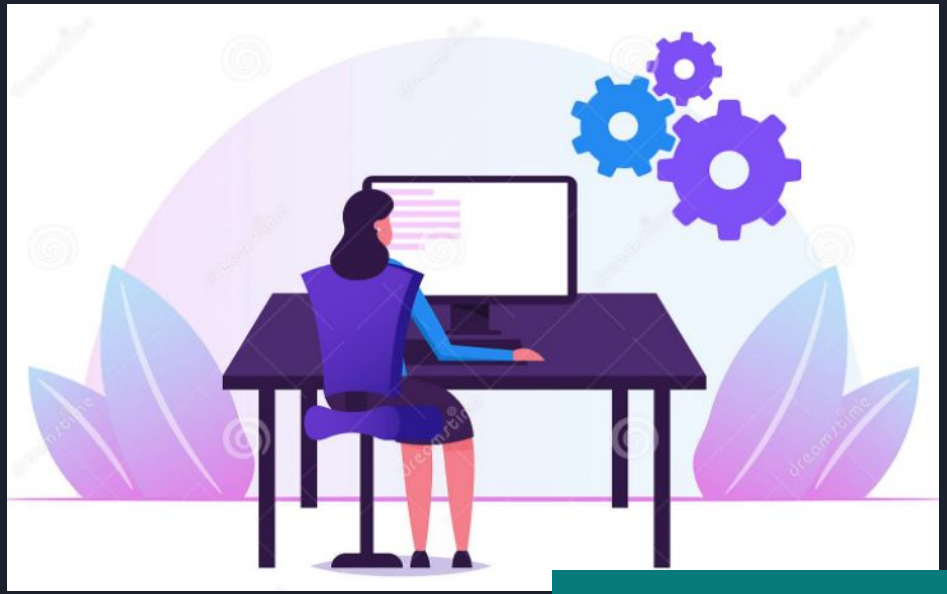
Welcome!

- We'll start in a moment :)
- We are NOT recording tonight's event. We may plan to take screenshots for social media.
    - ***If you want to remain anonymous***, change your name & keep video off.
- We'll introduce the hosts and break in-between for Q/A.
- We will make some time for Q&A at the end of the presentation as well.
- You can come prepared with questions. And, feel free to take notes.
- Online event best practices:
    - Don't multitask. Distractions reduce your ability to remember concepts.
    - Mute yourself when you aren't talking.
    - We want the session to be interactive.
    - Feel free to unmute and ask questions in the middle of the presentation.
    - Turn on your video if you feel comfortable.
    - Disclaimer: Speaker doesn't knows everything!

Check out:

- Technical Tracks and Digital Events
- Get updates – join the Digital mailing list
- Give us your feedback – take the Survey

WOMEN WHO CODE



WWCode Digital + **Backend Study Group**

January 20, 2022

WOMEN WHO CODE

Backend Study Group

- Welcome from WWCode!
- Our mission: Inspiring women to excel in technology careers.
- Our vision: A world where women are representative as technical executives, founders, VCs, board members and software engineers.

