UNIVERSIDADE FEDERAL FLUMINENSE

LEONARDO SOUSA LIMA RAMOS

`PhenoManager` : uma Abordagem para a Gerência de Hipóteses Científicas

Niterói

2019

UNIVERSIDADE FEDERAL FLUMINENSE

LEONARDO SOUSA LIMA RAMOS

`PhenoManager` : uma Abordagem para a Gerência de Hipóteses Científicas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Engenharia de Sistemas e Informação

Orientador:
Daniel Cardoso Moraes de Oliveira

Co-orientador:
Fabio Andre Machado Porto

Niterói

2019

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

R175p Ramos, Leonardo Sousa Lima
PhenoManager : uma Abordagem para a Gerência de Hipóteses Científicas / Leonardo Sousa Lima Ramos ; Daniel Cardoso Moraes De Oliveira, orientador ; Fabio Andre Machado Porto, coorientador. Niterói, 2019.
54 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2019.

DOI: http://dx.doi.org/10.22409/PGC.2019.m.14381925769

1. Workflow científico. 2. EScience. 3. Modelo Conceitual. 4. Hipótese Científica. 5. Produção intelectual. I. De Oliveira, Daniel Cardoso Moraes, orientador. II. Porto, Fabio Andre Machado, coorientador. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

LEONARDO SOUSA LIMA RAMOS

`PhenoManager` : uma Abordagem para a Gerência de Hipóteses Científicas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Engenharia de Sistemas e Informação

BANCA EXAMINADORA

Prof. D.Sc. Daniel Cardoso Moraes de Oliveira - Orientador, UFF
(Presidente)

Prof. D.Sc. Fabio Andre Machado Porto, LNCC

Prof^{a}. D.Sc. Aline Marins Paes Carvalho, UFF

Prof. D.Sc. Antônio Tadeu Azevedo Gomes, LNCC

Prof^{a}. D.Sc. Kary Ann del Carmen Ocaña Gautherot, LNCC

Niterói

2019

Dedico este trabalho aos meus orientadores pelo conhecimento adquirido, a minha esposa pela paciência e a meus pais e amigos pelo apoio.

Resumo

A pesquisa científica baseada em simulações computacionais é complexa uma vez que envolve o gerenciamento de enormes volumes de dados e metadados produzidos durante o ciclo de vida de um experimento científico, desde a formulação de hipóteses até sua avaliação final. Essa riqueza de dados precisa ser estruturada e gerenciada de uma maneira que faça sentido para os cientistas, de modo que o conhecimento relevante possa ser extraído para contribuir para o processo de investigação científica. Além disso, tratando-se do escopo do projeto científico como um todo, o mesmo pode estar associado a vários experimentos científicos diferentes, que por sua vez, podem requerer execuções de diferentes *Workflows* científicos, o que torna a tarefa bastante árdua. Tudo isso pode se tornar ainda mais difícil se considerarmos que as tarefas do projeto devem estar associadas à execução de tais simulações (que podem demorar dias ou semanas), que as hipóteses de um fenômeno carecem de validação e de reproduções e que a equipe do projeto pode se encontrar geograficamente dispersa. Este trabalho apresenta uma abordagem chamada `PhenoManager` que tem como objetivo auxiliar os cientistas a gerenciarem seus projetos científicos e o ciclo do método científico como um todo. O `PhenoManager` é capaz de auxiliar o cientista na estruturação, validação e reprodução de hipóteses de um fenômeno, por meio de modelos computacionais configuráveis na abordagem, além de se integrar com o sistema *SciManager*, que já trata da gerência de projetos científicos e suas tarefas. O `PhenoManager` é baseado em uma arquitetura de nuvem, o que faz com que esteja disponível para membros do projeto em locais dispersos. Para a avaliação deste trabalho foi utilizado o SciPhy, um *Workflow* científico da área de bioinformática, chegando a conclusão que a ferramenta proposta traz ganhos sem perdas consideráveis de performance.

Palavras-chave: Projeto científico, experimento científico, *workflow* científico, Modelo Conceitual, *eScience*, Hipótese Científica.

Abstract

Scientific research based on computer simulations is complex since it may involve managing the enormous volumes of data and metadata produced during the life cycle of a scientific experiment, from the formulation of hypotheses to its final evaluation. This wealth of data needs to be structured and managed in a way that makes sense to scientists so that relevant knowledge can be extracted to contribute to the scientific research process. In addition, when it comes to the scope of the scientific project as a whole, it may be associated with several different scientific experiments, which in turn may require executions of different scientific workflows, which makes the task rather arduous. All of this can become even more difficult if we consider that the project tasks must be associated with the execution of such simulations (which may take days or weeks), that the hypotheses of a phenomenon need validation and replication, and that the project team may be geographically dispersed. This work presents an approach called `PhenoManager` that aims to help scientists manage their scientific projects and the cycle of the scientific method as a whole. `PhenoManager` is able to assist the scientist in structuring, validating and reproducing hypotheses of a phenomenon through configurable computational models in the approach, as well as integrating with the SciManager system, which already deals with the management of scientific projects and their tasks. `PhenoManager` is based on a cloud architecture, which makes it available to project members in scattered locations. For the evaluation of this work was used SciPhy, a scientific workflow in the field of bioinformatics, reaching the conclusion that the proposed tool brings gains without considerable performance losses.

Keywords: Scientific project, scientific experiment, scientific workflow, Conceptual Model, eScience, Scientific Hypothesis.

Lista de Figuras

2.1 Relação entre os conceitos do método científico. 6

2.2 Ciclo de vida do experimento científico adaptado de Mattoso et al [41]. . . 10

2.3 Ciclo de vida de um projeto científico. 11

2.4 Ciclo de vida de exploração científica In-Silico [52] 12

2.5 Modelo Conceitual de Hipótese Científica [52] 13

3.1 Diagrama ER do `PhenoManager` 19

3.2 Arquitetura do `PhenoManager` 20

3.3 Tela de *login*. 24

3.4 Demonstração de criação de um ambiante *Cloud* com configuração específica para cada VM. 26

3.5 *Logs* de execuções de um modelo computacional exibidos em tempo real. . 27

3.6 Histórico de execuções de um modelo computacional. 28

3.7 Diagrama de sequência de execução de modelo computacional. 30

3.8 *Research Object de uma execução de um modelo computacional.* 31

4.1 O *Workflow* SciPhy. 33

4.2 Comparativo dos tempos de execução com a fila vazia. 35

4.3 Comparativo dos tempos de execução com cem execuções paralelas na fila. 35

4.4 Resultado da avaliação do treinamento. 38

4.5 Resultado da avaliação da facilidade de uso. 39

4.6 Resultado da avaliação da utilidade. 39

Lista de Tabelas

4.1 Tempos de execução dos experimentos (em minutos) 36

4.2 Questões utilizadas na avaliação do `PhenoManager` 37

Lista de Abreviaturas e Siglas

SGWfC	:	Sistema de Gerência de *Workflow* Científico
API	:	*HTTP* Application Programming Interface
VM	:	Máquina Virtual
JSON	:	*Extensible Markup Language*
UUID	:	*Universally Unique Identifier*
DOI	:	*Digital Object Identifier* - padrão para identificação de documentos em redes digitais
ORCID	:	*Open Researcher and Contributor ID* - código alfanumérico não proprietário para identificar exclusivamente cientistas e outros autores académicos e contribuidores

Sumário

1 Introdução **1**

2 Fundamentação Teórica **5**

2.1 Projeto, Experimento e *Workflow* Científicos 5

2.2 Proveniência de Dados 7

2.3 Ciclo de Vida do Experimento Científico 9

2.4 Ciclo de Vida do Projeto Científico 10

2.5 O Modelo de Dados de Hipóteses Científicas 12

2.6 Ambientes de Processamento de Alto Desempenho 14

2.7 *Research Objects* 16

3 Abordagem Proposta: `PhenoManager` **18**

3.1 Modelo de dados 18

3.2 Arquitetura do Sistema 19

3.2.1 O *SciManager* 22

3.3 Funcionalidades 23

3.3.1 Gerência de usuários e Grupos de usuários 23

3.3.2 Gerência de Projetos, Fenômenos, Hipóteses e Experimentos Científicos 24

3.3.3 Criação de modelos computacionais 25

3.3.3.1 Fluxo de processamento de mensagem de execução do modelo 27

3.3.4 Exportação de *Research Objects* 29

3.3.5 Integração com o *SciManager* 29

4 Avaliação Experimental **32**

4.1 Estudo de Caso: o *Workflow* Sciphy 32

4.1.1 Ambiente de Execução para Análise de *Overhead* 33

4.2 Avaliação do `PhenoManager` com Usuários 36

4.3 Consulta de dados no `PhenoManager` 40

5 Trabalhos Relacionados **43**

6 Conclusão **47**

Referências **49**

Capítulo 1

Introdução

A construção do conhecimento científico se dá pela utilização de rigorosos métodos científicos [11]. Cada hipótese criada a partir da observação de um determinado fenômeno deve ser minuciosamente testada por meio de experimentos de modo a garantir sua validade. Dessa forma, as teorias científicas são resultantes da obtenção dos dados da experiência adquiridos pela observação e também por análises [11].

O teste de hipóteses *in silico* envolve a execução de experimentos computacionais, representando os modelos matemáticos e confrontando dados simulados com observações coletadas [52]. Analisar dados das diferentes versões de execuções de um experimento, qualificando e ordenando os resultados obtidos, é uma tarefa que requer bastante disciplina e organização por parte do cientista. O ponto de partida de uma investigação científica é a descrição de um fenômeno, seja ele natural ou não. O fenômeno estudado ocorre em algum espaço-tempo, em que se observam quantidades físicas selecionadas [52]. As hipóteses científicas, por sua vez, interpretam conceitualmente o fenômeno estudado por meio de sua representação ou através de modelos matemáticos [52].

O processo de experimentação é uma das formas usadas para apoiar as teorias baseadas em um método científico [33]. De forma a considerar um corpo de conhecimento como sendo de fato "científico", sua veracidade e validade devem ser comprovadas [67, 41]. No método científico, um experimento científico consiste na montagem de um protocolo concreto a partir do qual se organizam diversas ações observáveis, direta ou indiretamente, de forma a corroborar ou refutar uma dada hipótese científica que foca em estabelecer relações de causa/efeito entre fenômenos [8, 25].

Em experimentos científicos tradicionais, como aqueles clássicos da química e da física, os cientistas realizam suas pesquisas em laboratórios ou no campo. Entretanto, o ato

de "fazer ciência" não se resume mais somente à pesquisa em laboratórios ou no campo. A evolução da ciência da computação nas últimas décadas permitiu a exploração de novos tipos de experimentos científicos baseados em simulação [67, 41]. Neste novo cenário, os experimentos científicos se tornaram dependentes da utilização maciça de recursos computacionais e de um aparato tecnológico especializado. Um experimento desta categoria é caracterizado pelo uso de simulações do mundo real, realizadas em ambientes virtuais e que são baseadas em modelos computacionais complexos. Em grande parte dos casos, estes modelos se referem ao encadeamento de programas que são utilizados durante as simulações. Cada execução de programa pode consumir ou produzir um grande volume de dados.

Um experimento de larga escala pode ser parte integrante de projetos científicos que podem ser conduzidos por equipes geograficamente dispersas. Um projeto científico inclui não somente a modelagem, execução e a análise dos dados de simulação, mas também toda a parte "burocrática" que envolve relatórios de andamento, análise de estatísticas, atribuição de responsabilidades, entre outros [53].

Até os dias de hoje, não é incomum um cientista implementar um *script* para modelar e executar seu experimento, de forma *ad hoc* [15], o que torna o experimento muito mais lento e caro, uma vez que o cientista neste cenário necessita ter um conhecimento computacional considerável, e o reuso de código não é algo trivial, visto que o código foi feito se pensando em uma determinada pesquisa, e por este motivo pode não ser reutilizado em outras pesquisas de forma simples. Além disto o cientista fica com a obrigação de modelar a captura de dados para responder perguntas como: Quem criou esse dado? Quem consumiu este dado? Quando ele foi alterado e por quem? Este processo não só consumia tempo, mas também era muito suscetível a erros [15]. Existem algumas soluções que auxiliam no uso de *scripts* para experimentos científicos como o noWorkflow [45], mas se limitam a experimentos sequenciais, ou seja, programas sequenciais com alguma dependência, de pequena escala. Atualmente, esses experimentos podem ser modelados como *workflows* científicos [41], e são especificados, executados e monitorados por sistemas complexos chamados de Sistemas de Gerência de *Workflow* Científicos (SGWfC) [68, 18, 38, 48] que possuem um motor de execução acoplado, responsável por invocar cada um dos programas do fluxo sem a necessidade do cientista se preocupar com a gerência do *workflow* em si. Exemplos de SGWfC são o Swift/T [71], o Pegasus [21], o Kepler [3] e o SciCumulus [18]. Apesar de eficazes, tais SGWfCs requerem uma grande curva de aprendizado para serem usados em larga escala, isto é, dependendo do tipo de processamento a ser realizado, tanto a etapa de configuração quanto o uso do SGWfC

pode não ser trivial.

Uma opção aos SGWfC existentes e aos *scripts* são as ferramentas para Computação Escalável e Intensiva em Dados (*Data-Intensive Scalable Computing* – DISC), que fornecem modelos de programação na qual os usuários desenvolvem a lógica para o processamento dos dados [9, 12]. Os modelos construídos são compilados durante a sua execução na forma de um grafo acíclico direcionado (*Directed Acyclic Graph* – DAG) constituído por operadores de dados paralelos [31]. Portanto, é possível realizar otimizações durante o escalonamento das execuções das atividades de processamento de dados [55]. A maioria das ferramentas DISC implementa o paradigma de programação *MapReduce*, que é inspirado em duas funções do modelo de programação funcional: o *Map* e o *Reduce* [19, 58, 69, 74]. Alguns exemplos de ambientes DISC são o Apache Hadoop [58] e o Spark [74].

Entretanto, a avaliação e a validação de uma hipótese científica pode necessitar da execução de diversos *Workflows*, *scripts* ou aplicações DISC distintos que podem estar executando em ambientes distribuídos e de alto desempenho, como as nuvens e supercomputadores. Dessa forma, gerenciar o projeto científico e qualificar os dados obtidos se torna uma tarefa ainda mais árdua do que quando consideramos um *workflow* ou *script* de forma isolada. Gerenciar projetos complexos não é uma exclusividade da área científica. Problemas de administração já são enfrentados há anos na área comercial. A área de Gerência de Projetos [54] tem como objetivo tratar tais problemas. Um exemplo da área comercial, no contexto de projetos de *software*, foi o desenvolvimento do Windows 7 que contou com mais de 2.000 funcionários espalhados por todo o mundo [1]. Na gerência de projetos assume-se a premissa de que a qualidade do produto final é influenciada pela qualidade do processo que levou à concepção de tal produto. Desta forma, se a condução de um projeto é bem gerenciada o produto final tem mais chances de ter uma qualidade aceitável para determinado fim. Podemos aplicar a mesma premissa no contexto de projetos científicos. Isto é: "se um projeto possuir um processo devidamente especificado e controlado; for executado seguindo procedimentos definidos; a análise dos dados seguir um protocolo; e a documentação for realizada, então o produto final (as conclusões acerca da hipótese científica) terá mais chance de qualidade maior".

Dessa forma, seria interessante que os cientistas tivessem acesso a uma abordagem que auxiliasse na gerência do projeto científico como um todo e no apoio ao método científico, ajudando na documentação, compartilhamento dos dados obtidos e na facilitação

[1] http://www.computerworld.com/article/2532600/operating-systems/microsoft-may-have-2-000-developersworking-on-windows-7.html

da reprodução dos experimentos realizados. Além disso, especificamente para a gerência de hipóteses que fogem da alçada de *Workflows* Científicos, ainda existe uma escassez de abordagens para tratar desse assunto, o que representa um desafio em aberto. Sendo assim, esta dissertação se propõe a tentar sanar ou, ao menos, amenizar as dificuldades encontradas no campo prático, implantando uma abordagem chamada de `PhenoManager` que visa apoiar a gerência e validação de uma gama maior de tipos de hipóteses científicas.

O `PhenoManager` aborda desde a etapa de concepção, de configuração do modelo de execução, até a validação e reprodução dos experimentos, sejam eles executados como *workflows*, *scripts* ou aplicações DISC, por meio dos dados de proveniência [24]. Além disso, para aproveitarmos o arcabouço de gerência de projetos, a solução proposta nesta dissertação foi feita de maneira a se integrar com o sistema *SciManager* [53] (que gerencia equipes em projetos científicos) em um mesmo ecossistema de *software*, de maneira que ambas as soluções sejam transparentes e aproveitem de suas funcionalidades. As aplicações se comunicam pelo protocolo *HTTP* e todas foram construídas utilizando o padrão arquitetural de microsserviços.

O restante desta dissertação se encontra organizado em 6 capítulos além desta introdução. No Capítulo 2 é apresentado o referencial teórico. No Capítulo 3 será discutido e apresentado o O `PhenoManager` . O Capítulo 4 apresenta a avaliação experimental. O Capítulo 5 discute os trabalhos relacionados e, por fim, o Capítulo 6 conclui esta dissertação e aponta trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo aborda os conceitos que serviram como alicerce para a construção desta dissertação.

2.1 Projeto, Experimento e *Workflow* Científicos

No contexto desta dissertação existem conceitos importantes a serem alucidados, e alguns deles são fundamentais para a compreensão deste trabalho, a saber: projetos científicos, experimentos científicos, hipótese científica, fenômeno, pesquisa e *workflows* científicos. Um projeto científico é a unidade de trabalho de mais alto nível e tem como objetivo principal funcionar como um roteiro de trabalho ou instrumento de planejamento, além de ser o elemento direcionador da pesquisa. Desta forma, um projeto científico deve definir qual(is) o(s) objetivo(s) da pesquisa e como a mesma será conduzida. Cada projeto conta com um grupo de pessoas para as quais as tarefas e responsabilidades relativas ao projeto serão atribuídas. Dependendo da quantidade de objetivos a serem alcançada, um projeto científico pode ser composto por um ou mais experimentos científicos.

