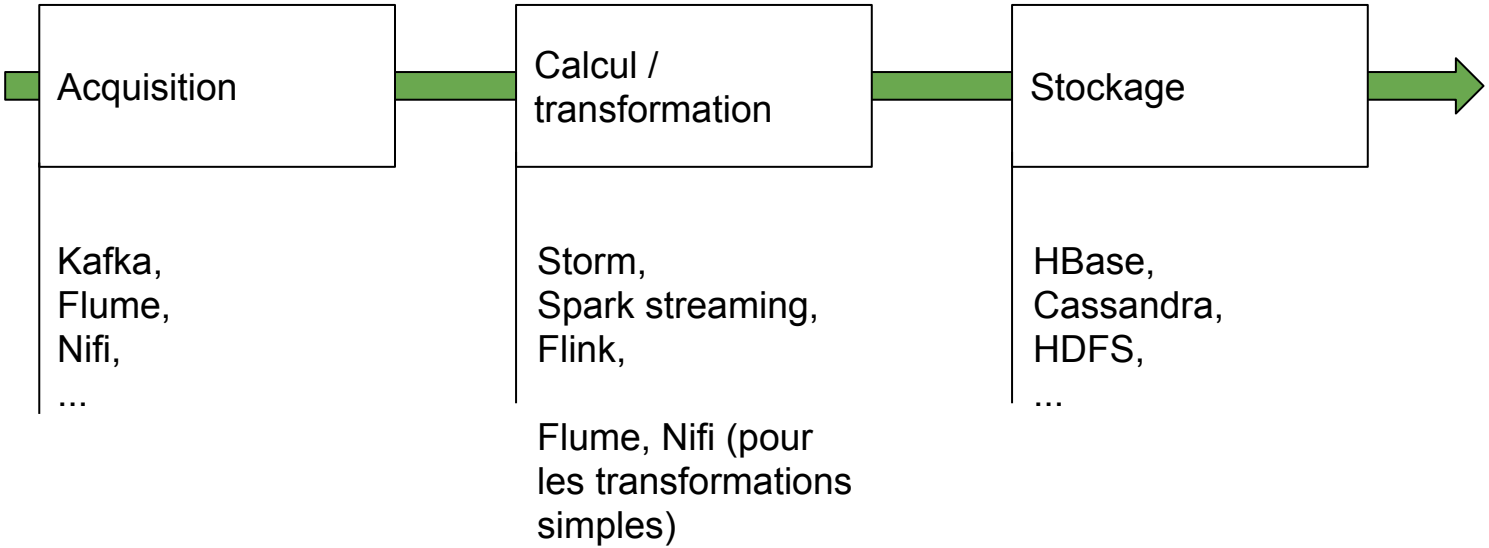
Hadoop & stream processing

Cours 2016-2017

Motivation stream processing dans Hadoop

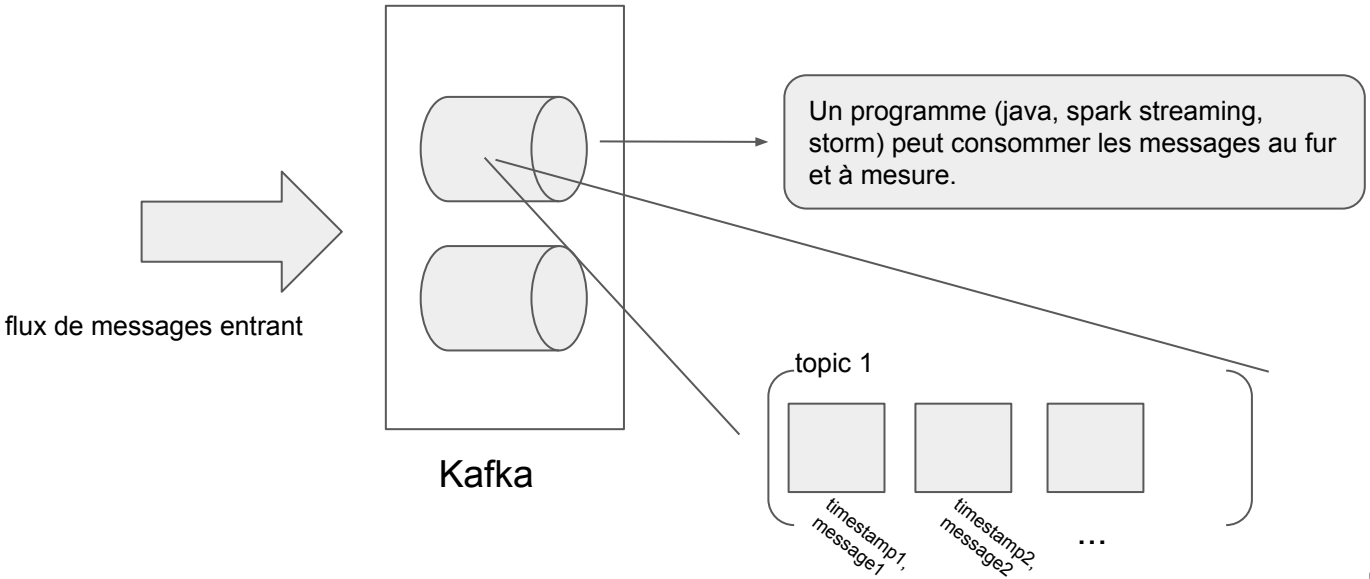
- **permet l'acquisition et le traitement au fil de l'eau de l'information**
- **transforme et analyse les données avant l'écriture les données sur disque**
- **exemple de projet stream processing dans Hadoop:**
