EPL660: Information Retrieval and Search Engines – Lab 8

University of Cyprus
Department of
Computer Science

Παύλος Αντωνίου

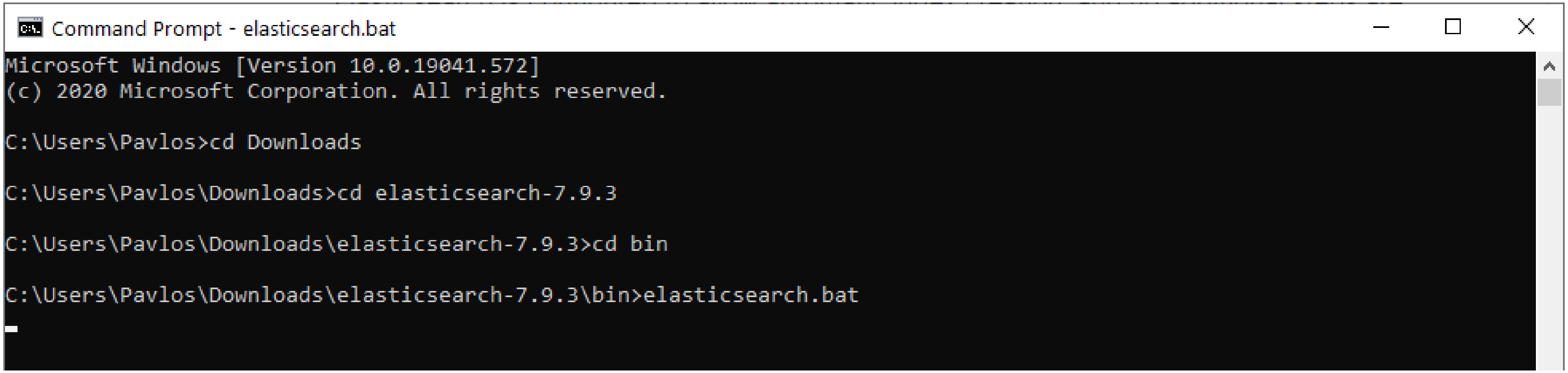
Γραφείο: B109, ΘΕΕ01

Hands on ElasticSearch

- Elasticsearch v7.8.0 installed on VM
- Kibana installed on VM
- Python client libraries for Elasticsearch installed
  - elasticsearch
    - more general but hides less the complexities of the API calls
  - elasticsearch-dsl
    - focused on the search capabilities and is more friendly for sending queries to ElasticSearch
- Activate Elasticsearch
  - sudo service elasticsearch start
- Activate Kibana
  - sudo service kibana start

Hands on ElasticSearch

- Install Elasticsearch on Windows
  - Download zip via https://www.elastic.co/guide/en/elasticsearch/reference/current/zip-windows.html
- Unzip and run \bin\elasticsearch.bat to start ES

