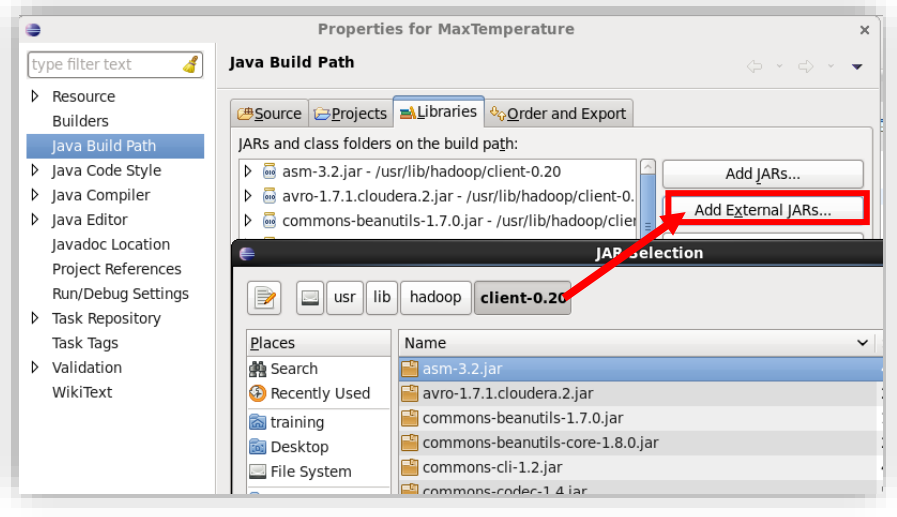
Atelier 2 : MapReduce

Initiation :

Les étapes de création d'un projet MapReduce sous Eclipse sont :

1. Créer un projet Java
2. Ajouter les fichiers jar existants dans le répertoire **/usr/lib/hadoop/client-0.20** au classpath du projet.



3. Créer les trois classes suivantes :
   a. Mapper
   b. Reducer
   c. Fonction pour exécuter le job
4. Exécuter le projet MapReduce.

Exemple :

Pour ce premier exercice, notre but est de chercher la température maximale par année, pour un fichier log dont les champs sont de la forme suivante :

```
0043012650999991949032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9+00781+9999
```

L'année est définie par les caractères dont la position est entre 15 et 19

La température est définie par les caractères dont la position est entre 88 et 92

2.1 Mapper

Le Mapper doit :

- Séparer les différents champs
- Extraire les éléments voulus à partir de ces champs, sous forme de clef/valeur

Pour calculer la température par année, le couple (clef, valeur) à extraire est (année, température). Pour faire cela, le code du **Mapper** se trouve sous le répertoire **/atelier2/code/MaxTemperature** dans le fichier MaxTemperatureMapper.java

2.2 Reducer

Dans l'exemple précédent, une fois que le Mapper extrait les couples (année, température), le **Reducer** aura comme tâche de calculer la température maximale par année. :

Le code du **Reducer** se trouve sous le répertoire **/atelier2/code/MaxTemperature** dans le fichier MaxTemperatureReducer.java

2.3 Créer le job driver de MapReduce

- Le rôle du job driver est de définir le fichier jar qui va contenir le driver, le mapper et le reducer.
- Démarrer le job mapreduce

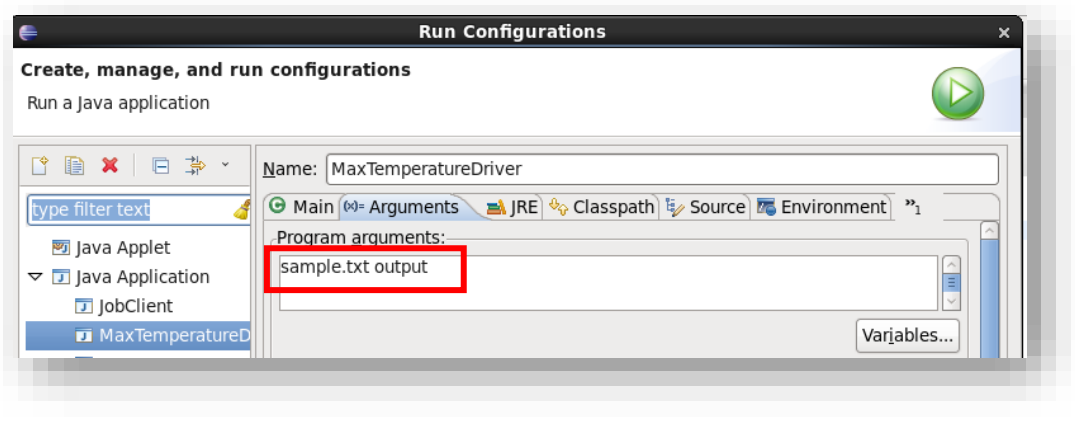
Le code du **Driver** se trouve sous le répertoire **/atelier2/code/MaxTemperature** dans le fichier MaxTemperatureDriver.java

2.4 Lancer un Job entier

Le code MapReduce peut être exécuté dans un cluster ou bien dans un IDE comme éclipse. La différence entre les deux modes d'exécution est que, dans le cas d'éclipse, le résultat est généré localement tandis que, pour l'exécution dans un cluster, ils seront générés dans HDFS.

