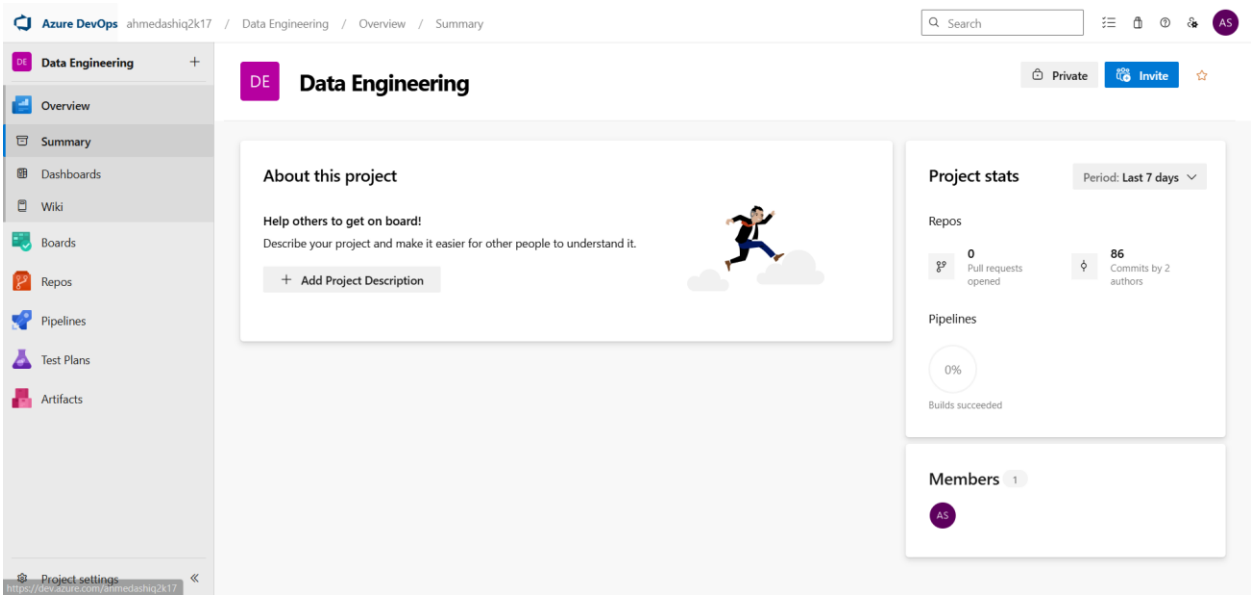
Week 5 – Automation via Azure DevOps

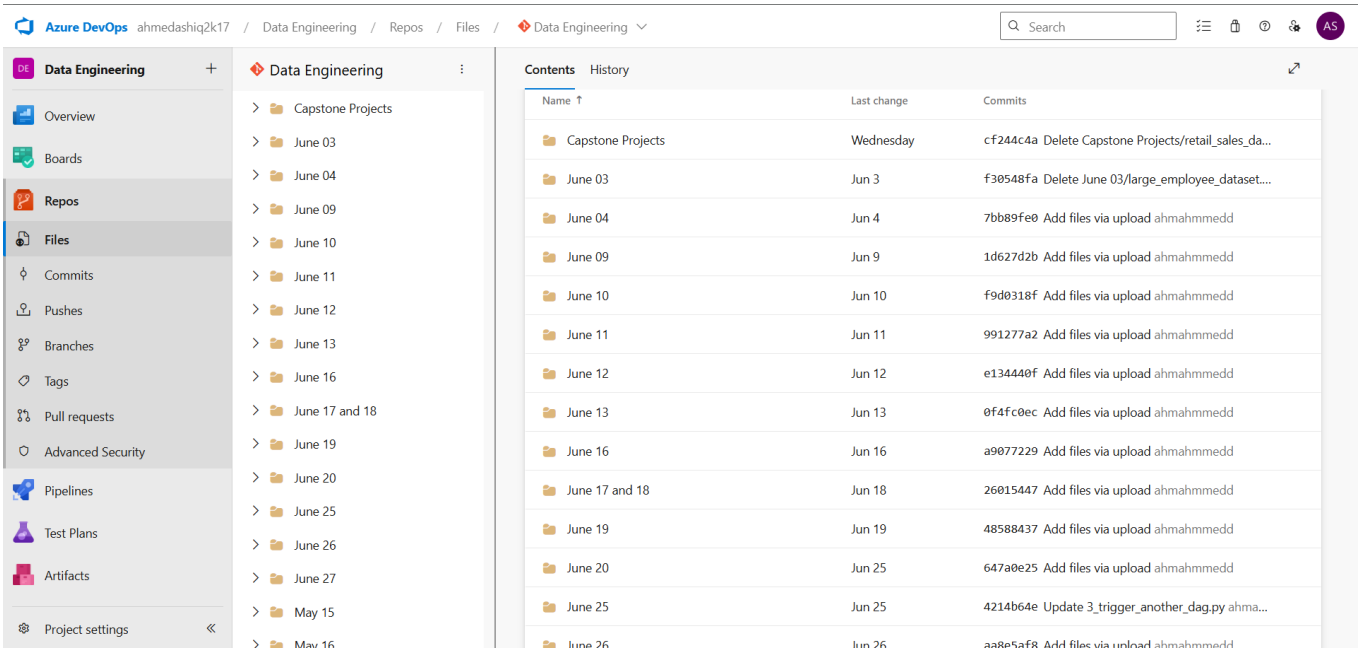
Step 1: Create a New DevOps Project:

1. **New Project -> Insert Name -> Create**



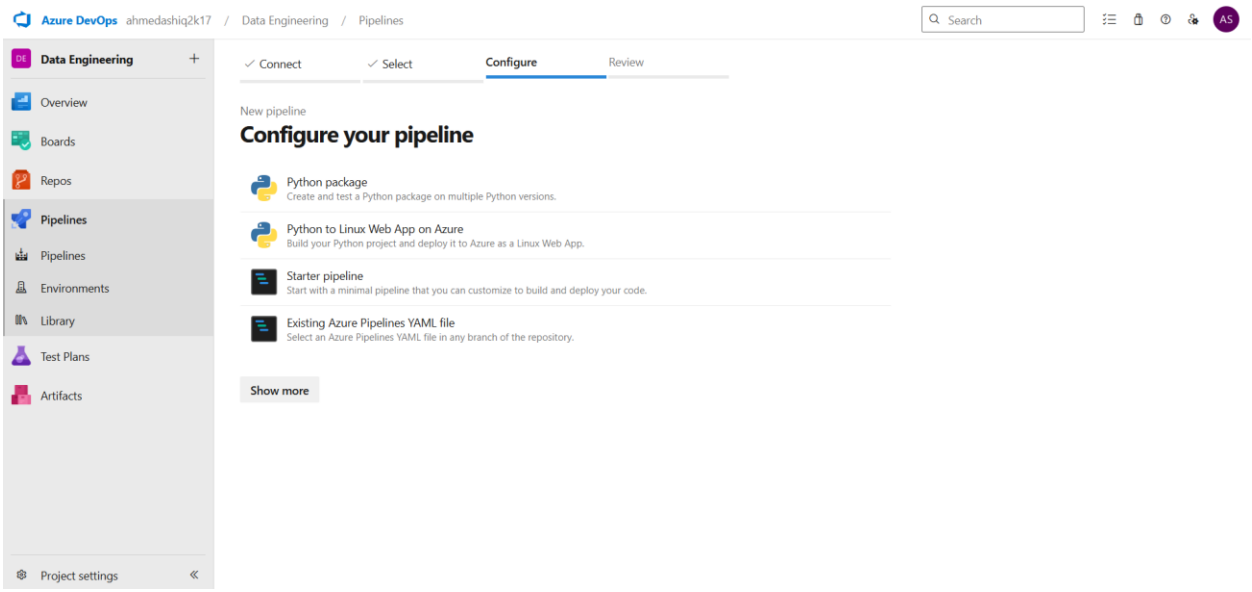
Step 2: Import repo into Azure DevOps

1. **Repos -> Import -> Insert Git link -> Import**



Step 3: Create Pipeline

1. **Pipelines → Create Pipeline -> Azure Repos Git -> Choose Repo -> Python Package**



