Gestão de dados (trabalho 3)
CGrid 2012/2013

Cristiano Monteiro [200601169] Luís Moreira [200606788]
Mário Pinto [200404637]

15 de Maio de 2013

1 Introdução

Com a evolução das disciplinas científicas aumentou também a quantidade de dados produzidos em experiências e que têm de ser analisados. Tendo em conta que estes dados e os meios computacionais necessários para os analisar se encontram geralmente dispersos geograficamente, existe a preocupação de fornecer um serviço de gestão eficiente e que suporte certas funções.

Podemos distinguir dois tipos de funções, as relacionadas com o armazenamento e as de meta-dados. O armazenamento são mecanismos de acesso, gestão e transferência dos dados, meta-dados são mecanismos de obtenção de informação sobre os dados armazenados.

Neste trabalho temos como foco os meta-dados, nomeadamente a descrição da proveniência dos dados, isto é, a informação anterior sobre um dado, desde os processos por que passou a informação sobre a sua execução e anotações de utilizadores.

1.1 Proveniência

A proveniência pode ser definida de várias formas, dependendo do contexto em que se enquadra ou está a ser aplicada. Num contexto de sistemas de base de dados, a "data proveniência" tem a ver com a descrição da origem dos dados que levaram à construção da base de dados, em quanto que num sistema de informação geográfica descreve os processos e transformações usadas para derivar os dados. Em sistemas grid definimos proveniência como a informação que descreve o historial de modificação desses dados desde a sua fonte inicial.

Esta capacidade de obtenção de informação de linhagem (proveniência) tem várias vantagens. Em meios científicos onde os trabalhos são colaborativos a proveniência dos dados é importante de maneira a garantir que se adequam e que são de fontes seguras e que os tratamentos a qual foram expostos foram os adequados e necessários para qualquer que seja o seu propósito de utilização. Uma analogia seria a citação em artigos científicos que atestam a proveniência de algumas informações usadas para a criação do novo artigo. Guardar o processo por qual os dados são passados consiste numa boa

maneira de posteriormente, na sua reutilização, garantir que são usados adequadamente por quem necessita, filtrando os dados considerados "maus".

Em resumo podemos considerar que com a proveniência garantimos informação que nos permite:

- Qualidade dos dados tendo em conta as suas características e processos;
- Auditoria para deteção de erro;
- Facilidade de repetição e replicação do processo;
- Atribuição e responsabilização pelos dados;
- Informação complementar sobre os dados;

A linhagem pode ser registada de várias formas diferentes em termos de representação e detalhe, dependendo do meio utilizado e do custo do detalhe.

1.2 Virtual Data System (VDS)

Num ambiente Grid, que é muito utilizado pela comunidade científica, os processos a serem executados sobre os dados podem ser descritos na forma de um DAG (*Directed acyclic graph*). Esta representação apresenta uma oportunidade de implementar linhagem colocando simplesmente a essa estrutura os logs obtidos a cada passo do processo (vulgo *"workflow"*), criando assim um DAG análogo. Em cada nó está presente a informação sobre esse processo, podendo ser adicionadas anotações criadas pelo próprio utilizador.

O VDS, implementado pelo sistema Chimera, permite-nos obter o registo dessa proveniência. Funciona submetendo ao sistema o "workflow" por qual os nossos dados vão correr e as instruções, como por exemplo o nível de detalhe. O processo submetido consiste em quatro estruturas que mantém toda a informação necessária para obter a linhagem.

- Ficheiros - dados;
- Transformações - processos a serem corridos sobre os dados;
- Derivações - chamada a cada transformação;
- Anotações - criadas pelo utilizador ou pelo sistema;
- Anotações em tempo de execução - stacktrace da execução da transformação.

O VDS executa então o programa usando o globus toolkit, mas recolhendo esta informação no VDC (Virtual Data Catalog).

Para verificar se esta informação realmente existe são corridas perguntas sobre o sistema para verificar se este é ou não capaz de as responder.

2 Sobre os artigos

Começamos por ler dois artigos genéricos sobre a gestão de dados em grids, e optamos então por aprofundar o tema de proveniência, cujo o artigo principal se encontra na página da disciplina.

Os restantes foram obtidos por analise das referências desse artigo e busca no google scholar, pelas palavras chave "data","management" e "provenance".

Escolhemos dois artigos sobre arquitetura de data grids[2, 1], um com uma visão geral e uma boa explicação sobre proveniência[4] e por ultimo o artigo da página da disciplina sobre um sistema especifico de proveniência[3].

Uma critica que fazemos é no artigo de [4] não se encontra a data em que foi realizado o que torna difícil de o posicionar no tempo, também a descrição do sistema VDS é genérica de mais e parece incompleta.

Referências

[1] Bill Allcock, Joe Bester, John Bresnahan, Ann L. Chervenak, Ian Foster, Carl Kesselman, Sam Meder, Veronika Nefedova, Darcy Quesnel, and Steven Tuecke. Data management and transfer in high-performance computational grid environments. *Parallel Computing*, 28(5):749–771, May 2002.

[2] Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury, and Steven Tuecke. The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications*, 23(3):187–200, July 2000.

[3] Ben Clifford, Ian Foster, and Jens-s Voeckler. Tracking provenance in a virtual data grid. (August 2007):565–575, 2008.

[4] Yogesh L Simmhan, Beth Plale, and Dennis Gannon. A Survey of Data Provenance Techniques Technical Report IUB-CS-TR618. pages 1–25.