

DataBird

Culture tech et data

Pourquoi de nouvelles infrastructures data ont-elles émergé ? 🤔

4 V's

Volume

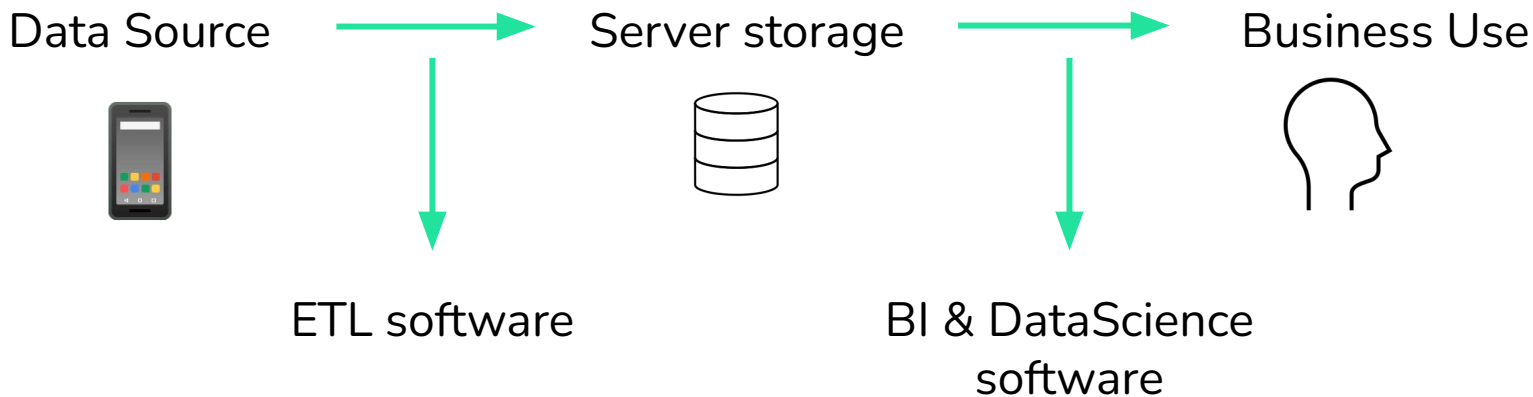
Velocity

Variety

Veracity



Rapide coup d'oeil aux infrastructures modernes



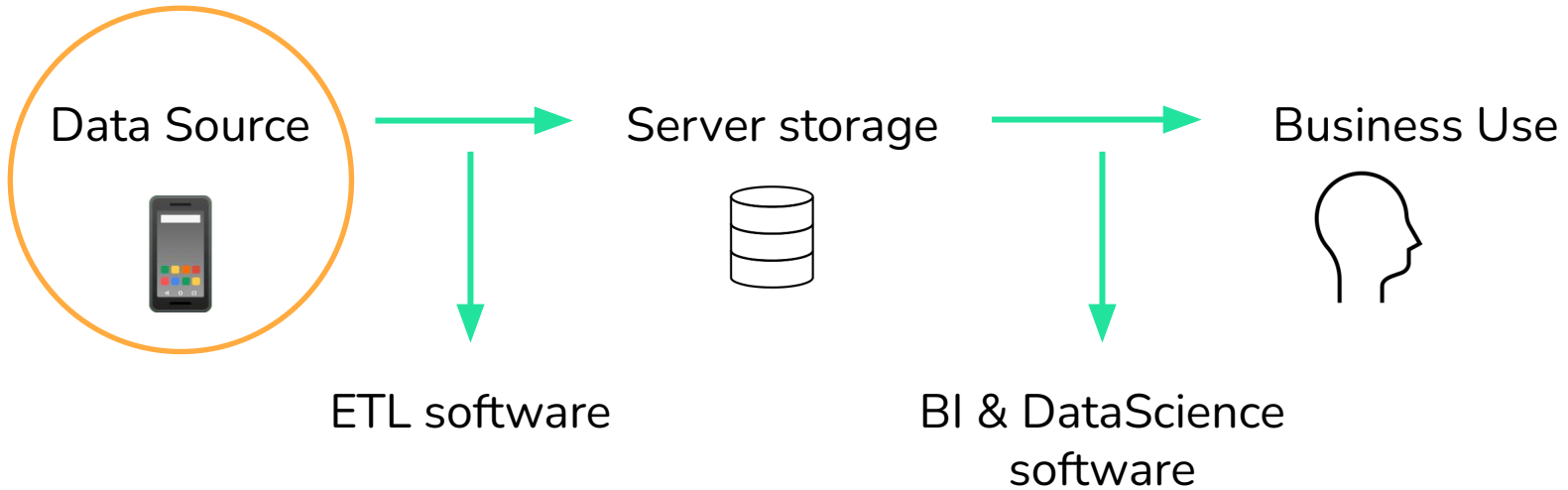
Today's agenda : A modern data architecture.



