Engenharia Informática

Sistemas de Informação - Parte 2

Jorge Ricardo Marques Duarte

Instituto Superior de Engenharia de Coimbra

2021110042

5 de dezembro de 2024

Conteúdo

1 Processo de negócio a modelar

O processo de negócio a modelar no contexto deste Data Warehouse (DW) está relacionado à gestão acadêmica e análise de desempenho de estudantes. Ele abrange o armazenamento e análise de dados sobre:

Informações demográficas e socioeconômicas dos estudantes (ex.: idade, renda familiar, acesso à internet). Matrículas em cursos e disciplinas, incluindo status financeiro e modos de inscrição. Estrutura de cursos e disciplinas, como tipo, duração e créditos (ECTS). Análise de desempenho acadêmico, incluindo notas finais e status de aprovação. Dados temporais e anos acadêmicos para acompanhar a evolução e conclusão de cursos. Esse modelo suporta decisões estratégicas, como alocação de recursos, identificação de padrões de sucesso e falhas, e políticas de inclusão socioeconômica.

2 Avaliação da granularidade

Na avaliação da granularidade, inicialmente estava voltado para um fato *F_EXAMS*, mas rapidamente notei que não representava adequadamente o desempenho acadêmico dos estudantes. Isso porque as notas finais não dependem exclusivamente dos exames, mas também de trabalhos práticos, projetos e outros componentes avaliativos. Usar apenas essa tabela implicaria assumir um erro ao simplificar o cálculo da nota final.

Por isso, optei por modelar a tabela *F_ACADEMIC_PERFORMANCE*, que captura o desempenho acadêmico de forma mais abrangente, incluindo a granularidade ao nível de cada estudante em cada disciplina. A granularidade deste Data Warehouse está, portanto, definida no nível de estudante, disciplina, curso, ano acadêmico e período de tempo, permitindo análises detalhadas e precisas sobre o desempenho e evolução dos estudantes.

3 Dimensões e os atributos

No modelo do Data Warehouse, as dimensões foram cuidadosamente definidas para permitir análises detalhadas do desempenho acadêmico. As principais dimensões e seus atributos são:

- **D_STUDENTS**: contém informações dos estudantes, como:
  - `STUDENT_ID` (identificador único do estudante),
  - `NAME` (nome do estudante),
  - `SOCIOECONOMIC_ID` (relacionado à dimensão socioeconômica),
  - `DEMOGRAPHIC_ID` (relacionado à dimensão demográfica).
- **D_SOCIOECONOMIC_DATA**: armazena dados socioeconômicos, como:
  - `SCHOLARSHIP_STATUS` (se possui bolsa de estudos),
  - `FAMILY_INCOME` e `INCOME` (renda familiar e individual),
  - `HAS_INTERNET_ACCESS` e `HAS_COMPUTER_ACCESS` (acessos básicos de tecnologia),
  - `WORKING_STATUS` (situação de trabalho).
- **D_STUDENT_DEMOGRAPHIC_DATA**: inclui dados demográficos, como:
  - `DATE_OF_BIRTH` (data de nascimento),
  - `NATIONALITY` (nacionalidade),
  - `GENDER` e `ETHNICITY` (dados de género e etnia),
  - `CITY_OF_BIRTH` e `COUNTRY_OF_BIRTH` (local de nascimento).
- **D_COURSES**: descreve os cursos, com:
  - `COURSE_ID` (identificador único),
  - `COURSE_NAME` (nome do curso),
  - `FIELD_OF_STUDY_ID` (área de estudo relacionada),
  - `DURATION_YEARS` (duração do curso em anos).
- **D_FIELDS_OF_STUDY**: define as áreas de estudo:
  - `FIELD_ID` (identificador único da área),
  - `FIELD_NAME` (nome da área de estudo).
- **D_TIME**: armazena dimensões temporais, como:

  - `DAY`, `MONTH`, `YEAR` e `SEMESTER` (dados temporais detalhados),
  - `WEEKDAY` (dia da semana),
  - `DATE` (data específica).

