ILUCTUS: uma alternativa para compartilhamento dos custos de processamento e armazenagem de dados

LUCAS EDUARDO BRETANA[1]; ALANA SCHWENDLER[2]; GERSON GERALDO H. CAVALHEIRO[3]

[1]*Universidade Federal de Pelotas – lebretana@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – aschwendler@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – gerson.cavalheiro@inf.ufpel.edu.br*

1. INTRODUÇÃO

Atualmente, diversas áreas de pesquisas promovem estudos que envolvem o processamento de um grande volume de dados. Este processamento tem por objetivo extrair informações dos dados manipulados, tornando-os aplicáveis em processos de tomadas de decisões ou na simples expansão dos limites do conhecimento da área a qual se aplica. Não raro, como resultado do processamento, os dados são evoluídos, gerando novos dados que podem, eles também, ser úteis em outras pesquisas. Quando as massas de dados a serem evoluídas são de grandes proporções, é comum explorar alternativas de processamento de alto desempenho, sendo que o processamento distribuído passa a ser uma das opções de implementação com maior retorno. Neste modelo de processamento, partes de um mesmo problema, denominadas tarefas, são computadas simultaneamente em diferentes sitios computacionais. Essas tarefas devem compartilhar os seus resultados parciais de modo a evoluir o problema por completo.

Em ambientes distribuídos, além do paralelismo permitir o aumento de desempenho, em termos de tempo de processamento, ele também permite a construção de aplicações que manipulem uma quantidade de dados que vai além da capacidade de armazenamento de um único sítio físico. Nestes ambientes, a colaboração entre as tarefas da aplicação deve se dar com apoio de algum mecanismo de comunicação, como troca de mensagens, RMI, DSM. Neste trabalho nosso foco fica voltado ao modelo de Espaço de Tuplas (Carriero Jr 1987) (TS, do inglês Tuple Space) que é utilizado para o desenvolvimento da biblioteca ILUCTUS (BRETANA, SCHWENDLER, CAVALHEIRO, 2017) desenvolvida previamente e que serve de alternativa para processamento distribuído de grandes quantidades de dados.

A biblioteca ILUCTUS conta com primitivas de manipulação de dados para um modelo de computação distribuída dando a abstração de memória compartilhada ao desenvolvedor. Para isso a ILUCTUS opera os dados dispostos no TS utilizando as primitivas descritas pela linguagem de coordenação Linda, implementadas na linguagem de programação Java, em conjunto com a nuvem do Dropbox num modelo PaaS. Neste esquema, a nuvem é utilizada como Espaço de Tuplas.

Projetos que geram grandes volumes de dados podem usufruir das ferramentas oferecidas pela ILUCTUS para manipular bases de dados compartilhadas e evoluir estes dados de maneira distribuída entre vários sítios de colaboração. No uso da ILUCTUS, os colaboradores de um mesmo projeto podem utilizar suas próprias heurísticas de processamento para evoluir uma base de dados.

Desta forma, é possível se aplicar um processo de refinamento dos dados permitindo a colaboração distribuída dos custos de processamento.

O desenvolvimento da biblioteca ILUCTUS está em curso e já produziu protótipos e versões iniciais. Estas versões da biblioteca foram aferidas e documentadas em trabalhos anteriores, assim como um estudo feito sobre as dificuldades características do desenvolvimento deste tipo de projeto (BRETANA, et. al., 2018). O desenvolvimento de diferentes trabalhos realçam várias perspectivas do funcionamento da ILUCTUS e com isso diferentes caminhos para futuro de seu desenvolvimento. ASsim, este trabalho se dedica a apresentar quais são os próximos passos a serem tomados pelos desenvolvedores e quais as futuras funcionalidades a serem incorporadas à biblioteca.

O artigo se organiza como segue. O modelo de operação da biblioteca, assim como a organização do nosso modelo de negócio, é descrito na Seção 2. Decisões de projeto de desenvolvimento são apresentados na Seção 3. Por fim, na Seção 4, é descrito o estado atual da pesquisa junto dos trabalhos futuros.

