Technologie cloud et systèmes distribués Plateformes Big Data

Mohamed-Lamine MESSAI

ICOM - Université Lumière Lyon 2

M2 BI&BD, 2021-2022

Sommaire

Sommaire

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.
- Puis partager ce travail avec les autres.

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.
- Puis partager ce travail avec les autres.
- Seule une poignée de sociétés avait le monopole de l'information sur internet.

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.
- Puis partager ce travail avec les autres.
- Seule une poignée de sociétés avait le monopole de l'information sur internet.
- Les données étaient statiques, fiables et peu nombreuses.

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.
- Puis partager ce travail avec les autres.
- Seule une poignée de sociétés avait le monopole de l'information sur internet.
- Les données étaient statiques, fiables et peu nombreuses.

Et puis un jour ....

- Les réseaux sociaux sont apparus.

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.
- Puis partager ce travail avec les autres.
- Seule une poignée de sociétés avait le monopole de l'information sur internet.
- Les données étaient statiques, fiables et peu nombreuses.

Et puis un jour ....

- Les réseaux sociaux sont apparus.
- D'autres services de partage ont vu le jour.

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.
- Puis partager ce travail avec les autres.
- Seule une poignée de sociétés avait le monopole de l'information sur internet.
- Les données étaient statiques, fiables et peu nombreuses.

Et puis un jour ....

- Les réseaux sociaux sont apparus.
- D'autres services de partage ont vu le jour.
- Les mobiles intelligents sont arrivés.

Le commencement

- Au départ l'homme voulait un ordinateur pour faire le travail à sa place.
- Puis partager ce travail avec les autres.
- Seule une poignée de sociétés avait le monopole de l'information sur internet.
- Les données étaient statiques, fiables et peu nombreuses.

Et puis un jour ....

- Les réseaux sociaux sont apparus.
- D'autres services de partage ont vu le jour.
- Les mobiles intelligents sont arrivés.
- Les objets connectés se sont démocratisé.

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.
- La majorité des données générées sont non structurées.

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.
- La majorité des données générées sont non structurées.
- Les sources :

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.
- La majorité des données générées sont non structurées.
- Les sources :
  - Les média sociaux

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.
- La majorité des données générées sont non structurées.
- Les sources :
  - Les média sociaux
  - Images et vidéos publiées sur internet

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.
- La majorité des données générées sont non structurées.
- Les sources :
  - Les média sociaux
  - Images et vidéos publiées sur internet
  - Transactions d'achats en ligne

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.
- La majorité des données générées sont non structurées.
- Les sources :
  - Les média sociaux
  - Images et vidéos publiées sur internet
  - Transactions d'achats en ligne
  - Signaux GPS de téléphones mobiles

Les faits

- Chaque jour nous produisons une quantité phénoménale de données.
- La majorité des données générées sont non structurées.
- Les sources :
  - Les média sociaux
  - Images et vidéos publiées sur internet
  - Transactions d'achats en ligne
  - Signaux GPS de téléphones mobiles

