Lab 1: Building a Data Lake and Exploring with Gemini - Due Aug 15, 2025 11:59 PM

○ Summer 2025 MGMT 59000-DY2-025 - Merge

Lab 1: Building a Data Lake and AI-Assisted Analytics

Overview

In this lab, you'll build your first cloud data lake using Google Cloud Platform (GCP) and perform AI-assisted analytics with Gemini. You'll follow step-by-step instructions to set up your environment, create a data lake, and apply the DIVE method for deeper analysis.

Duration: 2-3 hours
Format: Individual work

Learning Objectives

By completing this lab, you will be able to:

1. Set up a complete cloud analytics environment (GCP, GitHub, Gemini)
2. Build a data lake following industry best practices
3. Load and explore the Superstore dataset using AI assistance
4. Apply the DIVE method (Discover, Investigate, Validate, Extend) for analysis
5. Create professional documentation of your findings

Prerequisites

- Laptop with Chrome browser
- Purdue email address
- University GCP credits activated (no personal credit card needed)
- Stable internet connection

Important Resources

- **Step-by-Step Data Lake Tutorial:**https://scribehow.com/shared/MSDS506Lab_1_Building_a_Data_Lake_and_Exploring_with_Gemini__SNhoAaB0Ram3EC12OtMH8g
- **Office Hours:** via Zoom

Part 1: Environment Setup (45 minutes)

Task 1.1: Activate Your University GCP Credits

⚠️ **IMPORTANT: Use your Purdue-provided GCP credits, NOT personal credit card**

1. Check your Purdue email for GCP credit activation instructions
2. Follow the university-specific activation process
3. Verify you have $50 in credits available
4. Set up billing alerts at $10 threshold

📸 **Screenshot Required #1:** GCP console showing your available credits

Task 1.2: GitHub Setup

1. Create a GitHub account using your Purdue email
2. Choose a professional username (e.g., FirstnameLastname)
3. Fork the course repository: [URL to be provided]
4. Create this folder structure in your fork:

```
MGMT599-YourName/├── Labs/│   └── Lab1/├── Assignments/└── Final_Project/
```

📸 **Screenshot Required #2:** Your forked repository main page

Task 1.3: Gemini API Configuration

1. Navigate to https://makersuite.google.com/app/apikey
2. Sign in with your Google account
3. Click "Create API key"
4. Copy and store your API key securely (you'll need it for Colab)

Security Note: Never share or commit your API key to GitHub!

Task 1.4: Test Your Setup

Open Google Colab (https://colab.research.google.com) and run:

```
# Test your Gemini API
import google.generativeai as genai

# You'll enter your API key when prompted
api_key = input("Enter your Gemini API key: ")
genai.configure(api_key=api_key)

model = genai.GenerativeModel('gemini-pro')
response = model.generate_content("Say hello to MGMT 599!")
print(response.text)
```

✅ **Checkpoint 1:** All accounts created and API tested successfully

---

Part 2: Building Your Data Lake (45 minutes)

Task 2.1: Follow the Scribd Tutorial

📚 **Required Reading:** Complete the entire Scribd tutorial: https://scribehow.com/shared/MSDS506Lab_1_Building_a_Data_Lake_and_Exploring_with_Gemini__SNhoAaB0Ram3EC12OtMH8g

This tutorial will guide you through:

- Creating a GCP project specifically for the course
- Setting up Cloud Storage buckets with proper naming
- Configuring BigQuery datasets
- Loading the Superstore dataset
- Basic data exploration with Gemini

Important Notes:

- Use project name format: `mgmt599-[yourname]-lab1`
- Create bucket name: `mgmt599-[yourname]-data-lake`
- Follow the exact folder structure shown in the tutorial

📸 **Screenshot Required #3:** Your Cloud Storage bucket with uploaded data
📸 **Screenshot Required #4:** BigQuery dataset with loaded Superstore table

Task 2.2: Verify Your Data Lake

After completing the Scribd tutorial, verify your setup:

```
-- Run this in BigQuery to verify data loaded correctly
SELECT
  COUNT(*) as total_rows,
  COUNT(DISTINCT Customer_ID) as unique_customers,
  MIN(Order_Date) as earliest_order,
  MAX(Order_Date) as latest_order
FROM `your-project.your_dataset.superstore_sales`;
```

Expected results:

- Total rows: ~9,994
- Date range: 2020-2023

Part 3: AI-Assisted Data Exploration (45 minutes)

Task 3.1: Initial Data Understanding with Gemini

Create a new Colab notebook named `Lab1_AI_Analysis.ipynb` and complete these explorations:

Exploration 1: Understanding the Business Context

```
# Use this prompt with Gemini
prompt1 = """
I have a retail dataset called Superstore with columns including:
Sales, Profit, Quantity, Discount, Category, Sub-Category,
Customer ID, Segment, Region, State, City, Order Date, Ship Date

As a retail analyst, what are the 5 most important business questions
I should investigate with this data? For each question, explain why
it matters and which columns I should analyze.
"""
```

Exploration 2: Data Quality Assessment

```
# First, get basic statistics from your data
# Then use this prompt
prompt2 = """
```

```
Here are statistics from my Superstore dataset:
[Paste your data statistics here]

What data quality issues should I check for?
What patterns in these statistics might indicate problems?
Suggest specific validation queries I should run.
"""
```

Exploration 3: Quick Insights Generation

```
# After running some basic queries, use this prompt
prompt3 = """
Initial findings from Superstore data:
- Total sales: $X
- Profit margin: Y%
- Top category: Z

What do these numbers tell us about the business health?
What additional metrics would provide more context?
What might be concerning about these figures?
"""
```

Task 3.2: Create Initial Visualizations

Create at least 3 visualizations exploring different aspects:

1. Sales distribution by category
2. Profit trends over time
3. Regional performance comparison

For each visualization, ask Gemini to interpret the patterns you see.

Part 4: DIVE Method Application (45 minutes)

Task 4.1: Apply the DIVE Framework

Choose ONE business question from Task 3.1 and apply the DIVE method:

D - Discover (Basic Finding)

Start with a simple query and finding. Document:

- Your initial question
- The basic answer/metric

- Your first impression

I - Investigate (Dig Deeper)

Ask "why" questions about your discovery:

- Why does this pattern exist?
- What factors contribute to this?
- How does it vary across dimensions?

Use Gemini to help generate hypotheses and additional queries.

V - Validate (Challenge Assumptions)

Question your findings:

- What could make this conclusion wrong?
- What data limitations exist?
- Are there alternative explanations?

E - Extend (Strategic Application)

Transform insights into action:

- What should the business do?
- How can we measure impact?
- What are the risks?

Task 4.2: Document Your DIVE Journey

Create a markdown document (`dive_analysis.md`) with:

- Each stage clearly labeled
- Specific queries/prompts used
- Key findings at each stage
- Final recommendations

Part 5: Lab Submission (30 minutes)

Task 5.1: Compile Your Work

Organize all deliverables in your GitHub repository:

```
Labs/Lab1/
├── screenshots/
│   ├── gcp_credits.png
│   ├── github_repo.png
│   ├── storage_bucket.png
│   └── bigquery_dataset.png
├── Lab1_AI_Analysis.ipynb
├── dive_analysis.md
└── lab1_summary.md
```

Task 5.2: Create Lab Summary

Create `lab1_summary.md` with:

```
# Lab 1 Summary

## Environment Setup
- GCP Project ID: [your-project-id]
- GitHub Repository: [your-repo-url]
- Data Lake Bucket: [your-bucket-name]

## Key Findings
1. [Your most important discovery]
2. [Second key insight]
3. [Third key insight]

## DIVE Analysis Results
- Business Question Investigated: [question]
- Main Discovery: [finding]
- Strategic Recommendation: [action]

## Challenges and Solutions
- Challenge faced: [describe]
- How I solved it: [solution]

## Time Spent
- Environment setup: X minutes
- Data lake creation: Y minutes
- Analysis: Z minutes
- Total: [total] hours
```

Task 5.3: Submit to D2L

1. Ensure all files are pushed to GitHub

2. Submit to D2L:
    - Your GitHub repository URL
    - Direct link to your Lab1 folder

Tips for Success

1. **Start Early:** Don't wait until the due date - setup can take time
2. **Follow Instructions Exactly:** The Scribd tutorial has specific naming conventions
3. **Document As You Go:** Take screenshots immediately after each step
4. **Use Gemini Actively:** Don't just run queries - engage with AI for insights
5. **Check Your Work:** Verify all files are in GitHub before submitting

---

Lab 2 Preview

In Lab 2, you'll:

- Write complex SQL queries with Gemini assistance
- Build automated reports
- Create your first dashboard
- Apply DIVE to deeper analytics

Start thinking about what business problems interest you most!

Remember: This lab builds the foundation for the entire course. Take time to understand each component - you'll use these tools every week!

Lab 1: Building a Data Lake and Exploring with Gemini

Due August 15 at 11:59 PM