



# Pontifícia Universidade Católica do Paraná

## Plano de Ensino

Escola/ Câmpus:	Escola Politécnica			
Cursos:	Bacharelado em Ciência da Computação e Bacharelado em Cibersegurança	Ano/Semestre:	2023/1	
Código/Nome da disciplina:	Big Data			
Carga Horária:	80 h/a			
Requisitos:	Não há.			
Créditos:	4 créditos	Período: 5º	Turma: BCC	Turno: Noturno
Professor Responsável:	Jean Paul Barddal			

### 1. Ementa:

A disciplina de Big Data é de natureza teórico-prática e é ofertada para estudantes dos cursos de Bacharelado em Ciência da Computação, Bacharelado em Cibersegurança, Bacharelado em Engenharia de Software, Bacharelado em Sistemas de Informação e Tecnólogo em Segurança da Informação. Ao longo do semestre, o estudante identifica e caracteriza dados a partir de conceitos fundamentais, desenvolve algoritmos usando paradigmas específicos e aplica ferramentas para o armazenamento e processamento distribuído e eficiente de dados semi e não estruturados. Ao final da disciplina, o estudante é capaz de atuar em projetos complexos envolvendo grandes volumes de dados de estruturação não convencional. **Importante: esta disciplina é ofertada no nível I do programa Global Classes.**

#### (Syllabus)

Big Data is a theoretical and practical discipline that is offered to the Bachelor in Computer Science, Bachelor in Cybersecurity, Bachelor in Software Engineering, Bachelor in Information systems, and Technology in Information Security degrees students. Throughout the semester, the student identifies and characterizes data according to fundamental concepts, develops algorithms using specific paradigms, and applies tools for efficient semi-structured and unstructured distributed data storage and processing. At the end of the discipline, the student can work in complex projects that involve large data volumes that are unconventionally structured.

## 2. Relação com disciplinas precedentes e posteriores

### Disciplinas anteriores

Neste item, todos os cursos partilham o mesmo conjunto de disciplinas, que é descrito abaixo:

- Raciocínio Algorítmico
- Programação Orientada a Objetos
- Banco de Dados
- Resolução de Problemas Estruturados em Computação

### Disciplinas posteriores

- BCC: Data Science, Arquitetura de Sistemas Distribuídos Paralelos e Concorrentes
- BCS e TSI: Machine Learning e Experiência Criativa: Usando o Mindset do Adversário na Detecção e Prevenção de CiberAtaques
- BES: Arquitetura de Software, Gestão de Programas e Portfólios de TI, Controle Estatístico de Processos e Medição e Análise de Projetos e Processos de Software
- BSI: Analytics para Big Data, Sistemas de Informação Inteligentes

## 3. Temas de estudo

Abaixo são apresentados os principais temas da disciplina de Big Data, seguidos de seus respectivos tópicos teóricos e práticos.

- **Tema 1 (TE1): Fundamentos de Big Data**
- **Tema 2 (TE2): Dados semi e não estruturados**
- **Tema 3 (TE3): Armazenamento em Big Data**
- **Tema 4 (TE4): Processamento de Big Data**

#### 4. Resultados de Aprendizagem

##### **Bacharelado em Ciência da Computação**

Resultados de Aprendizagem	Temas de Estudo	Elemento de Competência (Internaliza, Mobiliza, Certifica) e Competência
RA1: Selecionar, de forma autônoma, soluções e formatos de processamento de bases de dados massivas, de acordo com os limites de computação e características de Big Data.	TE1, TE2, TE3 e TE4	<p>Competência A - Resolver problemas complexos não estruturados de solução algorítmica, considerando limites da computação, preceitos éticos e legais, de forma autônoma, crítica e inovadora.</p> <p>Elemento de competência A1 - Identificar as soluções algorítmicas viáveis para problemas complexos e não estruturados.</p>
RA2: Aplicar tecnologias para extração de informações a partir de arquivos persistidos em sistemas distribuídos específicos de dados.	TE2, TE3 e TE4	<p>Competência A - Resolver problemas complexos não estruturados de solução algorítmica, considerando limites da computação, preceitos éticos e legais, de forma autônoma, crítica e inovadora.</p> <p>Elemento de competência A2 - Selecionar algoritmos apropriados para problemas complexos e não estruturados.</p>
RA3: Implementar, de forma correta, soluções algorítmicas com dependabilidade para o processamento distribuído de dados semi e não estruturados em formato de Big Data.	TE1, TE2, TE3 e TE4	<p>Competência B - Projetar infraestrutura computacional sustentável, com segurança e dependabilidade, considerando tecnologias, estrutura organizacional e plano diretor de tecnologia da informação, implantando e monitorando sua execução de forma ética e resiliente.</p> <p>Elemento de competência B2 - Elaborar soluções tecnológicas eficientes para dados estruturados e não estruturados.</p>

