apache Spark

TECHNIQUES DE POINTE



MAP REDUCE

FONCTIONNEMENT MASTER / SLAVE

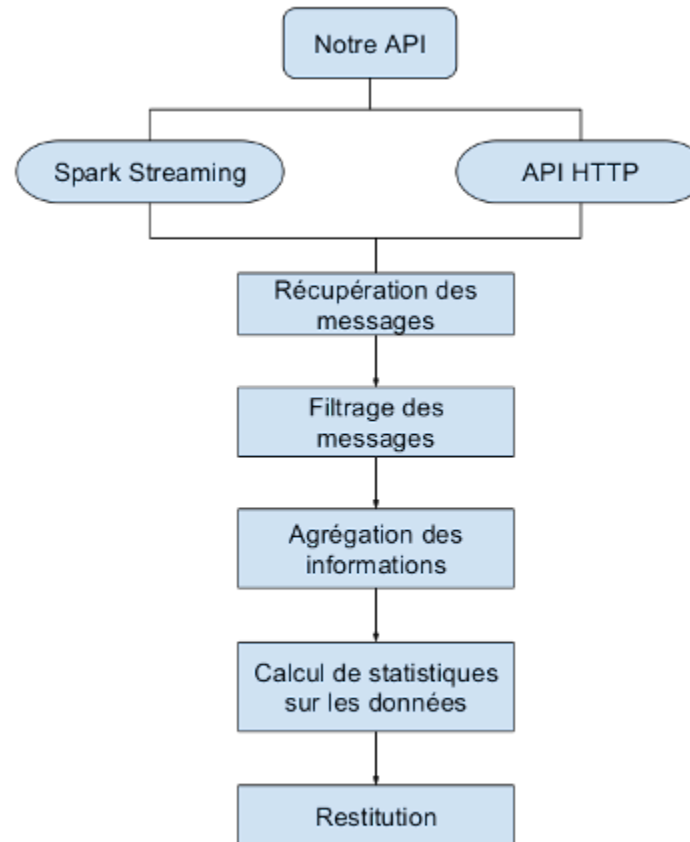
- Envoi des tâches aux *workers*
- Exécution des tâches sur les *workers*
- Récupération de tous les résultats
- Aggrégation des résultats sur le *master*

QUEUES DE MESSAGES

GESTION DES FLUX DE DONNÉES EN TEMPS RÉEL

- Envoi des données dans un *endpoint*
- Répartition des données dans plusieurs queues distribuées
- Traitement des données avec plusieurs *workers*

DEUX SOLUTIONS POSSIBLES



MÉTRIQUES DE COMPARAISON

- Le temps de réponse de l'ensemble de la chaîne
- Le temps de réponse de la tâche
- Vitesse de transmission des missions entre l'API Twitter et le prototype
- Nombre d'erreurs liées à la mémoire
- Pourcentage d'utilisation du CPU
- Nombre d'étape pour installer le prototype
- Nombre de configurations nécessaires pour pouvoir utiliser le prototype
- Nombre de machines utilisables pour effectuer les traitements de la chaîne

STREAMING HTTP (NON IMPLÉMENTÉ)

