

BI and Big Data – praktické ukážky

Miroslav Koseček – Aadaptive.biz
<http://aadaptive.biz/freelancer/>

Seminár Big Data International Workshop on Knowledge Management

*Vysoká škola manažmentu v Trenčíne
City University of Seattle, Bratislava, 20. 10. 2016*



BIG data



- Volume – objem (100+ TB)
- Variability
- Velocity – rýchlosť (s akou pribúdajú alebo ich treba čítať, či spracovať)
- Využitie – data mining
 - business analýzy
 - vedecké dáta (analýza genómu, LHC, analýza snímok...)

BIG data

- => masívne paralelné systémy
 - úložiská
 - HDFS
 - NoSQL
 - large column tables (Big Table, Cassandra)
 - key-value (Redis, Amazon Dynamo DB)
 - document store (MongoDB, Couch Base)
 - diagram (Neo4j)
 - spracovanie
 - map-reduce
 - Hadoop
 - Spark
 - Google App Engine



Platformy

- Hostované
 - prenájom virtuálnych strojov (Amazon AWS <https://aws.amazon.com/ec2/pricing/on-demand/> , Google, mnoho ďalších)
 - Prenájom kontajnerov (Google...)
 - Cloudové služby (Google – Big Query, App Engine <https://console.cloud.google.com/home/dashboard> ; Microsoft Azzure, Firebase...)
 - výhody – jednoduchá správa, on – demand
 - nevýhody – nemáte to celkom pod kontrolou
- Vlastné
 - z dlhodobého hľadiska lacnejšie
 - treba to manažovať

Ukážky - MongoDB

```
> show dbs
local  0.000GB
> use test
switched to db test
> db
test
>
> db.createCollection ('fakturyPrijata');
{ "ok" : 1 }
> show collections
fakturyPrijata
> db.createCollection('faktury')
{ "ok" : 1 }
> db.faktury.insert({Cislo_faktury : "1234"})
WriteResult({ "nInserted" : 1 })
> db.faktury.insert({Cislo_faktury : "1235", Suma : 23.80})
WriteResult({ "nInserted" : 1 })
> db.faktury.find()
{ "_id" : ObjectId("58026a33788d31e65179ab27"), "Cislo_faktury" : "1234", "Zakaznik" : "IBM s.r.o." }
{ "_id" : ObjectId("58026a78788d31e65179ab28"), "Cislo_faktury" : "1235", "Suma" : 23.8 }
>
_
```

Ukážky

- NoSQL – dokumentová db MongoDB

FAKTURA: 234200095

Dodávateľ : INMEDIA, spol. s r.o. Námestie SNP 11 960 01 Zvolen	IČO : 36019208 IČ DPH : SK2020066829 DIČ : 2020066825 Obchodný register : Obchodný register, OS B.Bystriča, Oddiel : Sro, vložka číslo : 4345/S	Doklad č Konštant Dátum vj Dátum sg Forma úl Dátum di					
Peňažný ústav : UniCredit Bank Slovakia a.s. Číslo účtu : 6619429000 /1111 Telefón : 045/5363008,5363024 Fax : 045/5326976	Odberateľ : Obecný úrad N 2568						
Miesto určenia :	Ul.Hlavná 1/1 013 02 Nededza						
Druh dodávky : 0,25 Spôsob dopravy : Č. objednávky : ozm Vnútrošný doklad : 340 200101 Hurišová Monika C. bez DPH :	DIČ: 2020677582 IČ DPH:	Teľ.čísľ IČ					
Balil : Hudec Miroslav Obch. zástupca : 411 Unilever Pecko	Prijemca :						
Pô	Kód	NÁZOV POLOŽKY	Množ.	Mj	Ďalenie	Cena/MJ bez DPH	[
1	071109	Cest.Tarhoňa 5kx2 Knorr	1,00	ks		20,820	
2	25401	Čaj Saga Čierne ribezlie 20x2g Knorr	6,00	ks	0.5 bal	0,750	
3	25201	Čaj Saga Jabiko 20x2g Knorr	6,00	ks	0.5 bal	0,750	
4	256011	Čaj Saga Lesné plody 20x2g Knorr	6,00	ks	0.5 bal	0,750	



```
1 - {
2   "Číslo_faktúry": "201110074",
3   "Dodávateľ": {
4     "Názov": "Cestné stavby Žilina s.r.o.",
5     "Adresa": {
6       "Ulica": "Štrková 17",
7       "Mesto": "Žilina",
8       "PSČ": "011 54"
9     }
10  },
11  "Dátum_vystavenia": "2012-01-01",
12  "Položky_faktúry": [
13    {
14      "Položka": "Stavebné práce uvedené podľa súpisu prác
15        uvedených v prílohe",
16      "Množstvo": 1,
17      "Cena": 55839.45
18    }
19  ],
20  "Príloha": {
21    "Obsah": "Tu bude štruktúrovaný text prílohy"
22  }
}
```

