

## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

---

Ce workshop vise à prendre en main le concept de l'ETL, qui consiste à extraire des données hétérogènes depuis n'importe quelle source, à les manipuler, puis à les stocker pour une éventuelle analyse.

Le WS est devisé en deux parties, dans la 1<sup>ère</sup>, vous apprendrez à prendre en main l'outil Talend TOSBD, et la 2<sup>ème</sup> partie consistera à construire le modèle multidimensionnel et à créer un Datalake.

Quelques figures sont présentes pour aider à la réalisation des différentes manipulations.

#### **Partie 1 :**

Cette partie consiste à prendre en main Talend en manipulant quelques composants.

Commencez par créer votre Job, puis ajoutez le composant « tRowGenerator » qui nous permettra de générer de nouvelles données. Nous souhaitons obtenir en sortie:

- Id : de type entier
- Nom
- Prénom
- Date de naissance : comprise entre 2000 et 2010
- Ville de naissance
- Une adresse mail sous forme [prenom.nom@cesi.fr](mailto:prenom.nom@cesi.fr)

Nous souhaitons également les trier par date de naissance et générer trois fichiers en sortie : un fichier XML et deux fichiers Excel. L'un des deux fichiers contiendra uniquement les personnes âgées de plus de 20 ans.

Voici un aperçu de la configuration :

- Le composant « tRowGenerator » est utilisé pour générer des données
- Le composant « tSortRow » est utilisé pour trier les données
- Le composant « tMap » est utilisé pour le traitement, le filtrage des données
- Les composants « tFileOutputExcel » et « tFileOutputXML » sont utilisés pour exporter les données vers des fichiers Excel et XML respectivement.

## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

---

#### **Partie 2 :**

Commencez par télécharger la base de données des ventes que notre client nous a envoyé à travers ce lien : [https://drive.google.com/drive/folders/1Vw2BgpVtvXZw85y1nEx1jbdI5gJ\\_Yul6?usp=sharing](https://drive.google.com/drive/folders/1Vw2BgpVtvXZw85y1nEx1jbdI5gJ_Yul6?usp=sharing)

Assurez-vous d'avoir dans le dossier trois fichiers au format csv ( *Features data set.csv* , *sales data-set.csv* , *stores data-set.csv* ).

La base de données concerne les données de vente, qui peuvent fournir des informations précieuses sur les performances passées, les tendances du marché, les préférences des clients et d'autres aspects commerciaux. Le but de cette partie est d'analyser ces données afin que les entreprises puissent prendre des décisions éclairées sur leur stratégie de vente, leur tarification, leur marketing, etc.

Commençons d'abord par l'extraction des données sources (**ETL**)

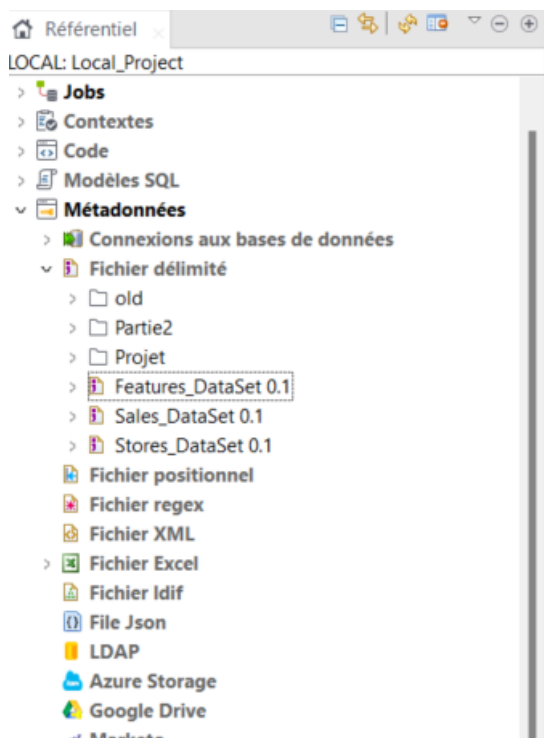
1. Lancez Talend et créez votre Projet/Workspace.
2. Récupérez et chargez les données sources des 3 fichiers CSV. Utilisez l'option "Fichier délimité" pour extraire les données sources. Allez dans le Référentiel, puis accédez à "Métadonnées", et cliquez avec le bouton droit sur "Fichier délimité" afin de charger les 3 fichiers CSV.

Assurez-vous d'avoir récupéré les trois fichiers de données.

## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

---



Nous allons maintenant passer à l'étape de traitement (**ETL**).

Pour rappel, le client nous a informé de l'existence de plusieurs factures erronées, des clients en double et plusieurs enregistrements vides !

**Remarque :** Pour ajouter des composants Talend sur le Designer, par exemple "tMap", vous pouvez soit passer par la "Palette", puis dérouler "Transformation" et glisser le composant "tMap" sur le Designer. Vous pouvez également rechercher n'importe quel composant dans la barre de recherche. Une autre manière de sélectionner un composant est de se mettre dans le Designer, de cliquer sur l'interface et d'écrire la lettre "t" suivie du nom du composant.

1. Créez votre premier Job.

## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

2. Déposez maintenant le premier fichier "Features\_DataSet" sur le Designer (choisissez tFileInputDelimited).
3. Commencez par analyser les dates. Que remarquez-vous ?
4. Rajoutez le composant "tMap" qui nous permettra d'effectuer la majorité des traitements.
5. Reliez les deux composants : cliquez avec le bouton droit sur le composant "Features\_DataSet" puis cliquez sur "Row", puis sur "Main", et reliez-le au "tMap".
6. Double-cliquez sur tMap. La partie de gauche représente les entrées et celle de droite les sorties. Sur la partie droite, cliquez sur "+" et choisissez un nom. Faites glisser les colonnes d'entrée vers la droite (les sorties) (Store, Temperature, Fuel\_Price, CPI, Unemployment, IsHoliday, date).
7. Trouvez la formule qui nous permettrait d'unifier les dates.
8. Ajoutez le composant "tLogRow" qui nous permettra d'afficher le résultat en sortie. Cliquez avec le bouton droit sur le "tMap", puis sélectionnez "Row", et choisissez le nom de sortie que vous avez mentionné dans le tMap, puis reliez-le au "tLogRow".
9. Exécutez le Job pour voir le résultat.

Nom	Valeur
Features_DataS...	"\n"
Features_DataS...	C:/User:
Features_DataS...	1
Features_DataS...	US-ASC
Features_DataS...	"
BDD_Ventes_A...	
BDD_Ventes_L...	postgre
BDD_Ventes_Sc...	public
BDD_Ventes_S...	localho
BDD_Ventes_P...	5432
BDD_Ventes_P...	Ventes

## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

---

10. Que remarquez-vous ? Avez-vous récupéré l'ensemble des enregistrements ? Que feriez-vous pour résoudre ce problème ?

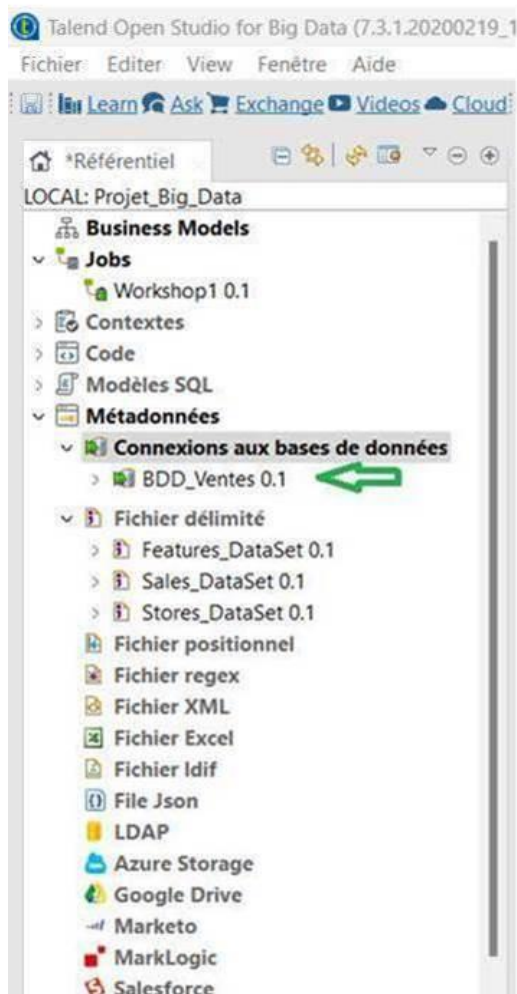
Nous allons maintenant passer à l'étape de chargement (**ETL**)