2. ARQUITETURA

O desenvolvimento da biblioteca ILUCTUS considera um modelo de negócio onde são desenvolvidos **Projetos Colaborativos**. Este modelo desenvolvido é ilustrado na Figura 1. A instanciação de um novo projeto é realizada por um **Administrador**, pela criação de uma base de dados inicial contendo os dados brutos que serão processados. Em seguida, um **Colaborador** interessado no projeto requisita acesso (*request access*) ao projeto para o Administrador, que concede o acesso (*grant access*), fazendo o compartilhamento da base de dados com o Colaborador. Após o acesso ser concedido, a base de dados é copiada para o ambiente de nuvem do Colaborador. O Administrador fornece uma aplicação ao Colaborador para fazer a evolução destes dados em nuvem. Esta aplicação executa nos recursos computacionais oferecidos pelo Colaborador e faz uso da ILUCTUS para acessar os dados na nuvem deste Colaborador. As soluções obtidas pela aplicação são salvas na nuvem em uma nova base de dados até que o processamento chegue ao fim. Os resultados obtidos são então compartilhados com o Administrador do projeto. Por fim, as cópias dos dados presentes na nuvem do Colaborador são apagadas, de forma a não onerar custos de armazenamento.

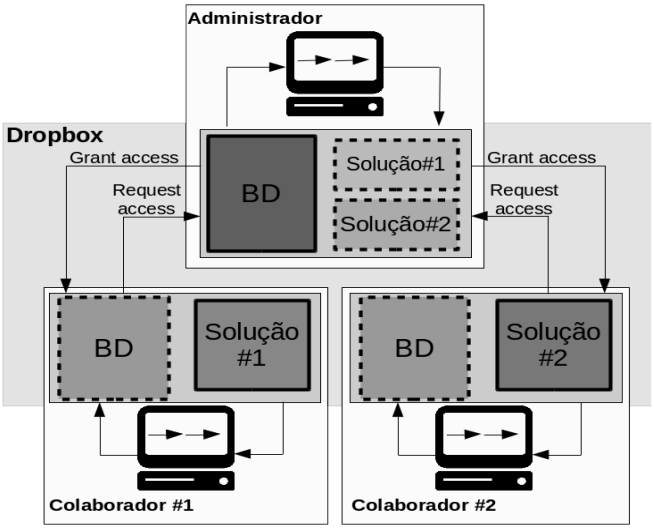

Figura 1: Modelo de negócio da ILUCTUS

O desenvolvimento da ILUCTUS se deu utilizando o padrão da linguagem de coordenação Linda (AHUJA et al. 1986) e a nuvem do Dropbox como substrato para armazenamento de dados. Sendo assim, destaca-se que, para o Dropbox, todo arquivo e diretório compartilhado irá onerar custos de armazenamento para cada um dos usuários que tenha acesso. Além disso, internamente um arquivo compartilhado é tratado como pertencente a um mesmo *namespace*. Todos os demais arquivos não compartilhados de um usuário pertencem ao seu *namespace root*. A Interface de Programação de Aplicativos do Dropbox, API Dropbox, se encontra ainda em desenvolvimento e por isso a própria ILUCTUS está sendo constantemente atualizada, utilizando dos novos recursos oferecidos e com isso aprimorando e desenvolvendo funcionalidades para a biblioteca.

3. DESENVOLVIMENTO

As atualizações da API do Dropbox permitem o aprimoramento do processo de autenticação da biblioteca ILUCTUS tornando-o mais robusto e eficiente. A fim de garantir que os dados de uma pesquisa serão acessados apenas pelos devidos membros do grupo é necessário que exista um sistema de autenticação dentro do projeto. Para fazer autenticação em um Projeto Colaborativo é necessária a troca de *tokens* de identificação entre Administrador e Colaborador. Da parte do Administrador, esses tokens, ou chaves, identificam o projeto e autorizam uma aplicação a ter acesso aos dados compartilhados na nuvem. Tokens são únicos para cada Colaborador em um Projeto Colaborativo. Pelo lado do Colaborador, o *token* identifica-o dentro do projeto, oferecendo, desta forma, acesso aos seus resultados na nuvem. Nesta troca inicial de informações entre o Administrador e Colaborador, além dos *tokens*, também é feita a troca de quaisquer outras informações necessárias para dar início ao processamento dos dados, tais como identificação da base de dados primária, lista de bases já processadas por Colaboradores pré-existentes, permissões sobre estas bases etc.

