# Project

## Objective

The objectives of this project can be categorized as

* Understand the architecture of a MapReduce job
* How to submit a job on YARN and monitor the progress
* How to interact with HDFS

## Before you start

### Data set

A sample data of three tables are posted in Slack channel.

### Two-way join code in MapReduce

The heart of this project is the source code of two-way SQL join we coded in MapReduce during the course. You need to modify and re-use it to implement circular join.

## Assignment requirements

Three tables are provided as

* R with fields of A and B
* S with fields of B and C
* T with fields of C and A

You need to implement a MapReduce program with two iterations; 1) join R and S as new table called RS 2) join RS with T as new table called RST.

RST is going to have three fields: A, B, and C.

As part of the project, you are required to present your approach in the class.

# Extra

The following problems would be interesting to execute in MapReduce:

* Run three-way join on STM data on STOPS, STOP\_TIMES and TRIPS
* Implement different SQL operations in MapReduce e.g. WHERE, GROUP BY and etc.