

DSC 207 - PYTHON FOR DATA SCIENCE

# HOW NETFLIX DATA SCIENCE TEAM WORKS

İLKAY ALTINTAŞ, PH.D.

UC San Diego

COMPUTER SCIENCE & ENGINEERING  
HALİCİOĞLU DATA SCIENCE INSTITUTE



By the end of this video, you will be able to:

- List common use cases of Jupyter notebooks inside Netflix.
- Describe how different teams inside Netflix collaborate with help of Jupyter Notebooks.



N SERIES

# STRANGER THINGS

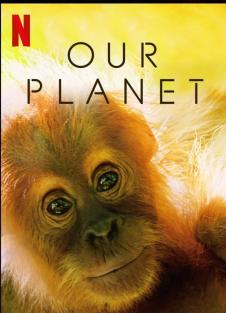
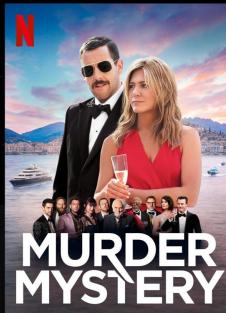
When a young boy vanishes, a small town uncovers a mystery involving secret experiments, terrifying supernatural forces and a strange little girl.

Play

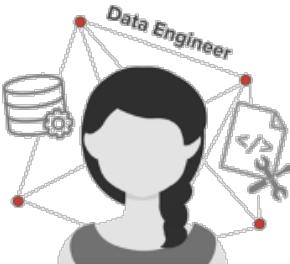
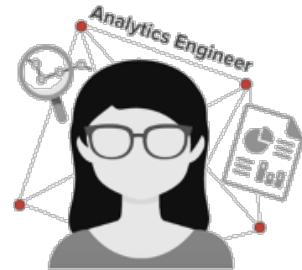
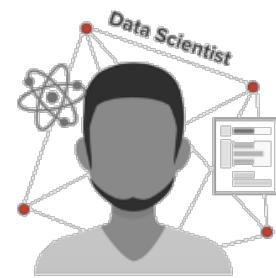
+ My List



## Popular on Netflix



# Data Empowers Netflix



## Data roles at Netflix

- business analyst
- data analyst
- quantitative analyst
- algorithm engineer
- analytics engineer
- data engineer
- data scientist
- machine learning scientist
- research scientist



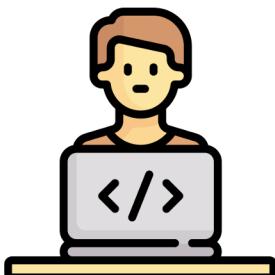
Common  
Interface  
?



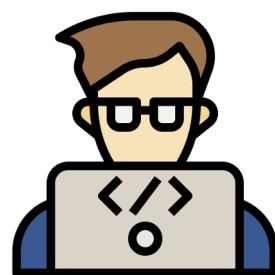
Common  
Execution  
Pattern  
?

## How do Different Roles at Netflix Collaborate

- Each role relies on a different set of tools and languages.
- Do they have any overlapping tasks? Is there a common pattern?
- Is there a tool which is precisely designed tackle this?



