Installation de Hadoop et Spark sous WSL et test

par TRAORE ELIE

Sommaire

1- Introduction

- Objectif
- Prérequis

2- Définition des termes

- WSL
- HADOOP
- SPARK

3- Installation des prérequis

- WSL
- Ubuntu ou autre distribution linux
- Java JDK

4- Installation de Hadoop

- Comment télécharger Hadoop
- Configuration des variables d’environnement
- Configuration d’Hadoop
- Démarrage de Hadoop et test

5- Installation de Spark

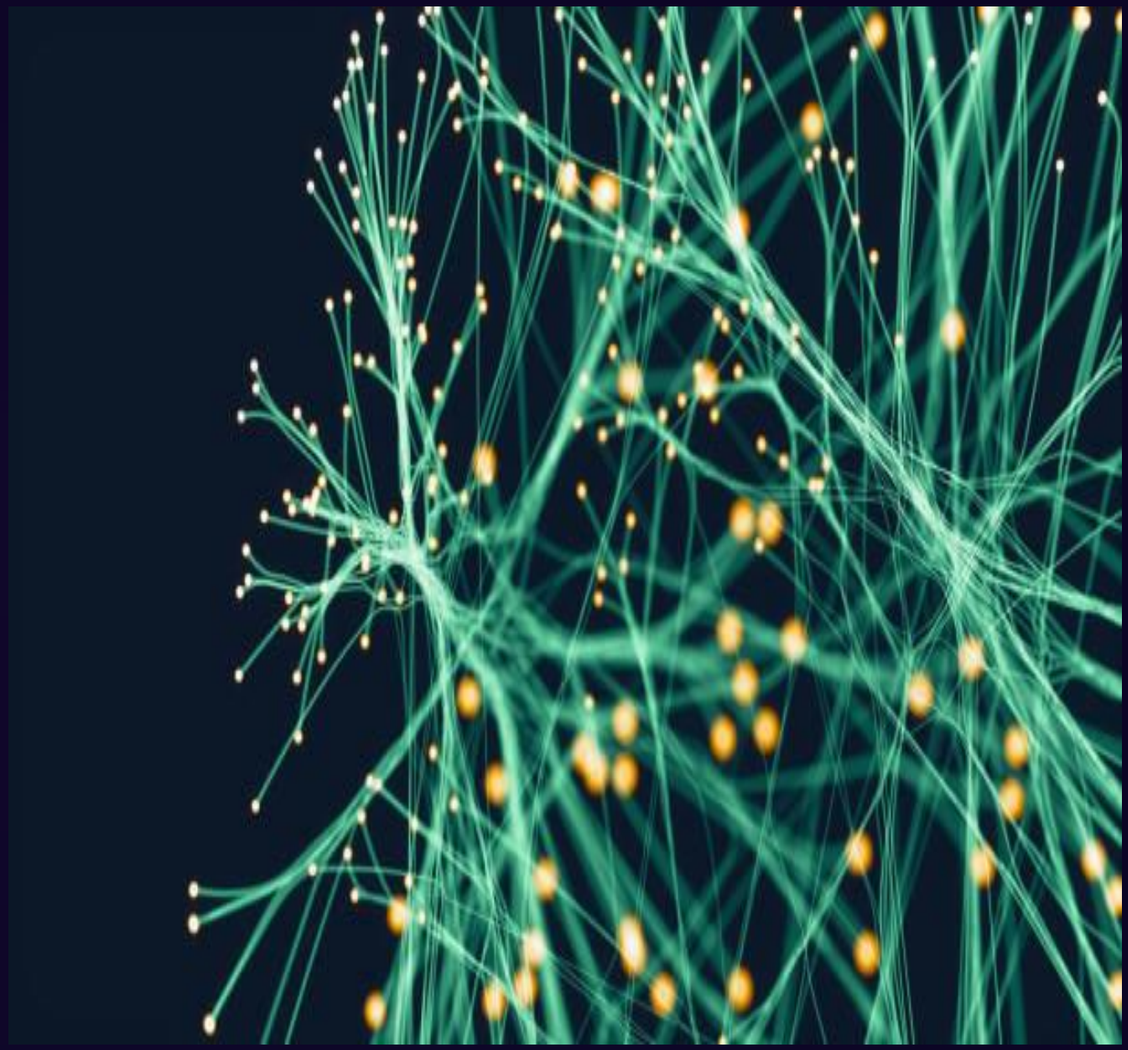
- Comment télécharger Spark
- Configuration des variables d’environnement
- Configuration de Spark et test

