INF2006 Cloud Computing and Big Data

Group Project Assignment 1 (Cloud Computing Focus)

Design and Implementation of Data Analytics as a Service (DAaaS) on AWS

---

1. Introduction

Cloud computing enables organisations to deliver scalable, on-demand services without the need to manage physical infrastructure. In this group project, students will apply cloud computing concepts taught in INF2006 to **design, implement, and evaluate a Data Analytics as a Service (DAaaS) platform deployed on Amazon Web Services (AWS)**.

This project focuses on **cloud-native data analytics**, requiring teams to design and deploy analytics capabilities as a cloud-hosted service. Student teams will:

- Source a real-world dataset from public data repositories
- Design analytics functions that extract meaningful insights from the dataset
- Implement these analytics as a **cloud-hosted service**
- Develop a web-based interface for users to interact with the analytics

The emphasis of this project is on **cloud architecture, service design, scalability, security, and operational considerations**, rather than advanced or sophisticated data science or machine learning techniques. Advanced data analytics and big data processing concepts will be covered in the second half of the module.

2. Project Overview

Your team is tasked with architecting and implementing a **Data Analytics as a Service (DAaaS)** platform on AWS. The platform should allow users to:

1. Access a curated dataset
2. Invoke one or more analytics functions
3. Visualise analytics results through a web application

The project is divided into three phases:

- **Phase 1: Cloud Solution Design**
- **Phase 2: Cloud Implementation**
- **Phase 3: Evaluation and Analysis**

Each phase requires clear documentation, justification of design choices, and reflection on cloud computing principles.

3. Dataset Selection

Each team must select **one primary dataset** obtained from a publicly available source. The dataset should be **non-trivial** (i.e. more than a few hundred records) and suitable for meaningful analytics.

Dataset Source and Scope

All teams are required to source their dataset from **data.gov.sg**, Singapore's official government open data portal.

To ensure relevance and consistency, each team must choose **one dataset from one of the following government agencies**:

- **Ministry of Health (MOH)**
- **Land Transport Authority (LTA)**
- **Ministry of Education (MOE)**

Within the selected agency, teams are free to choose **any dataset that they are interested in working on**, provided it supports meaningful analysis.

Examples of Dataset Themes

(Non-exhaustive; for illustration only)

- Healthcare statistics and trends (MOH)
- Public transport usage and mobility patterns (LTA)
- Education-related statistics and trends (MOE)

The chosen dataset, its source, and any licensing considerations must be **clearly stated and referenced** in the project report.

4. Analytics Requirements

Each team must design and implement **at least three (3) analytics functions**. These functions should go beyond simple data display and demonstrate value-added statistical analysis.

The analytics implemented in this project should be based on classical statistical and exploratory data analysis techniques, and must not rely on AI- or machine-learning-based analytics. The focus is on understanding the data and deriving insights through well-established statistical analytical methods.

Examples of Analytics Functions

(Teams are encouraged to be creative.)

- Descriptive statistics (e.g. averages, distributions, trends over time)
- Group-by analysis (e.g. by region, age group, category)
- Time-series analysis and visualisation

- Comparative analysis (e.g. across years or regions)
- Simple statistical-based projections or trend extrapolation (optional, not mandatory)

The analytics logic may be implemented using:

- Server-side application code (e.g. Python, Java, Node.js, PHP)
- SQL-based analytics
- AWS-managed analytics services (where appropriate)

You may use AI tools such as ChatGPT to help develop the application code. In so doing, you are required to **use them responsibly, declare their usage, and ensure that you can explain and justify all design and implementation decisions**. See below section on "AI Usage Declaration".