1. 3 grandes sources de données 🛠️
 1. Données internes à une entreprise
 2. Données liées aux événements
 - ↳ D'une utilisation d'un site internet
 - ↳ D'une utilisation d'un produit
 3. Données externes
 - ↳ Web Scraping
 - ↳ Les APIs de logiciels (salesforce, SAP, hubspot, mailing, ...)
2. Extract - Transform - Load data ❤️
 1. Zapier / DBT / Spark / Stitch & Talend / Fivetran / ...
3. Stocker les data 📱
 1. On premises
 2. On clouds
 - ↳ Data lakes
 - ↳ Data warehouses : Redshift / BigQuery / Snowflake / ...
4. Analytics & Dashboarding 🧐
 1. Data Science platform : Dataliku / Amazon SageMaker / Azure / ...
 2. Dashboarding : Looker / Google DataStudio / Tableau / Power BI / ...



1.0 L'origine des données



1.1 Data source : Les données internes

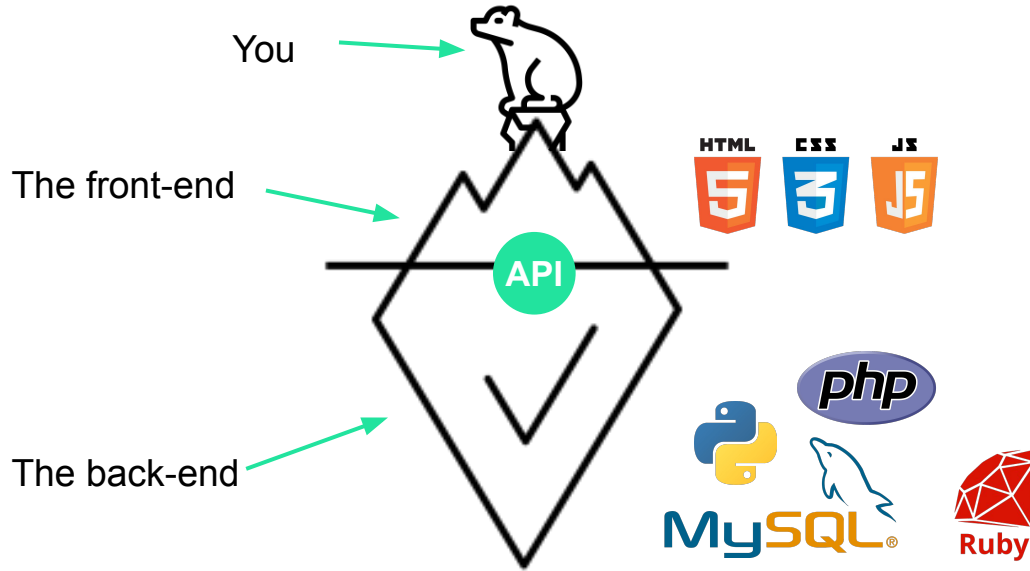
Les données internes à une entreprise stockées dans une base de données relationnelle (compte clients / stock / prix / RH / finance)

- Une base de données (OLTP) conçue pour utiliser un modèle relationnel
 - ↳ Le but étant de faciliter les opérations SQL “INSERT - UPDATE - DELETE”
 - ↳ Les données y sont normalisées : pour éviter les doublons et redondances, mise en place de contraintes pour éviter les valeurs aberrantes.



1.1 Data source : Données internes

Qui crée ces bases de données ? Développeur Back-end vs Front-end



1.2 Data source : Event data



Depuis son site web via des balises GTM par exemple ou sur un mobile via segment

de visiteurs

de pages consultées

..



Google Tag Manager



1.3.1 Data source : Données externes - web scraping

A quoi sert le Web Scraping ?

→ Récupérer automatiquement de l'information depuis des pages web avec un script qui va imiter un comportement humain.

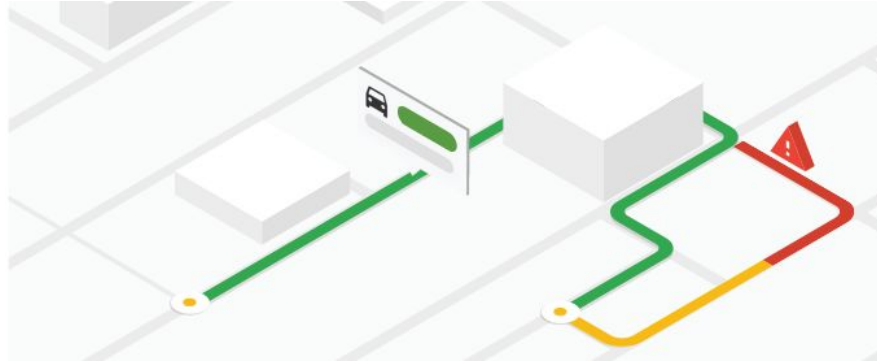
Si Uber veut lancer un nouveau service de covoiturage : comment analyser les prix de la concurrence ?



