Welcome!

- We'll start in a moment :)
- We are 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**
Backend Study Group

July 29, 2021

Introduction & Agenda

- 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.

Prachi Shah
Senior Software Engineer @ Metromile

Madhurima Nath
Data Scientist @ Slalom

- What is Backend Engineering?
- **Insights into data engineering, data science and machine learning engineering**
  - Data engineering [Part 1 of 2]
  - **<u>Data science and machine learning engineering</u> [Part 2 of 2]**
    + Introduction
    + Similarities/Differences
    + Day in a life of DS, MLE
    + Tech stack

WOMEN WHO CODE

Backend Engineering

- What is Backend Engineering?
- Design, build and maintain server-side web applications.
- Concepts: Client-server architecture, API, micro-service, database engineering, distributed systems, storage, performance, deployment, availability, monitoring, etc.

Software Design

- Defining the architecture, modules, interfaces and data.
- Solve a problem or build a product.
- Define the input, output, business rules, data schema.
- Design patterns solve common problems.
- 3 Types:
  - UI design: Data visualization and presentation.
  - Data design: Data representation and storage.
  - Process design: Validation, manipulation and storage of data.

WOMEN WHO CODE

Data Engineer (DE) vs Data Scientist (DS) vs Machine Learning Engineer (MLE)

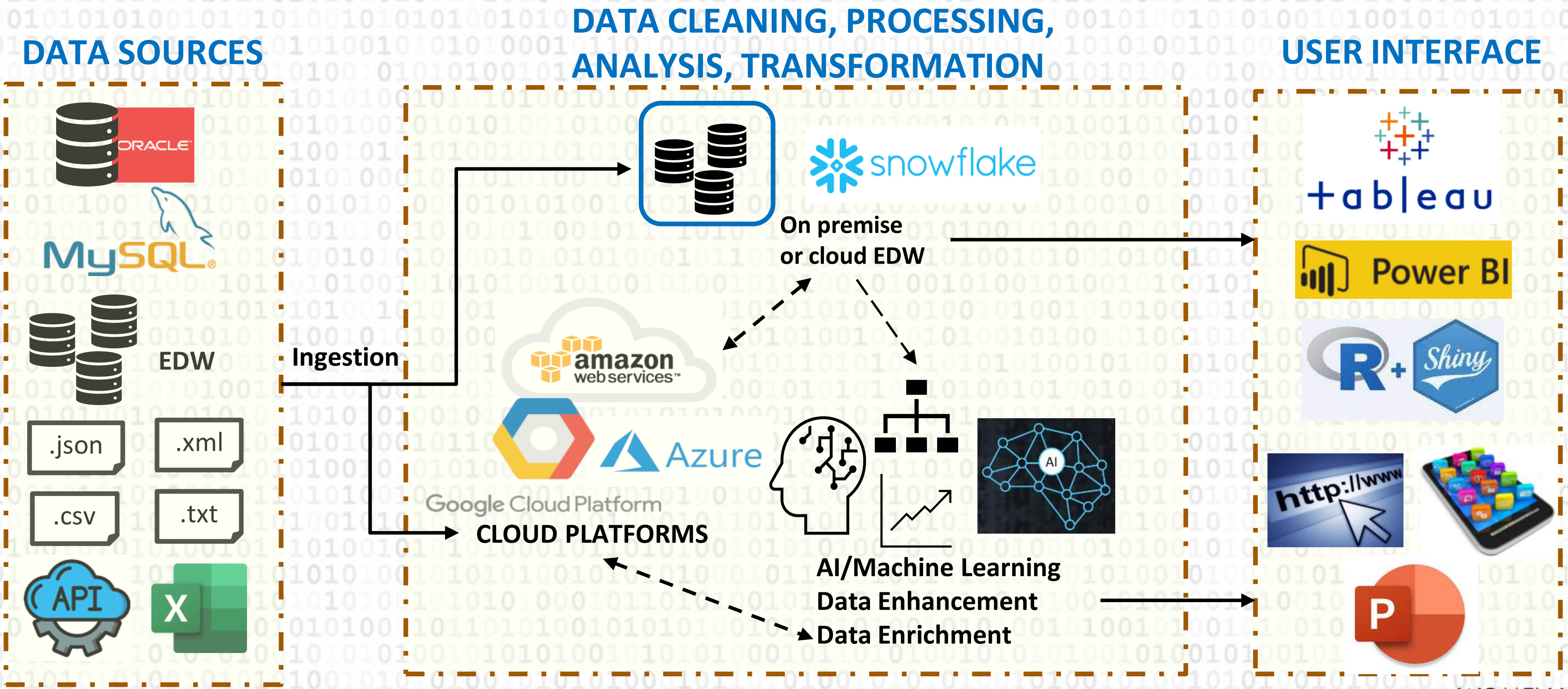
Data engineer:
builds and develops pipelines, and maintains of data infrastructure, either on-premises or in the cloud (or hybrid or multi-cloud), comprising of databases or data warehouses

