Data Science Product Development with AWS and Python

Building Clund 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 (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

🦆 **Quack** 🦆🐾

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

Quastion about Model/Model Registry?

The topic I am not covering:
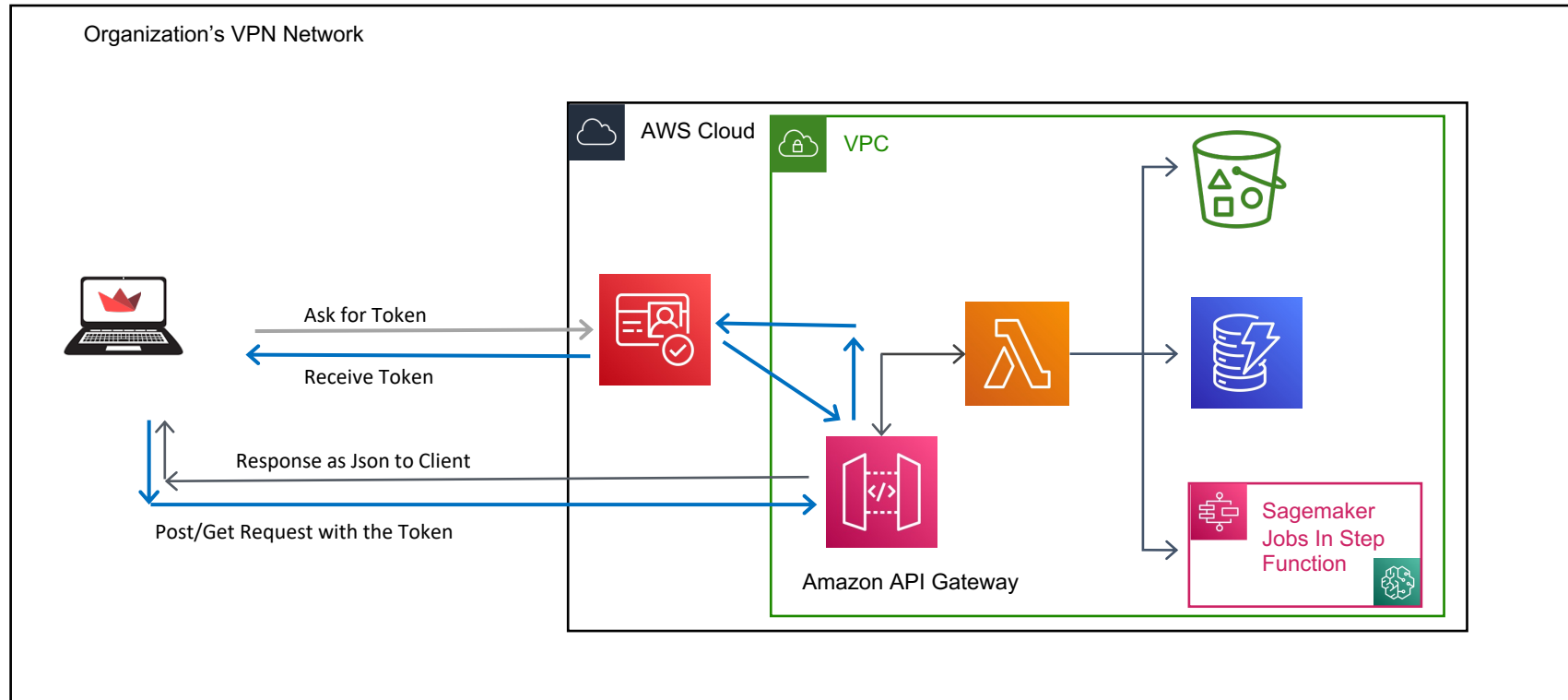
