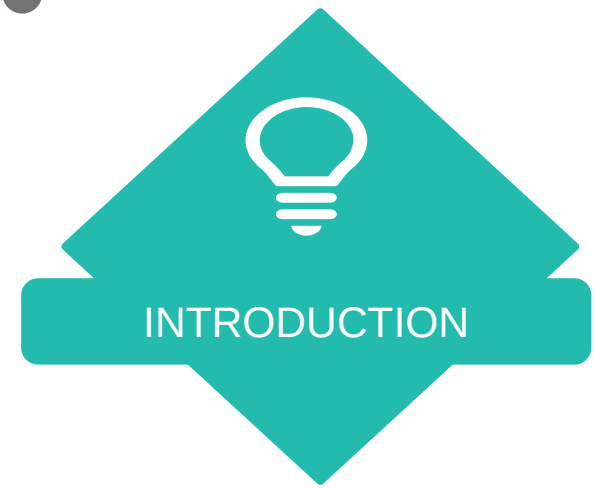
BIG DATA & DISTRIBUTED PROCESSING : MAPREDUCE AND PYSPARK

SPEAKER
José Arthur OUEDRAOGO
Data Science Research Engineer,
ED-ST / ERSIC

PLAN



ZiNDi

INTRODUCTION

ZiND!

INTRODUCTION

Artificial Intelligence (AI)

Data

Business Intelligence (BI)

Semantic Web

ZiND!

INTRODUCTION

DATA

STRUCTURED DATA

Relational databases, CSV files, JSON and XML data …

UNSTRUCTURED DATA

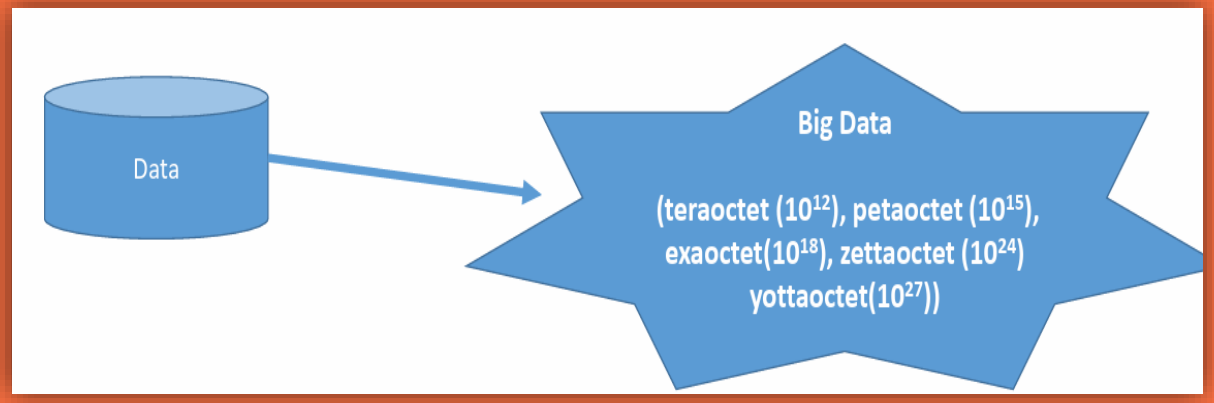
Sensor data, images, video, social medias posts …

ZiND!

GENERALITES

ZiND!

WHAT IS BIG DATA ?



Le Big Data fait référence à la gestion et à l'analyse de grandes quantités de données qui dépassent les capacités des outils traditionnels de traitement de données [3]

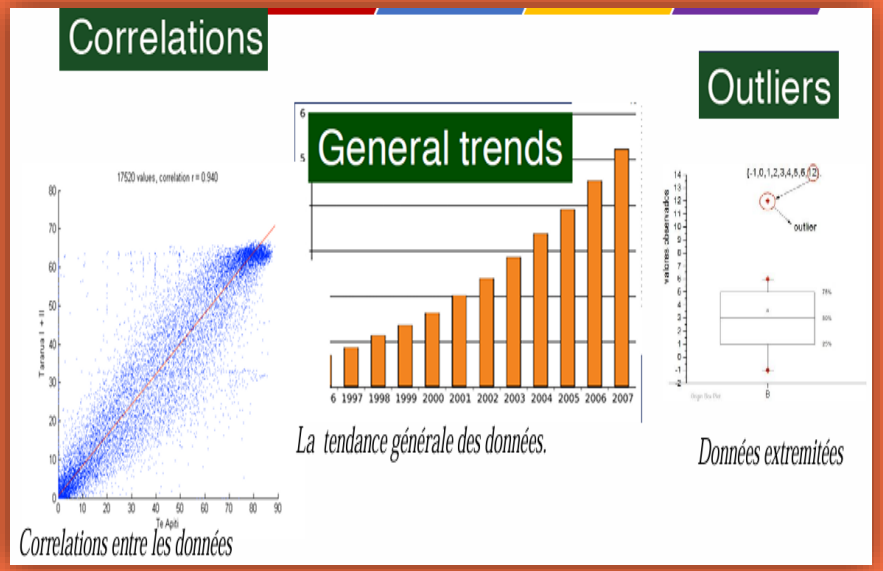
Ces données sont souvent caractérisées par leur volume, leur variété et leur vélocité.

