# Jupyter Notebook Ops

2021年02月26日

株式会社 Nowcast

隅田 敦

# 自己紹介

## 隅田 敦

**データエンジニア/**
**データサイエンティスト**

東京大学経済学部経済学科にて計量経済学を専攻. 経済現象の理解のためには高品質高頻度のデータが必要との想いから2018年よりナウキャストにてインターンを始める. エンジニアリング業務をこなす中で情報科学への関心が高まり, 2019年より東京大学大学院情報理工学系研究科コンピュータサイエンス専攻に進学し, 計算言語学/自然言語処理の研究を行う. 2021年4月より入社予定.

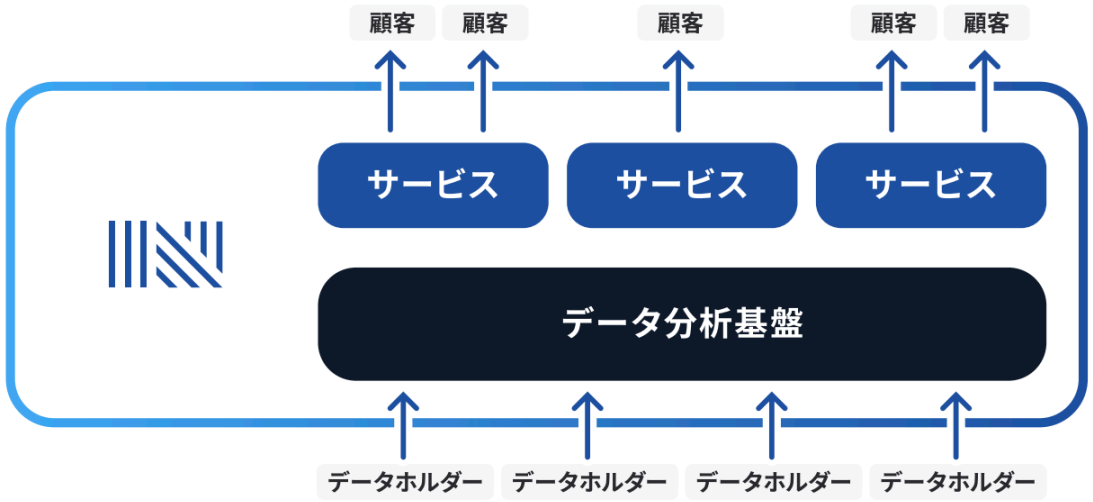
@yummydum

# 株式会社ナウキャストの紹介

## 事業の全体像

## 複数のデータソースで複数のサービスを展開し、様々な顧客セグメントに価値を提供

複数のAlternative Dataのソースを共通のデータ分析基盤で整形・分析し、金融業界を中心に様々な顧客にサービスを提供しています。






資料はこちら→

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

- Parametrization by **Papermill**
  - ノートブックをパラメタ化し使い回せるようにする
- Communication by **Commuter**
  - ノートブックを素早く手軽に共有する