  - suivi en temps réel des informations envoyées par des IOT
  - projet de détection d'attaques informatiques à partir de logs machines

Outils de gestion de flux



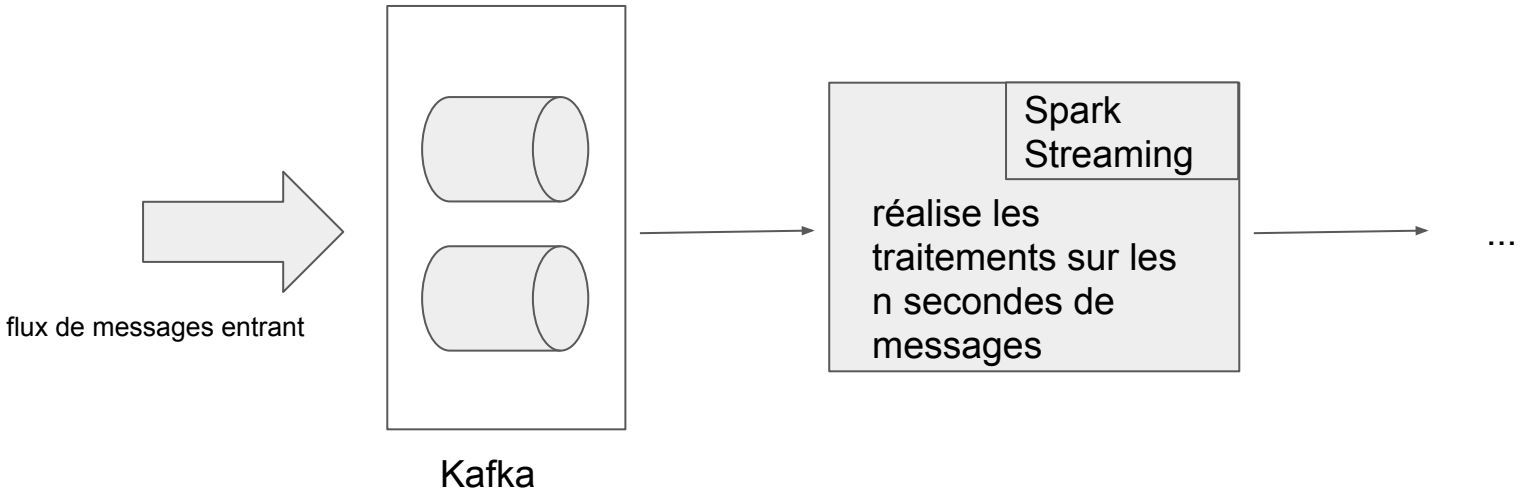
Kafka

- **Message Oriented Middelware, utilisé comme base de données tampon pour recueillir des flux**
- **Les flux sont séparés en “topics”**



