Faculdade IMPACTA

Pós-graduação

Big Data e Business Analytics

GRADE DE DISCIPLINAS DO CURSO		
Conteúdos	**Horas**	**Presencial**
Conceitos de Modelagem de Dados	24	24
Python para Data Science	24	24
Modelagem Dimensional e ETL	28	28
Análise Exploratória de Dados	24	24
Linguagens SQL e NoSQL	24	24
Data Mining & Machine Learning	24	24
Cloud Computing	20	20
Data Lab - Cases em Machine Learning	20	20
Arquitetura e Tecnologias em Big Data	20	20
Web Mining & Social NetWorking Analysis	28	28
Data Visualization & Storytelling	24	24
DataLab – Business Analytics	28	28
Big Data Governance	24	24
Corporate Performance Management – CPM	24	24
Projeto de Big Data Analytics	24	24
Subtotal	360	360
Total do Curso	**360**	**360**

Descrição das disciplinas

CONCEITOS DE MODELAGEM DE DADOS - 24 HORAS

A criação de qualquer solução de Banco de Dados (BD) eficaz passa pela completa compreensão do problema a ser resolvido. E o pleno entendimento de qualquer situação só é possível quando se tem acesso às informações que são necessárias a isso. Na maior parte dos casos essas informações não estão disponíveis para consulta, dependem da habilidade do profissional de banco de dados em extraí-las a partir do contato com as áreas de negócio envolvidas.

Um Banco de dados (BD) é um conjunto de dados integrados que tem por objetivo atender a uma comunidade de usuários. O modelo de dados é a descrição formal das estruturas de dados para representação de um BD, com suas respectivas restrições e linguagem para criação e manipulação de dados. Um Sistema Gerenciador de Banco de Dados (SGBD) é um software que incorpora as funções de definição, recuperação e alteração de dados em um BD. A disciplina de Conceitos de Modelagem de Dados tem por objetivos: (a) apresentar os conceitos de sistemas de Gerenciadores de Bancos de Dados (SGBD); (b) Construir o Projeto conceitual BD cuja ação produz o esquema de dados abstratos que descreve a estrutura de um BD de forma independente de um SGBD (esquema conceitual); (c) Construir o Projeto lógico BD cuja ação produz o esquema lógico de dados e representa a estrutura de dados de um BD de acordo com o modelo de dados subjacente a um SGBD. Na disciplina utiliza-se ferramentas gratuitas como Draw.io e Lucidchart e também Microsoft SQL Server para demonstrar a criação de um BD a partir dos diagramas criados.

PYTHON PARA DATA SCIENTE - 24 HORAS

Quando se fala em análise de dados, faz-se necessário conhecer uma linguagem de programação que ajuda a entender os dados, a ler as informações que estão ocultas até se estudar os dados.
Sendo assim, a utilização de Python para fazer essa tarefa é super interessante, por se tratar de uma linguagem de fácil aprendizado, que roda em qualquer sistema operacional e de grande utilização pelas empresas. Python se mantém na liderança da pesquisa realizada pela IEEE Spectrum como uma das principais linguagens de programação.
https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages

Além de atender sistemas embarcados, inclusive demanda de IoT, o interesse em grandes conjuntos de dados está cada vez mais voltado para suas aplicações em aprendizado de máquina - a existência de bibliotecas Python de alta qualidade para estatísticas e aprendizado de máquina tem tornado o Python flexível e mais atraente do que R, por exemplo, segundo a Spectrum. Nesta disciplina vamos conhecer: Variáveis e tipos de dados; Estruturas lógicas ou de controle de fluxo: estruturas sequenciais; Estruturas condicionais ou de seleção: Simples, Composta, Encadeada ou Aninhada; Estruturas lógicas de repetição: repetição com teste no início; Repetição com variável de controle; Métodos e funções em Python; Introdução às principais bibliotecas do Python para as etapas de ciência de dados: (1) Obtenção de dados, (2) Pré-processamento, (3) Análise e Modelagem e (4) Avaliação e Apresentação.

