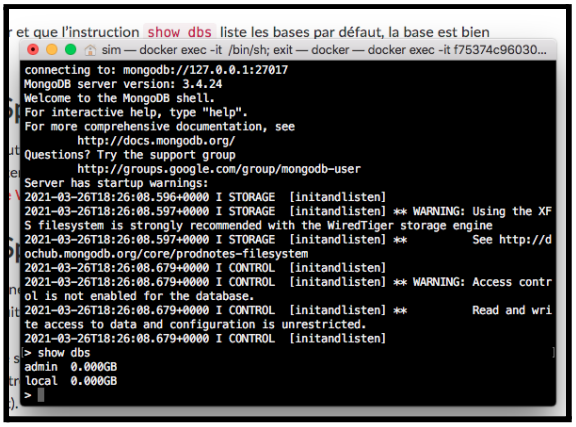
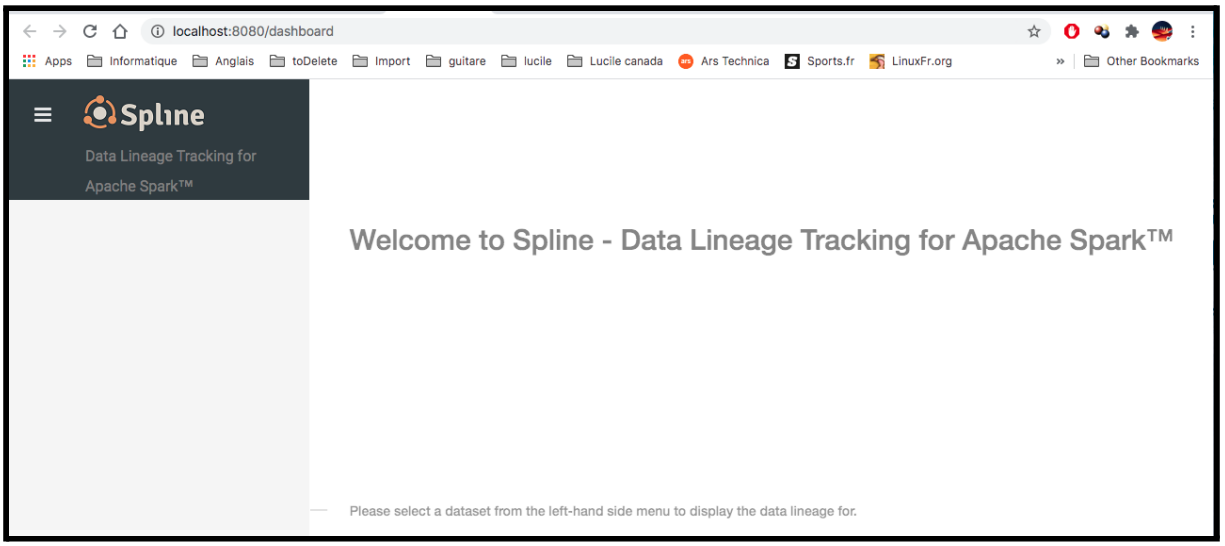
TD Data Lineage Partie 2: SPLINE

CASANOVA S., RAMOS Y.

1. Installer Docker Desktop (Windows) via ce lien: https://hub.docker.com/editions/community/docker-ce-desktop-windows .



2. Checker vos versions Java : **OK**
3. Installer Maven Apache via ce lien https://maven.apache.org/download.cgi (binary zip archive) : **OK**
4. Lancer les commandes suivantes pour l'intégration du serveur Spline dans Docker **OK**
5. Checker sur Docker si le conteneur spline est bien créé. Préciser les composants présents dans le conteneur. **OK**
6. Vous pouvez accéder à Spline Services avec ces URLs:
- Spline Web UI: http://localhost:8080
- Spline Server: http://localhost:9090



