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 Erstellen 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"

Jetzt gehts los. **DRT:** Kein Akademischer Ursprung. Dixon hats gemacht. Quellen sind dabei sein Blog[6] und YT[7]. Bei YT wird das Produkt von Pentaho vorgestellt. In diesem Rahmen wird der Begriff geboren. Dixon's Betrachtung beginnt datmit, dass Dixon Big-Data Szenarios betrachtet. Feststellen von Eigenschaften. Diese sind:

- bla
- lo
- li

Dazu noch im Wesentlichen die Eigenschafen, dass unbekannte Fragen beantw können und keine 1 mio und verscheidene Anwender Bisschen das kleinfieh Prinzip Also Datenvolumen.

Späeter wird er drutlicher und führt das folgende Bild an:
Wenn Data-Mart = Wasserflasche
Wasser aus Datenqeulle, rest fließt ab

Paradigma: Wissen nich wie wertvoll das ist, was da abgeht. Also: Wasser in See, daraus

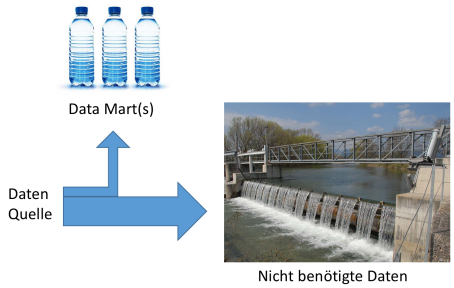

Abbildung 1: *DRT nach [7]*

Data Marts, auch Ad Hoc und DWH

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]

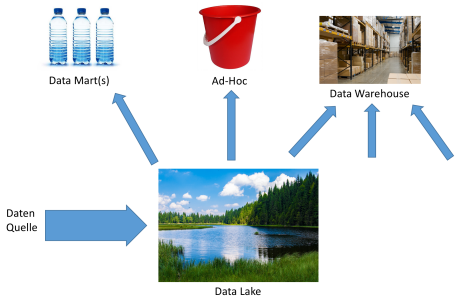

Abbildung 2: *DRT nach [7]*

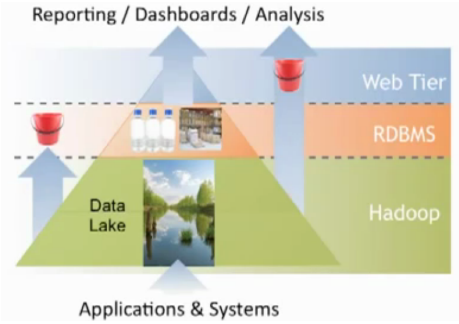

Abbildung 3: *DRT nach [7]*

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:
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, ?

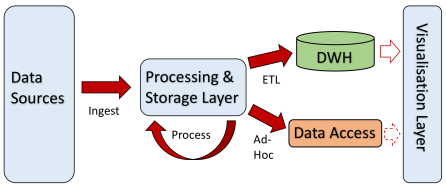

Abbildung 4: *DRT nach [11]*

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] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.

[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.