# Enjeux des Big Data IoT

Sihem Amer-Yahia
CNRS, LIG

**SIF – 23 mars 2016**

# Résumé

- Les *Volume*, *Variété* et *Vélocité* des Big Data ont mené à repenser la chaîne de leur traitement :
  - Ingestion des données volumineuses et hétérogènes à la vitesse à laquelle elles arrivent.
- Les profils des postes en techno de l'info ont changé :
  - *Data Scientists :* gouvernance des données et le développement de systèmes distribués.
  - Formation d'employés sur les technos Big Data.
- Aujourd'hui la valeur est dans la capacité à traiter la *Véracité* et la *Volatilité*, principales caractéristiques des Big Data IoT.
  - *Véracité*: les données IoT sont entâchées d'imprécisions – impact de l'environnement sur les capteurs, synchronisation imparfaite, sources diverses.
  - *Volatilité*: ce qui est observé à un instant $t$ n'est plus utile à $t+1$, ce qui est valable dans une région ne l'est pas dans une autre, ce qui est utile pour un ensemble de personnes ne l'est pas pour un autre.
- Les technos Big Data sont-elles adaptées aux IoT ?

# Plan

- Les Data Stack et les Data Lakes
- Volume, Variété et Vélocité
- Véracité et Volatilité

# Data Stack (classique)

**Tableaux de bord**

**Traitement des données**

**Stockage et Indexation**

| Ward No. | Ward name | Type | No. of Beds |
|---|---|---|---|
| 3 | Carey | Medical | 8 |
| 6 | Bracken | Medical | 16 |
| 7 | Brent | Surgical | 12 |
| 8 | Meavy | Surgical | 10 |

