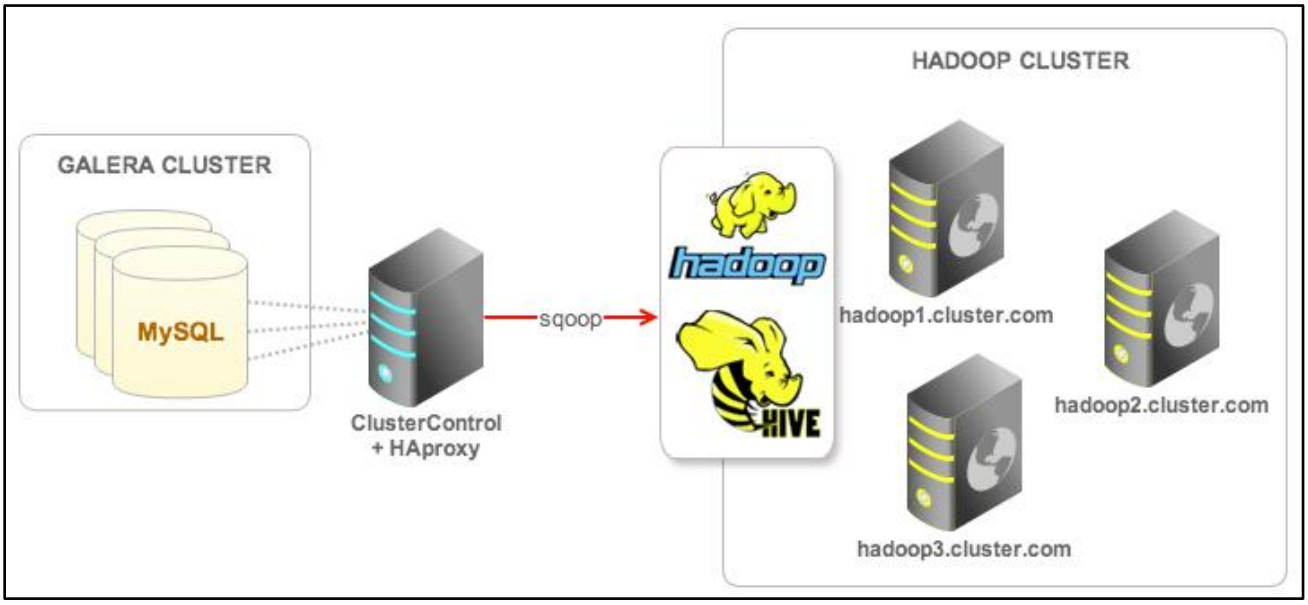
esprit Se former autrement HONORIS UNITED UNIVERSITIES	**Workshop Advanced Big Data** **Option :DS | SAE**

I. Introduction et Contexte

Dans le cadre de ce workshop, nous explorerons comment utiliser **Sqoop**, **Hive** et **MapReduce** pour gérer et analyser des données massives. L'objectif principal est de permettre aux participants de comprendre comment importer des données d'une base **MySQL** dans l'environnement **Hadoop**, tout en les intégrant dans des tables **Hive** pour une analyse plus approfondie. Les étudiants travailleront sur un exemple de données relatives à un système de gestion des élections, comprenant les informations sur les candidats, les électeurs, les districts, les votes, etc.

Les bases de données source sont présentes dans **MySQL**, et contiennent plusieurs tables importantes liées au processus électoral. Nous commencerons par l'importation des données à l'aide de **Sqoop**, sans métadonnées, puis nous évoluerons vers une intégration complète avec **Hive**.



esprit Se former autrement HONORIS UNITED UNIVERSITIES	**Workshop Advanced Big Data** **Option :DS | SAE**

II. Objectifs du Workshop

À la fin de ce workshop, les participants seront capables de :

- ✓ Comprendre les concepts de base de **Sqoop**, **Hive** et **MapReduce**.
- ✓ Importer des données d'une base de données **MySQL** dans **Hadoop** à l'aide de **Sqoop**.
- ✓ Créer des tables **Hive** et charger des données dans **Hive** avec métadonnées.
- ✓ Interroger les tables **Hive** à des fins d'analyse.
- ✓ Exécuter des jobs **MapReduce** simples pour traiter les données massives.

