DSCB420 – Datenanalyse und Business Intelligence 2

Aufgabenstellung Projektarbeit

Hochschule Karlsruhe
University of
Applied Sciences

Sommersemester 2024

Inhalt

1 Idee der Projektarbeit (wichtig, bitte lesen!) 2
2 Gegebene Datensätze 3
3 Aufgabenstellung 4
4 Organisation und einzureichende Dokumente 5
4.1 Ablauf 5
4.2 Einzureichende Dokumente und Prüfungsleistung 5
4.2.1 Projektbericht 5
4.2.2 Präsentation 6
4.2.3 Code und weitere Artefakte 6
4.2.4 Sonstige Hinweise 6

1 Idee der Projektarbeit (wichtig, bitte lesen!)

In diesem Modul sollen Ihre Kenntnisse und Kompetenzen aus den verschiedenen Modulen des Studiengangs „Data Science“ **integriert und ausgebaut** werden: Sie werden die Kompetenz aufbauen, verschiedenen analytische Komponenten (bspw. ETL mit SQL, OLAP-Modellierung, explorative Analysen mit pandas, Datentransformationen, Feature Engineering, Machine Learning etc.) in einer „modernen“ Gesamtarchitektur aufzubauen. Dies bedingt zum einen die Umsetzung bereits bekannter Konzepte in operativ lauffähigen Systemen (bspw. ETL-Pipelines). Dafür sind die Hintergründe entsprechender Systeme zu durchdringen und deren Anwendung zu üben („Data Engineering“ und Veranstaltungen im Teil „Küppers“). Zum anderen sollen aber auch weitere Konzepte insbesondere im Kontext der Datentransformation und Datenvorverarbeitung erlernt werden, um Datenströme und Analyseprozesse sinnvoll entsprechend der Anforderungen in verschiedenen Praxis-Anwendungsfällen anwenden zu können. Neben den bereits vorhandenen technischen Kompetenzen sollen dabei auch die betriebswirtschaftlichen und organisatorischen Perspektiven eingenommen werden.

In der Projektarbeit haben Sie neben der Bearbeitung von **vorgegebenen Fragestellungen (im Folgenden Teil 1)** die Aufgabe, aus den gegebenen Datenquellen weitere **interessante Anwendungsfälle (im Folgenden Teil 2)** zu erarbeiten. Es werden somit Datensätze vorgegeben, die Sie allerdings anreichern und Ihren Interessen entsprechend auswerten werden. Sie sind in diesem Teil also sehr frei in der Gestaltung des Anwendungsfalls und dürfen auch Daten (teilweise oder falls nötig komplett) selbst generieren. Insgesamt sollte die Gesamtkomplexität in dem „freien Teil“ jedoch angemessen gehalten werden. Wichtig ist uns, dass Sie einen konsistenten Gesamtanwendungsfall und eine passende Systematik erreichen.

Die im Rahmen des Projekts prototypisch aufzusetzenden Systeme und Methoden sind als eine Vorstufe zum operativen Betrieb zu sehen, d.h. Sie müssen nicht jedes einzelne Detail „durchautomatisieren“, aber die verwendeten Tools müssen dazu grundsätzlich in der Lage sein. Die Umsetzung wird daher insbesondere in der **Cloud** erfolgen, da eine lokale Umsetzung skalierbarer / operativer Systeme mit akzeptablem Aufwand nicht erreichbar ist.

Ihre Aufgabe ist insbesondere eine **Spezifikation der Anwendungsfälle** (Pflichtaufgaben und freier Teil) inkl. aller Prozesse und Stakeholder, der darin angestrebten „Datenflüsse“ inkl. „analytischer Aufgaben“, die innerhalb Ihrer Architektur zu erledigen sind. Es sind dabei insbesondere die anzuwendenden analytischen Methoden und deren Integration in einen sinnvollen Datenfluss von der Datenquelle bis zur Senke (bspw. Dashboard, DWH-Tabellen, Prozesse, einzelne Stakeholder usw.) festzuhalten. Die „Konsumenten“ (Stakeholder und Prozesse) sowie deren Anforderungen an eine Versorgung mit Daten sind festzuhalten.

Auf Basis dieses „Datenflusses“ sind Sie aufgefordert, eine **Analytics-Architektur zu gestalten**: Definieren Sie die Datenquellen, Data Lake-Strukturen, ETL-Prozesse, aufwändigere Analyseprozesse (bspw. wiederkehrende statistische Analysten und/oder Machine Learning), notwendige Datenbanken und Tabellen, ggf. kleinere visuelle Reports etc.

