DATA ENGINEERING @METISPROJECT

Guitar Appraisal App

Building a data pipeline for an ML-enabled web app

7%

16M Americans aged 16-34 learned to play guitar 2019-2021

62%

learners cite COVID as the reason for learning

92%

increase in sales of Fender guitars under $500

“

I want to be able to estimate the value of a used guitar using machine learning from my phone.

MIKE BERNARDO | IDEO

Web App Data Pipeline

DATA INGESTION (LOCAL)

Web Scrape **Reverb.com**
- Selenium, Beautiful Soup
- retrieves **130k guitars** listed

PROCESSING (LOCAL)

- Clean data
- feature engineering
- model building
- export data and model
- export to **AWS via boto 3**

DATA INGESTION(LOCAL UPDATE)

Scheduled **Reverb.com** web scrape
- PaperMill, Cron
- retrieves weekly updated listings

TESTING ROBUSTNESS)

Create updated **predictive LR model**
- retrieve master and updated data
- clean updated data
- combine data and rebuild model
- export **new model/data** via boto 3

DATA STORAGE (CLOUD)

AWS S3 via boto 3
- Scraped data stored in PostgreSQL
- accessed via psychopg2
- Linear Regression model as pkl
- different IAM users based on task

PROCESSING (CLOUD)

Create updated **predictive LR model**
- retrieve master and updated data
- clean updated data
- combine data and rebuild model
- export **new model/data** via boto 3

EMAIL

status of update

DEPLOYMENT (CLOUD)

- **Streamlit Cloud** front-end
- retrieves model from AWS S3
- user selects features
- LR model predicts value

Data Engineering

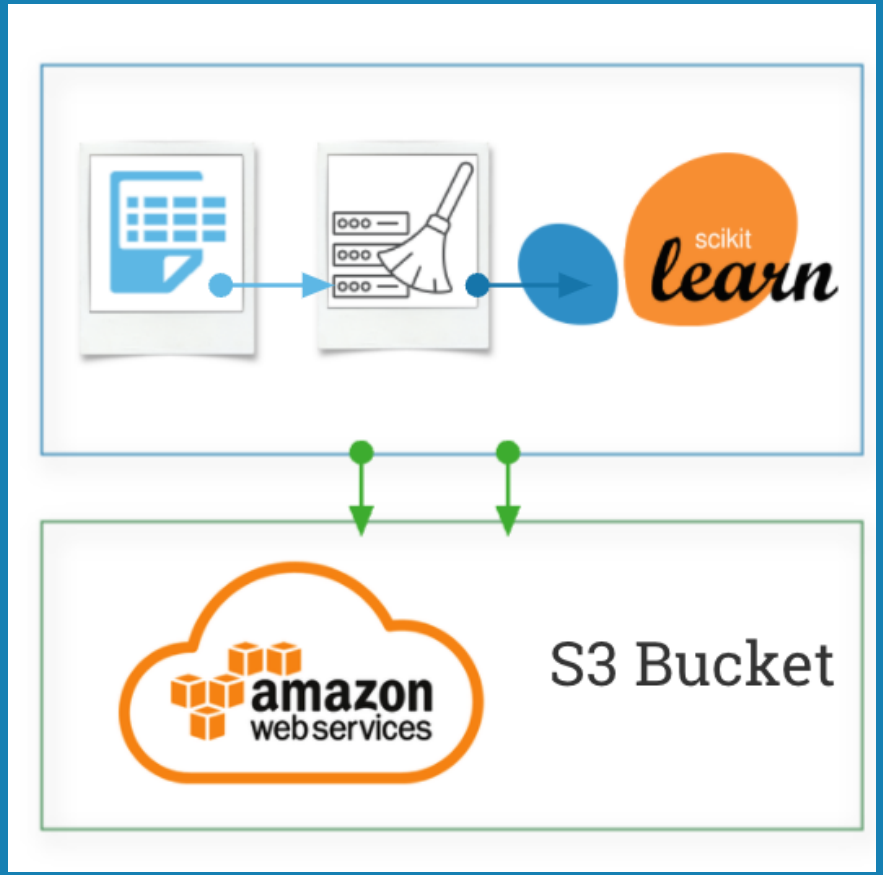
Local Data Ingestion and Processing

WEB SCRAPING

- **Selenium** and **Beautiful Soup**
- 1st pass to get URLs for each guitar
- 2nd pass to get details
- 130k initial records
- save **pandas** data frame

