

Great Expectations

Luiza Ramos Pascuotte
luizaramospascuotte@gmail.com
luiza.pascuotte@fatec.sp.gov.br

Resumo

Great Expectations (GX Core) é uma ferramenta poderosa de validação e monitoramento de dados que auxilia na garantia da qualidade de dados em pipelines analíticos e sistemas orientados a dados. Este capítulo aborda os fundamentos do GX Core, explicando como ele se integra em ambientes modernos de engenharia de dados, seus principais recursos, benefícios, e exemplos de uso prático. O objetivo é capacitar desenvolvedores e analistas com o conhecimento necessário para implementar práticas robustas de qualidade de dados utilizando esta ferramenta.

Palavras-chaves: *Great Expectations*, validação de dados, *pipeline* de dados, automação de testes, qualidade de dados.

Introdução

Em um mundo cada vez mais orientado por dados, garantir a qualidade e a integridade dos mesmos tornou-se essencial para o sucesso de organizações e projetos tecnológicos. Erros em dados podem levar a decisões incorretas, impactos financeiros negativos e até mesmo problemas reputacionais. Nesse contexto, surge o **Great Expectations**, uma poderosa ferramenta de validação de dados que se integra a diversos fluxos de trabalho, permitindo verificar a consistência, a completude e o formato dos dados automaticamente.

O **Great Expectations** permite definir "expectativas" específicas sobre os dados, como faixas de valores, formatos de datas ou a consistência entre colunas de um banco de dados. Ele se destaca pela flexibilidade de uso em diferentes fontes, como arquivos CSV, bancos de dados SQL e data lakes, além de ser amplamente utilizado em pipelines de ciência de dados e aprendizado de máquina.

Desenvolvimento

Nesta seção, abordaremos como utilizar o módulo GX Core do Great Expectations para validar dados de maneira programática, ilustrando com um exemplo prático baseado em um conjunto de dados de amostra.

O *Great Expectations* é uma ferramenta de código aberto voltada para validação, documentação e monitoramento de dados. Seu principal objetivo é ajudar equipes de dados a garantir que as informações processadas atendam a critérios pré-definidos de qualidade. O **GX Core**, a versão base do *framework*, foca na definição de "expectativas" – regras ou testes que os dados devem atender.

Principais funcionalidades do GX Core:

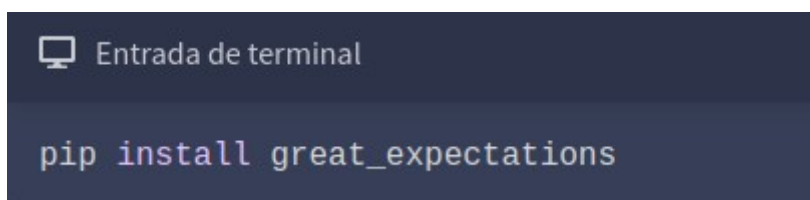
- **Validação de Dados:** Verifica automaticamente se os dados cumprem as regras especificadas.
- **Documentação Automatizada:** Gera documentação de dados em formato legível por humanos e máquinas.
- **Monitoramento Contínuo:** Identifica anomalias nos dados ao longo do tempo.
- **Flexibilidade e Integração:** Compatível com diversas fontes de dados, incluindo bancos SQL, arquivos CSV, JSON, Parquet, Data Lakes e armazenamento na nuvem, e sistemas de Big Data como Spark.

GX Core é uma biblioteca Python que você pode instalar com *pip install Python*, a versão utilizada dessa ferramenta para esse estudo é a versão 1.2.3 que suporta da versão 3.9 à 3.12 do Python.

A configuração do ambiente em que executaremos a ferramenta pode ser acompanhado através desse repositório do GitHub(https://github.com/LuizaPascuotte/tests_gx_core). Após a configuração do repositório em que desenvolveremos, a configuração do GX Core deve ser feita nessa ordem:

1. Instale a biblioteca GX Core através do seguinte comando no seu terminal na pasta do seu projeto:

Imagem1: Instalação da biblioteca GX Core.



