

4. Gestion des données urbaines dans les nuages informatiques

- Brève histoire des nuages informatiques
- Modèles de service et de déploiement
- Technologie clé : la virtualisation
- IaaS : les points de vue utilisateur et fournisseur
- PaaS : programmation et déploiement des applications
- Stockage de données
- Traitement de données
- **Traitement de flux de données**

Défis du traitement de flux de données

- Caractéristiques des flux de données
 - Volume
 - Variabilité
 - Vitesse

Obtenir de l'information utile en (presque)
temps-réel à partir de flux de données
continus



Définition d'un flux de données

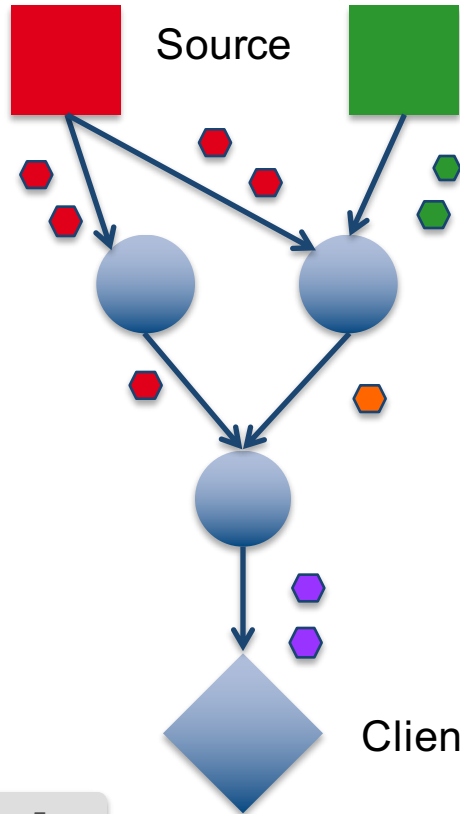
Un **flux de données** est une séquence (potentiellement) infinie de **tuples**.

<t1, t2, t3, ... >

Un **tuple** décrit un événement et est émis par une **source**.

(type, date, données spécifiques)

Description et structure d'une application de traitement de flux de données



Un traitement de flux de données est défini comme un graphe acyclique d'**éléments de calcul**.