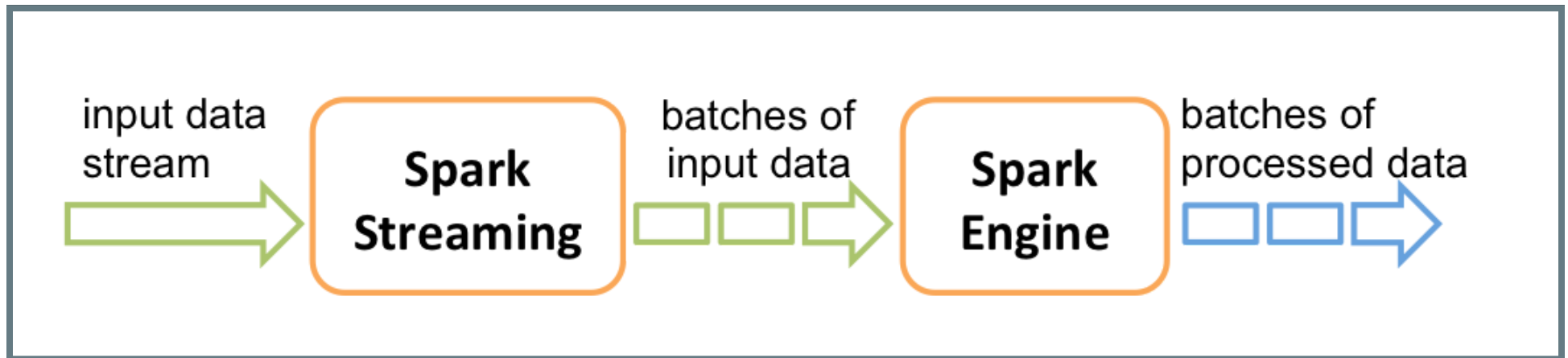
TÉLÉCHARGEMENT D'UN FICHIER INFINI

- Vieilles implémentations existantes
- Traitement des tweets à la chaîne
- Queue de messages à implémenter

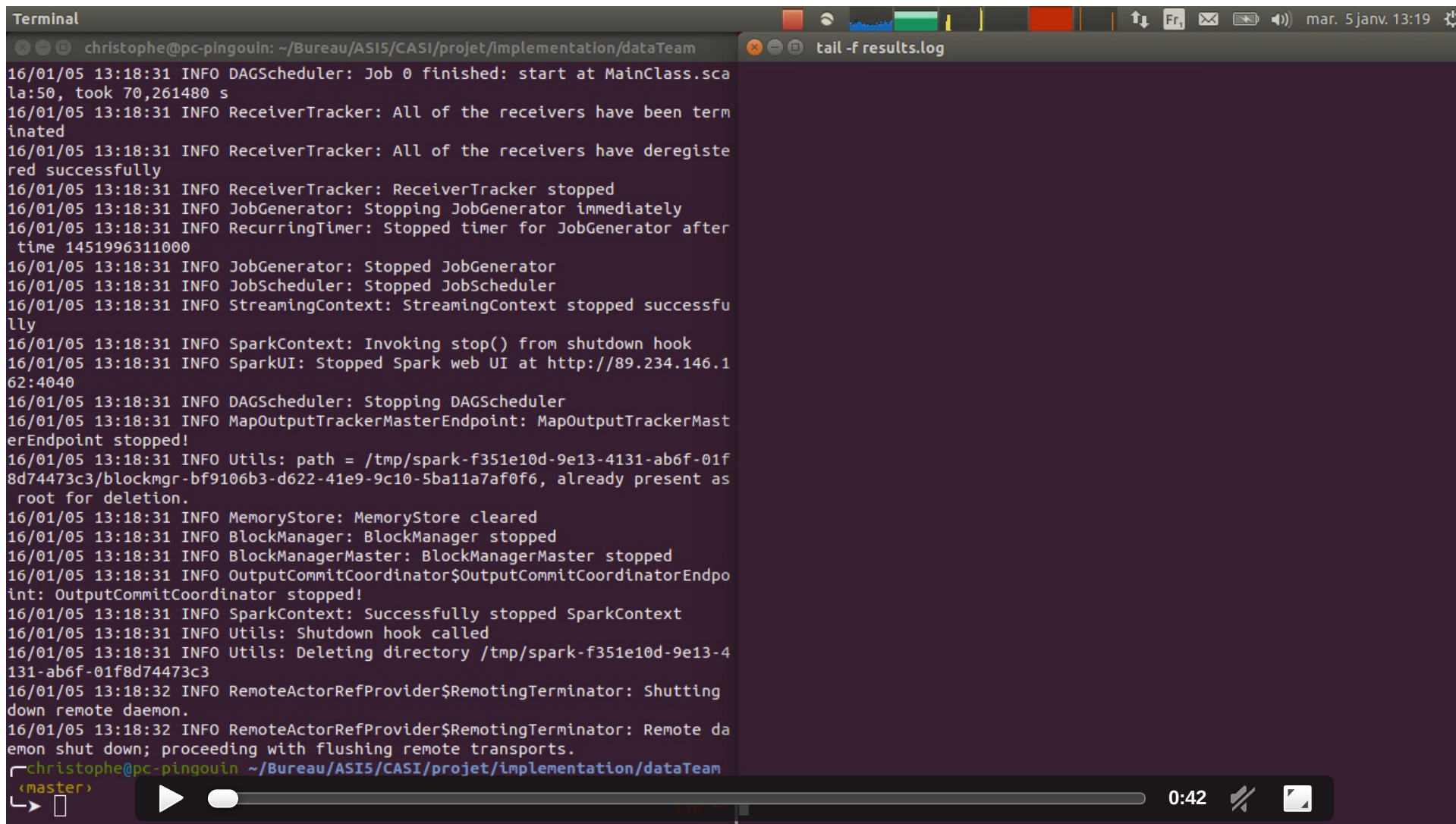
SPARK STREAMING (IMPLÉMENTÉ)

