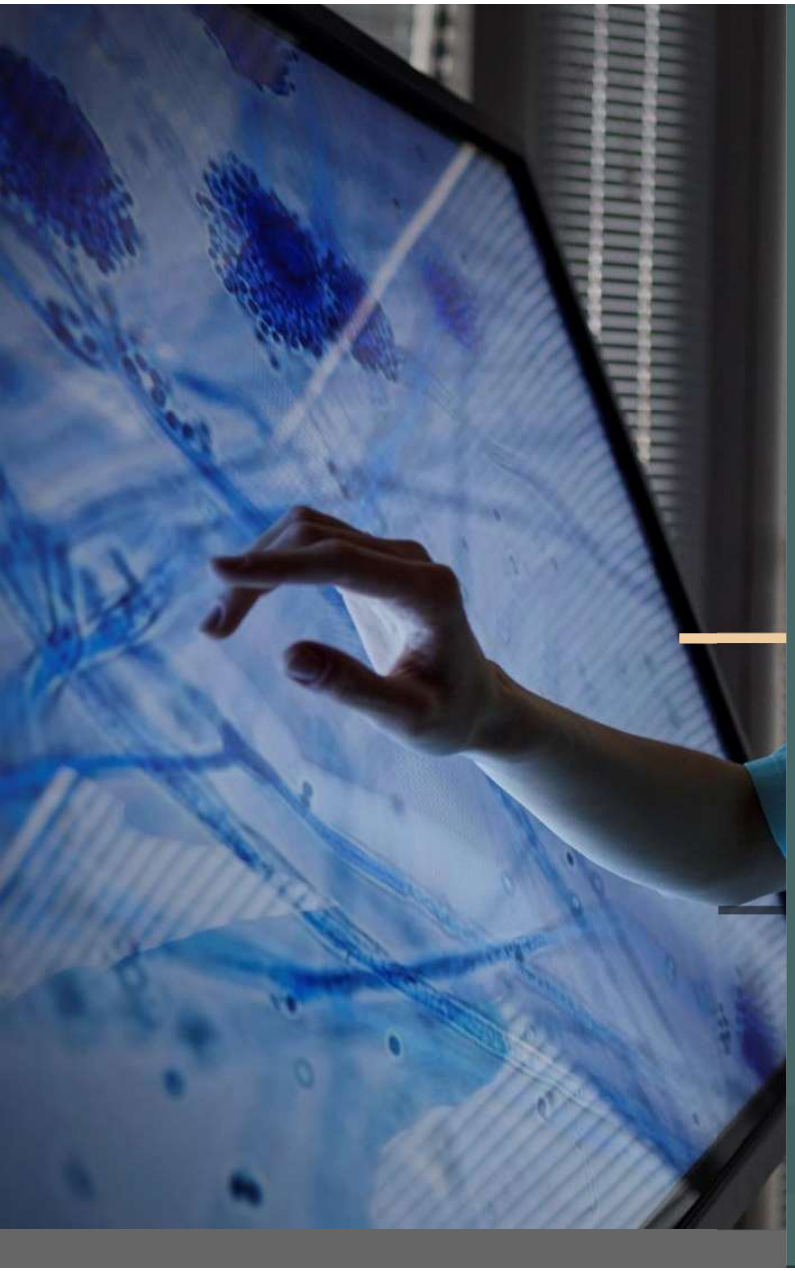


PROJET DE MISE EN PLACE D'UNE SOLUTION D'ANALYSE DE DONNÉES BIG DATA

Mars 2023

L'ÉQUIPE

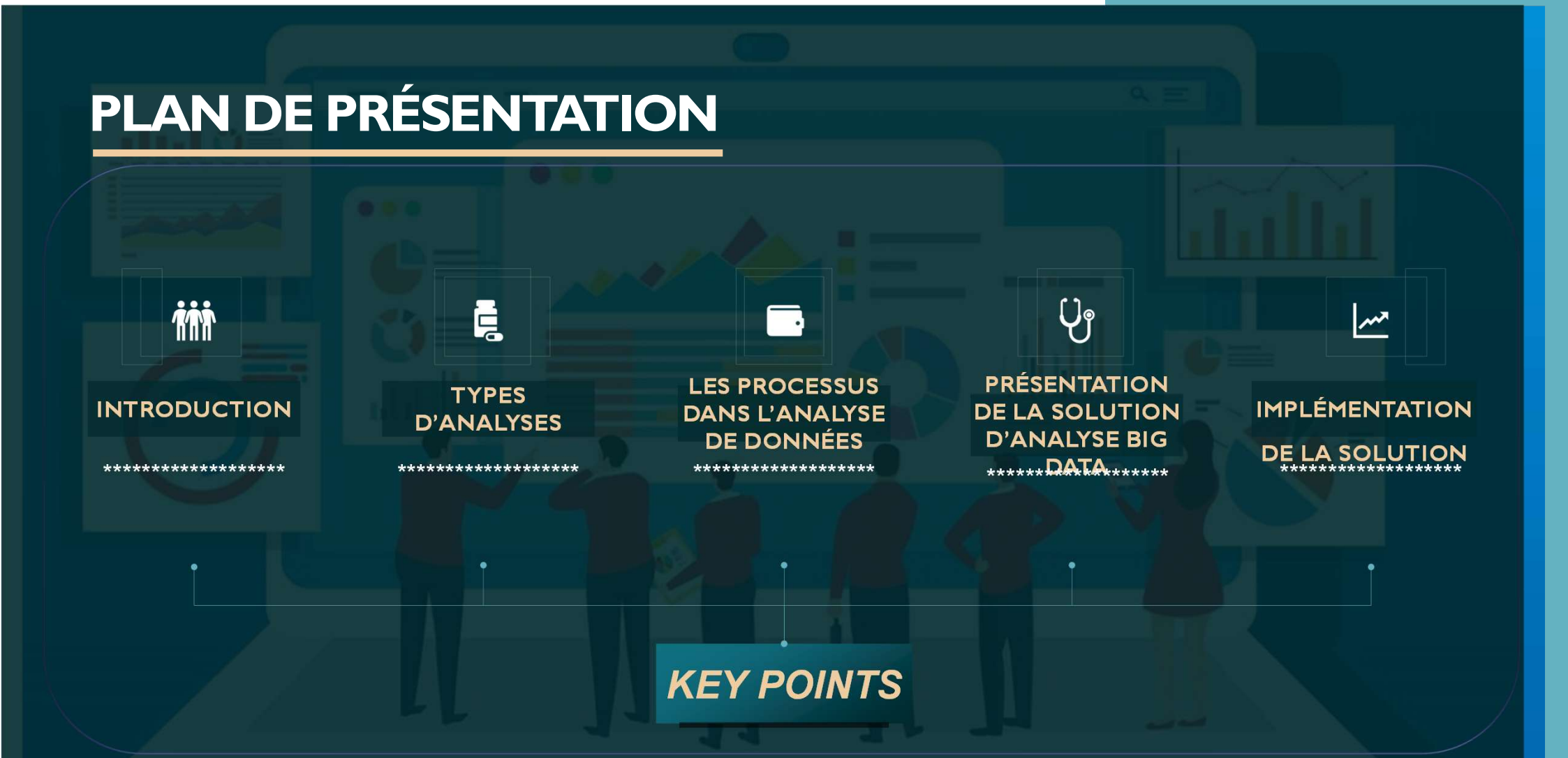
BAZIE Dinin Lothaire

SANOGO Steve

Sous la supervision

PREIRA Jean-Marie

PLAN DE PRÉSENTATION



INTRODUCTION

****

TYPES D'ANALYSES

****

LES PROCESSUS DANS L'ANALYSE DE DONNÉES

****

PRÉSENTATION DE LA SOLUTION D'ANALYSE BIG DATA

****

IMPLÉMENTATION DE LA SOLUTION

****

KEY POINTS

INTRODUCTION

- Data Analytics

- ✓ Processus de découverte, d'interprétation et de communication de modèles significatifs de données