**données structurées et semi-structurées**

```
<?xml version="1.0"?>
<quiz>
 <question>
  Who was the forty-second
  president of the U.S.A.?
 </question>
 <answer>
  William Jefferson Clinton
 </answer>
 <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

# Pourquoi de nouvelles architectures Big Data ?

- Parce que les besoins sont plus ambitieux.

- Parce que *Volume*, *Variété* et *Vélocité* nécessitent de :
  - Rendre la structure de coût plus flexible
  - Repenser la collecte des données
  - Restructurer le stockage des données
  - Permettre la fouille et le traitement de données en flux

# Des besoins plus ambitieux

- **Agriculture**
  - *Objectif* : réduire les coûts d'arrosage
  - *Approche* : corréler les données des capteurs d'humidité du sol et celles de la météo
  - *Difficulté* : traiter ensemble les différents types de données

- **Télécommunications**
  - *Objectif* : réduire le churn des clients
  - *Approche* :  prévoir les problèmes réseau et les corréler avec les activités de chaque client
  - *Difficulté* : synthétiser les rapports par client pour offrir un service personnalisé

- **Sécurité**
  - *Objectif* : identifier des activités suspectes dans des zones de pêche interdite
  - *Approche* : corréler des images géo-spatiales avec des tweets
  - *Difficulté* : analyser les images pour identifier des bateaux suspects et analyser les tweets

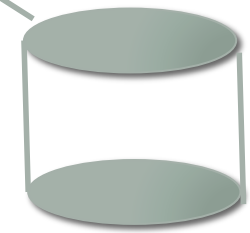
- **Santé (épidémiologie)**
  - *Objectif* : établir un lien entre la santé, la nutrition et l'activité physique des individus
  - *Approche* : corréler des tickets de caisse avec des tweets et des informations de déplacement
  - *Difficulté* : analyser les données dans le temps

# Data Lake = Data Stack + Intégration, Préparation, Fouille

**Analytique**
Traitement **et Fouille à Grande Echelle**

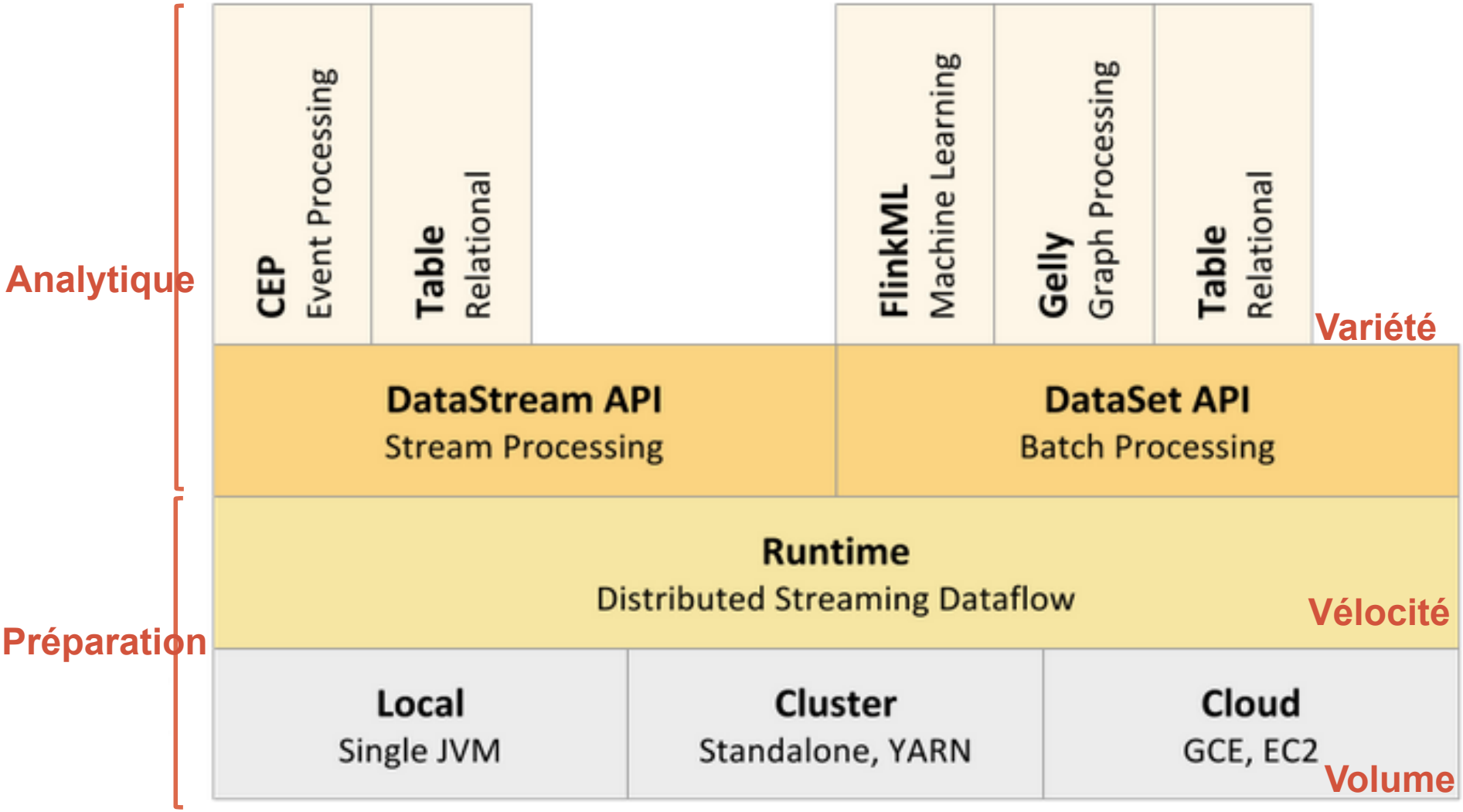
**Collecte et Préparation**
Stockage, **Intégration** et Indexation