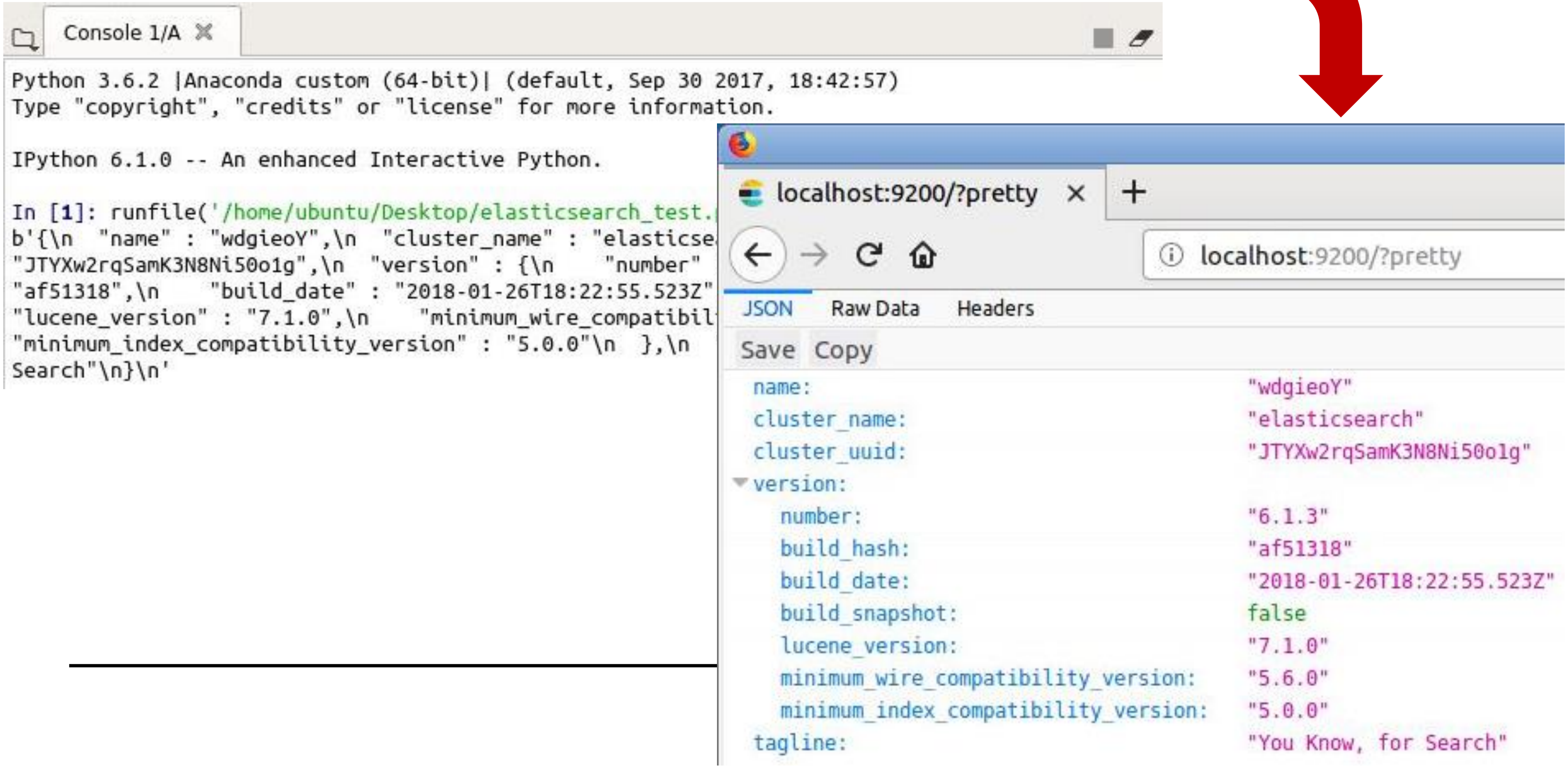


```
Microsoft Windows [Version 10.0.19041.572]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Users\Pavlos>cd Downloads

C:\Users\Pavlos\Downloads>cd elasticsearch-7.9.3

C:\Users\Pavlos\Downloads\elasticsearch-7.9.3>cd bin

C:\Users\Pavlos\Downloads\elasticsearch-7.9.3\bin>elasticsearch.bat
```

- Python libraries (if anaconda is in place):
  - `conda install -c conda-forge elasticsearch`
  - `conda install -c conda-forge elasticsearch-dsl`

Hands on ElasticSearch

- Check if Elasticsearch is working:
  - Run elasticsearch_test.py file in Spyder or Python IDLE
  - http://localhost:9200

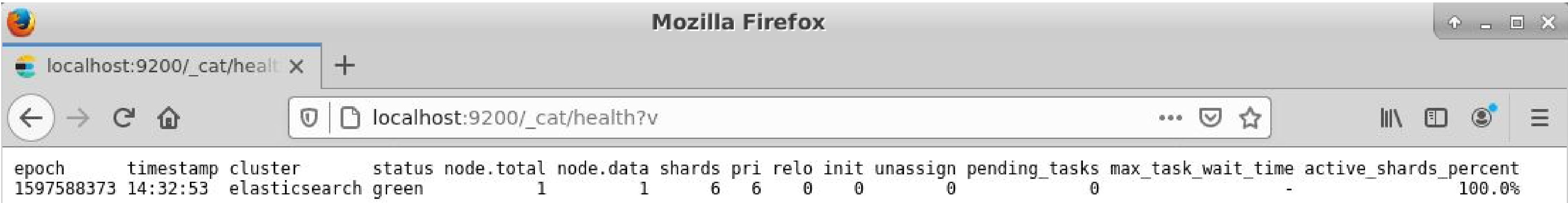


```
Python 3.6.2 |Anaconda custom (64-bit)| (default, Sep 30 2017, 18:42:57)
Type "copyright", "credits" or "license" for more information.

IPython 6.1.0 -- An enhanced Interactive Python.

In [1]: runfile('/home/ubuntu/Desktop/elasticsearch_test.
b'{\n  "name" : "wdgieoY",\n  "cluster_name" : "elasticse
"JTYXw2rqSamK3N8Ni50o1g",\n  "version" : {\n    "number"
"af51318",\n    "build_date" : "2018-01-26T18:22:55.523Z"
"lucene_version" : "7.1.0",\n    "minimum_wire_compatibil
"minimum_index_compatibility_version" : "5.0.0"\n  },\n
Search"\n}\n'
```

```
name:                                   "wdgieoY"
cluster_name:                           "elasticsearch"
cluster_uuid:                           "JTYXw2rqSamK3N8Ni50o1g"
version:
  number:                               "6.1.3"
  build_hash:                           "af51318"
  build_date:                           "2018-01-26T18:22:55.523Z"
  build_snapshot:                       false
  lucene_version:                       "7.1.0"
  minimum_wire_compatibility_version:   "5.6.0"
  minimum_index_compatibility_version:  "5.0.0"
tagline:                                "You Know, for Search"
```

