

# **BIG DATA**

## **SECTION D**

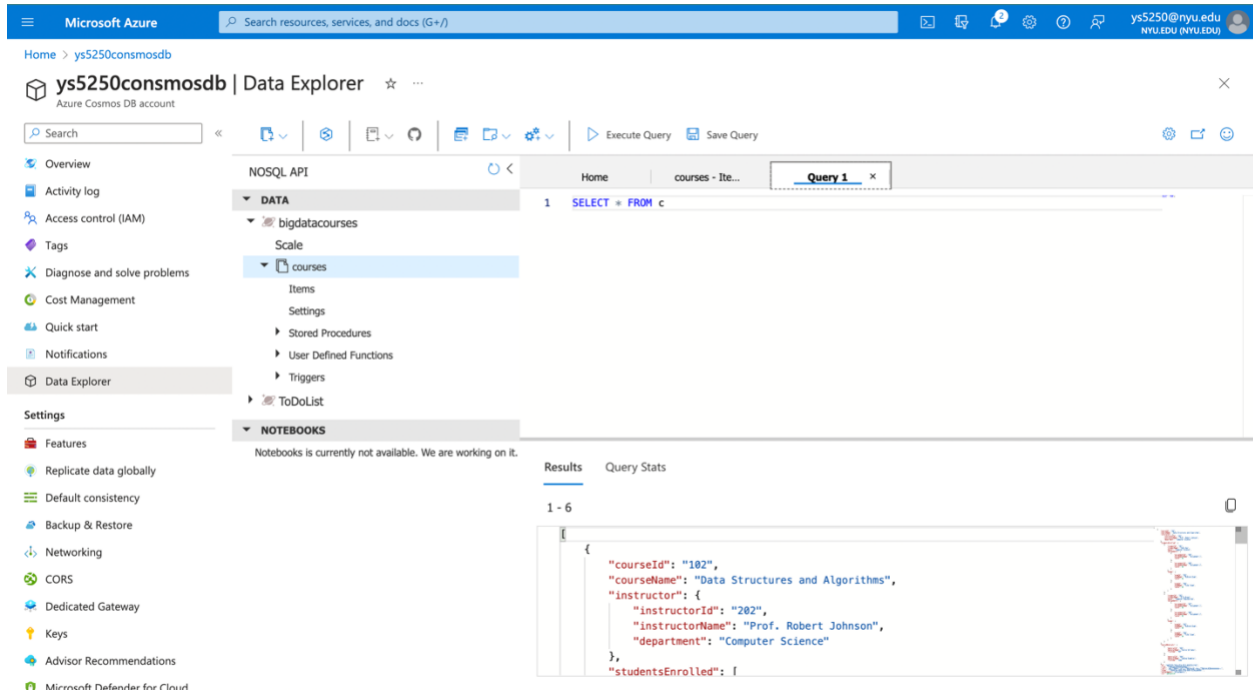
Fall'2023

Yogya Sharma

ys5250

GHW#3

Created database:



```
[
{
  "courseId": "102",
  "courseName": "Data Structures and Algorithms",
  "instructor": {
    "instructorId": "202",
    "instructorName": "Prof. Robert Johnson",
    "department": "Computer Science"
  },
  "studentsEnrolled": [
    {
      "studentId": "301",
      "studentName": "John Doe",
      "major": "Computer Science",
      "assignments": [
        {
          "assignmentId": "405",
          "assignmentName": "Assignment 1",
          "score": 95
        },
        {
          "assignmentId": "406",
          "assignmentName": "Assignment 2",
          "score": 88
        }
      ]
    }
  ],
  "exams": [
```

```
{
  {
    "examId": "505",
    "examName": "Midterm Exam",
    "score": 92
  },
  {
    "examId": "506",
    "examName": "Final Exam",
    "score": 85
  }
]
},
{
  "studentId": "303",
  "studentName": "Ella Brown",
  "major": "Computer Engineering",
  "assignments": [
    {
      "assignmentId": "407",
      "assignmentName": "Assignment 1",
      "score": 88
    },
    {
      "assignmentId": "408",
      "assignmentName": "Assignment 2",
      "score": 92
    }
  ],
  "exams": [
    {
      "examId": "507",
      "examName": "Midterm Exam",
      "score": 90
    },
    {
      "examId": "508",
      "examName": "Final Exam",
      "score": 94
    }
  ]
}
],
"courseMaterials": [
  {
    "materialId": "603",
    "materialName": "Course Syllabus",
    "fileType": "PDF"
  },
  {
    "materialId": "604",
    "materialName": "Coding Examples",
    "fileType": "Code"
  }
]
},
```

```

    "id": "4a873f83-7bfb-451b-8512-d64323c1f441",
    "_rid": "TTNHAJekiNIBAAAAAAAAA==",
    "_self": "dbs/TTNHAA==/colls/TTNHAJekiNI=/docs/TTNHAJekiNIBAAAAAAAAA==/",
    "_etag": "\"080091be-0000-0700-0000-6529a1ab0000\"",
    "_attachments": "attachments/",
    "_ts": 1697227179
  },
  {
    "courseId": "103",
    "courseName": "Big Data",
    "instructor": {
      "instructorId": "203",
      "instructorName": "Prof. Alice Wright",
      "department": "Computer Science"
    },
    "studentsEnrolled": [
      {
        "studentId": "303",
        "studentName": "Peter",
        "major": "Computer Science",
        "assignments": [
          {
            "assignmentId": "406",
            "assignmentName": "Assignment 1",
            "score": 90
          },
          {
            "assignmentId": "407",
            "assignmentName": "Assignment 2",
