



EKONOMICKÁ UNIVERZITA
Fakulta hospodárskej informatiky
Katedra aplikovanej informatiky



INŠTALÁCIA HADOOP 3.3.3 NA WINDOWS 10

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

Úvod.....	3
1.0 Postup inštalácie.....	4
Záver.....	9
Bibliografia.....	10

Úvod

V tomto zadaní si ukážeme a popíšeme postup akým sa inštaluje hadoop 3.3.3 na operačný systém Windows 10. Pred inštaláciou si je potrebné nainštalovať javu. My použijeme verziu 1.8.0_271, ktorá je zaručene touto verziou hadoopu podporovaná.

1.0 Postup inštalácie

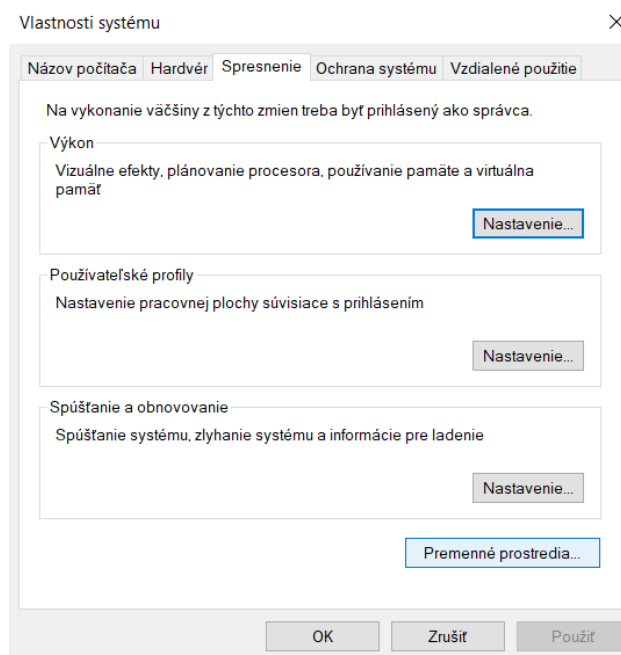
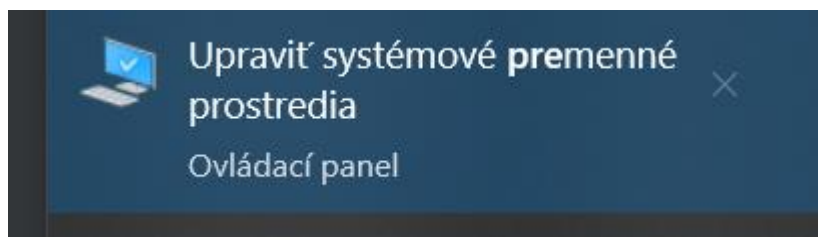
1. Pre inštaláciu hadoopu je potrebné skontrolovať či je Java nainštalovaná v systéme. Vykonať to môžeme napísaním príkazu `java -version` v príkazovom riadku. Pokiaľ dostaneme takýto výstup, tak Java nie je v našom systéme nainštalovaná.

