Atelier Collecte de données Outils

Animé par Sylvain Labasse

Au terme du module, vous saurez…

Enumérer les méthodes de collecte d'information

Identifier les sources répondant à un besoin

Utiliser Python pour constituer un jeu de données

Importer des jeux de données de toute source

Pre-Requis

Public

Développeuses/développeurs

Nécessaire

Formats de données

Fonctionnement des API restful

Python

DEROULEMENT

Support

Drive : Diapos, ateliers, vidéos, …

Notes de cours, Réalisation des ateliers

Drive : https://bit.ly/3ZIZ88k

i majuscule

Evaluations

Ateliers sur 4

Quiz individuel

0	1	2	3	4
Non remis	Hors sujet	<	=	>
		attentes		

Figure 1 - Barème des ateliers

L'ENVIRONNEMENT

Matériel

Mac / PC sous Windows ou Linux

Connexion Internet

Outils

Editeur/IDE Python

Compte Hugging Face

COLLECTE DE DONNEES (OUTILS)

Enjeux

- Analyse
- Architectures
- Sources

Techniques et outils

- Aspect légal
- Formats, API
- Scrapping

ENJEUX

OBJECTIFS

Besoins des entreprises en collecte de données

Architectures de collecte et mise à disposition

Sources de données publiques et privées

Enjeux

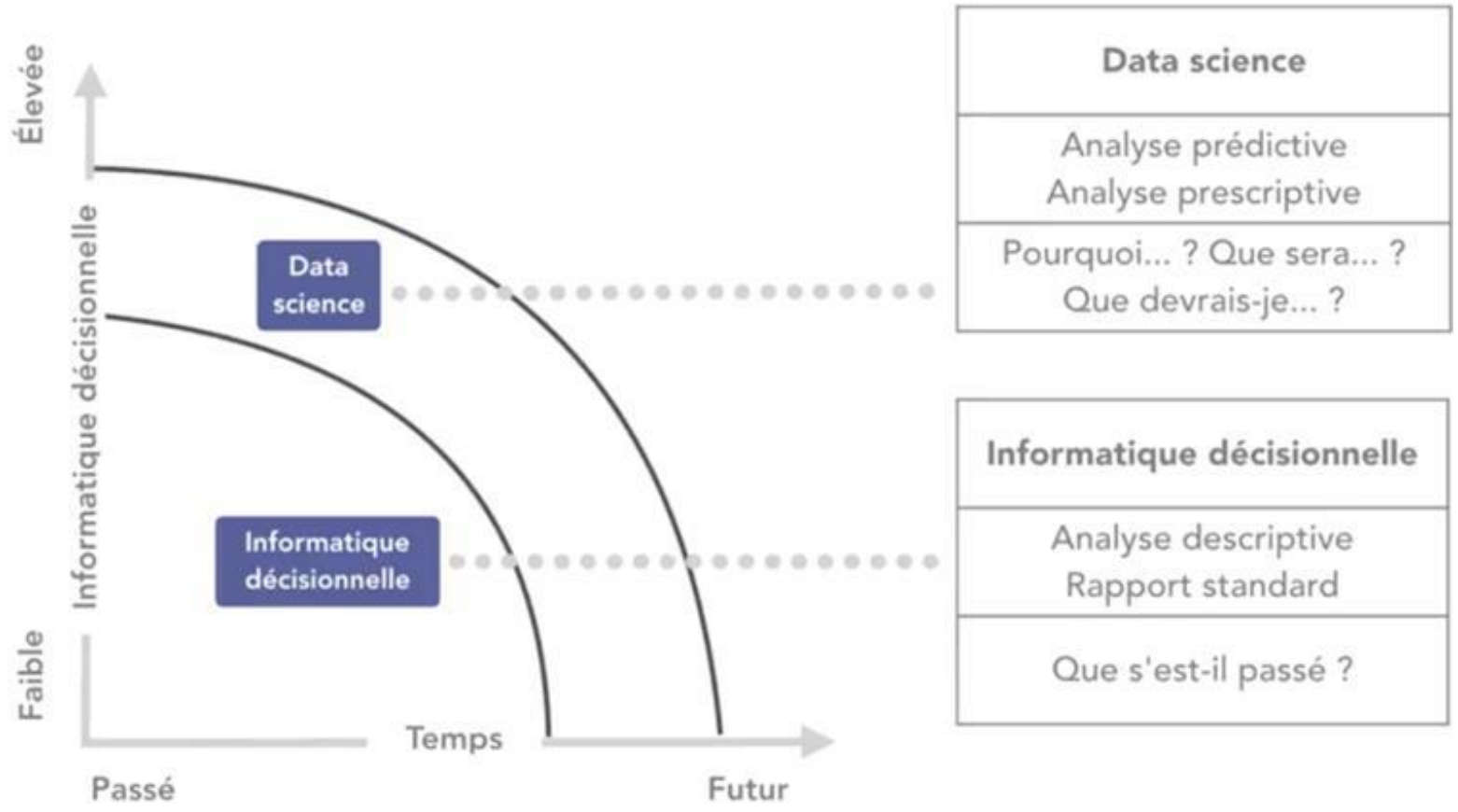
➔ Analyses

Architectures

Sources

Synthèse

ANALYSES


Élevée
Informatique décisionnelle
Faible
Data science
Informatique décisionnelle
Temps
Passé
Futur
Data science
Analyse prédictive
Analyse prescriptive
Pourquoi... ? Que sera... ?
Que devrais-je... ?
Informatique décisionnelle
Analyse descriptive
Rapport standard
Que s'est-il passé ?

Business Intelligence / Decisionnel

Décisionnel ≠ Opérationnel

Tableau de bord activité / marché

Interne ET externe

Prise de décision

Figure 2 - src : https://www.laboiteverte.fr/21-cockpits-davions/

Mise en œuvre

Inventaire et synthèse des données de toutes les applis

Agrégation selon de nombreuses dimensions (temps, géo., …)