Podemos definir um experimento científico, segundo o dicionário *Oxford*, como "um teste executado sob condições controladas, que é realizado para demonstrar uma verdade conhecida, examinar a validade de uma hipótese, ou determinar a eficácia de algo previamente não explorado" [64]. Um experimento científico está estritamente associado a um conjunto de ações controladas (*i.e.,* um protocolo) com um objetivo bem definido. Essas ações incluem variações de testes e seus resultados são, geralmente, comparados entre si para aceitar ou refutar uma hipótese científica [41]. Ou seja, de forma a confirmar ou refutar uma hipótese, experimentos devem ser definidos, e esses experimentos podem demandar a execução de diversas simulações implementadas de diferentes formas, conforme

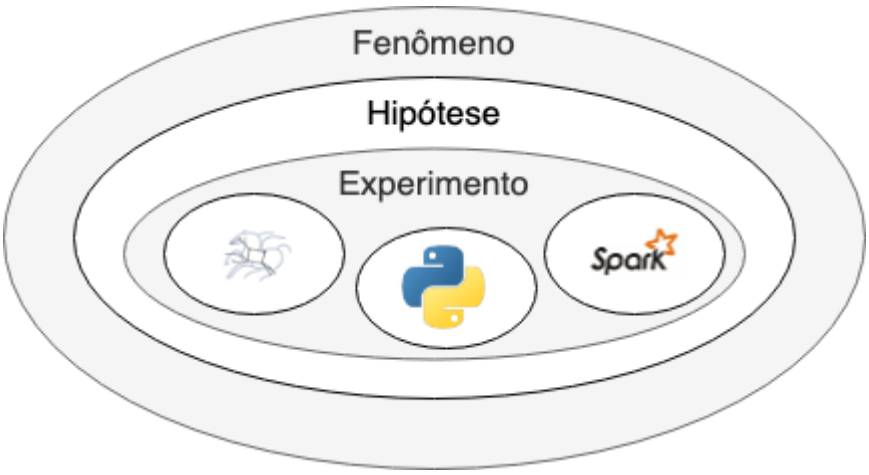


Figura 2.1: Relação entre os conceitos do método científico.

apresentado na Figura 2.1 (um workflow no SGWf Pegasus, um script Python e uma aplicação MapReduce implementada no Apache Spark).

Existem diversos tipos de experimentos científicos, a saber: *in vivo, in vitro, in virtuo e in silico* [67]. Nesta dissertação, o foco é o apoio aos experimentos *in silico*, *i.e.,* aqueles que são baseados em simulações computacionais. No contexto desta dissertação, o termo "experimento científico" será consistentemente utilizado para referenciar experimentos científicos baseados em simulação. Neste tipo de experimento, tanto o ambiente quanto os participantes do experimento são simulados em ambientes computacionais. Este tipo de experimento é utilizado nos mais diversos domínios científicos como, por exemplo, análises filogenéticas [47], genômica comparativa [39], processamento de sequências biológicas [36], estudos na área de saúde [2], prospecção de petróleo em águas profundas [35], mapeamento dos corpos celestes [32], ecologia [56], agricultura [17], busca de genes ortólogos dos tripanosomas causadores de doenças tropicais negligenciadas [13], dinâmica de fluidos computacional [28], estudos fisiológicos [70]. Todos esses exemplos podem ser considerados de larga escala por consumirem e produzirem um grande volume de dados e são um ponto de inflexão que merece o desenvolvimento de estudos específicos.

Experimentos em larga escala como os anteriormente citados podem ser executados milhares (ou em alguns casos, milhões) de vezes para produzir um resultado. Como cada execução destas pode demandar muitos recursos e tempo, esses experimentos usualmente requerem técnicas de paralelismo e execução em ambiente de processamento de alto desempenho (PAD), como *clusters* e supercomputadores, grades computacionais, ambientes de computação voluntária, nuvens de computadores, e mais recentemente os ambientes DISC. Esses experimentos são especialmente complexos de serem gerenciados pelo cien-

tista devido às questões de infraestrutura computacional como a grande quantidade de programas envolvidos na simulação e a quantidade de recursos computacionais necessários. Não é trivial controlar o volume de dados manipulados, evitar e contornar falhas de execução, *etc.* Essa tarefa se torna ainda mais complexa se necessitarmos capturar e armazenar descritores da execução para garantir a reprodutibilidade do mesmo [24]. Essa captura e armazenamento é uma característica fundamental para que um experimento seja considerado "científico" de fato.

Geralmente esses experimentos científicos representam o encadeamento e execução de diferentes programas, que podem consumir múltiplas combinações de parâmetros e grandes quantidades de dados. Assim, experimentos científicos podem ser modelados como *workflows* científicos. *Workflows* representam esse encadeamento por meio de artefatos executáveis e que representam uma das possíveis variações do experimento. Os *workflows* científicos são uma alternativa atraente para representar os encadeamentos de programas ao invés de usarmos uma abordagem ad hoc manual ou baseada em scripts. Os *workflows* científicos podem ser definidos como uma abstração para modelar o fluxo de atividades e de dados em um experimento. Em *workflows* científicos, essas atividades são geralmente programas ou serviços que representam algoritmos e métodos computacionais sólidos [40, 41].

Dessa forma, temos a seguinte interconexão entre os conceitos de projeto científico, experimento científico e *workflow*. Um *workflow* científico faz parte de um determinado experimento, que, por sua vez, se encontra no contexto de um projeto científico. Um *workflow* é a representação concreta de um experimento científico e simboliza um dos possíveis ensaios do experimento (*i.e., trials*). Já cada experimento segue um ciclo de vida bem definido que envolve as fases definidas por Mattoso *et al.* [41], apresentado na Seção 2.2. Além disso, será apresentado na Seção 2.3 que o projeto possui um ciclo de vida próprio independente do ciclo de vida do experimento.

2.2 Proveniência de Dados

A proveniência normalmente é utilizada no apoio aos experimentos científicos, dados de proveniência são metadados associados as diversas etapas do experimento científico, no contexto desta dissertação, do *workflow* científico, *script* ou aplicação MapReduce. Esses metadados são informações complementares que contêm respostas para as perguntas "como, quando, onde e por que um determinado dado foi obtido? E quem o obteve?" [10].

Ou seja, são informações relacionadas ao experimento, tais como a definição do *workflow* ou *script*, bem como os dados iniciais utilizados e os dados gerados durante a sua execução e como eles se relacionam. Assim, a proveniência é capaz de representar o histórico de execução do experimento. Além disso, ela auxilia o processo de investigação dos dados, tais como os processos de criação e validação dos mesmos. Dessa forma, a proveniência permite a reprodução do experimento por outro cientista e por esta razão é tida como fundamental para que o experimento seja considerado consistente e válido [10].

Nesta dissertação adotamos a definição de proveniência dada pelo W3C (*The World Wide Web Consortium*) *Provenance Working Group's* [44], que diz que a proveniência é um registro que descreve pessoas, instituições, entidades e atividades que produzem, influenciam ou entregam pedaços de dados e informações. De forma geral, a proveniência de dados facilita a repetição ou a reprodução de um resultado, além de permitir a avaliação da qualidade e a confiabilidade de uma determinada informação em um experimento científico [44, 37].

Através da análise dos dados de proveniência é possível, por exemplo [59]:

i) Mapear o fluxo dos dados;

ii) Determinar a utilização de recursos por cada etapa do *workflow*, *script* ou aplicação;

iii) Detectar erros na geração de dados ou no processo;

iv) Estimar a qualidade ou confiabilidade dos dados baseando-se na origem dos dados;

v) A replicação ou derivação de dados;

vi) A realização de consultas baseadas nos metadados de origem para a descoberta de dados.

Assim, a proveniência é amplamente utilizada, pois permite a reprodutibilidade do experimento científico (propriedade que possibilita a capacidade de outro cientista reproduzir o mesmo o experimento). Além disso, ela oferece um melhor entendimento sobre os experimentos científicos. Existem dois tipos de proveniência: a prospectiva e a retrospectiva [24, 14, 15]. A proveniência prospectiva é responsável por armazenar os passos do experimento que são necessários para gerar uma informação ou uma classe de informações. Já a retrospectiva captura os passos que foram executados assim como informações do ambiente em que o experimento foi executado. Em outras palavras a proveniência prospectiva diz respeito ao plano de execução do *workflow* ou *scripts*, enquanto a retrospectiva nos informa a respeito de suas execuções.

Para armazenar a proveniência retrospectiva devem ser utilizados, preferencialmente, modelos de dados que se baseiem na recomendação do W3C PROV [43], a qual propõe uma representação genérica de proveniência. A recomendação PROV optou por representar a proveniência por meio do modelo ER, isto é, por meio de um modelo de dados que descreve os dados ou aspectos de informação de um domínio de negócio ou seus requisitos de processo, de uma maneira abstrata que em última análise se presta a ser implementada de alguma forma, o PROV já possui ontologias associadas a sua representação, o que facilita a interoperabilidade. Tanto o PROV quanto o OPM, *Open Provenance Model*, que foi um modelo apresentado anteriormente ao PROV, expressam as relações causais entre Processos, Agentes, Artefatos e Papéis existentes em *workflows*, aplicações e *scripts*. No contexto destas representações, o Artefato representa um objeto físico ou digital. O Processo é definido como uma atividade ou série de atividades realizadas em artefatos ou causada por eles, e que resulte na produção de novos artefatos. O Agente é uma entidade que atua como um catalisador de um processo, permitindo, acelerando, controlando e afetando a sua execução.

Uma das principais vantagens de se utilizar recomendações como estas é garantir a interoperabilidade de descritores de proveniência oriundos de ambientes heterogêneos, independentemente da tecnologia e dos SGWfC utilizados ou da linguagem de *script* escolhida. Por esse motivo, já vem sendo utilizada por diversos SGWfC [30] e por soluções de captura de proveniência em *scripts* [45] como formato para exportação de proveniência e foi foco de diversos fóruns de discussão [65].

2.3 Ciclo de Vida do Experimento Científico

O ciclo de vida de um experimento científico em larga escala, segundo Mattoso *et al.* [41] consiste basicamente em múltiplas interações, a serem realizadas por cientistas, durante o curso de um experimento. Uma vez que podem existir diversos experimentos coexistindo dentro de um mesmo projeto científico, podemos ter mais de um ciclo de vida, superpostos ou interagindo entre si. Na Figura 2.2 é apresentada uma versão simplificada do ciclo proposto por Mattoso *et al.* [41] onde as principais fases podem ser identificadas: a **execução**, a **composição** e a **análise**.

Na fase de composição, os cientistas definem a estrutura do experimento científico, isto é, a sequência lógica das atividades, os tipos de dados de entrada/saída e parâmetros que devem ser fornecidos. A fase de execução é responsável por executar uma especifi-

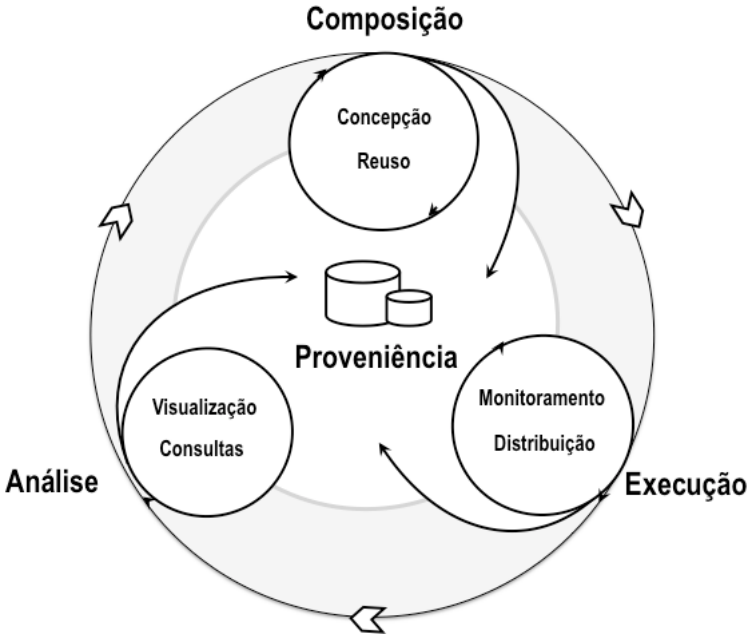


Figura 2.2: Ciclo de vida do experimento científico adaptado de Mattoso et al [41].

cação de *workflow* ou *script* (*trial*) em um determinado SGWfC, máquina de execução ou *framework*. Finalmente, a fase de análise apoia a interpretação e avaliação dos dados gerados pelas fases de composição e execução. Esta fase é altamente dependente de dados de proveniência [24] do experimento que foram gerados nas fases anteriores. Os dados de proveniência registram o histórico do experimento, desde sua especificação, os parâmetros utilizados, até os tempos de execução de cada atividade do mesmo. Com esses dados os cientistas são capazes de auditar execuções e garantir a reprodutibilidade do mesmo.

2.4 Ciclo de Vida do Projeto Científico

Da mesma forma que o ciclo de vida do experimento apresenta fases e características definidas, o ciclo de vida de um projeto científico (quando baseado em técnicas/metodologias de gerência de projetos) consiste em um conjunto de fases que são executadas pelos membros do projeto, conforme apresentado na Figura 2.3. Cada uma das fases apresentadas na Figura 2.3 possui diversas atividades realizadas por membros do projeto. Essas atividades são consideradas como "finalizadas" quando é feita a entrega dos produtos, comumente chamados de "entregáveis" (do inglês *deliverables*).

Na primeira fase deste ciclo, chamada de "Início do Projeto", são discutidos os objetivos do projeto, quais metodologias serão utilizadas e quem serão os membros (cientistas) e os gerentes (chefes de laboratório). Todas essas informações são requisitos de mais alto

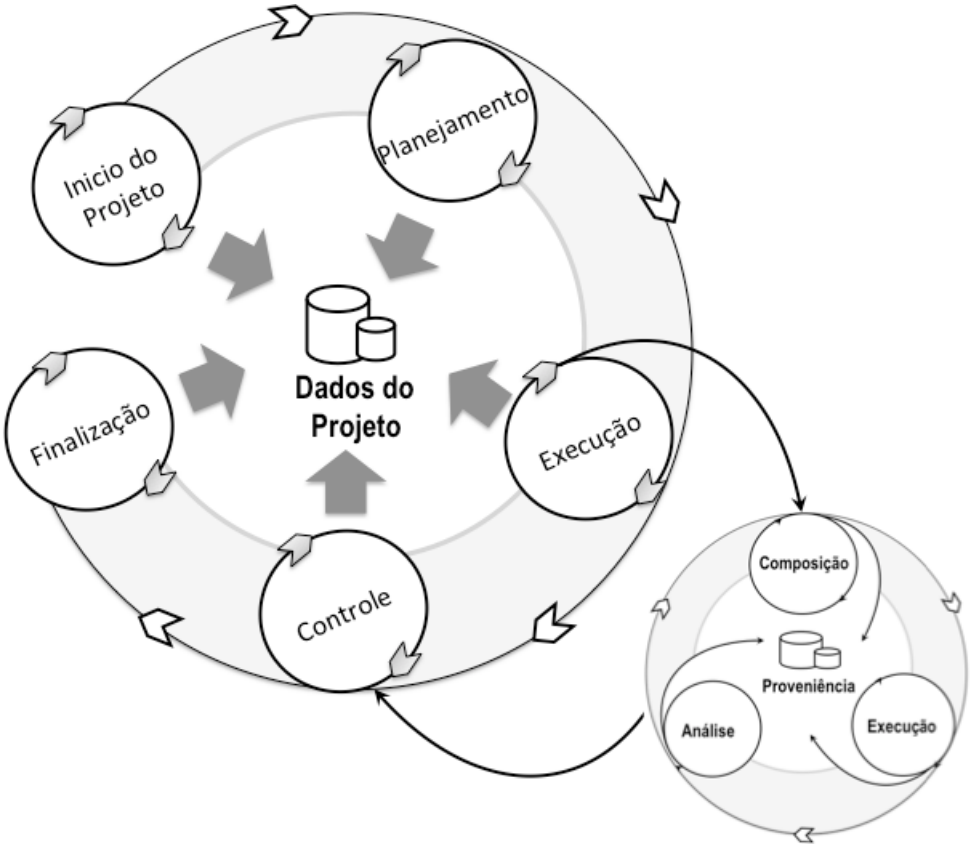


Figura 2.3: Ciclo de vida de um projeto científico.

nível do projeto. Como resultado desta fase é produzido um documento que representa um termo de abertura do projeto. O termo de abertura contém explicitamente os objetivos, os componentes da equipe, verba disponível, expectativas de HH (homem-hora) do projeto, etc. Até que o termo de abertura do projeto se encontre devidamente finalizado, várias interações podem se fazer necessárias.

A segunda fase, chamada de "Planejamento" do projeto, discute os requisitos de alto nível levantados na fase inicial a fim de se criar requisitos mais específicos que guiam a criação de tarefas. Esses requisitos dão origem a uma ou mais **hipóteses científicas**. Para confirmar ou refutar as hipóteses, uma ou mais tarefas são criadas e atribuídas aos membros do projeto. Por exemplo, em um projeto de bioinformática um requisito de alto nível é "Identificar fármacos para combater a Malária". Esse objetivo de alto nível pode ser detalhado em objetivos específicos como "Execução de Análise Filogenética" e "Estudo de Modelagem Molecular". Cada um desses objetivos mais específicos deve estar associado a um experimento e uma ou mais tarefas podem ser criadas associadas ao objetivo. Por exemplo, para a "Execução de Análise Filogenética" podemos criar tarefas intituladas "Criação do Esboço do *Workflow* de Análise Filogenética", "Seleção dos Dados de Entrada", "Estudo do Programa ModelGenerator", etc. Cada uma dessas tarefas está

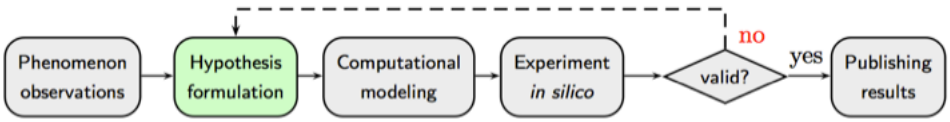


Figura 2.4: Ciclo de vida de exploração científica In-Silico [52]
.

associada a um experimento e a um *workflow* ou *script* e deve ser atribuída a um ou mais membros do projeto.