Data Engineering

Local Data Ingestion and Processing



CLEAN AND MODEL

- build functions to clean data
- feature engineering for model
- create predictive model
- upload cleaned data to AWS bucket using boto 3

Data Engineering

Local Data Ingestion and Processing

Predicted	Actual	brand	origin
1,454.47	1,299.99	Fender	Japan
2,064.38	1,699.99	Fender	United States
2,049.22	1,699.99	Fender	United States
1,454.47	1,299.99	Fender	United States
1,518.77	1,149.00	ESP LTD	Asia
1,488.31	1,299.00	ESP LTD	Asia
1,988.57	2,199.00	Parker	United States
1,407.72	1,399.00	Gibson	United States
2,064.38	1,699.99	Fender	United States
2,049.22	2,250.00	Fender	United States

INITIAL MODELING

- **feature engineering**
- used **sklearn** for logistic regression model (GLS instead of OLS)
- **predictive model** saved as **pkl** then uploaded to **AWS S3** bucket using **boto 3**

Data Engineering

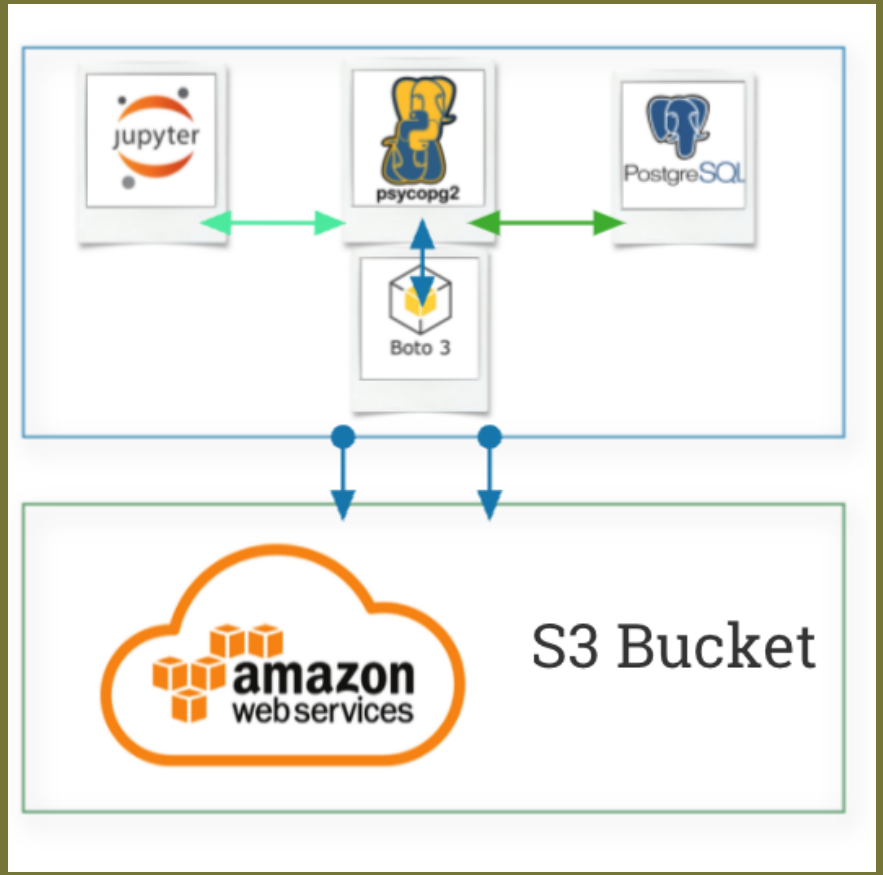
Local Data Ingestion and Processing

DATA UPDATES

- schedule via **CRON**
- run notebook via **papermill**
- scrape using **selenium/bs4**
- **rebuilds** data and model
- **PUT** new data/model to AWS S3
- backup data to **postgreSQL**

