

AWS-SPARK ON UNSTRUCTURED DATA STREAMING

01

Abstract



This project addresses the growing need to extract value from data that doesn't conform to traditional, structured formats. Unstructured data, such as text, images, audio, and video, is becoming increasingly prevalent, and organizations need effective ways to analyze it for insights. The primary goal is to design and implement a data engineering pipeline that effectively processes unstructured data streams in real-time. This pipeline leverages the power of Apache Spark in conjunction with various Amazon Web Services (AWS) to address the unique challenges posed by unstructured data.

Introduction

PROBLEM STATEMENT

Unlike structured data (e.g., relational databases) with predefined schemas, unstructured data lacks a rigid format. This inherent lack of structure makes it difficult to process, analyze, and extract meaningful insights using traditional data processing techniques.

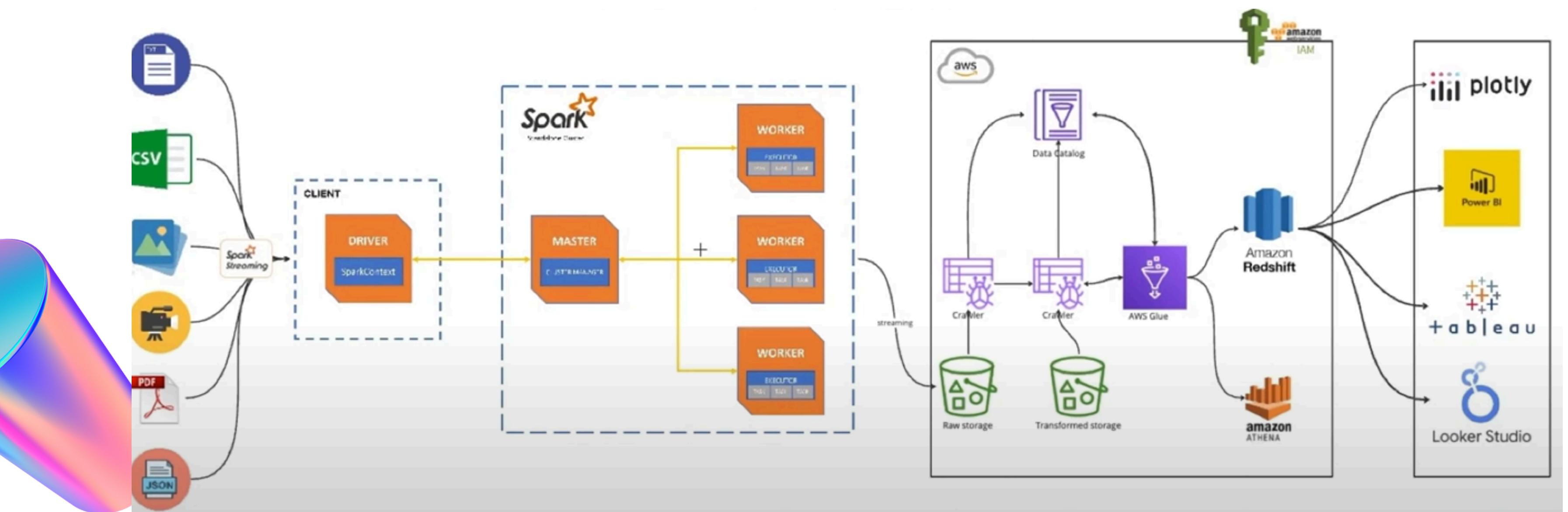
OBJECTIVE

To build a real-time data streaming pipeline that can ingest, process, and analyze unstructured data from various sources.

GOAL

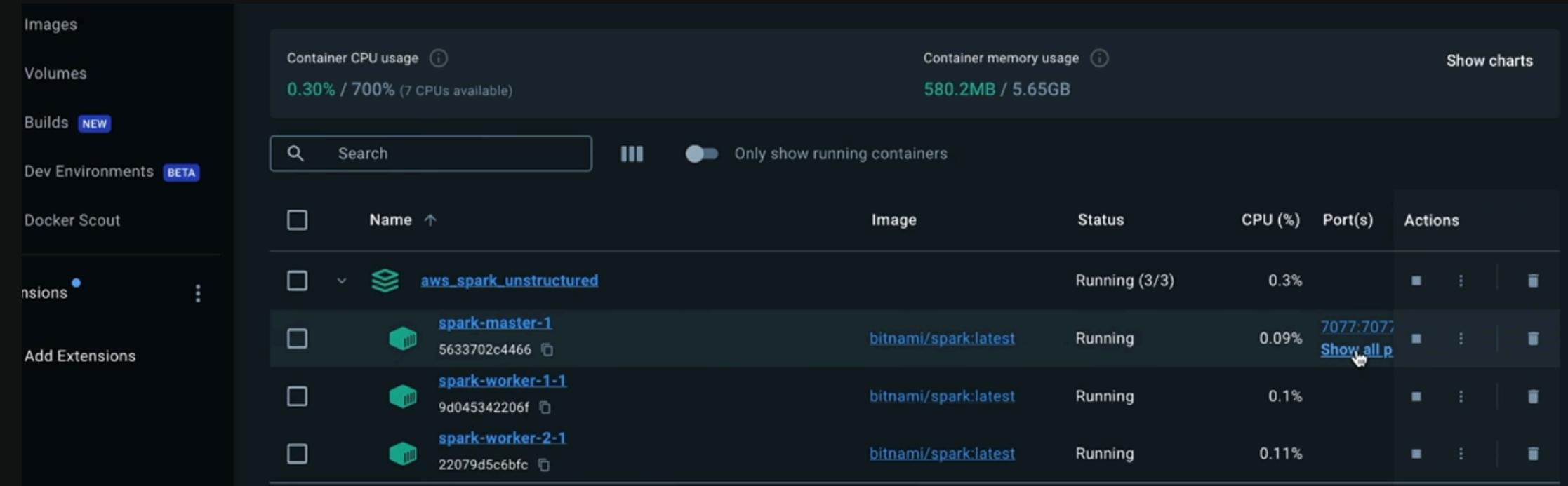
Specify the key functionalities and outcomes expected from the pipeline. Develop a system capable of handling various data types, extracting meaningful information, storing processed data in S3, cataloging data with Glue, and querying with Athena.

04 ARCHITECTURE



System Design

Development Environment:
Specifies the software and tools used for developing the project, including PyCharm, Python version 3.9.6, Docker



The screenshot shows a Docker interface with the following details:

Container CPU usage: 0.30% / 700% (7 CPUs available)

Container memory usage: 580.2MB / 5.65GB

Search: Search bar with placeholder "Search".

Filter: "Only show running containers" toggle switch.

Table Headers: Name, Image, Status, CPU (%), Port(s), Actions.

Table Data:

