

Festival Online de la Data



Synaltic

PROCHAINES DATES

Jeudi 18 Juin de 15h à 15h30

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

Victor Coustenoble, 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

Jeudi 9 Juillet de 15h à 16h

Pourquoi utiliser la DataViz pour analyser ses données ?

Franck Nguyen, Synaltic

SYNALTIC EN QUELQUES MOTS

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



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 !

TRIFACTA & SYNALTIC

Plus de 5 années de collaboration



Trifacta est le pionnier de l'industrie et le leader du marché mondial des technologies de préparation des données.



Offre un software qui aide les individus et les organisations à explorer, transformer et rassembler efficacement des données variées.



Apporte la couche "ETL" et son industrialisation du point de vue des métiers. Il s'agit aussi de permettre aux DSI de mieux suivre l'exploitation des données par les métiers.

Synaltic soutient Trifacta pour sa plateforme de préparation de données pour les métiers et les spécialistes de la donnée. Notre volonté est d'accompagner nos clients dans leur besoin d'exploitation de gros volumes de données à des fins d'analyse.





Préparation de Données dans les Clouds AWS, Azure et Google

18/06/2020

Franck Nguyen - Synaltic
Solution Architect
fnguyen@synaltic.fr

Victor Coustenoble - Trifacta
Director of Sales Engineering EMEA
vcoustenoble@trifacta.com

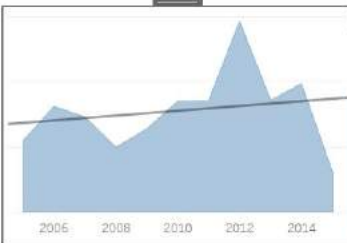
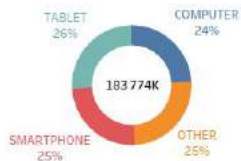
Agenda

- Besoin de Préparation de Données ?
- A propos de Trifacta
- Préparation de données dans les Clouds
 - Architectures
 - Démonstrations

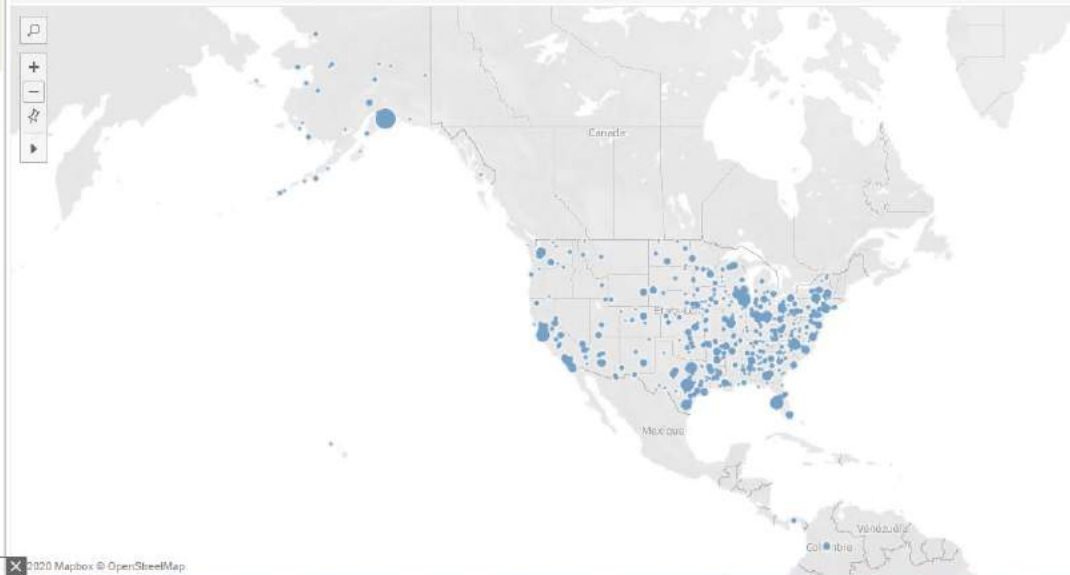
A quoi nous servent les données ?