- ✓ Permet aux organisations de gagner en visibilité et d'obtenir une compréhension plus approfondie de leurs processus et services

TYPES D'ANALYTICS

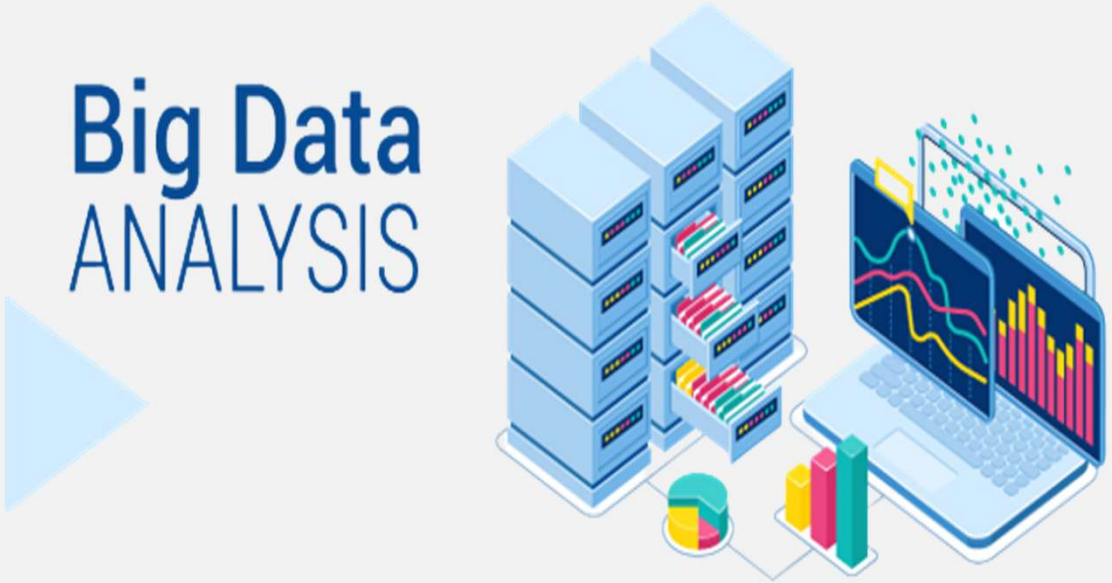
D'un point de vue technique, l'analyse des données peut être décrite comme le processus d'utilisation des données pour répondre à des questions, identifier des tendances et extraire des informations. Il existe de nombreux types d'analyse qui peuvent générer des informations pour stimuler l'innovation, améliorer l'efficacité et atténuer les risques.
Il existe quatre grands types d'analyse de données, chacun répondant à un type de question différent :