Data scientist:
builds and develops mathematical and statistical models -- called machine learning models, to find patterns and gain more insights from the data

Machine learning engineer:
design architecture and pipelines (or software) to integrate and automate the process of running the models developed by data scientists with the entire infrastructure

WOMEN WHO CODE

Data Architecture Diagram



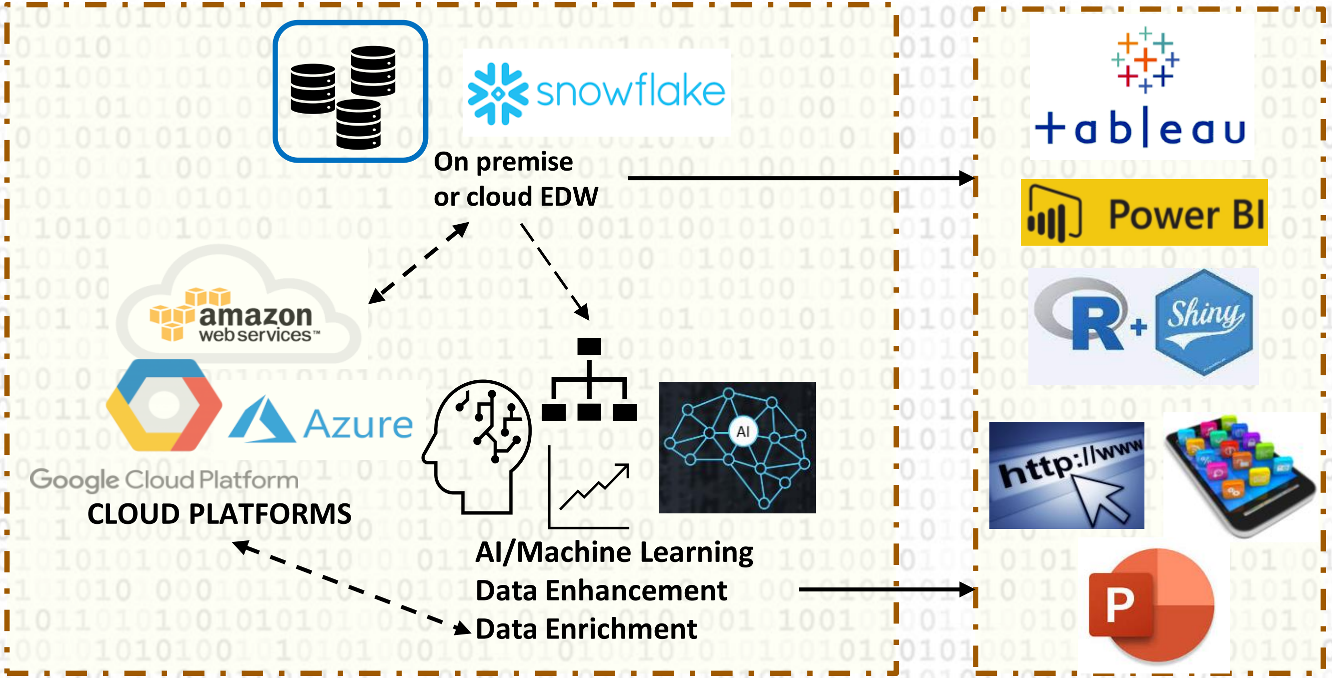
Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

