

## Contents

|   |   |
|---|---|
| Jour 1A .....   | 2 |
| Vocabulaire de la data .....  | 2 |
| 1.1 Présentation du métier de data analyst .....                              | 3 |
| 1.2 Missions du data analyst.....   | 4 |
| 2.1 Présentation du métier de data scientist.....                             | 4 |
| 2.2 Missions du data scientist.....   | 4 |
| 3.1 Présentation du métier de data engineer.....                              | 4 |
| 3.2 Missions du data engineer .....   | 4 |
| 4. Présentation du data hero (ambassador de data) .....                       | 4 |
| 5. Présentation du métier de data consultant.....                             | 5 |
| 6. Présentation du métier de DPO (déléguer à la protection des données) ..... | 5 |
| 7. Présentation du métier du Head of Data (manager d'équipe data).....        | 5 |
| 8. Présentation du métier de CTO (chief technology officer) .....             | 5 |
| JOUR 2M-Maturité .....  | 6 |
| Comment évaluer la maturité data d'une entreprise ? .....                     | 6 |
| L'évolution du niveau de maturité data d'une entreprise .....                 | 6 |
| La gouvernance des données.....   | 6 |
| Quelles sont les roadmaps d'un projet data ? .....                            | 7 |
| Comment devenir data driven ? .....   | 7 |
| Comment gérer les interactions des équipes tech et non tech ? .....           | 7 |
| Comment gérer une équipe de data analytics ?.....                             | 7 |
| Comment recruter des profils data ? .....                                     | 8 |
| Conseils pour se lancer dans la data.....                                     | 8 |
| The Modern Data Stack récap.....  | 8 |
| Grille de lecture - Data Maturity.....  | 8 |
| Data Maturity récap.....  | 8 |
| JOUR 2M- Infra.....   | 8 |
| 📌 Objectifs .....   | 8 |
| Support de la présentation .....  | 8 |

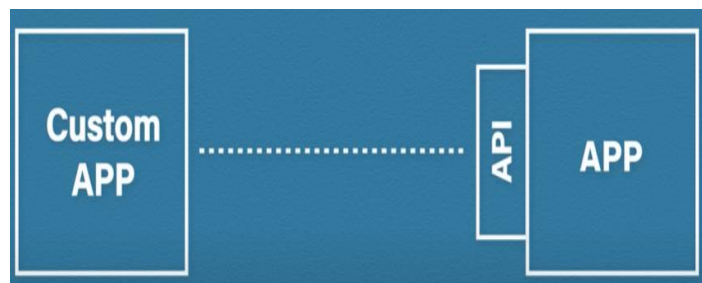
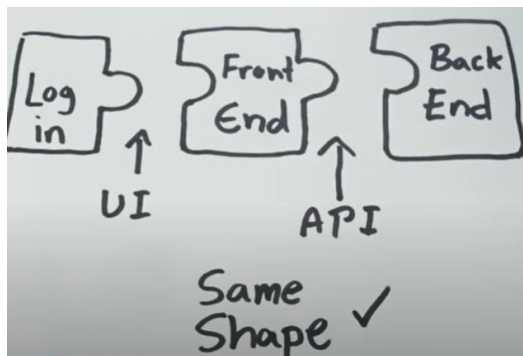
## Jour 1A

### Vocabulaire de la data

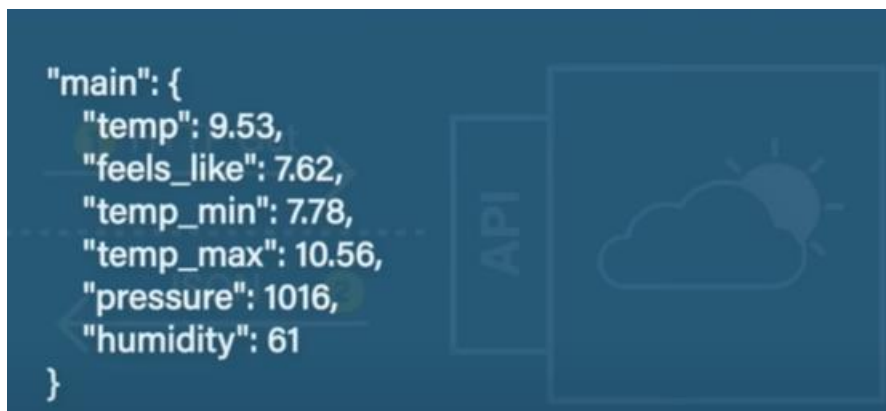
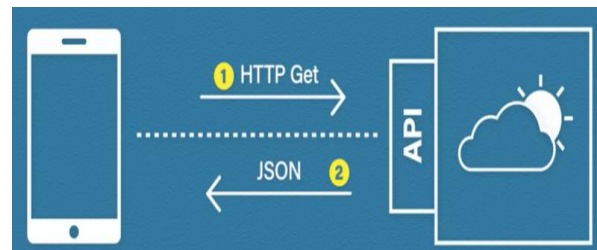
Les mots de vocabulaire abordés :