- **L'analyse descriptive** : "Que s'est-il passé ? "
- **L'analyse prédictive** : "Que pourrait-il se passer à l'avenir ?
- **L'analyse prescriptive** : "Que faut-il faire ensuite ?
- **L'analyse diagnostique** : "Pourquoi cela s'est-il produit ?"

PROCESSUS D'ANALYSE DES DONNÉES

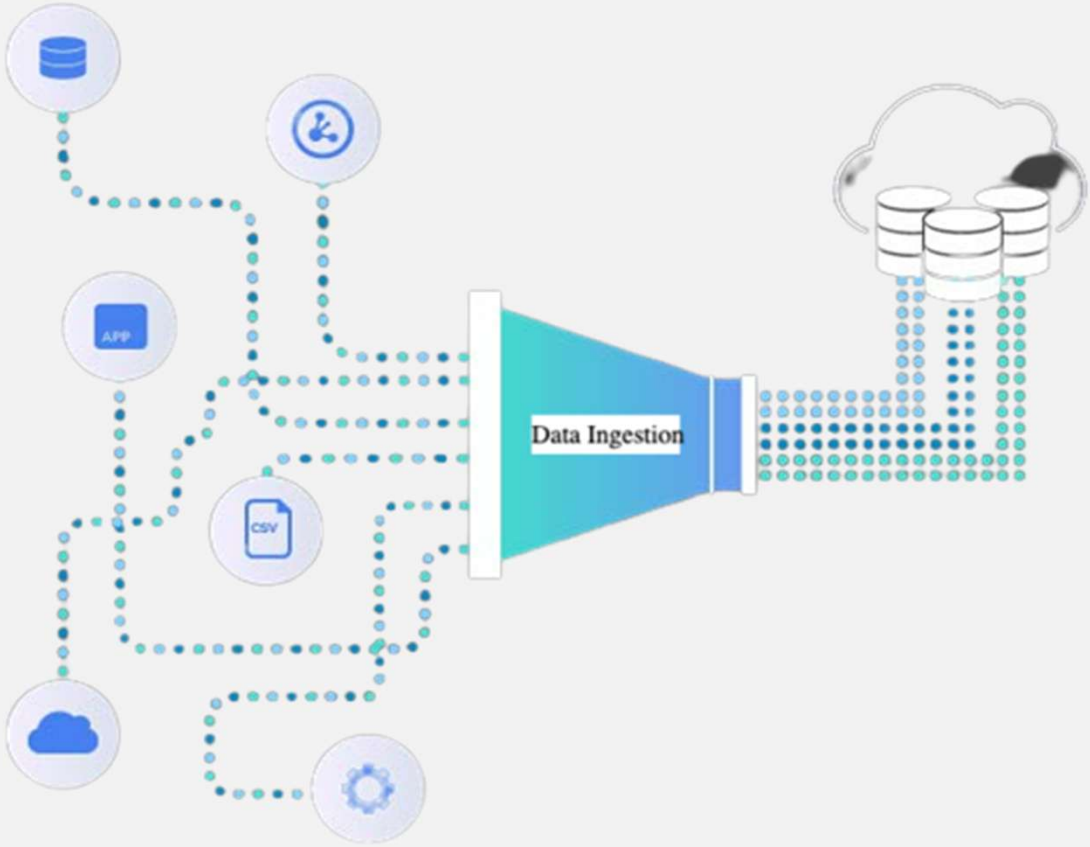
- **Poser les questions** : les questions auxquelles nous désirons répondent
- **Obtenir les données** : localisation et obtention des données pertinentes pour la ou les questions,
- **Étudier les données :** Cette étape consiste à déterminer si les données sont complètes et contiennent les informations pertinentes pour l'analyse.
- **Préparation des données** : C'est le processus de nettoyage des données
- **Analyse des données** : consiste à identifier les modèles, les corrélations et les relations contenus dans une ou L'analyse repose souvent sur des techniques statistiques et des outils logiciels tels que des feuilles de calcul et des applications de visualisation.
- **Présentation des résultats**

PRÉSENTATION DE LA SOLUTION



Architecture

DATA INGESTION



SOURCE DES DONNÉES

DATA SOURCING

OpenMeteo

Open-Meteo est une API météo open-source avec accès gratuit pour un usage non commercial. Aucune clé d'API n'est requise pour l'utiliser.
Elle fournit des données météorologiques en temps réel et des prévisions météorologiques à court et long terme.