6- Conclusion

1- Introduction

Objectif :

Maitriser l’installation de Hadoop et Spark sous Windows Subsystem for Linux (WSL) et pouvoir exécuter un test de fonctionnement.



Prérequis

WSL

WSL doit être installé et activé sur votre système Windows. Vous pouvez l'activer depuis les paramètres Windows.

Distribution Linux

Une distribution Linux comme Ubuntu ou une autre de votre choix est nécessaire sous WSL.

Java JDK

Assurez-vous que Java JDK (version 8 ou supérieure) est installé sur votre distribution Linux.

2- Définition des termes

Qu’est-ce que WSL ?

WSL (Windows Subsystem for Linux), ou en français "Sous-système Windows pour Linux", est une fonctionnalité intégrée à Windows 10 et Windows 11 qui permet d'exécuter des distributions Linux (comme Ubuntu, Debian, etc.) directement sur Windows, sans avoir besoin d'une machine virtuelle ou d'un double démarrage (dual boot).

Qu’est-ce que Hadoop ?

Hadoop est un framework logiciel dédié au stockage et au traitement de larges volumes de données. Il s’agit d’un projet open source, sponsorisé par la fondation Apache Software Foundation.

Qu’est-ce que Spark ?

Apache Spark, le framework d’analyse de données développée par l’université de Berkeley, est aujourd’hui considérée comme l’une des plateformes de big data les plus plébiscitées au monde. Elle compte parmi les « Top-Level Projects » (projets de niveau supérieur) d’Apache Software Foundation. Ce moteur analytique permet de traiter simultanément d’importants volumes de données et d’applications d’analyse de données dans des clusters informatiques distribués.

3- Installation des prérequis

Installation de WSL

1

Activer WSL

Ouvrez PowerShell en tant qu'administrateur et exécutez la commande `dism.exe /online /enable-feature /featurename:Microsoft-Windows-Subsystem-Linux /all /norestart`.

2

Mettre à jour WSL

Exécutez la commande `wsl --update` pour mettre à jour les composants de WSL.

3

Vérifier la virtualisation

Assurez-vous que la virtualisation est activée dans le BIOS. Vous pouvez le vérifier dans le Gestionnaire des tâches de Windows.

4

Activer la plateforme de machine virtuelle

Exécutez la commande `dism.exe /online /enable-feature /featurename:VirtualMachinePlatform /all /norestart` pour WSL 2.

5

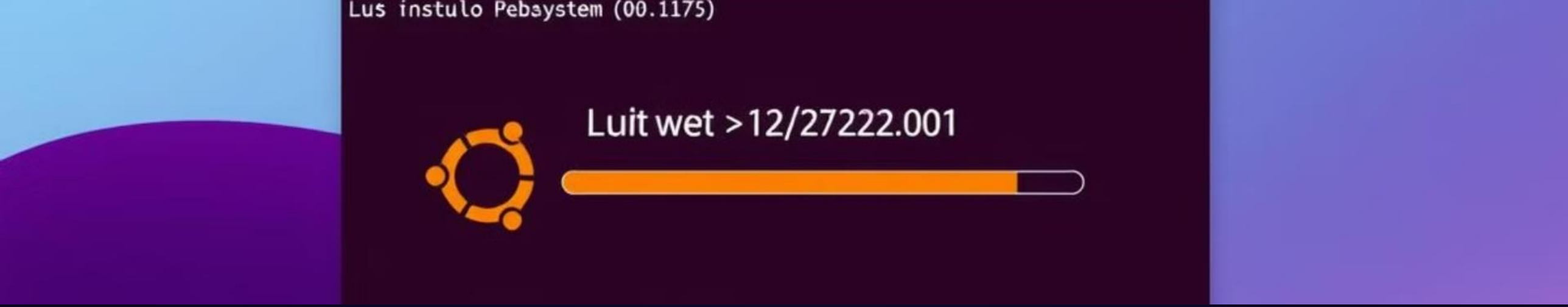
Redémarrer

Redémarrez votre ordinateur pour appliquer les changements.