Data Architecture Diagram – Data Scientist

DATA CLEANING, PROCESSING, ANALYSIS, TRANSFORMATION

USER INTERFACE

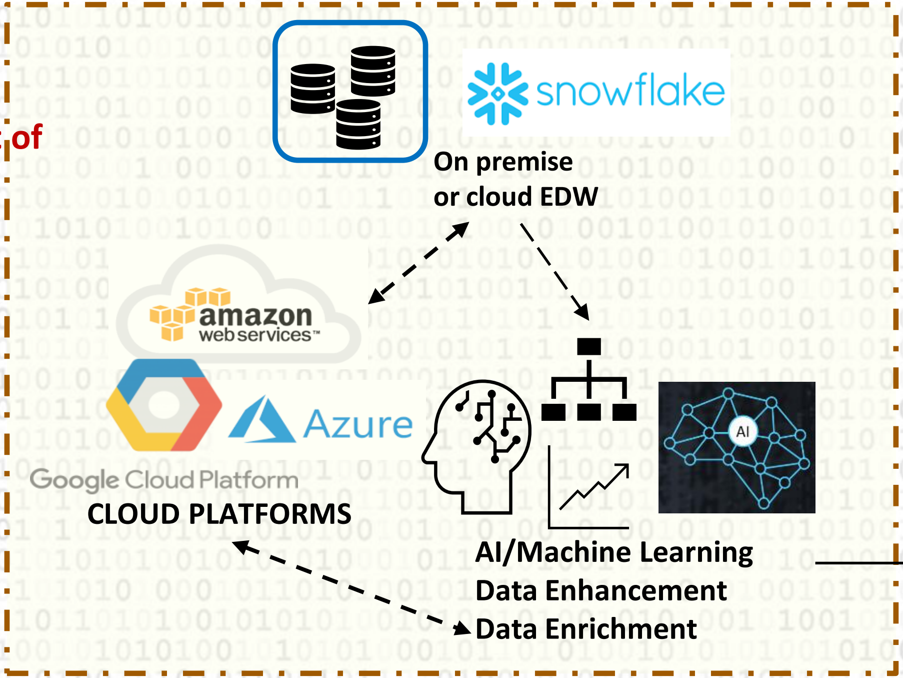


Disclaimer: These can change based on companies/industry

Data Architecture Diagram – ML Engineer

DATA CLEANING, PROCESSING, ANALYSIS, TRANSFORMATION

Automate and integrate with rest of the infrastructure



Data available for both analytics and visualization

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

How do Data Scientists (DS) and Machine Learning Engineers (MLE) differ from each other?

You can unmute and talk or use the chat.

How are they different/similar?

- Integration of the pipelines with the overall architecture

MLE: Take ML models and deploy it using automated pipelines

DS: Extensive data analysis, build ML models, perform statistical tests

- Data quality analysis
- SQL queries

- Test different ML models
- CI/CD*

*CI/CD: Continuous Integration Continuous Deployment

Disclaimer: These can change based on companies/industry

What is data science (DS) and machine learning engineering (MLE)?

<u>Data scientist (DS)</u>

- Finds patterns in the data **to obtain insights**
- Builds machine learning models to further enhance **understanding of data**

<u>Machine Learning Engineer (MLE)</u>

- Develops scripts to **integrate work of data scientists** with the larger framework
- **Automates** the work of data scientist such that models can be triggered to run without a data scientist

NOTE

Data engineers, data scientists and machine learning engineers make use of existing/pre-built modules or libraries or functions.

They write their own functions or custom codes as well; however, not same as a software engineer or software developer.

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

Tasks:

- convert business requirement into a data science/machine learning problem statement
- analyze of available data and data quality
- experiment and build the appropriate model and perform statistical tests
- visualize the outcomes (Python plots, simulations, R shiny, Tableau, Power BI, other custom dashboards)

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

Tasks:

- convert business requirement into a data science problem
- analyze of available data and data quality
- experiment and build the appropriate model and perform statistical tests
- visualize the outcomes

Which of these take up the most time?

You can unmute and talk or use the chat.

Disclaimer: These can change based on companies/industry

A Day in a Life of Data Scientist

Tasks:

- convert business requirement into a data science problem
- analyze of available data and data quality
- experiment and build the appropriate model and perform statistical tests
- visualize the outcomes

Time spent:

< 5%

~ 80%

- 80-85% data processing, analysis
  - best model with available data?
  - are the models doing, what is expected?
  - how long is training time?
  - model reusable?
- 15-20% coding on Jupyter/R notebook
  - use python/R libraries
  - build custom codes

~ 15%

Disclaimer: These can change based on companies/industry

A Day in a Life of Data Scientist

- convert business requirement into data science/machine learning problem

Examples of business requirements:

“We should give some incentives to our regular subscribers to stop them from leaving.”

“We think changing our app interface to blue when we are recommending new products would be great, everyone likes blue.”

“We need details on how we are spending, what products etc.”

