UNIVERSIDADE FEDERAL DE MINAS GERAIS
ICEx - DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
DISCIPLINA: ARMAZÉNS DE DADOS
Júnio Veras de Jesus Lima - junio.veras@dcc.ufmg.br
Vinícius Braga Freire - vinicius.braga@dcc.ufmg.br
João Pedro Macedo - joao.oliveira@dcc.ufmg.br
João Pedro Reis - jprtm@ufmg.br

Criação de Data Warehouse para análise do uso do Stack Overflow em 2021

1. Introdução

O Stack Overflow é uma plataforma gratuita de perguntas e respostas que aborda majoritariamente questões relacionadas com tecnologia e desenvolvimento, e é fortemente utilizada por estudantes e programadores.

Todo ano é feita uma pesquisa com alguns usuários da plataforma para conseguir dados e informações sobre eles. Para fazer a análise destes dados e obter informações relevantes, modelamos estes dados em um formato que propicia uma análise melhor e mais completa. Assim, criamos o data warehouse dos dados dos usuários que utilizam o Stack Overflow.

2. Base de Dados

A base de dados utilizada no projeto é o resultado da pesquisa anual do Stack Overflow em 2021. Essa pesquisa foi feita com 83439 pessoas de 181 países diferentes. Essa base de dados apresenta informações de diferentes âmbitos sobre os indivíduos, estando entre essas:

Dados pessoais: País, gênero, idade, sexualidade, …

Dados do trabalho: Salário, qual a ocupação, o tamanho da empresa, qual a moeda em que recebe o salário, …

Dados do indivíduo enquanto programador: Quais as linguagens de programação que utiliza, qual o sistema operacional principal, há quantos anos programa, há quantos anos programa profissionalmente, como o indivíduo lida quando está preso em algum problema, …

Dados do indivíduo enquanto usuário do Stack Overflow: Quais sites do stack overflow a pessoa utiliza, com qual frequência a pessoa acessa o Stack Overflow, em quais comunidades participa, …

Não foi encontrada nenhuma dificuldade para acessar essa base de dados, afinal a empresa disponibilizou os dados de forma unificada e em um padrão (.CSV) bastante utilizado.

Em contraposição, houveram diversos problemas na extração dos dados dessa base. Muitos desses problemas se devem ao fato que muitas perguntas do questionário permitem respostas abertas, principalmente quando se tratam de dados pessoais. Isso dificultou muito o tratamento dos dados que será abordado em uma seção posterior.

A base de dados está disponível em https://insights.stackoverflow.com/survey.

3. Dimensionamento

A partir do processo de design de 4 passos, a base de dados foi reestruturada para ser um modelo estrela, pois essa seria a melhor forma de utilizar as ferramentas do Pentaho.

- **Primeiro passo (Selecionar o processo de negócios a modelar):**

  A base de dados escolhida corresponde a respostas de questionários sobre o uso do Stack Overflow, então o processo a se modelar é o uso dessa plataforma.

- **Segundo passo (Declarar a granularidade do processo):**

  A granularidade desse processo são todos os tipos de dados de um indivíduo que usa a plataforma e os motivos de uso, que é a informação mais atômica encontrada na base de dados.

- **Terceiro passo (Escolher as dimensões que se aplicam a cada linha da tabela de fatos):**

  Para essa modelagem foram utilizadas as dimensões Pessoa (contendo dados pessoais), Programador (contendo dados sobre o perfil de desenvolvedor), Trabalho (contendo dados sobre o emprego do indivíduo) e Stack_Overflow (contendo dados sobre o uso da plataforma).

- **Quarto passo (Identificar os fatos numéricos que irão popular a tabela de fatos):**

  Como os dados da base foram divididos em cada uma das dimensões, e não haviam dados que pudessem ser fatos numéricos relevantes, optou-se pela realização de uma tabela de fatos sem fatos.

O esquema estrela pode ser visualizado a seguir:

Figura 1 - Modelo Estrela

Dimensão Pessoa:

- Country - país de origem
- US_State - estado de origem dos EUA
- UK_Country - país de origem do UK
- EdLevel - nível de educação
- Age - idade
- Gender - gênero
- Trans - identificação de transgênero
- Sexuality - sexualidade
- Ethnicity - etnia
- Accessibility - deficiências
- MentalHealth - saúde mental

Dimensão Programador:

- Age1stCode - faixa de idade que fez o primeiro código
- LearnCode - como aprendeu a programar
- YearsCode - tempo desenvolvendo códigos
- YearsCodePro - tempo desenvolvendo códigos profissionalmente
- DevType - tipo de desenvolvedor
- OpSys - sistema operacional mais utilizado
- Id_language - chave estrangeira para as linguagens utilizadas

Dimensão Trabalho:

- MainBranch - ocupação principal
- Employment - situação de emprego
- OrgSize - tamanho da organização que trabalha
- Currency - tipo de moeda que ganha
- CompTotal - renda
- CompFreq - período da renda

Dimensão Stack_Overflow:

- VisitFreq - frequência de visitação ao site
- Account - sinalização de posse de conta no site
- PartFreq - frequência de participação no site
- Comm - identificação de se considerar membro da comunidade do site
- OtherComms - identificação de participação de outras comunidades de outros sites
- Id_stuck - métodos para lidar com "obstáculos" durante o desenvolvimento

Um detalhe importante é que para as variáveis stuck e languages foram utilizadas *junk dimensions* contendo todos os agrupamentos possíveis desses dados. Foi criada uma tabela *languages* contendo todas as combinações das 12 linguagens de programação da base de dados, totalizando 4096 linhas, e uma tabela *stuck* contendo todos os 9 métodos para lidar com os "obstáculos" ao programar, totalizando 512 linhas. Com essa abordagem evita-se que as dimensões contenham muitas linhas, sendo apenas feito um mapeamento de id para essas tabelas.

