FACULDADE DE TECNOLOGIA DE SÃO JOSÉ DOS CAMPOS

FATEC PROFESSOR JESSEN VIDAL

Brendo Bubela, Bruna Gomes, Christian Dantas,

Davi Ramos, Jennifer Dominique, João Pedro,

Luara Goulart, Marcos Paulo, Mariana Araujo

DATA UNDERSTANDING

INTELIGÊNCIA ARTIFICIAL

São José dos Campos

2021

Sumário

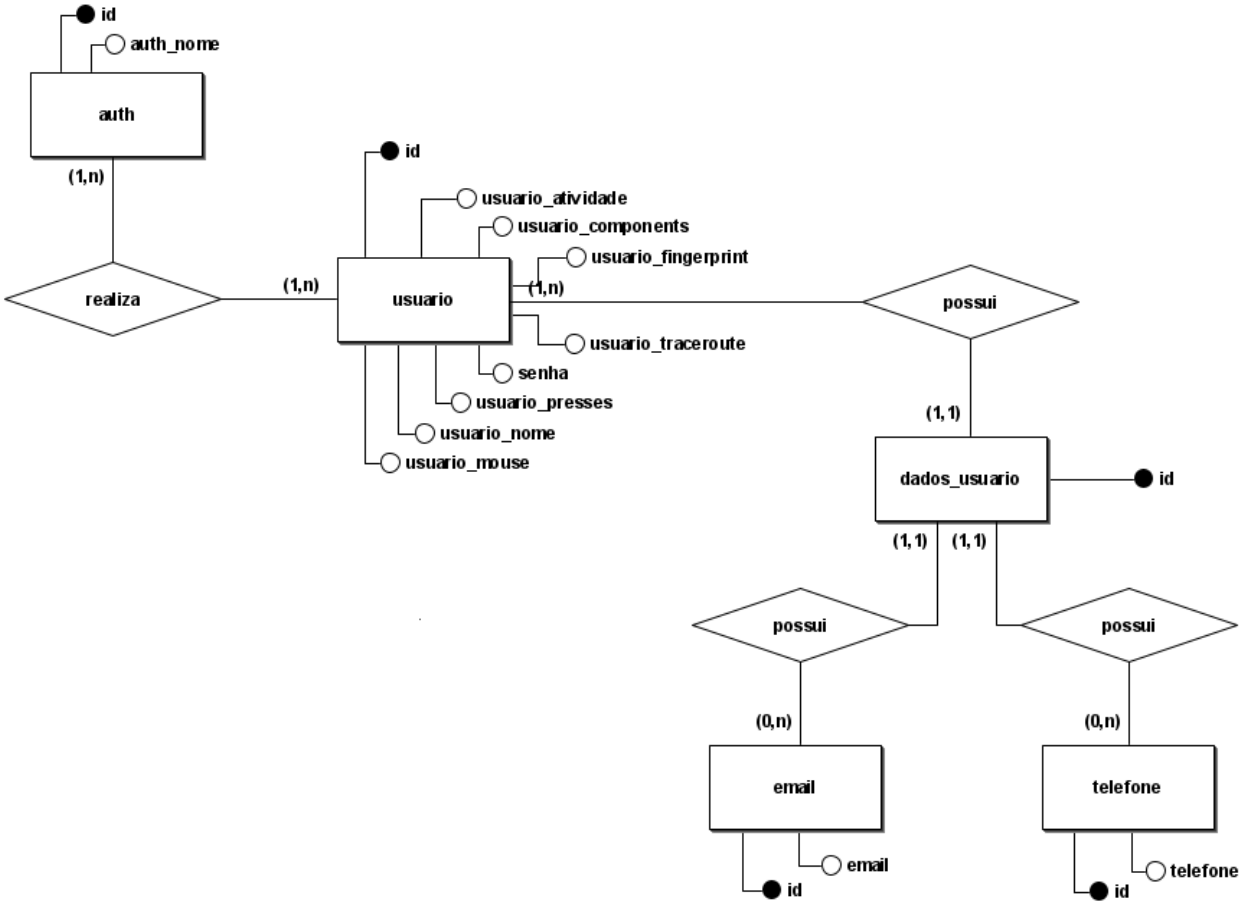
INTRODUÇÃO

A UOL é uma empresa brasileira fornecedora de conteúdo, produtos e serviços da Internet, dentre esses serviços tem o BOL - Brasil Online que é um portal de internet, serviço de webmail. O BOL é uma ferramenta muito utilizada, porém ela possui um obstáculo que a impede de funcionar com 100% da sua capacidade, o problema são os spammers.

