Obrázok, na ktorom je text, hodiny

Automaticky generovaný popisObrázok, na ktorom je logo

Automaticky generovaný popisEKONOMICKÁ UNIVERZITA

Fakulta hospodárskej informatiky

Katedra aplikovanej informatiky

# Analýza neštruktúrovaného a semištruktúrovaného zdroja pomocou Sparku

**Predmet:** Big Data

**Meno študentov:** Martin Jankech, Patrik Hajdučík

**Stupeň/ročník:** druhý/2.ročník

**Školský rok:** 2022/2023

**Meno prednášajúceho:** Ing. Jaroslav Kultán PhD.

**Meno cvičiaceho:** SCHMIDT, Peter, Ing. Mgr., PhD.

Obsah

[Analýza neštruktúrovaného a semištruktúrovaného zdroja pomocou Sparku 1](#_Toc132316444)

[Úvod 3](#_Toc132316445)

[Spark inštalácia 4](#_Toc132316446)

[Analáza log filu pomocou sparku 6](#_Toc132316447)

[Popis zdrojového kodu a dataset 6](#_Toc132316448)

# Úvod

V tomto zadaní sa budeme zaoberať analýzou logového súboru pomocou Apache Sparku v jazyku Python. Logové súbory sú neoddeliteľnou súčasťou moderných informačných technológií a zaznamenávajú informácie o rôznych udalostiach v systéme. Analýzou týchto súborov môžeme získať cenné informácie o používateľskom správaní, výkone aplikácie a ďalších dôležitých ukazovateľoch.

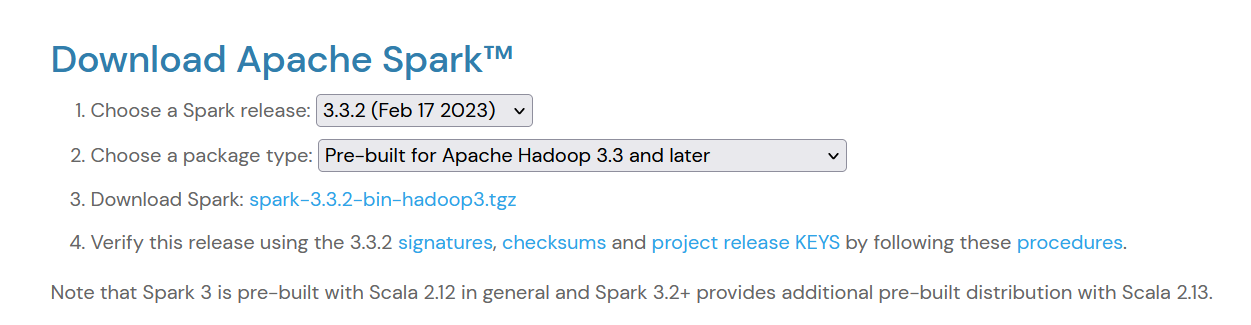
V rámci tohto projektu sa budeme zaoberať spracovaním veľkého objemu dát pomocou Apache Sparku, ktorý je distribuovaný výpočtový systém, umožňujúci rýchle a efektívne spracovanie veľkých dátových súborov. Výhodou Apache Sparku je jeho schopnosť paralelizovať úlohy a pracovať s dátami uloženými v pamäti, čo umožňuje výrazné zrýchlenie spracovania dát. V prvom ktorú si ukážeme ako apache Spark nainštalovať v prostredí windows. A potom sa budeme sústrediť na analýzu logového súboru pomocou Apache Sparku v jazyku Python. Python je jedným z najobľúbenejších programovacích jazykov v oblasti spracovania dát a má veľké množstvo knižníc a nástrojov, ktoré umožňujú rýchlu a efektívnu prácu s dátami.