MODELAGEM DIMENSIONAL E ETL- 28 HORAS

As necessidades das organizações para gerar *insights* de negócio e obter informações valiosas que apoiem na tomada de decisão, demandam cada vez mais processos e sistemas integrados dentro das empresas e até mesmo entre seus parceiros de negócios externos. Para atender a esta constante demanda por informações, muitas técnicas e tecnologias são empregadas para facilitar e agilizar o processo de disponibilização de dados. Uma das principais tarefas de sistemas de Business Analytics é a extração, limpeza e carga de dados de sistemas legados, internos e externos em uma área de armazenamento que servirá como base para o desenvolvimento de análises históricas, geração de dashboards, modelos de machine learning, entre outras aplicações analíticas. O modelo de dados desse componente deve satisfazer dois requisitos fundamentais: ser de fácil interpretação para usuários de negócio e otimizada para consultas de grandes volumes de dados. Tradicionalmente a área de armazenamento é denominada Data Warehouse e o modelo utilizado é o modelo dimensional: star schema ou snowflake. Com a presença de Big Data é comum o uso de banco de dados NoSQL (ou não relacionais). Com isso dois desafios são apresentados: o primeiro é conhecer esses modelos para extração desses dados nos sistemas de Business Analytics e o segundo é utilizar esses bancos de dados na área de armazenamento para possibilitar análises efetivas nesse contexto.

A disciplina de modelagem Dimensional e ETL tem como objetivos: (a) Entender os componentes do *framework* de Business Intelligence/Business Analytics. (b) Explicar a utilização de modelagem dimensional e como ela ajuda a construção de sistemas. (c) Conhecer e aplicar o processo de modelagem dimensional. (d) Implementar modelos dimensionais em banco de dados relacionais. (e) Entender as limitações da modelagem dimensional e como modelos NoSQL podem auxiliar a superá-las. (f) Entender e modelar banco de dados NoSQL nas categorias: chave-valor, documento, colunar e de grafo. (g) Demonstrar os principais conceitos do processo de ETL (*Extract, Transform and Load*). (h) Noções sobre coleta de dados de sistemas transacionais e transformação/padronização e união de informações. (i) Gerenciamento de dados mestre e montagem de fluxos e/ou códigos para extração e carga de dados em ambientes analíticos. As ferramentas utilizadas pela disciplina são o diagramador draw.io, o banco de dados relacional SQL Server e a ferramenta de acesso SQL Server Management Studio.

ANÁLISE EXPLORATÓRIA DE DADOS- 24 HORAS

Analisar o comportamento de dados tem se tornado extremamente importante porque a economia global não é apenas diversificada, mas também é rápida em frações de segundo. Ser capaz de coletar dados com mais eficiência e produzir hipóteses necessárias pode dar uma vantagem competitiva. O objetivo do curso é fornecer insumos teóricos e práticos sobre como organizar e descrever conjuntos de dados, dominar os fundamentos básicos de estatística descritiva, fazer inferências e realizar testes estatísticos a fim de habilitá-lo a atender as principais demandas de análise de dados típicas do dia a dia das empresas, tais como: 1) Calcular e interpretar medidas de posição e dispersão; 2) Construir distribuição de frequências, apresentá-las em tabelas e gráficos e calcular e interpretar medidas descritivas; 3) Discutir correlação de variáveis e causa e efeito; 4) Fazer estimativas de parâmetros populacionais com base em amostras para equações de regressão linear; 5) Estabelecer testes de hipóteses para parâmetros. É adotada uma abordagem prática voltada à aplicação em problemas reais. Vamos explorar as bibliotecas de dados da linguagem Python.

