

Data Science Product Development with AWS and Python SDKs



Building Cloud Native DS product for Internal Businesses

Sheikh Alam,
Data Science Product Developer,
@Novartis, CZ

Dec, 13, 2022 @Ataccama s.r.o
By Prague AWS user group

Product for Internal Businesses?

Nature

Products are not customer facing
Will be used by internal teams (Marketing, HR, Finance, Businesses)
Self service focused
Small User base but high value product

Criteria

Low developer cost
Small User base
Visual appeal is negotiable

Costing Model

| Role | Full Dev. Team | Only DS Team |
|-----------------------|----------------|--------------|
| PM (Data Science) | 5k | 5k |
| Data Scientist (2x) | 10K | 10k |
| API backend team (2x) | 8k | |
| Frontend UI Team (2x) | 7k | |
| PM (Full Project) | 5k | 5k |
| Cloud Ops | 5k | 5k |

40k vs 25k

Product Demo & Code Repo



Your Translation Buddy

The app can be used to translate English document into German.

Submission Guideline:

- First you have to create a project using the **Create Project** section
- Then you can navigate to the **Add Document** tab and submit the document by selecting the project you have created
- Next, you can navigate to the **Start Translation** section and initiate the translation process

Other Tabs:

- The status of the translation can be viewed in the **Job Status** tab
- History about your project can be viewed in the **History** tab

<https://github.com/DataPsycho/cloud-native-datascience>

Question about Model/Model Registry?

The topic I am not covering:

Which model we are using?

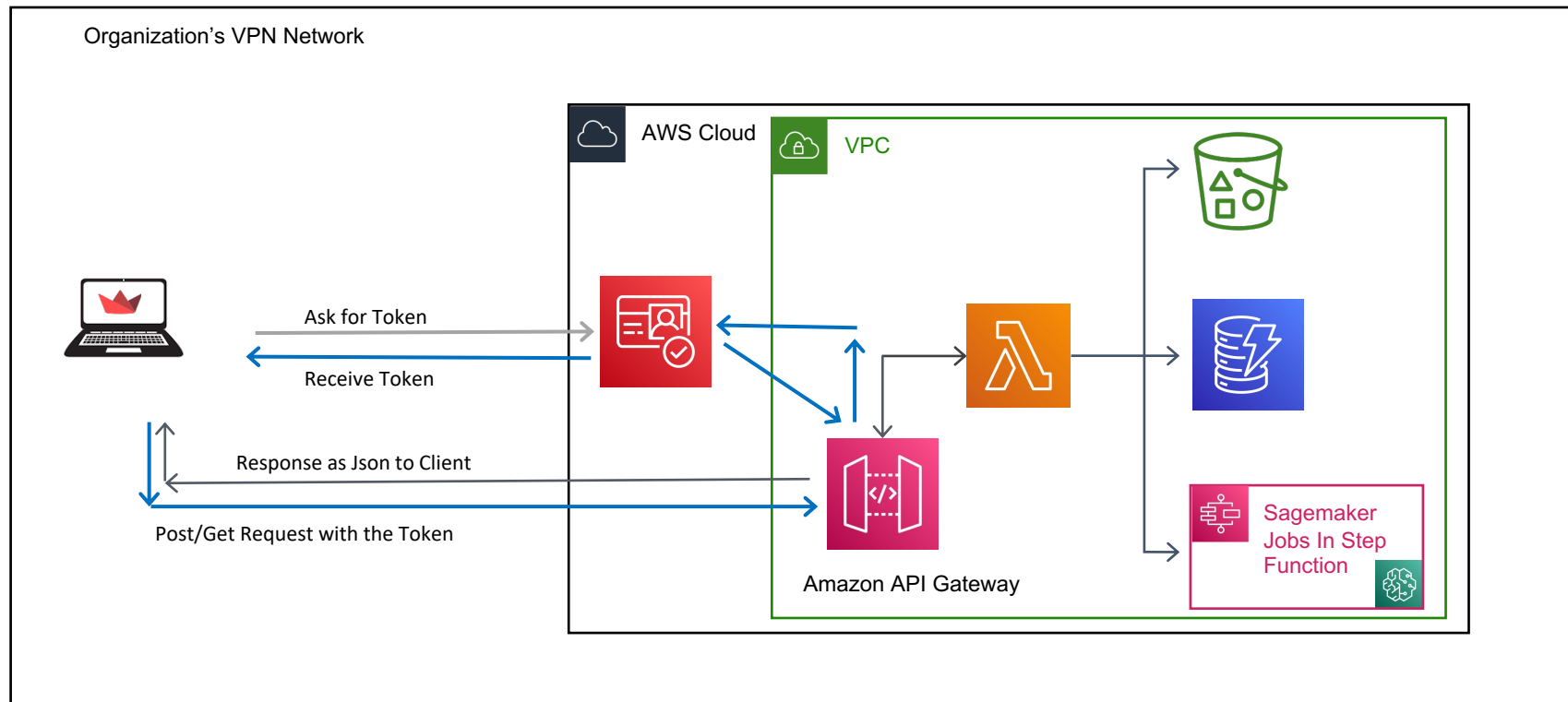
How frequent we are re training the model?