Name	Image	Status	CPU (%)	Port(s)	Actions
aws_spark_unstructured	bitnami/spark:latest	Running (3/3)	0.3%		⋮ ⌂
spark-master-1	bitnami/spark:latest	Running	0.09%	7077:7077 Show all p	⋮ ⌂
spark-worker-1-1	bitnami/spark:latest	Running	0.1%		⋮ ⌂
spark-worker-2-1	bitnami/spark:latest	Running	0.11%		⋮ ⌂

```
.py config.py udf_utils.py docker-compose.yml x

version: '3'

x-spark-common: &spark-common
  image: bitnami/spark:latest
  volumes:
    - ./jobs:/opt/bitnami/spark/jobs
  command: bin/spark-class org.apache.spark.deploy.worker.Worker spark://spark-master:7077
  depends_on:
    - spark-master
  environment:
    SPARK_MODE: Worker
    SPARK_WORKER_CORES: 2
    SPARK_WORKER_MEMORY: 1g
    SPARK_MASTER_URL: spark://spark-master:7077
  networks:
    - datamasterylab

services:
  spark-master:
    image: bitnami/spark:latest
    volumes:
      - ./jobs:/opt/bitnami/spark/jobs
    command: bin/spark-class org.apache.spark.deploy.master.Master
    ports:
      - "9092:8080"
      - "7077:7077"
```

```
main.py config.py udf_utils.py docker-compose.yml x

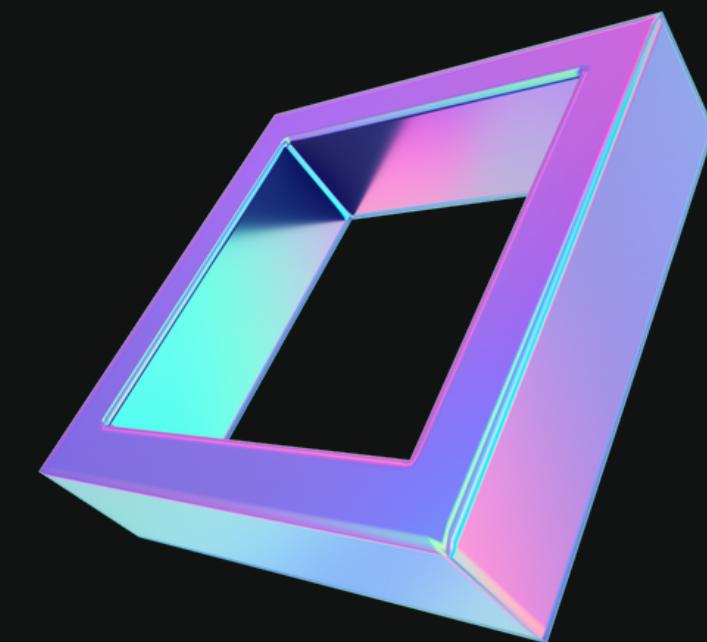
x-spark-common: &spark-common
  - datamasterylab

services:
  spark-master:
    image: bitnami/spark:latest
    volumes:
      - ./jobs:/opt/bitnami/spark/jobs
    command: bin/spark-class org.apache.spark.deploy.master.Master
    ports:
      - "9092:8080"
      - "7077:7077"
    networks:
      - datamasterylab

  spark-worker-1:
    <<: *spark-common
  spark-worker-2:
    <<: *spark-common
  spark-worker-3:
    <<: *spark-common
  spark-worker-4:
    <<: *spark-common

networks:
  datamasterylab:
```

Implementation



1. Setting up Spark Session:

```
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLev
Spark context Web UI available at http://host.docker.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1742143758386).
Spark session available as 'spark'.
Welcome to

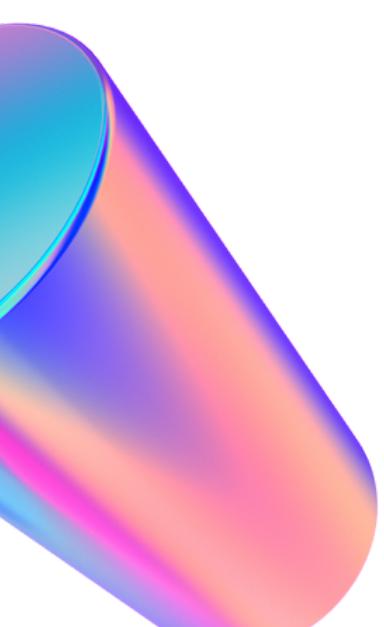
    ___      __
   / _ \ \  / /_ \ \ / \
  / \ \ \ / \ \ / \ \ / \
 / \ \ \ / \ \ \ / \ \ \ \  version 3.5.5
 / \ \ \ / \ \ \ / \ \ \ \ /
Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 21.0.5)
Type in expressions to have them evaluated.
Type :help for more information.
```

The screenshot shows the Apache Spark 3.5.0 Web UI running on port 9090. The main page displays the master node's configuration, including its URL (spark://172.18.0.2:7077), the number of alive workers (2), and memory usage (2.0 GiB Total, 0.0 B Used). Below this, the 'Workers' section lists two workers: worker-20240302214445-172.18.0.3-33059 and worker-20240302214445-172.18.0.4-38973, both marked as ALIVE. There are also sections for 'Running Applications' (0) and 'Completed Applications' (0).

Worker ID	Address	State	Cores	Memory	Resources
worker-20240302214445-172.18.0.3-33059	172.18.0.3:33059	ALIVE	2 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20240302214445-172.18.0.4-38973	172.18.0.4:38973	ALIVE	2 (0 Used)	1024.0 MiB (0.0 B Used)	

