PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTÍFICA - PIBIC PROJETO DE PESQUISA TÍTULO DO PROJETO DE PESQUISA:

Desenvolvimento de um Sistema para Explorar a Capacidade Relacional e Longitudinal da Base de Dados da PME/IBGE

Resumo de atividades:

No primeiro momento foi feito uma pesquisa para se avaliar qual seria o melhor Sistema de Gerenciamento de Banco de Dados (SGBD) para se utilizar no projeto, foram levados em conta vários fatores na escolha do SGBD, entre eles a facilidade de uso, a simplicidade, a segurança, a quantidade de recursos e a robustez. Escolhemos o “PostgreSQL” principalmente por sua robustez, qualidade imprescindível já que trabalhamos com um conjunto grande de dados, e também por ele ser Opensource, possuindo uma grande quantidade de documentação acessível e não gerando nenhum custo adicional ao projeto.

Depois da escolha do SGBD se deu a fase de importar os dados da PME/IBGE para o banco de dados. A PME é uma pesquisa realizada desde 1980 mensalmente em 6 regiões metropolitanas e a cada mês é gerado um arquivo txt com uma média de 100 000 (cem mil) linhas. Devido ao tamanho dos arquivos da PME a importação dos dados não é algo simples. Para auxiliar a importação dos dados foi utilizado o programa Spoon da comunidade Pentaho. O Spoon é um ETL Opensource desenvolvido em Java. ETL é a abreviação de Extract Transform Load (*Extraç*ão Transformação Carga), são ferramentas de software cuja função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e pôr fim a carga dos dados geralmente em um [Data Mart](http://pt.wikipedia.org/wiki/Data_Mart) e um [Data Warehouse](http://pt.wikipedia.org/wiki/Data_Warehouse). O spoon utiliza de uma interface gráfica que facilita muito o trabalho e ele também implementa algoritmos de concorrência que fazem com que a importação dos dados seja mais rápida.

Resolvido a parte de como importar os dados veio uma fase de análise do nosso modelo, por que não basta apenas importar os dados para o banco de dados é importante também que se tenha criado um “modelo relacional”. Não desejamos criar um banco de dados que apenas exibe os dados coletados na pesquisa, queremos definir relações entre os domicílios, pessoas e trabalhos descritos na PME, relações que não são claras na pesquisa original. Essa parte da pesquisa ainda está sendo desenvolvida, pois ainda existem questões para serem resolvidas. Uma das questões que ainda está sendo respondida é sobre o ID de pessoa. Na PME não existe nenhum campo que define unicamente um indivíduo, devido a isso na tentativa de se criar esse ID nós unimos mais de uma variável da PME em um único campo no banco de dados. Mas quais variáveis utilizar para criar esse campo? Utilizando como base o texto de Rafael Perez Ribas e Sergei Suarez Dillon Soares que analisa a PME/IBGE nós resolvemos utilizar como base para o nosso Identificador de pessoa as variáveis v203(sexo), v208(raça), v201(número de ordem), v204(dia de nascimento), v214(mês de nascimento), v224 (ano de nascimento). Utilizando essa nova variável fomos capazes de identificar com sucesso um bom número de indivíduos dentro da PME.

Além do problema do nosso modelo relacional outra atividade que está sendo desenvolvida é a definição da forma que trataremos os dados do banco de dados. Uma das formas propostas é utilizar apenas o SQL “puro” para se analisar os dados. O SQL é a linguagem de pesquisa padrão para banco de dados relacionais e o Postgres já dá suporte para ela, facilitando um pouco as coisas, pois não precisaremos instalar nenhuma nova ferramenta para analisar os nossos dados. Uma outra alternativa é utilizar o SQL junto com a linguagem R. R é uma [linguagem](http://pt.wikipedia.org/wiki/Linguagem_de_programa%C3%A7%C3%A3o) de programação para cálculos [estatísticos](http://pt.wikipedia.org/wiki/Estat%C3%ADstica) e [gráficos](http://pt.wikipedia.org/wiki/Gr%C3%A1fico) e que é muito usado atualmente no Laboratório de Econometria da UFC, ajudando a integração do nosso projeto com outros projetos da UFC. Uma terceira alternativa é usar SQL junto com alguma linguagem de alto nível como Python ou Java. Essa abordagem é interessante pois possibilitaria no futuro o desenvolvimento de programas mais complexos, com interfaces gráficas, páginas web que permitiriam exibir o conteúdo de uma forma “mais atraente” e mais simples para pessoas que não fazem parte do meio acadêmico.

A continuidade da pesquisa está se dando no teste de modelos relacionais simples na busca de um modelo mais complexo e definitivo e na extração de informações do banco de dados, as vezes utilizando o SQL puro e as vezes utilizando SQL em conjunto com python. Também estou estudando R e procurando maneiras de integra-lo com SQL. Existem bibliotecas como a **RODBC, a RMySQL, ROracle** e a [RJDBC](http://cran.r-project.org/web/packages/RJDBC/index.html) que permitem utilizar scripts SQL dentro do R.