A terceira fase é a "Execução" do projeto. Nessa fase, as tarefas especificadas na fase anterior são efetivamente executadas; os *workflows* ou *scripts* são implementados, testados, executados, e os dados são analisados, a fim de gerar as conclusões relativas do experimento sendo executado.

A quarta fase é a de "Controle e Monitoramento". Nela as tarefas executadas na fase de execução são verificadas e auditadas pelos gerentes do projeto. Uma vez que as tarefas tenham sido executadas, o projeto passa para a quinta e última fase, a de "Finalização". Nessa fase é obtido um documento final que apresenta tanto os resultados gerados como algumas estatísticas do projeto. Entre uma fase e outra do projeto são realizadas reuniões entre seus membros, com o objetivo de verificar o andamento e/ou conclusão do mesmo. Essas reuniões determinam se o projeto deve continuar para sua próxima fase e ajudam os cientistas a detectar e/ou corrigir erros que afetem o andamento do projeto.

2.5 O Modelo de Dados de Hipóteses Científicas

Nesta seção, apresentamos o modelo de dados para representação de hipóteses científicas utilizado como inspiração nesta dissertação. Esse modelo é caracterizado por seu modelo conceitual e lógico conforme definido em [52]. O modelo conceitual interpreta o papel dos dados *in silico*, destacando formulação e validação de hipóteses científicas, de acordo com o apresentado na Figura 2.4.

Uma das aplicações do modelo conceitual é sua implementação como uma base de dados que inclui dados e metadados sobre o ciclo de vida científico da exploração [52]. As gravações ali obtidas podem ser utilizadas como: caderno do processo científico; fonte de informação de dados de proveniência [24]; material de apoio à reprodutibilidade e à compreensão dos resultados e análise dos resultados, apenas para citar alguns.

A Figura 2.5 apresenta o modelo conceitual de hipóteses científicas por meio de um

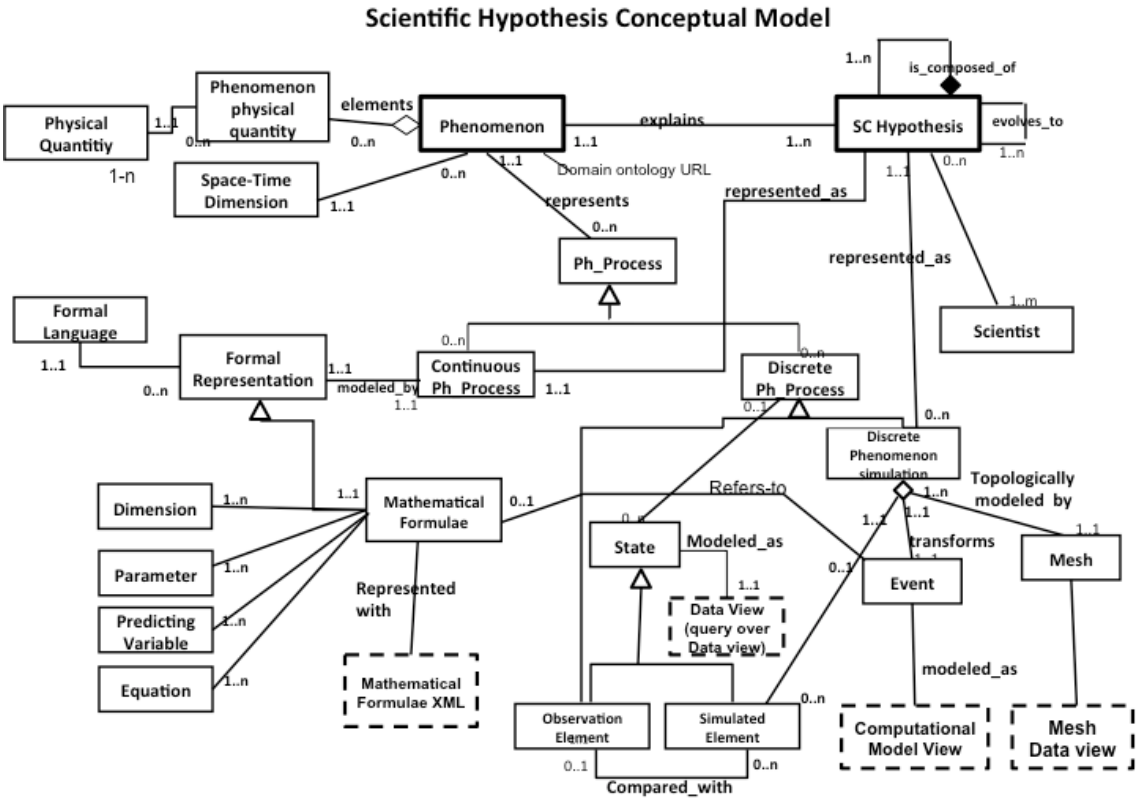


Figura 2.5: Modelo Conceitual de Hipótese Científica [52]

.

diagrama Entidade-Relacionamento [52]. Esse modelo reflete as entidades envolvidas durante um ciclo de vida de pesquisa científica *in silico*. O domínio é estruturado em torno dos seguintes conceitos principais: **Pesquisa, Fenômeno, Hipóteses, Processo de Hipóteses e Fenômenos** e **Processo abreviado**.

Cada **pesquisa** é distintamente identificada e oferece uma perspectiva sobre um progresso significativo importante na compreensão de um fenômeno particular. Ela engloba o fenômeno, foco da pesquisa e as hipóteses científicas concebidas para explicá-lo [52]. O ponto de partida de uma investigação científica é a especificação do fenômeno, como objeto da pesquisa [52].

Uma hipótese científica conceitualmente representa um modelo formal que fornece uma interpretação razoável para o fenômeno estudado. Uma hipótese é, possivelmente, formalmente expressa (isto é, quando existe conhecimento suficiente sobre o fenômeno estudado) e modelada em uma representação computacional. A formalização de um estudo científico pode ser fornecida por um modelo matemático. Na modelagem computacional, por exemplo, as Hipóteses podem modelar fenômenos contínuos naturais por meio de equações diferenciais, quantificando as variações de grandezas físicas no espaço-tempo.

A representação do modelo matemático é relevante como por ser a base para um mapeamento consistente de esquemas [29] entre sua representação matemática formal e sua representação de dados.

Como estamos interessados em modelar fenômenos naturais que ocorrem no espaço-tempo, nos inspiraremos na ontologia do Processo de Sowa [62], onde a especificação de uma hipótese científica é feita sob duas perspectivas, um processo contínuo e discreto. O primeiro refere-se ao modelo matemático, discutido anteriormente, representando o fenômeno estudado. O último corresponde à representação computacional da hipótese que induz transformações discretas do estado-fenômeno que levam à geração de dados de simulações.

Uma dada hipótese pode encontrar sua implementação em muitos modelos computacionais, usando diferentes métodos numéricos, por exemplo. Isso se reflete na cardinalidade da relação representada entre a Hipótese Científica e o Processo Discreto [52]. No entanto, está associado a um único modelo contínuo. Isso é explicado pelo fato que o modelo matemático é uma descrição formal e precisa da hipótese [52]. A questão das diferentes variações do formalismo matemático para a mesma representação conceitual é, no entanto, deixada a critério do cientista e uma investigação adicional está além do escopo desta dissertação.

É importante observar que o modelo computacional como apontado no diagrama da Figura 2.5 (*Computational Model View*) corresponde aos códigos de simulação que poderão executar em experimentos sobre o ambiente PAD. As Hipóteses devem ter estado (*validado*, *não validado*, *em validação*) e deveriam ser atualizadas a partir de condições sobre os resultados dos experimentos [52].

2.6 Ambientes de Processamento de Alto Desempenho

Conforme mencionado anteriormente, vários experimentos científicos necessitam de ambientes de PAD para executarem em tempo hábil. Entretanto, existem diversos tipos de ambiente de PAD que podem ser utilizados para esta execução e que são importantes para a presente dissertação, cada qual com vantagens e desvantagens associadas. Nesta seção discutimos os principais ambientes de PAD existentes.

Os *clusters* de computadores podem ser definidos como um conjunto de máquinas que, por meio de mensagens de comunicação, conseguem trabalhar como se fossem uma única máquina com alto poder de processamento e armazenamento. Uma característica

importante dos *clusters* é a homogeneidade de sua estrutura e a utilização de redes de alto desempenho com baixa latência (*i.e. infiniband*). As execuções de um *workflow* ou *script* em ambientes de *cluster* requerem a utilização de um escalonador [22], para que se consiga tirar alguma vantagem da estabilidade e das características de homogeneidade e do compartilhamento eficiente de recursos dos *clusters* no contexto do experimento.

As grades de computadores podem ser definidas como uma estrutura de malha que combina recursos distribuídos de *software* e *hardware* que são (em sua maioria) heterogêneos e fracamente acoplados. A utilização de um ambiente de grade normalmente depende de uma camada de *software* ou de um intermediário para gerenciar as execuções distribuídas. Um exemplo de intermediário é o Globus *Toolkit* [23]. A execução de *workflows* e *scripts* necessita também de um escalonador: porém, diferentemente dos *clusters*, esse escalonador necessita se preocupar com latência na transferência de dados, autenticações e questões de segurança [66, 72, 73].

Os ambientes de computação voluntária se baseiam na utilização de máquinas geograficamente dispersas, porém não dedicadas a uma determinada tarefa. A ideia principal é que exista um servidor central que controle a execução de pequenas tarefas em máquinas de terceiros, onde o tempo inativo das máquinas é disponibilizado para execuções distribuídas. Em termos de execuções de *workflows* e *scripts* nestes ambientes, podemos gerar facilmente milhões de tarefas uma vez que a quantidade de recursos disponível é consideravelmente maior do que em *clusters* e grades. O servidor central envia então, aos voluntários, pequenas tarefas de forma que os workflows possam ser processados de forma mais rápida do que seriam em supercomputadores (em teoria).

Mais recentemente o conceito de nuvem de computadores surgiu como um ambiente de PAD promissor, fato este que motivou o desenvolvimento desta dissertação. Estes ambientes são caracterizados pela diversidade dos recursos e pelo acesso através de uma camada de virtualização (máquinas virtuais). A grande vantagem das nuvens é que o usuário (no contexto desta dissertação, o cientista) tem um ambiente de alta capacidade de processamento, onde ele tem maior controle das ações, elasticidade dos recursos e, além disso, só necessita pagar pelo que é efetivamente usado, diferentemente dos *clusters* e grades onde é necessário um investimento inicial grande, seja monetário ou de configuração e manutenção.

Cada ambiente citado nesta seção pode ser implementado de diversas maneiras e seguindo diferentes arquiteturas para organização dos computadores e acesso de dados. No contexto desta dissertação, a arquitetura das máquinas não produz impactos severos, po-

rém a organização do acesso aos dados sim. As arquiteturas para acesso aos dados podem ser divididas em arquiteturas de discos compartilhados (do inglês *shared disk*) e arquiteturas de discos não compartilhados (do inglês *shared-nothing*) [65]. As arquiteturas de discos não compartilhados dependem exclusivamente dos discos locais em cada máquina. Já que nas arquiteturas de discos compartilhados todas as máquinas têm acesso aos discos, deixando a distribuição da execução mais flexível. Como veremos no decorrer da dissertação, optamos por utilizar uma arquitetura de discos compartilhados para avaliar a abordagem proposta.

2.7 *Research Objects*

A fim de facilitar o reuso, reprodução, unicidade e avaliação dos componentes envolvidos em um modelo computacional, foi proposta a entidade de documentação *Research Object*, *ROs* [5]. Através das *tags* semânticas, é possível descrever e documentar os dados de proveniência, tanto prospectiva quanto retrospectiva. Mais do que isso, o documento proposto visa descrever um determinado estado do modelo computacional a nível de publicação, e sendo assim, o documento contém dados como: descrição, resumo, cietístas envolvidos no estudo, dados sobre o ambiente computacional, além dos artefatos envolvidos, parâmetros utilizados e dos dados de proveniência [5].

Na computação, *Research Object* é um método para a identificação, agregação e troca de informações acadêmicas na *Web*. O objetivo principal da abordagem do *Research Object* é fornecer um mecanismo para associar recursos relacionados a uma investigação científica para que eles possam ser compartilhados usando um único identificador. Como tal, os objetos de pesquisa são uma forma avançada de publicação aprimorada. [7]

As implementações atuais se baseiam nas tecnologias e métodos da *Web* existentes, incluindo dados vinculados, *HTTP*, *URIs*, identificadores uniformes de recursos (*URIs*), Reutilização e Troca de Objetos da *Open Archives Initiative* (OAI-ORE) e o modelo *Open Annotation*, além de abordagens existentes para identificação e conhecimento representação no domínio científico, incluindo Identificadores de Objetos Digitais para documentos, identificadores ORCID para pessoas e o modelo de dados de Investigação, Estudo e Ensaio (ISA) [7].

A abordagem dos *Research Objects* é motivada principalmente pelo desejo de melhorar a reprodutibilidade das investigações científicas. O ponto central da proposta é a necessidade de compartilhar artefatos de pesquisa comumente distribuídos em repositó-

rios especializados na *Web*, incluindo dados de suporte, executáveis de software, código fonte, slides de apresentação e vídeos de apresentação. Os *Research Objects* não são uma tecnologia específica, mas são guiados por um conjunto de princípios. *Research Object* específicos são guiados por três princípios de identidade, agregação e anotação [6].

A Identidade digital diz respeito ao uso de identificadores exclusivos como nomes para itens, como DOIs para publicações ou dados e IDs do ORCID para pesquisadores. No contexto de agregação de dados é aconselhado que se use alguma forma de associação de itens relacionados que fazem parte de um estudo mais amplo e investigação para que outros possam descobrir mais facilmente esses recursos relacionados. Por fim, a anotação diz respeito a que se forneça metadados adicionais sobre os dados de estudo e como elas se relacionam, sua procedência e como foram produzidas [6].

Várias comunidades estão desenvolvendo ativamente o conceito de objeto de pesquisa, como por exemplo ROSC W3C. Um grupo comunitário do W3C intitulado Grupo de Comunidade de *Research Objects* para Comunicação Acadêmica (ROSC) foi iniciado em abril de 2013. O estatuto da comunidade declara que os objetivos da atividade do ROSC são: [3] "trocar requisitos e expectativas para apoiar uma nova forma de comunicação acadêmica". O grupo comunitário tem como objetivo produzir os seguintes tipos de entregas: casos de uso para representação, publicação e troca de *Research Objects* na *Web*; requisitos e anseios destilados dos casos de uso; uma pesquisa de trabalhos relacionados ao suporte à representação, publicação e troca de *Research Objects*; Melhores práticas e diretrizes para uma prática comunitária de compartilhamento, citação e intercâmbio de *Research Objects* [75].

Capítulo 3

Abordagem Proposta: `PhenoManager`

Neste capítulo, será apresentada a abordagem proposta para gerência de hipóteses científicas, o `PhenoManager` . O `PhenoManager` foi implementado como um sistema de código aberto e pode ser obtido diretamente no repositório do projeto no GitHub [1].

O sistema proposto tem como objetivo catalogar e definir hipóteses científicas, hierarquicamente, dentro do contexto de qual projeto e de qual fenômeno essa hipótese pertence e descreve. Além do mais, no que diz respeito à validação dos experimentos da hipótese, é possível, por meio de interface gráfica *Web*, criar *workflows* e *scripts*, sendo possível configurar sua execução, sua coleta de dados de proveniência e, por fim, ter um maior controle sobre os ambientes de execução dos modelos computacionais de estudo da hipótese. A fim de proporcionar compartilhamento e uma melhor troca de informação sobre os dados obtidos na ferramenta, foi desenvolvida a funcionalidade de exportação de *Research Objects* das execuções gerenciadas no sistema.

3.1 Modelo de dados

O modelo Entidade-Relacionamento das entidades fortes da ferramenta pode ser observado na Figura 3.1. Neste diagrama, temos os relacionamentos hierárquicos entre, projeto, que contém fenômenos, que por sua vez tem as definições de hipóteses. Hipóteses tem um estado atrelado . Abaixo do nível de abstração da hipótese, temos os experimentos que corroboram, ou não, com a hipótese levantada. Experimentos podem conter as fases de seu ciclo de vida, como definidas na Seção 2.3. No último nível, temos os modelos computacionais dos experimentos, que são a definição concreta dos experimentos

[1] https://github.com/UFFeScience/Phenomanager

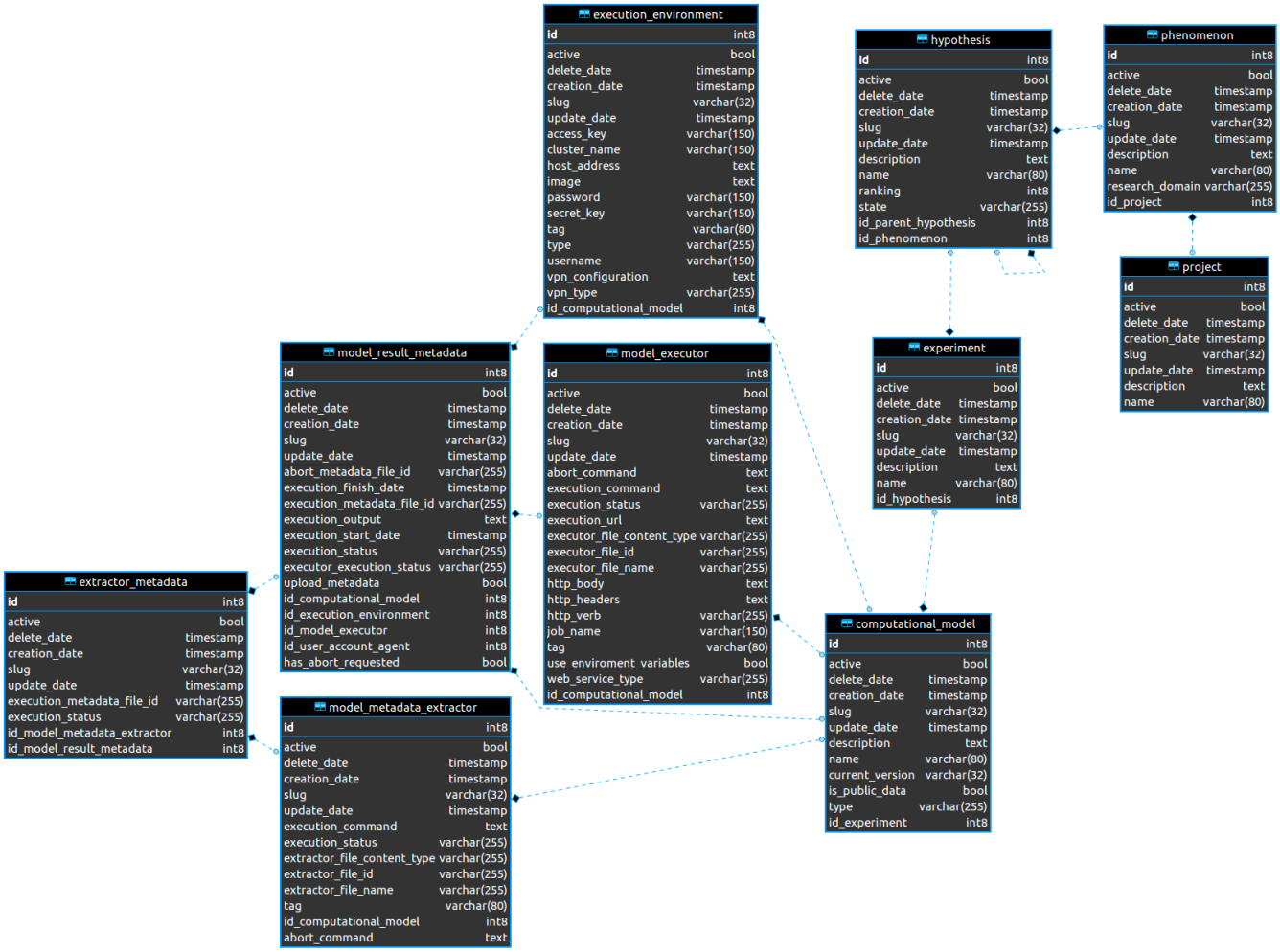


Figura 3.1: Diagrama ER do `PhenoManager` .

das hipóteses. Cada modelo deverá conter um ou mais executores e poderá conter diversos extratores de dados. Cada modelo deverá ter um ou mais ambientes computacionais, podendo este ambiente ser um ambiente de Cluster, *Cloud*, ou um ambiente com conexão *SSH* simples. Os modelos podem ser do tipo *WebService*, *Executable*(*script*) ou *workflow*..