            "score": 99
          }
        ]
      },
      {
        "examId": "506",
        "examName": "Midterm Exam",
        "score": 90
      },
      {
        "examId": "507",
        "examName": "Final Exam",
        "score": 70
      }
    ]
  },
  {
    "studentId": "304",
    "studentName": "Harry Potter",
    "major": "Computer Engineering",
    "assignments": [
      {
        "assignmentId": "406",
        "assignmentName": "Assignment 1",
        "score": 64
      }
    ]
  }
]

```

```

    },
    {
      "assignmentId": "407",
      "assignmentName": "Assignment 2",
      "score": 88
    }
  ],
  "exams": [
    {
      "examId": "506",
      "examName": "Midterm Exam",
      "score": 56
    },
    {
      "examId": "507",
      "examName": "Final Exam",
      "score": 89
    }
  ]
}
],
"courseMaterials": [
  {
    "materialId": "604",
    "materialName": "Course Syllabus",
    "fileType": "PDF"
  },
  {
    "materialId": "605",
    "materialName": "Coding Examples",
    "fileType": "Code"
  }
],
"id": "ae1018f1-63e0-4f8e-8162-27163d4b5b8c",
"_rid": "TTNHAJekiNICAAAAAAAAA==",
"_self": "dbs/TTNHAA==/colls/TTNHAJekiNI=/docs/TTNHAJekiNICAAAAAAAAA==/",
"_etag": "\"080082fd-0000-0700-0000-6529a3b80000\"",
"_attachments": "attachments/",
"_ts": 1697227704
},
{
  "courseId": "104",
  "courseName": "DSP",
  "instructor": {
    "instructorId": "204",
    "instructorName": "Prof. Minerva McGonnagal",
    "department": "Computer Science"
  },
  "studentsEnrolled": [
    {
      "studentId": "304",
      "studentName": "Draco Malfoy",
      "major": "Computer Science",
      "assignments": [

```

```

    {
      "assignmentId": "407",
      "assignmentName": "Assignment 1",
      "score": 65
    },
    {
      "assignmentId": "408",
      "assignmentName": "Assignment 2",
      "score": 74
    }
  ],
  "exams": [
    {
      "examId": "507",
      "examName": "Midterm Exam",
      "score": 69
    },
    {
      "examId": "508",
      "examName": "Final Exam",
      "score": 80
    }
  ]
},
{
  "studentId": "305",
  "studentName": "Luna Lovegood",
  "major": "Computer Science",
  "assignments": [
    {
      "assignmentId": "407",
      "assignmentName": "Assignment 1",
      "score": 80
    },
    {
      "assignmentId": "408",
      "assignmentName": "Assignment 2",
      "score": 89
    }
  ],
  "exams": [
    {
      "examId": "507",
      "examName": "Midterm Exam",
      "score": 77
    },
    {
      "examId": "508",
      "examName": "Final Exam",
      "score": 90
    }
  ]
}
],