1.3.2 Data source : Données externes - API

Si Uber veut trouver le trajet le plus court d'un point A à un point B.

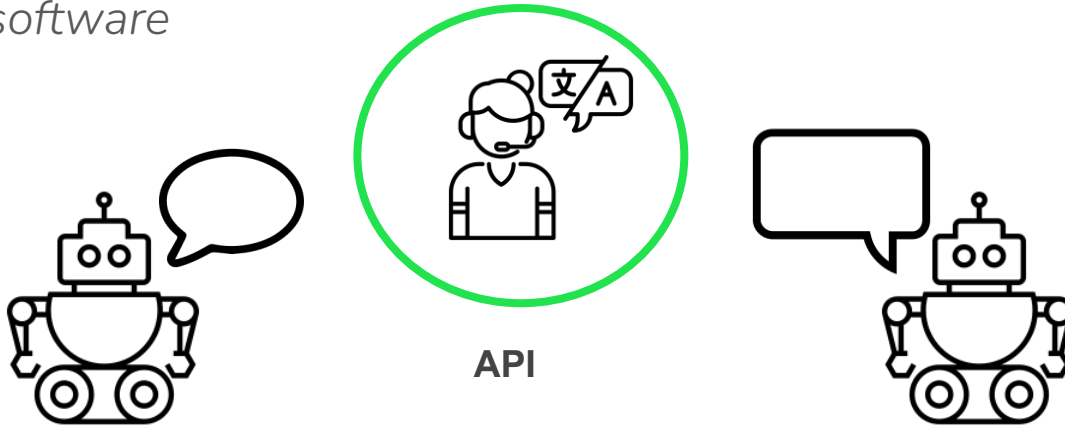
Comment faire cela rapidement ?



1.3.2 Data source : Données externes - API

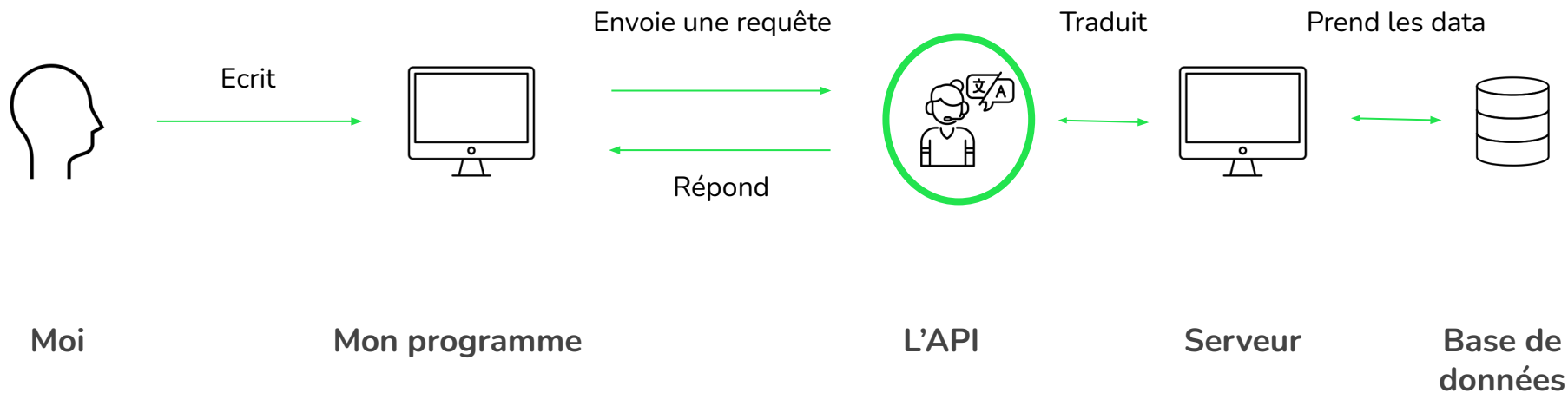
Qu'est-ce qu'une API ? 🚩

Une API (application programming interface) est un moyen de communication entre deux software



