

ec2-54-226-130-109.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb#

jupyterhub Untitled Last Checkpoint: 2 hours ago (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark

```
In [1]: from pyspark.sql import SparkSession

Starting Spark application

ID      YARN Application ID  Kind  State  Spark UI  Driver log  Current session?
0  application_1753689413660_0001  pyspark  idle  Link  Link  ✓

SparkSession available as 'spark'.

In [2]: spark = SparkSession.builder \
        .appName("LoadParquetFromS3") \
        .config("spark.hadoop.fs.s3a.impl", "org.apache.hadoop.fs.s3a.S3AFileSystem") \
        .config("spark.hadoop.fs.s3a.aws.credentials.provider", "com.amazonaws.auth.DefaultAWSCredentialsProviderChain") \
        .getOrCreate()

In [56]: s3_parquet_path = "s3a://final-project-bucket-group-5/master_data/part-00000-403c5616-918e-49ec-a9fe-e3c4f422e78e-c000.csv"

In [61]: df = spark.read.csv(s3_parquet_path, header=True, inferSchema=True)

In [62]: df.printSchema()

root
 |-- id: string (nullable = true)
 |-- case number: string (nullable = true)
 |-- block: string (nullable = true)
 |-- nibrs code: string (nullable = true)
 |-- primary type: string (nullable = true)
 |-- description: string (nullable = true)
 |-- location description: string (nullable = true)
```

ec2-54-226-130-109.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb#

jupyterhub Untitled Last Checkpoint: 2 hours ago (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark

```
-- fbi code: string (nullable = true)
-- x coordinate: string (nullable = true)
-- y coordinate: string (nullable = true)
-- year: string (nullable = true)
-- latitude: string (nullable = true)
-- longitude: string (nullable = true)
-- location: string (nullable = true)
-- weapon description: string (nullable = true)
-- vict age: string (nullable = true)
-- vict sex: string (nullable = true)
-- victim race: string (nullable = true)
-- date occ: string (nullable = true)
-- time occ: string (nullable = true)
-- date rptd: string (nullable = true)
-- time rptd: string (nullable = true)
-- date arrested: string (nullable = true)
-- time arrested: string (nullable = true)
-- premises desc: string (nullable = true)
-- district: string (nullable = true)
-- suspect age: string (nullable = true)
-- suspect sex: string (nullable = true)
-- suspect race: string (nullable = true)
-- case status: string (nullable = true)
-- crime category: string (nullable = true)
-- secondary description: string (nullable = true)
-- census tract: string (nullable = true)
-- zip code: string (nullable = true)
-- incident narrative: string (nullable = true)
-- priority level: string (nullable = true)
-- repeat offense flag: string (nullable = true)
```

ec2-54-226-130-109.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb#

jupyterhub Untitled Last Checkpoint: 2 hours ago (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted PySpark

```
In [64]: df.select("block").show(5)

+-----+
|      block      |
+-----+
|6666 Halsted St|
|1598 State St  |
|924 63rd St    |
|1706 King Dr   |
|4941 95th St   |
+-----+
only showing top 5 rows

In [65]: df.select(df["case_number"].alias("case_number")).show(5)

+-----+
|case_number|
+-----+
|CPD383874|
|CPD826714|
|CPD376812|
|CPD417703|
|CPD524301|
+-----+
only showing top 5 rows

In [47]:

In [49]: print(df.columns)
```

