


Tóp. de Banco de Dados V		
CENTRO UNIVERSITÁRIO CTC DEPARTAMENTO DE INFORMÁTICA		
INF2391	Tópicos de Bancos de Dados V: Ferramentas, plataformas e padrões para Integração de Dados em Ciência de Dados	
	CARGA HORÁRIA TOTAL:	CRÉDITOS: 3
	Pré-requisitos: Não há Prof. Antônio L. Furtado / Marco A. Casanova	

OBJETIVOS	Esta disciplina aborda problemas, técnicas e ferramentas para integração de grandes volumes de dados, com aplicações práticas em Ciência de Dados.
EMENTA	Introdução a integração de Dados; Arquiteturas: federações de bancos de dados, data lakes, bases de conhecimento integradas; Processos para integração de dados: ligação de entidades, extração de dados, fusão de dados, alinhamento de esquemas; ferramentas.
PROGRAMA	<ul style="list-style-type: none"> • Introdução: O que é Integração de Dados; Aprendizado de Máquina para Integração de Dados; Integração de Dados em Ciência de Dados • Arquiteturas: <ul style="list-style-type: none"> ○ Convencional x Web; ○ <i>Data Warehouse</i> x Integração Virtual; ○ <i>Data Lakes</i> • Processos para Integração de Dados: <ul style="list-style-type: none"> ○ Ligação de Entidades: estratégias baseadas em regras; estratégias supervisionadas; estratégias baseadas <i>Deep Learning</i>; ○ Extração de Dados: estratégias baseadas em regras; estratégias específicas para um domínio; estratégias probabilísticas; estratégias baseadas <i>Deep Learning</i>; ○ Fusão de Dados: estratégias baseadas em regras; estratégias específicas para um domínio; estratégias probabilísticas; estratégias baseadas <i>Deep Learning</i>; ○ Alinhamento de Esquemas: estratégias baseadas em Lógica de Descrição; estratégias para mapeamento semi-automático; <i>pay-as-you-go dataspace</i>s; estratégias baseadas em Lógicas e em <i>Deep Learning</i>; • Ferramentas: <ul style="list-style-type: none"> ○ Ferramentas genéricas para integração de dados: Oracle Data Integrator, Tableaux, Pentaho e outras; ○ Ferramentas genéricas para Data Lakes: Aurum, ConnectionLens, Apache Tika, SKLUMA; ○ Ferramentas para alinhamento de esquemas: Harmony, SMART; ○ Ferramentas para ligação de entidades: Magellan, DeepMatcher, Silk;

	<ul style="list-style-type: none"> ○ Ferramentas para reconhecimento de entidades nomeadas: Ambiverse, Flair, spaCy, Apache OpenNLP; ○ Ferramentas para limpeza e fusão de dados: CLAMS, Data X-Ray, OpenRefine.
AVALIAÇÃO	O desempenho do aluno será avaliado por seminários apresentados em sala.
BIBLIOGRAFIA PRINCIPAL	<p>AnHai Doan, Alon Halevy and Zachary Ives. <i>Principles of Data Integration</i>. Morgan Kaufmann, 1st edition (2012), 520 pages. Web page: https://research.cs.wisc.edu/dibook/</p> <p>Gerhard Weikum, Xin Luna Dong, Simon Razniewski and Fabian Suchanek (2021), “Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases”, <i>Foundations and Trends® in Databases</i>: Vol. 10, No. 2-4, pp 108–490. DOI: 10.1561/19000000064.</p> <p>(Artigos selecionados)</p>
BIBLIOGRAFIA COMPLEMENTAR	(Artigos selecionados)