Die Architektur ist **prototypisch umzusetzen**, d.h. alle von Ihnen definierten Komponenten sollten Sie von einer einfachen lokalen Umsetzung in Python (sofern nötig) in eine skalierbare Architektur überführen. Dies wird das Spezifizieren von ETL-Pipelines, Erzeugen von Datenbanken und Data Storage, sowie Implementierung komplexerer Analyse- und Transformationsschritte bedingen.

Abschließend sollen Sie Ihre Ergebnisse „reflektieren“ und die eingesetzten Methoden sowie die Architektur **evaluieren**: Welche Herausforderungen sind aufgetreten? Wie konnten diese gelöst werden oder waren alternative Lösungsansätze nötig? Welche Kosten würden unter bestimmten Annahmen (Datenvolumen, Anzahl Prozessdurchläufe etc.) für die Architektur (grob) entstehen? Welche Teile der Architektur sind skalierbar und konnten Sie dies beispielhaft zeigen (bspw. durch künstliche Verdoppelung des Datenvolumens)?

2 Gegebene Datensätze

Folgende Datensätze sind aufzunehmen (teilweise vorgegeben, teilweise ist die Datengrundlage selbst zu beschaffen bzw. zu errechnen):

(A) **Monatsstatistik im „Gastgewerbe“**
   a. Zeitraum: 01/1994 - 12/2023, Granularität: Monatsebene
   b. Quelle: Bereitgestellt vom Aufgabensteller hier (GDrive),
      Originalquelle: Hier (Statistisches Bundesamt)
   c. Deutschland (gesamt)
   d. Wichtig: Sie dürfen sich bei den Analysen auf einzelne Teilbereiche aus dem Gastgewerbe beschränken (Selektion über Code „WZ08W9“)

(B) **Feingranulare Wetterdaten** auf Tagesbasis
   a. Zeitraum: Variiert je nach Messstation, selektieren Sie zunächst passend zu (A)
   b. Quelle: Bereitgestellt vom Aufgabensteller hier (GDrive),
      Originalquelle: Hier (Deutscher Wetterdienst)
   c. Beschränken Sie sich zunächst auf die bereitgestellten Wetterstationen.

(C) **Aggregierte Wetterdaten** auf Monatsbasis
   a. Zeitraum, zeitliche und räumliche Granularität selbst zu bestimmen.
   b. Quelle: Selbst zu erzeugen aus (B)

(D) **Schulferien** in Deutschland / repräsentativem Bundesland / -ländern
   a. Zeitraum: Selbst zu ermitteln, anzupassen an weitere Datenquellen, Verfügbarkeit und Effizienz / Aufwand in der „Datensuche“ berücksichtigen, Granularität: Selbst zu ermitteln, anzupassen an weitere Datenquellen
   b. Quelle: Selbst zu ermitteln
   c. Deutschland (gesamt) bzw. Auswahl repräsentatives Bundesland / -länder (abhängig von der Verfügbarkeit)

3 Aufgabenstellung

Bitte setzen Sie folgende Aufgaben im Rahmen der Projektarbeit um:

(1) **Abruf, Aufbereitung und Kombination** der genannten Datenquellen
   a. Beachten Sie ein pragmatisches Vorgehen (bspw. müssen die Schulferien nicht den gesamten Zeitraum und auch nicht ganz Deutschland abdecken)

(2) Beantwortung **vorgegebener analytischer Fragen (Teil 1)**
   `Beschränken Sie sich für die folgenden Fragen zunächst auf die unter 2. (A) – (D) spezifizierten Daten. Falls Sie Erweiterungen für dringend geboten halten, sprechen Sie sie bitte mit den Aufgabenstellern ab.`
   a. Prüfen Sie, ob ein Zusammenhang zwischen der Durchschnittstemperatur und den Umsätzen der Campingplätze nachweisbar ist.
   b. Prüfen Sie zu (a) auch, ob es einen Zusammenhang mit zeitlicher Verzögerung gibt, und ermitteln Sie ggf. mit welchem Lag.
   c. Welcher der Wettermesswerte (= Spalten in den gegebenen Wetterdaten) hängt am stärksten mit den Campingumsätzen zusammen?
   d. Welcher Anteil der (saison- und trendbereinigten) Varianz der Campingumsätze lässt sich aus den verfügbaren Wetterinformationen erklären?

(3) Definition **weiterer analytischer Aufgaben und Datenquellen (Teil 2)**
   `Für diesen Aufgabenteil steht Ihnen frei, beliebige weitere Datenquellen zuzuziehen.`
   a. Festlegung der notwendigen Datengrundlage und Datenquelle(n)
   b. Erweiterung der o.g. Datenquellen und/oder Entwicklung eigener „fiktiver“ Datensätze
   c. Beispiel: Welche anderen erklärenden Variablen für die Umsatzschwankungen können Sie identifizieren? Wie könnte man diese Erkenntnisse operativ für einen Campingplatzbetreiber nutzen?

