INNOVATE

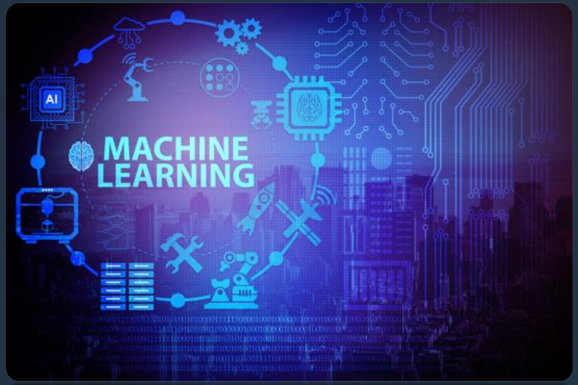
DATA AND AI/ML EDITION

Scalable data preparation & ML using Apache Spark on AWS

Suman Debnath

Principal Developer Advocate, Data Engineering
Amazon Web Services

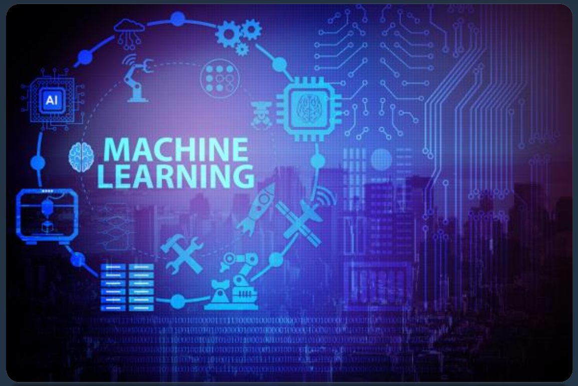
Personas



Personas

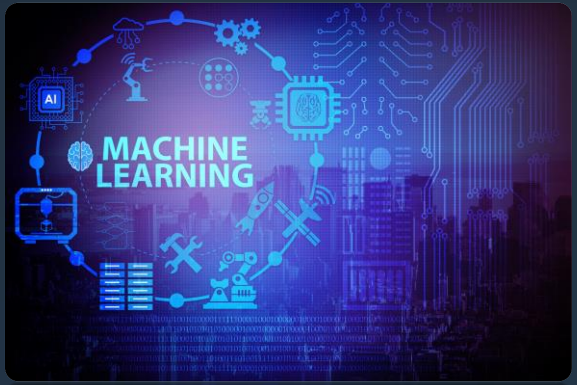
Data science

Python, R, TensorFlow, PyTorch



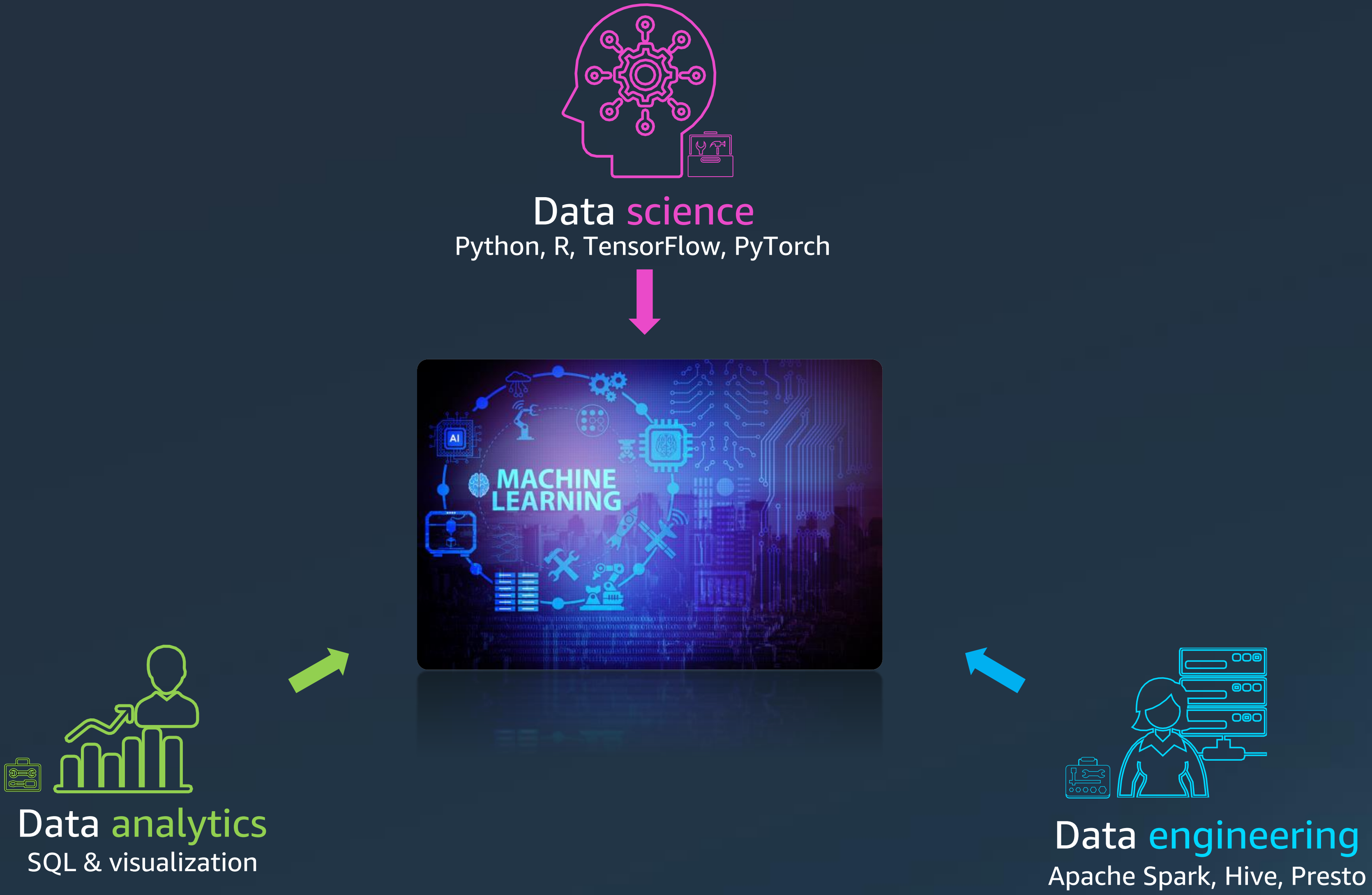
Personas

Data science
Python, R, TensorFlow, PyTorch



Data engineering
Apache Spark, Hive, Presto

Personas



Data science
Python, R, TensorFlow, PyTorch

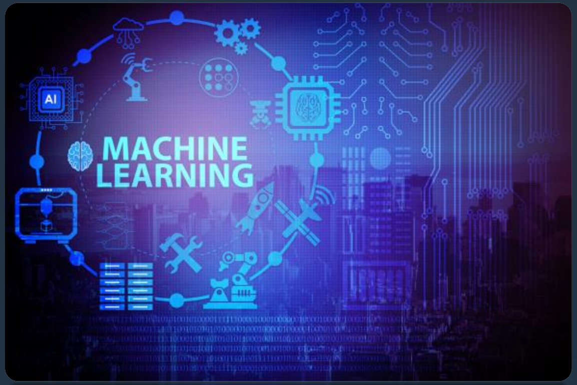
Data analytics
SQL & visualization

Data engineering
Apache Spark, Hive, Presto

Personas

Data science
Python, R, TensorFlow, PyTorch

Switching between multiple notebooks, tools, and interfaces reduces productivity



