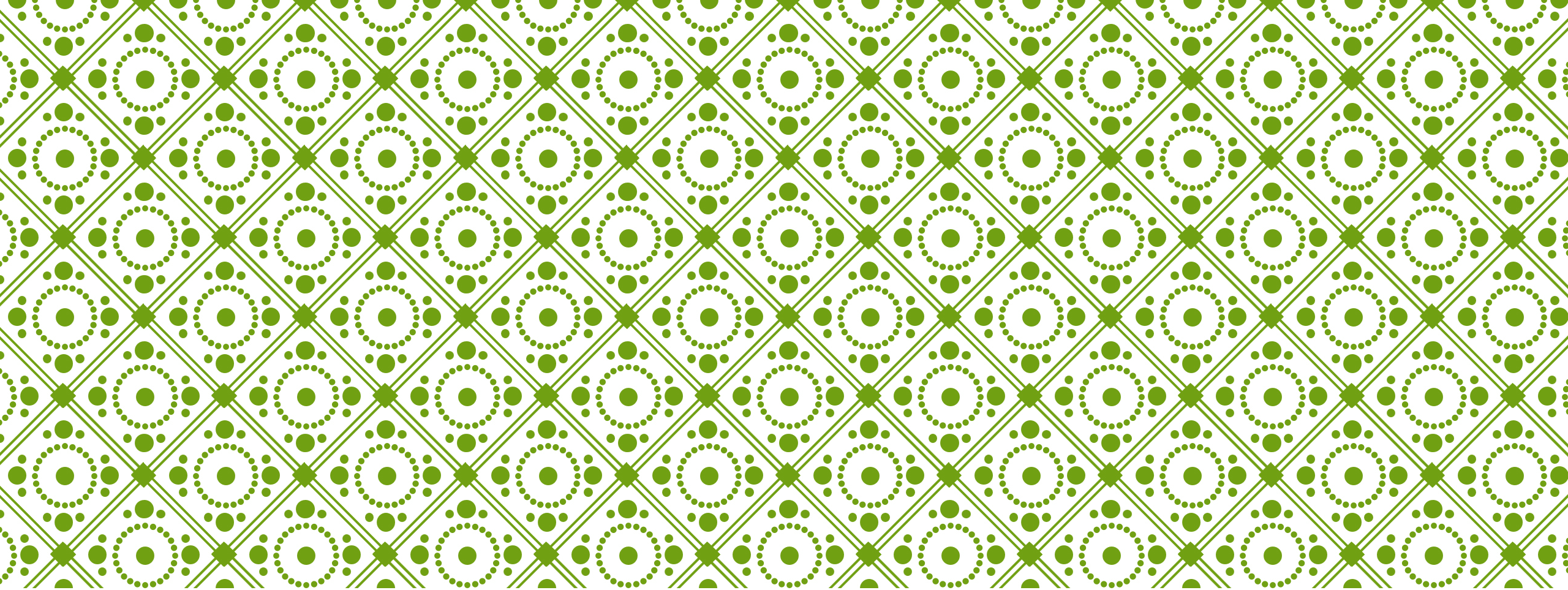


APRIORI PARALLÈLE

Forage de Données

Hélène de Fromont – Simon Roussel – Jérémie Pouillon

INTRODUCTION

Analyser les **hashtags** ou les **mots-clefs** utilisés par un utilisateur, pour proposer du contenu lié (donc susceptible de l'intéresser).

Paralléliser l'algorithme d'Apriori pour améliorer ses performances sur des gros volumes de données.

Faire la **comparaison** entre Apriori et Apriori Parallèle.

PRÉSENTATION D'HADOOP

Hadoop est un framework développé en Java, inspiré par MapReduce dont le but est de faciliter le développement d'applications distribués et échelonnables.

Il appartient depuis 2009 à Apache Software Foundation.

RÉCUPÉRATION DES DONNÉES

- Nombre d'utilisateurs limités en mode simple.
- Peu de contenu récupérable.