- **D_ACADEMIC_YEAR**: guarda informações sobre o ano letivo, com:
  - `ACADEMIC_YEAR_ID` (identificador único),
  - `ACADEMIC_YEAR` (descrição do ano letivo),
  - `START_DATE` e `END_DATE` (datas de início e fim).

Essas dimensões suportam a análise detalhada de desempenho acadêmico ao nível individual, temporal e institucional, proporcionando um modelo robusto para responder às necessidades do negócio.

4 Factos a serem incluídos

As tabelas de factos criadas para este Data Warehouse foram desenhadas para suportar análises específicas e detalhadas sobre o desempenho acadêmico e os indicadores de sucesso dos estudantes. As tabelas de factos são as seguintes:

4.1 F_ACADEMIC_PERFORMANCE

Esta tabela de factos registra o desempenho dos estudantes em cada disciplina. A granularidade está ao nível da inscrição de um estudante numa disciplina, para um determinado ano acadêmico. Os atributos desta tabela incluem:

- `ENROLLMENT_SUBJECT_ID`: identificador único da inscrição numa disciplina.
- `ENROLLMENT_ID`: identificador da inscrição no curso.
- `SUBJECT_ID`: identificador da disciplina.
- `STUDENT_ID`: identificador do estudante.
- `ACADEMIC_YEAR_ID`: identificador do ano acadêmico.
- `COURSE_ID`: identificador do curso.
- `TIME_ID`: identificador da dimensão temporal.
- `FINAL_GRADE`: nota final obtida pelo estudante na disciplina.
- `STATUS`: status da disciplina (ex.: aprovado, reprovado).

4.2 F_BENCHMARK_SUCCESS

Esta tabela de factos foca nos indicadores de sucesso dos estudantes, como conclusão de curso e situação profissional. A granularidade é por estudante e curso. Os atributos desta tabela incluem:

- `BENCHMARK_SUCCESS_ID`: identificador único do registo de sucesso.
- `STUDENT_ID`: identificador do estudante.
- `COURSE_ID`: identificador do curso.
- `ACADEMIC_YEAR_OF_COMPLETION_ID`: identificador do ano acadêmico de conclusão.
- `VERIFICATION_TIME_DATE_ID`: identificador da data de verificação.

- `WORKING_ON_FIELD_DATE_SINCE_ID`: identificador da data em que o estudante começou a trabalhar na sua área.
- `COURSE_CONCLUDED`: indicador booleano se o curso foi concluído.
- `IS_WORKING_ON_THE_FIELD`: indicador booleano se o estudante está a trabalhar na sua área de formação.

Estas tabelas de factos são alimentadas pelas dimensões relevantes, como `D_STUDENTS`, `D_COURSES`, `D_SUBJECTS`, `D_TIME` e `D_ACADEMIC_YEAR`, garantindo uma modelagem robusta para suportar análises e relatórios detalhados.