4. Arquitetura do Data Warehouse

Planejar a arquitetura do Data Warehouse é essencial para o bom funcionamento futuro. O Data Warehouse tem sua arquitetura em 4 etapas:

- Source System: a origem dos dados a serem utilizados no DW. No nosso caso o source system será a base de dados fornecida pelo Stack Overflow.
- Data Staging Area: seção onde os dados serão processados. Aqui é onde ocorrerão os tratamentos dos dados vindos da etapa anterior e é onde ocorrem os processos de ETL. Nessa etapa utilizaremos o PDI da Pentaho como ferramenta principal.
- The Data Warehouse: esta etapa é o data warehouse em si. Aqui serão salvos os dados, já tratados, em um banco de dados relacional que será o PostgreSQL. O banco de dados será modelado de acordo com a modelagem dimensional apresentada na seção 3.

  Devido a essa modelagem, utilizaremos apenas 1 data mart e por isso a tabela de BUS será como a apresentada a seguir:

	Pessoa	Programador	Trabalho	Stack_Overflow
Uso do Stack Overflow	X	X	X	X

Tabela 1 - Tabela de BUS do Data Warehouse

  Ainda nessa etapa é onde está a criação dos cubos. Para tal, utilizamos a ferramenta Schema Workbench da Pentaho.
- End User Data Access: Nessa etapa é onde as análises são retiradas sobre os dados. Aqui será utilizado o Mondrian e a ferramenta saiku do Pentaho.

5. ETL

Como dita a arquitetura do Data Warehouse implementada, é necessário tratar (extrair, transformar e carregar) os dados do sistema fonte através de um processo de ETL. Esta etapa é de suma importância, pois a partir dos dados externos ao processo podemos moldá-los de acordo com as necessidades e o dimensionamento do Data Warehouse, sempre mantendo a coerência e a fidelidade com os dados originais e os eventos que esses representam.

Para executar as tarefas de ETL foi utilizado a ferramenta KETTLE do grupo Pentaho.

5.1 Extração

A extração de dados foi feita a partir da base de dados apresentada anteriormente na seção 2. O Stack Overflow disponibilizou a base de dados em formato CSV e, portanto,

esse será o formato utilizado na extração de dados. O KETTLE suporta a extração de dados nesse formato, já que é um sistema em modelo de linhas e colunas.

5.2 Transformação e Limpeza

Para o tratamento dos dados vamos usar o arquivo extraído na seção anterior. Para deixar o processo mais rápido e organizado, dividimos o processo de tratamento em 5 seções:

Dimensão Programador:

A figura a seguir apresenta o diagrama do PDI utilizado para o processamento dos dados. A figura está dividida em 2 blocos: o de cima é o tratamento dos dados, enquanto o bloco de baixo cria a tabela de agregados de *languages*.