```
1381701784;10.66.17.30;66.249.75.151;80;63478;8762
1381701779;10.65.1.64;178.154.165.250;80;59708;7829
1381701808;192.134.0.49;10.64.1.43;53;32751;353
1381701766;212.198.5.116;10.66.164.10;33140;12211;216
1381701759;216.14.118.226;172.18.1.155;80;53635;20813
1381701781;172.16.1.127;79.89.31.97;80;51957;17236
1381701763;10.66.53.31;193.190.89.30;80;25755;388
1381701774;10.66.146.11;69.171.224.112;80;52652;30999
1381701793;88.179.58.121;10.66.53.31;54203;43855;292
1381701770;10.66.228.42;41.224.50.5;443;46614;40
```



**sources externes**

# Data Lake de Flink/Spark/Stratio/…



# Véracité et Volatilité

- Nécessitent une approche algorithmique
- Ne doivent pas compromettre Volume et Vélocité
- Un sujet de recherche ouvert

- Quelques pistes :
  - Ne pas tenter de “réparer” les données
  - Ne pas fixer une granularité de données
  - Permettre le test d’hypothèses dans le temps et l’espace

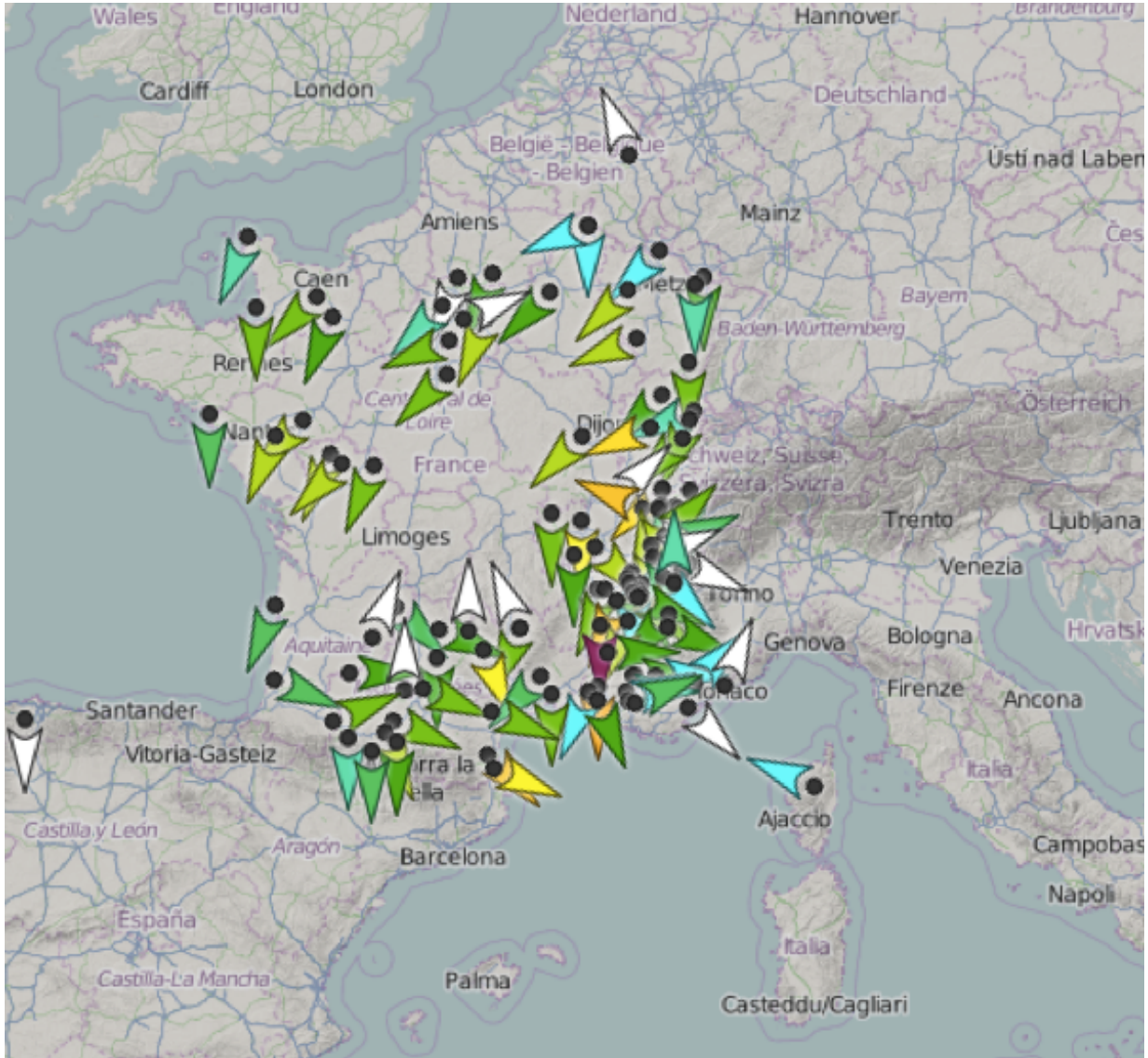
# Véracité

*repenser le traitement des données*

*raisonner de manière approximative*

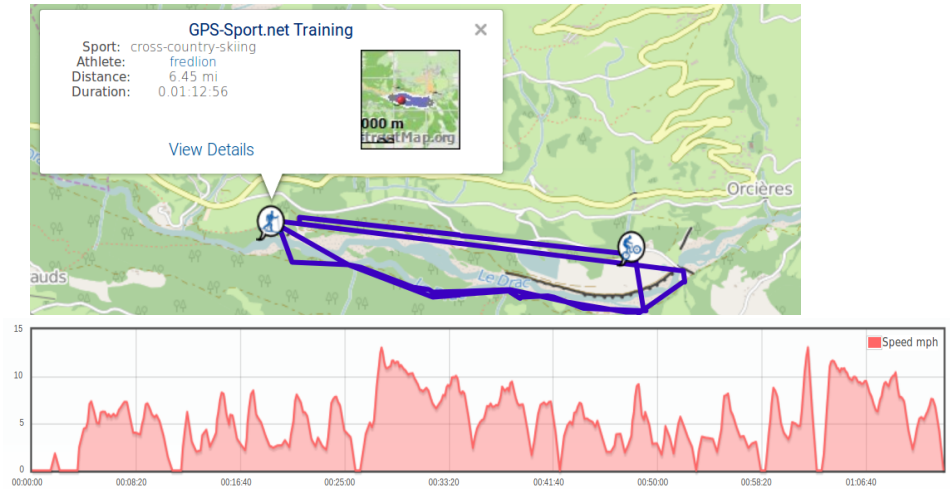
# Les capteurs ne sont pas d'accord

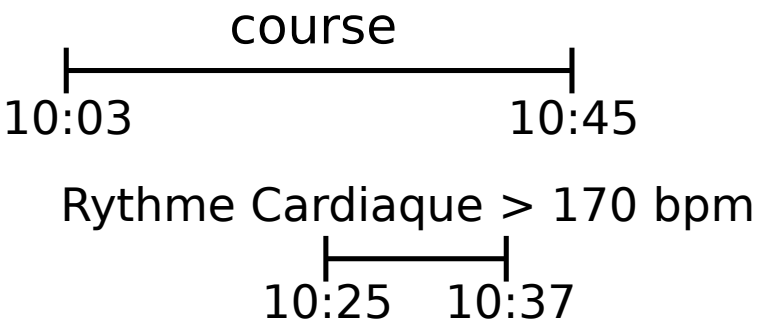
[http://pioupiou.fr/fr/map](http://pioupiou.fr/fr/map)



# Problème : observations asynchrones
# Solution : approximation du traitement

## Quantified Self




Connecter des données provenant de différentes sources:

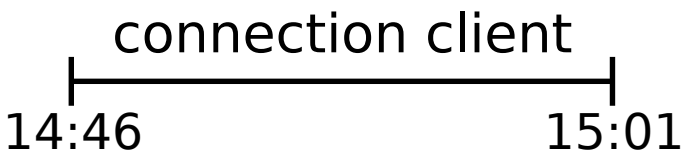
$$(x, y) \in C_1 \times C_2 \text{ s.t. } \mathit{contains}(x, y)$$

***Distributed Top-K Temporal Joins**, avec J. Pilourdault et V. Leroy, à paraître dans ACM SIGMOD 2016.*

# Problème : observations asynchrones
# Solution : approximation du traitement

## Données Réseau