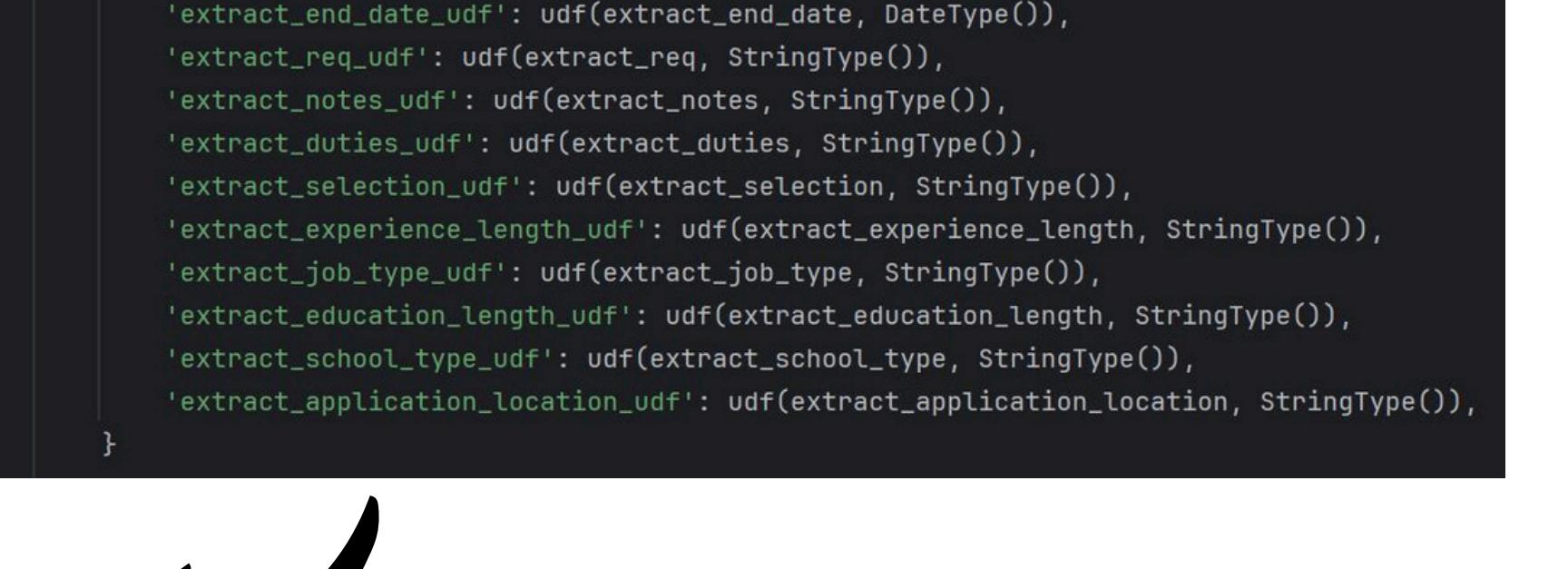
IMPLEMENTATION

2. Create main.py file:

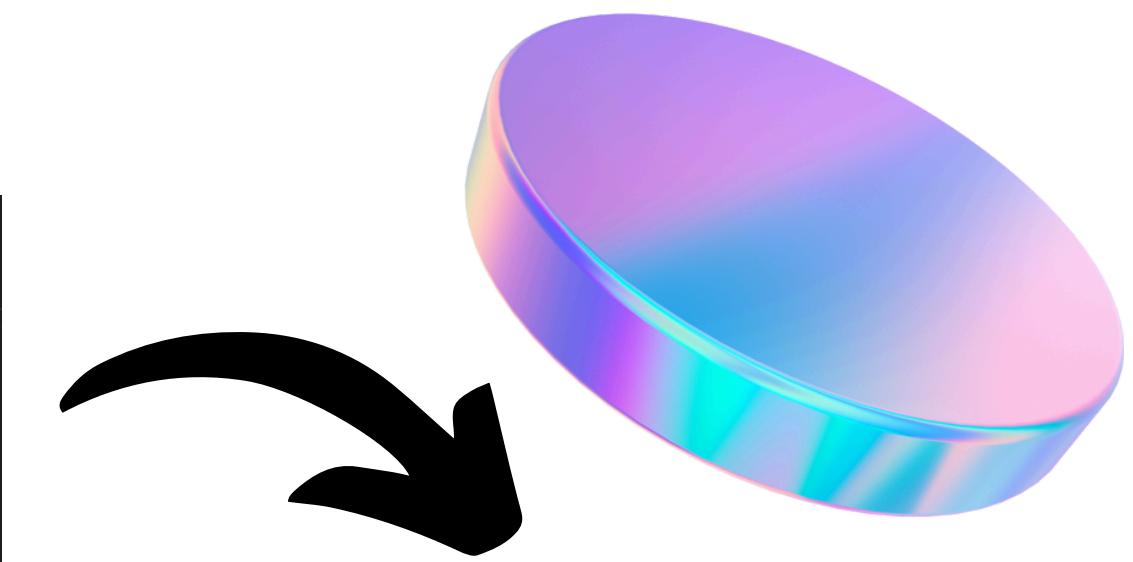


```
main.py x config.py udf_utils.py docker-compose.yml

1 from pyspark.sql import SparkSession, DataFrame
2 from pyspark.sql.functions import udf, regexp_replace, explode
3 from pyspark.sql.types import StructType, StructField, StringType, DoubleType, DateType, ArrayType
4
5 from config.config import configuration
6 from udf_utils import (
7     split_job_postings, extract_file_name, extract_position, extract_classcode,
8     extract_start_date, extract_end_date, extract_req, extract_notes,
9     extract_duties, extract_selection, extract_experience_length,
10    extract_job_type, extract_education_length, extract_school_type,
11    extract_application_location, extract_salary_start, extract_salary_end
12 )
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33     }
34
35 if __name__ == "__main__":
36     spark = (SparkSession.builder.appName('JobPostingProcessor')
37         .config('spark.jars.packages',
38             'org.apache.hadoop:hadoop-aws:3.3.1,com.amazonaws:aws-java-sdk:1.11.469')
39         .config('spark.hadoop.fs.s3a.impl', 'org.apache.hadoop.fs.s3a.S3AFileSystem')
40         .config("spark.hadoop.fs.s3a.access.key", "AKIAWCYX7WR77EMX2B5N")
41         .config("spark.hadoop.fs.s3a.secret.key", "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")
42         .config("spark.hadoop.fs.s3a.endpoint", "s3.amazonaws.com")
43         .config('spark.hadoop.fs.s3a.aws.credentials.provider',
44             'org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider')
45         .getOrCreate())
46
47     text_input_dir = 'file:///C:/Users/sahil/PycharmProjects/PythonProject2/input/input_text2'
48     csv_input_dir = 'file:///C:/Users/sahil/PycharmProjects/PythonProject2/input/input_csv'
49     image_input_dir = 'file:///C:/Users/sahil/PycharmProjects/PythonProject2/input/input_image'
50     json_input_dir = 'file:///C:/Users/sahil/PycharmProjects/PythonProject2/input/input_json2'
51     pdf_input_dir = 'file:///C:/Users/sahil/PycharmProjects/PythonProject2/input/input_pdf'
52     video_input_dir = 'file:///C:/Users/sahil/PycharmProjects/PythonProject2/input/input_video'
```



```
def define_udfs(): 1 usage
    return {
        'split_jobs_udf': udf(split_job_postings, ArrayType(StringType())),
        'extract_file_name_udf': udf(extract_file_name, StringType()),
        'extract_position_udf': udf(extract_position, StringType()),
        'extract_classcode_udf': udf(extract_classcode, StringType()),
        'extract_salary_start_udf': udf(extract_salary_start, DoubleType()),
        'extract_salary_end_udf': udf(extract_salary_end, DoubleType()),
        'extract_start_date_udf': udf(extract_start_date, DateType()),
        'extract_end_date_udf': udf(extract_end_date, DateType()),
        'extract_req_udf': udf(extract_req, StringType()),
        'extract_notes_udf': udf(extract_notes, StringType()),
        'extract_duties_udf': udf(extract_duties, StringType()),
        'extract_selection_udf': udf(extract_selection, StringType()),
        'extract_experience_length_udf': udf(extract_experience_length, StringType()),
        'extract_job_type_udf': udf(extract_job_type, StringType()),
        'extract_education_length_udf': udf(extract_education_length, StringType()),
        'extract_school_type_udf': udf(extract_school_type, StringType()),
        'extract_application_location_udf': udf(extract_application_location, StringType()),}
```