**Analytics Engineer**  
SparkSQL, Python



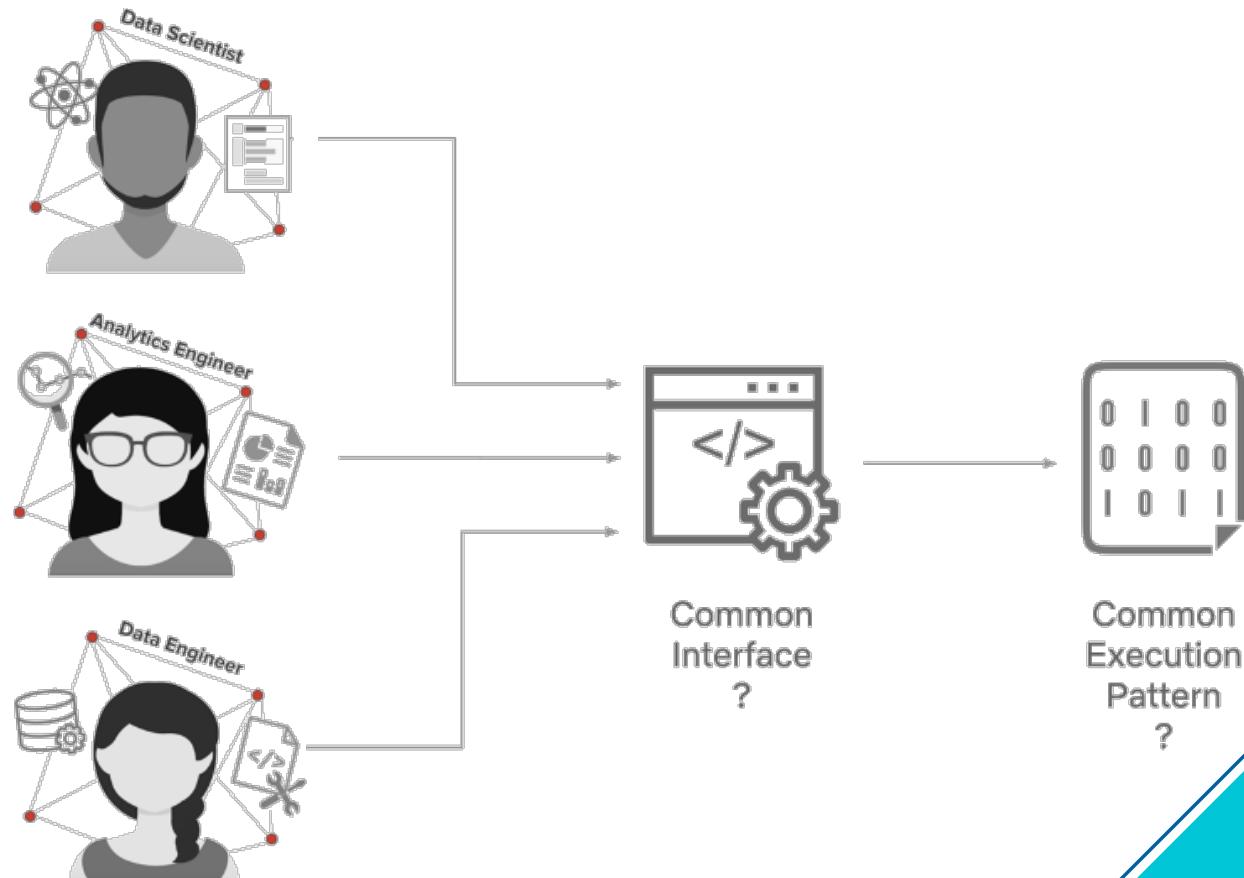
**Data Scientist**  
PySpark, Python



**Data Engineer**  
Scala, Spark, Python, SQL

# How is Jupyter most commonly used inside Netflix?

- Data Access
- Notebook Templates
- Scheduling Notebooks



# Fundamental Components of Netflix Data Platform: Storage

## EFS Volume

```
/efs  
  /users  
    /kylek  
    /m  
    /mufford  
      /MySparkJob.ipynb  
      /New_Titles.ipynb  
      /playback_agg.ipynb  
      /Untitled3.ipynb
```

## Directory Listing

|  |         |             |
|--|---------|-------------|
|  | kylek   | 2 days ago  |
|  | mpacer  | 1 day ago   |
|  | mufford | 6 hours ago |

```
[3] import kragle as kg
```

```
[4] %%sql
```

```
CREATE TABLE vid_agg_fr_2018_05_08 AS  
(  
  SELECT  
    device_type,
```