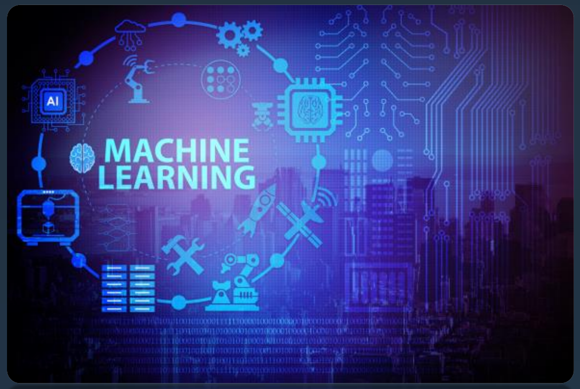
Data analytics
SQL & visualization

Data engineering
Apache Spark, Hive, Presto

Personas

Data science
Python, R, TensorFlow, PyTorch

Switching between multiple notebooks, tools, and interfaces reduces productivity



Data preparation and analytics are foundational components of ML workflows

Data analytics
SQL & visualization

Data engineering
Apache Spark, Hive, Presto

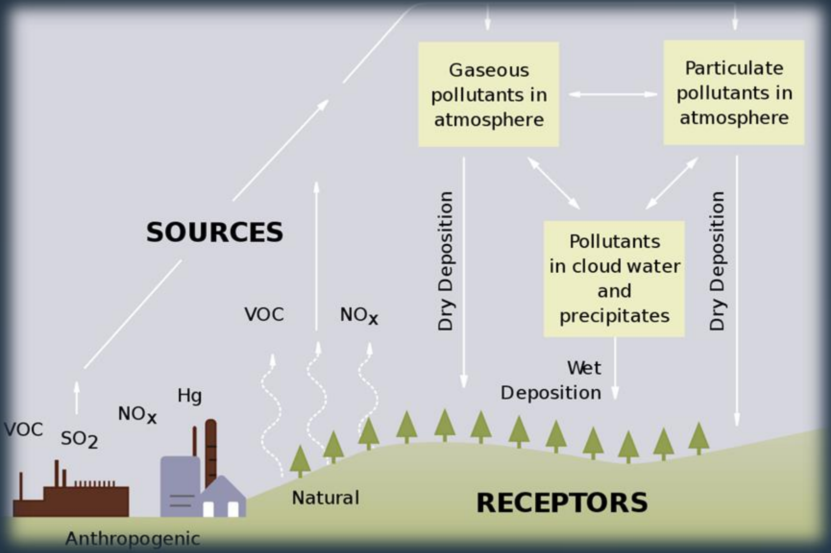
Challenge

BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

Challenge

BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

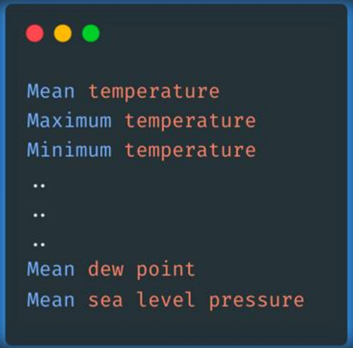
Build a machine learning model which can help to **predict the amount of NO_2** in the area **based on weather conditions**


Gaseous pollutants in atmosphere
