Avaliação – Engenheiro de Dados

Nesta avaliação utilizaremos uma base de dados do que contém informações de posição dos sensores de acelerômetros e giroscópios de celulares e *smartwatches*. Acesse o link https://archive.ics.uci.edu/ml/datasets/Heterogeneity+Activity+Recognition para mais informações sobre a base de dados. Usaremos PySpark no ambiente do Google Colab para realização das tarefas. Para configurar o ambiente e fazer o download dos arquivos, use o seguinte trecho de código:

```
!pip install pyspark

from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("seuapp").getOrCreate()

!wget https://archive.ics.uci.edu/ml/machine-learning-databases/00344/Activity%20recognition%20exp.zip

!unzip /content/Activity\ recognition\ exp.zip -d /content/
```

Assim, teremos dentro da pasta *content* a pasta *Activity recognition exp,* que contém quatro arquivos .csv com os dados de acelerômetro para os relógios e os celulares e os dados de giroscópio para os celulares e relógios, além de um arquivo readme.txt com mais informações sobre a base. Nas tabelas temos informações que identificam os usuários ("User"), o modelo do aparelho ("Model") e o aparelho específico que gerou o registro ("Device"), além das informações de posição e horário de criação e recepção de cada registro. Além disso, temos a variável "Gt", que informa qual atividade o usuário estava fazendo no momento em que o registro foi coletado (subindo escada, andando, pedalando, etc.).

Usando PySpark no Colab faça as seguintes tarefas:

1. Carregue as bases de dados como DataFrames do PySpark. Especifique os formatos dos dados manualmente, não use a opção *inferSchema = True.*
2. Faça uma análise inicial dos dados: quais problemas você encontrou? Como você trataria tais problemas?
3. Crie uma tabela com informações sumarizadas de cada usuário, como: quantos registros tal usuário possui, quais modelos de aparelho cada usuário operou, etc.
4. Os campos "Arrival_Time" e "Creation_Time" apresentam dados de data usando o formato unix_timestamp, com precisão de milissegundo e nanosegundo, respectivamente. Crie novas colunas que apresentem tais informações no formato *string,* de forma a ser facilmente interpretável por um ser humano.
5. Crie um campo que tenha o intervalo de tempo entre o registro atual e o anterior, para um mesmo usuário e device. Avalie a performance da sua *query*: quais estratégias você usaria para otimizar o tempo de processamento?

São esperadas as seguintes entregas, até 1h antes da entrevista, via e-mail:

- Códigos em formato **.ipynb**, estruturado e amplamente comentados.
- Resumo em formato **.ppt** ou **.pptx** dos resultados obtidos (máx. 5 slides), a ser apresentado durante a entrevista.

Durante a entrevista, serão dados 10 minutos para o candidato apresentar os resultados obtidos.

Bom trabalho!