Data Lakes

Raphael Drechsler

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

29.05.2018

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 verschiedenen Quellen nicht eindeutig beantwortet. Auch gibt es zum Zeitpunkt des Erstellens dieses Dokumentes in der deutschsprachigen Wikipedia noch 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] stellte Dixon damals eine 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 aus mehreren, durch Pentaho betrachtete Big-Data-Szenarien 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)*

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 folgende, von Dixon ausgeführte bildliche Vergleich macht diesen Umstand und die Vorstellung davon, was ein Data Lake ist, noch deutlicher.

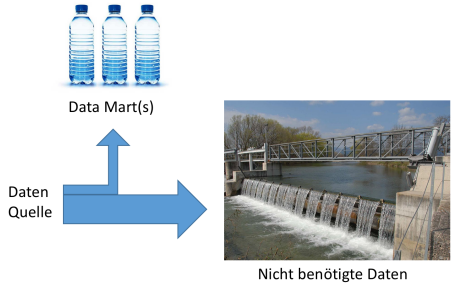

Abbildung 1: *Data Marts als Wasserflaschen nach [7]*

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

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 dabei wieder 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:

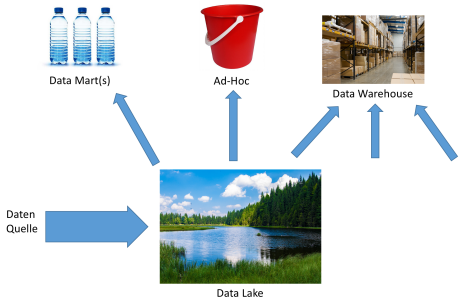

Abbildung 2: *Verbildlichung des Data Lakes nach [7]*

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

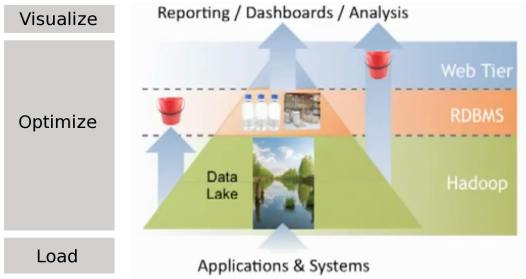

Abbildung 3: *Architektur Pentaho-Solution 2010 [8]*

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.*

Die Processing and Storage-Layer wird 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.

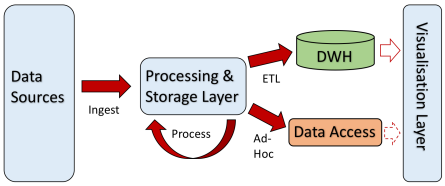

Abbildung 4: *Data-Lake Workflow nach [11]*

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 so weit aufbereitet, dass es für Analysen, Abfragen und schließlich Reports verwendet werden kann. Diese Aufbereitung obliegt der Rolle des sogenannten Data Scientist.[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 das Speichern des Data Lakes (Datenvolumens) bestehen die Anforderungen, dass zum Einen alle Daten gespeichert werden und dass zum Anderen die Daten getreu Dixons Definition in Rohform abgelegt werden müssen. Da daher sowohl strukturierte, semi-strukturierte und un-strukturierte Daten gespeichert werden müssen, ist eine sinnvolle Speicherung der Daten in einem RDBMS, welches vor dem Schreiben von Daten in die DB ein Schema voraussetzt (”Sceama on Write”), nicht ohne Weiteres möglich. Eine Lösung hierbei bietet der ”Sceama on Read”-Ansatz, bei dem jegliche Daten ohne definiertes Schema gespeichert werden und das Schema erst beim Lesen aus der DB über die Querys definiert wird.
Apache Hadoop folgt diesem Ansatz und hat sich als On-Premise-Speicher für Data-Lake-Solutions durchgesetzt. Es existieren auch Online-Speicher für Data-Lakes, welche auf Hadoop basieren (Google Cloud Platform, Amazon S3, Azure Data Lake). [10]