- API (application programming interface)  
Opendatagouv  
Connect apps together  
UI (user interface)

[https://www.youtube.com/watch?v=ZveW4\\_ZJtVY](https://www.youtube.com/watch?v=ZveW4_ZJtVY)



| HTTP METHODS |             | CRUD   |
|--------------|-------------|--------|
| POST         | HTTP GET    | CREATE |
| GET          | HTTP POST   | READ   |
| PUT          | HTTP PUT    | UPDATE |
| PATCH        | HTTP PATCH  | UPDATE |
| DELETE       | HTTP DELETE | DELETE |



- Back-end vs Front-end

- Les méthodes agiles  
C'est une série d'outils qui permet l'entreprise d'organiser l'équipe de projet  
<https://www.youtube.com/watch?v=gYsU2VTESE8>
- Environnement de test vs Mise en production (MEP)  
Test le modèle avant pour trouver des bugs et déploiement
- Git
- **ETL** (Extract transform load)
  - Pour les données à **faible** volume
  - Il est utilisé pour les données **structurées** et relationnelles
  - La mise en œuvre du processus ETL est relativement **simple**.
  - Apporter des modifications aux données dans l'étape de transformation prend du temps, et plus il y a de données, plus la transformation prend de temps.
  - Étant donné que dans le processus ETL, les données doivent être sélectionnées, puis la transformation et le chargement sont effectués sur les données sélectionnées, cela nécessite donc une **maintenance élevée**.
- **ELT** (Extract load transform)
  - Il peut être utilisé pour les données à volume **élevé**
  - Il est utilisé pour les sources de données **non structurées** et les Clouds.
  - Cela demande des **compétences particulières**
  - La **vitesse** n'a rien à voir avec le volume de données.
  - Comme les données sont toujours disponibles, le besoin de **maintenance est extrêmement faible**.  
<https://www.nemoudar.com/blog/elt-vs-etl/>
- Cloud (pour stocker les données)
- Composition d'une équipe produit  
Il a tout le rôle nécessaire pour un développement de produit. Il va faire la gestion de projet. Pour expliquer plus technique pour les développeurs.
- Intelligence artificielle vs Data Science vs Machine Learning  
ML( automatiser le machine), DS( statistique, optimisation), IA( les robots, permet de développer application de ML)
- Data Lake (centraliser la donnée)
- Data Warehouse (la base de données et le moteur qui permet de construire cette base de données pour obtenir une version plus exploitable des KPI (les indicateurs) pour tous les utilisateurs d'entreprise.

### 1.1 Présentation du métier de data analyst

- Le rôle du data analyst exploite les données de l'entreprise pour prendre des meilleures décisions
- Le data analyst travaille toujours pour le business

## 1.2 Missions du data analyst

Exemples de missions du data analyst :

- Réaliser un AB test
- Réaliser des dashboards
- Réaliser des analyses data ad hoc pour le business
- Les langages : SQL, Excel, python, tableau, domo, metabase, powerBI, data hekou

## 2.1 Présentation du métier de data scientist

- Traditionnellement, le data scientist ne s'occupait que de créer des modèles
- Aujourd'hui, on attend aussi de lui de mettre en production les modèles qu'il a développés
- Les data scientists sont essentiellement des ingénieurs
- Les langages : python, java, scala, tensorflow

## 2.2 Missions du data scientist

- Le data scientist va principalement se concentrer sur la création et la performance de modèles
- Il peut faire de la recommandation, de la prédiction ou de l'analyse d'image / texte / son
- Exemple de projet concret sur lequel le data scientist peut travailler :
  - Repérer automatiquement les défauts sur des rails de chemin de fer

## 3.1 Présentation du métier de data engineer

- Le data engineer crée, alimente et maintient les bases de données
- Le data engineer s'occupe de l'ETL
- Le data engineer est un "développeur" dans la data
- Les langages : SQL, DBT, Cloud, AWS, Azur, python, Java, Scala, Go
- cloud sur les optimisation technique

## 3.2 Missions du data engineer

- Une mission du data engineer pourrait être de tagger la donnée avec des segments définis par les data analysts et les équipes marketing

## 4. Présentation du data hero (ambassador de data)

- La data hero n'est pas un métier en tant que tel

- Le data hero est un membre d'une équipe métier qui s'est formé sur des aspects techniques
- Le data hero est un réel ambassadeur et facilitateur data

## 5. Présentation du métier de data consultant

- Un consultant data peut avoir n'importe quel rôle data dans l'entreprise
- Il permet de développer une expertise en interne
- Un exemple de mission sur lequel il pourrait intervenir : embaucher des consultants data pour traiter la mise en conformité de la RGPD

