Elastisch und skalierbar – The new Data Lake in der Oracle Cloud

Harald Erb
ORACLE Deutschland B.V. & Co. KG, Frankfurt/Main

Schlüsselworte

Data Management, Data Lake, Big Data, Cloud, Explorative Analyse, Analytics, Notebook, Streaming Data

Einleitung

Vielerorts ist der Umgang mit „Big Data“, und wie man neue Erkenntnisse aus dem Rohstoff „Daten“ zieht, noch ein schwieriges Geschäft. Zahlreiche neue Datenquellen sind in ihrer Vielfalt, Menge und zeitlichen Aufkommen adäquat entgegenzunehmen und zu einem vertretbaren Preis verläßlich aufzubewahren. Der Datenschatz soll dann aber nicht ungenutzt im Keller (Speicher) lagern, sondern schnell für die aktive Nutzung, also für vielfältigste Analysen bereitgestellt werden können. Dabei reicht das Analysespektrum von der „klassischen“ Self-service Business Intelligence Abfrage bis hin zu Deep Learning, um dringende – vielleicht sogar existenzielle – Geschäftsfragen und Initativen schnell beantworten bzw. umsetzen zu können. Auf Basis der Oracle Cloud-Plattform lässt sich für solche Zwecke das in diesem Artikel näher beschriebene „New Data Lake“-Konzept schnell aufbauen und abhängig vom Analyse- und Resourcenbedarf betreiben.

Warum über Data Management nachdenken?

Verglichen mit heute war Data Management noch eine überschaubare Aufgabe: Geschäftsanwendungen verarbeiteten in Echtzeit Transaktionen mit Hilfe hochoptimierter relationaler Datenbanken, deren reibungsloser Betrieb heute noch unternehmenskritisch ist. In das zentrale Enterprise Data Warehouse werden laufend ausgewählte Stamm- und Bewegungsdaten aus allen relevanten internen Informationsystemen geladen, die zuvor bereinigt, harmonisiert, angereichert und für den Langzeitdatenbestand historisiert wurden. Diese Vorgehensweise und die qualitativ sehr hochwertigen Daten sind auch heute noch für das Geschäft unverzichtbar (Finanzabschluß, Berichtspflichten, KPI-basierte Steuerung usw.) aber zunehmend nicht mehr ausreichend, wenn man die bisher zu wenig berücksichtigten Datenquellen in Abbildung 1 betrachtet.

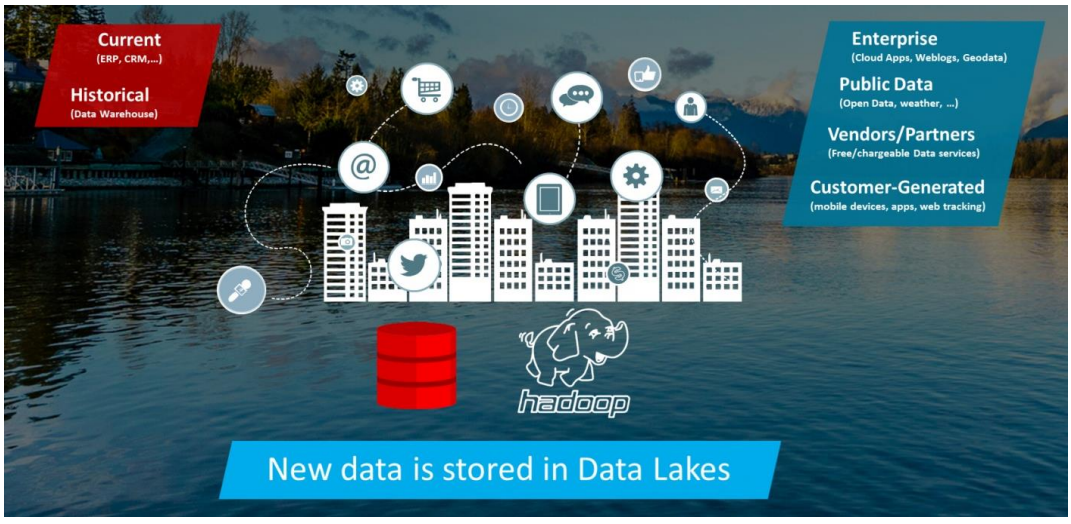

Abbildung 1: Wohin mit den neuen Daten?