Contraintes : Rapide, qualifié, fiable, …

Machine Learning

But des jeux de données

Entrainement

Test/Validation

Figure 3 - Autopilot Tesla – Démo novembre 2016

Caractéristiques

Discret / Continu

Littérales / Numériques

Pertinentes (lien), Représentatives, Complètes

Nombreuses et variées

ENJEUX

✓ Analyses

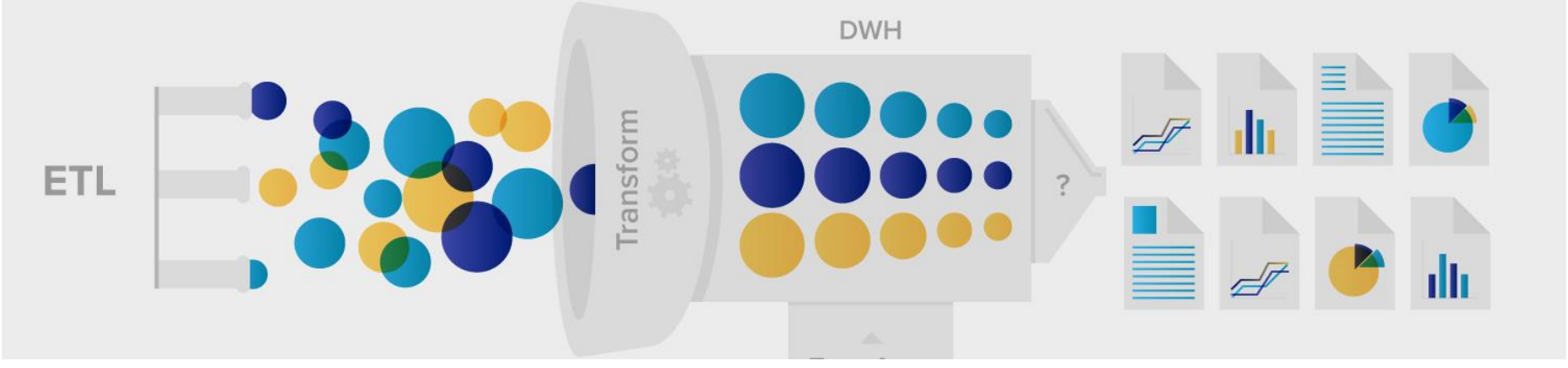
➔ Architectures

Sources

Synthèse

ARCHITECTURE BI CLASSIQUE

Chaîne



Src : https://www.xplenty.com/blog/etl-vs-elt/

Composants

ETL : Extract Transform Load = 80% de l'effort

DataWarehouse : Entrepôt avec magasins/domaine (datamarts)

Requêteurs ou Rapports = Dataviz

Extract Transform Loading

Extraction

Ciblage/profiling des données

Capture des valeurs ET modifications

Nettoyage et normalisation

Nettoyage, suivi des erreurs

Dédoublonnage et mise en conformité

Livraison et fourniture

Créateur de faits, dimensions, cubes

Générateur de clé de substitution

Outils

ETL “classiques”

Microsoft SSIS, Oracle DI, Talend, Pentaho DI, OpenRefine, ...

Cloud : AirByte, AWS Glue, Azure Fabric, Google DataFlow, ...

Langages

Python : dbt, Pandas (T)

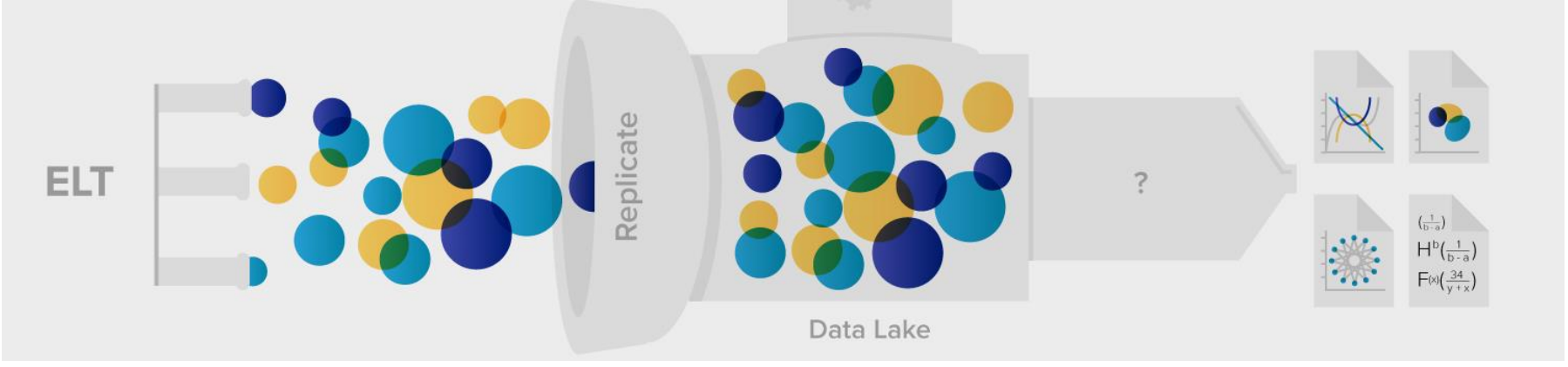
R : dplyr (E), tidyr (T)

Intégrés

Excel/PowerBI, Google Data Studio, …

ELT

Chaîne



Src : https://www.xplenty.com/blog/etl-vs-elt/

Composants (type ELK)

E : Collecte avec traitement minimal (Beats, LogStash)

Data Lake : Stockage bigdata faiblement structuré

Requêteurs ou Rapports (Elastic, Kibana)

« ETL TEMPS REEL »

Flux

Stockage de flux de données, tolérance de panne (réparti)

