




PDI et Pentaho MapReduce

EL-HAFED NAIMA

OUKAJA YOUSSEF MEHDI



PDI (Pentaho Data Integration) : est un logiciel ETL qui permet la conception ainsi que l'exécution des opérations de manipulation et de transformation de données très complexes.



MapReduce : Framework qui permet d'écrire des applications pour faire des traitements big data sur un cluster hadoop. MapReduce permet de manipuler de grandes quantités de données en les distribuant dans un cluster de machines(à des nœuds du cluster) pour pouvoir être traitées en parallèle.



PDI et Pentaho MapReduce : permet d'extraire des données d'un cluster Hadoop, de les transformer et de les transmettre au cluster.



Exemple : Utilisation de MapReduce de Pentaho pour générer un ensemble de données agrégées.

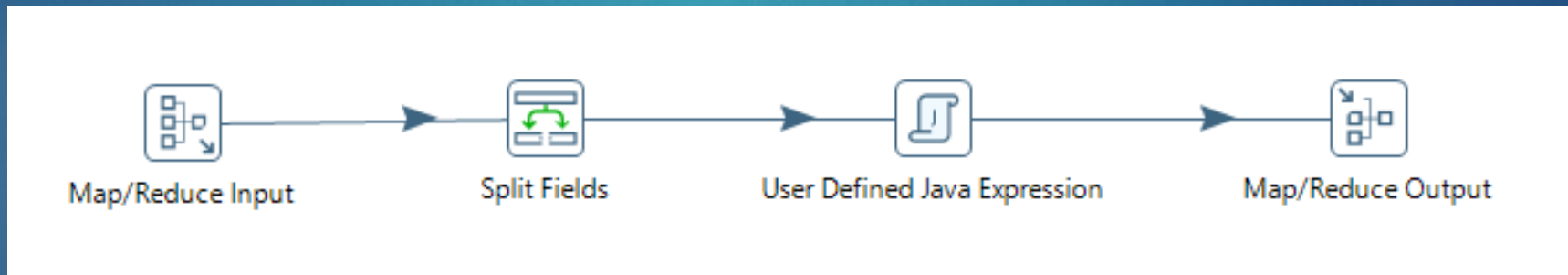
Les étapes de ce guide :

1. Chargement de l'exemple de fichier de données dans HDFS
2. Développer une transformation PDI qui servira de Mappeur.
3. Développer une transformation PDI qui servira de réducteur.
4. Développement d'un travail PDI qui appellera une étape Pentaho MapReduce qui exécute MapReduce à l'aide de la transformation mappeur et réducteur développée.
5. Exécution et révision de la sortie.

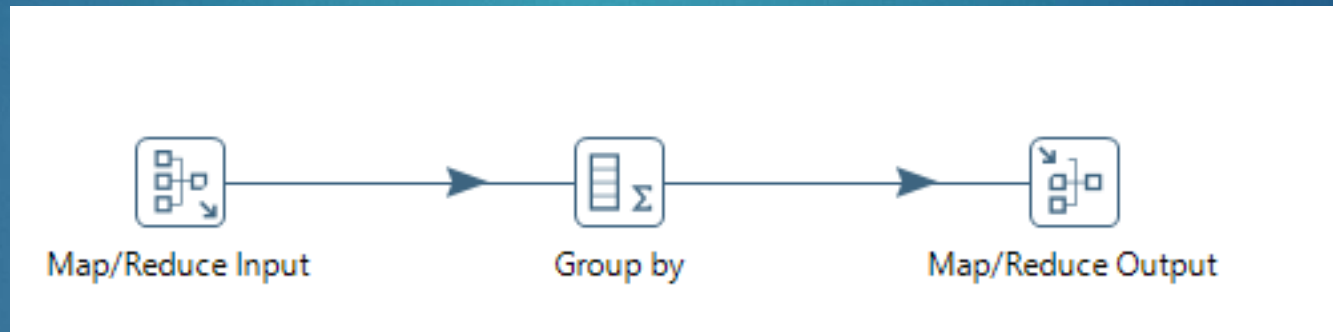
1. Chargement de fichier de données dans HDFS:

- `hadoop fs -mkdir -p /user/pdi/weblogs/parse`
- `hadoop fs -put weblogs_parse.txt
/user/pdi/weblogs/parse/`

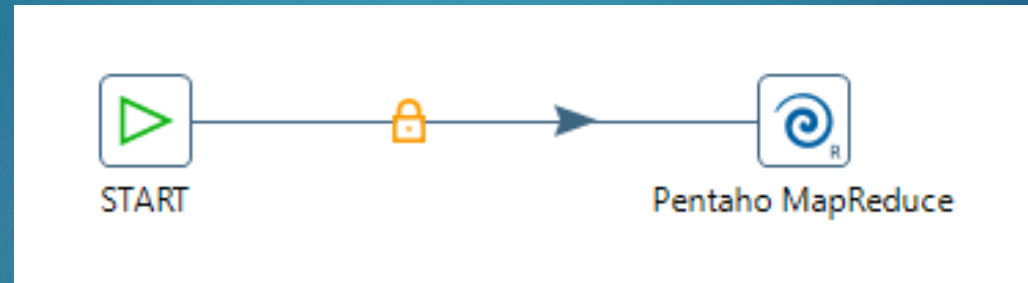
2. Développer une transformation PDI qui servira de Mappeur :



3. Développer une transformation PDI qui servira de réducteur :



4. étape Pentaho MapReduce :



5. Exécution et révision de la sortie :

Execution results				
History Logging Job metrics				
Job / Job Entry	Comment	Result	Reason	Files
▼ aggregate_mr				
Job: aggregate_mr	Start of job execution		start	
START	Start of job execution		start	
START	Job execution finished	Success		
Delete folders	Start of job execution		Followed unconditional link	
Delete folders	Job execution finished	Success		
Pentaho MapReduce	Start of job execution		Followed unconditional link	
Pentaho MapReduce	Job execution finished	Success		
Job: aggregate_mr	Job execution finished	Success	finished	



Exécutez la commande suivante pour afficher les résultats agrégés:

```
hadoop fs -cat  
/user/pdi/weblogs/aggregate_mr  
/part-00000 | head
```