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

- convert business requirement into data science/machine learning problem

Examples of business requirements:
"We should give some incentives to our regular subscribers to stop them from leaving."
"We think changing our app interface to blue when we are recommending new products would be great, everyone likes blue."
"We need details on how we are spending on what products, forecast future spending etc."

Can you identify what kind of data science/ machine learning problem these are?

You can unmute and talk or use the chat.

Disclaimer: These can change based on companies/industry

A Day in a Life of Data Scientist

- convert business requirement into data science/machine learning problem

Examples of business requirements:

“We should give some incentives to our regular subscribers to stop them from leaving.”
-- Classify users
-- Find what features important for subscribers
-- Find features contributing to losing membership

“We think changing our app interface to blue when we are recommending new products would be great, everyone likes blue.”

“We need details on how we are spending, what products, forecast future spending etc.”

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

- convert business requirement into data science/machine learning problem

Examples of business requirements:

“We should give some incentives to our regular subscribers to stop them from leaving.”

“We think changing our app interface to blue when we are recommending new products would be great, everyone likes blue.”
-- A/B tests

“We need details on how we are spending, what products, forecast future spending etc.”

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

- convert business requirement into data science/machine learning problem

Examples of business requirements:

"We should give some incentives to our regular subscribers to stop them from leaving."

"We think changing our app interface to blue when we are recommending new products would be great, everyone likes blue."

"We need details on how we are spending, what products, forecast future spending etc."
-- classify products into groups
-- time series of spending trends
-- time series for forecasting

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

- analyze of available data and data quality

From data ingested by data engineers:

- Check if data is enough to do the job
  - if yes, good
  - if no, find additional data – open source, other data sources
- Check data quality issues
  - nulls, missing, data formats
  - qualitative data, quantitative data, text data, image data

Qualitative data – non-numeric, e.g., name, state, yes/no responses
Quantitative data – numeric, e.g., price, temperature

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

- experiment and build the appropriate model and perform statistical tests

Example:
"We need details on how we are spending, what products, forecast future spending etc."
-- classify products into groups
-- time series of spending trends

How would you experiment/build a classification model for this requirement?

You can unmute and talk or use the chat.

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Data Scientist

- experiment and build the appropriate model and perform statistical tests

Example:
"We need details on how we are spending, what products, forecast future spending etc."
-- classify products into groups
-- time series of spending trends

- **Classification**:

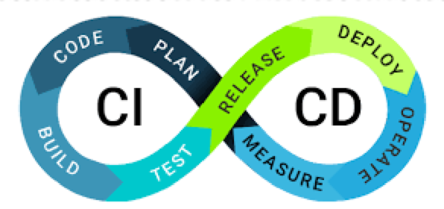
Q. Is the data quality good to start building a classification model?
Q. Is this a text classification problem?
Q. Is this a supervised or unsupervised, i.e., can I provide examples to train or not?
Q. Do we know what would be the target classification groups? – Yes, supervised?
Q. Should we try some unsupervised methods like clustering?
Q. Which supervised or unsupervised models to use for classification?
Q. Is this a binary classification or multi-class classification?
Q. How to represent the model output for the end users?

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

Tech stack for data scientist

Languages: **Python/R**, **SQL/NoSQL**, Scala
Jupyter/R notebooks, CI/CD* framework



