

Thème 7 -- Batch

Emma GRATZMULLER, Emmanuel MENEGHETTI,
Audrey POISSON, Hugo TENG



Plan du cours

- Introduction
- Traitement de données en Batch
 - Avantages
 - Limites
- Plateformes, infrastructures et langages de programmation
 - Hadoop
 - Spark
 - Spring et autres
- Cas d'usage et exemples

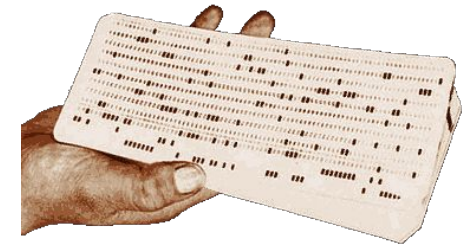
Définition



Le **Batch processing** ou **traitement de données par lots** est un mode de traitement de données dans lequel les programmes à exécuter ou les données à traiter sont groupés en lots. Plus exactement, le batch consiste en l'exécution de **travaux répétitifs** contenant un **volume important de données** avec peu ou pas d'intervention des utilisateurs.

Le traitement par lots a commencé avec des cartes perforées, qui ont été compilées en instructions pour les ordinateurs. Des jeux entiers, ou des lots, de cartes étaient traités en une seule fois. Ce système, créé par Herman Hollerith, **remonte à 1890**.

Aujourd'hui la plupart des fonctions de traitement par lots sont activées sans interaction, le programme est **autonome** et s'exécute seul, même en l'absence de l'utilisateur, c'est pourquoi on les utilise souvent pour exécuter de nuit des travaux sur le système.



Exemples d'utilisation

On peut s'en servir par exemple pour :

- Télécharger des jeux ou logiciels
- Renommer en masse des fichiers ou dossiers informatiques
 - Remplacer une partie du nom ou l'effacer.
 - Ajouter une séquence numérique ou alphabétique (001, 002, 003... a, b, c, d...).
 - Ajouter la date et l'heure (actuelle ou prise du fichier ou dossier).
- Automatiser des processus de facturation
- Envoie de mail commerciaux

Fichier Batch

Un fichier batch est un fichier texte non formaté qui contient une ou plusieurs commandes et porte l'extension de nom de fichier **.bat** ou **.cmd**.

Exemple de fichier Batch

exemple de traitement par lots de la ligne de commande Windows

```
cd C:\Répertoire
```

```
FOR %%f IN (*.doc *.txt) DO XCOPY C:\Répertoire\"%f" C:\Back-up-Répertoire\Textes /d /y
```

```
FOR %%f IN (*.jpg *.png *.bmp) DO XCOPY C:\Répertoire\"%f" C:\Back-up-Répertoire\Images /d /y
```

- Effectue une copie des éléments du Répertoire vers un dossier Back-up du Répertoire
- Répartit les fichiers dans les sous-dossiers selon leur extension
- /d = seulement les fichiers actualisés
- /y = confirmation automatique d'écrasement du fichier cible existant, inhibition permanente de l'invite de confirmation