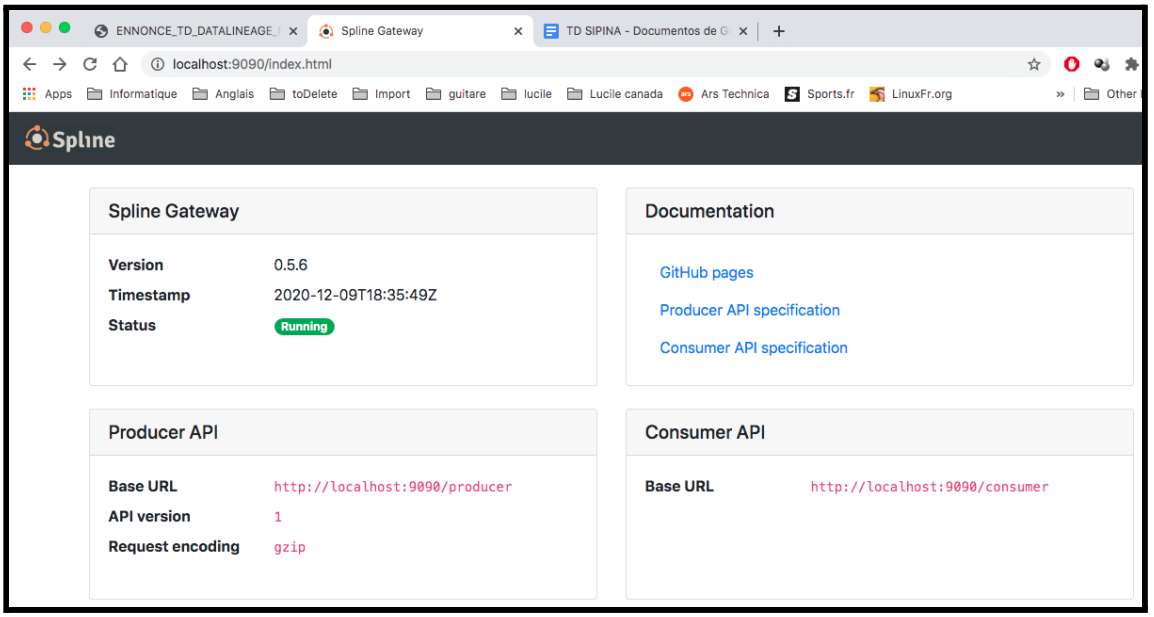
Traçage du lineage avec Spline

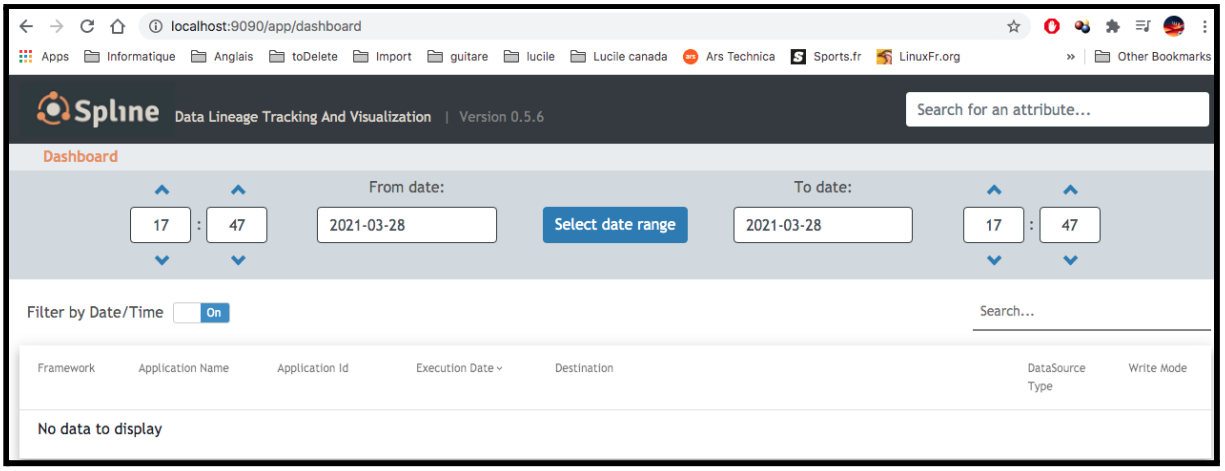
1. Cloner le code source Spline de github avec git : **OK**