ec2-54-226-130-109.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb#

jupyterhub Untitled Last Checkpoint: 2 hours ago (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted PySpark

```
In [49]: print(df.columns)

['id', 'case_number', 'block', 'nibrs_code', 'primary_type', 'description', 'location_description', 'arrest', 'domestic', 'beat', 'area', 'ward', 'community_area', 'fbi_code', 'x_coordinate', 'y_coordinate', 'year', 'latitude', 'longitude', 'location', 'weapon_description', 'vict_age', 'vict_sex', 'victim_race', 'date_occ', 'time_occ', 'date_rptd', 'time_rptd', 'date_arrested', 'time_arrested', 'premises_desc', 'district', 'suspect_age', 'suspect_sex', 'suspect_race', 'case_status', 'crime_category', 'secondary_description', 'census_tract', 'zip_code', 'incident_narrative', 'priority_level', 'repeat_offense_flag']

In [67]: from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType

In [69]: from pyspark.sql.types import BooleanType

In [71]: from pyspark.sql.types import LongType

In [72]: # Define schema based on your sample
schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("case_number", StringType(), True),
    StructField("block", StringType(), True),
    StructField("nibrs_code", StringType(), True),
    StructField("primary_type", StringType(), True),
    StructField("description", StringType(), True),
    StructField("location_description", StringType(), True),
    StructField("arrest", BooleanType(), True),
    StructField("domestic", BooleanType(), True),
    StructField("beat", StringType(), True),
])
```

```
StructureField("community_area", StringType(), True),
StructureField("fbi_code", StringType(), True),
StructureField("x_coordinate", LongType(), True),
StructureField("y_coordinate", LongType(), True),
StructureField("year", IntegerType(), True),
StructureField("latitude", DoubleType(), True),
StructureField("longitude", DoubleType(), True),
StructureField("location", StringType(), True),
StructureField("weapon_description", StringType(), True),
StructureField("vict_age", IntegerType(), True),
StructureField("vict_sex", StringType(), True),
StructureField("victim_race", StringType(), True),
StructureField("date_occ", StringType(), True),
StructureField("time_occ", StringType(), True),
StructureField("date_rptd", StringType(), True),
StructureField("time_rptd", StringType(), True),
StructureField("date_arrested", StringType(), True),
StructureField("time_arrested", StringType(), True),
StructureField("premises_desc", StringType(), True),
StructureField("district", StringType(), True),
StructureField("suspect_age", IntegerType(), True),
StructureField("suspect_sex", StringType(), True),
StructureField("suspect_race", StringType(), True),
StructureField("case_status", StringType(), True),
StructureField("crime_category", StringType(), True),
StructureField("secondary_description", StringType(), True),
StructureField("census_tract", StringType(), True),
StructureField("zip_code", StringType(), True),
StructureField("incident_narrative", StringType(), True),
StructureField("priority_level", StringType(), True),
StructureField("repeat_offense_flag", BooleanType(), True)
```

```
df = spark.read.format("csv") \
    .option("header", True) \
    .schema(schema) \
    .load(s3_parquet_path)

In [73]: df.show(5)
```

id	case_number	block	nbrs_code	primary_type	description	location_description	arrest	domestic	b
eat	area	ward	community_area	fbi_code	x_coordinate	y_coordinate	year	latitude	longitude
location	weapon_de	script	vict_age	vict_sex	victim_race	date_occ	time_occ	date_rptd	time_rptd
date_arrested	time	premises_desc	district	suspect_age	suspect_sex	suspect_race	case_status	crime_category	secondary_description
census_tract	zip_code	incident_narrative	priority_level	repeat_offense_flag					
100000000	CPD383874	6666 Halsted St	1310	Criminal Damage	Criminal Damage t...	Public Transit	false	true	2
522	West Side	W40	Austin	290	1150092	1899985	2017	41.8699	-87.6385
410505	75	null	White	2017-07-16T00:00:...	2113	2017-07-16T00:00:...	2113	null	n
ull	Parking Lot	D17	59	M	null	Open	Property	Vandalism	
10000001	CPD826714	1598 State St	0810	Theft	Theft	Street	false	true	1

