Festival Online de la Data

Synaltic

PROCHAINES DATES

Jeudi 4 juin de 15h à 15h30

Debezium, CDC Open Source en temps réel

Yabir Canario de la Mota, consultant #Synalteam

Jeudi 11 Juin de 15h à 15h30

Knowage, toute la BI en Open Source

Stefano Scamuzzo, Engineering

Charly Clairmont, Synaltic

Jeudi 18 Juin de 15h à 15h30

Préparation de Données rapide et agile dans les Clouds AWS, Azure et Google

Victor Coustonoble, Trifacta

Franck Nguyen, Synaltic

Jeudi 25 juin de 15h30 à 16h

Comment mobiliser business et IT autour de leurs données les plus précieuses

Jean-Michel Franco, Talend

Romain Legrand, Synaltic

SYNALTIC EN QUELQUES MOTS

Acteur innovant, dénicheur de solutions et une équipe de collaborateurs engagés

Synaltic

30

30 collaborateurs formés et certifiés contribuant aux communautés d'utilisateurs.

15

Spécialiste en Data Management depuis plus de 15 ans et plus de **250 projets** réalisés.

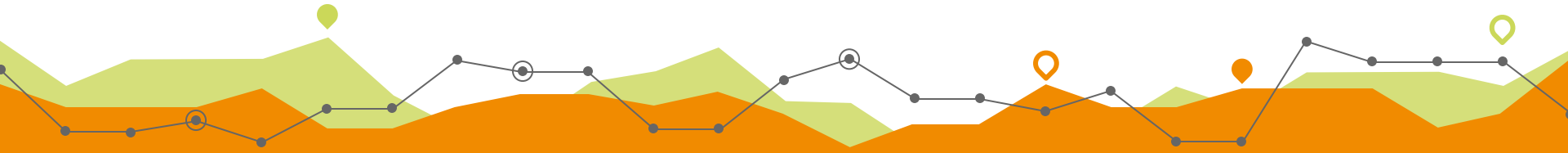
160

Plus de **160 clients** dont certains s'appuient sur nous avec succès depuis plus de **8 ans**.

Experts en Data Management, passionnés d'Open Source et d'Innovation !