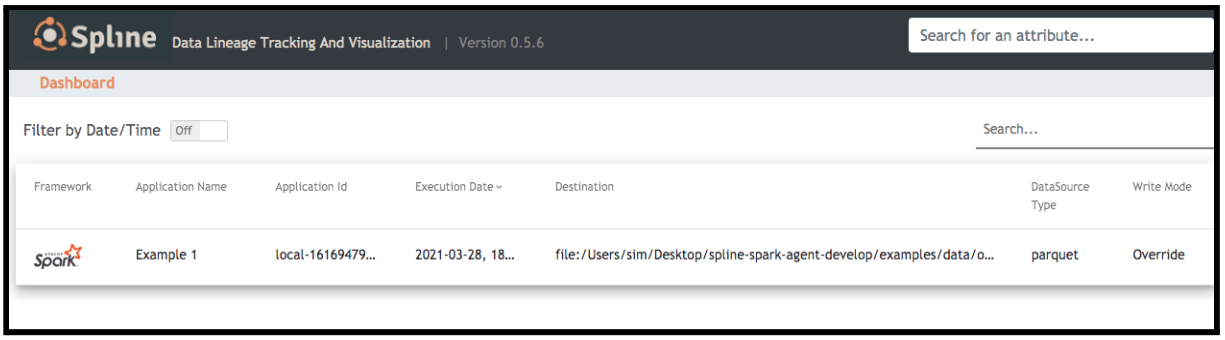
2. Lancer le code Example1Job.java suivant cette commande:
mvn test -P examples -D exampleClass=za.co.absa.spline.example.batch.JavaExampleJob