3.2 Arquitetura do Sistema

A figura 3.2 apresenta a arquitetura do PhenoManager e seus componentes principais. O `PhenoManager` pode ser dividido em seis camadas funcionais: (i) Camada de Autenticação, (ii) Camada de Gerência do Ambiente, (iii) Camada de Execução, (iv) Camada de Dados, (v) Camada de Consulta, e (vi) Portal Web. A seguir, será discutido em detalhes cada uma dessas camadas.

A Camada de Autenticação é a responsável por gerenciar as credenciais de acesso ao

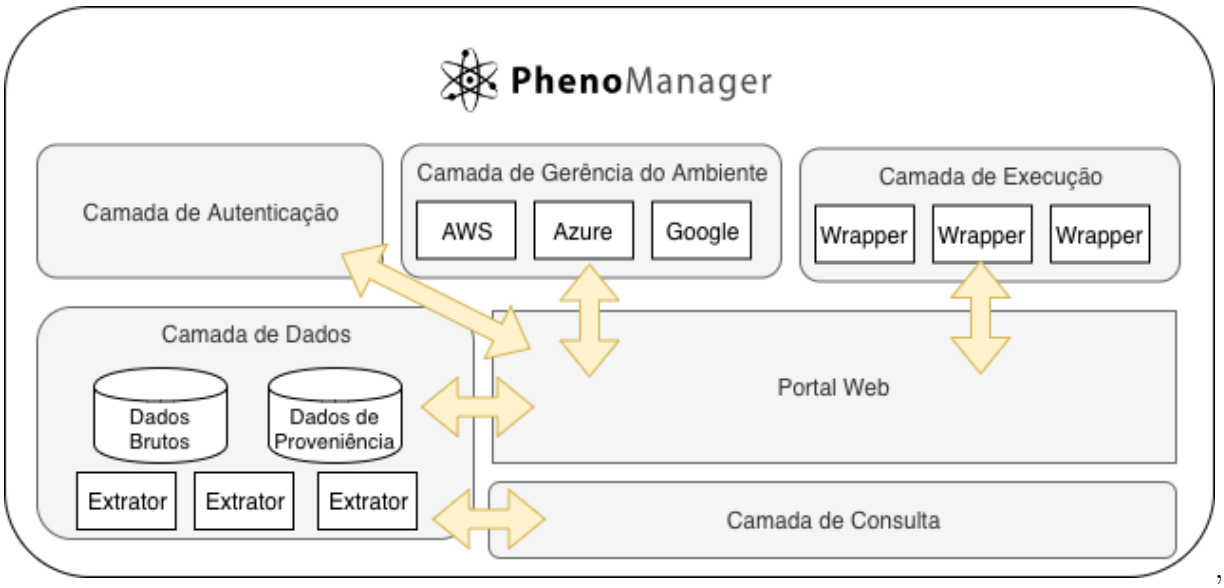


Figura 3.2: Arquitetura do `PhenoManager`

`PhenoManager` . Essa camada é fundamental, uma vez que dados de pesquisas não publicados serão manipulados pelo sistema. O controle de credenciais é realizado em dois níveis. No nível do "Perfil Pessoal", cada usuário da ferramenta configura/preenche suas informações pessoais e carrega credenciais para acesso aos diversos ambientes (e.g., Amazon AWS). No nível de "Grupos de Usuários", o(s) usuário(s) administrador(es) (com privilégios para criar grupos) procuram, selecionam e agrupam perfis de usuários existentes, e, a partir da criação do grupo, compartilham os mesmos privilégios no ambiente. Cada usuário ou grupo poderá ter acesso aos dados e funcionalidades providos pelo sistema de acordo com os seguintes privilégios: (i) Permissão para leitura (*"READ"*): o usuário e/ou grupo pode apenas visualizar os dados, porém não pode editar e/ou cadastrar qualquer dado; (ii) Permissão para escrita (*"WRITE"*): o usuário e/ou grupo pode ler e cadastrar informações dentro do contexto especificado, tornando-o membro ativo do sistema; (iii) Permissão de administrador (*"ADMIN"*): além das permissões anteriormente citadas, o usuário e/ou grupo pode cadastrar permissões para outros usuários e grupos;

A Camada de Gerência do Ambiente é a responsável por configurar ambientes distribuídos para a execução das simulações. Para cada ambiente diferente, um componente deve ser desenvolvido com as chamadas para a *API* específica do mesmo. O `PhenoManager` já provê nativamente a integração com três tipos de ambientes diferentes: *Cluster*, *Cloud* (*Amazon AWS*) e *SSH*. Além disso, é possível configurar uma conexão *VPN* para estes ambientes, podendo selecionar entre *Cisco VPN* e *VPN default*. Para o ambiente *Cloud*, é possível configurar no detalhe, os tipos e imagens das máquinas virtuais que serão construídas no ambiente da *Amazon AWS*.

A Camada de Execução é a responsável por invocar SGWfs, *scripts* ou aplicações externas nos ambientes de alto desempenho. Para cada sistema diferente ou aplicação a ser invocada, um *wrapper* específico deve ser provido (já que o `PhenoManager` necessita conhecer o processo de invocação da aplicação externa). A chamada aos *wrappers* é assíncrona, logo o serviço dessa camada pode ser escalado em mais instâncias, aumentando, dessa forma, o *throughput* de execuções paralelas de simulações científicos para diferentes usuários. Sendo assim, por meio desse artifício, procuramos garantir o paralelismo e a alta disponibilidade do `PhenoManager` .

A Camada de Dados contém o banco de dados de proveniência (com todos os metadados registrados na ferramenta) e os dados brutos produzidos pelas simulações computacionais. Além disso, essa camada contém uma série de extratores responsáveis por acessar a base de proveniência ou o log da aplicação externa e carregar as informações no banco de dados de proveniência do `PhenoManager` . Em sua versão atual, o banco de dados de proveniência se encontra modelado no PostgreSQL e os dados brutos são carregados, opcionalmente, no Google Drive.

A Camada de Consulta é a responsável por permitir que o cientista possa submeter consultas aos dados gerenciados pelo `PhenoManager` . Essa camada provê uma *API* que abstrai as consultas aos dados de modo a facilitar sua manipulação por cientistas e outras aplicações consumidoras deste serviço. A *API* permite que o cientista submeta consultas contendo filtros, ordenações, funções de agregação e projeções de campos em todas as entidades expostas no modelo de dados do `PhenoManager` . Outro ponto importante é que a *API* só responde com sucesso se as credenciais corretas forem passadas no cabeçalho da solicitação. Além disso, a Camada de Consulta é responsável por exportar pacotes *Research Objects* (RO), que contém tanto os metadados consultados quanto os dados brutos produzidos pela simulação. Por meio dos ROs, os cientistas são capazes de reproduzir uma determinada simulação, um experimento ou verificar se uma hipótese foi efetivamente validada.

Finalmente, o Portal *Web* é o responsável por toda a interface com o cientista e a integração com as demais camadas. Nele o cientista registra os fenômenos observados, as hipóteses associadas, seus experimentos e os modelos que executam as simulações de cada experimento (e.g., workflow, *script* ou aplicação). Além disso, por meio do Portal *Web*, o cientista é capaz de executar efetivamente suas simulações e consultar os dados de proveniência coletados de forma integrada, isto é, se um mesmo experimento for composto de diversos workflows e aplicações, as consultas à base de proveniência consideram todas as

simulações como parte do mesmo experimento, o que não ocorre nas ferramentas existentes que gerenciam os modelos computacionais de maneira isolada [20].

Em termos de implementação, o `PhenoManager` foi desenvolvido na linguagem Java e segue o padrão arquitetural de APIs como microsserviços, ou seja, cada componente é um serviço *Web* autônomo e pequeno que disponibiliza apenas uma funcionalidade [46]. Todos os microsserviços foram construídos por meio do *framework Spring Boot*, que já oferece apoio para desenvolvimento de aplicações nesse padrão de uma maneira rápida e pouco verbosa. Para segurança de dados e autenticação entre os componentes, foi utilizado o arcabouço *Spring Security*. O desenvolvimento das interfaces, templates e telas do Portal *Web* foi realizado com AngularJs. Como cada serviço que compõe a arquitetura do `PhenoManager` é completamente isolado dos demais, a escalabilidade se torna um dos pontos chave deste ecossistema. Para garantir a invocação de aplicações externas de forma assíncrona, foi escolhido o *message Broker* de código aberto RabbitMQ.

3.2.1 O *SciManager*

No contexto deste trabalho, um outro sistema importante a ser explorado, e que faz parte do mesmo ecossistema do `PhenoManager` , é o *SciManager*. A aplicação *SciManager* é uma ferramenta que auxilia o cientista na gerência do projeto científico como um todo e não só particularmente na execução de um determinado experimento ou workflow (para tanto já existem os SGWfCs e agora o `PhenoManager` , que tem objetivo similar). O objetivo do *SciManager* é ser sistema que visa tirar proveito de todo o arcabouço de gerência de projeto já existente na Engenharia de Software e em outras áreas. Especificamente para a gerência de projetos científicos, ainda existia uma escassez de abordagens que gerenciassem esse tipo de projeto baseado em técnicas de gerência de projetos, o que representava um desafio em aberto. Desta forma, o *SciManager* foi desenvolvido visando preencher esta lacuna e apoiar a gerência de projetos científicos em larga escala com a colaboração entre equipes geograficamente dispersas [53]. Esta ferramenta é baseada em técnicas e ferramentas existentes de gerência de projetos e possui uma arquitetura baseada no conceito de computação em nuvem, similar ao `PhenoManager` .

No contexto da gerência de projetos, o *SciManager* provê como funcionalidades o controle das atividades aferidas a um projeto, disponibilizando de gráficos de acompanhamentos, quadros de tarefas e de maneira bem mais simplificada, o acompanhamento do status e da execução de Workflows de Experimentos Científicos. Diferente do `PhenoManager` , o *SciManager* apenas provê execução de *workflows* científicos e suas execuções são feitas

por meio de integração com o *SciCumulus* [18].

O `PhenoManager` não provê acompanhamento de tarefas e nem contém um quadro de tarefas e nem gráficos de acompanhamento de metas por usuários dos projetos, pois o enfoque se dá no ciclo de vida da hipótese e não na gerência do projeto em si. Por esse motivo, foi implementada a integração de ambos os sitemas, de modo a termos gerência do projeto cietífico, a gerência do ciclo das hipótese e dos dados dos modelos computacionais de maneira mais eficiente e uma interface de execução de *workflows* e scripts bem mais robusta e genérica, sem as limitações do *SciManager* nesse aspecto.

A integração do `PhenoManager` com o *SciManager* é feita de forma natural, e os projetos criados no `PhenoManager` podem ser portados ao *SciManager*, criando uma integração entre ambos os dados. Os dados se relacionam por meio de um UUID, um código serial e único, que é compartilhado entre as bases de dados de ambas as aplicações. Os usuários do *SciManager* podem também ser portados para o `PhenoManager` da mesma forma.

3.3 Funcionalidades

Ao acessar o sistema, a primeira tarefa que o usuário deve realizar é se autenticar pela tela de *login*, conforme apresentado na Figura 3.3. Após efetuar a autenticação, o usuário terá acesso ao *dashboard* do `PhenoManager` , que apresenta todas as execuções de simulações em andamento, finalizadas, e simulações com erro, para controle do cientista. Além disso, são apresentados atalhos para os modelos computacionais e projetos de mais uso pelo cientista.

3.3.1 Gerência de usuários e Grupos de usuários

O `PhenoManager` foi projetado para atender projetos científicos com múltiplos usuários geograficamente dispersos. Assim, se faz necessário um controle desses usuários na ferramenta. Esse controle de usuários é realizado em dois níveis. No nível do "Perfil Pessoal", cada usuário da ferramenta configura/preenche suas informações pessoais. Em seu perfil, o usuário define informações básicas como nome, e-mail de contato, instituição e, opcionalmente, pode realizar o upload de uma foto.

No nível de "Grupos de Usuários" (menu "*Teams*"), o(s) usuário(s) administrador(es) (com privilégios para criar grupos) procuram, selecionam e agrupam perfis de usuários existentes, e, a partir da criação do grupo, irão partilhar os mesmos privilégios na ferra-

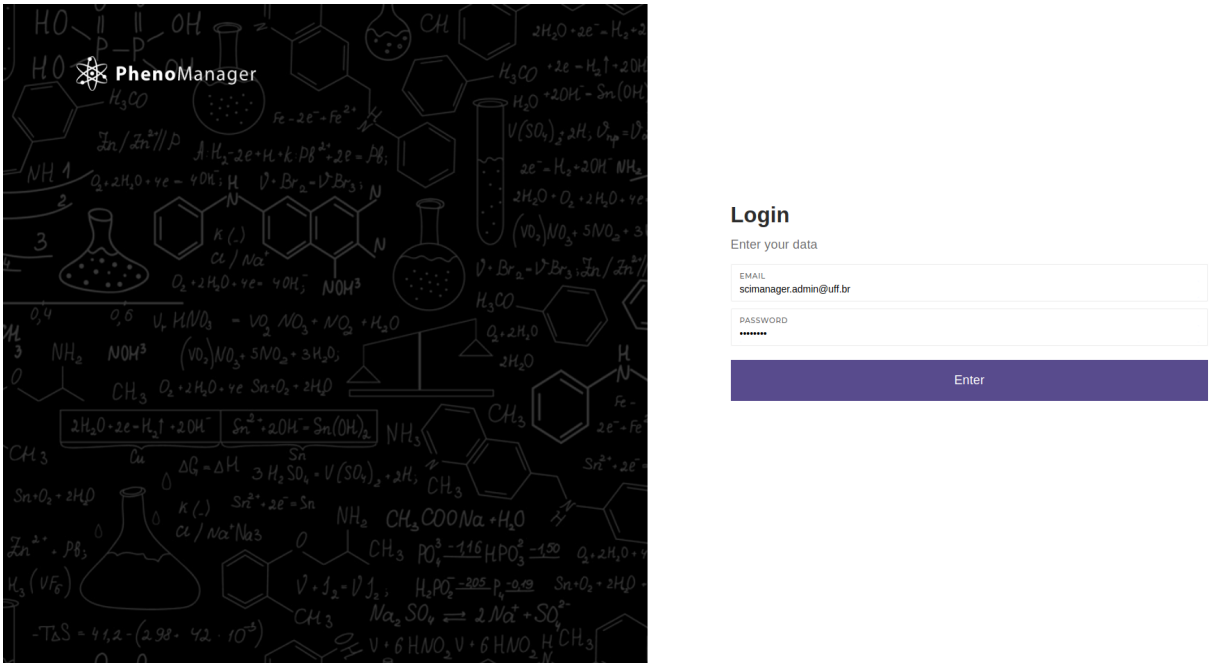


Figura 3.3: Tela de *login*.

menta.

Para garantir que cada componente do sistema seja visualizado e manuseado apenas por quem possui as devidas credenciais, foi implementado um sistema de cadastro de permissão para usuários e grupos do sistema. Cada usuário ou grupo poderá ter acesso aos dados e funcionalidades providos pelo sistema de acordo com os seguintes privilégios: "*READ*", "*WRITE*"e "*ADMIN*", conforme já explicado em sessão anterior. Todas as entidades do sistema requerem permissão para serem acessadas, e a *API* do `PhenoManager` retorna *HTTP Status 401* caso haja violação de permissionamento.

3.3.2 Gerência de Projetos, Fenômenos, Hipóteses e Experimentos Científicos

Após definir os perfis de acesso e grupos de usuários, as funcionalidades-chave da ferramenta são habilitadas. Assim, a primeira tarefa que o usuário pode realizar é cadastrar um projeto científico no sistema que é a unidade de trabalho de mais alto nível.

Um projeto possui um nome, uma descrição/documentação e, fenômenos associados ao domínio tratado. Nele, os administradores, e usuários com permissão de escrita, além de registrar o projeto no sistema, podem editar, bem como podem criar/editar os fenômenos que o mesmo se propõe a estudar (dependendo do nível de permissão que o usuário tiver). Já os cientistas membros, com permissão apenas de leitura têm a capacidade apenas de

visualizar as informações do projeto.

