



?

Choisir ses services Data dans Azure

A-t-on encore besoin d'une base de données en 2021 ?



**Rejoindre Microsoft Tech
Community dans Slack**

<https://bit.ly/3wy1pl3>

Annonce

- Comment ne pas se sentir perdu.e devant le nombre de services Azure dédiés à la donnée ?
 - Sont-ils compatibles ?
 - Y a-t-il des recouvrements de fonctionnalités ?
 - Combien cela va-t-il coûter ?
 - Est-ce qu'on a encore besoin des bases de données en 2021 ?
- Au travers de démonstrations concrètes, nous aborderons les bases d'une **architecture data Azure**, basée sur les **services managés** et répondant aux deux scénarios principaux que sont le *dashboarding* et le *Machine Learning*.

Paul Péton

MVP Data Platform & AI



- Meetup organizer and speaker
- Fan of Big Data Hebdo ([podcast](#))
- Formed as a Data Miner (old buzzword...)
- Manager expert Analytics @AVANADE



• <https://www.linkedin.com/in/paul-peton-datascience>



• <http://methodidacte.org>



• <https://paul-peton.medium.com/>



• <https://github.com/methodidacte/>

Télécharger ce support

2021 Data Engineering Ecosystem

Presented by  lakeFS

Data Discovery



Data Quality & Observability



Lineage, Management, Governance



MLOps



Analytics Workflow



Notebooks



Virtualization



Formats



Data Lifecycle Management



Compute



Object Storage



Metastores



Orchestration



Analytics Engine



Ingest - Tech



Ingest - SaaS



+
◦ •

Faut-il construire
toute l'architecture
avant de
commencer ?

Non !

+
• ◦

Evaluer en premier le potentiel de valeur de vos données



Par une étude préalable, exploratoire, confirmant ou non le potentiel explicatif, prédictif voire prescriptif de vos données.

Cette étape permettra d'évaluer **l'investissement nécessaire** et le **ROI attendu**.

Une approche en quatre étapes



Ingestion



Stockage



Calcul



Exposition

INGESTION

STORAGE

COMPUTE

EXPOSE

?

?

?

AI

BI



DevOps

?

?

MLOps

Les concepts qui nous guideront



Cloud computing



Platform as a Service



Serverless

A gallery wall with several framed paintings. The wall is a deep teal color. The paintings include a portrait of a young boy in the top left, a landscape with figures in the top center, a portrait of a woman in the top right, a portrait of a woman in the bottom left, a large religious scene in the bottom center, and a portrait of a woman in the bottom right. A person in a dark coat is seen from behind, looking at the large religious painting. Two horizontal white lines are overlaid on the image, one above and one below the word 'Exposition'.

Exposition

Expose:

la partie « émergée » du nuage



Dashboarding

- « *regard dans le rétroviseur* »
- sous forme visuelle, interactive
- donnant une marge de manœuvre aux utilisateurs (« *self service BI* »)



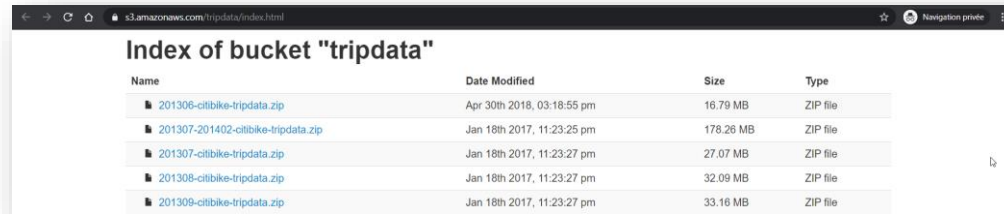
Modèle prédictif

- « *regard vers l'avant* »
- sous forme d'une API, intégrée dans une application
- réponse en temps réel ou en mode *batch*

*L'exposition doit permettre la **création de valeur** à partir des données.
Cette création de valeur doit, à plus ou moins long terme, **dépasser le coût d'investissement et de fonctionnement** de la plateforme data.*