## 6. Présentation du métier de DPO (délégué à la protection des données)

- Le DPO conseille les entreprises dans la mise en conformité RGPD (règlement européen pour la protection des données)
- Il accompagne dans le respect des règles européennes mais aussi dans la mise en place des documents en cas de contrôle
- Un exemple concret dans lequel le DPO agirait serait le partage d'informations sensibles dans le cadre d'une entreprise opérant dans le recrutement

## 7. Présentation du métier du Head of Data (manager d'équipe data)

- Son rôle est de recruter et de former les équipes data
- Il va être aussi l'ambassadeur data dans l'entreprise
- Il est vraiment le pilier central de la stratégie "data-driven" de l'entreprise

## 8. Présentation du métier de CTO (chief technology officer)

- Son rôle est de donner la direction technique/technologique de l'entreprise. Il peut être :
  - Le manager des équipes techniques
  - L'expert technique de l'entreprise
  - Les deux

## JOUR 2M-Maturité

### 📌 Objectifs

- Evaluer la maturité data d'une entreprise
- Développer une approche data driven
- Quels sont les profils à recruter

### Comment évaluer la maturité data d'une entreprise ?

- Pour analyser la maturité data d'une entreprise on peut regarder :
  - Les typologies et les rôles des profils présents dans l'entreprise
  - La taille des équipes data
  - Les technologies utilisées (cloud, tooling, ...)
  - Les process mis en place

<https://www.nemoudar.com/blog/data-maturity-model/>

چهار مرحله مدل بلوغ داده

Data Aware (آگاهی به داده)

Data Proficient (چیرگی بر داده ها)

Data Savvy (درک داده ها)

Data Driven (داده محور)

### L'évolution du niveau de maturité data d'une entreprise

- Stade 0 : Utilisation de simples fichiers Excel sur un Drive
- Stade 1 : Recruter une première personne en data et mise en place d'une base de données et d'un outil de visualisation open source
- Stade 2 : Formation des équipes au SQL et au dashboarding pour utiliser et valoriser la donnée
- Stade 3 : Apparition de problématiques plus complexes. Nécessité d'acculturation et de formation en Python par exemple
- Stade 4 : Intégration d'outils nouveaux et plus performants

### La gouvernance des données

La gouvernance de données peut se définir en 5 points :

- Les data contracts : définir la donnée et son emplacement
- Définir les KPIs principaux
- L'accès aux données : les "data catalog"

- La sécurité : qui a accès à quoi ?
- La compliance (RGPD)

### Quelles sont les roadmaps d'un projet data ?

- Roadmap d'un projet **data sciences** : prédiction de volume pour la supply chain
  - Cadrer votre projet
  - Développer vos algorithmes
  - Déployer vos modèles
  - Piloter vos résultats
- Roadmap d'un projet **data analyse** : manque de visibilité sur les habitudes des consommateurs
  - Cadrer votre projet (5 Why's)
  - Analyser, récupérer et agréger vos données
  - Créer votre dashboard et restituer vos analyses
- Roadmap d'un projet **data engineer** : l'équipe finance n'a pas de visibilité sur certains indicateurs
  - Identifier où est la donnée
  - Mettre en place un data contract
  - Récupérer et transformer la donnée
  - Monitorer la performance du projet

### Comment devenir data driven ?

- Identifier des cas d'usage
- Former des ambassadeurs
- Former et acculturer vos collaborateurs à la data

### Comment gérer les interactions des équipes tech et non tech ?

- Enjeu de vocabulaire : il faut parler la même langue
- Adapter la complexité des analyses et des recommandations à son public
- Instaurer des data hero et des process

### Comment gérer une équipe de data analytics ?

- 2 flows de travail
  - Des gros projets
  - Des analyses ad hoc avec un système de ticketing

- Les outils comme Trello ou Jira peuvent être utilisés

Commenter recruter des profils data ?

- Matérialiser et valoriser l'impact d'un profil data sur le business
- Etre clair sur le rôle en question et la montée en compétences
- Embaucher des profils curieux et débrouillards

Conseils pour se lancer dans la data

The Modern Data Stack récap

PDF

Grille de lecture - Data Maturity


PDF

Data Maturity récap

PDF

## JOUR 2M- Infra

🔖 Objectifs

Pour ce module c'est [Alexandre](#), cofondateur et directeur pédagogique de Databird , qui va t'expliquer les infrastructures data modernes.

