Chapitre 3

Outils et technologies utilisés en Data Science

Pré[paré-senté] par :
Dr. Bilal Dendani

جامعة باجي مختار - عنابة
BADJI MOKHTAR - ANNABA UNIVERSITY
Dr. DENDANI Bilal

Chapitre 3 : Outils et technologies utilisés en Data Science

- Les outils de stockage de données
- Les outils de préparation de données
- Les outils de visualisation de données
- Les outils IDE notebooks
- Les plateformes complètes de Data science

Introduction

- En science de données, la maîtrise des outils et technologies est essentielle pour transformer les données brutes en informations exploitables.
- Les Data Scientists utilisent un ensemble varié d'outils pour chaque étape du cycle de vie de sciences de données, depuis le stockage jusqu'à l'analyse et la visualisation.
- Ces outils facilitent la manipulation de grandes quantités de données, automatisent les tâches de préparation, et permettent des analyses avancées.

Les outils de stockage de données

Quelle solution de stockage de données choisir ?

Base de données SQL

Données structurées, requêtes complexes, relationnelles.
Exemple : MySQL, PostgreSQL, Oracle

Base de données NoSQL

Données non structurées/semi-structurées, flexibles, big data.
Exemple : MongoDB, Cassandra, Redis

Stockage dans le cloud

Stockage évolutif, flexible, accès à distance, grandes données.
Exemple : Amazon S3, Google Cloud Storage, Azure Blob Storage.

Lac de données (Data Lakes)

Données brutes, structurées/non structurées, post-traitement.
Exemple : Hadoop HDFS, AWS Lake Formation.

Bases de données relationnelles (SQL)

- Utilisent un modèle structuré basé sur des tables et des relations entre elles. Elles sont largement utilisées pour stocker et interroger des données bien organisées.
- Organisent les données en tables avec des colonnes et des lignes, chaque table représentant une entité (par exemple, Clients, Ventes).
- Le langage de requête utilisé est le Structured Query Language (SQL) pour manipuler et interroger les données.
- MySQL, PostgreSQL, Oracle.

Bases de Données NoSQL (Not Only SQL)

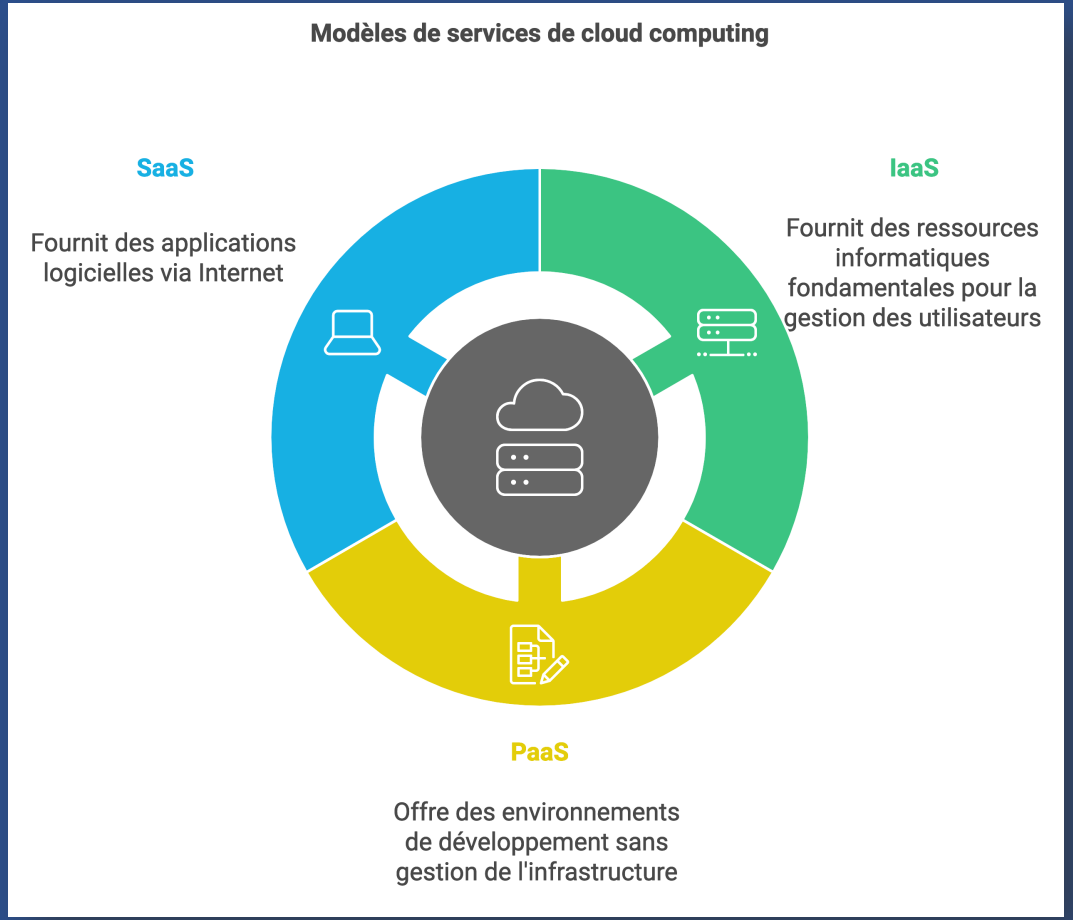
- Offrent une grande flexibilité, permettant de stocker des données non structurées ou semi-structurées, comme les documents JSON, les publications sur les réseaux sociaux ou les données de capteurs de manière plus scalable.
- NoSQL est utilisé pour les données à grande échelle, caractérisé par sa vitesse et sa flexibilité élevées.
- Elles sont conçues pour la scalabilité et la flexibilité, ce qui les rend adaptées aux applications de big data et aux analyses en temps réel.
- Exemples : MongoDB, Redis, Cassandra.

https://www.datacamp.com/blog/nosql-databases-what-every-data-scientist-needs-to-know

Stockage dans le cloud

- Le stockage dans le cloud permet un accès distant, une haute disponibilité et une scalabilité selon les besoins.

- Exemples: Amazon S3, Google Cloud Storage, Azure Blob Storage.



Lacs de Données (Data Lakes)

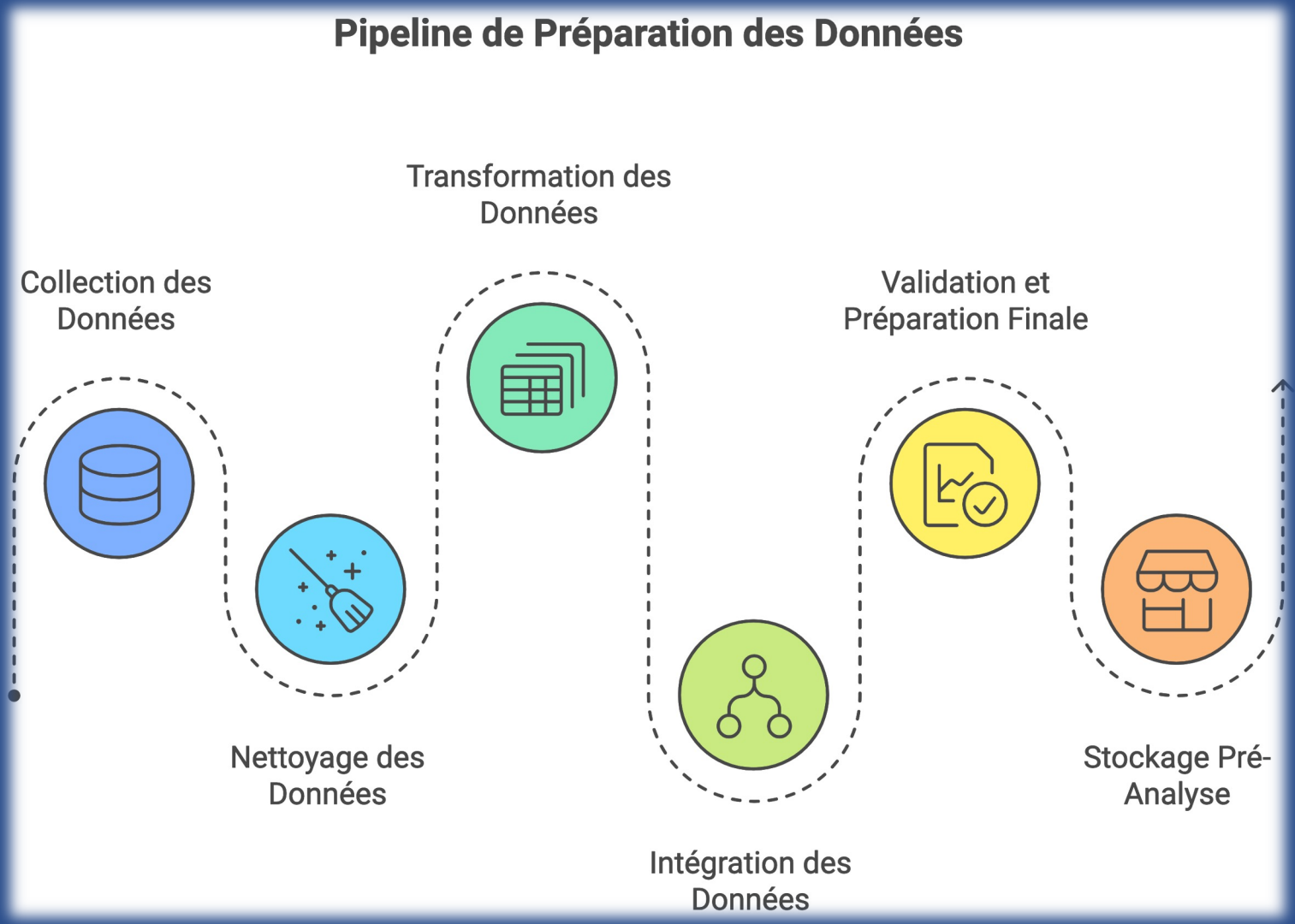
- Les lacs de données sont utlisés pour le stockage de données brutes structurés ou non-structurés.
- Utilisation pour les analyses avancées et le machine learning.

- Technologies associées : HDFS, AWS Lake Formation, Azure Data Lake.

Outils de Préparation de Données en Science des Données

- La préparation des données est une étape clé du processus de science des données.
- Elle vise à nettoyer, transformer, et intégrer les données pour assurer leur qualité et leur utilité.
- Dans ce partie, nous explorerons différents outils utilisés pour chacune de ces étapes

Pipeline de préparation de données



Outils de collect de données

Outils de Collecte de Données

Stockage de Données

Bibliothèques pour lire et traiter des données à partir de divers formats de fichiers.
- Bibliothque Pandas : lire fichier CSV
- Spark : Charger de données de grandes quantités

Bases de Données SQL

Outils pour extraire des données des bases de données relationnelles en utilisant des requêtes structurées.