Cas d'application: *les vélos en location à New York City*

- Données
 - <https://www.citibikenyc.com/system-data>



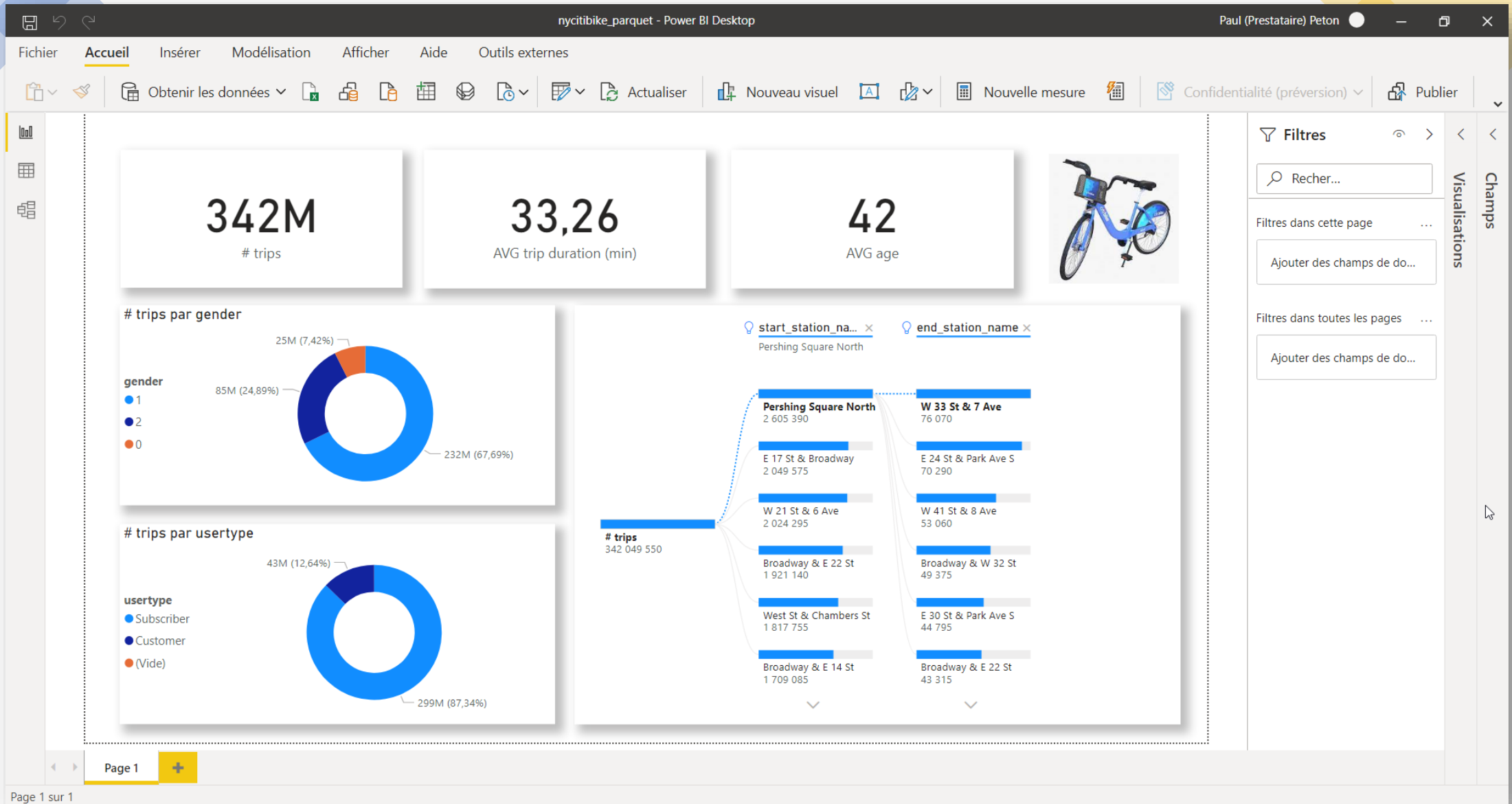
Index of bucket "tripdata"