Les principaux caractéristiques de cette API sont :

- Free API
- High Resolution
- Fast Update
- Historical Data
- Easy To Use
- Open Source

DONNÉES

Nous avons récupéré les données climatiques de la ville de Dakar pour les années de 2002 à 2022 pour chaque heure pour un total de 184.080 lignes et 8 colonnes.
Les caractéristiques des données récupérés sont :

Name	Unit	Format
time	iso8601	Date
temperature	°C	Float
apparentTemperature	°C	Float
relativeHumidity	%	Float
windSpeed	km/h	Float
pressure	hPa	Float
weathercode	WMO code	ategorial
rain	mm	Float

Data end-point and Parameters

```
https://archive-api.open-meteo.com/v1/archive
```

```
parameters = {
            'latitude': '14.69',
            'longitude': '-17.44',
            'start_date': '2002-01-01',
            'end_date': '2022-12-31',
            'hourly': "temperature_2m,relativehumidity_2m,
            }
```

FILEBEAT

Filebeat est un outil de transfert de données léger et efficace qui collecte et transfère des données de fichiers de logs et de métriques vers une sortie de votre choix. Filebeat est développé par **Elastic**, la même entreprise qui développe **Elasticsearch**, **Logstash** et **Kibana**, et est distribué sous licence Apache.
Filebeat est facile à installer et à configurer et prend en charge de nombreux formats de fichiers :

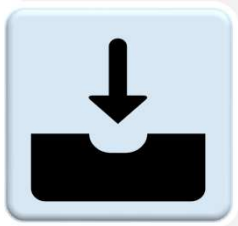
- **logs,**
- **Fichiers système,**
- **JSON,**
- **CSV**
- **Apache.**

Il est disponible pour de nombreuses plateformes, y compris **Linux**, **Windows** et **Mac OS X.**

Principe de fonctionnement de Filebeat

Filebeat est composé de deux éléments principaux fonctionnant ensemble pour suivre les fichiers et envoyer les données d'événement à une sortie spécifié:

- **Inputs :**
  Ce sont les sources de données que Filebeat surveille. Des inputs configurés dans un fichier de configuration pour collecter des données de différents types de sources telles que des fichiers de logs, HTTP JSON, AWS S3. Ces inputs déterminent les données qui seront collectées et transférées.



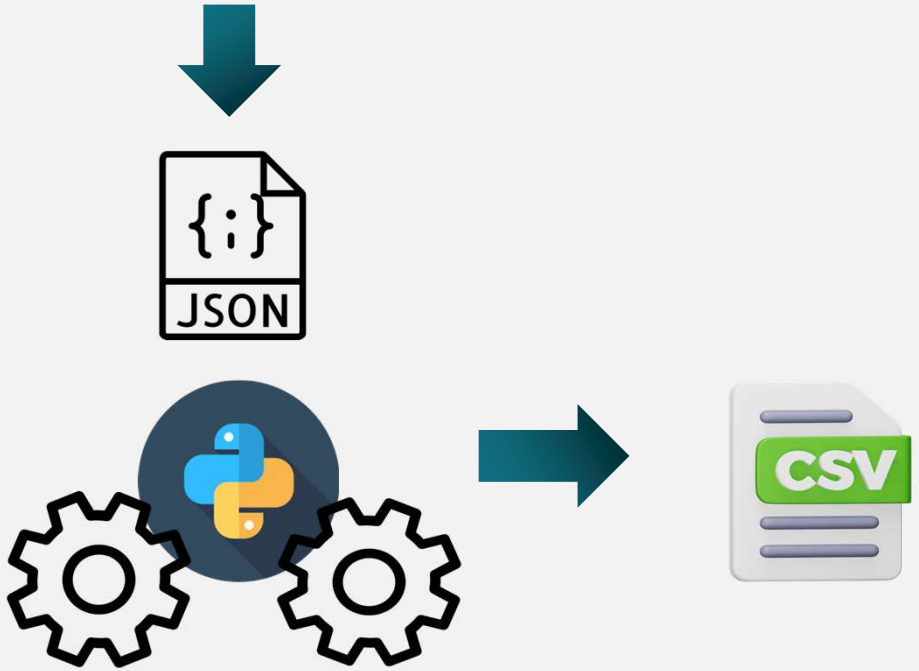
- **Harvesters:**
  Filebeat utilise des harvesters pour surveiller en continu les fichiers de sources de données et les transmettre à la chaîne de traitement des données. Les harvesters sont configurés pour détecter les changements dans les fichiers de logs afin de garantir que toutes les données sont capturées.