Outils ETL(Extract, Transform, Load)

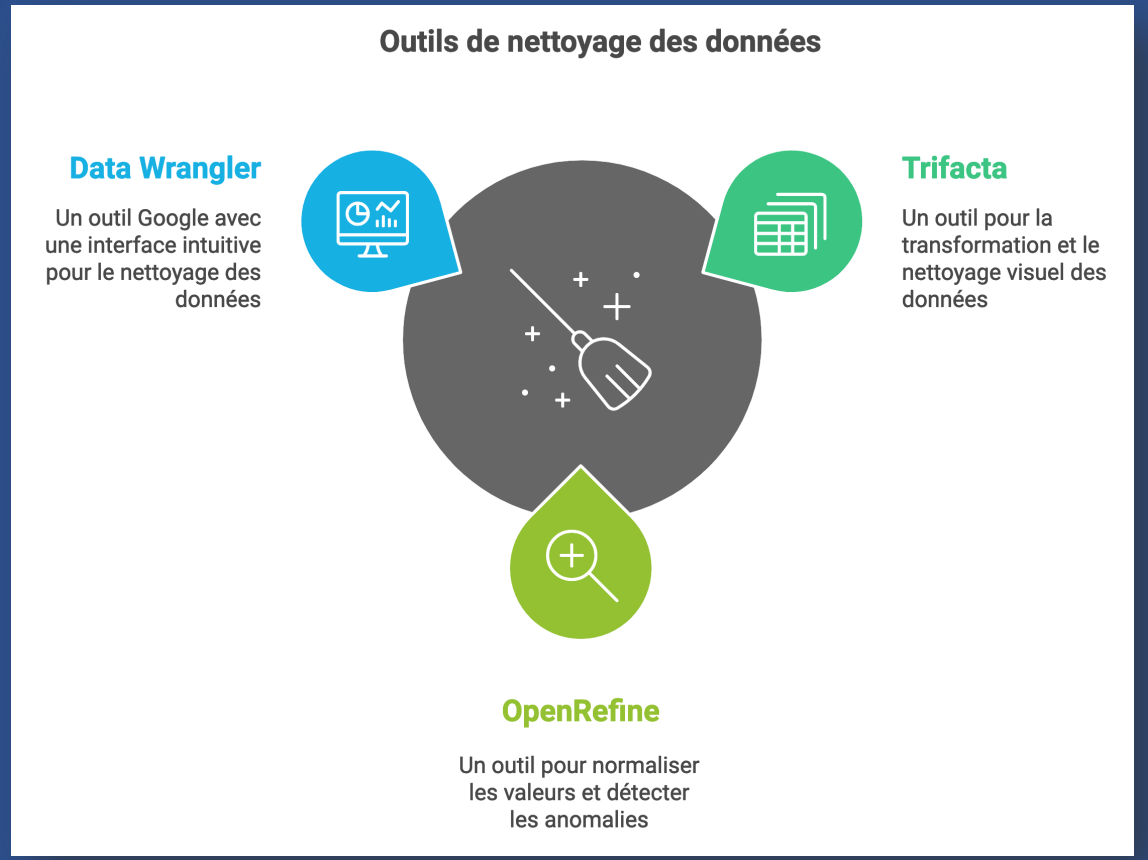
- Logiciels pour automatiser la collecte, la transformation et l'intégration des données.
Outils :Talend, Apache Nifi

APIs

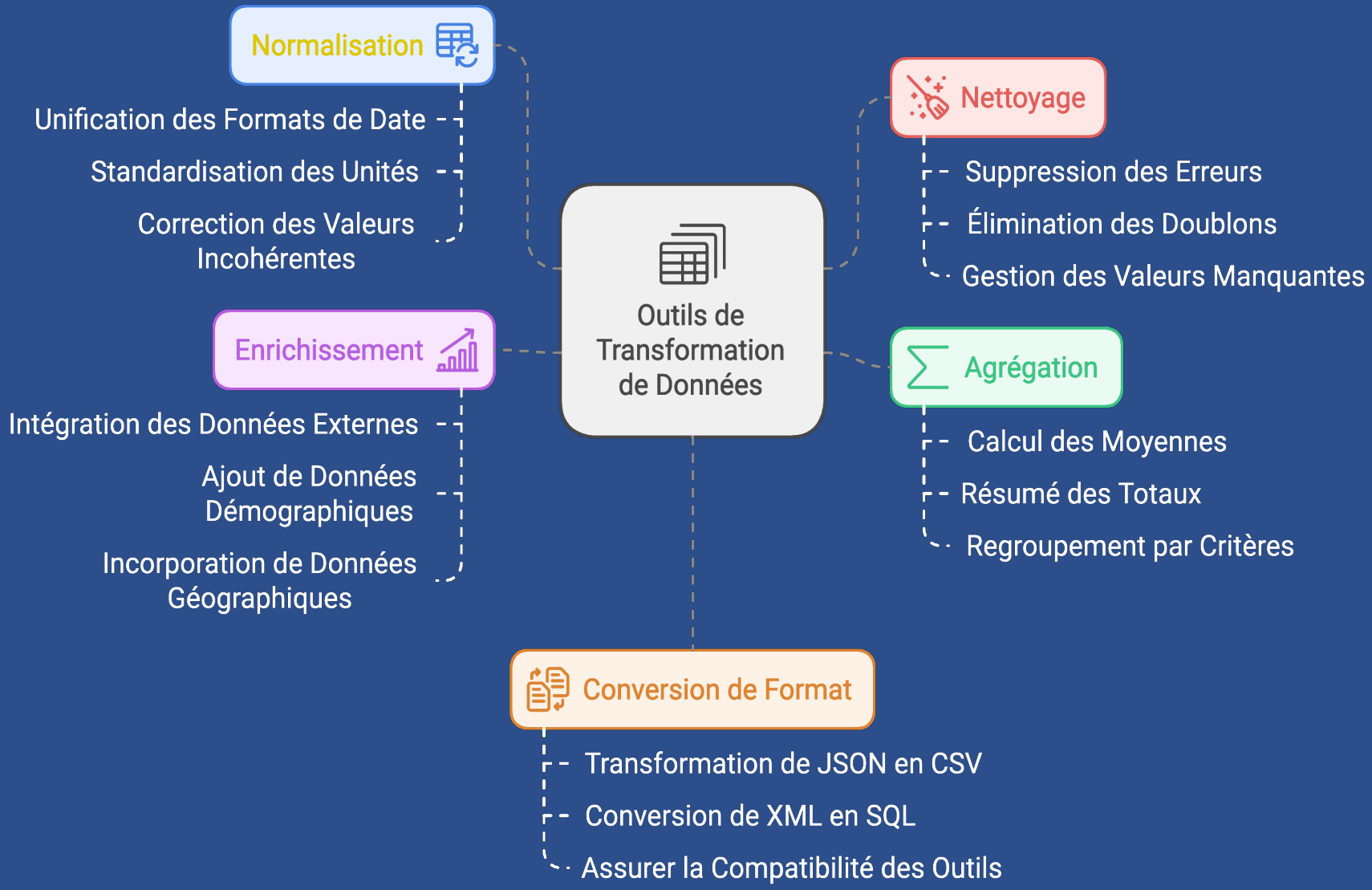
- Bibliothèques pour la collecte de données en temps réel via des requêtes HTTP.
- Bibliothèque Python Request pour collecter des données en temps réels

Outils de Nettoyage des Données

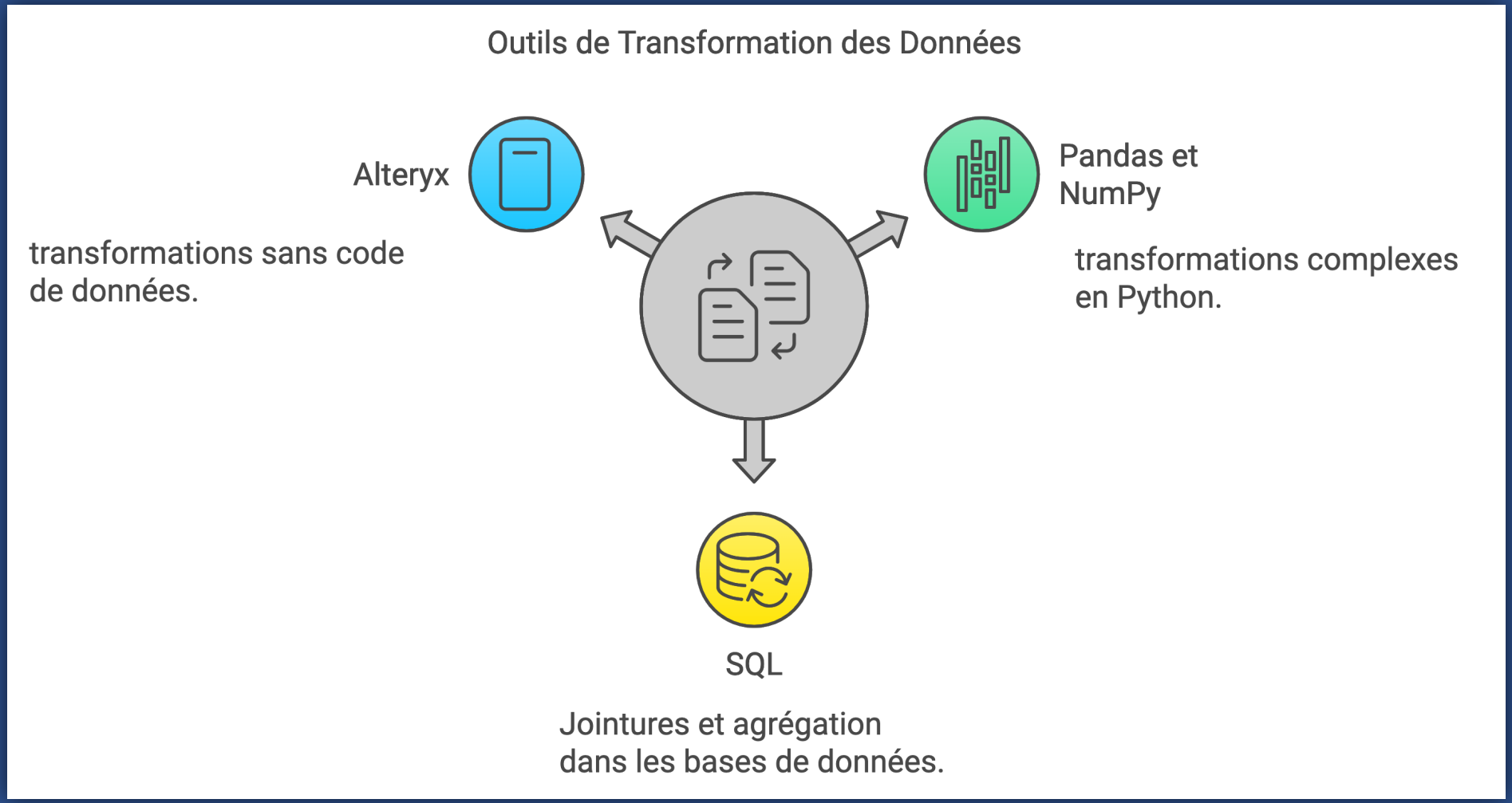
- Le nettoyage de données est une étape essentielle dans le processus de science des données.
- Le nettoyage de données permet d'améliorer la Qualité des Données.
- Le nettoyage permet d'éliminer ou de corriger les valeurs aberrantes ou erronées qui peuvent biaiser les résultats des modèles de machine learning et des analyses statistiques.