Name	Date Modified	Size	Type
201306-citibike-tripdata.zip	Apr 30th 2018, 03:18:55 pm	16.79 MB	ZIP file
201307-201402-citibike-tripdata.zip	Jan 18th 2017, 11:23:25 pm	178.26 MB	ZIP file
201307-citibike-tripdata.zip	Jan 18th 2017, 11:23:27 pm	27.07 MB	ZIP file
201308-citibike-tripdata.zip	Jan 18th 2017, 11:23:27 pm	32.09 MB	ZIP file
201309-citibike-tripdata.zip	Jan 18th 2017, 11:23:27 pm	33.16 MB	ZIP file

- De 2017 à mars 2021 : 
Table : nycitibike (342,049,550 lignes)

- Objectifs :
 - Analyser les données visuellement
 - Prédire le temps de trajet (« *trip_duration* »)







+ New

🏠 Home

Author

📄 Notebooks

⚙️ Automated ML

🧩 Designer

Assets

📁 Datasets

🧪 Experiments

🔗 Pipelines

📦 Models

🔗 Endpoints

Manage

💻 Compute

📦 Datastores

📝 Data Labeling

🔗 Linked Services

[Home](#) > [Endpoints](#) > tripduration-service

tripduration-service

Details

Test

Consume

Deployment logs

Input data to test real-time endpoint

Test

```
{  
  "data": [  
    [3621, 3094, 37793, 1991, 1],  
    [3081, 3048, 26396, 1969, 0]  
  ]  
}
```