III. Première Partie : Création des Tables dans MySQL et Importation avec Sqoop

1. Étape 1 : Connexion à MySQL et création des tables

Avant de passer à l'importation des données, vous devez vous connecter à la base de données **MySQL** sur Cloudera et créer les tables qui serviront de source pour l'importation.

Connexion à MySQL

- ➢ Pour commencer, ouvrez votre terminal et utilisez la commande suivante pour vous connecter à la base de données **MySQL**:

```
mysql -u root -p
```

- ➢ Ensuite, entrer le mot de passe *cloudera.*

esprit Se former autrement HONORIS UNITED UNIVERSITIES	**Workshop Advanced Big Data** **Option :DS | SAE**

Exécution du script

- Avant de lancer le script, assurez-vous qu'il est bien copié dans un emplacement accessible sur votre système local. Une fois connecté, vous pouvez exécuter le script en utilisant la commande suivante :

```
SOURCE  /emplacement_du_script/Elections_Tunisie_2024.sql;
```

Vérification de la création de la base et des tables

- Après avoir exécuté le script, vous devez vérifier que la base de données et les tables ont été créées avec succès. Utiliser les commandes suivantes :

```
SHOW DATABASES;  -- pour vérifier la création de la base de données
SHOW TABLES;     -- pour vérifier la création des tables
```

Exploration des données

Une fois que vous avez confirmé la création de la base de données et des tables, vous aurez la possibilité de consulter le contenu des tables et de découvrir leur structure ainsi que les liens entre elles.

2. Étape 2 : Importation avec Sqoop (sans métadonnées)

Dans cette étape, nous allons importer l'ensemble de la base de données *elections_tunisie_2024* depuis **MySQL** vers **Hive** en utilisant **Sqoop**. Bien que nous allons importer toutes les tables, nous nous concentrerons principalement sur la table « candidats » pour illustrer le processus.

Création de la base de données sous Hive

- Lancer votre terminal **Hive** et exécuter les commandes suivantes pour créer une base de données destinée à stocker nos données importées :

```
CREATE DATABASE elections; -- Crée la base de données
USE elections; -- Sélectionne la base de données pour les commandes suivantes
```

Vérification de la création du dossier correspondant sous HDFS

- Avant de procéder à l'importation, il est bon de vérifier que le dossier correspondant à notre base de données existe sous **HDFS**. Utiliser la commande suivante :

```
hdfs dfs -ls /user/hive/warehouse/
```

Importation de la table candidats avec Sqoop

Nous allons maintenant importer l'ensemble de la base de données *elections_tunisie_2024*.

- Lancer la commande **sqoop** suivante :

```
sqoop import-all-tables \
--connect jdbc:mysql://localhost:3306/elections_tunisie_2024 \
--username cloudera \
--password cloudera \
--target-dir /user/hive/warehouse/elections.db \
-m 1
```

Note : En fonction du volume des données à importer, cette commande peut prendre un certain temps pour se terminer, car elle lancera des opérations (jobs) **MapReduce** pour transférer les données de **MySQL** vers **HDFS.**

Vérification de l'importation des fichiers dans elections.db

- Pour vérifier que les fichiers ont bien été importés dans *elections.db*, exécuter la commande suivante :

```
hdfs dfs -ls /user/hive/warehouse/elections.db
```

Bien que nous ayons importé toutes les tables, nous allons nous concentrer sur la création de la table « candidats ». Nous ne verrons pas de table « candidats » dans notre base de données **Hive**, car nous avons seulement importé les données sans métadonnées.

esprit Se former autrement HONORIS UNITED UNIVERSITIES	**Workshop Advanced Big Data** **Option :DS | SAE**

Création de la table candidats dans Hive

- Lancer la commande suivante pour créer la table « candidats » dans **Hive** :

```
CREATE TABLE candidats (id_candidat INT, nom STRING, parti STRING,age INT, experience STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION '/user/hive/warehouse/elections.db/candidats';
```

Note : Les noms des colonnes peuvent être obtenus à partir de la table candidats dans MySQL que nous avons créée précédemment.

