Université Abdelmalek Essaadi

-Faculté Polydisciplinaire de Larache-

PRÉSENTATION SUR APACHE SQOOP

- **Réalisé par: HAFIANI Salma**
- **Encadré par: Pr.Hicham Gibet Tani**

04 Dec, 2024

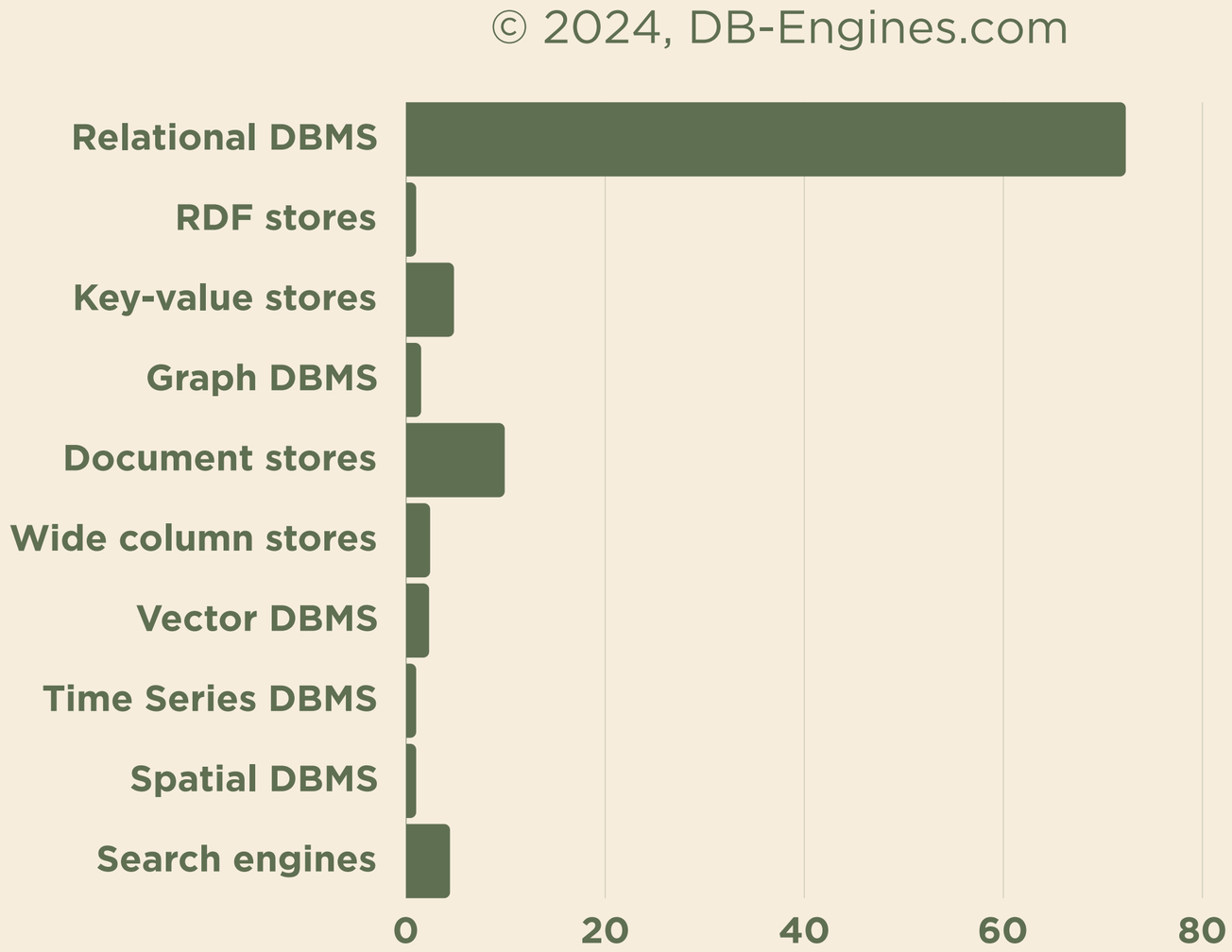
Master DevOps & Cloud Computing

PLAN:

INTRODUCTION

De nos jours, les entreprises expriment des besoins croissants en matière de big data. Malgré tout, SQL reste la technologie la plus utilisée pour stocker des données et il paraît peu probable que les entreprises soient prêtes à migrer durablement tout leur système vers Hadoop.

