Der neue Jobhit? „Data Scientist“, meist in Verbindung mit dem Schlagwort Big Data ist wohl der neue „Star“ im Berufshimmel

Aktuelle Trends in den Stellenangeboten im IT-Markt zeigen, dass die Suche nach Data Scientists sich zunehmend ausweitet. Kein Wunder, sprach der Harvard Business Review 2012 in dem Zusammenhang etwa vom "sexiest job in the 21st century".

Leute werden Brachenübergreifend gesucht. Das Anwendungsgebiet ist derart breit gefächert, so dass es in fast allen größeren Unternehmen, wie im Sektor der Automobilindustrie, der Luftfahrt, dem Finanzwesen oder der Versicherungsbranche hohen Bedarf gibt.

Nicht alle Unternehmen nutzen exakt die Bezeichnung „Data Scientist“. Stattdessen suchen sie häufig unter dem Titel „Big Data Engineer“. Der Grund für die Begriffsvielfalt liegt darin, dass es in der IT-Welt keine eindeutige Definition für Data Science gibt. So können auch die Anforderungen je nach Branche und Einsatzgebiet sehr variieren.

Häufig werden die Begriffe Data Science, Big Data und NoSQL vermischt. Das liegt daran, dass zumindest für die ersten beiden keine genaue Definition existiert. Man kann Big Data als den Informatik-lastigen Teil der Data Science ansehen, während man es im mathematischen Teil eher mit Begriffen wie Machine Learning oder Predictive Analytics zu tun hat. NoSQL dagegen ist nicht mit Data Science gleichzusetzen. Vielmehr handelt es sich um einen technischen Aspekt von Big Data und somit der Data Science, der sich mit dem nicht-relationalen Persistieren von Daten beschäftigt.

NoSQL-Kenntnisse allein bringen im Data-Science-Umfeld allerdings noch keinen Mehrwert. Stattdessen ist die Auswahl der passenden NoSQL-Technik für eine konkrete Aufgabenstellung und das Einbinden

eines NoSQL-Produkts in die Realisierung des Gesamtsystems von Interesse.

Zusammenfassend kann man sagen, dass es um das Erfassen und Auswerten von Datenmassen in jeglicher digitalen Form geht. Ganz gleich, ob die Daten als Datenbank oder als Dateien vorliegen.

Ein Data Scientist benötigt (mindestens) Kenntnisse in zwei klassischen Fächern: Mathematik und Informatik. Dazu kommt idealerweise noch Wissen aus dem jeweiligen Anwendungsgebiet. Kernaufgabe eines Data Scientist ist, aus diversen Datenquellen Antworten auf Fragen zu finden, die dem (internen oder externen) Kunden einen Mehrwert für einen konkreten Themenkomplex gibt. Im Projektalltag kommen dem Data Scientist unterschiedliche Aufgaben zu. Darunter fallen u.a.:

- Datensuche (welche Daten stehen zur Verfügung beziehungsweise welche lassen sich zusätzlich besorgen),
- Datenbereinigung (Aufbereitung der Daten für die anschließende Analyse),
- Offline-Datenanalyse (wie lassen sich aus den vorliegenden Daten die gewünschten Informationen extrahieren) und
- Überführen der Ergebnisse in ein produktives System zur Online-Analyse.

Eine Liste konkreter Techniken und Tools zu erstellen, die ein Data Scientist beherrschen sollte, ist aufgrund des weiten Aufgabengebiets jedoch schwierig. Es gibt z.B. Aufgaben aus den Bereichen Mathematik und Statistik, Business Intelligence, Mustererkennung oder Maschinelles Lernen, aber noch einige Andere.

Wer eine kompakte Intensivausbildung in der deutschen Big-Data-Hauptstadt Berlin sucht, sollte sich das Angebot des Data Science Retreat ansehen. Die Besonderheit des Programms ist die Tatsache, dass jeder Teilnehmer eine intensive persönliche Betreuung erhält. Jedem der maximal zehn Teilnehmern in den drei Monate dauernden Kursen stellen die Organisatoren einen Mentor zur Seite. Bei ihnen handelt es sich um Chief Data Scientists aus bekannten Big-Data-Unternehmen, die durch eigene Erfahrungen die für den Praxiseinsatz benötigten Fähigkeiten in das Programm einbringen.