Which model we are using?
How frequent we are re training the model?
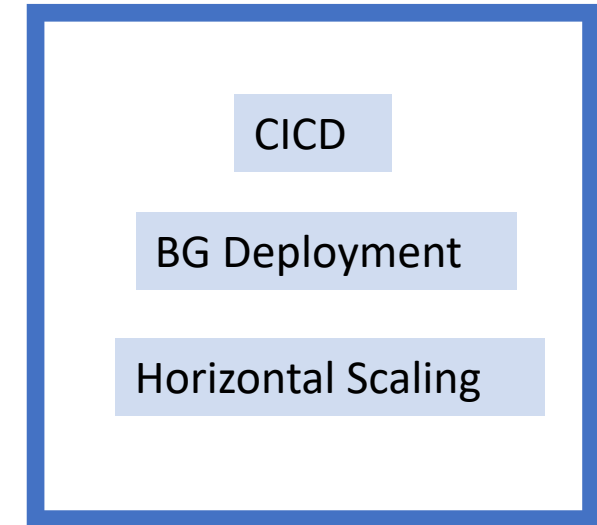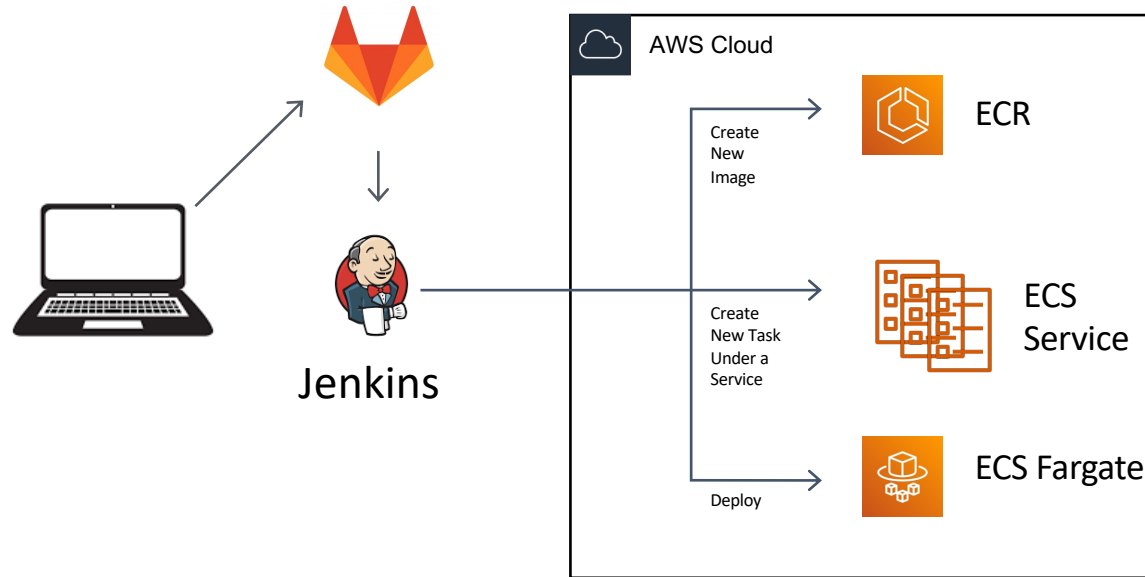Why not using 3rd party translation service like Deepl?
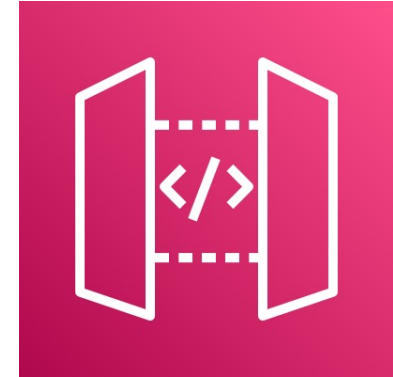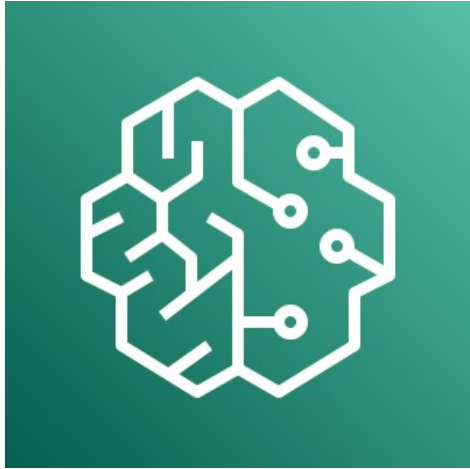
Natalia Klyueva,
Data Scientist @novartis

