Big Data – Practical Exam

Winter Semester 2023/2024,

Cooperative State University Baden-Wuerttemberg

www.marcel-mittelstaedt.com

Practical Exam

Topics, Deliverables, Presentation…

Practical Exam - **Grading**

Grade/Percentage:

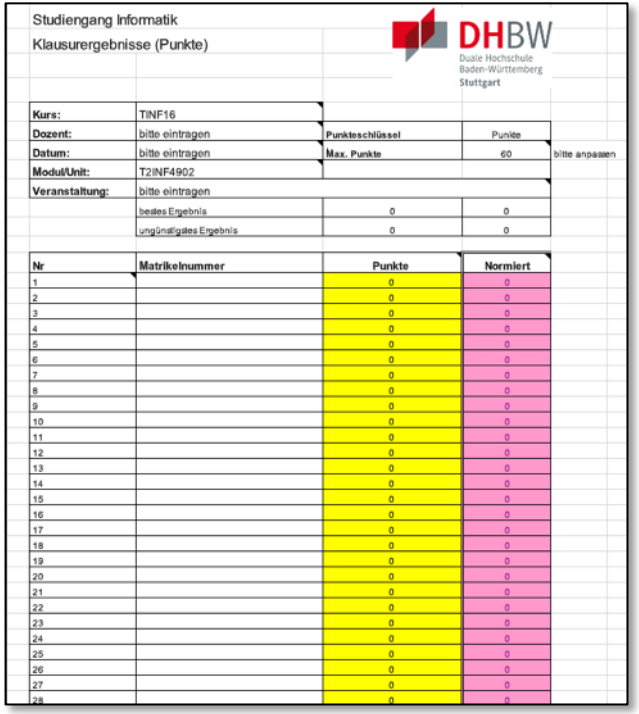
- grade will be mixed with grade of another lecture
- therefore, the rating won‘t be a grade (1-6) but a percentage:

Percentage	Grade
100%	1.0
…	…
50%	4.0
…	…

Timeline:

05.07.2024 23:59 All deliverables (next slide) will be delivered to following email address (DropBox, Google-Drive…):
contact@marcel-mittelstaedt.com

tbd - Presentation of Practical Exam



Studiengang Informatik
Klausurergebnisse (Punkte)

DHBW Duale Hochschule Baden-Württemberg Stuttgart

Kurs:	TINF16			
Dozent:	bitte eintragen	Punkteschlüssel	Punkte	
Datum:	bitte eintragen	Max. Punkte	60	bitte anpassen
Modul/Unit:	T2INF4902			
Veranstaltung:	bitte eintragen			
	bestes Ergebnis	0	0	
	ungültigstes Ergebnis	0	0	

Nr	Matrikelnummer	Punkte	Normiert
1		0	0
2		0	0
3		0	0
4		0	0
5		0	0
6		0	0
7		0	0
8		0	0
9		0	0
10		0	0
11		0	0
12		0	0
13		0	0
14		0	0
15		0	0
16		0	0
17		0	0
18		0	0
19		0	0
20		0	0
21		0	0
22		0	0
23		0	0
24		0	0
25		0	0
26		0	0
27		0	0
28		0	0

Practical Exam - **Deliverables**

Deliverables:

- A simple Documentation:
  - Explanation of whole ETL Workflow
  - List of Jobs/Transformations in Case of PDI or DAGs and Steps in Case of Airflow
  - Short description of the purpose of each job/transformation or task and applied business rules
  - all PDI Jobs, Transformations and related files (ktr, kjb, kettle.properties, shared.xml… files)
  - All Airflow DAGs and tasks
- All Scripts (e.g. Download) or other external applications called within PDI
- All Airflow DAGs, Python Files etc.
- All DDLs (`CREATE Table…`):
  - One file for each table
  - Table name = File Name, e.g.:

- Depending on Exam Type:
  - Code of Frontend Application and related Database (DDLs) or
  - Calculated KPIs

Practical Exam - **Presentation**

Procedure:

1. Start ETL Workflow

2. During execution:
   - Quickly explain data source
     - API
     - Data Structure
     - Approach for gathering data
   - Quickly Explain whole ETL Workflow
     - Explain Idea and purpose of each Job/Transformation
     - External ressources/scripts (e.g. download)
     - Explain Data Model (Raw Layer, Final Layer, simple Frontend)

3. After execution:
   - Depending on Exam:
     - Demo of simple Frontend application or
     - Explanation of calculated KPIs

Use GeoLite2 To Create A Searchable IP and GeoLocation Database

Practical Exam

Goal

Maxmind.com provides regulary exports of worldwide IP and Geolocation data:

- https://dev.maxmind.com/geoip/geolite2-free-geolocation-data

```
curl -s http://ifconfig.me
88.130.59.75
```

1

```
network,geoname_id,registered_country_geoname_id,represented_country_geoname_id,is_anonymous_proxy,is_satellite_provider,postal_code,latitude,longitude,accuracy_radius
88.130.59.0/24,2939623,2921044,,0,0,85221,48.2600,11.4340,50
[…]
```