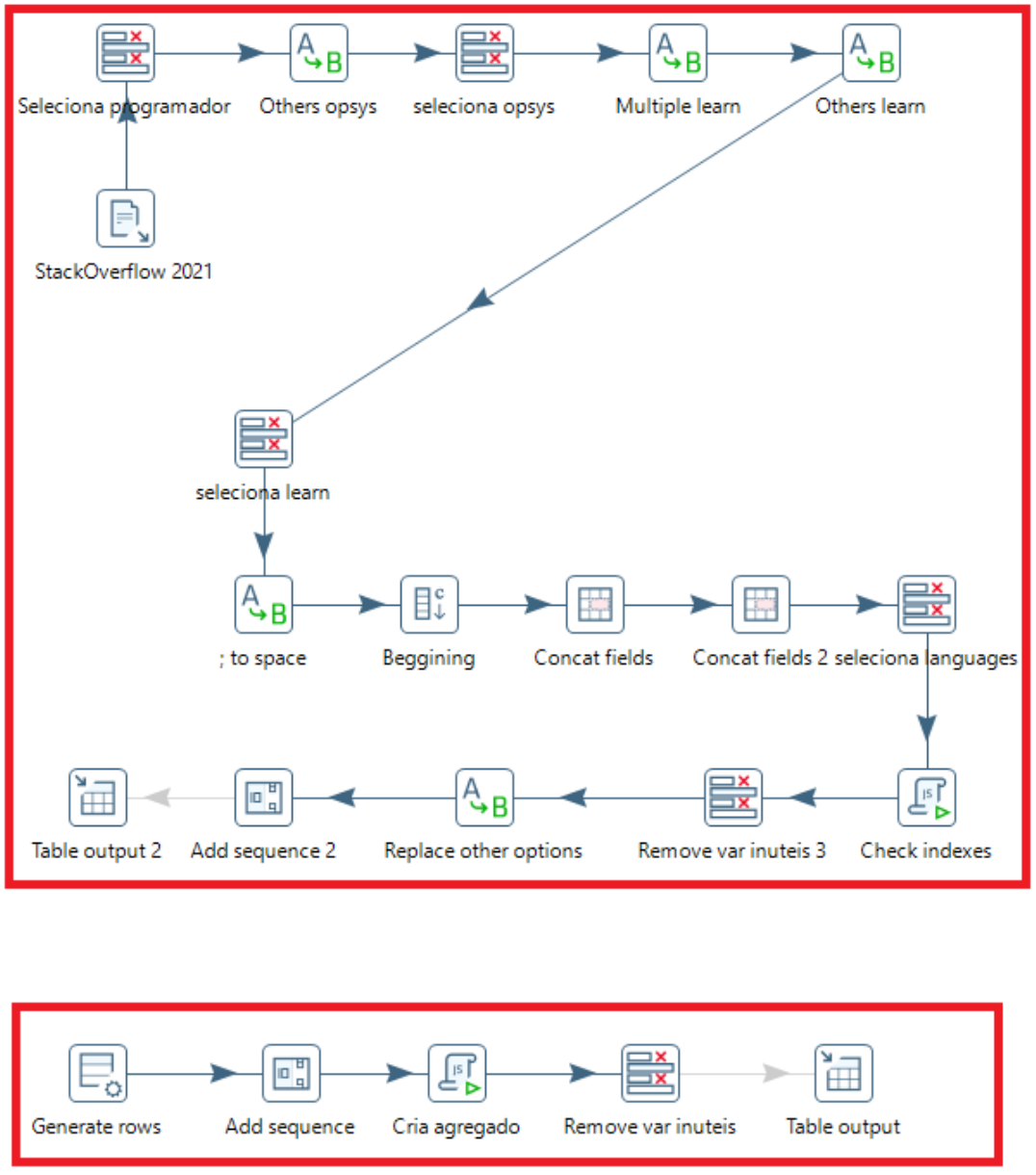

Figura 2 - Tratamento dos dados de programação

Os principais tratamentos executados nessa dimensão foram:

- Exclusões de respostas nulas (a pesquisa do Stack Overflow permitia a inserção de respostas abertas sobre o título de 'Outras opções' e muitos usuários escolhiam este campo mas o deixavam vazio). Essas respostas foram transformadas em 'NA' ou não aplicável. Esse caso ocorreu para todas as dimensões.
- Transformar uma string onde todas as linguagens de programação usadas pelo usuário estavam concatenadas e apenas separadas por ';'. Após separá-las, foi necessário criar um link (uma chave estrangeira) para o agregado de linguagens criado no bloco inferior da tela. Como dito na seção 3, esse agregado possui 4096 linhas.

Esta dimensão oferece muitos desafios para o tratamento de dados, pois a criação do agregado não é um processo trivial, onde foi necessário criar funções (em JavaScript) para tratamento de bits e assim criar uma "tabela verdade das linguagens".

Dimensão Trabalho:

A figura a seguir apresenta um diagrama do tratamento dos dados sobre o trabalho. Diferente da dimensão anterior, não foi necessário muito tratamento nesta dimensão.

A parcela mais difícil desta parte foi transformar os símbolos das moedas de cada um dos países dos quais os indivíduos recebiam para o padrão internacional (3 letras e em caixa alta, como BRL).

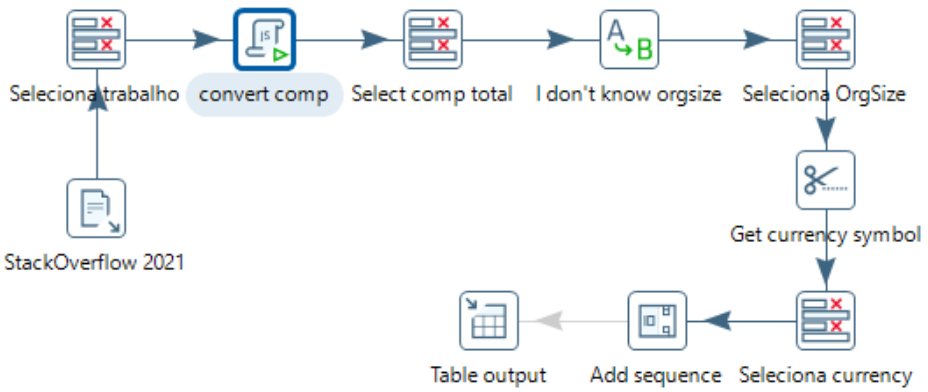

Figura 3 - Tratamento dos dados de trabalho

Dimensão Pessoa:

A dimensão pessoa foi tratada através do processo da imagem a seguir. Esta dimensão foi a mais desafiadora para se tratar, afinal ela possui muitos campos e esses campos refletem a intimidade do indivíduo, o que obrigou a pesquisa a ter múltiplas respostas.