6

Définir WSL 2 comme version par défaut

Exécutez la commande `wsl --set-default-version 2` pour définir WSL 2 comme version par défaut.



Installation d'Ubuntu

1 Lister les distributions disponibles

Exécutez la commande `wsl --list --online` pour voir quelles distributions peuvent être installées.

2 Installer Ubuntu 22.04 LTS

Utilisez la commande `wsl --install -d Ubuntu-22.04` pour installer Ubuntu 22.04 LTS.

3 Vérifier l'installation

Exécutez la commande `wsl --list --verbose` pour confirmer que l'installation d'Ubuntu 22.04 LTS avec WSL 2 est réussie.

Installation de Java JDK

Mise à jour

Mettez à jour la liste des paquets avec la commande `sudo apt update`.

Installation

Installez Java JDK version 8 avec la commande `sudo apt install openjdk-8-jdk`.

Vérification

Vérifiez la version installée de Java avec la commande `java -version`.

4- Installation de Hadoop

Installation de Hadoop

Téléchargement

Téléchargez Hadoop depuis le site officiel d'Apache en utilisant `wget https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz`.

Extraction

Extrayez les fichiers de l'archive avec la commande `tar -xzvf hadoop-3.4.1.tar.gz`.

Déplacement

Déplacez le dossier extrait vers `/usr/local` avec `sudo mv hadoop-3.4.1 /usr/local/hadoop`.