Ces données sont appelées **Big Data** ou **Données Massives**

Quelques chiffres

- Google traite 20 PO par jour (2008). Données stockées (2014) 15000 PO

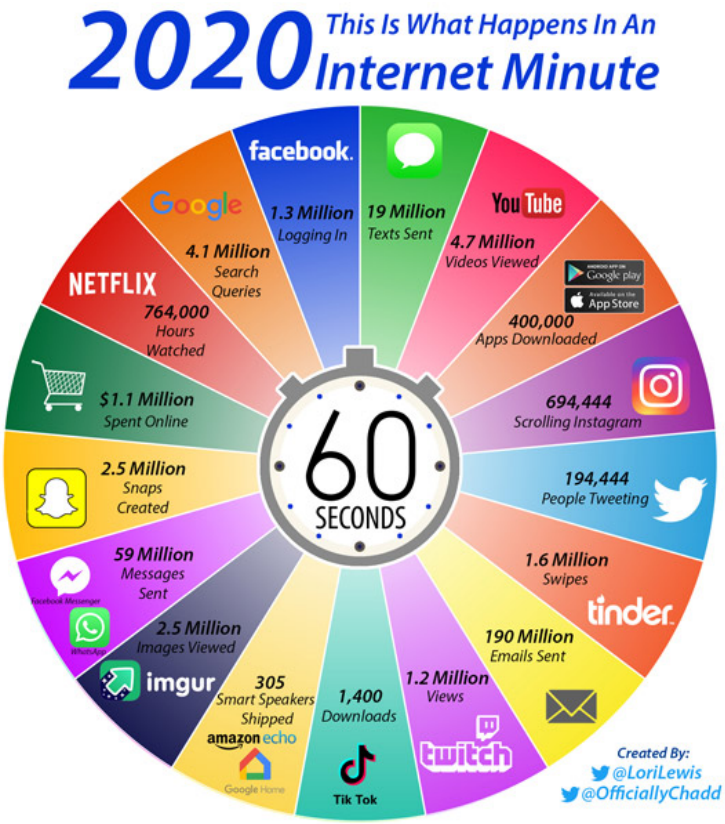
Quelques chiffres

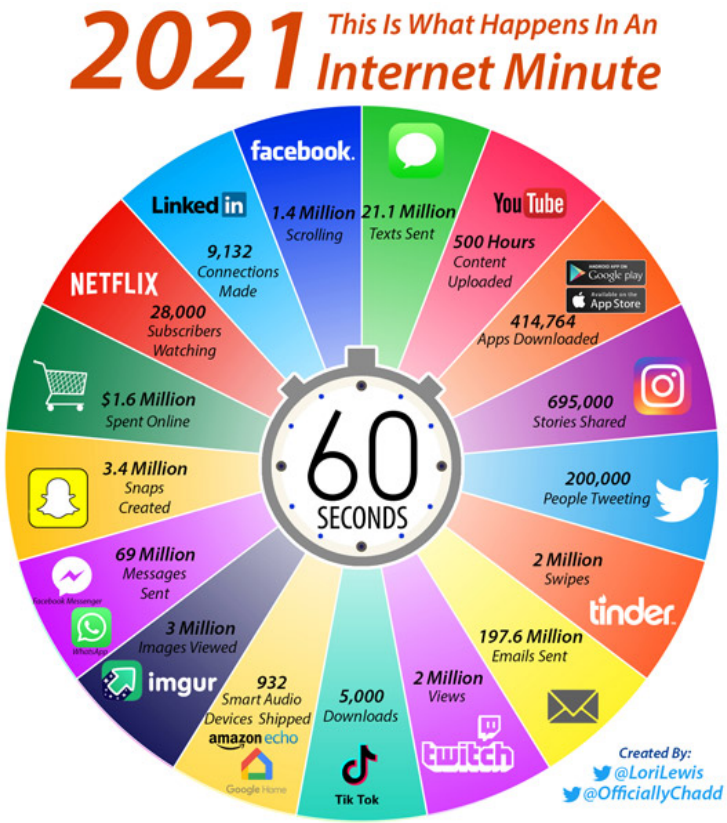
- Google traite 20 PO par jour (2008). Données stockées (2014) 15000 PO
- Facebook a 2,5 PO de données utilisateur par jour (2009). Données stockées (2014) 300 PO

Quelques chiffres

- Google traite 20 PO par jour (2008). Données stockées (2014) 15000 PO
- Facebook a 2,5 PO de données utilisateur par jour (2009). Données stockées (2014) 300 PO
- eBay a 6,5 PO de données utilisateur par jour (2009). Données stockées (2014) 90 PO

Quelques chiffres





Les premiers confrontés à de grande volumétries de données

- Google, Yahoo, Twitter, LinkedIn...etc
- Innovation portée principalement sur deux technologies :
  - Les bases de données NoSQL
  - Développement de plateformes de stockage et de traitement des données

Ce que peut apporter le Big Data

- Prise de décision en apportant d'autres axes d'analyse (ex. les sentiments ou les réactions du client)

Ce que peut apporter le Big Data

- Prise de décision en apportant d'autres axes d'analyse (ex. les sentiments ou les réactions du client)
- Plus de compétitivité

Ce que peut apporter le Big Data

- Prise de décision en apportant d'autres axes d'analyse (ex. les sentiments ou les réactions du client)
- Plus de compétitivité
- Une meilleur compréhension des informations

Les enjeux du Big Data

- Réunir un grand volume de données variées

Les enjeux du Big Data

- Réunir un grand volume de données variées
- Traiter ces données

Les enjeux du Big Data

- Réunir un grand volume de données variées
- Traiter ces données
- Protéger ces données

Les enjeux du Big Data

- Réunir un grand volume de données variées
- Traiter ces données
- Protéger ces données
- Trouver les informations utiles

Les enjeux du Big Data

- Réunir un grand volume de données variées
- Traiter ces données
- Protéger ces données
- Trouver les informations utiles
- Visualiser ces informations / données

Les 3-Vs

- Volume

Les 3-Vs

- Volume
- Variété

Les 3-Vs

- Volume
- Variété
- Vitesse

Les 3-Vs

- Volume
- Variété
- Vitesse

La définition Wikipedia

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Les 3-Vs

- Volume
- Variété
- Vitesse

La définition Wikipedia

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

La définition qui résume un peu tout

Le **Big Data** c'est quand les données font partie du problème (Mike Loukides - O'Reilly Radar)

Les 5-Vs

- Véracité

Les 5-Vs

- Véracité
- Valeur

L'approche traditionnelle

Les besoins métier guident la conception de la solution

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités
- Le responsable métier éxécute les requêtes pour répondre aux questions

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités
- Le responsable métier éxécute les requêtes pour répondre aux questions
- De nouvelles éxigences nécessitent de nouvelles conceptions / maintenances

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités
- Le responsable métier éxécute les requêtes pour répondre aux questions
- De nouvelles éxigences nécessitent de nouvelles conceptions / maintenances

Appropriée pour :

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités
- Le responsable métier éxécute les requêtes pour répondre aux questions
- De nouvelles éxigences nécessitent de nouvelles conceptions / maintenances

Appropriée pour :

- Des données structurées

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités
- Le responsable métier éxécute les requêtes pour répondre aux questions
- De nouvelles éxigences nécessitent de nouvelles conceptions / maintenances

Appropriée pour :

- Des données structurées
- Opérations et processus répétitifs

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités
- Le responsable métier éxécute les requêtes pour répondre aux questions
- De nouvelles éxigences nécessitent de nouvelles conceptions / maintenances

Appropriée pour :

- Des données structurées
- Opérations et processus répétitifs
- Sources relativement stables

L'approche traditionnelle

Les besoins métier guident la conception de la solution

- Le responsable métier définit les besoins
- Le service informatique conçoit la solution avec un ensemble de structures et de fonctionnalités
- Le responsable métier éxécute les requêtes pour répondre aux questions
- De nouvelles éxigences nécessitent de nouvelles conceptions / maintenances

Appropriée pour :

- Des données structurées
- Opérations et processus répétitifs
- Sources relativement stables
- Besoins bien compris et bien cadrés

L'approche Big Data

Les sources d'information guident la découverte

L'approche Big Data

Les sources d'information guident la découverte

- Le responsable métier et le service informatique identifient les sources de données disponibles

L'approche Big Data

Les sources d'information guident la découverte

- Le responsable métier et le service informatique identifient les sources de données disponibles
- Le service informatique fournit la plateforme qui permettra l'exploration des données disponibles

L'approche Big Data

Les sources d'information guident la découverte

- Le responsable métier et le service informatique identifient les sources de données disponibles
- Le service informatique fournit la plateforme qui permettra l'exploration des données disponibles
- Le responsable métier détermine la question à poser en explorant les données et les liens entre les données