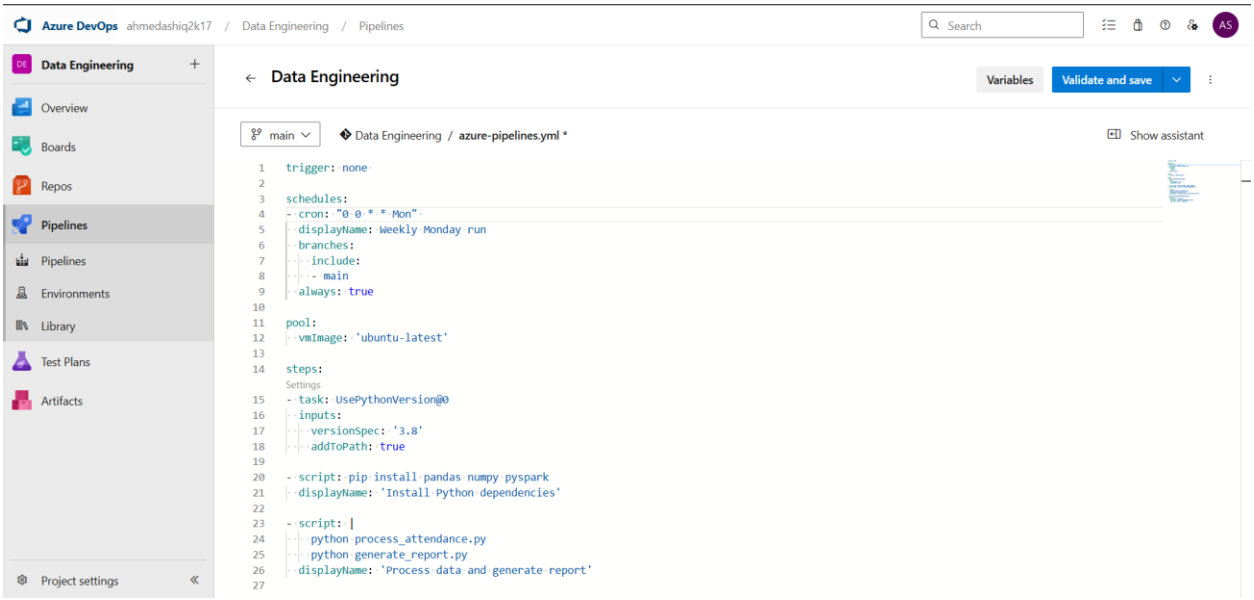
Step 4: Insert the yaml file code:

```
trigger: none


schedules:
- cron: "0 0 * * Mon"
  displayName: Weekly Monday run
  branches:
    include:
    - main
  always: true
```

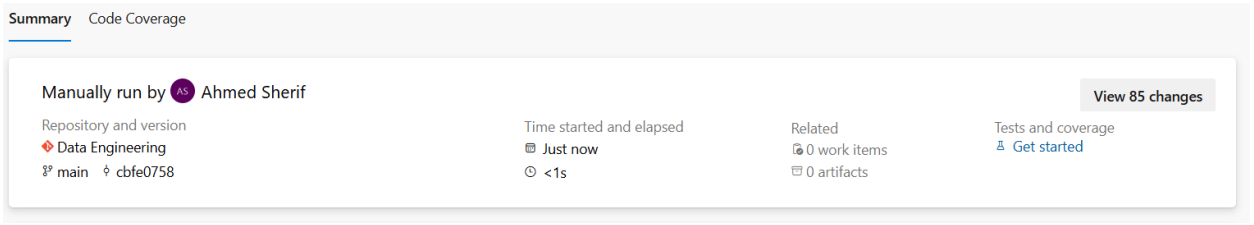
```
pool:
  vmImage: 'ubuntu-latest'

steps:
- task: UsePythonVersion@0
  inputs:
    versionSpec: '3.8'
    addToPath: true

- script: pip install pandas numpy pyspark
  displayName: 'Install Python dependencies'

- script: |
    python process_attendance.py
    python generate_report.py
  displayName: 'Process data and generate report'

- task: PublishPipelineArtifact@1
  inputs:
    targetPath: 'reports'
    artifact: 'weekly_attendance_report'
    publishLocation: 'pipeline'
```

Step 5: Run the yaml code

1. **Validate and save -> Run**



Step 6: Summary of Pipeline



Capstone Tasks:

- Set up a DevOps pipeline to automate weekly processing

```
schedules:
- cron: "0 0 * * Mon"
  displayName: Weekly Monday run
  branches:
    include:
    - main
  always: true
```

- Schedule the pipeline to run every Monday

```
schedules:
- cron: "0 0 * * Mon"
  displayName: Weekly Monday run
  branches:
    include:
    - main
  always: true
```

- Output a report with top 5 absentees or lowest performing departments

```
# 1. Top 5 Absentees
top_absentees = cleaned_df.filter(col("Clock_out").isNull()) \
    .groupBy("Employee_ID") \
    .agg(count("*").alias("Absence_Days")) \
    .orderBy(desc("Absence_Days")) \
    .limit(5)
top_absentees.show()


# 2. Lowest Performing Departments
dept_performance = cleaned_df.groupBy("Department") \
    .agg(avg("work_hours").alias("Avg_Hours"),
         avg(when(col("Status") == "completed",
1).otherwise(0)).alias("Completion_Rate")) \
    .orderBy("Completion_Rate") \
    .limit(5)
dept_performance.show()
```