Sprint 1

# Summary

You are a BI developer working for a public transportation organization. Your organization wants to produce some operational reports based on the bus/metro schedule as well as providing analysts with tools to better understand the hidden patterns and anomalies with the bus and metro schedules.

# Requirements

Your architect asked you to analyse the data source in order to better understand the quality and the content of the data. He’s asking you to use whatever tool you want in order to:

* Better understand the data source
* Provide a data quality assessment containing:
  + Candidate Primary Keys for each file
  + Relationships between each file
  + Column statistics (minimum, maximum, average and standard deviation)

Your architect want a report for each file that lists the above profiling elements as well as a SQL Server database that contains a dump of each file provided. The database should have:

* A table per file with all the data loaded.
* Each table should have a primary key, defined thanks to you data analysis
* Primary Key – Foreign Key relationship defined between the tables

Note: for this sprint, there is no need to convert the columns into their actual data types. Storing everything as varchar or nvarchar is sufficient.

# Provided documentation and artifacts

* A zip file, gtfs\_stm.zip, available in the GitHub project
* Documentation about the files, available [here](http://www.stm.info/en/about/developers/available-data-description) and [here](https://developers.google.com/transit/gtfs/reference?hl=en-FR)

# Expected deliverables

* A data quality assessment report, providing, for each table / column:
  + Candidate Primary Keys for each file
  + Relationships between each file
  + Column statistics (minimum, maximum, average and standard deviation)
* Each file loaded into a database, named “STM\_SRC”. The database should have:
  + A table per file
  + Columns should be varchar or nvarchar
  + Each table should have a primary key
  + Tables should have Foreign Key – Primary Key relationship
* A package that automates the loading process
  + The architect should be able to empty the database and reload it at will.

# Advice and hints

* You’re asked to produce results, you can use any means necessary to make it happen
* Use any tool you want for the data quality assessment and the data loading: SQL, Data Profiler, Other tool.
* Read the stm and GTFS documentation very carefully, all the answers are there
* Spend the necessary time to browse the data, you might discover mistakes in the documentation after browsing the data
* Use the best tool for each task
* <http://msdn.microsoft.com/en-us/library/bb895263(v=sql.110).aspx>
* <http://sqlage.blogspot.ca/2013/12/ssis-how-to-unzip-zip-files-in-ssis-how.html>
* <http://taskunzip.codeplex.com/>