

## Question 1 and 2

The screenshot shows a Jupyter Notebook titled "SparkAssignment3" running on a local host. The notebook contains two questions and their corresponding code solutions.

**Question 1:** Import the data from people.json file using JSON python library.

```
In [3]: 1 datafile = "file:///Users/himanideshpande/miniforge3/envs/ENV_SPARK/SparkInstallation/spark-3.2.1-bin-hadoop2.7/example-people.json"
2 input_json = sc.textFile(datafile)
3
```

**Question 2:** Import people.json file using HiveContext or sqlContext into a DataFrame. Print schema information using printSchema() function. Next register the dataframe as a temporary table. Display distinct names from the people.json file by firing a SQL query on the temporary table created. Submit your code and the screenshot of the results.

```
In [5]: 1 hiveCtx = HiveContext(sc)
2 person_json_data = hiveCtx.read.json(datafile)
3
4 person_json_data.printSchema()
5
```

Warning: HiveContext is deprecated in Spark 2.0.0. Please use SparkSession.builder.enableHiveSupport().getOrCreate() instead.

```
root
 |-- age: long (nullable = true)
 |-- name: string (nullable = true)
```

**In [6]:**

```
1 person_json_data.registerTempTable("persons")
2 results = hiveCtx.sql("SELECT distinct(name) FROM persons order by name ASC")
```

Warning: Deprecated in 2.0, use createOrReplaceTempView instead.

## Question 2 and 3

One.IU | All IU Campu x miniforge3/envs/ENV x SparkAssignment3 - x Client Projects: Visu x Assignment 3 x miniforge3/envs/ENV x SparkAssignment3 - x +

localhost:8888/notebooks/miniforge3/envs/ENV\_SPARK/SparkInstallation/spark-3.2.1-bin-hadoop2.7/PYTHON/SparkAssignment3.ipynb#

Jupyter SparkAssignment3 Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
/Users/himanideshpande/miniforge3/envs/ENV_SPARK/lib/python3.8/site-packages/pyspark/sql/dataframe.py:138: FutureWarning: Deprecated in 2.0, use createOrReplaceTempView instead.
warnings.warn(

In [7]: 1 results.show()

+-----+
|   name|
+-----+
|   Andy|
|  Justin|
| Michael|
+-----+

1 Question 3.
2 Import file people.txt located at the same location as people.json into a rdd using built-in CSV library. Import
the CSV library at the beginning and use csv.reader() to load the data. Display the data in your rdd. Submit your
code and screenshot of the results.

In [8]: 1 xtf = "/Users/himanideshpande/miniforge3/envs/ENV_SPARK/SparkInstallation/spark-3.2.1-bin-hadoop2.7/examples/src
2
3

In [10]: 1
2
3 def loadCSVRecord(line):
4
5     input = StringIO(line)
6     # reader = csv.DictReader(input, fieldnames=["name", "age"])
7     reader = csv.reader(input, delimiter=',')
8     return next(reader)
9
10 input_csv = sc.textFile(datatxtfile).map(loadCSVRecord)

In [11]: 1 input_csv.collect()

Out[11]: [['Michael', ' 29'], ['Andy', ' 30'], ['Justin', ' 19']]
```