(4) Spezifikation **eines integrierten Anwendungsfalls** (Kombination von Teilen 1 und 2)
   a. Definition (fiktiver) **Geschäftsprozesse oder eines Geschäftsmodells**. Der Prozess oder das Geschäftsmodell sollen auf Basis der genannten Daten, Fragestellungen und analytischen Prozesse gestaltet werden (von „einfachen“ Reports bis hin zu optimierten/automatisierten Prozessen ist alles denkbar)
   b. Festlegung von **Stakeholdern und Konsumenten der Daten** (Festlegung verschiedener Anspruchsgruppen, bspw. Kapazitätsplanung eines Campingplatzbetreibers, Data Scientists, Gastronomie-Einkäufer mit Reporting-Anforderungen bzw. Self-Service BI-Nutzer, automatisierbare Prozesse, etc.)
   c. Spezifikation der **analytischen Anforderungen** (notwendige analytische Berechnungen, Berechnungen, Modelle, Transformationsschritte etc.) – **WICHTIG**: Aus Teil 1 dürfen die gegebenen Fragestellungen (teilweise) verwendet und mit den eigenen Ideen aus Teil 2 integriert werden.
   d. Zusammenführung der genannten Aspekte in einer **Modellierung des Datenflusses**
   e. **WICHTIG**: Aus den behandelten Bereichen im Teil „Küppers“ sind in dem Datenfluss und der Architektur die Aspekte „Data Warehousing / OLAP“ (durch ein entsprechendes Datenmodell und Ladeprozesse), „Big Data-Historisierung“ (=Data Lake-Aufbau), „Batch Processing“ (bspw. unter Verwendung von PySpark) sowie „Stream Processing“ (bspw. unter Verwendung von PubSub und Cloud Functions oder Spark Streaming) abzudecken.
   f. **WICHTIG**: Halten Sie alle Punkte einfach, um den Modulrahmen nicht zu sprengen!

(5) **Umsetzung der Architektur**
   a. Ggf. Entwicklung von Jupyter-Notebooks / lokalem Python- / SQL-Code zur Umsetzung der methodischen / analytischen Anforderungen
   b. Festlegung der zu verwenden Architekturkomponenten (Datenbanken, ETL-Tools, Analytics-Tools, Infrastruktur etc.)
   c. (Teilweise) Überführung des lokalen Codes in Architekturkomponenten
   d. Ggf. Entwicklung von Simulatoren für die Datenquelle(n) (bspw. Eingang von Wettermeldungen mittels PubSub-Streaming)

4 Organisation und einzureichende Dokumente

4.1 Ablauf

Die Übung ist durch Projektgruppen bestehend aus **4 Studierenden** (im Ausnahmefall auch 3 oder 5, bitte vorher abklären) umzusetzen.

- **Gruppencalls mit Betreuern (pro Gruppe 15min):**
  Kalenderwoche 19 (6.5.-10.5.) „Meilenstein 1st Draft“
  Präsentation der bisherigen Ergebnisse zu (1) und (2), Diskussion von ersten Ideen zu (3)-(5)
  Kalenderwoche 24 (10.6.-14.6.) „Statuscall Anwendungsfall und weitere Ideen“
  Finalisierung weiterer Datenquellen, konkreter Datenfluss, analytischer Aufgaben und eingesetzten Methoden, sowie des Architekturentwurfs (d.h. (3)-(5))

  WICHTIG: Der in diesen Calls gezeigte Fortschritt der Projektbearbeitung geht mit in die Benotung ein.
- **Individueller Test**: **Mittwoch, 26.06.2024**, ab 14:00 Uhr in E104 (60 Minuten)
- **Abgabe Projektbericht und „Deliverables“: Montag, 01.07.2024, 20 Uhr**
  Über ILIAS muss der Projektbericht von der Gruppe als PDF-Datei bis zur Deadline eingereicht werden (Abschnitt „Übungsaufgaben“ → „Baustein Übung“ – dort können Sie die Datei mit dem Projektbericht je Gruppe hochladen). Darüber hinaus müssen Sie in einer zip-Datei sämtlichen Code einreichen. Bitte fügen Sie im Anhang des Projektberichts eine entsprechende Ordner-Übersicht mit Erklärung der einzelnen Dateien ein. Cloud-Pipelines sollten Sie mit Screenshots dokumentieren und im Anhang abbilden.
- **Präsentation: Mittwoch, 03.07.2024, ab 11:30 Uhr** in 2,5 Vorlesungsblöcken (pro Gruppe 15 Minuten Präsentation, ca. 5 Minuten Diskussion). Achtung: die Präsentationen sind an dem Tag bis 11:30 Uhr in ILIAS als PDF einzureichen.