*CI/CD: Continuous Integration Continuous Deployment

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

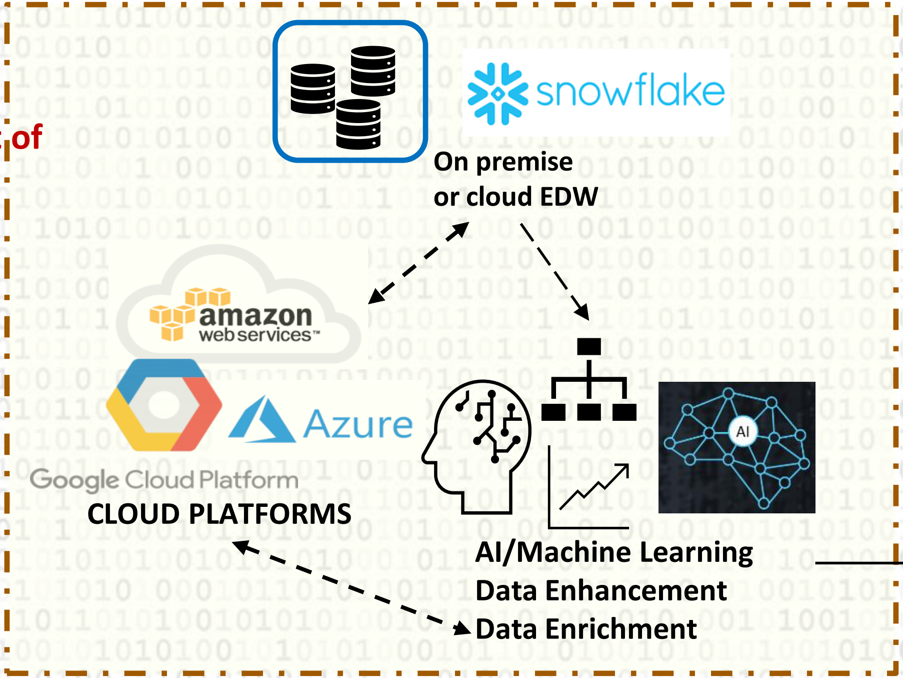
Data Scientist Role

WOMEN WHO CODE

Data Architecture Diagram – ML Engineer

DATA CLEANING, PROCESSING, ANALYSIS, TRANSFORMATION

Automate and integrate with rest of the infrastructure



Data available for both analytics and visualization

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

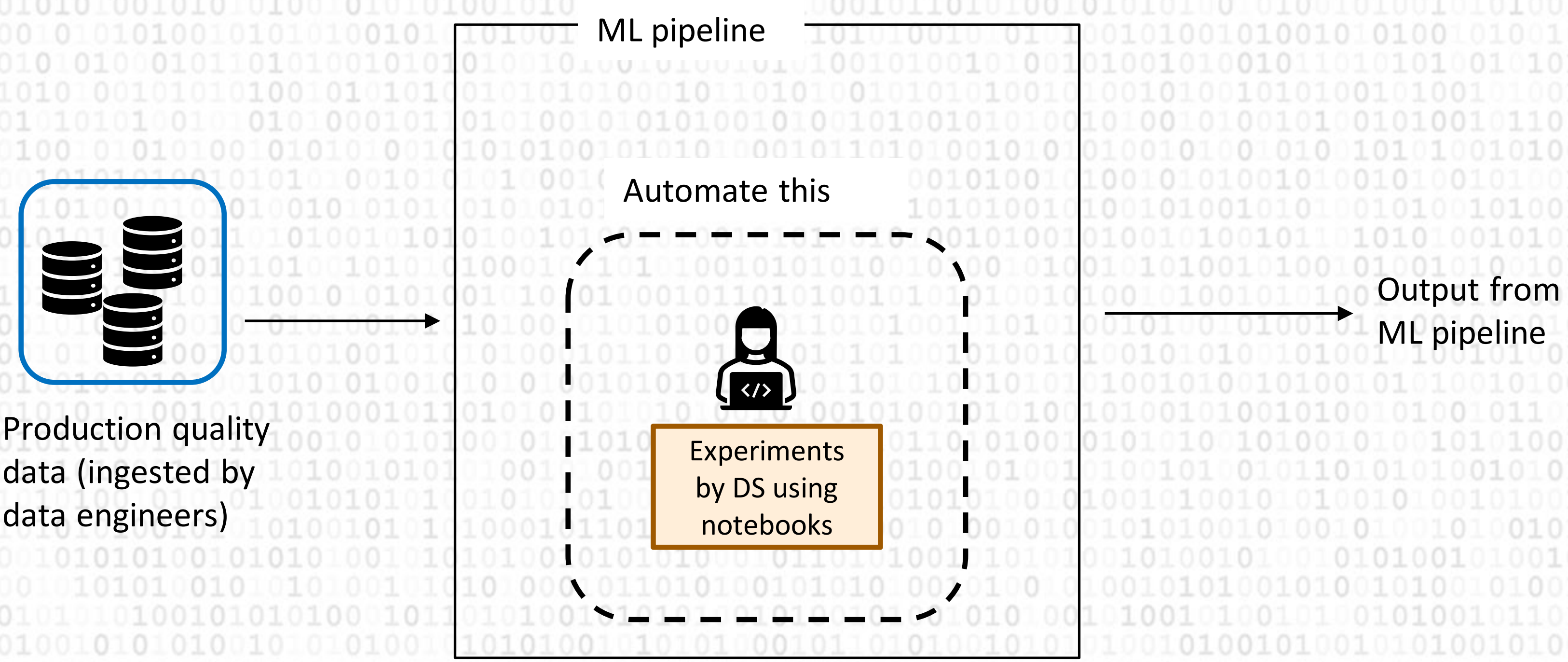
Tasks:

- design pipelines to integrate the ML pipeline with the larger infrastructure
- provision workspaces for data scientist to develop and deploy models
- automate work of data scientist

Disclaimer: These can change based on companies/industry

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure



Disclaimer: These can change based on companies/industry

A Day in a Life of Machine Learning Engineer

- provision workspaces for data scientist to develop and deploy models and automate

What are the steps necessary for model deployment and automation? Thoughts?