WOMEN WHO CODE

Harini Rajendran

Software Engineer, Confluent

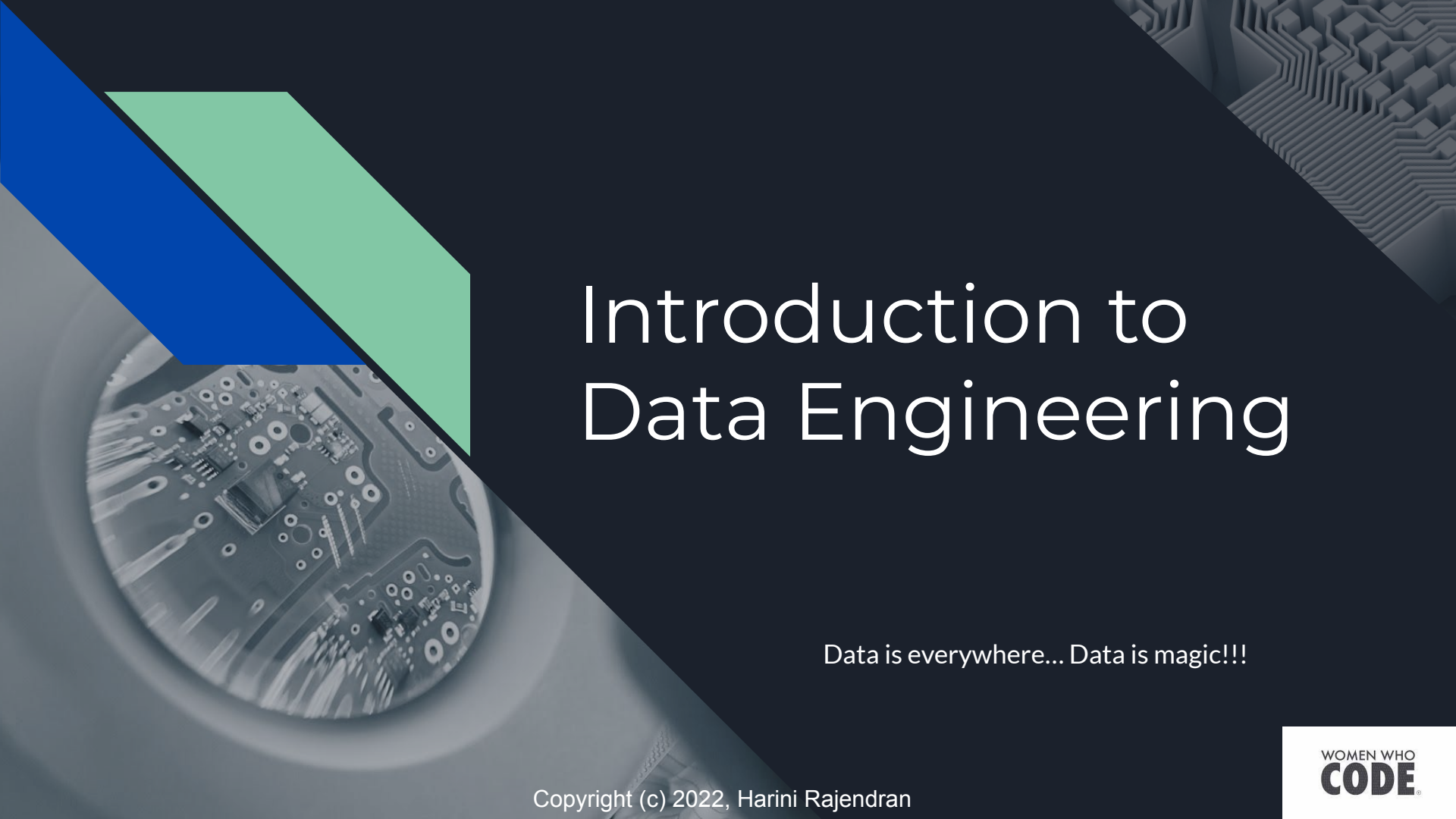
https://www.linkedin.com/in/hrajendran/

Prachi Shah

Director, Women Who Code San Francisco

https://www.linkedin.com/in/prachisshah/

WOMEN WHO CODE



Introduction to Data Engineering

Data is everywhere… Data is magic!!!