Outils de Transformation des Données

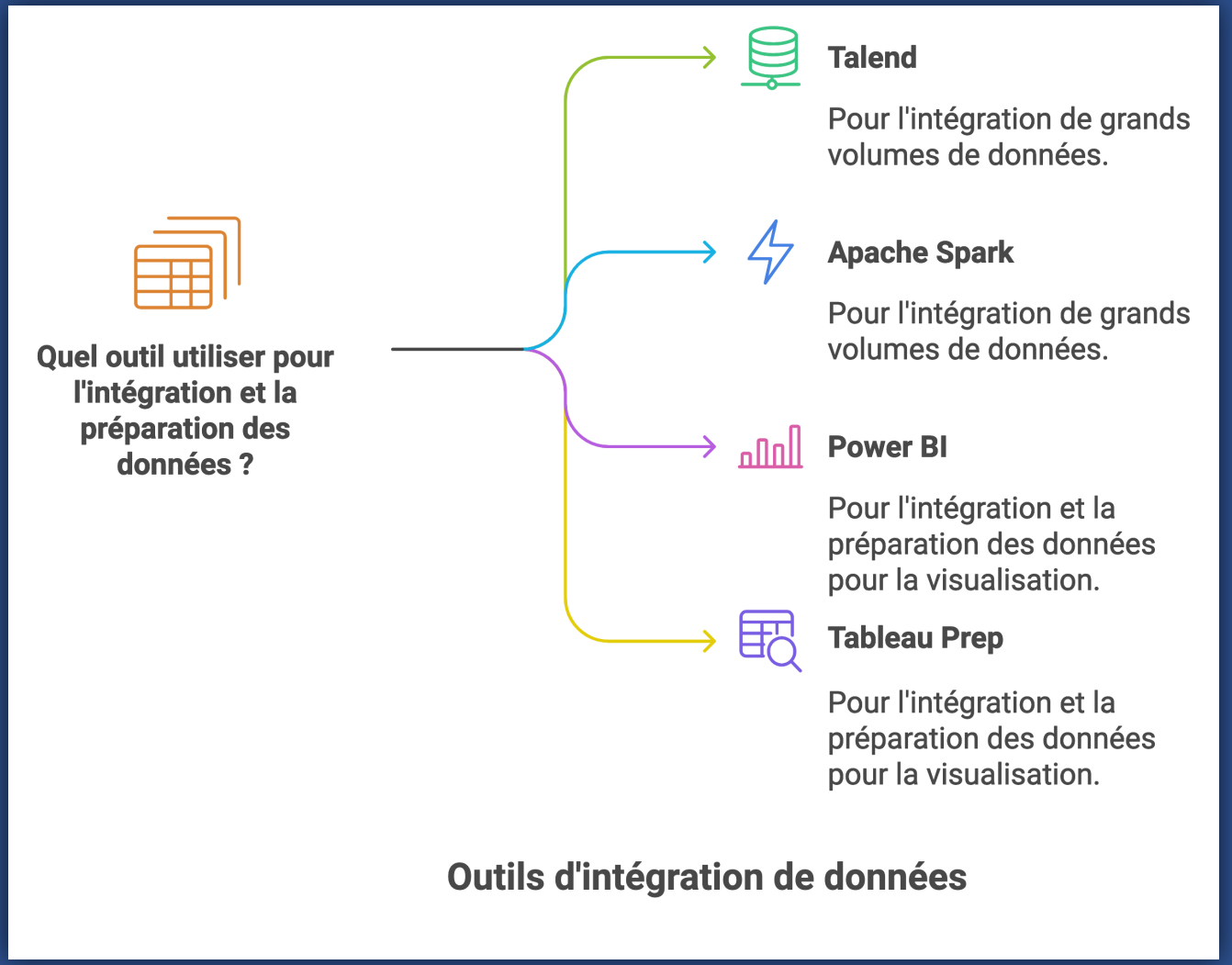


Outils de transformation de données



Outils d’integration de données

- L'intégration de données est le processus de combiner des données provenant de différentes sources pour créer une vue unifiée et cohérente.
- Cela implique la collecte, la transformation et la consolidation des données afin qu'elles puissent être utilisées efficacement pour l'analyse et la prise de décision.



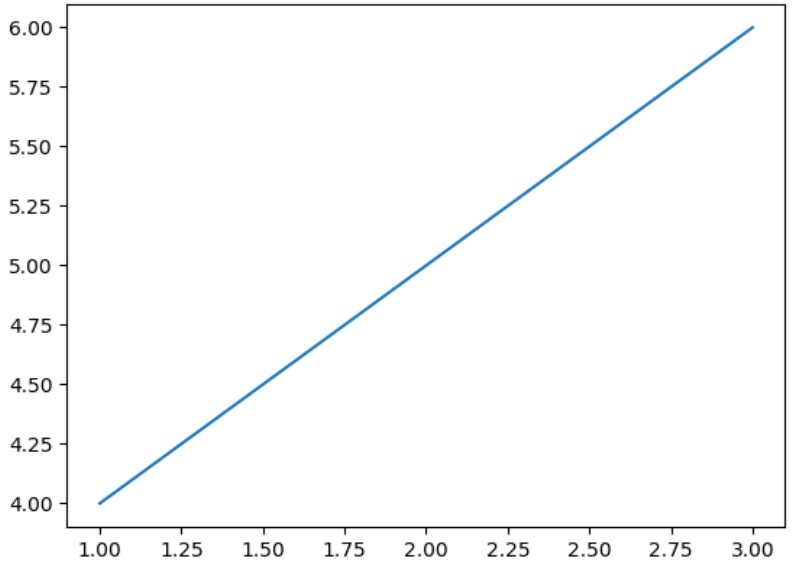
Outils de visualisation de données

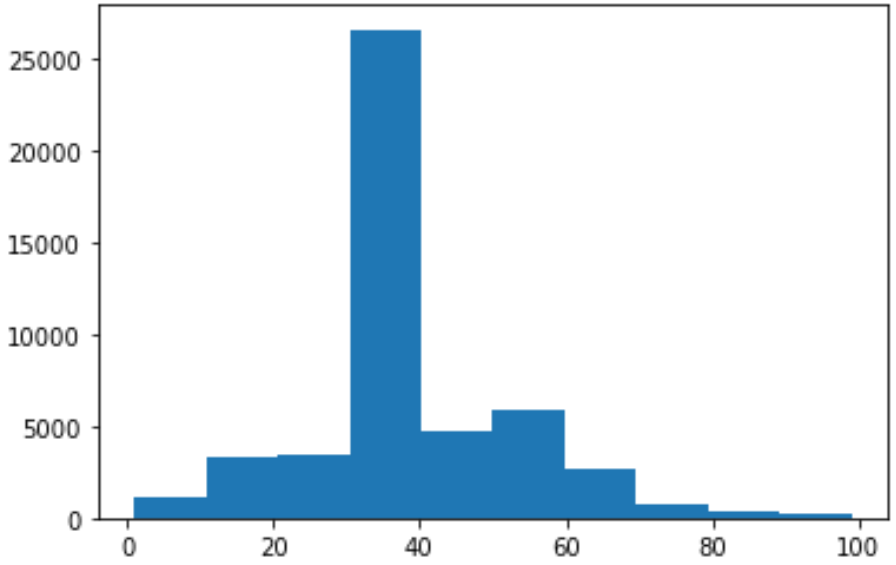
- En data science, les outils de visualisation de données jouent un rôle crucial pour explorer les données, identifier les tendances et communiquer les résultats. Voici les principaux outils utilisés, classés en fonction de leur type et de leurs usages :
- 1. Bibliothèques de visualisation en Python :

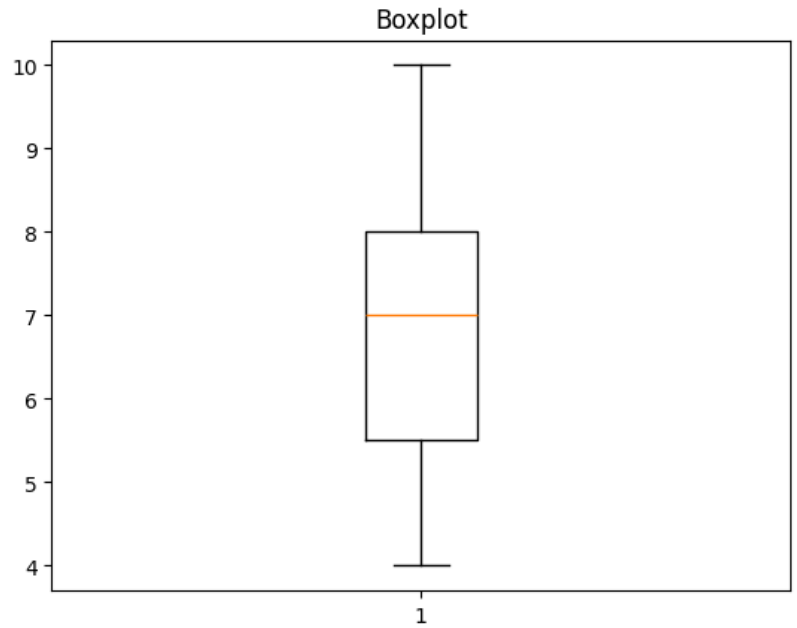
Python est l'un des langages les plus populaires en data science, et il offre une riche collection de bibliothèques pour la visualisation.