Suivi des ventes

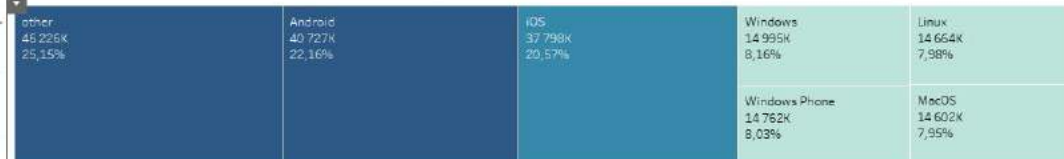
CA Total
183 774K



Répartition des clients



Répartition des clients par OS



Top Clients

Courtney Chairat	552K
Zia Richard	535K
Tarik Odonnell	477K
Ivy Fulton	437K
Nerea Gilbert	436K
Adam Oneal	429K
Maisie Neal	426K
Kasimir Cardenas	416K
Jackson Cox	407K

Top Produits

1105142400000-70	55K
1234137600000-728	55K
1218499200000-601	54K
1322265600000-555	52K
1134518400000-537	50K
1106956800000-578	50K
1154736000000-602	47K
1200700800000-421	46K
1139356800000-170	46K

Le chemin pour y arriver est parfois long et difficile... voir pénible !

OS Brand

```
IF CONTAINS([OS Clean], 'Windows') THEN 'Microsoft'  
ELSEIF [OS Clean] = 'MacOS' or [OS Clean] = 'iOS' then 'Apple'  
ELSEIF [OS Clean] = 'Android' then 'Google'  
ELSE [OS Clean]  
END
```

Dot Website

```
SPLIT(  
  SPLIT(  
    SPLIT([Url], '  
    , '.', -1)  
  )  
  , '.', -1)  
)
```

Product ID Clean

```
SPLIT([Url], 'pid=', -1)
```

Date Event Time Clean

```
DATETIME(  
  IF CONTAINS([Event Time], 'Nov') = FALSE THEN  
    SPLIT([Event Time], '/', 2)  
    + '/' +  
    SPLIT([Event Time], '/', 1)  
    + '/' +  
    SPLIT([Event Time], '/', -1)  
  ELSE [Event Time]  
END  
)
```

OS Clean

```
IF ISNULL([Os Id]) = TRUE then 'other'  
ELSEIF [Os Id] = 'Window' then 'Windows'  
ELSE [Os Id]  
END
```

Device Type Clean

```
UPPER (  
  IF ISNULL([Device Type]) = TRUE THEN 'other'  
  ELSE [Device Type]  
END  
)
```

Domain Name Clean

```
SPLIT(  
  SPLIT([Url], '.', 1)  
) , '///', -1
```

```
[customerfirstname] + ' ' + [customerlastname]
```

Des solutions qui posent un certain nombre de problèmes...

Ces outils ne sont pas fait pour cela !

- Difficultés à réaliser certaines opérations de nettoyage (fonctions manquantes, manipulation de code...)
- Impact sur les performances
- Difficultés d'industrialisation & de maintenance
- Problèmes de gouvernance & de partage

Le nettoyage et la préparation de données peut occuper jusqu'à 80% du temps d'un Data Analyst !

Data Platforms



Databases



Spreadsheets



Log Files



IoT Sensors



Apps

“The hardest part of AI is the **data wrangling.**”

—SWAMI SIVASUBRAMANIAN, VP AWS MACHINE LEARNING

#reInvent2018

80%

“It’s impossible to overstate this:
80% of the work in any data project
is in **cleaning the data.**”

— DJ Patil, Data Jujitsu, O’Reilly Media 2012

Analysis



Enterprise Data Warehouse



AI



Business Intelligence

Data Platforms



Databases



Spreadsheets



Log Files



IoT Sensors



Apps



- Donnez aux experts métiers les capacités d'une **exploration visuelle et interactive avec des recommandations de transformation automatiques** pour faciliter et accélérer la découverte et la préparation des données.
- Permettre au service informatique de **gérer et d'opérer de manière collaborative** les pipelines de transformations des données créés par des experts métiers.
- **Plate-forme unique à l'échelle de l'entreprise** accéder à de diverses sources de données, pour un grand nombre d'utilisateurs et de cas d'utilisation.