Configuration des variables d'environnement

1

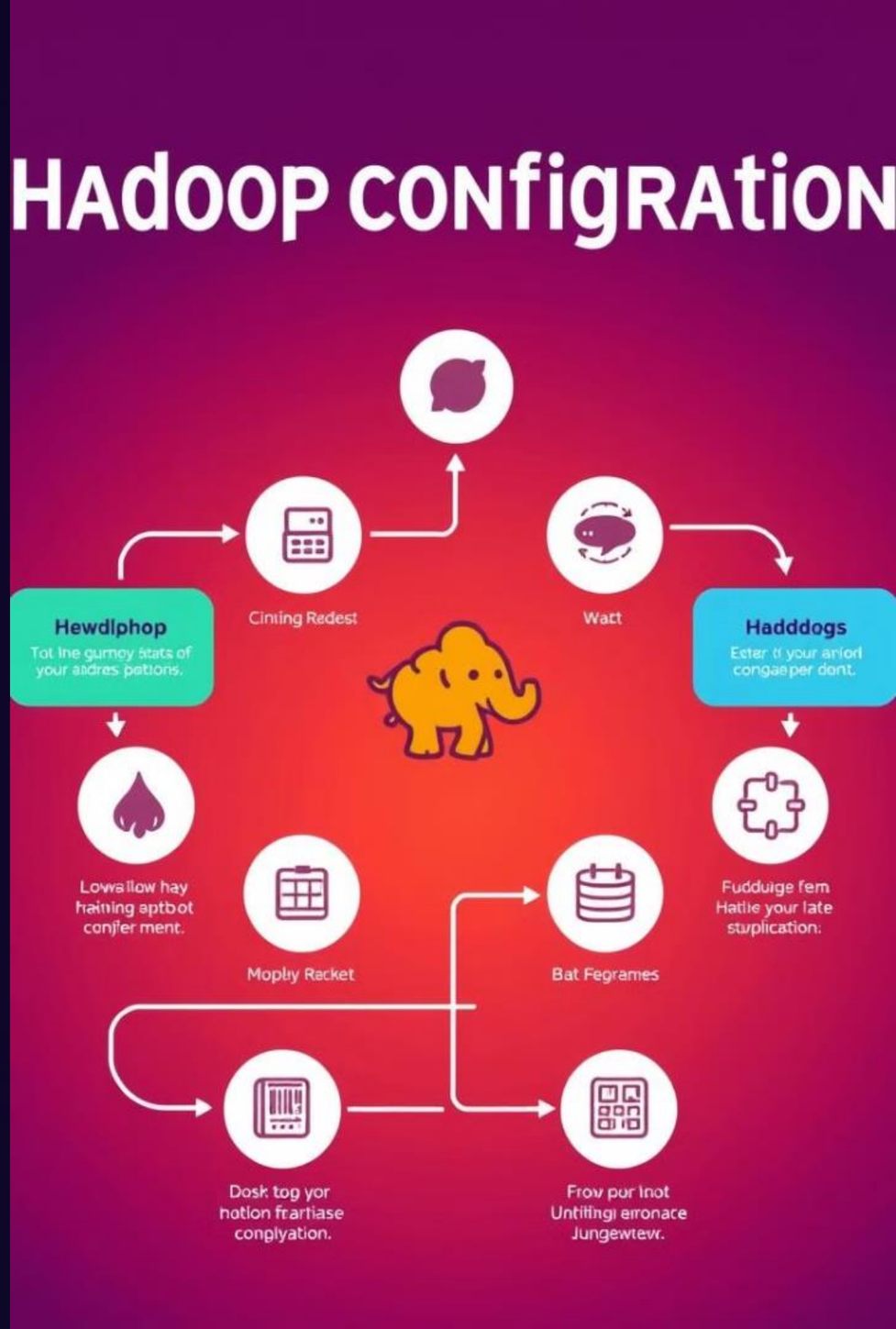
Variables d'environnement

Ajoutez les variables d'environnement nécessaires au fichier `~/.bashrc`.

2

Fichier 'log4j.properties'

Modifiez le fichier `log4j.properties` pour configurer les paramètres de journalisation d'Hadoop.



Configuration de Hadoop

1

Fichier 'hadoop-env.sh'

Créez le fichier `hadoop-env.sh` et configurez l'environnement Java.