Synaltic

Streaming processing



De la base de données classique au streaming avec Debezium

Juin 2020
Yabir CANARIO DE LA MOTA

Yabir CANARIO DE LA MOTA

Data Engineer

Sommaire

Le contexte

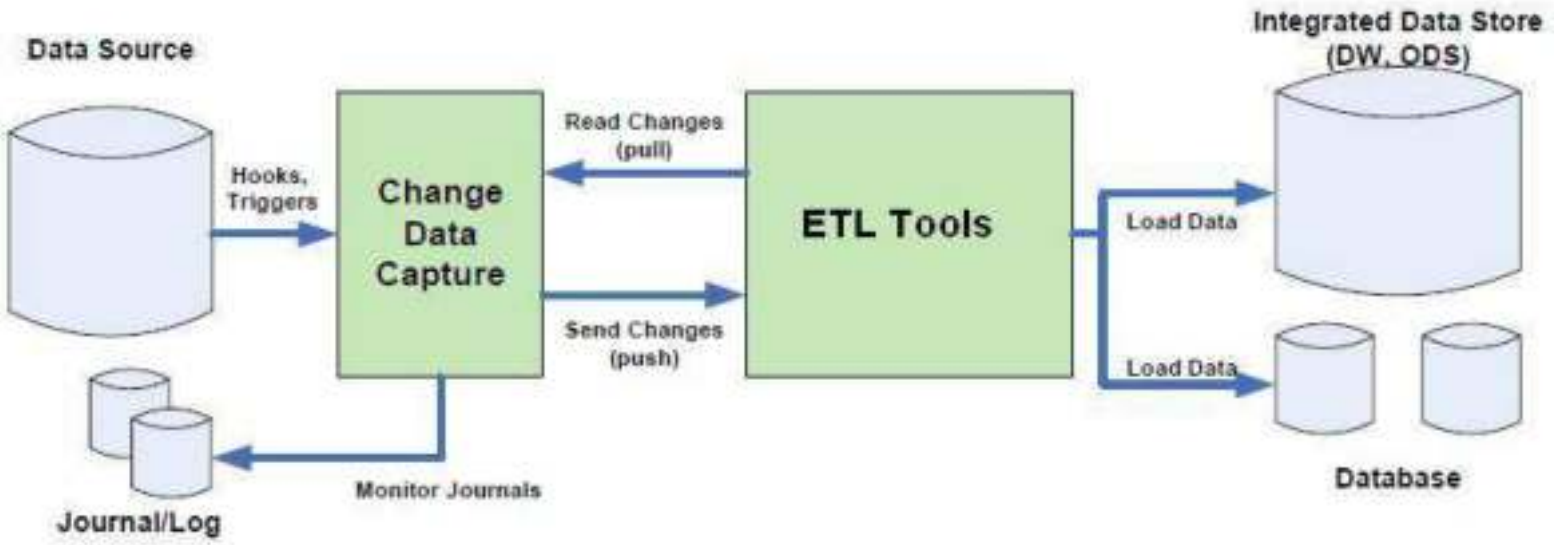
Debezium

Démo

Cas d’utilisation

Fonctionnement

Le contexte : Change Data Capture (CDC)



Le contexte : Event Driven

mai 2019

CDC par trigger

- impact sur les performances
- pas de solution générique

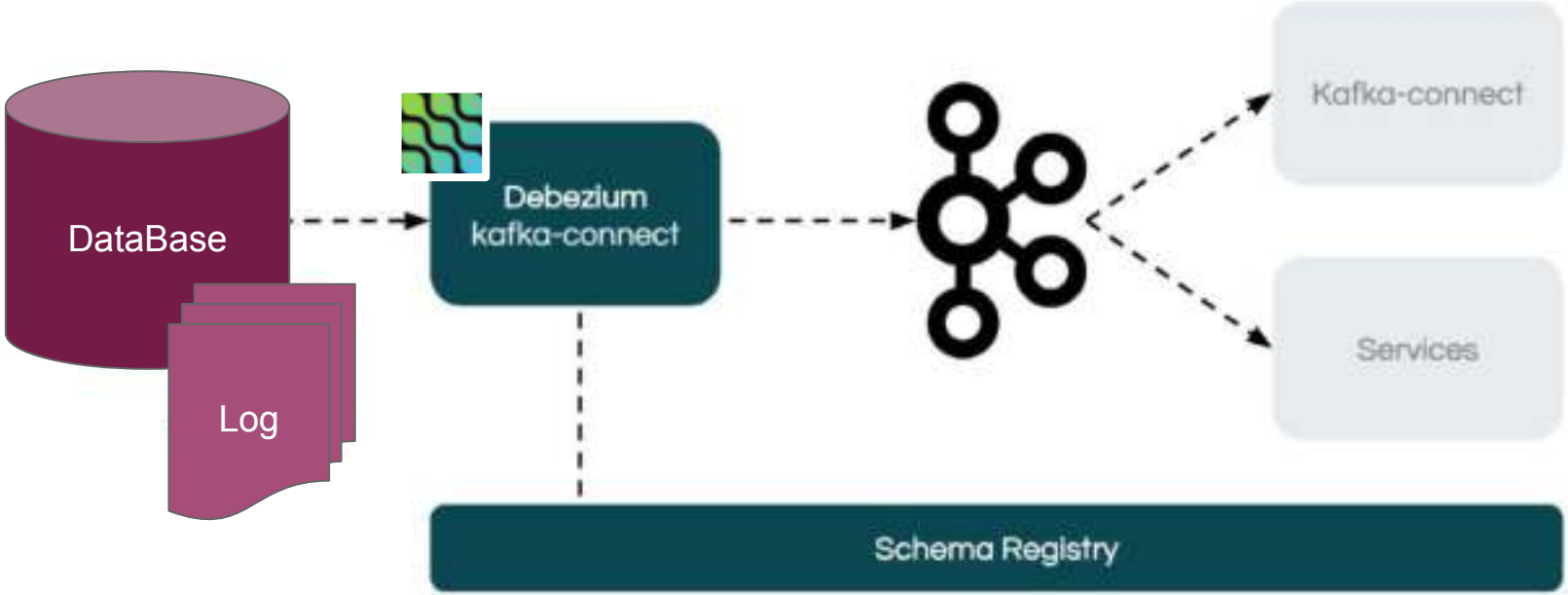
CDC par lecture du log

- pas de reprise sur erreur
- pas de solution générique

CDC par Debezium

Sortie de Debezium
(licence Apache)

Qu'est-ce que Debezium ?