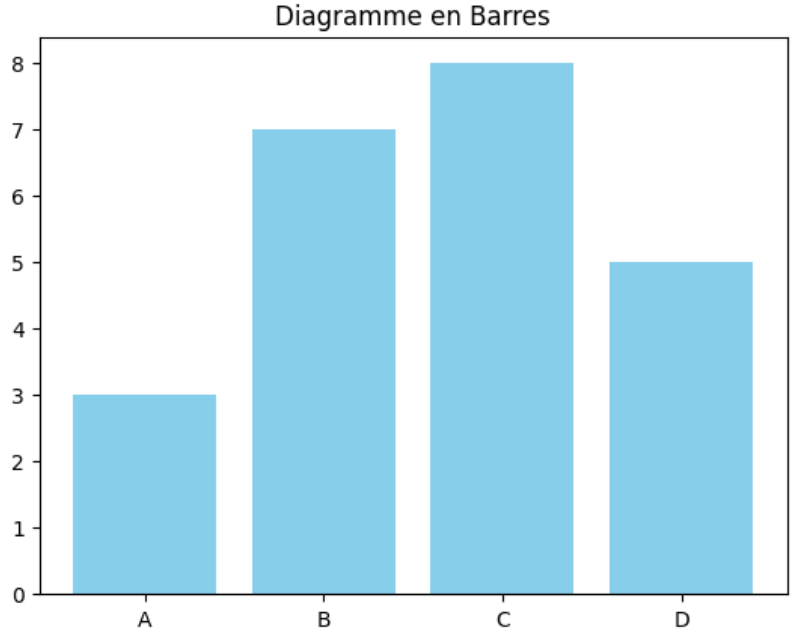
Matplotlib

- Une bibliothèque fondamentale pour créer des graphiques statiques, animés ou interactifs.
- Utilisées pour contruire des courbes, histogrammes, graphiques à barres, nuages de points.
- Leur points forts est sa haute personnalisation, adaptable à d'autres bibliothèques comme pandas et NumPy.







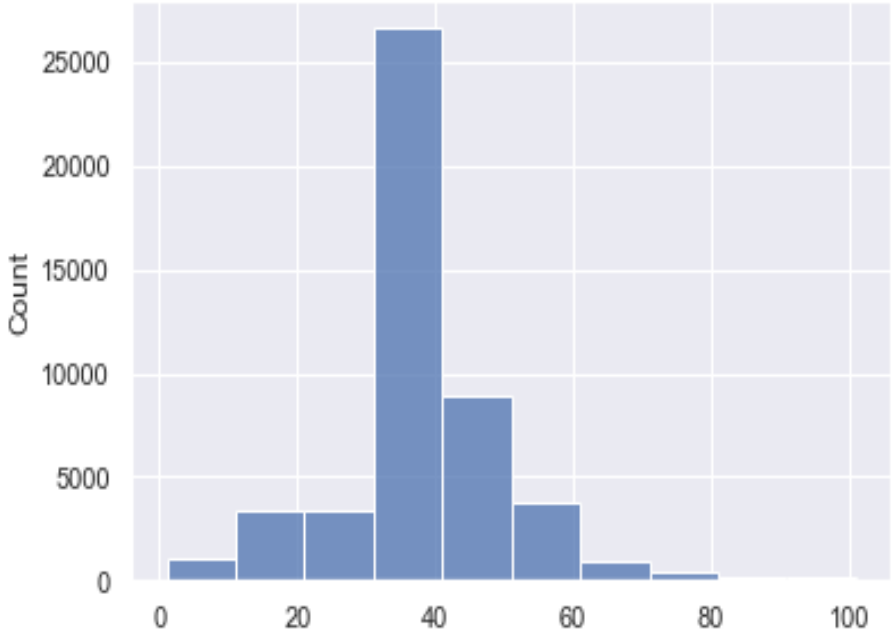


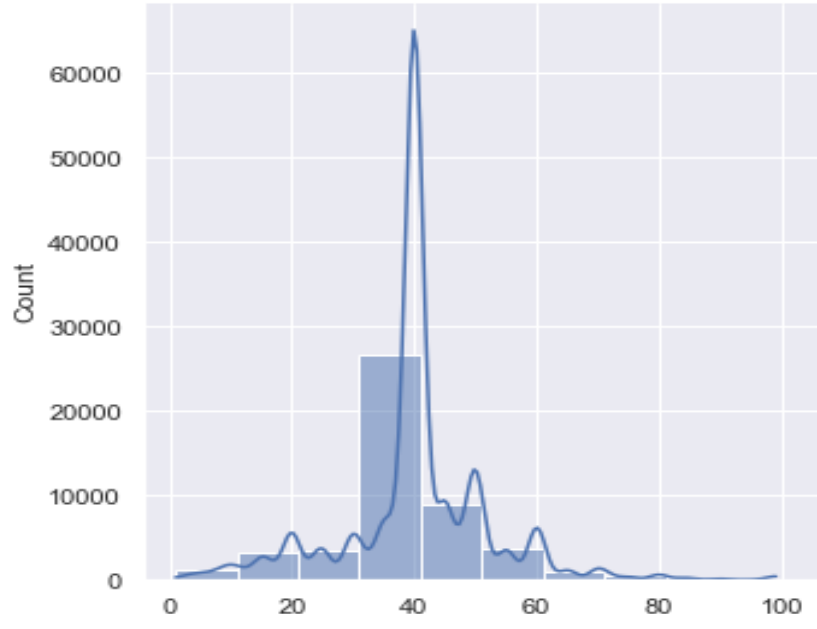
)

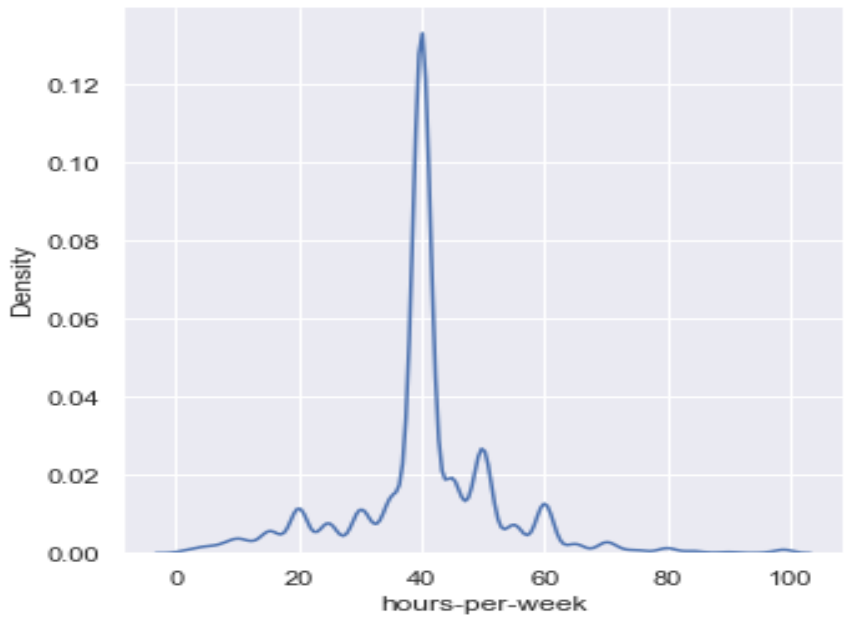
https://colab.research.google.com/drive/1n4uImR1BCV8ZJl1AJCHJvQMMcMANWuy5

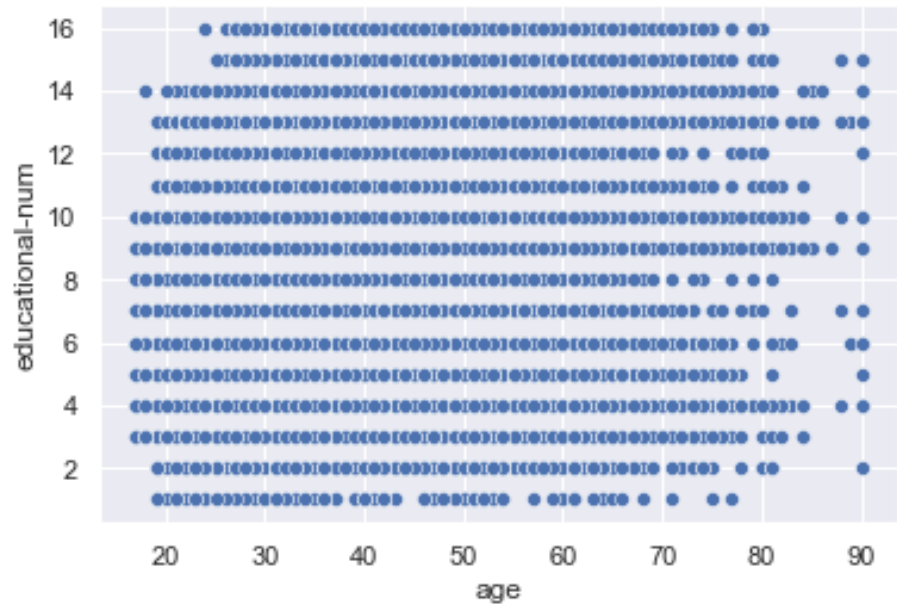
Seaborn

- Basé sur Matplotlib, Seaborn est conçu pour des graphiques statistiques avancés.
- Usages Visualisation de distributions, heatmaps, régressions.
- Points forts : Esthétiques par défaut, simplification des tâches complexes.







