

ADVANCED DB TAKE AWAY CAT

NAME: EMMANUEL MWANIA

REG NO: SCT221-0410/2021

UNIT: ADVANCED DATABASE MANAGEMENT SYSTEMS

UNIT CODE: ICS2404

COURSE: BIT

YEAT: 3.1

TAKE AWAY CAT:

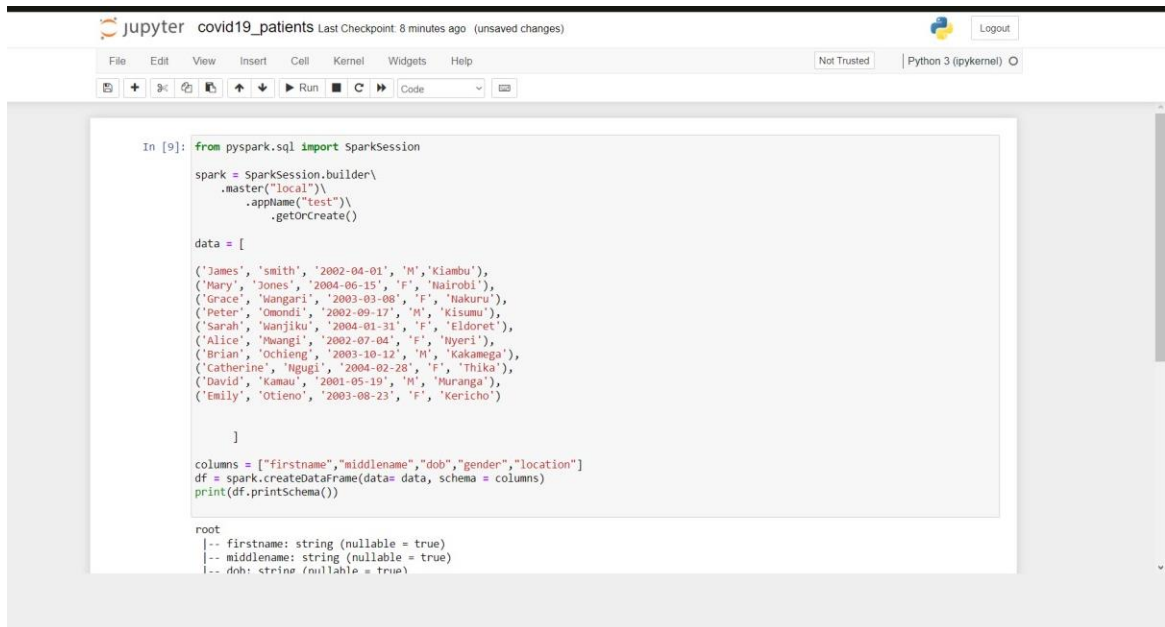
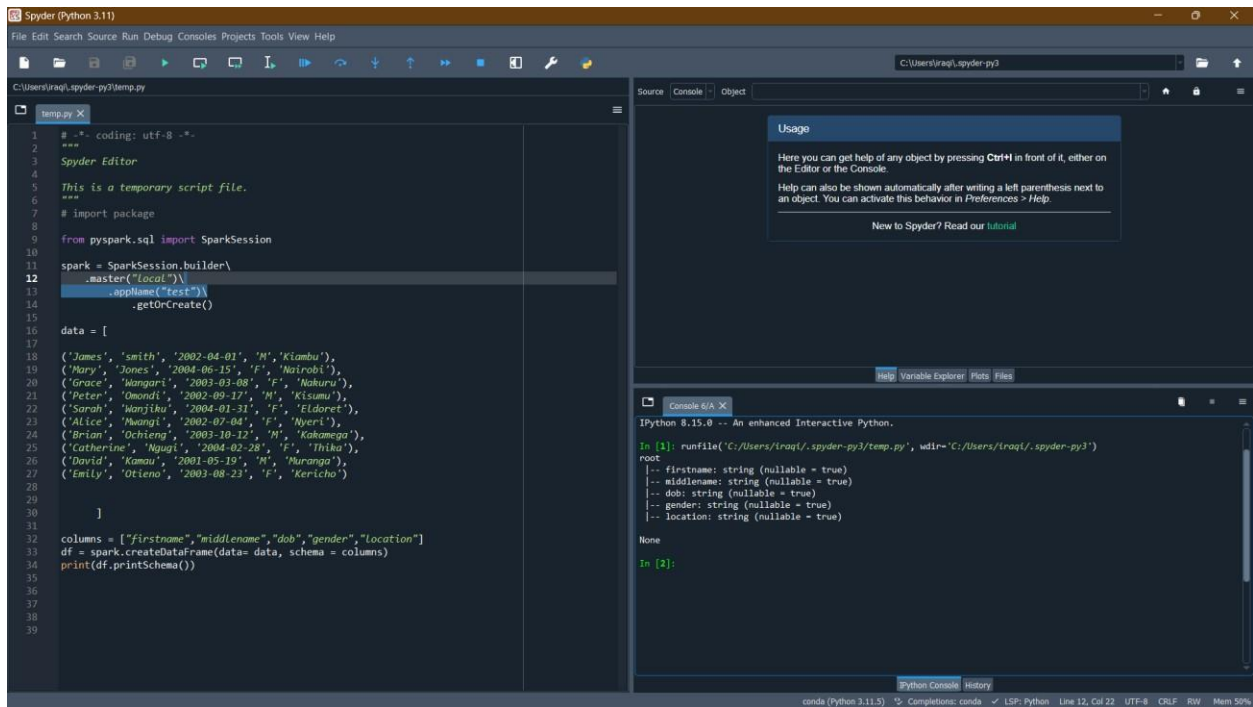
```
1. dianza@haperf101:~ (ssh)
-bash-4.1$ git clone https://@gitlab.cern.ch:8443/db/hadoop-tutorials-2016.git
Initialized empty Git repository in /afs/cern.ch/user/d/dianza/hadoop-tutorials-2016/.git/
remote: Counting objects: 340, done.
remote: Compressing objects: 100% (215/215), done.
remote: Total 340 (delta 172), reused 183 (delta 92)
Receiving objects: 100% (340/340), 1.74 MiB, done.
Resolving deltas: 100% (172/172), done.
-bash-4.1$ ls
cerndb-infra-flume-ng-audit-db      it-puppet-environments             private
cerndb-infra-monitoring-racmon     it-puppet-hostgroup-playground    public
copy-data-from-meetup              jstatd.all.policy                 repo.sh
create-vm-puppet-flume-htutorials.sh map-files                          rpmbuild
create-vm-puppet-kristina-summer-student.sh mapfiles-to-parquet-and-avro      target
create-vm-puppet.sh                nohup.out                         tmp
hadoop-tutorials-2016              os.sh                             tmpaaa
hbase-Hadalytic.ops               prepare-test.sql
-bash-4.1$ cd hadoop-tutorials-2016/
-bash-4.1$ ls
1_sql_and_data_formats  2_data_ingestion  README.md
-bash-4.1$ cd 2_data_ingestion/
-bash-4.1$ l
```