SCRIPT

```
filebeat.inputs:
- type: httpjson
  request.url: https://archive-api.open-meteo.com/v1/archive?
  json.keys_under_root: true
  json.overwrite_keys: true
processors:
- decode_json_fields:
    fields: ["message"]
    target: ""
    overwrite_keys: false
- drop_fields:
    fields: ["*"]
    ignore_missing: true
- include_fields:
      fields: ["hourly"]
output.file:
  path: "/home/hadoop/"
  filename: dakar_weather_2002_2022
```

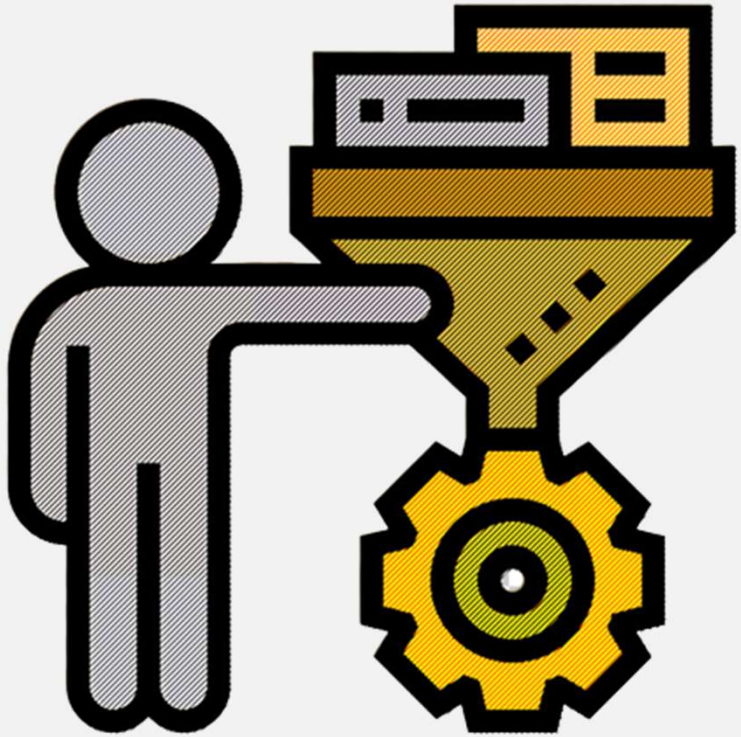
SCRIPT execution

```
/etc/filebeat-8.6.2-linux-x86_64/filebeat -c /home/hadoop/sript.yml
```



DATA INGESTION

DATA PROCESSING

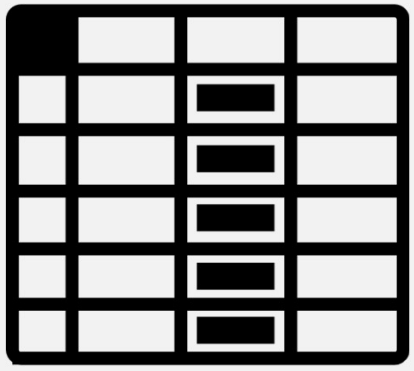


TRAITEMENT DES DONNEES

Apache Spark est un Framework de traitements Big Data open source développé en 2009 construit pour effectuer des analyses sophistiquées. Spark permet des calculs 10 à 100 fois plus rapide que sur Hadoop grâce à son mode de calcul IN-MEMORY. Le traitement des données réalisé avec Spark à essentiellement porté sur :