Spark streaming

- **Spark : framework de calcul distribué, utilisation BATCH**
- **Spark streaming : Spark lancé toutes les n secondes avec les N données récupérées. MICROBATCH**



HBase

Définition : NoSQL

- Base de données avec un schéma de stockage non tabulaire
- Différents types de bases NoSQL : documents, clé-valeur, colonne, graphes
- Modèles de données dits schema-less
- Scalables

HBase

Base de données NoSQL orientées colonnes

- utilise HDFS
- peut être géré par YARN

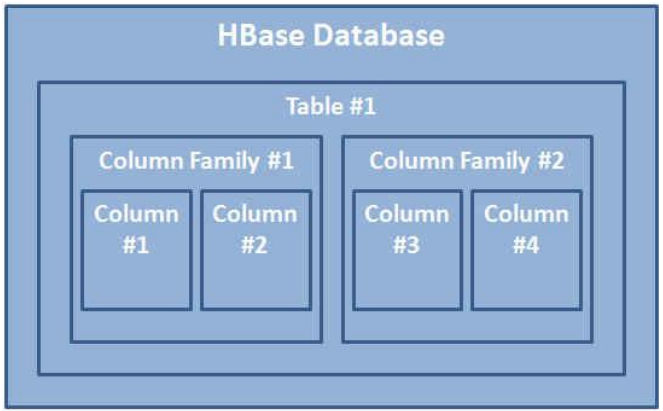
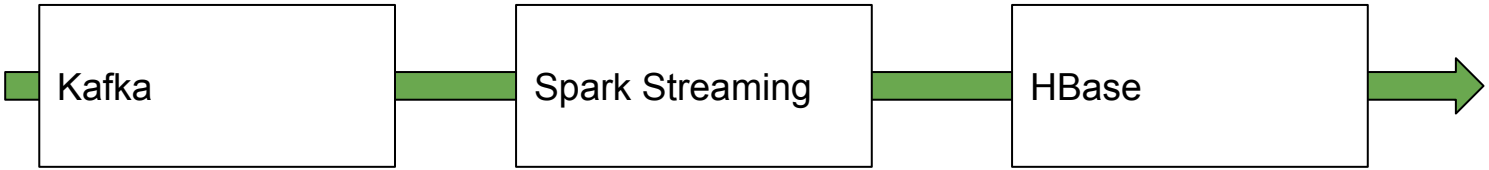


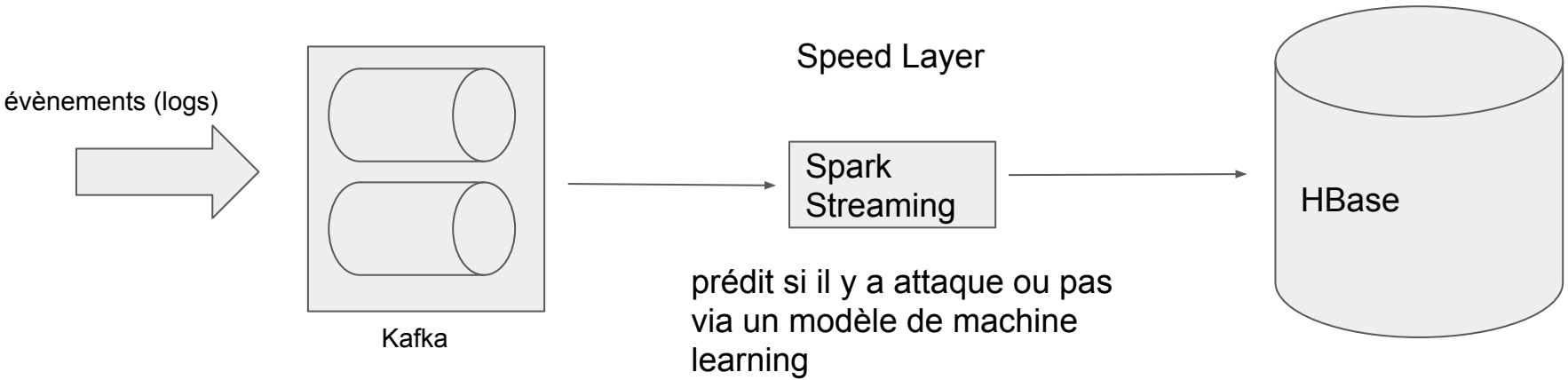
Figure 1 - HBase Data Organization

```
row-key => {
    cf1:column-key1: value1,
    cf1:column-key2: value2,        Column family cf1
    cf2:column-key3: value3,
    cf2:column-key1: value4,        Column family cf2
}
```