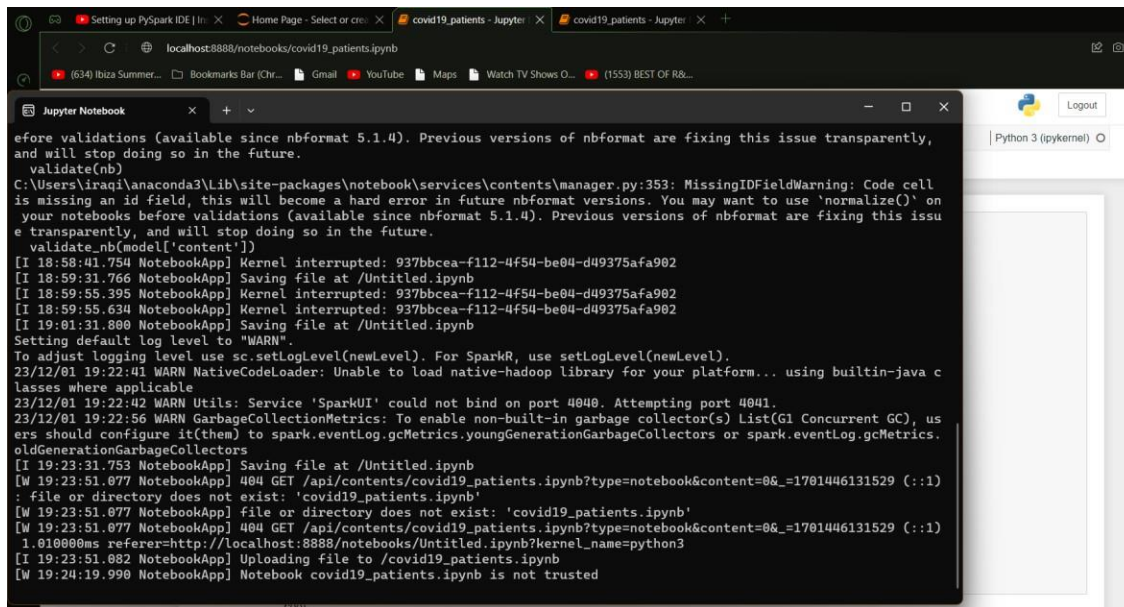
```
1. dianza@haperf101:~ (ssh)
1_flume_chat_gateway  3_meetup_to_kafka      pom.xml
-bash-4.1$ cd 0_batch_ingestion/
-bash-4.1$ ls
kite sqoop
-bash-4.1$ cd kite/
-bash-4.1$ ls
0_get_data  2_create_part_file  4_load_data  6_clean
1_get_schema  3_create_datastore  5_show_data  run_all
-bash-4.1$ ./0_get_data

@# GETTING CSV DATA:
>>>

#source: http://files.grouplens.org/datasets/movielens/ml-latest-small.zip
hdfs dfs -get /tmp/ratings.csv .
head -10 ratings.csv
<<<

userId,movieId,rating,timestamp
1,16,4.0,1217897793000
1,24,1.5,1217895807000
1,32,4.0,1217896246000
1,47,4.0,1217896556000
1,50,4.0,1217896523000
1,110,4.0,1217896150000
1,150,3.0,1217895940000
1,161,4.0,1217897864000
1,165,3.0,1217897135000
-bash-4.1$
```