11. Une fois que vous avez récupéré les 8190 éléments, procédons maintenant à l'alimentation des tables,
- a- Préparons la nouvelle base de données qui accueillera les données du client
  - b- Sous le SGBD PostgreSQL, Lancez « PgAdmin 4 » puis cliquez avec le bouton droit sur « Databases » puis sur « Create » puis sur « Database » et donnez-lui un nom par exemple « Ventes ». Enfin, validez en cliquant sur « Save »
  - c- Une fois que la BDD Ventes est créée et prête à être alimentée.
12. Passons maintenant à connecter Talend vers la BDD Ventes. Toujours dans le Référentiel, cliquez avec le bouton droit sur "Connexions aux bases de données", puis sur "Créer une connexion". Remplissez les informations concernant la base de données et cliquez sur "Next". Choisissez "PostgreSQL" comme SGBD, puis complétez les différents champs relatifs à la base de données "Ventes" créée au préalable. Cliquez ensuite sur "Tester la connexion" pour vous assurer que la connexion est établie. Si ce n'est pas le cas, vérifiez les informations que vous avez saisies. Ensuite, cliquez sur "Exporter en tant que contexte", cochez "Créer un nouveau contexte dans le référentiel" et cliquez sur "Finish".

## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

---



13. Vous allez maintenant ajouter un nouveau composant qui vous permettra de créer et d'alimenter les nouvelles tables : tDBOutput PostgreSQL. Vous avez deux options :
- Soit vous supprimez le composant « tLogRow » et depuis le « tMap », vous reliez la sortie vers tDBOutput,
  - Soit vous reliez directement le composant « tLogRow » au « tDBOutput » .

**Remarque :** Afin de créer la table Feature, il va falloir modifier la taille du champs « unemployment » de 5 à 11

## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

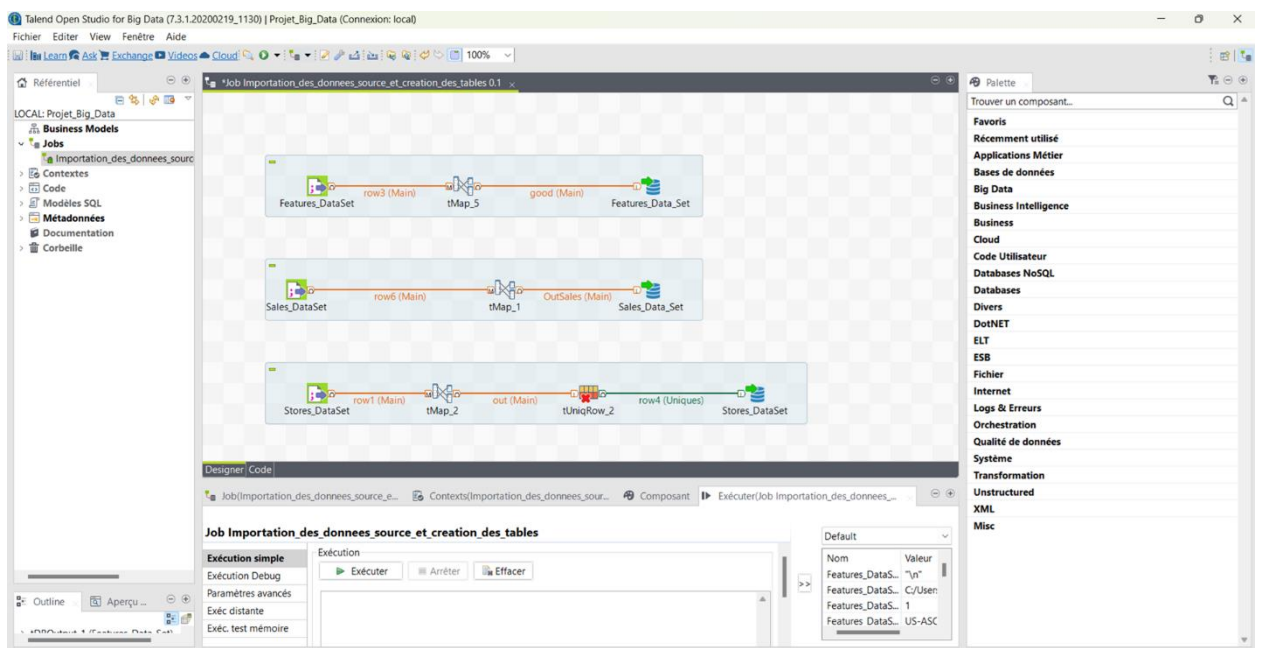
Schéma de Features\_DataSet

Features\_DataSet

Colonne	Clé	Type	<input checked="" type="checkbox"/> N..	Modèle de date (C...	Length	Precision	Défaut	Comment...
Markdown2	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0		
Markdown3	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0		
Markdown4	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0		
Markdown5	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0		
CPI	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		11	8		
Unemployment	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		11	4		
IsHoliday	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		5	0		

OK Cancel

14. Refaites la manip, afin de récupérer les trois fichiers et d'alimenter la BDD  
Faites attention aux enregistrements vides et aux doublons



## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

---

15. Vérifiez la création des trois tables dans PostgreSQL

### **Partie 3 :**

Assurez-vous d'avoir téléchargé et mis en place la VM Cloudera, pour plus d'informations sur l'installation de Cloudera, référez-vous au manuel d'installation.