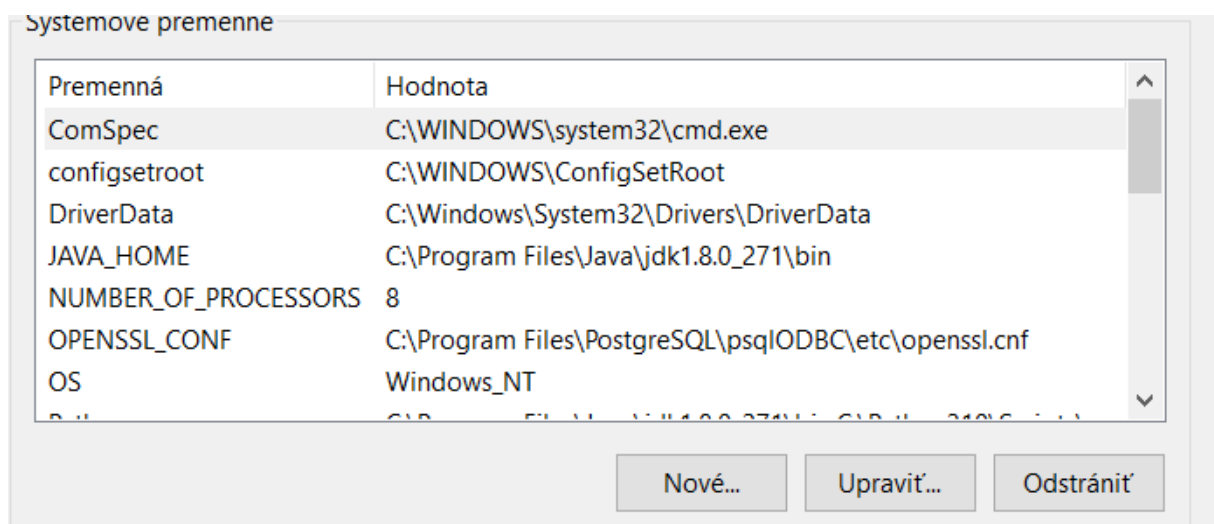
```
java -version
```

```
C:\Users\boskom>java -version
'java' is not recognized as an internal or external command,
operable program or batch file.

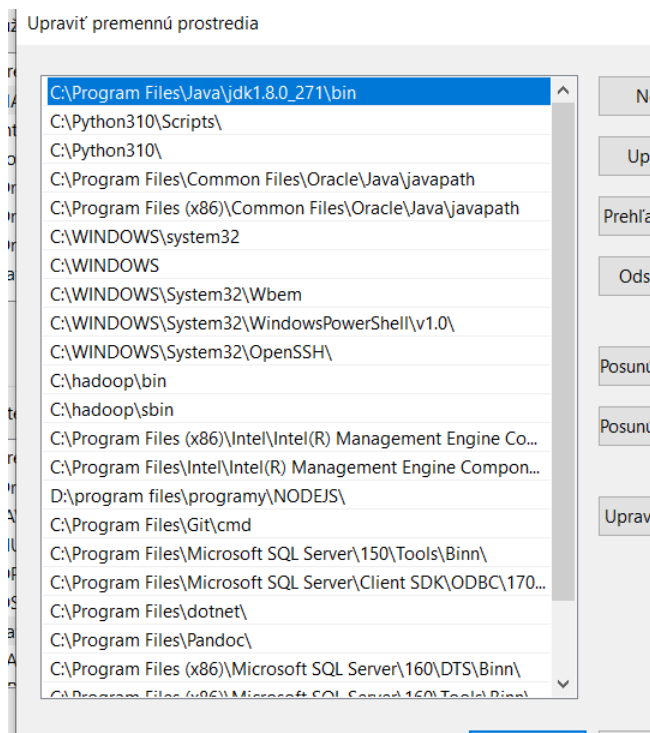
C:\Users\boskom>
```

2. Pokiaľ nemáme javu nainštalovanú tak ju treba stiahnuť napr. so stránky Oracle a nainštalovať. Odporúčame do koreňového adresára C: (spôsobuje najmenej problémov).
3. Po inštalácii je potrebné nastaviť systémové premenné prostredia.
4. Vytvoríme premennú JAVA_HOME a pridáme cestu k nainštalovanej jave.





