DataTalks.Club

Data Engineering Zoomcamp

January 17, 2022

Plan

- Instructors
- Course
  - Is it for me?
  - Syllabus
- Course logistics
- DataTalks.Club slack
- Questions

Ankush Khanna

- Senior Data Engineer at Wayfair
- Anything Streaming or Batch

Sejal Vaidya

- Data & ML Engineer
- Course Topics: Platform Infrastructure and Data Ingestion

Victoria Perez Mola

- Team lead of the Analytics platform at Tier Mobility
- Not a data engineer
- Course topics: Analytics engineering

Alexey Grigorev

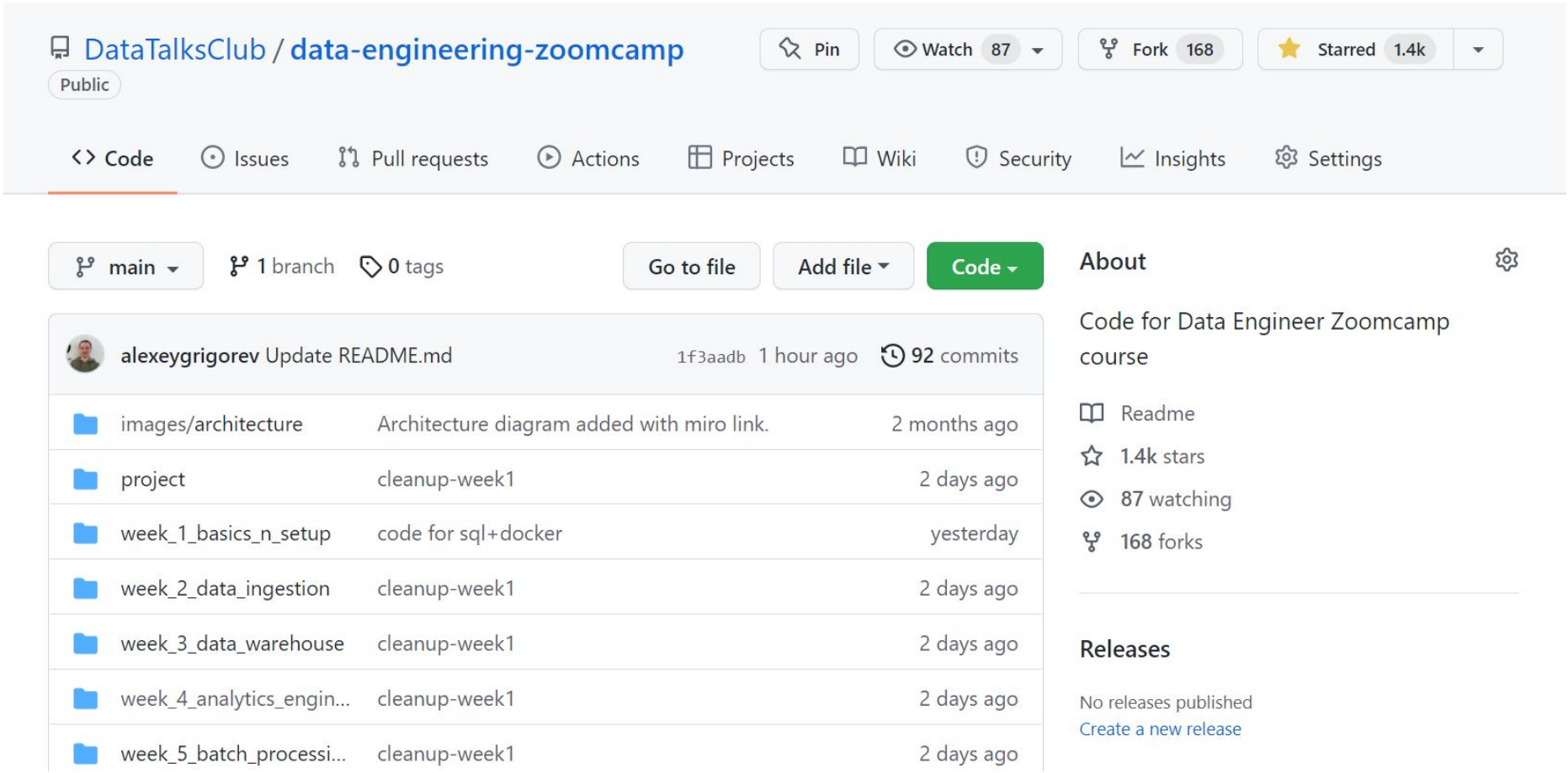
- Principal Data Scientist at OLX Group
- Not a data engineer =)
- Instructor for ML Zoomcamp