Ingestion Für das Aufnehmen der Daten in die Processing and Storage-Layer ist es erforderlich, dass Daten auf jede Art (also per Batch und per Streaming) aufgenommen werden können. Apache bietet für beide Arten der Datenaufnahme entsprechende Processing-Systeme an. Als Beispiele seien hier Apache MapReduce, Squoop und Spark als Batch-Processing-Systeme und Apache Flink, Storm und Flume als Streaming-Systeme genannt.[10]
Für den Fall, dass im Rahmen der Informationsgewinnung aus dem Data Lake Echtzeiteinsichten bezüglich Streams gewünscht sind, gilt es einen Konflikt zu lösen, der zwischen Verfügbarkeit und Konsistenz der Daten besteht. Die Stream-Daten müssen für spätere Auswertungen im Data-Lake persistiert werden, was jedoch Zeit kostet und somit die Möglichkeit auf Echtzeiteinsichten verwehrt. Durch den Einsatz einer Lambda-Architektur lässt sich dieser Konflikt dadurch auflösen, dass ein Batch-Processing-Tool (die Batch-Layer) eine Serving-Layer mit Daten beliefert. Eingehende Anfragen werden zu großen Teilen aus dieser Serving-Layer beantwortet. Das Delta zur Echtzeitinformation wird durch ein zweites, parallel laufendes Streaming-Tool (der Speed-Layer) aufgefüllt. Somit ist sowohl das Persistieren der Daten als auch eine Echtzeitauswertung möglich. (*Siehe Abbildung 5*) [13]

Eine Alternative zur Lambda-Architektur bietet die Kappa-Architektur. In dieser wird lediglich ein Stream-basiertes Processing-Tool

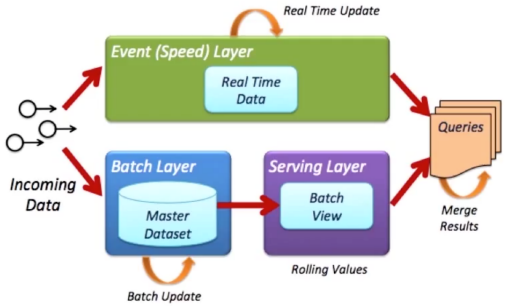

Abbildung 5: *Lambda-Architektur [14]*

eingesetzt. Dieses ist im Fall einer fehlerhaften Verarbeitung in der Lage mit teilweise persistierte Daten einen sogenannten "Replay" auszuführen. Dabei wird ein paralleler Streaming-Job gestartet, bis der Fehler ausgeglichen ist. Dies bietet den Vorteil, dass die entsprecheden Jobs für die Verarbeitung nur noch mit einem Tool implementiert werden müssen.[15]
Die Tools für die Datenaufnahme sowie die Lambda- bzw. Kappa-Architektur finden auch in den späteren Prozessen des Workflows (insbesondere Process und ggf. Consumption) Anwendung.

Process Sind die Daten in den Data-Lake gelangt, muss der Datensee für die Verwendung aufbereitet werden. Die entsprechenden Arbeiten werden von einer Rolle ausgeführt, die im Konzept des Data-Lakes als Data-Scientist bezeichnet wird. Der Data-Scientist muss die Daten zunächst im Schritt der Daten-Vorbereitung untersuchen. Er führt ein Profiling der Daten durch und hält seine Ergebnisse sinnvollerweise in einem Metadaten-Katalog fest. Hierbei ist es wichtig zu verstehen, worum es sich bei den vorliegenden Daten handelt und wie ihr ursprüngliches Schema definiert war. Darüber lässt sich erkennen, wie Daten verschiedener Quellen integriert werden können. Diese Vorbereitung ist notwendig, um getreu dem "Schema on Read"-Ansatz ein sinnvolles Schema für das Lesen der Daten definieren zu können.
Anschließend kann die Analyse der eigentlichen Daten und daraufhin die Bereitstellung der Daten für das Konsumieren erfolgen. Dabei können über den gesamten Arbeitsschritt des Processings mehrere Iterationen notwendig sein, um einen Mehrwert in Form eines Informationsgewinnes zu erzeugen.
Ist ein Mehrwert erzeugt worden, ist es im Hinblick auf die sich ändernden oder hinzukommenden Daten sinnvoll, die auf die Daten angewandten Operationen als sogenannten WorkFlow zu arrangieren und diesen WorkFlow periodisch oder Ereignis-gesteuert auf die Daten anzuwenden.
Ebenfalls kann es dem Data-Scientist von Nutzen sein, wiederkehrende, zusammenhängende Operationen als sogenannten DataFlow zu arrangieren und diesen DataFlow künftig als Tools für das Vorbereiten und die Analyse der Daten zu nutzen.
[10][16]

