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.

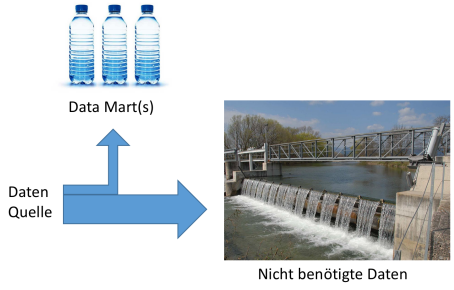

Abbildung 1: *Data Marts als Wasserflaschen nach [7]*

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 der Daten, welcher abfließt, wertvolle Informationen enthalten kann, wird das Datenvolumen als Data Lake persistiert. Aus diesem lassen sich die Data Marts beliefern. Zusätzlich ist es durch das Speichern möglich, per Ad-Hoc-Query oder Report direkt auf das Datenvolumen zuzugreifen und somit zuvor unbekannte Fragen beantworten zu können. Zudem können Data Lakes wiederum als Datenquellen für Data Warehouses genutzt werden. Es ergibt sich das folgende Bild:

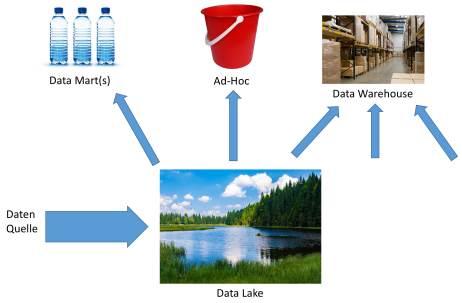

Abbildung 2: *Verbildlichung des Data Lakes nach [7]*

Diesem Prinzip folgend stellt Dixon die folgende Architektur der Pentaho-Solution vor.

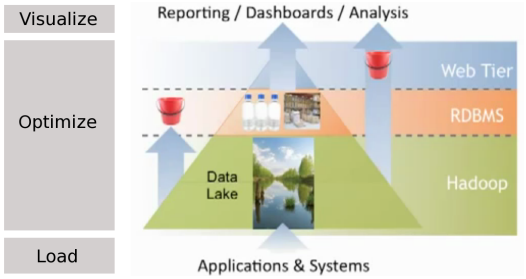

Abbildung 3: *Architektur Pentaho-Solution 2010 [8]*

Dabei finden sich die Elemente des Prinzips in den drei Schichten der Architektur (Load, Optimize und Visualize) wieder.[7][8]
Im weiteren Verlauf der Video-Strecke zur Solution geht Dixon auf die einzelnen Komponenten und deren Funktionsweisen ein. Im Wesentlichen ist die Definition des Data Lakes durch Dixon bzw. Pentaho an diesem Punkt abgeschlossen.
Da die Definition einigen Raum für Interpretation lässt, wurde der Begriff im Laufe der folgenden Jahre von verschiedenen Seiten unterschiedlich aufgefasst und teilweise neu interpretiert. Heute gibt es keinen einheitlichen Begriff des Data Lakes mehr.[9]

II. Wie funktioniert ein Data Lake?

Über die verschiedenen Lösungen und Konzepte, die zu Data-Lake-Solutions existieren, gibt es einige Gemeinsamkeiten. Diese sollen im folgenden betrachtet werden.

Aufbau und Workflow Der Aufbau einer Data-Lake-Solution ist zu der von Dixon dargestellten Architektur analog. Die Architektur besteht aus den folgenden drei Schichten.[10] [11]

- Data Sources: *Umfasst die Quell-Systeme bzw. Data-Streams inkl. der Daten, die das Daten-Volumen (den Data Lake) bilden*
- Processing and Storage-Layer: *Schicht zum Speichern und weiterverarbeiten des Daten-Volumens/Data Lakes*
- Visualisation-Layer: *Schicht in welcher die Daten aus dem DWH visualisiert werden oder/und eine Oberfläche für das Abfragen von Ad-Queries bereitgestellt wird. Weitere Komponenten und Formen der Visualisierung sind hierbei denkbar.*

Dabei wird die Processing and Storage-Layer gelegentlich als der Data Lake bezeichnet (vgl. bspw. [12]), was von Dixons Definition des Data Lakes als Datenvolumen (und nicht als Speicherort) abweicht.

Der Workflow in einer Data-Lake-Solution lässt sich wie folgt skizzieren. Dabei können die der Processing and Storage-Layer nachgelagerten Komponenten je nach betrachteter Solution variieren.

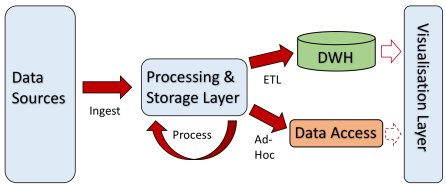

Abbildung 4: *DRT nach [11]*

Die Daten durchlaufen in diesem Prozess die folgenden Schritte.

