COMM034 Coursework 2025

Author: [Andrew Crossan](mailto:a.crossan@surrey.ac.uk?subject=COMM034%20Coursework%20Query)

# Logistics

* **Coursework counts for 100% of your grade.**
* **Handout**: Wednesday, 23rd October
* **Due**: Monday, 1st December
* We’ll be using Google Cloud & AWS for our coursework.
* Check your GCP credits, make sure they’re close to the starting level of $50.
* **Task A**
  + You’ll first develop an application to analyse text in a file.
  + You’ll deploy on cloud architecture to trigger in an event driven manner
  + You’ll extend your solution to 2 or more files.
  + You’ll ensure that your solution uses the scalability cloud resources to efficiently process the files.
* **Task B**
  + You’ll deploy an orchestrated “Jump Start Solution”.
  + You’ll examine and modify the solution and capture your learnings along the way.
  + You’ll submit a report on your findings, along with code and screenshots.
* **Task C**
  + You will submit a short video explaining how your solution to Task A and Task B works
* Use the given template for your report.

# Task A (12 hours)

Task A of this coursework requires you to design, implement, and deploy a scalable, event-driven cloud service capable of performing linguistic analysis on batches of uploaded text files. The primary goal is to demonstrate proficiency in serverless architecture, inter-service communication, and the optimization of compute tasks through parallel processing and scaling.

For your solution, you may choose to use either Google Cloud or AWS or a combination of both as the task is possible on both platforms. Please choose the platform that you are most comfortable with.

The task is divided into 3 sections:

A1. Write code to analyse a text file

A2. Deploy your solution to the cloud with an event driven trigger

A3. Extend the solution to analyse and compare multiple text files efficiently

## Assessed Content

**A1. Analyse a text file – around 3 hours of effort**

**Marks: 12%**

Write 3 functions that takes the name of a text file. Your code should perform the following analysis

* **Word frequency analysis** – Determine the 20 most frequently used words in the text file excluding the stop words shown in Figure 1 below.
* **Sentence start words** – Determine the 10 most frequently used words to start a sentence
* **Sentence length distribution** – Determine the mean, median and standard deviation of the lengths of the sentences

There are multiple ways to complete this task and you must note any choices that you make when designing your algorithms. Provide the code for each function. In the report, describe how your algorithms work and the choices that you have made during implementation.

**Report: Complete section A1 (300 words ±10% not including the code).**

**A2. Deploy your solution – around 3 hours of effort**

**Marks: 8%**

Provide endpoints to trigger your functionality over a web connection in an event driven manner. The should be customisable such that a user can choose which analysis to perform or select any combination of analyses. In the report, describe what services you used, and how you allowed the user to select different combinations of analyses.

**Report: Complete section A2 (200 words ±10%).**

**A3. Extend your solution to allow analysis of multiple text files efficiently – around 6 hours of effort**

**Marks: 20%**

Provide an interface to allow the analysis to be performed on multiple files simultaneously. The ideal solution should take advantage of the scalability of cloud services to efficiently process the files using parallelisation. Describe what services you used, how you extended your solution and justify the efficiency of your solution.