Analysis



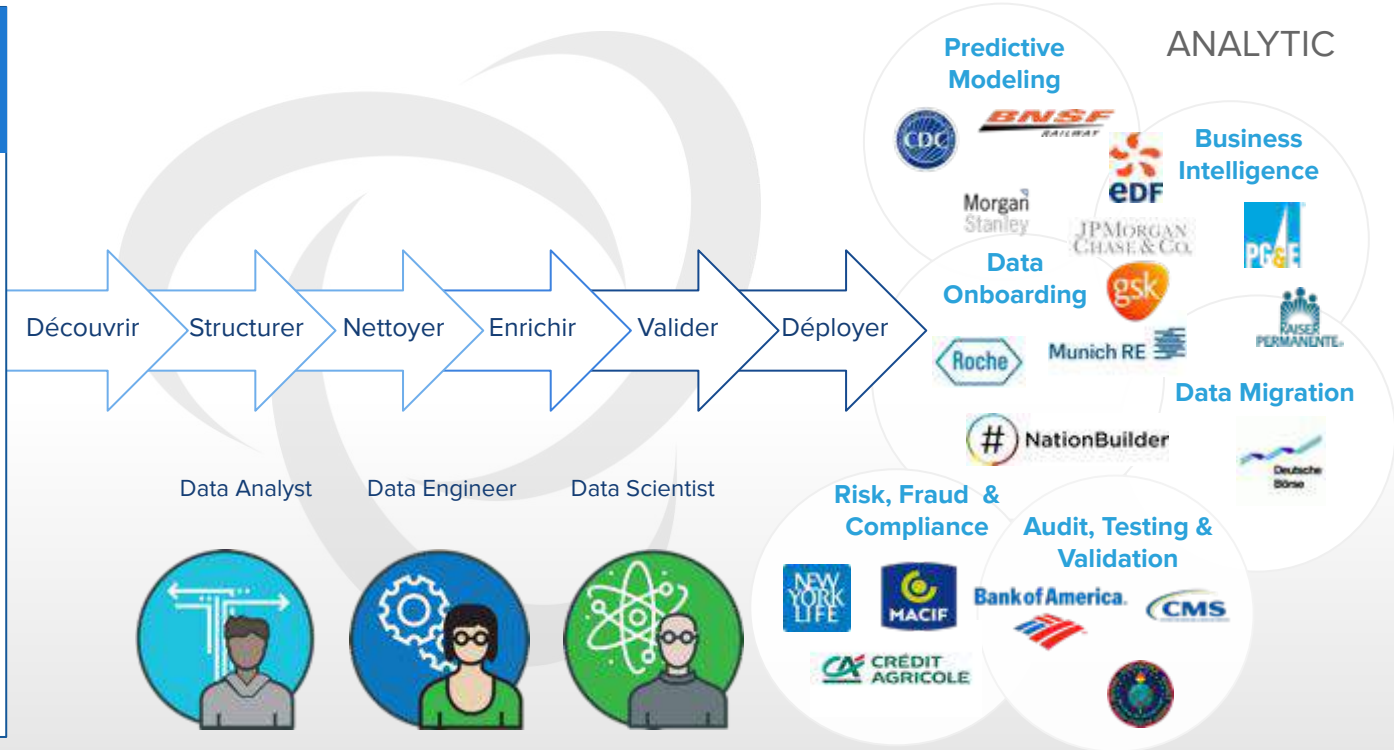
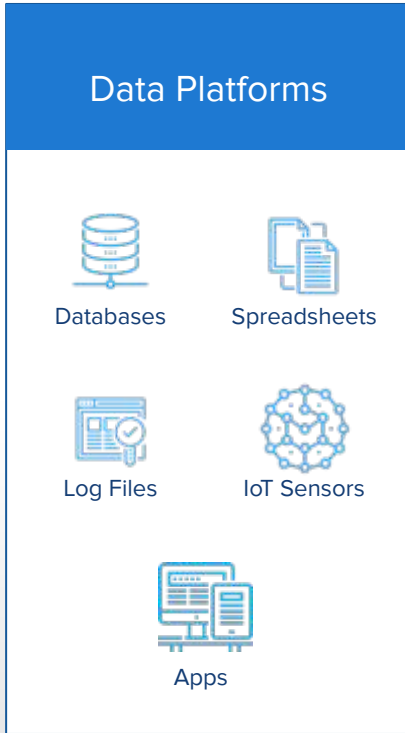
Data Warehouse