- Ingestion: *(engl. für Aufnehmen). Die Daten werden aus den Quell-Systemen bzw. Data-Streams in die Processing and Storage-Layer geladen.*
- Processing: *Das persistierte Datenvolumen wird soweit aufbereitet, dass es für Analysen, Abfragen und schließlich Reports verwendet werden kann. Die Aufbereitung wird durch die Rolle des Data Scientist, welche für das .....DRT [10]*
- Bereitstellung für konsumierende Systeme *Die aufbereiteten Daten werden nun den nachgelagerten Systemen bereitgestellt.*

Wie sich an diesen Prozess die Visualisierung anschließt variiert - je nach eingesetzten Komponenten - von Solution zu Solution.

Im Folgenden sollen einige Detailfragen, die die Beschaffenheit der Komponenten einer Data-Lake-Solution und deren Zusammenspiel betreffen, näher beleuchtet werden.

Storage Für Storage Anforderung: Alle Daten speichern und getreu nach Dixon's Definition in Rohform. Daher Struct, Semi, un speichern. Folglich kann man da den Sceama on Read, bei dem das Schema beim Lesen also Abfragen aufgesetzt wird und Persistentschicht erstmal alles ok findet. Technologie allem vorran Hadoop zu nennen. Hat sich als on premis breitgemacht weil günstig, skalierbar und viele Tools. Ebenso Online Lösungen Hadoop Basiert. opt: File Based in On-Prmise und Object in online Hadoops vorteil ist Skalieren nach Scale up. Alles was dem Prinzip folgt ist für DL geeignet. [10]

Ingestion Aufnehmen der Daten in Datalake. Anforderung hierfür, dass sowohl Batch-basiert und Stream- Basiert. Tools bringt Apache mit ...(sb), ...(bb).[10] Zudem für Echtzeiteinsichten Kombination: Über Lambda-Architektur werden 2 Layers. Das Löst den Konflikt. [13] Alternativ Kappa. Da macht man das mit nur einem Stream bsaed, das in der Lage ist Replay zu machen. Das bietet den Vorteil, dass nur eine Sprache und weniger Last.[14] Diese Kombinationen sind dann auch für Processing genutzt - zieht sich ja bis zur Abfrage durch.

Process Daten im See. Jetzt Data Scientist Daten für Verwendung aufbereiten. Also erstmal schauen was das ist Data Profiling und festhalten im MetadatenKtalog. Daten zusammenbringen - integration. Wichtig jetzt: Schema hinzufügen. Erst dann kann man die eigentlichen Daten Analysieren. Dabei kann Machine Learning eingesetzt werden. Dann für das Konsumieren bereitstellen. Iterativer Prozess. Wird Mehrwert generiert muss Prozedur wiederholt werden wenn sich Daten ändern. Also abspeichern als WorkFlow. Dann noch sich wiederholende Aufgaben abspeichern als DataFlow für die Werkzeugkiste des Data Scientist. Opt: Auch überlegen ob Codierung zu colum seperated Files oder Einsatz von NoSQL DB als intergraler Bestandteil von P und S. Ggf. Mehr Administrationsaufwand und Speicher aber auszahlen in Zugriffsgeschwindigkeit und Zugriffsmöglichkeiten.[10][15]

Consumption Themen wie Visualisierung und Weboberfläche für Selfservice.[10][15]

Monitoring Einsatz von Moniitoring-Tool wie Apache Ambari zum Überwachen der Systemlandschaft.[10]

Data Governance Umfasst ein großes Umfeld und ist in Data Lake ein wunder Punkt. Hierzu im folgenden Abschnitt mehr.

III. Data Swamps: Kritik am Data Lake

Welche Kritikpunkte am Data-Lake-Konzept bzw. an Data-Lake-Solutions bestehen, wird deutlich, wenn man die existierenden Verbildlichungen von pathologischen Data Lakes betrachtet.

- Der Data Lake als Sumpf: *Der gespeicherte Data Lake ist nicht zu durchschauen und die Aufbereitung zu abgepackten Mineralwasserflaschen ist unverhältnismäßig aufwendig bis unmöglich.[3]*
- Der Data Lake als Finnische Seenplatte: *Der Data Lake ist stark heterogen. Die aus mehreren Quellen im Data Lake vereinigten Datenmengen bilden mehrere voneinander abgetrennte Teil-Seen die nur schwer oder nicht zu integrieren sind[16]*
- Der Data Lake als Flohmarkt: *Hier findet man alles. Es stellt sich jedoch die Frage, wie man effizient sucht und welche Qualität die angebotenen Waren (Daten) haben.[15]*

Gartner [2] beschreibt den Hype von Data Lakes darin begründet, dass das Konzept scheinbar eine Antwort auf die Frage nach mehr Agilität und Verfügbarkeit von Datenanalysen darstellt. Jedoch sei das Konzept lückenhaft. So kritisiert Gartner im Bericht *"The Data Lake Fallacy: All Water and Little Substance."*, dass das Aufnehmen sämtlicher Daten aus mehreren Quellen zu einem Data Lake führt, für den sich die benötigten Metadaten nicht ohne Weiteres erstellen oder gewinnen lassen, wodurch die gesammelten Daten ihren Wert verlieren. Zudem führt Gartner als wesentlichen Kritikpunkt an, dass das Konzept Data Lake keine Vorgaben zum Thema Data Governance macht.[3]