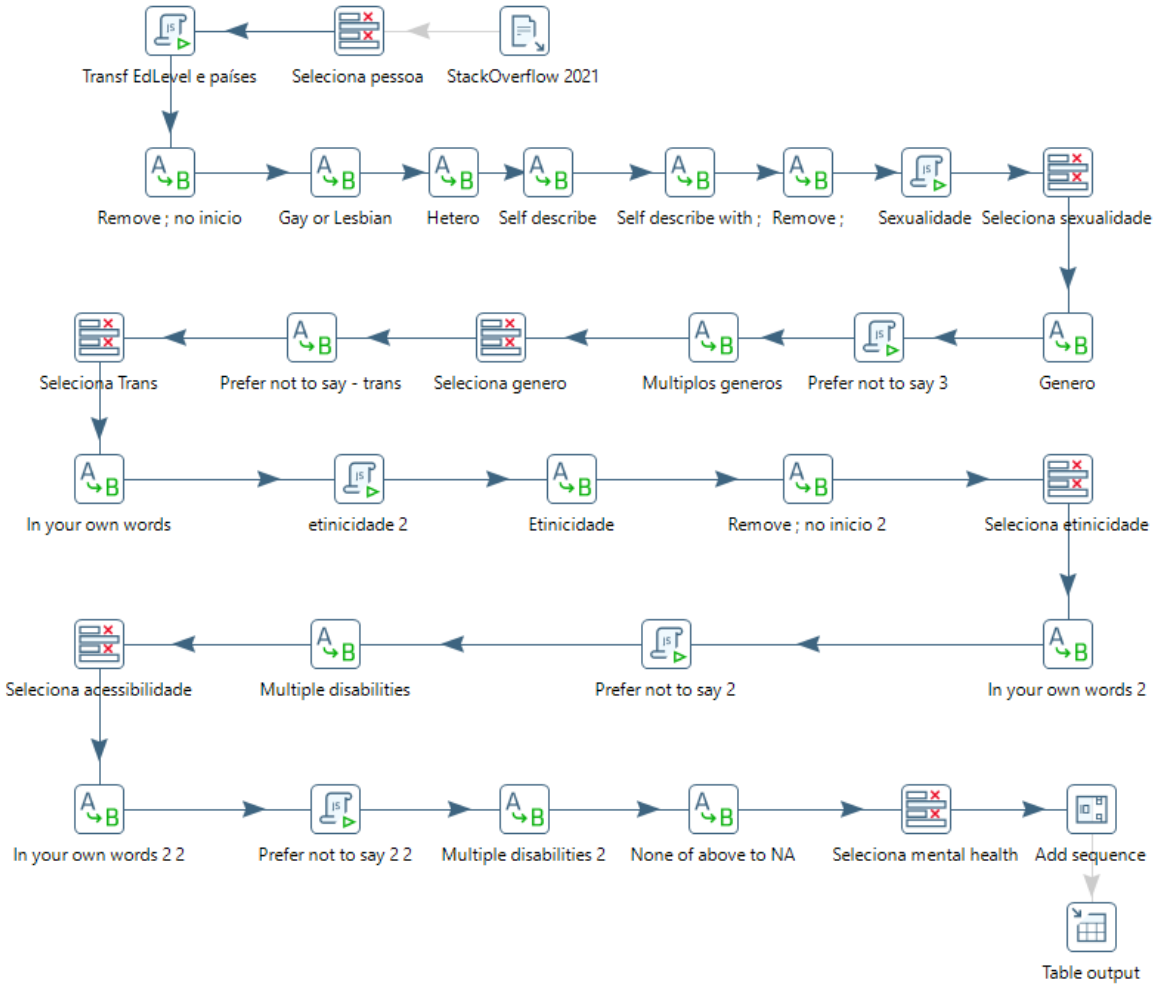

Figura 4 - Tratamento dos dados de pessoa

Entre as operações executadas, as principais foram:

- Tratar a sexualidade das pessoas. Para tal, foi feito uma mudança nas nomenclaturas utilizadas (e.g. 'Gay or Lesbian' passaram a ser 'Homosexual'). Além disso, foi necessário tratar as pessoas que escolheram múltiplas sexualidades (o que não faz muito sentido, pois se nenhuma das sexualidades apresentadas no formulário representam a pessoa bastava escolher que nenhuma opção se aplicava) e para esses casos transformou-se tais sexualidades em 'Prefer not to say'.

- Como na dimensão programador, foi necessário transformar os dados vazios em ‘Prefer not to say’.
- Para a etnia das pessoas, foi necessário criar uma classe 'Múltiplas etnias' para os casos onde as pessoas possuíam mais de 1 etnia.

O tratamento dessa dimensão foi bastante difícil, tanto pelo tamanho como pela complexidade das decisões que foram tomadas. Decidir se iríamos ocultar (e.g. ‘Prefer not to say’) as múltiplas sexualidades, decidir em criar uma nova etnia, múltiplas deficiências, entre outros, tomou uma considerável parcela do tempo de planejamento do nosso projeto.

Dimensão Stack Overflow:

Nesta dimensão foram executadas as transformações apresentadas na figura a seguir.