created a batch file in notepad

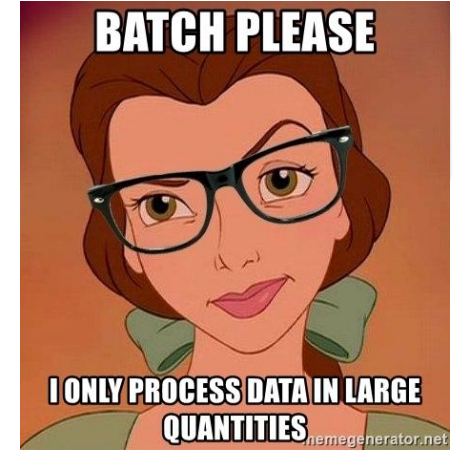


Démonstration



- **Efficacité :**
 - Traitement des ressources **dès que disponibles**
 - Possibilité de donner des **priorités**
 - Possibilité de tourner **hors ligne**
- **Fonctionnalités hors ligne :**
 - Permet le fonctionnement **hors des périodes d'activité** et minimise l'utilisation des processeurs
- **Simplicité :**
 - **Moins complexe** qu'un système en stream (voir prochain cours),
 - Nécessite **moins de maintenance**
 - Pas de prise en charge système/matérielle spéciale
- **Processus non-interventionniste :**
 - **Travail automatique** qui permet aux opérateurs de travailler sur autre chose.
 - Des **alertes** sont lancées en cas de problème.

- **Données de meilleure qualité :**
 - Automatisation
 - Processus non-interventionniste
 - Moins d'erreur
 - Meilleure précision
 - Exactitude accrue
- **Rapidité - Business Intelligence accélérée :**
 - Permet de traiter rapidement de **gros volumes**
 - Traite **simultanément** de nombreux enregistrements
 - Fournit des données qui permettent de **prendre rapidement des mesures**
- **Moins cher :**
 - Coûts opérationnels (main-d'œuvre, équipement...) réduits.





Inconvénients

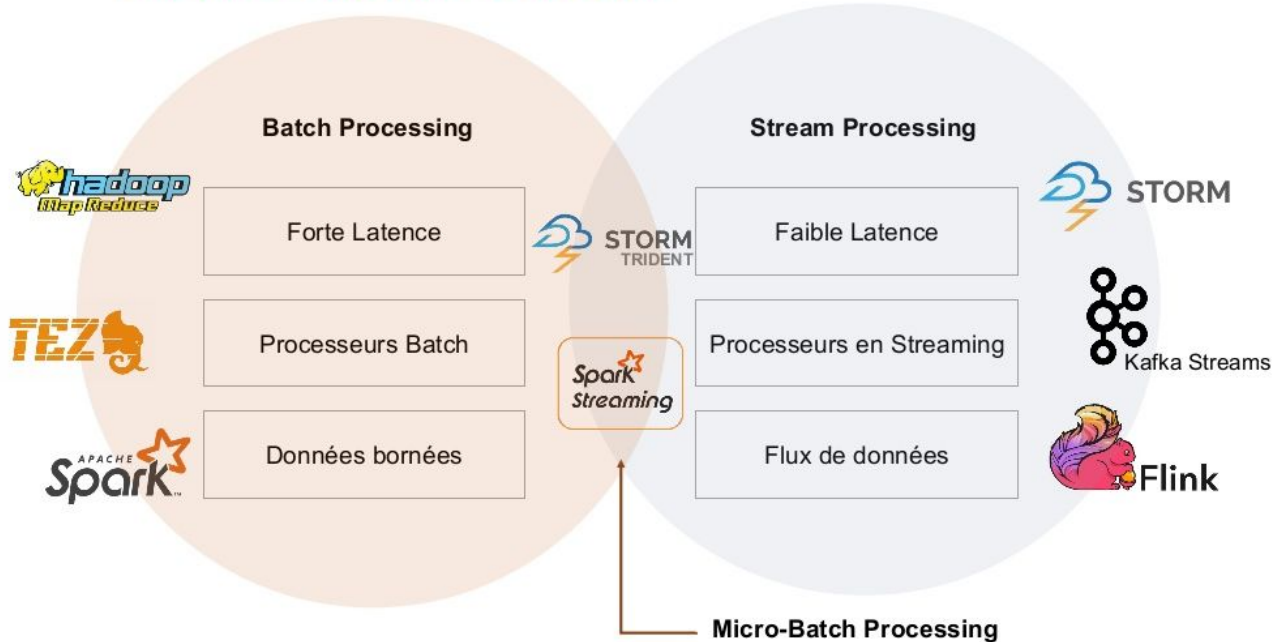
- **Déploiement et formation :**
 - Gérer les systèmes en batch nécessite de savoir comment fonctionnent les lots,
 - Savoir comment les **programmer**,
 - Savoir comment **gérer les exceptions et les problèmes**.
- **Débogage complexe :**
 - Si une erreur apparaît, il faudra généralement faire appel à un spécialiste car déboguer des systèmes batch est souvent complexe.
- **Coût des infrastructures batch :**
 - Les infrastructures pour les systèmes batch peuvent être élevés
 - Exemple : pour le traitement par Microsoft Azure, les tarifs vont de 30 € à 5000 € par mois
- **Forte latence**

