Data Development

III — Big Data, Calcul distribué : les projets Hadoop et Spark

Patrick S. Kanmeugne
SupInfo

Plan(1) 

Le contexte du Big Data ?

ce qui change par rapport à l'analyse de données classique

Big Data ~ La nature et la taille des données à analyser obligent à adapter les ressources (calcul et stockage) et à repenser les méthodes !

Les quatre Vs

Volume, Variété, Vitesse, Véracité

La *complexité* de l'analyse de données évolue simultanément selon :

- Volume ⟶ une grande quantité de données
- Variété (diversité) ⟶ plusieurs types de données
- Vitesse ⟶ besoin d'interactivité dans l'analyse
- Véracité ⟶ crédibilité des sources, qualité des données

Plan(2)

Le projet Hadoop

un projet de la fondation Apache

Hadoop Framework

- Framework Open-Source de «*Big Data Processing*» distribué
- Calculs distribués sur des clusters de plusieurs ordinateurs
- Système de fichiers adapté pour stockage distribué
- Programmes écrits suivant un modèle simple

Map/Reduce

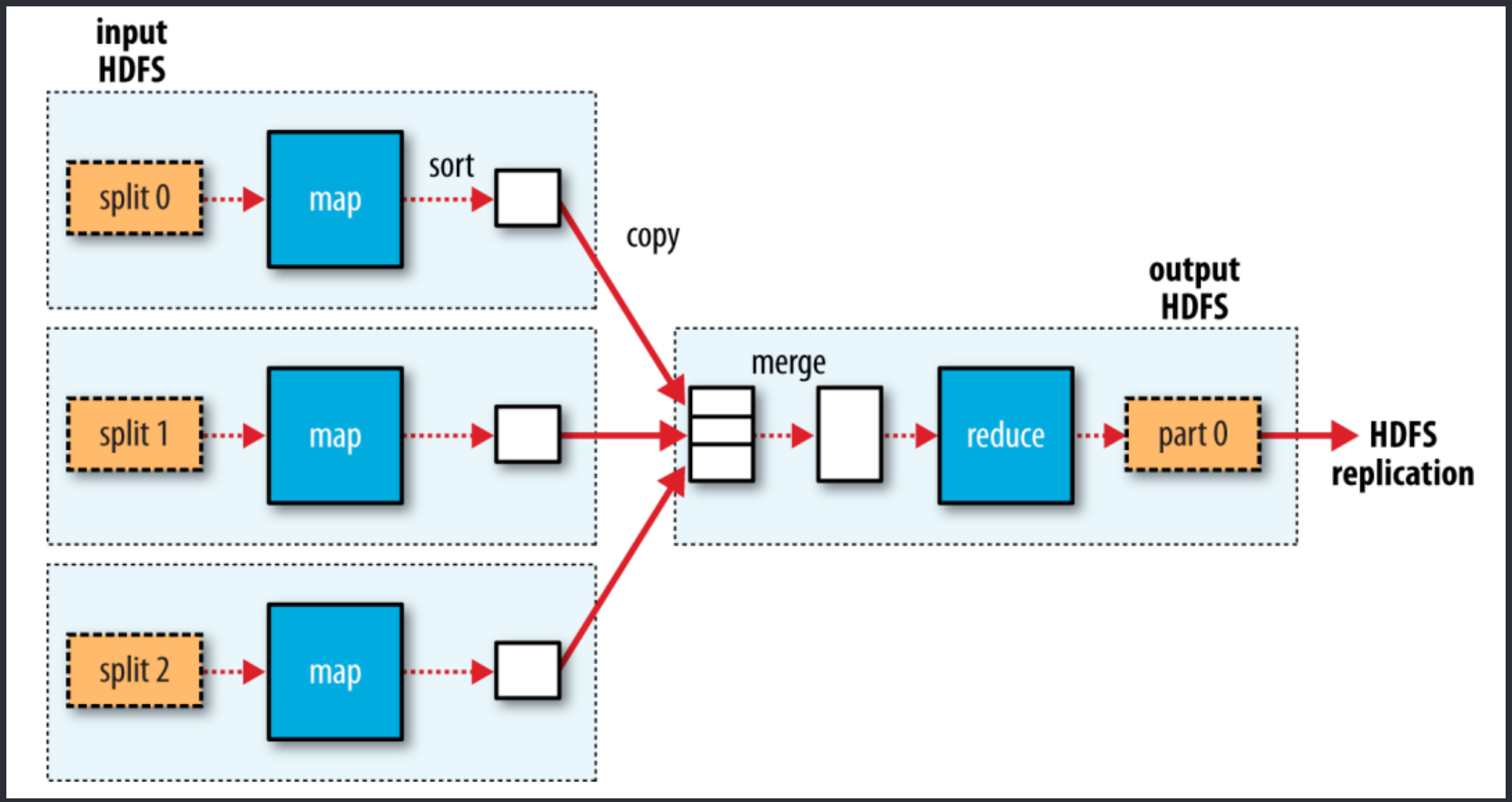
Un modèle de programme simple pour le framework Hadoop