Requêtes continues

Éléments de calcul et fenêtres

- **Exemples**

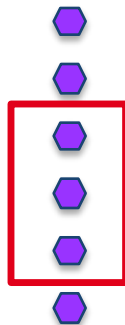
- Filtre, map, agrégation

Élément de calcul

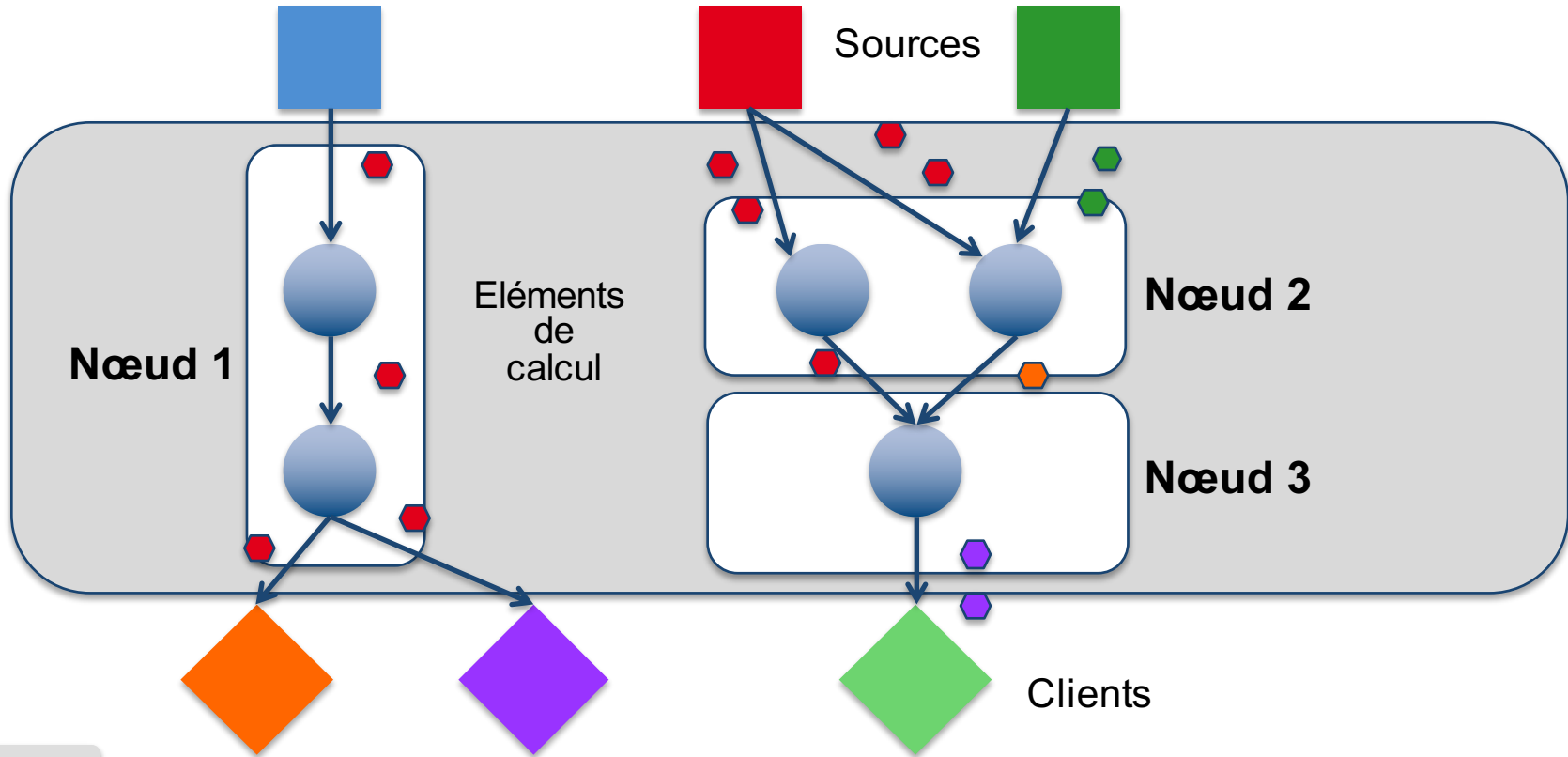
- Applique un traitement sur un tuple ou une **fenêtre** de tuples
- Est **avec** ou **sans état**

Fenêtre

- Portion finie d'un flux infini
 - Plage de temps
 - Nombre de tuples



Architecture d'un environnement de traitement de flux de données



Qualité de service

- **Traitement des situations de surcharge**
 - Une partie des données n'est pas traitée
 - ✓ Aléatoirement
 - ✓ Selon une politique de priorité
 - ✓ Selon l'impact sur la qualité de service
 - Elimination ou stockage en vue d'un traitement différé
- **Traitement des défaillances**
 - Défaillance d'éléments de traitement
 - Perte de données

Traitement de flux de données dans les clouds

- **Personnalisation de la configuration** pour les opérateurs
- Possibilités d'**adaptation** étendues

Adaptation de l'application

- Nombre d'instances de machine virtuelle
- Quantité de ressources de la machine virtuelle
- **Adaptation automatique** (*auto-scaling*)