La conséquence est qu’au sein d’une organisation, des bases de données SQL sont éparpillées sur différents serveurs (chacune ayant un rôle différent), empêchant une analyse big data de l’ensemble des données en même temps.

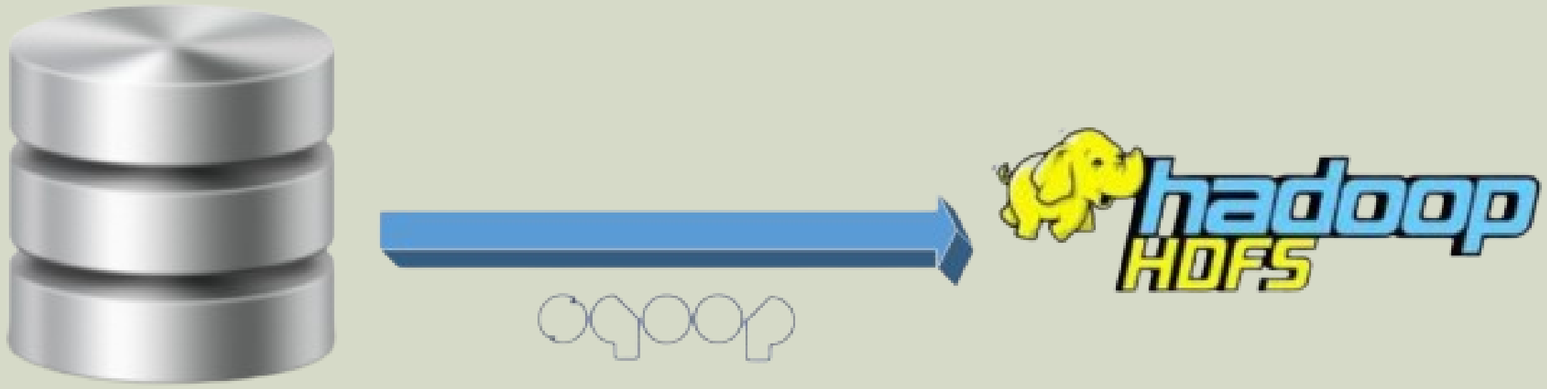


APACHE SQOOP

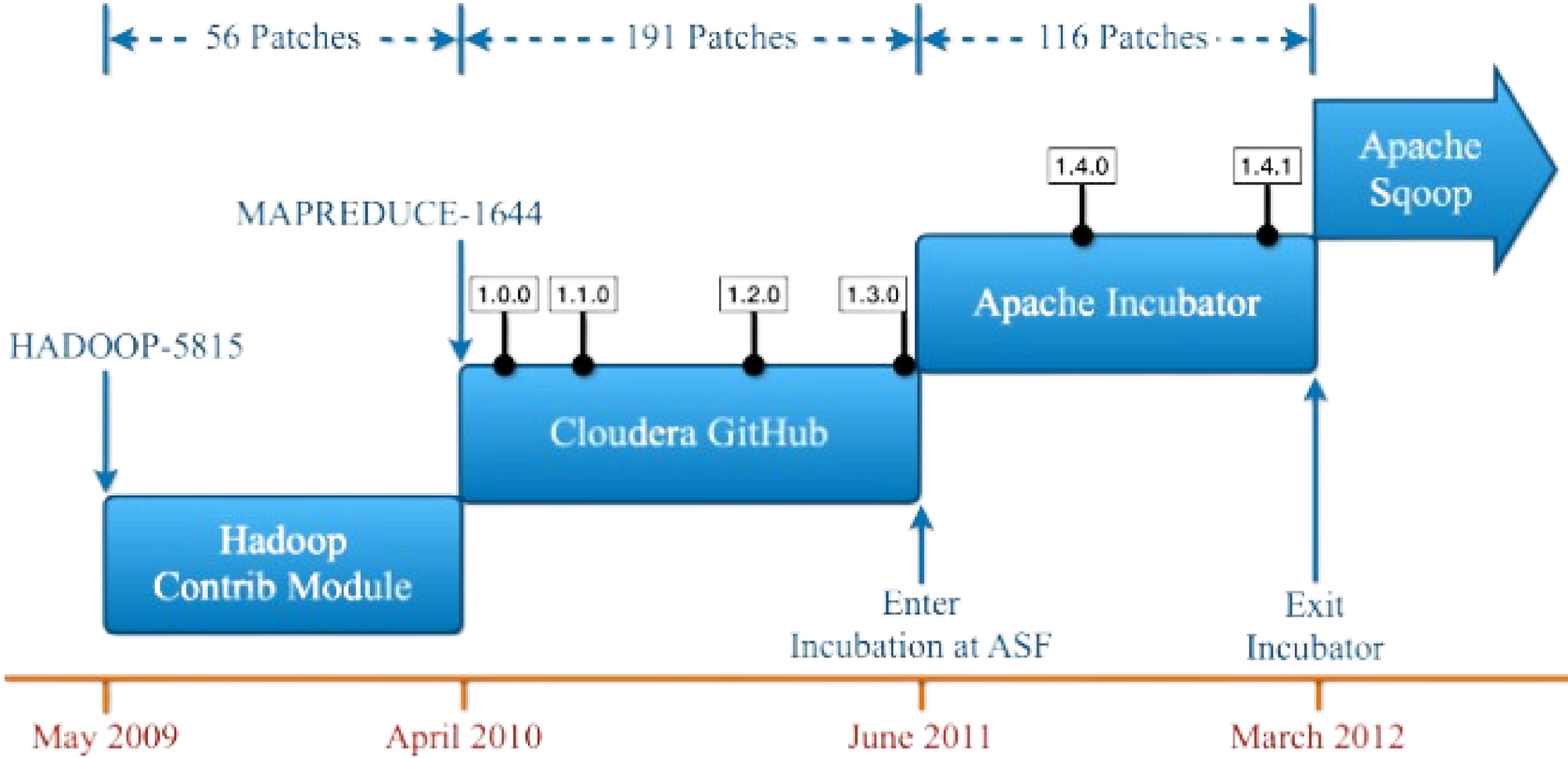
Définition:

Apache-sqoop est un outil open-source conçu pour transférer efficacement les données volumineuses entre les datastores externes comme les bases de données relationnelles (SGBDR), les data warehouses et l'écosystème Hadoop (HDFS, HIVE,..)

- Sqoop a obtenu son nom d'une combinaison des mots **SQ**L et Had**oop**, ce qui illustre sa mission principale : faire le lien entre les bases de données relationnelles et l'écosystème Hadoop.



HISTORIQUE :


56 Patches
191 Patches
116 Patches
MAPREDUCE-1644
HADOOP-5815
1.0.0
1.1.0
1.2.0
1.3.0
1.4.0
1.4.1
Apache Incubator
Apache Sqoop
Cloudera GitHub
Hadoop Contrib Module
Enter Incubation at ASF
Exit Incubator
May 2009
April 2010
June 2011
March 2012

POURQUOI APACHE SQOOP:

Intégration transparente :

Permet de connecter et de transférer facilement de grandes quantités de données entre différentes sources et Hadoop.

Gain de temps et d'efficacité :

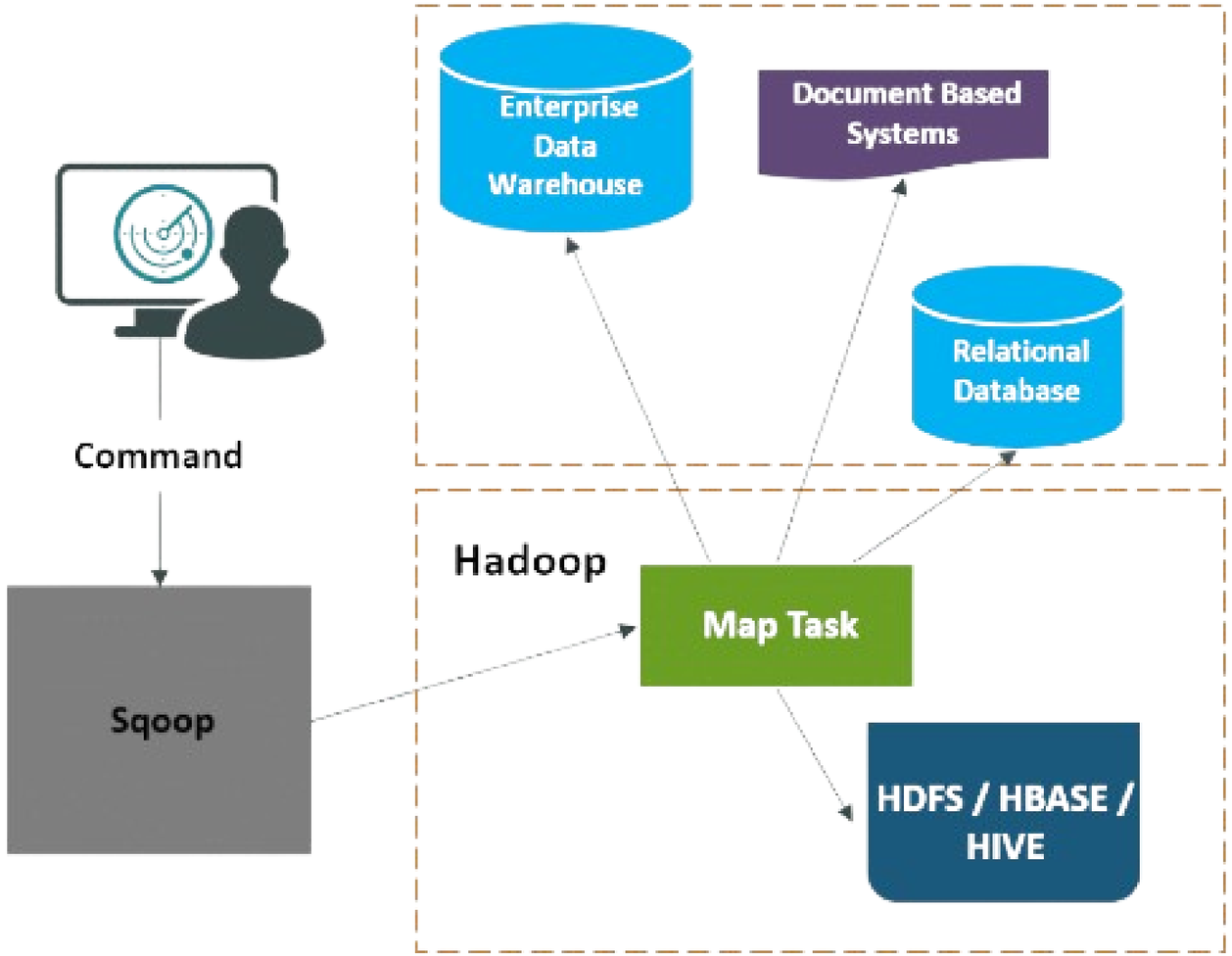
Automatise l’intégration des données, libérant les entreprises pour se concentrer sur l’analyse.

Importations/exportat-ions incrémentielles :

Transfère uniquement les données nouvelles ou mises à jour, réduisant la charge de travail.

ARCHITECTURE DE SQOOP

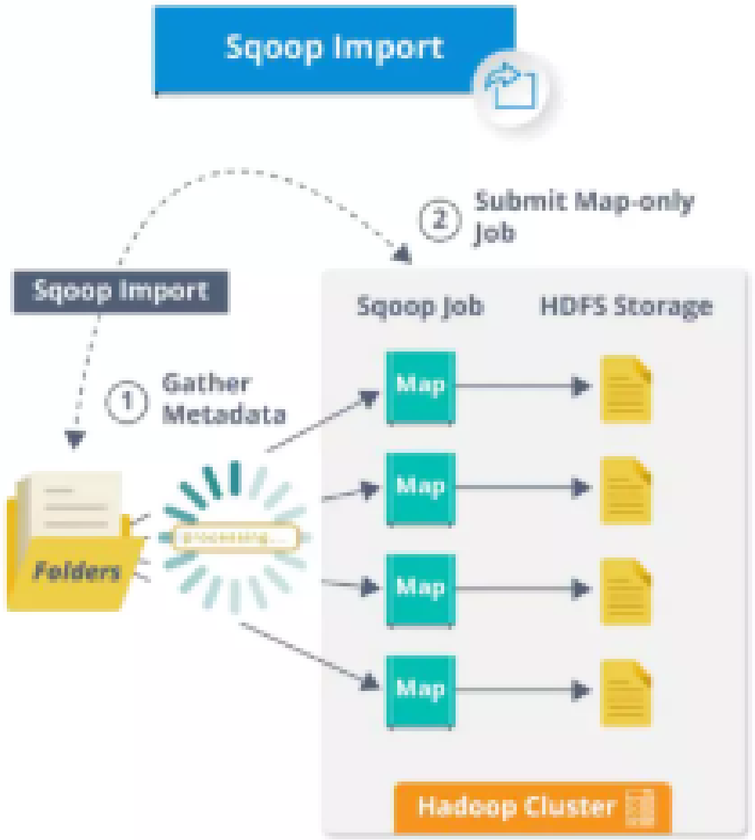
Apache Sqoop est conçu selon une architecture client/serveur. Le Sqoop s’exécute sur le cluster Hadoop et communique avec le serveur Sqoop qui se trouve sur le système de stockage relationnel.