ec2-54-226-130-109.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb#

jupyterhub Untitled Last Checkpoint: 2 hours ago (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark

```
In [77]: from pyspark.sql.functions import to_timestamp
df = df.withColumn("date_occ", to_timestamp("date_occ", "yyyy-MM-dd HH:mm:ss"))

In [78]: df.printSchema()

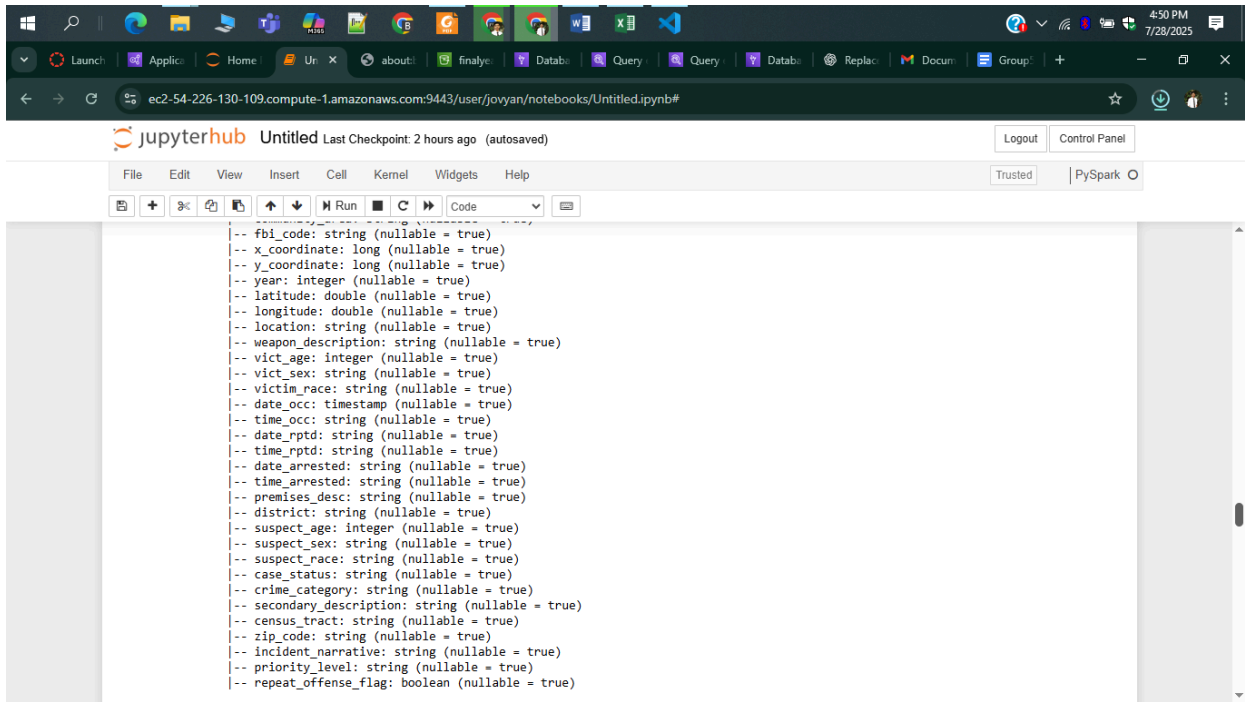
root
 |-- id: integer (nullable = true)
 |-- case_number: string (nullable = true)
 |-- block: string (nullable = true)
 |-- nibrs_code: string (nullable = true)
 |-- primary_type: string (nullable = true)
 |-- description: string (nullable = true)
 |-- location_description: string (nullable = true)
 |-- arrest: boolean (nullable = true)
 |-- domestic: boolean (nullable = true)
 |-- beat: string (nullable = true)
 |-- area: string (nullable = true)
 |-- ward: string (nullable = true)
 |-- community_area: string (nullable = true)
 |-- fbi_code: string (nullable = true)
 |-- x_coordinate: long (nullable = true)
 |-- y_coordinate: long (nullable = true)
 |-- year: integer (nullable = true)
 |-- latitude: double (nullable = true)
 |-- longitude: double (nullable = true)
 |-- location: string (nullable = true)
 |-- weapon_description: string (nullable = true)
 |-- vict_age: integer (nullable = true)
```

ec2-54-226-130-109.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb#

jupyterhub Untitled Last Checkpoint: 2 hours ago (autosaved) Logout Control Panel