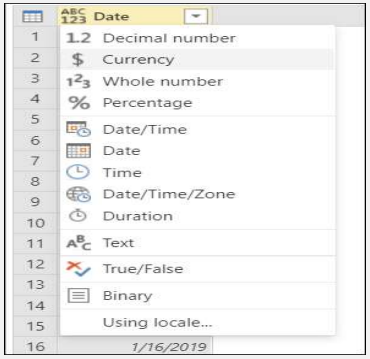
❑ La création de colonnes

- ✓ day
- ✓ month
- ✓ Year



❑ Le typage des colonnes

- ✓ String
- ✓ Int
- ✓ Float
- ✓ Date



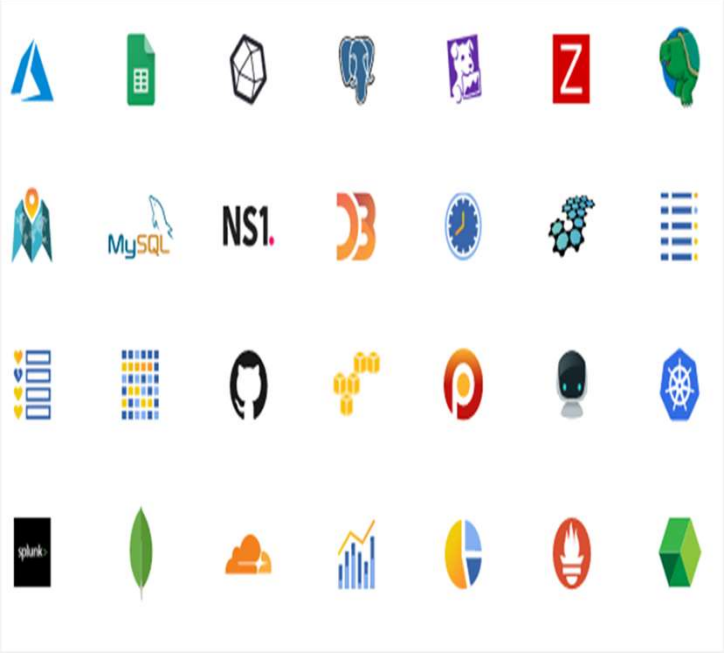
DATA VISUALIZATION

VISUALISATION DES DONNÉES

Grafana

Grafana est une application Web d'analyse et de visualisation multiplateforme qui agit comme une seule fenêtre pour afficher toutes vos données métriques, où qu'elles se trouvent.

Connecter vos outils et vos équipes avec Grafana plugins. Ces plugins vous permettent de vous connecter à des base de données existantes via des API et de rendre les données en temps réel sans avoir besoin de les migrer dans une de tes bases de données.



GRAFANA CONFIGURATION

1. Création du fichier de configuration

sudo vim /etc/yum.repos.d/graphana.repo

```
[grafana]
name=grafana
baseurl=https://rpm.grafana.com
repo_gpgcheck=1
enabled=1
gpgcheck=1
gpgkey=https://rpm.grafana.com/gpg.key
sslverify=1
sslcacert=/etc/pki/tls/certs/ca-bundle.crt
```