**Report: Complete section A3 (400 words ±10%).**

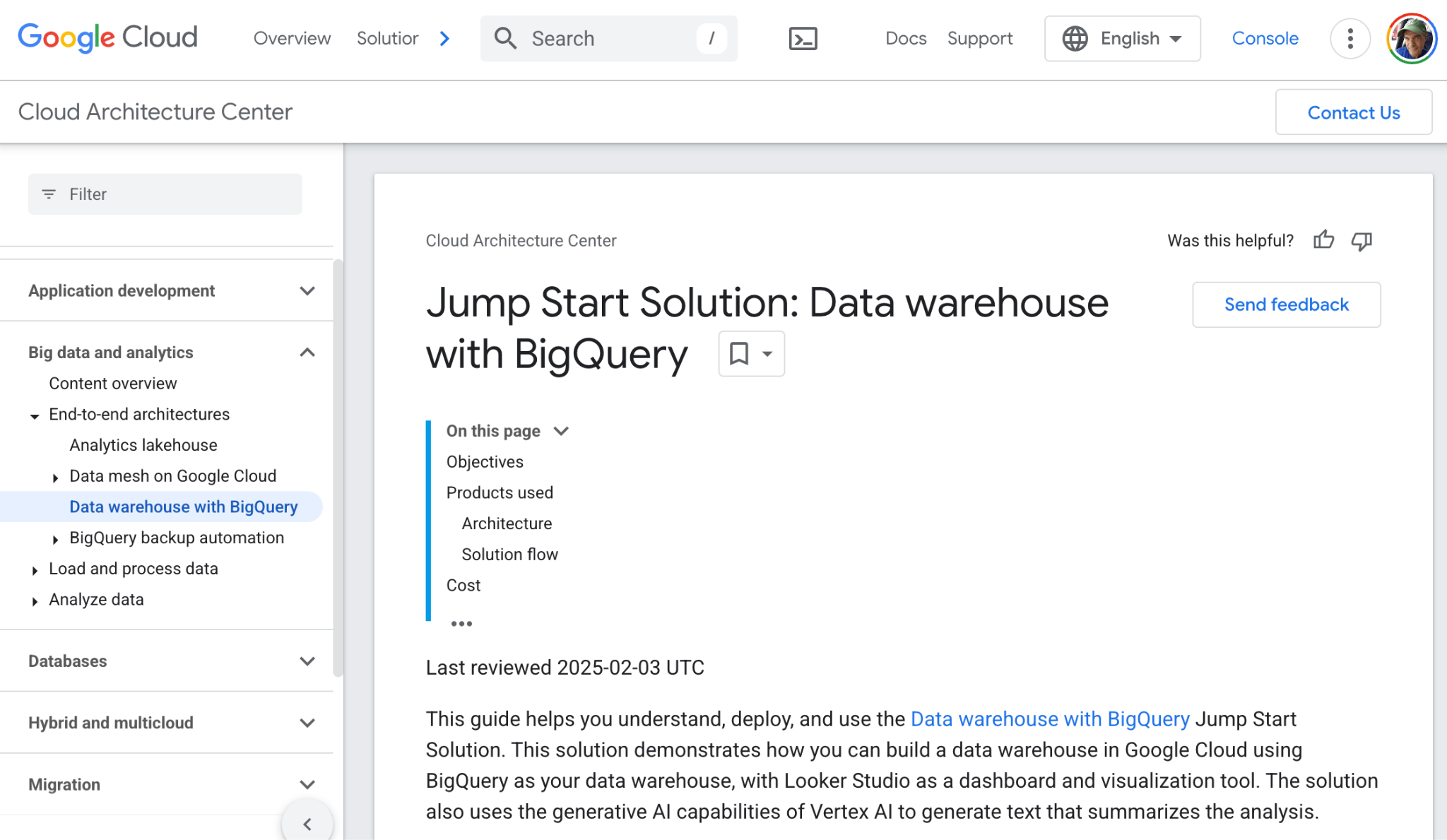
["i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", "yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself", "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these", "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do", "does", "did", "doing", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "while", "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "before", "after", "above", "below", "to", "from", "up", "down", "in", "out", "on", "off", "over", "under", "again", "further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each", "few", "more", "most", "other", "some", "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very", "s", "t", "can", "will", "just", "don", "should", "now"]

Figure . A list of English stop words that should be excluded from the analysis in Task A1

# Task B (18 hours)

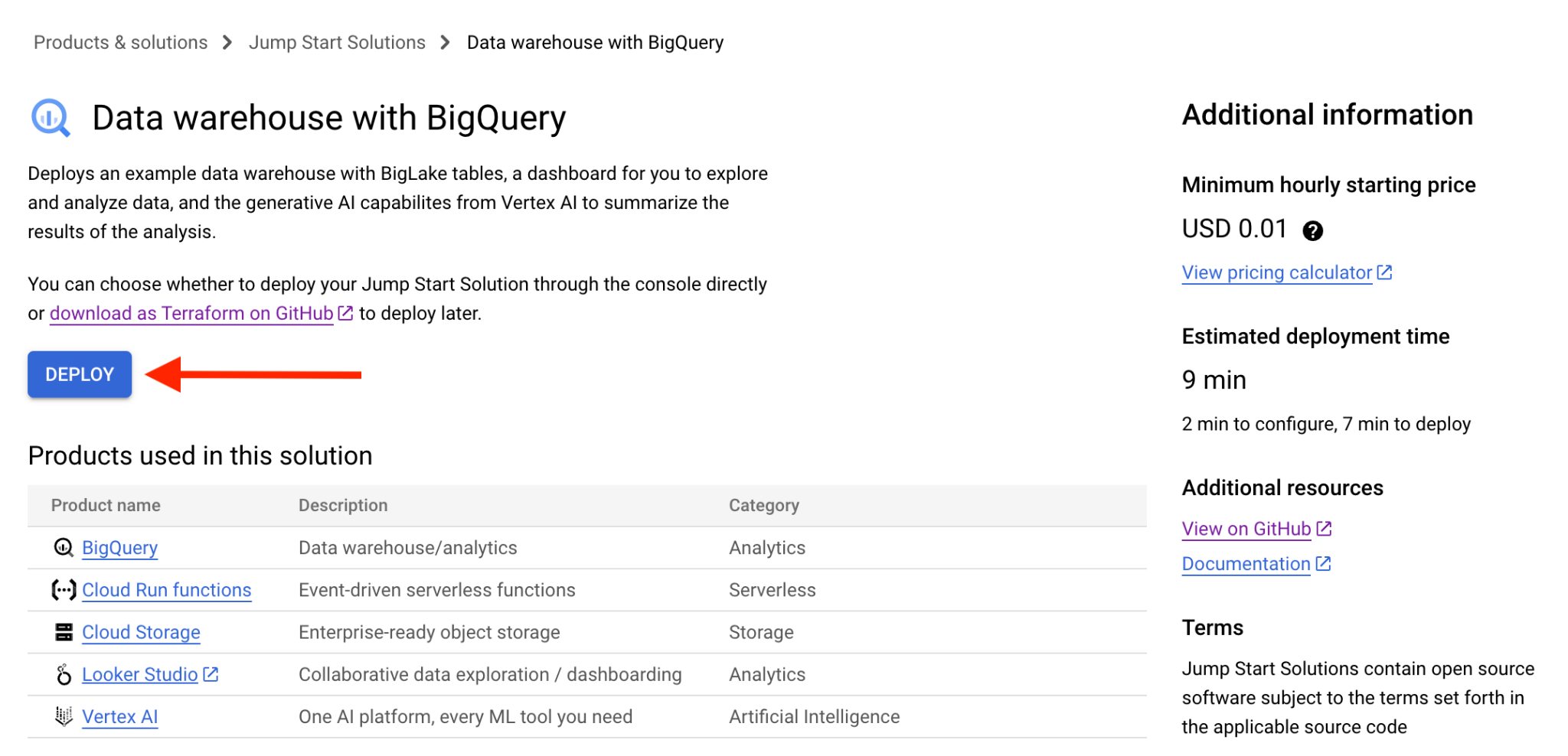
# Jump Start Solution - Data Warehouse with BigQuery