Hands on ElasticSearch

- Check cluster health:
  - http://127.0.0.1:9200/_cat/health?v



```
epoch      timestamp cluster       status node.total node.data shards pri relo init unassign pending_tasks max_task_wait_time active_shards_percent
1597588373 14:32:53  elasticsearch green           1         1      6   6    0    0        0             0                  -                100.0%
```

  - Elasticsearch provides a handy "traffic lights“ classification of cluster health:

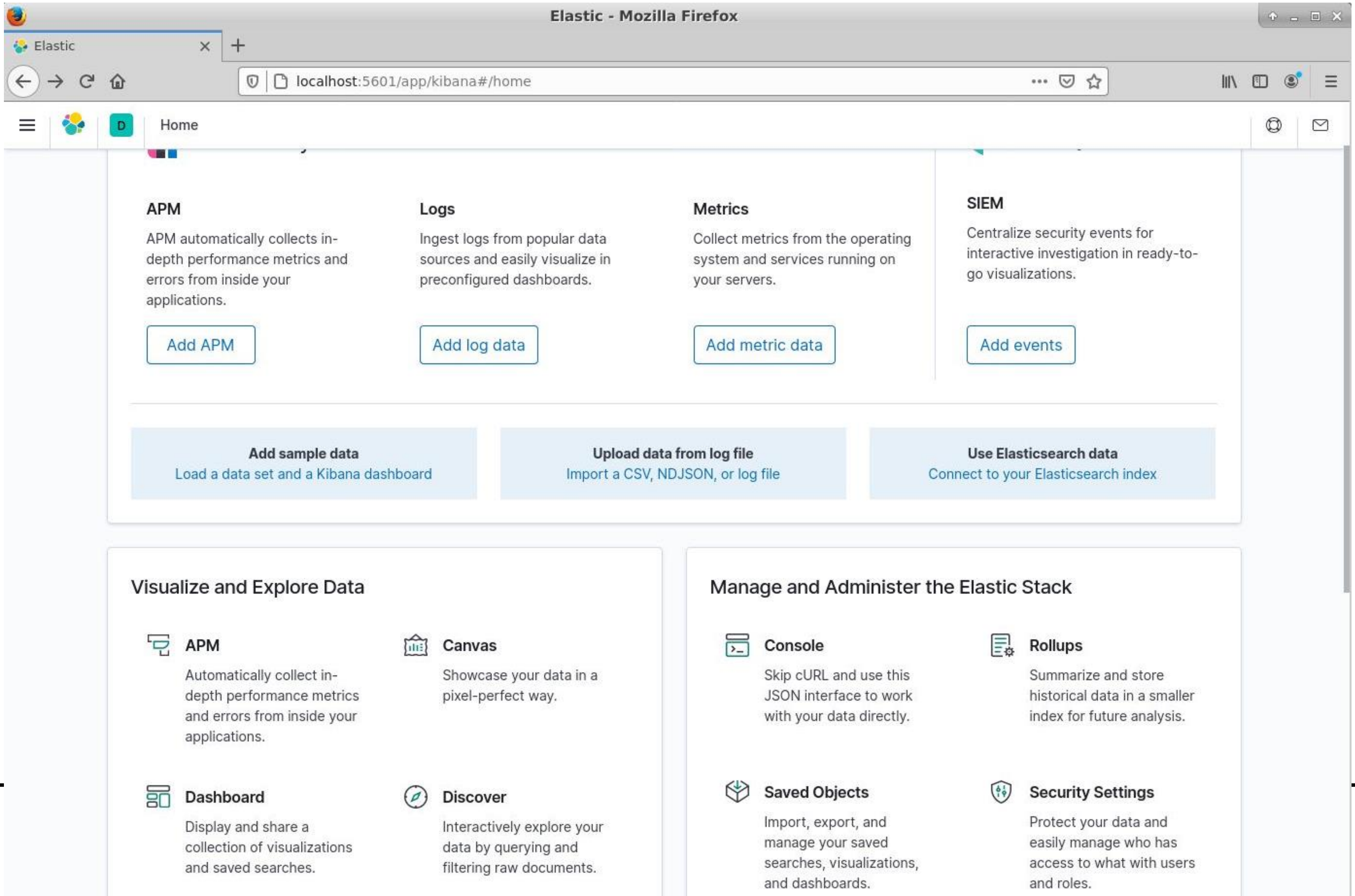
● RED: Some or all of (primary) shards are not ready

● YELLOW: Elasticsearch has allocated all of the primary shards, but some/all of the replicas have not been allocated. Your cluster is fully operational.