Instagram

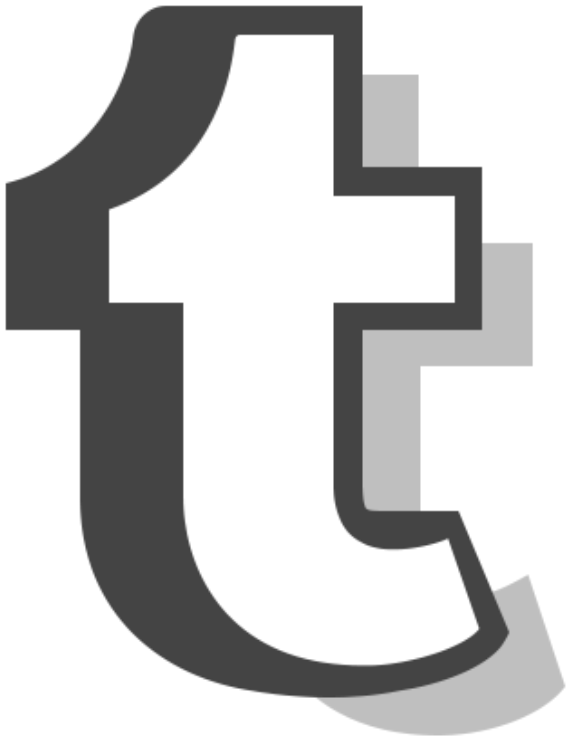
- Nombres de caractères limités par post.
- Peu de hashtags.

Twitter

- Moins d'utilisateurs que les deux autres.
- Généralement beaucoup de hashtags par post.

Tumblr

L'API DE TUMBLR




```
{
   "meta": {
      "status": 200,
      "msg": "OK"
   }, "response": [
      {
         "blog_name": "captaindadpool13",
         "id": 143090889352,
         "post_url": "http://captaindadpool13.tumblr.com/post/143090889352",
         "type": "photo",
         "date": "2016-04-20 02:09:50 GMT",
         "timestamp": 1461118190,
         "state": "published",
         "tags": ["deadpool", "marvel", "attack on titan", "comic book", "nerd", "geek"],
         "note_count": 0,
         "caption": "",
         "photos": [ … ],
         "can_send_in_message": true,
         "can_reply": false
      }, ... ]
}
```

BASE DE DONNÉES

#deadpool

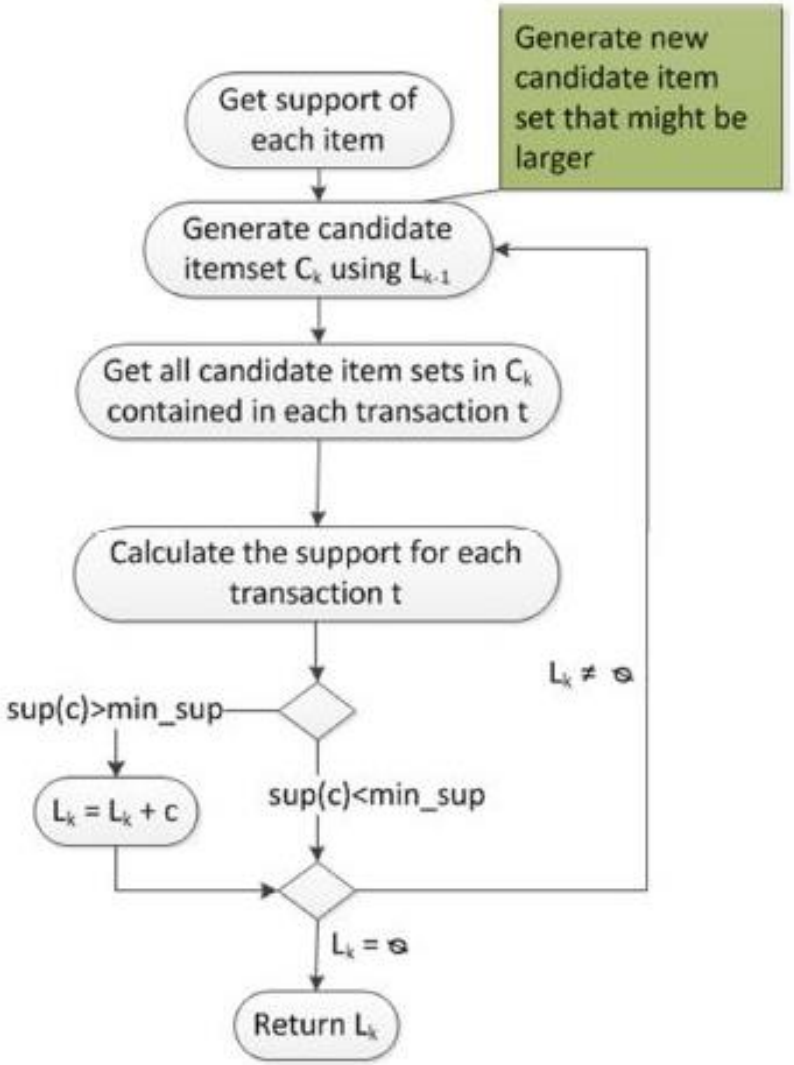
Obtention des hashtags de 60.000 posts.