Plan

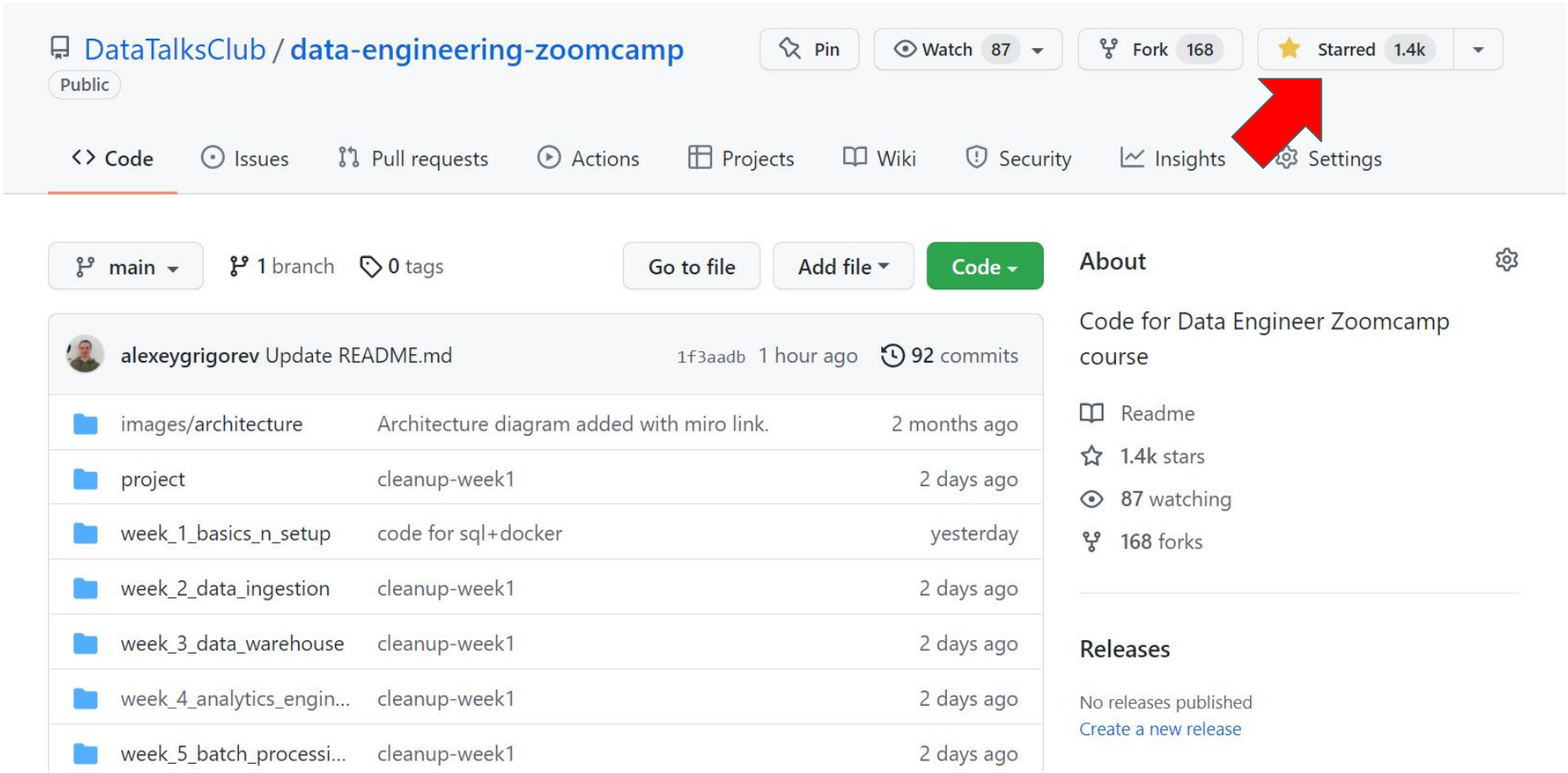
- Instructors
- **Course**
  - Is it for me?
  - Syllabus
- Course logistics
- DataTalks.Club slack
- Questions