Ce que nous apporte Debezium

- Simplifie la mise en place du CDC
- Accélère la propagation des changements
- Enrichi les événements (insertion, update, ...)
- DB Agnostique
- Tolérant à la panne
- Du Batch au Streaming

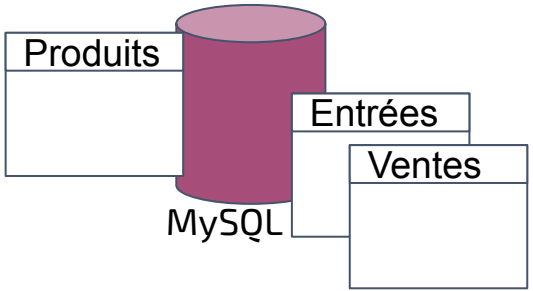
https://debezium.io/

Démonstration

E-commerce

Synaltic

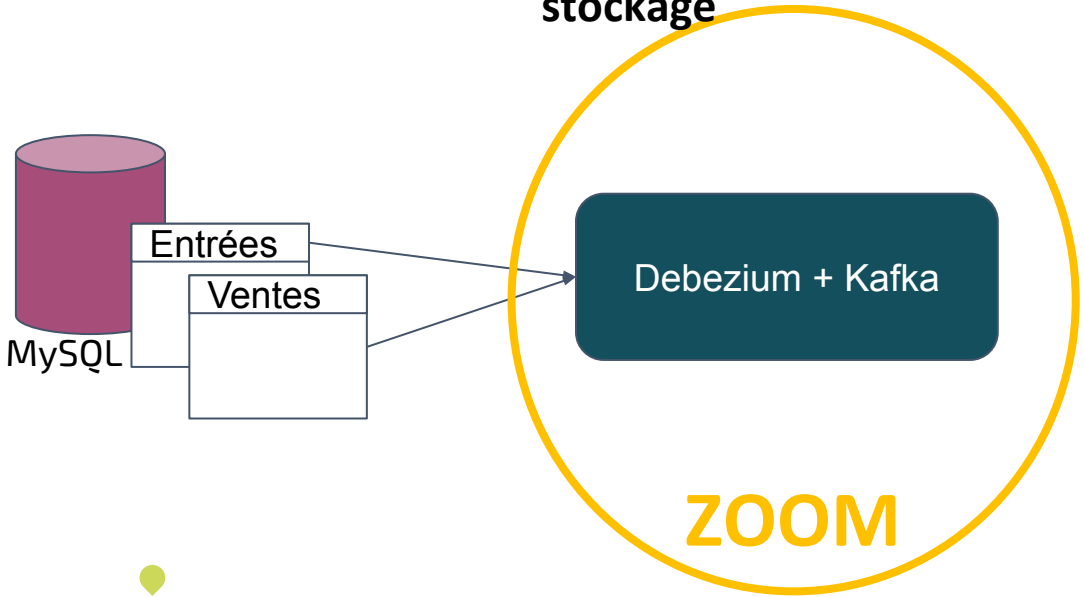