(Possibilité d'avoir des caractères spéciaux dans les hashtags de Tumblr.)

```
3    deadpool      deadpool 2016     deadpool spoilers   wade wilson    i love deadpool you guys
4    me   deadpool      sara rambles
5    💀💩l      deadpool      dopinder
…
44   Spider-Man    DeadpoolSpidey    Eric B. & Rakim Paid In Full      Marvel
```

APRIORI

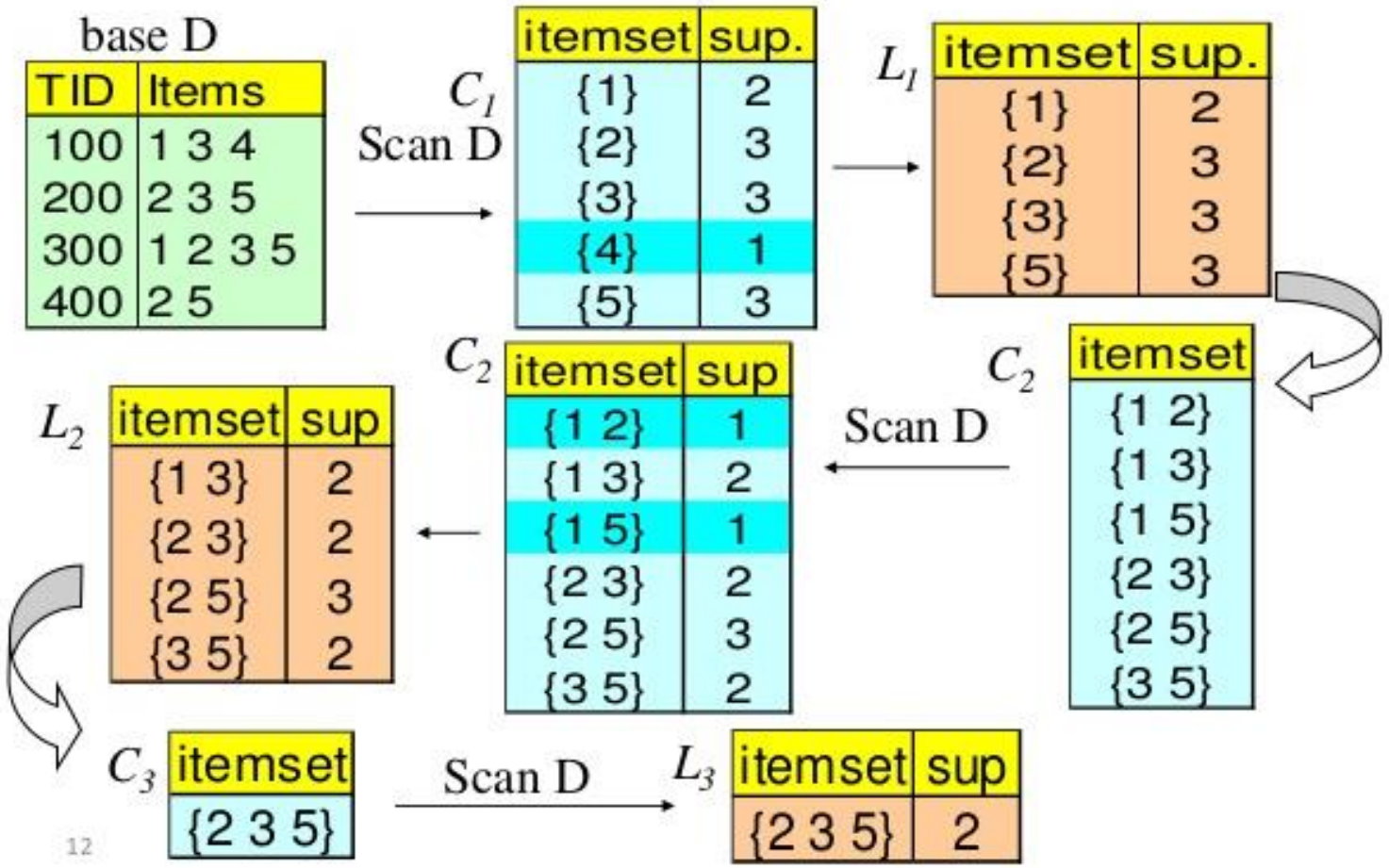
1. Comptez les occurrences de chaque valeurs/items.
2. Générer les candidats :
   - Générer les itemsets de taille k ;
   - Supprimer ceux dont le support est inférieur à la valeur fixée.
3. Garder les candidats qui ne sont pas déjà inclus dans un de taille (k + 1).