Résumé

| Avantages | Inconvénients |
|--|---|
| <ul style="list-style-type: none">• Efficacité• Fonctionnalité hors ligne• Simplicité• Processus non-interventionniste• Données de meilleure qualité• Rapidité - Business Intelligence accélérée• Moins cher | <ul style="list-style-type: none">• Déploiement et formation• Débogage complexe• Coût des infrastructures batch• Forte latence |



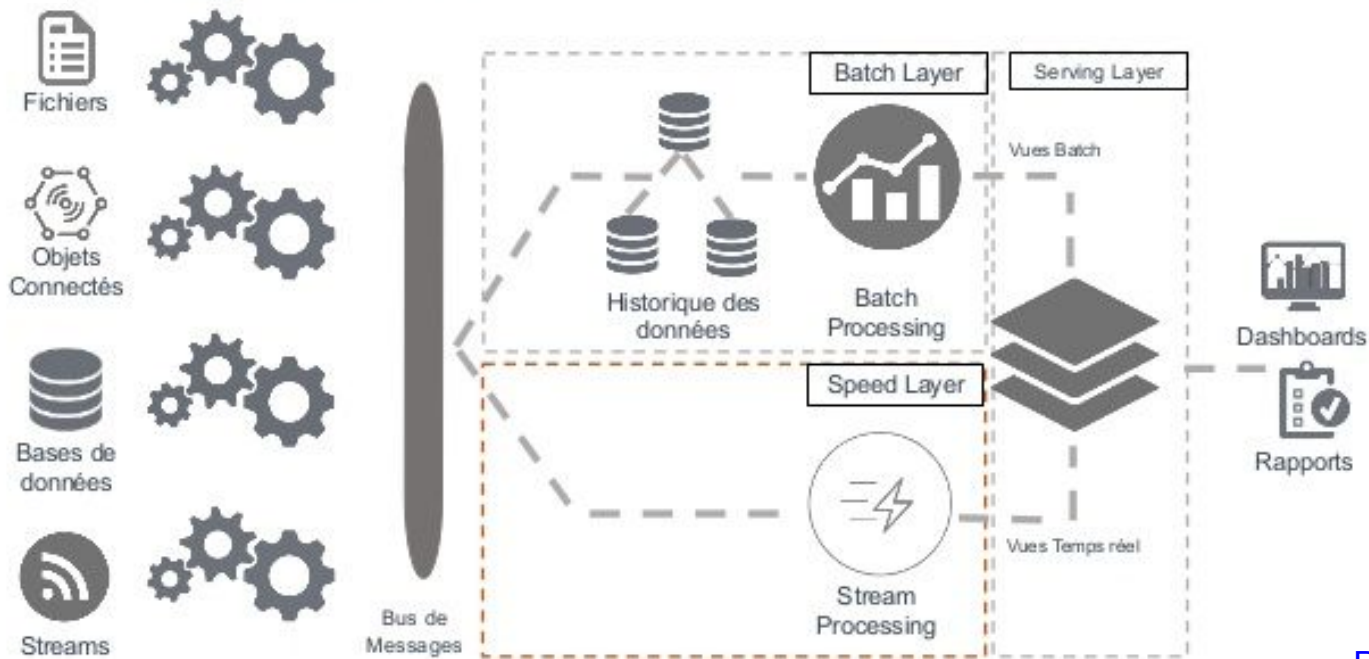
BATCH VS STREAM PROCESSING



Vue “historique” :