5 Modelo em estrela - Constelação

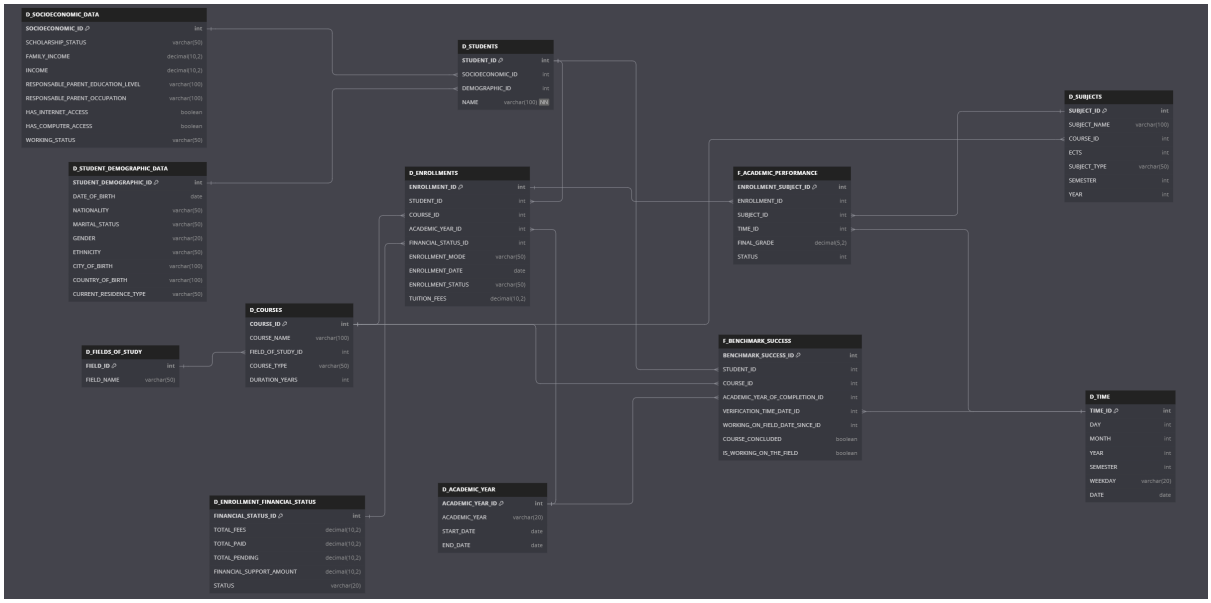

Figura 1: Modelo em estrela inicial

Este é o modelo em estrela inicial, onde podemos observar que neste momento ainda não continha os campos *COURSE_ID, STUDENT_ID, e ACADEMIC_YEAR_ID* na tabela de fato *F_ACADEMIC_PERFORMANCE*.

Posteriormente, de modo a melhorar a performance das queries, acabei por defini-los no modelo onde ficou como podemos observar na Figura 2.