Unternehmen, die den Wert ihrer eigenen Daten erkannt haben, überwinden mittlerweile interne Hürden sowie Datensilos und stellen diese unternehmensweit für Experimente (Data Labs, Hackathons) bzw. Umsetzung neuer Ideen (neue Services, Geschäftsmodelle) zur Verfügung. Besteht die Chance, bisher nicht realisierbare Projekte bewältigen zu können, kooperieren Unternehmen auch in Form von Joint Ventures miteinander und bringen dazu u.a. eigene Daten „mit". Allgemein bekannt und akzeptiert ist, dass es eine Vielzahl frei zugänglicher Datenquellen gibt (die wir mit Steuergeldern schon bezahlt haben), z.B in Europa von der Europäischen Union, Open Data-Initativen bis runter auf Städte- und Kreisebene, Forschungsinstitutionen usw. Die Datenbeschaffung ist aufgrund gängiger Datenformate (csv, XML, JSON) und Schnittstellen einfach, mit der Datenqualität und –vollständigkeit muss man sich dann allerdings selbst helfen. Komfortabler ist es dagegen, auf freie oder kostenpflichtige Datenprovider zurückzugreifen, die „Data as a Service" (DaaS) anbieten, wie von Oracle's Data Management Plattform (basierend auf Oracle BlueKai) selbst! Möchte man im Business-to-Consumer Marketing für eine Kampagnenplanung Internetaktivitäten und „digitale Spuren" der Konsumenten berücksichtigen, die heute von mehreren Geräten (Computer, Smartphone, Tablet) generiert werden, dann ist die Sammlung, Konsolidierung (möglichst pro Nutzer-ID) und Anreicherung der Daten eine zu bewältigende Herausforderung. Erst danach sind diese externen Informationen in Kombinationen mit eigenen Unternehmensdaten für eine bessere Kundensegmentierung – und ansprache auf den „richtigen Kanälen" effektiv nutzbar. Es gibt viele weitere Szenarien und Erfolgsstories, die datengetriebene Lösungen und Anwendungen zum Thema haben. Man kann sie grob in drei Kategorien einteilen:

» **Erweiterung bestehender und Erstellen brandneuer Anwendungen**: hier muss man nur auf sein eigenes Smartphone schauen oder an die intelligenten Geräte im Haushalt denken und hat schnell beliebig viele Anwendungsbeispiele parat, die insbesondere mit aktuellen Ort- und Zeitinformationen sowie Daten arbeiten, die ihre (im Idealfall identifizierten) Nutzer in großer Menge selbst generieren

» **Verbesserung der Analytik**: Hierüber liest man schon viel in Zeitungen oder Fachartikeln, wie Unternehmen mit mehr Detailinformationen zum Kaufverhalten und besseren mathematischen Modellen erfolgreicher Kaufabschlüsse tätigen, das Risiko der Kundenabwanderung minimieren usw.

» **Erfüllung regulatorischer Vorgaben**: Einzuhaltende Aufbewahrungsfpflichten, z.B. für Kassenbons mittels elektronischer Archivierung, werden heute eher so gelöst, dass die Daten einerseits revisionssicher vorliegen und andererseits produktiv genutzt werden können. Die besagten Kassenbons sind gleichzeitig auch die Datengrundlage für Warenkorbanalysen und helfen, das (sich über die Zeit ändernde) Kaufverhalten der Filialkunden besser zu verstehen.