Via ce module, tu vas donc :

- apprendre de quoi est constitué une infrastructure data moderne
- apprendre quelle est le rôle de chaque outils utilisé
- quelles sont les grandes tendances de marchés sur "l'infra" et le "tooling" data

Support de la présentation

PDF Culture tech et data



## Intro - Modern data Stack

- Une infrastructure data moderne permet de centraliser des données éparses (internes / d'évènements / externes) via des ETL (ou ELT)
- Afin de fédérer les données on peut utiliser des APIs qui vont agir comme des connecteurs de données
- Une fois les données centralisées dans un data warehouse, des analyses peuvent être faites, soit en SQL, soit via des plateformes de data analytics ou de dashboarding

## Pourquoi de nouvelles infrastructures data ont-elles émergé ? 🤔

4 V's  
0:45

Volume

Velocity

Variety

Veracity



## Rapide coup d'oeil aux infrastructures modernes

1:00



## 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 / ...



### Les sources de données

Nous avons segmenté les sources de données en 3 :

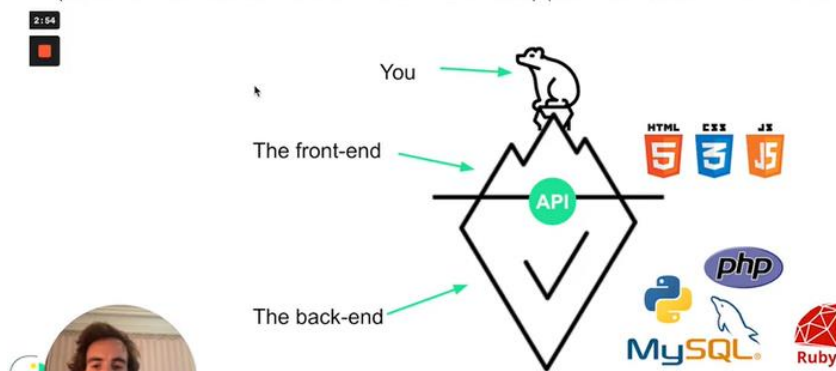
- Les données internes à l'entreprise (compte client / RH / Finance ...)
- Les données d'événements (comportement utilisateur sur un site ou un produit)
- Les données externes (WebScraping ou contenues dans des SaaS)

Les APIs :

- Sont un moyen de communiquer entre deux programmes informatiques
- Agissent comme des agents de sécurité et des traducteurs
- Suivent les principes REST à 80% d'entre elles

## 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

À 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.

6:58

Comment faire cela rapidement ?

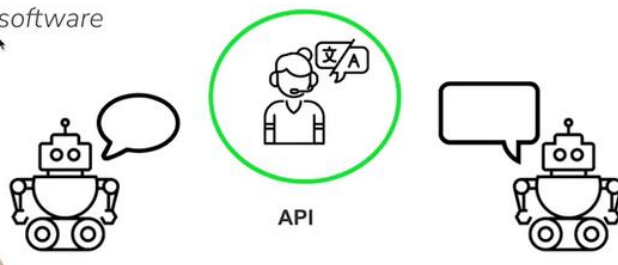


## 1.3.2 Data source : Données externes - API

Qu'est-ce qu'une API ? 📢

8:17

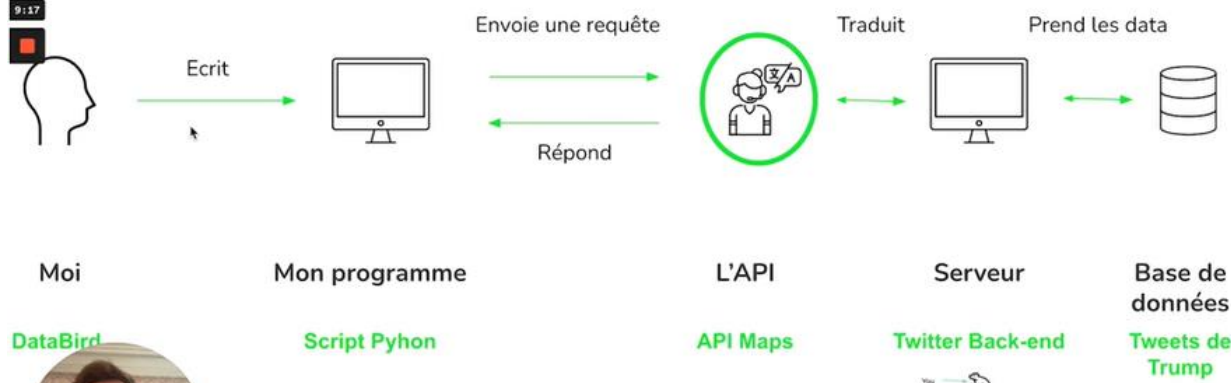
Une API (application programming interface) est un moyen de communication entre deux logiciels



## 1.3.2 Data source : Données externes - API

Qu'est-ce qu'une API ? 📢

9:17



## 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 rédigées strictement indépendamment du client et du serveur



## Les APIs

- GET permet de récupérer de l'info dans la base de données qu'on interroge
- POST permet de créer une nouvelle ligne dans la base de données qu'on interroge
- PUT permet de mettre à jour une ligne dans la base de données qu'on interroge
- DELETE permet de supprimer une ligne dans la base de données qu'on interroge