Particulate pollutants in atmosphere
SOURCES
Dry Deposition
Pollutants in cloud water and precipitates
Dry Deposition
VOC
NOx
Wet Deposition
Hg
NOx
VOC
SO2
Natural
RECEPTORS
Anthropogenic

Challenge

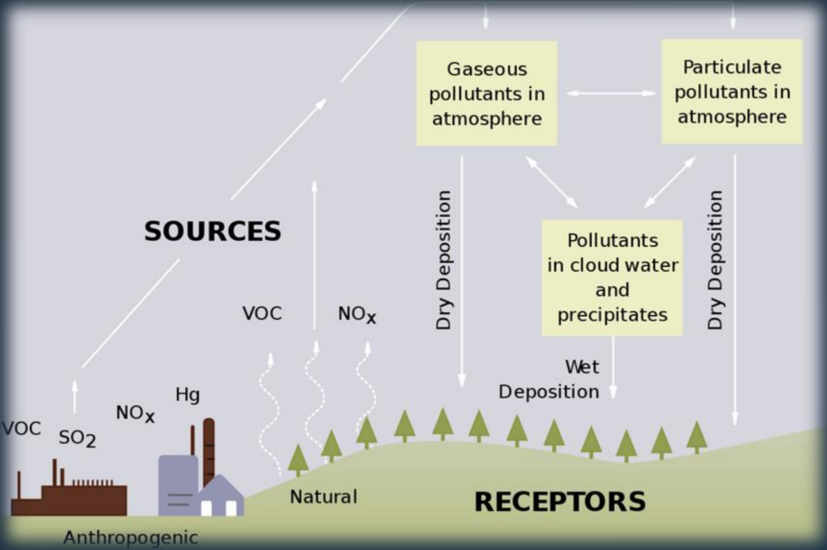
BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

Build a machine learning model which can help to **predict the amount of NO_2** in the area **based on weather conditions**



```
Mean temperature
Maximum temperature
Minimum temperature
..
..
..
Mean dew point
Mean sea level pressure
```

Input

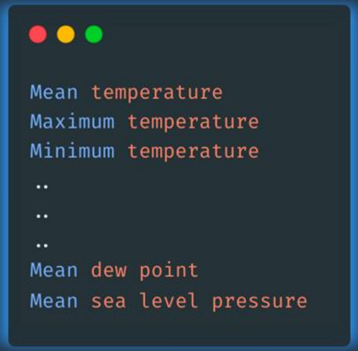


Gaseous pollutants in atmosphere
Particulate pollutants in atmosphere
SOURCES
Dry Deposition
Pollutants in cloud water and precipitates
Dry Deposition
VOC
NOx
Wet Deposition
Hg
NOx
VOC
SO2
Natural
RECEPTORS
Anthropogenic

