Big Data Developer - Project - Sprint 1

Summary

You have joined a start-up company in Montreal that provides public transport services. One of the products is a trip planning application and the other one is consultancy to BIXI. You are responsible to create a data warehouse for the company to achieve the goal.

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
- Relationship between files

Your architect wants a UML diagram to show the relationship between files and **optionally** a dump of one feed of data in a Hive database.

The only types that you can use are those supported by Hive. Also note that files start with an array which should be dropped. That makes the table to have one column and one row which is not correct data model.

Note that the data understanding in these situations has two aspects:

1. Understand the standards or protocols that your source is following
2. Understand how the source actually implemented such a standard and protocol

Provided documentation and artifacts

- Bixi is using General Bikeshare Feed Specification (GBFS) that you can find the documentation on GitHub (https://github.com/NABSA/gbfs/blob/master/gbfs.md)
- To get online feed of Bixi visit https://www.bixi.com/en/open-data (moved to https://bixi.com/en/page-27)

Expected deliverables

- A UML (or ERD) diagram that documents the data model used in GBFS and the relationships between files
- A UML (or ERD) diagram that documents the data model used in BIXI open data and the relationships between files

Optional:

- Download a feed of data (JSON file) and load into a staging table on Hive. Create a database called “[group]_[name]” that contains one table per file.
- Implement a Shell script/Scala program that automates this process
    - Download the feed
    - Drop tables and create them again
    - Load the feed to database

It is encouraged to do the parts in “optional” section as well as reviewing the program courses and projects in the first sprint to speed up delivery of the projects in the coming sprints.