L'approche Big Data

Les sources d'information guident la découverte

- Le responsable métier et le service informatique identifient les sources de données disponibles
- Le service informatique fournit la plateforme qui permettra l'exploration des données disponibles
- Le responsable métier détermine la question à poser en explorant les données et les liens entre les données
- Intégrer les nouvelles connaissances à l'approche traditionnelle

L'approche Big Data

Les sources d'information guident la découverte

- Le responsable métier et le service informatique identifient les sources de données disponibles
- Le service informatique fournit la plateforme qui permettra l'exploration des données disponibles
- Le responsable métier détermine la question à poser en explorant les données et les liens entre les données
- Intégrer les nouvelles connaissances à l'approche traditionnelle

La question :

L'approche Big Data

Les sources d'information guident la découverte

- Le responsable métier et le service informatique identifient les sources de données disponibles
- Le service informatique fournit la plateforme qui permettra l'exploration des données disponibles
- Le responsable métier détermine la question à poser en explorant les données et les liens entre les données
- Intégrer les nouvelles connaissances à l'approche traditionnelle

La question :

- Comment faire fonctionner ensemble les deux approches ?

Méga-Octet, Tera-Octet, Peta-Octet et au delà ?

- C'est quoi ?

Méga-Octet, Tera-Octet, Peta-Octet et au delà ?

- C'est quoi ?
- Des préfixes grecs

Méga-Octet, Tera-Octet, Peta-Octet et au delà ?

- C'est quoi ?
- Des préfixes grecs
- Exa-Octet (EO)

Méga-Octet, Tera-Octet, Peta-Octet et au delà ?

- C'est quoi ?
- Des préfixes grecs
- Exa-Octet (EO)
- Zetta-Octet (ZO)

Méga-Octet, Tera-Octet, Peta-Octet et au delà ?

- C'est quoi ?
- Des préfixes grecs
- Exa-Octet (EO)
- Zetta-Octet (ZO)
- Yotta-Octet (YO)

Sommaire

Google

Aux commencement, tout est parti de deux publications de Google

- Google File System (GFS) : une solution évolutive de système de fichiers distribué pour les applications de données intensives réparties.

Google

Aux commencement, tout est parti de deux publications de Google

- Google File System (GFS) : une solution évolutive de système de fichiers distribué pour les applications de données intensives réparties.
- MapReduce : traitement de donnée simplifié sur un grand nombre de machines.

Google

Aux commencement, tout est parti de deux publications de Google

- Google File System (GFS) : une solution évolutive de système de fichiers distribué pour les applications de données intensives réparties.
- MapReduce : traitement de donnée simplifié sur un grand nombre de machines.
- Ces systèmes décrivent la nouvelle infrastructure de Google pour traiter les grandes quantités de données. Ces systèmes sont restés propriétaires.

Yahoo

- Doug Cutting a été embauché par Yahoo pour poursuivre ses travaux sur le dévéloppement d'un framework de calcul distribué.

Yahoo

- Doug Cutting a été embauché par Yahoo pour poursuivre ses travaux sur le dévéloppement d'un framework de calcul distribué.
- Naissance de Hadoop (une implémentation libre de MapReduce et d'un système de fichier distribué).

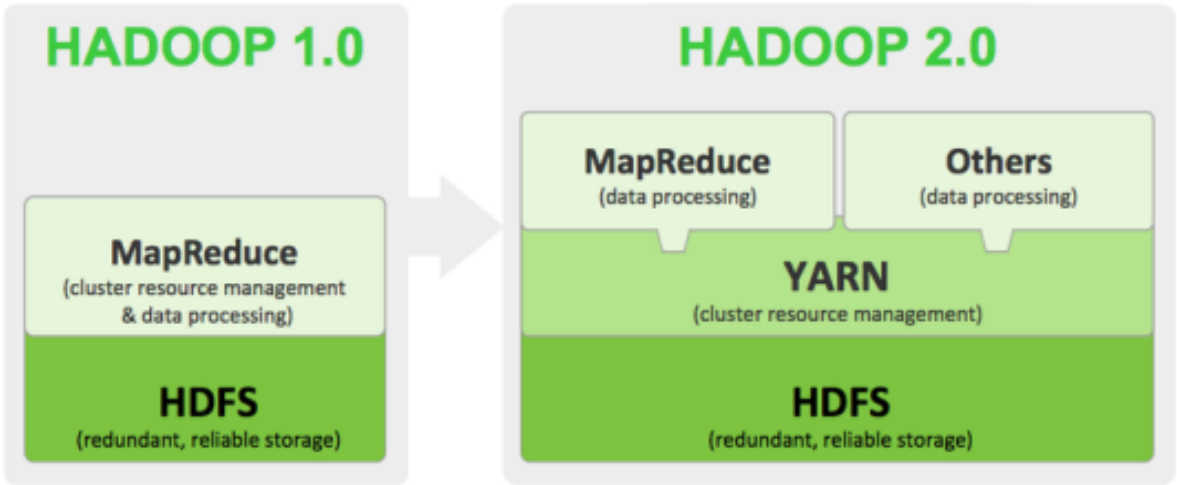
Yahoo

- Doug Cutting a été embauché par Yahoo pour poursuivre ses travaux sur le dévéloppement d'un framework de calcul distribué.
- Naissance de Hadoop (une implémentation libre de MapReduce et d'un système de fichier distribué).
- Yahoo donne Hadoop à la fondation Apache.