Input



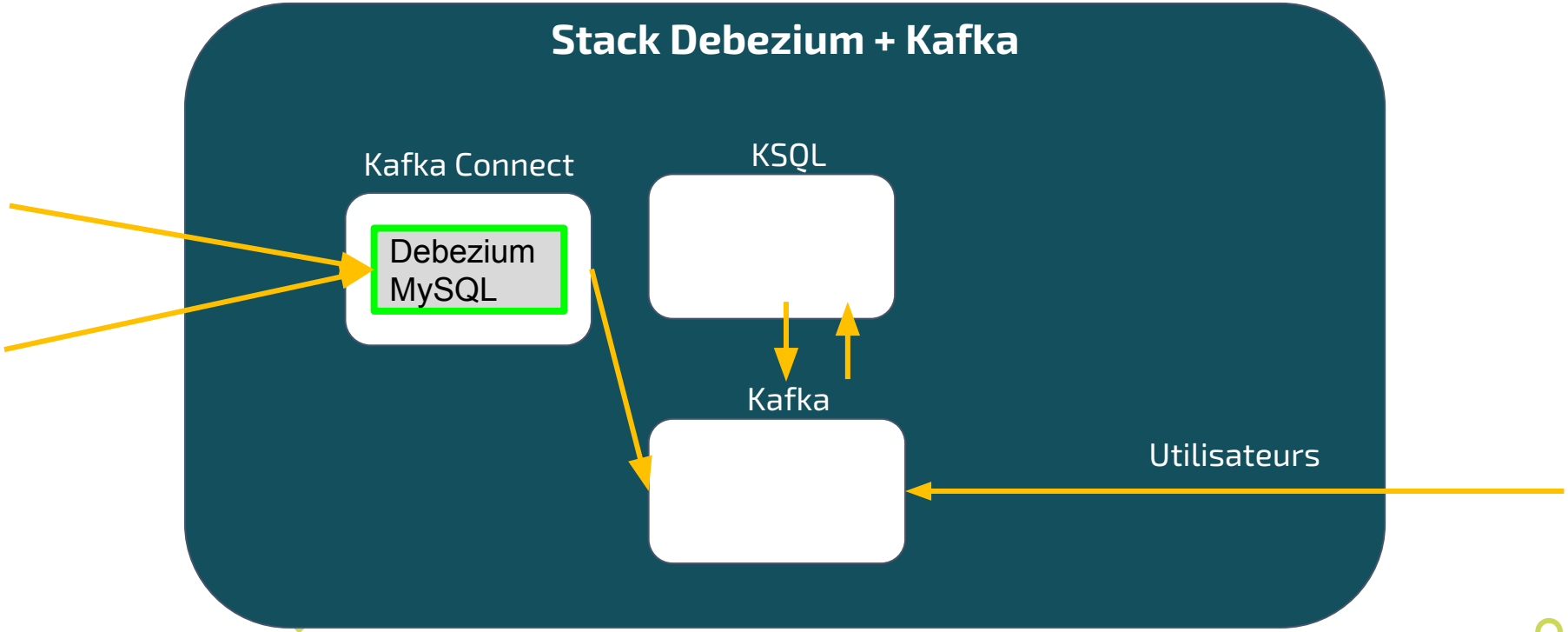
E-commerce

Input

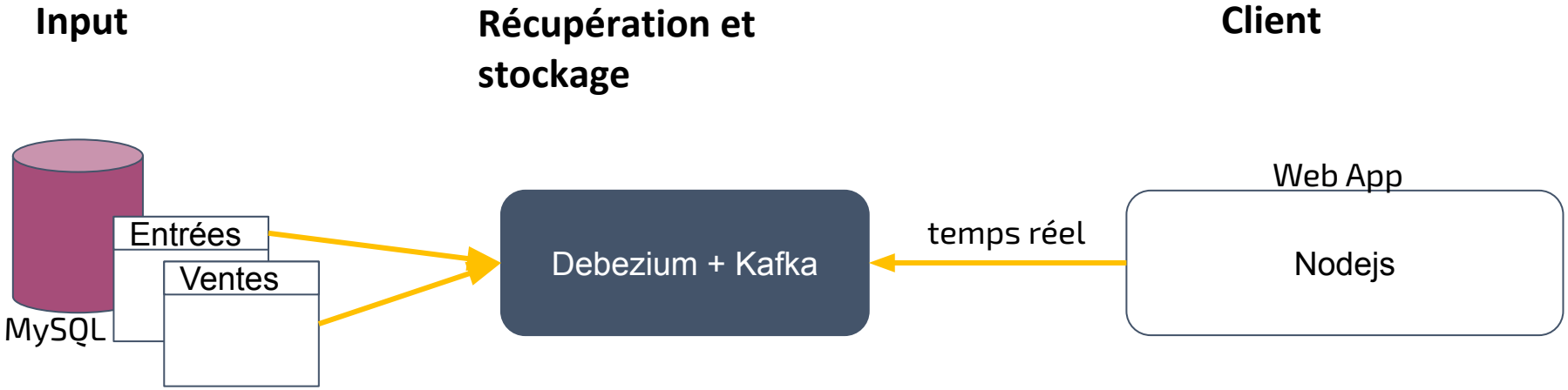
Récupération et stockage



E-commerce


Stack Debezium + Kafka
Kafka Connect
Debezium
MySQL
KSQL
Kafka
Utilisateurs

E-commerce


Input
Récupération et stockage
Client
Entrées
Ventes
MySQL
Debezium + Kafka
temps réel
Web App
Nodejs