Jump Start Solutions are pre-configured, readily deployable solutions created by Google and built right into the GCP console. Review the documentation on the [Data Warehouse with BigQuery](https://cloud.google.com/architecture/big-data-analytics/data-warehouse) Jump Start Solution, which is the solution we’ll be using in this coursework.

[](https://cloud.google.com/architecture/big-data-analytics/data-warehouse)

# Deploy the Solution

[Deploy the solution using the console](https://cloud.google.com/architecture/big-data-analytics/data-warehouse#deploy-from-console):



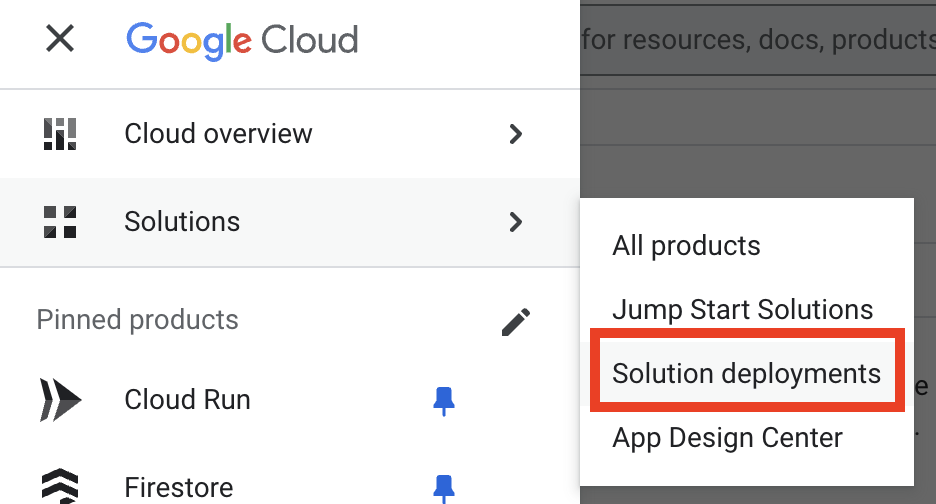
Don’t bother with the Terraform CLI deployment option, unless you’re interested in learning more about [Terraform](https://www.terraform.io/).

