Week 10 Activity: A resource tagging based cost governance simulator

The **CloudMart Resource Tagging Dataset** represents a simulated multi-department cloud environment designed to study the relationship between **resource tagging**, **cost visibility**, and **governance** in cloud cost management.

It models the operations of **CloudMart Inc.**, an e-commerce company that uses various cloud services (e.g., EC2, S3, RDS, Lambda, CloudFront, etc.) across multiple regions and environments (Prod, Dev, Test).

Lab Objectives

By the end of this lab, you will be able to:

- Understand the structure and importance of resource tagging in cloud environments.
- Measure tagging compliance and cost visibility.
- Identify untagged resources and quantify their hidden costs.
- Visualize cloud costs across departments, services, and environments.
- Simulate tag remediation and observe its effect on cost reporting.

Dataset:

This dataset is intended for:

- Performing **Exploratory Data Analysis (EDA)** on cloud cost and tagging practices.
- Understanding the **impact of missing tags** on financial accountability.
- Demonstrating **best practices for cost allocation** and **tag remediation** workflows.
- Building **interactive dashboards** (e.g., using Streamlit) for **cloud cost visibility**.

Attribute	Description
ResourceID	Unique identifier for each cloud resource (e.g., EC2 instance ID, S3 bucket).
Service	The type of cloud service (e.g., EC2, S3, RDS, Lambda, EBS, CloudFront).
Region	Geographic region where the resource is hosted (e.g., us-east-1, eu-west-1).

Department	Internal business unit using the resource (e.g., Marketing, Sales, Analytics, Finance, Engineering).
Project	Project or application associated with the resource (e.g., CampaignApp, CRMTool, DataLake).
Environment	Operational environment (e.g., Prod, Dev, Test).
Owner	Responsible person or team (often via email address).
CostCenter	Accounting or budget code used for financial tracking.
CreatedBy	Indicates automation or provisioning source (e.g., Terraform, Jenkins, CloudFormation, Manual).
MonthlyCostUSD	Monthly estimated cost in U.S. dollars.
Tagged	Indicates whether the resource is properly tagged (Yes or No).

Task Set 1 – Data Exploration

#	Task	Hints / Questions
1.1	Load the dataset in Python or Streamlit and display the first 5 rows.	Use pd.read_csv() or upload via Streamlit.
1.2	Check for missing values in the dataset.	df.isnull().sum()
1.3	Identify which columns have the most missing values.	Look for Department, Project, or Owner.
1.4	Count total resources and how many are tagged vs untagged.	Use df['Tagged'].value_counts().
1.5	What percentage of resources are untagged?	Compute (untagged / total) * 100.

Task Set 2 – Cost Visibility

#	Task	Hints / Questions
2.1	Calculate total cost of tagged vs untagged resources.	Group by Tagged and sum MonthlyCostUSD.
2.2	Compute the percentage of total cost that is untagged.	(untagged_cost / total_cost) * 100.
2.3	Identify which **department** has the most untagged cost.	Group by Department and Tagged.
2.4	Which **project** consumes the most cost overall?	Use .groupby('Project')['MonthlyCostUSD'].sum().
2.5	Compare **Prod vs Dev** environments in terms of cost and tagging quality.	Group by Environment and Tagged.

Task Set 3 – Tagging Compliance

#	Task	Hints / Questions
3.1	Create a “Tag Completeness Score” per resource.	Count how many of the tag fields are non-empty.
3.2	Find top 5 resources with lowest completeness scores.	Sort by the new score column.
3.3	Identify the most frequently missing tag fields.	Count missing entries per column.
3.4	List all untagged resources and their costs.	Filter where Tagged == 'No'.
3.5	Export untagged resources to a new CSV file.	Use df[df['Tagged']=="No"].to_csv('untagged.csv').

Task Set 4 – Visualization Dashboard

#	Task	Hints / Questions
4.1	Create a pie chart of tagged vs untagged resources.	Use plotly.express.pie().
4.2	Plot a bar chart showing cost per department by tagging status.	Use barmode='group'.
4.3	Show a horizontal bar chart of total cost per service.	Group by Service.
4.4	Visualize cost by environment (Prod, Dev, Test).	Pie or bar chart works.
4.5	Add interactive filters in Streamlit (Service, Region, Department).	Use st.selectbox() or st.multiselect().

Task Set 5 – Tag Remediation Workflow

#	Task	Hints / Questions
5.1	In Streamlit, create a table where untagged resources can be edited.	Use st.data_editor().
5.2	Fill missing tags (Department, Project, Owner) manually.	Simulate remediation.
5.3	Download the updated dataset.	Use st.download_button().
5.4	Compare cost visibility before and after remediation.	Recalculate tagging metrics after updates.
5.5	Discuss how improved tagging affects accountability and reports.	Write a short reflection.

Deliverables

At the end of this lab show the demo and submit:

1. Your EDA notebook and Streamlit dashboard link.
2. The “before and after” datasets (original.csv and remediated.csv).
3. A short report summarizing:
   - % of untagged resources
   - Total untagged cost
   - Departments with missing tags
   - Recommendations for governance improvement