

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

main

1 branch

0 tags

Go to file

Add file

Code

About



Code for Data Engineer Zoomcamp course

Readme

1.4k stars

87 watching

168 forks

Releases

No releases published

[Create a new release](#)

alexeygrigorev Update README.md

1f3aadb 1 hour ago 92 commits

images/architecture	Architecture diagram added with miro link.	2 months ago
project	cleanup-week1	2 days ago
week_1_basics_n_setup	code for sql+docker	yesterday
week_2_data_ingestion	cleanup-week1	2 days ago
week_3_data_warehouse	cleanup-week1	2 days ago
week_4_analytics_engin...	cleanup-week1	2 days ago
week_5_batch_processi...	cleanup-week1	2 days ago

main

1 branch

0 tags

Go to file

Add file

Code

About

Code for Data Engineer Zoomcamp course

Readme

1.4k stars

87 watching

168 forks

Releases

No releases published

[Create a new release](#)



alexeygrigorev Update README.md

1f3aadb 1 hour ago

92 commits



images/architecture

Architecture diagram added with miro link.

2 months ago



project

cleanup-week1

2 days ago



week_1_basics_n_setup

code for sql+docker

yesterday



week_2_data_ingestion

cleanup-week1

2 days ago



week_3_data_warehouse

cleanup-week1

2 days ago



week_4_analytics_engin...

cleanup-week1

2 days ago



week_5_batch_processi...

cleanup-week1

2 days ago



Taxi & Limousine Commission

Kreyòl Ayisyen ▶ Translate | ▼ Text-Size



About

Passengers

Drivers

Vehicles

Businesses

TLC Online

Search



About TLC

Data and Research

TLC Initiatives

Contact TLC

Data

Pilot Programs

Industry Reports

Factbook

[TLC Trip Record Data](#)