Neben der Vermittlung der fachlichen Kenntnisse wird im Kursprogramm des Data Science Retreat auch viel Wert auf die benötigten Soft-Skills (Präsentations- und Kommunikationstechniken) und ausreichend Praxisbezug gelegt. Im Mittelpunkt des Programms steht für jeden Teilnehmer ein Portfolio-Projekt, in dem er, zusammen mit seinem Mentor, von Beginn des Kurses an das wahre Leben eines Data Scientist kennen lernt. Beispiele für Projekte vergangener Kurse findet man auf dem Blog des Data Science Retreat.

Das Projekt und damit auch der Kurs enden mit einer Präsentation der Ergebnisse am sogenannten Hiring Day. An ihm nehmen Personalverantwortliche aus diversen Unternehmen, größtenteils aus dem Berliner Raum, mit dem klaren Ziel teil, aus dem Kreis der frischgebackenen Data Scientists neue Mitarbeiter für ihr Unternehmen zu gewinnen. Die auf der Webseite des Programms veröffentlichten Zahlen zeigen, dass die Teilnehmer der bisherigen drei Kursdurchläufe dabei meist die Qual der Wahl hatten.

Für wen eine solche Intensivausbildung nicht in Frage kommt, dem bleibt meist nur der Weg über ein Selbststudium. Material dafür (meist in englischer Sprache und kostenlos oder zumindest kostengünstig) ist in

Form von Büchern und Online-Kursangeboten reichlich vorhanden. Wer also ein guter Autodidakt ist, hat an dieser Stelle eine große Auswahl. Im Hadoop-Bereich findet man entsprechende Angebote unter anderem bei den Anbietern der Hadoop-Distributionen Hortonworks, Cloudera und MapR, meist auch in Verbindung mit Sandboxen. Letztere bieten die Möglichkeit, die Einarbeitung in Testumgebungen in einem VMware- oder VirtualBox-Image durchzuführen.

In der Regel bieten die Unternehmen hinter derartigen Online-Kursen eine anschließende Zertifizierung zum "XYZ-zertifizierten" Hadoop-Entwickler an. Inwieweit solche Zertifizierungen schon ausreichen, um sich erfolgreich auf eine entsprechende Stellenausschreibung zu bewerben, ist fraglich.

Es gibt kein definiertes Vorgehen für das Durchführen von Big-Data-Projekten. Stattdessen kommt es viel auf Erfahrung, Ausdauer bei der Arbeit mit den großen Datenmengen und eine ausgewogene Mischung aus Domänen- und Technik-Wissen an. Diese Eigenschaften lassen sich nicht durch Zertifizierungen nachweisen. Ein Nachweis über den praktischen Umgang mit den Themen in Projekten oder die Veröffentlichung von eigenen Arbeiten, etwa im persönlichen GitHub-Profil, sagen meist mehr aus. Trotzdem kann die Teilnahme an solchen Online-Zertifizierungen sinnvoll sein, beispielsweise dann, wenn man für die Einarbeitung in Data Science ein Ziel benötigt. Nur darf man nach Erreichen dieses (Zwischen-)Ziels nicht glauben, dass man am Ende des Weges ist. Jose Quesada, der Initiator des Berliner Data Science Retreat, hat dies so formuliert: "A good GitHub profile is ten times better than any certification".

Fazit

Die Entscheidung, sich im Bereich Data Science zu spezialisieren, verspricht interessante und vielfältige Aufgaben. Material für ein Selbststudium ist in diesem Bereich genügend vorhanden. Es bleibt die persönliche Frage nach den dafür benötigten autodidaktischen Fähigkeiten und der zur Verfügung stehenden Zeit. Eine gute Lösung würde darin bestehen, dass ein Unternehmen ein Data-Science-Pilotprojekt startet, das dessen Mitarbeiter zum Wissensaufbau nutzen können. Auch für das Unternehmen ergäben sich daraus (mindestens) zwei Pluspunkte: Es hätte danach Mitarbeiter mit Data-Science-Know-how und – idealerweise – aus dem Pilotprojekt Erkenntnisse darüber gewonnen, ob und wenn ja in welcher Form Data-Science-Projekte der Firma Mehrwert bringen.

Nicht nur wegen der Aussicht auf attraktive Verdienstmöglichkeiten lohnt sich ein Einstieg oder eine Weiterbildung als Data Scientist. Es lockt eine eigenverantwortliche Tätigkeit im Bereich Big Data, der über die Einstufung als reines Hype-Thema hinweg ist und zukünftig wahrscheinlich noch viele Anwendungsgebiete zu Tage bringen wird, in denen zurzeit noch gar nicht über den Einsatz von Data Scientists nachgedacht wird.

Carla Marques Alvito
Pascal Rosenberg

www.alvitopersonalservices.com