Is it for me?

- Pre-requisites
  - Experience with programming (Python)
  - Being comfortable with command line (Git, etc)
  - Exposure to SQL
- Not required
  - Previous experience with data engineering

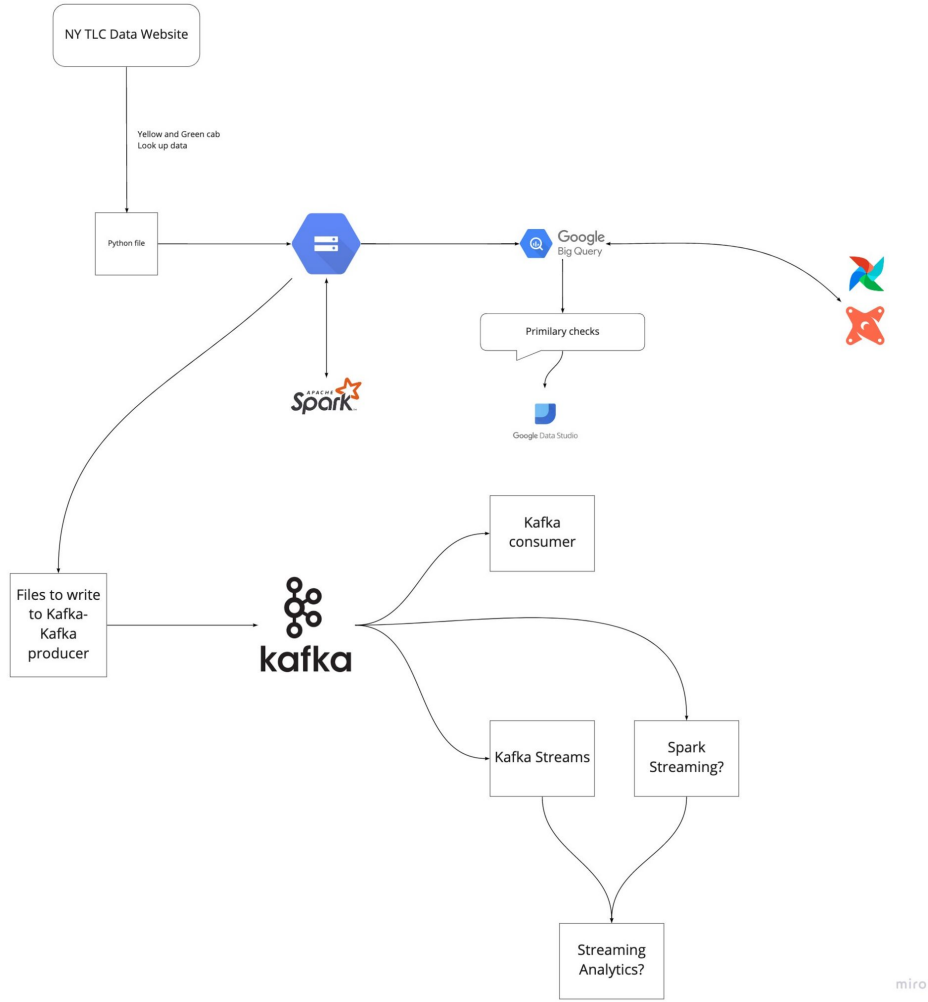


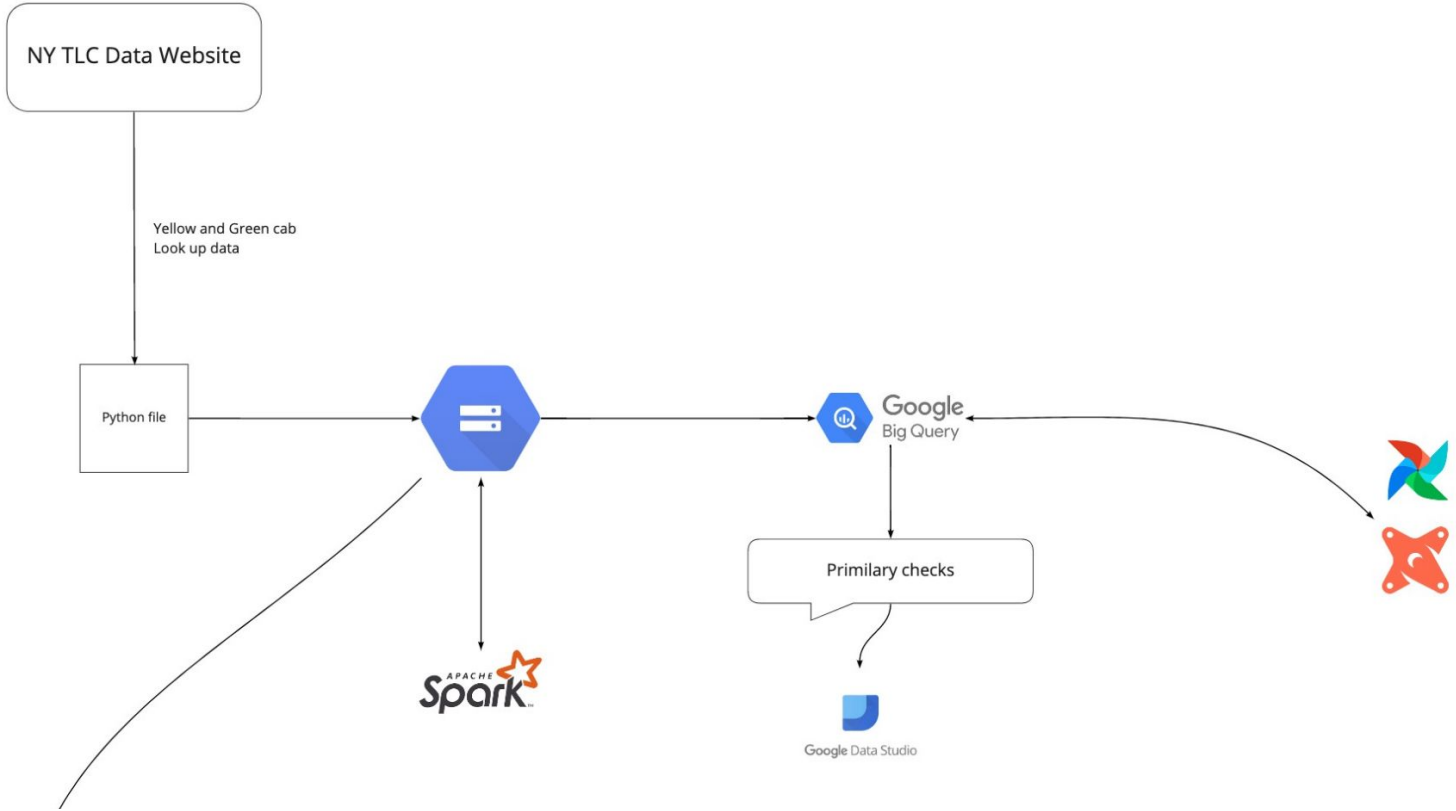
https://github.com/DataTalksClub/data-engineering-zoomcamp