Por tratar-se de um ambiente de nuvem passivo, onde não é explorado o poder de processamento da nuvem, mas sim apenas sua capacidade de armazenamento, a primitiva **Eval** foi adaptada do modelo proposto por Ahuja et al. (1986). Porém as novas perspectivas, derivadas dos experimentos realizados, levaram a re-avaliação do comportamento implementado para determinadas características da ILUCTUS, destacando-se a implementação da Eval. Esta primitiva passa a ter a semântica de agendamento de tarefas, onde o dado a ser escrito no TS é a tarefa agendada, que será executada posteriormente, de acordo com a aplicação desenvolvida.

O uso de tecnologias de nuvem junto com o modelo de Espaço de Tuplas permitiu a criação de um modelo que permite a evolução dos dados de uma forma distribuída. Conforme mencionado, aplicações manipulando grande volume de dados podem se beneficiar deste modelo, permitindo tanto o compartilhamento dos dados pela nuvem com outros colaboradores, como o também o processamento destes dados em seus próprios recursos de computação. Sendo assim, os colaboradores compartilham não apenas os custos de armazenamento dos dados, mas também de seu processamentos.

Dentro os exemplos de aplicações que podem usar esse modelo de programação podemos destacar a pesquisa de padrões em séries temporais representando, por exemplo, a variação de determinado sinal em um certo objeto de

estudo. Nessas pesquisas os sinais são coletados com determinada frequência por meio de sensores, gerando uma grande quantidade de dados. O processo de análise destes dados consiste em aplicar técnicas para busca por padrões nestas séries. Este processo de análise dos dados, muitas vezes, pode ser coordenado de forma a ser processado em diferentes núcleos de execução distribuídos, buscando e calculando diversos tipos de informações. Integrar a ILUCTUS num trabalho deste tipo é a próxima tarefa a ser desenvolvida neste projeto.

4. ESTADO ATUAL E PERSPECTIVAS

Atualmente o desenvolvimento do projeto está focado no desenvolvimento de um trabalho colaborativo com outras áreas de pesquisa. O objetivo desta colaboração é a implementação de uma aplicação capaz de calcular, de forma colaborativa, diversos fatores sobre um determinado objeto de estudo. Neste trabalho o objeto de estudos são plantas, das quais são medidos diversos dados brutos por meio de sensores, para que então seja feito o cálculo das informações mais relevantes, e.g. o cálculo da entropia aproximada (SARAIVA, et. al., 2017). Este trabalho condiz bem com o modelo de negócio da ILUCTUS e por isso pode valer-se dos benefícios oferecidos, além de comprar o funcionamento das novas funcionalidades da biblioteca.

Como trabalhos futuros destaca-se a concretização deste trabalho em colaboração com outra área de pesquisa, que tende a propiciar novos refinamentos à biblioteca. Novas publicações servirão como documentação do desenvolvimento deste trabalho e, de novas atualizações e funcionalidades a serem desenvolvidas.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AHUJA, Sudhir; CURRIERO, N.; GELERNTER, David. Linda and friends. **Computer;(United States)**, v. 19, n. 8, 1986.

BRETANA, L. E., SCHWENDLER, A., CAVALHEIRO, G. G. H. Computação distribuída: Desafios do uso do Dropbox como suporte ao espaço de tuplas. **XVIII Escola Regional de Alto Desempenho.** Porto Alegre, p.125-128, 2018.

BRETANA, L. E., SCHWENDLER, A., CAVALHEIRO, G. G. H. ILUCTUS: Uma Biblioteca para o Apoio ao Processamento Colaborativo de Dados. **Simpósio em Sistemas Computacionais de Alto Desempenho (WSCAD)**, 2017.

CARRIERO Jr, N. J. **Implementation of tuple space machines**, 1987.

SARAIVA, G. F. R.; FERREIRA, A. S.; SOUZA, G. M. Osmotic stress decreases complexity underlying the electrophysiological dynamic in soybean. **Plant Biology**, v. 19, n. 5, p. 702-708, 2017.