IMPLEMENTATION

3. Setting up udf_utils.py:

```
data_schema = StructType([
    StructField( name: 'file_name', StringType(), nullable: True),
    StructField( name: 'position', StringType(), nullable: True),
    StructField( name: 'classcode', StringType(), nullable: True),
    StructField( name: 'salary_start', DoubleType(), nullable: True),
    StructField( name: 'salary_end', DoubleType(), nullable: True),
    StructField( name: 'start_date', DateType(), nullable: True),
    StructField( name: 'end_date', DateType(), nullable: True),
    StructField( name: 'req', StringType(), nullable: True),
    StructField( name: 'notes', StringType(), nullable: True),
    StructField( name: 'duties', StringType(), nullable: True),
    StructField( name: 'selection', StringType(), nullable: True),
    StructField( name: 'experience_length', StringType(), nullable: True),
    StructField( name: 'job_type', StringType(), nullable: True),
    StructField( name: 'education_length', StringType(), nullable: True),
    StructField( name: 'school_type', StringType(), nullable: True),
    StructField( name: 'application_location', StringType(), nullable: True),
])
])
```



```
.py config.py udf_utils.py docker-compose.yml
import re
from datetime import datetime
from pyspark.sql.types import ArrayType, StringType

def split_job_postings(file_content): 2 usages
    """Splits multi-job file into individual postings"""
    try:
        return re.split( pattern: r'\n(?=Job Posting: job\d+\.\txt)', file_content)[1:]
    except Exception as e:
        raise ValueError(f"Error splitting jobs: {str(e)}")

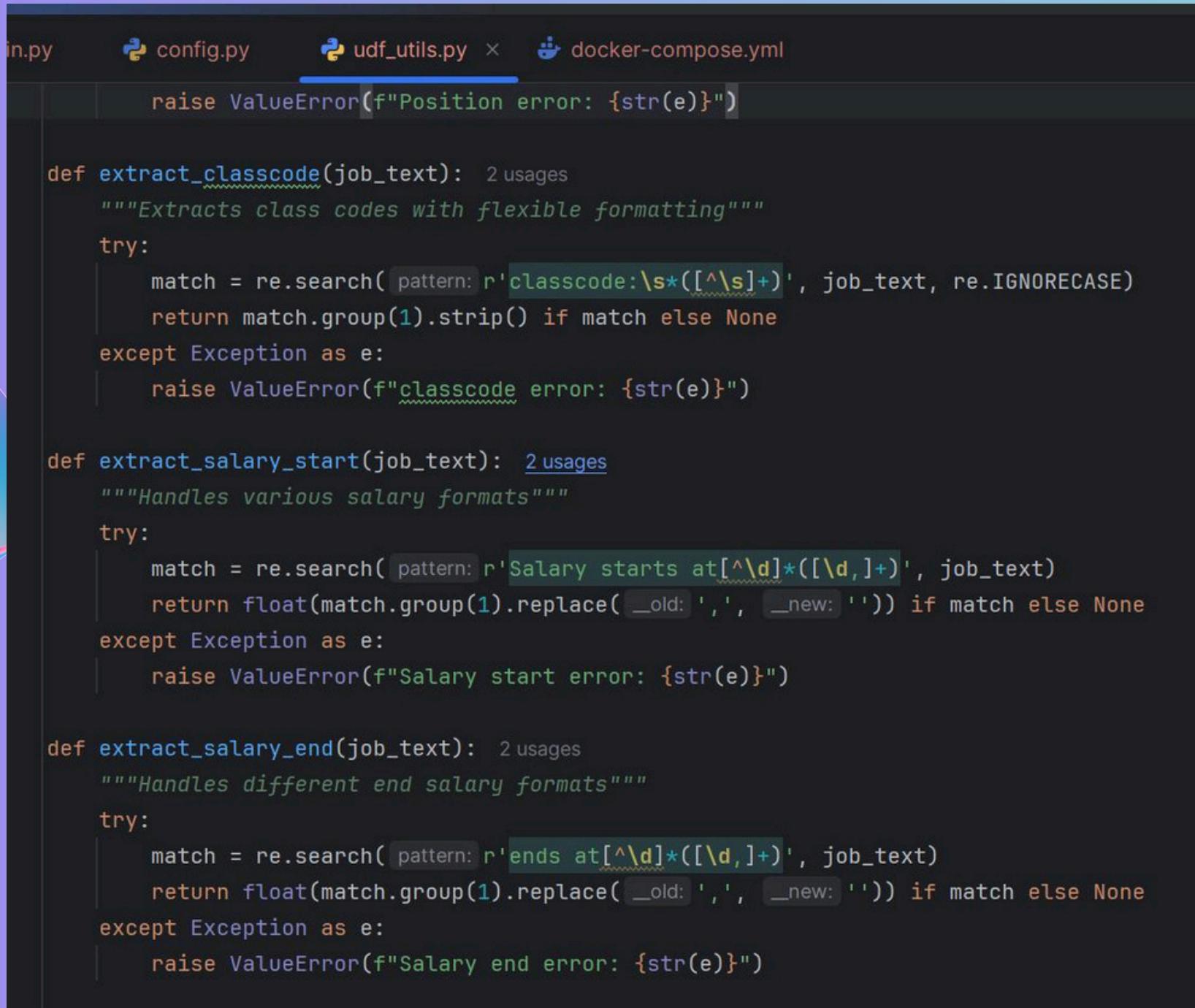
def extract_file_name(file_content): 2 usages
    """Gets main category from first line"""
    return file_content.split('\n')[0].strip()

def extract_position(job_text): 2 usages
    """Extracts all positions from job text"""
    try:
        matches = re.findall( pattern: r'Position:\s*(.+?)\n', job_text)
        return ', '.join(matches) if matches else None
    except Exception as e:
        raise ValueError(f"Position error: {str(e)}")

def extract_classcode(job_text): 2 usages
    """Extracts class codes with flexible formatting"""
    try:
```

Implementation

3. Create udf_utils.py file:



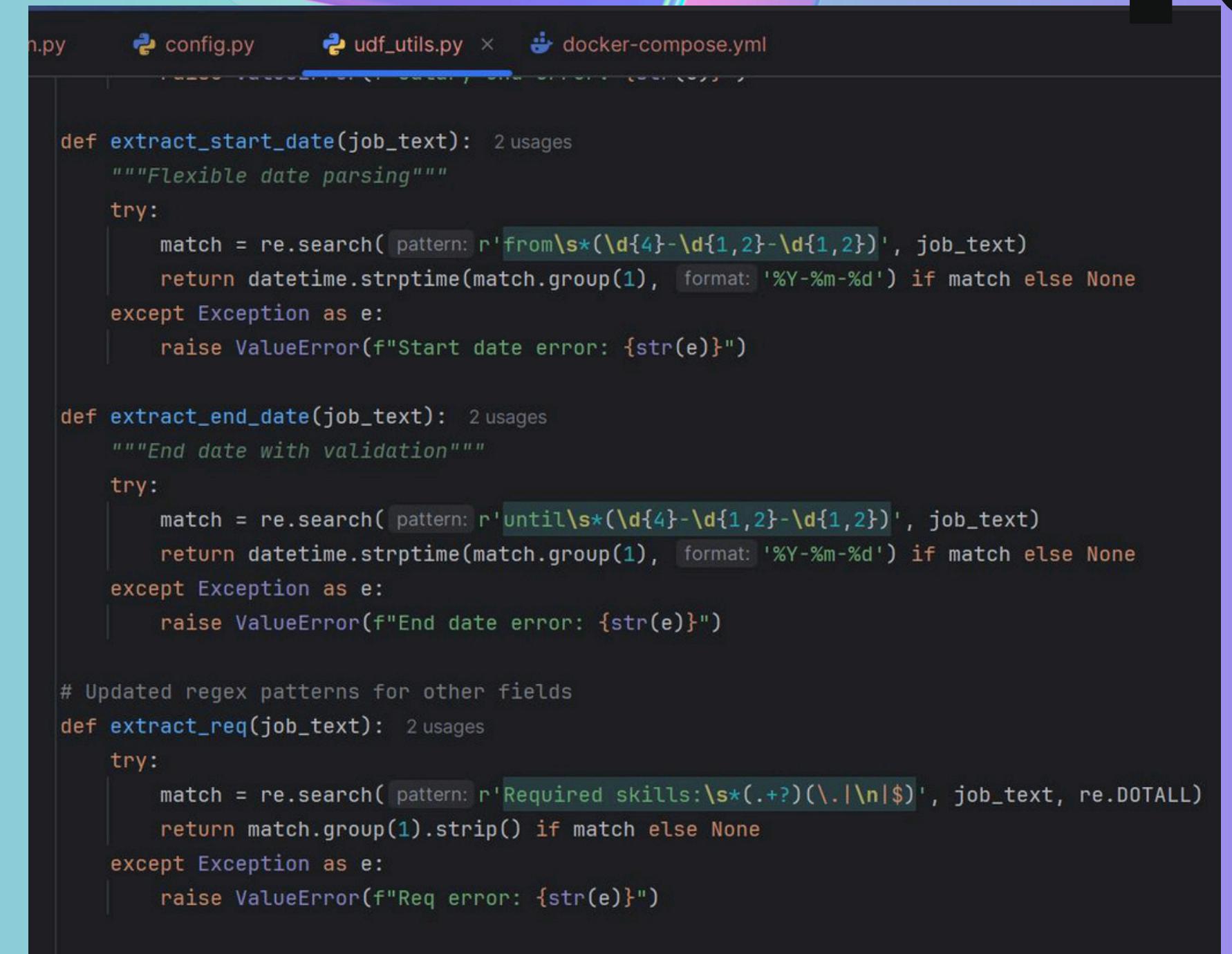
```
in.py config.py udf_utils.py docker-compose.yml

raise ValueError(f"Position error: {str(e)}")

def extract_classcode(job_text): 2 usages
    """Extracts class codes with flexible formatting"""
    try:
        match = re.search(pattern: r'classcode:\s*(\^|\s)+', job_text, re.IGNORECASE)
        return match.group(1).strip() if match else None
    except Exception as e:
        raise ValueError(f"Classcode error: {str(e)}")

def extract_salary_start(job_text): 2 usages
    """Handles various salary formats"""
    try:
        match = re.search(pattern: r'Salary starts at[\^|\d]*([\d,]+)', job_text)
        return float(match.group(1).replace(',', '')) if match else None
    except Exception as e:
        raise ValueError(f"Salary start error: {str(e)}")

def extract_salary_end(job_text): 2 usages
    """Handles different end salary formats"""
    try:
        match = re.search(pattern: r'ends at[\^|\d]*([\d,]+)', job_text)
        return float(match.group(1).replace(',', '')) if match else None
    except Exception as e:
        raise ValueError(f"Salary end error: {str(e)}")
```



```
in.py config.py udf_utils.py docker-compose.yml

def extract_start_date(job_text): 2 usages
    """Flexible date parsing"""
    try:
        match = re.search(pattern: r'from\s*(\d{4}-\d{1,2}-\d{1,2})', job_text)
        return datetime.strptime(match.group(1), format: '%Y-%m-%d') if match else None
    except Exception as e:
        raise ValueError(f"Start date error: {str(e)}")

def extract_end_date(job_text): 2 usages
    """End date with validation"""
    try:
        match = re.search(pattern: r'until\s*(\d{4}-\d{1,2}-\d{1,2})', job_text)
        return datetime.strptime(match.group(1), format: '%Y-%m-%d') if match else None
    except Exception as e:
        raise ValueError(f"End date error: {str(e)}")

# Updated regex patterns for other fields
def extract_req(job_text): 2 usages
    try:
        match = re.search(pattern: r'Required skills:\s*(.+?)(\.|\\n|$', job_text, re.DOTALL)
        return match.group(1).strip() if match else None
    except Exception as e:
        raise ValueError(f"Req error: {str(e)}")
```