- Tâche de fond
- Données historiques
- Consolidation
- Précision
- Grosses quantités de données

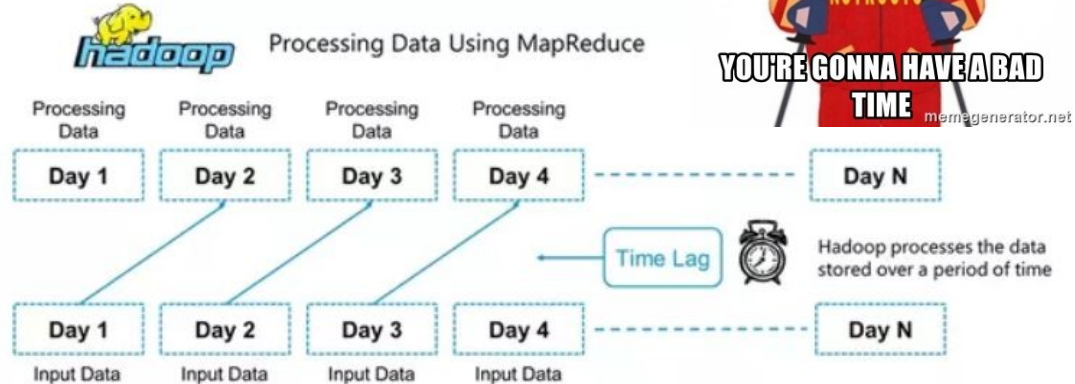
ARCHITECTURE LAMBDA



[Plus d'infos ici](#)



- Framework libre open source
- Java
- Petas octets de données
- 2009, Apache
- Time lag important (jour / mois)
- Très utilisé pour le Big Data



Hadoop permet de traiter de manière distributive et résiliente une très grande quantité de données non structurées assez rapidement, de lancer des applications sur des grappes de machines standards.

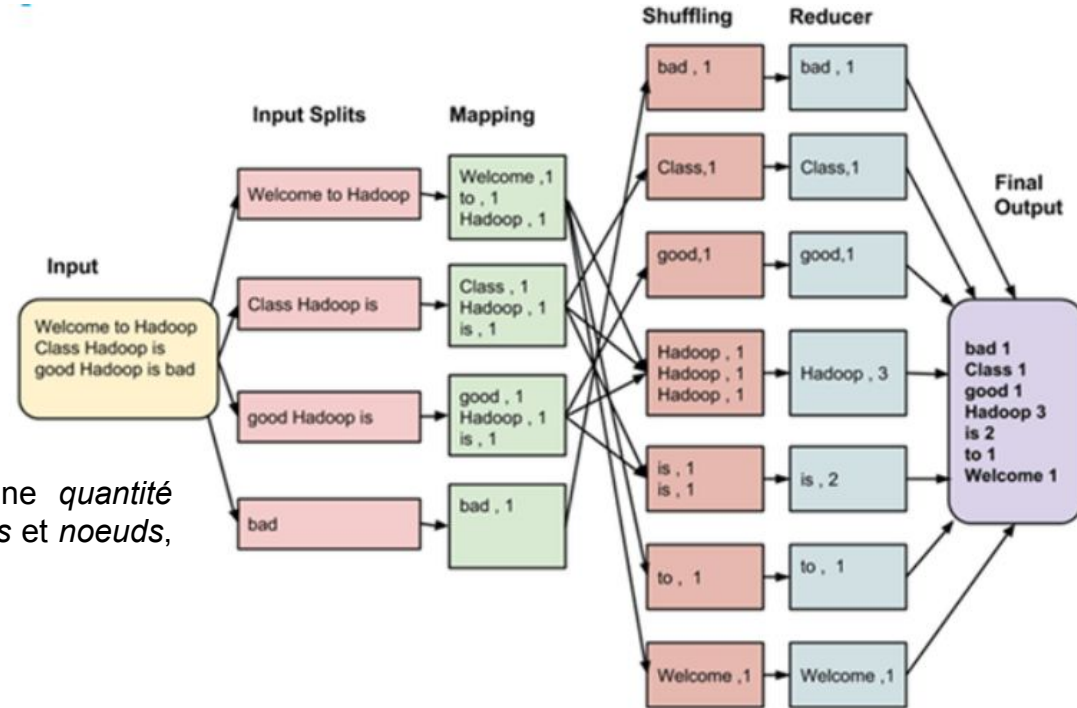
Lien [doc](#)