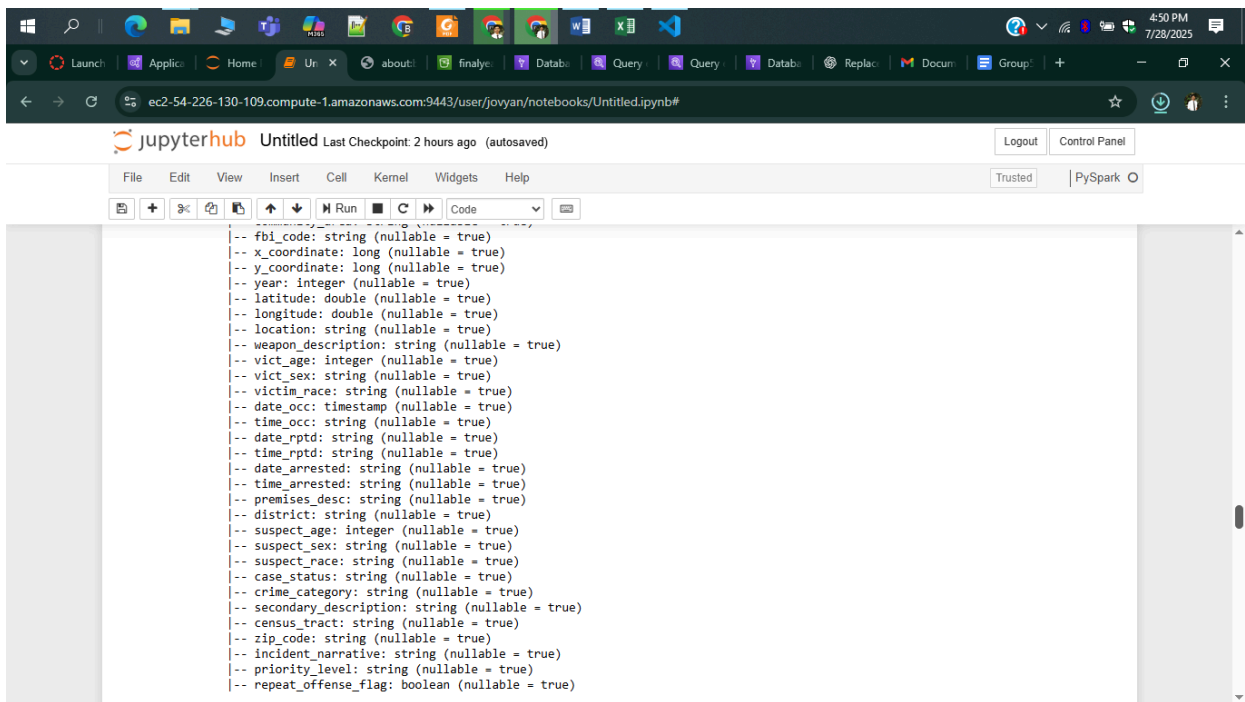
File Edit View Insert Cell Kernel Widgets Help Trusted | PySpark

```
-- fbi_code: string (nullable = true)
-- x_coordinate: long (nullable = true)
-- y_coordinate: long (nullable = true)
-- year: integer (nullable = true)
-- latitude: double (nullable = true)
-- longitude: double (nullable = true)
-- location: string (nullable = true)
-- weapon_description: string (nullable = true)
-- vict_age: integer (nullable = true)
-- vict_sex: string (nullable = true)
-- victim_race: string (nullable = true)
-- date_occ: timestamp (nullable = true)
-- time_occ: string (nullable = true)
-- date_rptd: string (nullable = true)
-- time_rptd: string (nullable = true)
-- date_arrested: string (nullable = true)
-- time_arrested: string (nullable = true)
-- premises_desc: string (nullable = true)
-- district: string (nullable = true)
-- suspect_age: integer (nullable = true)
-- suspect_sex: string (nullable = true)
-- suspect_race: string (nullable = true)
-- case_status: string (nullable = true)
-- crime_category: string (nullable = true)
-- secondary_description: string (nullable = true)
-- census_tract: string (nullable = true)
-- zip_code: string (nullable = true)
-- incident_narrative: string (nullable = true)
-- priority_level: string (nullable = true)
-- repeat_offense_flag: boolean (nullable = true)
```