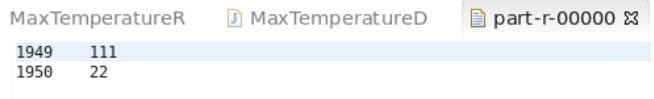
2.4.1 Exécution dans eclipse :

Pour lancer le job mapreduce, il suffit d'exécuter le projet java. Mais avant nous devons spécifier le fichier d'entrer qui sera utilisé durant la phase de Mapping (dans ce cas : sample.txt) et la sortie à générer (Par exemple output)



Le répertoire de sortie **OUTPUT**, après exécution, contiendra un fichier appelé part-00000, représentant la sortie désirée.

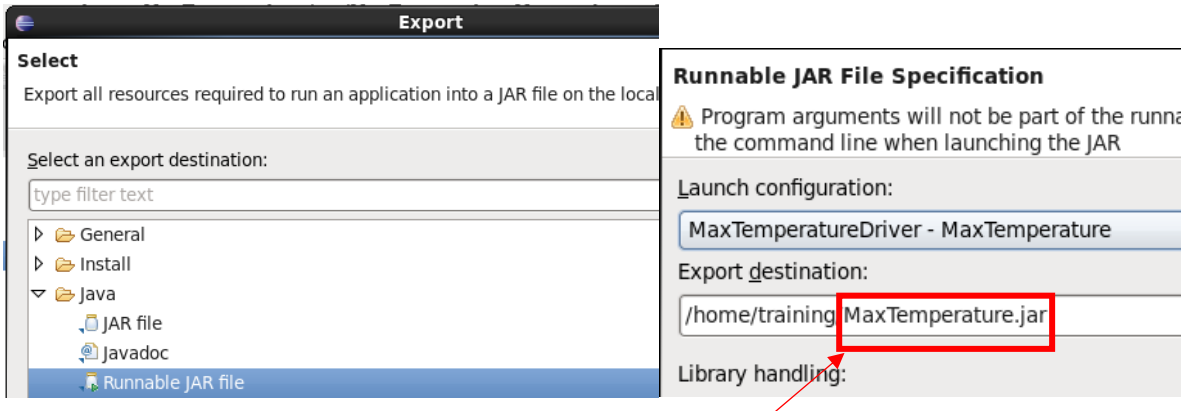
Dans cet exemple, le résultat est le suivant :



2.4.2 Exécution sous HDFS :

Tout d'abord, nous devons créer le fichier jar :

Clic droit sur le nom du projet puis cliqué sur EXPORT.



Lancer un job entier sur Hadoop implique qu'on fera appel au mapper puis au reducer sur une entrée volumineuse, et obtenir à la fin un résultat, directement dans HDFS. Pour faire cela :

Créer un répertoire nommé 'myinput' dans HDFS

hadoop fs -mkdir myinput

Copier le fichier sample.txt sous le répertoire myinput dans HDFS.

hadoop fs -put atelier/sample.txt myinput

Exécuter l'instruction suivante :

hadoop jar MaxTemperature.jar myinput/sample.txt joboutput

Cette instruction donne en paramètres le fichier jar généré, et les répertoires contenant le fichier d'entrée (**myinput**) et la sortie à générer (**joboutput**). Le répertoire de sortie, après exécution, contiendra un fichier appelé part-00000, représentant la sortie désirée.

hadoop fs -ls joboutput

hadoop fs -cat joboutput/part-r-00000

Pour copier le fichier dans le disque local

hadoop fs -get joboutput/part-r-00000 mylocalfile.txt

cat mylocalfile.txt

Remarque : Le répertoire de sortie ne doit pas exister avant l'exécution de l'instruction.

Application :

Le fichier **purchases.txt** stocké sous le répertoire **/Atelier2/data** présente un fichier qui stocke des produits ainsi que leurs couts de vente par magasin et par date.

Une ligne de ce fichier est composée de plusieurs champs séparés par une tabulation (/t)

Elle a la forme suivante :

Date temps magasin produit coût paiement

Ecrire le Mapper et le Reducer permettant de déterminer le total des ventes par magasin.

Remarque : Le couple (clef, valeur) à extraire est (magasin, coût).