AI



Business Intelligence





Préparation de données dans les Clouds

Architectures & Démonstrations

Pourquoi et comment utiliser la préparation des données dans les Data Lake Cloud ?

Les solutions modernes de préparation de données peuvent aider à accélérer l'adoption des Data Lake dans le Cloud.

Six étapes à suivre:

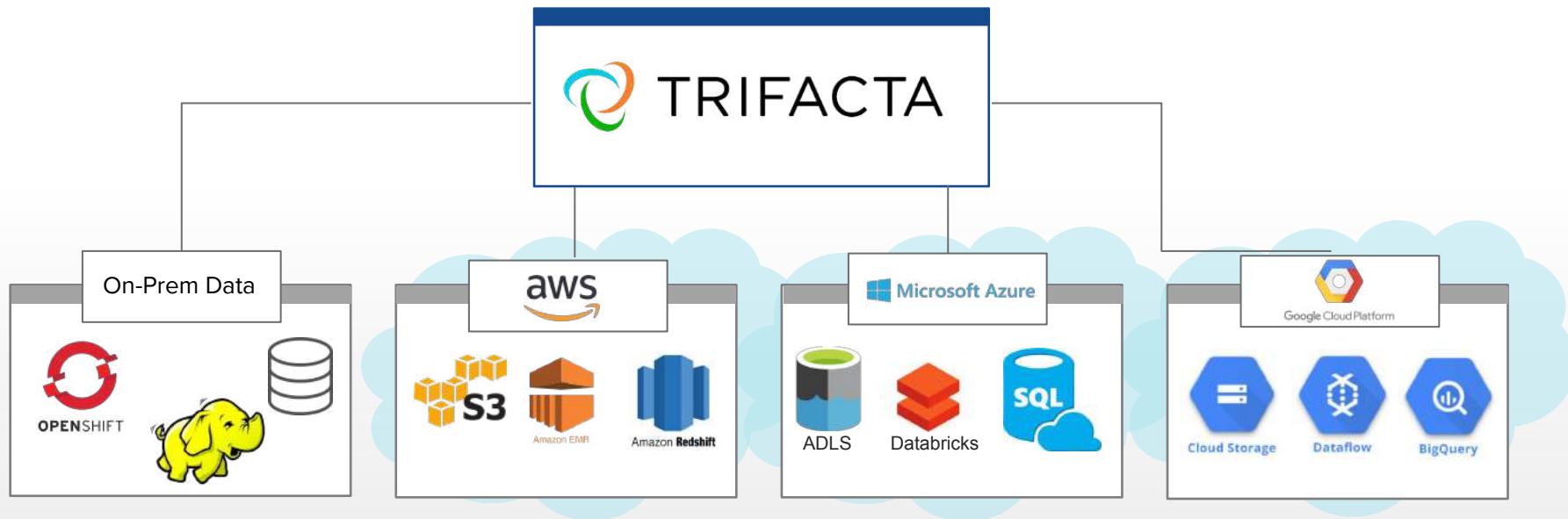
1. Choisir une solution **nativement intégrée** dans les plateformes Cloud.
2. Utiliser la data prep pour **augmenter l'efficacité** et la collaboration des équipes.
3. Choisir les **bons cas d'utilisation**, ne pas refaire ce qui fonctionne déjà bien.
4. Permettez aux utilisateurs de disposer d'un **accès aux données Cloud en libre-service** tout en maintenant la gouvernance des données.
5. S'assurer de la **qualité des données à l'échelle** avec une validation continue.
6. **Publier et mettre en opération** pour l'analytique et le machine learning.