**Once your solution is deployed, keep a close eye on your costs by visiting the Billing -> Credits page.**

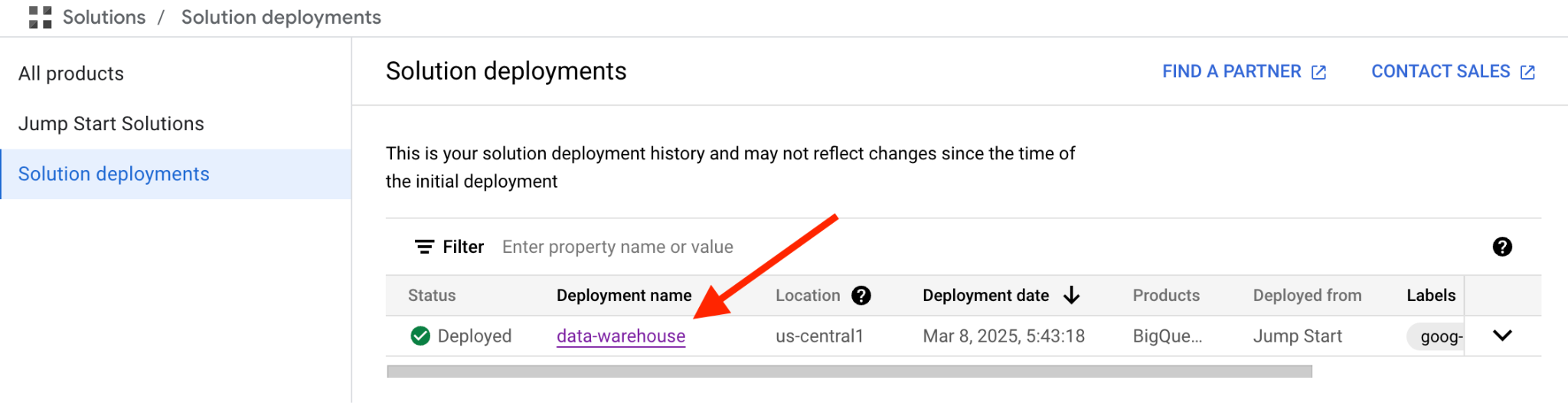
# 

# Execute Tour and Tutorials

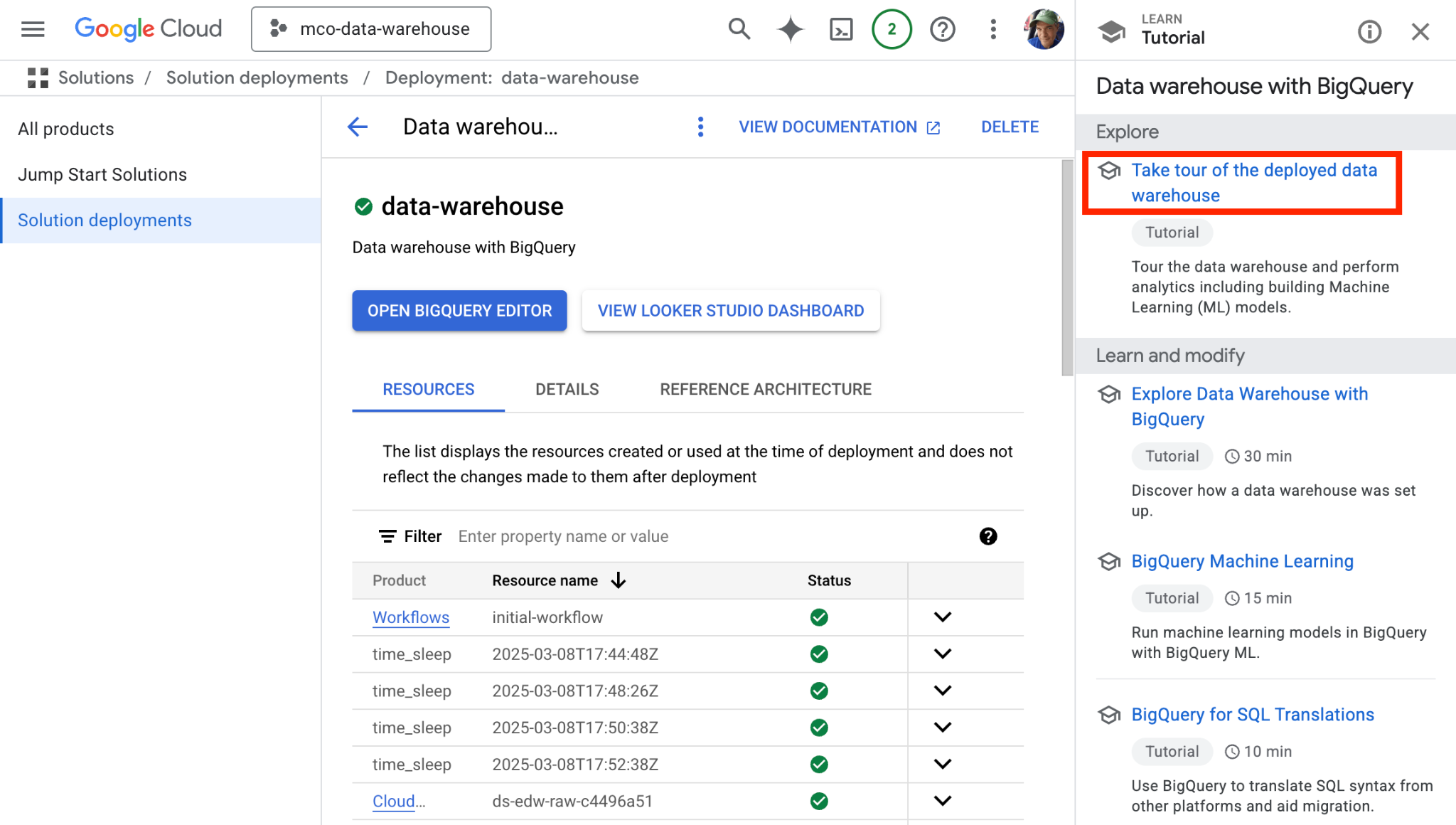
If you leave the console and return later, you can always find your deployment under Solutions -> Solution deployments:



Drill down into your deployed solution:

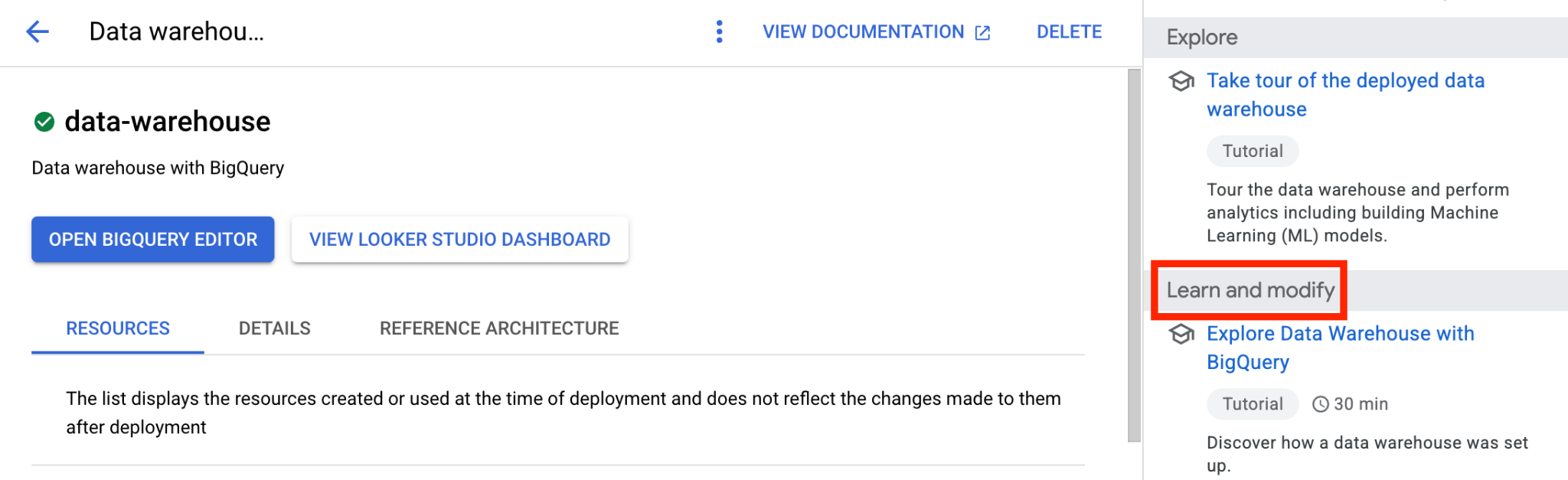


and take the guided tour of the deployed data warehouse:

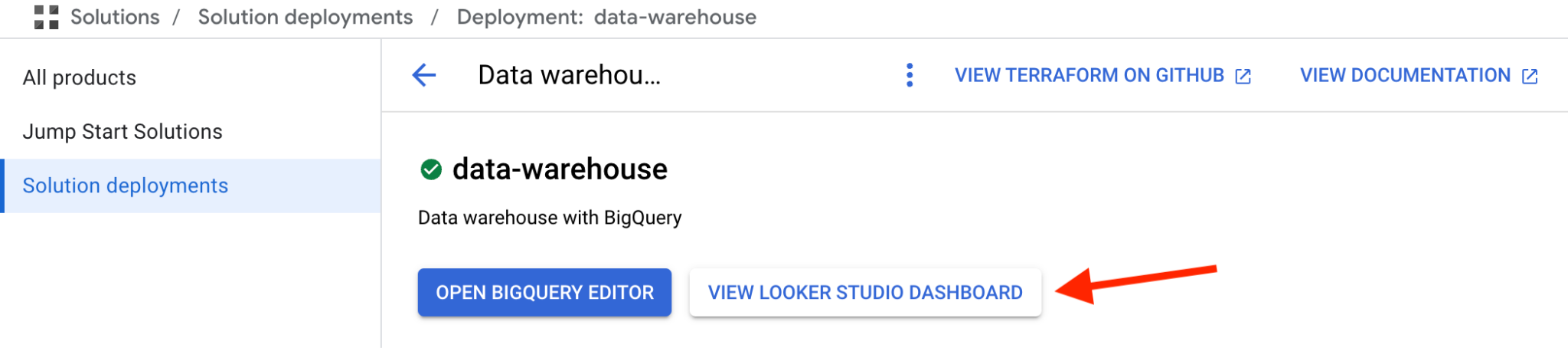
****

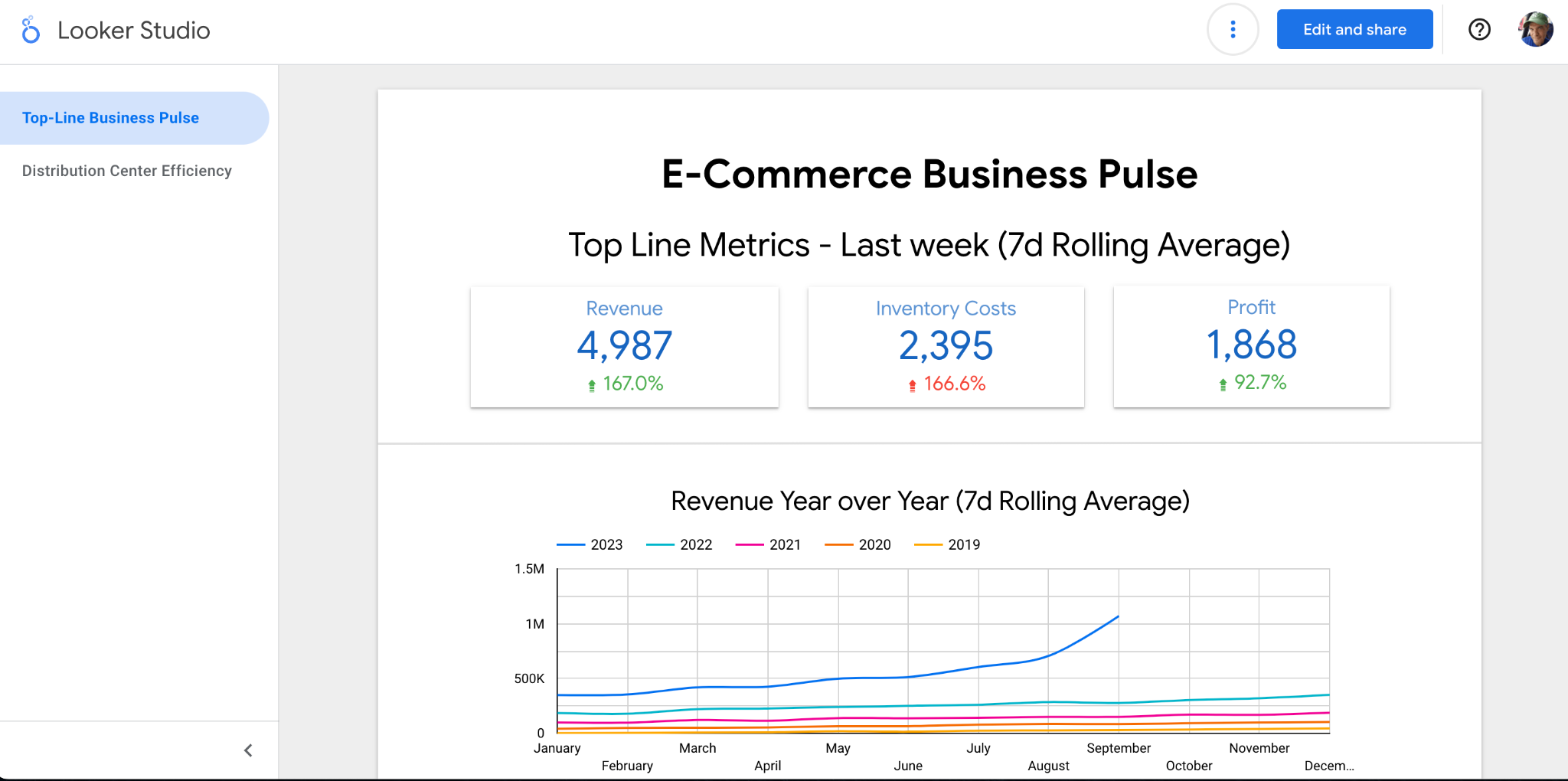
Take the following “Learn and modify” tutorials:

* Explore Data Warehouse with BigQuery (~30 minutes)
* BigQuery Machine Learning (~15 minutes)
* BigQuery for SQL Translations (~10 minutes)
* BigQuery Fine Grained Security (~10 minutes)
* BigQuery for Analytics (~10 minutes)
* Using Generative AI with BigQuery Machine Learning (~10 minutes)
* Connecting BigQuery data to apps (~5 minutes)



Examine the Looker Studio dashboard:





# Task B Rubric (16 hours)

## Preparation

**B1. Deploy the Solution & Explore the Environment – around 2 hours of effort**

**Marks: Not assessed**

Familiarize yourself with the coursework, deploy the solution, verify Cloud Storage objects, BigQuery tables, and Looker dashboard.

**B2. Familiarise yourself with the Solution – around 3 hours of effort**

**Marks: Not assessed**

Complete all tutorials mentioned in the “Execute Tour and Tutorials” section above.

## Assessed Content

**B3. Query the Dataset – around 4 hours of effort**

**Marks: 20%**

Identify a business aim for your analysis and formulate and run three SQL queries of your own design that relate to this aim. Capture the queries, the results, explain why you chose them, what the queries do, and how they work.