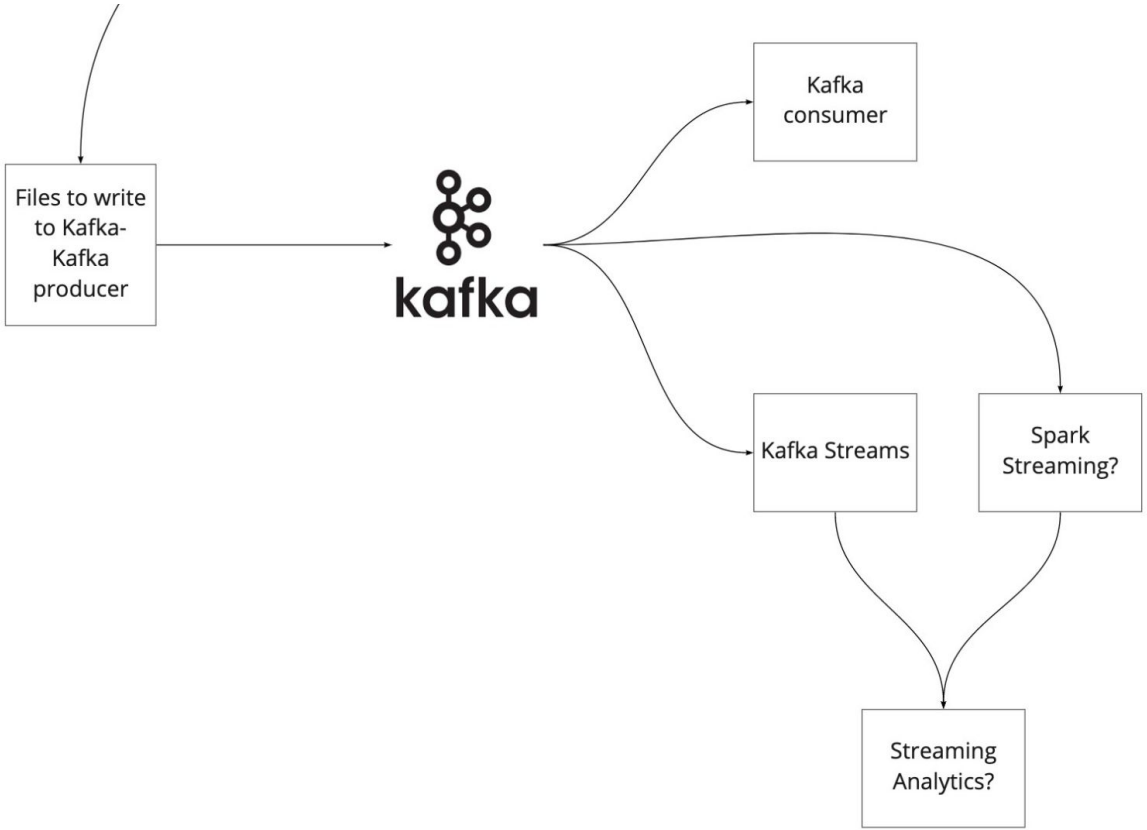


https://github.com/DataTalksClub/data-engineering-zoomcamp

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page


NY TLC Data Website
Yellow and Green cab
Look up data
Python file
Google Big Query
Primilary checks
Google Data Studio
Spark
Files to write to Kafka- Kafka producer
kafka
Kafka consumer
Kafka Streams
Spark Streaming?
Streaming Analytics?
miro


NY TLC Data Website
Yellow and Green cab
Look up data
Python file
Google
Big Query
Primilary checks
APACHE
Spark
Google Data Studio


Files to write to Kafka- Kafka producer
kafka
Kafka consumer
Kafka Streams
Spark Streaming?
Streaming Analytics?