```
1381701784;10.66.17.30;66.249.75.151;80;63478;8762
1381701779;10.65.1.64;178.154.165.250;80;59708;7829
1381701808;192.134.0.49;10.64.1.43;53;32751;353
1381701766;212.198.5.116;10.66.164.10;33140;12211;216
1381701759;216.14.118.226;172.18.1.155;80;53635;20813
1381701781;172.16.1.127;79.89.31.97;80;51957;17236
1381701763;10.66.53.31;193.190.89.30;80;25755;388
1381701774;10.66.146.11;69.171.224.112;80;52652;30999
1381701793;88.179.58.121;10.66.53.31;54203;43855;292
1381701770;10.66.228.42;41.224.50.5;443;46614;40
```



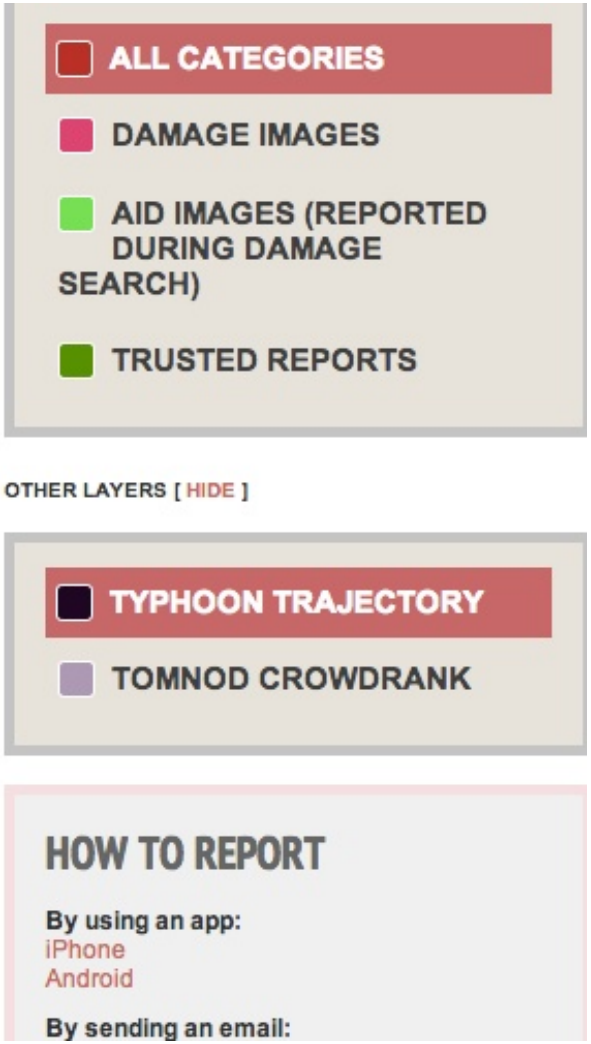

Suivre le traffic provenant de deux pays:

$$(x, y) \in C_1 \times C_2 \text{ s.t. } \mathit{justBefore}(x, y)$$

*avec J. Pilourdault et V. Leroy.* ***Distributed Top-K Temporal Joins****, à paraître dans ACM SIGMOD 2016.*

# Véracité : au delà des IoT, les IoE

## Capteurs humains sur CrowdMap (Ushahidi)

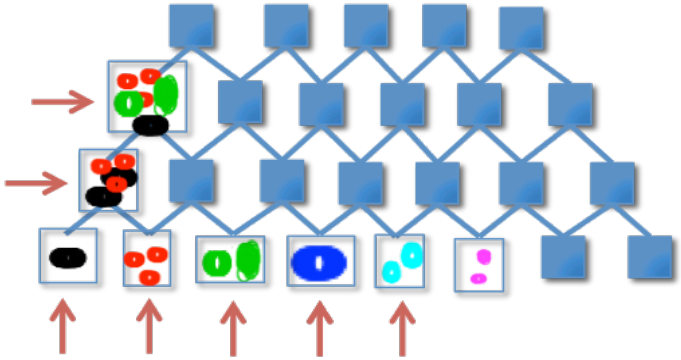


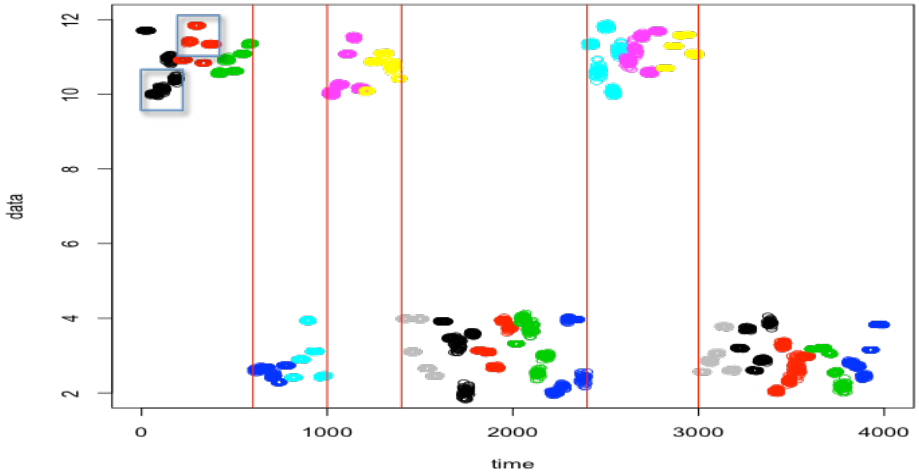
# Volatilité

*raisonner à plusieurs granularités*

*permettre une analytique évolutive*

# Raisonner à plusieurs granularités
## sémantiques, temporelles et spatiales






*avec S. Kleisarchaki, A. Douzal, V. Christophides: **Querying Temporal Drifts at Multiple Granularities**. CIKM 2015.*
*avec S. Mishra, V. Leroy: **Discovering Characterizing Regions for Consumer Products**. DSAA 2015.*

# Permettre une analytique évolutive

- Pour prédire à 2h d'avance la pluviométrie d'un champ en particulier, le plus fiable n'est pas d'utiliser un an d'historique :
  - 24h sur  un rayon de 100km serait plus utile.

- Repenser la fouille de données
  - Interactive
  - Continue
  - Rétroactive

# Conclusion

- *Volume*, *Variété* et *Vélocité* ont imposé une refonte de la chaîne de traitement des données.

- *Véracité* et *Volatilité* nécessitent un changement de paradigme de leur analytique.
  - Traitement approximatif pour une gestion flexible de la véracité.
  - Fouille interactive et continue pour une analytique évolutive.

- L'implication des experts du métier est indispensable.
  - Leur fournir les outils nécessaires pour manipuler/explorer les données.
  - SAP développe ces outils mais seulement pour la préparation des données.