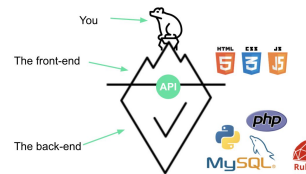
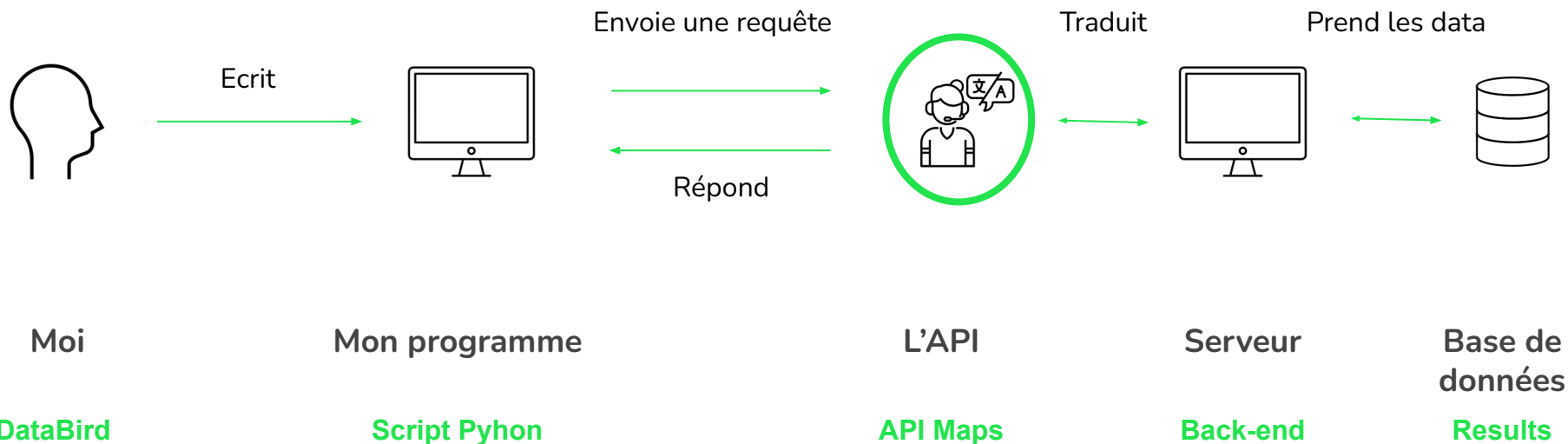
1.3.2 Data source : Données externes - API

Qu'est-ce qu'une API ? 📌



1.3.2 Data source : Données externes - API

Qu'est-ce qu'une API ? 🚩



1.3.2 Data source : Données externes - API

Utiliser une API : focus sur les APIs “REST”

1. Séparation client - serveur

- ↳ Chacun a un rôle spécifique
- ↳ Ils utilisent les protocoles HTTP pour transférer les données
- ↳ Peu importe comment les deux sont codés, ils peuvent communiquer

2. Système en couches

- ↳ Un client se connectant à un serveur ne sait pas ce qu’il se trouve derrière l’API
- ↳ Positif pour la propriété intellectuelle, la sécurité et le contrôle des émissions

3. Scalable

- ↳ Les requêtes doivent être très précises et rédigés strictement indépendamment du client et du serveur



1.3.2 Data source : Données externes - API

Les méthodes principales pour requêter une API

Methods	API verb in HTTP
CREATE	POST
READ	GET
UPDATE	PUT
DELETE	DELETE

- Quelles méthodes utiliser pour récupérer les tweets ? Toutes les méthodes seront-elles disponibles ?
- Dans quel format seront envoyées les données ?



Exemple d'une requête API avec Postman

<https://www.pappers.fr/>

<https://www.postman.com/>

→ https://api.pappers.fr/v2/recherche?api_token=d6ef957beb40690c8884e6cae24967d71068f7dbbb36bfcd&par_page=2

GET	▼	https://api.pappers.fr/v2/recherche?api_token=d6ef957beb40690c8884e6cae24967d71068f7dbbb36bfcd&par_page=2
-----	---	---

Chemin vers l'API

Clef d'accès

Paramètres



Exemple d'une requête API avec Postman

GET https://api.pappers.fr/v2/recherche?api_token=d6ef957beb40690c8884e6cae24967d71068f7dbbb36bfcd&par_page=2&siren=883538001 Send

Params Authorization Headers (6) Body Pre-request Script Tests Settings Cookies

Query Params

	KEY	VALUE	DESCRIPTION	...	Bulk Edit
<input checked="" type="checkbox"/>	api_token	d6ef957beb40690c8884e6cae24967d71068f7dbbb...			
<input checked="" type="checkbox"/>	par_page	2			
<input checked="" type="checkbox"/>	siren	883538001			

Body Cookies Headers (16) Test Results Status: 200 OK Time: 755 ms Size: 2.24 KB Save Response