## 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



### APIs Partie 2

Les sites internet et outils utilisés :

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

<https://www.pappers.fr/api/documentation#tag/Recherche>

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

La requête API utilisée sur **PostMan** :

[https://api.pappers.fr/v2/recherche?api\\_token=d6ef957beb40690c8884e6cae24967d71068f7dbbb36bfcd&siren=883538100](https://api.pappers.fr/v2/recherche?api_token=d6ef957beb40690c8884e6cae24967d71068f7dbbb36bfcd&siren=883538100)

Key takeaways :

- Un outil comme Postman permet d'interroger des APIs
- Le Token fourni vous identifie
- Des paramètres supplémentaires peuvent être fournis
- Les APIs utilisent le protocole HTTP pour envoyer de l'info et souvent des données au format JSON



Il faut aller dans la documentation : API → documentation : <https://www.pappers.fr/api/documentation>

Pappers

entreprise >

association >

arche >

estions / Autocomplete >

tes annuels >

ments >

Pappers API (2.5.0)

Download OpenAPI specification:

Download

Support technique Pappers: support@pappers.fr

L'API de Pappers vous permet de récupérer des informations et documents sur les entreprises françaises à partir de leur numéro SIREN ou SIRET.

Vous devez indiquer votre clé d'API dans les requêtes, en utilisant le paramètre `api_token`.

L'URL d'accès à l'API est <https://api.pappers.fr/v2/>

Lien vers la documentation de la V1 : <https://www.pappers.fr/api/documentation/v1>

L'historique des modifications (changelog) est accessible à l'uri suivante :  
<https://www.pappers.fr/api/changelog>

Fiche entreprise

Récupère l'ensemble des informations disponibles sur une entreprise.

Vous devez fournir soit le SIREN, soit le SIRET. Si vous indiquez le SIREN, tous les établissements associés à ce SIREN seront renvoyés dans la clé `etablissements`. Si vous indiquez le SIRET, seul l'établissement associé sera renvoyé dans la clé `etablissement`.

QUERY PARAMETERS

api\_token

string

GET /entreprise

Response samples

200

Content type  
application/json

Copy Expand all Collapse all

- association >
- archie >
- estions / Autocomplete >
- tes annuels >
- ments >

Lien vers la documentation de la V1 : <https://www.pappers.fr/api/documentation/v1>

L'historique des modifications (changelog) est accessible à l'url suivante : <https://www.pappers.fr/api/changelog>

## Fiche entreprise

### 🔗 Récupère l'ensemble des informations disponibles sur une entreprise.

Vous devez fournir soit le SIREN, soit le SIRET. Si vous indiquez le SIREN, tous les établissements associés à ce SIREN seront renvoyés dans la clé `etablissements`. Si vous indiquez le SIRET, seul l'établissement associé sera renvoyé dans la clé `etablissement`.

#### QUERY PARAMETERS

|  |  |
|--|--|
| <div>api_token</div> <div>required</div> | <div>string</div> <div>Exemple: <code>api_token=votre_clé_ici</code></div> <div>Clé d'utilisation de l'API</div> |
| <div>siren</div>                         | <div>string</div> <div>Exemple: <code>siren=443061841</code></div> <div>SIREN de l'entreprise</div>              |
| <div>siret</div>                         | <div>string</div> <div>Exemple: <code>siret=44306184100047</code></div> <div>SIRET de l'entreprise</div>         |

entreprise

Récupère l'ensemble des informations disponibles sur une entreprise.

association

archie

estions / Autocomplete

xtes annuels

ments

une entreprise.

Vous devez fournir soit le SIREN, soit le SIRET. Si vous indiquez le SIREN, tous les établissements associés à ce SIREN seront renvoyés dans la clé `etablissements`. Si vous indiquez le SIRET, seul l'établissement associé sera renvoyé dans la clé `etablissement`.

QUERY PARAMETERS

- `api_token` (required) string  
Exemple: `api_token=votre_clé_ici`  
Clé d'utilisation de l'API
- `siren` string  
Exemple: `siren=443061841`  
SIREN de l'entreprise
- `siret` string  
Exemple: `siret=44306184100047`  
SIRET de l'entreprise
- `format_publications_bodacc` string  
Enum: "objet" "texte"  
Exemple: `format_publications_bodacc=objet`  
Format attendu pour les publications BODACC. Valeur par défaut: "objet".
- `marques` boolean  
Si vrai, le retour inclura les marques éventuelles de l'entreprise. Valeur par défaut: `false`.