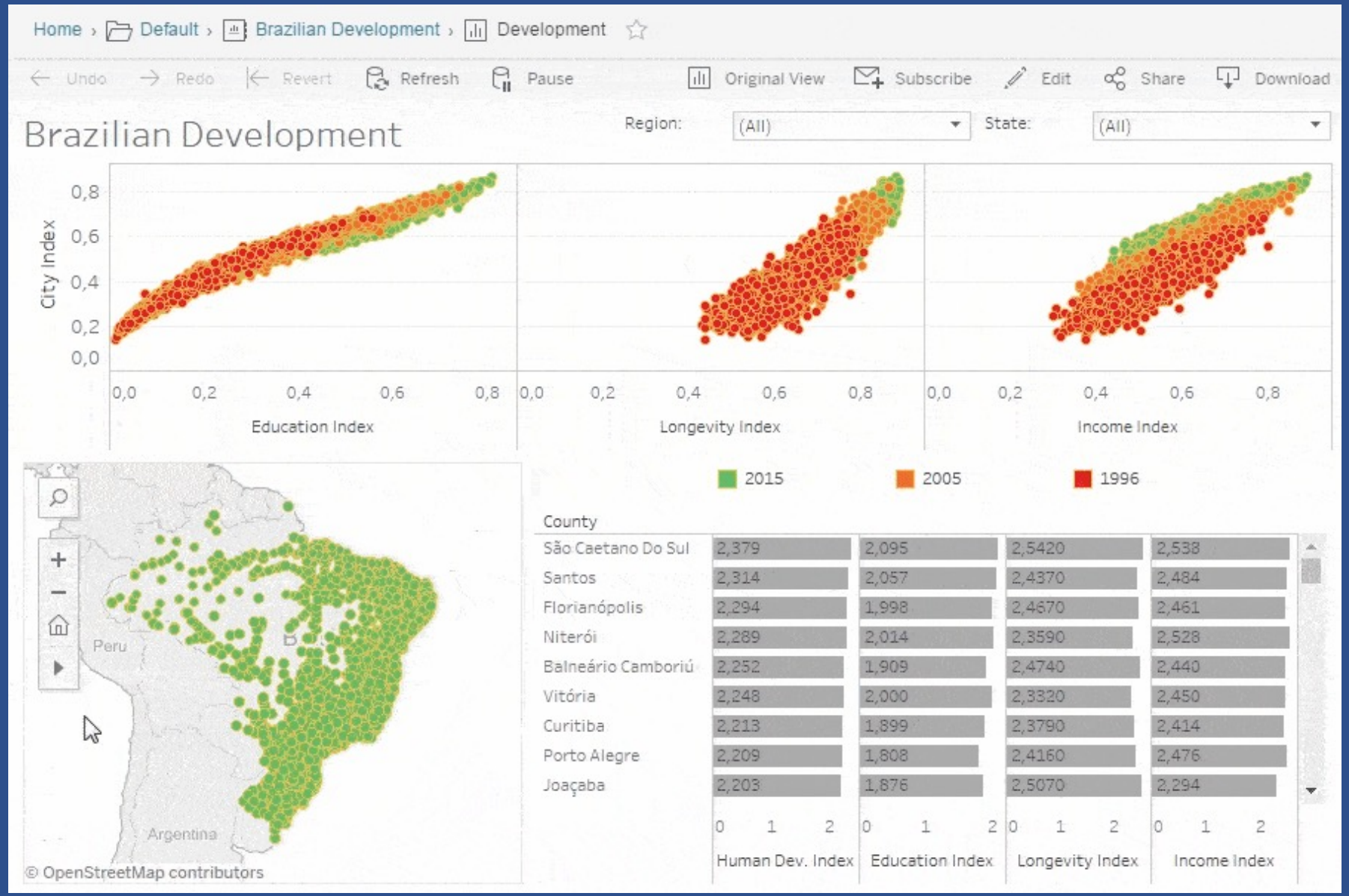


https://colab.research.google.com/drive/1n4uImR1BCV8ZJl1AJCHJvQMMcMANWuy5

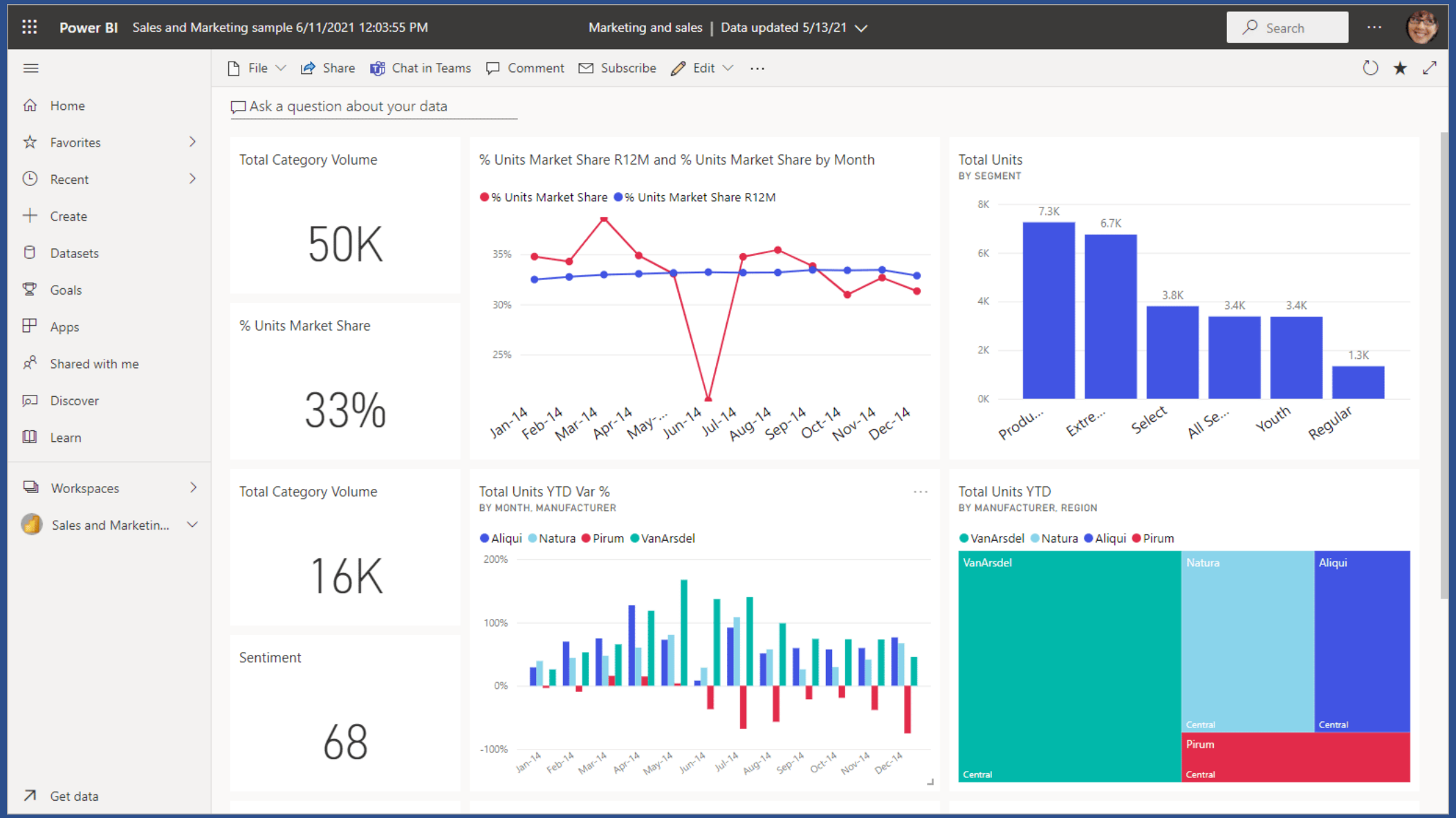
Plotly

- Outil interactif pour créer des visualisations dynamiques et des tableaux de bord.
- Utilisés pour créer des cartes interactives, des graphiques 3D, des diagrammes complexes.
- Les points forts de cette bibliothèque réside dans sa compatibilité avec Dash pour créer des tableaux de bord.