hadoop

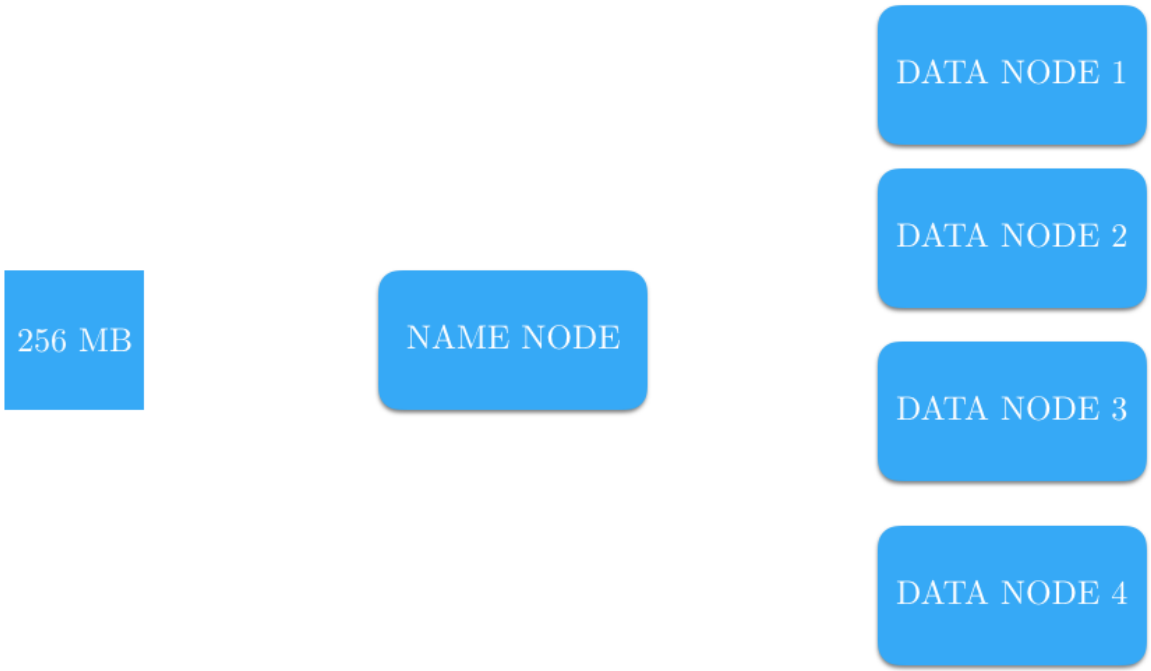
- Logiciel open source conçu pour le stockage et le traitement de données à grande échelle.
- Cutting a nommé le logiciel comme le jouet éléphant de son fils.
- Qui utilise Hadoop ?