Why not using 3rd party translation service like Deepl?



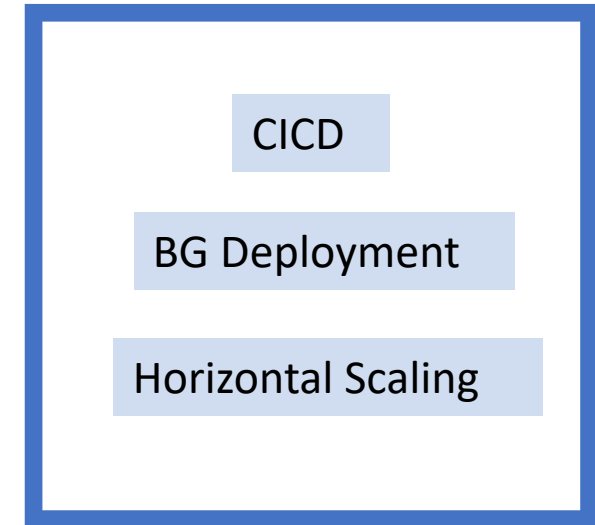
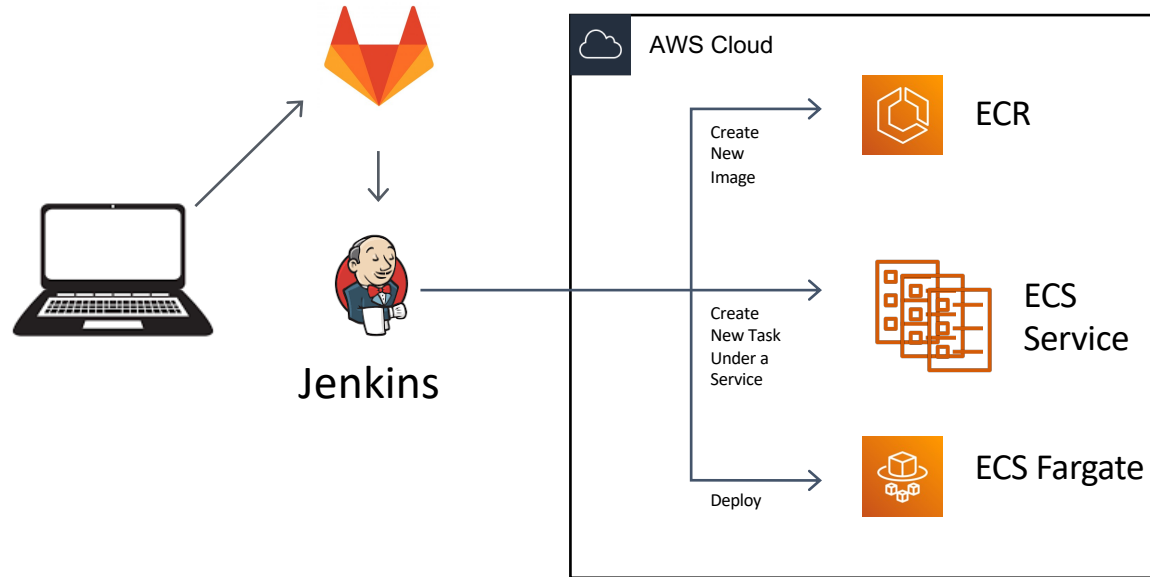
Natalia Klyueva,
Data Scientist @novartis