Gestion dynamique des ressources par le fournisseur de cloud

Migration de machine virtuelle

Logiciels de traitement de flux de données

Grands fournisseurs de clouds

Kinesis, MillWheel, Azure stream, IBM streams



Grands acteurs du traitement de données

- Storm
- Mantis



Communautés open source

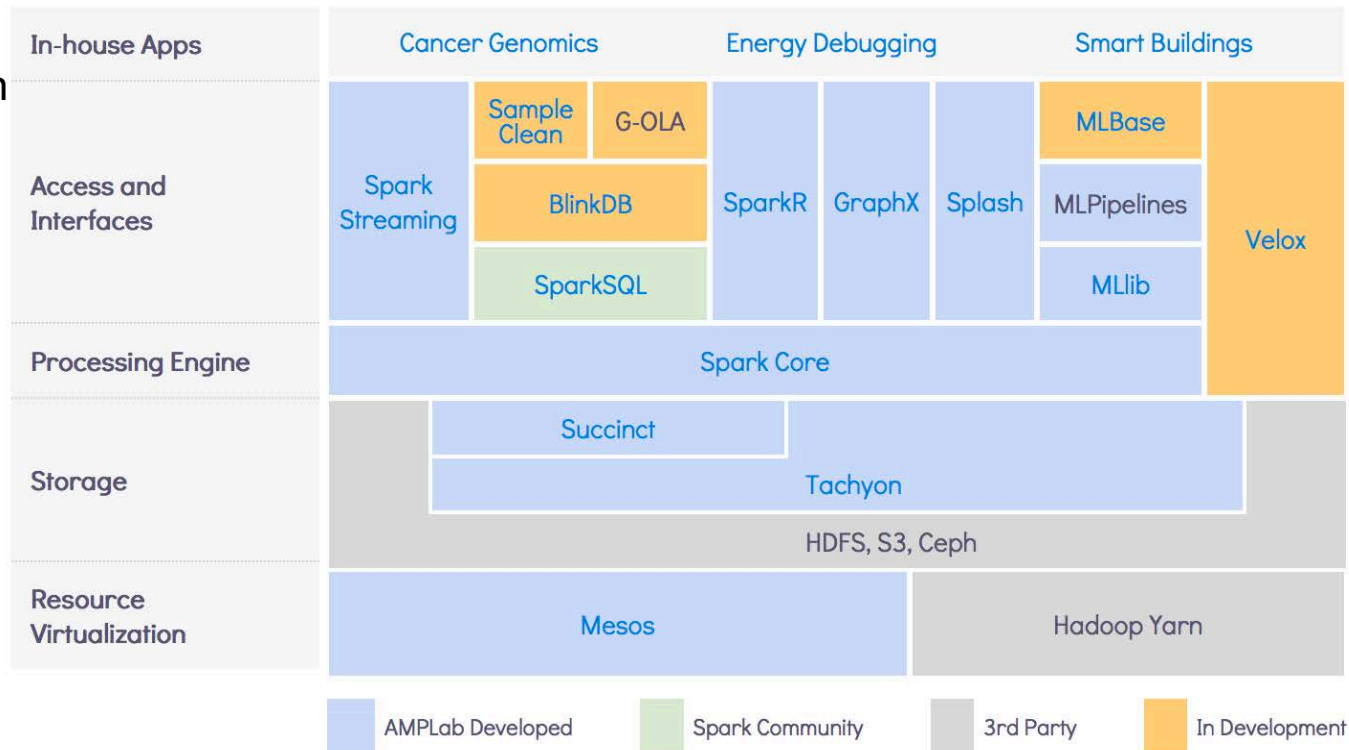
- S4, Samza (Apache)
- Spark streaming (UC Berkeley)
- Heka (Mozilla)