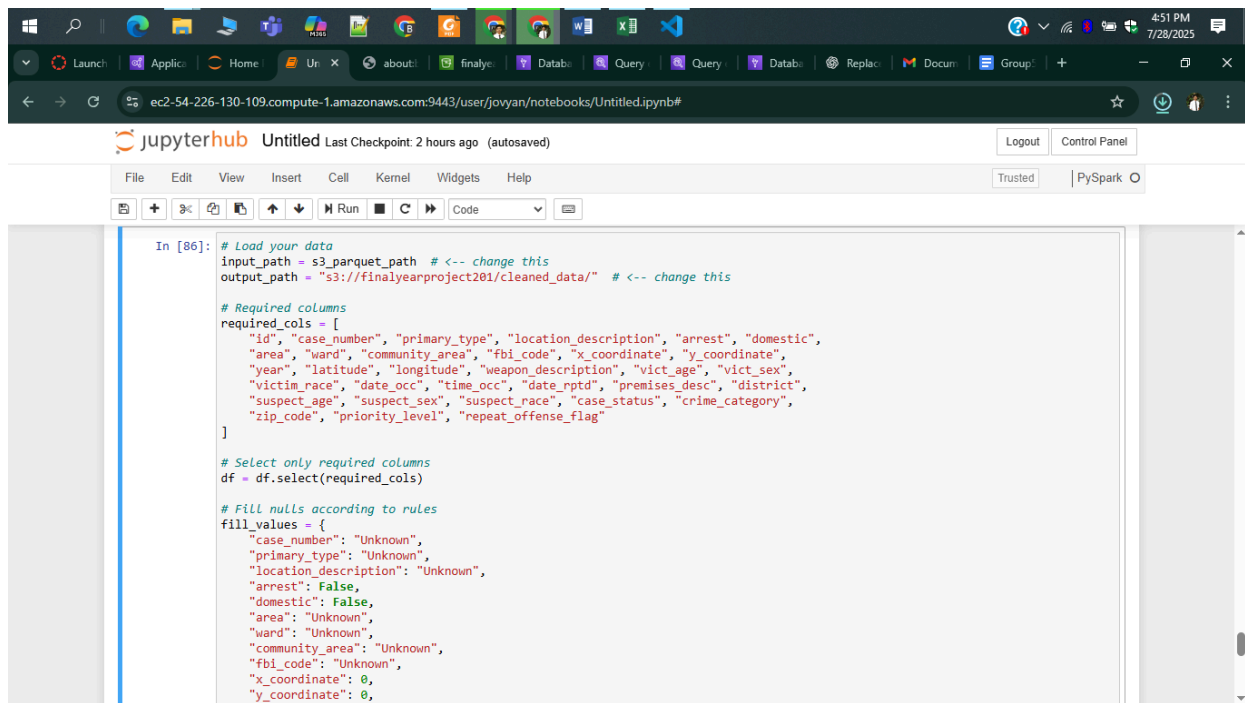
The screenshot shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running cells, and code execution. The notebook is titled "Untitled" and shows a schema definition for a crime data table. The schema is defined using a series of comments and data types, indicating nullable fields.

```
-- fbi_code: string (nullable = true)
-- x_coordinate: long (nullable = true)
-- y_coordinate: long (nullable = true)
-- year: integer (nullable = true)
-- latitude: double (nullable = true)
-- longitude: double (nullable = true)
-- location: string (nullable = true)
-- weapon_description: string (nullable = true)
-- vict_age: integer (nullable = true)
-- vict_sex: string (nullable = true)
-- victim_race: string (nullable = true)
-- date_occ: timestamp (nullable = true)
-- time_occ: string (nullable = true)
-- date_rptd: string (nullable = true)
-- time_rptd: string (nullable = true)
-- date_arrested: string (nullable = true)
-- time_arrested: string (nullable = true)
-- premises_desc: string (nullable = true)
-- district: string (nullable = true)
-- suspect_age: integer (nullable = true)
-- suspect_sex: string (nullable = true)
-- suspect_race: string (nullable = true)
-- case_status: string (nullable = true)
-- crime_category: string (nullable = true)
-- secondary_description: string (nullable = true)
-- census_tract: string (nullable = true)
-- zip_code: string (nullable = true)
-- incident_narrative: string (nullable = true)
-- priority_level: string (nullable = true)
-- repeat_offense_flag: boolean (nullable = true)
```