WOMEN WHO CODE

Agenda

Why is Data important

What is Data Engineering

What does a Data Engineer do

What technologies are used

Data Engineering vs Data Science

Skill sets of Data Engineer

Q & A

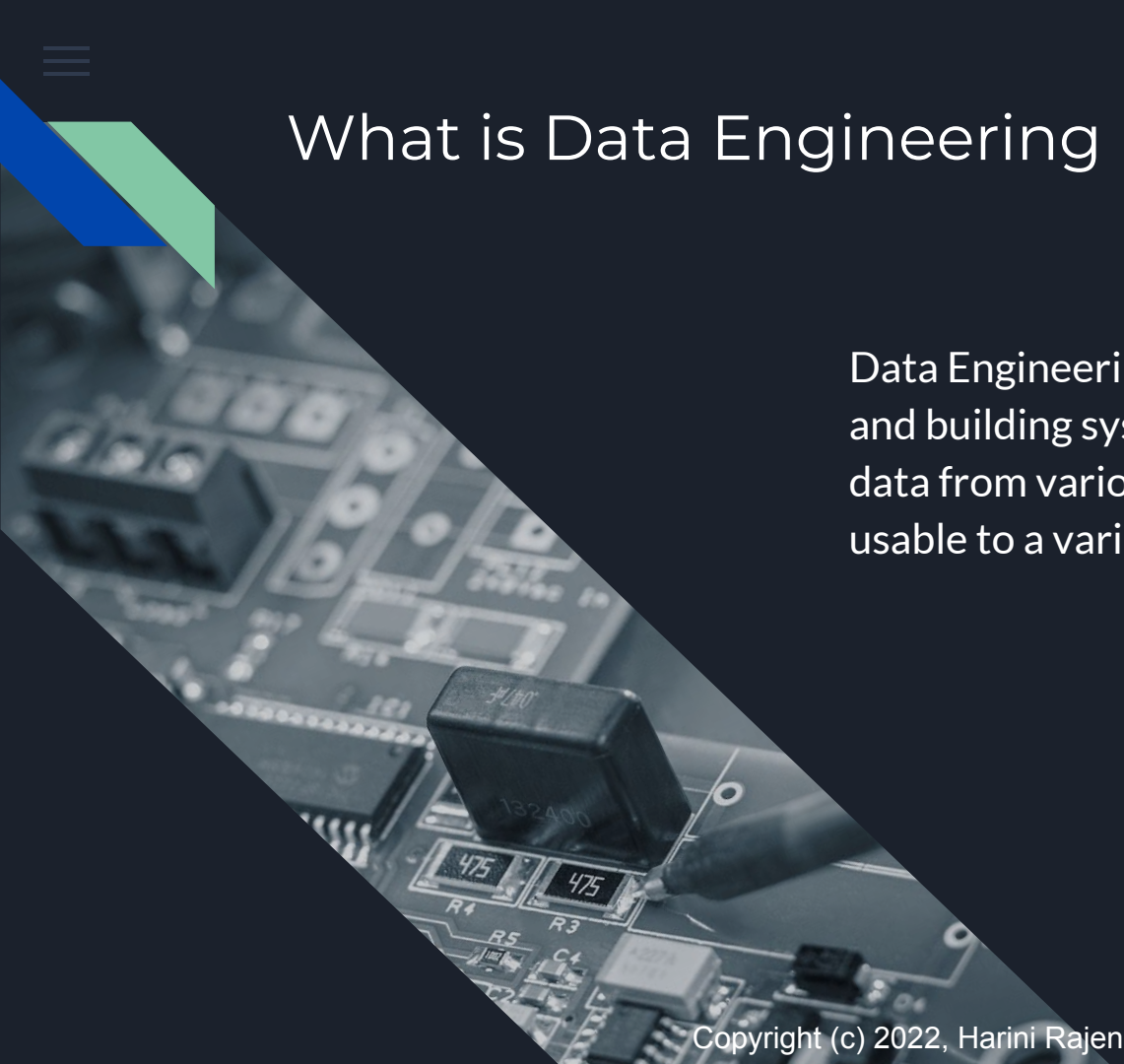
WOMEN WHO CODE

Why is Data important

Credit

WOMEN WHO CODE

What is Data Engineering



Data Engineering is the process of designing and building systems that would make raw data from various different data sources usable to a variety of stakeholders.