Test result

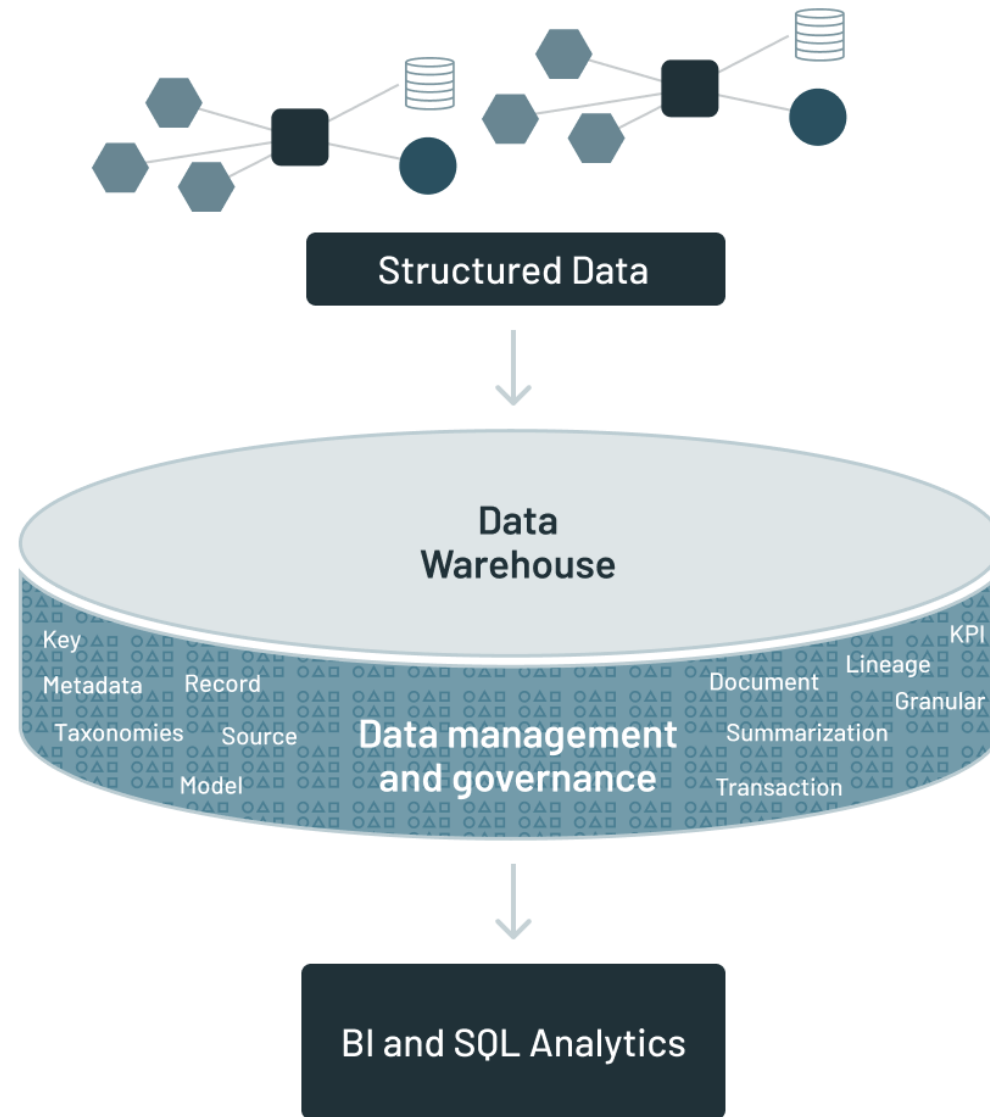
parsed

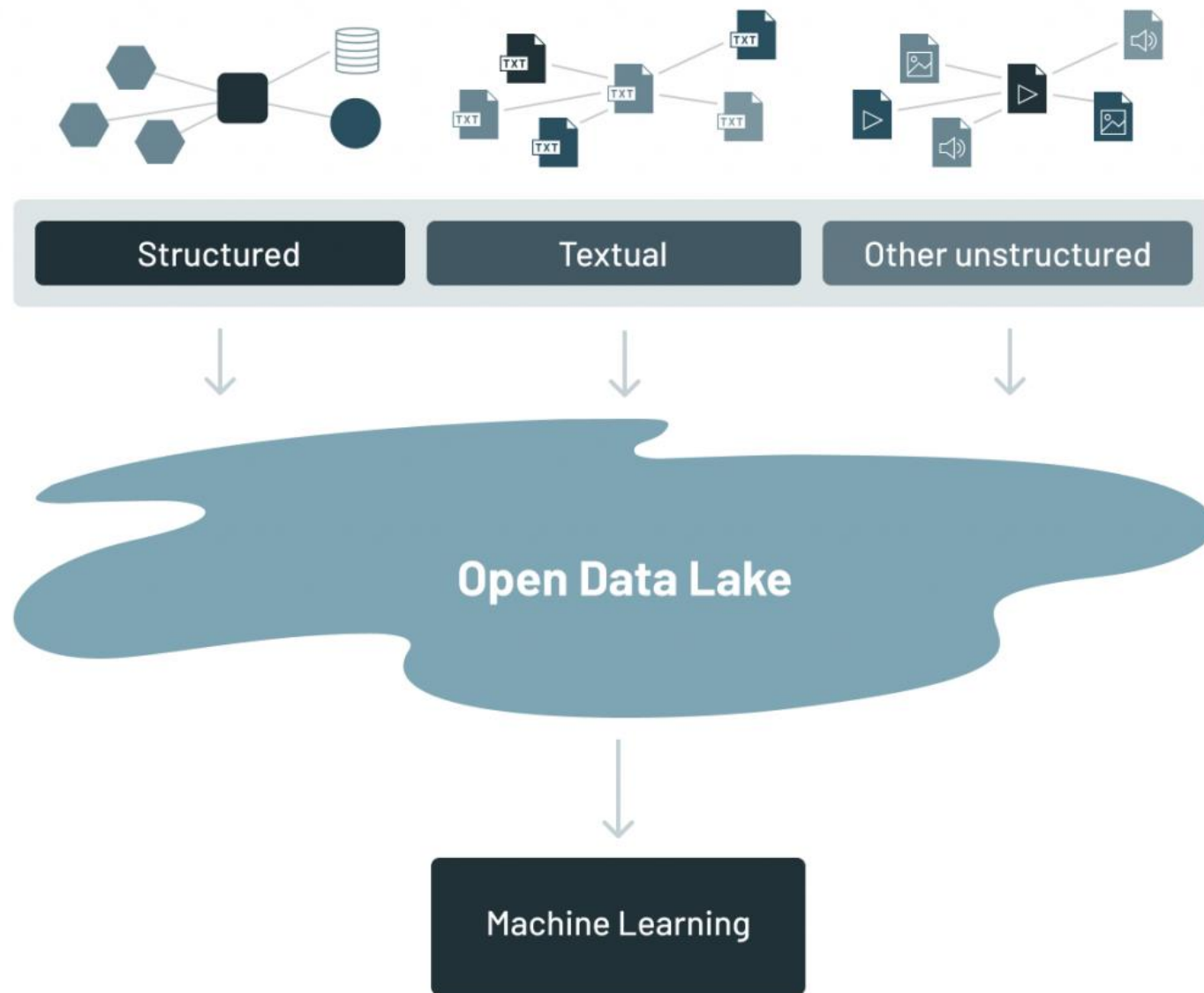
raw

```
[  
  842.778606974003,  
  828.0094733252218  
]
```