**Bacharelado em Cibersegurança e Tecnologia em Segurança da Informação**

Resultados de Aprendizagem	Temas de Estudo	Elemento de Competência (Internaliza, Mobiliza, Certifica) e Competência
RA1: Selecionar, de forma autônoma, soluções e formatos de processamento de bases de dados massivas, de acordo com os limites de computação e características de Big Data.	TE1, TE2, TE3 e TE4	Competência A – Desenvolver sistemas seguros, seguindo padrões e boas práticas, de forma criativa, colaborativa e autorregulada.  Elemento de Competência A1 - Analisar o contexto do problema e recursos de hardware e software. Elemento de Competência A2 - Projetar sistemas de forma criativa.
RA2: Aplicar tecnologias para extração de informações a partir de arquivos persistidos em sistemas distribuídos específicos de dados.	TE2, TE3 e TE4	Competência A – Desenvolver sistemas seguros, seguindo padrões e boas práticas, de forma criativa, colaborativa e autorregulada.  Elemento de Competência A3 - Implementar sistemas seguros, seguindo padrões e boas práticas de forma colaborativa. Elemento de Competência A4 - Testar o sistema de forma autorregulada.
RA3: Implementar, de forma correta, soluções algorítmicas com dependabilidade para o processamento distribuído de dados semi e não estruturados em formato de Big Data.	TE1, TE2, TE3 e TE4	Competência A – Desenvolver sistemas seguros, seguindo padrões e boas práticas, de forma criativa, colaborativa e autorregulada.  Elemento de Competência A3 - Implementar sistemas seguros, seguindo padrões e boas práticas de forma colaborativa. Elemento de Competência A4 - Testar o sistema de forma autorregulada.

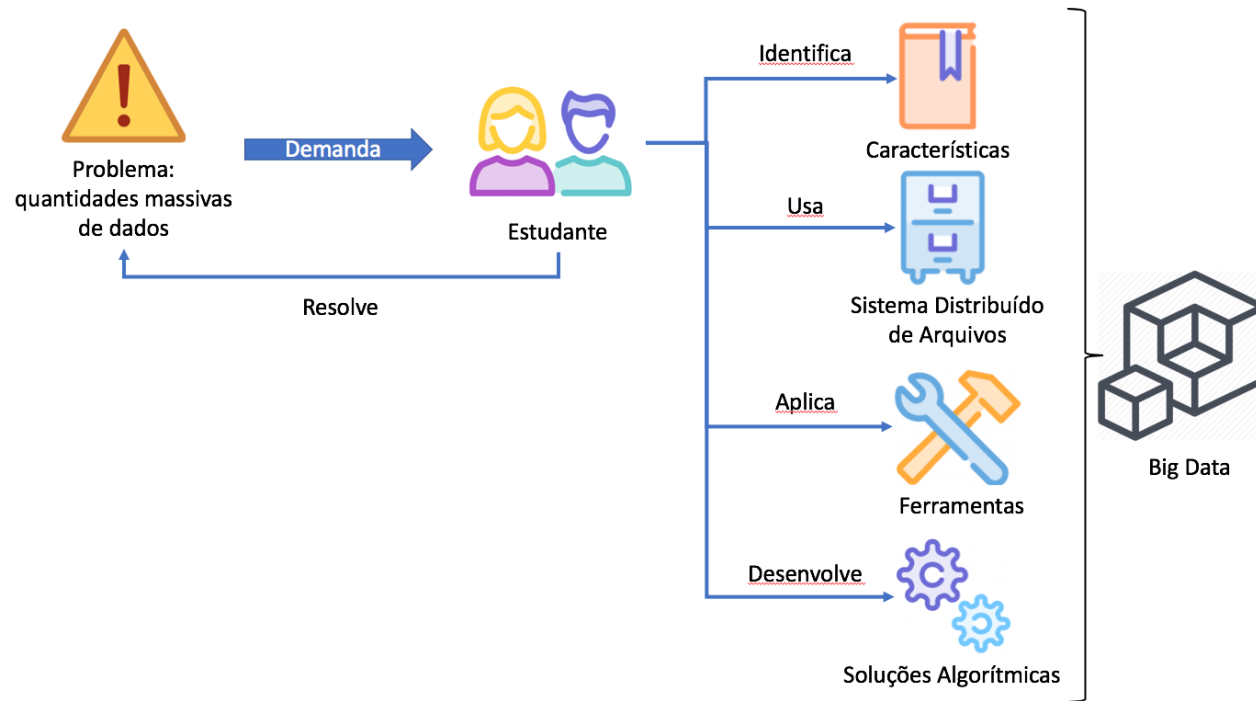
## **Bacharelado em Engenharia de Software**