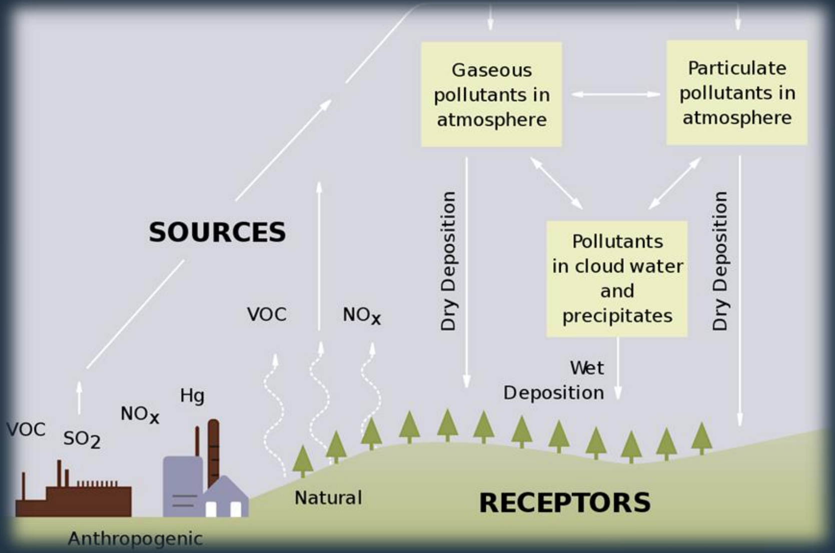
Challenge

BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

Build a machine learning model which can help to **predict the amount of NO_2** in the area **based on weather conditions**



```
Mean temperature
Maximum temperature
Minimum temperature
..
..
..
Mean dew point
Mean sea level pressure
```



Gaseous pollutants in atmosphere
Particulate pollutants in atmosphere
Dry Deposition
Pollutants in cloud water and precipitates
Dry Deposition
SOURCES
VOC
NOx
Wet Deposition
Hg
NOx
VOC
SO2
Natural
RECEPTORS
Anthropogenic

```
no2_avg
```

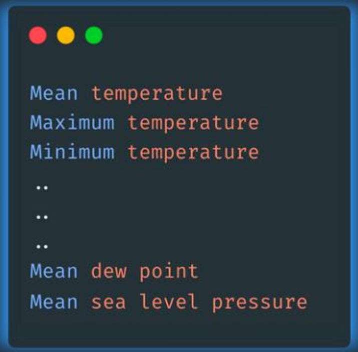
Input

Prediction

Challenge

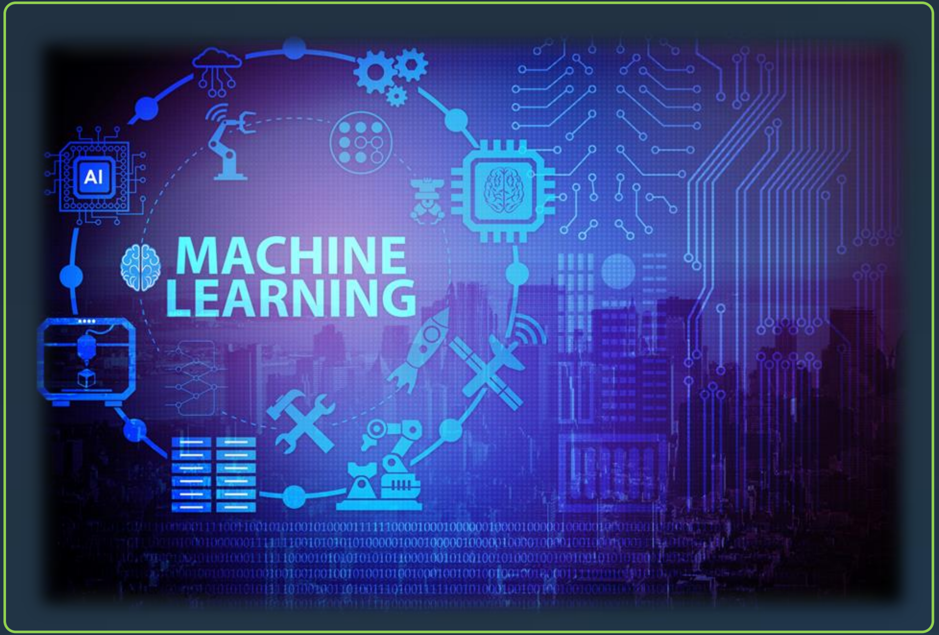
BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