2. Installation

sudo yum install grafana

3. Installation de polkit

sudo yum install polkit

4. Start graphana

sudo service grafana-server start

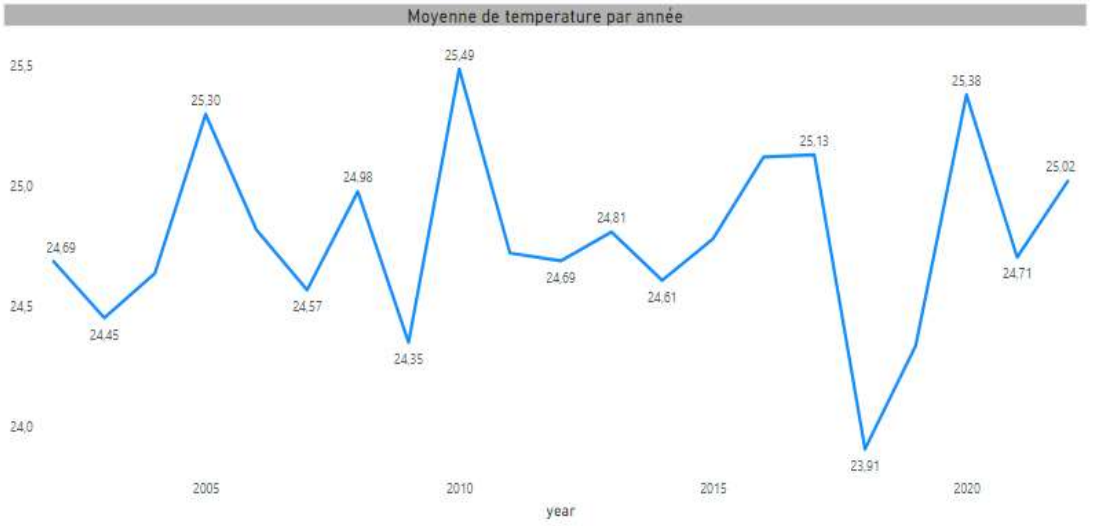
5. Connexion

http://34.229.72.125:3000

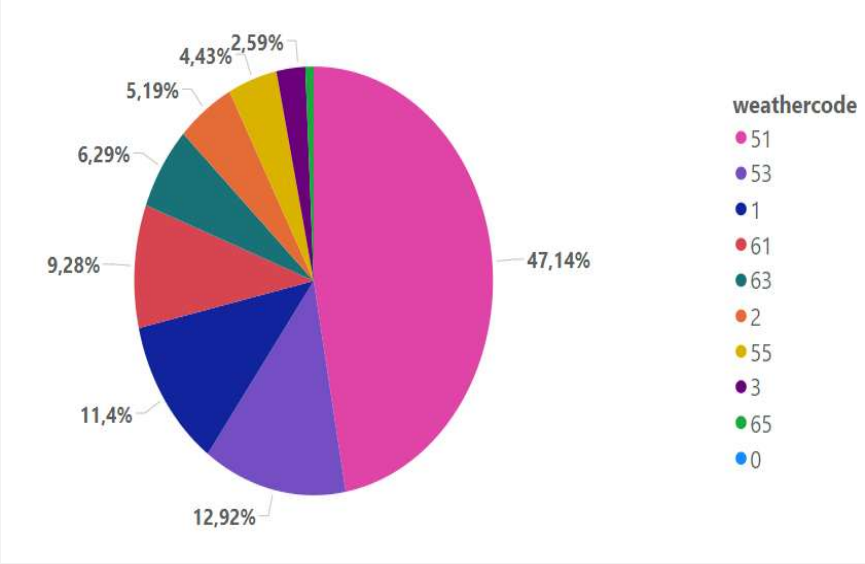
6. Data Source Link

HTTP

VISUALISATION

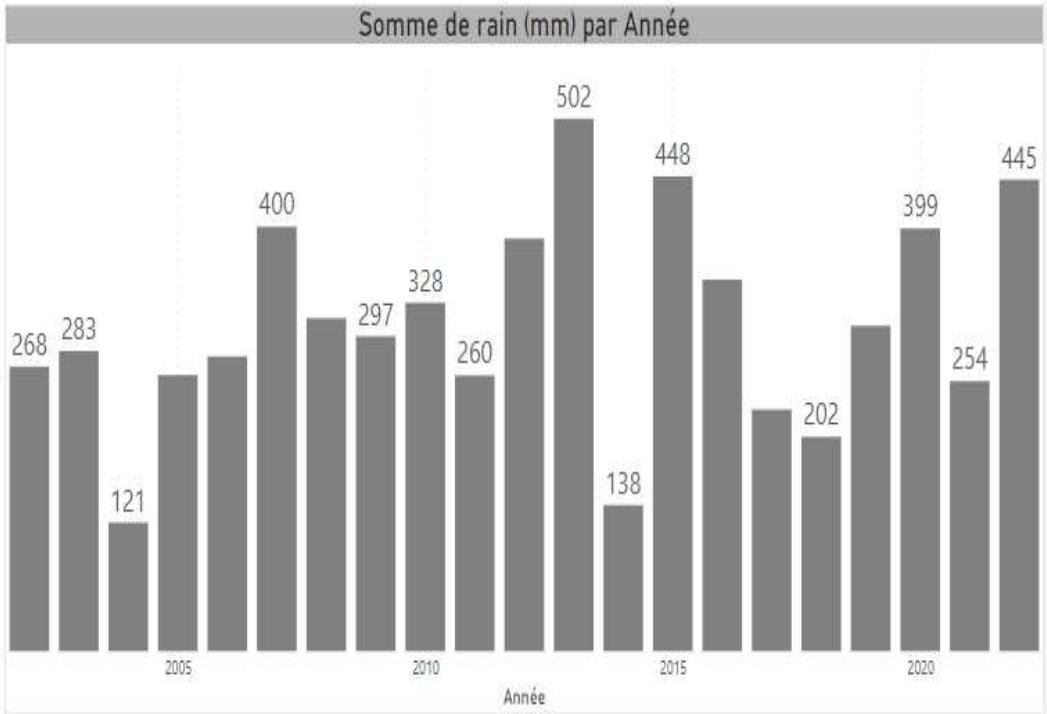


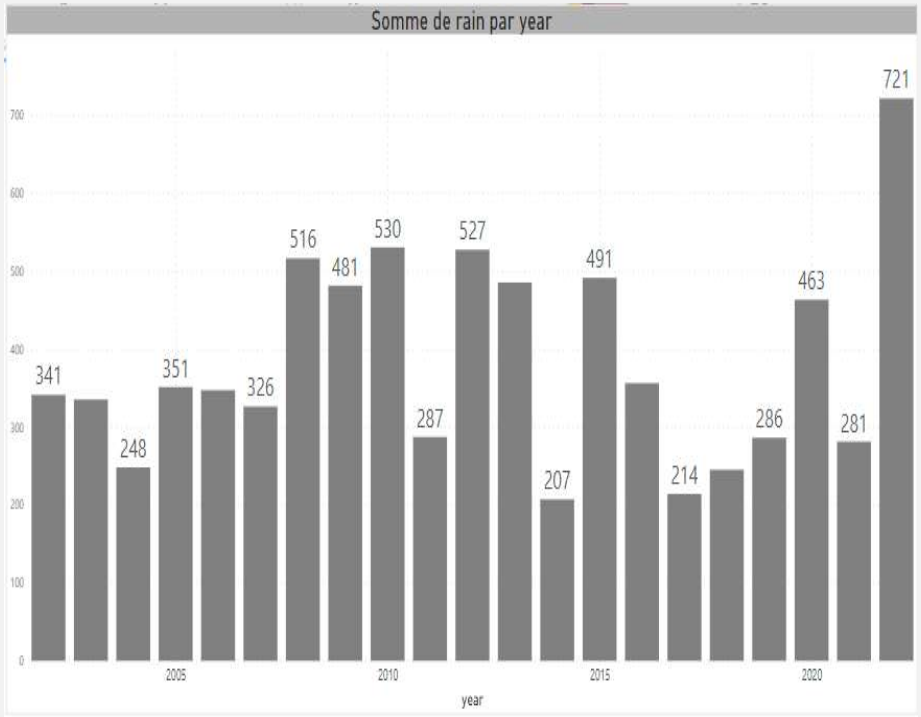
WM0	Meaning
1	Les nuages se dissolvent généralement ou deviennent moins développés
53	Bruine, non gelée, continue modéré lors de l'observation
51	Bruine, non gelée, continue
61	Pluie, non gelée, continue
63	Pluie, non gelée, continue, lourd au moment de l'observation