**Report: Complete section B3 in the report (500 words ±10% not including the query text).**

**B4. Update the Dashboard – around 3 hours of effort**

**Marks: 10%**

Change the title and add a new page of your choosing to your dashboard. The page should provide related visualisations that analyse a different business aspect from section B3 above.

**Report: Complete section B4 in the report. Explain what the new page shows and why you chose those elements to display (300 words ±10%).**

**B5. Querying through Cloud Run – around 4 hours of effort**

**Marks: 20%**

Add a simple Cloud Run service that supports two different queries of the table and give results on demand via a web user interface.

**Report: Complete section B5 describing how you created the service and what happens in the cloud when users access it (300 words ±10%).**

# Task C (2 hours)

**C1. Create a Video Demonstrating Your Working Systems – around 2 hours of effort**

**Marks: 10%**

Create a video in MP4 format of between 4 and 5 minutes that demonstrates the functionality of your solutions to the assessed sections of Tasks A and B. You should explain what is happening at each stage of your solution. Comment on what part of your solution is functional and any parts that are not.

# 

# Marking Rubric

**Each section has a hard upper limit on the number of words. Eg., Section C specifies 600** ±10% **words. The hard upper limit here will therefore be 330 words and a 10% penalty will be applied to the marks for a section where the word count is above this upper limit. There is no penalty for being below, with the lower limit provided as a guide to how much detail is required.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section A1** | |  | **Text file analysis** |
|  | ≥ 85% | | All analyses successful completed to a high standard and included. An outstanding discussion identifying how the analyses work and the choices that were made when building the algorithm including discussion on the strengths and weaknesses of the algorithm. |
|  | ≥ 70% | | All analyses successful completed to a high standard and included. An excellent discussion identifying how the analyses work and the choices that were made when building the algorithm including discussion some of the strengths and weaknesses of the algorithm. |
|  | ≥ 60% | | All analyses successful completed. A very good very discussion identifying how the analyses work and the choices that were made when building the algorithm including some reference to the strengths and weaknesses of the algorithm. |
|  | ≥ 50% | | All code provided for each analysis that to some extent completes the task. A good discussion identifying how the analyses work and the choices that were made when building the algorithm that might including some reference to the strengths and weaknesses of the algorithm. |
|  | ≥ 40% | | Partial solution to the analyses provided. Some discussion identifying how the analyses work and the choices that were made when building the algorithm that might including some reference to the strengths and weaknesses of the algorithm. This discussion might lack technical depth. |
|  | < 40% | | A limited attempt that fulfills at most a few of the requirements specified with limited justification and discussion. |
| **Section A2** | |  | **Deploy Your Solution** |
|  | ≥ 85% | | Your code has been successful deployed in a sensible manner and meets all requirements. An outstanding discussion the services used and how you provided the optionality to the user of the types of analyses with excellent technical depth. |
|  | ≥ 70% | | Your code has been successful deployed in a sensible manner and meets all requirements. An excellent discussion the services used and how you provided the optionality to the user of the types of analyses with excellent technical depth. |
|  | ≥ 60% | | Your code has been successful deployed in a sensible manner and meets the requirements although potentially in a less optimal manner. A very good discussion the services used and how you provided the optionality to the user of the types of analyses. |
|  | ≥ 50% | | Your code has been successful deployed in a sensible manner and to a large extent meets the requirements in a less optimal manner. A good discussion the services used and how or to what extent flexibility has been provided for the user to run different analyses. |
|  | ≥ 40% | | Your code has been successful deployed that meets some of the requirements. A some discussion the services used and how or to what extent flexibility has been provided for the user to run different analyses. |
|  | < 40% | | A limited attempt that fulfills at most a few of the requirements specified with limited justification and discussion. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Section A3** | |  | **Extending the solution to multiple files** |
|  | ≥ 85% | | An excellent solution that meets all requirements. The solution uses cloud technologies to efficiently complete the task for multiple files. An outstanding discussion with justification of the architecture and choices made with excellent technical depth. |
|  | ≥ 70% | | An excellent solution that meets all requirements. The solution uses cloud technologies to efficiently complete the task for multiple files. An excellent discussion with justification of the architecture and choices made with excellent technical depth. |
|  | ≥ 60% | | A fully working solution that meets all requirements. The solution provides some scalability to efficiently complete the task for multiple files although this might not be optimal. A very good discussion with justification of the architecture and choices made with some technical depth. |
|  | ≥ 50% | | A solution that largely meets all requirements. The solution may not perform the task efficiently for multiple files. A good discussion with justification of the architecture and choices made the might lack technical depth. |
|  | ≥ 40% | | A solution that partially meets the requirements. There has been limit attempt to perform the task efficiently for multiple files. Some discussion with justification of the architecture and choices made the lacks technical depth. |
|  | < 40% | | A limited attempt that fulfills at most a few of the requirements specified with limited justification and discussion. |