Resultados de Aprendizagem	Temas de Estudo	Elemento de Competência (Internaliza, Mobiliza, Certifica) e Competência
RA1: Selecionar, de forma autônoma, soluções e formatos de processamento de bases de dados massivas, de acordo com os limites de computação e características de Big Data.	TE1, TE2, TE3 e TE4	<p>Competência A - Resolver problemas complexos não estruturados de solução algorítmica, considerando limites da computação, preceitos éticos e legais, de forma autônoma, crítica e inovadora.</p> <p>Elemento de competência A1 - Identificar as soluções algorítmicas viáveis para problemas complexos e não estruturados.</p>
RA2: Aplicar tecnologias para extração de informações a partir de arquivos persistidos em sistemas distribuídos específicos de dados.	TE2, TE3 e TE4	<p>Competência A - Resolver problemas complexos não estruturados de solução algorítmica, considerando limites da computação, preceitos éticos e legais, de forma autônoma, crítica e inovadora.</p> <p>Elemento de competência A2 - Selecionar algoritmos apropriados para problemas complexos e não estruturados.</p>
RA3: Implementar, de forma correta, soluções algorítmicas com dependabilidade para o processamento distribuído de dados semi e não estruturados em formato de Big Data.	TE1, TE2, TE3 e TE4	<p>Competência B - Projetar infraestrutura computacional sustentável, com segurança e dependabilidade, considerando tecnologias, estrutura organizacional e plano diretor de tecnologia da informação, implantando e monitorando sua execução de forma ética e resiliente.</p> <p>Elemento de competência B2 - Elaborar soluções tecnológicas eficientes para dados estruturados e não estruturados.</p>

### **Bacharelado em Sistemas de Informação**

Resultados de Aprendizagem	Temas de Estudo	Elemento de Competência (Internaliza, Mobiliza, Certifica) e Competência
RA1: Selecionar, de forma autônoma, soluções e formatos de processamento de bases de dados massivas, de acordo com os limites de computação e características de Big Data.	TE1, TE2, TE3 e TE4	<p>Competência A - Projetar soluções computacionais de acordo com especificações de requisitos, utilizando diretrizes da Engenharia de Software, considerando as tecnologias atuais de forma autorregulada.</p> <p>Elemento de competência A1 - Interpretar especificações variadas de requisitos de sistemas de informação.</p>
RA2: Aplicar tecnologias para extração de informações a partir de arquivos persistidos em sistemas distribuídos específicos de dados.	TE2, TE3 e TE4	<p>Competência A - Resolver problemas complexos não estruturados de solução algorítmica, considerando limites da computação, preceitos éticos e legais, de forma autônoma, crítica e inovadora.</p> <p>Elemento de competência A2 - Selecionar métodos, modelos e ferramentas apropriados para o projeto do sistema de informação utilizando diretrizes da Engenharia de Software.</p>
RA3: Implementar, de forma correta, soluções algorítmicas com dependabilidade para o processamento distribuído de dados semi e não estruturados em formato de Big Data.	TE1, TE2, TE3 e TE4	<p>Competência A - Resolver problemas complexos não estruturados de solução algorítmica, considerando limites da computação, preceitos éticos e legais, de forma autônoma, crítica e inovadora.</p> <p>Elemento de competência A3 - Elaborar soluções tecnológicas eficientes para dados estruturados e não estruturados.</p>