Le **weathercode** est un système de classification numérique utilisé pour représenter les conditions météorologiques dans les rapports et les observations météorologiques. Les codes de temps WMO sont des nombres à deux chiffres qui décrivent les conditions météorologiques d'une région à un moment donné.

VISUALISATION





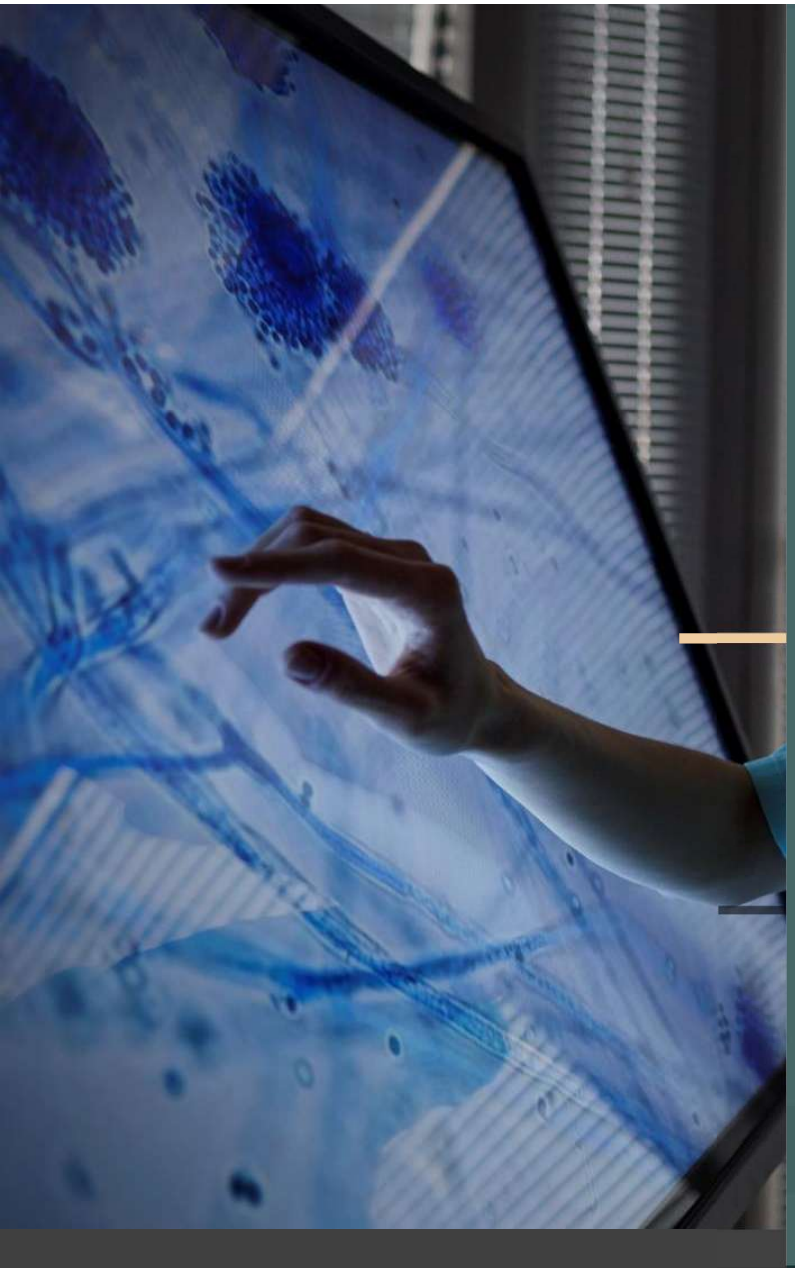
RESSOURCES

Cours Plateforme Big Data By Jean-Marie Preira

Google

ChatGPT


THANK À
VOUS
ESMT
INGC3@2023
Lothaire, Steve



PROJET DE MISE EN PLACE D'UNE SOLUTION D'ANALYSE DE DONNÉES BIG DATA

Mars 2023