MongoDB a Python

The screenshot shows an IDE window with a Python script named 'faktury.py' and an IPython console. The script defines a MongoDB client, connects to a database, and inserts a document. The console shows the execution of the script, displaying the inserted document's details.

```
4
5 @author: Kosecek
6 """
7
8 from pymongo import MongoClient
9 import pprint
10
11 cl = MongoClient()
12
13 db = cl["test"]
14
15 faktury = db["faktury"]
16
17 for fa in faktury.find():
18     pprint.pprint(fa)
19
20 fakt1 = {
21     "Číslo faktúry": "201110074",
22     "Dodávateľ": {
23         "Názov": "Cestné stavby Žilina s.r.o.",
24         "Adresa": {
25             "Ulica": "Štrková 17",
26             "Mesto": "Žilina",
27             "PSČ": "011 54"
28         }
29     },
30     "Dátum vystavenia": "2012-01-01",
31     "Položky faktúry": [
32         {
33             "Položka": "Stavebné práce uvedené podľa súpisu prác uvedených v prílohe",
34             "Množstvo": 1,
35             "Cena": 55839.45
36         }
37     ],
38     "Príloha": {
39         "Obsah": "Tu bude štruktúrovaný text prílohy"
40     }
41 }
42
43 faktury.insert_one(fakt1)
44
45 for fa in faktury.find():
46     pprint.pprint(fa)
```

IPython console output:

```
In [1]:
In [1]:
In [1]:
In [1]: runfile('C:/Users/Miro/Documents/Verziovane/Projctcs/Seminar big data/faktury.py', wdir='C:/Users/Miro/Documents/Verziovane/Projctcs/Seminar big data')
{'_id': ObjectId('580879baf629d4cf2ca39c0e'), 'cislo_fatkury': '123'}
{'_id': ObjectId('58087a4ff629d4cf2ca39c0f'),
 'cislo_fatkury': '124',
 'zakaznik': 'IBM s.r.o'}
{'_id': ObjectId('580879baf629d4cf2ca39c0e'), 'cislo_fatkury': '123'}
{'_id': ObjectId('58087a4ff629d4cf2ca39c0f'),
 'cislo_fatkury': '124',
 'zakaznik': 'IBM s.r.o'}
{'Dodávateľ': {'Adresa': {'Mesto': 'Žilina',
                          'PSČ': '011 54',
                          'Ulica': 'Štrková 17'},
               'Názov': 'Cestné stavby Žilina s.r.o.'},
 'Dátum vystavenia': '2012-01-01',
 'Položky faktúry': [{'Cena': 55839.45,
                     'Množstvo': 1,
                     'Položka': 'Stavebné práce uvedené podľa súpisu
prác '
                     'uvedených v prílohe'}],
 'Príloha': {'Obsah': 'Tu bude štruktúrovaný text prílohy'},
 '_id': ObjectId('5814e2b807d9352724e92e77'),
 'Číslo faktúry': '201110074'}
In [2]:
```

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 35 Column: 25 Memory: 66%