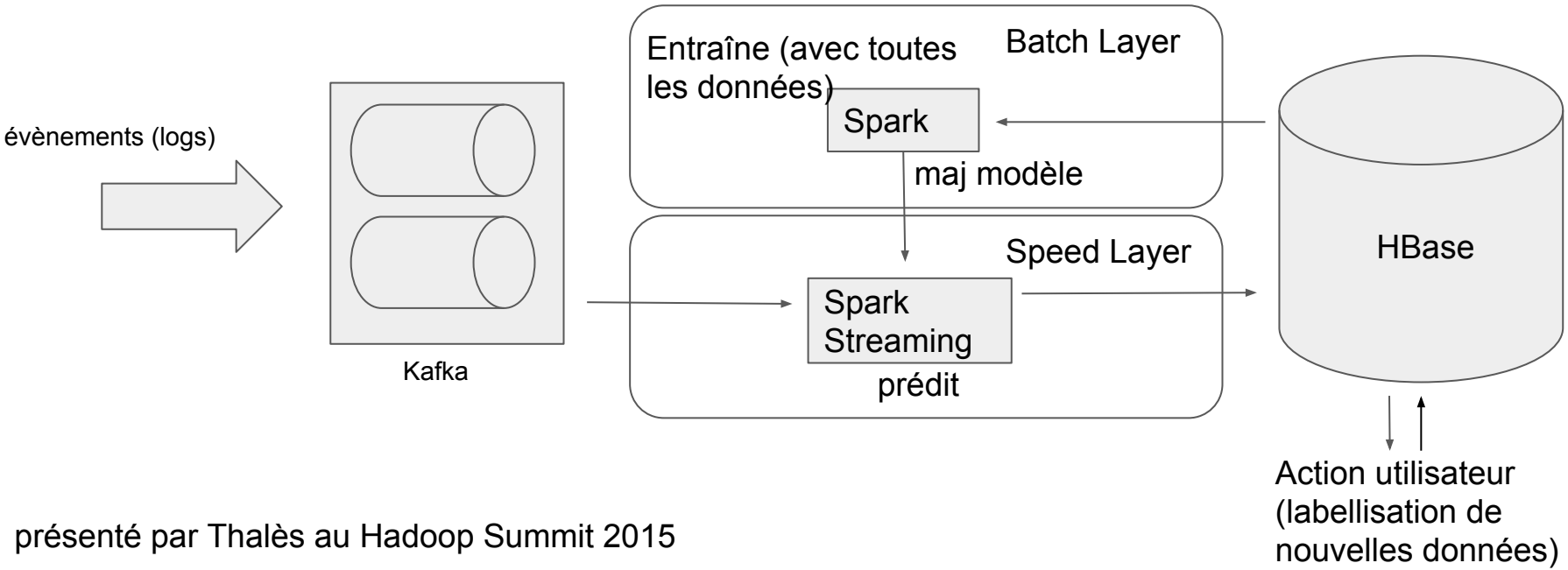
Exemple de gestion de flux



Exemple : détection d'attaques



Détection d'attaques : mise à jour du modèle via la lambda architecture



présenté par Thalès au Hadoop Summit 2015