2

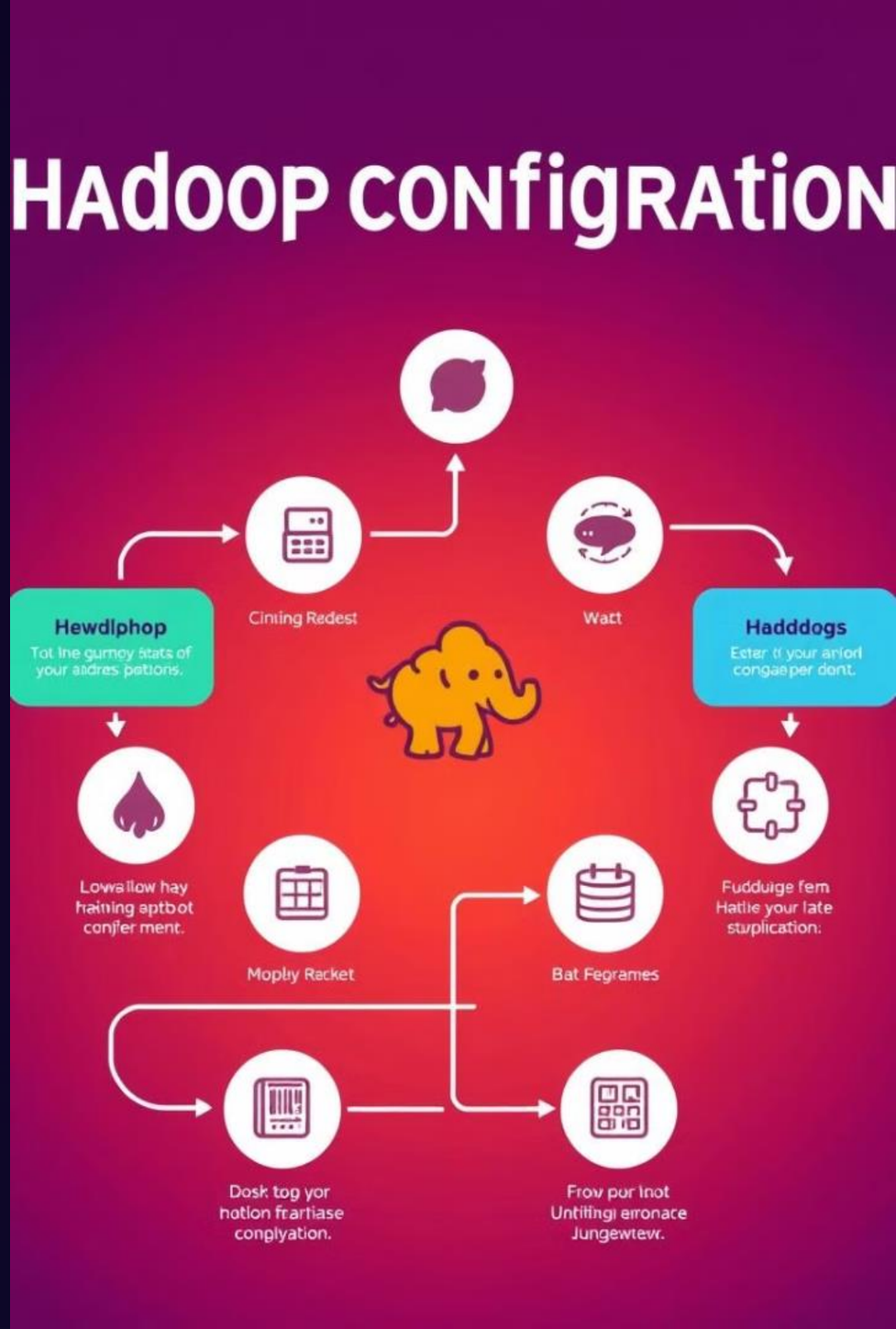
Fichiers 'core-site.xml'

Editer le fichier `core-site.xml` pour le système de fichiers HDFS.

3

Fichier 'hdfs-site.xml'

Editer le fichier `hdfs-site.xml` pour le système de fichiers HDFS.



Démarrage de Hadoop et test

1

Démarrer Hadoop

Exécuter la comande suivante : start-dfs.sh

2

Vérification du fonctionnement :

Accéder à l'interface Web de Hadoop : Ouvrir un navigateur et aller sur 'http://localhost:9870'

3

Crée un répertoire nommé /test dans HDFS

Exécuter la comande suivante : hdfs dfs -mkdir /test

4

Lister le contenu du répertoire racine

Exécuter la comande suivante : hdfs dfs –ls /

5- Installation de Spark

Installation de Spark

Téléchargement

Téléchargez Spark depuis le site officiel d'Apache en utilisant `wget https://downloads.apache.org/spark/spark-3.5.3/spark-3.5.3-bin-hadoop3.tgz `.

Extraction

Extrayez les fichiers de l'archive avec la commande `tar xvf spark-3.5.3-bin-hadoop3.tgz `.

Déplacement

Déplacez le dossier extrait vers `/usr/local` avec `tar xvf spark-3.5.3-bin-hadoop3.tgz`.

Configuration des variables d’environnement

1

Variables d’environnement

Ajoutez les variables d'environnement nécessaires au fichier `~/.bashrc`.

Configuration de Hadoop et test

1

Fichier ‘test_spark.py’

Créez le fichier `test_spark.py` et l’éditer.

2

Fichier ‘test2_spark.py’

Créez le fichier `test_spark.py` et l’éditer.

Pour le test exécuter la commande :
spark-submit test_spark.py

6- Conclusion

Conclusion

Vous avez maintenant une infrastructure Hadoop et Spark opérationnelle sous WSL. N'hésitez pas à explorer et à expérimenter avec ces outils pour analyser vos données volumineuses et découvrir de nouvelles possibilités.

MERCI !

Pour plus d’informations :

TRAORE ELIE

elietraore79@gmail.com