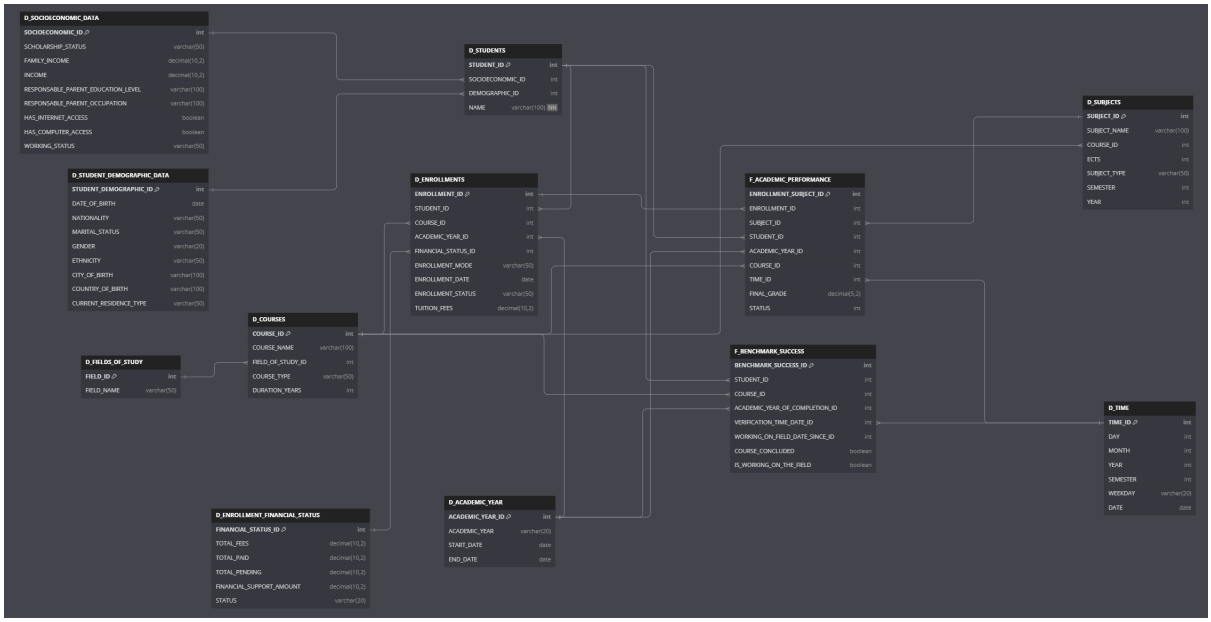

Figura 2: Modelo em estrela - Final

6 Cálculos aproximados do tamanho que ocupará o modelo concebido

Nota: Suponha que as condições seguintes serão observadas: terá de registar dados de 5 anos e a existem cerca de 30 cursos. Contemple apenas cursos de licenciatura (3 anos). Suponha que uma turma só pode ter 40 alunos no máximo.

6.1 Estimativa de registos

Primeiramente, calculei o número total de alunos que estarão presentes no sistema ao longo de 5 anos:

- Para 30 cursos, com 40 alunos por turma, temos um total de $30 \times 40 = 1200$ alunos por ano.
- Com 5 anos de dados, o total de alunos é $1200 \times 5 = 6000$.

6.2 Cálculo de Tamanho por Tabela

Abaixo, apresento os cálculos detalhados para cada tabela, considerando os campos, tipos de dados e número de registos:

6.2.1 Tabela `D_STUDENTS`

A tabela `D_STUDENTS` contém as seguintes colunas principais:

- `STUDENT_ID` (4 bytes)
- `SOCIOECONOMIC_ID`, `DEMOGRAPHIC_ID` (4 bytes cada)
- `NAME` (100 caracteres, média de 50 bytes por nome)

O tamanho total por registo é $4 + 4 + 4 + 50 = 62$ bytes. Com 6000 registos:

$$6000 \times 62 = 372\,\text{KB}.$$

6.2.2 Tabela `D_COURSES`

Esta tabela possui:

- `COURSE_ID` (4 bytes)
- `COURSE_NAME` (100 bytes)
- `FIELD_OF_STUDY_ID`, `DURATION_YEARS` (4 bytes cada)

Com $4 + 100 + 4 + 4 = 112$ bytes por registo e 30 registos:

$$30 \times 112 = 3.36\,\text{KB}.$$

6.2.3 Tabela `F_ACADEMIC_PERFORMANCE`

Esta tabela é a mais volumosa, pois contém um registo para cada disciplina frequentada por cada aluno. Os campos principais incluem:

- IDs (`ENROLLMENT_SUBJECT_ID`, `ENROLLMENT_ID`, etc.) com 4 bytes cada.
- `FINAL_GRADE` (5 bytes) e `STATUS` (4 bytes).

O total por registo é $7 \times 4 + 5 + 4 = 37$ bytes. Considerando que cada aluno frequenta 10 disciplinas por ano:

$$6000 \times 3 \times 10 = 180,000\,\text{registos}.$$

O tamanho total da tabela é:

$$180,000 \times 37 = 6.66\,\text{MB}.$$

6.2.4 Tabela `F_BENCHMARK_SUCCESS`

Por fim, esta tabela possui campos principais como:

- IDs (`BENCHMARK_SUCCESS_ID`, `STUDENT_ID`, etc.) com 4 bytes cada.
- `COURSE_CONCLUDED`, `IS_WORKING_ON_THE_FIELD` (1 byte cada).

O total por registo é $6 \times 4 + 2 = 26$ bytes. Com 6000 registos:

$$6000 \times 26 = 156\,\text{KB}.$$

6.3 Tamanho Total do Modelo

Somando todas as tabelas principais, o tamanho aproximado do modelo é:

$$372\,\text{KB} + 3.36\,\text{KB} + 6.66\,\text{MB} + 156\,\text{KB} = 7.19\,\text{MB}.$$

6.4 Explicação

O tamanho total do modelo é dominado pela tabela `F_ACADEMIC_PERFORMANCE`, devido à alta granularidade dos dados. Este modelo foi projetado para suportar consultas analíticas detalhadas, sendo eficiente para armazenar informações de 5 anos de atividades acadêmicas.