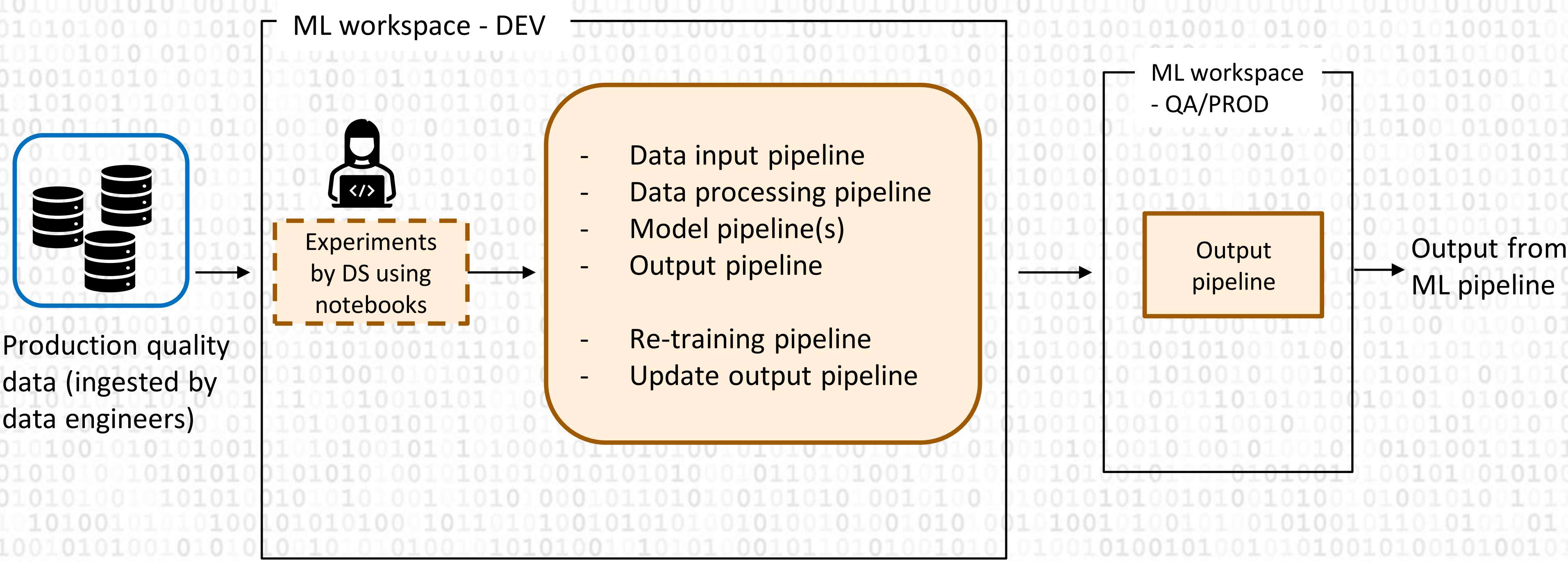
You can unmute and talk or use the chat.