Map/Reduce

- Modèle de programmation pour Hadoop
- 2 principales fonctions : *Map* et *Reduce*

Map/Reduce

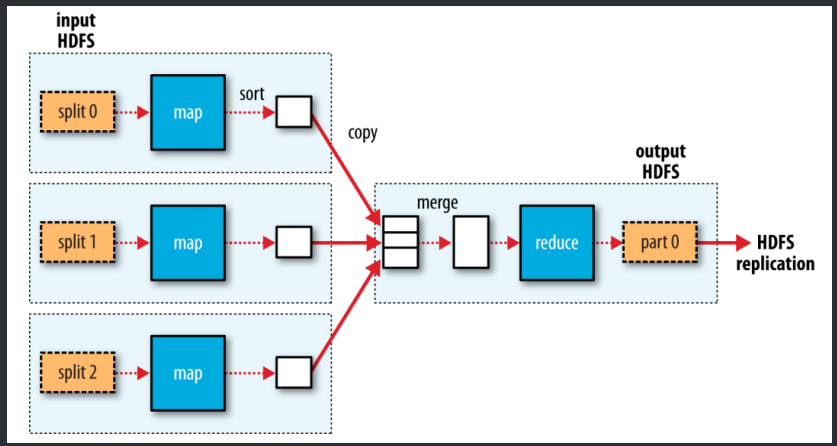
Architecture Map/Reduce [1] : avec une fonction «Map» et une focntion «Reduce»



Map/Reduce

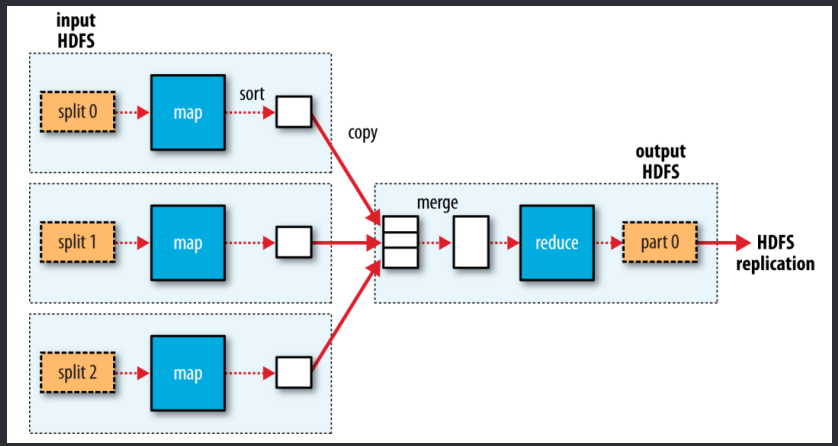
Map phase

- 1^{ere} phase du MapReduce
- Données d'entrée transformées en pairs : *< key, value >*