7 Implementação modelo em estrela no Oracle

Para implementar o modelo em estrela no SQL Server, utilizei o Liquibase para definir as tabelas dimensionais e factuais com suas devidas relações. O uso do Liquibase permite a gestão de alterações no esquema do banco de dados de forma controlada e rastreável, facilitando atualizações e rollback quando necessário.

7.1 Benefícios do Uso de Liquibase

- **Rastreabilidade**: Cada alteração no esquema é registrada, permitindo auditorias. - **Automação**: Scripts podem ser aplicados automaticamente em diferentes ambientes. - **Rollback**: Facilita reverter mudanças em caso de erro.

Com essa abordagem, assegurei uma implementação consistente do modelo em estrela, alinhada aos requisitos do Data Warehouse.

O script liquibase está presente no projeto datawarehouse, no ficheiro *V1__create_constellation.yml*.

8 Sumário dos Valores introduzidos por tabela

Os seguintes resultados mostram a contagem de registos inseridos em cada tabela:

Tabela	**Total de Registos**
D_STUDENTS	33150
D_STUDENT_DEMOGRAPHIC_DATA	33150
D_SOCIOECONOMIC_DATA	33150
D_COURSES	58
D_FIELDS_OF_STUDY	48
D_SUBJECTS	2006
D_TIME	7016
D_ACADEMIC_YEAR	21
F_ACADEMIC_PERFORMANCE	1126081
F_BENCHMARK_SUCCESS	7115

Tabela 1: Sumário dos Registos nas Tabelas

9 Dicionário de Dados

Para a apresentação e catalogação das informações de colunas, tabelas e as suas relações criei um projeto chamado Data Dictionary, onde apresenta as relações e informações configuradas através do ficheiro *metadata.json*