# Spark inštalácia

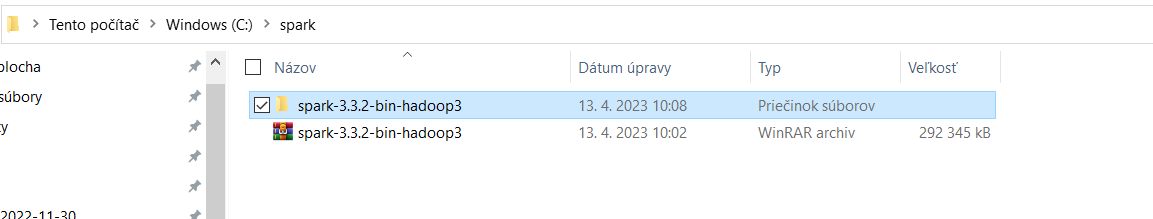
Treba mat nainštalovanú javu a python a hadoop

Stiahneme Apache Spark s oficiálnej stránky : <https://spark.apache.org/downloads.html>

My sme použili verziu pre hadoop 3.3.0



Na C disku vytvoríme priečinok spark a nakopíruje do neho stiahnutý Spark a extrahujeme ho



Ďalej je potrebné nastaviť užívateľské a systémové premenne prostredia

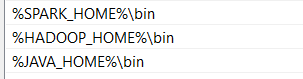
Používateľské premenne prostredia takto :



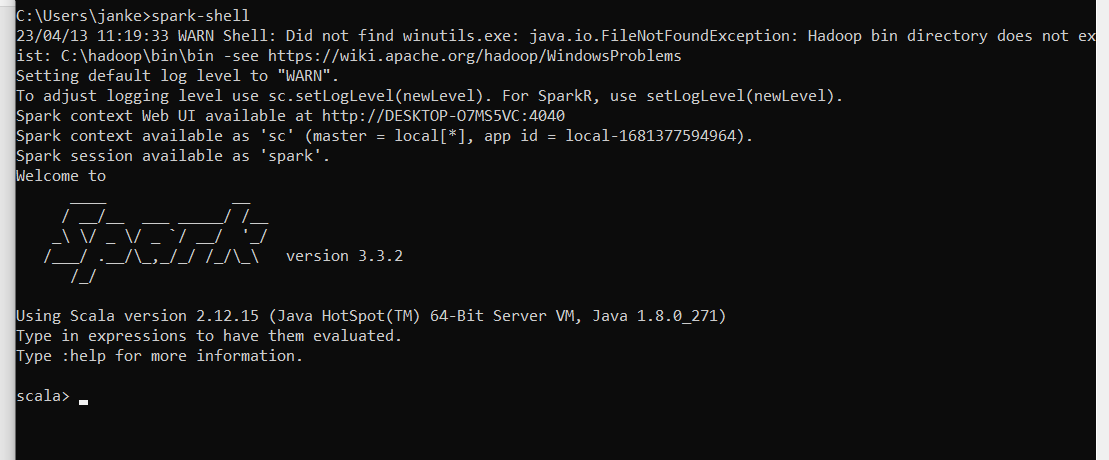




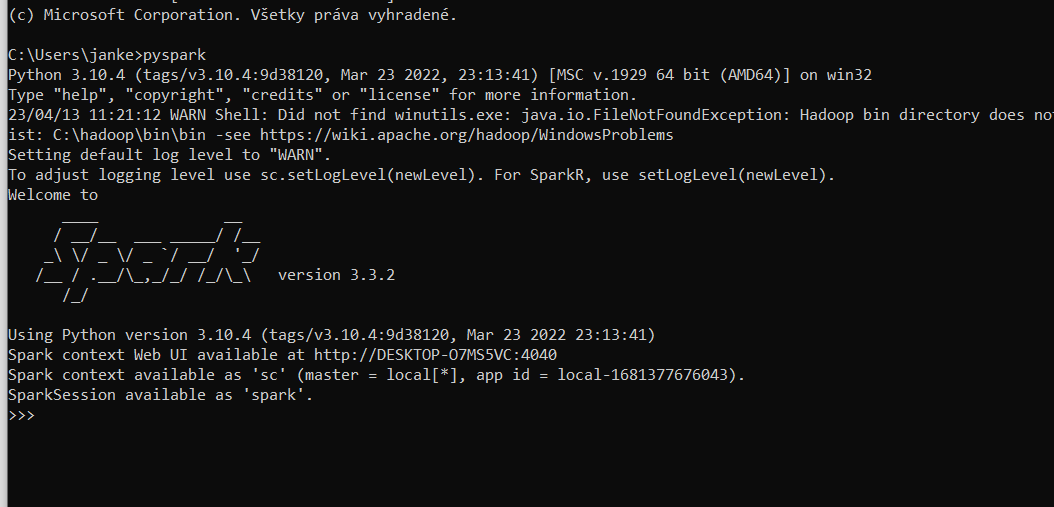
A do path v systémových premenných prostredia pridáme :