DEMO

https://ecommerce.dev.synaltic.app

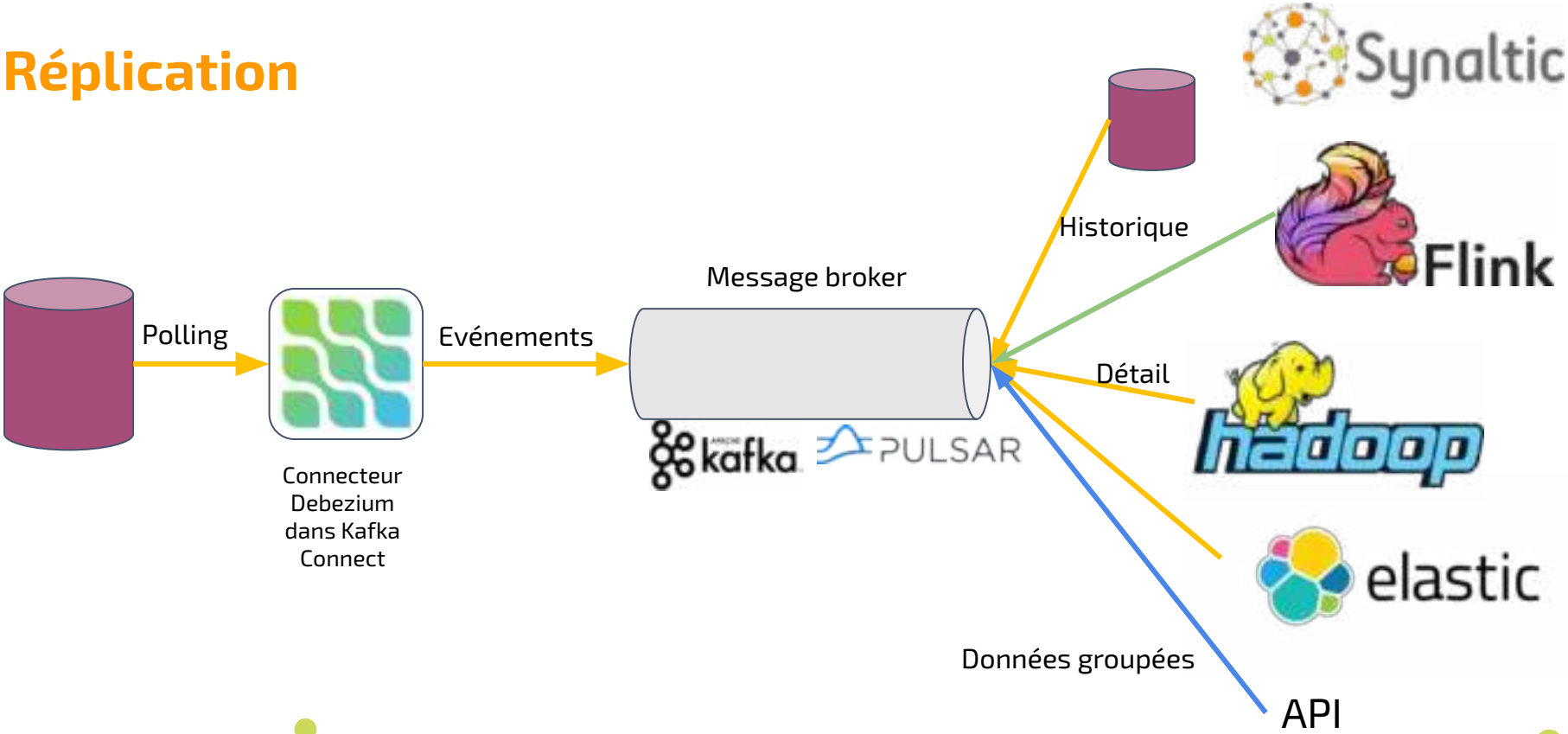
E-commerce

KSQL

```
CREATE STREAM entrees_stream … WITH (KAFKA_TOPIC='dbserver1.synalticdb.entrees,
VALUE_FORMAT='AVRO');

CREATE STREAM ventes_stream … WITH (KAFKA_TOPIC='dbserver1.synalticdb.ventes',
VALUE_FORMAT='AVRO');

CREATE STREAM S_UNION_ENTREES_VENTES WITH (PARTITIONS=1) AS \
SELECT 'STOCK' as SRC, * FROM entrees_stream_key;
INSERT INTO S_UNION_ENTREES_VENTES \
SELECT 'VENTES' AS SRC, * FROM ventes_stream_key;

CREATE TABLE T_STOCKS AS \
        SELECT id_produit as idproduit, SUM(quantite) as quantite \
        FROM S_UNION_STOCK_VENTES \
        WITH (KAFKA_TOPIC=T_STOCKS, VALUE_FORMAT='delimited', KEY='idproduit') \
        GROUP BY id_produit;

CREATE TABLE T_EVENTS AS \
   SELECT ..., sum(quantite) as quantite, count(*) as count \
   FROM ...
   WINDOW TUMBLING (SIZE 60 SECONDS) \
   GROUP BY id_produit, client, SRC;

CREATE TABLE T_ALERTS AS \
   SELECT ..., sum(quantite) as quantite, count(*) as count \
   FROM ...
   GROUP BY client, SRC \
   HAVING sum(quantite) > 50 OR sum(quantite) < -50;
```

Uses cases

Réplication



D'autres Use Cases ...

- Audit / Historisation
- Mise à jour full-text search via Elasticsearch, Solr, etc
- Mise à jour CQRS modèle de lecture
- Mise à jour de la UI en temps réel
- Pouvoir requêter les données en streaming

Comment suivre tous les changements ?