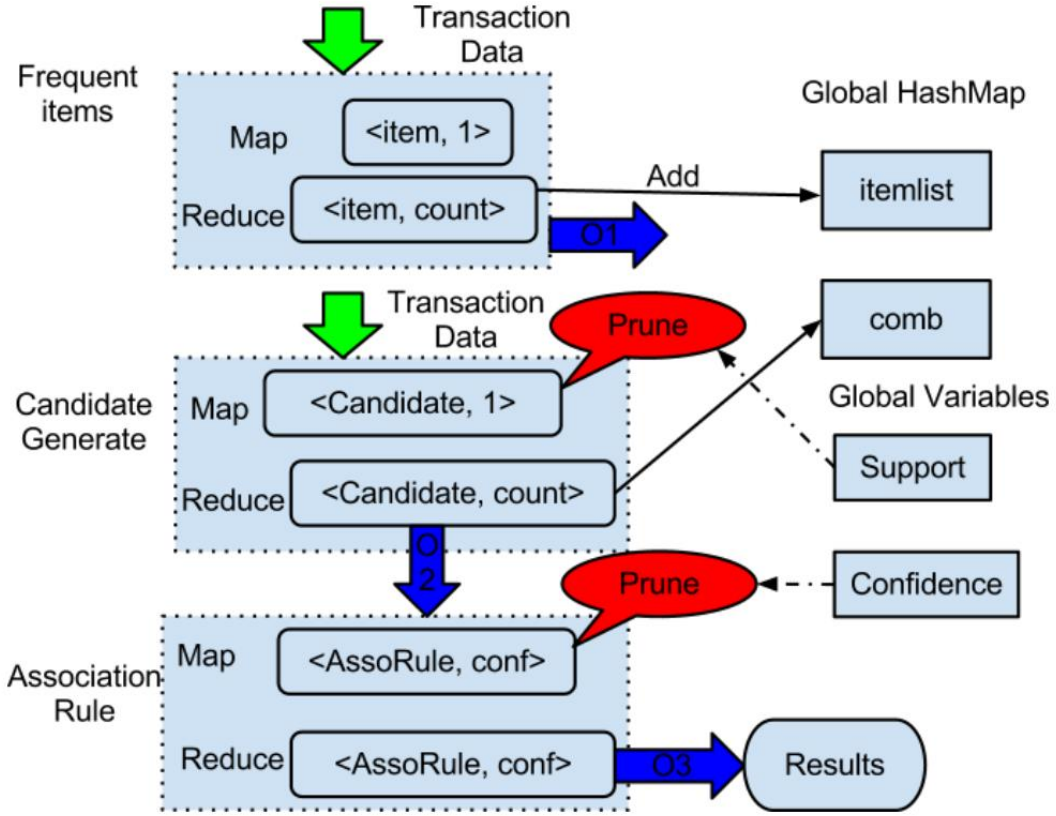
EXEMPLE D'APRIORI

Support minimum : 2

- Génération des candidats de taille k.
- Elimination des candidats dont le support est inférieur.
- Jusqu'à ce qu'on ne trouve plus de candidats pour une certaine taille n.



APRIORI PARALLÈLE



1. Chaque machine obtient une partie du fichier source.
2. Compte de la fréquence des items avant de mettre en commun dans l'itemlist.
3. Générer les candidats en parallèle sur les machines, qu'on met en commun à la fin.
4. On peut générer les règles d'associations.

GÉNÉRATION DES RÈGLES

En utilisant les candidats fournis par Apriori ou/et Apriori Parallèle :

- On peut définir les items représentatif ;
- qui servent ensuite à générer les règles d'associations.

{deadpool} → {sara rambles, wade wilson} 75%

EXEMPLE DE RÉSULTAT

#deadpool

april19th bryansinger deadpool fox hughjackman jenniferlawrence jlaw marvel wolverine xmen xmenfranchise xmenmovies xmennewcovers xnews

On peut donc vous conseiller de regarder les films Wolverine, Xmen et les films du studio Marvel.

ENJOY !