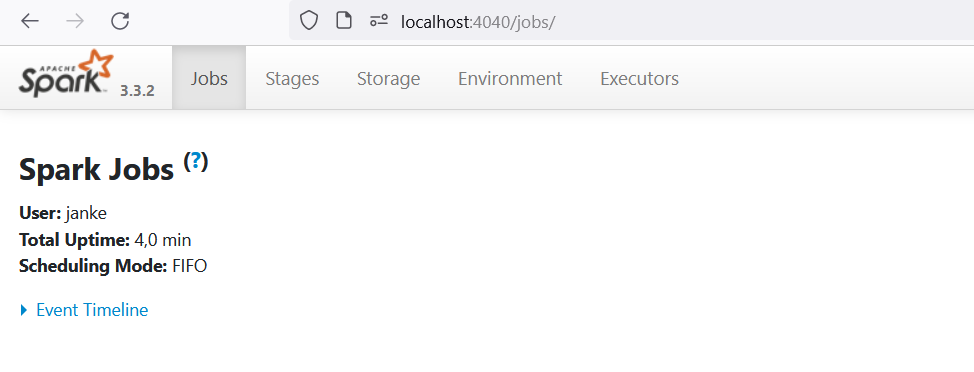
Do cmd dáme príkaz spark-shell a mal by sa nám spustiť spark



Príkazom pyspark pustíme Pyspark



Taktiež môžeme navštíviť adresu <http://localhost:4040/jobs/> kde môžeme pristúpiť k webovému UI sparku



# Analýza log súboru pomocou sparku

Apache Spark je vynikajúci a ideálny rámec na spracovanie, analýzu a modelovanie štruktúrovaných a neštruktúrovaných údajov - v širokom rozsahu! V tomto zošite sa zameriame na jednu z najpopulárnejších prípadových štúdií v odvetví - analýzu logov.

Logy serverov sú zvyčajne veľmi častým zdrojom údajov v podnikoch a často obsahujú zlatú baňu využiteľných poznatkov a informácií. Údaje protokolov pochádzajú z mnohých zdrojov v podniku, napríklad z webu, klientskych a výpočtových serverov, aplikácií, obsahu generovaného používateľmi, plochých súborov. Možno ich použiť na monitorovanie serverov, zlepšenie obchodných a zákazníckych informácií, budovanie odporúčacích systémov, odhaľovanie podvodov a mnoho ďalšieho.

Spark umožňuje lacno vypisovať a ukladať logy do súborov na disku, pričom stále poskytuje bohaté API na vykonávanie analýzy údajov v širokom rozsahu. V tejto praktickej ukážke sa dozviete, ako používať Apache Spark na reálnych produkčných logoch z NASA, a naučíte sa pracovať s dátami a základné, ale výkonné techniky prieskumnej analýzy dát.

# Popis zdrojového kódu a dataset

Popis zdrojového kódu sa nachádza v priloženom pdf súbore a dataset je zabalený v rar súbore.

# Zdroje

https://www.youtube.com/watch?v=OmcSTQVkrvo