- Relies on Amazon S3 and EFS Cloud Storage
- Personal workspace for notebooks

# Fundamental Components of Netflix Data Platform: *Compute*

Container Parameters

**Name:**  Max 75 chars

**CPUs:**  Must be between 1 and 8

**Memory:**  GB Must be between 4 and 256

**Disk Space:**  GB Must be between 10 and 128

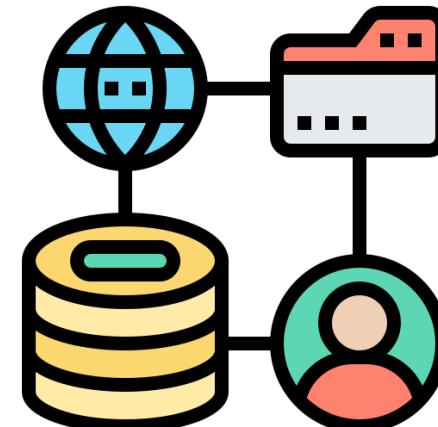
**GPUs:**  Must be between 0 and 8

**Network Bandwidth:**  Mbps Must be between 128 and 2000

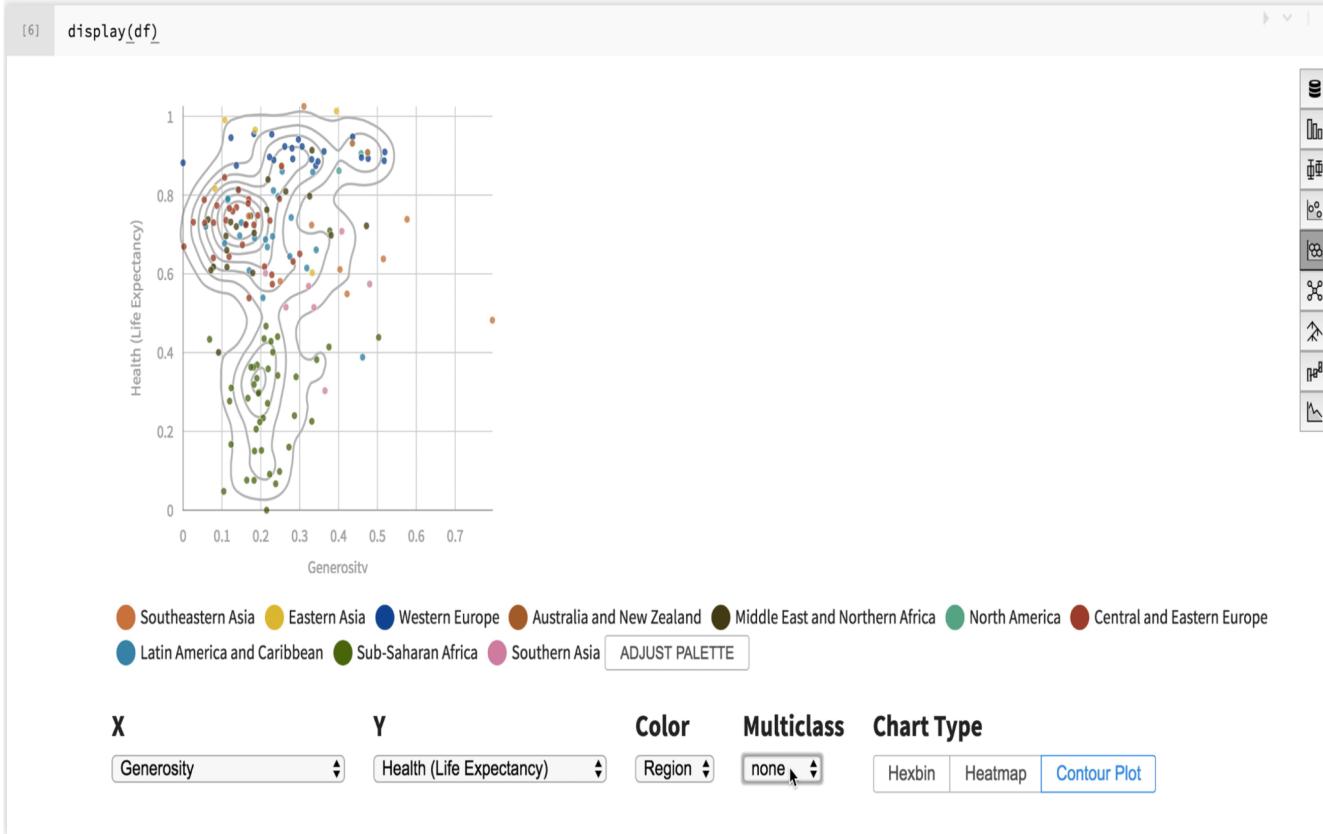
**Environment Variables:**   
one per line and without quotes

**Launch** **Cancel**

- All jobs on the Data Platform run on containers
- User-specified container parameters