Após realizada a configuração do projeto, resta ao cientista configurar os fenômenos e as hipóteses da qual esse fenômeno descrito remete. Assim como o projeto científico, o fenômeno e a hipótese têm um nome e uma descrição.

No caso da hipótese, a mesma pode conter inúmeras outras hipóteses filhas que descrevem derivações de um estudo. De acordo com os resultados obtidos durante a observação das execuções dos experimentos de dada hipótese, o cientista pode mudar o estado das mesmas. Uma hipótese pode assumir os seguintes estados no sistema: *Formulated*: uma hipótese recém criada; *Validated*: uma hipótese validada por experimentos; *Confirmed*: hipótese mostrou-se verdadeira, porém, carece de validação; *Improved*: uma hipótese que teve uma melhoria na sua formulação; *Refuted*: uma hipótese que foi refutada;

A próxima etapa da linha de configuração do sistema é cadastrar os experimentos científicos que farão o papel de validar ou não a hipótese. O experimento científico, assim como os dados anteriores, contém nome, descrição, e também os modelos computacionais que são a representação concreta do experimento. Além disso, como ele é uma modelagem conceitual de um modelo de execução, o mesmo poderá ter uma diretriz dos parâmetros que o modelo computacional, por sua vez, usa para a sua execução. Também é possível criar e configurar as fases do ciclo de vida do experimento científico, conforme visto na Seção 2.3.

Além disso, é possível criar pontos de verificação para a validação de um experimento. Itens de validação são entidades que têm como objetivo determinar uma diretriz de como um experimento poderá ser validado. Um mesmo experimento pode conter inúmeros itens de validação e ao selecionar um item como validado, o cientista tem a oportunidade de realizar o *upload* de arquivos que reforcem a evidência que esse ponto de verificação foi de fato validado. Pontos de verificação auxiliam e servem para balizar o cientista sobre o estado em que uma hipótese se encontra.

3.3.3 Criação de modelos computacionais

A criação e configuração de um modelo de execução dentro do `PhenoManager` deve seguir as seguintes etapas:

i) Cadastro do Ambiente computacional em que o modelo irá executar (*SSH*, *Cluster*, *Amazon*), conforme Figura 3.4;

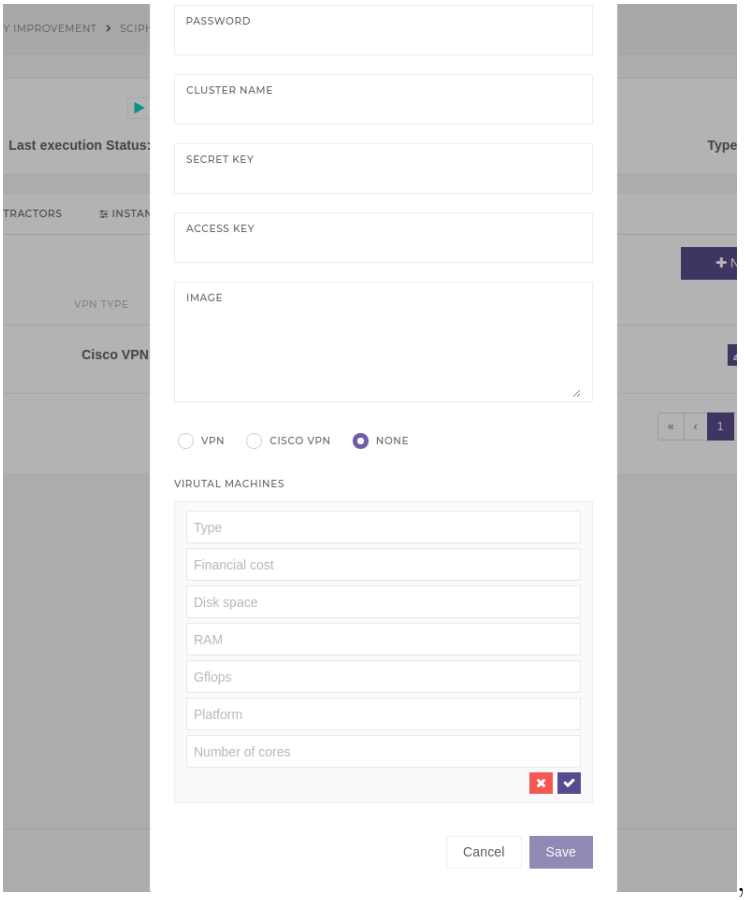


Figura 3.4: Demonstração de criação de um ambiante *Cloud* com configuração específica para cada VM.

ii) Cadastro do artefato/conector de execução, que é o componente que será executado no ambiente configurado na etapa anterior;

iii) Cadastro dos parâmetros de instância, que simbolizam as entradas usadas pelo artefato de execução;

iv) Cadastro do(s) extrator(es) de dados da execução, responsável por realizar a extração dos dados de proveniência gerados pela execução (etapa opcional). É possível ter muitos extratores diferentes ativos e também é possível não ter nenhum ativo;

Na etapa de configuração do ambiente de execução, o cientista tem a opção de configurar integração com três tipos de ambiente diferente: *Cluster*, *Cloud (Amazon AWS)* e *SSH*. Além disso, é possível configurar conexão *VPN* para estes ambientes, podendo selecionar entre *Cisco VPN* e *VPN default*. Para o ambiente *Cloud*, é possível configurar no detalhe, os tipos e imagens das máquinas virtuais que serão construídas no ambiente da *Amazon AWS* (Figura 3.4). Para o tipo *Cluster*, o usuário tem como especificar os recursos que serão alocados e utilizados dentro do arquivo *batch* do *job*.

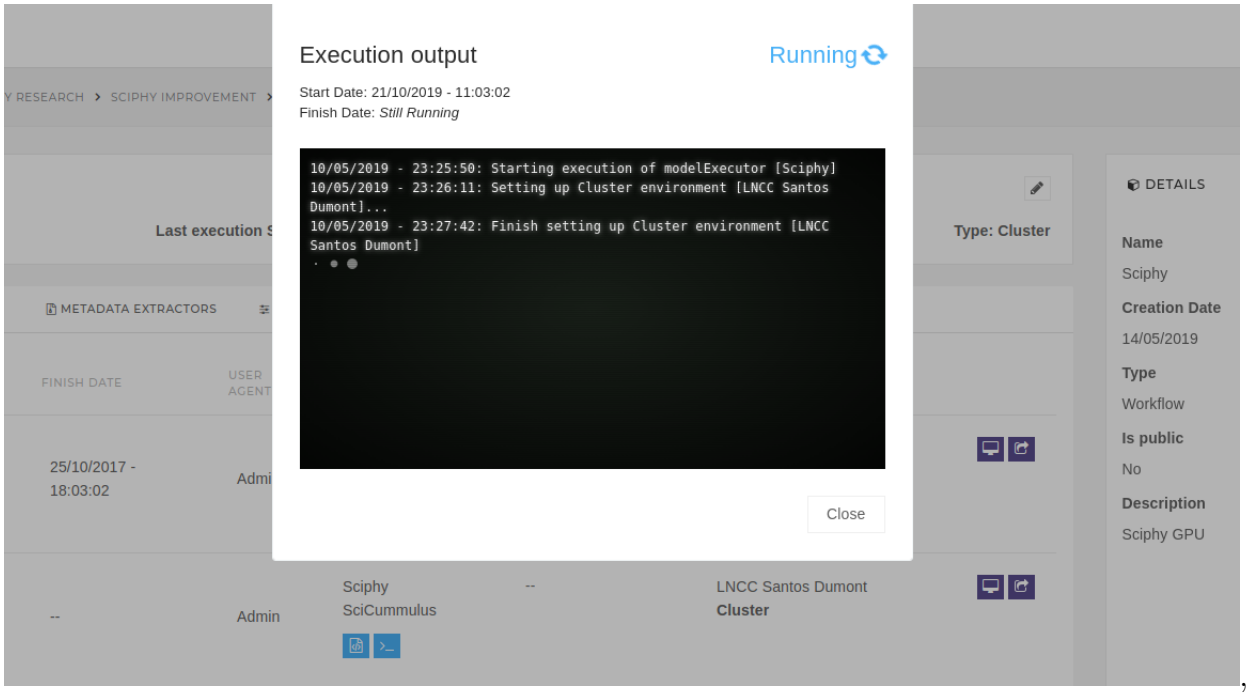


Figura 3.5: *Logs* de execuções de um modelo computacional exibidos em tempo real.

Durante a configuração do executor, podemos escolher três tipos de Executor: "*WebService*", "*Workflow*", "*Command*" e "*Executable*". Um executor *WebService* pode realizar chamadas *REST* e *SOAP*, por meio de qualquer verbo *HTTP*. Para o tipo "*Workflow*", é esperado um arquivo de formato *.zip*, com os programas e com o SGWfC responsáveis pela chamada do *Workflow*. Já no tipo *Executable*, não é esperado um formato *.zip* do executável do programa. Por fim, o tipo "*Command*", espera um texto com a linha de comando que será executada no ambiente configurado.

Após a configuração do modelo ser plenamente realizada, já é possível iniciar a execução do modelo (ou parar a mesma, caso o cientista deseje) e o serviço de execução assíncrono se encarregará de orquestrar a execução. Os dados de proveniência e *logs* da execução são exibidos em tempo real conforme a execução dá seguimento, como podemos observar na Figura 3.5. Na Figura 3.6, vemos o histórico de execuções de um dado modelo computacional.

3.3.3.1 Fluxo de processamento de mensagem de execução do modelo

Nessa seção, será explicada de maneira mais detahada o fluxo de processamento assíncrono para consumir uma mensagem de execução de um determinado modelo computacional.

A Figura 3.7, simboliza o diagrama de sequência do fluxo de execução assíncrono de

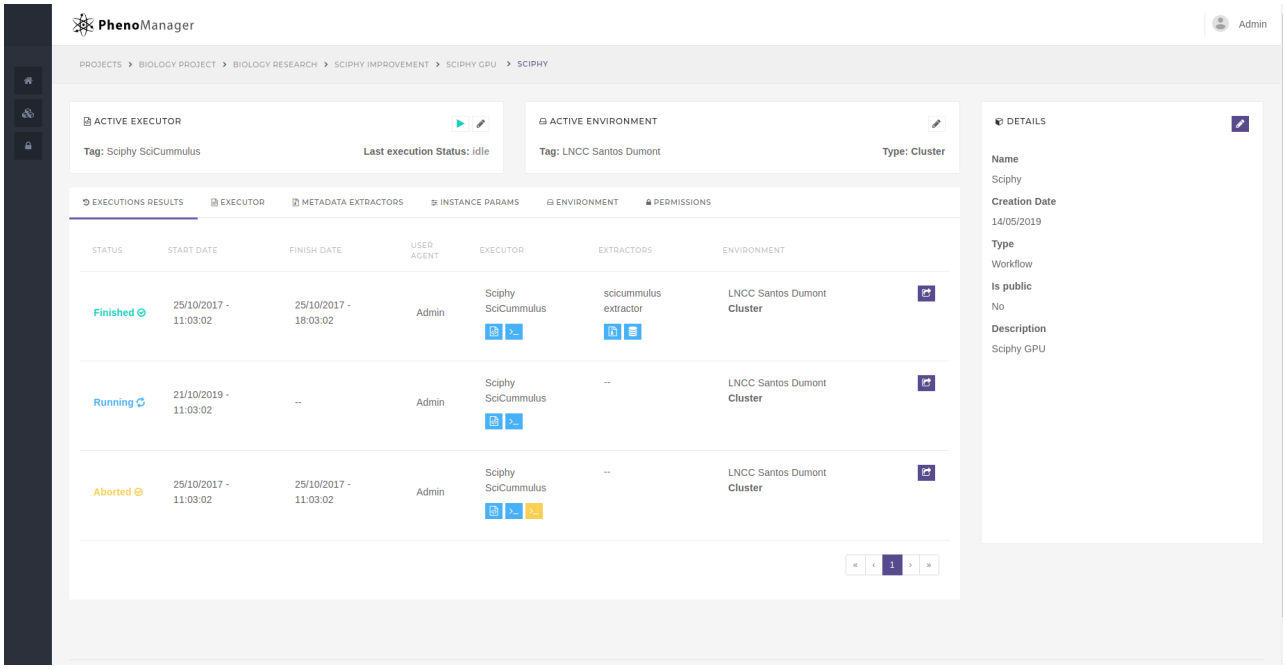


Figura 3.6: Histórico de execuções de um modelo computacional.

um modelo computacional devidamente configurado. A execução em si do modelo se dá a partir do consumo de uma mensagem, pelo serviço consumidor. Para cada um dos quatro possíveis diferentes cenários de ações que podem ocorrer de acordo (início de execução de um executor; cancelamento de execução de um executor; início de execução de um extrator, cancelamento da execução de um extrator) com o corpo da mensagem, há uma chave diferente de roteamento. Dito isto, cada ação diferente pode ter um consumidor diferente. mesmo que cada ação utilize a mesma fila. Dessa forma, podemos escalar mais instâncias de consumidores para as ações que mais demandam processamento e que necessitam de um *throughput* maior. Após receber a mensagem, o serviço consumidor segue os seguintes passos no *pipeline* do consumo da mensagem:

i) Recebimento da mensagem;

ii) Verifica status do executor/extrator (não pode cancelar uma execução que não está em andamento e nem iniciar de novo um processo que ainda está em andamento);

iii) É baixado do *Google Drive* o executor/extrator que foi solicitada a execução (caso seja um modelo executor que não seja do tipo *WebService*);

iv) É feita a conexão com o ambiente do modelo computacional e é estabelecida conexão VPN, se o mesmo estiver configurada para tal;

v) O executor/extrator é copiado para o ambiente cuja conexão foi estabelecida na etapa anterior;

vi) É feita a execução propriamente dita e os metadados de saída da execução são exibidos em tempo real no sistema e posteriormente é feito o upload para o repositório do *Google Drive*, se a execução foi configurada para tal, caso contrário o dado bruto da execução permanecerá apenas no ambiente de execução;

vii) Se a execução da etapa anterior foi realizada por um executor, todos os extratores de dados ativos do modelo serão executados em sequência, e é feito o *upload* das saídas de dados para o repositório do *Google Drive*, se a execução foi configurada para realizar essa operação de upload;

viii) Mensagem é marcada como processada;

3.3.4 Exportação de *Research Objects*

O `PhenoManager` permite que seja exportado um *JSON* com o *Research Object* referente a uma execução específica do modelo computacional de um experimento (como um *SNAPSHOT* de um estado do modelo)(Figura 3.8). O mesmo conterá os dados de entrada, os dados de saída, o programa que realizou a execução e os dados dos usuários envolvidos nesse estado do modelo. Um ponto importante é que se o modelo computacional estiver configurado como "público", literalmente qualquer pessoal com o *link* do *Research Object* poderá visualizar estes dados.

3.3.5 Integração com o *SciManager*

Para podermos reaproveitar todo o arcabouço de gerência de projetos que o *SciManager* provê, foi implementada a sincronização de projetos e experimentos do `PhenoManager` para com o *SciManager* e vice versa. Dessa forma, podemos utilizar a gama de execuções distintas que o `PhenoManager` se propõe, além do já consolidado controle de quadro de tarefas de projetos científicos que pode ser utilizado no *SciManager*. Além disso, usuários e grupos também podem ser portados, fazendo com que uma mesma credencial de acesso possa ser utilizada em ambos sistemas do mesmo ecossistema. Importante salientar que todos os dados disponibilizados em um projeto serão portados para o *SciManager*, desde os experimentos e fases, até os usuários e seus dados de acesso.