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

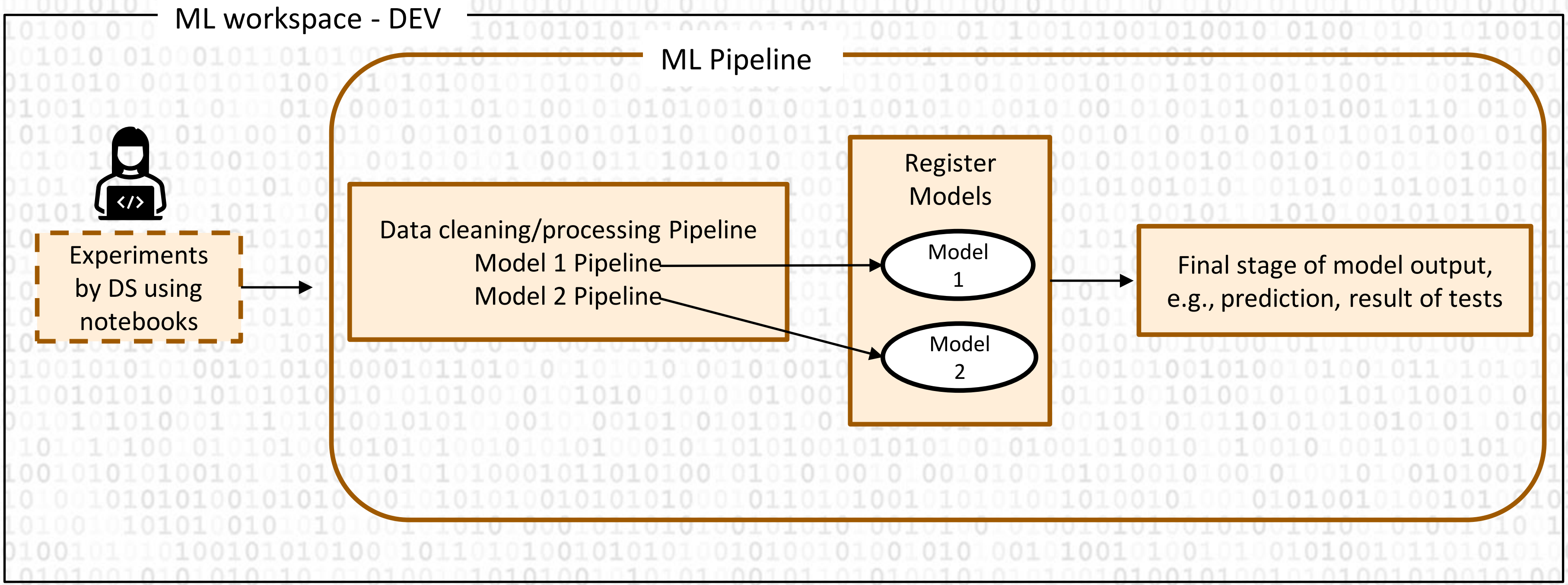


Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

- provision workspaces for data scientist to develop and deploy models and automate



Disclaimer: These can change based on companies/industry

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

MLOps Directory Structure

Directory	Description
cicd	CICD pipeline files (configuration files)
experiments	jupyter notebook (other files) for experimentation
infra	infrastructure-as-code: DevOps template files
integration_test	scripts and files for integration testing
mlops	MLOps package source and unit tests
workspace	workspace resources scripts and configuration files

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

Directory	Description
cicd	CICD pipeline files (configuration files)
experiments	jupyter notebook (other files) for experimentation
infra	infrastructure-as-code: DevOps template files
integration_test	scripts and files for integration testing
mlops	MLOps package source and unit tests
workspace	workspace resources scripts and configuration files

cicd:
- all configuration files
- setting up DEV/QA/PROD environments
- deploying DEV → QA → PROD

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

Directory	Description
cicd	CICD pipeline files (configuration files)
experiments	jupyter notebook (other files) for experimentation
infra	infrastructure-as-code: DevOps template files
integration_test	scripts and files for integration testing
mlops	MLOps package source and unit tests
workspace	workspace resources scripts and configuration files

experiments:
all notebooks (Python or R)
used by DS

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

Directory	Description
cicd	CICD pipeline files (configuration files)
experiments	jupyter notebook (other files) for experimentation
infra	infrastructure-as-code: DevOps template files
integration_test	scripts and files for integration testing
mlops	MLOps package source and unit tests
workspace	workspace resources scripts and configuration files

infra:
- all DevOps files that will manage infrastructure with configuration files rather than through a graphical user interface
- this makes things faster by eliminating manual processes

Disclaimer: These can change based on companies/industry

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

Directory	Description
cicd	CICD pipeline files (configuration files)
experiments	jupyter notebook (other files) for experimentation
infra	infrastructure-as-code: DevOps template files
integration_test	scripts and files for integration testing
mlops	MLOps package source and unit tests
workspace	workspace resources scripts and configuration files

integration_test:
python scripts to integrate ML pipeline with the larger architecture

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

Directory	Description
cicd	CICD pipeline files (configuration files)
experiments	jupyter notebook (other files) for experimentation
infra	infrastructure-as-code: DevOps template files
integration_test	scripts and files for integration testing
mlops	MLOps package source and unit tests
workspace	workspace resources scripts and configuration files

mlops:
- all custom-built functions/modules/packages to be used by the model
- all unit test cases and files