The screenshot shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running cells, and code execution. The notebook is titled "Untitled" and shows a schema definition for a crime data table. The schema is defined using a series of comments and data types, indicating nullable fields.

```
-- fbi_code: string (nullable = true)
-- x_coordinate: long (nullable = true)
-- y_coordinate: long (nullable = true)
-- year: integer (nullable = true)
-- latitude: double (nullable = true)
-- longitude: double (nullable = true)
-- location: string (nullable = true)
-- weapon_description: string (nullable = true)
-- vict_age: integer (nullable = true)
-- vict_sex: string (nullable = true)
-- victim_race: string (nullable = true)
-- date_occ: timestamp (nullable = true)
-- time_occ: string (nullable = true)
-- date_rptd: string (nullable = true)
-- time_rptd: string (nullable = true)
-- date_arrested: string (nullable = true)
-- time_arrested: string (nullable = true)
-- premises_desc: string (nullable = true)
-- district: string (nullable = true)
-- suspect_age: integer (nullable = true)
-- suspect_sex: string (nullable = true)
-- suspect_race: string (nullable = true)
-- case_status: string (nullable = true)
-- crime_category: string (nullable = true)
-- secondary_description: string (nullable = true)
-- census_tract: string (nullable = true)
-- zip_code: string (nullable = true)
-- incident_narrative: string (nullable = true)
-- priority_level: string (nullable = true)
-- repeat_offense_flag: boolean (nullable = true)
```