Pretty Raw Preview Visualize JSON

```
1  [
2    "resultats": [
3      {
4        "siren": "883538001",
5        "siren_formate": "883 538 001",
6        "nom_entreprise": "DATABIRD",
7        "personne_morale": true,
8        "denomination": "DATABIRD",
9        "nom": null,
10       "prenom": null,
11       "sexe": null,
12       "siege": {
13         "siret": "88353800100016",
14         "siret_formate": "883 538 001 00016",
15         "nic": "00016",
16         "numero_voie": 10,
17         "indice_repetition": null,
18         "type_voie": "RUE",
19         "libelle_voie": "GRENETA",
20         "complement_adresse": null,
21         "adresse_ligne_1": "10 RUE GRENETA",
22         "adresse_ligne_2": null
23       }
24     }
25   ]
26 }
```

JSON

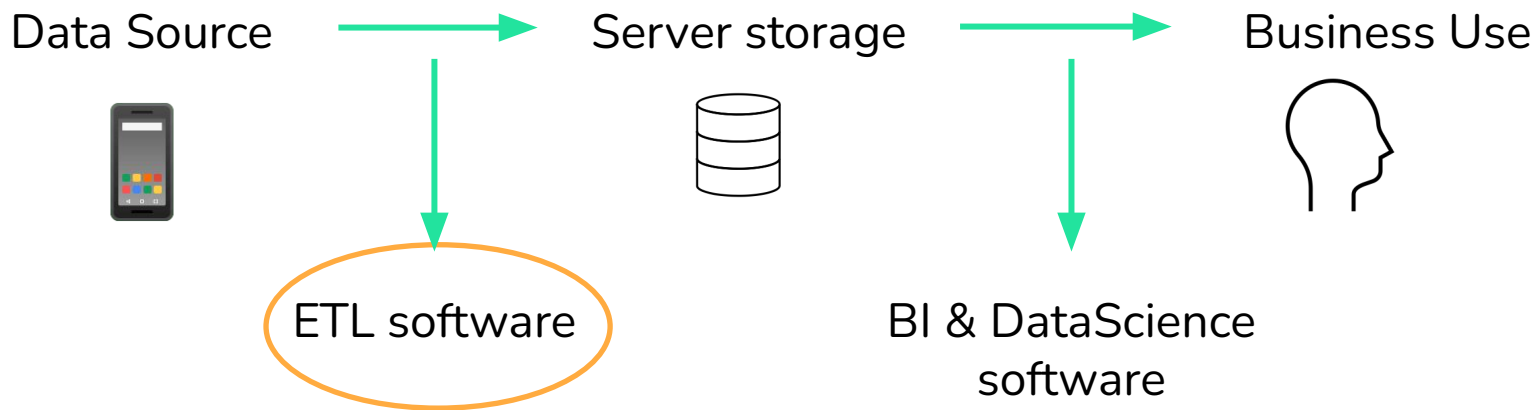


1.3.2 Data source : Données externes - API

Logiciels tiers : lesquels ? Pour quelles données ?



2.0 Software ETL



2. Extract - Load - Transform data ❤️

- Anciennement ETL avec Spark.
- Aujourd'hui changement de paradigme, on transforme les données à la fin.
 - Possible grâce à la diminution des coûts de stockage.
- Plus de stabilité et plus résilient au changement.

Logiciels d'ETL nouvelle génération

👉 ☁️ **Talend & Stitch** : Software qui connecte des APIs entre elles pour centraliser la donnée

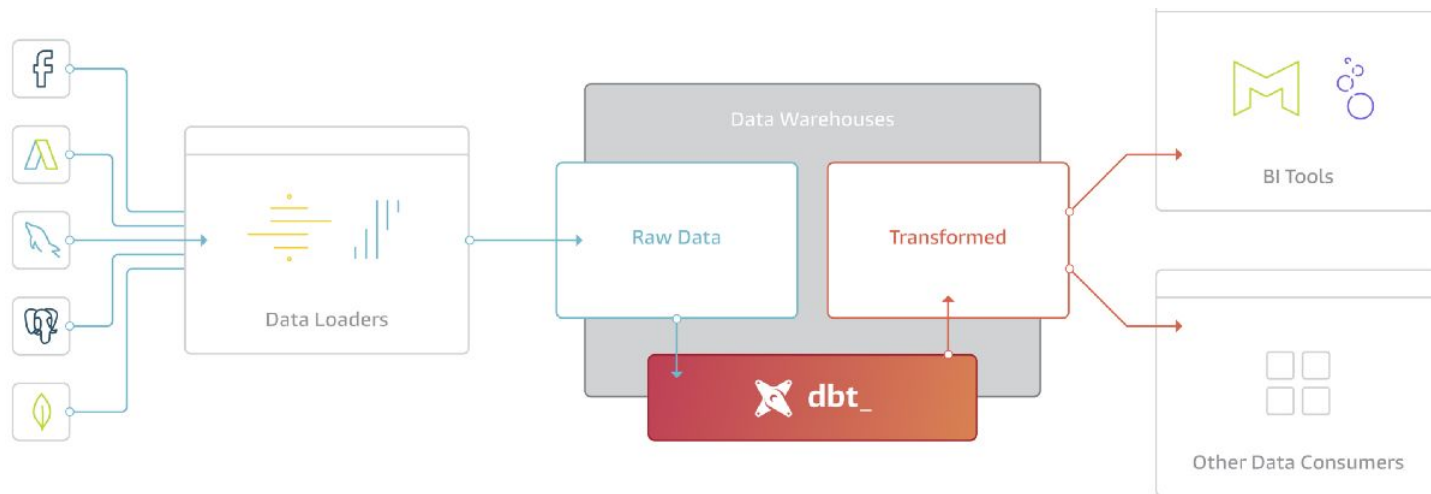
👉 ☁️ **Fivetran** : Même chose que précédemment avec quelques transformations SQL possibles