Synaltic

Monitoring de la base de données

Applications écrivant à la base de données
-> les modifications sont stockées dans le fichier log
-> les tables sont mises à jour

Utilisons le fichier log de la base de données

- MySQL -- Binary log / binlog
- Postgres -- write-ahead log
- MongoDB -- op log

Infrastructure du messaging

Kafka correspond bien aux objectifs CDC

- Messages identifiés par clé
- Ordre garanti
- Basée par pull request
- Support la compaction
- Scalabilité horizontale

Kafka Connect

- Framework pour connecteurs source et sink
- Suivi des offsets
- Support schema
- Clustering
- Variété de connecteurs

KSQL

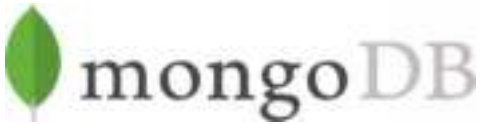
- Création de tables et streaming
- Jointures

Connecteurs CDC Debezium

Synaltic

- MySQL
- Postgres
- MongoDB
- SQL Server

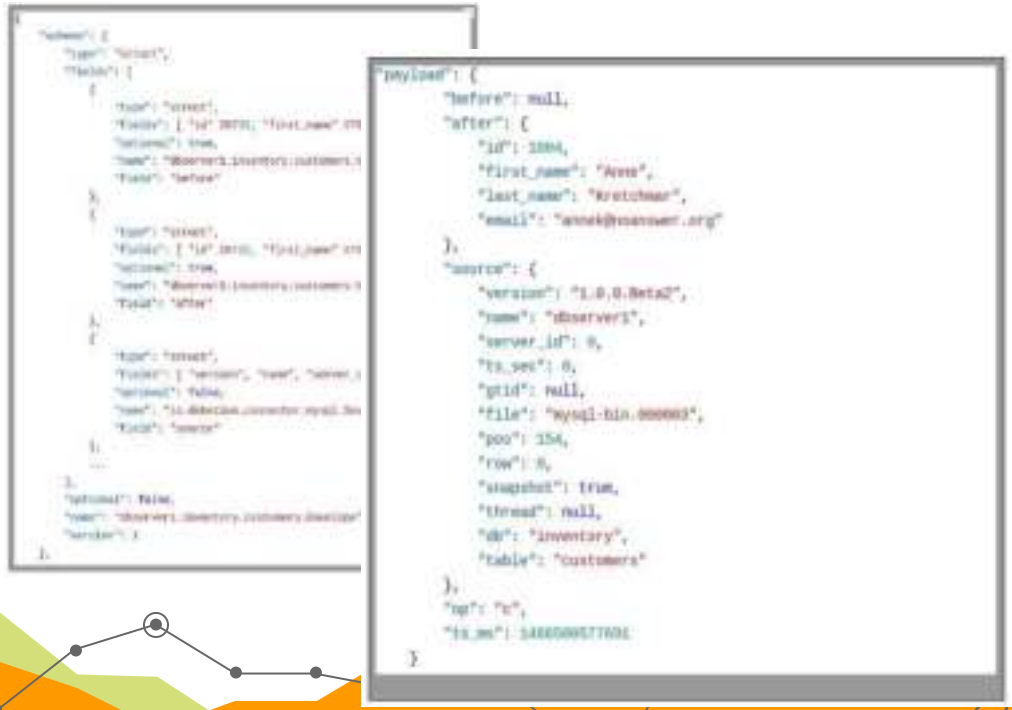
- Oracle (incubating)
- Cassandra (incubating)



Microsoft SQL Server

Structure de message

- Pair clé-valeur (dont clé = Primary key)
- Payload :
  - Before
  - After
  - Source
  - Operation
  - Date
- Types :
  - JSON
  - Avro