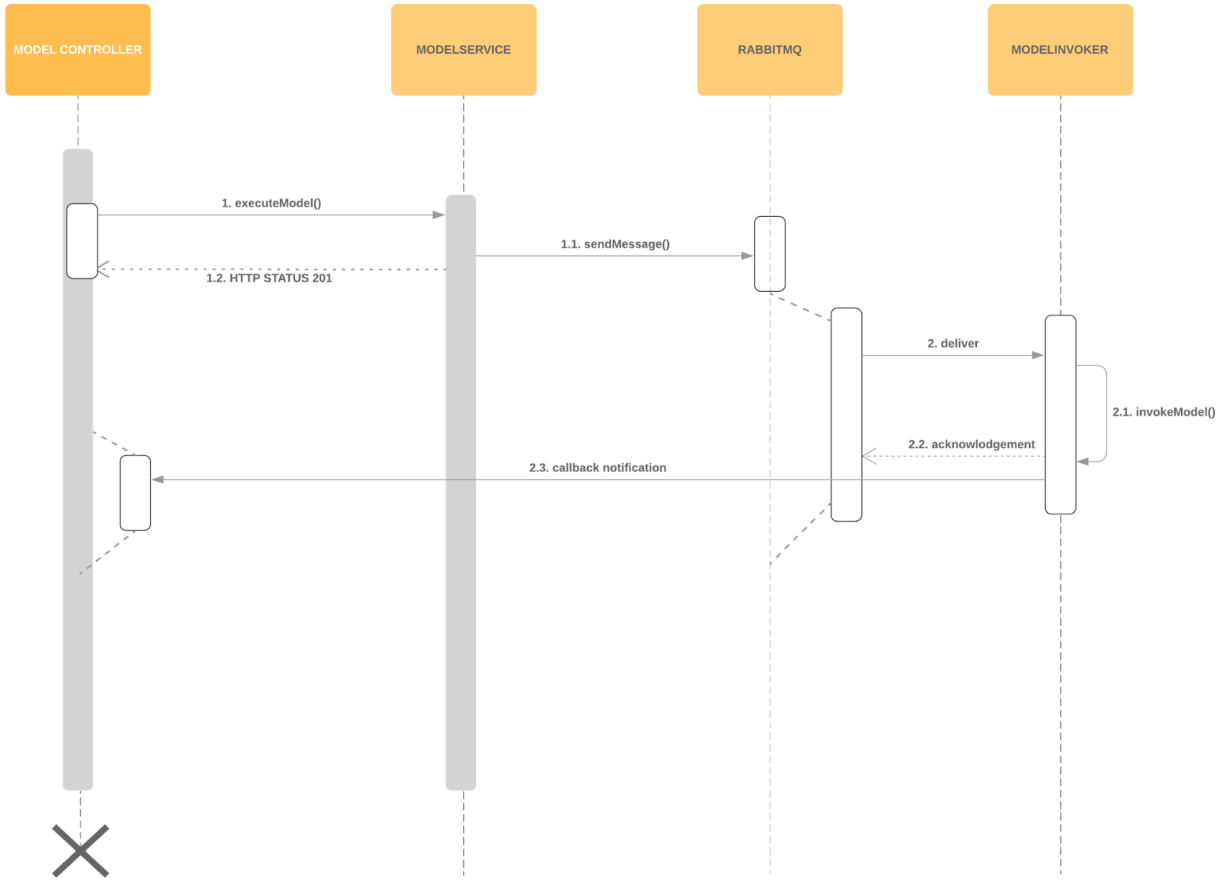


Figura 3.7: Diagrama de sequência de execução de modelo computacional.

```
{
    "@context":{
        "schema":"http://schema.org/",
        .
        .
        .
    },
    "@graph":[{
            "@type":[
                "ro:ResearchObject",
                "ore:Aggregation"
            ],
            "@id":"4B471432FCD146018593817458D6E21D"
        },{
            "schema:name":"Sciphy"
        },{
            "dc:creator":"QWE123987POEIWQPEWQ12687EWQEWQEF"
        },{
            "dc:abstract":"Sciphy GPU"
        },{
            "dc:contributor":["QWE123987POEIWQPEWQ12687EWQEWQEFP"]
        },{
            "dc:title":"Sciphy"
        },{
            "Ore:aggregates":[{
                    "@type":"ro:Resource",
"@id":"http://localhost:9500/PhenoManagerApi/v1/computational_models/4B471432FCD146018593817
458D6E21D/instance_params/3EE92B89774345BD9A8CA4DF77FB148A/value_file"
                },{
                    "@type":"ro:Resource",
"@id":"http://localhost:9500/PhenoManagerApi/v1/computational_models/4B471432FCD146018593817
458D6E21D/instance_params/E4A3F7035B0D45D29ABA0754E89FE8F5/value_file"
                },{
                    "@type":"ro:Resource",
"@id":"http://localhost:9500/PhenoManagerApi/v1/computational_models/4B471432FCD146018593817
458D6E21D/model_executors/4C6212B68E2C47929F509B88A037583B/executor"
                },{
                    "@type":"ro:Resource",
"@id":"http://localhost:9500/PhenoManagerApi/v1/computational_models/4B471432FCD146018593817
458D6E21D/model_result_metadatas/NDSJDKHAJKD789HDSJGAJD"
                }
            ]
        },{
            "prov:wasAttributedTo":["QWE123987POEIWQPEWQ12687EWQEWQEF"]
        },{
            "schema:contributor":["QWE123987POEIWQPEWQ12687EWQEWQEF"]
        },{
            "foaf:name":"Admin",
            "@@type":"foaf:Person",
            "@id":"QWE123987POEIWQPEWQ12687EWQEWQEF",
            "schema:name":"Admin"
        }
    ]
}
```


,

Figura 3.8: *Research Object de uma execução de um modelo computacional.*

Capítulo 4

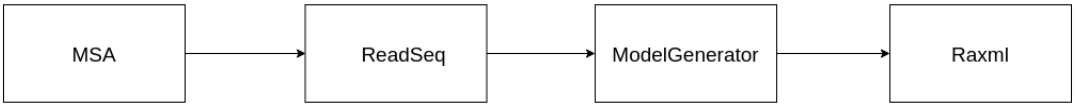
Avaliação Experimental

Este capítulo apresenta a avaliação experimental da abordagem proposta. Utilizamos um *Workflow* científico da área de bioinformática como estudo de caso para a avaliação do `PhenoManager` . Os experimentos foram separados em três partes. Na primeira, avaliamos o *overhead* do `PhenoManager` no momento da execução de um *workflow*. Na segunda, avaliamos o `PhenoManager` quanto ao seu uso, utilizando o *Technology Acceptance Model* (TAM) e na terceira, utilizamos a *API* do `PhenoManager` para consultar, extrair e manipular os dados obtidos durante o experimento. Assim, a ideia central deste capítulo é avaliar a eficiência e praticidade da solução implementada assim como suas limitações.

4.1 Estudo de Caso: o *Workflow* Sciphy

O Sciphy é um *workflow* científico que foi projetado para gerar árvores filogenéticas com máxima verossimilhança. Ele foi projetado inicialmente para trabalhar com sequências de aminoácidos, podendo ser estendido para outros tipos de sequências biológicas. O SciPhy é composto por quatro atividades, sendo elas:

- *MSA:* Constrói alinhamentos individuais. Ele recebe um arquivo multi-fasta contendo sequências de DNA e RNA como entrada, produzindo como saída um alinhamento (*MSA*). Neste ponto o SciPhy obtém o alinhamento individual;
- ReadSeq: Converte o alinhamento individual para o formato *PHYLIP*;
- *Model Generator:* Nesta atividade cada *MSA* é testado para encontrar o melhor modelo evolutivo;



,

Figura 4.1: O *Workflow* SciPhy.

- *RaXml:* Nesta atividade tanto o modelo evolutivo quanto o *MSA* são utilizados para gerar árvores filogenéticas para cada um dos programas de *MSA* que foram eleitos;

Como os cientistas não conhecem a *priori* qual o método de alinhamento que produz o melhor resultado final, eles precisam executar o SchiPhy várias vezes, uma para cada método MSA. Estas atividades, respectivamente, executam as seguintes aplicações de bioinformática: programas de alinhamento genético (permitindo ao cientista a escolher entre o *MAFFT*, o *Kalign*, o *ClustalW*, o *Muscle*, ou o *ProbCons*), o *ReadSeq* [26], o *ModelGenerator* [34], o *RAxM*L [63] e um script Perl. A Figura 4.1 apresenta a visão conceitual do SciPhy.

4.1.1 Ambiente de Execução para Análise de *Overhead*

O ambiente que escolhemos para realizar os teste foi o Supercomputador Santos Dumont no LNCC. O Santos Dumont possui capacidade instalada de processamento na ordem de 1,1 *Petaflop/s* (1,1 x 1015 float-point operations per second), apresentando uma configuração híbrida de nós computacionais, no que se refere à arquitetura de processamento paralelo disponível. Além disso, ele possui um total de 18.144 núcleos de CPU, distribuidos em 756 nós computacionais (24 núcleos por nó), dos quais são compostos, na sua maioria, exclusivamente por CPUs com arquitetura multi-core. Há, no entanto, quantidade adicional significativa de nós que, além das mesmas *CPUs multi-core*, contém tipos de dispositivos com a chamada arquitetura many-core: GPU e MIC. Ademais, ele também é dotado de um nó diferenciado, o *MESCA2*, com número elevado de núcleos (240) e arquitetura de memória compartilhada de grande capacidade (6 Tb em um único espaço de endereçamento).

Para nosso experimento, iremos utilizar dois cenários de execução com uma situação comparativa para cada cenário. O primeiro cenário irá mensurar a média dos tempos de dez execuções com as filas do *RabbitMq* vazias e apenas um serviço consumidor. O segundo cenário irá mensurar a média dos tempos de dez execuções com a fila de execuções do *RabbitMq* com mil mensagens paralelas e dez consumidores com dez *threads* cada.

Em cada um dos cenários, será comparado a média do tempo de execução do usuário cientista executando o *Workflow* manualmente utilizando o SGWfC *SciCumulus* [18] e depois comparando com o tempo de execução da mesma ação, utilizando-se do `PhenoManager` para encapsular o processo. A medição de tempo no `PhenoManager` se dará desde o momento que o cientista apertará o botão de "*Play*"na ferramenta `PhenoManager` , até a atualização do status de execução do executor para "Finalizada".

Para obtermos dados comparativos justos, a execução na plataforma `PhenoManager` se dará sem a extração dos dados, já que manualmente, o cientista teria que realizar essa etapa de maneira separada também. Além disso, a execução se dará sem que haja upload dos *logs* e metadados da execução pelo mesmo motivo. Para a execução manual, foi criado um *script.sh* que envelopa a execução do *SciCumulus* e realiza a cronometragem da execução do *Workflow*, desde seu início até a fim da execução de seu último programa.

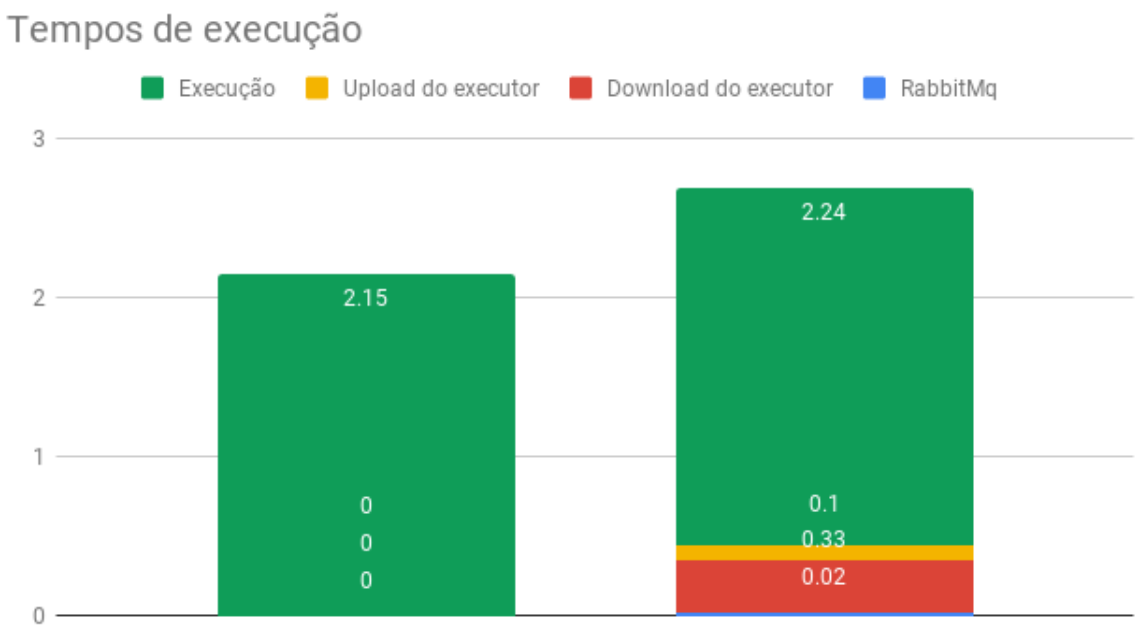
Além disso, para melhor entendermos os pontos em que o `PhenoManager` trará atraso no tempo de execução, foram medidas cada etapa do processo da execução:

- Tempo de envio da mensagem para o *RabbitMq* até o seu consumo;
- Tempo de download do executor do modelo que se encontra no *Google Drive*;
- Tempo de upload do executor para o ambiente configurado (Santos Dumont no LNCC);
- Tempo da execução propriamente dita;

Na Figura 4.2, podemos observar o comparativo dos tempos envolvidos na execução tanto manual quanto pelo `PhenoManager` . Observa-se que neste cenário, o tempo da execução propriamente dita em si tem uma variação pequena e pouco relevante.

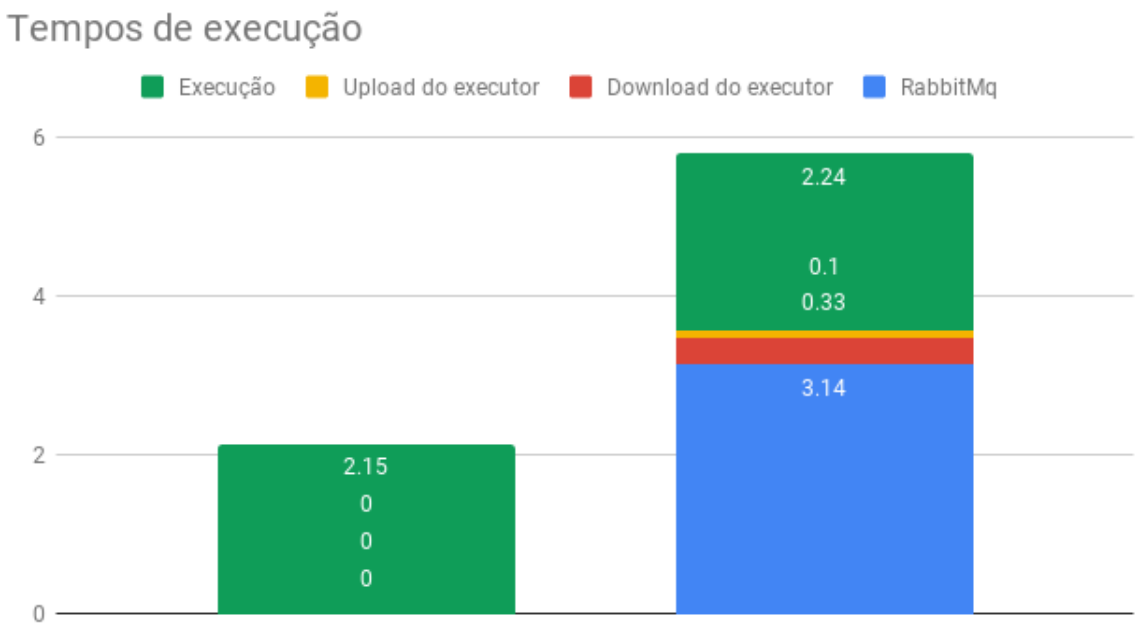
Já na Figura 4.3, no cenário que temos mais execuções paralelas acontecendo no consumidor, podemos observar que há um aumento maior no tempo total da execução do experimento. Neste caso, o modelo computacional permaneceu com estado "*SCHEDULED*"pelo tempo em que o mesmo aguardava a disponibilidade de um consumidor para realizar sua execução. Esse tempo poderia ser severamente diminuído se tívessemos utilizado mais instâncias do consumidor, porém, é claro, instanciar mais consumidores implica em custo financeiro de infraestrutura.

A Tabela 4.1.1 compila os resultados gerais dos tempos de execução medidos nos experimentos.



,

Figura 4.2: Comparativo dos tempos de execução com a fila vazia.



,

Figura 4.3: Comparativo dos tempos de execução com cem execuções paralelas na fila.

Tempos de execução (em minutos)	Manual	PhenoManager (Fila vazia)	PhenoManager (Fila com execuções)
RabbitMq	0:00	0:02	3:14
Download do executor	0:00	0:33	0:33
Upload do executor	0:00	0:10	0:10
Execução	2:15	2:24	2:24

Tabela 4.1: Tempos de execução dos experimentos (em minutos)

4.2 Avaliação do PhenoManager com Usuários

De forma a avaliar qualitativamente o PhenoManager , foi utilizado o modelo de avaliação denominado TAM *(Technology Acceptance Model)* [16]. Este modelo já foi utilizado em trabalhos semelhantes [53, 60]. O experimento proposto por de Souza *et al.* [61] foi a base para a avaliação utilizada nessa dissertação. A ideia principal é avaliar a receptividade/comportamento de um usuário no que se refere a utilidade e a facilidade da tecnologia/ferramenta que está sendo proposta, neste caso, o PhenoManager . A utilidade refere-se ao quanto o usuário acredita que a abordagem proposta o auxiliará em suas tarefas e a facilidade se refere ao quão fácil/simples será utilizar tal abordagem.

Dessa forma, para realizar a avaliação, o TAM sugere a criação de questionários cujas perguntas estejam relacionadas à facilidade e à utilidade da abordagem proposta. Seguindo uma escala de Likert [16], para cada pergunta do questionário, o usuário pode selecionar uma e, somente uma, das opções a seguir: (a) Discordo Totalmente, (b) Discordo Parcialmente, (c) Neutro, (d) Concordo Parcialmente e (e) Concordo Totalmente. A Tabela 4.2 apresenta as questões utilizadas para a avaliação do PhenoManager . As questões cujo identificador se inicia com a letra T se referem a avaliação do treinamento. As questões cujo identificador se inicia com a letra F se referem a avaliação da facilidade. Os demais se referem a avaliação da utilidade.

Para poder avaliar o PhenoManager , cada um dos usuários escolhidos passou por um treinamento simples e presencial de aproximadamente 1 hora. A avaliação do PhenoManager contou com a participação de 9 avaliadores, todos eles estudantes de graduação, mestrado e doutorado da Universidade Federal Fluminense. Do total de alunos avaliadores, 5 eram alunos de graduação e 4 de pós-graduação. Os alunos de pós-graduação já se encontram familiarizados com o ambiente de pesquisa e o funcionamento de um projeto científico. Os alunos de graduação, apesar de não terem tanta experiência, cursaram a disciplina de e-Science oferecida como eletiva para a graduação e se encontram familiarizados com os

ID	Questão
T1	O treinamento oferecido pelos avaliadores foi completo?
T2	O treinamento me ofereceu subsídios para poder utilizar o PhenoManager ?
T3	O treinamento foi adequado em relação ao tempo e ao detalhamento?
F1	Você gostou de utilizar o PhenoManager ?
F2	O acesso ao PhenoManager é simples?
F3	Utilizar o PhenoManager é uma boa ideia?
F4	As funcionalidades no PhenoManager são o simples de serem compreendidas?
F5	é fácil encontrar a informação que desejo no PhenoManager ?
F6	O PhenoManager possui uma interface atraente e amigável?
F7	Mesmo antes de clicar em uma funcionalidade, você já consegue prever o que vai acontecer?
U1	O uso do PhenoManager agrega valor aos experimentos que eu executo?
U2	Utilizar o PhenoManager é útil para ensinar iniciantes a trabalhar em projeto?
U3	Utilizar o PhenoManager pode melhorar o desempenho do meu projeto científico?
U4	O uso do PhenoManager pode facilitar meu trabalho?
U5	Seria factível integrar o uso do PhenoManager na minha rotina de trabalho?
U6	Eu recomendaria o PhenoManager no desenvolvimento de outros projetos?
U7	Os conceitos de validação de hipóteses e projeto científico foram abordados por completo no PhenoManager ?

Tabela 4.2: Questões utilizadas na avaliação do PhenoManager

conceitos envolvidos no PhenoManager . Os projetos e experimentos utilizados na avaliação são alguns já existentes e cujos avaliadores já se encontravam familiarizados com o conteúdo, a saber um experimento de análise filogenética.