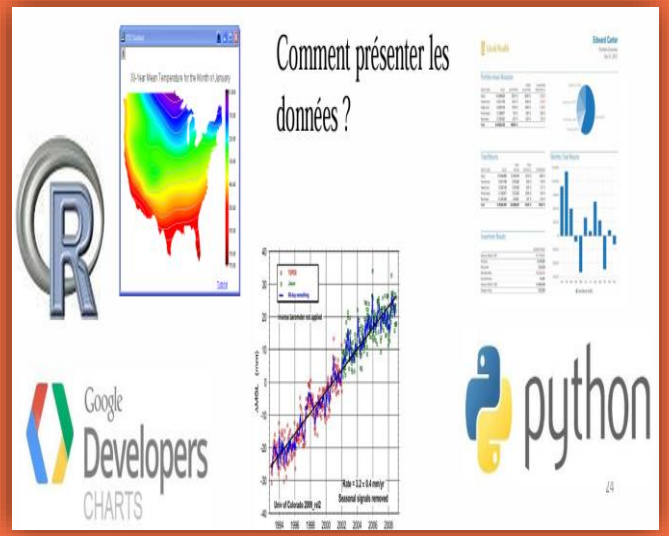
Le Big Data offre de nombreuses opportunités pour l'analyse et la prise de décision, mais nécessite des outils et des techniques spécifiques pour être exploité efficacement

ZiND!

BIG DATA WORKFLOW

Acquire → Prepare → Analyze → Report





ZiND!

BIG DATA TOOLS

cassandra

APACHE Spark™

hadoop

mongoDB

Flink

APACHE STORM™

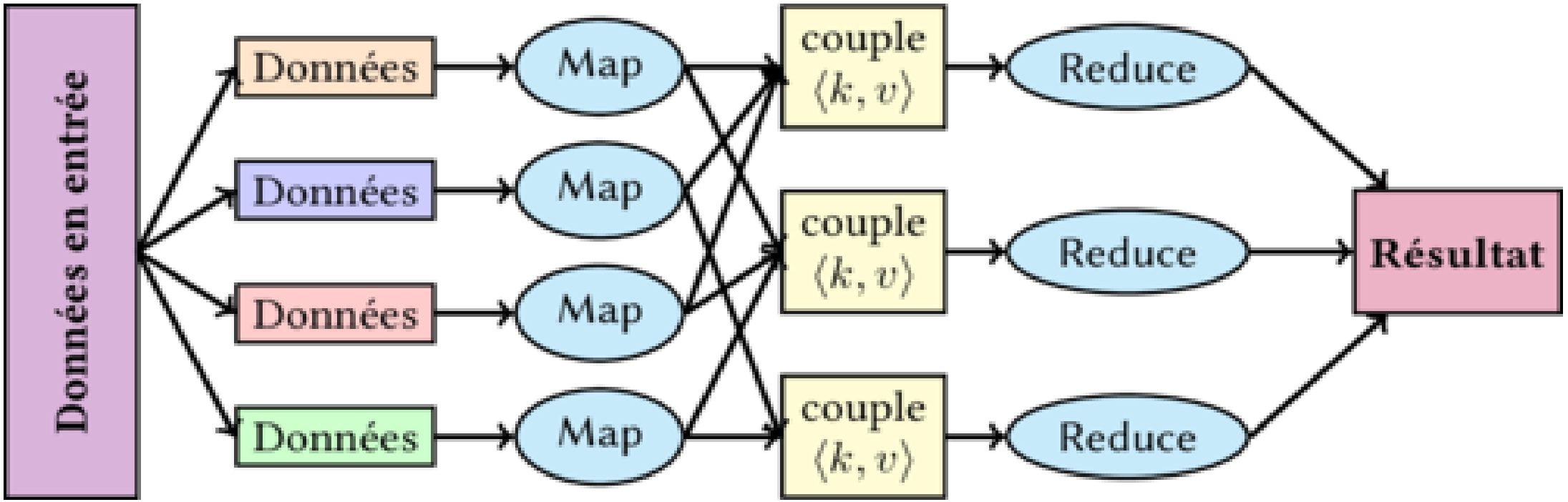
ZiND!