```

```

"courseMaterials": [
  {
    "materialId": "605",
    "materialName": "Course Syllabus",
    "fileType": "PDF"
  },
  {
    "materialId": "606",
    "materialName": "Coding Examples",
    "fileType": "Code"
  }
],
"id": "4ed89504-79b6-4b25-8feb-fda754297870",
"_rid": "TTNHAJekiNIDAAAAAAAAA==",
"_self": "dbs/TTNHAA==/colls/TTNHAJekiNI=/docs/TTNHAJekiNIDAAAAAAAAA==/",
"_etag": "\"0900091e-0000-0700-0000-6529a4c60000\"",
"_attachments": "attachments/",
"_ts": 1697227974
},
{
  "courseId": "105",
  "courseName": "Data Engineering",
  "instructor": {
    "instructorId": "205",
    "instructorName": "Prof. Severus Snape",
    "department": "Computer Science"
  },
  "studentsEnrolled": [
    {
      "studentId": "305",
      "studentName": "Ron Weasley",
      "major": "Computer Engineering",
      "assignments": [
        {
          "assignmentId": "408",
          "assignmentName": "Assignment 1",
          "score": 88
        },
        {
          "assignmentId": "409",
          "assignmentName": "Assignment 2",
          "score": 78
        }
      ]
    },
    {
      "examId": "508",
      "examName": "Midterm Exam",
      "score": 90
    },
    {
      "examId": "509",
      "examName": "Final Exam",
      "score": 88
    }
  ]
}

```

```

    }
  ]
},
{
  "studentId": "306",
  "studentName": "Hermonie Granger",
  "major": "Computer Engineering",
  "assignments": [
    {
      "assignmentId": "408",
      "assignmentName": "Assignment 1",
      "score": 99
    },
    {
      "assignmentId": "409",
      "assignmentName": "Assignment 2",
      "score": 100
    }
  ],
  "exams": [
    {
      "examId": "508",
      "examName": "Midterm Exam",
      "score": 100
    },
    {
      "examId": "509",
      "examName": "Final Exam",
      "score": 100
    }
  ]
}
],
"courseMaterials": [
  {
    "materialId": "606",
    "materialName": "Course Syllabus",
    "fileType": "PDF"
  },
  {
    "materialId": "609",
    "materialName": "Coding Examples",
    "fileType": "Code"
  }
],
"id": "7827ccf2-c6fe-4cf1-b54f-ee3e3533fbd9",
"_rid": "TTNHAJekiNIEAAAAAAAAA==",
"_self": "dbs/TTNHAA==/colls/TTNHAJekiNI=/docs/TTNHAJekiNIEAAAAAAAAA==/",
"_etag": "\"09009038-0000-0700-0000-6529a59e0000\"",
"_attachments": "attachments/",
"_ts": 1697228190
},
{
  "courseId": "106",