A Figura 4.4 apresenta os resultados relativos ao treinamento oferecido. Em relação à completude do treinamento, 88% dos avaliadores consideraram o treinamento completo. Apenas um avaliador queixou-se de que nem todos os conceitos de experimentação científica foram explicados. Esse problema foi devido ao curto espaço de tempo da sessão de treinamento, de forma que alguns conceitos não puderam ser explicados com detalhes. Em relação ao subsídio necessário para se utilizar a ferramenta, todos os avaliadores consideraram que o treinamento ofereceu tais subsídios de forma satisfatória. Em relação a duração do treinamento, 77% dos avaliadores consideraram o tempo satisfatório e dois deles mencionaram que 1 hora não é suficiente para apresentar todas as funcionalidades da ferramenta. Apesar das críticas, consideramos que o treinamento foi satisfatório, pois a maioria dos avaliadores concordou com o treinamento dado. Além disso, procuraremos melhorá-lo em próximas apresentações do PhenoManager .

A Figura 4.5 apresenta os resultados relativos à facilidade de uso percebida com o PhenoManager . Em relação à facilidade de acesso e de uso do PhenoManager , 100%

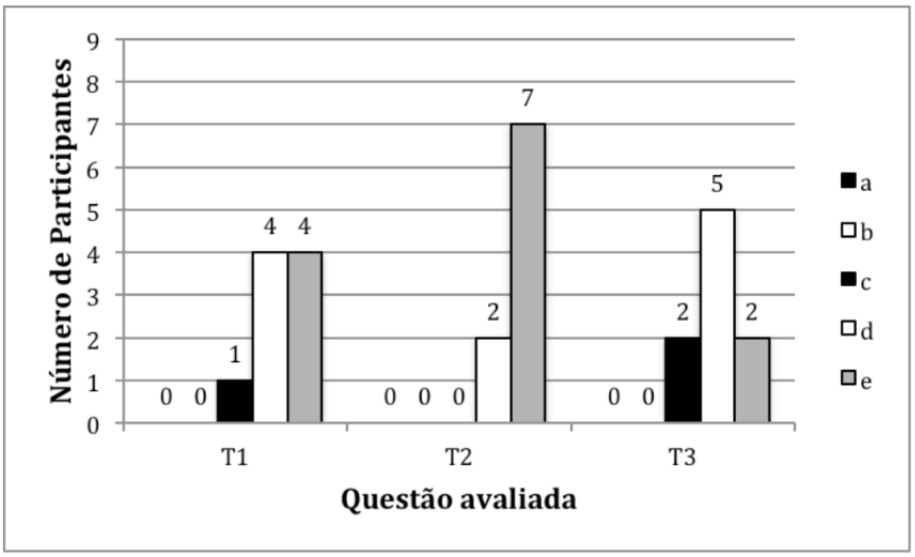


Figura 4.4: Resultado da avaliação do treinamento.

dos avaliadores concordaram (totalmente ou parcialmente) que o sistema é de fácil acesso e fácil de ser utilizado. O mesmo aconteceu na avaliação da interface em que 100% dos avaliadores concordaram que a interface é amigável e atraente. As únicas perguntas em que alguns dos avaliadores se mantiveram neutros foram as perguntas F4 e F5, onde avaliamos a simplicidade de compreensão das funcionalidades e a facilidade de se encontrar informação, respectivamente. Na pergunta F4, um usuário se manteve neutro e alegou que alguns títulos de menus não estavam claros, o que já foi modificado na ferramenta e será liberado em versões futuras. Em relação a pergunta F5, 33% dos avaliadores alegaram que mais filtros de usuários,*workflows*, etc. se fazem necessários no sistema, pois quando a quantidade de informações crescer se tornará inviável buscar os dados nas listas gerenciadas.

A Figura 4.6 apresenta os resultados relativos a utilidade do `PhenoManager` . Nas perguntas em que se verificava se o uso do `PhenoManager` agrega valor ao trabalho e se o uso do `PhenoManager` facilita o trabalho, 100% dos avaliadores concordaram. Em relação ao uso do `PhenoManager` com pesquisadores iniciantes, 88% dos avaliadores concordaram que utilizar a ferramenta ajuda membros iniciantes do projeto a entender o processo. Apenas um avaliador mencionou que se o treinamento não for bem realizado, pode se tornar um empecilho, o que corrobora a crítica em relação ao tempo de treinamento.

Em relação a verificação se o uso do `PhenoManager` pode melhorar o desempenho do serviço, também, 88% dos avaliadores concordaram e um dos avaliadores mencionou que mensurar a melhora do desempenho pode ser complexo, pois dependendo do problema, as

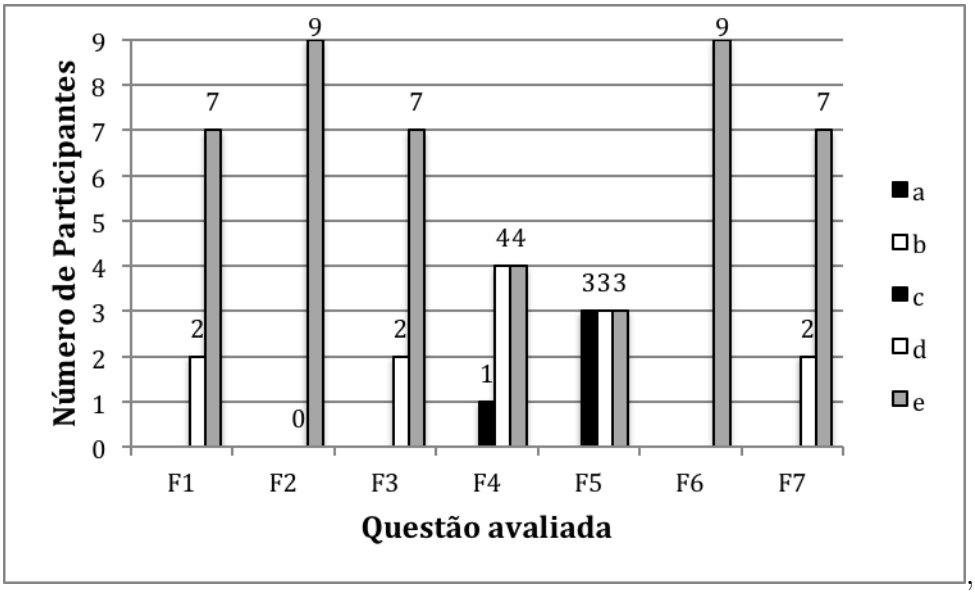


Figura 4.5: Resultado da avaliação da facilidade de uso.

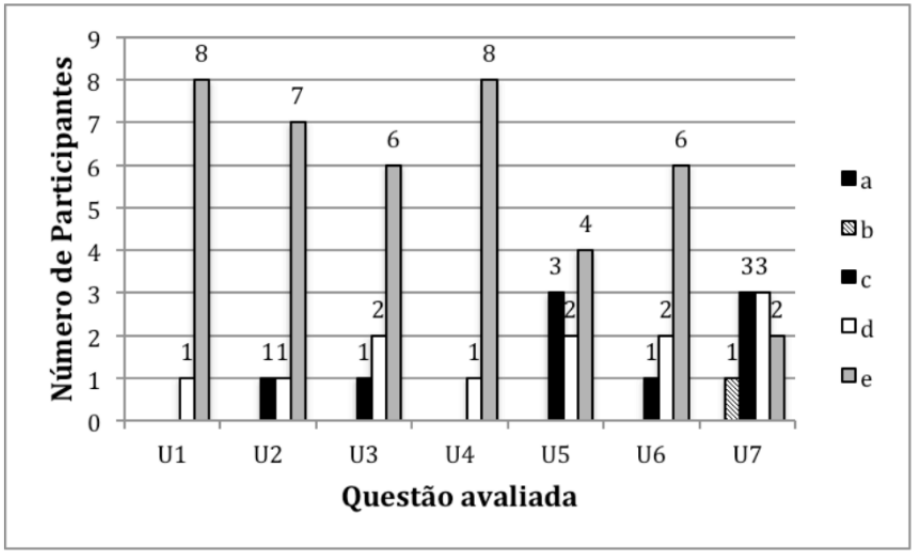


Figura 4.6: Resultado da avaliação da utilidade.

tarefas são bastante demoradas, o que faz com que a contribuição do PhenoManager seja difícil de ser mensurada. Na questão U5 que trata da incorporação do PhenoManager no trabalho, 33% dos avaliadores se mantiveram neutros, pois alegaram que inserir uma ferramenta de gerência em projetos em andamento seria demasiadamente complexo, porém, consideraram o uso em projetos futuros. Na questão U6, 88% dos avaliadores recomendariam o PhenoManager e o avaliador que se manteve neutro alegou que precisaria de mais tempo para estudar e experimentar a tecnologia. Na questão U7, 22% dos avaliadores se mantiveram neutros e 11% discordaram parcialmente. Os avaliadores que discordaram (11%) alegaram que existem conceitos específicos de domínio que não foram tratados na ferramenta (*e.g.* tamanho da sequência de DNA dada como entrada, sequências pertencentes ao banco de dados, *etc.*). Entretanto, essa customização é complicada de ser generalizada e, em um primeiro momento, foi optado por manter a ferramenta mais genérica possível.

Todas as considerações foram anotadas e serão consideradas nas próximas versões do PhenoManager . Apesar de os resultados serem promissores, novas avaliações devem ser realizadas no PhenoManager a fim de confirmar sua utilidade para a comunidade científica. Além disso, segundo Travassos *et al.* [67] devemos explicitar quais são as ameaças a validade do experimento realizado. Uma das ameaças se refere ao uso de alunos como avaliadores. Tentou-se diminuir o impacto dessa ameaça selecionando alunos que já conhecessem os conceitos de experimentação científica e já trabalhassem em projetos científicos reais (seja no nível de IC, Mestrado ou Doutorado). Além disso, Svahnberg *et al.* [13] defendem que o uso de alunos em experimentos é válido desde que o experimento tenha sido conduzido da forma correta. Outra ameaça a validade foi o curto espaço de tempo para se realizar o experimento e o treinamento. Os avaliadores tiveram cerca de uma hora de treinamento e duas horas cada um para realizar a análise. Tal item foi questionado na avaliação e um dos avaliadores mencionou que necessitava de mais tempo para avaliar a ferramenta de forma satisfatória.

4.3 Consulta de dados no PhenoManager

Por último, a última avaliação que restou foi avaliar o uso da ferramenta de consulta que a API do PhenoManager provê. Para isso, foi proposto uma consulta simples, buscando pelos campos slug, que é o UUID da entidade, a versão, o nome de todos os modelos computacionais que contenham a palavra chave *"scyphy"* e também filtrando para que os modelos pertençam a uma hipótese com status como validada. Desta maneira, buscaremos

todos os experimentos que foram realizados nas etapas anteriores da validação e que foram utilizados nesta sessão desta dissertação.

A semântica da APi de consulta do PhenoManager se dá da seguinte forma:

```
http://phenomanager.ic.uff.br/PhenoManagerApi/v1/computational_models?projection=[name,
currentVersion]&filter=[experiment.hypothesis.name=like=sciphy;experiment.hypothesis
.state=VALIDATED]
```

A consulta trouxe dois resultados para o filtro utilizado, validando o teste proposto.

```
{
    "records": [
        {
            "slug": "QERTYUI789321POIUTB8764YUI09POI543",
            "name": "Workflow Sciphy GPU",
            "currentVersion": "1.0"
        },
        {
            "slug": "UTB8764YUI09POI543QERTYUI789321POI",
            "name": "Script Sciphy",
            "currentVersion": "1.2"
        }
    ],
    "metadata": {
        "totalCount": 2,
        "pageOffset": 0,
        "pageSize": 20
    }
}
```

Ainda utilizando a API, poderíamos realizar um teste mais específico, trazendo dados sobre a execução de um dado modelo computacional. Para isto, será utilizado o UUID de um dos registros retornados na consulta anterior. Nesta nova avaliação, buscaremos o nome de todos os usuários responsáveis por execuções, a data de término de todas as execuções com sucesso e o status da execução do modelo computacional do campo slug igual a *"QERTYUI789321POIUTB8764YUI09POI543"*. A dada consulta se dará da seguinte forma:

```
http://phenomanager.ic.uff.br/PhenoManagerApi/v1/computational_models/QERTYUI789321POIU
TB8764YUI09POI543/model_result_metadatas?projection=[userAgent.name,executionFinishDate]
&filter=[executionStatus=FINISHED]
```

A consulta sobre o modelo computacional especifico trouxe os seguinte resultados abaixo, validando o teste proposto.


```
{
    "records": [
        {
            "userAgent": {
                "name": "Leonardo Ramos"
            },
            "executionStatus": "FINISHED",
            "executionFinishDate": "2019-06-03 23:57:20"
        },
        {
            "userAgent": {
                "name": "Administrador"
            },
            "executionStatus": "FINISHED",
            "executionFinishDate": "2019-06-05 19:41:32",
        }
    ],
    "metadata": {
        "totalCount": 2,
        "pageOffset": 0,
        "pageSize": 20
    }
}
```

Capítulo 5

Trabalhos Relacionados

Este capítulo apresenta trabalhos relacionados com esta dissertação, ou seja, propostas de trabalhos que tenham como objetivo auxiliar no método científico, seja ajudando na reprodução de experimentos, na gerência, modelagem ou execução de modelos computacionais.

Existem poucas iniciativas que visam viabilizar execuções de tarefas de maneira simples e agnóstica. Mais escassas ainda são as soluções que, além disso, se propõe a auxiliar na reprodução de um dado experimento.

Para o propósito de compartilhamento de *Research Objects* e dados de experimentos, podemos citar dois trabalhos que se assemelham com a proposta desta dissertação: *myExperiment* [27] e *Galaxy Project* [1].

O Galaxy [1] é um gerenciador de *workflows* científico, integrador de dados, e plataforma de publicação e persistência para análise de dados que visa tornar a bioinformática acessível a pesquisadores que não possuem experiência em programação de computadores ou administração de sistemas. Embora tenha sido inicialmente desenvolvido para pesquisa em genômica, é uma ferramenta que pode ser usado independente de domínio e pode ser aplicada, em teoria, a qualquer domínio científico. Nos dias de hoje, é usado como um sistema geral de gerenciamento de *workflows* de bioinformática. O Galaxy suporta uma variedade de formatos de dados biológicos amplamente utilizados e a tradução entre esses formatos. Há tambem uma interface *Web* com muitos utilitários de manipulação de texto, permitindo que os pesquisadores façam sua própria reformatação e manipulação personalizadas sem precisar fazer nenhuma ou mínima programação. O Galaxy inclui utilitários de manipulação de intervalo para realizar operações teóricas definidas (por exemplo, interseção e união) em intervalos. Muitos formatos de arquivo biológico incluem dados de

intervalo genômico (um quadro de referência, por exemplo, nome do cromossomo e posições de início e parada), permitindo que esses dados sejam integrados [1]. Um contraponto ao Galaxy é que o mesmo carece de uma maior flexibilidade com relação à configuração dos ambientes dos modelos de execução e da gestão do projeto como um todo, o que difere da proposta do trabalho do contexto desta dissertação.

O *MyExperiment* é uma plataforma social para compartilhamento de pesquisa e de *Research Objects* como os de *workflows* científicos.[27] Diferentemente do *Galaxy*, seu objetivo é apenas compartilhar *Research Objects* de pesquisas e não se propõe a execução e confuguração de modelos de dados em si. Esta ferramenta poderia ser facilmente integrada com o `PhenoManager` , contanto que o modelo computacional seja configurado como público, e, desta forma, poderíamos ter os *Research Objects* da ferramenta desta dissertação sendo compartilhados e reusados no *MyExperiment*. Para o mesmo propósito de compartilhamento de *Research Objects*, pode ser citado também o *Wf4Ever Toolkit* [49], uma ferramenta de semelhante propósito, que, porém, é centrada no armazenamento e na categorização de *Research Objects* e seus metadados. Assim como o *MyExperiment*, esta ferramenta poderia ser facilmente integrada com o `PhenoManager` , tirando proveito dos benefícios de ambas as propostas.

A respeito da modelagem da execução propriamente dita de modelos computacionais, uma ferramenta que que pode ser citada é a ferramenta *Rabix*. Esse projeto funciona como um modelador de *Workflow*, possibilitando, por meio de *Common Workflow Language (CWL)* [4], a configuração de cada etapa de um *workflow* científico, se encarregando da execução e do armazenamento dos dados para posterior reprodução. CWL é um padrão aberto para descrever *workflows* e ferramentas de análise de maneira a torná-los portáteis e escaláveis em uma variedade de ambientes de software e hardware, das estações de trabalho aos ambientes de cluster, nuvem e computação de alto desempenho (PAD). O CWL foi projetado para atender às necessidades de ciência com muitos dados, como Bioinformática, Imagem Médica, Astronomia, Física e Química [4]. O *Rabix*, em relação ao ambiente de execução, também pode ser agnóstico, permitindo que os *jobs* sejam executados em ambientes passíveis de configuração. Essa solução não é disponível em nuvem, portando sendo requirido que o usuário a utilize em sua máquina pessoal, ou ambiente de execução de sua preferência em que o próprio usuário deveria configurar e arcar com os custos de tempo de configuração e de aprendizado, diferente da ferramenta proposta, que se encarrega de realizar as devidas configurações, além de prover ambiente em nuvem em alta disponibilidade.

Semelhante a CWL, e inserido no contexto desta dissertação, podemos citar outra abordagem no que diz respeito a modelagem de *workflows* científicos e sua reprodutividade no trabalho apresentado em [42]. Geralmente, os SGWfCs oferecem mecanismos fundamentais de automação como despacho de tarefas e movimentação de dados, mas trabalhos recentes têm demonstrado também o interesse no suporte a atributos não-funcionais, como confiabilidade e rastreabilidade, nesses sistemas [42]. A configuração desses atributos nos SGWfCs atuais é, contudo, limitada em vários aspectos, em particular no que se refere às interações entre tarefas. Essa limitação confere menor expressividade às linguagens de modelagem de *workflows* providas por esses SGWfCs. Objetivando o aprimoramento dessa expressividade, o trabalho analisado apresenta a proposta *Open Scientific Connectors* (OSC), que usa conceitos arquiteturais para representar interações entre tarefas como elementos de primeira classe. Esta abordagem propicia uma maior capacidade de composição e de reuso na modelagem de *workflows*, particularmente nas configurações dos mecanismos que lidam com atributos não-funcionais. Este trabalho trata da modelagem em etapas anteriores a proposta do `PhenoManager` , o que tornaria o uso de ambas e até a integração dos dois trabalhos uma opção plausível.