- Téléchargez la 2<sup>ème</sup> base de données à travers ce lien :  
[https://drive.google.com/drive/folders/1KyQBu4N8tbNslnamrch9mpnDaPoAc6Xe?usp=drive\\_li nk](https://drive.google.com/drive/folders/1KyQBu4N8tbNslnamrch9mpnDaPoAc6Xe?usp=drive_li nk)
- Assurez-vous d'avoir dans le dossier les deux fichiers au format csv (*Customer.csv* & *Product.csv*)

Notre second client souhaite récupérer le chiffre d'affaire par produit et par client afin d'anticiper l'agencement des produits par région et de minimiser les pertes tout en récompensant leurs meilleurs clients

Nous allons d'abord analyser sa requête puis construire le schéma décisionnel, et enfin passer par la suite à l'analyse des données

Commençons par les besoins du client. Pour construire le schéma décisionnel, le choix sera le model en étoile vue qu'on manipule des données venant d'un seul métier en l'occurrence les ventes.

1. Elaborez le schéma décisionnel ? (Les dimensions, la table de faits et les mesures)

**Remarque :** la dimension temps est indispensable dans la majorité des cas.

- Vous allez maintenant analyser les données sources, puis commencer par alimenter d'abord les différentes dimensions. Ce n'est qu'une fois que toutes les dimensions auront été créées que vous passerez à l'étape de la création de la table de faits, qui accueillera l'ensemble des clés de substitution des dimensions ainsi que les mesures.

2. Créez les dimensions

Voici quelques remarques et recommandations à suivre :

- a. Sur Talend, commencez par récupérer d'abord les données sources sous forme de fichier délimité.
- b. Ensuite, créez un Job et placez sur le Designer les deux fichiers Customers et Products. Utilisez les composants tMap, tHDFS(Output/Input), et tAggregateRow.



## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES

---

- c. Alimentez d'abord les différentes dimensions, puis la table de faits. Concernant la dimension temps, agrégez la date en jour, mois et année pour une analyse plus significative. Vous avez à votre disposition une centaine de fonctions disponibles dans le composant tMap.
  - d. Assurez-vous de vérifier l'existence des données dans le Datalake dans HDFS après l'exécution du Job.
- 
- Si vous exécutez plusieurs fois le job, vous risquez sans doute de rencontrer une erreur ! Pour résoudre ce problème, vous avez deux options :
    - ✓ Soit vous renommez à chaque fois le nom du fichier de destination.
    - ✓ Soit, dans la partie "Action", vous modifiez le paramètre par défaut qui est "Create". Vous pouvez le changer soit en "Ecraser" pour remplacer l'ancien fichier, soit en "Ecrire après" si vous souhaitez compléter l'ancien fichier avec de nouvelles données.

Nom de fichier: "/user/cloudera/Clients.txt"

Type de fichier: Fichier texte

Action: Ecraser

Séparateur de: La variable associée à ce paramètre est : \_FILE\_ACTION\_ context.ClusterHadoop\_HDFS\_Hdfs