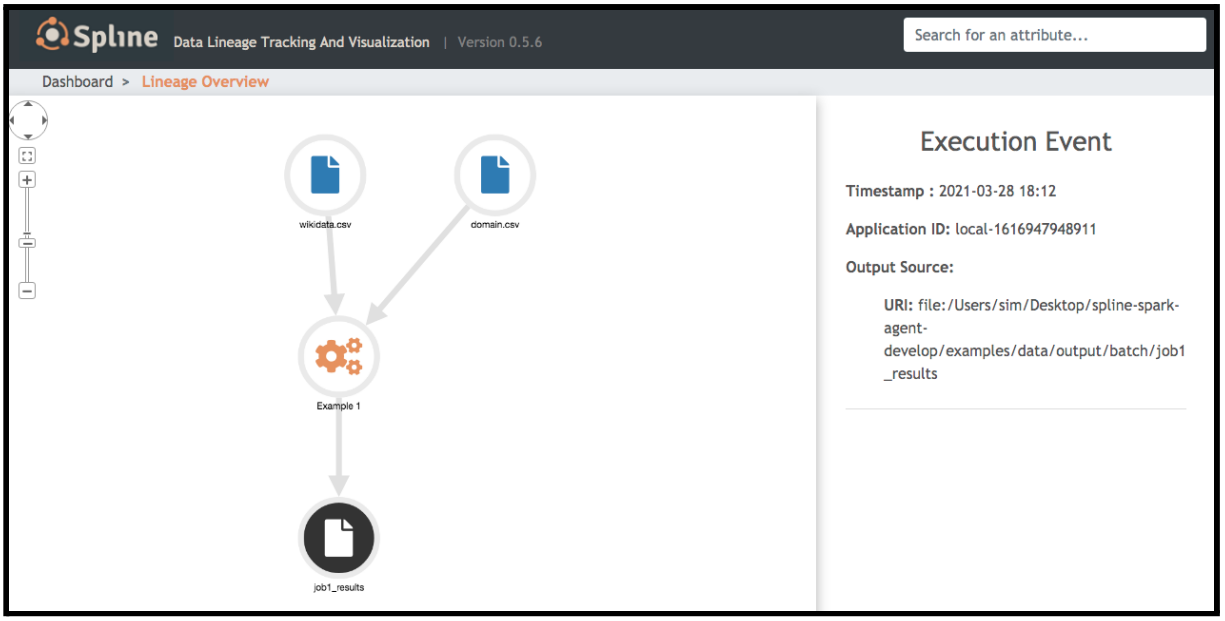
```
examples — sim@sims-MacBook-Pro — ..elop/examples — -zsh — 149×37
main:

main:

runClass:
     [echo] Running za.co.absa.spline.example.batch.Example1Job
[echoproperties] #Ant properties
[echoproperties] #Sun Mar 28 18:12:26 CEST 2021
[echoproperties] http.nonProxyHosts=local|*.local|169.254/16|*.169.254/16
[echoproperties] spark.version=2.4.2
[echoproperties] spline.mode=BEST_EFFORT
[echoproperties] spline.producer.url=http\://localhost\:8080/producer
     [java] Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
     [java] 21/03/28 18:12:27 INFO SparkContext: Running Spark version 2.4.2
     [java] 21/03/28 18:12:27 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appl
cable
     [java] 21/03/28 18:12:27 INFO SparkContext: Submitted application: Example 1
     [java] 21/03/28 18:12:27 INFO SecurityManager: Changing view acls to: sim
     [java] 21/03/28 18:12:27 INFO SecurityManager: Changing modify acls to: sim
     [java] 21/03/28 18:12:27 INFO SecurityManager: Changing view acls groups to:
     [java] 21/03/28 18:12:27 INFO SecurityManager: Changing modify acls groups to:
     [java] 21/03/28 18:12:27 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(si
); groups with view permissions: Set(); users  with modify permissions: Set(sim); groups with modify permissions: Set()
     [java] 21/03/28 18:12:28 INFO Utils: Successfully started service 'sparkDriver' on port 50032.
     [java] 21/03/28 18:12:28 INFO SparkEnv: Registering MapOutputTracker
     [java] 21/03/28 18:12:28 INFO SparkEnv: Registering BlockManagerMaster
     [java] 21/03/28 18:12:28 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
     [java] 21/03/28 18:12:28 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
     [java] 21/03/28 18:12:28 INFO DiskBlockManager: Created local directory at /private/var/folders/tg/glt8xxn94z554wqtpm_170s80000gn/T/blockmgr-b7
6f6a8-50b8-4d32-b785-9612497b814c
     [java] 21/03/28 18:12:28 INFO MemoryStore: MemoryStore started with capacity 2004.6 MB
     [java] 21/03/28 18:12:28 INFO SparkEnv: Registering OutputCommitCoordinator
     [java] 21/03/28 18:12:28 INFO Utils: Successfully started service 'SparkUI' on port 4040.
     [java] 21/03/28 18:12:28 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.2.8:4040
     [java] 21/03/28 18:12:28 INFO Executor: Starting executor ID driver on host localhost
     [java] 21/03/28 18:12:29 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 50033.
     [java] 21/03/28 18:12:29 INFO NettyBlockTransferService: Server created on 192.168.2.8:50033
```