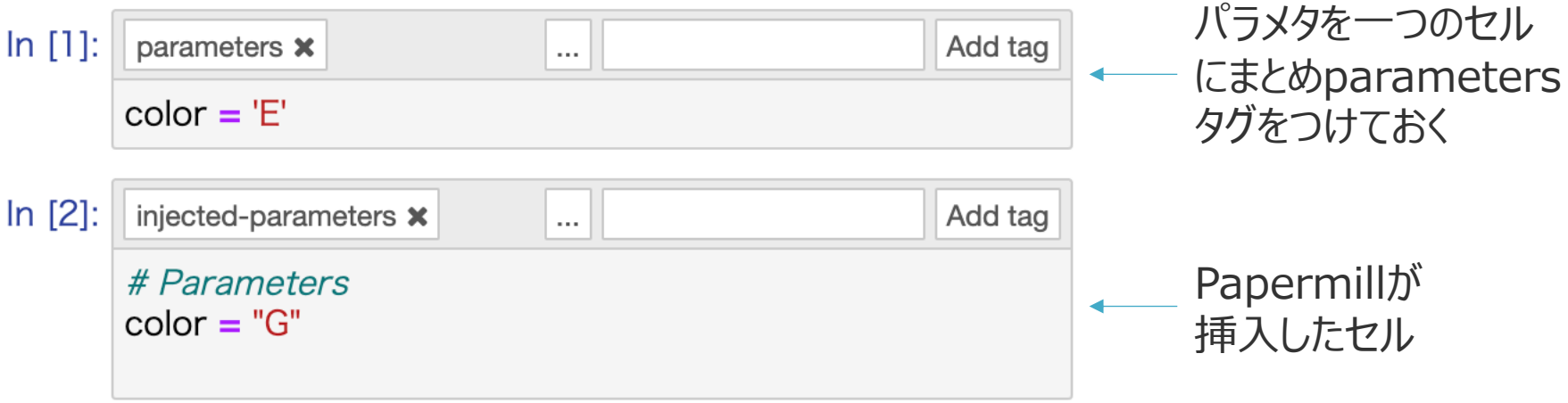
まだあるよ→

# Parametrization by Papermill

- ナウキャストではPOSデータやクレジットカードデータを用いて企業の売上予測をしています
  - 証券コード毎に詳細な分析(企業，事業，商品，イベント…)
  - 対象とする証券コードが200個，1つのノートブックの実行に15分なら50時間かかってしまう
  - データセットはどんどん新しくなるので定期的に再実行する必要がある

**ノートブックをパラメタ化して並列分散処理しよう!**

## 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



# Parametrization by Papermill

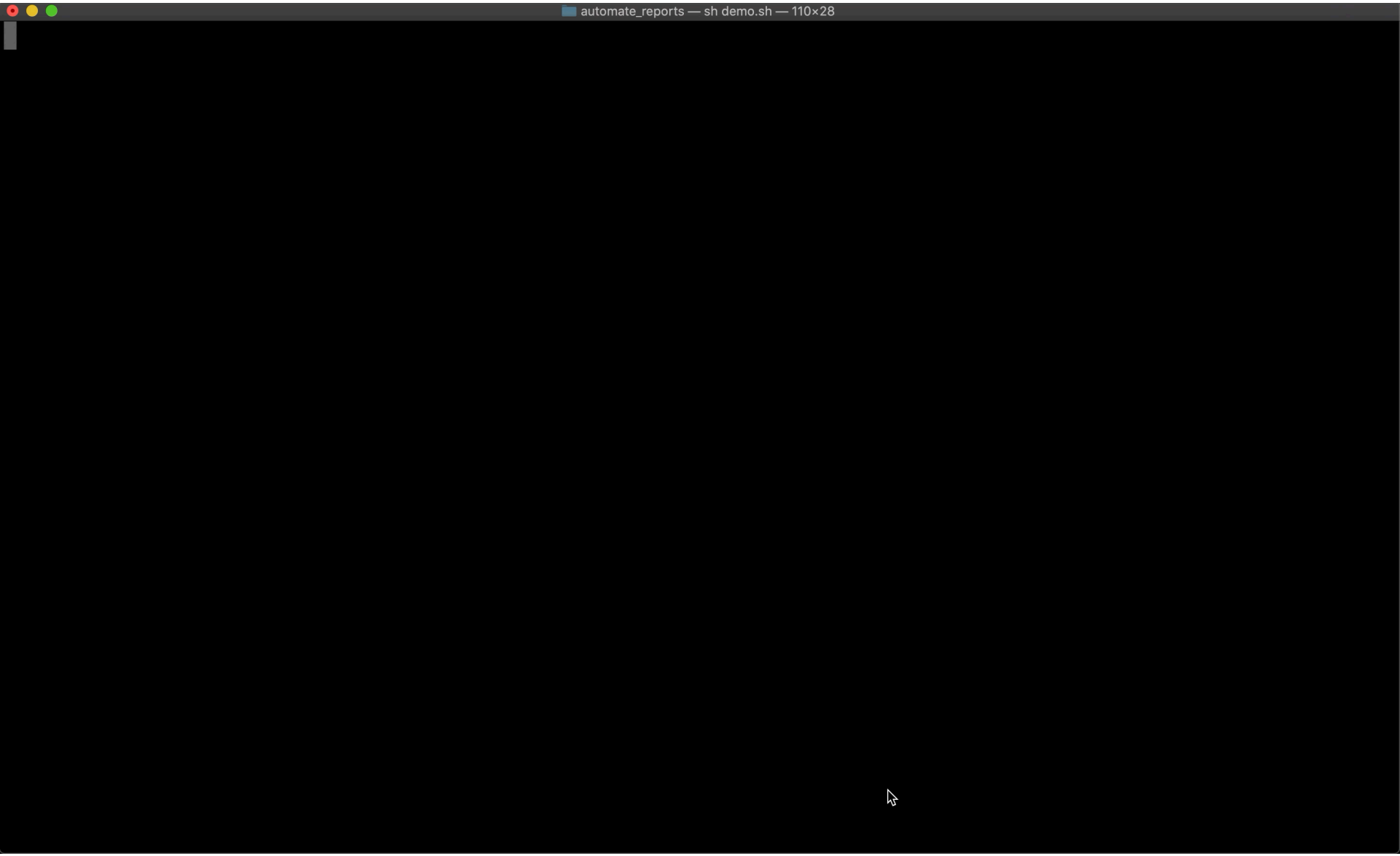
```python
import sys
import papermill as pm

# 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},
)
```

