V Workshop de Dissertações de Mestrado do PPgSI (2018)

FICHA DE PESQUISA

DADOS GERAIS				
Título do projeto de pesquisa	Abordagem para integração automática de dados estruturados e não estruturados com uso de técnicas de aprendizagem em um contexto Big Data.			
Orientando	Keylla Ramos Saes			
Orientador (es)	Luciano Vieira de Araújo			
Semestre no curso, na data do workshop	[ ] 2º semestre	[ ] 3º semestre	[X] 4º semestre	[ ] 5º semestre
Qualificação	[X] Qualificação já realizada em: 13/04/2018 [ ] Realização da qualificação planejada para: dd/mm/aaaa			
Defesa	Prazo máximo para depósito: 03/09/2019 Realização da defesa planejada para: 10/2018			
Linha e Área de pesquisa	Gestão e Desenvolvimento de Sistemas: [X] BD [ ] Engenharia de Software [ ] Gestão de TI [ ] IHC	Inteligência de Sistemas: [ ] IA [ ] Processamento Gráfico [ ] Reconhecimento de Padrões		
Área de aplicação	[ ] Ambientes corporativos [ ] Bioinformática [ ] Biometria [ ] Dispositivos móveis [ ] Educação	[ ] Educação a distância [ ] Governo eletrônico [ ] Internet [ ] Jogos [ ] Jogos sérios	[ ] Língua Natural [ ] Linguística [ ] Processos de Negócio [ ] Quimioinformática	[ ] Redes Sociais [ ] Robótica [ ] Saúde [X] [Big Data]
Publicações associadas ao projeto de mestrado	1) Integration of structured and unstructured data in a Big Data context: a survey (artigo em elaboração com envio acordado para ACM) 2) Approach to automatic integration of structured and unstructured data with use of learning techniques in the context of Big Data (artigo em elaboração com envio acordado para VLDB)			

DESCRIÇÃO DO PROJETO DE PESQUISA	
Contextualização / motivação	Essa pesquisa é parte de uma pesquisa maior sobre Data DNA.
Problema de pesquisa	Integrar dados com modelos de representação homogêneos não é um processo trivial, e a complexidade é ampliada ao considerar na integração modelos de representação de dados heterogêneos. Devido a variedade de modelos de representação, integrar dados de maneira manual pode tornar-se inviável. Como proposta para este desafio de integração, esse trabalho apresenta uma abordagem automática para apoiar a integração de dados com grande variedade de modelos de representação e alto volume de dados, ou seja, em um contexto Big Data.
Objetivo geral da pesquisa	Este estudo propõe uma abordagem para ampliação dos cenários de integração de dados automática entre modelos de representação de dados estruturados e não estruturados, como forma de apoio a geração de conhecimento, através de agilidade na disponibilização de dados integrados com redução de processos manuais.
Trabalhos relacionados	• Proposta de um processo de integração de dados heterogêneos utilizando uma virtualização de dados que possui um tradutor entre linguagem SQL e NOSQL. A proposta só aceita dados armazenados em SQL e dados armazenados no MongoDB, o que é uma limitação da solução. (LAWRENCE,2014). • Uma abordagem semiautomática propõe um mediador com algoritmos inteligentes com uso de Machine Learning para realizar a integração de dados contidos em esquemas distintos. A baixa acurácia da solução é uma limitação que será endereçada em trabalhos futuros. (DOAN; DOMINGOS, 2001) • O SEMINT é a proposição de um modelo de integração para dados estruturados com a utilização de algoritmos inteligentes. O trabalho apresenta como contribuição um processo semiautomático para integração de dados, porém limitados aos dados estruturados. (LI; CLIFTON, 2000)
Proposta para solução	*(i)* proposta uma nova abordagem;
Dados	Dados públicos extraídos de bases federadas como CKAN, Data One, e outras bases como Kaggle, e bases do governo, como bolsa família.
Validação	A validação da abordagem para integração de dados heterogêneos utilizará dados públicos extraídos de bases federadas como CKAN, Data One, e outras bases como Kaggle, e bases do governo para realizar integração de dados de maneira automática.

Limitações	A abordagem proposta permitirá a inclusão de componentes para os diversos modelos de representação, entretanto os experimentos serão limitados a alguns modelos de representação que estão sendo definidos.
Resultados esperados	Contribuições científicas: A abordagem proposta visa contribuir com a integração de dados de maneira automática. Diversas áreas da ciência necessitam analisar dados heterogêneos em suas pesquisas, e esta abordagem automática reduz esforço de integração do dado. Com isso, o pesquisador ou analista de dados podem se concentrar mais em atividades de interpretação do dado ao invés de utilizar tempo para entender como integrar os dados para posterior análise. Além desse benefício, a integração automática de dados pode gerar novas estruturas de dados que proporcionarão novas ideias para análises científicas.
	Contribuições tecnológicas: Apresentação de uma abordagem com uma arquitetura de integração de dados estruturados e não estruturados, implementado como em uma prova de conceito.

MÉTODO DE PESQUISA			
Gênero	[ ] Pesquisa teórica	[ ] Pesquisa prática	[ ] Pesquisa empírica [X] Pesquisa metodológica
Natureza	[ ] Pesquisa básica/pura	[X] Pesquisa aplicada	
Objetivo	[X] Pesquisa exploratória	[ ] Pesquisa descritiva	[ ] Pesquisa explicativa [ ] [outro – escrever]
Abordagem	[ ] Pesquisa quantitativa	[X] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Procedimento técnico principal	[X] Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto* [ ] Pesquisa de levantamento	[ ] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação [ ] Pesquisa etnográfica	[ ] Pesquisa netnográfica [ ] Teoria fundamentada em dados (*Grounded theory*) [ ] Ciência do projeto (*Design science*) [ ] [outro – escrever]
Técnica(s) / Instrumento(s) de coleta de dados	[ ] Medição [ ] Questionário [ ] Entrevista [ ] Grupos focais	[ ] Formulário [ ] *Benchmark*	[ ] Observação (direta / participante) [ ] Diário de campo / notas de campo [ ] Análise documental (ou de artefatos) [X] [Dados públicos]
Técnica(s) de análise de dados	Análise quantitativa: [ ] Estatística descritiva [X] Estatística inferencial [X] [outro – escrever]		Análise qualitativa: [X] Análise de conteúdo [ ] Análise do discurso [X] [outro – escrever]

CRONOGRAMA

	2017												2018												2019											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Estudo da literatura	X	X																																		
Revisão sistemática	X	X	X	X	X	X	X	X	X	X	X	X																								
Definição da proposta	X	X	X	X	X	X																														
Execução da proposta							X	X	X	X	X	X																								
Coleta/obtenção de dados											X	X	X	X	X	X	X	X																		
Análise dos dados												X	X	X	X	X	X	X	X	X	X															
Validação da proposta												X	X	X	X																					
Preparo da qualificação												X	X	X	X																					
Exame de qualificação																X																				
Escrita da dissertação																	X	X	X	X	X															
Escrita de artigo					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X															
Depósito da dissertação																						X														