IMPLEMENTATION

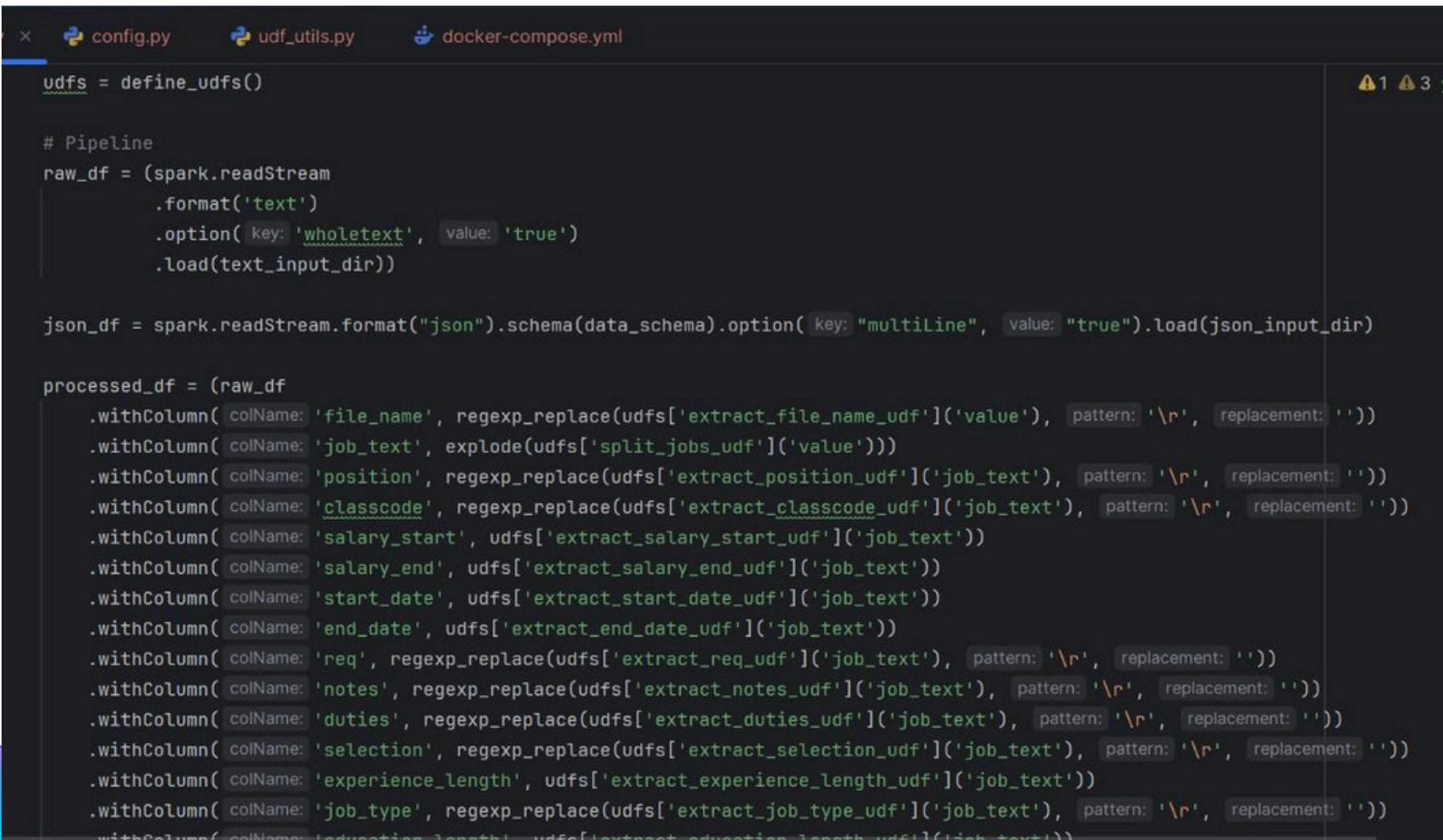


```
main.py config.py udf_utils.py docker-compose.yml
1
2 def extract_notes(job_text): 2 usages
3     try:
4         match = re.search( pattern: r'Additional notes:\s*(.+?)(\.\|\n|$)', job_text, re.DOTALL)
5         return match.group(1).strip() if match else None
6     except Exception as e:
7         raise ValueError(f"Notes error: {str(e)}")
8
9 def extract_duties(job_text): 2 usages
10    try:
11        match = re.search( pattern: r'Primary duties:\s*(.+?)(\.\|\n|$)', job_text, re.DOTALL)
12        return match.group(1).strip() if match else None
13    except Exception as e:
14        raise ValueError(f"Duties error: {str(e)}")
15
16 def extract_selection(job_text): 2 usages
17    try:
18        match = re.search( pattern: r'Selection process:\s*(.+?)(\.\|\n|$)', job_text, re.DOTALL)
19        return match.group(1).strip() if match else None
20    except Exception as e:
21        raise ValueError(f"Selection error: {str(e)}")
22
23 def extract_experience_length(job_text): 2 usages
24    try:
25        match = re.search( pattern: r'Experience required:\s*(\d+)[+]?\s*years', job_text)
26        return int(match.group(1)) if match else None
27
```

```
in.py config.py udf_utils.py docker-compose.yml
def extract_job_type(job_text): 2 usages
    return match.group(1).strip() if match else None
except Exception as e:
    raise ValueError(f"Job type error: {str(e)}")
def extract_education_length(job_text): 2 usages
    try:
        match = re.search( pattern: r'Educational requirement:\s*(\d+)\s*years', job_text)
        return int(match.group(1)) if match else None
    except Exception as e:
        raise ValueError(f"Edu length error: {str(e)}")
def extract_school_type(job_text): 2 usages
    try:
        match = re.search( pattern: r'Preferred school type:\s*(.+?)(\.\|\n|$)', job_text)
        return match.group(1).strip() if match else None
    except Exception as e:
        raise ValueError(f"School type error: {str(e)}")
def extract_application_location(job_text): 2 usages
    try:
        match = re.search( pattern: r'Application location:\s*(.+?)(\.\|\n|$)', job_text)
        return match.group(1).strip() if match else None
    except Exception as e:
        raise ValueError(f"Location error: {str(e)}")
```