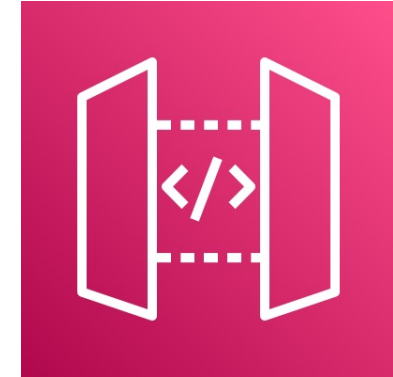
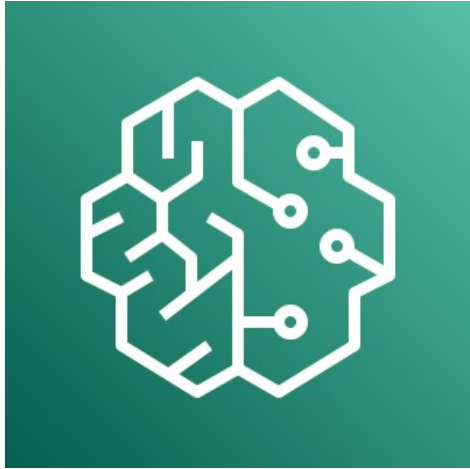
Server Side & Client Side Integration



💡 *API Gateway must need Cognito to authenticate each request, when using Client facing api for the Business user*

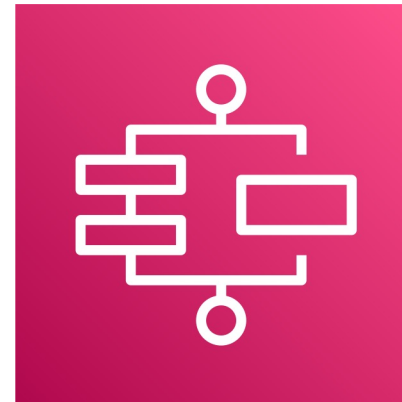
Client Side: Hosting Streamlit with ECS Fargate