Server Side & Client Side Integration

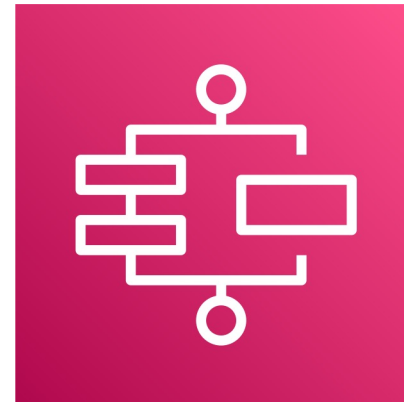API Gateway must need Cognito to authenticate each request, when using Client facing api for the Business user

Client Side: Hosting Streamlit with ECS Fargate

Dive Into AWS

Roles and Policies

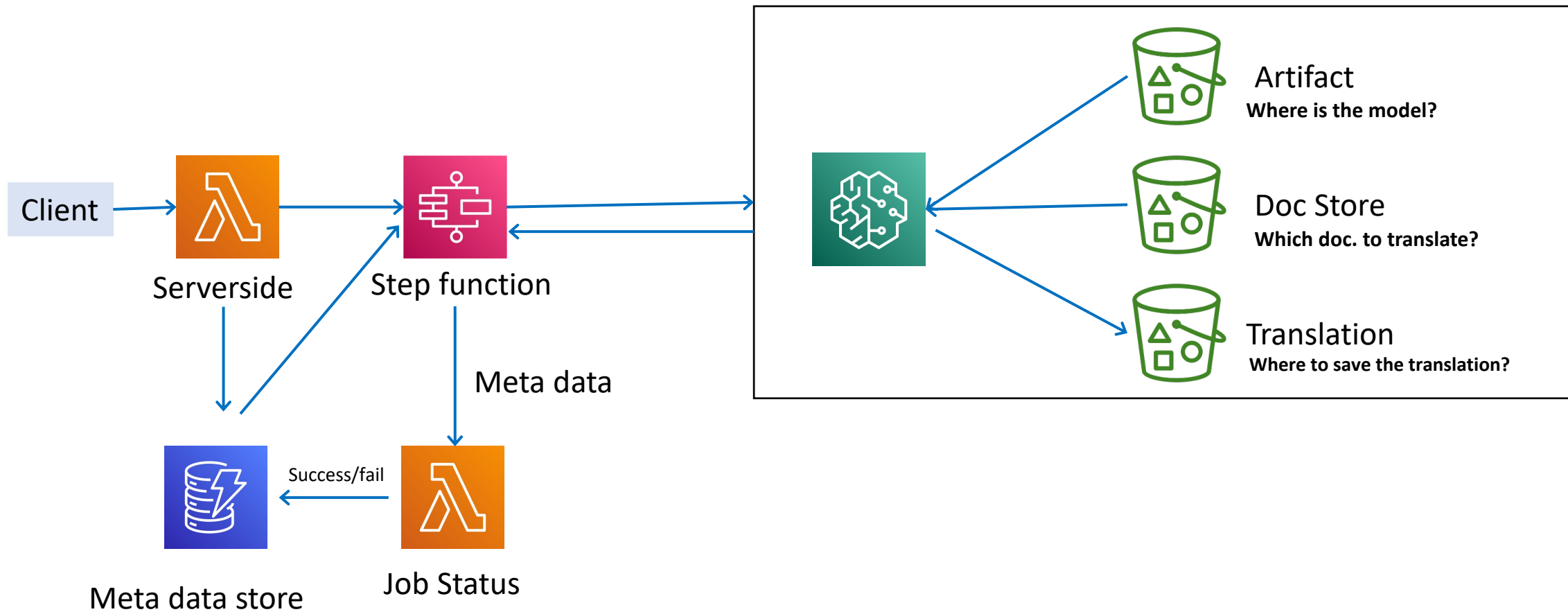# Metadata Management with Lambda & DynamoDB (Serverside)