hadoop



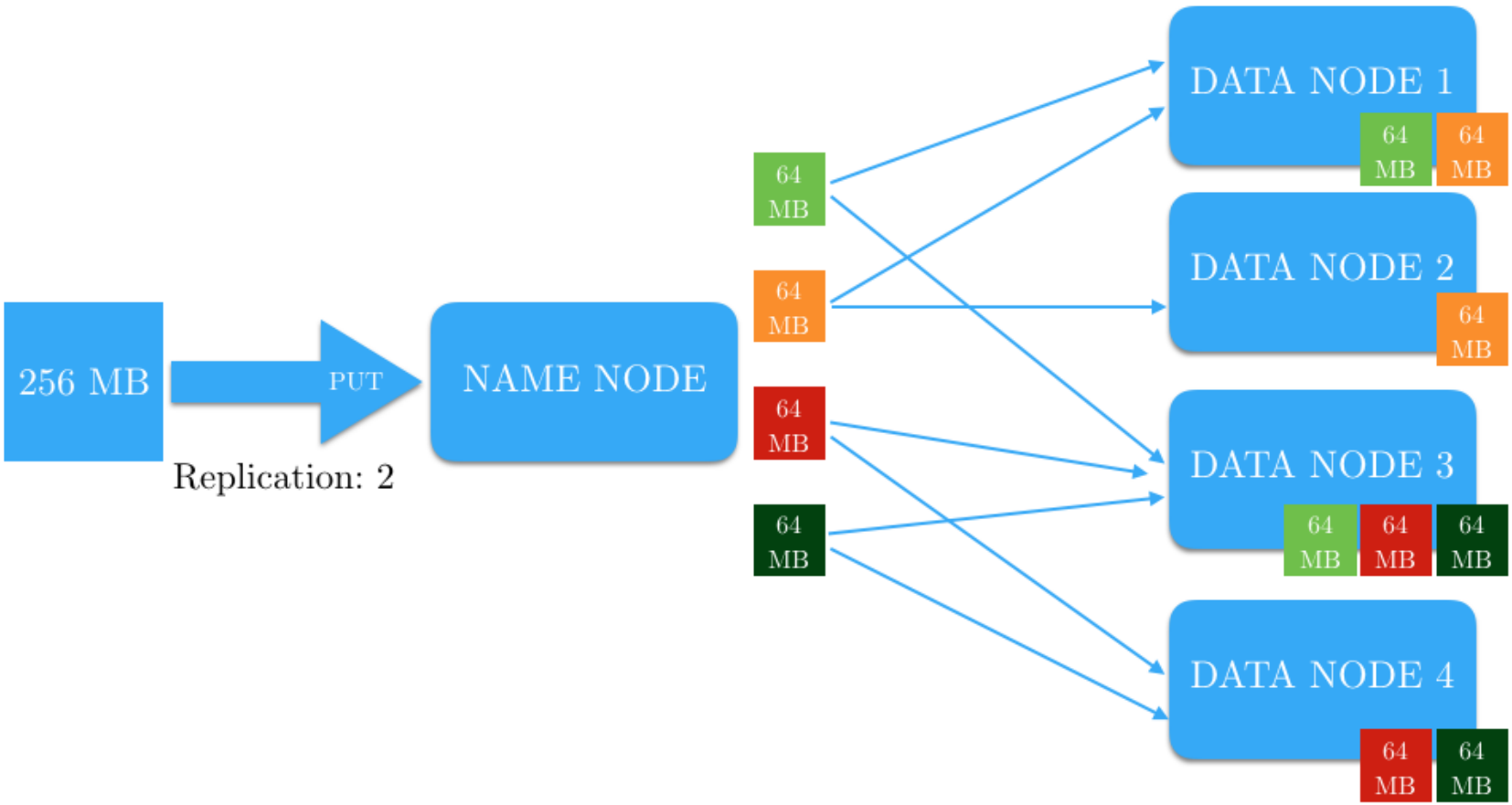
Le stockage distribué

- Hadoop Distributed File System



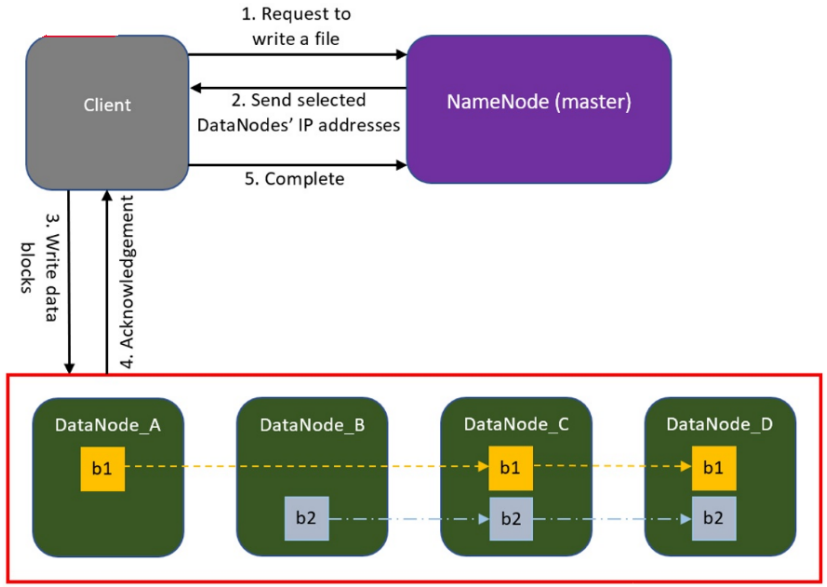
Le stockage distribué

- Hadoop Distributed File System



Le stockage distribué

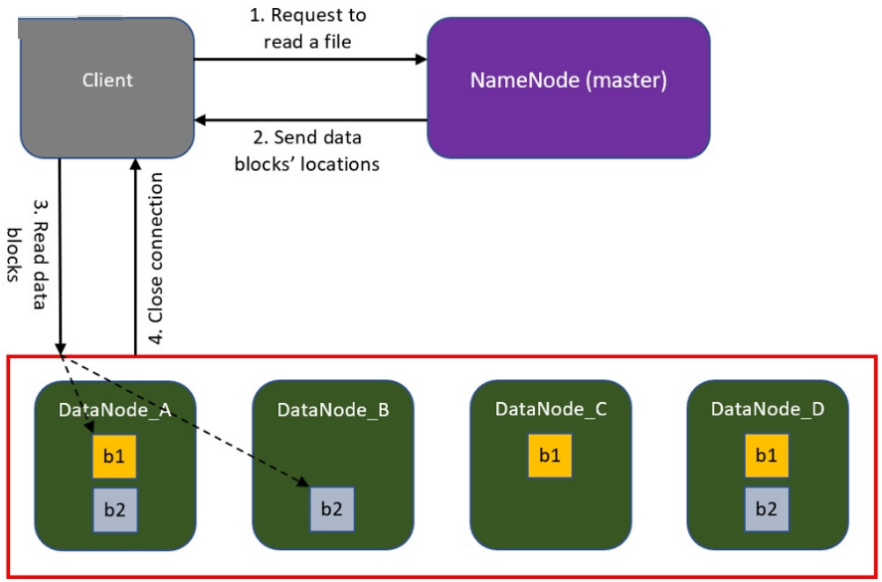
- Un client envoie une demande d'écriture d'un fichier au *Name-Node* :



- Un modèle WORM (*Write Once Read Many*).

Le stockage distribué

- Un client envoie une demande de lecture de fichier au *NameNode*, ce dernier consulte le fichier de métadonnées (FsImage) et indique au client les *DataNodes* qui ont les blocs de données.



Le traitement distribué

- Conséquences
  - Pour le système : besoin d'un réseau de communication.

Le traitement distribué

- Conséquences
  - Pour le système : besoin d'un réseau de communication.
  - **Pour les algorithmes : données partielles -> résultats partiels.**
- Nouveau paradigme de programmation

Le traitement distribué

- Conséquences
  - Pour le système : besoin d'un réseau de communication.
  - **Pour les algorithmes : données partielles -> résultats partiels.**
- Nouveau paradigme de programmation
  - Solution :

Le traitement distribué

- Conséquences
  - Pour le système : besoin d'un réseau de communication.
  - **Pour les algorithmes : données partielles -> résultats partiels.**
- Nouveau paradigme de programmation
  - Solution :
    - Google MapReduce