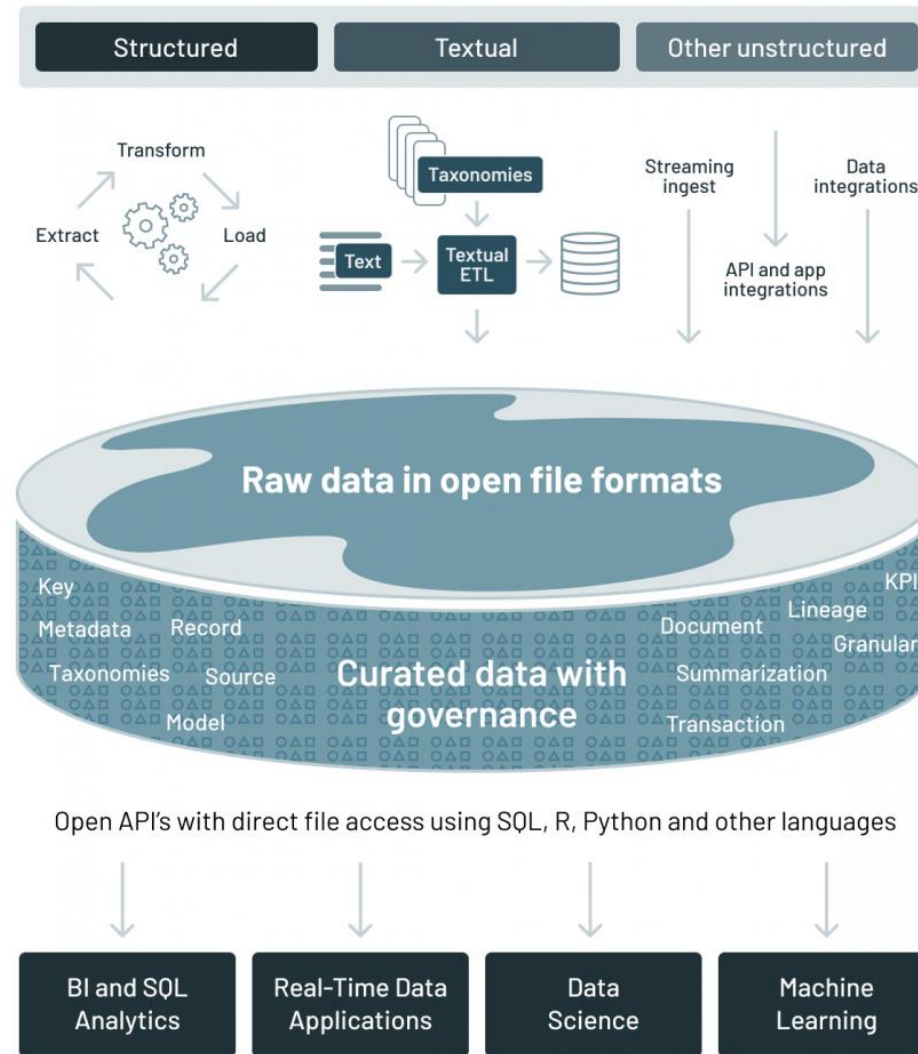



Storage





Data Lakehouse



L'approche Lakehouse: *un lac multi-couches*

- Les données sont archivées dans le *data lake* dans plusieurs couches
 - RAW : les données brutes (sans transformation) telles qu'elles sont transmises
 - CLEAN : les données nettoyées et enregistrées dans un format optimisé (ex.: parquet)
 - AGG : les données agrégées selon les besoins de reporting



- La BI peut tirer profit des couches CLEAN et AGG
- La Data Science utilisera les couches RAW ou CLEAN selon le besoin

