Data Engineering Prozesse

Eine Übersicht über die 5 wichtigsten Schritte im Data Engineering

1. Rohdaten sammeln

Im ersten Schritt müssen die Daten gesammelt werden, die später verarbeitet und analysiert werden sollen. Dazu gibt es im wesentlichen zwei verschiedene Möglichkeiten:

Zugriff via API (push/pull)
Die Daten liegen hier bereits in strukturierter Form vor und werden entweder in die Applikation eingespeist (push) oder von der Applikation selbst von fremden Servern o.ä. regelmäßig abgefragt (pull).

Screen Scraping (PDF, Website)
Hier liegen die Daten noch nicht in strukturierter Form vor, sondern müssen von Websites, PDF-Dateien oder ähnlichem ausgelesen werden. Als Beispiel dient hier das “Euros für Ärzte”-Projekt von correctiv, bei dem aus 40 unterschiedlichen Quellen Daten ausgelesen werden mussten.

2. Daten aufarbeiten

Im zweiten Schritt müssen die Daten durch verschiedene Prozesse vorbereitet werden, um sie im nächsten Schritt effizient speichern zu können:

Vereinheitlichen (zB verschiedene Strukturen)
Zunächst müssen die Daten in eine einheitliche Struktur gebracht werden, wenn dies noch nicht geschehen ist. Hier hilft nur ein Blick auf die Daten, um dann individuell die Datensätze zu verändern, sodass sie eine gemeinsame Struktur haben. Bei dem “Euros für Ärzte”-Projekt dauerte dies beispielsweise fünf Tage. Eventuell ist es hier auch notwendig, die Inhalte zu verändern, um so zum Beispiel Geoinformationen in einheitlicher Struktur vorliegen zu haben.

Duplikaterkennung
Eventuell ist auch eine Duplikaterkennung und -eliminierung notwendig, wenn man die Datenquelle nicht unter eigenem Zugriff hat.

Filtern
Noch vor der eigentlichen Speicherung können die Daten bereits vorgefiltert werden. Hierbei muss natürlich darauf geachtet werden, keine Daten zu löschen, die später noch relevant werden können, deren Wichtigkeit jetzt aber noch nicht bekannt ist.

3. Daten abspeichern

Im dritten Schritt werden die Daten so aufbereitet gespeichert, dass wir im nächsten Schritt diese einfach und schnell analysieren können.

Speichertechnologie wählen
Es muss eine Speichertechnologie gewählt werden, die es ermöglicht, für die gegebenen Daten schnell und unkompliziert Anfragen stellen zu können. Hierbei muss natürlich insbesondere auf Menge und Komplexität der Daten geachtet werden, dies entscheidet die Wahl für eine Technologie maßgeblich.

Speicherort wählen
Anschließend muss ein Speicherort gewählt werden. Das kann von der normalen HDD im lokalen Computer bis zu einem Hana-Cloud-Cluster alles sein, je nach gewählter Speichertechnologie und benötigtem Speicherplatz.

4. Daten analysieren

Da die Daten jetzt in passender Form vorliegen, können sie analysiert werden. Das ist häufig der aufwendigste Teil, da darüber nachgedacht werden muss, welche Fragen die Daten beantworten können, und hier auch kreative Lösungen gefragt sind.

Was wollen wir?
Es müssen Fragen formuliert werden, die mit dem vorliegenden Datensatz beantwortet werden sollen.

Lösungen für die Fragen finden
Auf die Fragen muss dementsprechend eine Lösung gefunden werden, also beispielsweise eine SQL-Abfrage formuliert werden.

5. Daten visualisieren

Damit der ganze Aufwand nicht umsonst war, muss das Ergebnis aus Punkt 4 nun dargestellt werden.

Zielgruppengerechte Darstellung
Die erste Frage ist die nach der Zielgruppe. Sollen die Ergebnisse nur dem Chef vorgestellt werden, oder der interessierten Öffentlichkeit? Je nach Zielgruppe kommen unterschiedliche Formate in Frage. Für die Öffentlichkeit bietet sich, wie bei dem “Euros für Ärzte”-Projekt, eine Website an.

Verschiedene Diagrammarten
Verschiedene Daten lassen sich mit unterschiedlichen Diagrammen am Besten darstellen. Hier ist etwas Geschick gefragt, die richtige Diagrammart auszuwählen, ein Tool wie beispielsweise *tableau* kann dabei helfen.

Interaktiv oder Statisch?
Zu guter letzt kann die Darstellung statisch oder interaktiv sein. Interaktivität bietet die Möglichkeit, mehr Informationen darzustellen, die auf den ersten Blick nicht sichtbar sind (z.B. bei Hover auf Datenpunkte), interaktive Websites oder Visualisierungen sind aber aufwendiger zu erstellen.

Florian Schmidt
5. Semester Bachelor

Data Engineering in der Praxis | Ringvorlesung
Dr. Ralf Krestel, Prof. Dr. Emmanuel Müller,
Prof. Dr. Felix Naumann, Dr. Matthias Uflacker

Hasso Plattner Institut
Digital Engineering · Universität Potsdam