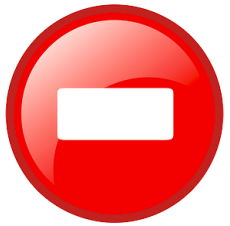
RÉSULTATS OBTENUS

Taille Input – Support Min	Apriori	Apriori Parallèle
100 – 3	42 sec	22 sec
100 – 4	0 sec	8,5 sec
500 – 3	3 min 34 sec	47 sec
500 – 4	6 sec	19,5 sec

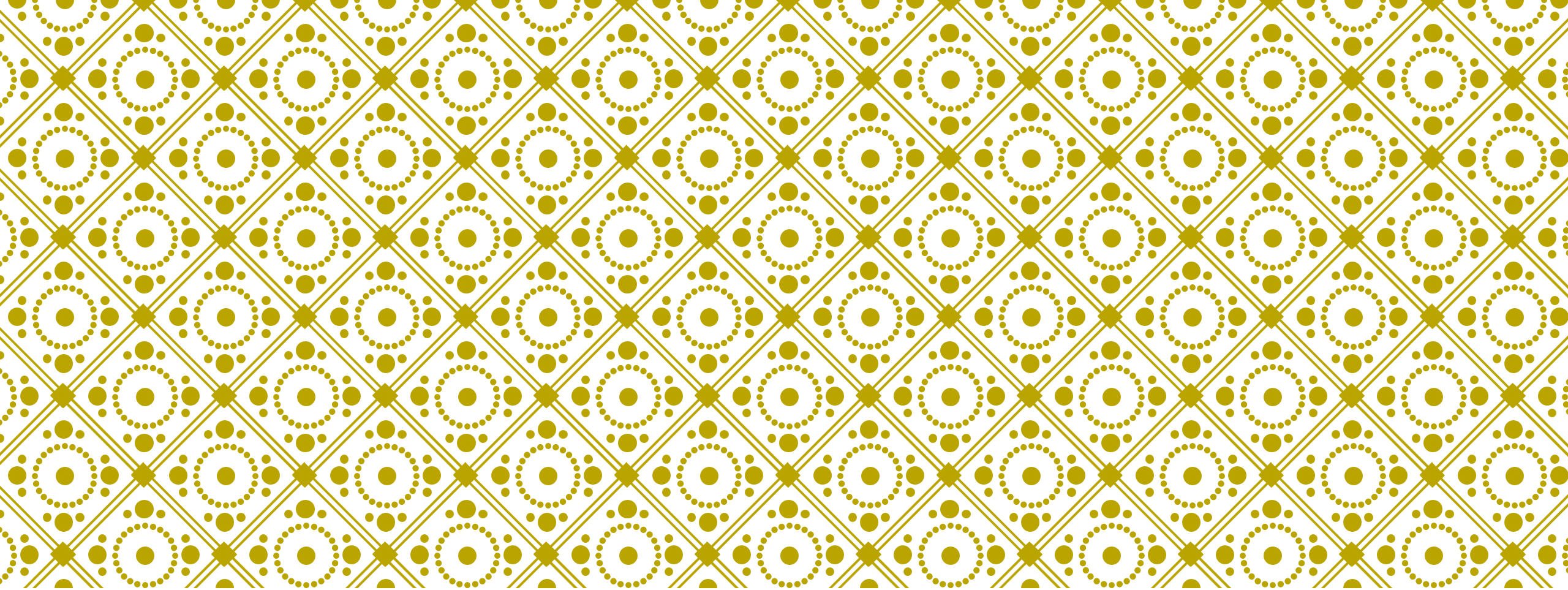
- Meilleures performances de l'**Apriori Parallèle** lorsqu'il y a un gros volume de données à traiter.
- L'**Apriori Standard** est plus efficace sur un plus petit volume de données.

CONCLUSION

- **Comparaison** entre Apriori et sa version parallélisée.
- Découverte de l'environnement **Hadoop**.



- **Manque de puissance** pour exécuter nos algorithmes sur des grosses bases de données.



QUESTIONS ?

Merci de votre attention !

SOURCES

- T. Hardes, Association Rule Mining ; http://thardes.de/big-data-englisch/association-rule-mining/
- S. Zhao and R. Du, Distributed Apriori in Hadoop MapReduce Framework ;
- S.A. Itkar and U.V. Kulkarni, Distributed Algorithm for Frequent Pattern Mining using HadoopMap Reduce Framework ;
- R. Agrawal and J.C. Shafer, Parallel Mining of Association Rules.