Tendances récentes du traitement de flux de données dans le cloud

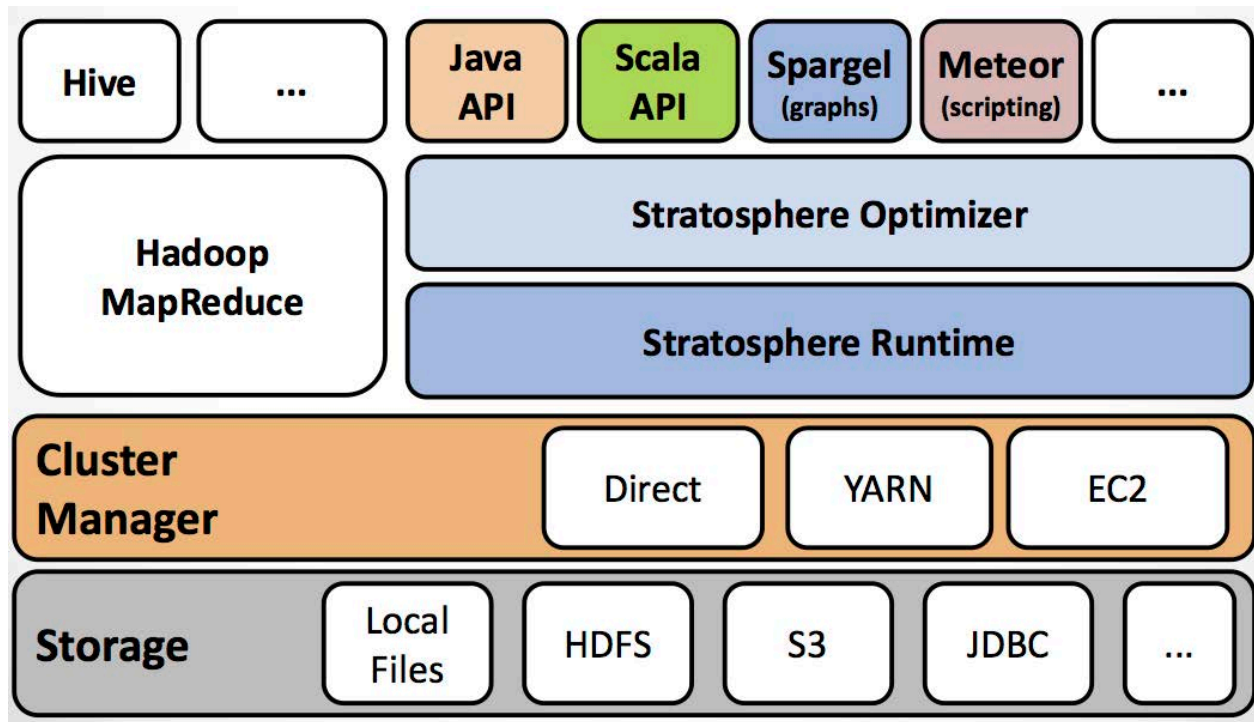
- *Stream (Data) processing as a Service* (DaaS)
 - Contrat de service
- **Optimisation de l'utilisation des ressources** dans les clouds IaaS
 - Grand nombre de requêtes et de flux au comportement imprévisible
- **Optimisation du coût économique** des traitements de flux de données
- **Traitement combiné** de données stockées sur disque et de flux de données

Un exemple de pile logicielle de traitement de données : Berkeley Data Analytics Stack (BDAS)

Projet américain
mené par
UC Berkeley



Un exemple de pile logicielle de traitement de données : Stratosphere



Projet européen mené par
TU Berlin,
Humboldt University et
Hasso Plattner Institute

Un grand défi du cloud pour la ville intelligente

Traitement de données massives en temps-réel
dans le contexte de l'**Internet des objets**
pour fournir efficacement des **informations pertinentes**
à de très **nombreux utilisateurs mobiles**
en fonction de **leur localisation** et du **contexte**



Illustrations & photos : crédits

p. 2 : © Julien Eichinger, Fotolia ; By Camelia.boban CC BY-SA 3.0, via Wikimedia Commons ; Domaine public, Pixabay.

p. 3 : Domaine public, Pixabay.

p. 13 : droits réservés, projet Berkeley Data Analytics Stack, <https://amplab.cs.berkeley.edu/software/>

p. 14 : droits réservés, projet européen Stratosphere piloté par TU Berlin, Humboldt University et Hasso Plattner Institute

p.15 : © weedeesign, Shutterstock