Disclaimer: These can change based on companies/industry

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

Directory	Description
cicd	CICD pipeline files (configuration files)
experiments	jupyter notebook (other files) for experimentation
infra	infrastructure-as-code: DevOps template files
integration_test	scripts and files for integration testing
mlops	MLOps package source and unit tests
workspace	workspace resources scripts and configuration files

workspace:
- all steps used by the DS
- files to run these steps connected way
- files to run these as API calls using endpoints

Disclaimer: These can change based on companies/industry

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

workspace directory

```
└── workspace
    ├── pipelines
    │   ├── deploy_as_endpoint.py
    │   ├── run_pipeline.py
    │   └── template
    │       ├── endpoint.json
    │       ├── requirements.txt
    │       └── script.py
    ├── workspace.py
    └── workspace.json
```

Experiments by DS using notebooks →

- Data input pipeline
- Data processing pipeline
- Model pipeline(s)
- Output pipeline

- Re-training pipeline
- Update output pipeline

Disclaimer: These can change based on companies/industry

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

ML workspace - DEV

workspace directory

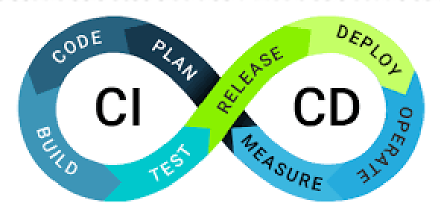
```
└── workspace
    ├── pipelines
    │   ├── deploy_as_endpoint.py   → run the pipeline using endpoint (API call)
    │   ├── run_pipeline.py         → run connected pipeline on command line
    │   └── template                → generalized template directory (follow these for each pipeline)
    │       ├── endpoint.json       → details on steps in the pipelines, input/output parameters etc.
    │       ├── requirements.txt    → all the pre-defined functions/libraries required to run a pipeline
    │       └── script.py           → code that DS used
    ├── workspace.py                → code to set up workspace, what datasets, resources etc. to use
    └── workspace.json              → details on steps for workspace, defined parameters
```

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

Tech stack for machine learning engineer

Languages: **Python/R**, **SQL/NoSQL**, Scala
Jupyter/R notebooks, CI/CD* framework
Shell/bash scripting, command line



*CI/CD: Continuous Integration Continuous Deployment

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE

Backend Study Group

- WWCode Presentation and Demo
- WWCode YouTube channel:
  - March 25, 2021 session recording: Backend Engineering
  - April 8, 2021 session recording: Java Microservice and REST API Demo
  - April 22, 2021 session recording: Creational Design Patterns
  - May 20, 2021 session recording: Structural Design Patterns
  - June 3, 2021 session recording: Behavioral Design Patterns

- **Resources**:

mlops on azure, mlops on aws, mlops on gcp
free online resources to learn mlops
coursera mlops course
mlops infrastructure page
ml aws certification, mle gcp certification
azure ds associate, azure ai fundamentals certification, azure ai engineer certification

WOMEN WHO CODE

A Day in a Life of Machine Learning Engineer

- design pipelines to integrate the ML pipeline with the larger infrastructure

workspace.json

```
{ "workspace":
  {
    "name": "wwcodesf-$:environment-mlworkspace",
    "environment": "$:environment",
    ...
    ...,
    "data_storage": [
      {
        "name": "some_name_$:environment",
        "attributes": {
          "account_name": "some_account_name",
          "account_key": "$:storage_account_key",
          "create_if_not_exists": true
        },
        "options": {
          "create_in_environment": [
            "dev", "qa", "prod"
          ]
        }
      }
    ],
```

```
"mlenvironment": [
  {
    "name": "name_env",
    "attributes": {
      "pip_packages": [
        "package1",
        "package2",
        "pandas"
      ],
      "pip_wheel_paths": [
        "$CustomFunctionPath"
      ]
    }
  },
  {
    "name": "model_training_env",
    "attributes": ....
  },
  {
    "name": "model_predict_env",
    "attributes": ....
  },
],
```

```
    "compute_resources": [
      {
        "name": "endpoint1",
        "compute_type": "type_of_compute_resource",
        "attributes": {
        "size": "large",
        "num_nodes": 2
        }
      }
    ]
  }

}
```

Disclaimer: These can change based on companies/industry

WOMEN WHO CODE