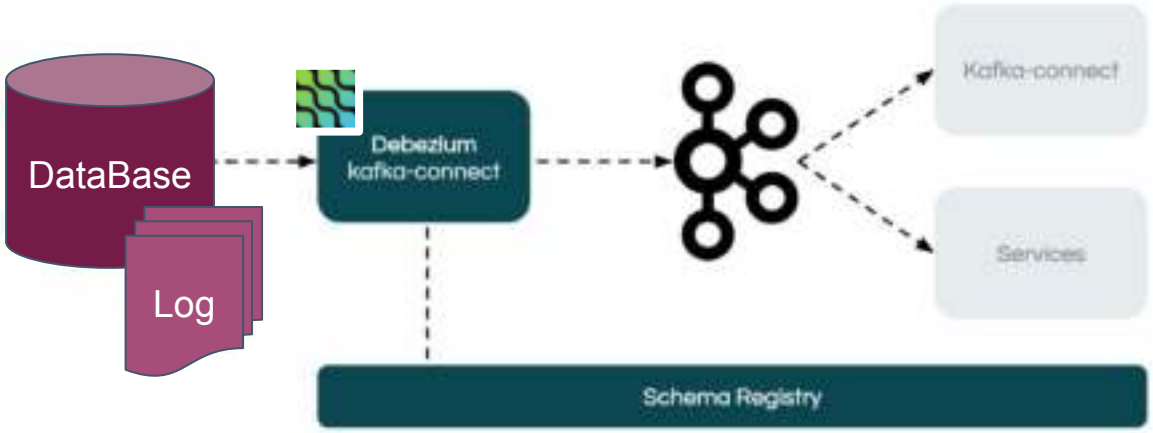
Fonctionnalités Debezium

- Standardisation des messages, peu importe la base de données source
- Garanti de la récupération de tous les changements de la base de données
- Produit des événements de changement avec un délai très faible (par exemple, plage de ms pour MySQL ou Postgres)
- Récupération de données supprimées
- Filters -> whitelist/blacklist

- Masking
- Snapshots
- Monitoring a travers JMX
  - Statut du Snapshot
  - Dernier event/snapshot
- Embedded Engine
- Transformation

Pour conclure, Debezium :

- Simplifie la mise en place du CDC
- Accélère la propagation des changements
- Enrichi les événements
- DB Agnostique
- Tolérant à la panne
- Du Batch au Streaming



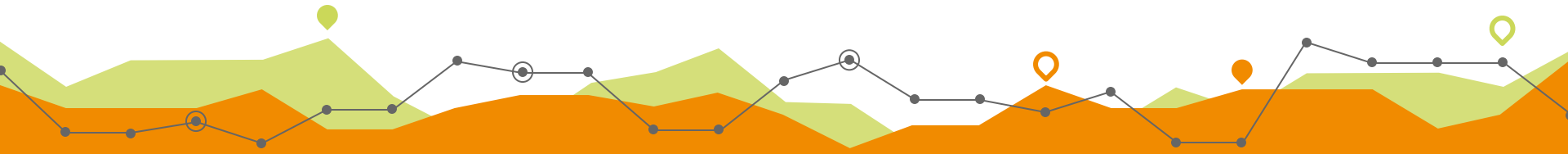
Pour aller plus loin !

Tutorat Debezium
https://debezium.io/documentation/reference/1.0/tutorial.html

Exemples docker-compose
https://github.com/debezium/debezium-examples

Démo Synaltic
https://ecommerce.dev.synaltic.app

Accompagnement Synaltic
https://www.synaltic.fr/



Questions ?

contact@synaltic.fr

Festival Online de la Data

Synaltic

PROCHAINES DATES

Jeudi 4 juin de 15h à 15h30

Debezium, CDC Open Source en temps réel

Yabir Canario de la Mota, consultant #Synalteam

Jeudi 11 Juin de 15h à 15h30

Knowage, toute la BI en Open Source

Stefano Scamuzzo, Engineering

Charly Clairmont, Synaltic

Jeudi 18 Juin de 15h à 15h30

Préparation de Données rapide et agile dans les Clouds AWS, Azure et Google

Victor Coustonoble, Trifacta

Franck Nguyen, Synaltic

Jeudi 25 juin de 15h30 à 16h

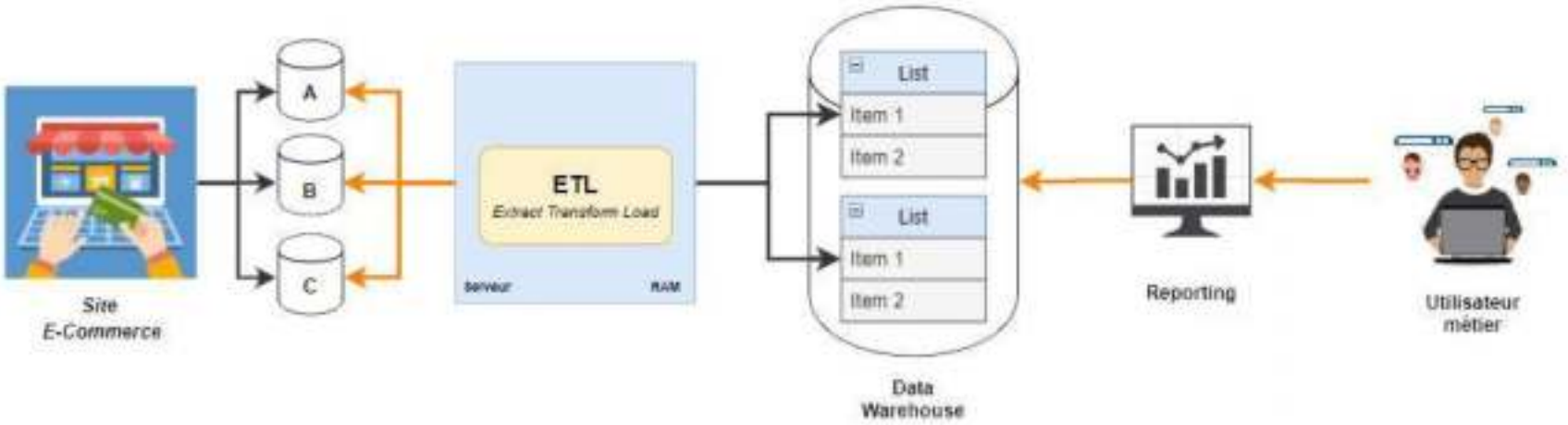
Comment mobiliser business et IT autour de leurs données les plus précieuses