Course overview

- Week 1: Introduction & Prerequisites
- Week 2: Ingestion and orchestration
- Week 3: Data warehouse (Big query)
- Week 4: Analytics engineering (dbt)
- Week 5: Batch processing (Spark)
- Week 6: Streaming (Kafka)
- Weeks 7-10: Project

Week 1: Introduction & Prerequisites

- Setting up the environment
  - Google Cloud account
  - Docker
  - Terraform
- Running Postgres in Docker
- Taking a look at the NY taxi dataset
- SQL refresher

Week 2: Ingestion and orchestration

- Data Lake
  - What is a Data Lake, ELT vs. ETL, Using GCS
- Orchestration
  - What is an Orchestration Pipeline, Data Ingestion, Introducing & Using Airflow
- Demo:
  - Setting up Airflow with Docker
  - Data ingestion DAG
    - Extraction, Pre-processing (parquet, partitioning), Loading, Exploration with BigQuery, etc.
- Best Practices

Week 3: Data Warehouse

- What is Data warehouse
- BigQuery?
  - Partitioning and Clustering
  - With Airflow
  - Best practices

Week 4: Analytics Engineering

- What is dbt and how does it fit the tech stack?
- Using dbt:
  - Anatomy of a dbt model
  - Seeds
  - Jinja, Macros and tests
  - Documentation
  - Packages
- Build a dashboard in Google data studio

Week 5: Batch processing

- Spark internals
- Broadcasting
- Partitioning
- Shuffling
- Spark + Airflow
- Apache Flink as alternative to Spark

Week 6: Stream processing

- Basics of Kafka
- Consumer-Producer
- Kafka Streams
- Kafka Connect

Project

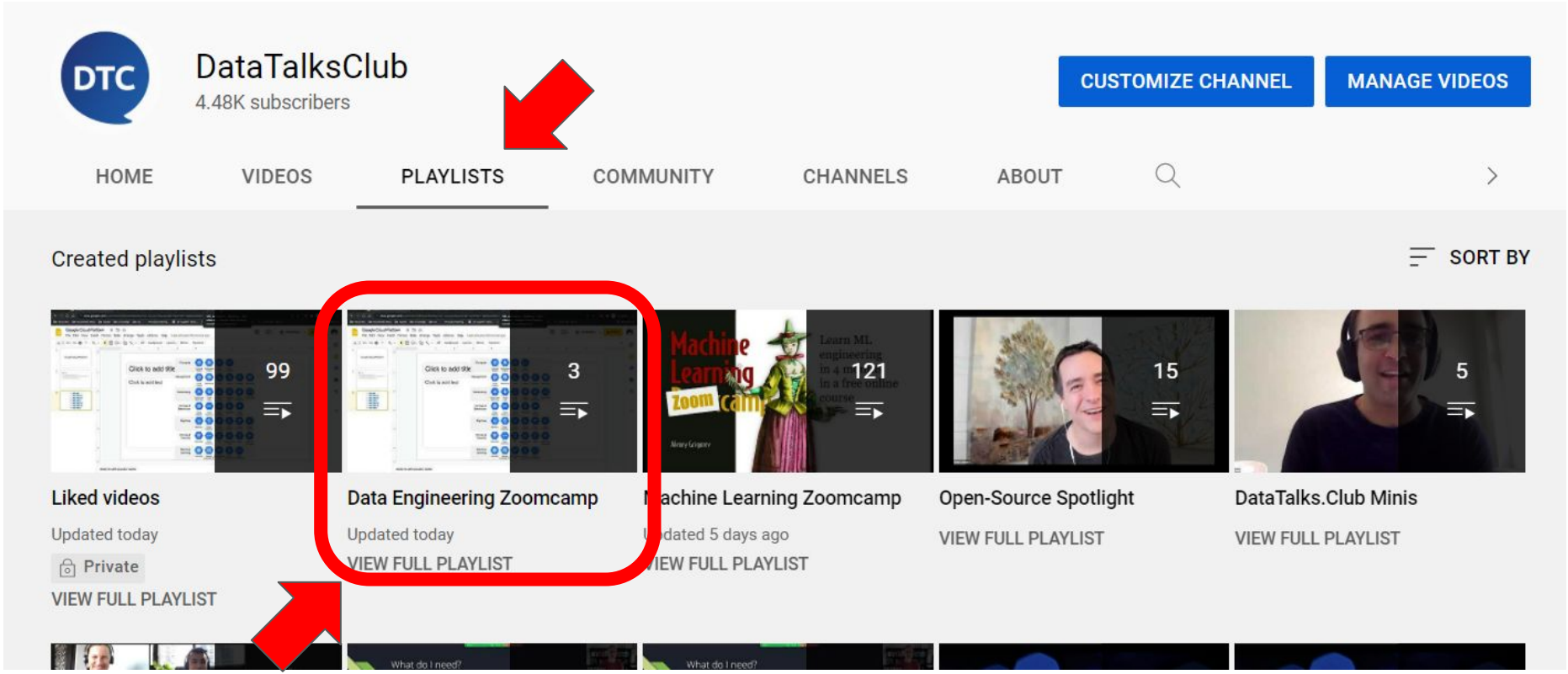
- Putting everything we learned in practice

