Sprint 2

Summary

In this sprint, you need to use your expertise in Hive/HDFS to create the dimensions to be used as reference data. It is a set of Hive tables that are normally cold and updated not frequently. They are:

- System information
- Station information

Skillset

- HDFS
- Hive
- JSON
- CSV
- Parquet
- Scala
- SQL
- REST API

Requirements

Implement a program to run ETL for system information and station information automatically as

1. Drop table
2. Transform JSON files to CSV
3. Enrich stations information data with system information. Note that the system information has only one record. You can simply use cross join to accomplish this task.

Step 1

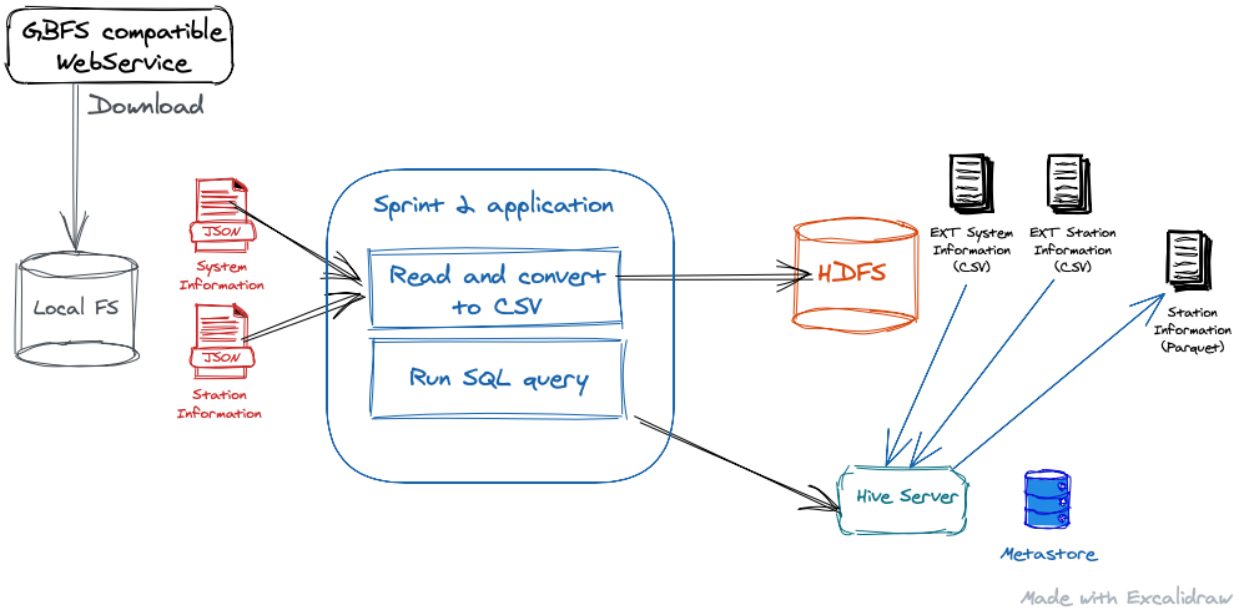
Download files to the staging area which is locate on the local file system.

Step 2

Convert JSON files to CSV and write them on a directory on the HDFS. In this approach, you need to use an appropriate JSON library such as Circe to parse JSON documents into CSV records. You need to create external tables to enabled Hive accessing your data.

Step 3

Use Hive JDBC to execute a cross join between your two external tables and load data into the final table of station_information. Note that this table has the format of Parquet.



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/page-27
- To get to know how to use Google web service to get address information using latitude and longitude information, check out (https://developers.google.com/maps/documentation/geocoding/intro#ReverseGeocoding)

Expected deliverables

- Hive database of your name
- External table of ext_station_information
- External table of ext_system_information
- Managed table of station_information
- The main point of this section is automation. You should be able to run the application and get the same results as the data would not change.

Item	Value
Hive database name	[group_name]_[name]
HDFS base folder for external tables	/user/[group name]/[name]/external/[table name]

Enriched Station Information

Field name	*Type*
system_id	String
timezone	String
station_id	Integer
name	String
short_name	String
lat	Double
lon	Double
capacity	Integer

Bonus

- Download JSON files automatically. To complete the “extract” phase of the application properly, a better way is to download the JSON files from the feeds using a web client. It could be achieved in two steps:
    - Download the main feed which is a JSON document and has links to the other feeds
    - For each feed in the main document, download the data

  You should find an HTTP client library of the language you have chosen to implement the project.