3. Vérifier si le lineage est bien tracé ou pas (en consultant http://localhost:9090)



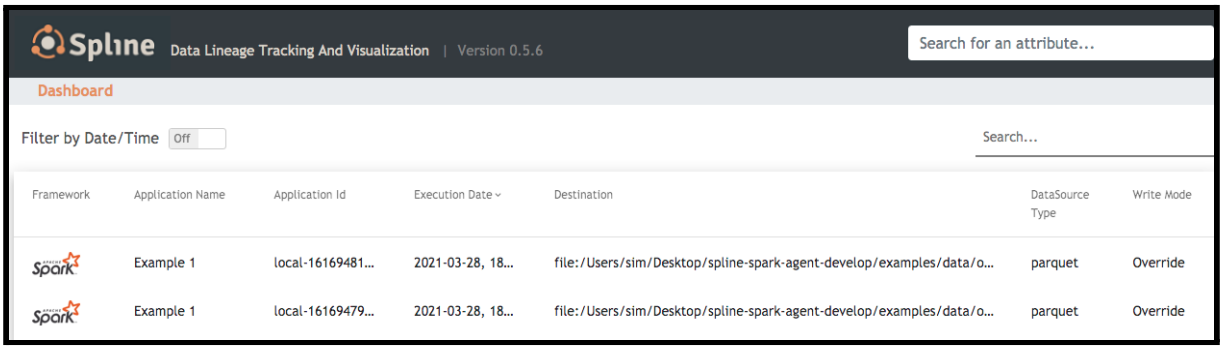

localhost:9090/app/dashboard
Apps
Informatique
Anglais
toDelete
Import
guitare
lucile
Lucile canada
Ars Technica
Sports.fr
LinuxFr.org
Other Bookmarks
Spline
Data Lineage Tracking And Visualization
Version 0.5.6
Search for an attribute...
Dashboard
From date:
To date:
17
47
2021-03-28
Select date range
2021-03-28
17
47
Filter by Date/Time
On
Search...
Framework
Application Name
Application Id
Execution Date
Destination
DataSource Type
Write Mode
No data to display

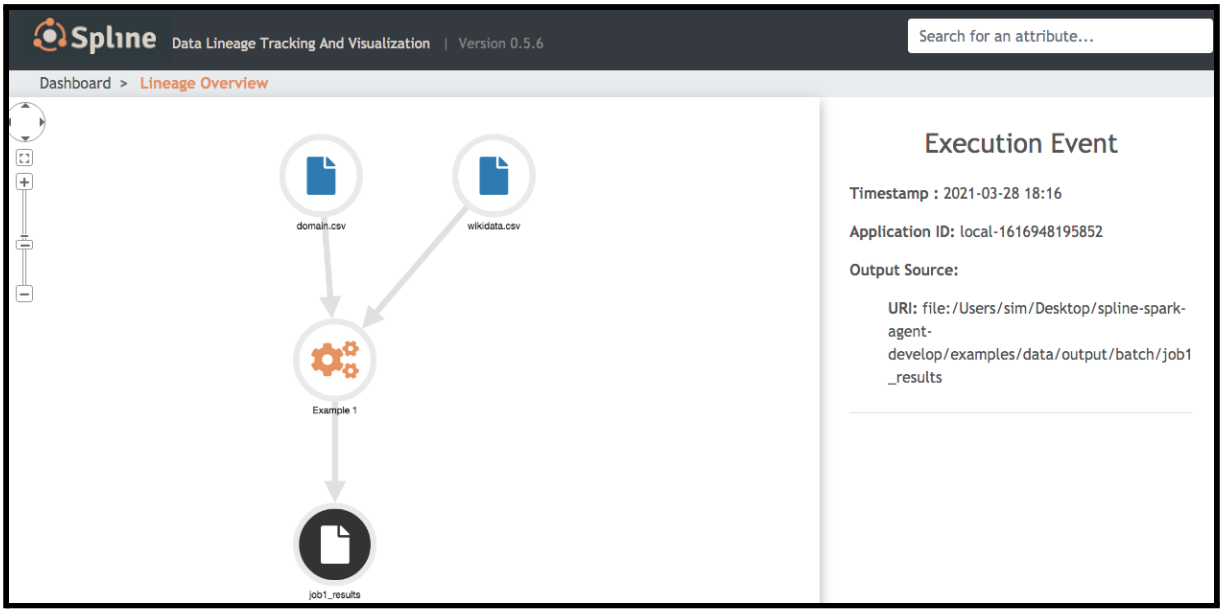

Spline
Data Lineage Tracking And Visualization
Version 0.5.6
Search for an attribute...
Dashboard
Filter by Date/Time
Off
Search...
Framework
Application Name
Application Id
Execution Date
Destination
DataSource Type
Write Mode
Spark
Example 1
local-16169479...
2021-03-28, 18...
file:/Users/sim/Desktop/spline-spark-agent-develop/examples/data/o...
parquet
Override