GeoLite2-City-Blocks-IPv4.csv

2

```
geoname_id,locale_code,continent_code,continent_name,country_iso_code,country_name,subdivision_1_iso_code,subdivision_1_name,subdivision_2_iso_code,subdivision_2_name,city_name,metro_code,time_zone,is_in_european_union
3207535,de,EU,Europa,DE,Deutschland,BY,Bayern,,,Höhenkirchen-Siegertsbrunn,,Europe/Berlin,1
2939623,de,EU,Europa,DE,Deutschland,BY,Bayern,,,Dachau,,Europe/Berlin,1
3207410,de,EU,Europa,DE,Deutschland,BY,Bayern,,,Rödental,,Europe/Berlin,1
3207412,de,EU,Europa,DE,Deutschland,BY,Bayern,,,Röslau,,Europe/Berlin,1
3208324,de,EU,Europa,DE,Deutschland,BY,Bayern,,,Asbach-Bäumenheim,,Europe/Berlin,1
[…]
```

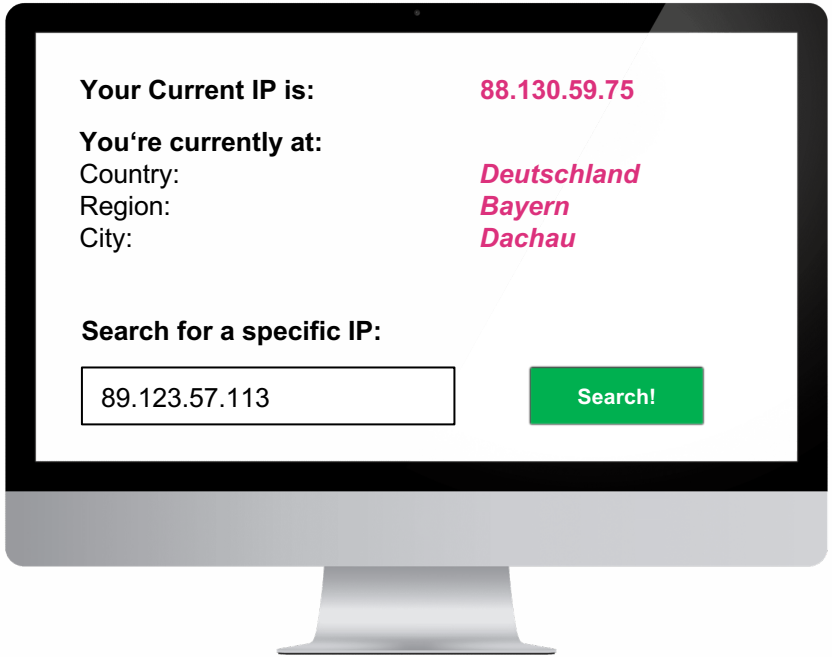
GeoLite2-City-Locations-[XX].csv

Goal

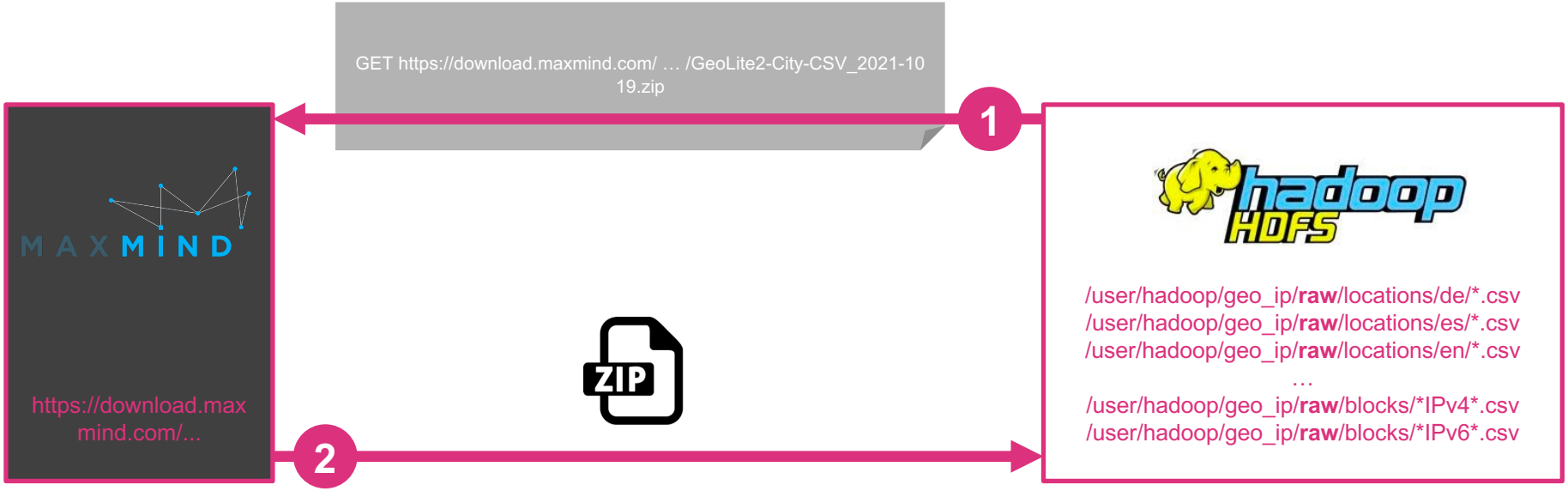
We want to make use of this data to build a real time IP-Geolocation resolution as well as a searchable database for Ips and related Geolocations.

