Jupyter Notebook Ops

2021年02月26日

株式会社 Nowcast

隅田 敦

自己紹介

サンプルコード

https://github.com/yummydum/automate_reports

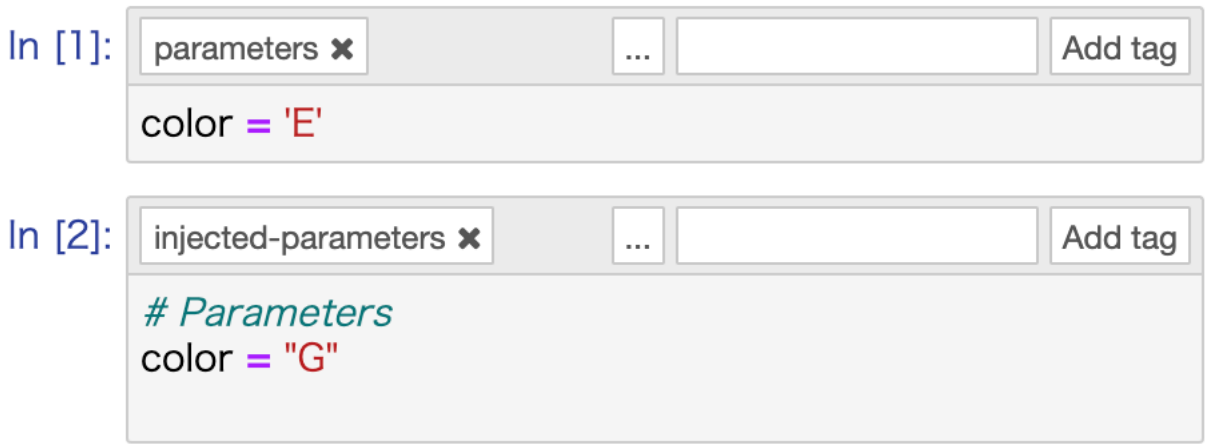
Jupyter Notebookによる分析や実験を効率よく運用・管理したい

- Parametrization
  - ノートブックをパラメタ化し使い回せるようにする
- Reproducibility and Scalability
  - ノートブックを複製可能な環境で同じように実行出来る
- Communication
  - ノートブックを素早く手軽に共有する

Parametrization by Papermill

- ノートブックをパラメタ化すれば…
  - 店舗Aの売上分析用ノートブックを店舗B，店舗Cにも流用
  - 機械学習モデルの実験用ノートブックを異なるハイパーパラメタで学習
  - 日時で更新されるデータセットに同じノートブックを適用して分析

- **Papermill**
  - ノートブックにパラメタを設定し実行してくれるライブラリ



パラメタを一つのセルにまとめparametersタグをつけておく

Papermillが挿入したセル

Parametrization by Papermill

```python
import sys
import papermill as pm

# Check if the parameter is properly set
nb_path = 'notebooks/example.ipynb'
params = pm.inspect_notebook(nb_path)
assert 'color' in params

# Execute the notebook with the parameter
color = sys.argv[1]
pm.execute_notebook(
    nb_path,
    f'notebooks/example_{color}.ipynb',
    parameters={'color': color},
)
```

Reproducibility and Scalability by Docker

- ノートブックをコンテナ化すれば…
  - 再現性を担保できる
  - ローカルで書いたノートブックをワークステーションで実行
  - パラメタ化してコンテナオーケストレーションで並列分散実行

- **Docker**

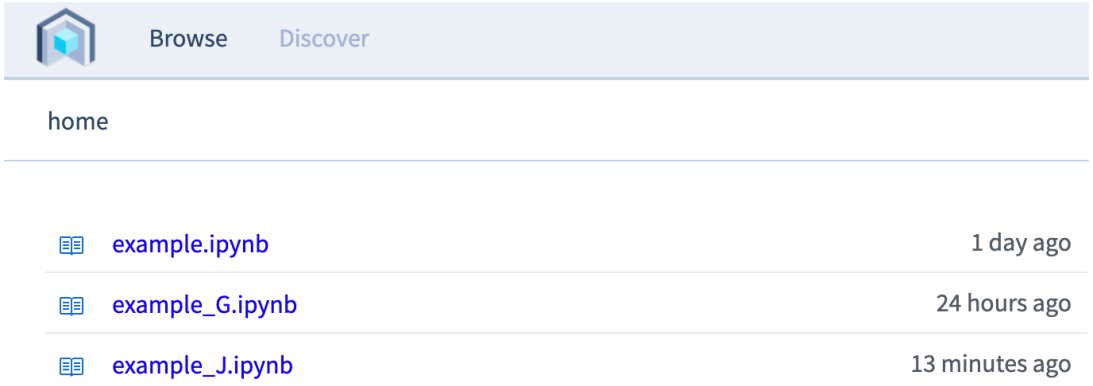
```
1    FROM jupyter/base-notebook
2    RUN pip install papermill seaborn
3    ENTRYPOINT ["python","run_notebook.py"]
```