# 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
  - 数百のノートブックを手渡しするのは…

- **Commuter**
  - ローカル・S3からノートブックを読み込みhtmlに変換してくれるwebサーバー
  - 誰でもブラウザからノートブックを閲覧出来る!

# Demo: diamond EDA for different colors



# サンプルコード

https://github.com/yummydum/jupyter-notebook-ops

手元で動かしてみよう!

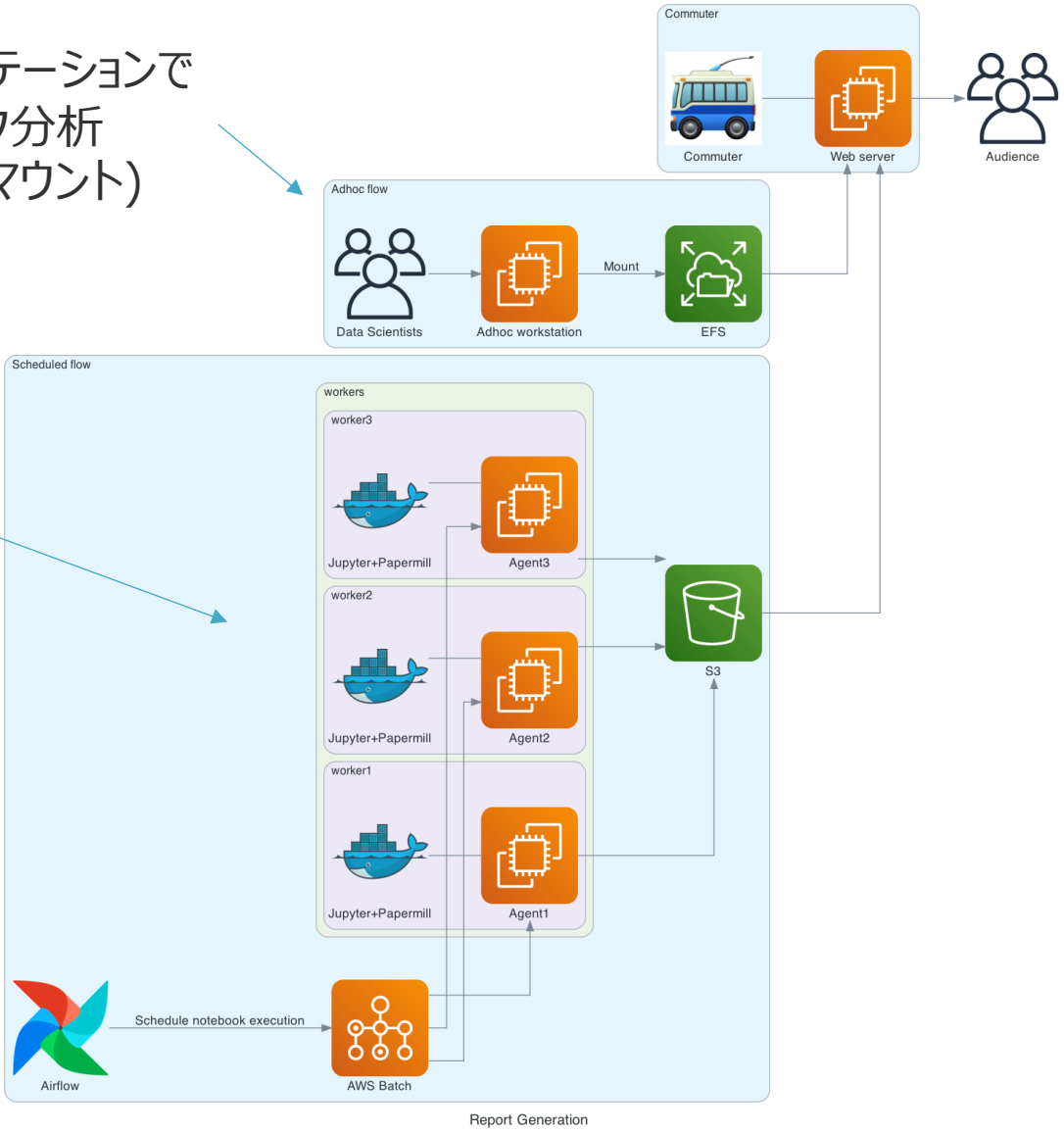
# Notebook infrastructure example

ワークステーションで
アドホック分析
(EFSをマウント)