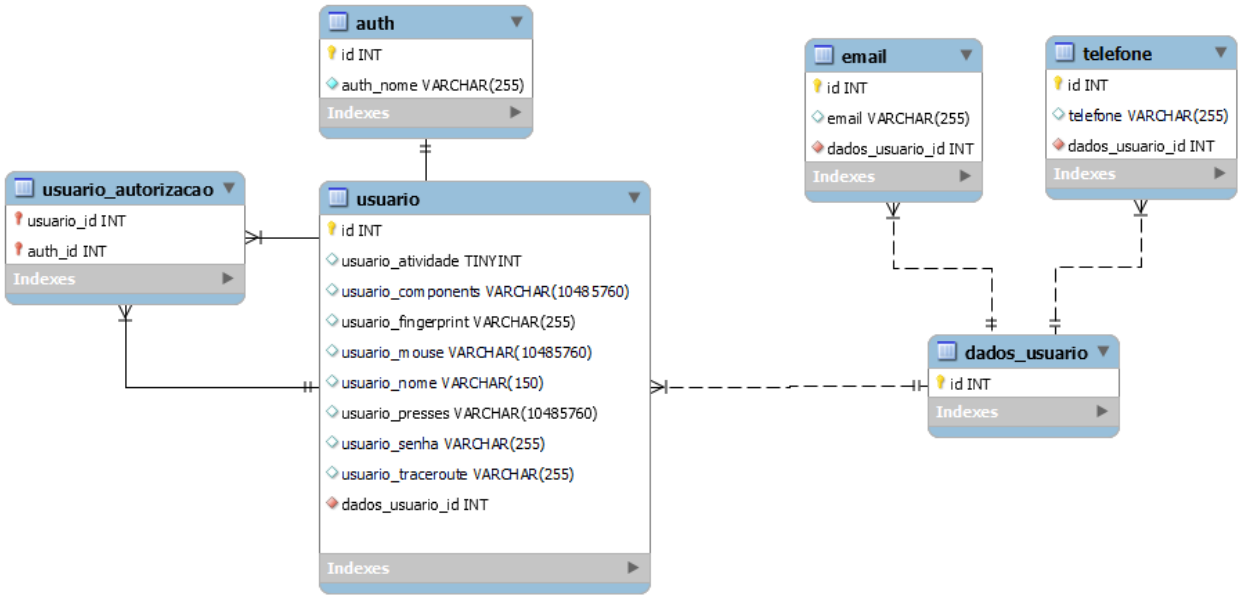
MODELOS DO BANCO DE DADOS

Antes de começar a coletar dados, foi feito o modelo conceitual e lógico do banco de dados.