LINGUAGENS SQL E NoSQL- 24 HORAS

Cerca de 80% do Big Data são dados não estruturados. Armazenar e processar esses dados em bancos relacionais não é uma tarefa viável, uma vez que não foram concebidos com esse objetivo. Exatamente aí os bancos de dados NoSQL estão sendo usados cada vez mais, para atender aplicações analíticas criadas na era do Big Data. Os bancos de dados NoSQL oferecem alta velocidade operacional e maior flexibilidade para desenvolvedores de software e outros usuários quando comparados a bancos de dados tabulares (ou SQL) tradicionais. As estruturas de dados usadas pelos bancos de dados NoSQL são: valor-chave, coluna-larga, gráfico ou documento, e diferem dos bancos relacionais. Os bancos de dados NoSQL podem ser dimensionados em milhares de servidores, embora às vezes com perda de consistência de dados. Mas o que torna os bancos de dados NoSQL especialmente relevantes hoje é que eles são particularmente adequados para trabalhar com grandes conjuntos de dados distribuídos, o que os torna uma boa opção para grandes projetos de dados e análises. Ainda assim, os bancos de dados relacionais que utilizam SQL continuam com sua importância no Big Data, principalmente quando considera-se a entrega de informação para o usuário final. O objetivo dessa disciplina é apresentar os conceitos e princípios da linguagem de consulta a banco de dados (SQL), prover os alunos de conhecimentos necessários para criação de objetos (tabelas, views), prover os alunos de experiência para o desenvolvimento de consultas simples e especialmente consultas analíticas, transmitir aos alunos os conceitos e princípios de NoSQL, ensinar como utilizar bancos de dados e ferramentas de NoSQL, principalmente MongoDB.

DATAMINING & MACHINE LEARNING - 24 HORAS

Como implementamos processos e algoritmos personalizados que satisfazem as necessidades de dados de nossos negócios? Essa disciplina tem os seguintes objetivos: (1) Conhecer os principais conceitos de Mineração de Dados; (2) Conhecer e aplicar as etapas do processo de mineração de dados: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento; (3) As principais tarefas ou métodos de Mineração de Dados: algoritmos de regressão e classificação, algoritmos de agrupamento e associação, algoritmos de indução de regras e árvores de decisão; (4) Saber aplicar Mineração de Dados em problemas reais; (5) Conhecer e aplicar metodologias de desenvolvimento de projetos de Mineração de Dados. Nessa disciplina usamos as ferramentas Scikit-learn e Rapidminer, líder no quadrante mágico do Gartner para plataformas de ciência de dados e de aprendizado de máquina.

CLOUD COMPUTING - 20 HORAS

Espera-se que haja entendimento do modelo de Cloud Computing, das tecnologias envolvidas, do seu impacto nos sistemas corporativos, as tendências de desenvolvimento e suas limitações. Para tanto, nessa disciplina os assuntos abordados são: (1) Conceitos: nuvem, computação em nuvem, nuvem como serviço - XaaS, tipos de serviços em nuvem (modelo de software como serviço - SaaS, plataforma como serviço - PaaS, modelo de infraestrutura como serviço - IaaS); (2) Fundamentos de computação distribuída para computação em nuvem: computação paralela, distribuída, grids e cluster; (3) Fundamentos básicos de sistema operacional Linux; (4) Tecnologias para computação em nuvem: aprender a fazer o provisionando

de recursos para a computação em nuvem, como a criação de VMs, configuração das propriedades de rede, bem como acesso remoto à máquina virtual.

DATALAB – CASES EM MACHINE LEARNING - 20 HORAS

Essa disciplina aborda práticas de como os diferentes algoritmos vistos na disciplina de Data Mining e Machine Learning podem ser selecionados e implementados para satisfazer as necessidades de dados de diferentes áreas de negócios em um ciclo de vida completo de análise de dados. Assim, o aluno vai desenvolver aplicações de Analytics bem difundidas no mercado, como para as áreas de: crédito (regressão e classificação), marketing (regressão e indução de regras) e cobrança (regressão, classificação e indução de regras). Fazer cada projeto com a definição do problema, análise das bases e desenvolvimento da solução de Analytics. Conceber um projeto completo de Analytics com orientação, acompanhamento e apresentação dos projetos. Nessa disciplina usamos as ferramentas Scikit-learn e Rapidminer, MongoDB, SQL Server.

