UNIVERSIDADE FEDERAL DE PERNAMBUCO
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

DataScript

ANDREZZA DE MELO BONFIM
ATHOS PUGLIESE
JORDAN KALLIURE SOUZA CARVALHO

PARADIGMAS DE LINGUAGENS DE PROGRAMAÇÃO

RECIFE
2025

Proposta/Objetivo do Projeto

Este projeto tem como objetivo desenvolver uma DSL que permitirá ao usuário carregar, analisar, filtrar e visualizar as características fundamentais de um conjunto de dados de forma rápida e intuitiva.

A DSL será uma extensão da linguagem imperativa 2 do JavaCC, de forma que sua utilização seja intuitiva, permitindo que pessoas não técnicas em dados escrevam scripts em uma linguagem de alto nível e expressiva, sem precisarem conhecer a fundo cada detalhe da API do Pandas.

Gerenciamento de Dados

Carregar Dados: Ler e interpretar conjuntos de dados a partir de arquivos no formato .csv.

Identificar Tabelas: Atribuir nomes (aliases) aos conjuntos de dados carregados para fácil referência.

Análise Estatística Univariada

Medidas de Tendência Central: Calcular a média, mediana e moda de uma coluna numérica.

Medidas de Dispersão: Calcular o desvio padrão, variância, valor mínimo, valor máximo e a amplitude (diferença entre máximo e mínimo).

Medidas de Posição: Determinar os quartis (Q1, Q2, Q3) de uma coluna.

Manipulação de Dados

Contagem: Obter o número total de registros (linhas) em uma tabela.

Filtragem: Criar novos subconjuntos de dados baseados em condições lógicas (ex: idade > 30, curso == "Computação").

A ideia é permitir que o usuário escreva, por exemplo:

```
LOAD "funcionarios.csv" AS func;
LOAD "vendas.csv" AS vendas;


-- Análise estatística dos funcionários
MEAN func.salario AS media_salarial;
MEDIAN func.salario AS mediana_salarial;
```

```
MODE func.departamento AS departamento_mais_comum;
STD func.idade AS desvio_idade;
MIN func.salario AS menor_salario;
MAX func.salario AS maior_salario;
RANGE func.idade AS amplitude_idades;
QUARTILES func.salario AS quartis_salario;

-- Contagem de registros
COUNT func AS total_funcionarios;
COUNT vendas AS total_vendas;

-- Filtragem de dados
FILTER func WHERE idade > 30 AS funcionarios_seniores;
FILTER func WHERE departamento == "TI" AS func_ti;
FILTER vendas WHERE valor > 1000 AS vendas_grandes;

-- Análise nos dados filtrados
MEAN funcionarios_seniores.salario AS media_seniores;
COUNT funcionarios_seniores AS total_seniores;

-- Visualização básica
SHOW func LIMIT 10;
SHOW STATS func.salario;
SHOW STATS func.idade;

SAVE funcionarios_seniores AS "seniores.csv";
SAVE func_ti AS "ti_funcionarios.csv";
```