Consumption Für diesen Abschnitt des Workflows ist zu überlegen, welche Tools auf welche Daten des Data-Lakes zugreifen sollen. Fertig aufbereitete Daten können beispielsweise an nachgelagerte Systeme wie Data Marts, Data-Warehouses (via ETL-Prozess) und an Datenbanken von Unternehmensanwendungen weitergeleitet werden. Für die Visualisierung von Daten des DWHs bzw. von Data Marts lassen sich dann beispielsweise BI-Selfservice-Tools einsetzen.
Für den Zugriff auf die unaufbereiteten oder nur teilweise aufbereiteten Daten beispielsweise per Web-Oberfläche besteht eine Designfrage darin, welchen Anwendern man hierbei welchen Zugriff ermöglicht und welchen Mehrwert das bietet. Die Antworten auf diese Frage bedingen stark die Beschaffenheit der entsprechenden Web-Oberfläche für etwaige Self-Service Analysen. [10][16]

Monitoring Der Einsatz von Monitoring-Tools ermöglicht einen Überblick über die Vitaldaten der einzelnen Komponenten innerhalb der Systemlandschaft. Daher ist der Einsatz eines entsprechenden Tools (wie beispielsweise Apache Ambari) eine sinnvolle Maßnahme.[10]

Data Governance Diese Detailfrage ist für den Erfolg einer Data-Lake-Solution entscheidend. Der folgende Abschnitt soll erläutern weshalb.

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[17]*
- 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.[16]*

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. [18] 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.[19] 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 verlieren 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 für diesen Abstract 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[20], Hitachi[21] und Hortonworks[22].

Auch eine Suche nach erfolgreich umgesetzten Data-Lake-Solutions liefert Ergebnisse. Zu nennen sind hierbei beispielsweise die Success-Storys der Unternehmen Nissan[23], UC Irvine Health[24] und Pinsight Media[25] als Kunden von Hortonworks. Auch auf der Website von Zaloni - "the Data Lake Company" finden sich kurze, positive Statements der Kunden CDS Global und Enterprise Strategy Group bezüglich der umgesetzten Lösungen.[26].

Die wesentliche Frage ist jedoch, ob all diese umgesetzten und angebotenen Lösungen das Label einer Data-Lake-Solution tragen sollten. Inwiefern folgen die Lösungen dem ursprünglichen Konzept von Dixon bzw. Pentaho? Inwiefern weichen Sie davon ab? Und ist das ausschlaggebend dafür, dass eine Solution als Data-Lake-Solution gilt? Kurzum: Ohne genaue Definition der Begriffe Data-Lake und Data-Lake-Solution ist es nicht zweifelsfrei möglich Solutions diesen Begriffen unterzuordnen.
Zu einem ähnlichen Schluss kommt auch der Blogeintrag "Are Data Lakes Fake-News?". Hier heißt es, dass der Begriff "Data Lake" einige nützliche Konzepte (Data Reservoir und self-service analytics) fasst, jedoch letzenendes zu einer "cath-all-phrase" für alle Lösungen geworden ist, die nicht zum Thema Data-Warehousing gehören.[5]

Es ist also festzuhalten, dass Lösungen, die der grundlegenden Idee des Data Lakes folgen, existieren. Ob eine solche Solution aus diesem Grund das Label Data-Lake-Solution tragen sollte und welcher Mehrwert sich daraus ergibt obliegt der Einschätzung des Betrachters.

[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] Satyam Rai. Big data lambda architecture. https://www.youtube.com/watch?v=1CG01JmKp2Y&t=2s. Upload: 30.09.2015, Zugriff: 04.05.2018.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] Matt Spillar. How nissan is harnessing big data to provide value to customers. https://de.hortonworks.com/blog/nissan-harnessing-big-data-provide-value-customers/. Veröffentlicht: 13.11.2017, Zugriff: 03.05.2018.

[24] Hortonworks. Hortonworks kunden: Uc irvine health. https://de.hortonworks.com/customers/uc-irvine-health/. Zugriff: 03.05.2018.

[25] Tom Hastain. Pinsight media connects brands to audiences to be

first fueling intelligent ad decisions. https://de.hortonworks.com/blog/pinsight-media-connects-brands-audiences-first-fueling-intelligent-ad-decisions/. Veröffentlicht: 03.04.2017, Zugriff: 03.05.2018.

[26] Zaloni. Zaloni homepage: Solutions: Data lake in a box. https://www.zaloni.com/analytics-ready-data-lake/. Zugriff: 03.05.2018.