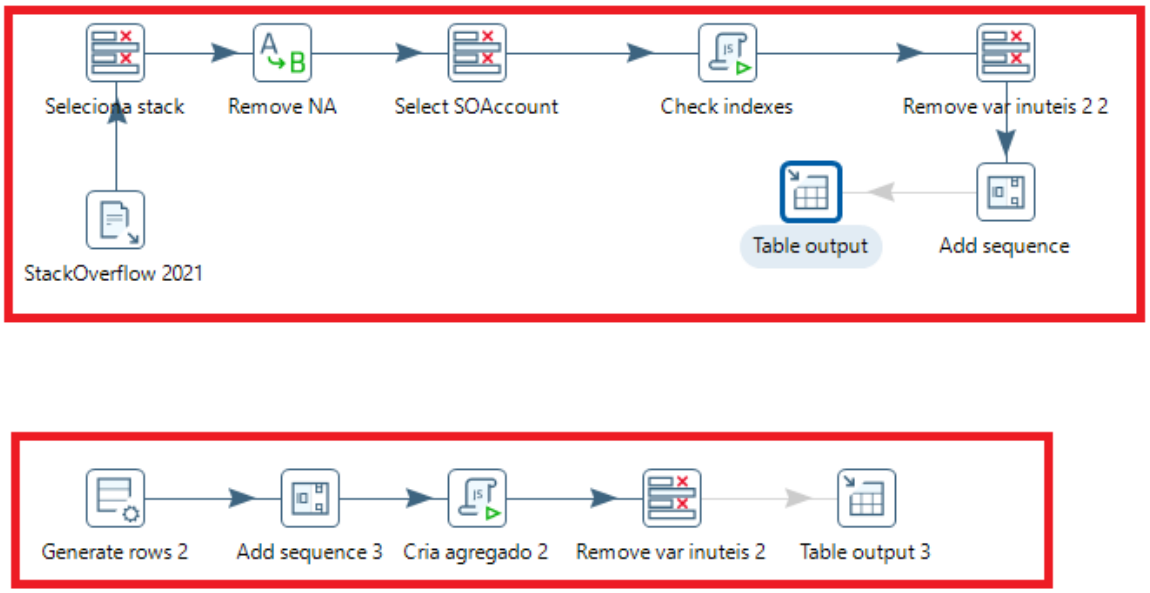

Figura 5 - Tratamento dos dados do Stack Overflow

Da mesma forma como a dimensão programador, foi necessário criar um agregado (bloco de baixo) de stuck que irá armazenar as maneiras como as pessoas reagem quando ficam presas em um certo problema. O restante das outras transformações foram apenas mudanças pequenas.

Tabela Fato:

Na tabela de fato a única transformação necessária foi juntar os ID 's de cada uma das outras dimensões e exportar essa tabela resultado para o banco de dados. A transformação é mostrada na figura a seguir.

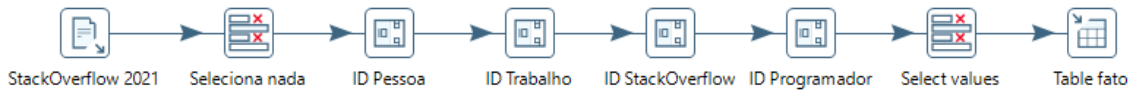

Figura 6 - Tratamento dos dados da tabela de fatos

5.3 Carga de Dados

Após o tratamento do dado através dos processos explicados anteriormente, foi necessário fazer a carga desses dados remodelados no banco de dados utilizado. O banco de dados escolhido foi o PostgreSQL. Como mostra a figura 6 o KETTLE disponibiliza uma bloco para injeção de dados direto em um Data Base. Porém para fazer isso, foi necessário, primeiramente, criar as tabelas no banco de dados seguindo o modelo dimensional da seção 3.

6. Banco de Dados

Como apresentado anteriormente, o banco de dados utilizado para implementação do Data Warehouse foi o PostgreSQL. Foi decidido usar esse SGBD pois este é open source e possui uma interface (pgAdmin) bastante intuitiva.

A criação das tabelas do banco de dados foram feitas em função do modelo dimensional apresentado na seção 3, enquanto a população dessas tabelas foram feitas pelas transformações do KETTLE apresentadas na seção 5.

Assim, para integrar as ferramentas do Pentaho com os dados foi necessário fazer uma conexão entre tais ferramentas e o PostgreSQL. Como esse SGBD é bastante utilizado, há o pleno suporte deste dentro das ferramentas do Pentaho.

7. Criação de Cubos

Como especificado na seção 4 a ferramenta utilizada para a criação dos cubos dimensionais foi o Pentaho Schema Workbench. Por contar com uma única tabela de fatos, nosso DW pôde ser condensado em um único cubo dimensional. Quatro dimensões principais foram adicionadas ao cubo, que correspondem às dimensões do modelo discutidas em detalhes anteriormente: Pessoa, Programador, Trabalho e Stack_Overflow.

A única medida adicionada no cubo utiliza um contador como agregador. O campo da tabela de fato utilizado por esta medida é um booleano que representa a participação de um usuário na pesquisa do Stack Overflow, ou seja, este campo sempre apresenta o valor 1, logo, nossa medida é um contador de participantes da pesquisa. Medidas mais complexas

são utilizadas em análises posteriormente, mas elas são resultantes de membros calculados através de consultas MDX e não foi necessário codificá-las no cubo.

Cada um dos dados nas dimensões se tornou um nível dentro de sua própria hierarquia durante a criação do cubo. Apesar de ser possível adicionar múltiplos níveis por hierarquia optamos por mantê-los isolados, isto se deu pelo simples fato de que os dados nas nossas dimensões não possuem uma relação hierárquica entre si.