CONSOMMATION DES TWEETS EN DISTRIBUÉ

- Bibliothèque existante
- *Scaling* facile et intégré
- Traitement par lots



DÉMO



The image shows a terminal window with a dark background and light-colored text. The terminal output consists of a series of log messages from a Spark application, including information about job completion, receiver termination, and the shutdown of various components like the DAGScheduler, JobGenerator, and SparkContext. The messages are timestamped and include details about the state of the application. Overlaid on the right side of the terminal is a video player interface. It features a dark purple background with a play button icon on the left, a progress bar in the center, and a timestamp '0:42' on the right. The video player also includes a volume icon and a close button.

```
Terminal
christophe@pc-pingouin: ~/Bureau/ASIS/CASI/projet/implementation/dataTeam
16/01/05 13:18:31 INFO DAGScheduler: Job 0 finished: start at MainClass.sca
la:50, took 70,261480 s
16/01/05 13:18:31 INFO ReceiverTracker: All of the receivers have been term
inated
16/01/05 13:18:31 INFO ReceiverTracker: All of the receivers have deregiste
red successfully
16/01/05 13:18:31 INFO ReceiverTracker: ReceiverTracker stopped
16/01/05 13:18:31 INFO JobGenerator: Stopping JobGenerator immediately
16/01/05 13:18:31 INFO RecurringTimer: Stopped timer for JobGenerator after
time 1451996311000
16/01/05 13:18:31 INFO JobGenerator: Stopped JobGenerator
16/01/05 13:18:31 INFO JobScheduler: Stopped JobScheduler
16/01/05 13:18:31 INFO StreamingContext: StreamingContext stopped successfu
lly
16/01/05 13:18:31 INFO SparkContext: Invoking stop() from shutdown hook
16/01/05 13:18:31 INFO SparkUI: Stopped Spark web UI at http://89.234.146.1
62:4040
16/01/05 13:18:31 INFO DAGScheduler: Stopping DAGScheduler
16/01/05 13:18:31 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMast
erEndpoint stopped!
16/01/05 13:18:31 INFO Utils: path = /tmp/spark-f351e10d-9e13-4131-ab6f-01f
8d74473c3/blockmgr-bf9106b3-d622-41e9-9c10-5ba11a7af0f6, already present as
root for deletion.
16/01/05 13:18:31 INFO MemoryStore: MemoryStore cleared
16/01/05 13:18:31 INFO BlockManager: BlockManager stopped
16/01/05 13:18:31 INFO BlockManagerMaster: BlockManagerMaster stopped
16/01/05 13:18:31 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpo
int: OutputCommitCoordinator stopped!
16/01/05 13:18:31 INFO SparkContext: Successfully stopped SparkContext
16/01/05 13:18:31 INFO Utils: Shutdown hook called
16/01/05 13:18:31 INFO Utils: Deleting directory /tmp/spark-f351e10d-9e13-4
131-ab6f-01f8d74473c3
16/01/05 13:18:32 INFO RemoteActorRefProvider$RemotingTerminator: Shutting
down remote daemon.
16/01/05 13:18:32 INFO RemoteActorRefProvider$RemotingTerminator: Remote da
emon shut down; proceeding with flushing remote transports.
christophe@pc-pingouin ~/Bureau/ASIS/CASI/projet/implementation/dataTeam
(master)
▶ 0:42
```

RÉSULTATS

Lot	Nombre de tweets	Temps de traitement	Temps d'exécution
1	39 tweets	9ms	28,1s
2	27 tweets	21ms	29,2s
3	49 tweets	19ms	30,2s

Traitement de 30,59 tweets par seconde, 10% d'utilisation du CPU (2 x 2.00GHz), aucune tâche échouée.

CONCLUSION

POSEZ VOS QUESTIONS !