Le traitement distribué

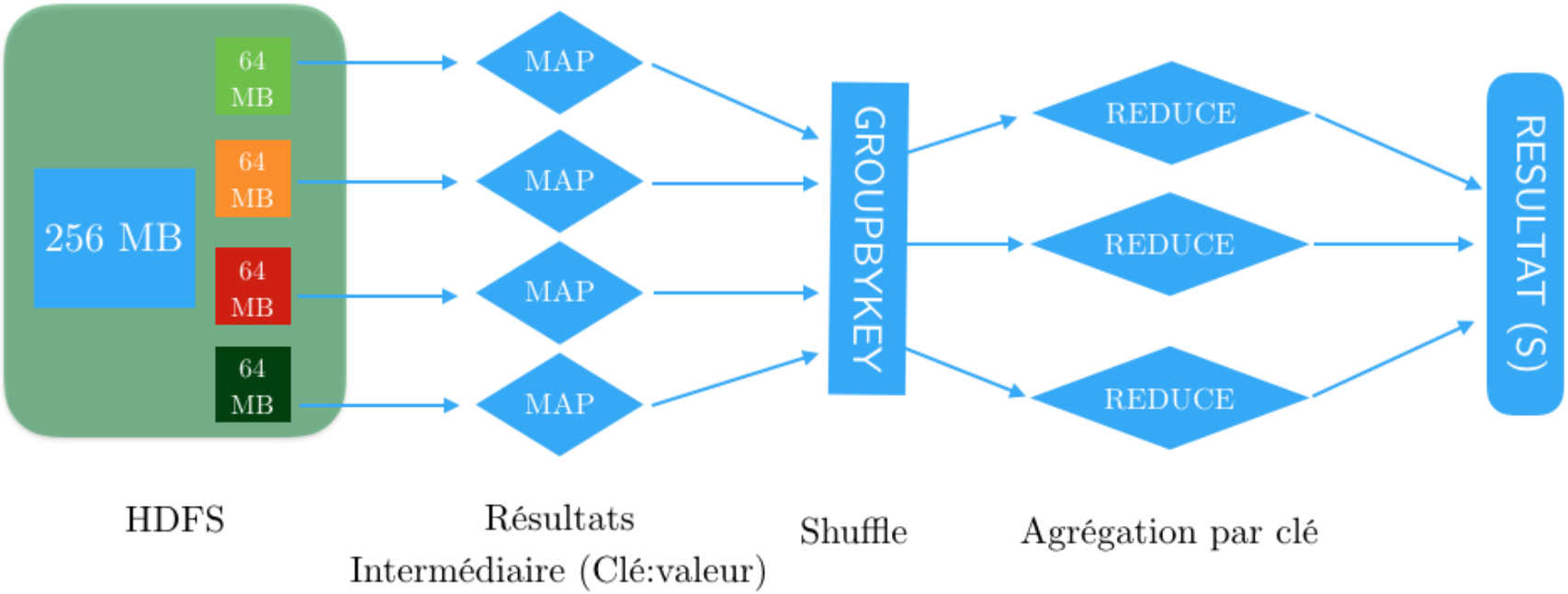
- Conséquences
  - Pour le système : besoin d'un réseau de communication.
  - **Pour les algorithmes : données partielles -> résultats partiels.**
- Nouveau paradigme de programmation
  - Solution :
    - Google MapReduce
    - Hadoop : l'implémentation la plus connue

MapReduce

- Illustration des *Map* et *Reduce*

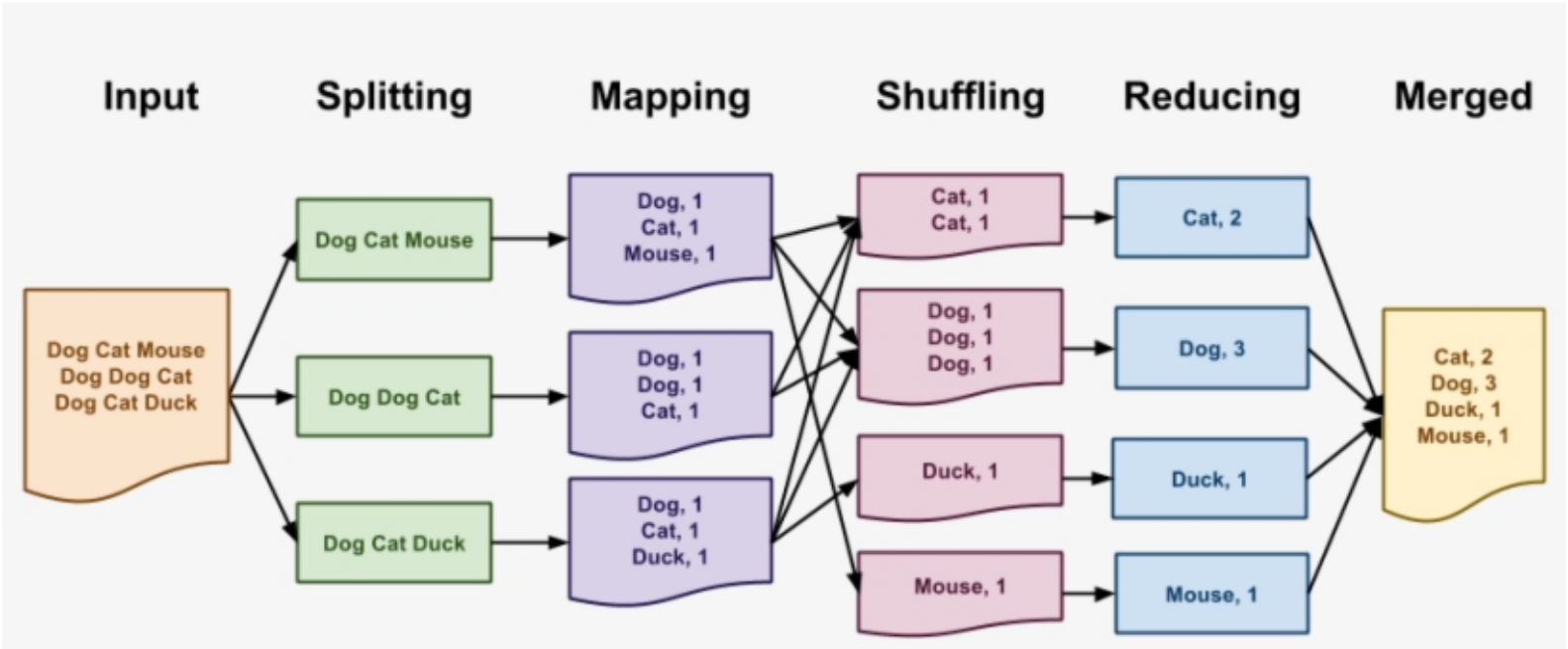
MapReduce

- Illustration des *Map* et *Reduce*



MapReduce

- Exemple : Word count



MapReduce

- Un modèle de programmation inspiré de la programmation fonctionnelle.
- Hadoop est capable d'exécuter des programmes MapReduce écrits dans divers langages : Java, Ruby, Python, C++, ...
- Rôle des programmeurs : écrire des *mappers* et des *reducers*.