## 5. Mapa Mental



## 6. Metodologia e Avaliação

Os resultados de aprendizagem desta disciplina serão trabalhados de acordo com a tabela apresentada abaixo. Nesta tabela, os resultados de aprendizagem (RAs) são relacionados com os seus respectivos indicadores de desempenho (IDs), os métodos utilizados para a apresentação conceitual e prática dos conteúdos, assim como os processos de avaliação via metodologias diagnósticas, formativas e somativas; a serem realizadas, sempre visando feedback rápido e efetivo para os estudantes.

Focando nas avaliações, estas podem ser diagnósticas, formativas ou somativas, conforme detalhamento abaixo:

- **Avaliações diagnósticas:** atividades com objetivo de acompanhar o aprendizado dos temas discutidos até o momento visando feedback imediato e de forma coletiva. A sugestão aqui é utilizar métodos como a linha da vida para engajar os estudantes a responder, mesmo que de forma não verbal, questões relacionadas ao conteúdo ministrado, permitindo ao professor elucidar conceitos que possivelmente estejam causando dúvidas e que os estudantes se auto-avaliem e identifiquem tópicos a serem verificados antes das avaliações somativas.
- **Avaliações formativas:** atividades com enfoque prático que objetiva, por meio do trabalho em equipe, aplicar os conceitos ministrados pelo professor em problemas do mundo real. Estas avaliações permitem ao professor prover feedback aos alunos durante a própria aula, ainda de forma coletiva, mas focada em grupos menores de estudantes. Este tipo de atividade também permitirá que avaliação por pares seja realizada, onde equipes validem os trabalhos realizados pelas demais.
- **Avaliações somativas:** atividades com o objetivo de quantificar a performance dos estudantes, de forma individual, nos conceitos teóricos e práticos que envolvem a disciplina. Sugere-se aqui a aplicação de avaliações no formato de provas. O feedback para este tipo de atividade será realizado de forma colaborativa, no sentido de que o professor apresentará a resposta para todas as questões da avaliação para todos os estudantes, seguido por um feedback individual para cada estudante. Objetiva-se realizar a devolutiva destas avaliações em um prazo de uma a duas semanas após sua realização.

Alinhamento Construtivo			
Resultado de aprendizagem	Indicadores de desempenho	Processos de Avaliação	Métodos ou técnicas empregados**
<b>RA1:</b> Selecionar, de forma autônoma, soluções e formatos de processamento de bases de dados massivas, de acordo com os limites de computação e características de Big Data.	<b>ID1.1:</b> Identifica, de forma autônoma, os Vs de Big Data em diferentes tipos de problemas computacionais.  <b>ID1.2:</b> Categoriza dados de acordo com seu nível de estruturação em	<b>Avaliação diagnóstica:</b> resolução de exercícios com cunho conceitual-teórico durante as aulas. <u>Feedback:</u> coletivo e imediato, durante as aulas.  <b>Avaliação formativa:</b> resolução de exercícios com simulação realística. <u>Feedback:</u> coletivo, em sala de aula.	<b>Métodos empregados:</b> PBL  <b>Meios de interação:</b> Zoom, Kahoot e Blackboard (caso a situação sanitária demande)