Build a machine learning model which can help to **predict the amount of NO_2** in the area **based on weather conditions**

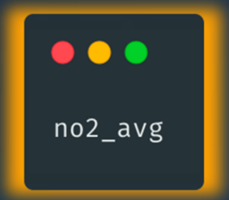


```
Mean temperature
Maximum temperature
Minimum temperature
..
..
..
Mean dew point
Mean sea level pressure
```

Input



We need to ***build this machine learning*** **model**



```
no2_avg
```

Prediction

Our tasks

BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

Our tasks

BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

Clean and prepare data using Apache Spark for use in machine learning

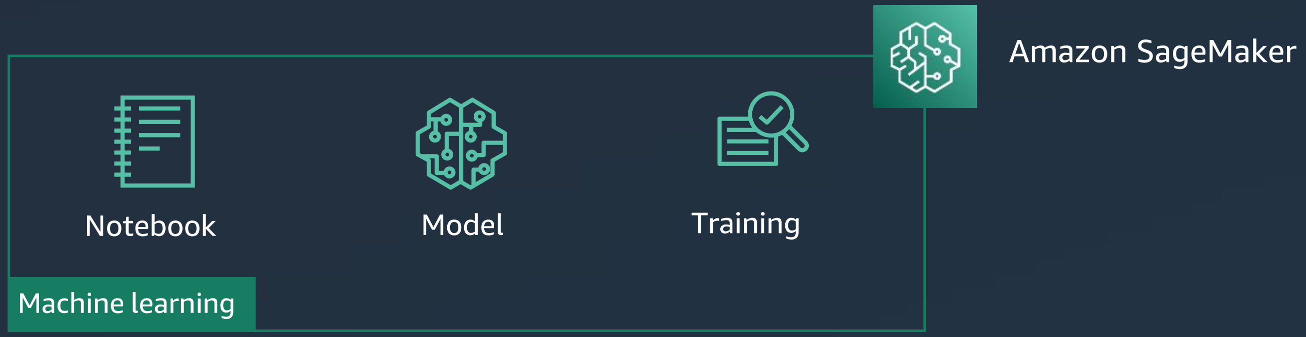
Our tasks

BUILD END-TO-END DATA PREPARATION AND ML WORKFLOWS ON AMAZON SAGEMAKER STUDIO

Clean and prepare data using Apache Spark for use in machine learning

Train the ML model using SageMaker to predict the NO_2 level of air

Our solution



Amazon SageMaker

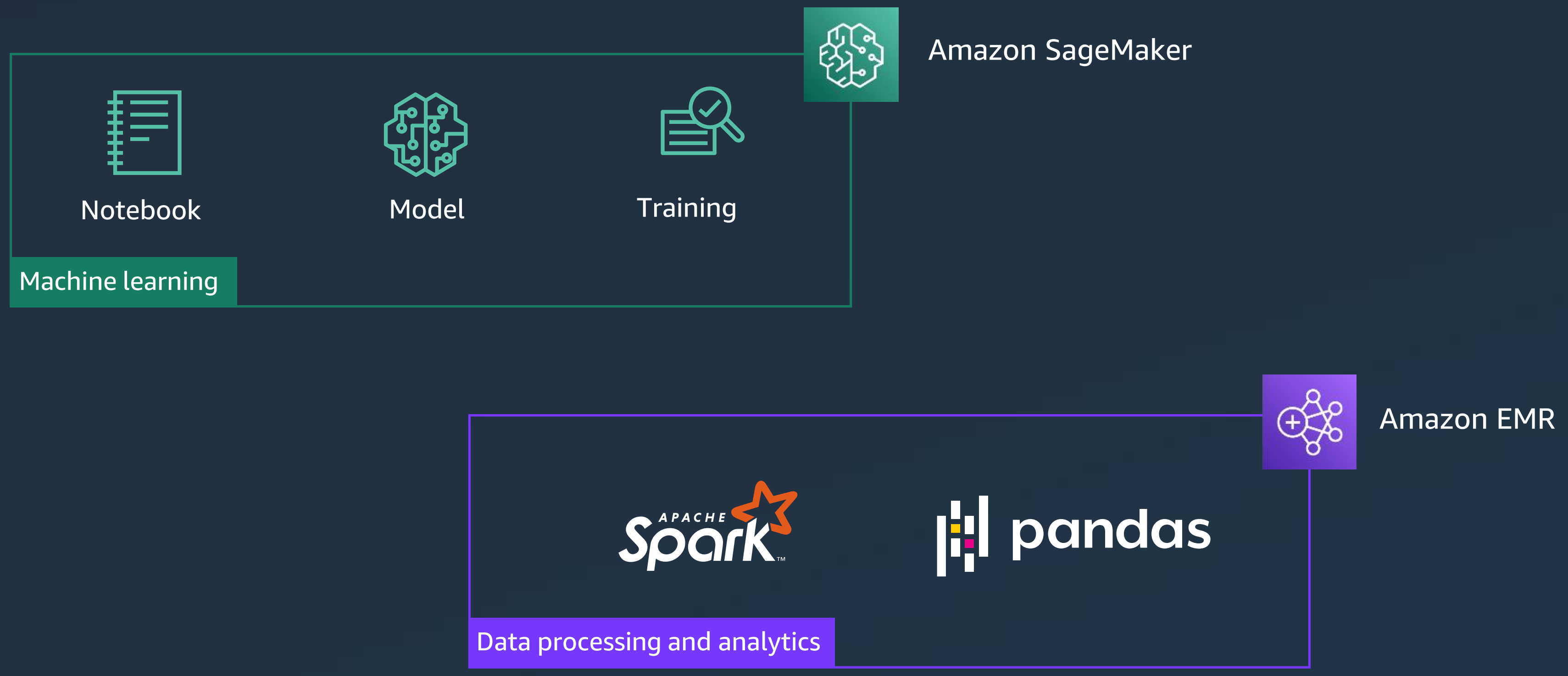
Notebook

Model

Training

Machine learning

Our solution



Amazon SageMaker