Publish/Suscribe : File de Message, RabbitMQ

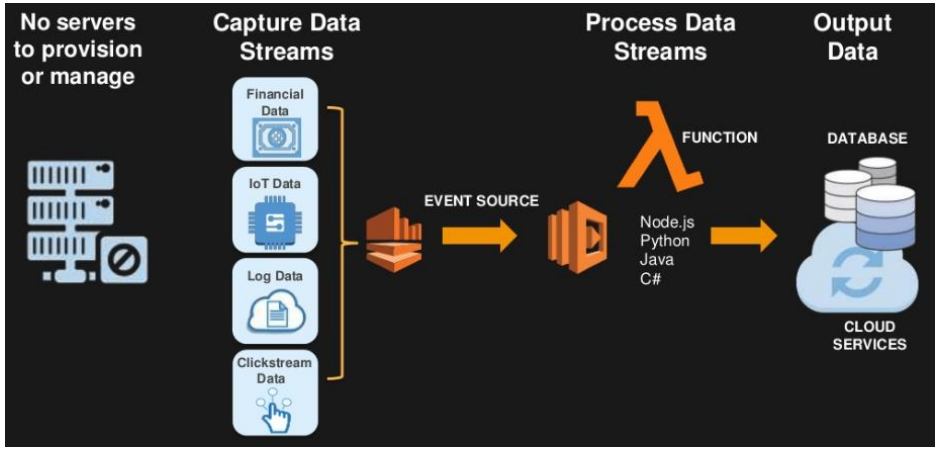
Producteur/consommateur : Apache Kafka

Serverless

Déclenché par évènement

Traitement éphémère sans état

AWS lambda, Azure functions, …



ENJEUX

- ✓ Analyses
- ✓ Architectures
- ➔ Sources
- Synthèse

SOURCES PRIVEES

Patrimoine informationnel

Fichiers, bases de données, applications

Prestataires de données / Numérisation

Utilisateurs

Manuel : Formulaires, saisie (peu normalisé)

Automatique : Carte fidélité, analytics, app, IoT, service gratuit

Ludification[1] : BLAP vs RECIPE

[1] https://scottnicholson.com/pubs/recipepreprint.pdf : Badge, Level/Leaderboard, Achievements, Points vs Recul, Exposition, Choix, Info, Play, Engagement

SOURCES PUBLIQUES

Open data (BI)

Nations : data.gouv.fr, data.europa.eu, data.gov, …

Organisations : data.worldbank, census.gov, …

Datasets (ML)

Répertoires : Liste dans Wikipedia[2]

Moteurs : Google data search, Kaggle, VisualData

Universités/Recherche : UCI, CMU, Laion (240TB laion-5b)

[2] https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Enjeux

- ✓ Analyses
- ✓ Architectures
- ✓ Sources
- ➔ Synthèse

RESUME

Besoins des entreprises en collecte de données

Architectures de collecte et mise à disposition

Sources de données publiques et privées

TECHNIQUES ET OUTILS

OBJECTIFS

Droits et contraintes légales sur les données

Formats de mise à disposition « prêts à l'emploi »

Contraintes et bonnes pratiques de scraping

TECHNIQUES ET OUTILS

➔ Aspect légal

Formats, API

Scrapping

Synthèse

Cadre

Copyright / Droit « sui generis » des bases

Domaine publique (~> 70 ans) / CC0

EU[3] : Exceptions pour institutions publiques sans but lucratif

US : « Fair Use » favorable à l'utilisation technique

RGPD / Ccpa

Accord explicite pour le traitement de données personnelles

Solution : Anonymisation[4]

[3] https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32019L0790&from=EN#d1e976-92-1

[4] https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

JURISPRUDENCE

Etats-Unis

Linkedin/**hiQ Labs** (2019) : Plusieurs cours dont cour suprême
Autorise scraping de données publiques (malgré CGU)

UE

Ryanair/**Opodo** (2010) : Transf. suffisante, invest. substantiel
LeBonCoin/EntreParticuliers (2021) : Invest. subst. du plaignant