James Dixon bezieht 2014 zu dieser Kritik Stellung. Hierbei wird besonders ersichtlich, dass das von Gartner kritisierte Konzept einer Data-Lake-Solution von seiner ursprünglichen Definition aus dem Jahre 2010 abweicht. [17] So weist Dixon insbesondere darauf hin, dass der Data Lake nach seiner ursprünglichen Definition exakt eine Daten-Quellen akzeptiert und verweist für eine Solution, die mehrere Datenquellen aufnimmt, auf den sogenannten Wassergarten und die entsprechende Wassergartenarchitektur.[18] Bezüglich der fehlenden Metadaten merkt Dixon an, dass zum Data Lake nicht zwingend keine Metadaten vorliegen müssen. Genauer geht Dixon an dieser Stelle nicht auf die kritisierten Punkte ein, weswegen sich sie Kritik an einer ungenauen, lückenhaften Definition hält.

Sean Martin (Cambridge Semantics[3]) beschreibt, dass viele Firmen sämtliche Daten, in der Hoffnung sie später nutzen zu können, in Hadoop speichern. Jedoch verleiren Sie anschließend den Überblick darüber, was alles gespeichert ist. Bei einem Blick in die Praxis ist festzustellen, dass diese Gefahr einen Data-Swamp zu erzeugen bekannt geworden ist und sich daher ein Trend etabliert hat: Vorsichtiger werden. Primäre Aufgabe einer Data-Lake-Solution ist es nicht mehr alle Daten in Hadoop zu speichern. Stattdessen liegt der Fokus nun drauf, aus der gespeicherten Datenmenge einen Mehrwert zu erzeugen und nicht in der Datenmenge unterzugehen. [1] Diese Entwicklung kann als Paradigmenwechsel aufgefasst werden, da die neue Herangehensweise vom ursprünglichen Konzept (Alle Daten -wenn auch von nur einer Quelle- speichern) abweicht.

In jedem Fall rücken Data Governance und insbesondere die Beachtung der Metadaten als Schlüssel zu einer erfolgreichen Data-Lake-Solution in den Fokus. Dabei sind Data-Catalogue-Tools (Beispielsweise *Smart Data Catalog von Waterline* und *AWS Glue*) und spezielle Tools für Data Governance (wie *Apache Atlas* und *Cloudera Navigator*) sinnvolle Tools, um die für Data Governance relevanten Themen wie Data Lineage, Meatadaten-Suche, Datenquali-

[2] Gartner Inc. - Marktforschung und Analyse von IT-Entwicklungen

[3] Firma für Big-Data-Management und explorative Datenanalyse mit Sitz in Boston, Massachusetts

tät, Data Lifecycle-Management, Data Security und Data Integration anzugehen.[10]

IV. Fake-News! Existieren Data Lakes überhaupt?

Uli Bethke (CEO von Sonra[4]) stellte August 2017 in einem Blogeintrag[5] die Frage "Are Data Lakes Fake-News?" und beantwortete Sie mit ja. Das soll als Motivation dienen, um abschließend die Frage zu untersuchen, ob Data Lakes überhaupt existieren.

Nach einer kurzen Recherche im Internet lassen sich einige Firmen finden, welche Solutions anbieten, die den Gegenstand "Data Lake" im Titel tragen. Unter anderem zu nennen sind: HVR, Podium Data, Snowflake, Zaloni[19], Hitachi[20] und Hortonworks[21].

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.

[4] Unternehmen für IT und Services mit Sitz in Dublin

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] Matt Kalan. The future of big data architecture. https://www.mongodb.com/blog/post/the-future-of-big-data-architecture.

Veröffentlicht: 13.01.2017, Zugriff: 30.04.2018.

[13] Nathan Marz. How to beat the cap theorem. http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html. Veröffentlicht: 13.10.2011, Zugriff: 29.04.2018.

[14] Jay Kreps. Questioning the lambda architecture. https://www.oreilly.com/ideas/questioning-the-lambda-architecture. Veröffentlicht: 02.07.2014, Zugriff: 29.04.2018.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Timothy King. 4 data lake tools vendors to watch in 2018. https://solutionsreview.com/data-management/4-data-lake-tools-vendors-to-watch-in-2018/. Veröffentlicht: 17.04.2018, Zugriff: 02.05.2018.

[20] Hitachi Vantara. Hitachi website: Enterprise data lake. https://www.hitachivantara.com/de-de/solutions/data-analytics/enterprise-data-lake.html. Zugriff: 02.05.2018.

[21] Shaun Connolly. Enterprise hadoop and the journey to a data lake. https://de.hortonworks.com/blog/enterprise-hadoop-journey-data-lake/. Veröffentlicht: 15.03.2014, Zugriff: 02.05.2018.