Data Lakes

Raphael Drechsler

HTWK Leipzig
Fakultät Informatik, Mathematik und Naturwissenschaften
Studiengang Informatik Master - Matrikelnr. 69872

30.30.2018

Zusammenfassung

Der Begriff des Data Lakes ist 2010 entstanden und wurde in den letzten Jahren stark "gehyped".[1] [2] [3] Es haben sich viele verschiedene Konzepte und Ansichten zum Thema entwickelt. Im Internet findet man bei einer Recherche zum Thema Data Lake von einem existierneden Unternehmen, welches sich "the Data-Lake-Company" nennt[4], bis hin zu einem Blogeintrag, der die Frage "Are Data Lakes Fake-News?" mit ja beantwortet[5] eine ganze Menge. Dabei wird die Frage danach, was ein Data Lake ist, von den verscheidenen Quellen nicht eindutig beantwortet. Auch gibt es zum Zeitpunkt des Erstellens dieses Dokumentes in der deutschsprachigen Wikipedia nock keinen Eintrag zu diesem Thema. Die Motivation dieses Abstracts besteht also darin, die bestehenden Unklarheiten zu beleuchten; zu klären was ein Data-Lake ist und sich mit der Frage "Are Data Lakes Fake-News" auseinanderzusetzen.

I. Definitionsfrage "Data Lake"

Der Begriff des Data Lakes wurde erstmalig von James Dixon (CTO von Pentaho [1]) geprägt. Auf seinem Blog [6] und in mehren auf Youtube veröffentlichten Videos [7] stellt Dixon damals die von Pentaho angebotene Hadoop-basierte Big-Data-Lösung vor. Im Rahmen dieser Vorstellung stellt er auch das Prinzip vor, auf welchem die Solution basiert: Den Data Lake.
Dixon's Erläuterung des Prinzips beginnen damit, dass er durch Pentaho betrachtete Big-Data-Szenarien betrachtet und folgende gemeinsame Eigenschaften ableitet.

- Es liegt ein großes Datenvolumen vor, welches zu analysieren ist
- Die Daten entspringen einer Quelle
- Die Daten liegen in ihrer rohen Form vor (können also strukturiert, semi-strukturiert und un-strukturiert sein)
- ggf. sind die Daten angereichert (bspw. Anreichern von Weblogs um Geocodes)

[1] Pentaho gehört seit September 2017 dem Unternehmen Hitachi Vantara an

Liegt ein Daten-Volumen vor, auf welches diese Eigenschaften zutreffen, handelt es sich Dixon nach um einen Data Lake. Im Weiteren nennt Dixon zusätzliche Eigenschaften eines solchen Data Lakes. Im Kern der Betrachtung steht dabei, dass der Data Lake als Datenvolumen verschiedenen Anwendern über verschiedene Unternehmensbereiche bekannte und unbekannte (wenn auch kleinere) Fragen beantworten kann und es daher sinnvoll ist, dieses Datenvolumen für spätere Analysen abzuspeichern.
Der von Dixon ausgeführte bildliche Vergleich macht diesen Umstand und die Vorstellung davon, was ein Data Lake ist, noch deutlicher.

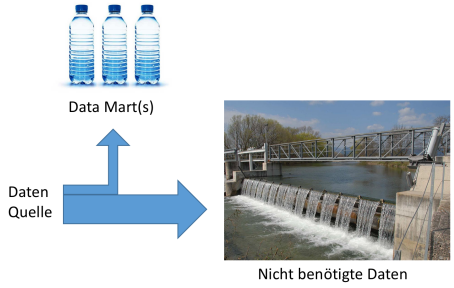

Abbildung 1: *DRT nach [7]*

Die Verbildlichung setzt bei den Data Marts an und stellt diese als fertig abgefüllte Mineralwasser-Flaschen dar. Das Wasser für diese Flaschen wurde aus einer Datenquelle gewonnen, bereinigt, aufbereitet und für den finalen Verwendungszweck abgepackt. Der Teil des Wassers (der Großteil), welcher nicht in die Data Marts eingegangen ist, fließt einfach ab. (Siehe Abb. 1)
Das Konzept des Data Lakes setzt an dieser Stelle an. Unter der Annahme, dass auch der Teil des Wassers, welcher abfließt, wertvolle Informationen enthalten kann, wird das Datenvolumen als Data Lake persistiert. Aus diesem lassen sich dann die Data Marts beliefern. Zusätzlich ist es dadurch möglich per Ad-Hoc-Query oder Report direkt auf das Datenvolumen zuzugreifen und somit zuvor unbekannte Fragen beantworten zu können. Zudem können Data Lakes wiederrum als Datenquellen für Data Warehouses genutzt werden. Es ergit sich das folgende Bild:

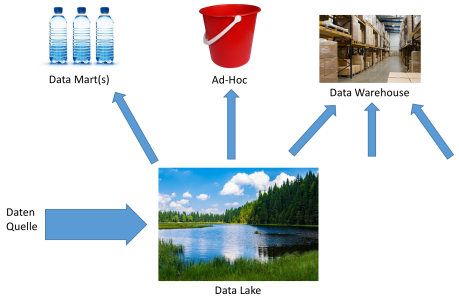

Abbildung 2: *DRT nach [7]*

Anmerkung zu extra-Pfeilen Data-Mart

Entsprechend folgt die Pentaho-Architektur 2010.

Die drei Schichten werden hier erstmalig gezeigt und sind klar.

Damit wars das an Definition. Niocht sehr genau. Weitere Unterkonzepte und Lösungen entstanden. Es gibt keinen einheitlichen Begriff.[9]

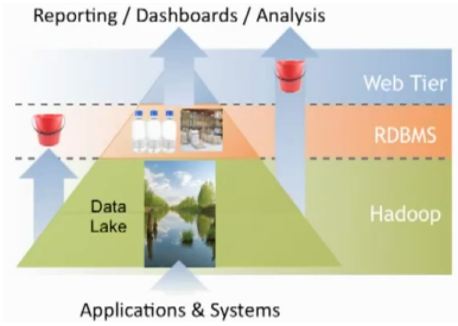

Abbildung 3: *DRT nach [8]*

II. Wie funktioniert ein Data Lake?

Begriffe die sich gefestigt haben.

Aufbau und Workflow Aufbau Analog zu Dixon Aufbau in drei Schichten.[10] [11]

- F
- P u S
- Viz

Dabei wird P und S gelegentlich synonym als Data Lake bezeichnet, was von Dixon abweicht.
Workflow
Zusammengefasst wiefolgt:

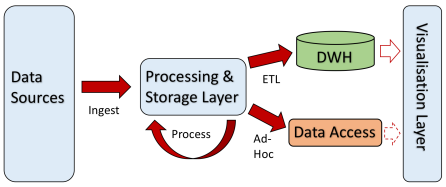

Abbildung 4: *DRT nach [11]*

Brauchen Daten nach DL
Brauchen aufbereiten des Wassers. Hierbei spielt die Rolle des Data Scientist eine Rolle.[10]
Dann zur Verfügung stellen
Oben: Visaulisierung. Was dabei visualisiert wird variiert von Sol zu Sol.

Storage QUELLEN : 8, ?

Ingestion QUELLEN : 8, 10, 11

Process QUELLEN : 8, 12

Consumption QUELLEN : 8, 12

Monitoring QUELLEN : 8, ?

Data Governance QUELLEN : 8, 12

III. Data Swamps: Kritik am Data Lake

Mögliche Darstellungen die es so gibt:

- Sumpf: findest nix und gehst unter [3]
- Finnland: heterogen, nicht zu inregreiren [12]
- Flohmarkt: Findest alles aber wie sucht man?, wem kann man vertrauen? Qualität? [13]

Gartners wesentliche Punkte
Aufstieg Data Lake durch scheinbar Löung im Problem. Konzept hat aber Lücken und wenig Substanz. Es kommt zu undergoverned und Meta-Daten-losen Hadoop-Clustern. Dies ist im wesentlichen das was unter dem Begriff Data Swamp verstanden wird.[3]

Battle: Gartner vs. Dixon und
Dixon setzt sich zur Wehr. Insbesondere Anzahl Quellen: Wassergartenarchitektur.[14] Auch Metadaten. Macht dazu keine Angaben, aber sagt, dass es nicht heißt, dass nicht.[15] Auf jeden Fall festhalten: ungenaue Definition.

