Assignment Report: Running Apache Beam Locally on Google Colab

Student Name: Sakina Banu

Platform Used: Google Colab

Assignment Objective: Use Apache Beam with Python to read a local file, filter for employees in the "Accounts" department, and count their attendance days.

# Step-by-Step Execution

## Step 1: Check PIP Version

Command:  
!pip --version

Purpose: Verifies the package manager is available and up-to-date.

## Step 2: Install Apache Beam

Commands:  
!pip install apache-beam[gcp]  
!pip install apache-beam

This installs the core Apache Beam SDK.

## Step 3: Create Input File

Command:  
with open('input.txt', 'w') as f:  
 f.write("""...""")

This file mimics structured data from a company’s attendance log.

## Step 4: Apache Beam Pipeline Code

This script reads the file, filters for “Accounts” department, groups and counts attendance per employee, and writes results.

import apache\_beam as beam  
from apache\_beam.options.pipeline\_options import PipelineOptions  
  
input\_file = 'input.txt'  
output\_prefix = 'output/results'  
  
with beam.Pipeline(options=PipelineOptions()) as p:  
 (  
 p  
 | 'Read lines' >> beam.io.ReadFromText(input\_file)  
 | 'Split row' >> beam.Map(lambda record: record.split(','))  
 | 'Filter Accounts Dept' >> beam.Filter(lambda record: record[3].strip() == 'Accounts')  
 | 'Map employee to 1' >> beam.Map(lambda record: (record[1], 1))  
 | 'Sum counts' >> beam.CombinePerKey(sum)  
 | 'Format output' >> beam.Map(lambda result: str(result))  
 | 'Write to file' >> beam.io.WriteToText(output\_prefix)  
 )

## Step 5: View Output

Command:  
!cat output/results-00000-of-00001

Expected Output:  
('John Doe', 3)  
('Emily Davis', 4)

# Conclusion

This assignment demonstrated how to execute an Apache Beam pipeline using Python in a local runtime environment (Google Colab). The pipeline was successfully executed without any cloud setup, achieving the desired result of filtering and counting department-specific data from a local file.

# Tools and Technologies Used

- Google Colab

- Python 3

- Apache Beam

- PIP (Python Package Installer)