Avaliação de sistemas NoSQL para o gerenciamento de dados em *workflows* científicos

Proposta de trabalho

Trabalho de formatura supervisionado
(MAC0499)

Gabriel Ferreira Guilhoto
Orientadora: Profa. Dra. Kelly Rosa Braghetto

Bacharelado em Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

19 de abril de 2015

Sumário

1 Contextualização do tema **3**

1.1 *Workflows* científicos 3

1.2 Sistemas NoSQL 4

2 Objeto do estudo **4**

3 Planejamento **5**

3.1 Atividades previstas 5

3.2 Cronograma 6

1 Contextualização do tema

A computação tem sido relevante em pesquisas científicas por permitir a análise de extensas quantidades de dados cuja análise manual seria inviável. Para se descrever experimentos de análise de dados a serem executados computacionalmente, geralmente se usam *workflows* (fluxos de trabalho) que abstraem os processos computacionais envolvidos para facilitar o uso por cientistas de diversas áreas. Para armazenar e processar esses grandes volumes de dados, existem plataformas computacionais de alto desempenho distribuídas e diferentes sistemas para o gerenciamento de dados, entre os quais estão os sistemas NoSQL. Entretanto, sistemas gerenciadores de *workflows* científicos geralmente não têm integração com sistemas NoSQL.

1.1 *Workflows* científicos

Um modelo de *workflow* define as atividades que devem ser realizadas e estabelece relações de precedência ou transferência de dados entre elas por meio de conectores. *Workflows* científicos tipicamente são modelados como fluxos de dados, o que significa que atividades produzem dados que serão manipulados por outras atividades [BC14].

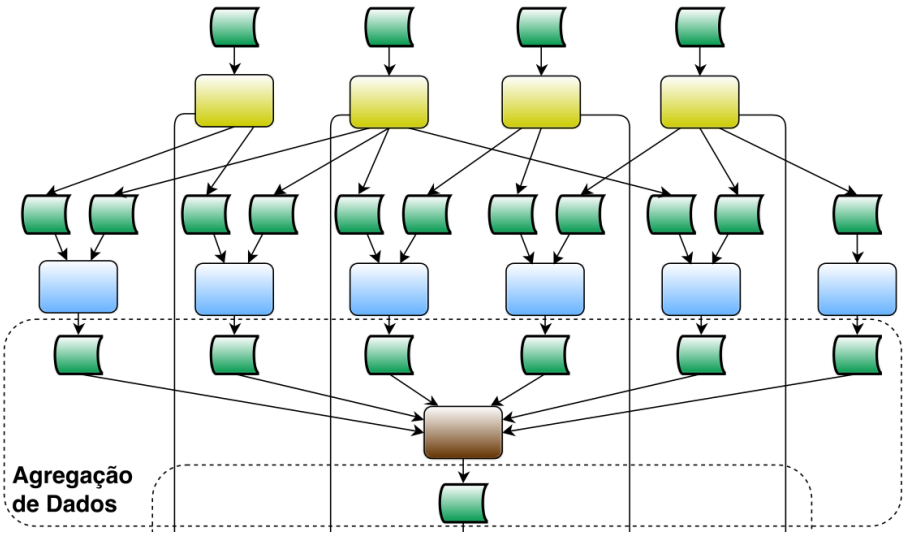

Figura 1: Exemplo de parte de um modelo de *workflow* científico. Os retângulos amarelos, azuis e marrom são atividades, as formas verdes são dados de entrada/saída e as setas são conectores. Imagem obtida de [BC14].

Existem sistemas de gerenciamento de *workflows* científicos (SGWCs) que

disponibilizam interfaces para interação dos usuários e controlam os dados e o ambiente computacional usados na execução dos *workflows*. Um SGWC pode adotar estratégias de paralelismo de dados para reduzir o tempo de execução do *workflow*. Se uma atividade processa um grande volume de dados, o SGWC pode criar cópias da atividade a serem executadas paralelamente e atribuir a cada uma delas uma partição ou uma réplica do conjunto de dados [PA06].

1.2 Sistemas NoSQL

Os sistemas NoSQL surgiram como alternativa aos sistemas gerenciadores de bancos de dados relacionais, que podem ser ineficientes para o tratamento de grandes quantidades de dados não estruturados [LdOdSP11]. Os sistemas NoSQL são preparados para explorar a elasticidade proporcionada, por exemplo, por ambientes de computação em nuvem, que permitem que recursos sejam adicionados ou removidos dinamicamente de acordo com a demanda [HYA+15]. Geralmente adotam modelos de dados flexíveis e priorizam a disponibilidade e o desempenho em detrimento da consistência forte dos dados. Uma característica comum é a capacidade de replicar e particionar os dados em vários servidores, de forma a permitir o paralelismo de operações e a consequente melhoria de desempenho [LdOdSP11].

2 Objeto do estudo

Por serem aplicados a diferentes domínios, os *workflows* científicos têm formatos de dados variados, o que pode tornar apropriada a utilização de um sistema NoSQL para gerenciamento de seus dados. Um sistema NoSQL também é conveniente por viabilizar o paralelismo de dados por meio de replicação e particionamento, para que seja possível manipular grandes volumes de dados eficientemente.

