Projektarbeit

Deadline

Upload nach ILIAS bis : So 29.05.2022 23:59:59.999 (MESZ)

Sprache

Die komplette Arbeit kann entweder in Englisch oder Deutsch erfasst werden.

Format

Alles in einem Zip-File.

- Akzeptierte Formate für Text:
    - Jupyter Notebook Files (`ipynb`)
    - Office Documents (`docx`, `xlsx`, `pptx`)
    - Portable Document Format (`pdf`)
    - Markdown files (`md`)
- Akzeptierte Formate für Code:
    - Jupyter Notebook Files (`ipynb`)
    - Code Files (`sql`, `py`, `sh`)

Legen Sie ausserdem einen Auszug der Daten bei.

Bei mehreren Dokumenten sollte die Reihenfolge, in der die Files gelesen werden sollten, ersichtlich sein.

Mögliche Ordnerstruktur

```
.
├── data_sample
│   ├── movie_lens_100_lines.csv
├── src
│   ├── 1_Pre_Processing
│   │   └── 1_preprocessor.ipynb
│   ├── 2_DDL
│   │   ├── 1_staging_ddl.ipynb
│   │   ├── 2_compressed_parquet_ddl.ipynb
│   ├── 3_Prototyping
│   │   ├── 1_create_data_sample.py
│   │   ├── 2_basic_checks.ipynb
│   ├── 4_Analysis
│   │   ├── 1_rated_movies_during_corona.py
│   │   ├── 2_...
│   ├── 5_Results
│   │   ├── 1_rated_movies_during_corona.py
│   │   ├── 2_...
```

```
├── doc
│   ├── 1_report.docx
```

Tools

I schlage vor

- git (und github / gitlab)
- JupyterLab (macht es übrigens auch einfach Code und Dokumentation unter einen Hut zu bringen)
- HDFS
- Spark

zu benutzen.

Punkte

Thema	Max Punkte
Abstract	2
Cluster	4
Dataset	4
Prototyping	4
Dataflow	2
Analysis	12
Learnings	2

Vorgehen

- Bilden Sie 3er oder 4er Teams. Eine Person meldet mir bis am 28.04.2022 alle Teammitglieder.
- Suchen Sie nach einem geeigneten Dataset. Fragen Sie mich bei Bedarf.
    - Dataset
- Denken Sie sich interessante Fragen aus, welche der Datensatz beantworten könnte. Auch hier: melden Sie sich bei Fragen oder wenn Sie Unterstützung brauchen.
- Erstellen Sie einen Cluster-Verbund.
- Bringen Sie die Daten ins HDFS.
- Erstellen Sie einen Prototype mit weniger Daten.
    - Wo braucht es ein pre-processing?
    - Reichen die Tools und Techniken aus, um die BigData Fragen zu beantworten?
- Erstellen Sie das Projekt mit allen Daten.

Inhalt der Dokumentation

Quellenangaben bei Code, Text und Bildern sind zwingend anzugeben.

Abstract

Schreiben Sie ein Abstract über die gesamte Arbeit.

Cluster

- Zeigen Sie die Cluster-Topologie als Bild. Wie heissen die Server, welche Services laufen auf welcher Maschine.
- Zeigen Sie, wie Sie die Ressourcen (CPU / Memory) aufgeteilt haben.
- Erklären Sie, warum Sie sich für diese Topologie entschieden haben und warum Sie die eingesetzten Frameworks und Tools benutzen.
- Wo gab es Probleme, was haben Sie neu dazugelernt?
- Wie haben Sie getestet, dass die Services funktionieren?

Dataset

- Wieso haben Sie sich für diesen Datensatz entschieden?
- Wie haben Sie den Datensatz heruntergeladen? (api, save as, csv, xml, json, ...)
- Wie haben Sie die Daten ins HDFS geladen? Musste die Blocksize von HDF angepasst werden?
- Ist Ihr Projekt ein Big-Data-Problem?
- Erklären Sie Ihre Daten. Was bedeuten die Felder, welches Schema sollten die Daten haben, gibt es primary keys und foreign keys?

Fragestellungen an den Datensatz (mindestens eine Frage)

- Welche Fragen möchten Sie durch die Daten beantworten?

Prototyping

- Erklären Sie den ersten Kontakt zu den Daten. Hat es Überraschungen gegeben?
- Welche pre-processing Schritte haben Sie gemacht?
- Was haben Sie durch das prototyping gelernt?

Dataflow

- Erklären Sie, wie die Daten End-to-End durch das System laufen.
    - Wie speichern Sie die Rohdaten?
    - Haben Sie die Daten partitioniert?
    - In welchem Format haben Sie die prozessierten Daten gespeichert. Wieso?

Analysis

- Erweitern Sie nun den Prototype mit allen Daten.
- Beantworten Sie Ihre gestellten Fragen mit Hilfe von queries, charts.
- Zeigen Sie Ihre Resultate und Ergebnisse
- Machen Sie eine Plausibilitätskontrolle.
- Gibt es durch die Beantwortung der Fragen neue Fragestellungen?

Learnings

- Was waren die grössten Challenges im Projekt?
- Was würden Sie beim nächsten Mal anders machen?

- Was sind Ihre grössten Learnings?