TECHNIQUES ET OUTILS

✓ Aspect légal

➔ Formats, API

Scrapping

Synthèse

Fichiers

Texte standard

CSV, INI, …

XML avec notion de schéma XSD et transformation XSL

JSON, YAML

Bureautique

Excel (Open XML), ODS (Open Document = XML)

Formulaire PDF (AcroForms)

Bdd fichier : Access, SQLLite, …

SERVEUR DE DONNEES

Relationnel

SQL : tables, colonnes, enregistrements

Accès centralisé, structure rigide

Ex : MySQL, Oracle, SQLServeur, Snowflake

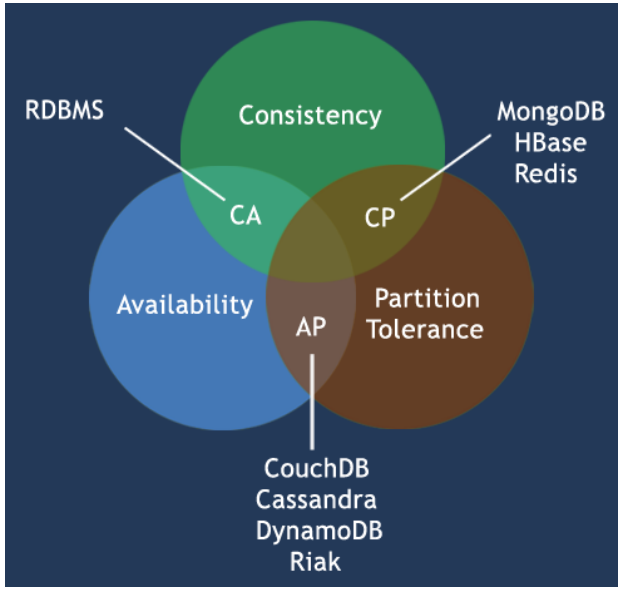


Figure 4 - Théorème CAP (http://www.abramsimon.com)

NoSQL

Orientés : Document, colonne, graphe, objets, flux (dsms)

3V : Volume, Variété, Vitesse

Ex : MongoDB, CouchDB, HBase, Neo4j

APPLICATIF

RPC et ORB

Appel de composants applicatifs à distance

Procédures : Remote Procedure Call

Objet : Object Request Broker : Corba, DCOM (Microsoft)

Services Web

2000 – Service Web XML (W3C) : SOAP, WSDL et UDDI

2010 – API Rest : http, json/xml + WebHooks

2015 – GRPC basé sur HTTP2, IDL Protobuf

DONNEES NON STRUCTUREES

Texte

Métadonnées : Source, titre, auteur, …

Contenu : Structure (html), NLP (cf Apache Open NLP)

Media

Métadonnées : XMP, EXIF / Dates, Géolocalisation, Auteur, Tag

Image : Dimensions, couleurs, prétraitements

Son : Taux d'échantillonnage, canaux, durée

TECHNIQUES ET OUTILS

- ✓ Aspect légal
- ✓ Formats, API
- ➔ Scrapping
- Synthèse

SCRAPING

Mise en œuvre

Python - BeautifulSoup[5]

Ethique : robots.txt/cgu, stricte nécessaire, rgpd, ~~republication~~

Cloud

Proxy, supervision, captchas, …

Paiement au volume[6]

[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[6] https://research.aimultiple.com/web-scraping-tools/

TECHNIQUES ET OUTILS

- ✓ Aspect légal
- ✓ Formats, API
- ✓ Scrapping
- ➔ Synthèse

RESUME

Droits et contraintes légales sur les données

Formats de mise à disposition « prêts à l'emploi »

Contraintes et bonnes pratiques de scraping

BILAN

VOUS SAVEZ MAINTENANT…

Enumérer les méthodes de collecte d'information

Identifier les sources répondant à un besoin

Utiliser Python pour constituer un jeu de données

Importer des jeux de données de toute source

POUR ALLER PLUS LOIN…

DIAE506 - Conception des modèles de données IA

DIAE504 - Machine learning et IA