IMPLEMENTATION

4. Checking Data Stream:

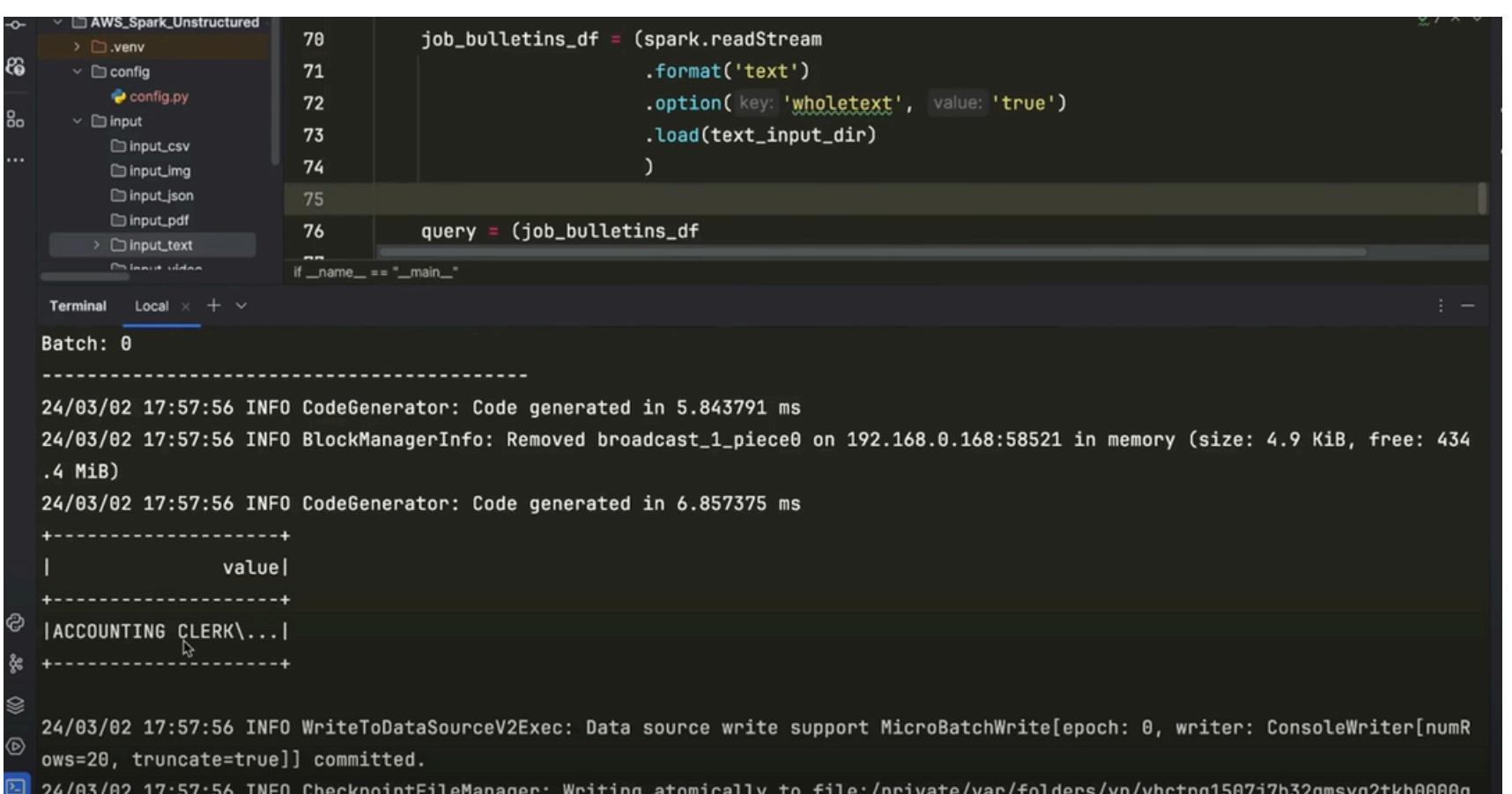


```
udfs = define_udfs()

# Pipeline
raw_df = (spark.readStream
    .format('text')
    .option(key: 'wholetext', value: 'true')
    .load(text_input_dir))

json_df = spark.readStream.format("json").schema(data_schema).option(key: "multiline", value: "true").load(json_input_dir)

processed_df = (raw_df
    .withColumn(colName: 'file_name', regexp_replace(udfs['extract_file_name_udf']('value'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'job_text', explode(udfs['split_jobs_udf']('value')))
    .withColumn(colName: 'position', regexp_replace(udfs['extract_position_udf']('job_text'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'classcode', regexp_replace(udfs['extract_classcode_udf']('job_text'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'salary_start', udfs['extract_salary_start_udf']('job_text'))
    .withColumn(colName: 'salary_end', udfs['extract_salary_end_udf']('job_text'))
    .withColumn(colName: 'start_date', udfs['extract_start_date_udf']('job_text'))
    .withColumn(colName: 'end_date', udfs['extract_end_date_udf']('job_text'))
    .withColumn(colName: 'req', regexp_replace(udfs['extract_req_udf']('job_text'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'notes', regexp_replace(udfs['extract_notes_udf']('job_text'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'duties', regexp_replace(udfs['extract_duties_udf']('job_text'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'selection', regexp_replace(udfs['extract_selection_udf']('job_text'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'experience_length', udfs['extract_experience_length_udf']('job_text'))
    .withColumn(colName: 'job_type', regexp_replace(udfs['extract_job_type_udf']('job_text'), pattern: '\r', replacement: ''))
    .withColumn(colName: 'education_length', udfs['extract_education_length_udf']('job_text')))
```



```
job_bulletins_df = (spark.readStream
    .format('text')
    .option(key: 'wholetext', value: 'true')
    .load(text_input_dir))

query = (job_bulletins_df
    .selectExpr("if __name__ == '__main__'"
```

Batch: 0

```
24/03/02 17:57:56 INFO CodeGenerator: Code generated in 5.843791 ms
24/03/02 17:57:56 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 192.168.0.168:58521 in memory (size: 4.9 KiB, free: 434.4 MiB)
24/03/02 17:57:56 INFO CodeGenerator: Code generated in 6.857375 ms
+-----+
|      value|
+-----+
|ACCOUNTING CLERK\...|
+-----+
```

```
24/03/02 17:57:56 INFO WriteToDataSourceV2Exec: Data source write support MicroBatchWrite[epoch: 0, writer: ConsoleWriter[numRows=20, truncate=true]] committed.
24/03/02 17:57:56 INFO CheckpointFileManager: Writing atomically to file:/private/var/folders/vn/vhctog1507i7h32qmsvn2tkh0000g
```