Responses

Response samples

200

Content type: application/json

Copy Expand all Collapse all

```
{
  "siren": "443061841",
  "siren_format": "443 061 841",
  "nom_entreprise": "GOOGLE FRANCE",
  "personne_morale": true,
  "denomination": "GOOGLE FRANCE",
  "nom": null,
  "prenom": null,
  "sexe": null,
  "code_naf": "70.102",
  "libelle_code_naf": "Activités des sièges",
  "domaine_activite": "Activités des sièges",
  "conventions_collectives": [
    {
      "code": "1515",
      "libelle": "Convention collective de la métallurgie"
    }
  ],
  "date_creation": "2002-05-16",
  "date_creation_format": "16/05/2002",
  "entreprise_cessee": true,
  "date_cessation": "2002-05-16",
  "entreprise_employeuse": true,
  "categorie_juridique": "5499",
  "forme_juridique": "Société à responsabilité limitée",
  "effectif": "Entre 500 et 999 salariés",
  "effectif_min": 500,
  "effectif_max": 999
}
```

Et sur le **PostMan** ca donne ca ;

HubSpot

New Collection

GET

https://api.pappers.fr/v2/recherche?api\_token=d6ef957beb40690c8884e6cae24967d71068f7dbb36b&par\_page=2&siren=883538001

Send

Params

Authorization

Headers (6)

Body

Pre-request Script

Tests

Settings

Cookies

Query Params

| KEY   | VALUE  | DESCRIPTION |
|---|--|-------------|
| <input checked="" type="checkbox"/> api_token | d6ef957beb40690c8884e6cae24967d71068f7dbb36b |             |
| <input checked="" type="checkbox"/> par_page  | 2  |             |
| <input checked="" type="checkbox"/> siren     | 883538001                                    |             |
| Key   | Value  | Description |

Body

Cookies

Headers (16)

Test Results

Status: 200 OK Time: 483 ms Size: 2.24 KB Save Response

Pretty

Raw

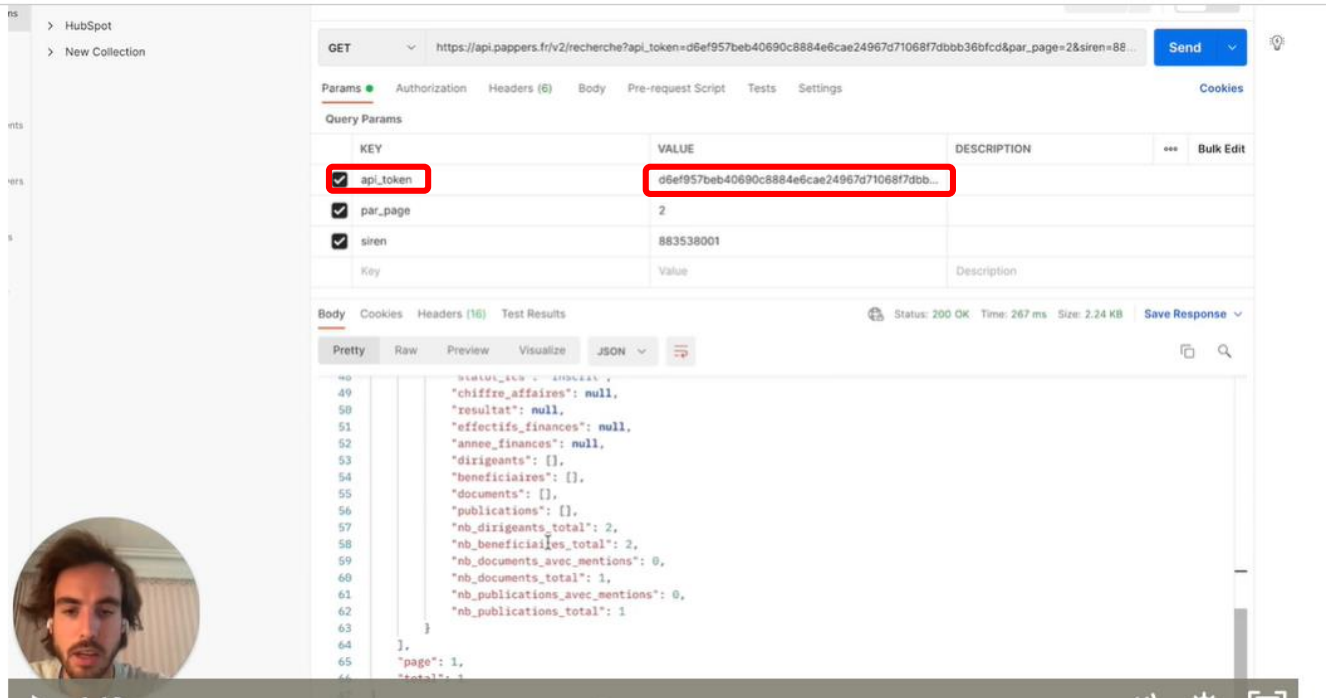
Preview

Visualize

JSON

```
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
{
  "resultats": [
    {
      "siren": "883538001",
      "siren_format": "883 538 001",
      "nom_entreprise": "DATABIRD",
      "personne_morale": true,
      "denomination": "DATABIRD",
      "nom": null,
      "prenom": null,
      "sexe": null,
      "siege": {
        "siret": "88353800100016",
        "siret_format": "883 538 001 00016",
        "nic": "00016",
        "numero_voie": 10,
        "indice_repetition": null,
        "type_voie": "RUE"
      }
    }
  ]
}
```





GET https://api.pappers.fr/v2/recherche?api\_token=d6ef957beb40690c8884e6cae24967d71068f7dbb36bfd&par\_page=2&siren=88...