Logiciels de visualisation de données : Tableau



Logiciels de visualisation de données : PowerBI



Google Data Studio

Google News Initiative Resources About English My Account

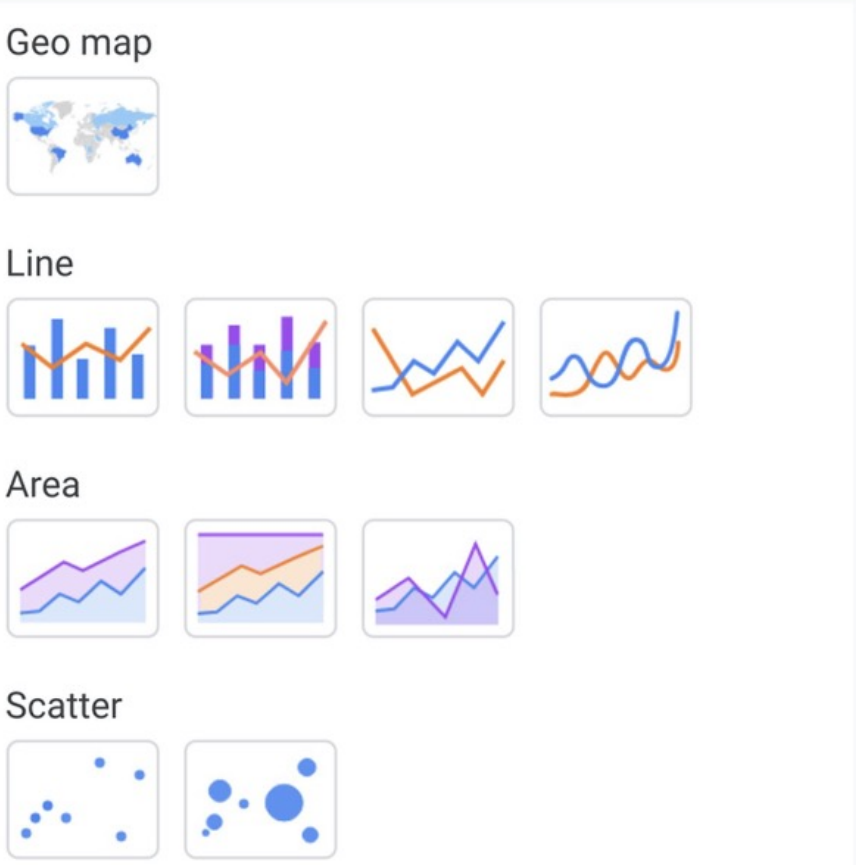
< DATA JOURNALISM

Lesson 11 of 13 | Data Studio: Make interactive data visualizations

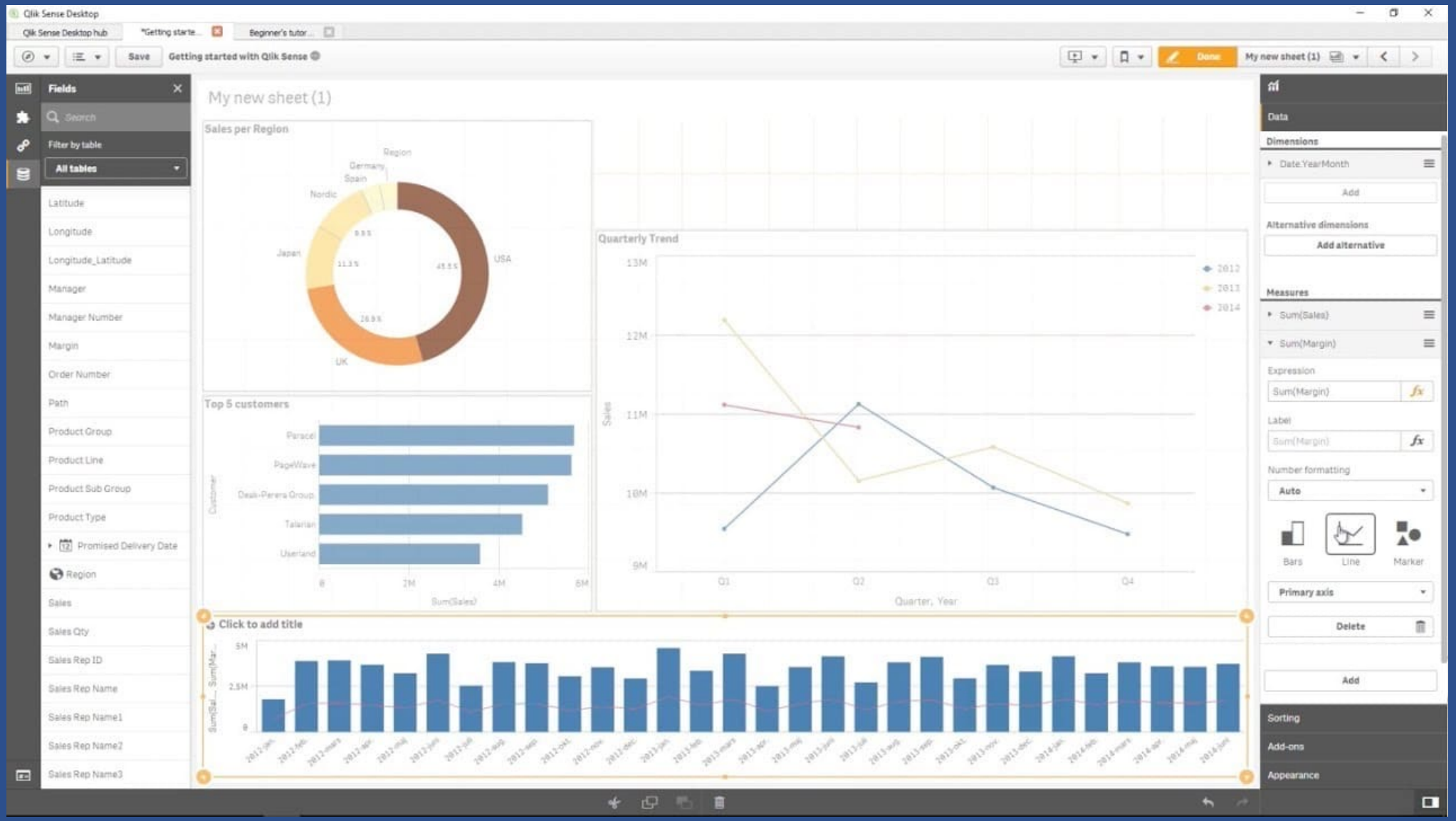
Data Studio: Make interactive data visualizations

Give life to your datasets by creating powerful interactive visualizations with an easy-to-use studio.

Download Lesson



Qlik Sense


Qlik Sense Desktop
Qlik Sense Desktop hub
Save
Getting started with Qlik Sense
My new sheet (1)
Fields
Search
Filter by table
All tables
Latitude
Longitude
Longitude_Latitude
Manager
Manager Number
Margin
Order Number
Path
Product Group
Product Line
Product Sub Group
Product Type
Promised Delivery Date
Region
Sales
Sales Qty
Sales Rep ID
Sales Rep Name
Sales Rep Name1
Sales Rep Name2
Sales Rep Name3
Sales per Region
Region
Germany
Spain
Nordic
Japan
USA
UK
Top 5 customers
Parace
PageWave
Deak-Perera Group
Talarian
Userland
Customer
Sum(Sales)
Quarterly Trend
2012
2013
2014
Sales
Quarter, Year
Click to add title
Data
Dimensions
Date.YearMonth
Add
Alternative dimensions
Add alternative
Measures
Sum(Sales)
Sum(Margin)
Expression
Label
Number formatting
Auto
Bars
Line
Marker
Primary axis
Delete
Add
Sorting
Add-ons
Appearance

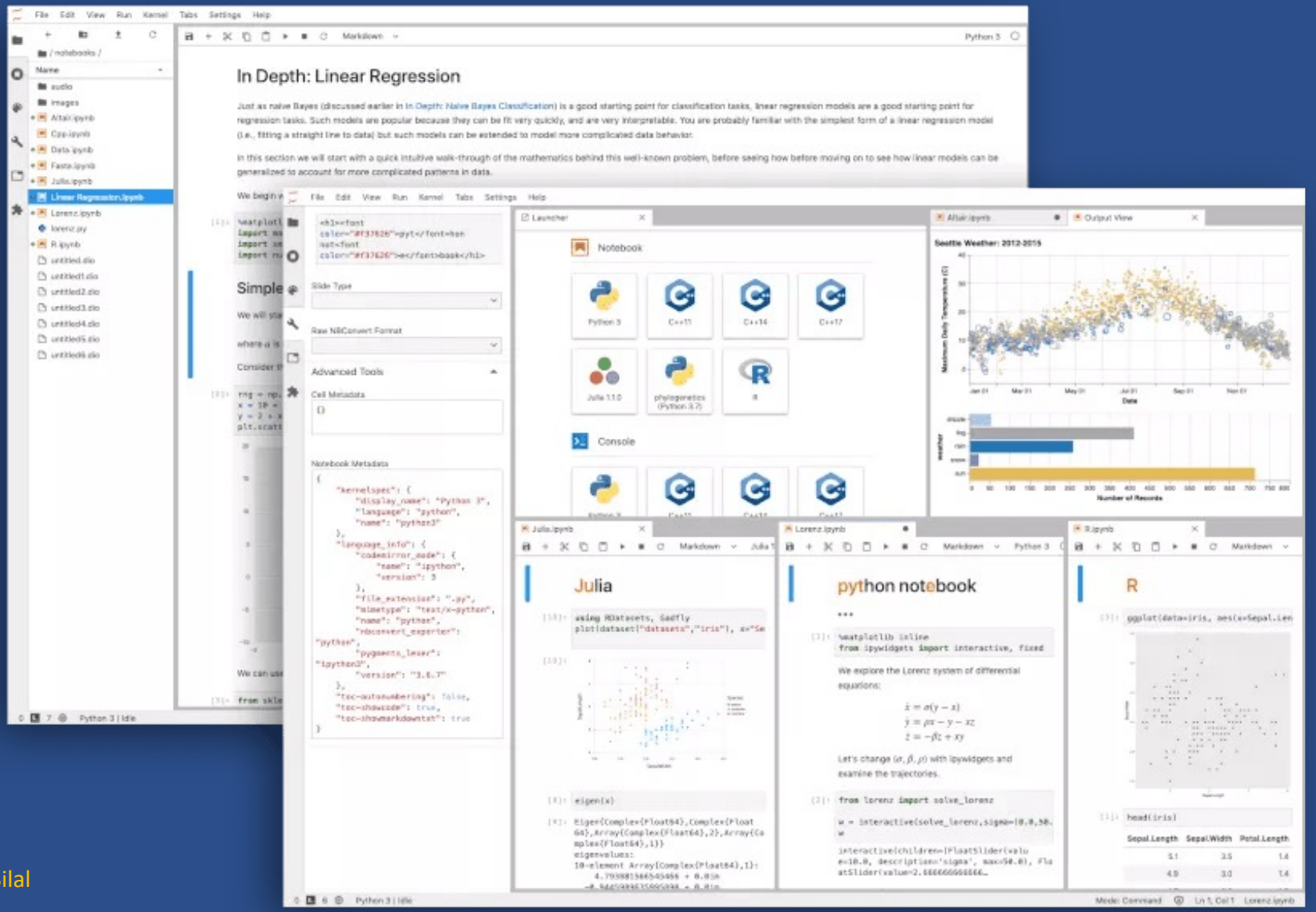
D3.js


Animated treemap
Temporal force-directed gra...
Connected scatterplot
The wealth & health of natio...
Scatterplot tour
Bar chart race
2000
Stacked-to-grouped bars
Streamgraph transitions
Smooth zooming
Zoom to bounding box
Orthographic to equirectang...
World tour
Walmart's growth
Hierarchical bar chart
Zoomable treemap
Zoomable circle packing
Collapsible tree
Zoomable icicle
Zoomable sunburst
Sortable bar chart
Icelandic population by age,...
Pie chart update
Arc tween

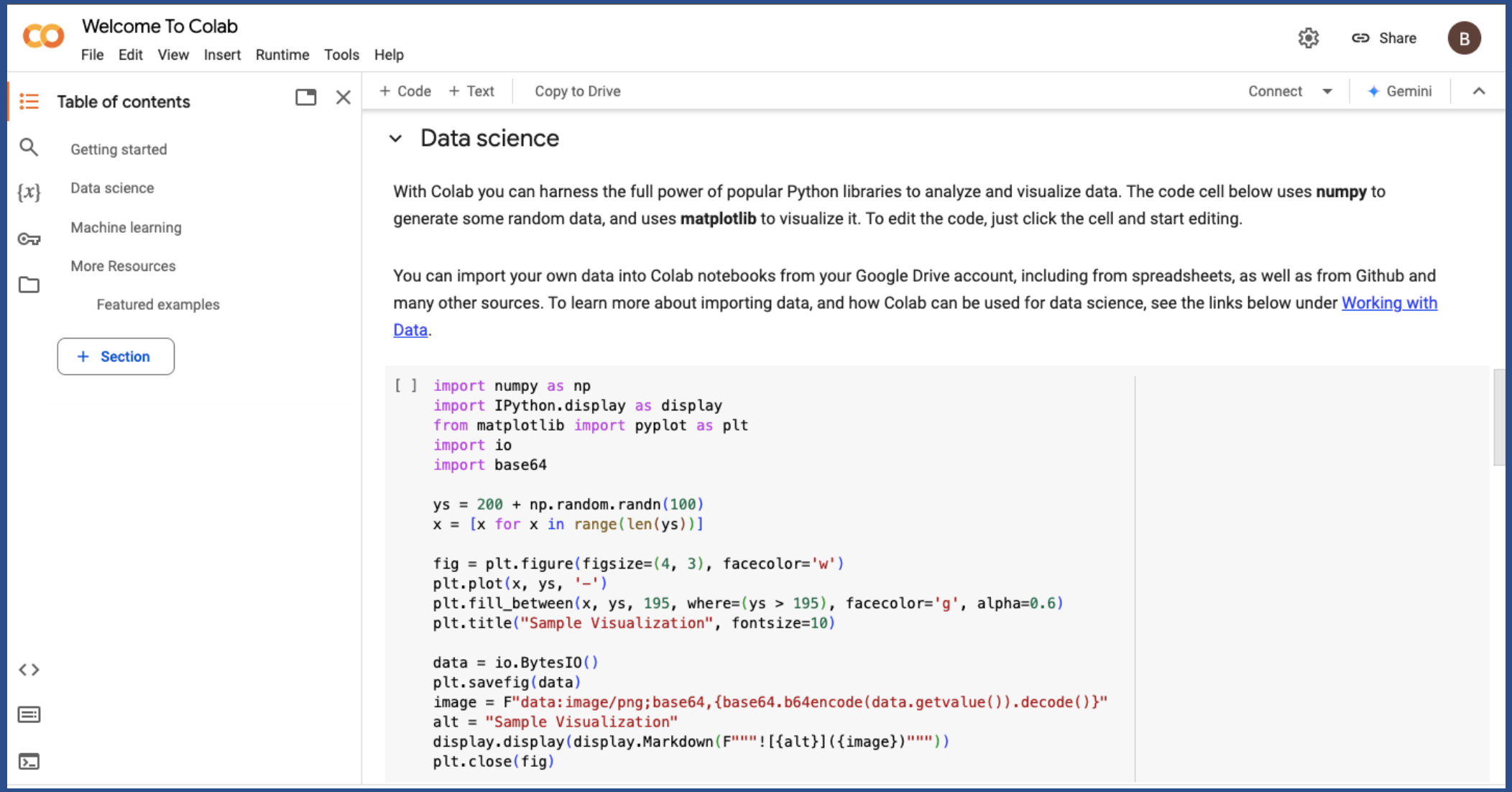
Les outils de dévelopement IDE Notebook

- Les outils IDE notebooks (Integrated Development Environment Notebooks) sont des environnements interactif.
- Permettant d'écrire et d'exécuter du code tout en intégrant des visualisations, des explications textuelles et des analyses de données.
- Ils sont particulièrement utiles en science des données pour explorer des datasets, tester des algorithmes, et présenter des résultats.