Le Data Lake:

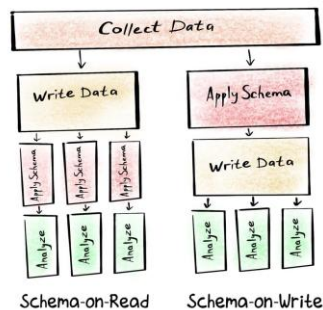
entre décharge publique et mine d'or

- Anticiper la gouvernance
 - En particulier, les règles de nommage
- Réfléchir à une stratégie de *backup / restore*
 - Sans doute partielle
 - Identifier où se trouve la valeur
 - Evaluer le temps nécessaire pour reproduire Silver & Gold à partir de Bronze
 - Accepter de perdre les données des couches Raw



A quoi sert une base de données ? (dans une approche analytique)

- C'est une ressource de calcul
 - Toujours disponible ? Donc toujours facturée !
 - Privilégier la scalabilité automatisée
 - *up* pendant les traitements batch, *down* sinon
 - le mode *serverless* pour les requêtes adhoc des analystes
- Les données sont typées
 - Voire contrôlées par des mécanismes de *triggers*
- *Schema on write versus schema on read*



@luminousmen.com

Schema-on-Write	Schema-on-Read
<ul style="list-style-type: none">- fast reads- slower loads- not agile- structured- fewer errors- SQL	<ul style="list-style-type: none">- slower reads- fast loads- very agile- structured/unstructured- more errors- NoSQL

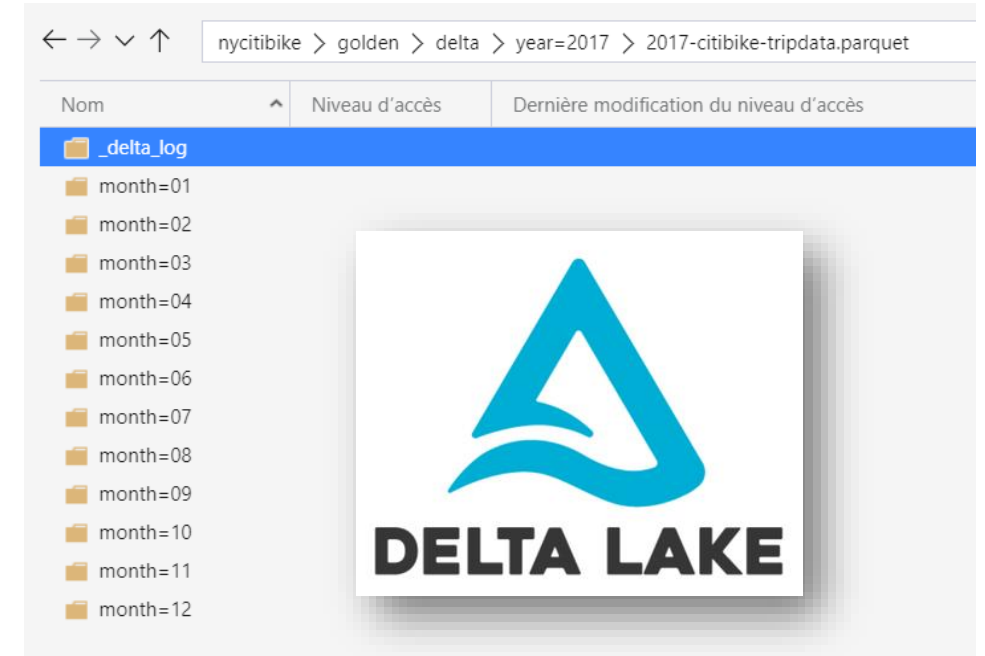
@luminousmen.com

<https://luminousmen.com/post/schema-on-read-vs-schema-on-write>

Delta Lake:

un parquet avec des propriétés ACID

- ACID : Atomicité, Cohérence, Isolation et Durabilité
- CSV (raw) -> Parquet (clean)
- « surcouche du format Parquet »
- Historique JSON
 - *Time travel*
 - Transactions
 - Supporte merge, update et delete
- Alternatives : Apache Iceberg, Apache Hudi