[CquoiHadoop](#)

Lien [image](#)

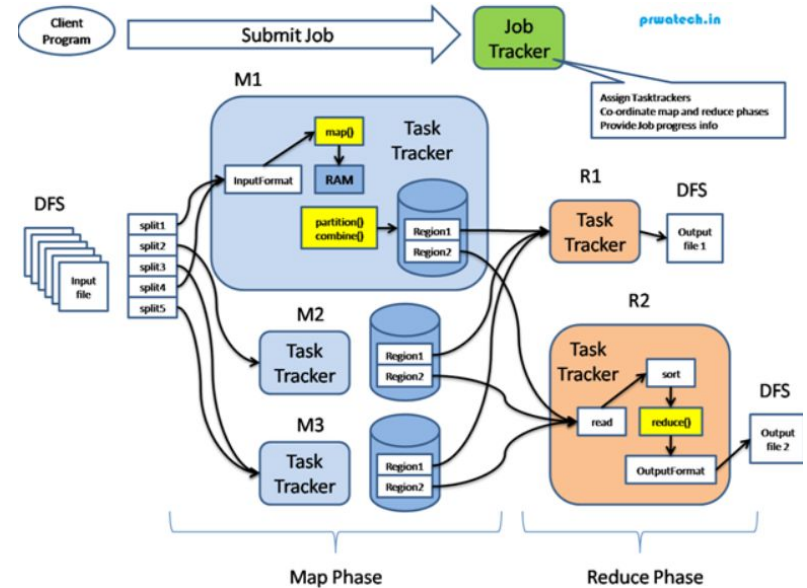
MapReduce permet de filtrer et morceler le travail entre les nodes (fonction “mapper”), puis à les organiser en une réponse (fonction “reducer”).

Hadoop MapReduce permet donc de gérer une *quantité énorme de données* à travers de nombreux *serveurs* et *noeuds*, et de les *combiner* pour donner notre réponse.




Sûr : Si un noeud **tombe**, la tâche est donné à un **autre**.
Copies des données stockées automatiquement.

MapReduce ne convient **PAS** aux tâches **analytiques** **itérative** et **interactive**



lien [image](#)

APACHE Streaming

- Langages : Java, Python, Scala ,.NET et R
- Spark fonctionne par **Micro Batch**, qui sont des batches rapides (minute)
- Streaming 



Machine Learning, queries interactive...

Lien [image](#)



Micro Batch

Lien [doc](#)

lien [image](#)