As sub-dimensões Languages e Stuck são acessadas a partir das dimensões as quais estão conectadas, isto é, Programador e Stack_Overflow, respectivamente. Este acesso é feito de uma forma simples: cada um dos dados nas sub-dimensões é ligado a uma dimensão utilizando uma tag join, assim como no caso dos níveis os joins foram mantidos isolados em suas respectivas hierarquias.

8. Navegação pelas Dimensões

Com o cubo gerado a partir do Schema Workbench, foram feitas navegações pelas dimensões e cruzamentos de dados para identificar características interessantes a respeito da base de dados. A navegação foi feita através do uso da ferramenta Saiku no servidor Pentaho.

8.1. Como as pessoas lidam quando ficam presas em um problema em função da idade

A primeira análise foi o cruzamento entre os dados de idade dos usuários e as maneiras que as pessoas lidam quando ficam presos em um problema.

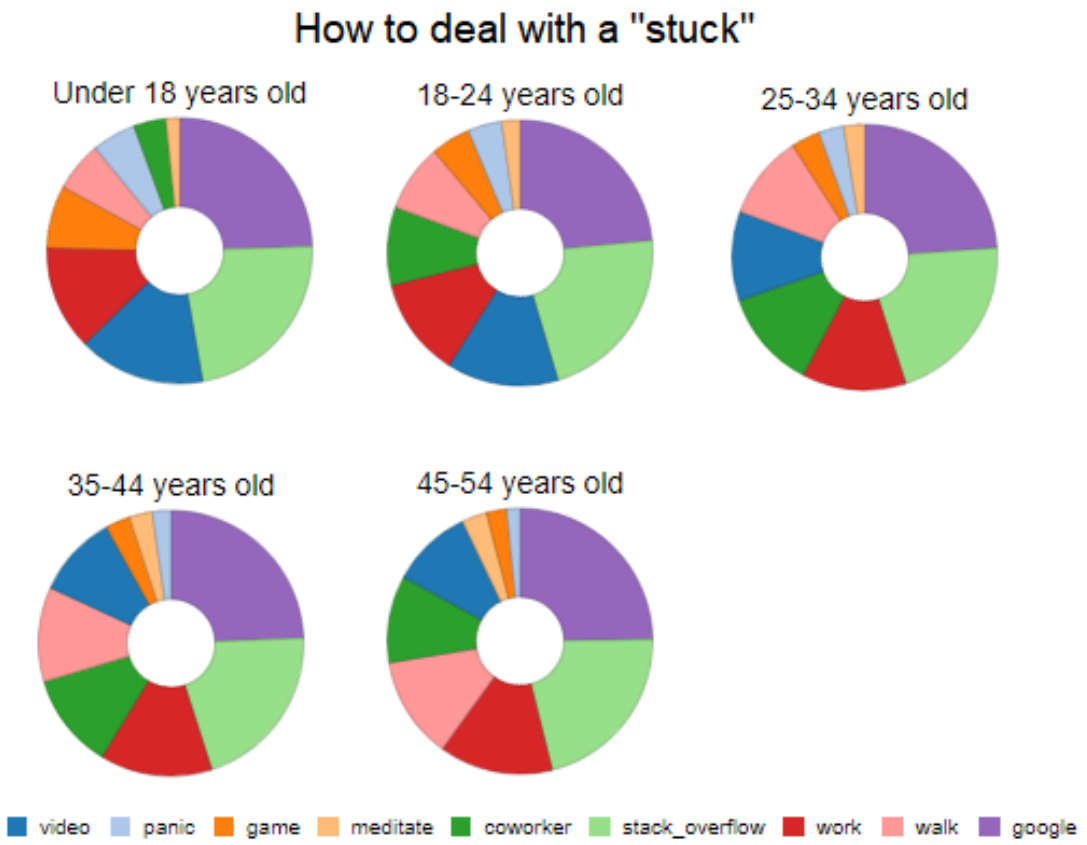

Figura 7 - Gráfico “ How to deal with a ‘stuck’ ”

A figura 7 foi gerada a partir de uma consulta utilizando a linguagem MDX na ferramenta saiku.

Com a figura, é possível observar que algumas formas de lidar com problemas apresentam uma variação conforme a idade.

- A porcentagem de indivíduos que entram em pânico ao se deparar com algum obstáculo é bem maior entre as faixas etárias menores e decresce conforme a idade aumenta.
- Pedir ajuda a um coworker é uma maneira de lidar com problemas que aumenta conforme a idade.
- Jogar algum jogo para desestressar é mais comum entre os jovens.
- Caminhar é mais utilizado entre os grupos mais velhos.

Um dado interessante é que o uso do stack overflow como ferramenta para buscar como resolver um problema é algo recorrente em todas as faixas etárias.

8.2. Principais tecnologias utilizadas por funcionários de diferentes tamanhos de organizações

Esta análise mostra a relação entre o tamanho das organizações, em número de funcionários, e a utilização de linguagens e tecnologias pelos funcionários destas organizações.