Notebook

Model

Training

Machine learning

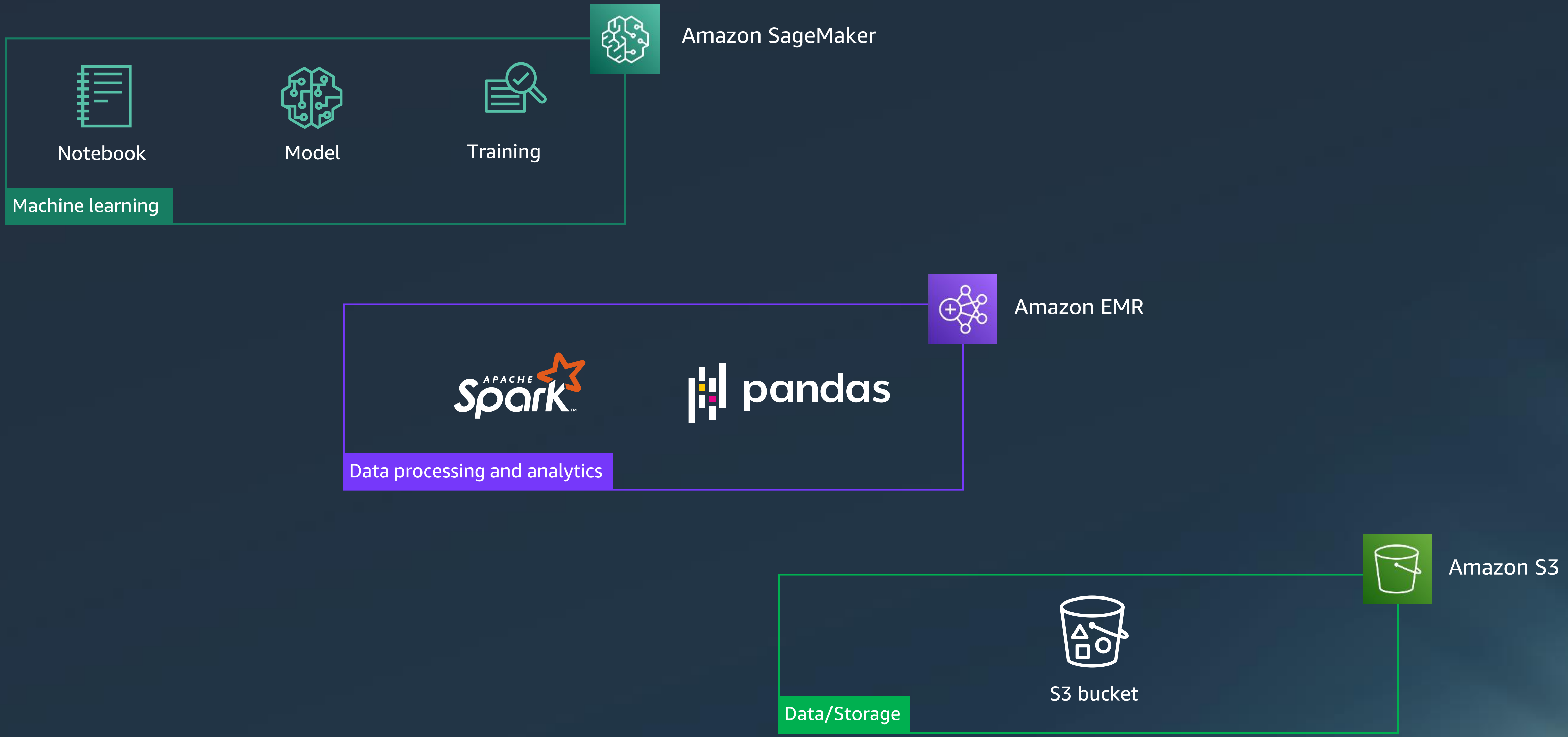
Amazon EMR

Apache Spark

pandas

Data processing and analytics

Our solution



Amazon SageMaker

Notebook

Model

Training

Machine learning

Amazon EMR

APACHE Spark

pandas

Data processing and analytics

Amazon S3

S3 bucket

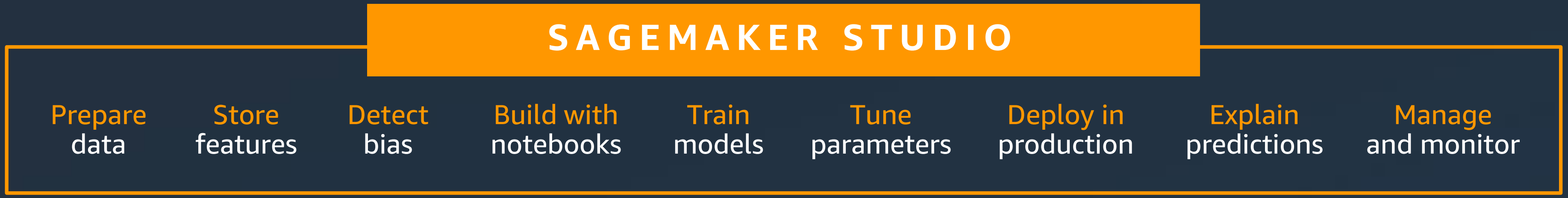
Data/Storage

Amazon SageMaker Studio

FULLY INTEGRATED DEVELOPMENT ENVIRONMENT (IDE) FOR MACHINE LEARNING

Amazon SageMaker Studio

FULLY INTEGRATED DEVELOPMENT ENVIRONMENT (IDE) FOR MACHINE LEARNING



SAGEMAKER STUDIO

Prepare data

Store features

Detect bias

Build with notebooks

Train models

Tune parameters

Deploy in production

Explain predictions

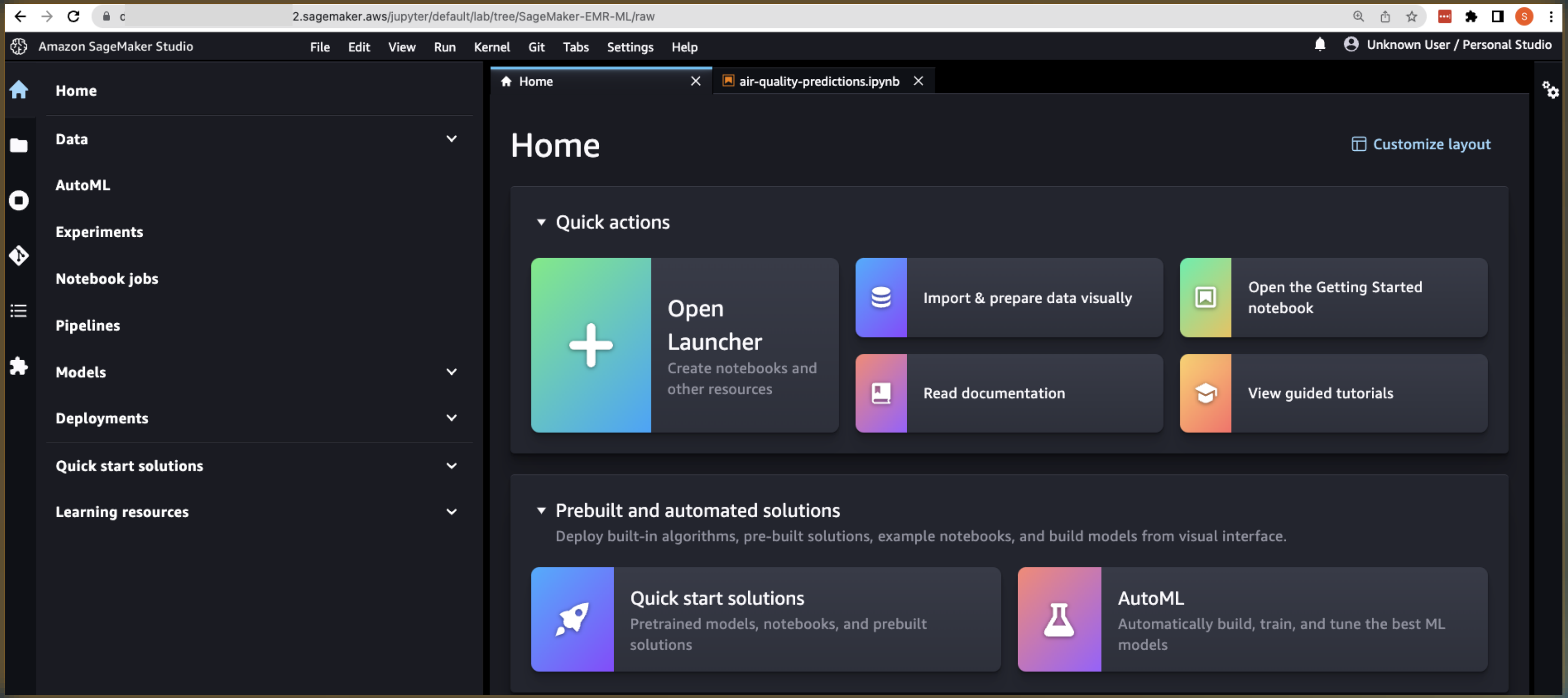
Manage and monitor

Amazon SageMaker Studio

FULLY INTEGRATED DEVELOPMENT ENVIRONMENT (IDE) FOR MACHINE LEARNING

SAGEMAKER STUDIO

Prepare data | Store features | Detect bias | Build with notebooks | Train models | Tune parameters | Deploy in production | Explain predictions | Manage and monitor