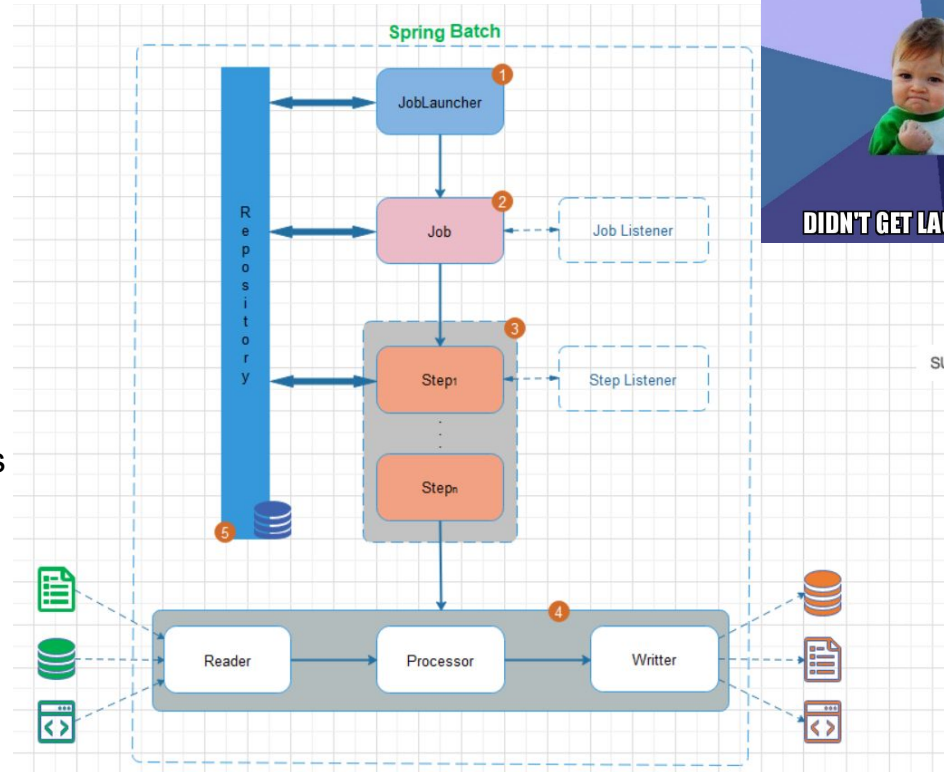
| | Hadoop | Spark |
|------------------|---|--|
| Catégorie | Basic Data processing engine | Data analytics engine |
| Usage | pour de grandes quantité de données | Micro Batch, pour des traitement de petits paquets rapidement, presque en temps réel |
| Latence | Très grande latence (heures / jours) | Faible latence (minutes) |
| Data | Batch Processing mode, MapReduce | De façon itérative |
| Facilité d'usage | Complexe , gestion d'API de bas niveau | Plus facile , permet l'usage des langages de haut niveau |
| Scheduler | Un job scheduler extérieur est nécessaire | Calculs en mémoire interne, job scheduler non nécessaire |
| Sécurité | Très sûr | Moins sûr |
| Cout | Moins, grâce au MapReduce | Plus chère, à cause des calculs en mémoire interne |



- XML, Java
- *light*
- Traitement en chunk / Tasklet
- **micro batch**
- Composants et fonctionnalités réutilisables
- **Moins de code**
- Des librairies à foison
- Plus populaire que Spark

Lien [doc](#)

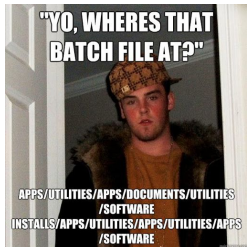
Lien [tuto](#)



Lien [image](#)



Azure Batch



[Meme](#)



IBM
workload
automation



Batch Patch



AWS Batch

- **Payroll :**
 - Système de paie en entreprise
 - Par semaine, par mois... → cycle
 - Batch processing très souvent utilisé
- **Pourquoi? :**
 - + rapide
 - Pas de nécessité de hardware supplémentaire
 - Fonctionnement hors périodes d'activité
 - Coûts plus faibles
 - Possibilité de process simultanés pour des cycles différents



- **Processus général :**

- 1 - Data

2 fichiers:

- Master File (classé selon primary key Employee No) : toutes les informations par employé (N°, noms, taux horaire, département, somme déjà versée cette année)
- Transaction File (non classé initialement): informations de la dernière "période" (N°, type de transaction, valeur)
- Transactions → heures travaillées, ajout d'un nouvel employé si recrutement, suppression d'un employé si démission...