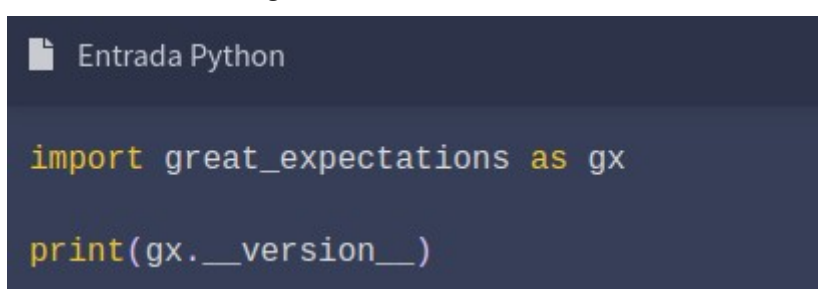
```
Entrada de terminal

pip install great_expectations
```

Autoria(https://docs.greatexpectations.io/docs/core/introduction/try_gx?procedure=instructions)

2. Verifique se o GX Core foi instalado com sucesso executando o comando abaixo no seu interpretador *Python*, IDE, *notebook* ou *script*:

Imagem2: Verificar versão instalada



```
Entrada Python

import great_expectations as gx

print(gx.__version__)
```

Autoria(https://docs.greatexpectations.io/docs/core/introduction/try_gx?procedure=instructions)

Cenário

No exemplo a seguir, validamos os dados de uma tabela que contém informações sobre viagens de táxi. O objetivo é garantir que a coluna *passenger_count* (número de passageiros) contenha apenas valores entre 1 e 6, conforme esperado.

Os dados foram carregados de um arquivo CSV hospedado em um repositório público e processados utilizando o **Pandas**, com validações realizadas via **Great Expectations**.

Etapas do Processo

1. Criação do Contexto de Dados

O **Contexto de Dados** é o ponto de entrada principal para trabalhar com o **Great Expectations**. Ele permite configurar fontes de dados, criar expectativas e validar lotes de dados. No exemplo, foi utilizado o comando:

Imagem3: Criando contexto

```
context = gx.get_context()
```

Autoria própria(2024)

2. Importação dos Dados

Os dados foram carregados diretamente de um arquivo CSV online utilizando o Pandas:

Imagem4: Importando os dados

```
df = pd.read_csv(  
    "https://raw.githubusercontent.com/great-expectations/gx_tutorials/main/data/yellow_tripdata_sample_2019-01.csv"  
)
```

Autoria própria(2024)

O arquivo contém colunas como *vendor_id*, *pickup_datetime*, *dropoff_datetime*, e *passenger_count*, entre outras, conforme mostrado abaixo (exemplo reduzido):

vendor_id	pickup_datetime	dropoff_datetime	passenger_count	trip_distance	fare_amount
1	2019-01-15 03:36:12	2019-01-15 03:42:19	1	1.0	6.5

3. Configuração do *Data Source* e *Batch*

Para validar os dados, é necessário criar uma **Fonte de Dados** (*DataSource*) e associá-la a um **Data Asset**. Em seguida, define-se um **Batch** (lote de dados), que representa uma porção dos dados a ser validada.

Imagem5: Configuração do data source e batch

```
data_source = context.data_sources.add_pandas("pandas")  
data_asset = data_source.add_dataframe_asset(name="pd dataframe asset")  
  
batch_definition = data_asset.add_batch_definition_whole_dataframe("batch definition")  
batch = batch_definition.get_batch(batch_parameters={"dataframe": df})
```

Autoria própria(2024)

Esses comandos configuram o Pandas *DataFrame* como fonte de dados e associam o *DataFrame* completo ao lote de validação.