ARQUITETURA E TECNOLOGIAS EM BIG DATA - 20 HORAS

As tecnologias de Big Data possibilitam a utilização de informações disponíveis por meio de um sistema integrado e distribuído, possibilitando a coleta de dados nos formatos estruturados, não estruturados e/ou semiestruturados, consolidando-os em um repositório denominado *Data Lake*. Estes dados podem ser transações, interações NRT (*Near Real Time*) e sistemas de observações com as seguintes características: Dados de transações: podem ser dados históricos de transações dos principais sistemas de aplicativos de negócios, Dados estruturados em repouso e Dados que são capturados, mas não utilizados.

Dados de Interação NRT: Dados de mídia sociais - dados do LinkedIn, Twitter, Facebook, etc, e conteúdo da Web - fluxo de cliques, registros da web, vídeo, imagens entre outros. Dados de observações: “Internet das Coisas” - dados de transmissão e dados gerados por máquina - dados gerados a partir de sensores, GPS, RFID, dispositivos móveis, etc. Todas estas capacidades disponibilizadas por soluções de Big Data são possibilitadas por apresentarem economias de escala, agilidade e, principalmente, desempenho computacional.

Os objetivos desta disciplina são: comparar as arquiteturas de BI tradicional com as novas arquiteturas para big data, apresentar os principais componentes do ecossistema Hadoop (HDFS, Sqoop, Hive), Arquitetura Lambda e cloud computing para coleta e processamento de dados, conhecer os princípios teóricos que norteiam as arquiteturas técnicas de Big Data (ACID, CAP, multi-tenancy, replication factor), além de discutir os principais tópicos técnicos que afetam provisionamento, desempenho e manutenção das soluções de Big Data.

WEB MINING & SOCIAL NETWORKING ANALYSIS - 28 HORAS

Os objetivos da disciplina são: (1) Conhecer os conceitos e as técnicas que envolvem em Text Mining, Web Mining e Social Network Analysis (SNA); (2) Projetar e implementar robôs de busca e processamento de dados não estruturados (crawler e scraping); (3) Projetar e implementar algoritmos computacionais de classificação de textos, recuperação e extração de informação e conhecimento em bases de dados

textuais, não estruturadas e presentes na web; (4) Desenvolver aplicações em tempo real para busca de sites, wikis, tweets e APIs; (5) Desenvolver projeto de mineração de texto e mineração web na plataforma RapidMiner e com Python usando frameworks como NLTK, Spacy BeatifulSoup e SKLearn, com SQLServer e MongoDB.

DATA VISUALIZATION & STORYTELLING - 24 HORAS

Essa disciplina aborda como você pode visualizar e apresentar dados para diferentes públicos. Introduz a visão multidimensional, permitindo que as informações sejam dispostas em diferentes perspectivas e possam ser utilizadas para contar uma história, tornando a decisão mais natural, fácil e intuitiva, auxiliando o tomador de decisão já que se pode verificar tendências nos dados a fim de extrair o máximo de insights sobre eles, assim como estabelecer relevância para os mesmos. Aprender como organizar as diferentes formas de apresentar a visualização dos dados por meio de gráficos, utilizando uma ferramenta atual e de grande demanda pelo mercado, com base em especificações de projeto. Storytelling com Dados e com técnicas de comunicação que permita alavancar a visualização de dados e o raciocínio lógico. Conhecer a biblioteca Python Seaborn baseada em Matplotlib de alto nível para visualização de dados e a biblioteca para visualização de dados interativa Bokeh, voltada para execução em Web Browser. As ferramentas de mercado escolhidas para a disciplina são o Tableau e o Power BI, por serem os líderes de mercado pelo quadrante mágico do Gartner.