```



```
"courseName": "Data Science",
"instructor": {
  "instructorId": "206",
  "instructorName": "Prof. Horace Slughorn",
  "department": "Computer Science"
},
"studentsEnrolled": [
  {
    "studentId": "307",
    "studentName": "Neville Longbottom",
    "major": "Computer Science",
    "assignments": [
      {
        "assignmentId": "409",
        "assignmentName": "Assignment 1",
        "score": 78
      },
      {
        "assignmentId": "410",
        "assignmentName": "Assignment 2",
        "score": 83
      }
    ]
  },
  {
    "examId": "509",
    "examName": "Midterm Exam",
    "score": 88
  },
  {
    "examId": "510",
    "examName": "Final Exam",
    "score": 98
  }
]
},
{
  "studentId": "308",
  "studentName": "Ginny Weasley",
  "major": "Computer Science",
  "assignments": [
    {
      "assignmentId": "409",
      "assignmentName": "Assignment 1",
      "score": 76
    },
    {
      "assignmentId": "410",
      "assignmentName": "Assignment 2",
      "score": 91
    }
  ],
  "exams": [
    {
```

```

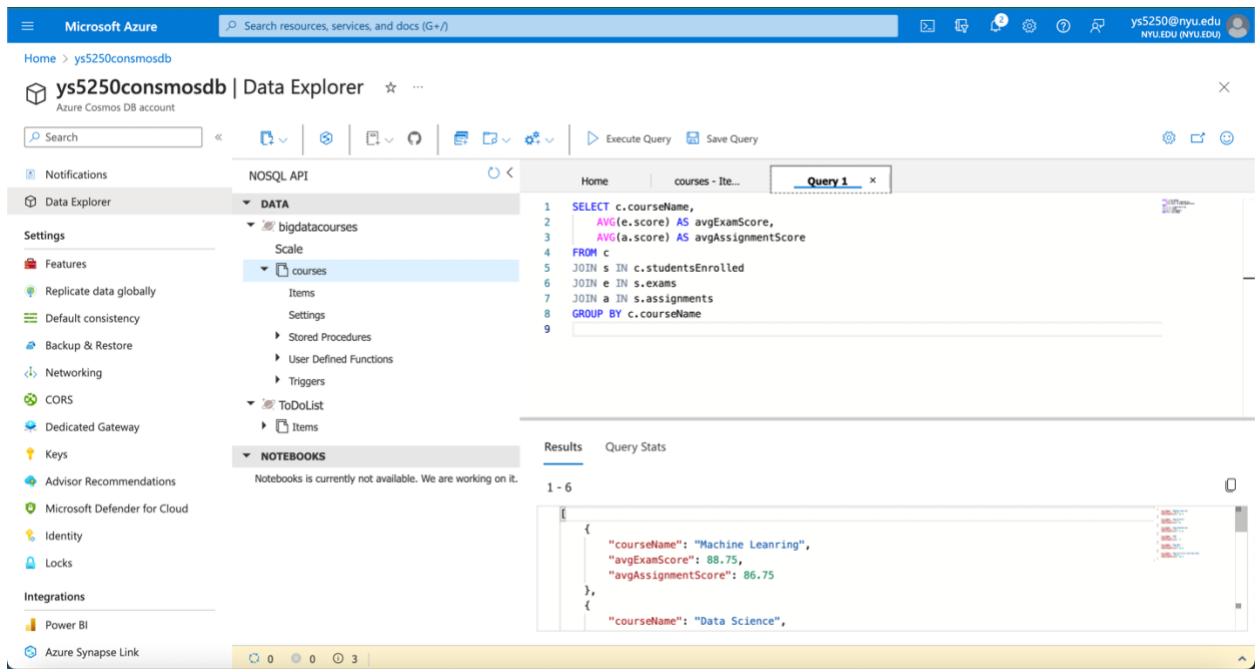
        "examId": "509",
        "examName": "Midterm Exam",
        "score": 89
    },
    {
        "examId": "510",
        "examName": "Final Exam",
        "score": 76
    }
]
}
],
"courseMaterials": [
    {
        "materialId": "607",
        "materialName": "Course Syllabus",
        "fileType": "PDF"
    },
    {
        "materialId": "610",
        "materialName": "Coding Examples",
        "fileType": "Code"
    }
],
"id": "56dd01e1-fa8e-4d9c-aa6f-c3f7baebd13e",
"_rid": "TTNHAJekiNIFAAAAAAAAA==",
"_self": "dbs/TTNHAA==/colls/TTNHAJekiNI=/docs/TTNHAJekiNIFAAAAAAAAA==/",
"_etag": "\"0900996f-0000-0700-0000-6529a7720000\"",
"_attachments": "attachments/",
"_ts": 1697228658
},
{
    "courseId": "107",
    "courseName": "Machine Learning",
    "instructor": {
        "instructorId": "207",
        "instructorName": "Prof. Rubeus Hagrid",
        "department": "Computer Science"
    },
    "studentsEnrolled": [
        {
            "studentId": "309",
            "studentName": "Priya Patel",
            "major": "Computer Science",
            "assignments": [
                {
                    "assignmentId": "410",
                    "assignmentName": "Assignment 1",
                    "score": 79
                },
                {
                    "assignmentId": "411",
                    "assignmentName": "Assignment 2",
                    "score": 73
                }
            ]
        }
    ]
}