Doc Store

Client

Meta Data:
- Project ID
- S3 location
- Who uploaded
- Timestamp
- Version

Optimistic Lock

Example: Storing a Document
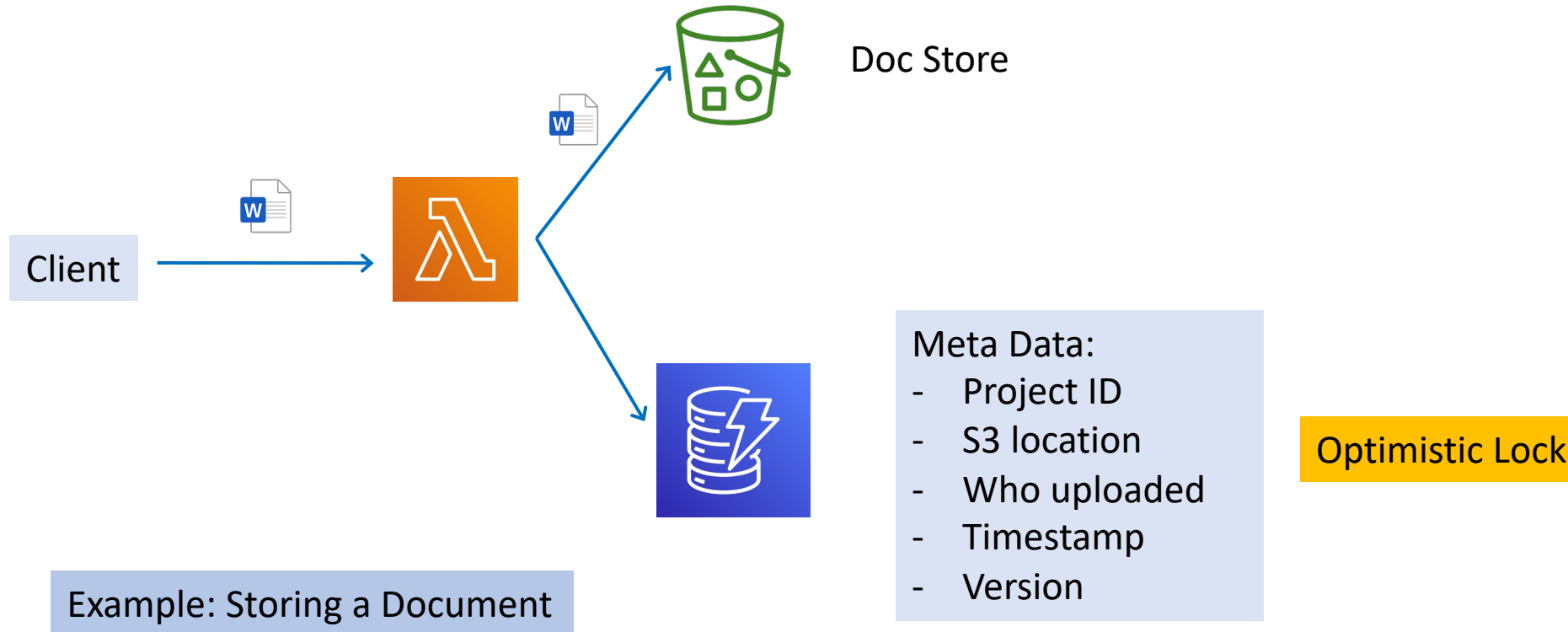
Batch Translation with Sagemaker & Stepfunction (ML)

The Wall of Texts

Initial Notification on Planned Downtime for AIRFLOW ▇▇▇▇▇▇

Dear Colleagues,

We would like to inform you that Unix team has planned a scheduled downtime for ▇▇▇▇▇▇▇▇ 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	▇▇▇▇▇▇▇▇▇▇
Applications not Impacted	N/A
Downtime Description	▇▇▇▇▇▇▇▇▇▇
Impacted Sites	▇▇▇▇▇▇▇
Downtime Start Time	17-12-2022 07:00 CET
Downtime End Time	17-12-2022 21:00 CET
Duration	14 Hours
Business Impact	▇▇▇▇▇▇▇▇

Step Function (The Airflow Killer)

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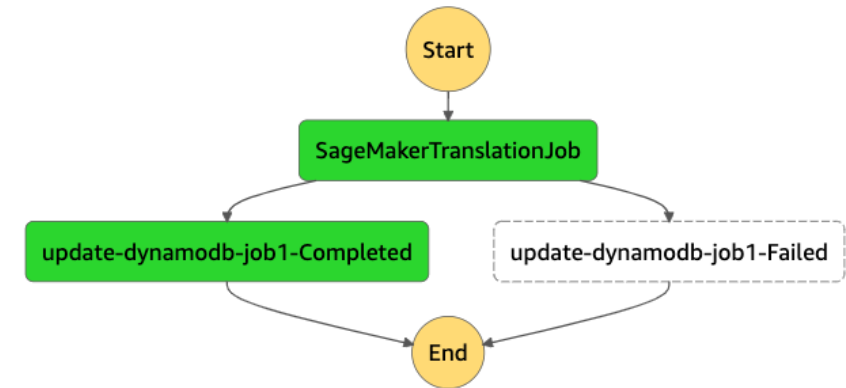
```python
# Next, we define the workflow
basic_workflow = Workflow(
    name="MyWorkflow_Simple", definition=basic_path, role=workflow_execution_role
)

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

Example from official Git Repo

UI with Streamlit

When to Use?

Quick Development

Single Alalytics 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 a

Learnpub

Thanks You !
ধন্যবাদ !

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

@MrDataPsycho