5. Combining Structured and Unstructured data together and processing it:

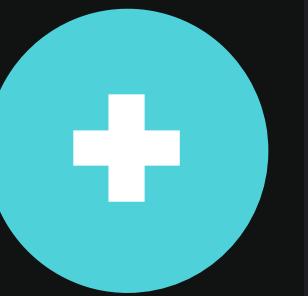


```
processed_df = processed_df.select('file_name', 'position', 'classcode', 'salary_start', 'salary_end',
    'start_date', 'end_date', 'req', 'notes', 'duties', 'selection',
    'experience_length', 'job_type', 'education_length', 'school_type',
    'application_location')

json_df = json_df.select('file_name', 'position', 'classcode', 'salary_start', 'salary_end',
    'start_date', 'end_date', 'req', 'notes', 'duties', 'selection',
    'experience_length', 'job_type', 'education_length', 'school_type',
    'application_location')

union_dataframe = processed_df.union(json_df)

def streamWriter(input: DataFrame, checkpointFolder, output): 1 usage
    return(input.writeStream.
        format('parquet')
        .option( key: 'checkpointLocation', checkpointFolder)
        .option( key: 'path', output)
        .outputMode('append')
        .trigger(processingTime='5 seconds')
        .start()
    )
```



```
x config.py udf_utils.py docker-compose.yml
def streamWriter(input: DataFrame, checkpointFolder, output): 1 usage
    format('parquet')
    .option( key: 'checkpointLocation', checkpointFolder)
    .option( key: 'path', output)
    .outputMode('append')
    .trigger(processingTime='5 seconds')
    .start()
)

# query = (union_dataframe.writeStream
#           .outputMode('append')
#           .format('console')
#           .option('truncate', False)
#           .start())

query = streamWriter(union_dataframe, checkpointFolder='s3a://aws-spark-unstructured-project/checkpoints/',
                     output='s3a://aws-spark-unstructured-project/data/spark-unstructured')

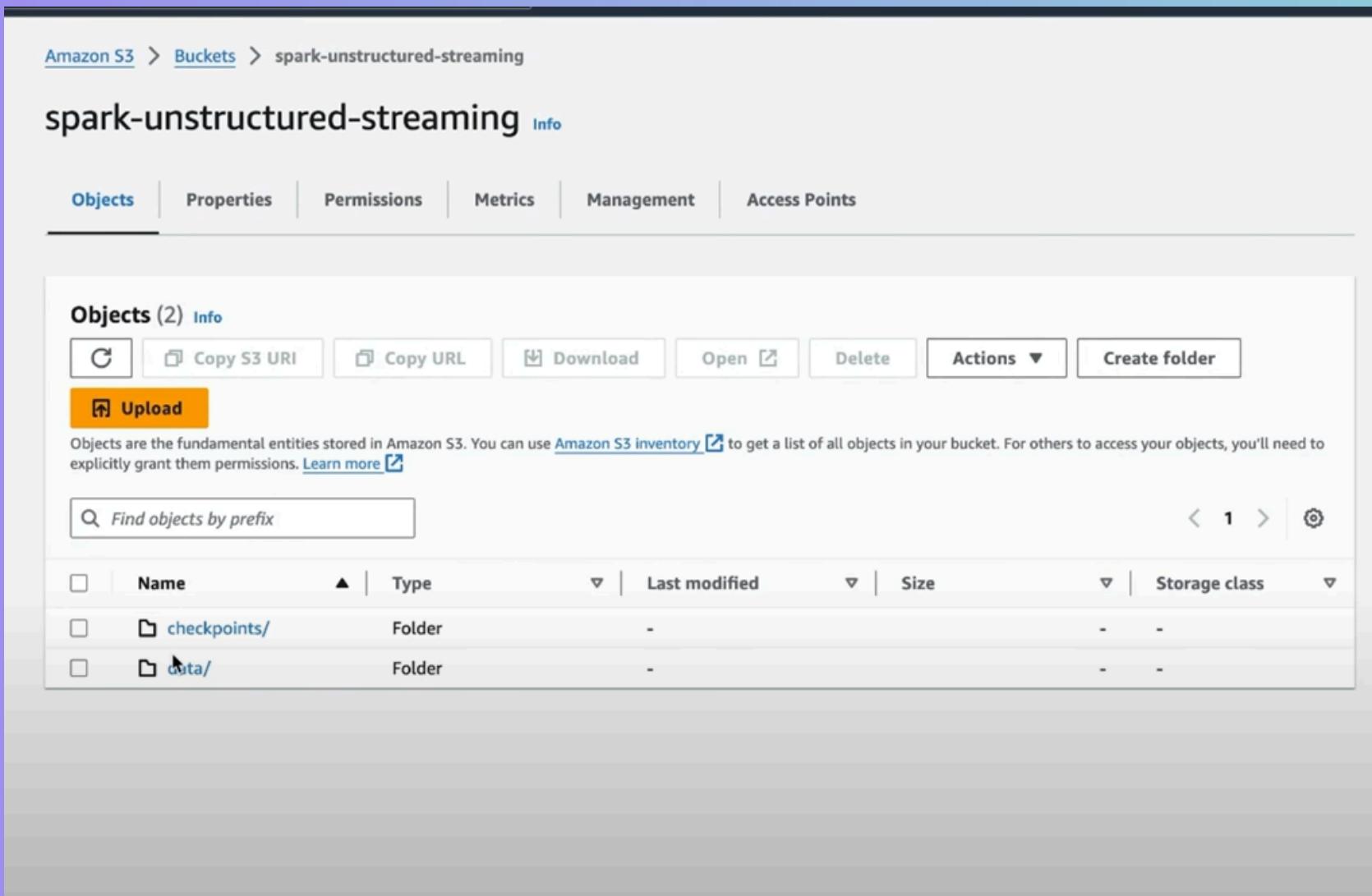
query.awaitTermination()

spark.stop()
```

```
if __name__ == "__main__":
    Terminal Local Local (2) + -
24/03/02 22:01:12 INFO TaskSetManager: Starting task 548.0 in stage 0.0 (TID 548) (172.18.0.3, executor 0, partition 548, PROCESS_LOCAL, 44023 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 543.0 in stage 0.0 (TID 543) in 115 ms on 172.18.0.3 (executor 0) (545/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 549.0 in stage 0.0 (TID 549) (172.18.0.4, executor 1, partition 549, PROCESS_LOCAL, 44017 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 545.0 in stage 0.0 (TID 545) in 94 ms on 172.18.0.4 (executor 1) (546/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 550.0 in stage 0.0 (TID 550) (172.18.0.4, executor 1, partition 550, PROCESS_LOCAL, 44034 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 546.0 in stage 0.0 (TID 546) in 96 ms on 172.18.0.4 (executor 1) (547/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 551.0 in stage 0.0 (TID 551) (172.18.0.3, executor 0, partition 551, PROCESS_LOCAL, 44056 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 547.0 in stage 0.0 (TID 547) in 85 ms on 172.18.0.3 (executor 0) (548/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 552.0 in stage 0.0 (TID 552) (172.18.0.3, executor 0, partition 552, PROCESS_LOCAL, 44021 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 548.0 in stage 0.0 (TID 548) in 78 ms on 172.18.0.3 (executor 0) (549/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 553.0 in stage 0.0 (TID 553) (172.18.0.4, executor 1, partition 553, PROCESS_LOCAL, 44015 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 554.0 in stage 0.0 (TID 554) (172.18.0.4, executor 1, partition 554, PROCESS_LOCAL, 44017 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 549.0 in stage 0.0 (TID 549) in 86 ms on 172.18.0.4 (executor 1) (550/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 550.0 in stage 0.0 (TID 550) in 69 ms on 172.18.0.4 (executor 1) (551/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 555.0 in stage 0.0 (TID 555) (172.18.0.3, executor 0, partition 555, PROCESS_LOCAL, 43997 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 556.0 in stage 0.0 (TID 556) (172.18.0.3, executor 0, partition 556, PROCESS_LOCAL, 44015 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 552.0 in stage 0.0 (TID 552) in 51 ms on 172.18.0.3 (executor 0) (552/650)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 551.0 in stage 0.0 (TID 551) in 68 ms on 172.18.0.3 (executor 0) (553/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 557.0 in stage 0.0 (TID 557) (172.18.0.4, executor 1, partition 557, PROCESS_LOCAL, 44002 bytes)
24/03/02 22:01:12 INFO TaskSetManager: Finished task 553.0 in stage 0.0 (TID 553) in 52 ms on 172.18.0.4 (executor 1) (554/650)
24/03/02 22:01:12 INFO TaskSetManager: Starting task 558.0 in stage 0.0 (TID 558) (172.18.0.4, executor 1, partition 558, PROCESS_LOCAL, 44000 bytes)
```

IMPLEMENTATION

6. AWS INTEGRATION



Amazon S3 > Buckets > spark-unstructured-streaming

spark-unstructured-streaming [Info](#)

Objects (2) [Info](#)