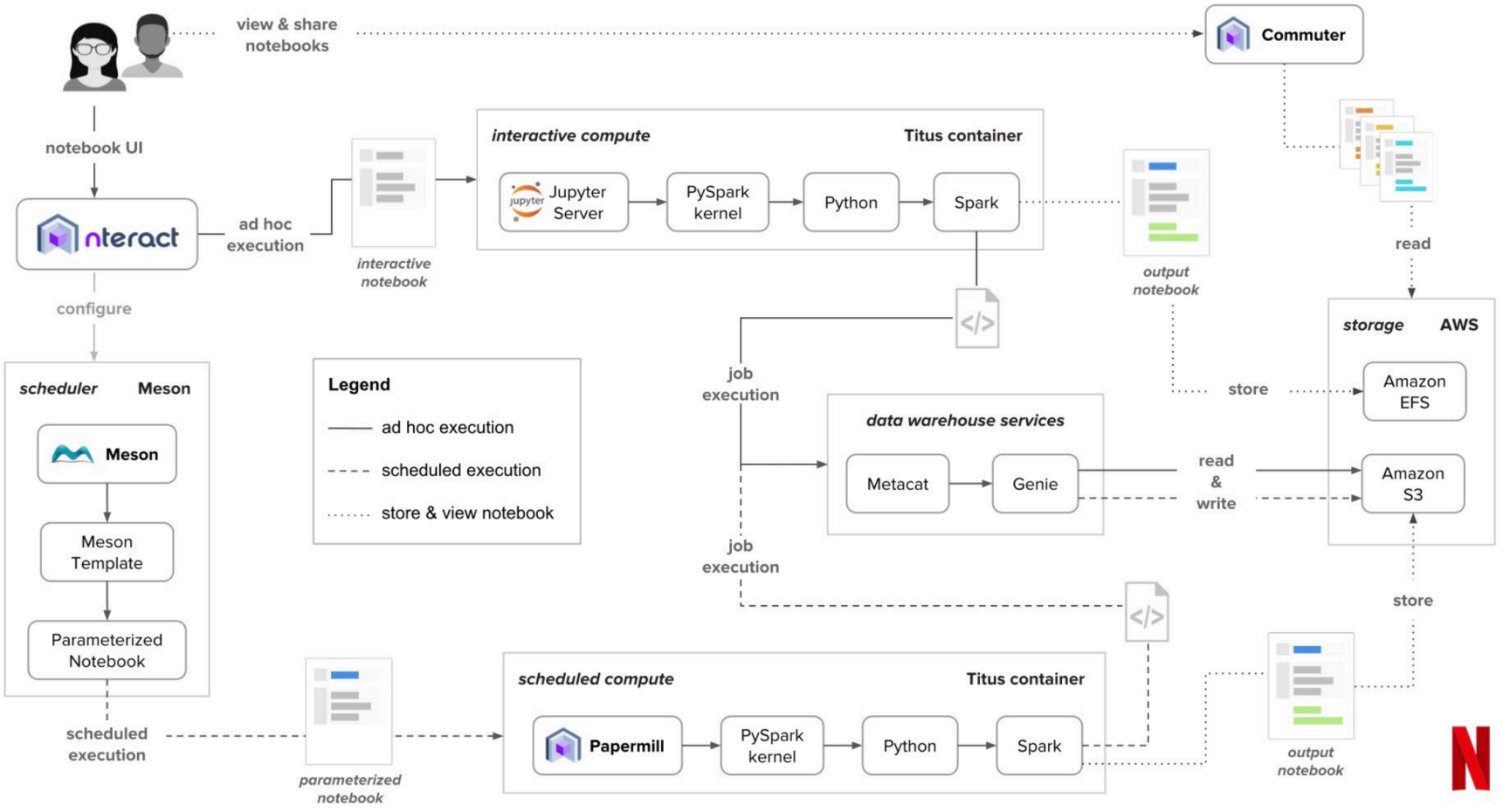
パラメタ化+
並列分散処理

データの更新に
合わせてノートブック
を定期的に更新

ブラウザから最新の
ノートブック一覧を
いつでも見れる!



# 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!

資料はこちら(大事なことなので以下略)→