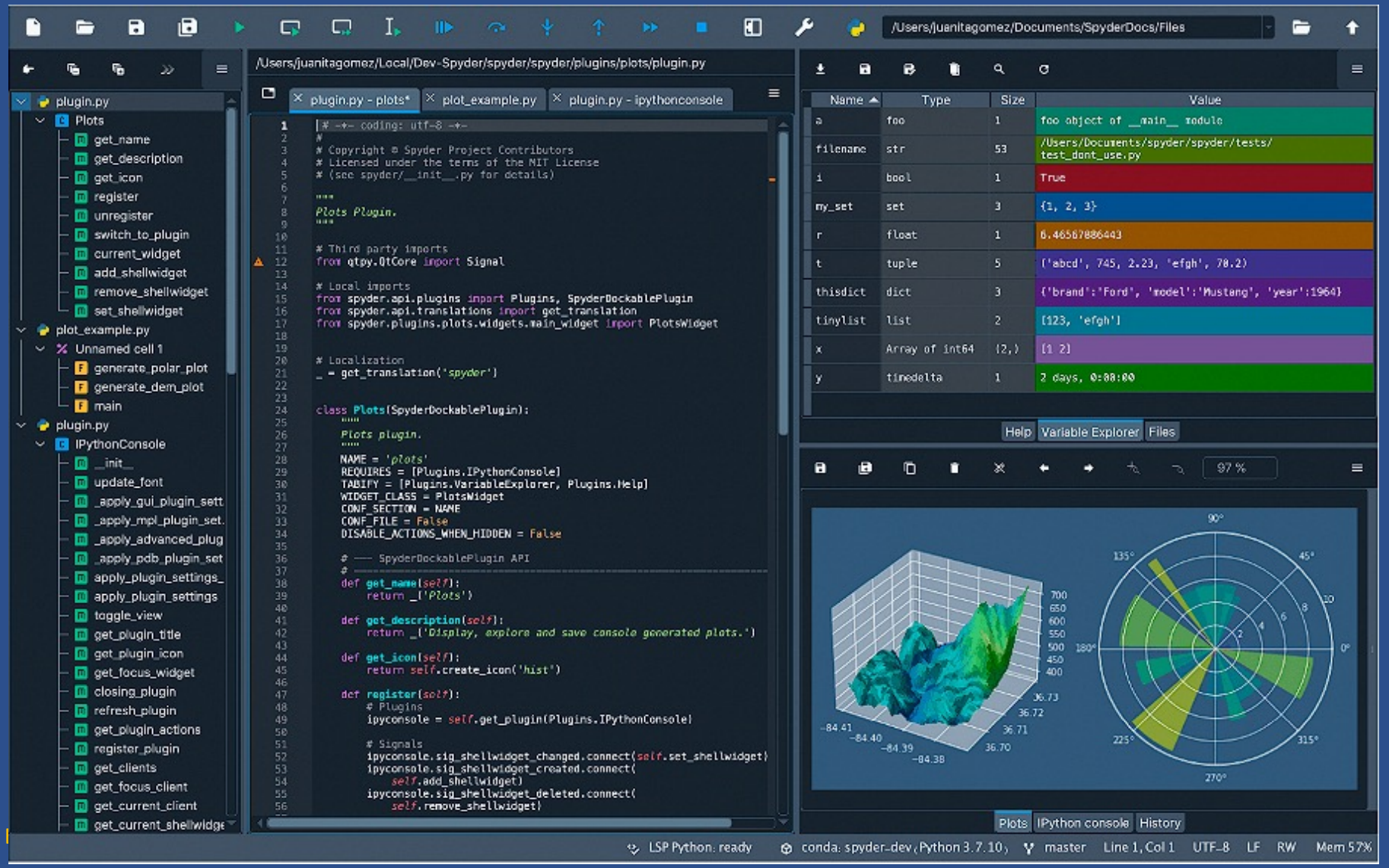
Les outils IDE notebooks: Jupyter notebook et Jupyter Lab



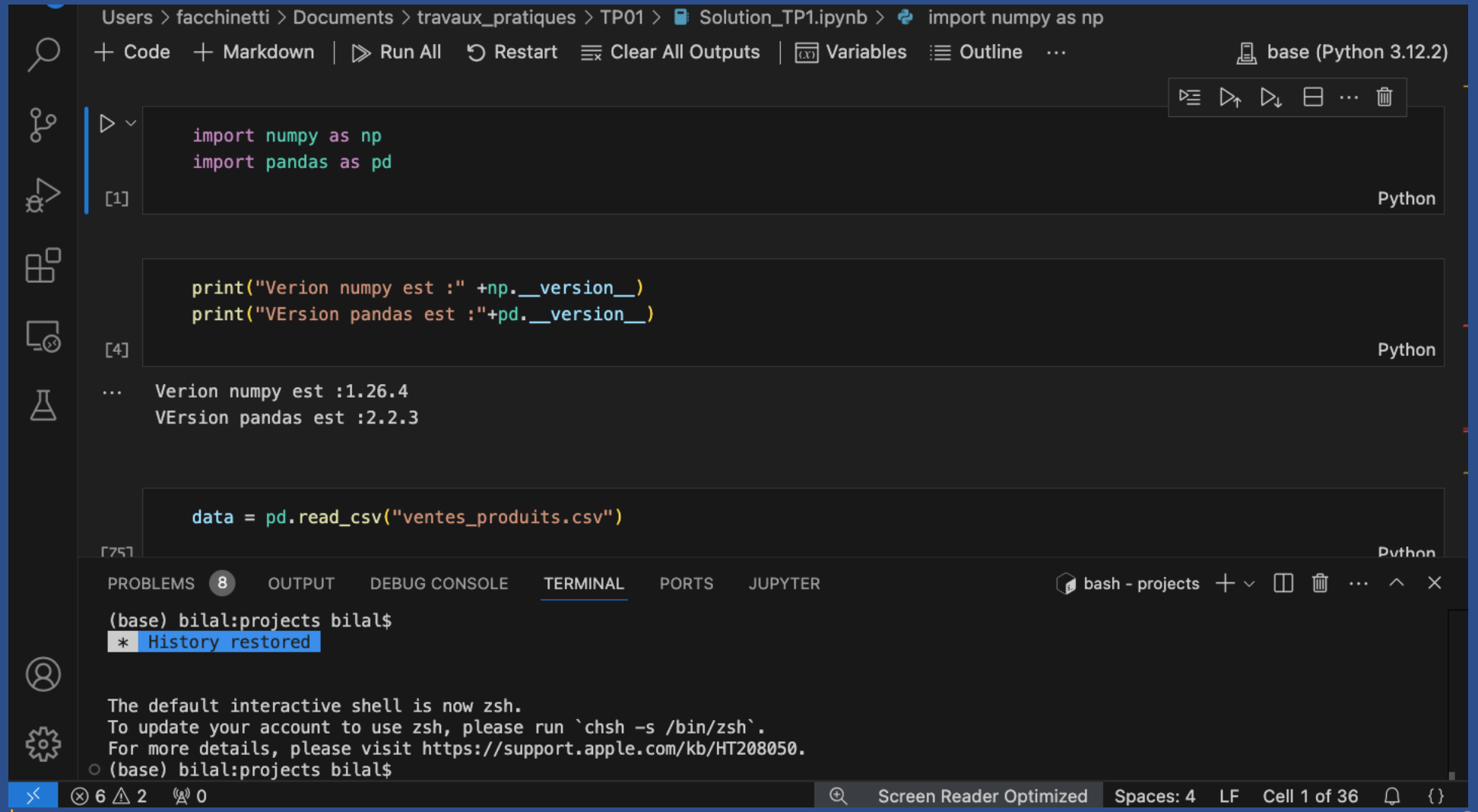
Google Colab



Spyder


/Users/juanitagomez/Documents/SpyderDocs/Files
/Users/juanitagomez/Local/Dev-Spyder/spyder/spyder/plugins/plots/plugin.py
plugin.py - plots*
plot_example.py
plugin.py - ipythonconsole
Name
Type
Size
Value
Help
Variable Explorer
Files
Plots
IPython console
History
LSP Python: ready
conda: spyder-dev (Python 3.7.10)
master
Line 1, Col 1
UTF-8
LF
RW
Mem 57%