MapReduce

- Avantages :

MapReduce

- Avantages :
  - Une parallélisation et distribution du traitement qui est tolérant aux fautes.

MapReduce

- Avantages :
  - Une parallélisation et distribution du traitement qui est tolérant aux fautes.
  - Une simple abstraction de données adaptable à tout algorithme.

MapReduce

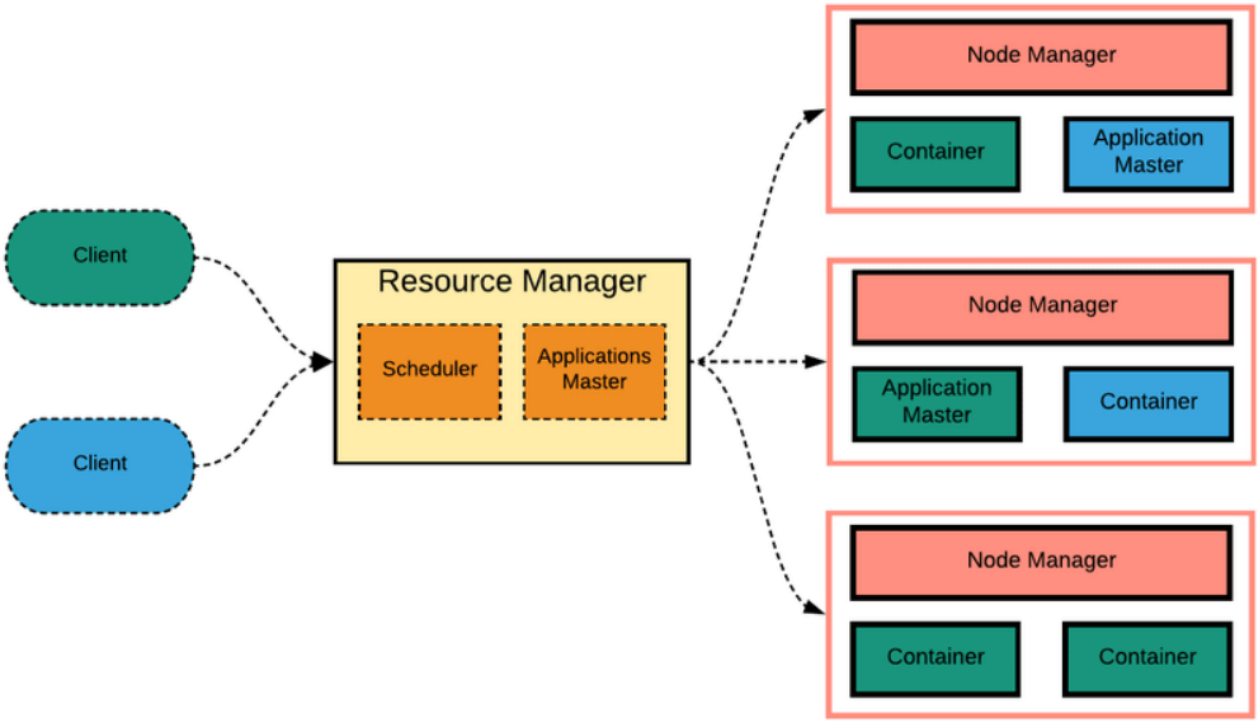
- Avantages :
  - Une parallélisation et distribution du traitement qui est tolérant aux fautes.
  - Une simple abstraction de données adaptable à tout algorithme.
- Inconvénients :

MapReduce

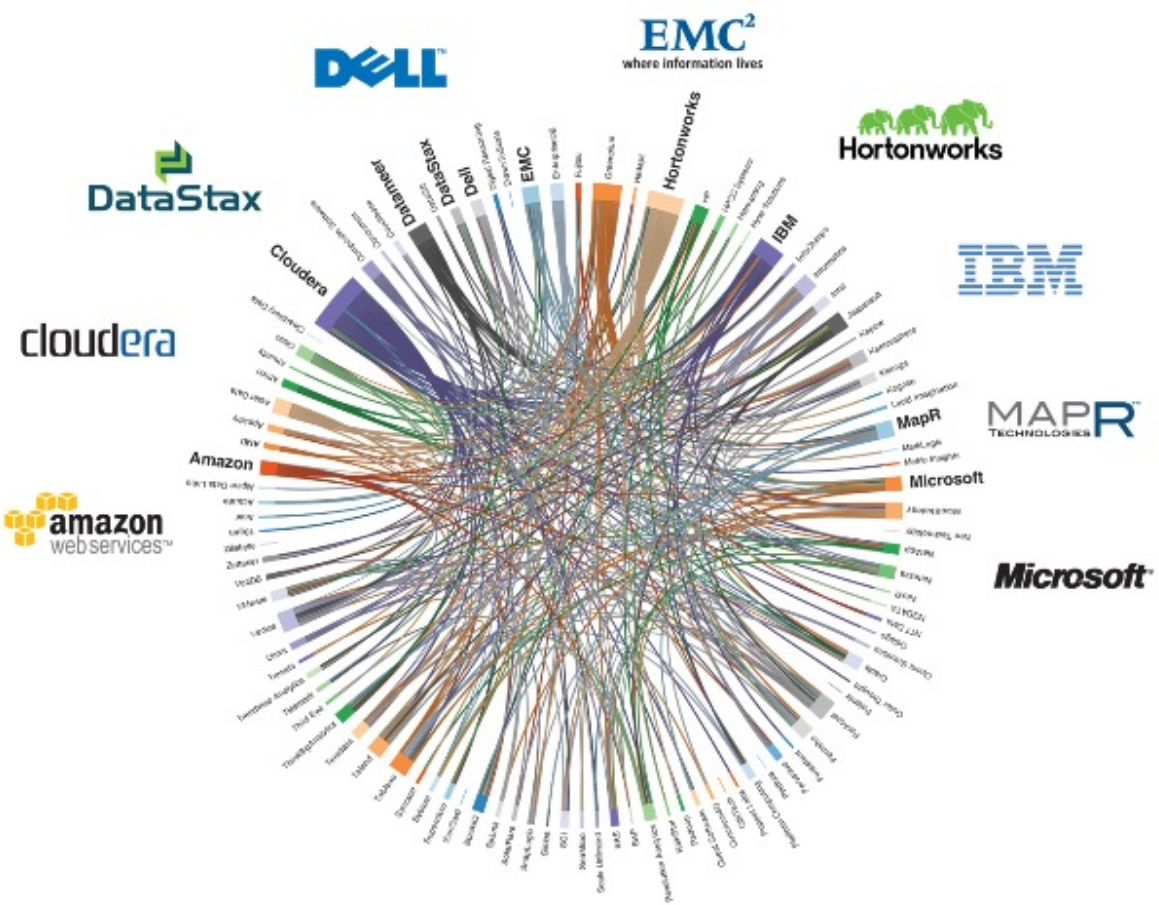
- Avantages :
  - Une parallélisation et distribution du traitement qui est tolérant aux fautes.
  - Une simple abstraction de données adaptable à tout algorithme.
- Inconvénients :
  - *Mappers* et *Reducers* sont construits indépendamment.