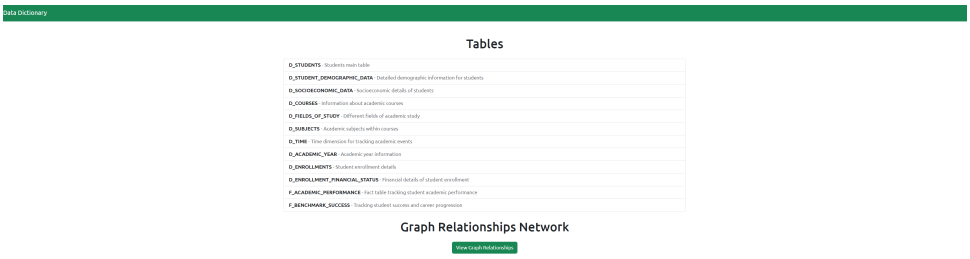

Figura 3: Data Dictionary - Página Principal

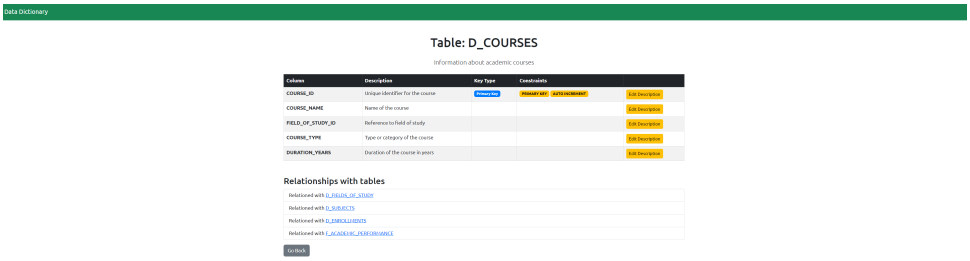

Figura 4: Visualização de Informação de Tabela

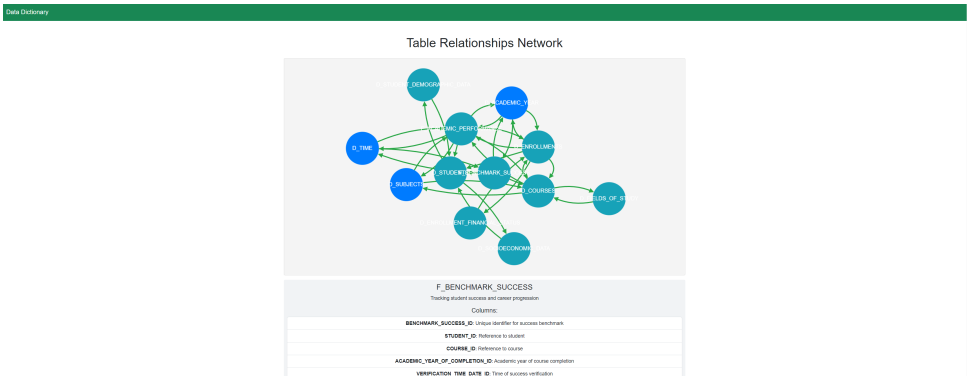

Figura 5: Data Dictionary - Visualização de relações em Graph

10 Power BI Dashboard

No presente trabalho, criei um relatório em Power BI que contém várias visualizações relacionadas com os dados de matrícula e desempenho acadêmico. O documento é composto pelas seguintes páginas:

- Figura 6: Relatório Geral, que apresenta informações gerais como por exemplo o número de alunos e cursos.

- Figura 7: Informação de Matrículas, exibindo dados detalhados sobre o total de matrículas e inscrições.

- Figura 8: *Performance de Alunos Acadêmica, com a análise do desempenho dos alunos nas disciplinas e estatísticas de sucesso.

- Figura 9: Sucesso Acadêmico, que mostra as taxas de sucesso dos alunos por terem frequentado o curso.

Essas páginas fornecem uma visão abrangente dos dados de performance e sucesso dos alunos em diferentes áreas do sistema educacional.