WOMEN WHO CODE

What is Data Engineering

Datasources

Mobile App

Website

Microservices

Stakeholders

Data Scientist

Analysts

Customers

Executives

WOMEN WHO CODE

What does a Data Engineer do

- Collecting, cleaning, organizing and exposing the data in a standardized format for consumption.
- Ensuring security and lifecycle for data
- Defining and managing data access policies (Data governance)
- Collaborating with different stakeholders like BI Analysts, Executives, Data Scientist, etc and help build solutions that solves their needs
- Making sure all the big data applications are healthy [This is a very crucial one and the most challenging]

What is Big Data

Large sets of diverse data generated in huge volumes at high speed

Characteristics of big data

- Volume
- Velocity
- Variety
- Veracity (Accuracy and Reliability)

Building big data pipelines

- Instrumentation
  - Instrumenting the services (web-site, app, microservices, etc) to produce the required data
- Ingestion
  - Gathering all the data instrumented in step 1 in a centralized place. This is the first part of the big data pipeline. Mostly big data technologies like kafka, s3, etc are used here.
- Processing
  - Gathered data is cleaned, transformed and processed for easy and fast retrieval. Distributed data processing platforms like hadoop, spark, etc are used here
- Storing
  - The processed data is stored for efficient retrieval. This is where distributed storage systems like s3, NoSQL databases, etc comes into picture

Building big data pipelines

- Access
  - Entities with relevant permissions should be able to access the data in an easy and straightforward manner. Data access should be fast and the retrieval time should mostly be in the order of milliseconds to under a few seconds based on the needs of the system. In case of batch data access, sub second latencies are not the norm.

Types of big data pipelines

- Stream Processing
  - Real-time processing on continuous stream of data. Data is processed as and when it is generated.

- Batch Processing
  - Data is batched in groups and processed periodically. This is not real-time.

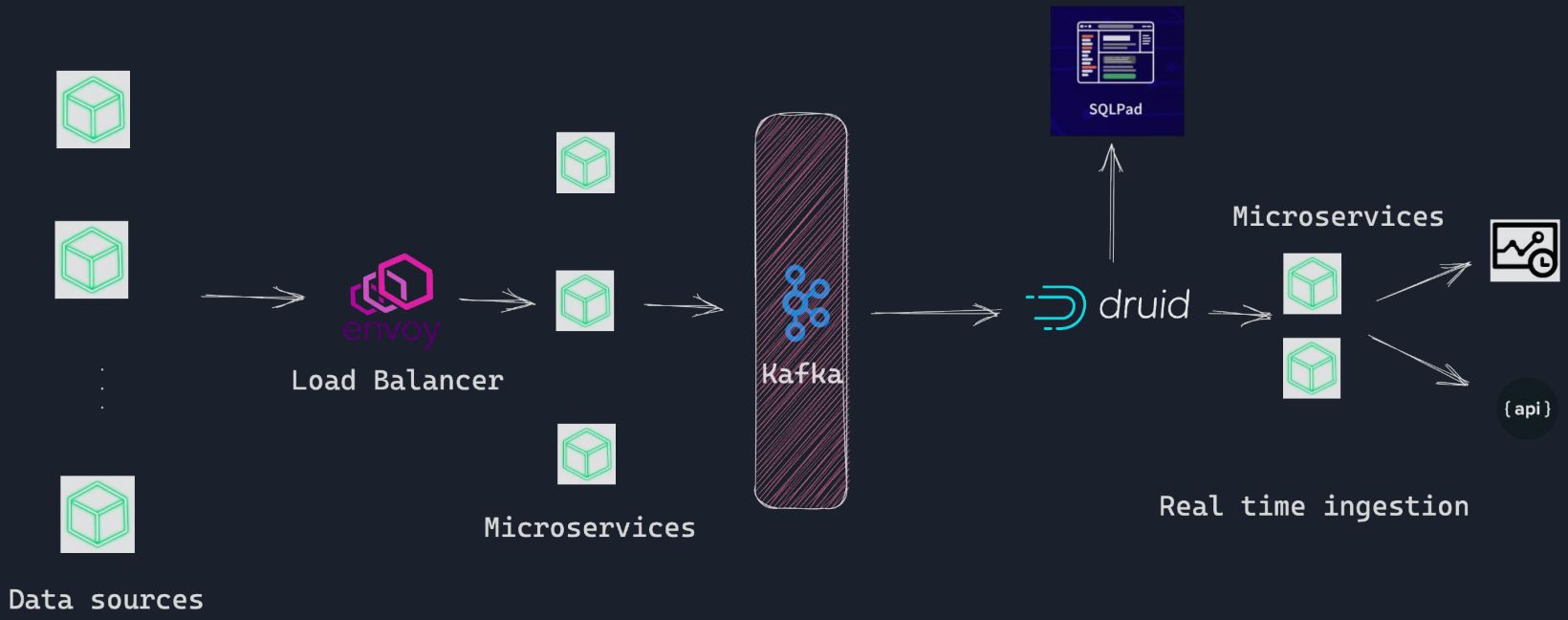
- Micro batch processing
  - Data is batched but the batches are so small that it feels as if it is stream processing