Query Params

| KEY       | VALUE  | DESCRIPTION |
|-----------|--|-------------|
| api_token | d6ef957beb40690c8884e6cae24967d71068f7dbb... |             |
| par_page  | 2  |             |
| siren     | 883538001                                    |             |

Body

```

{
  "chiffre_affaires": null,
  "resultat": null,
  "effectifs_finances": null,
  "annee_finances": null,
  "dirigeants": [],
  "beneficiaires": [],
  "documents": [],
  "publications": [],
  "nb_dirigeants_total": 2,
  "nb_beneficiaires_total": 2,
  "nb_documents_avec_mentions": 0,
  "nb_documents_total": 1,
  "nb_publications_avec_mentions": 0,
  "nb_publications_total": 1
}

```

## Les APIs de SaaS

- Tous les SaaS qu'une entreprise utilise possède une API
- Interroger l'API de votre SaaS vous permet de récupérer de manière automatique toute la donnée qu'il contient
- Beaucoup de logiciels d'ETL se résument maintenant à des connecteurs d'APIs

### 1.3.2 Data source : Données externes - API

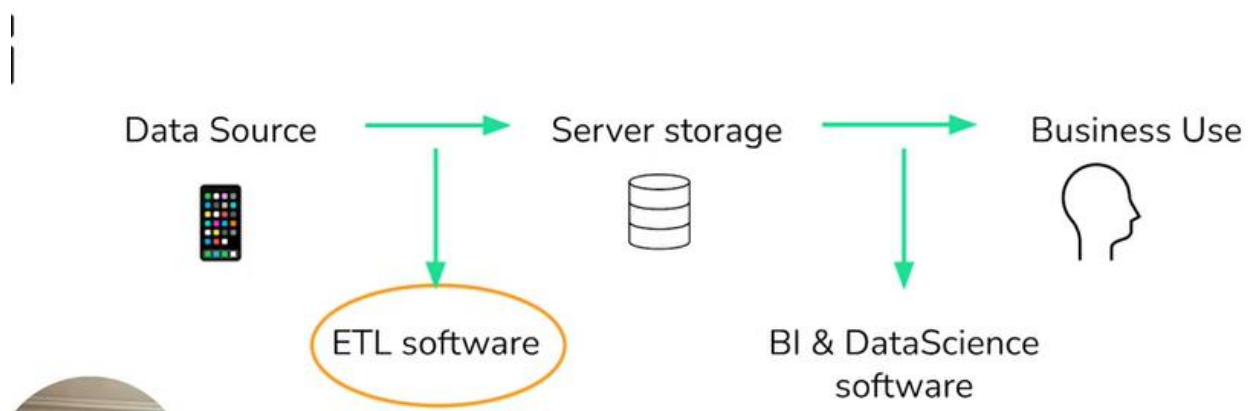
Logiciels tiers : lesquels ? Pour quelles données ?



## Les ETL

- On passe d'ETL à ELT dû aux coûts de stockage
- Dans un monde idéal, c'est le consommateur de la donnée qui va la modéliser car personne ne sait mieux que lui ce dont il a besoin
- Des logiciels comme DBT permettent de faire des transformations de données en SQL (à la place de Python, Java ou Scala)

## 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 possibles

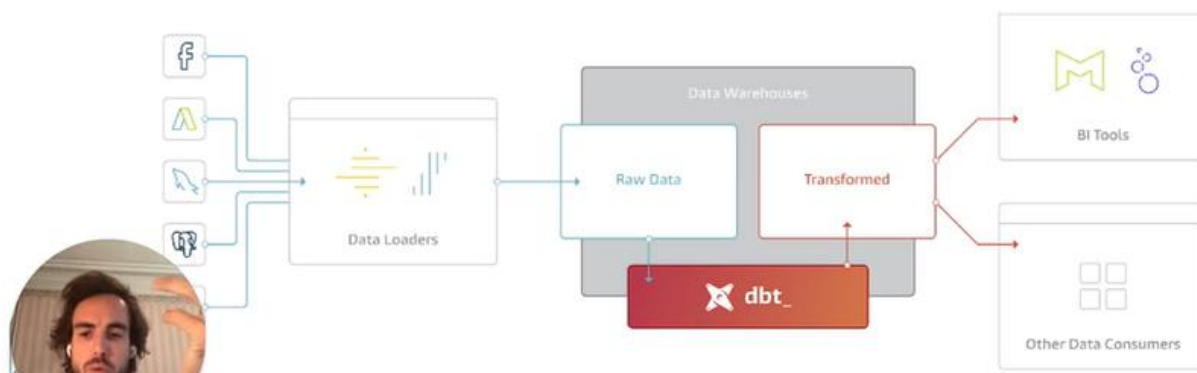
## 2. Extract - Load - Transform data ❤️