Data Engineering

Data Storage **(AWS S3)**

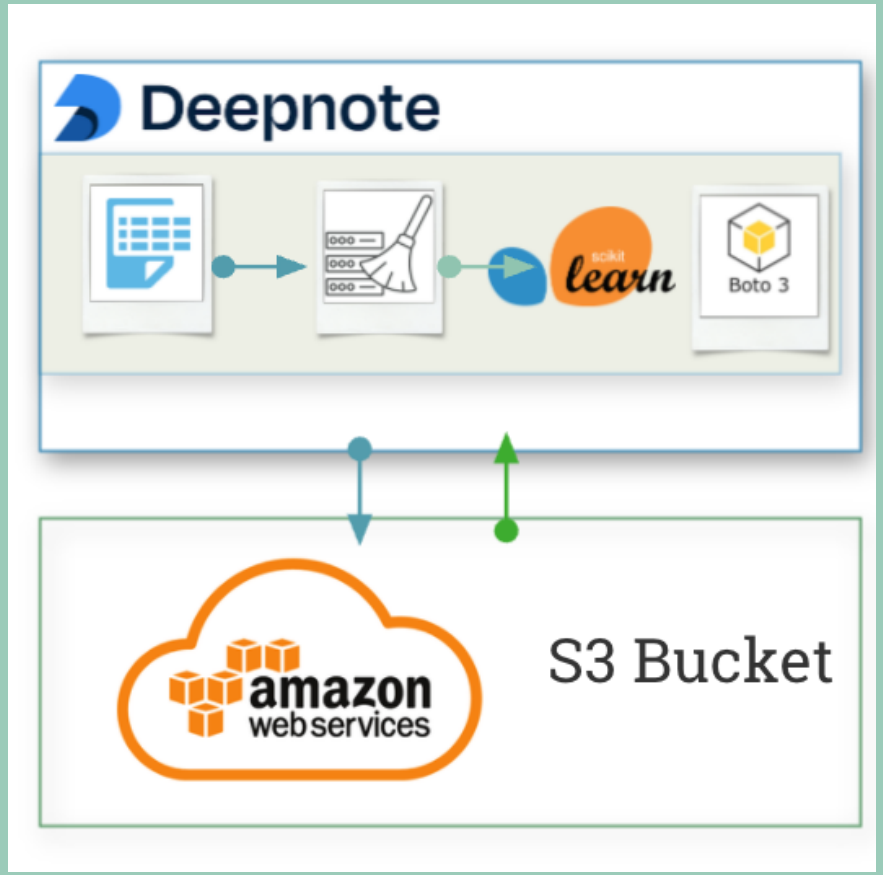


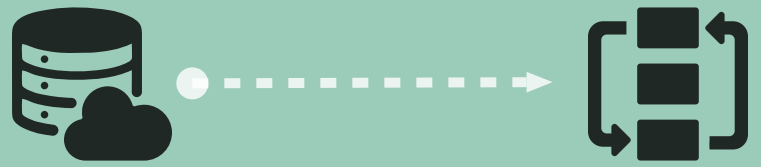
CLOUD STORAGE

- data on **AWS S3** via **boto 3**
- data in **postgres** via **psychopg2**
- linear regression model **pkl file**
- role-based **IAM** users
- built-in data **versioning**