Neben diesem Problem ansehen, was in der Praxis passiert: Hier ist Sean Martin zu zitieren.
Es kommt generell zu dem Trend des Vorsichtiger werdens und die Flut kommen sehen. Paradigmenwechsel.[1]

IV. Fake-News! Existieren Data Lakes überhaupt?

Blogeintrag nur nennen und erste Zeile zitieren.[5]

Nach Recherche lassen sich da schon ein paar Firmen finden, die Data Lake Lösungen anbieten. Unter anderem zu nennen sind : Firma[**?**], Firma[**?**], Firma[**?**],...

Nach weiterer Recherche auch success-stories auffindbar. Zu nennen sind hierbei die Storys von Firma[**?**], Firma[**?**], Firma[**?**]. Auch Zaloni hat Testemonial [**?**].

Im Bsp von UCI Health ist Lösung gut, weil [1][**?**]

Also irgendwie schon.

Die wesentliche Frage ist allerdings die Definitionsfrage.
Selber Schluss im Blogeintrag. Lösung die dem Paradigma grundlegend folgen gibt es. Jetzt im Auge des Betrachters ob man das Kind beim Namen nennt oder nicht.

Literatur

[1] Alan Morrison Brian Stein. Data lakes and the promise of unsiloed data. Technical report, PricewaterhouseCooper, 2014.

[2] James Ovenden. Say goodbye to your data lake in 2017. `https://channels.theinnovationenterprise.com/articles/say-goodbye-to-your-data-lake-in-2017`. Veröffentlicht: 10.01.2017, Zugriff: 29.04.2018.

[3] Rob van der Meulen Janessa Rivera. Gartner says beware of the data lake fallacy. `https://www.gartner.com/newsroom/id/2809117`. Veröffentlicht: 28.07.2014, Zugriff: 29.04.2018.

[4] Zaloni. Zaloni homepage. `https://www.zaloni.com`. Zugriff: 30.04.2018.

[5] Uli Bethke. Are data lakes fake news? `https://sonra.io/2017/08/08/are-data-lakes-fake-news/`. Veröffentlicht: 08.08.2017, Zugriff: 29.04.2018.

[6] James Dixon. James dixon's blog: Pentaho, hadoop, and data lakes. `https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/`.

Veröffentlicht: 14.10.2010, Zugriff: 29.04.2018.

[7] James Dixon. Pentaho hadoop series part 1: Big data architecture. https://www.youtube.com/watch?v=tR_yLsr87Uk. Upload: 24.10.2012, Zugriff: 29.04.2018.

[8] James Dixon. Pentaho hadoop series part 3: Overview. https://www.youtube.com/watch?v=_lCyXUA1iag&t=6s. Upload: 24.10.2012, Zugriff: 30.04.2018.

[9] Lance Weaver. Why companies are jumping into data lakes. https://blog.equinix.com/blog/2016/11/10/why-companies-are-jumping-into-data-lakes/. Veröffentlicht: 10.11.2016, Zugriff: 29.04.2018.

[10] Christian Mathis. Data lakes. *Datenbank-Spektrum*, 17(3):289–293, 2017.

[11] Bhushan Satpute. Enterprise data lake: Architecture using big data technologies. https://www.youtube.com/watch?v=hsq4s_l9ZDM&t=380s. Upload: 28.03.2016, Zugriff: 29.04.2018.

[12] Martin Willcox. What is a data lake, anyway. https://www.youtube.com/watch?v=NO0r452uQM0&t=835s. Upload: 10.02.2015, Zugriff: 29.04.2018.

[13] Alex Gorelik. How to build a successful data lake: Talk at hadoop summit 2016. https://www.youtube.com/watch?v=zHokpz3qNJ8&t=610s. Upload: 29.06.2016, Zugriff: 29.04.2018.

[14] James Dixon. Pentaho hadoop series part 5: Big data and data warehouses. https://www.youtube.com/watch?v=1CG01JmKp2Y&t=2s. Upload: 24.10.2012, Zugriff: 29.04.2018.

[15] James Dixon. James dixon's blog: Data lakes revisited. https://jamesdixon.wordpress.com/2014/09/25/data-lakes-revisited/. Veröffentlicht: 25.09.2014, Zugriff: 29.04.2018.