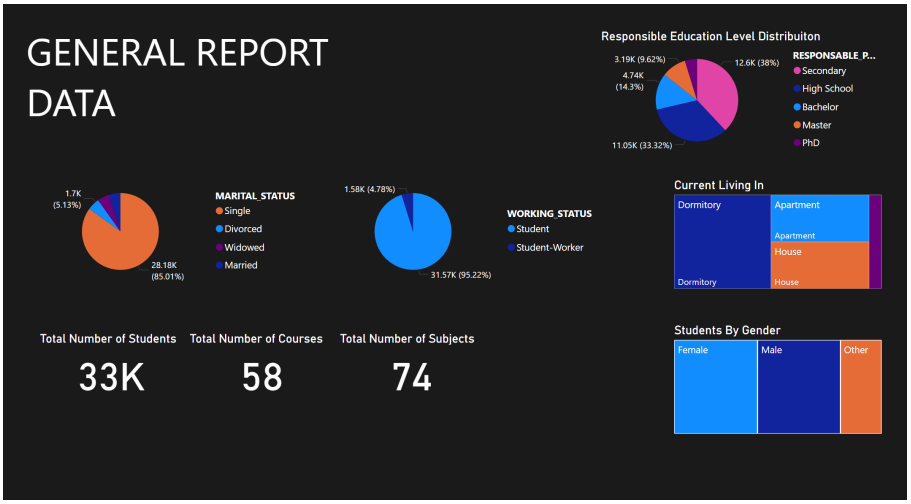

Figura 6: General Relatório Geral

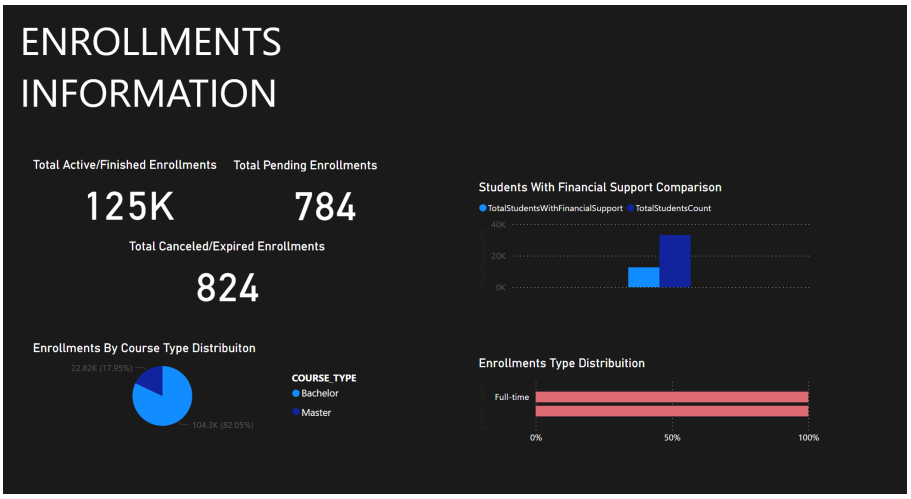

Figura 7: Informação de matriculas

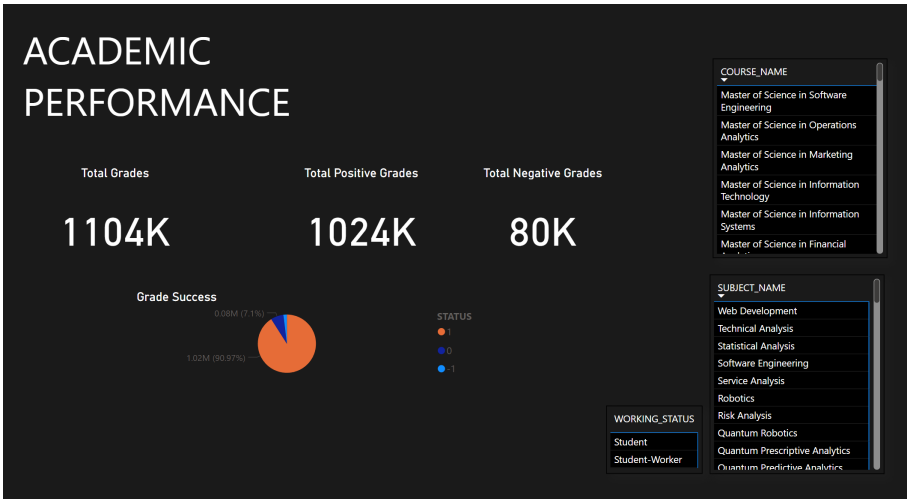

Figura 8: Performance de Alunos Académica

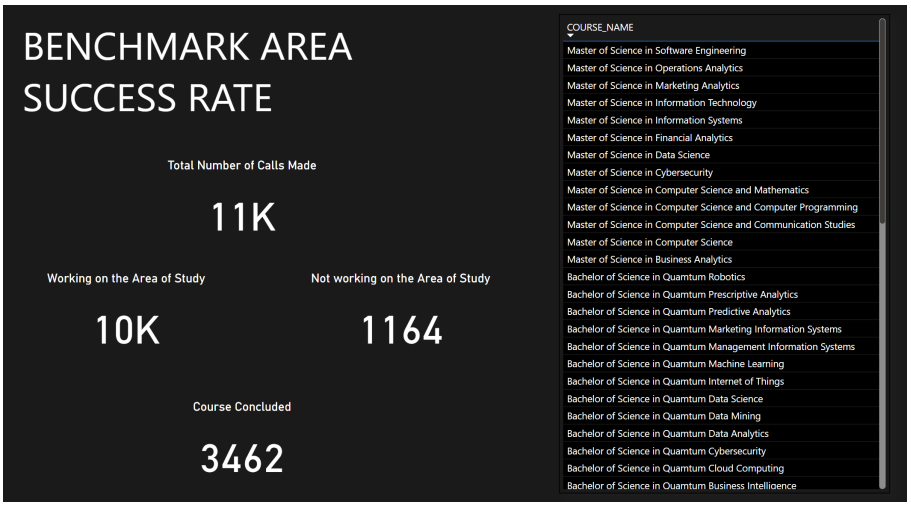

Figura 9: Sucesso Académico