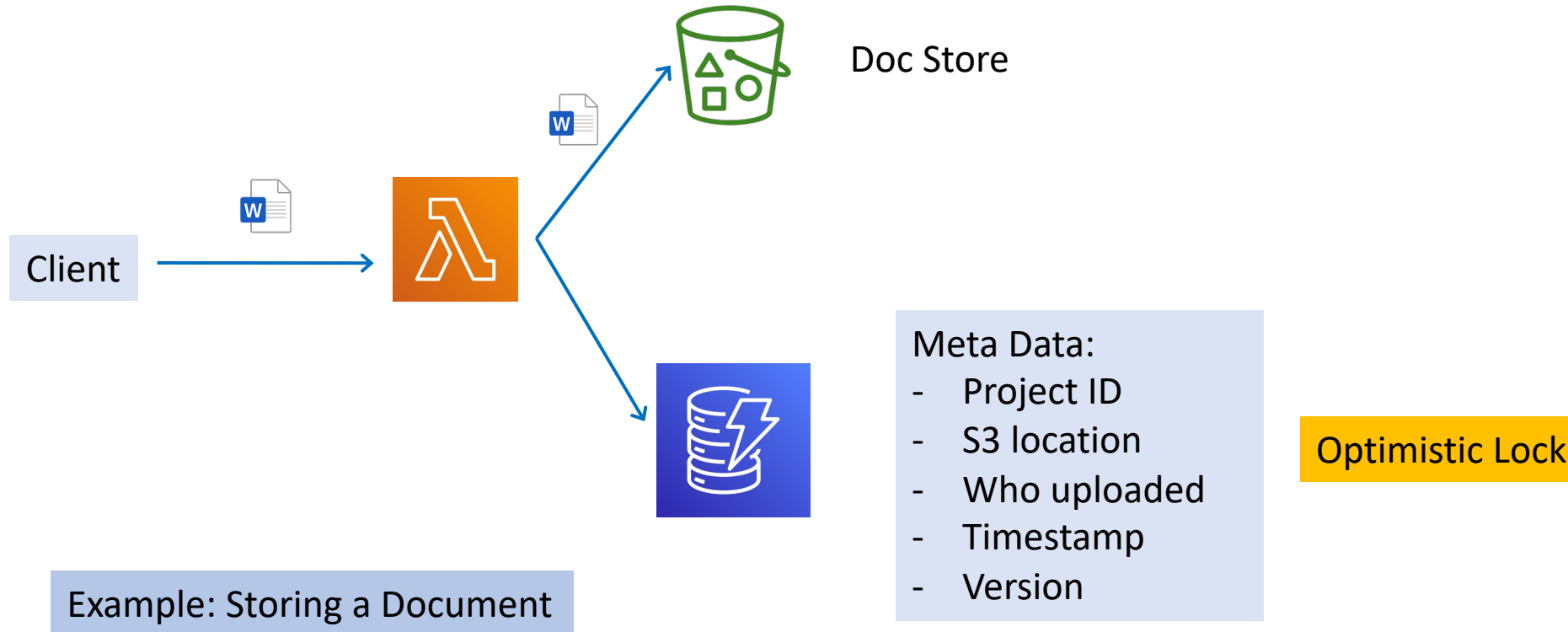


Dive Into AWS

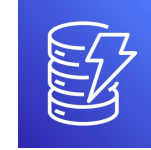
Roles and Policies



Metadata Management with Lambda & DynamoDB (Serverside)



Single Table Design with DynamoDB

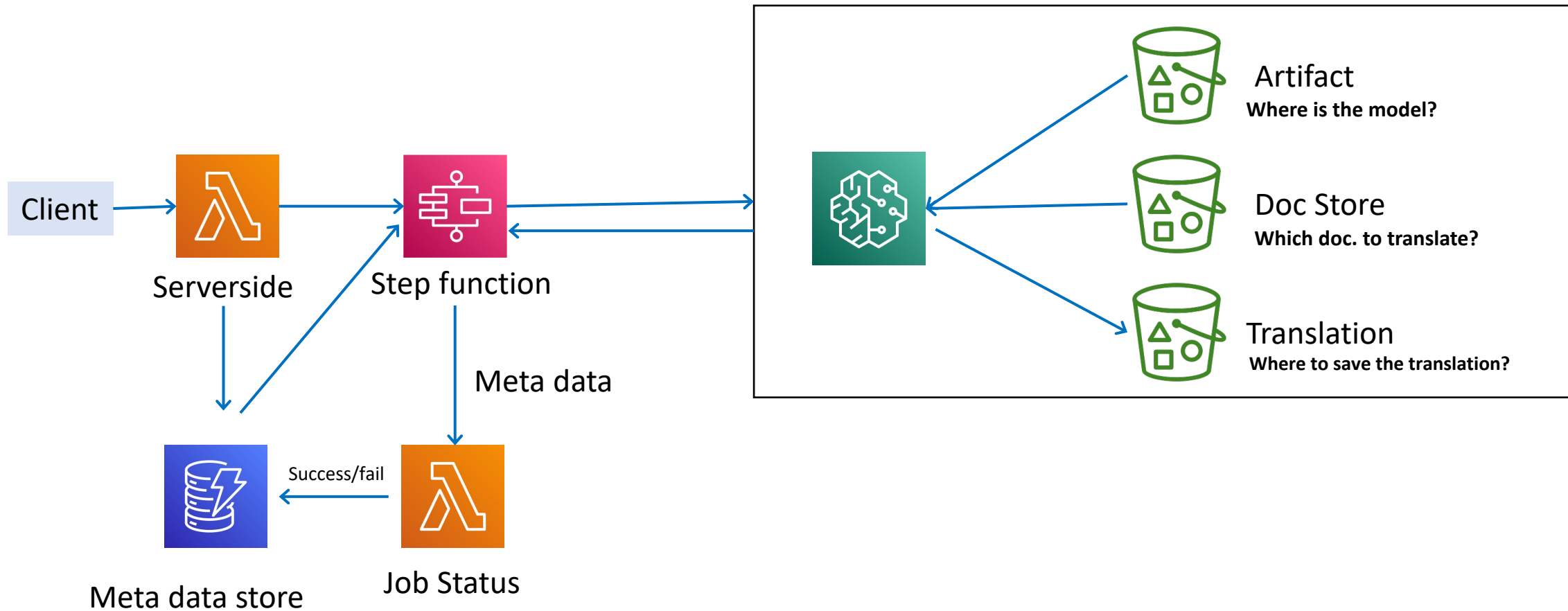


| | | | | | | | | |
|--------------------------|--------------|---|------------------------|------|----------------|----------------|---------------|---------|
| <input type="checkbox"/> | proj#project | | test-01-2022-12-07T... | true | quack-trans... | docstore/do... | demo.docx | project |
| <input type="checkbox"/> | proj#project | | test-02-2022-12-12T... | true | quack-trans... | docstore/do... | english_de... | project |
| <input type="checkbox"/> | proj#project |   | test-03-2022-12-12T... | true | null | null | null | project |
| <input type="checkbox"/> | proj#project | | test-04-2022-12-12T... | true | quack-trans... | docstore/do... | english_de... | project |