公式イメージを使いましょう(自分でやると色々とハマります)
デフォルトユーザーのjovyan君はよわよわ権限

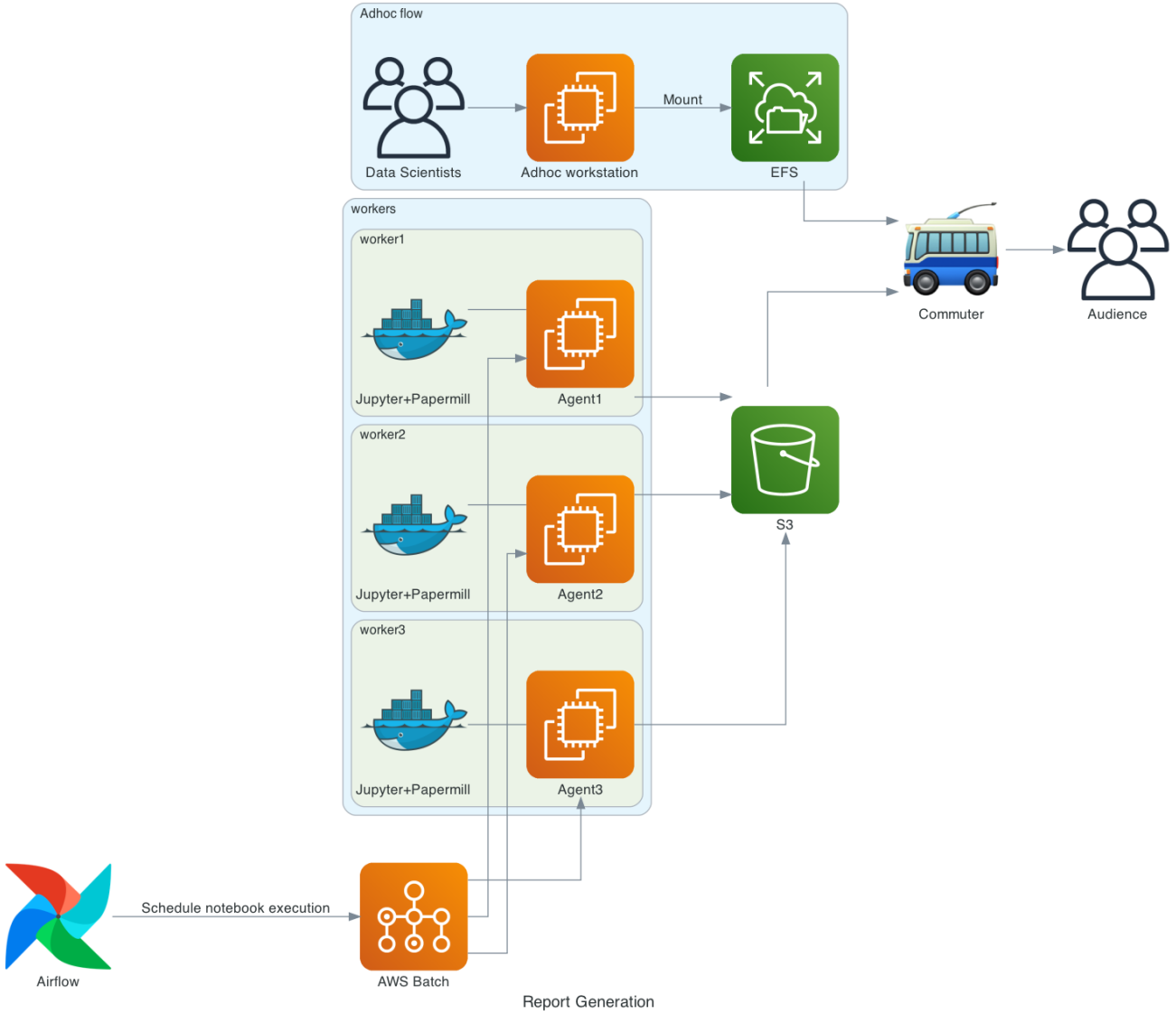
Communication by Commuter

- ノートブックの共有は地味に面倒くさい
  - 誰もが.ipynbを開けるとは限らない
  - ノートブックを開くたびにファイルの差分が生じるのでGitと相性が悪い
  - Githubに上げるにはファイルサイズが大きい