Das Data Lake-Konzept

Der interessante Aspekt bei den oben eben skizzierten Beispielen ist die Fähigkeit, bei Bedarf in kurzer Zeit unternehmenseigene Daten mit den neuen Daten kombinieren, analysieren oder anderweitig produktiv verwerten zu können. Ein modernes Data Lake-Konzept muss unter Berücksichtigung betrieblicher Rahmenbedingungen genau diese Anforderung unterstützen und nicht nur für die kostengünstige Ablage neuer Daten (im Sinne einer erweiterten Staging Area für das existierende Data Warehouse) zuständig sein. Abbildung 2 hebt daher als wesentlichen Funktionsbereiche die agile Datenaufbereitung (Prepare) und vielfältige Analysefähigkeiten (Analyze) samt Data Lake Anwender (Data Consumers) besonders hervor.

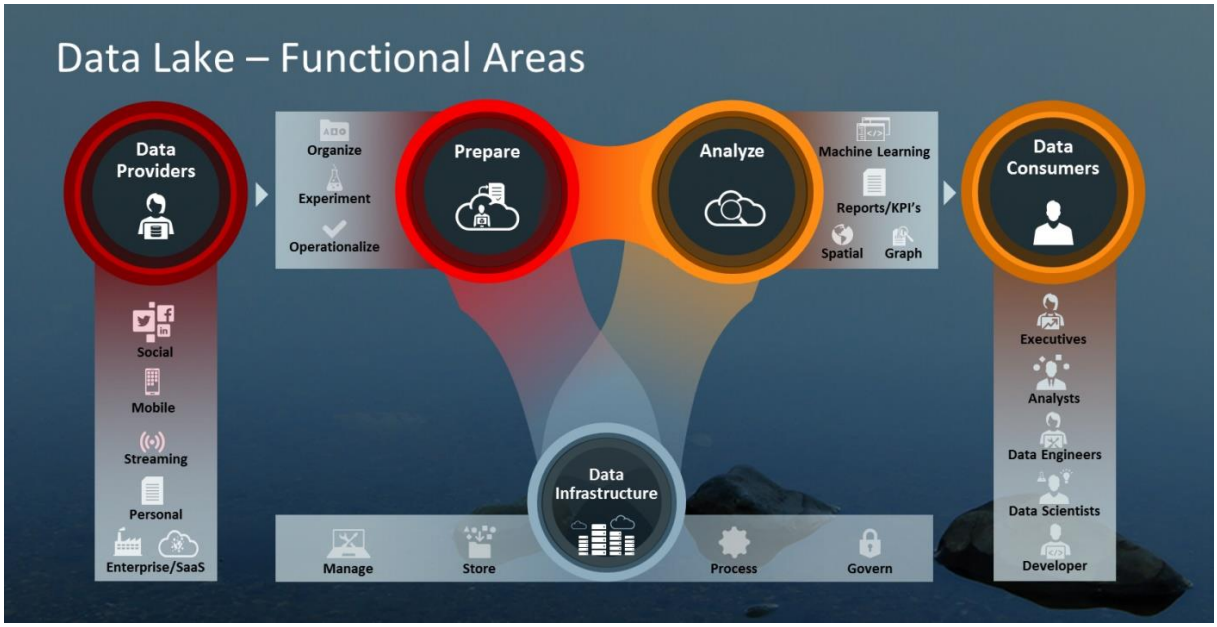

Abbildung 2: Funktionsbereiche eines Cloud-basierten Data Lakes

Im Vergleich zum klassischen Data Warehouse werden in einem Data Lake die eingehenden Daten nicht mit komplexen Datenqualitäts- und Integrationsverfahren in definierte Strukturen überführt, sondern direkt in ihrer Ursprungsform abgelegt. Damit können beliebige Daten schnell und einfach für Analysen nutzbar gemacht und beliebig verknüpft werden. Manchmal ist es aber auch erforderlich, Daten schon vor dem Speichern zu analysieren. Zum Beispiel um Echtzeitanforderungen (Warnung bei bestimmten Zustandsinformationen) bis zum vollautomatischen Prozess umzusetzen oder wenn sich eine vollumfängliche Speicherung technologisch/wirtschaftlich nicht rechtfertigen lässt oder eine Vorverdichtung stattfinden soll.