No que se refere a reprodutividade de experimentos, podemos citar o trabalho em [57], que foca no ambiente e da infra estrutura a ser reproduzida, propondo uma abordagem baseada em ontologia. Diferente da proposta deste trabalho, e também muito importante, o trabalho [57], foca no compartilhamento dos recursos de ferramentas e de infraestrutura necessários para reproduzir e conduzir um dado experimento. No contexto da gerência do projeto científico, este é um fator muito relevante, que, porém não é abordado nesta dissertação.

Na cerne da gerência do projeto científico como um todo, existem algumas iniciativas que têm como objetivo oferecer apoio aos cientistas na gerência de projetos científicos. O LabGuru [1] tem como objetivo auxiliar o cientista na gerência do seu laboratório. Apesar de possuir funcionalidades para distribuição de tarefas, o LabGuru tem como objetivo maior alocação de recursos (pessoas e equipamentos) nas tarefas do dia-a-dia e a compra de equipamentos para que o laboratório possa continuar funcionando. Não há uma preocupação com a execução do experimento e com a análise dos dados no LabGuru. A importância de se ter métodos e técnicas que ajudem a gerenciar projetos científicos foi também levantada por Portny e Austin [51]. Nesse mesmo artigo, Portny e Austin discutem sobre a falta de ferramentas para apoiar tal gerência. Pinheiro *et al.* [50] propõem uma metodologia de gestão de projetos de pesquisa, que se baseia no uso de técnicas de

[1] http://www.labguru.com/features/research-project-management

gerência de projetos na pesquisa para a obtenção de produtos. Apesar de não apresentarem nenhuma ferramenta, os autores discutem a importância delas para o processo. Além disso, ferramentas comerciais já existentes como o Trello [2], Tasker [3], Redmine [4] e o Jira [5] podem ser utilizadas para gerência de projetos científicos. Entretanto, como essas ferramentas não tem como objetivo atender tal categoria de projetos, não há associação com o conceito de experimento científico e nem mecanismos para especificar um *workflow* ou executá-lo em um SGWfC existente. Além do mais, algumas dessas ferramentas, como o Jira por exemplo, são pagas. Ademais, podemos citar o *SciManager* [53], que já foi citado anteriormente nesta dissertação e inclusive provê integração com o `PhenoManager` e vice-versa.

[2]https://trello.com/
[3]http://www.tasker.com.br/
[4]http://www.redmine.org/
[5]https://www.atlassian.com/software/jira

Capítulo 6

Conclusão

Este capítulo é uma apresentação das conclusões desta dissertação, além de também apresentar as principais perspectivas de trabalhos futuros e melhorias que podem ser desenvolvidos como consequência direta a partir desta dissertação.

Gerenciar um projeto científico é uma tarefa nada trivial. Um projeto científico pode englobar vários experimentos e cada experimento pode necessitar de várias execuções de diferentes modelos computacionais para que uma hipótese seja verificada. A complexidade de configuração e execução de modelos computacionais nos diversos ambientes e a falta de *know-how* dos envolvidos podem ser dificultadores extra. Essa gerência pode se tornar ainda mais complexa à medida que a equipe do projeto se encontra distribuída separadamente no espaço geográfico.

Dessa forma, distribuir tarefas, medir esforços despendidos, controlar a equipe e ter controle sobre os ambientes computacionais, assim como os dados de proveniência envolvidos, são tarefas muito onerosas que requerem uma especificação do cientista que ultrapassa sua área de pesquisa natural. A área de gerência de projetos já trata de problemas semelhantes em diversas situações como no desenvolvimento de software. Entretanto, a gerência do ciclo da ciência possui peculiaridades e características únicas que precisam ser tratadas com mais atenção. Visando prover maior suporte aos cientistas em experimentos científicos em ambientes de alto desempenho, dado que este tema tem sido abordado em diversos congressos e conferências, seja ela nacional ou internacional, esta pesquisa de dissertação apresentou esta abordagem com seus resultados para prover apoio a esta necessidade detectada.

Isto posto, a necessidade de ferramentas específicas para a gerência do projeto científico, assim como a configuração e execução de modelos computacionais nos mais diversos

ambientes de maneira transparente, fácil e agnóstica fica evidente. Nesta dissertação foi apresentado o `PhenoManager` , um sistema de informação cujo foco é apoiar a método científico, desde a criação e modelagem do domínio envolvido, até a configuração do ambiente, do modelo de execução e do acompanhamento de resultados. O `PhenoManager` foi desenvolvido utilizando software livre e deve ser hospedado em um ambiente de nuvem para garantir alta disponibilidade de seus serviços.

Com a avaliação experimental foi possível perceber que o *overhead* causado pela ferramenta é pouco significante, haja visto o benefício e a praticidade que ela proporciona. Com o uso da ferramenta, o cientista terá atenuada a responsabilidade de conhecer e admistrar diversas ferramentas, linguagens e arquiteturas diferentes tão detahadamente e tão a fundo para configurar um modelo de execução. Para exercer as mesmas tarefas que antes ele exercia, agora basta o cientista saber manusear a ferramenta proposta, independente de que tipo de programa ele deseja executar.

Trabalhos futuros incluem implementar avaliação de desempenho de uma hipótese perante as demais, de forma a ordená-las por ordem de ranqueamento de acordo com alguma métrica estipulada. Esse fluxo de avaliação de resultados poderia entrar no final do *pipeline* de execução de um modelo, de maneira paralela. Outra possível melhoria seria possibilitar o monitoramento da saúde do ambiente computacional em que o modelo efetuará sua execução, para que fosse possível detectar possíveis problemas inesperados de falta de recursos ou de problemas de configuração de ambiente. Por fim, uma funcionalidade que agregaria bastante valor seria viabilizar a configuração de diversos executores, sincronizando suas execuções com suas saídas de dados, de modo a ser possível criar um *Workflows* pelo próprio sistema, sem a necessidade de utilizar um *SGWfC* como um executor em um arquivo *.zip*.

Por fim, espera-se que o `PhenoManager` seja uma ferramenta que possa trazer benefícios ao cotidiano do cientista e que os usuários finais tenham suas vidas facilitadas.

Referências

[1] E. Afgan, D. Baker, M. Van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, 44(W1):W3–W10, 2016.

[2] F. C. Ahlert, L. Moura, G. Borba, D. Silva, and D. Silva. Gestão de serviços na área da saúde: a simulação computacional no auxílio à tomada de decisão. *Encontro Nacional de Engenharia de Produção, XXIX*, 2009.

[3] I. Altintas, B. Ludaescher, S. Klasky, and M. A. Vouk. Introduction to scientific workflow management and the kepler system. In *SC'06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, page 205, 2006.

[4] P. Amstutz, M. R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, et al. Common workflow language, v1. 0. 2016.

[5] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.

[6] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan. Research objects: Towards exchange and reuse of digital knowledge. 2010.

[7] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. M. Gómez-Pérez, S. Bechhofer, et al. Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics*, 32:16–42, 2015.

[8] W. I. B. Beveridge. *The art of scientific investigation.* Edizioni Savine, 2017.

[9] R. E. Bryant. Data-intensive scalable computing for scientific applications. *Computing in Science & Engineering*, 13(6):25–33, 2011.

[10] P. Buneman, S. Khanna, and T. Wang-Chiew. Why and where: A characterization of data provenance. In *International conference on database theory*, pages 316–330. Springer, 2001.

[11] A. F. Chalmers and R. Fiker. *O que é ciência afinal?* Brasiliense São Paulo, 1993.

[12] C. P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[13] S. M. S. da Cruz, P. M. Barros, P. M. Bisch, M. L. M. Campos, and M. Mattoso. Provenance services for distributed workflows. In *2008 Eighth IEEE International Symposium on Cluster Computing and the Grid (CCGRID)*, pages 526–533. IEEE, 2008.

[14] S. M. S. da Cruz, M. L. M. Campos, and M. Mattoso. Towards a taxonomy of provenance in scientific workflow management systems. In *2009 Congress on Services-I*, pages 259–266. IEEE, 2009.

[15] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1345–1350. ACM, 2008.

[16] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[17] B. Y. C. L. de Mello and L. L. Caimi. Simulação na validação de sistemas computacionais para a agricultura de precisão. *R. Bras. Eng. Agric. Ambiental, Campina Grande Nov./Dec*, 12(6):666–675, 2008.

[18] D. de Oliveira, E. Ogasawara, F. Baião, and M. Mattoso. Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. In *International Conference on Cloud Computing*, pages 378–385. IEEE, 2010.

[19] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[20] E. Deelman, D. Gannon, M. Shields, and I. Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future generation computer systems*, 25(5):528–540, 2009.

[21] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, et al. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming*, 13(3):219–237, 2005.

[22] J. Dias, E. Ogasawara, and M. Mattoso. Paralelismo de dados científicos em workflows usando técnicas p2p. In *IX Workshop de Teses e Dissertações em Banco de Dados*, pages 85–91, 2010.

[23] I. Foster and C. Kesselman. Globus: A metacomputing infrastructure toolkit. *The International Journal of Supercomputer Applications and High Performance Computing*, 11(2):115–128, 1997.

[24] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21, 2008.

[25] H. G. G. Gauch Jr, Hugh G. and H. G. G. Jr. *Scientific Method in Practice*. Oxford University Press, 2002.

[26] D. Gilbert. Sequence file format conversion with command-line readseq. *Current protocols in bioinformatics*, (1):A–1E, 2003.

[27] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, et al. myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(suppl_2):W677–W682, 2010.

[28] M. d. N. Gomes, L. Isoldi, C. Olinto, L. Rocha, and J. Souza. Computational modeling of a regular wave tank. In *2009 3rd Southern Conference on Computational Modeling*, pages 60–65. IEEE, 2009.

[29] B. Gonçalves and F. Porto. Managing large-scale scientific hypotheses as uncertain data with support for predictive analytics. *CoRR*, 2014.

[30] A. J. Hey, S. Tansley, K. M. Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[31] M. Interlandi, K. Shah, S. D. Tetali, M. A. Gulzar, S. Yoo, M. Kim, T. Millstein, and T. Condie. Titian: Data provenance support in spark. *Proceedings of the VLDB Endowment*, 9(3):216–227, 2015.

[32] J. C. Jacob, D. S. Katz, G. B. Berriman, J. Good, A. C. Laity, E. Deelman, C. Kesselman, G. Singh, M.-H. Su, T. A. Prince, et al. Montage: a grid portal and software toolkit for science-grade astronomical image mosaicking. *arXiv preprint arXiv:1005.4454*, 2010.

[33] R. D. Jarrard. Scientific methods. *University of Utah, Utah*, 2001.

[34] T. M. Keane, C. J. Creevey, M. M. Pentony, T. J. Naughton, and J. O. Mclnerney. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC evolutionary biology*, 6(1):29, 2006.

[35] F. I. Leal and S. de Oliveira Junior. Modelagem e simulação de mecanismos artificiais de elevacão em plataformas offshore de prospecção de petróleo. *TecMec 2006*, 2006.

[36] S. Lifschitz. Gerenciadores de dados biológicos: Genéricos ou ad-hoc? In *XXXIV Seminário Integrado de Software e Hardware (SEMISH) do XXVII Congresso da Sociedade Brasileira de Computação*, pages 2085–2099, 2007.

[37] C. Lim, S. Lu, A. Chebotko, and F. Fotouhi. Prospective and retrospective provenance collection in scientific workflow environments. In *2010 IEEE International Conference on Services Computing*, pages 449–456. IEEE, 2010.

[38] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, 2006.

[39] A. Mattos, F. Silva, N. Ruberg, and M. Cruz. Gerência de workflows científicos: uma análise crítica no contexto da bioinformática. *COPPE/UFRJ*, 2008.

[40] M. Mattoso, C. Werner, G. Travassos, V. Braganholo, and L. Murta. Gerenciando experimentos científicos em larga escala. *SBC-SEMISH*, 8:121–135, 2008.

[41] M. Mattoso, C. Werner, G. H. Travassos, V. Braganholo, E. Ogasawara, D. Oliveira, S. Cruz, W. Martinho, and L. Murta. Towards supporting the life cycle of large scale scientific experiments. *International Journal of Business Process Integration and Management*, 5(1):79–92, 2010.

[42] V. Medeiros and A. T. A. Gomes. Towards fully configurable support to non-functional attributes in scientific workflows. In *Proceedings of the 8th IEEE International Conference on eScience (eScience 2012)*, pages 800–145, 2012.

[43] P. Missier, K. Belhajjame, and J. Cheney. The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776. ACM, 2013.

[44] L. Moreau, P. Missier, K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, et al. Prov-dm: The prov data model. *Retrieved July*, 30:2013, 2013.

[45] L. Murta, V. Braganholo, F. Chirigati, D. Koop, and J. Freire. noworkflow: capturing and analyzing provenance of scripts. In *International Provenance and Annotation Workshop*, pages 71–83. Springer, 2014.

[46] S. Newman. *Building microservices: designing fine-grained systems*. "O'Reilly Media, Inc.", 2015.

[47] K. A. Ocaña, D. de Oliveira, E. Ogasawara, A. M. Dávila, A. A. Lima, and M. Mattoso. Sciphy: a cloud-based workflow for phylogenetic analysis of drug targets in protozoan genomes. In *Brazilian Symposium on Bioinformatics*, pages 66–70. Springer, 2011.

[48] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.

[49] K. Page, R. Palma, P. Holubowicz, G. Klyne, S. Soiland-Reyes, D. Cruickshank, R. G. Cabero, E. G. Cuesta, D. De Roure, and J. Zhao. From workflows to research objects: an architecture for preserving the semantics of science. In *Proceedings of the 2nd International Workshop on Linked Science*, volume 10, page 2012. Citeseer, 2012.

[50] A. A. Pinheiro, A. C. Siani, J. d. F. Guilhermino, M. d. G. M. d. Henriques, C. M. Quental, A. P. B. Pizarro, et al. Management methodology for product-oriented r&d projects: a proposal. *Revista de Administração Pública*, 40(3):457–478, 2006.

[51] S. E. Portny and J. Austin. Project management for scientists. *Science Careers (July 12)*, 2002.

[52] F. Porto, R. G. Costa, A. M. d. C. Moura, and B. Gonçalves. Modeling and implementing scientific hypothesis. *Journal of Database Management (JDM)*, 26(2):1–13, 2015.

[53] L. S. Ramos, K. A. Ocaña, and D. de Oliveira. Um sistema de informação para gerência de projetos científicos baseados em simulações computacionais. In *Anais do XII Simpósio Brasileiro de Sistemas de Informação*, pages 216–223. SBC, 2016.

[54] K. H. Rose. A guide to the project management body of knowledge (pmbok® guide)—fifth edition. *Project management journal*, 44(3):e1–e1, 2013.

[55] R. Sakellariou and H. Zhao. A hybrid heuristic for dag scheduling on heterogeneous systems. In *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, page 111. IEEE, 2004.

[56] M. M. Salvatierra et al. Modelagem matemática e simulação computacional da presença de materiais impactantes tóxicos em casos de dinâmica populacional com competição inter e intra-específica. 2005.

[57] I. Santana-Perez and M. S. Pérez-Hernández. Towards reproducibility in scientific workflows: An infrastructure-based approach. *Scientific Programming*, 2015, 2015.

[58] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*, pages 1–10. Ieee, 2010.

[59] C. Silva. *Captura de Dados de Proveniência de Workflows Científicos Em Nuvens Computacionais*. PhD thesis, Dissertação de M. Sc., COPPE/UFRJ, Rio de Janeiro, Brasil, 2011.

[60] I. Souza, P. Oliveira, E. Junior, A. Inocêncio, and P. Júnior. Tese-um sistema de informacao para gerenciamento de projetos experimentais em engenharia de software. In *Anais do XI Simpósio Brasileiro de Sistemas de Informação*, pages 563–570. SBC, 2015.

[61] I. Souza, P. Oliveira, E. Junior, A. Inocêncio, and P. Júnior. Tese-um sistema de informacao para gerenciamento de projetos experimentais em engenharia de software. In *Anais do XI Simpósio Brasileiro de Sistemas de Informação*, pages 563–570. SBC, 2015.

[62] J. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA, ©2000., 1999.

[63] A. Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[64] A. Stevenson. *Oxford dictionary of English*. Oxford University Press, USA, 2010.

[65] M. Stonebraker. The case for shared nothing. *IEEE Database Eng. Bull.*, 9(1):4–9, 1986.

[66] D. Thain, T. Tannenbaum, and M. Livny. Condor and the grid. *Grid computing: Making the global infrastructure a reality*, pages 299–335, 2003.

[67] G. H. Travassos and M. O. Barros. Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. In *2nd Workshop on Empirical Software Engineering the Future of Empirical Studies in Software Engineering*, pages 117–130, 2003.

[68] G. Vossen and M. Weske. The wasa approach to workflow management for scientific applications. In *Workflow Management Systems and Interoperability*, pages 145–164. Springer, 1998.

[69] L. Wang, J. Tao, R. Ranjan, H. Marten, A. Streit, J. Chen, and D. Chen. G-hadoop: Mapreduce across distributed data centers for data-intensive computing. *Future Generation Computer Systems*, 29(3):739–750, 2013.

[70] T. Weiss, C. Tamura, and E. L. Krüger. Uso de simulação computacional como suporte a um estudo de iluminação natural em câmara climática. *ENCONTRO NACIONAL*, 13, 2015.

[71] J. M. Wozniak, T. G. Armstrong, M. Wilde, D. S. Katz, E. Lusk, and I. T. Foster. Swift/t: Scalable data flow programming for many-task applications. In *ACM SIGPLAN Notices*, volume 48, pages 309–310. ACM, 2013.

[72] J. Yu and R. Buyya. Scheduling scientific workflow applications with deadline and budget constraints using genetic algorithms. *Scientific Programming*, 14(3-4):217–230, 2006.

[73] J. Yu, R. Buyya, and C. K. Tham. Cost-based scheduling of scientific workflow applications on utility grids. In *First International Conference on e-Science and Grid Computing (e-Science'05)*, pages 8–pp. Ieee, 2005.

[74] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

[75] J. Zhao. Research object for scholarly communication (rosc) community group charter. `http://www.w3.org`, 2013.