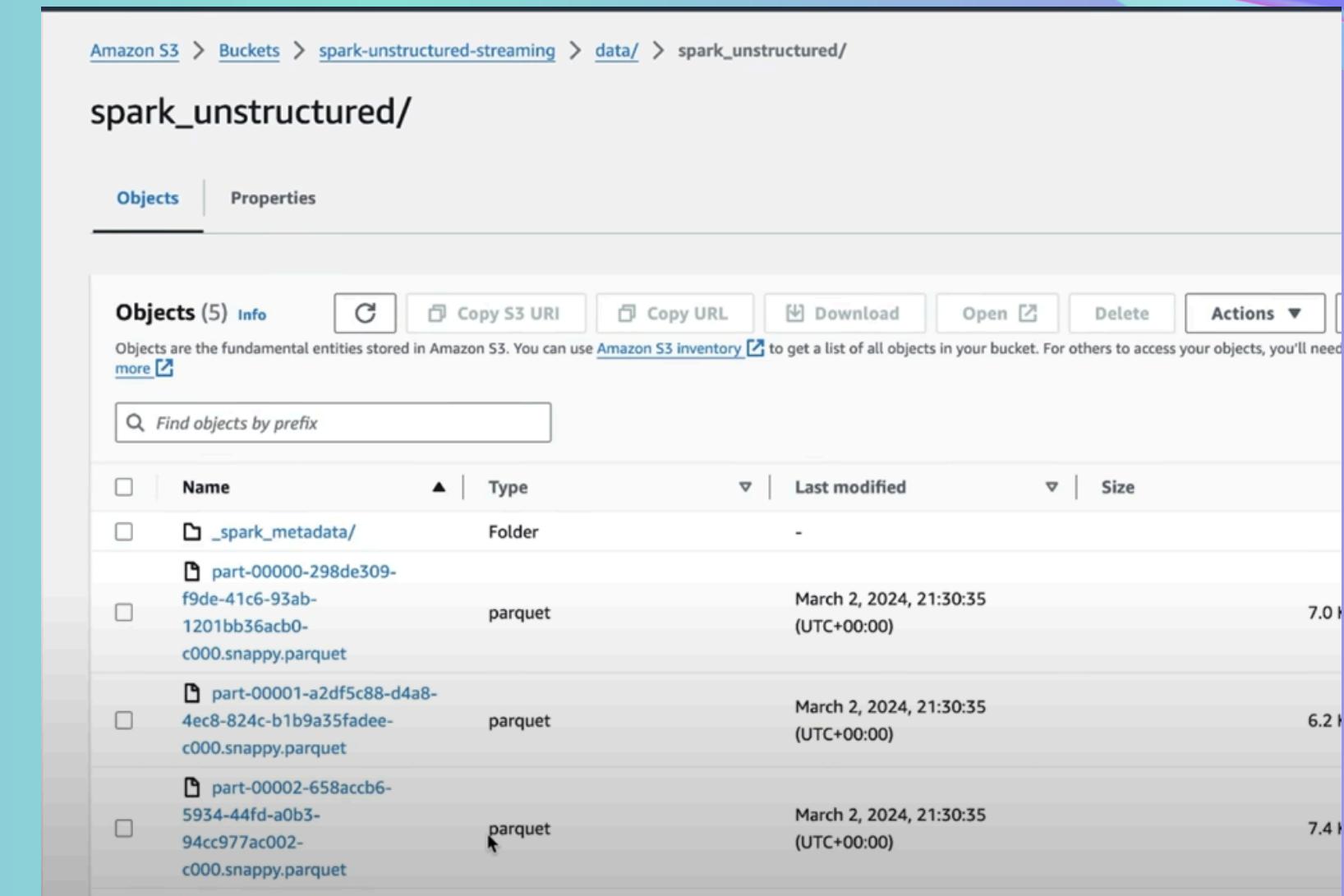
[C](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#)

[Upload](#)

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](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	checkpoints/	Folder	-	-	-
<input type="checkbox"/>	data/	Folder	-	-	-



Amazon S3 > Buckets > spark-unstructured-streaming > data/ > spark_unstructured/

spark_unstructured/

Objects (5) [Info](#)

[C](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

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](#)

<input type="checkbox"/>	Name	Type	Last modified	Size
<input type="checkbox"/>	_spark_metadata/	Folder	-	-
<input type="checkbox"/>	part-00000-298de309-f9de-41c6-93ab-1201bb36acb0-c000.snappy.parquet	parquet	March 2, 2024, 21:30:35 (UTC+00:00)	7.0 K
<input type="checkbox"/>	part-00001-a2df5c88-d4a8-4ec8-824c-b1b9a35fadee-c000.snappy.parquet	parquet	March 2, 2024, 21:30:35 (UTC+00:00)	6.2 K
<input type="checkbox"/>	part-00002-658accb6-5934-44fd-a0b3-94cc977ac002-c000.snappy.parquet	parquet	March 2, 2024, 21:30:35 (UTC+00:00)	7.4 K

AWS INTEGRATION

The screenshot shows the AWS Glue Data Catalog Tables interface. The left sidebar includes sections for Getting started, ETL jobs, Data Catalog tables (selected), Databases, Stream schema registries, Data Integration and ETL, and ETL jobs. The main area displays a table titled "Tables (2)". The table has columns for Name, Database, Location, Classification, Deprecated, View data, and Data quality. It lists two entries: "spark_unstructured" (smartcity, s3://spark-streaming-da, Parquet) and "tbl_spark_unstructured" (spark_unstructured_db, s3://spark-unstructured, Parquet). A large white arrow points from the top right towards this interface.

The screenshot shows the AWS Glue Crawler configuration page for "spark_unstructured_crawler". The top bar includes "Last updated (UTC)" (March 2, 2024 at 21:38:31), "Run crawler", "Edit", and "Delete" buttons. The "Crawler properties" section shows the crawler's name, IAM role (AWSGlueServiceRole-Spark-Unstructured), database (spark_unstructured_db), state (READY), and table prefix (tbl_). The "Crawler runs" tab is selected, showing one run from March 2, 2024 at 21:38:43, which is currently running. A large white arrow points from the top right towards this interface.

The screenshot shows the AWS Glue Crawler results page. The top bar includes "Copy" and "Download results" buttons. The main area is titled "Results (10)" and contains a table with 10 rows of data. The columns are labeled #, file_name, start_date, end_date, salary_start, salary_end, classcode, and req. The data includes various job titles and their corresponding details. A large white arrow points from the bottom right towards this interface.

#	file_name	start_date	end_date	salary_start	salary_end	classcode	req
1	SENIOR PHOTOGRAPHER	2016-04-15	2016-05-05	60969.0	94106.0	1795	
2	SENIOR UTILITY SERVICES SPECIALIST	2018-11-30	2018-12-13	117199.0	145596.0	3753	Two years of
3	PARKING MANAGER	2014-02-07	2014-02-20	73936.0	106300.0	9170	
4	FIREARMS EXAMINER	2016-06-24		81160.0	118661.0	2233	
5	COMMERCIAL FIELD SUPERVISOR	2016-12-16	2016-12-29	86025.0	106884.0	1603	1. Two years
6	DIRECTOR OF MAINTENANCE AIRPORTS	2016-04-15	2016-04-28	108576.0	171654.0	7270	
7	AQUARIST	2014-05-02	2014-05-15	55624.0	76817.0	2400	
8	SENIOR TRANSPORTATION ENGINEER	2017-10-20	2017-11-02	105444.0	149981.0	9262	1. Two years

Benefits

- Real-time Insights
- Improved Decision-Making
- Handles Variety of Data
- Unified Data Processing
- Faster Data Access
- Scalable Processing
- Reliable Storage
- Easy Querying
- Simplified Data Management
- Streamlined Connectivity
- Centralized Configuration
- Automated Data Extraction
- Real-Time Data Updates
- Scalable and Real-Time Processing

Conclusion

This project—the AWS Spark Unstructured Data Streaming project—demonstrates an end-to-end solution for processing and analyzing unstructured data in real time using Apache Spark integrated with AWS. The system ingests data from diverse sources (text and JSON), applies custom UDFs to parse and extract critical fields from the unstructured input, and unifies the data streams for further processing. The processed output is then stored persistently (e.g., in AWS S3 as parquet files) for downstream analytics.

Throughout the project, key challenges such as dependency management, environment configuration, and the handling of unstructured data formats were addressed. Overall, the project showcases how Spark Streaming, in combination with AWS, can efficiently handle complex, real-time data pipelines in an unstructured data environment.





**THANK
YOU**

SAHIL PATIL