Architecture

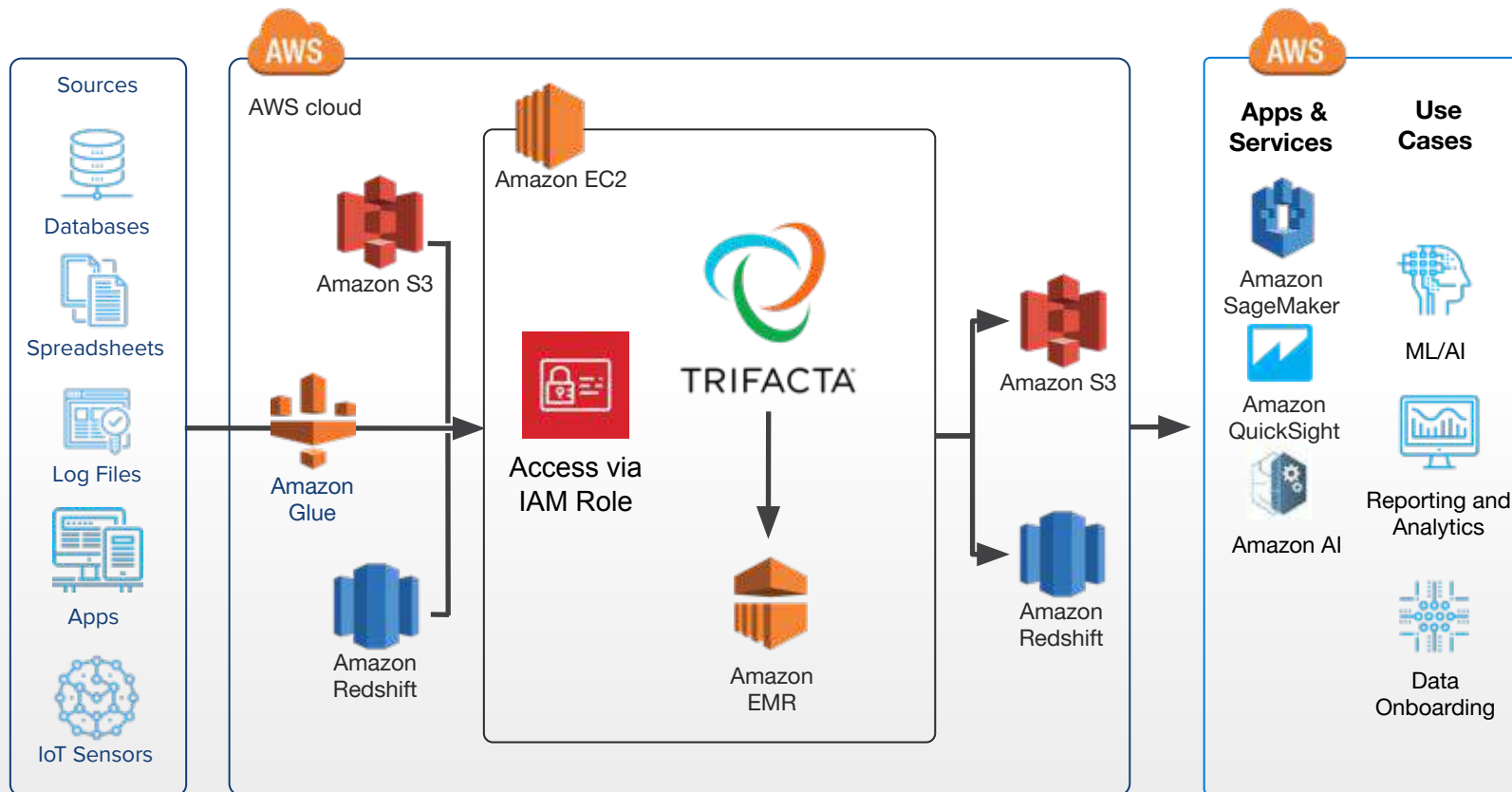
Wrangle Once. Run Everywhere.