4. Criação da Expectativa

As expectativas são declarações sobre o que os dados devem obedecer. No exemplo, foi criada uma expectativa para validar que os valores na coluna *passenger_count* estejam no intervalo entre 1 e 6:

Imagem6: Criando as expectativas

```
expectation = gx.expectations.ExpectColumnValuesToBeBetween(  
    | column="passenger_count", min_value=1, max_value=6  
    | )
```

Autoria própria(2024)

5. Validação dos Dados

O lote de dados foi validado contra a expectativa definida:

Imagem7: Validação dos dados

```
validation_result = batch.validate(expectation)
```

Autoria própria(2024)

O resultado da validação é um relatório detalhado em formato JSON, conforme mostrado a seguir:

Imagem8: Resultados

```
Calculating Metrics: 100%|
.51it/s]
{
  "success": true,
  "expectation_config": {
    "type": "expect_column_values_to_be_between",
    "kwargs": {
      "batch_id": "pandas-pd dataframe asset",
      "column": "passenger_count",
      "min_value": 1.0,
      "max_value": 6.0
    },
    "meta": {}
  },
  "result": {
    "element_count": 10000,
    "unexpected_count": 0,
    "unexpected_percent": 0.0,
    "partial_unexpected_list": [],
    "missing_count": 0,
    "missing_percent": 0.0,
    "unexpected_percent_total": 0.0,
    "unexpected_percent_nonmissing": 0.0,
    "partial_unexpected_counts": [],
    "partial_unexpected_index_list": []
  },
}
```

Autoria própria(2024)

Análise dos Resultados

- **success: true:** A validação foi bem-sucedida. Todos os valores da coluna *passenger_count* estavam dentro do intervalo de 1 a 6.
- **element_count: 10000:** O número total de elementos validados foi 10.000.
- **unexpected_count: 0:** Não houve valores fora do intervalo especificado.
- **missing_count: 0:** Nenhum valor estava ausente na coluna validada.

Conclusão

O exemplo demonstrou como configurar um Contexto de Dados, carregar e validar dados diretamente de um Pandas *DataFrame* utilizando o *GX Core* do *Great Expectations*. Essa abordagem programática é eficiente para integrar validações de dados em pipelines e fluxos de trabalho automatizados.

Caso surjam inconsistências nos dados, o *Great Expectations* fornece relatórios detalhados para identificação e resolução dos problemas.

Considerações Finais

Neste capítulo, foi apresentado um exemplo estudo rápido e prático do módulo *GX Core* do **Great Expectations**. Apesar de ter sido demonstrado apenas um fluxo básico, a ferramenta oferece inúmeras possibilidades de integração, incluindo bancos de dados relacionais, plataformas como *Snowflake* e armazenamentos em nuvem, além do suporte à sua interface própria, o **GX Cloud**. Além disso, é possível criar uma ampla gama de expectativas para atender às mais diversas necessidades de validação de dados.

O **Great Expectations (GX)** é uma estrutura poderosa para descrever dados de maneira expressiva e validar se eles atendem aos critérios definidos por meio de testes. O módulo *GX Core*, por sua vez, fornece uma interface programática em *Python*, permitindo construir e executar fluxos de validação de dados de forma flexível e automatizada.

Os fluxos de trabalho criados com *GX Core* são altamente personalizáveis e permitem validar dados em diferentes contextos, desde pipelines simples até processos complexos de grandes volumes de dados. No entanto, trabalhar com *GX Core* exige familiaridade com a linguagem *Python* e conceitos de manipulação de dados, o que torna a ferramenta especialmente indicada para profissionais com conhecimento técnico avançado.

De maneira geral, a experiência de desenvolvimento com o *GX Core* é enriquecedora, especialmente para cenários onde a qualidade dos dados é crítica. No futuro, explorar mais módulos e funcionalidades da ferramenta, como a criação de relatórios automatizados ou a integração com ferramentas de monitoramento, pode abrir ainda mais possibilidades para validação e governança de dados.

Bibliografia

GREAT EXPECTATIONS. Documentation. Disponível em: <https://docs.greatexpectations.io/>.

Acesso em: 13 nov. 2024.

Data Cloud | Snowflake Brasil. Disponível em: <https://www.snowflake.com/pt_br/>. Acesso em: 06 dez. 2024.

great-expectations/great_expectations: Always know what to expect from your data. Disponível em: <https://github.com/great-expectations/great_expectations/tree/develop>. Acesso em: 6 dez. 2024.

TLC Trip Record Data - TLC. Disponível em: <<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>>.