ARCHITECTURE DE SQOOP

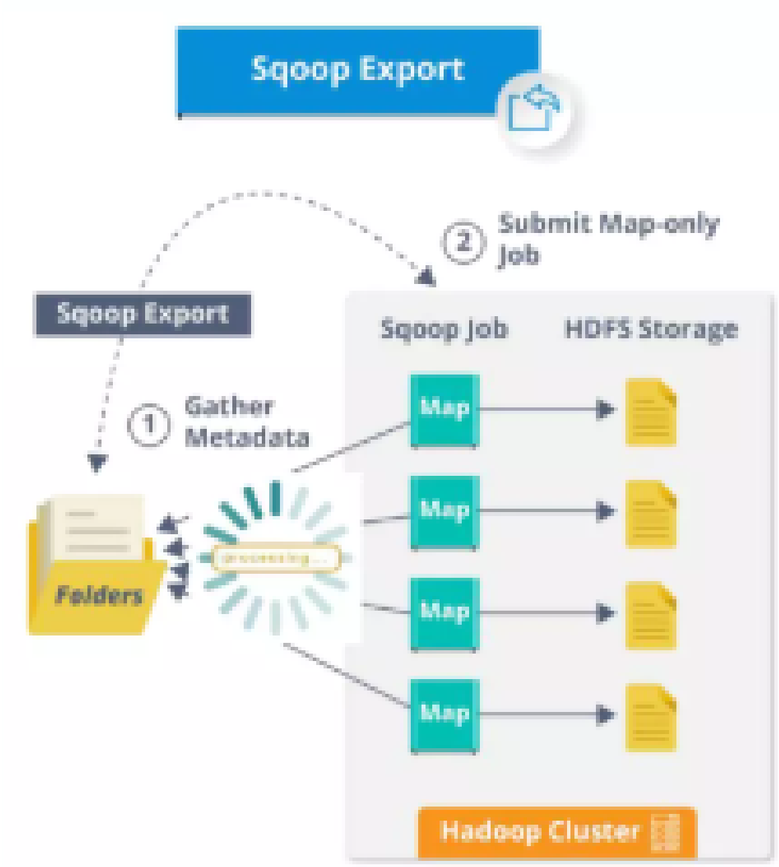
Soumission de commande : Une commande Sqoop est utilisée pour importer ou exporter des données.
Récupération des données : Sqoop récupère les données depuis diverses sources, tq les entrepôts de données d'entreprise, les bases de données relationnelles ou les systèmes basés sur des documents.
Exécution des mappers : Sqoop exécute plusieurs mappers pour charger ou exporter les données vers ou depuis HDFS, Hive ou HBase.

ARCHITECTURE SQOOP IMPORT:



- **Analyse des bases de données:** Sqoop commence par analyser les bases de données et récupérer les métadonnées nécessaires pour accéder aux données.
- **Utilisation de Map-Only MapReduce:** Sqoop utilise un job MapReduce sans phase de réduction (car il n'y a pas d'agrégation) pour transférer les données.
- **Transfert vers HDFS:** Les données sont transférées dans HDFS sous forme de fichiers associés à la table récupérée.
- **Personnalisation du répertoire:** Sqoop permet de spécifier un répertoire de destination personnalisé dans HDFS pour stocker les données.

ARCHITECTURE SQOOP IMPORT:



- Les tables doivent exister dans le RDBMS.
- Les données sont transférées via des tâches "Map" en plusieurs transactions.
- Une table de staging peut être utilisée pour sécuriser les transferts et éviter les pertes de données en cas d’échec.

LES FONCTIONNALITÉS CLÉS DE SQOOP

Importation et exportation de données :

Transfert bidirectionnel entre bases relationnelles et Hadoop via MapReduce, Hive ou Spark.