4.2 Einzureichende Dokumente und Prüfungsleistung

4.2.1 Projektbericht

- Erstellen Sie eine Titelseite Bezeichnung „DSCB420 – Projektbericht SoSe2024“ sowie die Matrikelnummern der Gruppenteilnehmer (keine Namen!)
- **12-14 Seiten (A4) inkl. Abbildungen**, die Titelseite zählt nicht dazu, verlagern Sie Details in den Anhang, beschränken Sie sich auf Kernaussagen und achten Sie auf Systematik
  - Schriftart Arial, Schriftgrad 11, Zeilenabstand „mehrfach 1,2“
  - Seitenränder 2,5cm
  - Abstand zwischen Absätzen 3 Punkte (Einstellung unter „Absatz“ in Word)
  - Abstand nach Überschriften 6 Punkte (es sollten Kapitel / Überschriften zur Strukturierung verwendet werden)
- Erstellen Sie einen Executive Summary (1/2 Seite, zählt nicht zu den 12-14 Seiten).
- Erstellen Sie ein Inhalts- und Abbildungs- sowie Tabellenverzeichnis (zählt nicht zu den 12-14 Seiten).
- Achten Sie auf Systematik im Argumentationsaufbau, präzise Formulierungen, ein durchgängiges Abbildungsdesign mit hoher Abbildungsqualität, sorgfältige Prüfungen der formellen Aspekte, eine zielgruppengerechte Aufbereitung der Inhalte, etc.
- Reichen Sie den Projektbericht als PDF-Datei ein.

4.2.2 Präsentation

Alle Gruppen müssen die Ergebnisse präsentieren und bei sämtlichen Präsentationen besteht Anwesenheitspflicht. Es müssen alle Teilnehmer der jeweiligen Gruppe präsentieren und Fragen beantworten können.

Die finale Präsentation ist eine Art „Pitch“: Nehmen Sie an, Sie müssen Ihren selbst definierten „fiktiven Auftraggebern“ aus dem Anwendungsfall Ihre Ergebnisse vermitteln.

Formelle Anforderungen an die Präsentation

- Zeitlicher Rahmen der Präsentation: 15 Minuten Präsentation, ca. 5 Minuten Diskussion
    - Genaues Timing ist ein Bewertungskriterium!

Wichtige Hinweise zu den Bewertungskriterien: Achten Sie insbesondere auf

- eine saubere Struktur und Systematik in der Präsentation (unterstützt bspw. durch eine Agenda),
- übersichtliche und leicht nachvollziehbare Folien (es gibt keine Formatvorlage, hier können Sie selbst kreativ sein),
- eine zielgruppengerechte Aufbereitung der Inhalte (Zielgruppe: „fiktive Auftraggeber aus Ihrem Anwendungsfall sowie den Pflichtfragen“),
- eine Darstellungsform, die auf eine Präsentation zugeschnitten ist (nicht 1:1 Kopie aus dem Projektbericht!) und klares Foliendesign,
- lesbare Abbildungen, Tabellen, etc.,
- einen flüssigen Vortragsstil ohne ablesen (üben!), sowie
- Nutzung „interaktiver“ Inhalte – sofern möglich und angebracht. Beispiele:
    - Darstellung von ETL-Pipelines (ggf. als Screencast mit „Zeitraffer“)
    - Vorstellung der Cloud-Komponenten aus der Architektur
    - Nutzung von Jupyter-Notebooks, um die Elemente der „methodischen Komponenten“ vorzustellen (entweder „lokal“ oder bspw. pyspark-Code).
    - Animationen können bei der systematischen Herleitung komplexer Inhalte helfen (aber nicht übertreiben!)
- Live-Vorstellung von Artefakten und ggf. „Dashboards“ für die Stakeholder, die das „Ende Ihrer Pipeline“ darstellen (nicht zu viel Aufwand in Dashboards stecken!)

4.2.3 Code und weitere Artefakte

Reichen Sie zusätzlich – abhängig von Ihrem Anwendungsfall und der Architektur – die erstellten Artefakte ein (vgl. „Ablauf“). Diese können sein:

- ETL Pipeline (bspw. in Spark/SQL/Data Fusion) – dokumentiert über Code bzw. Screenshots im Anhang
- pandas Code (Prototyping)
- Python Spark / Dask Code (operative Systeme)
- Transformationen in weiteren Tools – dokumentiert über Code bzw. Screenshots im Anhang
- Sämtlicher Code zu angewandten Methoden, Simulationen etc.

4.2.4 Sonstige Hinweise

Zur Individualisierung der Leistung findet ein Test statt (60 Minuten, 26.06.2024, ab 09:50 Uhr, E104).

Es gibt keine weitere Prüfungsleistung / Klausur.