Plan

- Instructors
- Course
  - Is it for me?
  - Syllabus
- **Course logistics**
- DataTalks.Club slack
- Questions

Course logistics

- Lessons
  - Pre-recorded
  - Published on our YouTube channel
- Office hours
  - Live on Mondays 17:00 CET
  - Homework solutions and answering questions
- Project
  - 1-2 weeks - working on the project
  - 3 week - peer reviewing



https://www.youtube.com/c/DataTalksClub

Course logistics

- Certificate - for passing the project
- Leaderboard
  - Scores for homework
  - Project
  - Learning in public

Learning in public

- LinkedIn
- Twitter
- Blogs

Thinam Tamang • 1st
Data Science | Machine Learning | Deep Learning | Practitioner
2w •

🏆 Day 251 of **#300DaysOfData**!

📑 Channels and Features :
Channels and Features are largely used interchangeably and refer to the size of the second axis of a weight matrix which is the number of activations per grid cell after a convolution. Channels refer to the input data i.e colors or activations inside the network. Using a stride 2 convolution often increases the number of Features at the same time because the number of activations in the activation map decrease by the factor of 4.

🎯 On my Journey of Machine Learning and Deep Learning, I have read and implemented from the book "Deep Learning for Coders with Fastai and PyTorch". Here, I have read about Convolutional Neural Network, Refactoring, Channels and Features, Understanding Convolution Arithmetic, Biases, Receptive Fields, Convolution over RGB Image, Stochastic Gradient Descent and few more topics related to the same from here.

📗 Notebook:
🔰 Convolutional Neural Network : **https://lnkd.in/dmdtzCzH**

📈 I have presented the implementation of Convolutional Neural Network and Training the Learner using Fastai and PyTorch here in the snapshot. I hope you will gain some insights and work on the same. I hope you will also spend some time learning the topics from the Book mentioned below. Excited about the days ahead !!

Learning in public links

Your answer

Plan

- Instructors
- Course
  - Is it for me?
  - Syllabus
- Course logistics
- **DataTalks.Club slack**
- Questions

Plan

- Instructors
- Course
  - Is it for me?
  - Syllabus
- Course logistics
- DataTalks.Club slack
- **Questions**

Join at

slido.com

#DEZ