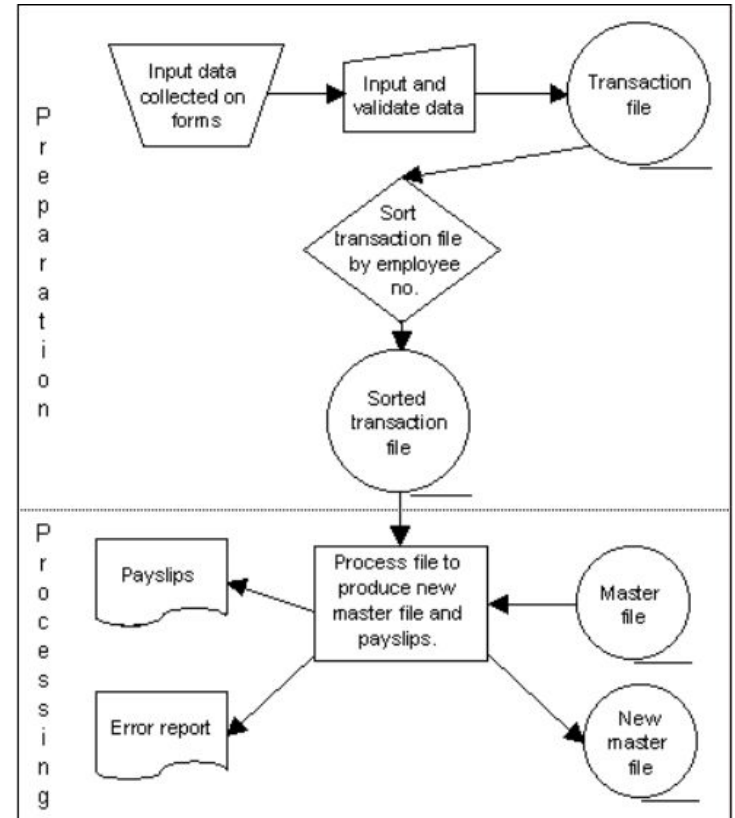
| Record In Master File | | Record In Transaction File | |
|-----------------------|-----------|----------------------------|--------------|
| Employee No. | 12A004 | Employee No. | 12A004 |
| Surname | Jones | Transaction Type | Hours Worked |
| Forenames | Anna Jane | Value | 25 |
| Pay Rate | £7.50 | | |
| Department | Accounts | | |
| Pay This Year | £5575.00 | | |

Lien [image](#)

○ 2 - Processing

- Classer le transaction file selon le n° employé
- Process
- ★ Calcul de la somme à payer à l'aide du nb d'heure (T file) et du taux horaire (M file) (d'où l'importance de les classer)
- ★ Nouveau masterfile créé avec la somme annuelle mise à jour
- ★ L'ancien masterfile devient un back-up
- ★ Possibilité d'impression automatique de fiche de paie
- ★ Rapport d'erreur

Nombreuses plateformes proposent des solutions de payroll par traitement batch



Lien [image](#)



Quand l'utiliser ?

- Pour du traitement **régulier**.
 - Le batch processing est une technique qui permet d'automatiser et de traiter plusieurs transactions comme si elles n'étaient qu'un seul groupe Traitement régulier
- Pour des **consolidations** ponctuelles.
 - Le batch processing permet de traiter des tâches telles que la paie, la réconciliation de fin de mois ou le règlement de transactions pendant la nuit.
- Pour des **grosses quantités de données**, qui serait plus coûteuses à traiter en Stream qu'en Batch

Attention aux limites !