● GREEN: Elasticsearch is able to allocate all shards and replicas to machines within the cluster.

Hands on Kibana

- Check if Kibana is working:
  - http://localhost:5601



Elastic - Mozilla Firefox

Elastic

localhost:5601/app/kibana#/home

Home

APM

APM automatically collects in-depth performance metrics and errors from inside your applications.

Add APM

Logs

Ingest logs from popular data sources and easily visualize in preconfigured dashboards.

Add log data

Metrics

Collect metrics from the operating system and services running on your servers.

Add metric data

SIEM

Centralize security events for interactive investigation in ready-to-go visualizations.

Add events

Add sample data
Load a data set and a Kibana dashboard

Upload data from log file
Import a CSV, NDJSON, or log file

Use Elasticsearch data
Connect to your Elasticsearch index

Visualize and Explore Data

APM
Automatically collect in-depth performance metrics and errors from inside your applications.

Canvas
Showcase your data in a pixel-perfect way.

Dashboard
Display and share a collection of visualizations and saved searches.

Discover
Interactively explore your data by querying and filtering raw documents.

Manage and Administer the Elastic Stack

Console
Skip cURL and use this JSON interface to work with your data directly.

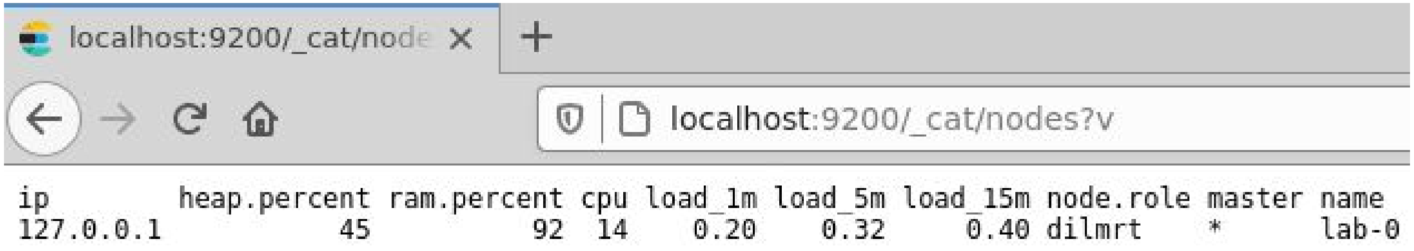
Rollups
Summarize and store historical data in a smaller index for future analysis.

Saved Objects
Import, export, and manage your saved searches, visualizations, and dashboards.

Security Settings
Protect your data and easily manage who has access to what with users and roles.

RESTful API Calls

- Access ElasticSearch via Restful API on browser
  - View nodes: http://127.0.0.1:9200/_cat/nodes?v



```
ip        heap.percent ram.percent cpu load_1m load_5m load_15m node.role master name
127.0.0.1           45          92  14    0.20    0.32     0.40 dilmrt    *      lab-0
```

  - View all indices: http://127.0.0.1:9200/_cat/indices?v
  - View shards: http://127.0.0.1:9200/_cat/shards?v
  - View segments: http://127.0.0.1:9200/_cat/segments?v

Today’s lab

- Datasets
  - `20_newsgroups`: Text from 20 usenet groups on various topics, a classic corpus in IR evaluation, from here.
  - `novels`: A number of random novels and other texts in English from the Gutenberg project, with a tendency towards late 19th and early 20th centuries.

Today’s lab

- Download lab8.zip and unzip it
- Create folder e.g. `/home/ubuntu/datasets`
- Move 20_newsgroups.tar.gz and novels.zip into the datasets folder and unzip them
  - `tar xzvf 20_newsgroups.tar.gz`
  - `unzip novels.zip`
- In this lab:
  - You will learn how to use the ElasticSearch database
  - How to index a set of documents
  - How to ask simple queries about indexed documents
- Go through Lab8.pdf to run the examples
- Submit results to Moodle by Nov. 19 @ 15:00