Data Engineering

CLOUD Data Processing



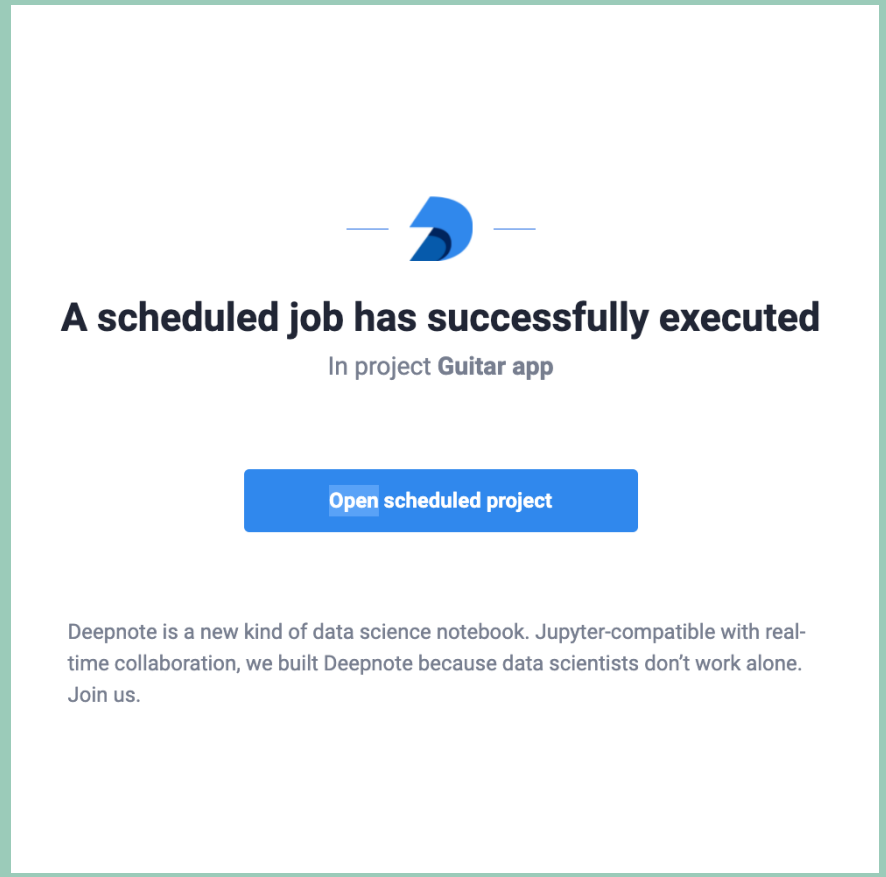


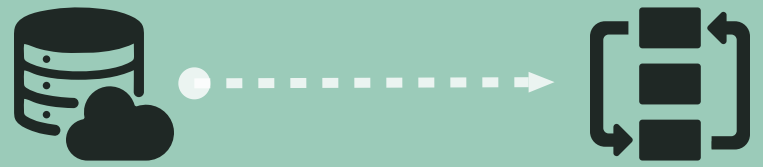
UPDATE DATA

- scheduled **Deepnote** execution
- **GET** existing data from S3
- **GET** updated data from S3
- Combine data / Rebuild model
- **PUT** new data and model to S3