Als Schlüsseltechnologie für Big Data wurde in den letzten Jahren üblicherweise das Open Source Software Framework Hadoop angesehen, weil damit beliebige Datenarten in großer Menge verarbeitet und Berechnungen über viele Knoten eines Clusters verteilt werden können. Die große Idee hinter Hadoop war und ist es, die Daten mittels des Hadoop Distributed File Systems (HDFS) zu verteilen und die Analysen (z.B. mit Map Reduce) direkt bei den Daten durchzuführen. Datenspeicherung (Storage) und Rechenleistung (Compute) wurden dabei zusammengeführt mit dem Vorteil, dass man bei Analysen mit sehr großen Datenmengen das Bewegen eben dieser Daten vermeidet. Aber selbst bei frühen Data Lake-Konzepten erstreckete sich die Datenspeicherung durchaus über mehrere Data Stores (Repositories), vor allem war aber von Hadoop, relationalen Technologien und NoSQL-Datenhaltungen die Rede. In diesem Kontext bietet Oracle seit dem Jahr 2011 mit der Oracle Big Data Appliance ein Engineered System an, das u.a. die Cloudera Distribution of Hadoop (CDH) und weitere auf Oracle-Technologie abgestimmte Softwarekomponenten enthält und zum Aufbau eines Data Lakes eingesetzt werden kann.

Abbildung 3 zeigt die Konzeptsicht eines Data Lakes. Entgegengenommen Rohdaten (Raw Data) werden dort teilweise aufbereitet, um Fachanwendern die Analysen zu erleichtern. Dabei kommen Verfahren zur Datenqualitätssicherung (Datenbereinigung, -harmonisierung, Abgleich mit/Verwendung von Referenzdaten) und Datenanreicherung (z.B. Geocodierung) zum Einsatz. Diese aufbereiteten Daten (Curated & Transformed Data) werden dann oft einer großen Gruppe von Anwendern bereitgestellt, die bevorzugt über eine SQL-Schnittstelle und Abfragewerkzeugen auf die neuen Datasets zugreifen. Data Lakes adressieren üblicherweise eher explorative Anwendungsfälle, bei denen neue Fragestellungen häufig in Discovery Labs

(oder: Data Labs) unter Einbeziehung bisher nicht genutzter Daten untersucht und im Erfolgsfall in die produktive IT-Infrastruktur überführt werden.

