WBA0753 v1.0

Linguagens de programação para ciência de dados (*Python* com *Spark*)

Autoria do Desafio Profissional: Danilo Rodrigues Pereira

Leitor Crítico: Adriano Thomaz

Desafio Profissional

1. Caso – *Pipeline* para extração de informações em arquivos de *logs Apache* utilizando *Python* e *Apache Spark*

A quantidade de dados criados e armazenados globalmente continua crescendo a cada ano. Essa crescente criação de dados pelas mídias sociais, aplicativos de negócios e telecomunicações e vários outros domínios está levando à formação de *Big Data* (CHAMBERS, 2018). Entretanto, não é a quantidade de dados disponíveis que importa para as empresas, mas sim o que vamos fazer com esses dados para conseguir extrair informações importantes (conhecimentos).

Diante desse cenário, uma empresa de Tecnologia da Informação (TI) deseja analisar informações (*erros 404*, *host*, problemas de autenticação e modificações de perfil de usuário e permissões em arquivos de *logs Apache*). Os arquivos de *logs* são muito importantes para a segurança de dados das empresas, pois é nesses arquivos que fica registrado tudo o que acontece nos servidores de aplicação e nos sistemas.

Para gerenciar efetivamente um servidor da web, é necessário obter feedback sobre as atividades, desempenho do servidor e quaisquer problemas que possam estar ocorrendo ou que ocorreram. O servidor HTTP *Apache* fornece recursos de log muito abrangentes e flexíveis.

Diante das informações anteriores, você deverá propor um *pipeline* (etapas) para extração de informações de arquivos de *logs Apache* usando *Python* e *Apache Spark*.

Para resolver este desafio profissional, você deverá ler com atenção os conteúdos da disciplina (disponível no ambiente virtual) e aprofundar os estudos mediante leituras complementares.

2. Papel do aluno na resolução do problema

Imagine que você é o cientista de dados e seu objetivo é implementar soluções para gerenciamento de arquivos de *logs* de forma eficiente. Neste estudo de caso, o trabalho do cientista de dados da empresa é criar um *pipeline* para extração de informações dos arquivos de *logs Apache*. A automatização deste processo será responsável por uma grande redução de custos e tempos, e por um significativo aumento de produtividade da equipe de desenvolvedores.

Para a criação do pipeline, você poderá seguir os seguintes passos:

1. Você precisa descobrir quais os tipos de *logs* que o servidor *Apache* disponibiliza.
2. Em seguida, você deve decidir se o processamento desses arquivos de *logs* será feito em lote (*batch*) ou em tempo real (*streaming*).
3. Listar todas as bibliotecas *Python* (*PySpark*, *Pandas*, *Matplotlib*, *SciPy*, *Numpy*, entre outras), que poderão utilizadas durante o desenvolvimento.
4. Criar *scripts Python* para leitura dos arquivos de *logs*.
5. Filtrar nos arquivos as informações que a empresa deseja analisar.
6. Gerar relatórios com análise estatística.

3. Resolução do Desafio Profissional

Caro(a) aluno(a)!
Lembre-se de que o conteúdo da disciplina deverá ser considerado no processo de resolução do desafio. Além disso, a **Biblioteca Virtual** está à disposição para pesquisas complementares.
Outro ponto importante é que o trabalho desenvolvido por você, no processo de resolução do desafio, deverá ser submetido à um processo de autoavaliação. O objetivo é estimular a autocrítica e reflexão sobre o próprio desempenho a fim de aprimorar sua autonomia e envolvimento pelo próprio aprendizado.
Para isso, você deverá levar em consideração os itens dispostos na grade de autoavaliação que se encontra disponível a seguir.

4. Grade de autoavaliação

Como o objetivo é estimular a autocrítica e a reflexão sobre o próprio desempenho a fim de aprimorar sua autonomia e envolvimento pelo próprio aprendizado, leve em consideração os itens dispostos na grade de autoavaliação e pontue o seu desempenho na resolução deste Desafio Profissional.

	Tema	Objetivos Gerais	Objetivos Específicos	Peso
1	Utilização dos referenciais teóricos	Verificar se os pressupostos teóricos presentes na Leitura Fundamental foram utilizados para o cumprimento da proposta.	1) Os pressupostos teóricos foram apreendidos? 2) A problematização do caso contribuiu para sua aprendizagem? 3) A problematização estimulou enriquecimento teórico/prático em relação à temática?	20
2	Execução da tarefa	Verificar se a execução da tarefa ocorreu de forma eficiente, conforme sua proposta.	1) Você atingiu os objetivos propostos? 2) O Desafio Profissional foi resolvido com base na fundamentação teórica e em pesquisas complementares? 3) Você considera sua capacidade de articulação dos conceitos mobilizados satisfatória? 4) Você se sentiria capaz de se posicionar e argumentar caso a situação apresentada fosse real?	30
3	Estrutura do trabalho final	Avaliar se o produto final apresentado como resolução do Desafio Profissional é satisfatório.	1) A resolução contempla as etapas explicitadas pelo Desafio Profissional? 2) O resultado final apresentado corresponde ao desafio apresentado? 3) O produto final elaborado por você é condizente com a proposta de solução?	30
4	Desafio	Avaliar se os objetivos de aprendizagem foram alcançados.	1) Você aplicou os conhecimentos teóricos da disciplina? 2) Considera que o trabalho final expressa o conhecimento construído por você em termos práticos e teóricos? 3) O trabalho final demonstra as habilidades e competências desenvolvidas a partir dos objetivos propostos pelo Desafio Profissional?	20
	TOTAL			**100**

Bons estudos!