	<p>diferentes problemas computacionais.</p> <p><b>ID1.3:</b> Associa o formato adequado de processamento de dados de acordo com o problema a ser solucionado.</p>	<p><b>Avaliação somativa:</b> Questões em provas escritas também com temática de simulação realística em formato de PBL.</p> <p><u>Feedback:</u> individual.</p> <p><u>Recuperação:</u> ocorrerá no formato de trabalhos extra-classe ou de prova, de acordo com a necessidade de cada estudante.</p>	
<p><b>RA2:</b> Aplicar tecnologias de forma ética para extração de informações a partir de arquivos persistidos em sistemas distribuídos específicos de dados.</p>	<p><b>ID2.1:</b> Armazena bases de dados massivas em um sistema distribuído de arquivos de acordo com sua estruturação.</p> <p><b>ID2.2:</b> Aplica SQL estendido, de forma ética, para extrair informações a partir de arquivos dispostos em sistemas de arquivos para Big Data.</p> <p><b>ID2.3:</b> Cria rotinas usando paradigma de programação distribuída para extrair informações a partir de dados semi e não estruturados dispostos em sistemas distribuídos específicos de arquivos.</p>	<p><b>Avaliação diagnóstica:</b> resolução de exercícios com cunho conceitual-teórico durante as aulas.</p> <p><u>Feedback:</u> coletivo e imediato, durante as aulas.</p> <p><b>Avaliação formativa:</b> resolução de exercícios com simulação realística.</p> <p><u>Feedback:</u> coletivo em sala de aula.</p> <p><b>Avaliação somativa:</b> trabalho de implementação de projeto realizado em equipes.</p> <p><u>Feedback:</u> coletivo, para a equipe, com vistoria de código individualizada.</p> <p><u>Recuperação:</u> ocorrerá no formato de trabalhos extra-classe ou de prova, de acordo com a necessidade de cada estudante.</p> <p><b>Avaliação somativa:</b> questões em provas escritas.</p> <p><u>Feedback:</u> individual.</p> <p><u>Recuperação:</u> ocorrerá na semana de recuperação.</p>	<p><b>Métodos empregados:</b> PBL, Peer Review</p> <p><b>Meios de interação:</b> Zoom, Kahoot e Blackboard (caso a situação sanitária demande)</p>
<p><b>RA3:</b> Implementar, de forma correta, soluções algorítmicas com dependabilidade para o processamento distribuído de dados semi e não estruturados em formato de Big Data.</p>	<p><b>ID3.1:</b> Codifica soluções algorítmicas eficientes para processamento distribuído de dados.</p>	<p><b>Avaliação diagnóstica:</b> resolução de exercícios com cunho conceitual-teórico durante as aulas.</p> <p><u>Feedback:</u> coletivo e imediato, durante as aulas.</p>	<p><b>Métodos empregados:</b> PBL, Peer Review</p> <p><b>Meios de interação:</b> Zoom, Kahoot e Blackboard (caso a situação sanitária demande)</p>

	<p><b>ID3.2:</b> Testa, de forma crítica, soluções algorítmicas usando técnicas de depuração e amostragem.</p>	<p><b>Avaliação formativa:</b> resolução de exercícios com simulação realística.  <u>Feedback:</u> coletivo em sala de aula.</p> <p><b>Avaliação somativa:</b> trabalho de implementação de projeto realizado em equipes.  <u>Feedback:</u> coletivo, para a equipe, com vistoria de código individualizada.  <u>Recuperação:</u> ocorrerá no formato de trabalhos extra-classe ou de prova, de acordo com a necessidade de cada estudante.</p> <p><b>Avaliação somativa:</b> questões em provas escritas.  <u>Feedback:</u> individual.  <u>Recuperação:</u> ocorrerá na semana de recuperação.</p>	
--	--	--	--

Para **aprovação** nesta disciplina, assume-se os seguintes critérios de aprovação:

- Caso o estudante **não** atinja a nota necessária em um RA durante uma das atividades propostas, haverá uma oportunidade de recuperação ainda durante o semestre. A única ressalta é a última prova, por questões de cronograma.
- A recuperação de um RA está disponível para notas inferiores a 7,0 (sete). Notas superiores a 7,0 (sete) inviabilizam a realização da recuperação.
- O estudante poderá recuperar, **no máximo**, nota 7,0 (sete) para o RA em questão. Isso se dá de forma independentemente se a recuperação ocorre durante ou no final do semestre.
- Qualquer atividade somativa perdida somente poderá ser feita mediante Recuperação do RA. Provas de segunda chamada só serão realizadas após indicação formal da coordenação.
- Caso, mesmo após a recuperação ofertada durante o semestre, o estudante **não** atinja a nota mínima esperada em um ou mais RAs, será oportunizada uma semana de Recuperação, onde o estudante terá a oportunidade de recuperar os resultados não atingidos por meio de atividades específicas.
- O mapeamento entre cada RA e os seus respectivos pesos é fornecido na tabela abaixo.

<b>Resultado de Aprendizagem (RA)</b>	<b>Avaliação somativa – Trabalho MapReduce (Peso no RA)</b>	<b>Avaliação somativa – Trabalho Spark (Peso no RA)</b>	<b>Avaliação somativa – Trabalho SparkSQL (Peso no RA)</b>	<b>Avaliação somativa – Prova 01 (Peso no RA)</b>	<b>Avaliação somativa – Prova 02 (Peso no RA)</b>	<b>Nota do RA (0 a 10)</b>
<b>RA1</b>	2,00	N/A	N/A	8,00	N/A	Soma das notas obtidas nas atividades
<b>RA2</b>	2,00	2,00	2,00	2,00	2,0	Soma das notas obtidas nas atividades
<b>RA3</b>	2,50	2,50	2,50	N/A	2,50	Soma das notas obtidas nas atividades

- A nota final do estudante na disciplina é calculada de acordo com a média das notas obtidas em cada um dos RAs, conforme a equação abaixo:

$$\text{Nota Final} = (0.2 * RA_1) + (0.3 * RA_2) + (0.5 * RA_3)$$

## 7. Cronograma de atividades

O cronograma previsto para a realização desta disciplina é fornecido abaixo. Mudanças no cronograma poderão ocorrer de acordo com o andamento da disciplina e o cronograma abaixo serve para guiar o planejamento do curso.

<b>Período</b> (horas aula totais, dia, semana, quinzena, mês)	<b>RAs</b>	<b>Atividades pedagógicas</b> (* = atividades que vão gerar entregas para atribuir frequência)	<b>Em aula/ Pré-aula/ Pós-aula/ TDE</b>	<b>Carga horária da atividade</b>
9 de março de 2023	RA1	Apresentação do plano de ensino, ementa, formato de avaliação. Acolhida institucional.	Em aula	4h/a
16 de março de 2023	RA1, RA2, RA3	Contextualização sobre Big Data. Os Vs de Big Data. Principais ferramentas e ecossistemas para Big Data. Características e estruturação de dados. Formatos de processamentos de dados: em lote (batch) ou em stream. O ecossistema Hadoop. Sistema distribuído de arquivos. Visão geral sobre frameworks de Big Data.	Em aula/TDE	4h/a
23 de março de 2023	RA1, RA2, RA3	Paradigma MapReduce. Setup do ambiente para programação MapReduce. Entendimento e Implementação em MapReduce.	Em aula/TDE	4h/a
30 de março de 2023	RA2, RA3	Implementação em MapReduce. Writables e Custom Writables.	Em aula/TDE	4h/a
13 de abril de 2023	RA2, RA3	Otimização de rotinas MapReduce com Combiner.	Em aula/TDE	4h/a
20 de abril de 2023	RA2, RA3	Concatenação de rotinas MapReduce. Atividade prática de MapReduce.	Em aula/TDE	4h/a

