Milestone 2 - Student Guide

1. CPU Utilization Calculation

Formula: **cpu_utilization = cpu_used / cpu_total**

- cpu_total → Maximum available CPU for the resource.
- cpu_used → CPU used by that resource on a given day.

Since the dataset doesn't have info about CPU cores, we consider **cpu_total** as the **maximum observed CPU usage** for that resource.

2. Storage Efficiency Calculation

Formula: **storage_efficiency = storage_used / storage_allocated**

But since we don't have **storage_allocated** in our dataset, we assume:
storage_allocated = peak (maximum) storage used for each resource.

Final formula becomes:
storage_efficiency = storage_used / max(storage_used_per_resource)

Example Dataset:

Resource ID	Date	Storage Used (GB)
VM-1	2025-08-01	120
VM-1	2025-08-02	150
VM-1	2025-08-03	200
VM-2	2025-08-01	80
VM-2	2025-08-02	100

Step 1: Find peak storage per resource.

Resource ID	Max Storage (GB)
VM-1	200
VM-2	100

Step 2: Calculate storage efficiency for each resource.

Resource ID	Date	Storage Used (GB)	Max Storage (GB)	Storage Efficiency
VM-1	2025-08-01	120	200	0.60 (60%)
VM-1	2025-08-02	150	200	0.75 (75%)
VM-1	2025-08-03	200	200	1.00 (100%)
VM-2	2025-08-01	80	100	0.80 (80%)
VM-2	2025-08-02	100	100	1.00 (100%)

3. Python Code Example

```python
import pandas as pd

data = {
'resource_id': ['VM-1','VM-1','VM-1','VM-2','VM-2'],
'date': ['2025-08-01','2025-08-02','2025-08-03','2025-08-01','2025-08-02'],
'storage_used': [120,150,200,80,100]
}

df = pd.DataFrame(data)
df['max_storage'] = df.groupby('resource_id')['storage_used'].transform('max')
df['storage_efficiency'] = df['storage_used'] / df['max_storage']
print(df)
```