XVI EAMC - Serviços de Atualização no Gateway BioinfoPortal: Suporte ao Bancos de dados de Proveniência

Marco Cabral[1,2], Antônio Tadeu Azevedo Gomes[1], Marcelo Galheigo[1], Kary Ocaña[1]
[1] Laboratório Nacional de Computação Científica (LNCC)
[2] Universidade Federal do Rio de Janeiro (UFRJ)

{macabral, atagomes, galheigo, karyann}@lncc.br

Introdução

O *gateway* científico BioinfoPortal gerencia a submissão de dados científicos e a execução automática de *software* de bioinformática por meio de uma interface *web* amigável e interativa. A arquitetura do BioinfoPortal está acoplada ao supercomputador Santos Dumont e ao Sistema Nacional de Ambientes de Computação de Alto Desempenho (SINAPAD), o que permite as execuções paralelas e distribuídas de *software* e *workflows* científicos de bioinformática. O BioinfoPortal utiliza, via *Web services* RESTful, o *middleware* CSGrid para permitir a extração, gerenciamento e processamento de dados em cada submissão de tarefas.

Objetivos

- Atualização das camadas de banco de dados e de gerência de execuções da arquitetura do BioinfoPortal, por meio do desenvolvimento de serviços específicos para integrar dados contidos nessas camadas.
- Análise, extração e gerência de informações de dados científicos e de proveniência extraídas das camadas da arquitetura do BioinfoPortal e de *software* de bioinformática.
- Implementação e validação de um banco de dados que centralize informações do BioinfoPortal e do ambiente computacional.
- Desenvolvimento de sistemas para criar inteligência em análise de coleta de dados e tomada de decisão, tal que melhore a eficiência do *gateway* em termos de velocidade, execução e armazenamento.

Metodologia

- Na primeira etapa, o projeto físico utilizou o PostgreSQL v10 como SGBD relacional *Open Source*.
- A segunda etapa envolve a utilização de serviços *RESTful* para o desenvolvimento dos sistemas de tomada de decisão inteligentes. A linguagem de programação utilizada é PHP.

Referência

[1] Ocaña, K.A.C.S., et al. (2020). BioinfoPortal: A scientific gateway for integrating bioinformatics applications on the Brazilian national high-performance computing network. *In Future Generation Computer Systems*, Rio de Janeiro, v. 107, p. 23, Janeiro 2020.

[2] KIM, S.-H. et al. (2017). Science Gateway Cloud With Cost-Adaptive VM Management for Computational Science and Applications. *IEEE Systems Journal* , v. 11, n. 1, p. 173-185, Março 2017. ISSN 1932-8184.

[3] LESK, A. M (2019). Bioinformatics,*Britannica*, Pennsylvania, Fevereiro 2019.

[4] Gesing S, Krüger J, Grunzke R, Herres-Pawlis S, Hoffmann A. (2016). Using Science Gateways for Bridging the Differences between Research Infrastructures,*Journal of Grid Computing*, 2016;14:545–57.

Resultados I: Arquitetura do Banco de dados

O modelo conceitual de banco de dados do BioinfoPortal foi implementado, como apresentado na Figura 1. Iniciou-se a o mapeamento dos dados na arquitetura do *gateway* para a implementação do modelo lógico.