5. Taktiež pridáme cestu k jave v premennej path rovnako ako na obrázku.



6. Po inštalácii javy sa môžeme pustiť na inštaláciu hadoopu.
7. Najprv si stiahneme hadoop so stránky Apache hadoop. My použijeme verziu 3.3.3
8. Rozbalený adresár nahráme na cieľové miesto. V princípe môže byť v inom adresári alebo na inom mieste, ale najmenej problémov sa vyskytne keď je v roote na bootovacom disku v našom prípade disk C. Názov adresára Hadoop 3.3.0 sme premenovali na Hadoop. Pozn. Je to dobré urobiť ešte pred tým než nastavíme cesty k súborom.
9. V adresári data vytvoríme 2 adresáre datanode a namenode

Tento počítač > Windows (C:) > hadoop >

	Názov	Dátum úpravy	Typ	Veľkosť
	bin	17. 2. 2023 11:25	Priečinok súborov	
	data	17. 2. 2023 11:53	Priečinok súborov	
	datanode	17. 2. 2023 11:24	Priečinok súborov	
	etc	17. 2. 2023 11:23	Priečinok súborov	
	include	17. 2. 2023 11:22	Priečinok súborov	
	lib	17. 2. 2023 11:23	Priečinok súborov	
	libexec	17. 2. 2023 11:22	Priečinok súborov	
	licenses-binary	17. 2. 2023 11:22	Priečinok súborov	
	logs	17. 2. 2023 11:55	Priečinok súborov	
	namenode	17. 2. 2023 11:24	Priečinok súborov	
	sbin	17. 2. 2023 11:22	Priečinok súborov	
	share	17. 2. 2023 11:22	Priečinok súborov	
	LICENSE	24. 3. 2020 18:23	Textový dokument	16 kB
	LICENSE-binary	4. 7. 2020 19:29	Súbor	23 kB
	NOTICE	24. 3. 2020 18:23	Textový dokument	2 kB
	NOTICE-binary	24. 3. 2020 18:23	Súbor	27 kB
	README	24. 3. 2020 18:23	Textový dokument	1 kB

10. Stiahneme si konfiguračný súbor pre Hadoop 3.3.0 - 3.3.0 Configuration Files Apache Hadoop.zip, ktorý obsahuje jeden adresár bin. Obsah tohto adresára nakopírujeme do pôvodného adresára bin, a pôvodné súbory prepíšeme. Stiahnuť to môžeme napr. odtiaľto

https://github.com/brainmentorspvtltd/BigData_RDE/blob/master/Hadoop%20Configuration.zip

alebo

<https://kai.ponuky.info/index.php?menu=6>

11. Ďalej si otvoríme adresár etc kde budeme konfigurovať 1 cmd a 4 xml súbory

a. core-site.xml b. hadoop-env.cmd c. hdfs-site.xml d. mapred-site.xml e. yarn-site.xml

12. core-site.xml

doplníme konfiguráciu na

```
<configuration>
```

```
<property>
```

```
<name>fs.defaultFS</name>
```

```
<value>hdfs://localhost:9000</value>
```

```
</property>
```

```
</configuration>
```

Nastavíme hadoop-env.cmd set JAVA_HOME=C:\Java\jdk1.8.0_341

My sme museli cestu upraviť takto

```
@rem The java implementation to use. Required.
set JAVA_HOME=C:\Progra~1\Java\jdk1.8.0_271\
```

13. Nasleduje hdfs-site.xml, kde nastavíme konfiguráciu na

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:///C:/hadoop/data/namenode</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value> file:///C:/hadoop/data/datanode</value>
</property>
</configuration>
```

14. Mapred-site.xml

```
<configuration>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>
```

15. Yarn-site.xml upravíme na

```
<configuration>
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
<property>
  <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

Nezabudnite všetky konfiguračné zmeny uložiť

16. ďalej pridáme cestu k hadoopu ako používateľskú premennú

Premenná	Hodnota
HADOOP_HOME	C:\hadoop\bin
IntelliJ IDEA	C:\Program Files\JetBrains\IntelliJ IDEA 2020.2.4\bin;
kotlin	C:\kotlin\bin
OneDrive	C:\Users\janke\OneDrive - Ekonomická univerzita v Bratislave
OneDriveCommercial	C:\Users\janke\OneDrive - Ekonomická univerzita v Bratislave
OneDriveConsumer	C:\Users\janke\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\janke\...