Home

Quick actions

- Open Launcher: Create notebooks and other resources
- Import & prepare data visually
- Open the Getting Started notebook
- Read documentation
- View guided tutorials

Prebuilt and automated solutions

Deploy built-in algorithms, pre-built solutions, example notebooks, and build models from visual interface.

- Quick start solutions: Pretrained models, notebooks, and prebuilt solutions
- AutoML: Automatically build, train, and tune the best ML models

github.com/debnsuma/sagemaker-studio-emr-spark.git

Demo

Visit the Data & AI/ML resource hub

Dive deeper into these resources, get inspired and learn how you can use AI and machine learning to accelerate your business outcomes.

- 6 steps to machine learning success e-book
- 7 leading machine learning use cases e-book
- Machine learning at scale e-book
- Achieving transformative business results with machine learning e-book
- Tackling our world’s hardest problems with machine learning e-book
- Accelerating machine learning innovation through security e-book
- … and more!

https://bityl.co/FqdC

Visit resource hub

AWS Training and Certification

Access the AI & ML learning plan courses built by AWS experts on AWS Skill Builder

- Get started with digital self-paced, on-demand training and ramp-up guides to help you grow your technical skills

- Learn how to apply machine learning, artificial intelligence, and deep learning to unlock new insights and value in your role

- Take the steps today, towards validating your expertise with an AWS Certified Machine Learning – Specialty Certification

https://bit.ly/3FnxDH7

Learn your way explore.skillbuilder.aws »

Thank you for attending AWS Innovate – Data & AI/ML Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

- aws-apj-marketing@amazon.com
- twitter.com/AWSCloud
- facebook.com/AmazonWebServices
- youtube.com/user/AmazonWebServices
- slideshare.net/AmazonWebServices
- twitch.tv/aws

Thank you!