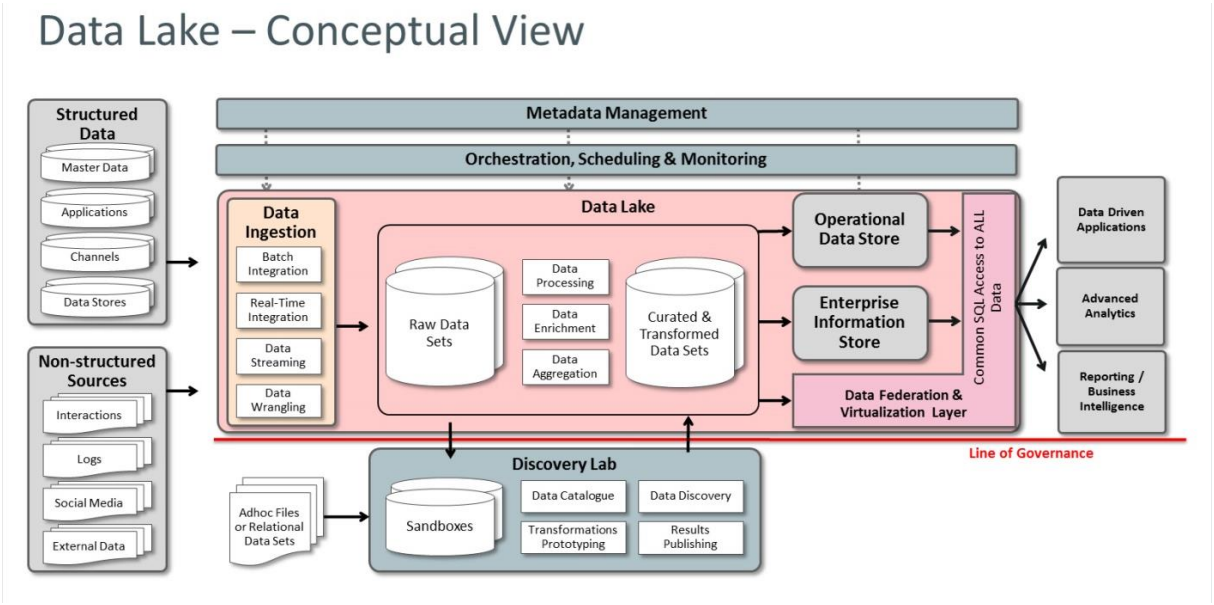

Abbildung 3: Data Lake – in der Konzeptansicht

Der aufbereitete Teil des Data Lakes ist von seinem Konzept her nicht weit von dem des Data Warehouse entfernt. Ziel ist vor allem aber die schnellere Umsetzung neuer Anforderungen, um auf dynamische Veränderungen im Geschäftsumfeld schnell agieren zu können. Die so gewonnene Agilität ist allerdings eine große Herausforderung aus Governance-Sicht. Diese beschränkt sich dabei nicht nur auf die allgemeine Zugriffssicherheit, sondern umfasst auch Aspekte wie Nachvollziehbarkeit der Datenflüsse, Dokumentation der Dateninhalte und Interpretationen (Data Catalog) oder aber auch Maskierung von Daten für bestimmte Bennutzergruppen bis hin zur Einrichtung eines Zonenkonzepts für unterschiedlich eingestufte Datenklassen. Effektive Governance erfordert daher einen ganzheitlichen Ansatz über den gesamten Prozess und Technologiegrenzen hinweg, um ein komplettes Bild über den Datenschatz zu erhalten.

Elastisch und skalierbar - Aufbau eines Data Lakes in der Oracle Cloud

In den letzten Jahren ist die Diversifizierung im Bereich der Technologien eher gestiegen. Beispiele dafür sind Graph-Datenbanken, um stark vernetzte Informationen abzuspeichern, darzustellen und analysieren zu können (Wer sind die Meinungsmacher zu aktuellen Themen in den sozialen Netzwerken?). Ein erkennbarer Trend der letzten Jahre ist auch, dass Object Storage und Apache Spark Hadoop zunehmend Konkurrenz machen, da die feste Verbindung von Compute und Storage Einschänkungen in Bezug auf flexible Skalierung und Elastizität angeht. Dagegen würde eine Trennung von Storage und Compute das dynamische Hinzufügen von Rechenkapazität für sehr komplexe Berechnungen ermöglichen. Genau diesen Aspekt adressiert Object Storage. Dieser ist sehr einfach administrierbar und gleichzeitig sehr kostengünstig und skalierbar und ermöglicht das effiziente Speichern auch sehr großer Datenmengen. In Kombination mit Apache Spark, das als Framework für Cluster Computing für die Parallelisierung der Berechnung sorgt, lassen sich damit sehr gut komplexe analytische Berechnungen durchführen. Dabei müssen die Daten allerdings in die entsprechenden Spark Compute-Knoten geladen werden und somit bewegt werden. Da aber häufig nur kleinere Datenausschnitte betrachtet werden, hat dies oftmals keine negativen Performance-Auswirkungen und wird durch die In-Memory Verarbeitung von Spark überkompensiert.