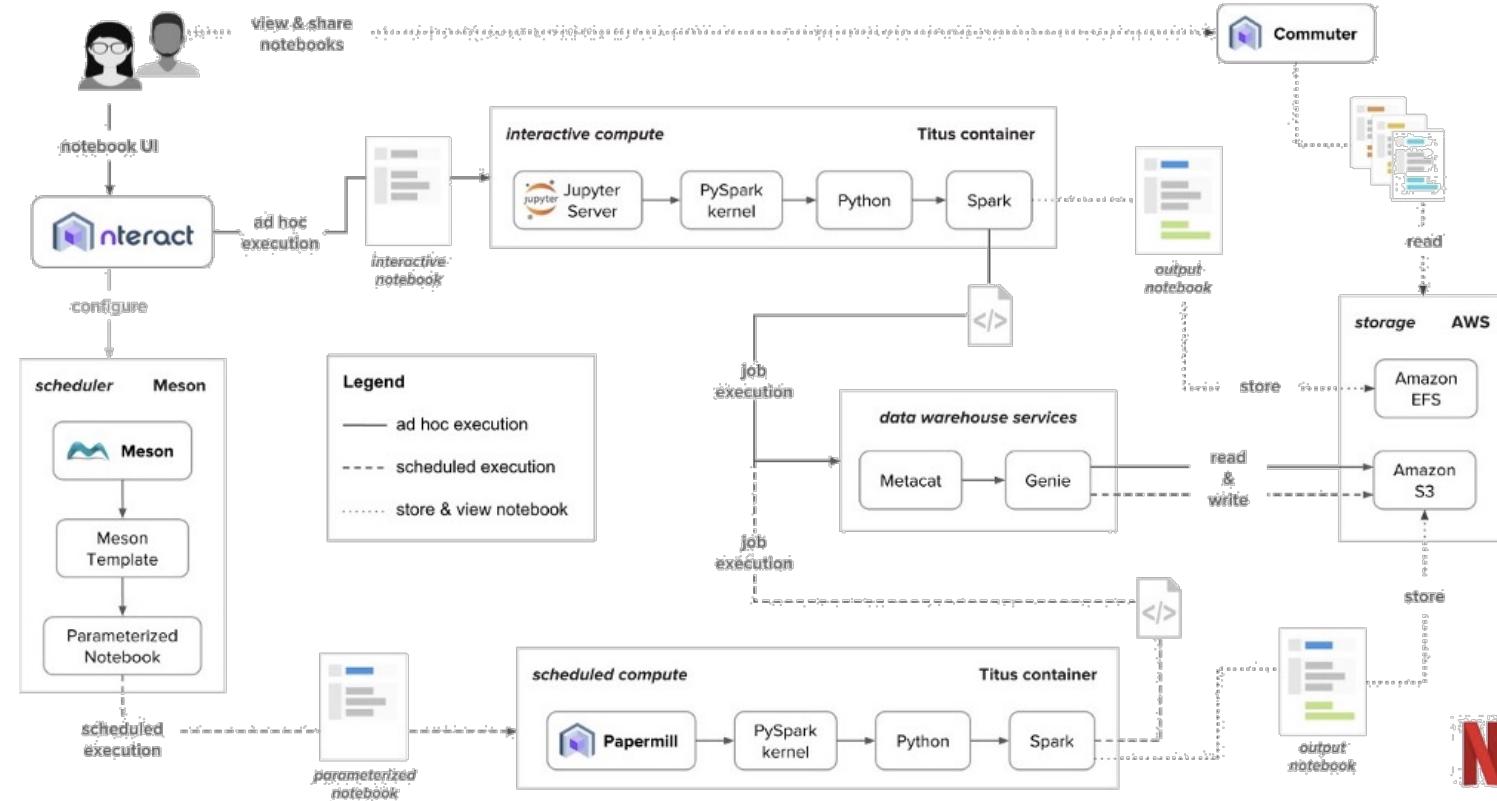


# Fundamental Components of Netflix Data Platform: *Interface*



- interact provides ideal building blocks to which emphasize simplicity and composability!

# Notebook Infrastructure



- **Interact**
- **Papermill**
- **Commuter**
- **Titus**