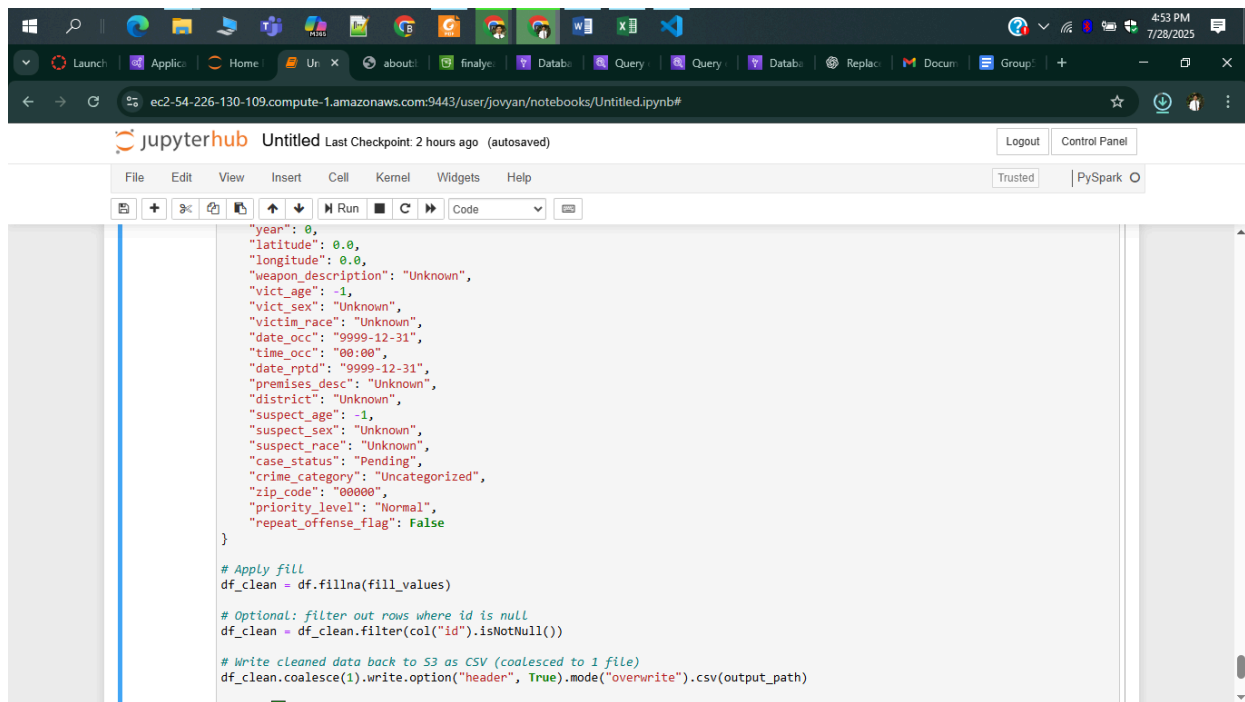
The screenshot shows a Jupyter Notebook interface with a dark theme. The browser address bar at the top displays the URL: `ec2-54-226-130-109.compute-1.amazonaws.com:9443/user/jovyan/notebooks/Untitled.ipynb#`. The Jupyter Notebook header includes the logo, the text "Untitled Last Checkpoint: 2 hours ago (autosaved)", and buttons for "Logout" and "Control Panel". Below the header is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". A toolbar contains icons for file operations and execution. The main area shows a code cell with the following Python code:

```
In [86]: # Load your data
input_path = s3_parquet_path # <-- change this
output_path = "s3://finalyearproject201/cleaned_data/" # <-- change this

# Required columns
required_cols = [
    "id", "case_number", "primary_type", "location_description", "arrest", "domestic",
    "area", "ward", "community_area", "fbi_code", "x_coordinate", "y_coordinate",
    "year", "latitude", "longitude", "weapon_description", "vict_age", "vict_sex",
    "victim_race", "date_occ", "time_occ", "date_rptd", "premises_desc", "district",
    "suspect_age", "suspect_sex", "suspect_race", "case_status", "crime_category",
    "zip_code", "priority_level", "repeat_offense_flag"
]

# Select only required columns
df = df.select(required_cols)

# Fill nulls according to rules
fill_values = {
    "case_number": "Unknown",
    "primary_type": "Unknown",
    "location_description": "Unknown",
    "arrest": False,
    "domestic": False,
    "area": "Unknown",
    "ward": "Unknown",
    "community_area": "Unknown",
    "fbi_code": "Unknown",
    "x_coordinate": 0,
    "y_coordinate": 0,
```