The screenshot shows a Jupyter Notebook interface with a terminal window open. The terminal displays various logs and warnings, including a 'MissingIDFieldWarning' from nbformat and several 'WARN' messages from Spark. The logs indicate that the notebook is saving files and uploading them to the Jupyter server. The warnings suggest that the notebook is using an older version of nbformat and that the Spark UI service could not bind on port 4040.

```
efore validations (available since nbformat 5.1.4). Previous versions of nbformat are fixing this issue transparently, and will stop doing so in the future.
validate(nb)
C:\Users\iraqi\anaconda3\lib\site-packages\notebook\services\contents\manager.py:353: MissingIDFieldWarning: Code cell is missing an id field, this will become a hard error in future nbformat versions. You may want to use 'normalize()' on your notebooks before validations (available since nbformat 5.1.4). Previous versions of nbformat are fixing this issue transparently, and will stop doing so in the future.
validate(nb[model['content']])
[I 18:58:41.754 NotebookApp] Kernel interrupted: 927bbcea-f112-4f54-be04-d49375afa902
[I 18:59:21.766 NotebookApp] Saving file at /Untitled.ipynb
[I 18:59:55.395 NotebookApp] Kernel interrupted: 927bbcea-f112-4f54-be04-d49375afa902
[I 18:59:55.634 NotebookApp] Kernel interrupted: 927bbcea-f112-4f54-be04-d49375afa902
[I 19:01:31.800 NotebookApp] Saving file at /Untitled.ipynb
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/12/01 19:22:41 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/12/01 19:22:42 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
23/12/01 19:22:56 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users should configure it(them) to spark.eventLog.gcMetrics.youngGenerationGarbageCollectors or spark.eventLog.gcMetrics.oldGenerationGarbageCollectors
[I 19:23:31.753 NotebookApp] Saving file at /Untitled.ipynb
[W 19:23:51.077 NotebookApp] 404 GET /api/contents/covid19_patients.ipynb?type=notebook&content=08_1701446131529 (::1) : file or directory does not exist: 'covid19_patients.ipynb'
[W 19:23:51.077 NotebookApp] file or directory does not exist: 'covid19_patients.ipynb'
[W 19:23:51.077 NotebookApp] 404 GET /api/contents/covid19_patients.ipynb?type=notebook&content=08_1701446131529 (::1) 1.010000ms referer=http://localhost:8888/notebooks/Untitled.ipynb?kernel_name=python3
[I 19:23:51.082 NotebookApp] Uploading file to /covid19_patients.ipynb
[W 19:24:19.990 NotebookApp] Notebook covid19_patients.ipynb is not trusted
```

(iv) Describe pre-processing tasks/techniques used to prepare the data: Handle Missing Data:

You can use PySpark functions like `na.drop()` or `na.fill()` to handle missing values. Data Cleaning and Transformation:

Remove irrelevant columns or rows using PySpark DataFrame operations.

Use functions like `filter()` or `drop()`.

Feature Engineering:

Create new features or modify existing ones based on domain knowledge.

Use PySpark DataFrame transformations.

Scaling/Normalization:

If your algorithm requires it, use PySpark's StandardScaler or MinMaxScaler for feature scaling.

Reasoning:

The choice of pre-processing tasks depends on the characteristics of your data and the requirements of your predictive model.

Handling missing data is crucial to avoid biases in your analysis.

Data cleaning and transformation ensure that the data is in a suitable format for analysis.

Feature engineering enhances the model's ability to capture patterns.

Scaling is essential for algorithms sensitive to the scale of features.