[Présentation de Delta Lake - Azure Synapse Analytics | Microsoft Docs](#)

[Synapse/Hitchikers Guide to Delta Lake - Python.ipynb at main · Azure-Samples/Synapse · GitHub](#)

[How Interchangeable Are Delta Tables Between Azure Databricks and Azure Synapse Analytics? – Welcome to the Community Blog of Paul Andrew \(mrpaulandrew.com\)](#)



Compute

Calcul:

la puissance se paie, au juste prix

- L'illusion du PaaS : à l'origine, tout est machine (virtuelle)
 - Attention aux quotas, aux disponibilités...
- Choisir les machines adaptées
 - GPU uniquement pour des approches par Deep Learning
- Un cluster Spark se compose :
 - D'un nœud « *driver* »
 - Demande souvent plus de mémoire
 - D'un ou plusieurs nœuds « *workers* »



Ingestion Orchestration



Ingestion, orchestration: *de l'huile dans les rouages*

- Choisir entre les approches *no code*, *low code* ou *full code*
- Disposer de nombreux connecteurs
- Exploiter les API pour piloter les différents services
 - Exemple : <https://methodidacte.org/2021/01/azure-data-factory-pilote-les-actualisations-power-bi/>
- Ordonnancer dans le temps ou déclencher sur événement
 - Par exemple, à l'arrivée d'un nouveau fichier sur le compte de stockage



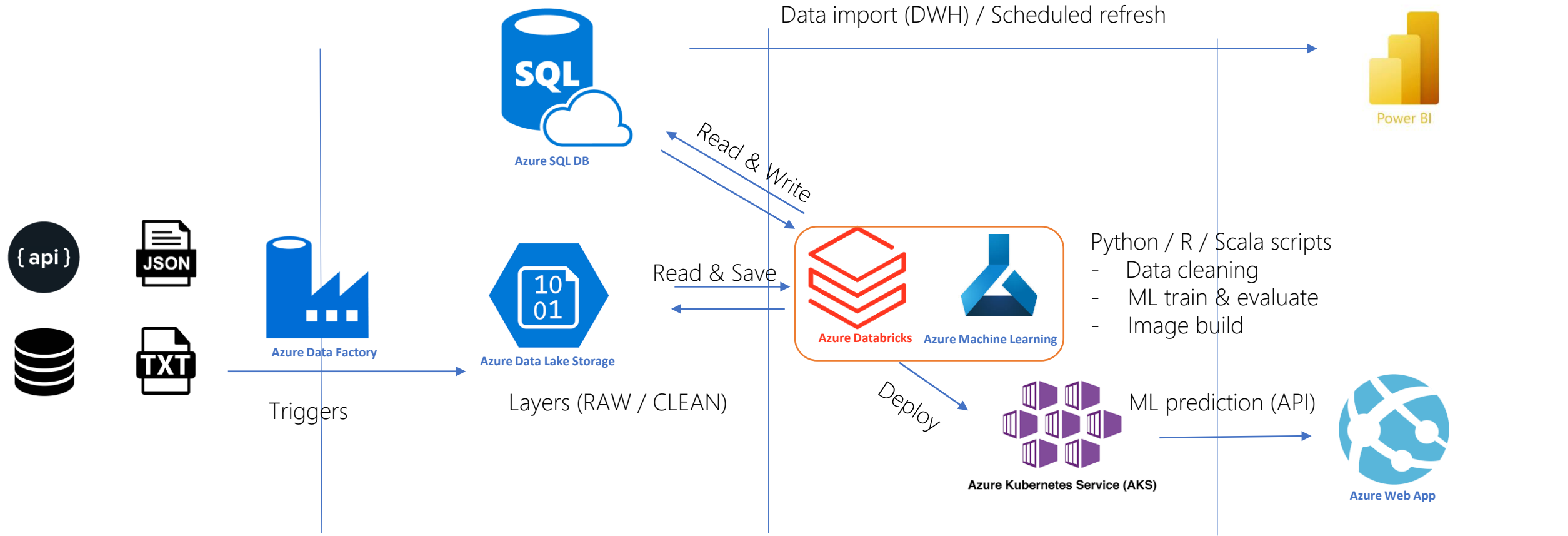
Big Picture(s)