| <input type="checkbox"/> | PK | SK | active | bucket | bucket_key | document | entity_type | finished_at | jid |
|--------------------------|-------------------------|----------|--------|----------------|------------------|----------|-------------|----------------|---------------|
| <input type="checkbox"/> | artf#artifact | job#job1 | | quack-trans... | artifacts/job... | | job | | |
| <input type="checkbox"/> | artf#artifact | model | | quack-trans... | models/v0.0... | | model | | |
| <input type="checkbox"/> | proj#test-01-2022-12... | job1#v0 | | quack-trans... | job/test-01-... | | job1 | 2022-12-12T... | 88c52bc4-b... |
| <input type="checkbox"/> | proj#test-01-2022-12... | job1#v1 | | quack-trans... | job/test-01-... | | job1 | 2022-12-10T... | f927027f-f... |
| <input type="checkbox"/> | proj#test-01-2022-12... | job1#v2 | | quack-trans... | job/test-01-... | | job1 | 2022-12-10T... | f2872184-a... |
| <input type="checkbox"/> | proj#test-01-2022-12... | job1#v3 | | quack-trans... | job/test-01-... | | job1 | 2022-12-12T... | bbaab119-... |

```
AttributeDefinitions:
  - AttributeName: PK
    AttributeType: S
  - AttributeName: SK
    AttributeType: S
  - AttributeName: parent_entity_pid
    AttributeType: S
  - AttributeName: status_jid
    AttributeType: S
  - AttributeName: parent_entity_type
    AttributeType: S
  - AttributeName: entity_type
    AttributeType: S
```

Batch Translation with Sagemaker & Stepfunction (ML)



The unwanted Emails

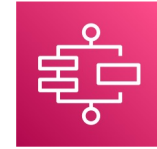
Initial Notification on Planned Downtime for AIRFLOW [REDACTED]

Dear Colleagues,

We would like to inform you that Unix team has planned a scheduled downtime for [REDACTED] vers patching as part of the security patch cycle, so services will not be available on 17-December-2022 during patching window.

| | |
|--|----------------------|
| Application(s) Impacted [REDACTED] | [REDACTED] |
| Applications not Impacted | N/A |
| Downtime Description | [REDACTED] |
| Impacted Sites | [REDACTED] |
| Downtime Start Time | 17-12-2022 07:00 CET |
| Downtime End Time | 17-12-2022 21:00 CET |
| Duration | 14 Hours |
| Business Impact | [REDACTED] |

Step Function (The Airflow Killer)



```
import stepfunctions
import logging

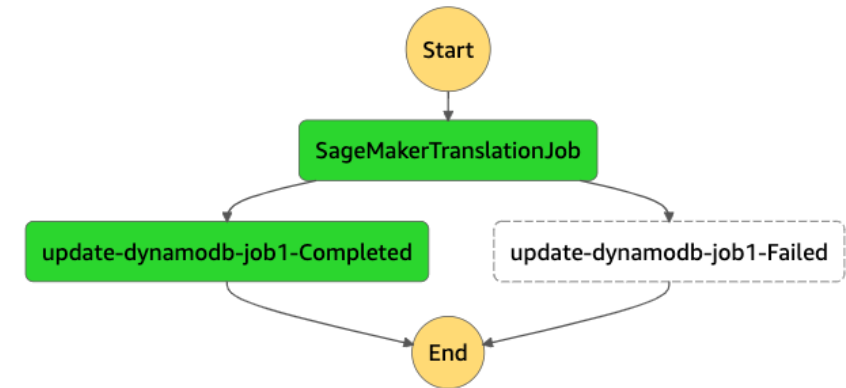
from stepfunctions.steps import *
from stepfunctions.workflow import Workflow

stepfunctions.set_stream_logger(level=logging.INFO)

workflow_execution_role = "<execution-role-arn>" #
```

```
# Next, we define the workflow
basic_workflow = Workflow(
    name="MyWorkflow_Simple", definition=basic_path, role=workflow_execution_role
)

# Render the workflow
basic_workflow.render_graph()
```