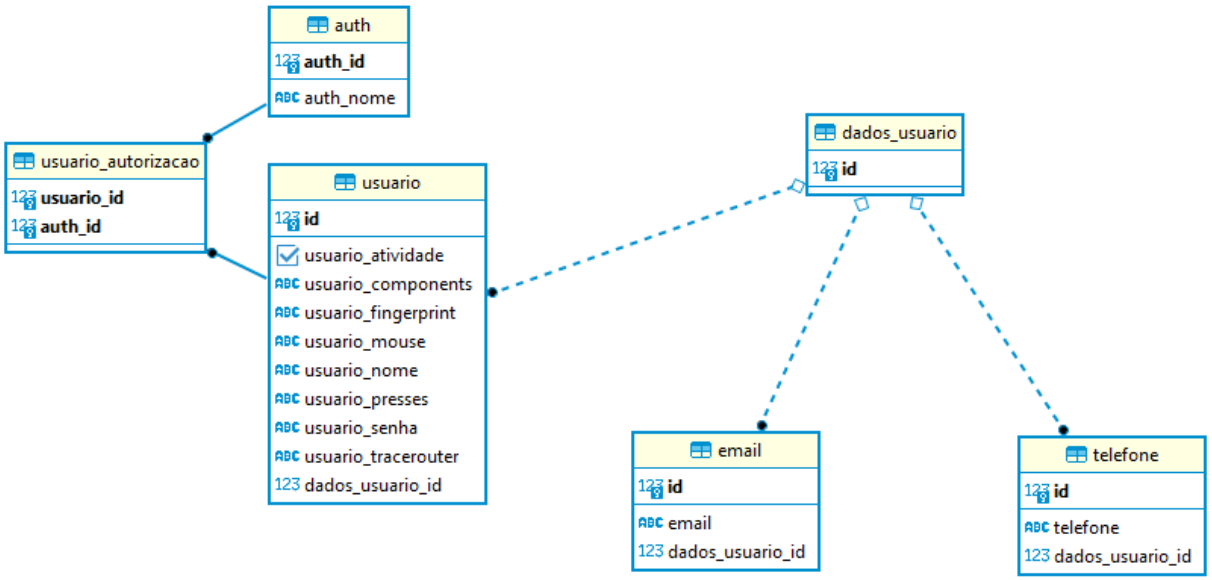
O Modelo Conceitual é uma descrição mais abstrata da realidade, onde os fatos do mundo real são descritos de uma forma mais natural, bem como suas propriedades e relacionamentos. O Modelo Conceitual deve ser sempre a primeira etapa de um projeto de um Banco de Dados.



O Modelo Lógico tem por objetivo representar as estruturas que irão armazenar os dados dentro de um Banco de Dados, a partir deste momento é que são definidas com maior propriedade as entidades e os seus atributos. O Modelo Lógico é iniciado somente a partir da estruturação do Modelo Conceitual.



Depois de criados ambos os modelos, foi iniciado o código do banco de dados. O banco de dados usado foi o Postgres e através do gerenciador DBeaver foi possível visualizar o ER Diagram.



DADOS COLETADOS

Para treinar a Inteligência artificial, foram coletados apenas dados comportamentais do usuário do teclado e mouse no cadastro na plataforma BOL. O Device Fingerprint e o Traceroute foram obtidos para comparação de dados. Os dados coletados são:

- **Device fingerprinting:**

{"visitorId":"cdcdea882a83da64003c0d50d8625a10",
"timezone":{"value":"America/Sao_Paulo","duration":0}

- **HTTP headers;**
- **Resolução do monitor;**
- **Tempo de atividade do dispositivo;**
- **Nível de bateria;**
- **Mime types;**
- **Informação do Sistema Operacional;**
- **Versão do navegador;**
- **Extensões do navegador;**
- **Propriedades do hardware;**
- **Fontes do computador;**
- **Canvas;**
- **Webgl;**
- **Hardware benchmarking;**
- **Geolocalização;**
- **ISP; e**
- **Endereço de IP local.**
- **Dados do teclado:**

{"current":["CapsLock","M","CapsLock","a","r","i","a","n","a","1","2","9","8","1","0","0","6","7","7","2","m","a","r","i","a","n","a","Shift","@","u","s","e","r",".","c","o","m","ArrowLeft","ArrowLeft","ArrowLeft","ArrowLeft","Backspace","Backspace","Backspace","Backspace","ArrowLeft","a","d","m","i","n","ArrowRight","a","d","m","i","n","1","2","3","4"]}

- **Dados do mouse:**

{"current":[{"x":600,"y":280,"click":false},{"x":600,"y":280,"click":true},{"x":602,"y":280,"click":false},...,{"x":604,"y":280,"click":false}]}