Map/Reduce

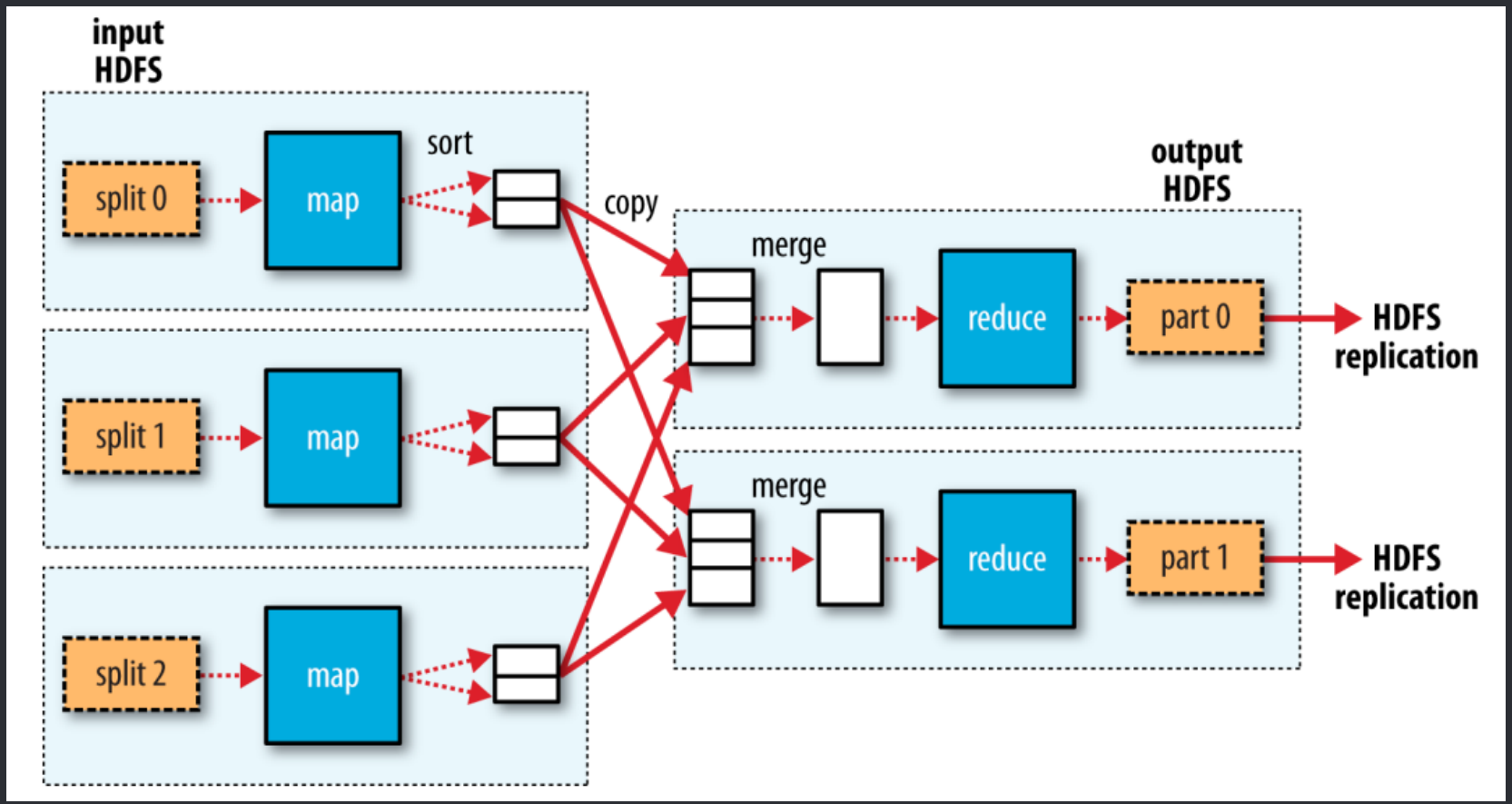
Reduce phase



- 2^{eme} phase du MapReduce
- les données < *key*, *value* > sont agrégées
- Production du résultat final

Map/Reduce

Architecture Map/Reduce [1] : avec ++ fonctions «Reduce»



Plan(3) 

Analyse de données et Informatique

L'informatique fournit les outils…

- Programmation : SQL [1], R, Python…
- Administration de base de données
- Réseaux informatiques
- HPC [2], Cloud