17. A tak isto aj systémovú premennú v path

```
C:\WINDOWS\System32\WindowsPowerShell\v1.
C:\WINDOWS\System32\OpenSSH\
C:\hadoop\bin
C:\hadoop\sbin
```

18. Ďalej naformátujeme namenode

```
C:\Users\janke>hdfs namenode -format
2023-02-28 12:32:52,233 INFO namenode.NameNode: STARTUP_MSG:
/*****
```

19. Potvrdíme Y

```
Re-format filesystem in Storage Directory root= C:\hadoop\data\namenode; location= null ? (Y or N) _
```

```
2023-02-28 12:34:04,453 INFO common.Storage: Will remove files: [C:\hadoop\data\namenode\current\edits_inprogress_00000000000000000001, C:\hadoop\data\namenode\current\fsimage_00000000000000000000
namenode\current\fsimage_00000000000000000000.mds, C:\hadoop\data\namenode\current\seen_txid, C:\hadoop\data\namenode\current\VERSION]
2023-02-28 12:34:04,513 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been successfully formatted.
2023-02-28 12:34:04,569 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namenode\current\fsimage.ckpt_00000000000000000000 using no compression
2023-02-28 12:34:04,716 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\current\fsimage.ckpt_00000000000000000000 of size 400 bytes saved in 0 seconds .
2023-02-28 12:34:04,733 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-02-28 12:34:04,741 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-02-28 12:34:04,742 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-07H5SVC/192.168.56.1
*****/
```

20. Prikazom C:\Users\EU>start-dfs.cmd spúšťame namenode a datanode a príkazom C:\Users\EU> start-yarn.cmd spúšťame yarn

Príkazom JPS môžeme otestovať či sa vykonávajú všetky služby

Či nám služby fungujú môžeme overiť aj ich otvorením v prehliadači na localhoste a portoch 8088 a 9870

(P., 2023)

The screenshot displays the Hadoop web interface in a browser. The top section, titled "All Applications", shows cluster metrics and a table of applications. The "Cluster Metrics" section includes a table with columns: Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, Memory Used, and Memory Total. The "Cluster Nodes Metrics" section includes a table with columns: Active Nodes, Decommissioning Nodes, Decommissioned Nodes, and Lost Nodes. The "Scheduler Metrics" section includes a table with columns: Scheduler Type, Scheduling Resource Type, and Minimum Allocation. The "Applications" table has columns: ID, User, Name, Application Type, Application Tags, Queue, Application Priority, StartTime, LaunchTime, FinishTime, State, FinalStatus, Running Containers, and Allocated CPU VCore. The table is currently empty, showing "No data available in table".

The bottom section, titled "Overview", shows the status of the Hadoop cluster. It includes a table with columns: Started, Version, Compiled, Cluster ID, and Block Pool ID. The "Started" column shows "Tue Feb 28 12:36:58 +0100 2023". The "Version" column shows "3.3.0, raa96f1871bfd858f9bac59cd2a81ec470da649af". The "Compiled" column shows "Mon Jul 06 20:44:00 +0200 2020 by brahma from branch-3.3.0". The "Cluster ID" column shows "CID-17ae0cc1-52ba-4297-b751-6f593cd2d659". The "Block Pool ID" column shows "BP-185534152-192.168.56.1-1677584044418".

Záver

V tomto zadaní sme si ukázali a popísali ako nainštalovať Hadoop na Windows. Až na jednu zmenu, ktorú sme museli spraviť pri zadávaní cesty k jave sme postupovali rovnako ako v návode, ktorý nám bol poskytnutí. S potešením môžeme prehlásiť že inštalácia podľa návodu bola úspešná v všetko funguje ako má.

Bibliografia

P., S. (28. 2. 2023). *Ako nainštalovať Hadoop*. Dostupné na Internete:
<https://kai.ponuky.info/index.php?menu=6>