MAPREDUCE

WHAT IS MAPREDUCE

> MapReduce est un modèle de programmation et un framework de traitement de données largement utilisé dans le domaine du Big Data. Il permet de traiter de grandes quantités de données en parallèle sur un cluster de machines. Le modèle MapReduce se compose de deux étapes principales : la phase de "map" qui effectue des opérations de transformation sur les données, et la phase de "reduce" qui agrège les résultats des opérations de "map". [1,2,3]

WHAT IS MAPREDUCE ?

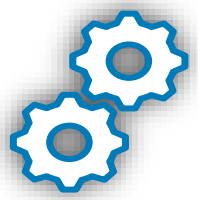


WHAT IS MAPREDUCE?

Traitement distribué

Le **traitement distribué des données** est un ensemble de calcul (traitement + analyse) des données dont l'exécution est parallélisée entre plusieurs machines.

ZiND!



CAS D'UTILISATION DE MAPREDUCE & PySPARK

Analyse de données

- Analyses sur de grandes quantités de données, telles que l'analyse de tendances, la détection d'anomalies ou la segmentation de clients.

Traitement de données en temps réel

- Traiter des flux de données en temps réel, tels que les données de capteurs ou les données de médias sociaux

Machine Learning

- Traitement de données en machine learning, telles que l'entraînement de modèles de prédiction ou la classification de données.

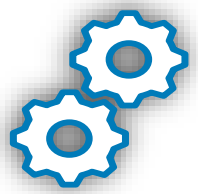
REFERENCES BIBLIOGRAPHIQUE

[1] J. Dean et S. Ghemawat, “MapReduce : simplified data processing on large clusters,” Communications of the ACM, t. 51, no 1

[2] A. Sere, J. S. D. Ouattara, D. Bassole, **J. A. Ouedraogo** et M. Kabore, “A Framework for Data Research in GIS Database using Meshing Techniques and the Map-Reduce Algorithm,” International Journal of Advanced Computer Science and Applications, t. 12, no 3, 2021.
doi : 10.14569/IJACSA.2021.0120374. adresse : http://dx.doi.org/10.14569/IJACSA.2021.0120374.

[3] A. Sere, **J. A. Ouedraogo**, B. Zerbo et O. Sie, “Post classification in the social networks using the map-reduce algorithm,” International Journal of Advanced Computer Science and Applications (IJACSA), t. 11, no 12, 2020, issn : 21565570.
doi : 10.14569/IJACSA.2020.0111293.
adresse: https://thesai.org/Publications/ViewPaper?Volume=11&Issue=12&Code=IJACSA&SerialNo=93

ZiND!



REFERENCES BIBLIOGRAPHIQUE

https://www.infoq.com/fr/articles/apache-spark-introduction/

https://www.talend.com/fr/resources/what-is-mapreduce/

https://insatunisia.github.io/TP-BigData/tp1/

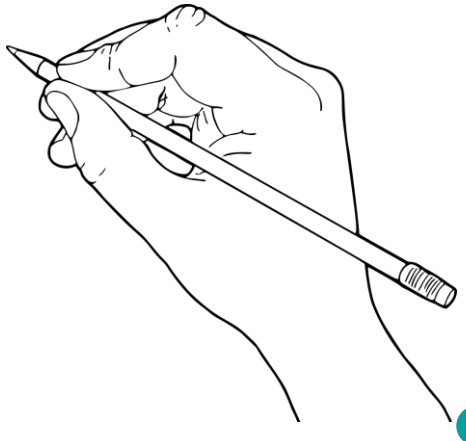
https://k21academy.com/big-data-hadoop-dev/spark-vs-mapreduce/

https://nyu-cds.github.io/python-bigdata/02-mapreduce/

https://medium.com/@haataa/pyspark-basics-6543795fd093

ZiND!

Merci pour votre attention



josearthur.oued@outlook.com

https://www.linkedin.com/in/jos%C3%A9-arthur-ouedraogo-137104167/

ZiND!