Compatibilité multi-sources :

Prend en charge divers SGBD (MySQL, Oracle, PostgreSQL, SQL Server, etc.) sans nécessiter de code spécifique.

Parallélisme et tolérance aux pannes :

Exécute des tâches en parallèle pour une efficacité accrue et assure la reprise en cas d’échec.

LES COMMANDES SQOOP

La syntaxe générale d'une commande Sqoop est la suivante :

sqoop <commande> [options]

Commande	Description
version	Afficher les information de version
eval	Evaluer une instruction SQL et afficher les résultats
export	Exporter un répertoire HDFS vers une table de base de données
help	Lister les commandes disponibles
import	Importer une table d'une base de données vers HDFS
Import-all-tables	Importer des tables d'une base de données vers HDFS
list-databases	Lister les bases de données disponibles sur un serveur
list-tables	Lister les tables disponibles dans une base de données
codegen	Générer du code pour interagir avec les enregistrement de base de données

LES COMMANDES SQOOP

La commande d'importation :
Syntaxe :

Sqoop import (generic-args) (eval-args)

Les arguments de la commande d'importation :

Arguments	Description
connect	Fournir la chaine de connexion (connect contient l'URL de JDBC)
username	Définit le nom de l'utilisateur
P	Demande le mot de passe dans la console
password	Définit le mot de passe
table	Définit le de la table à importer
target-dir	Définit le répertoire HDFS de destination
fields-terminated-by	Définit le séparateur des valeurs des données importées dans HDFS (par défaut « , »)
hive-import	Importe une table dans Hive
create-hive-table	Importe la définition d'une table dans Hive
warehouse-dire	HDFS parent pour destination de la table

LES COMMANDES SQOOP

La commande d’exportation :
Syntaxe :

Sqoop export (generic-args)(export-args)

• Les arguments de la commande d’exportation :

Arguments	Description
connect	Fournit la chaine de connexion
username	Définit le nom de l’utilisateur
p	Demande le mot de passe dans la console, sinon vous pouvez utiliser
password	Définit le mot de passe
table	Définit le nom de la table à charger
-m, -num-mappers	Utilise n tâche du « Map » pour exporter en parallèle
export-dir	Chemin de la source HDFS pour l’exportation
columns	Définit les colonnes à exporter dans une base de données

LES AVANTAGES ET LES LIMITES D'APACHE SQOOP

LES AVANTAGES

Transferts efficaces:

Permet l'importation et l'exportation de gros volumes de données entre bases relationnelles et Hadoop rapidement et efficacement.

Réduction de la charge manuelle :

Automatise les transferts de données, évitant l'écriture de scripts complexes.

Flexibilité :

Permet des importations/exportations incrémentielles, limitant le transfert aux nouvelles données uniquement.

LES LIMITES

Maintenance réduite:

Le projet Sqoop n'est plus activement développé par la communauté Apache, ce qui peut poser des problèmes de compatibilité avec les technologies modernes.

Performance limitée pour les données complexes:

Moins efficace pour des formats de données très complexes ou non structurés.

APACHE SQOOP VS APACHE FLUME

Critère	Apache Sqoop	Apache Flume
Objectif principal	Importation/exportation de données entre RDBMS et Hadoop	Collecte et transport de données en temps réel vers Hadoop
Type de données	Données structurées (tables RDBMS)	Données non structurées ou semi-structurées (logs, événements)
Mode de transfert	Par lots (batch)	En temps réel (streaming)
Cas d'utilisation	Transfert de données massives depuis/vers RDBMS	Collecte de données en temps réel (logs, IoT, événements)
Architecture	Client-serveur (client Sqoop, serveur Hadoop)	Décentralisée (sources, canaux, puits)
Performance	Optimisé pour les transferts massifs de données	Optimisé pour les flux de données en temps réel
Complexité de gestion	Relativement simple (connexion et transfert)	Plus complexe (gestion des flux de données en continu)
Parallélisme	Utilise le parallélisme pour accélérer les transferts	Gère des événements multiples en continu

CONCLUSION

En résumé, Apache Sqoop facilite l'importation et l'exportation de données entre les bases relationnelles et Hadoop, avec des transferts efficaces. Pour les besoins en temps réel, il peut être couplé avec d'autres outils comme Apache Flume.

MERCI DE VOTRE ATTENTION!