INGESTION

STORAGE

COMPUTE

EXPOSE



Dev

DevOps

Logs

MLOps

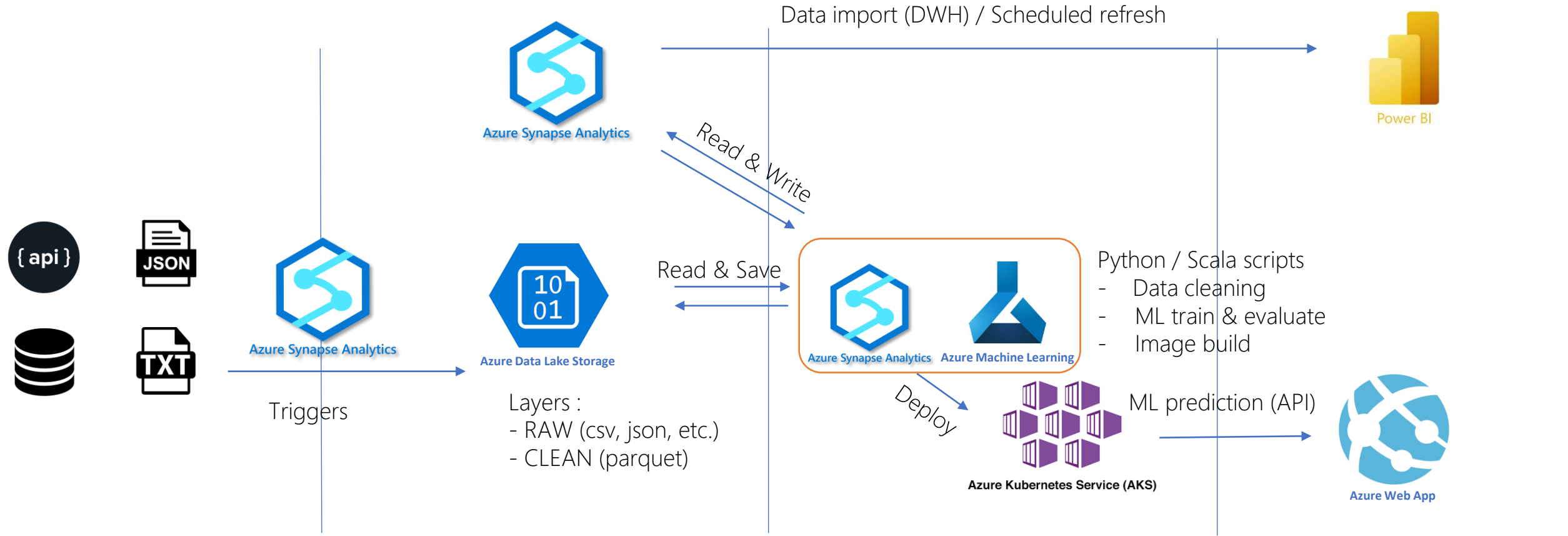
Security

INGESTION

STORAGE

COMPUTE

EXPOSE



Dev

DevOps

Logs

MLOps

Security

Approche DevOps: les outils complémentaires

Versioning

- Git
- Vault

Tests, qualimétrie

- Pytest ou autre
- SonarQube

Continuous X

- Integration
- Deployment

Monitoring

- (Essayer de) tout centraliser dans Log Analytics

Encadrement et optimisation des coûts

- Bien dimensionner les ressources de calcul

Synapse Analytics & Azure ML



Charles-Henri Sauget

MVP Data Platform chez Air Liquide



charleshenrisauget



@SaugetCh



www.sauget-ch.fr



Paul Peton

Manager expert analytics chez Avanade



paul-peton-datascience



@paulpeton

Meetup



LES GENTILS DEVELOPPEURS DATA PLATFORM.

YouTube Live
Mercredi 2 Juin 2021 à 17h

<https://www.meetup.com/fr-FR/Les-gentils-developpeurs-Data-Platform/events/278004460/>



#GDDP



Les-gentils-developpeurs-Data-Platform

Unified Analytics Pipeline

Delta Architecture with Databricks