* Example from official Git Repo

Monitoring with CloudWatch Insights



```
▶ 2022-12-12T18:10:48.583+01:00 2022-12-12 17:10:47 [INFO] src.logger - Model loaded from local path successfully.
▶ 2022-12-12T18:10:48.583+01:00 2022-12-12 17:10:47 [INFO] src.logger - File read successfully from /opt/ml/proces
▶ 2022-12-12T18:10:48.583+01:00 /miniconda3/lib/python3.7/site-packages/transformers/generation_utils.py:1364: Use
▶ 2022-12-12T18:11:35.065+01:00 2022-12-12 17:11:34 [INFO] src.logger - File created successfully to /opt/ml/proce
▼ 2022-12-12T18:11:35.065+01:00 2022-12-12 17:11:34 [INFO] __main__ - {'execution_time': 0.948}
2022-12-12 17:11:34 [INFO] __main__ - {'execution_time': 0.948}
```

1) Actual Log

| # | bin(10m) | exec_time |
|-----|-------------------------------|-----------|
| ▶ 1 | 2022-12-10T14:20:00.000+01:00 | 0.944 |
| ▶ 2 | 2022-12-12T18:10:00.000+01:00 | 0.948 |
| ▶ 3 | 2022-12-09T15:50:00.000+01:00 | 1.095 |
| ▶ 4 | 2022-12-12T18:00:00.000+01:00 | 0.972 |
| ▶ 5 | 2022-12-09T15:30:00.000+01:00 | 1.367 |
| ▶ 6 | 2022-12-12T17:40:00.000+01:00 | 0.92 |
| ▶ 7 | 2022-12-09T16:10:00.000+01:00 | 0.957 |
| ▶ 8 | 2022-12-12T16:10:00.000+01:00 | 0.982 |

3) Aggregated Log

/aws/sagemaker/ProcessingJobs X

```
1 fields @timestamp, @message
2 | filter @message like "INFO"
3 | parse @message "*" [*] * - {"execution_time": *} as @dt, @level, @thred, @execution_time
4 | filter @thred = "__main__"
5 | stats avg(@execution_time) as exec_time
6 by bin(10m)
7 | sort @timestamp desc
8 | limit 20
```

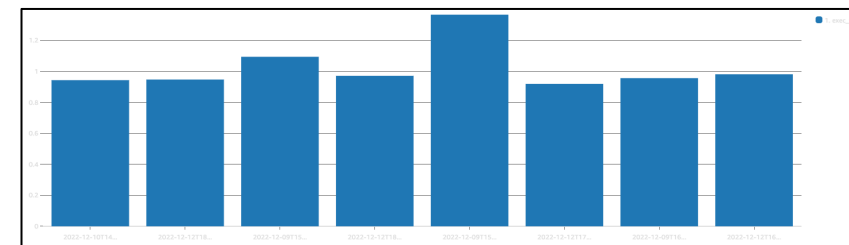
Run query

Cancel

Save

History

2) Query To transform the log



4) Chart from Log

UI with Streamlit

When to Use?

Quick Development

Single Analytics Projects

Simple Application

When not to?

Complex Application

Multipage Application

SAML, SSO Authentication

Building Analytics Platform

Few Pain Points

AWS Roles and Policies

Debugging Integration of Lambda & Stepfunction

Authentication with Streamlit

Network Load Balancer setup for VPC

It's Me

From Pandas to PySpark DataFrame

39 Lessons 54 Playgrounds 27 Illustrations

Takeaway Skills

- ✓ A working knowledge of Apache Spark and the PySpark library for Python
- ✓ A strong understanding of the advantages of using PySpark instead of Pandas for processing large datasets
- ✓ The ability to calculate some Metrics or produce aggregated analytics reporting solutions
- ✓ The ability to write Production Code in PySpark

Course at Educative



MrDataPsycho

55 Followers

Data Science | Dev | Author @ Learnpub,
Educative: Pandas to Pyspark

[Edit profile](#)

Medium

BOOKS

#1



From Pandas to PySpark DataFrame

MrDataPsycho

You already know Pandas? But it is not enough, for started quickly with PySpark DataFrame API if you :

Learnpub

Thanks You !
ধন্যবাদ !



linkedin.com/in/mr-data-psycho/



@MrDataPsycho