**PySpark Exercise**

1. Create Account/Login to [Databricks - Community Version](https://community.cloud.databricks.com/login.html)
2. After logging in the account
   1. Open Databricks
   2. Go to Compute Section on left bar
   3. Click on button ‘Create Cluster’
   4. Rename the cluster as per your choice.
   5. Set ‘Terminate after 30 minutes of inactivity’ in configuration.
   6. Click on button ‘Create Cluster’
3. After creating a cluster
   1. In the left bar on the screen, go to ‘Workspace’ 🡪 ‘Users’ 🡪 Click on dropdown button against your email id.
   2. In dropdown menu, click on ‘Create’ 🡪 ‘Notebook’
   3. Enter Name ‘PySpark\_Excercise\_01’
4. Create Database with name ‘kiva\_crowdfunding’
5. Go to this link, [kiva\_crowdfunding](https://www.kaggle.com/datasets/kiva/data-science-for-good-kiva-crowdfunding/code?datasetId=12414&sortBy=voteCount)
   1. Click on download button.
   2. Extract all (4 files in total) the file from downloaded zip file.
6. Visit Azure databricks portal, go to ‘Data’ 🡪 Click on button ‘Create a Table’
   1. Under ‘Upload File’ section, select file to upload
   2. Click on ‘Create Table with UI’. One shall also explore ‘Create Table in Notebook’
   3. Choose Cluster
   4. Click on ‘Preview Table’
   5. Select ‘kiva\_crowdfunding’ as a database
   6. Check for ‘First row is header’
   7. Check for ‘Infer Schema’
   8. Try changing the schema for each column (from dropdown button in Table Preview Section), if it does not seem fit. FYI, column names can also be changed from here.
   9. Click on ‘Create Table ’ button

Step 6 to be done for all 4 datasets

1. Download the ipynb notebook sent along in the mail.
   1. In the left bar on the screen, go to ‘Workspace’ 🡪 ‘Users’ 🡪 Click on dropdown button against your email id.
   2. Click on ‘Import’ button
   3. Click ‘browse’ 🡪 select ipynb notebook and upload
2. Open the Exercise Notebook
   1. Do as per instructions in the notebook.