☐ Encodage

☐ Compression

☐ Compresser les données

☒ Inclure l'en-tête

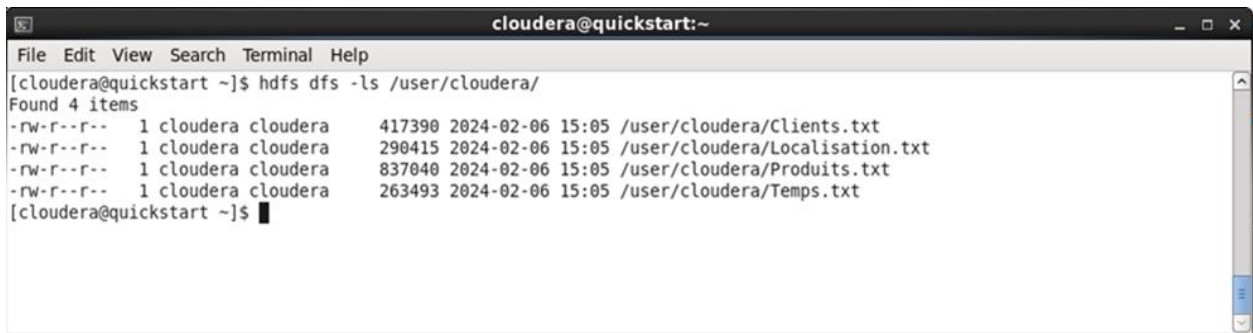
3. Que remarquerez-vous en analysant le fichier « Sales » ? Y a-t-il des incohérences ?
4. Dans la VM Cloudera, vérifiez les 4 fichiers dans HDFS

Nous pouvons consulter maintenant l'existence des 4 fichiers dans la VM Cloudera dans HDFS :

Soit via la commande : « dhfs dfs -ls /user/cloudera/ »

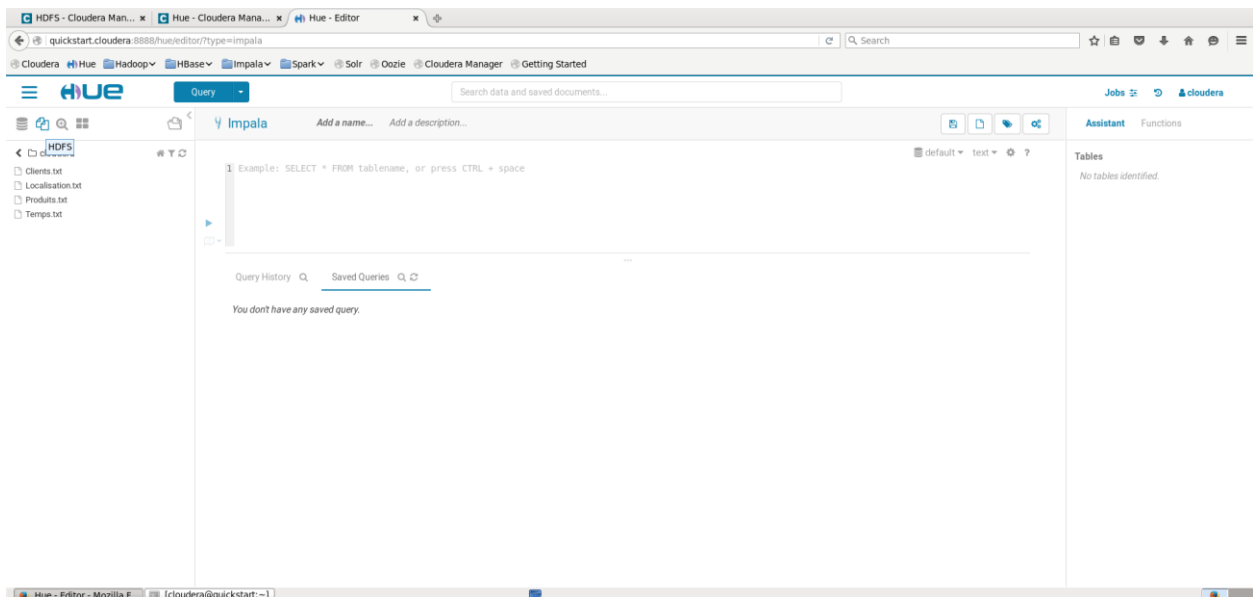
## BLOC : BIG DATA

### WORKSHOP : INTEGRATION DE DONNEES



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/  
Found 4 items  
-rw-r--r-- 1 cloudera cloudera 417390 2024-02-06 15:05 /user/cloudera/Clients.txt  
-rw-r--r-- 1 cloudera cloudera 290415 2024-02-06 15:05 /user/cloudera/Localisation.txt  
-rw-r--r-- 1 cloudera cloudera 837040 2024-02-06 15:05 /user/cloudera/Produits.txt  
-rw-r--r-- 1 cloudera cloudera 263493 2024-02-06 15:05 /user/cloudera/Temps.txt  
[cloudera@quickstart ~]$
```

Ou bien, via l'interface « HUE » :



5. Passez maintenant à la création de la table des faits à travers des différentes dimensions?
6. Dans la VM Cloudera, vérifiez la table de faits dans HDFS



# BLOC : BIG DATA

WORKSHOP : INTEGRATION DE DONNEES

---