Workflow:

- **Gather data** from maxmind.com
- **Save raw data** (*CSV files*) to HDFS (partitioned by country code, e.g. *de*, *es*, *en*…)
- **Optimize**, **reduce** and **clean raw data** and save it to **final** directory on HDFS
- **Export** Geolite2 data **to end-user database** (e.g. MySQL, MongoDB…)
- Provide a simple **HTML Frontend** which is able to:
  - **determine** a user's IP address, **lookup** and **show Geolocation**
  - process user input (IP…) and check against end-user database
  - Display result Geolocation
- The whole data workflow **must be implemented** within an ETL **workflow tool** (e.g. Pentaho Data Integration or Airflow) and **run automatically**



Dataflow: 1. Get Geolite2 Data



Dataflow: 2. Raw To Final Transfer

/user/hadoop/geo_ip/**raw**/locations/de/*.csv
/user/hadoop/geo_ip/**raw**/locations/es/*.csv
/user/hadoop/geo_ip/**raw**/locations/en/*.csv
…
/user/hadoop/geo_ip/**raw**/blocks/*IPv4*.csv
/user/hadoop/geo_ip/**raw**/blocks/*IPv6*.csv



1

- move data from *raw* to *final* directory
- optimize and reduce data structure for later query purposes if necessary
- remove duplicates if necessary
- …

/user/hadoop/geo_ip/**final**/locations/de
/user/hadoop/geo_ip/**final**/locations/es/
/user/hadoop/geo_ip/**final**/locations/en/
…
/user/hadoop/geo_ip/**final**/blocks/*IPv4*
/user/hadoop/geo_ip/**final**/blocks/*IPv6*

Dataflow: 3. Enhance Data And Save Results

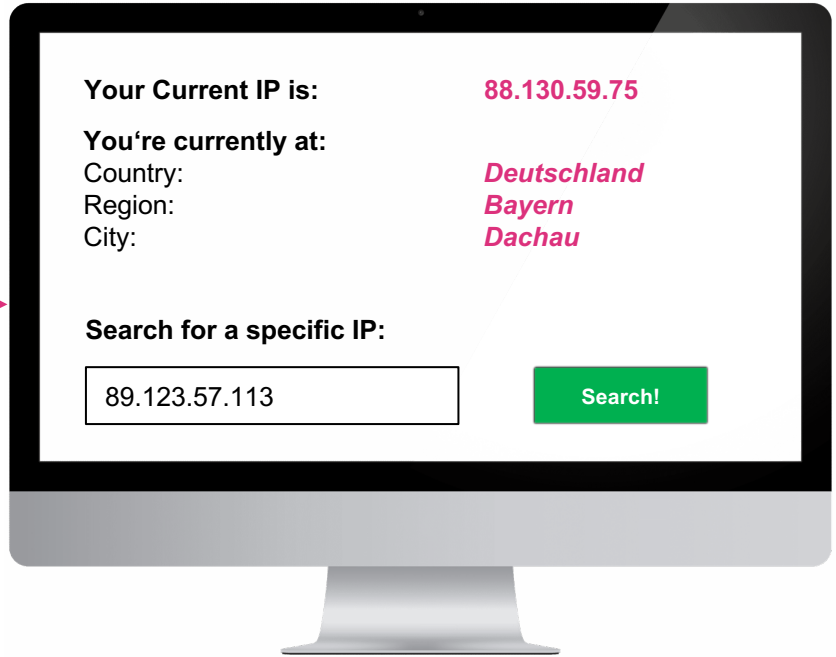
/user/hadoop/geo_ip/**final**/locations/de
/user/hadoop/geo_ip/**final**/locations/es/
/user/hadoop/geo_ip/**final**/locations/en/
…
/user/hadoop/geo_ip/**final**/blocks/*IPv4*
/user/hadoop/geo_ip/**final**/blocks/*IPv6*

1

- enhance data (e.g. for later querying)
- use *Hive*, *Python*, *Spark* or *PySpark*
- save everything to a end-user database (e.g. *MySQL*, *MongoDB*)

Dataflow: 4. Provide Simple Web Interface

1



- Provide a simple **HTML Frontend** which is able to:
  - **determine** a user's IP address, **lookup** and **show** **Geolocation**
  - process user input (IP…) and check against end-user database
  - Display result Geolocation