27 de abril de 2023	RA1, RA2, RA3	Avaliação somativa em formato de prova 01. Entrega do trabalho de MapReduce.	Em aula/TDE	4h/a
4 de maio de 2023	RA2, RA3	Feedback da prova 01. Visão geral sobre Spark. Setup do ambiente de programação para Spark. Spark e Paired Resilient Distributed Dataset. Programação Spark: RDDs, transformações e operações.	Em aula/TDE	4h/a
11 de maio de 2023	RA2, RA3	Spark: Programação com RDDs.	Em aula/TDE	4h/a
18 de maio de 2023	RA2, RA3	Spark: Programação com PairRDDs.	Em aula/TDE	4h/a
25 de maio de 2023	RA2, RA3	SparkSQL.	Em aula/TDE	4h/a
1 de junho de 2023	RA2, RA3	Avaliação somativa em formato de prova 02. Entrega do trabalho de SparkSQL.  Recuperação do RA1.	Em aula/TDE	4h/a
15 de junho de 2023	RA1, RA2, RA3	Feedback da prova 02 e feedbacks dos projetos de Spark e SparkSQL.	Em aula/TDE	4h/a
22 de junho de 2023	RA2, RA3	Recuperação dos RAs 2 e 3.	Em aula/TDE	4h/a
29 de junho de 2023 (Semana de recuperação estendida)	De acordo com a necessidade de cada estudante	Semana de recuperação estendida.	Em aula	4 h/a

## 8. Bibliografia

Todas as atividades contarão com o apoio de material preparado pelo professor sob a forma de slides, exercícios implementados em linguagem de programação durante as aulas e disponibilizados no ambiente virtual de aprendizagem, assim como referências para livros e artigos científicos publicados em conferências e revistas especializadas na área.

Os livros apresentados como básicos abaixo podem ser consultados na biblioteca da PUCPR pelos estudantes como fonte complementar:

### Básica

ROSANGELA MARQUESONE. BIG DATA: TÉCNICAS E TECNOLOGIAS PARA EXTRAÇÃO DE VALOR DOS DADOS. SÃO PAULO: CASA DO CÓDIGO, 2018.

CEZAR TAURION. BIG DATA – BRASPOST, 2015.

TOM WHITE. HADOOP: THE DEFINITIVE GUIDE (INGLÊS) - O'REILLY, 2015.

VIKTOR MAYER-SCHÖNBERGER, KENNETH CUKIER. BIG DATA – CAMPUS-ELSEVIER, 2013.

### Complementar

FELIPE NERY RODRIGUES MACHADO. BIG DATA - O FUTURO DOS DADOS E APLICAÇÕES – ERICA/SARAIVA, 2018.

ATOS RAMOS. INFRAESTRUTURA BIG DATA COM OPENSOURCE - CIÊNCIA MODERNA, 2015.

JUDITH HURWITZ, MARCIA KAUFMAN. BIG DATA PARA LEIGOS – ALTA BOOKS, 2015.

EDWARD CAPRIOLO, DEAN WAMPLER, JASON RUTHERGLEN. PROGRAMMING HIVE (INGLÊS). O'REILLY, 2017.

MARK HAMSTRA, MATEI ZAHARIA, HOLDEN KARAU. LEARNING SPARK: LIGHTNING-FAST BIG DATA ANALYSIS (INGLÊS) – O'REILLY, 2016.

LAZYPROGRAMMER. BIG DATA, MAPREDUCE, HADOOP, AND SPARK WITH PYTHON: MASTER BIG DATA ANALYTICS AND DATA WRANGLING WITH MAPREDUCE FUNDAMENTALS USING HADOOP, SPARK, AND PYTHON (INGLÊS). DISPONÍVEL EM FORMATO AMAZON KINDLE. 2017.

DONALD MINER, ADAM SHOOK. MAPREDUCE DESIGN PATTERNS: BUILDING EFFECTIVE ALGORITHMS AND ANALYTICS FOR HADOOP AND OTHER SYSTEMS. O'REILLY, 2012.

OFER MENDELEVITCH, CASEY STELLA, DOUGLAS EADLINE. PRACTICAL DATA SCIENCE WITH HADOOP AND SPARK: DESIGNING AND BUILDING EFFECTIVE ANALYTICS AT SCALE (ADDISON-WESLEY DATA & ANALYTICS) (ENGLISH EDITION). 1ST EDITION – ADDISON-WESLEY PROFESSIONAL, 2016.

## **9. Acessibilidade**

Não houve necessidade de adaptação.

## **10. Adaptações para práticas profissionais**

Não houve necessidade de adaptação.