Data Engineering

CLOUD Data Processing



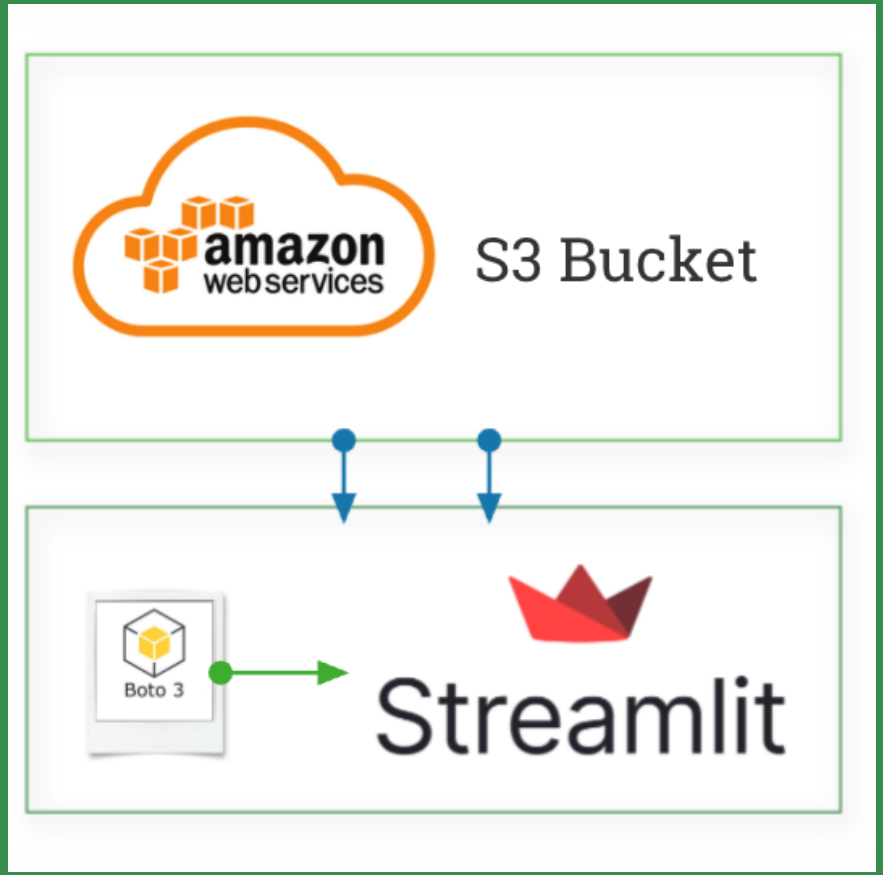


MONITORING

- executed notebook saved
- see errors in execution if any
- email notification

Data Engineering

Model Deployment

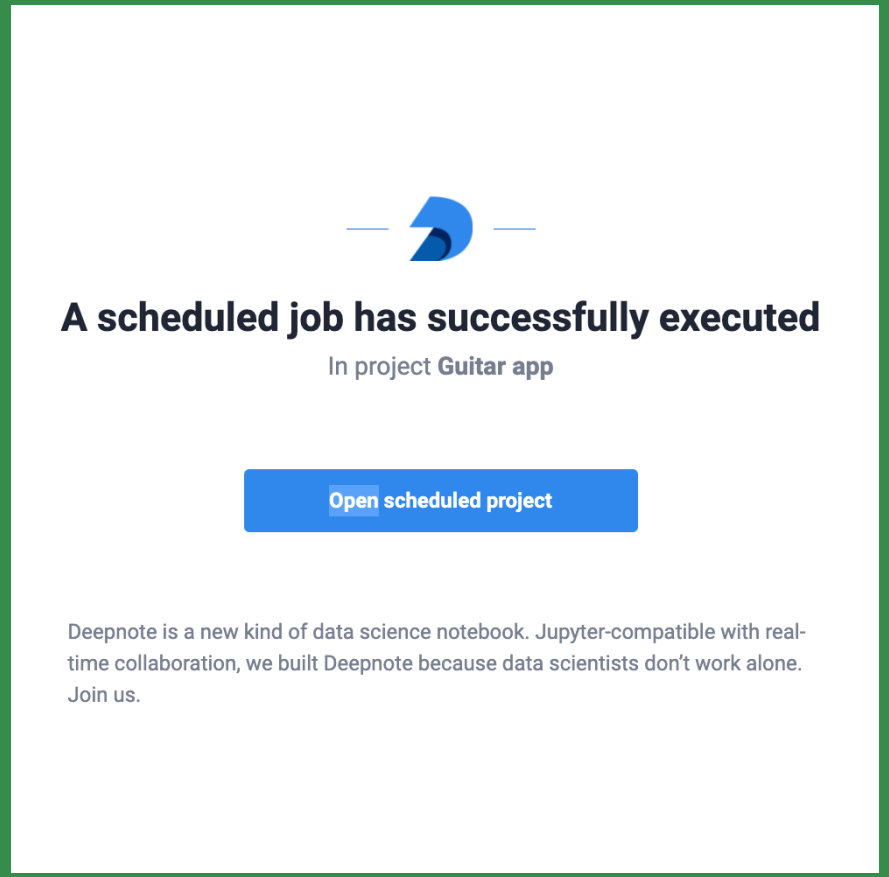


STREAMLIT CLOUD APP

- **STRAMLIT** front-end via **Github**
- **GET** latest LR model from AWS S3
- user selects **features**
- LR model **predicts** appraisal value

Data Engineering

Model Deployment



MONITORING

- executed notebook saved
- see errors in execution if any
- email notification

The Live App