- **Traceroute:**

1 1 ms 1 ms 1 ms 192.168 .15 .1 & 2 * * * null & 3 24 ms 7 ms 7 ms 152 - 255 - 151 - 108. user.vivozap.com.br[152.255 .151 .108] & 4 56 ms 9 ms 7 ms 187 - 100 - 193 - 201. dsl.telesp.net.br[187.100 .193 .201] & 5 * * * null & 6 10 ms 14 ms 21 ms 187.110 .147 .152.nipfiber.com.br[187.110 .147 .152]

VARIÁVEIS

As Features, atributos do projeto são:

- **Device Fingerprint:**
  - comparação dos dados dos dispositivos dos usuários para eliminar suspeitas. O “visitorId” é o id do Fingerprint e o “timezone” é o local onde o dispositivo estava na hora do cadastro. São do tipo VARCHAR.

usuario_componentes: “visitorId”, “timezone”

- **Movimentos do mouse:**
  - Possui o ponto X e o Y que correspondem as coordenadas do mouse na tela do monitor, também tem o click, para saber se o usuário apertou o botão do mouse ou não. Tem o objetivo de encontrar semelhanças de rotas entre as suspeitas, resultando em porcentagem de similaridade. São do tipo VARCHAR.

usuario_mouse: “x”, “y”, “click”

- **Teclas pressionadas na digitação do usuário:**
  - Possui todas as teclas do teclado. Verifica a similaridade ou padronização entre o cadastro suspeito. São do tipo VARCHAR.

usuario_presses: “current”

- **Traceroute:**
  - Similaridade entre rotas de IP. Possui IP e domínio. São do tipo VARCHAR.

usuario_traceroute: “rota”

- **Variável da IA:**
  - É a variável que guardará o resultado das buscas do usuário único.

usuario_identificado: “score”, “classe_do_suspeito”

VARIÁVEL OBJETIVO

A variável escolhida para ser a target é o **Score**, possui quatro classes que são:

- Baixo - se o valor for menor que 6;
- Atenção - se o valor for igual a 6 ou 7;
- Alto - se o valor for 8 ou 9;
- Muito alto - se o valor for 10.

ANÁLISE DE RISCOS

Como parte do processo de levantamento de informações, foram analisados os possíveis riscos dentro do contexto da aplicação solicitada pelo cliente.

Dentro desta análise foram separados os riscos e categorizados pelo nível de impacto dentro do processo de criação da Inteligência Artificial.

Baixo Impacto

Riscos de baixo nível são aqueles que impactam o processo de construção de uma inteligência artificial, porém, não interrompe o processo de desenvolvimento, e pode ser rediscutido mais tarde dentro das implementações do projeto. São eles:

- Dados insuficientes para análise

A falta de dados pode acarretar em uma análise de baixo percentual de confiabilidade pela IA, pois não são fornecidos para a ferramenta, os dados de referência necessários para a tomada de decisão dentro das análises realizadas.

- Falta de padronização da "fonte" de dados utilizada nas análises

Quando se utiliza mais de uma base de dados, sem uma padronização ou estrutura que siga uma ordem da origem dos dados, pode gerar conflitos de informações, acarretando em problemas para os desenvolvedores responsáveis pelo suporte e criação da Inteligência artificial.

- **Utilização da IA para análises simples**

Algumas tarefas possuem um nível de complexidade baixo, não havendo a necessidade da utilização da Inteligência artificial para o processo, podendo alternar para um processo de automação.

Impacto Parcial

Riscos de impacto parcial, são aqueles que impactam o processo de construção de uma inteligência artificial, e podem gerar problemas nos processos de desenvolvimento, exigindo que o time redesenhe a proposta da implementação. São eles:

- **Não definir a regra de negócio, o objetivo principal da utilização da IA**

Não ter de forma clara o objetivo principal da utilização da IA, faz com que o time de desenvolvimento perca tempo desenvolvendo uma proposta mal planejada

- **Título**

Descrição

Alto Impacto

Riscos de alto nível, são aqueles que impactam todo processo de construção de uma inteligência artificial, desde o início da sua implementação, pois pode bloquear o time de desenvolvimento caso o mesmo não possua um objetivo claro para a utilização da ferramenta.

-