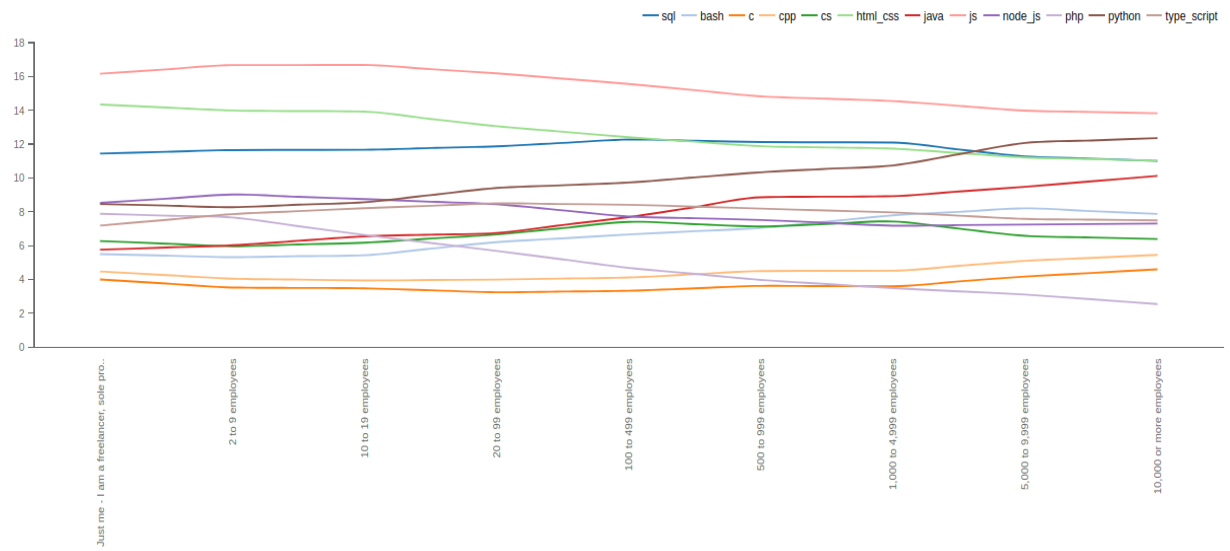

Figura 8 - Gráfico “ Tamanho da Organização X Popularidade das Tecnologias ”

A partir da imagem é possível observar algumas tendências, em especial:

- Java não é uma escolha popular para desenvolvedores solo, mas é uma das mais utilizadas em grandes corporações
- JavaScript é a linguagem mais utilizada independente do tamanho da organização

- Algumas tecnologias comumente associadas, como JavaScript e HTML/CSS possuem uma curva semelhante
- PHP é a linguagem que sofre a queda mais acentuada de utilização conforme o tamanho da organização aumenta

8.3. Países que mais contribuíram para a pesquisa

Esta análise mostra a distribuição de países para os participantes da pesquisa.

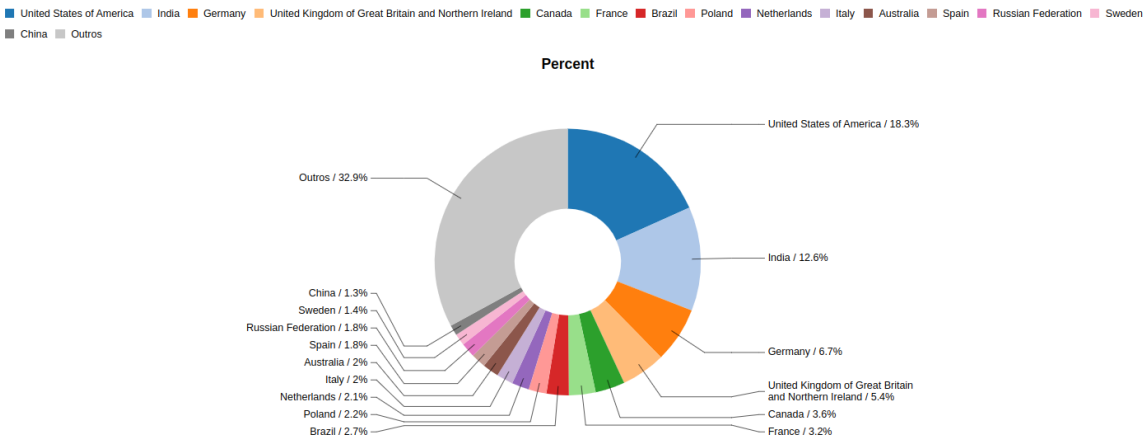

Figura 9 - Gráfico “ Participantes por país ”

A partir da imagem e de dados que foram apresentados anteriormente é possível realizar algumas inferências:

- Na pesquisa houveram usuários de 181 países diferentes, porém, quase 70% desses usuários vieram dos top 15 países mostrados no gráfico.
- Estados Unidos e Índia são os dois países com mais desenvolvedores, abrangendo cerca de 30% do total.
- O Brasil é o sétimo país com mais desenvolvedores, com 2.7% do total.

8.4. Principal ocupação por idade

Nessa análise foi feito o cruzamento entre os dados de idade dos usuários e as ocupações dos usuários.

