# Project – Course 3

## Objective

In this project, you will exercise your knowledge of HDFS from the course. The project is also a practice of ETL pipeline generalization. In this project, you will achieve:

- ✓ Configuring Hadoop API
- ✓ Work with Hadoop HDFS API
- ✓ Understand how to change a pipeline to work with multiple data sources

## Before you start

### Data set

This time we use STM GTFS data set. In this project, we continue the practice of enrichment (the most common data transformation task to do). To download the dataset, visit http://www.stm.info/en/about/developers

You have already analyzed the structure of the data. Note the path you have saved the dataset for further use in the application.

## Project requirements

### Preparation

Put the files of **trip**, **route** and **calendar** on HDFS under **/user/[group name]/[your name]/stm/** where **[group name]** is your class group for example *summer2019* and **[your name]** is a name of your choice all in lowercase without space or any non-alphabetic character.

### Enricher

Use your code base of the project of the course 2 (create a new project called **course3project**) to enrich **route** with **trip** based on *route_id* and then enrich the result with **Calendar** based on *service_id*. At the end, you should write the final result in a CSV file with a proper header.

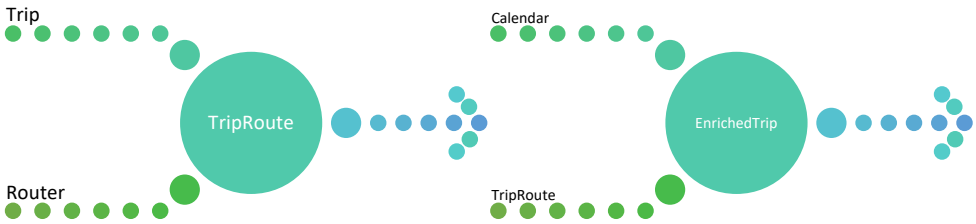
Apply required changes to read files from HDFS and write the result back to the HDFS. The followings should be applied:

- You may read all the files in memory once and do the join as these files are small. It is encouraged to use streaming as a bonus.
- The destination directory on HDFS is **/user/[group name]/[your name]/course3/**
- Delete the destination directory and its content if exists at the beginning of the project
- There should not be any manual work needed and every time you run your project, you get the same result
- You could write intermediate data back on HDFS for debugging but you shouldn't keep them. At the end of the project, there is only file in the destination directory, and it is the final result.

## Reminder from the course 2 project

Parse **trips** data and enrich it with **routes** and **calendar** information. Note that here we create a new object that has information of trip, routes and calendar. It's possible to do both joins at the same time. But, for simplicity, do one at a time. Hence, you will have an intermediate class as well.

a. Trip
b. Calendar
c. Route
d. TripRoute (*intermediate class*)
e. EnrichedTrip (*final class*)




For the **trip/route** join use Map Join and for the **calendar** use nested loop join.