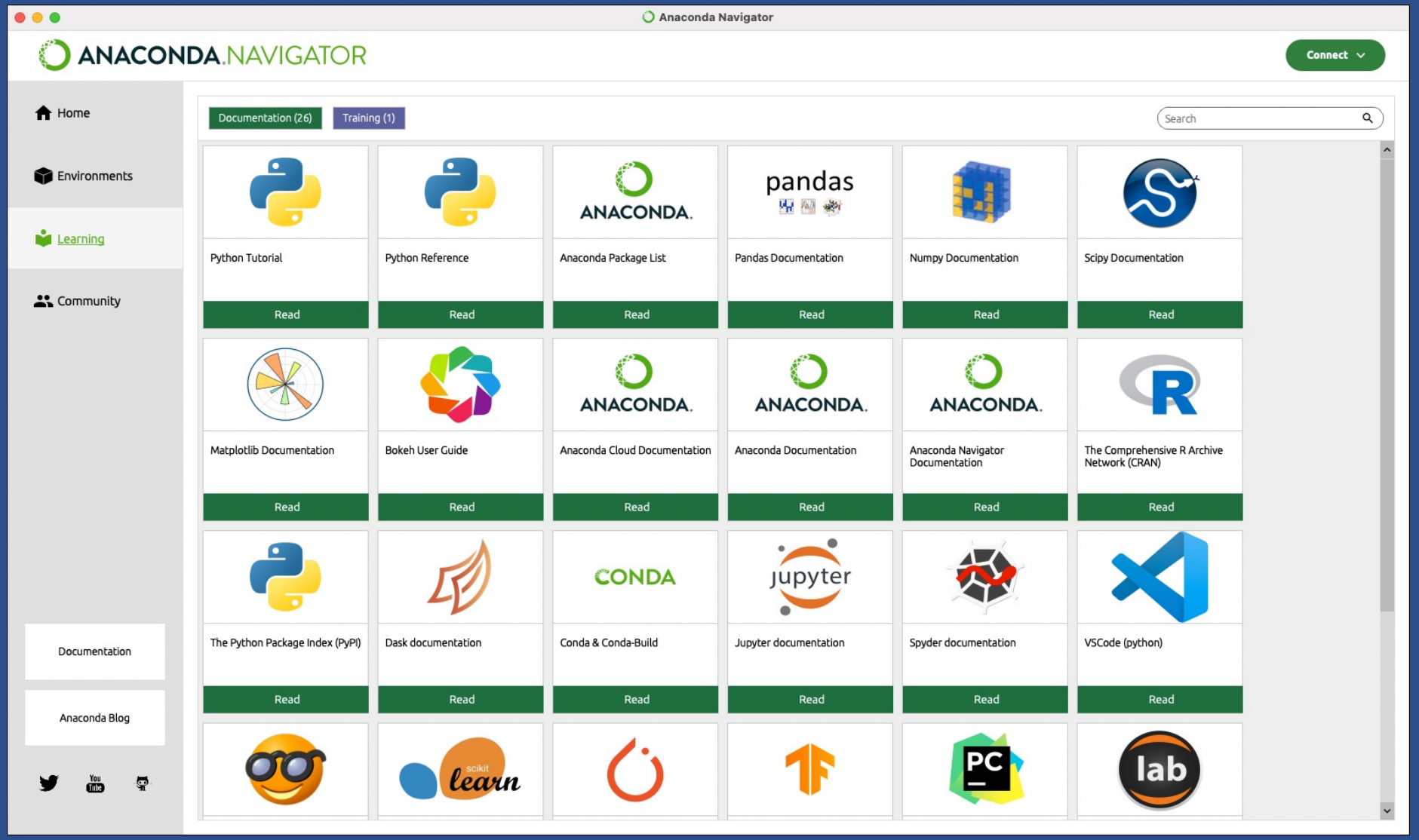
VS Code avec Jupyter Extension



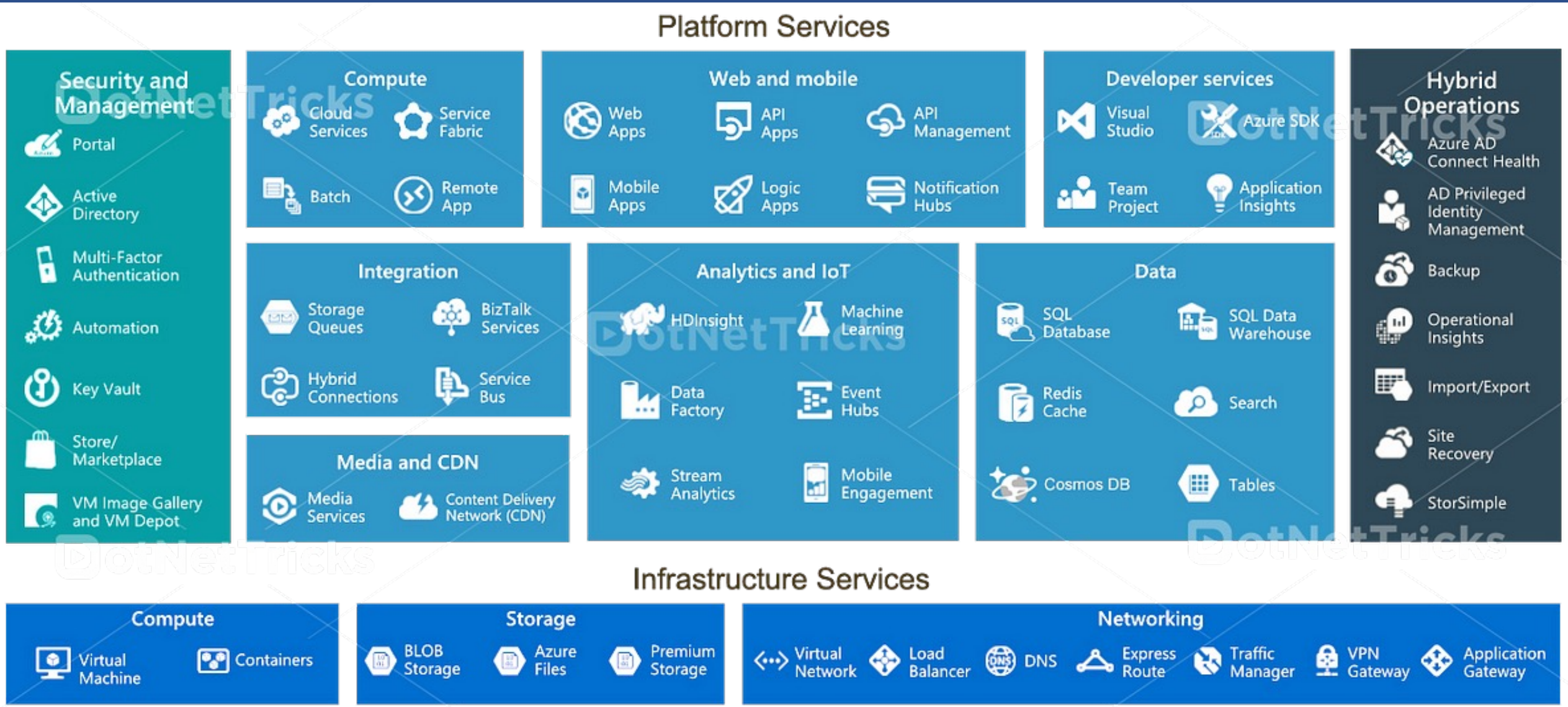
Les plateformes complètes de Data science

- Les plateformes complètes de Data Science sont des environnements intégrés qui regroupent tous les outils nécessaires pour réaliser l'ensemble du processus de data science, de la collecte des données jusqu'à la visualisation et le déploiement des modèles. Ces plateformes permettent aux data scientists de collaborer, d'automatiser des workflows, et de gérer des projets complexes en un seul endroit.

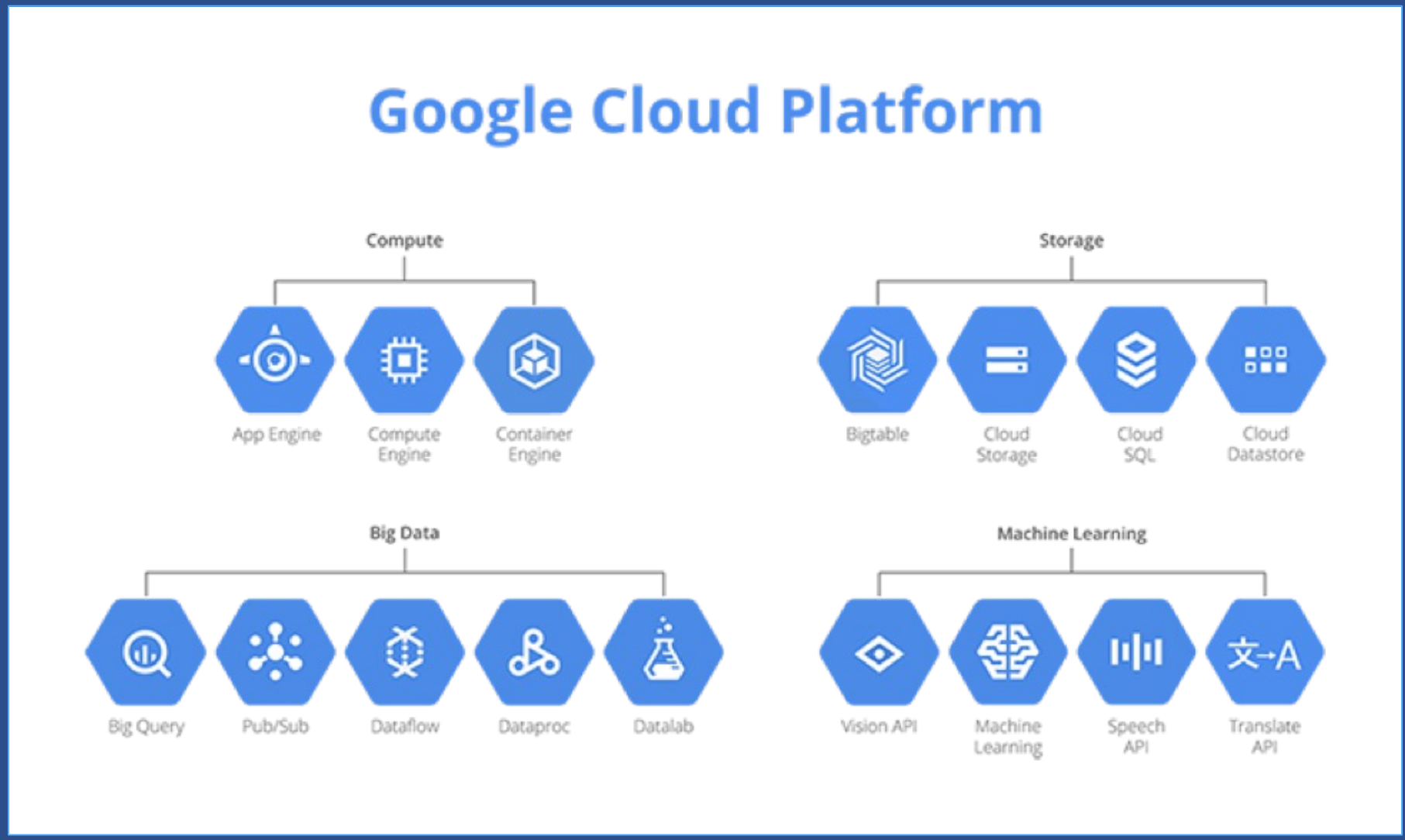
Anaconda


Anaconda Navigator
ANACONDA.NAVIGATOR
Connect
Home
Environments
Learning
Community
Documentation
Anaconda Blog
Documentation (26)
Training (1)
Search
Python Tutorial
Python Reference
Anaconda Package List
Pandas Documentation
Numpy Documentation
Scipy Documentation
Matplotlib Documentation
Bokeh User Guide
Anaconda Cloud Documentation
Anaconda Documentation
Anaconda Navigator Documentation
The Comprehensive R Archive Network (CRAN)
The Python Package Index (PyPI)
Dask documentation
Conda & Conda-Build
Jupyter documentation
Spyder documentation
VSCode (python)
Read

Microsoft Azure



Google cloud platform



Data science on AWS