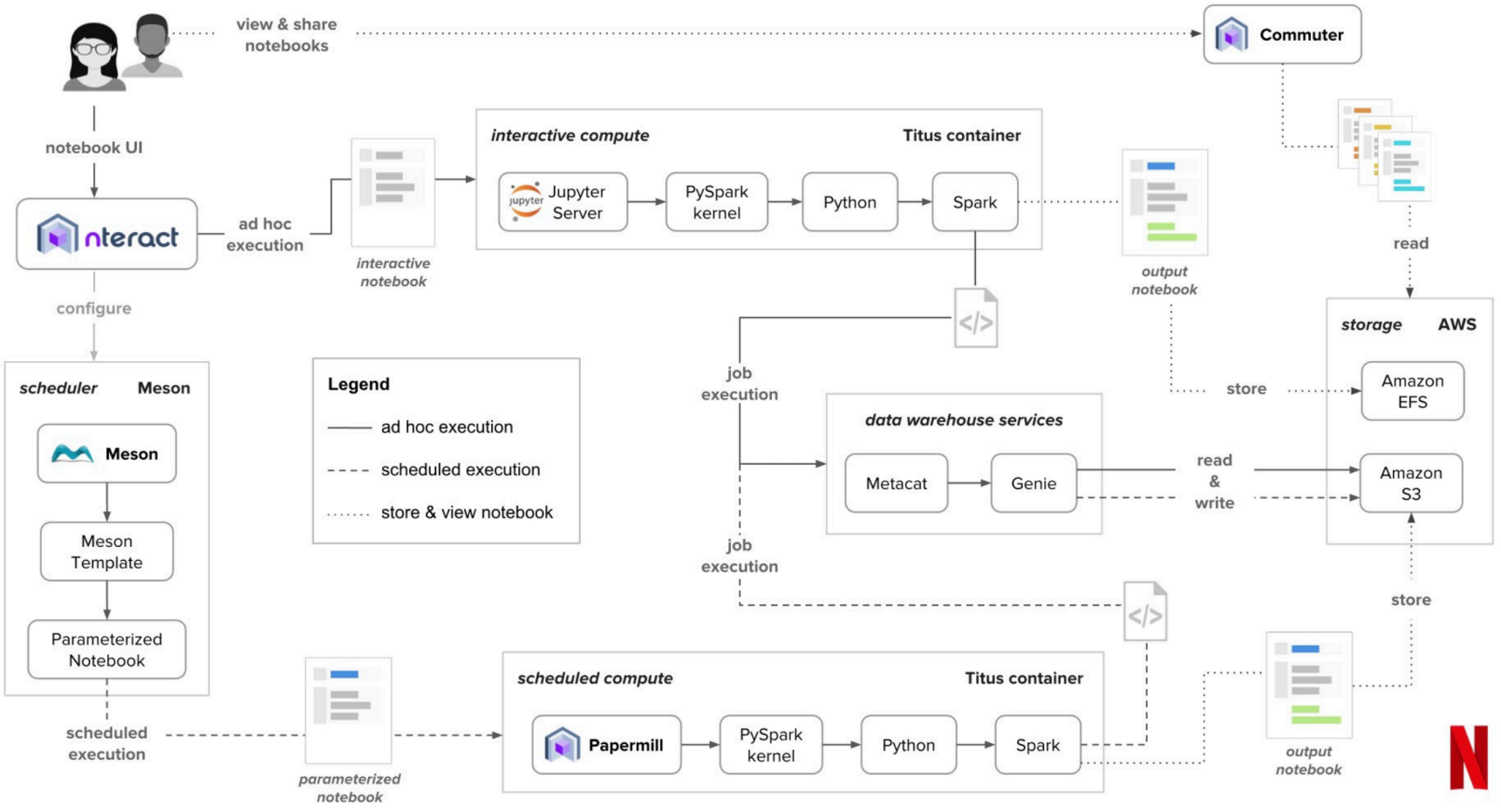
- **Commuter**
  - ローカル，S3からノートブックを読み込みread onlyで表示してくれる
  - Elastic Searchによる検索機能なども



快適なノートブック生活



参考: Netflixのノートブックインフラ



Notebook Infrastructure at Netflix

https://netflixtechblog.com/notebook-innovation-591ee3221233?gi=19cdf66a04b4

We are hiring!