Apache Spark Analytics Workbench

with JupyterLab and Spark Connect by Code Creator on Ubuntu 24.04

User Friendly Setup and Usage Instructions

Access Method	Use the instance public IP address. A domain name is not required.
Open Ports	22 for SSH and 80 for the web interface.

Quick Start

1. Launch the EC2 instance and keep it running for at least **about 2 to 3 minutes on first boot.**
2. Confirm that your AWS security group allows inbound access on ports 22 and 80.
3. SSH into the server using your key pair.

ssh -i your-key.pem ubuntu@YOUR-PUBLIC-IP

4. Once logged in, display the first boot access details:

cat /home/ubuntu/FIRST_LOGIN.txt

*This file shows your public URL, **your JupyterLab token,** and optional SSH tunnel commands for Spark internal interfaces.*

Open the Web Interface

Open your browser and go to your instance public IP address:

http://YOUR-PUBLIC-IP

When prompted, enter the JupyterLab token shown in **FIRST_LOGIN.txt**.

What Starts Automatically

- JupyterLab
- Apache Spark Master
- Apache Spark Worker
- Spark Connect
- Spark History Server

Use Spark in JupyterLab

Spark Connect is already configured. In JupyterLab, create a notebook and run the following test:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
print(spark.range(10).count())
```

If everything is working correctly, the output should be **10**.

Test 2

```
cd /opt/apache-spark-stack

sudo docker exec -it spark-master /opt/spark/bin/spark-submit \
  --master spark://spark-master:7077 \
  /opt/spark/examples/src/main/python/pi.py 10
```

The job should complete successfully.

Optional Spark Interfaces Through SSH Tunnels

The Spark internal interfaces are intentionally kept private. Use SSH tunnels from your local computer to reach them safely.

Interface	SSH Tunnel Command	Open In Browser
Spark Master UI	ssh -i your-key.pem -L 8080:127.0.0.1:8080 ubuntu@YOUR-PUBLIC-IP	http://127.0.0.1:8080
Spark Worker UI	ssh -i your-key.pem -L 8081:127.0.0.1:8081 ubuntu@YOUR-PUBLIC-IP	http://127.0.0.1:8081
History Server	ssh -i your-key.pem -L 18080:127.0.0.1:18080 ubuntu@YOUR-PUBLIC-IP	http://127.0.0.1:18080
Spark Connect	ssh -i your-key.pem -L 15002:127.0.0.1:15002 ubuntu@YOUR-PUBLIC-IP	Client tools connect to 127.0.0.1:15002

Manage the Services

Check container status:

```
cd /opt/apache-spark-stack && sudo docker compose ps
```

Start the services:

```
cd /opt/apache-spark-stack && sudo docker compose up -d
```

Stop the services:

```
cd /opt/apache-spark-stack && sudo docker compose down
```

Important Notes

- Use the instance public IP address for access. A domain name is not required.
- Allow a few minutes on first boot for the services to initialize.
- JupyterLab is the main public interface for this product.
- Spark internal interfaces are intended to be accessed through SSH tunnels.