Nesse sentido, neste trabalho de conclusão de curso serão avaliados e comparados sistemas NoSQL no gerenciamento dos dados de *workflows* científicos. O trabalho está relacionado com o mestrado em andamento de Elaine Naomi Watanabe, que propõe uma extensão de um SGWC para permitir o uso de um sistema NoSQL com o objetivo de prover execuções mais eficientes para os *workflows* científicos. O mecanismo proposto no mestrado gera um modelo NoSQL a partir de uma descrição de um *workflow*, seus dados de entrada e anotações sobre o tipo de processamento que cada atividade realiza sobre seus dados, essenciais para possibilitar o paralelismo de dados [Wat15].

Existe um conjunto de modelos de *workflows* científicos sintéticos gerados a partir de aplicações reais que exemplificam várias construções de fluxos de dados bastante frequentes e são usados para testar ferramentas [BCD+08]. Com base nesses *workflows*, serão definidas operações de consulta e manipulação de dados relevantes para a execução de *workflows* científicos.

Neste trabalho, serão avaliados diferentes sistemas NoSQL quanto ao tempo de execução das operações e ao consumo de processador e memória. O objetivo do trabalho é verificar a adequação de sistemas NoSQL à execução de *workflows* científicos e identificar qual sistema é mais apropriado.

3 Planejamento

3.1 Atividades previstas

1. Escolha de um conjunto de *workflows* sintéticos como estudo de caso para o trabalho (a partir do conjunto disponível em [MJ])

2. Identificação dos tipos de operações de gravação, replicação, particionamento e consultas de dados realizados com mais frequência, com base nos *workflows* sintéticos

3. Estudo de sistemas NoSQL, para identificar:

   (a) quais recursos eles possuem para replicação/particionamento de dados

   (b) como cadastrar/recuperar dados pela interface de acesso ao sistema

   Pretende-se investigar sistemas NoSQL do tipo chave-valor (como Redis [Red] e Riak [Tec]), orientados a documentos (como MongoDB [Mon] e CouchDB [Foub]) e orientados a colunas (como Cassandra [Foua]) e verificar quais são mais adequados.

4. Definição de uma API para os tipos de operações identificadas no item 2

5. Implementação da API do item 4 para os sistemas NoSQL avaliados como viáveis de serem usados com os *workflows* científicos (de acordo com o item 3)

6. Avaliação do desempenho das implementações para os diferentes sistemas NoSQL, usando como base os *workflows* sintéticos

7. Escrita da monografia

3.2 Cronograma

	Atividades						
Meses	1	2	3	4	5	6	7
Abril	x	x					x
Maio	x	x	x				x
Junho			x	x			x
Julho					x		x
Agosto					x		x
Setembro					x	x	x
Outubro						x	x
Novembro							x

Referências

[BC14] Kelly Rosa Braghetto e Daniel Cordeiro. Introdução à modelagem e execução de workflows científicos. *Atualizações em Informática. 1ed. Porto Alegre: SBC*, páginas 1–40, 2014. Disponível em `http://www.ime.usp.br/~kellyrb/files/jai2014-cap-workflows.pdf`. Acessado em 15 de abril de 2015. 1.1, 1

[BCD+08] S. Bharathi, A. Chervenak, E. Deelman, G. Mehta, Mei-Hui Su, e K. Vahi. Characterization of scientific workflows. Em *Workflows in Support of Large-Scale Science, 2008. WORKS 2008. Third Workshop on*, páginas 1–10, novembro 2008. 2

[Foua] The Apache Software Foundation. The Apache Cassandra Project. `http://cassandra.apache.org/`. Acessado em 19 de abril de 2015. 3

[Foub] The Apache Software Foundation. Apache CouchDB. `http://couchdb.apache.org/`. Acessado em 19 de abril de 2015. 3

[HYA+15] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, e Samee Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47(0):98 – 115, 2015. 1.2

[LdOdSP11] Bernadette Farias Lóscio, Hélio Rodrigues de Oliveira, e Jonas César de Sousa Pontes. NoSQL no desenvolvimento de aplicações web

colaborativas. *Anais do VIII Simpósio Brasileiro de Sistemas Colaborativos*, outubro 2011. 1.2

[MJ] Gaurang Mehta e Gideon Juve. WorkflowGenerator. Disponível em `https://confluence.pegasus.isi.edu/display/pegasus/WorkflowGenerator`. Acessado em 15 de abril de 2015. 1

[Mon] MongoDB, Inc. MongoDB. `https://www.mongodb.org/`. Acessado em 19 de abril de 2015. 3

[PA06] C. Pautasso e G. Alonso. Parallel computing patterns for grid workflows. Em *Workflows in Support of Large-Scale Science, 2006. WORKS '06. Workshop on*, páginas 1–10, junho 2006. 1.1

[Red] Redis. `http://redis.io/`. Acessado em 19 de abril de 2015. 3

[Tec] Basho Technologies. Riak. `http://basho.com/riak/`. Acessado em 19 de abril de 2015. 3

[Wat15] Elaine Naomi Watanabe. Gerenciamento de grandes volumes de dados na execução de workflows científicos em nuvens computacionais. janeiro 2015. Qualificação de mestrado. 2