MapReduce/YARN

- MapReduce v2/YARN

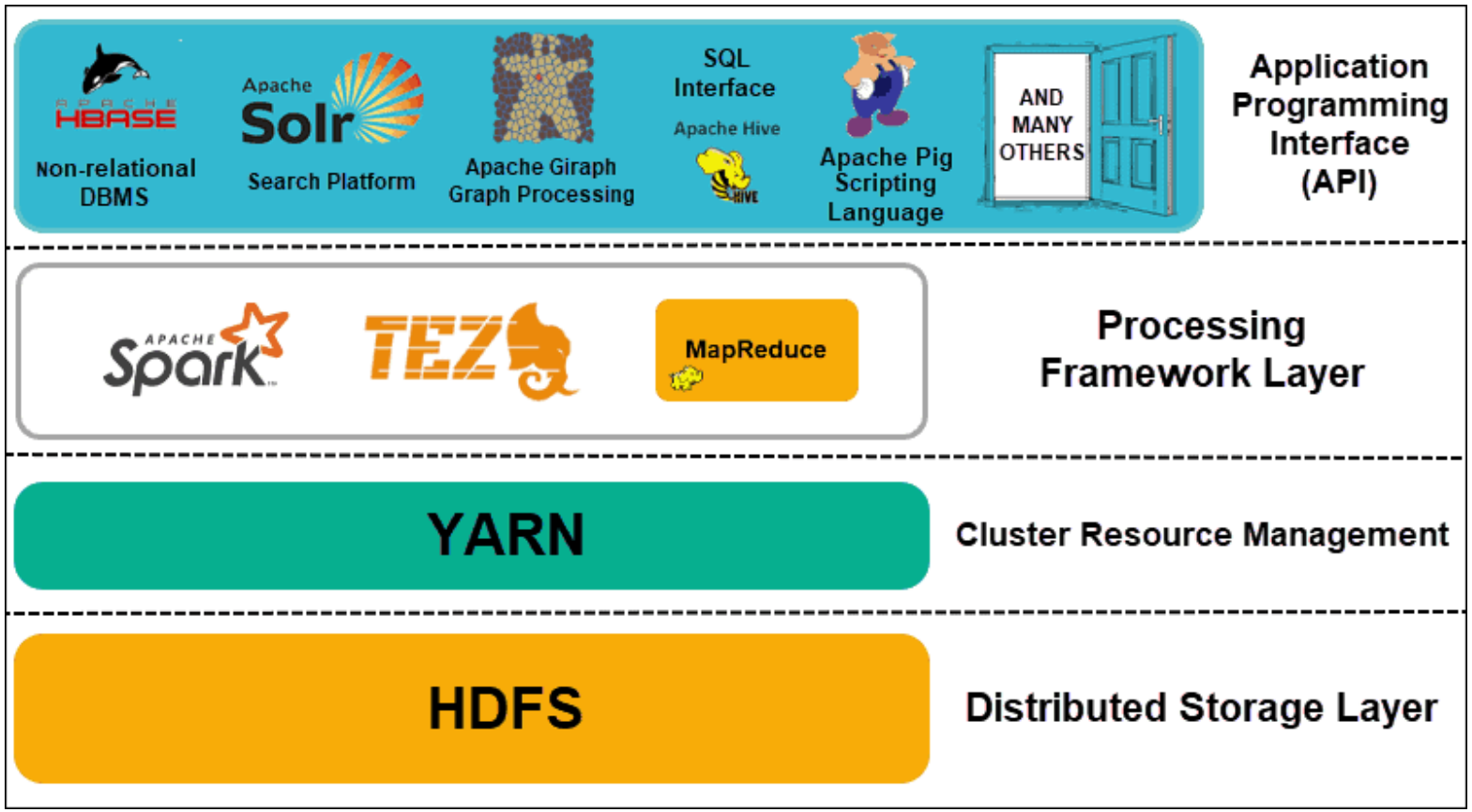


Les distributions de Hadoop



Source : https ://www.lebigdata.fr/hadoop-top-des-distributeurs

Présentation : écosystème Hadoop



Débat : les métiers du Big Data

- Big Data solutions architect

Débat : les métiers du Big Data

- Big Data solutions architect
- Big Data engineer

Débat : les métiers du Big Data

- Big Data solutions architect
- Big Data engineer
- Big data manager

Débat : les métiers du Big Data

- Big Data solutions architect
- Big Data engineer
- Big data manager
- Big data visualizer

Débat : les métiers du Big Data

- Big Data solutions architect
- Big Data engineer
- Big data manager
- Big data visualizer
- ...

Débat : les métiers du Big Data

Références

- Cours de R. Rado
- Livre : Hadoop Illuminated.
- Livre : Hadoop In Pratice.
- Livre : Hadoop In Action.
- https ://ressources.esri.ca