Delta Project Documentation

Spark version 3.3.2
Github link

File Structure

- delta_project
  - FileStore
  - modules
    - create_delta.py
    - spark.py
  - myfuctions.py
  - test_myfunctions.py
  - test_driver.py
  - fill_delta.py
  - Requirements.txt

File Description

- **FileStore –** Location where delta tables will be stored while running in local.
- **modules -** Contains reusable code
  - **create_delta.py -** Contains create_delta(*List, Schema, Path, Mode*)
    - *List* - List of tuples containing data.
    - *Schema* – Schema of data to write.
    - *Path* – Path where delta table will be stored in local, in case of databricks instead of path table name can be provided as well.
    - *Mode* - Mode to write data i.e. Overwrite, append etc.
    - **spark.py -** Creates spark session with delta options configured in local. Not needed while running on databricks as it already provides a spark session with delta enabled.

- **myfunctions.py -** Contains functions for testing purposes.
- **test_myfunctions.py -** Contains testing scenarios and uses modules.myfunctions.py.
- **test_driver.py -** Run this file to start testing using pytest library.
- **fill_delta.py -** This file imports sample housing data from sckit learn library and stored in delta table using modules.create_delta and modules.spark.
- **requirements.txt** - List of libraries required to run this project.

Steps to run locally

- Make sure python, spark and java are installed correctly
- Clone the repo to local.
- Open cmd and change directory to project folder
  cd <path to project directory>
- Run *pip install requirements.txt* to install required libraries.
- Run the following command to create and load data into delta table.
  python fill_delta.py OR *python3 fill_delta.py*
- The delta table will stored in the project folder under Filestore/housing_data.
- Run the following command to run tests. (Total 5 tests present currently)
  p*ython test_driver.py* OR *python3 test_driver.py*

Steps to run on Databricks

- Attach the Github repo to Databricks workspace.
- Create a cluster.
- Attach the cluster in fill_delta notebook.
- In fill_delta add a cell and run the following commands
  %run <path to your repo in databricks>/modules/spark
  %run <path to your repo in databricks>/modules/create_delta
- Comment out lines in create delta notebook as indicated in the notebook if you want to store in delta as a named tabled, the delta table will be stored in default database, otherwise in the filepath provided.
- Now you can run the fill_delta notebook.
- Open test_myfunctions.py and run the following command
  %run <path to your repo in databricks>/myfucntions
- *Open test_driver notebook and run the notebook to run tests.*