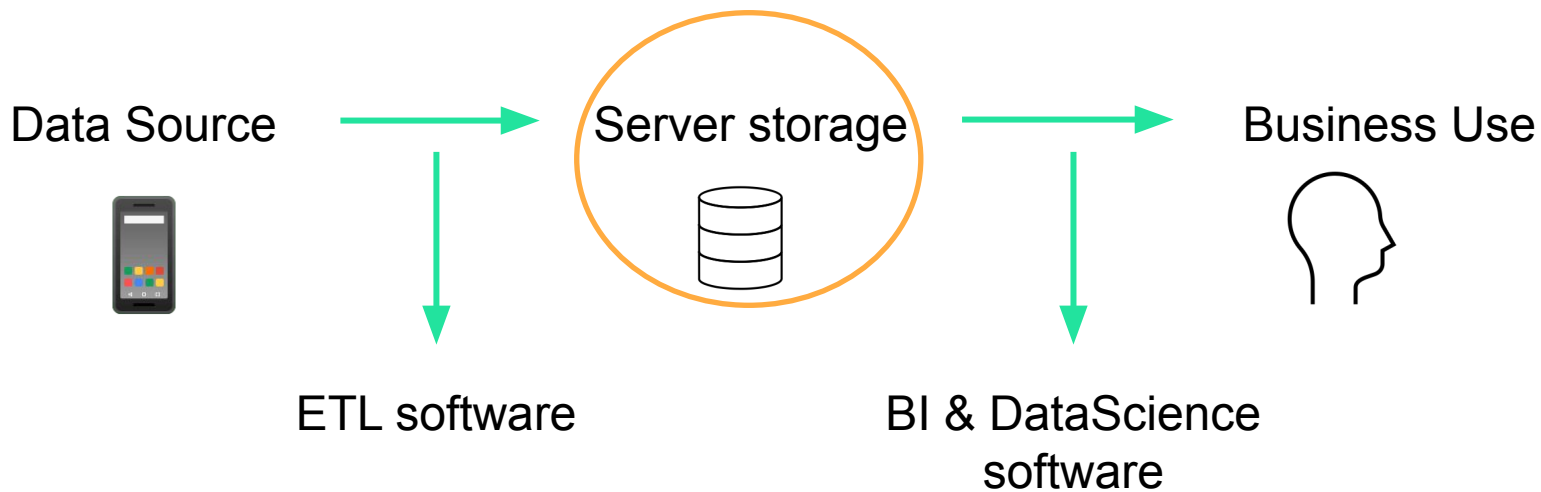
2. Extract - Load - Transform data ❤️

☁️ 🥩 DBT : Des transformations SQL au sein du Data Warehouse

→ Souvent combiné avec les connecteurs mentionnés la slide précédente



3.0 Stockage



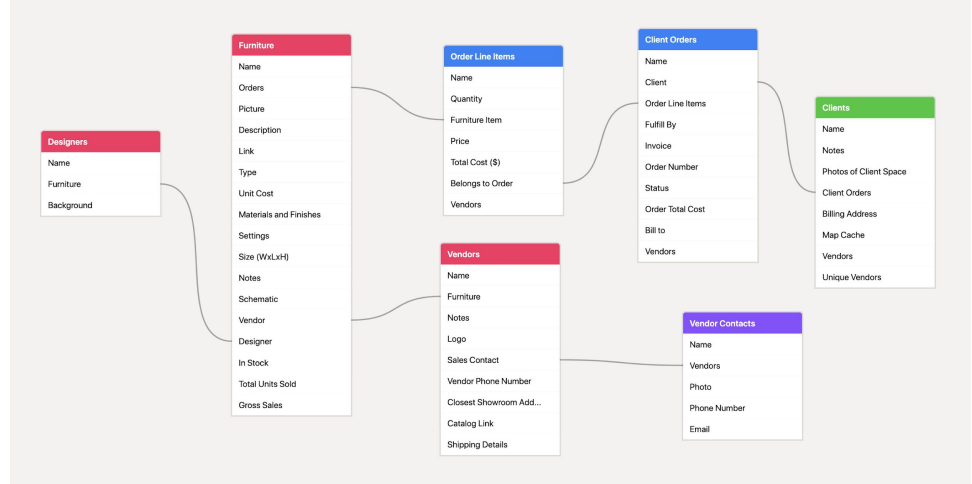
3.0 Stocker ses données : version cheap

Base de données très simple



	A	B	C
1	Name	first name	email
2	Miny	Alexandre	alexandre@data-bird.co
3	Grignola	Antoine	Antoine@data-bird.co
4			
5			
6			
7			