Technologies Used

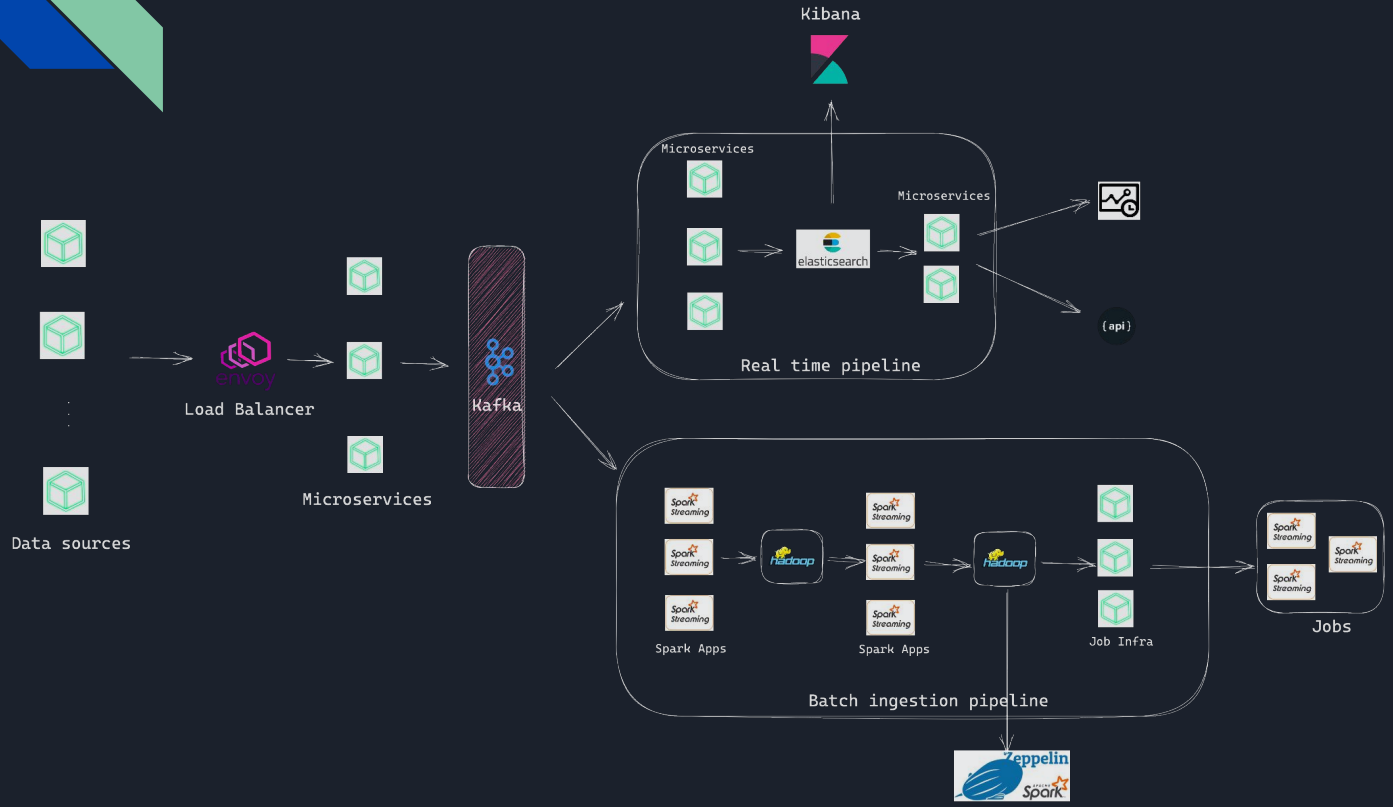
Stream Processing	Apache Kafka, Amazon Kinesis, Apache Spark, Apache Flink
Batch Processing	Hadoop MapReduce, Apache Spark(Micro batching)
Databases	Relational DB: Sql, Postgres Datastore:Amazon S3, Azure blob, HDFS NoSQL DB: MongoDB, Elastic Search, Cassandra, Couch DB, Influx DB, Apache Druid