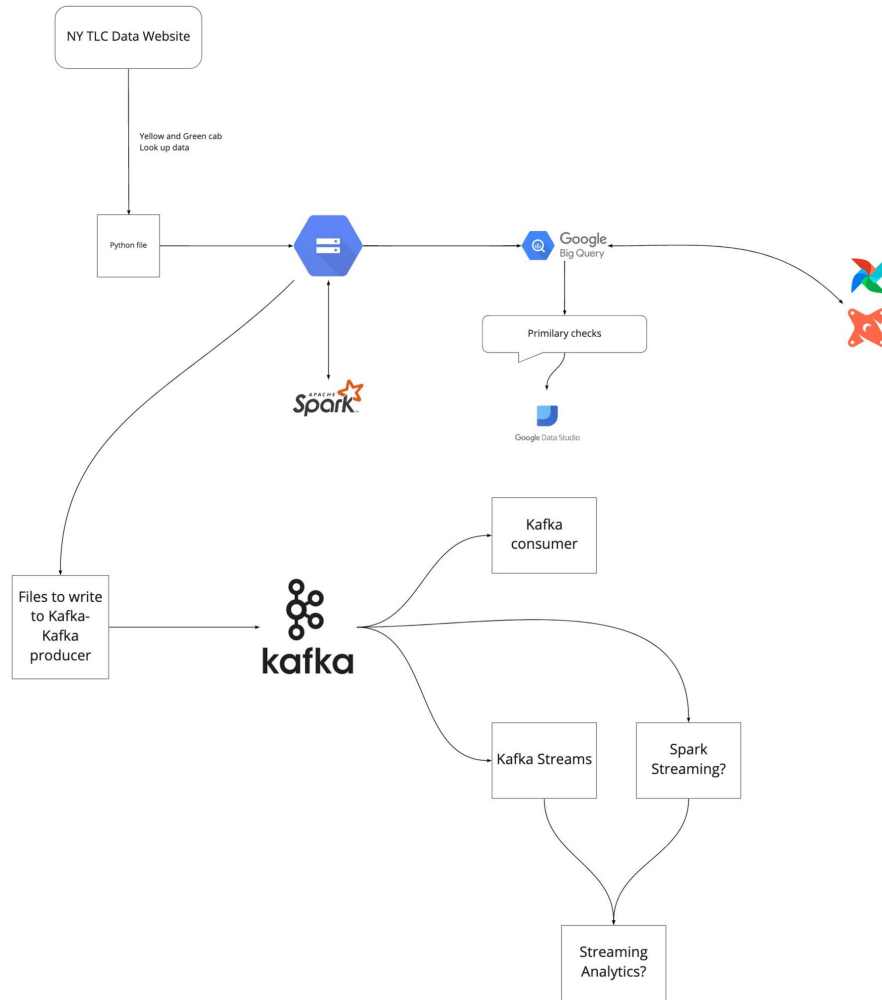
Request Data

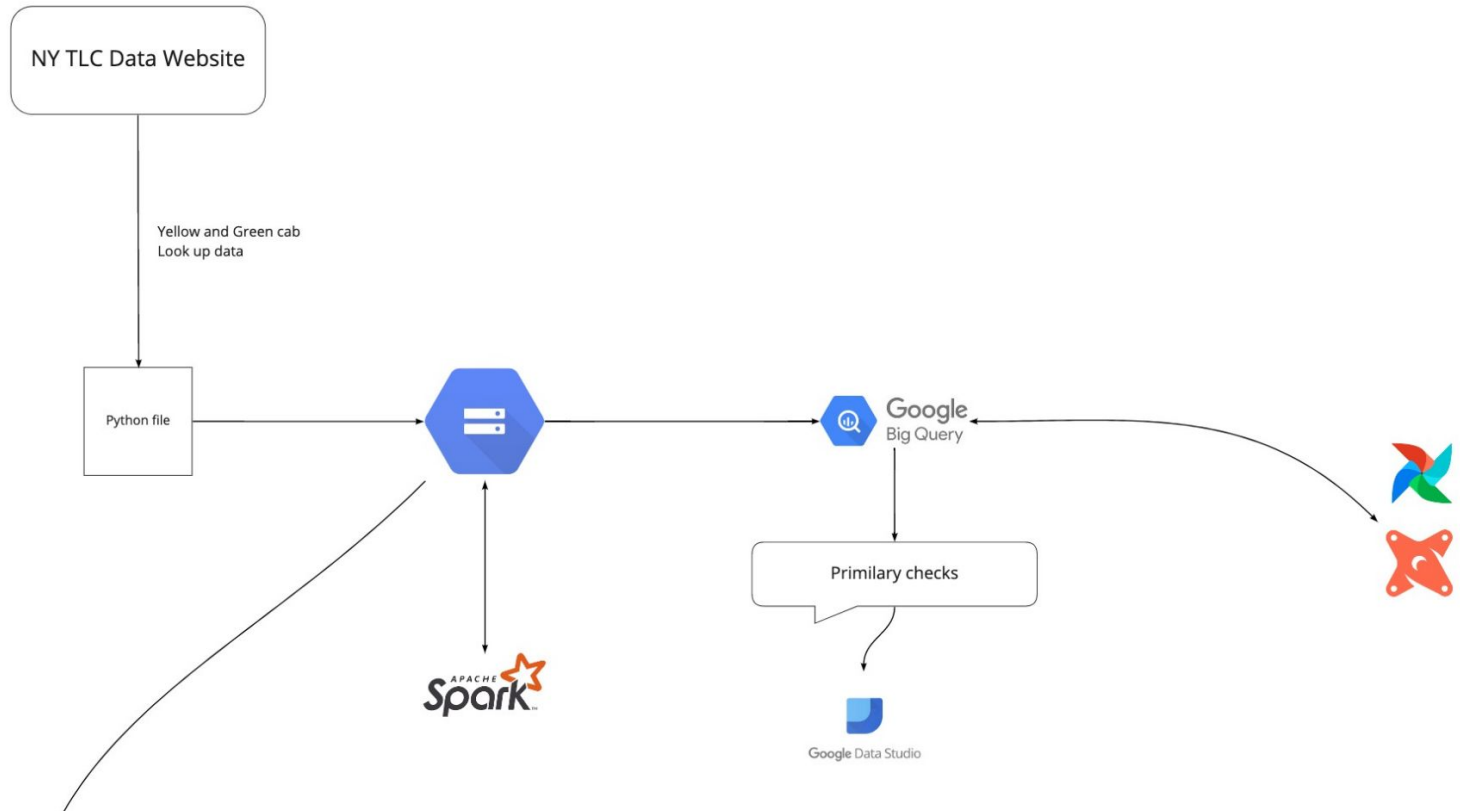
TLC Trip Record Data

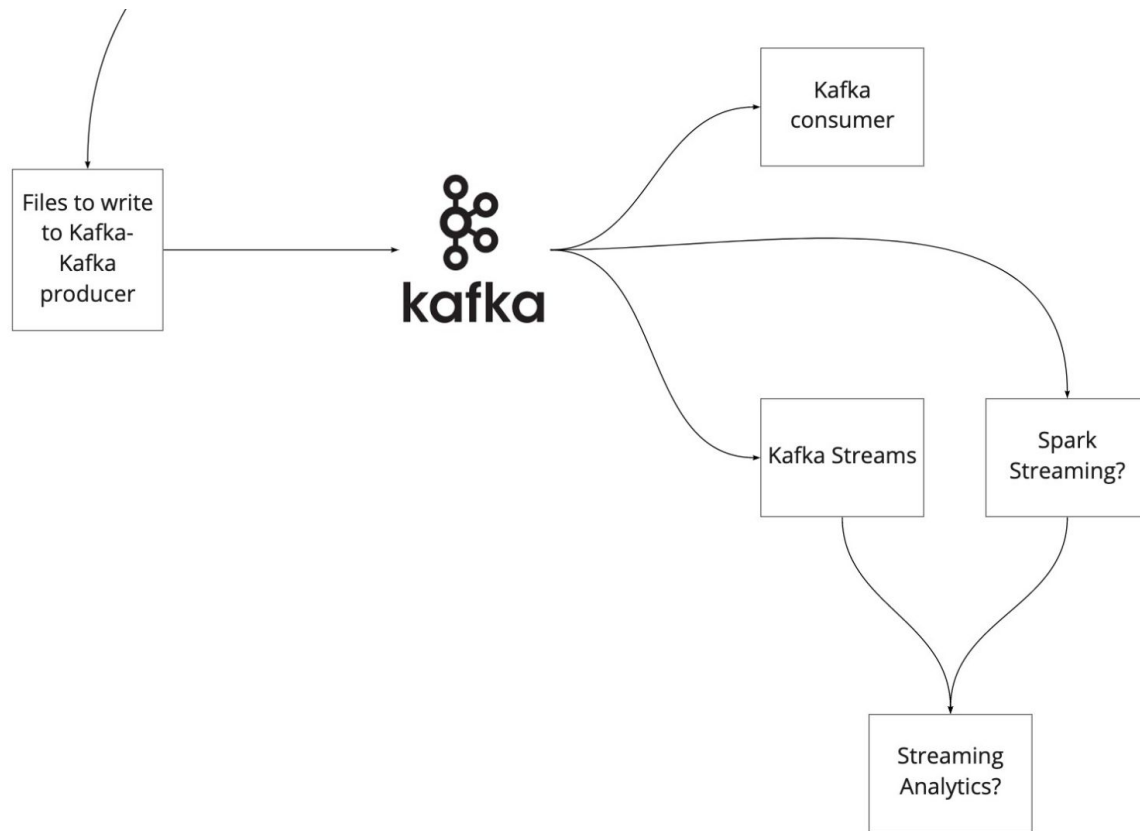
The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

The For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information.

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







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



DataTalksClub

4.48K subscribers

CUSTOMIZE CHANNEL

MANAGE VIDEOS

HOME

VIDEOS

PLAYLISTS

COMMUNITY

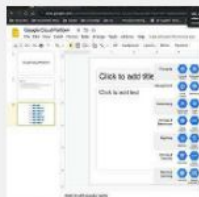
CHANNELS

ABOUT

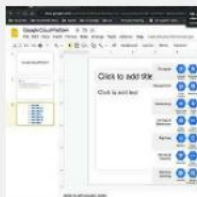


Created playlists

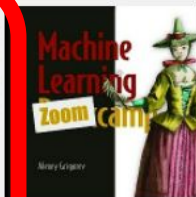
≡ SORT BY



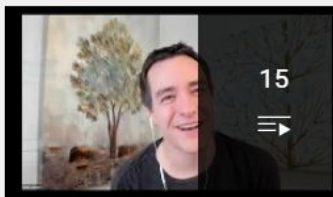
99



3



121



15



5



Liked videos

Updated today



Private

VIEW FULL PLAYLIST

Data Engineering Zoomcamp

Updated today

VIEW FULL PLAYLIST

Machine Learning Zoomcamp

Updated 5 days ago

VIEW FULL PLAYLIST

Open-Source Spotlight

VIEW FULL PLAYLIST

DataTalks.Club Minis

VIEW FULL PLAYLIST

<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