Inspiré de SQL : Créer des relations entre tables



3.1 Stocker des données sur ses serveurs

Pourquoi avoir ses propres serveurs ?

1. Brèche de sécurité ?
2. Réglementation - données sensibles ?
3. Visibilité - où sont mes données ?
4. Accessibilité et latence - Ai-je un accès rapide ?
5. Confiance - Quelle configuration ?



3.2 Stocker ses données sur les services cloud

Pourquoi avoir ses données sur le cloud ?

1. Flexible et scalable
2. Facilité d'utilisation
3. Moins coûteux à gérer (réplication, sécurité)
4. Cryptage outsourcé
5. Mise à jour des logiciels de gestion



3.2 Stocker ses données sur le cloud : data lakes

Les data lakes sont tels des dépôts centralisés pour stocker toutes les données structurées ou non.

Exemples de data lake : HDFS, Hive, MongoDB, Cassandra, ElasticSearch



3.2 Stocker ses données sur le cloud : data warehouses

Les data warehouses sont des base de données sur lesquelles on effectue des analyses. On requête les data warehouse avec du SQL.

→ Un système de stockage + super calculateur

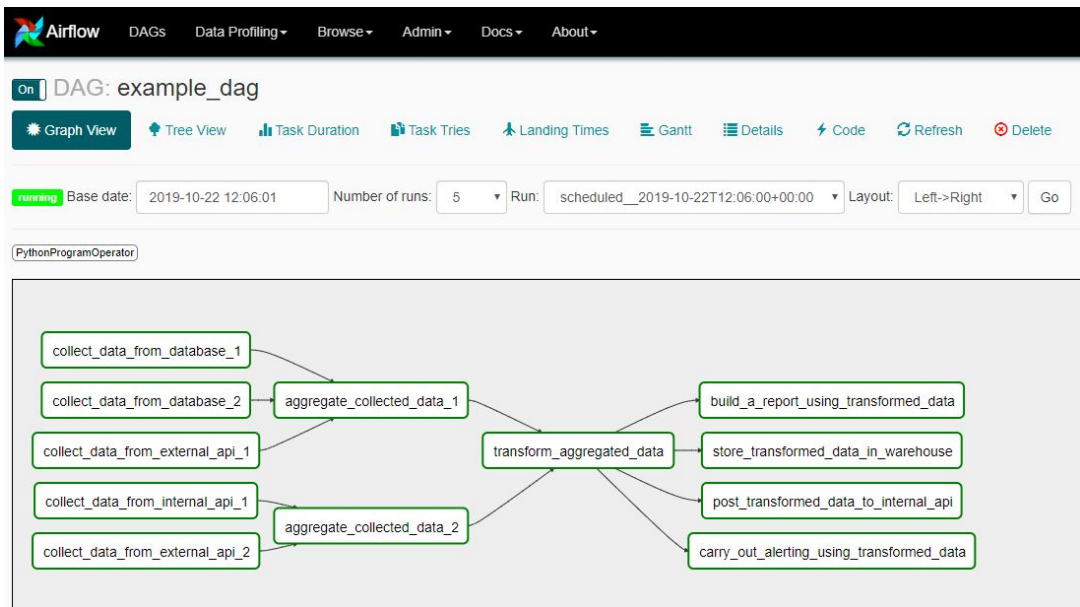
→ Principaux data warehouses : BigQuery / Redshift / Azure / Snowflake

- Les Data warehouses facilitent le SELECT en SQL
 - ↳ Travailler avec les informations historiques et archivées -> Pas de temps réel
 - ↳ Base de données OLAP - online analytical processing databases
 - ↳ Les bases OLTP (comme celles pour le stockage des donnée de l'entreprise) sont trop lentes pour accéder aux données