```

```
    }
  ],
  "exams": [
    {
      "examId": "510",
      "examName": "Midterm Exam",
      "score": 89
    },
    {
      "examId": "511",
      "examName": "Final Exam",
      "score": 77
    }
  ]
},
{
  "studentId": "310",
  "studentName": "Cho Campbell",
  "major": "Computer Science",
  "assignments": [
    {
      "assignmentId": "410",
      "assignmentName": "Assignment 1",
      "score": 99
    },
    {
      "assignmentId": "411",
      "assignmentName": "Assignment 2",
      "score": 96
    }
  ],
  "exams": [
    {
      "examId": "510",
      "examName": "Midterm Exam",
      "score": 90
    },
    {
      "examId": "511",
      "examName": "Final Exam",
      "score": 99
    }
  ]
}
],
"courseMaterials": [
  {
    "materialId": "608",
    "materialName": "Course Syllabus",
    "fileType": "PDF"
  },
  {
    "materialId": "611",
    "materialName": "Coding Examples",
```

```
      "fileType": "Code"
    }
  ],
  "id": "f511c9d9-bd6d-4973-bfb6-11c8c277cc24",
  "_rid": "TTNHAJekiNIGAAAAAAAAA==",
  "_self": "dbs/TTNHAA==/colls/TTNHAJekiNI=/docs/TTNHAJekiNIGAAAAAAAAA==/",
  "_etag": "\"0900b489-0000-0700-0000-6529a84e0000\"",
  "_attachments": "attachments/",
  "_ts": 1697228878
}
]
```

1. Fetch the course name and the average exam and average assignment score for all courses.



Output:

```
[  
  {  
    "courseName": "Machine Learning",  
    "avgExamScore": 88.75,  
    "avgAssignmentScore": 86.75  
  },  
  {  
    "courseName": "Data Science",  
    "avgExamScore": 87.75,  
    "avgAssignmentScore": 82  
  },  
  {  
    "courseName": "Data Engineering",  
    "avgExamScore": 94.5,  
    "avgAssignmentScore": 91.25  
  },  
  {  
    "courseName": "DSP",  
    "avgExamScore": 79,  
    "avgAssignmentScore": 77  
  },  
  {  
    "courseName": "Big Data",  
    "avgExamScore": 76.25,  
    "avgAssignmentScore": 76.25  
  }  
]
```

```
    "avgAssignmentScore": 85.25
  },
  {
    "courseName": "Data Structures and Algorithms",
    "avgExamScore": 90.25,
    "avgAssignmentScore": 90.75
  }
]
```

- Calculate the average exam score for students majoring in "Computer Science" who are enrolled in the "Data Structures and Algorithms" course.

The screenshot shows the Microsoft Azure Data Explorer interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information for 'ys5250@nyu.edu'. The main header displays 'ys5250consmosdb | Data Explorer'. The left sidebar contains a navigation menu with options like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Cost Management, Quick start, Notifications, Data Explorer, and Settings. The central pane shows the 'DATA' section with a tree view containing 'bigdatacourses', 'Scale', 'courses', 'Items', 'Settings', 'Stored Procedures', 'User Defined Functions', 'Triggers', and 'ToDoList'. The 'courses' folder is selected. The right pane shows a query editor with the following SQL query:

```
1 SELECT AVG(e.score) AS avgExamScore
2 FROM c
3 JOIN s IN c.studentsEnrolled
4 JOIN e IN s.exams
5 WHERE c.courseName = "Data Structures and Algorithms" AND s.major = "Computer Science"
6
```

Below the query editor, the 'Results' tab is active, displaying the query results in a JSON format:

```
[
  {
    "avgExamScore": 88.5
  }
]
```

The bottom of the interface shows a 'Filter' dropdown set to 'All' and a 'Clear Notifications' link.

Output:

```
[
  {
    "avgExamScore": 88.5
  }
]
```

3. Find the courses where at least one student's major is "Computer Engineering" and they scored below 70 on exams.

The screenshot shows the Microsoft Azure Data Explorer interface. The left sidebar contains navigation options like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Cost Management, Quick start, Notifications, Data Explorer, Settings, Features, Replicate data globally, Default consistency, Backup & Restore, Networking, CORS, Dedicated Gateway, Keys, Advisor Recommendations, and Microsoft Defender for Cloud. The main area displays the 'Data Explorer' for the 'ys5250consmosdb' account. The 'DATA' section is expanded, showing 'bigdatacourses' and 'courses'. The 'courses' collection is selected, and a query is executed. The query is as follows:

```
1 SELECT c.courseName
2 FROM c
3 JOIN s IN c.studentsEnrolled
4 JOIN e IN s.exams
5 WHERE s.major = "Computer Engineering" AND e.score < 70
6
```

The query results are displayed in the 'Results' tab, showing one item: 'Big Data'.

```
[
  {
    "courseName": "Big Data"
  }
]
```

A notification at the bottom indicates: '4:56 PM Successfully fetched 1 item for container courses'.

