Data Engineering Challenge

Für die Bewerbung als Data Engineer (d/w/m) im Bereich HR Data Science

Hintergrund: Im Team HCD wird aktuell ein Data-Lake-Projekt durchgeführt. Nach Konzeptionsphase und Aufbau der technischen Infrastruktur in Zusammenarbeit mit der konzerninternen IT sollen nun Datenquellen angebunden und für Anwendungsfälle aufbereitet zur Verfügung gestellt werden. In den Data Lake werden Daten aus verschiedenen internen und externen Datenquellen überführt, verarbeitet, gespeichert und zur Analyse und Exploration bereitgestellt.

Aufbereitung des Datensatzes

Untersuche den bereitgestellten Datensatz „Kündigungen_2017.csv“ hinsichtlich der Datenqualität:

- Welche Besonderheiten gibt es in der Datenstruktur?
- Können interessante Kennzahlen abgeleitet werden?

Automatisierte Verarbeitung

Implementiere einen automatisierten Workflow, mit dem die Daten erst aufbereitet und anschließend mehrere Kennzahlen berechnet werden. Dafür kannst du entweder deine eigenen Kennzahlen aus dem vorherigen Abschnitt oder die folgenden benutzen:

- Anzahl jährliche Kündigungen je Bereich
- Durchschnittliche monatliche Arbeitsstunden je Bereich

Bitte gehe dabei auf folgende Fragestellungen ein:

- Wie werden voneinander abhängige Prozessschritte orchestriert?
- Wie werden die Kennzahlen automatisch aktualisiert, sobald Daten für ein neues Jahr bereitstehen?
- Wie kann der Workflow einfach in einer AWS-Cloudumgebung betrieben werden?
- Wie speicherst du den bereitgestellten Datensatz, die berechneten Kennzahlen und wie können perspektivisch weitere strukturierte und unstrukturierte Daten aus anderen Quellen integriert werden?

Bereitstellung der Kennzahlen

Für eine Dashboard-App sollen die zuvor generierten Kennzahlen für ein angefragtes Berichtsjahr über einen REST-Service im JSON-Format zur Verfügung gestellt werden. Da die Kennzahlen potenziell vertrauliche Informationen enthalten, soll der Service nur eine Antwort liefern wenn die Dashboard-App im Request-Header eine Authentifizierung sendet.

Technische Rahmenbedingungen

- Deine Lösung ist in Python implementiert.
- Es steht dir frei, beliebige Frameworks und Libraries zu benutzen.
- Bitte erläutere die Gründe für deine Auswahl und mögliche Vor- und Nachteile dieser.

Bereitstellung deiner Lösung

Bitte stelle uns deine Lösung als Git-Repository (bspw. in einem Zip-Archiv oder als privates Repository auf GitHub o. ä.) zur Verfügung. Eine Beschreibung, wie der Code durch uns lokal ausgeführt werden kann, sollte enthalten sein.

Bewertungskriterien

Was für uns zählt, ist einen Eindruck davon zu bekommen wie du an Probleme herangehst und du deinen Code und deine Lösung strukturierst.

Hierauf achten wir beim Review der Data Engineering Challenge:

- Ist die Aufgabenstellung gelöst? Fehlt ein Teil der Umsetzung und wenn ja warum?
- Sind APIs und der Code ausreichend dokumentiert? Gibt es eine Anleitung, wie das Projekt aufgesetzt werden kann?
- Welche Technologien und Methoden wurden für die Umsetzung genutzt und sind diese sinnvoll? Begründe ggf. in der Dokumentation, warum du eine bestimmte Entscheidung getroffen hast.
- Ist der Code verständlich und kann er durch andere Kolleg:innen im Team gewartet werden? Welche Software Design Patterns wurden verwendet? Wurde ein Linter verwendet?
- Gibt es automatisierte Tests für den Code? Welche Art von Tests wurde gewählt und warum?

Präsentation

Zum Ende eines Sprints stellen wir unsere Arbeit dem Team und unseren Stakeholdern vor.

- Bitte bereite dich auf eine kurze Ergebnispräsentation (ca. 15 Minuten) vor.
- Die Wahl der Mittel ist dir freigestellt (bspw. Demo deiner Lösung und vorhandener Dokumentation, Powerpoint, PDF, Whiteboard, …)
- Du trägst gegenüber deinen Kolleg:innen, Management und Recruiting vor – deine Präsentation sollte also auch für nicht-technisches Publikum verständlich sein.
- Im Anschluss wird es eine kurze Diskussion und Raum für (technische) Rückfragen geben.