Spline
Data Lineage Tracking And Visualization
Version 0.5.6
Search for an attribute...
Dashboard > Lineage Overview
wikidata.csv
domain.csv
Example 1
job1_results
Execution Event
Timestamp : 2021-03-28 18:12
Application ID: local-1616947948911
Output Source:
URI: file:/Users/sim/Desktop/spline-spark-agent-develop/examples/data/output/batch/job1_results

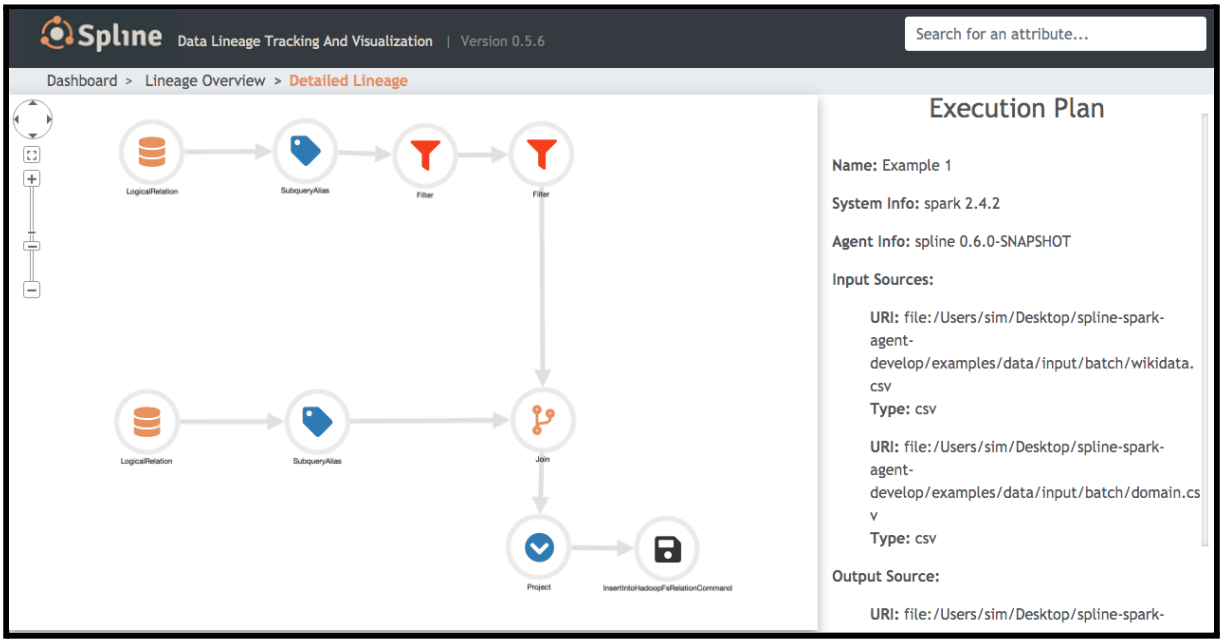
4. Ajoutez à la commande précédente: -D spline.mode=REQUIRED et vérifier encore une fois.

```
examples — sim@sims-MacBook-Pro — ..elop/examples — -zsh — 149×37
main:

main:

runClass:
     [echo] Running za.co.absa.spline.example.batch.Example1Job
[echoproperties] #Ant properties
[echoproperties] #Sun Mar 28 18:16:33 CEST 2021
[echoproperties] http.nonProxyHosts=local|*.local|169.254/16|*.169.254/16
[echoproperties] spark.version=2.4.2
[echoproperties] spline.mode=REQUIRED
[echoproperties] spline.producer.url=http\://localhost\:8080/producer
     [java] Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
     [java] 21/03/28 18:16:34 INFO SparkContext: Running Spark version 2.4.2
     [java] 21/03/28 18:16:34 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appli
cable
     [java] 21/03/28 18:16:34 INFO SparkContext: Submitted application: Example 1
     [java] 21/03/28 18:16:34 INFO SecurityManager: Changing view acls to: sim
     [java] 21/03/28 18:16:34 INFO SecurityManager: Changing modify acls to: sim
     [java] 21/03/28 18:16:34 INFO SecurityManager: Changing view acls groups to:
     [java] 21/03/28 18:16:34 INFO SecurityManager: Changing modify acls groups to:
     [java] 21/03/28 18:16:34 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(sim
); groups with view permissions: Set(); users  with modify permissions: Set(sim); groups with modify permissions: Set()
     [java] 21/03/28 18:16:35 INFO Utils: Successfully started service 'sparkDriver' on port 50048.
     [java] 21/03/28 18:16:35 INFO SparkEnv: Registering MapOutputTracker
     [java] 21/03/28 18:16:35 INFO SparkEnv: Registering BlockManagerMaster
     [java] 21/03/28 18:16:35 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
     [java] 21/03/28 18:16:35 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
     [java] 21/03/28 18:16:35 INFO DiskBlockManager: Created local directory at /private/var/folders/tg/glt8xxn94z554wqtpm_170s80000gn/T/blockmgr-ad8
fc391-b721-490a-a386-182771813c3e
     [java] 21/03/28 18:16:35 INFO MemoryStore: MemoryStore started with capacity 2004.6 MB
     [java] 21/03/28 18:16:35 INFO SparkEnv: Registering OutputCommitCoordinator
     [java] 21/03/28 18:16:35 INFO Utils: Successfully started service 'SparkUI' on port 4040.
     [java] 21/03/28 18:16:35 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.2.8:4040
     [java] 21/03/28 18:16:35 INFO Executor: Starting executor ID driver on host localhost
     [java] 21/03/28 18:16:35 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 50049.
     [java] 21/03/28 18:16:35 INFO NettyBlockTransferService: Server created on 192.168.2.8:50049
```