L'analyse de données et le machine learning

Machine Learning

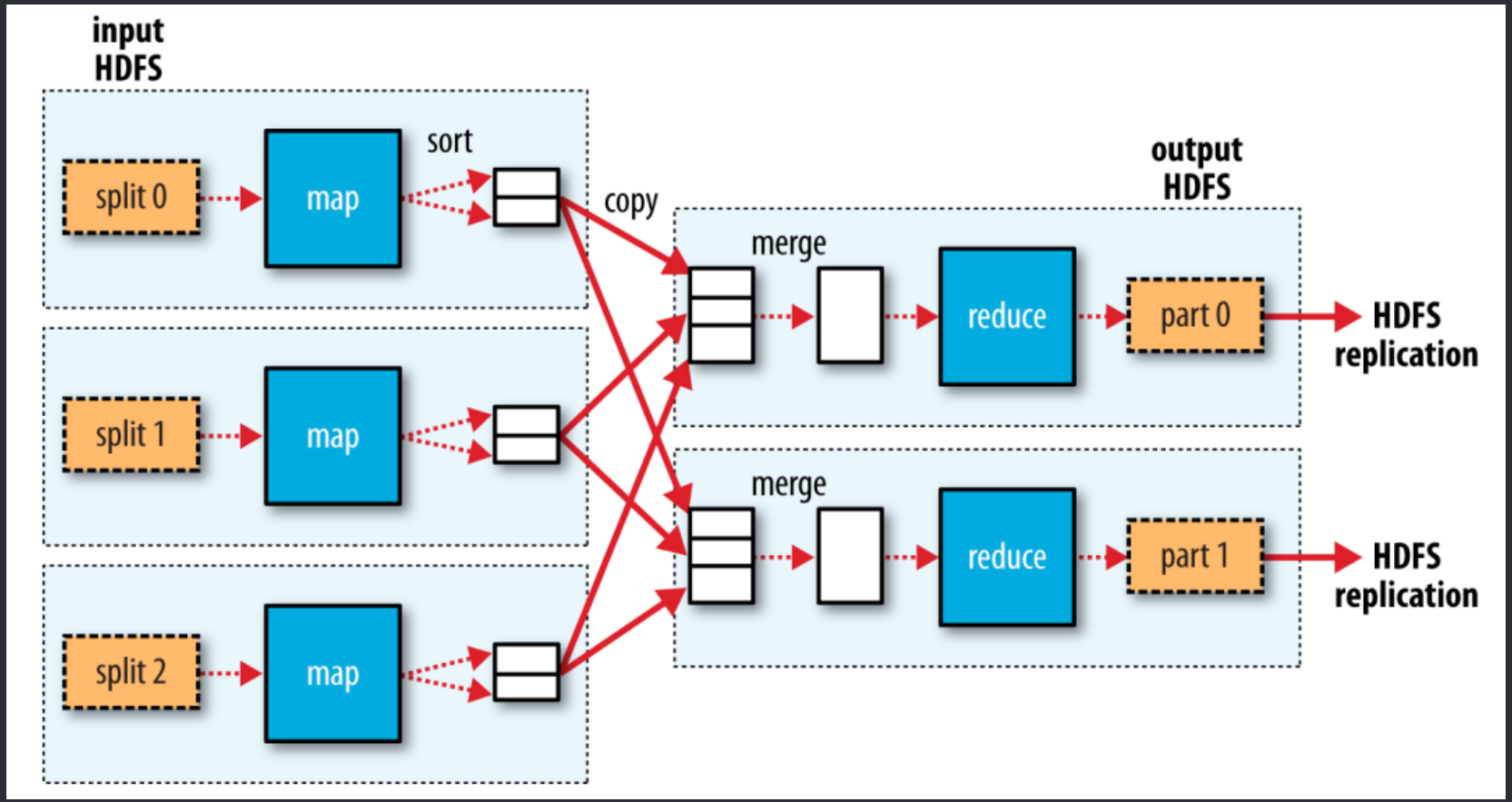
Le ML est un domaine d'études qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés.
Arthur Samuel (1959)

3 sous-domaines de l'apprentissage :

- *de catégories* à partir de données étiquetées (supervisé)
- *de similarités* entre les données (non supervisé)
- *de politique d'actions* (renforcement)

Analyse de données et Intelligence Artificielle

L'analyse de données mène naturellement vers l'IA



Plan(4)

Bibliographie

[1] Tom White. *Hadoop The Definitive Guide, 4th Edition*. O'Reilly.

[2] Bill Chambers and Matei Zaharia. *Spark : The Definitive Guide*. O'Reilly.