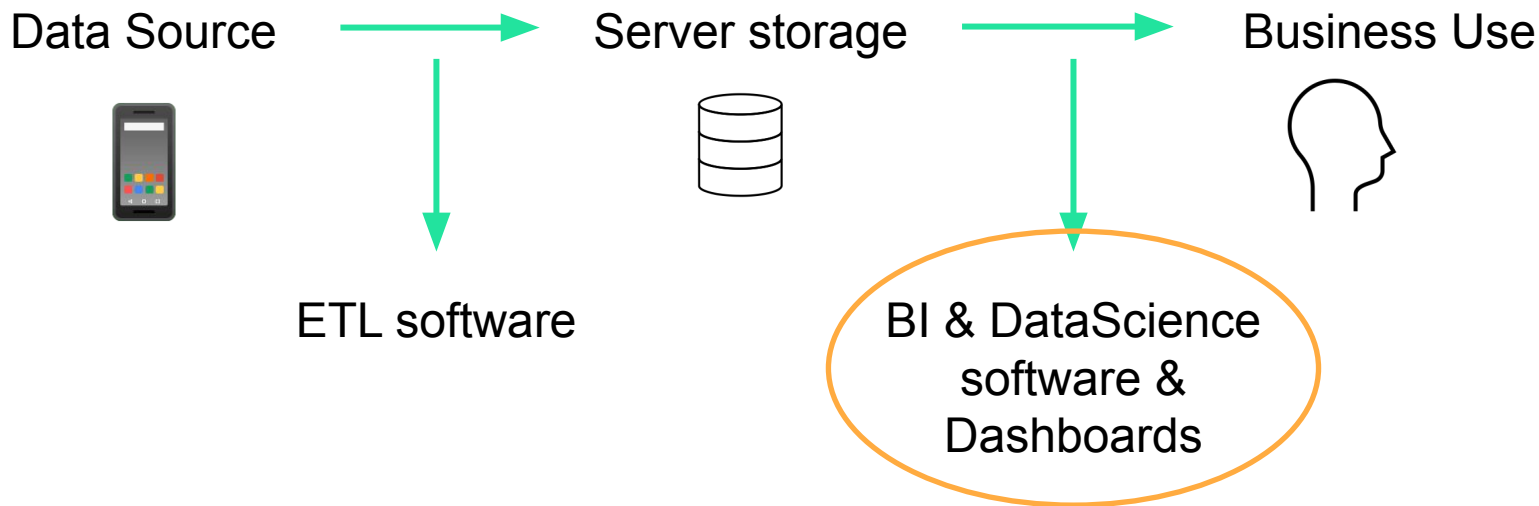


3.3 Airflow en chef d'orchestre

Airflow permet d'organiser et lancer automatiquement les scripts selon les dépendances existantes.



4.0 BI & DataScience software



4.1 Analytics & Dashboarding : plateformes DS

Les plateformes permettent de faciliter le travail des data scientists & analysts

- ☁️ 🌐 📺 Alteryx : dashboarding & no-code self-service data analytics platform
- ☁️ 🌐 📺 Knime : similar to alteryx but with a open-source version
- ☁️ 🌐 📺 Dataliku DSS : replaces existing tools rather than to integrate with them.
- ☁️ 🥩 🌐 DataBricks : Spark platform + workflow orchestration
- ☁️ 🌐 📺 DataRobot : Automated ML → automatically finds a good model
- ☁️ 🥩 🌐 Sagemaker : All-in-one tool → Infra + automated ML + platform



4.2 Analytics & Dashboarding : Dashboards

L'intérêt d'un dashboard est qu'il se connecte au cloud et donc se met à jour automatiquement avec de nouvelles données.

 : appartient à Google et est très complet.
Looker




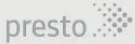



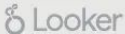



 Data Studio : appartient à Google et est intégré pour les données marketing.

 : très intuitif avec une logique de “drag & drop”

 : appartient à Microsoft et assez customizable
Power BI

→ Mais encore : Chartio / Mode / Metabase / Sisense / Qlick view / Periscope / ...

5. Un marché ultra dynamique

On Prem → Cloud Data Warehouse	Data warehouses are moving to the cloud with increased flexibility, scale, and ease of use—allowing any company to be a data company	 snowflake  Google Big Query
Hadoop → Next-gen Data Lakes	Data lakes and related systems are becoming more performant and reliable, adding RDBMS-like features including ACID transactions and interactive SQL queries	 databricks  presto
ETL → ELT	Brittle ETL processes (extract-transform-load) are being replaced with more flexible and consistent ELT pipelines (extract-load-transform)	 Fivetran  dbt
Workflow Manager → Dataflow Automation	Data flow automation systems are helping to orchestrate thousands of data pipelines with a cleaner abstraction and modern executor integrations	 PREFECT  DAGSTER  Apache Airflow
Analyst Teams → Self-serve Insights	Reporting, dashboarding, and automated analysis tools are becoming more available to non-technical users	 Looker  Superset
Endpoint Protection → Global Data Governance	Data security and privacy measures (e.g., access controls) are becoming centralized on the data platform as use of data is increasingly regulated and user endpoints are harder to protect	 Collibra  PRIVAGERA