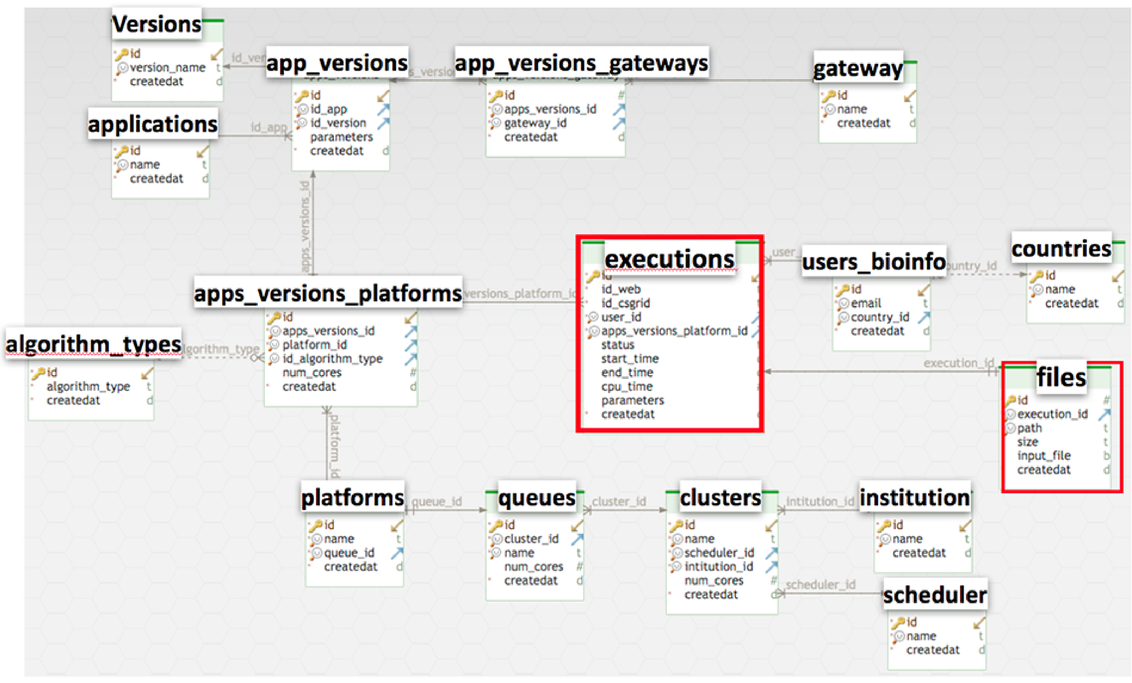

Figura 1. Esquema Conceitual Entidade-Relacionamento do Banco de Dados do Bioinfo.

Dentre as entidades do modelo conceitual ER do BioinfoPortal (Figura 1), *Files* e *Executions* são entidades originais, as demais entidades pertencem à nova versão do banco de dados.

Resultados II: Sistemas inteligentes

Os serviços RESTful propostos interagem dinamicamente com o CSGrid, SDumont e a arquitetura BioinfoPortal. O sistema de Autenticação (Figura 2A) usa métodos LDAP e RSA para conexão. O sistema de Mapeamento (Figura 2B) utiliza dados da Autenticação para mapear informações, como nome dos dados, fila, plataforma computacional, versões de *software*, *clusters* ou outras informações das diversas camadas do BioinfoPortal.

```php
<?php
    $url = ('*******');
    $data = array(
     'username'=> '******',
     'password'=> '******',
     'service'=> '*******',
     'uuid' => $uuid

    );
    $headers = array(
     #'Accept: application/xml'
      'Accept: application/json'
    );

    $handle = curl_init();
    curl_setopt($handle, CURLOPT_URL, $url);
    curl_setopt($handle, CURLOPT_HTTPHEADER, $headers);
    curl_setopt($handle, CURLOPT_RETURNTRANSFER, true);
    curl_setopt($handle, CURLOPT_SSL_VERIFYHOST, false);
    curl_setopt($handle, CURLOPT_SSL_VERIFYPEER, false);
    curl_setopt($handle, CURLOPT_POST, true);
    curl_setopt($handle, CURLOPT_POSTFIELDS, http_build_query($data));

    $response = curl_exec($handle);
    $obj = json_decode ( $response );
    $b = $obj->{'uuid'};

?>
```

Figura 2A: Sistema de Autenticação *LDAP*

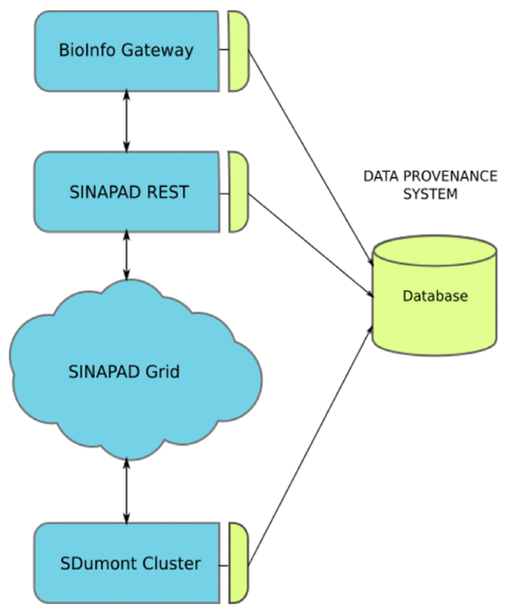

Figura 2B: Mapa de extração e armazenamento de dados

Agradecimentos

CNPq
SINAPAD
LNCC
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

Conclusão

A integração dos sistemas ao banco de dados centralizado permitirá melhor armazenamento, gerencia de metadados e informações científicas obtidas das camadas do BioinfoPortal, tais como arquivos, tempos de execução ou número de nós usados. O banco de dados acoplado à implementação dos sistemas RESTful propostos permitirá uma melhor gerência das submissões e execuções do gateway. Atualmente encontram-se em desenvolvimento os sistemas para mapear os dados de localização do usuário (IP, País) e para mapear as informações dos dados de envio. Essas informações em conjunto permitirão que técnicas como aprendizado de máquina sejam acopladas para análise preditivas de submissão dentro do Bioinfo-Portal.