Output:

```
[
  {
    "courseName": "Big Data"
  }
]
```



4. Select any two courses where there is at least one student's major that is "Computer Science" (There must be at least three students in different courses, who's major is Computer Science)

The screenshot shows the Microsoft Azure Data Explorer interface. The left sidebar contains navigation options like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Cost Management, Quick start, Notifications, Data Explorer, Settings, Features, Replicate data globally, Default consistency, Backup & Restore, Networking, CORS, Dedicated Gateway, Keys, Advisor Recommendations, and Microsoft Defender for Cloud. The main pane displays a query in the 'Query 1' tab. The query is a SQL statement that selects the top 2 course names from the 'bigdatacourses' collection, joined with the 'studentsEnrolled' collection, where the student's major is 'Computer Science'. The results pane shows two JSON documents: one for 'Machine Learning' and one for 'Data Science'.

```
1 SELECT TOP 2 c.courseName
2 FROM c
3 JOIN s IN c.studentsEnrolled
4 WHERE s.major = "Computer Science"
5 GROUP BY c.courseName
6
```

```
{
  "courseName": "Machine Learning"
},
{
  "courseName": "Data Science"
}
```

Output:

```
[
  {
    "courseName": "Machine Learning"
  },
  {
    "courseName": "Data Science"
  }
]
```

- The screenshot displays the Azure portal's Data Explorer for an Azure Cosmos DB account. The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Cost Management, Quick start, Notifications, Data Explorer, Settings, and Features. The main area shows the 'courses' collection under the 'DATA' section. A query editor is open with the following SQL query:

```
SELECT s.studentName,
IF(EXISTS(SELECT VALUE e.score FROM e IN s.exams WHERE e.score < 70), 'Failed', 'Passed')
AS passGrade
FROM c
JOIN s IN c.studentsEnrolled
```

The query results are displayed below, showing a single record for 'John Doe' with a 'passGrade' of 'Passed'. A message at the bottom states: 'We have detected you may be using a subquery. Non-correlated subqueries are not currently supported. Please see Cosmos sub query documentation for further information.'

```
[
  {
    "studentName": "John Doe",
    "passGrade": "Passed"
  },
  {
    "studentName": "Ella Brown",
    "passGrade": "Passed"
  },
  {
    "studentName": "Peter",
    "passGrade": "Passed"
  },
  {
    "studentName": "Harry Potter",
    "passGrade": "Failed"
  },
  {
    "studentName": "Draco Malfoy",
    "passGrade": "Failed"
  },
  {

```

```
    "studentName": "Luna Lovegood",  
    "passGrade": "Passed"  
  },  
  {  
    "studentName": "Ron Weasley",  
    "passGrade": "Passed"  
  },  
  {  
    "studentName": "Hermonie Granger",  
    "passGrade": "Passed"  
  },  
  {  
    "studentName": "Neville Longbottom",  
    "passGrade": "Passed"  
  },  
  {  
    "studentName": "Ginny Weasley",  
    "passGrade": "Passed"  
  },  
  {  
    "studentName": "Priya Patel",  
    "passGrade": "Passed"  
  },  
  {  
    "studentName": "Cho Campbell",  
    "passGrade": "Passed"  
  }  
]  

```

## Learning:

In this assignment I've gained valuable insights into the world of data modeling and querying using Azure Cosmos DB. The task revolved around creating a structured database for educational courses, instructors, and students, showcasing the importance of sound data modeling practices. I learned how to design a NoSQL database schema that accommodates complex relationships within the data. The assignment further improved my proficiency in crafting Cosmos DB queries to extract meaningful information from the dataset, including calculating averages, filtering data, and assessing pass/fail status based on exam scores. Understanding the data import and export process, involving Azure services like Azure Data Factory, proved invaluable for potential real-world data migration and analysis tasks.