On-Premise or Cloud:

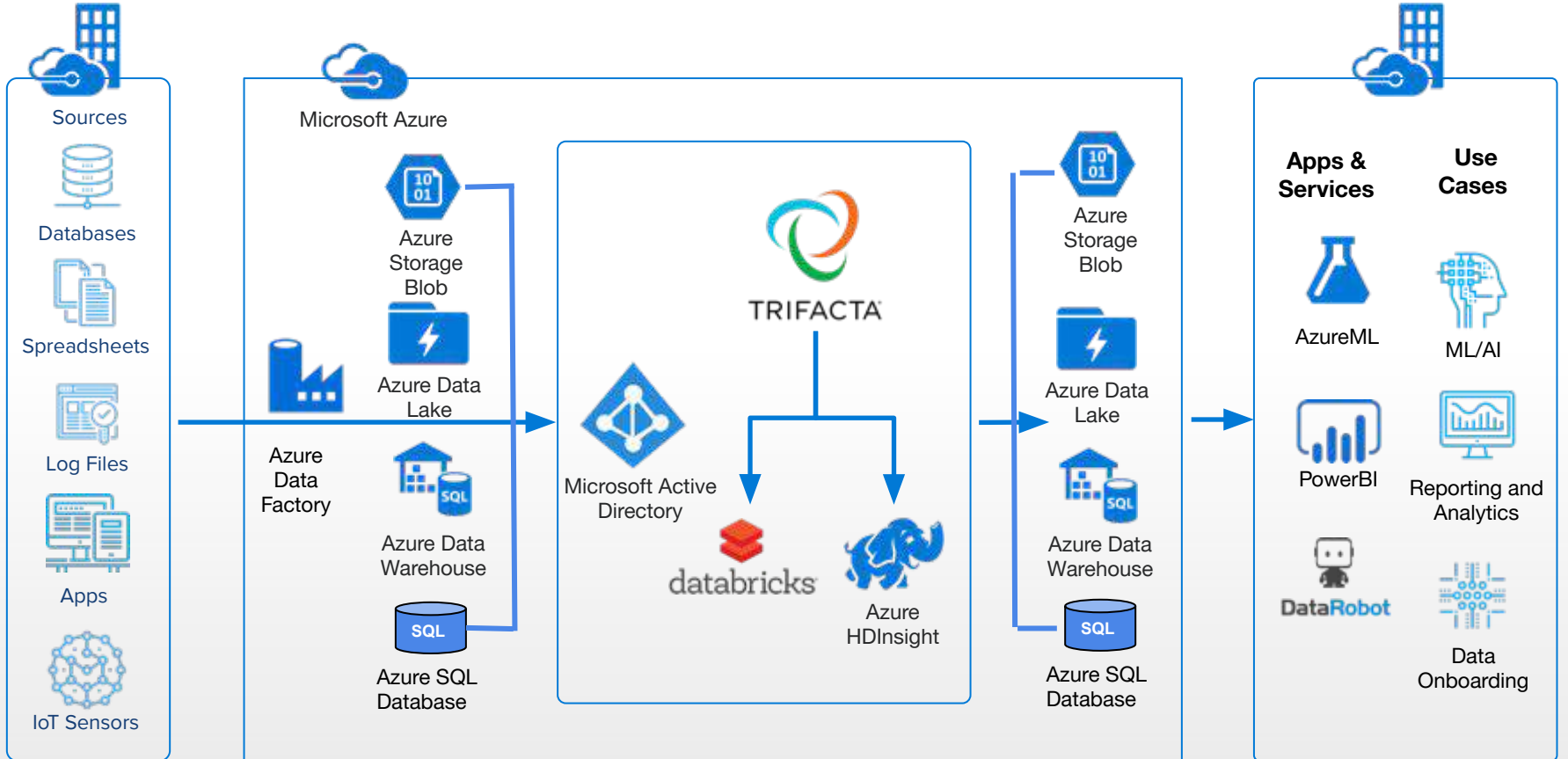
- Native Storage
- Native Processing
- Native Security