DATALAB – BUSINESS ANALYTICS - 28 HORAS

Essa disciplina aborda práticas avançadas de como os diferentes algoritmos vistos na disciplina de Data Mining e Machine Learning e em Datalab - Cases de Machine Learning, podem ser selecionados e implementados para satisfazer as necessidades de dados de diferentes áreas de negócios em um ciclo de vida completo de análise de dados.

Desenvolver aplicações de Analytics bem difundidas no mercado, como para as áreas de: crédito (regressão e classificação), marketing (regressão e indução de regras), cobrança (regressão, classificação e indução de regras).

Cada projeto conta com a definição do problema, análise das bases e desenvolvimento da solução de Analytics. O aluno vai conceber um projeto completo de Analytics com orientação, acompanhamento e apresentação dos mesmos. Nessa disciplina usamos as ferramentas Scikit-learn e Rapidminer, MongoDB, SQL Server.

BIG DATA GOVERANCE - 24 HORAS

A Governança de Dados é uma coleção de práticas e processos que ajudam a garantir o gerenciamento formal de ativos de dados dentro de uma organização.

Os objetivos da disciplina são: (a) Reconhecer a importância da qualidade de dados em organizações e implicações da qualidade em projetos convencionais e de Big Data; (b) Definir conceitos fundamentais da qualidade de dados; (c) Definir gestão de dados e funções associadas; (d) Explicar as funções de gestão de dados presentes no DMBOK (Data Management Body of Knowledge) fornecidos pela DAMA (Data Management Association International International), com ênfase em Master Data Management (MDM) e qualidade e governança de dados; (e) Planejamento, implementação e desenvolvimento de projetos que assegurem a qualidade de dados,

dados mestre e governança de dados; (f) Diferenciar, comparar e julgar soluções de mercado e metodologias de implementação para governança de dados em projetos Big Data e convencionais.

A ferramenta usada é a aplicação de análise de qualidade de dado Quadient® DataCleaner, com excel e MongoDB.

CORPORATE PERFORMANCE MANAGEMENT - CPM - 24 HORAS

Apresentar e discutir os principais conceitos e sistemas de Gestão de Performance Corporativa, CPM (Corporate Performance Management) ou EPM (Enterprise Performance Management), focando no aprendizado dos conceitos e de tecnologias disponíveis para tomada de decisão.

Mostrar como utilizar Business Analytics e as soluções de EPM nos ambientes empresariais, para a adoção de métodos computacionais mais elaborados, com o objetivo de melhorar o suporte à decisão, sempre em conjunto com o julgamento humano.

Compreender como realizar a gestão de desempenho por meio de medidas, relatórios, dashboards e visualização de dados em gráficos, Balanced Scorecard e Key Performance Indicators em um contexto de fluxo contínuo de dados.

Estabelecer a relação de CPM com MDM (Gestão de Dados Mestres) em Data Governance (Governança de Dados) e como a utilização do Analytics Strategy (Estratégia Analítica) vem sendo incorporada nas organizações como ferramenta para capturar e explorar estratégias emergentes de acordo com as tendências do mercado.

Entender como as empresas estão adequando suas estratégias para atender a GDPR (Regulamento Geral de Proteção a Dados), a mais significativa legislação europeia referente à proteção de dados vigente dos últimos 20 anos.

PROJETO DE BIG DATA ANALYTICS - 24 HORAS

Desenvolver um projeto com um cliente real, como parte inicial do projeto de conclusão de curso, que pode envolver a modelagem, o projeto e a utilização de algoritmos computacionais de classificação de textos, recuperação e extração de informação e conhecimento em bases de dados textuais não estruturadas e presentes na web.