WOMEN WHO CODE

Data Pipeline Example - 1



Data Pipeline Example - 2



WOMEN WHO CODE

Data Engineering vs Data Science

Data Engineer	Data Scientist
Work on raw data	Work on clean data provided by data engineers
Collect, clean, organize data	Perform analysis on clean data and find insights
Need strong distributed systems, data modeling and strong system designing skills.	Need in-depth knowledge on various ML and data mining techniques and understanding of statistics
Versed in Sql, NoSql, cloud and other big data tools and frameworks	Versed in Python, R, statistics and ML techniques
Collaborate with tech and non-tech stakeholders to build the data platform	Communicate the results of analysis to various tech and non-tech audience

Data Engineering Skill Sets

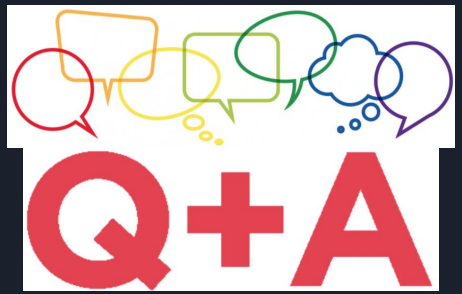
- Software Engineering best practices
- Programming languages
  - Python, Java/Scala
- Sound understanding of Distributed Systems design principles
- Data Modeling
- Big data ecosystem
  - Batch processing tools and frameworks
    - HDFS, Hadoop MapReduce, Pig, Hive
  - Real time processing tools and frameworks
    - Apache Kafka, Amazon Kinesis, Apache Spark, Apache Flink, etc
- Cloud platforms: AWS, GCP and Azure
- Databases
  - Relational Databases
    - MySQL, Postgres, Oracle Database
  - NoSQL Databases
    - Elastic Search, MongoDB, Cassandra, Druid, etc
- Soft Skills
  - Great team player
  - Strong communication skills
  - Decision making skills (sometimes with Incomplete data)
  - Critical thinking

WOMEN WHO CODE

Backend Study Group

WWCode Slack Handle: Harini Rajendran

https://www.linkedin.com/in/hrajendran/



Resources and References:

- https://quanthub.com/what-is-data-engineering/
- https://www.dremio.com/data-lake/data-engineering/
- https://www.precisely.com/glossary/data-engineering
- https://towardsdatascience.com/introduction-to-data-engineering-e16c9942dc2c

Backend Study Group:

• Presentations and session recordings found here: WWCode YouTube channel

You can unmute and talk or use the chat.

WOMEN WHO CODE