Amazon Web Services



Microsoft Azure

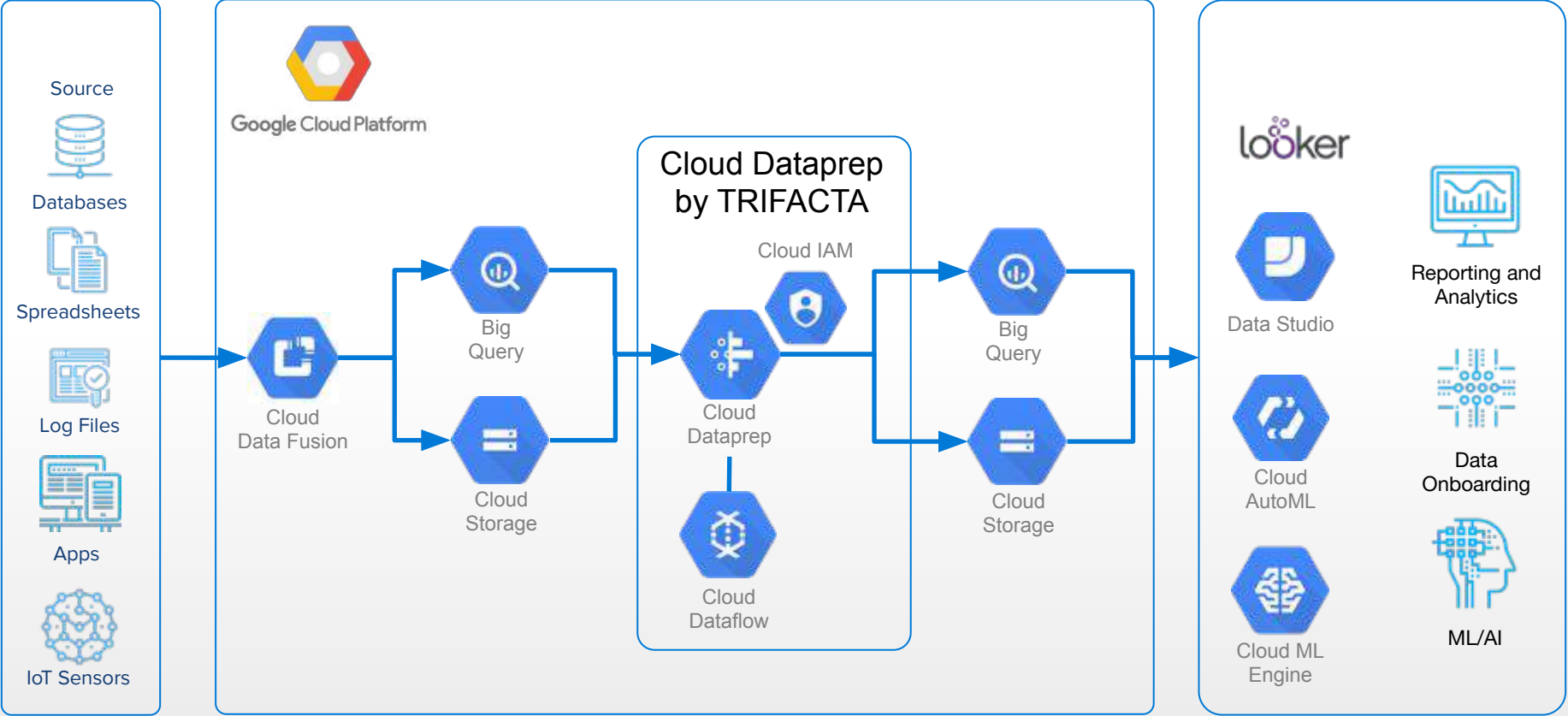


Démonstration

Google Cloud



Google Cloud Platform





Essayer et utiliser Trifacta gratuitement

<https://www.trifacta.com/start-wrangling/>

**Merci
Questions ?**

Festival Online de la Data



Synaltic

PROCHAINES DATES

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

Jeudi 9 Juillet de 15h à 16h

Pourquoi utiliser la DataViz pour analyser ses données ?

Franck Nguyen, Synaltic