|  |  |  |
| --- | --- | --- |
| **Section B3** |  | **Query the Dataset** |
|  | ≥ 85% | A well chosen technically challenging query that extracts useful data from the dataset. An excellent justification with a solid understanding and insightful discussion conveying how it works. |
|  | ≥ 70% | A well chosen technically challenging query that extracts useful data from the dataset. An excellent justification with a solid understanding conveyed of how it works. |
|  | ≥ 60% | A well chosen query that extracts useful data from the dataset. A very good justification with a good understanding conveyed of how it works. |
|  | ≥ 50% | A reasonable choice of query that extracts some useful data from the dataset. A good justification that might lack some detail with a reasonable understanding conveyed of how it works at a high level. |
|  | ≥ 40% | A straight-forward query that extracts some basic data from the dataset. A limited justification that lacks detail with a limited understanding conveyed of how it works that lacks technical details. |
|  | < 40% | A basic query or an attempt that does not fulfill the basic requirements of this section. A poor or limited discussion of the attempt. |

|  |  |  |
| --- | --- | --- |
| **Section B4** |  | **Update the Dashboard** |
|  | ≥ 85% | Clear evidence of a successfully completed new page. The page has been well chosen and the justification is excellent. The section is written at an exceptional standard. |
|  | ≥ 70% | Clear evidence of a successfully completed new page. The page has been well chosen and the justification is excellent. |
|  | ≥ 60% | Clear evidence of a successfully completed new page. The page has been well chosen and the justification is very good. |
|  | ≥ 50% | Clear evidence of a successfully completed new page. The page has been reasonably well chosen and the justification is reasonable but lacks some detail. |
|  | ≥ 40% | Evidence that the task has been partially completed. The new page has been chosen with limited justification provided. |
|  | < 40% | No evidence that the task has been completed or an attempt that does not meet the requirements specified. Limited or no justification for the choices. |

|  |  |  |
| --- | --- | --- |
| **Section B5** |  | **Querying through Cloud Run** |
|  | ≥ 85% | Evidence of a successfully deployed Cloud Run service that completes the task. An excellent description of how to create the service. A clear, concise and technically sound discussion on what happens in the cloud when a user accesses the service. The section is written at an exceptional standard. |
|  | ≥ 70% | Evidence of a successfully deployed Cloud Run service that completes the task. An excellent description of how to create the service. A clear, concise and technically sound discussion on what happens in the cloud when a user accesses the service. |
|  | ≥ 60% | Evidence of a successfully deployed Cloud Run service that completes the task. A very good description of how to create the service. A clear discussion on what happens in the cloud when a user accesses the service that conveys technical understanding. |
|  | ≥ 50% | Evidence of a successfully deployed Cloud Run service that completes the task. A reasonable description of how to create the service. A reasonable discussion on what happens in the cloud when a user accesses the service that may lack technical detail in some places. |
|  | ≥ 40% | Limited evidence of a successfully deployed Cloud Run service that completes the task with a reasonable description of how to create the service and a reasonable discussion on what happens in the cloud when a user accesses the service or evidence of a successfully deployed Cloud Run service that completes the task. Some vague description of how to create the service. A limited discussion on what happens in the cloud when a user accesses the service that lacks technical detail. |
|  | < 40% | No evidence of a successfully deployed Cloud Run service that completes the task. Missing or extremely vague discussion of how to create the service with little attempt to describe what happens when a user accesses the service. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Section C6** | |  | **Video Demonstration** |
|  | ≥ 85% | | A video that clearly and concisely shows each section of the solutions working with commentary that explains in an authoritative manner what is happening at each stage with an excellent level of technical depth. It explains clearly what is complete and working and what is not. |
|  | ≥ 70% | | A video that clearly and concisely shows each section of the solutions working with commentary that explains in a clear manner what is happening at each stage with an excellent level of technical depth. It explains clearly what is complete and working and what is not. |
|  | ≥ 60% | | A video that relatively clearly shows each section of the solutions working with commentary that explains in a clear manner what is happening at each stage with a very good level of technical depth. It explains to a large extent what is complete and working and what is not. |
|  | ≥ 50% | | A video that relatively clearly shows a large part of the solutions working with commentary that explains to some extent what is happening at each stage with some technical depth. It to some extent explains what is complete and working and what is not. |
|  | ≥ 40% | | A video that provides a slight vague overview of the solutions which may not be complete, with commentary that explains parts of what is happening at each stage at a high level. It attempts to explain what is complete and working and what is not. |
|  | < 40% | | A video that does not meet the requirements stated in the section. |

**The submission format will be a document reporting all of your findings using** [**this template**](https://surreyac-my.sharepoint.com/:w:/r/personal/mc0107_surrey_ac_uk/Documents/COMM034%20Coursework%20Report%20Template%201.docx?d=wfda1018fc74c4a70a55a5c33c579e517&csf=1&web=1&e=cydJiJ)**.**