5. System Architecture Expectations

Your DAaaS platform should be designed using **core AWS services**, such as (but not limited to):

- **Compute**: EC2, Lambda, or container-based services
- **Storage**: S3, EBS
- **Database**: RDS, DynamoDB, or other suitable AWS database services
- **Networking**: VPC, subnets, security groups
- **Application Delivery**: Load balancer, API endpoints
- **Monitoring**: CloudWatch

The architecture should demonstrate:

- Clear separation of concerns (frontend, backend, data layer)
- Secure configuration and access control
- Consideration of scalability and availability

6. Project Phases and Tasks

Phase 1: Cloud Solution Design

1. **Problem Definition and Use Case**
    - Describe the dataset and the problem your analytics service addresses
    - Identify target users and usage scenarios
2. **Cloud Service Model Selection**
    - Identify whether your solution primarily leverages IaaS, PaaS, or a combination
    - Justify your choices

3. **Architecture Design**
    - Propose a detailed cloud architecture diagram
    - Explain data flow and service interactions
4. **Data Storage and Management**
    - Design how raw and processed data are stored
    - Address data durability and access patterns
5. **Security and Access Control**
    - Describe the **security and access control mechanisms** implemented for the system, including **network-level and application-level security measures**.

Assumptions (e.g. number of users, data size, budget constraints) must be explicitly stated.

Phase 2: Cloud Implementation

1. **AWS Environment Setup**
    - Configure networking, compute, and storage services
2. **Data Ingestion**
    - Upload and manage the dataset within AWS
    - Perform any necessary preprocessing
3. **Web Application Development**
    - Develop a frontend that allows users to interact with the analytics
    - Display results using text, charts, tables, or dashboards
4. **Scalability and Availability Features**
    - Implement auto-scaling, load balancing, or serverless designs where appropriate

Phase 3: Evaluation and Analysis

1. **Performance Evaluation**
    - Evaluate system performance under different workloads. You may use tools such as Apache JMeter to stress test the web application under different scenarios and workloads.
2. **Cost Analysis**
    - Analyse AWS resource usage and estimated costs
    - Discuss cost-performance trade-offs

3. **Scalability Assessment**
    - Explain how the system scales with increasing data size or user load
4. **Reliability and Fault Tolerance**
    - Discuss how failures are handled and mitigated

7. Deliverables

1. Written Report

- Format: **PDF only** (any other format is not acceptable)
- Length: **15 ± 5 pages** (excluding references and appendices)
- Filename: Report_Gxx.pdf (where xx is the group number)

The report must cover all three phases and include a **team contribution section** as the final chapter.

2. Video Presentation

- Duration: **8 ± 3 minutes**
- Upload video to **YouTube only**. **DO NOT upload video to the xSITE LMS**. Penalty may be imposed if video is uploaded to LMS.
- Submit a text file named Video_Gxx.txt (where xx is the group number) containing:
    - Team member names
    - YouTube link

Each member must briefly articulate their contributions in the video.

8. Assessment Criteria

The project will be assessed based on:

1. Quality and clarity of cloud architecture design
2. Correctness and robustness of AWS implementation
3. Effectiveness and relevance of analytics functions
4. Consideration of scalability, security, and cost
5. Quality of evaluation and analysis, including performance assessment, scalability discussion, and cost considerations
6. Quality of report writing and video presentation
7. Evidence of teamwork and individual contribution

9. AI Usage Declaration

If AI-assisted tools (e.g. ChatGPT, GitHub Copilot) are used in any part of the project, teams must include a short subsection titled "AI Usage Declaration" in the report.

The declaration should clearly state:

- Which AI tools were used
- Which parts of the project were AI-assisted (e.g. code generation, debugging, documentation)
- How the team validated, modified, and integrated the AI-generated outputs

Students remain fully responsible for understanding, explaining, and defending all submitted work. Undeclared or inappropriate use of AI tools may be treated as an academic integrity issue.

10. AWS Service Availability for Assessment

To avoid unnecessary consumption of AWS credits, teams are **not required to keep their application and AWS services running continuously** while waiting for the project to be assessed.

The instructor will separately notify each team of the specific date and time window during which the deployed application will be accessed for evaluation. Teams are expected to ensure that all required AWS services and the application are properly deployed and accessible during this assessment window.

Outside of the stated assessment window, teams are encouraged to stop or terminate AWS resources when they are not in use, as part of good cloud cost management practice.

11. Notes

- The focus is on **cloud computing principles**, not sophisticated data science
- Use of external libraries and sample code is allowed, but must be cited
- Plagiarism or misrepresentation of contributions will be penalised

---

All the best, and we look forward to seeing your cloud-based analytics services come to life!