The screenshot shows the same Jupyter Notebook interface, but with the second cell of code visible. The code continues from the previous cell:

```
    "year": 0,
    "latitude": 0.0,
    "longitude": 0.0,
    "weapon_description": "Unknown",
    "vict_age": -1,
    "vict_sex": "Unknown",
    "victim_race": "Unknown",
    "date_occ": "9999-12-31",
    "time_occ": "00:00",
    "date_rptd": "9999-12-31",
    "premises_desc": "Unknown",
    "district": "Unknown",
    "suspect_age": -1,
    "suspect_sex": "Unknown",
    "suspect_race": "Unknown",
    "case_status": "Pending",
    "crime_category": "Uncategorized",
    "zip_code": "00000",
    "priority_level": "Normal",
    "repeat_offense_flag": False
}

# Apply fill
df_clean = df.fillna(fill_values)

# Optional: filter out rows where id is null
df_clean = df_clean.filter(col("id").isNotNull())

# Write cleaned data back to S3 as CSV (coalesced to 1 file)
df_clean.coalesce(1).write.option("header", True).mode("overwrite").csv(output_path)
```


Clean Data:

Amazon S3

General purpose buckets

Directory buckets

Table buckets

Vector buckets

Access Grants

Access Points (General Purpose Buckets, FSx file systems)

Access Points (Directory Buckets)

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight

cleaned_data/

Objects

Properties

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

Name	Type	Last modified	Size	Storage class
_SUCCESS	-	July 28, 2025, 16:18:05 (UTC+05:30)	0 B	Standard
part-00000-e6933299-a3b0-4b8e-8770-419ee1585a2a-c000.csv	csv	July 28, 2025, 16:11:32 (UTC+05:30)	3.5 GB	Standard

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Zero-ETL integrations

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Interactive Sessions

Data classification tools

Announcing new optimization features for Apache Iceberg tables

Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

masterdata123

Database properties

Name	Description	Location	Created on (UTC)
masterdata123	-	-	July 28, 2025 at 10:54:55

Tables (1)

View and manage all available tables.

Filter tables

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
cleaned_data	masterdata123	s3://finalyearproject20	CSV	-	Table data	View data quality	View statistics

us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/5e59bf7f-d5b0-49a0-a7b5-13bf28ec4f25

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings Workgroup primary

Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences. [Edit preferences](#)

Data

Data source: AwsDataCatalog

Catalog: None

Database: masterdata123

Tables and views: [Create](#)

Filter tables and views

Tables (1): cleaned_data

Views (0)

Query 1: `SELECT * FROM "AwsDataCatalog"."masterdata123"."cleaned_data" limit 10;`

SQL Ln 1, Col 72

[Run again](#) [Explain](#) [Cancel](#) [Clear](#) [Create](#)

[Query results](#) [Query stats](#)

Reuse query results up to 60 minutes ago

us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/5e59bf7f-d5b0-49a0-a7b5-13bf28ec4f25

Amazon Athena > Query editor

[Query results](#) [Query stats](#)

Completed Time in queue: 71 ms Run time: 724 ms Data scanned: 2.13 MB

[Copy](#) [Download results CSV](#)

Results (10)

Search rows

#	id	case_number	primary_type	location_description	arrest	domestic	area	ward	community_area
1	29019099	DPD707893	Arson	Street	false	false	Oak Cliff	W3	Oak Lawn
2	29019100	DPD754795	Theft	Parking Lot	true	true	North Dallas	Unknown	Bishop Arts
3	29019101	DPD571163	Trespassing	Apartment	true	false	North Dallas	W3	Preston Hollow
4	29019102	DPD683659	Drug Offense	Retail Store	true	false	Deep Ellum	W11	Cedars
5	29019103	DPD499012	Robbery	Gas Station	false	false	Pleasant Grove	Unknown	Bishop Arts
6	29019104	DPD978360	Assault	Street	false	false	Oak Cliff	W13	Cedars
7	29019105	DPD025986	Robbery	Hotel	false	false	Deep Ellum	W14	Bishop Arts
8	29019106	DPD721979	Theft	Bar	false	false	North Dallas	W6	Cedars
9	29019107	DPD224615	Theft	School	false	true	Oak Cliff	W5	Bishop Arts
10	29019108	DPD644250	Battery	Park	true	true	Deep Ellum	W5	Oak Lawn