Jean-Michel Franco, Talend

Romain Legrand, Synaltic

Annexes

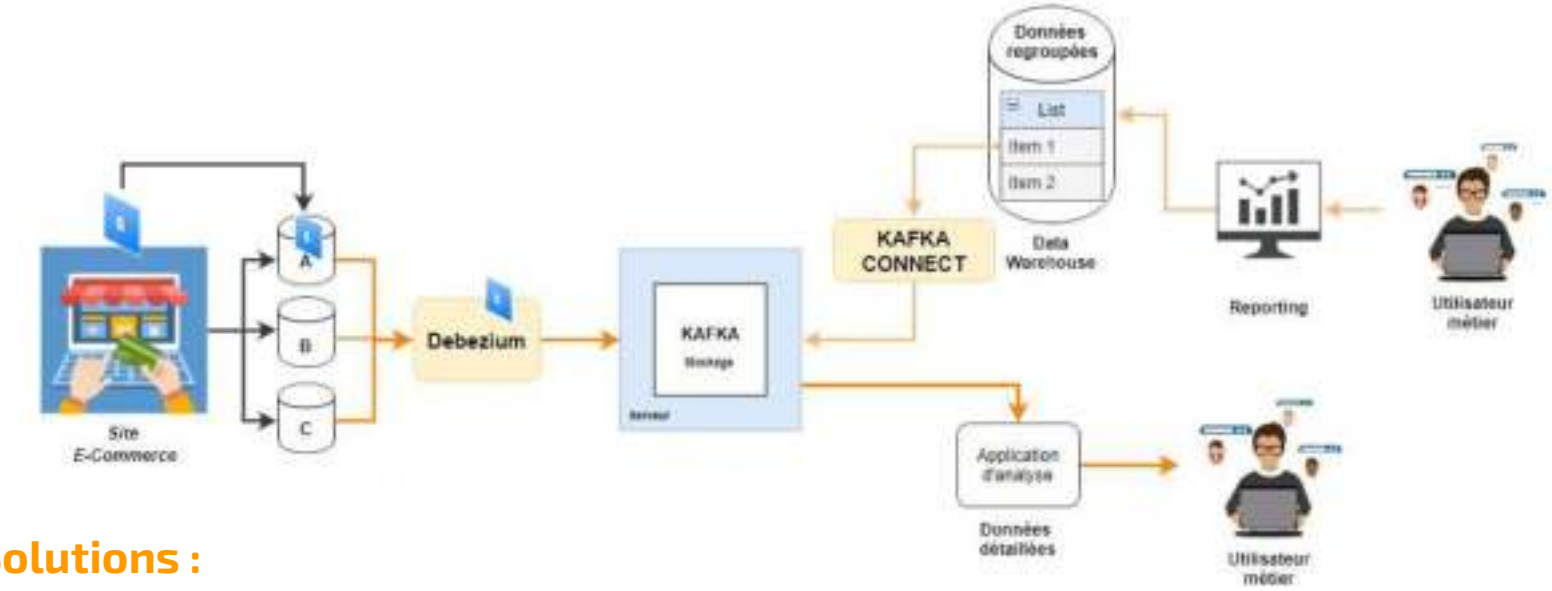
Flux orienté ETL



Inconvénients :

- **Batch :** récupération une fois par jour
- **Dépendant d'un développeur :** nécessite un développement important pour récupérer des données
- **Non scalable :** ne répond pas à une demande utilisateur métier rapidement

Flux orienté Streaming avec Debezium



Solutions :

- **Streaming :** en temps réel
- **Scalabilité :** mémoire des données lus en cas de panne
- **Pas de développement particulier :** l'utilisateur final récupère l'information qu'il a besoin rapidement