Vérification dans Hive

- Vérifier les données importées en interrogeant la table « candidats » dans **Hive** :

```
SELECT * FROM candidats;
```

3. Étape 3 : À Vous de Jouer !

Maintenant que vous avez importé la base de données *elections_tunisie_2024* sans métadonnées et créé la table candidats dans **Hive**, il est temps de passer à l'étape suivante.

Votre mission :

1. **Créer une base de données Hive appelée electionsMeta.**
2. **Importer les données de la base de données elections_tunisie_2024 depuis MySQL en utilisant Sqoop**, mais cette fois-ci en incluant les métadonnées.
3. **Créer les tables dans la base de données Hive** en exploitant les données et métadonnées importées.
4. **Vérifier vos résultats** pour vous assurer que tout est en ordre.

Bon travail !

Cette activité vous aidera à mieux comprendre le processus d'importation avec métadonnées et à acquérir de l'expérience dans le travail avec des tables **Hive**. Prenez votre temps, et n'hésitez pas à poser des questions si vous en avez !

IV. Deuxième Partie : Analyse et Traitement des Données avec Hive et MapReduce

Après avoir importé avec succès vos données électorales dans **Hive** à l'aide de **Sqoop** et créé les tables contenant des informations sur les candidats, les électeurs, les districts électoraux et les votes, vous êtes maintenant prêts à plonger dans la partie traitement et analyse de ces données. Ces données sont précieuses et peuvent fournir des insights essentiels sur les tendances électorales et le comportement des électeurs.

1. Étape 1 : Traitement des Données avec Hive

Votre première mission consiste à effectuer une analyse de base sur ces données en utilisant **Hive**. Imaginer que vous êtes des analystes électoraux qui doivent déterminer combien de votes chaque candidat a reçus dans chaque district. Pour cela, vous allez exécuter une requête **Hive** qui compte le nombre total de votes par candidat et par district.

En exécutant cette requête, vous allez découvrir des résultats qui vous aideront à comprendre la popularité des candidats dans différentes régions. C'est l'occasion de tirer des conclusions sur qui mène dans les différentes circonscriptions.

2. Étape 2 : Traduire le Traitement en MapReduce

Une fois que vous avez extrait ces informations en utilisant **Hive**, vous êtes prêts à aller plus loin. Vous devez vous transformer en développeurs en **Python**, prêts à répliquer ce traitement à l'aide de **MapReduce**.

Vous allez créer deux fichiers **Python**, **mapper.py** et **reducer.py**, qui effectueront la même tâche que votre requête **Hive**. En utilisant les principes de **MapReduce**, vous devez mapper les votes aux candidats et aux districts, puis réduire ces informations pour obtenir un total par candidat et par district.

Cette étape vous permettra de voir comment les mêmes analyses peuvent être réalisées avec des outils différents, renforçant ainsi votre compréhension des différents écosystèmes de traitement de données.

V. Conclusion du Workshop

Ce workshop vous a permis d'explorer l'importation et le traitement des données dans un environnement **Hadoop**. Vous avez appris à vous connecter à **MySQL**, à importer la base de données *elections_tunisie_2024*, à créer des tables **Hive**, et à traiter les données avec **Hive** et **MapReduce**. Ces compétences vous fourniront une base solide pour travailler avec des données dans un écosystème **Big Data**.

Validation en Classe

Le travail que vous allez réaliser sera validé en classe, avec une session de questions-réponses pour chaque groupe. Cela nous permettra de vérifier si vous avez bien assimilé les concepts abordés.