- Il est nécessaire de pouvoir assurer la **maintenance** des systèmes de Batch processing et de pouvoir les **déboguer**
- Ces systèmes peuvent permettre **d'économiser** de l'argent et de la main-d'œuvre au **fil du temps**, mais ils peuvent être **coûteux** à concevoir et à mettre en œuvre dès le **départ**.
- Il faut garder en tête la forte **latence** ! Ce n'est pas fait pour pouvoir traiter rapidement des données, mais plutôt pour des opérations régulières, que l'on pourrait laisser tourner la nuit par exemple (télécharger WoW par exemple)

<https://www.talend.com/fr/resources/batch-processing/>

<https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203563-batch-definition-traduction/>

<http://www.marche-public.fr/Terminologie/Entrees/traitement-par-lots.htm>

https://fr.wikipedia.org/wiki/Traitement_par_lots

<https://www.ionos.fr/digitalguide/serveur/outils/creer-un-fichier-batch/>

<https://fr.wiktionary.org/wiki/batch>

<https://www.investopedia.com/terms/b/batch-processing.asp#:~:text=Batch%20processing%20started%20with%20punch,data%20from%20the%20U.S.%20Census>

<https://www.bmc.com/blogs/what-is-batch-processing-batch-processing-explained/>

[https://docs.microsoft.com/en-us/previous-versions/windows/it-pro/windows-xp/bb490869\(v=technet.10\)?redirectedfrom=MSDN](https://docs.microsoft.com/en-us/previous-versions/windows/it-pro/windows-xp/bb490869(v=technet.10)?redirectedfrom=MSDN)

<https://azure.microsoft.com/fr-fr/pricing/details/batch/windows-virtual-machines/>

<http://ictsmart.tripod.com/ict4/print/partprpy.htm>

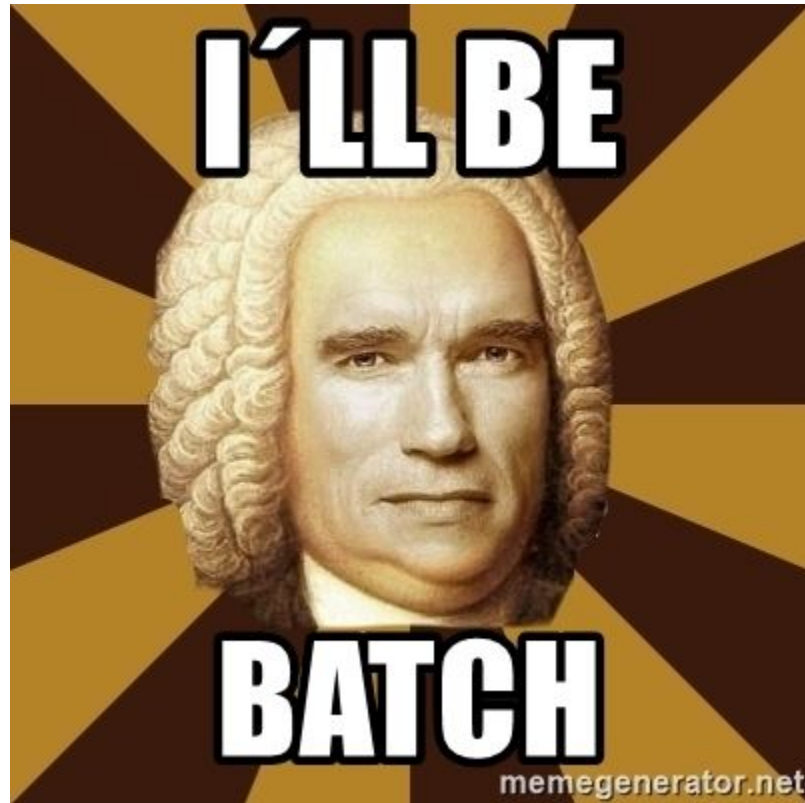
<https://blog.k2datascience.com/batch-processing-apache-spark-a67016008167>

<https://www.educba.com/hadoop-vs-spark/>

<https://gkemayo.developpez.com/tutoriels/java/tutoriel-sur-mise-oeuvre-spring-batch-avec-spring-boot/>

<https://www.slideshare.net/mehdibenissa/drpf-for-big-data-stream-processing-architectures>

<https://slides.com/pcourbin/esilv-a5-iot-cloud-1#/5/6/2>



Qu'avez-vous retenu de ce cours sur le Batch ?