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

2:18



→ Souvent combiné avec les connecteurs mentionnés la slide précédente



### Le stockage

- Un data lake permet de stocker toute sorte de données
- Un data warehouse est à la fois un endroit de stockage et un super calculateur
- **Airflow** permet d'orchestrer toutes les opérations

## 3.0 Stocker ses données : version cheap

Base de données très simple

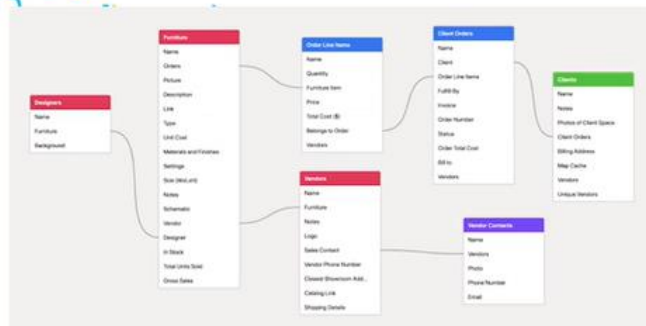
0:32



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



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

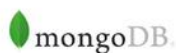
Pourquoi avoir ses données sur le cloud ?

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



Google  
BigQuery



Copyright © DataBird, All rights reserved.

## 3.3 Airflow en chef d'orchestre

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



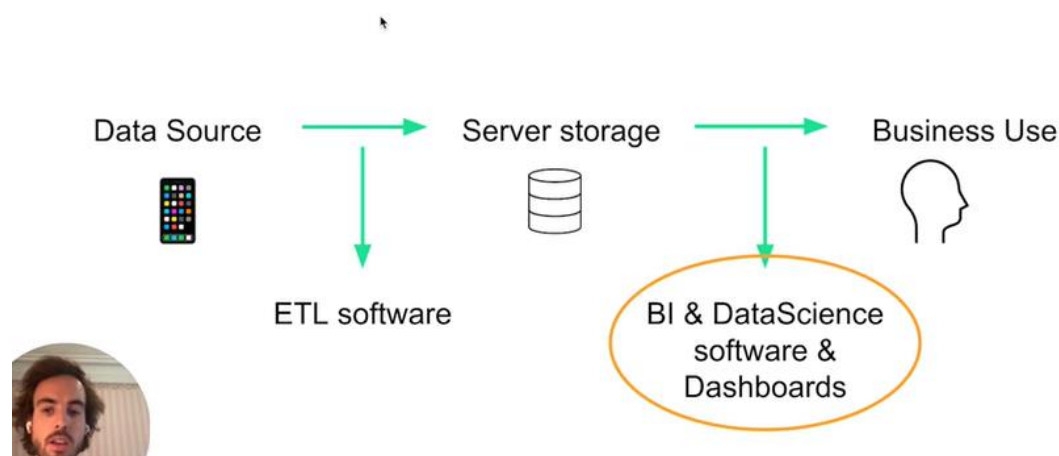
Copyright © DataBird, All rights reserved.



## Analytics Platform

- Les plateformes de data sciences permettent de simplifier et fluidifier le travail des data scientists
- Les logiciels de dashboarding se connectent directement au cloud et se mettent à jour automatiquement, ce qui permet un reporting plus facile
- C'est un marché dynamique et qui a vocation à encore grandir

## 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.



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



+ a b l e a u : très intuitif avec une logique de "drag & drop"



: appartient à Microsoft et assez customizable



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

## 5. Un marché ultra dynamique

45

|  |  |                                |
|--|--|--------------------------------|
| 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                               | prestodb                       |
| ETL → ELT                                    | Brittle ETL processes (extract-transform-load) are being replaced with more flexible and consistent ELT pipelines (extract-load-transform)   | etltran 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 PRIVA@ERA             |



## Recap

Pour résumer ce module, lis ces 3 articles (classés par ordre d'importance) afin de bien comprendre les outils évoqués !

- <https://technically.dev/posts/what-your-data-team-is-using>
- <https://towardsdatascience.com/modern-unified-data-architecture-38182304afcc>
- <https://medium.com/castor-app/what-if-you-had-to-build-a-data-stack-from-scratch-9700c8ec558c>

## JOUR 2A

PDF : Cas\_Macro\_ParisCode-220207-121217