5. Utiliser dans l'interface Spline UI le lineage détaillé et décrivez les différentes transformations ainsi que les sources et les outputs.



Transformations :

En regardant la data lineage généré par spline, nous trouvons :

- Une première branche qui correspond aux 3 actions. Ces actions sont :
  - Lecture ("wiki.csv")
  - filtrage (total_response_total>1000),
  - filtrage (counts_views>10)
- Une deuxième branche qui correspond à 1 action de lecture ("domains.csv")
- Une jointure de type left_outer entre les 2 premieres branches
- Une projection sur : page_title, domain et count_views, et finalement,
- Ecriture de résultats

Tous les actions de data lineage correspondent au code exécuté :

```
object Example1Job extends SparkApp("Example 1") {

  // Initializing library to hook up to Apache Spark
  spark.enableLineageTracking()

  // A business logic of a spark job ...

  val sourceDS = spark.read
    .option("header", "true")
    .option("inferSchema", "true")
    .csv("data/input/batch/wikidata.csv")
    .as("source")
    .filter($"total_response_size" > 1000)
    .filter($"count_views" > 10)

  val domainMappingDS = spark.read
    .option("header", "true")
    .option("inferSchema", "true")
    .csv("data/input/batch/domain.csv")
    .as("mapping")

  val joinedDS = sourceDS
    .join(domainMappingDS, $"domain_code" === $"d_code", "left_outer")
    .select($"page_title".as("page"), $"d_name".as("domain"), $"count_views")

  joinedDS.write.mode(SaveMode.Overwrite).parquet("data/output/batch/job1_results")
}
```

Sources :
URI: file:/Users/sim/Desktop/spline-spark-agent-develop/examples/data/input/batch/wikidata.csv
Type: csv
URI: file:/Users/sim/Desktop/spline-spark-agent-develop/examples/data/input/batch/domain.csv
Type: csv

Outputs:
URI: file:/Users/sim/Desktop/spline-spark-agent-develop/examples/data/output/batch/job1_results
Type: parquet