TDWI ROUNDTABLE

Freiburg | 26. April 2018

tdwi VERANSTALTUNGEN

Data Lake – wie grenzt er sich das vom Data Warehouse ab?

Der 9. TDWI Roundtable Freiburg fand am 26. April hoch über den Dächern der Freiburger Altstadt in den Räumen der Allgeier Enterprise Services statt. Zum Fachvortrag des Abends, fanden sich trotz des frühsommerlichen Wetters wieder über zwanzig Zuhörerinnen und Zuhörer ein, darunter auch viele neue Gesichter aus Anwenderunternehmen der Region Südbaden.

Das Vortragsthema von Dr. Alexander Schätzle lautete „Data Lake – The Magic (and Madness) beyond the Gate“. Als Big Data Architect bei der badenIT GmbH, der IT-Tochter des örtlichen Energieversorgers badenova AG & Co. KG, stellen Dr. Schätzle und sein Team Infrastruktur und Services rund um Big-Data-Anwendungen bereit, wobei der Schwerpunkt auf Anwendungen im Machine-Learning-Bereich liegt. In seinem Fachvortrag ging Alexander Schätzle grundlegend auf das Data-Lake-Konzept ein. Zunächst zeigte er Use Cases auf und führte allgemein in das Hadoop-Ökosystem und den Zoo von Tools in diesem Umfeld ein. Dabei wurde deutlich, dass die Komplexität dieses Umfelds, alleine was die Vielzahl an Tools und Frameworks anbelangt, bei einem Produktiveinsatz im Unternehmensumfeld nicht unterschätzt werden darf.

Im weiteren Verlauf wurden Ähnlichkeiten und Unterschiede zum etablierten Data-Warehouse-Ansatz sehr schön herausgearbeitet. So wurde deutlich, dass Data Lake und Data Warehouse zwei sich ergänzende Konzepte sind, die sich vor allem hinsichtlich der Art der darin gespeicherten Daten und der Einsatzszenarien aber auch Anwendergruppen voneinander unterscheiden.

Möchte man eine Data-Lake-Architektur für unternehmenskritische Daten realisieren, so entstehen naturgemäß viele Anforderungen, die aus der Data-Warehouse-Welt wohlbekannt sind: Metadaten, Sicherheit, Auditing, Berechtigungskonzept, Datenqualität und Data Governance seien als hier als Stichworte genannt, die dem Data Warehouse Professional natürlich nicht fremd sind, aber in der Hadoop-Welt erst in den letzten Jahren adressiert worden sind und auch heute noch teilweise eine Herausforderung in der Umsetzung darstellen. Durch die eher explorative Verwendung des Data Lakes durch Data Scientists ergeben sich oft Unterschiede in der Realisierung dieser Anforderungen beim Data Lake. Auch gibt es in der Regel nicht nur eine Best Practice, sondern mehrere Wege zum Ziel, die immer spezifisch zum Projekt passen müssen.

Foto©: Prof. Dr. Tobias Hagen

Foto©: Prof. Dr. Stephan Trahasch

Der Referent betonte mehrfach, wie wichtig eine frühzeitige Beschäftigung mit diesen Fragestellungen ist, damit aus dem „schönen Data Lake" nicht im Handumdrehen ein Data Swamp – ein „Datensumpf" – wird, in dem zwar viele Daten gespeichert wurden, diese Daten aber nicht mehr in nutzbringende Informationen gewandelt werden können, zum Beispiel weil sie schlichtweg nicht mehr aufgefunden werden.

Überraschend für manche Teilnehmer war das Marktvolumen am Produkten und Services im Data Lake-/ Hadoop-Umfeld, das derzeit schon in der Größenordnung des BI-Markts liegt und enorme Wachstumsraten aufweist.

Im Anschluss an der Vortrag folgte eine rege Diskussion insbesondere zu Gemeinsamkeiten und Abgrenzungen zwischen Data Lake und Data Warehouse.

Beim anschließenden Networking wurde die Diskussion fortgeführt und die Teilnehmer der Community in Südbaden konnten Kontakte knüpfen und vertiefen – bei Sonnenuntergang und wunderbarer Aussicht über die Stadt. Die Organisatoren bedanken sich beim Referenten für den hervorragenden Vortrag und bei Allgeier Enterprise Services dafür, dass der Roundtable Freiburg die Räumlichkeiten am Fahnenbergplatz nutzen darf.

Der nächste TDWI Roundtable Freiburg ist für September geplant.

Prof. Dr. Tobias Hagen