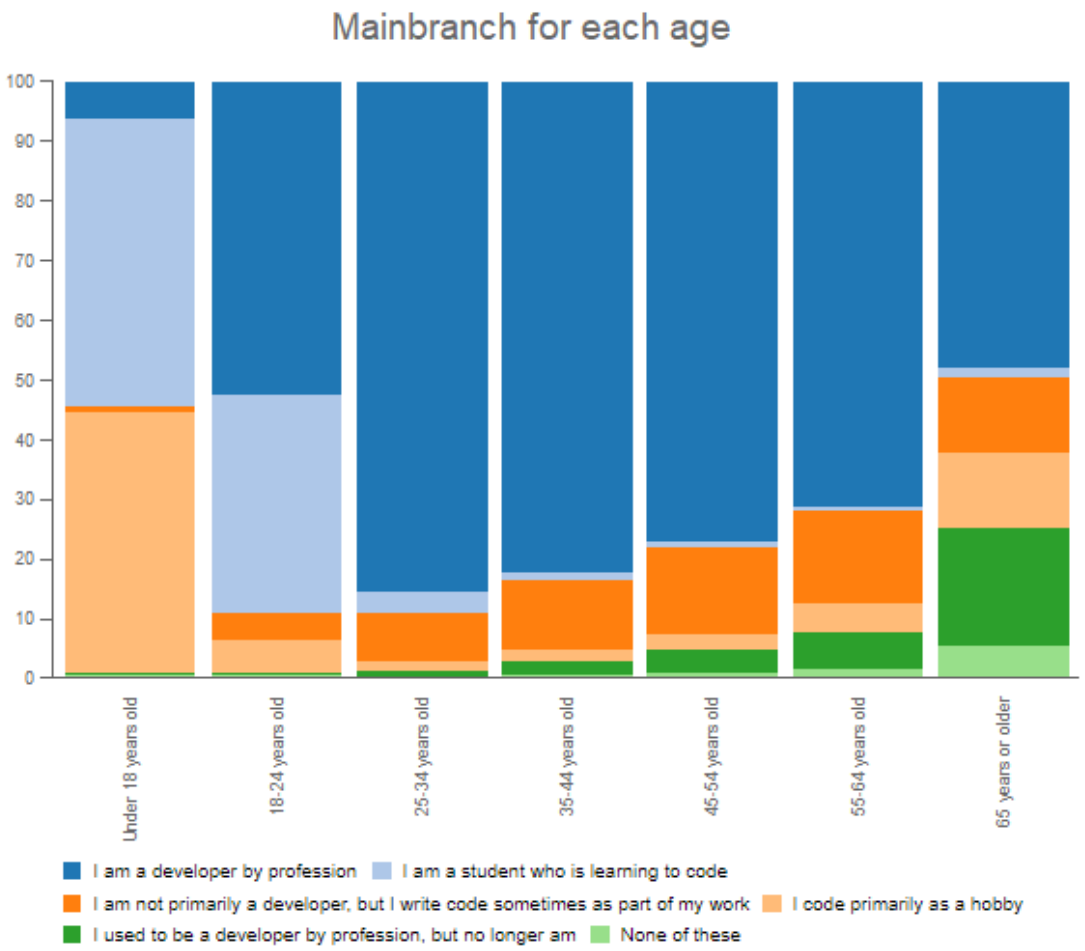

Figura 10 - Gráfico “ Mainbranch for each age ”

A figura 10 foi gerada a partir da navegação pelo cubo com a ferramenta saiku.

Com a figura, é possível observar a distribuição das ocupações em cada faixa etária.

- É notável que as faixas etárias mais jovens (até 24 anos) possuem uma quantidade significativa de estudantes.
- Usuários com idade menor que 18 anos desenvolvem código para estudo ou hobby, não existindo muitos profissionais na área.
- Conforme a idade aumenta, as pessoas que saem da área de desenvolvimento aumentam, podendo ser por questões de criação de novas tecnologias e obsolescência de outras, e por questões de aposentadoria.

9. Conclusão

Como a tendência atual é que o volume de dados gerados nas mais diversas áreas e para diferentes fins aumente significativamente, é necessária uma maneira eficiente de analisar esses dados, produzir relatórios e extrair informações interessantes a respeito disso. Uma forma de fazer isso é com o uso de armazéns de dados.

Este trabalho foi feito utilizando algumas das ferramentas do Pentaho (PDI, Saiku, Schema Workbench, Mondrian) e utilizando o SGBD postgresql. Os dados utilizados provinham de uma pesquisa realizada em 2021 na plataforma Stack Overflow, que basicamente continha perguntas a respeito do uso da plataforma e algumas questões relacionadas ao perfil do usuário. A partir das ferramentas, foi possível realizar a extração dos dados, realização de cubos, navegação desses cubos e a realização de análises a partir de cruzamento de dados.

As ferramentas utilizadas facilitam muito o processo de análise e manipulação dos dados, porém existem algumas funcionalidades que são pouco intuitivas e necessitam de um entendimento maior, como a utilização da linguagem MDX por exemplo. Além disso, a instalação das ferramentas pode ser bem complicada.

Ao final do processo, foi possível extrair informações interessantes sobre a base de dados, como por exemplo as diferentes formas de lidar com problemas entre os desenvolvedores conforme a idade.

Com isso, foi perceptível a capacidade e o poder de um armazém de dados e de algumas ferramentas Pentaho.