HTTP requests to NASA Kennedy Space Center WWW server

Datasets: Os dois conjuntos de dados não estruturados possuem todas as requisições HTTP para o servidor da NASA Kennedy Space Center WWW na Flórida para os períodos de Julho e Agosto de 1995.
Objetivo: Fazer análise exploratória dos dados com a finalidade de encontrar informações relevantes a respeito das requisições HTTP feitas na época.

Obtenção dos dados:

1. Fazer o download de cada um dos datasets por vez e salvar em um diretório de fácil acesso.

Para facilitar etapas futuras é interessante já deixar preparado o caminho para esses diretórios, para isso basta abrir o terminal do Ubuntu (CTRL + ALT + T), navegar até a pasta onde os arquivos foram salvos, digitar pwd e copiar o endereço indicado e colar em um bloco de notas.
Após fazer isso adicione o nome do arquivo ao final.

Exemplo de endereço a ser obtido:

/home/thais/Área de Trabalho/nasa_challenge/NASA_access_log_Aug95.gz

Ambiente
A análise exploratória foi feita utilizando o shell do Spark no Ubuntu versão 18.04.

Como utilizar o shell:

1. Acessar https://spark.apache.org/downloads.html
2. Baixar a release 2.4.4 (Aug 30 2019) que é a versão estável mais recente com o pacote Pre-built for Apache Hadoop 2.7.
3. Salvar em um diretório de sua preferência e descompactar
4. Abrir o terminal do Ubuntu e navegar até a pasta onde o executável spark-shell está baixado.
   Exemplo:
   /home/thais/Área de Trabalho/spark-2.4.4-bin-hadoop2.7/bin
5. Digitar ./spark-shell para acessar o ambiente interativo

Como rodar o código:

1. Fazer o download o código nasa_challenge.scala e salvar na mesma pasta em que os dados baixados no passo 1 da Obtenção dos dados foram salvos.
   Abrir o código e buscar pela função main para trocar o path de july_file e aug_file para os respectivos caminhos obtidos na etapa de obtenção de dados.
   "file:///…./…./Documentos/nasa_challenge/NASA_access_log_Jul95.gz"

2. Voltar ao spark shell e digitar :load -v endereço_do_código baixado

Exemplo

```
:load -v /home/thais/Documentos/nasa_challenge/nasa_challenge.scala
```

Funções Scala:

```
/*
```

A função cleanFile recebe uma String e retorna um DataFrame formatado, isto é, com colchetes desnecessários retirados, com a quantidade de colunas corretas e o restante das devidas correções para montar o DataFRame.

```
/*
```

A função replaceBracket recebe uma string e substitui todos os colchetes por um espaço. A função retorna a string substituída.

```
*/
def replaceBracket(bracket:String):String = {
   try{
        return bracket.replace("[", "")
   } catch {
        case e: Exception => bracket
   }
}
```

```
/*
```

A função formatData recebe uma string e retorna ela formatada em 8 colunas diferentes.

```
*/
def func(corrupt:String):Seq[String] = {
   try {
        val arr = corrupt.split(" ")
        val tam = arr.size-1
        return Seq(arr(0), arr(1), arr(2), arr(3), arr(4), arr.slice(5, tam-2).mkString(" "), arr(tam-1), arr(tam))
   } catch {
        case e: Exception => Seq("","","","","","","","")
   }
}
```