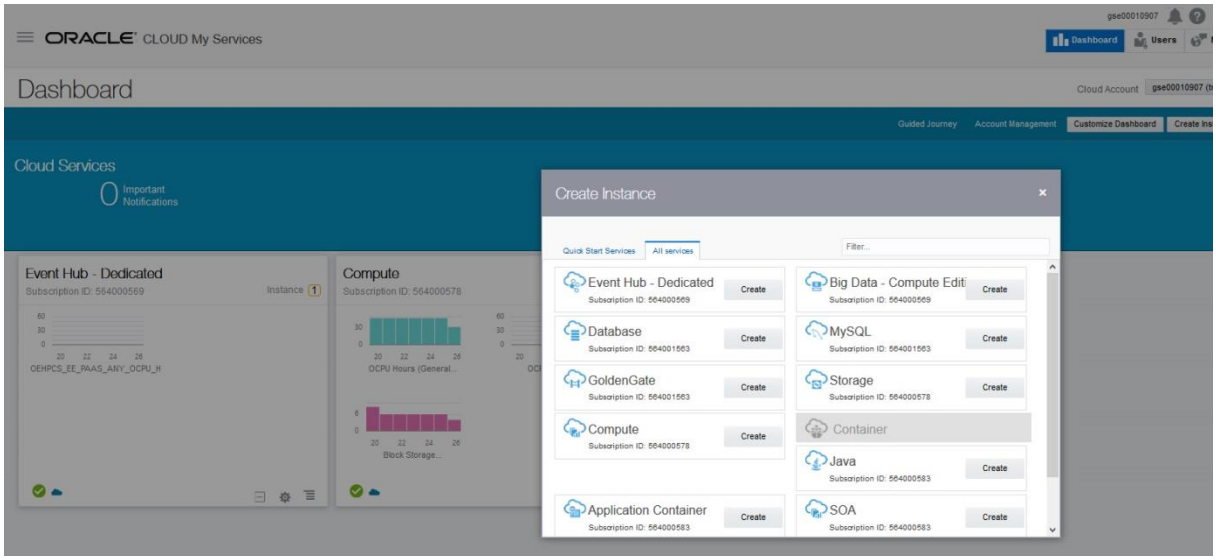

Abbildung 4: Service Console zur Verwaltung der Oracle Cloud Services

Wer über Flexibilität bei der Nutzung von Speicherplatz, Rechenleistung oder die benötigte Anwendungssoftware für die Datenaufbereitung und –analyse nachdenkt, wägt sicher auch den Aufbau einer eigenen IT-Infrastruktur gegen das Beziehen von Cloud Computing Services miteinander ab und sollte sich unbedingt das modulare Angebot in der Oracle Cloud näher ansehen. Abbildung 4 zeigt, wie sich die erforderlichen Oracle Cloud Services für den Aufbau eines Data Lakes auswählen lassen, um die eben besprochenen Big Data Schlüsseltechnologien Hadoop, Spark usw. zusammen mit den Vorteilen von Object Storage zu nutzen:

» **Oracle Big Data Cloud Service – Compute Edition:** mit diesem Service können Big Data Cluster innerhalb weniger Minuten bereitgestellt, bei Bedarf um weitere Rechenkapazität erweitert aber auch wieder verkleinert und als Datendrehscheibe bzw. Processing Engine mit anderen Cloud Services gekoppelt werden. Ausgefallene Clusterkomponenten und -knoten werden im Rahmen der kontinuierlichen Überwachung ohne menschliches Eingreifen automatisch korrigiert. Administrationsaufgaben sind variabel über grafische Benutzeroberflächen oder per Maschine-zu-Maschine-Kommunikation über REST API's durchführbar.

» **Oracle Event Hub Cloud Service:** eine von Oracle verwaltete Streaming-Plattform, die mit Apache Kafka eine weitere Big Data Schlüsseltechnologie verwendet. Vor allem, wenn es um die Verarbeitung von Protokolldaten geht, z.B. basierend auf Aktivitäten in sozialen Netzwerken oder Sensordaten, werden Messaging-Systeme beim Sammeln, Analysieren und Verteilen dieser Datenströme vor große Herausforderungen gestellt. Apache Kafka besticht in diesem Feld vor allem durch sehr hohen Datendurchsatz (mehrere Tausend Nachrichten pro Sekunde sind kein Problem). Kafka-Komponenten verbinden Arbeitsspeicher, Cache von Speichersystemen und die Speicherverwaltung des lokalen Betriebssystems miteinander, sind im Cluster-Betrieb auf mehreren Rechnerknoten verteilbar, so dass eine effiziente Verteilung der Rechen- und Speicheraufgaben ermöglicht wird.

» **Oracle Storage Cloud Service:** Neben anderen Speicher-Services (z.B. für Backup und Archivierung) bietet Oracle den besagten Object Storage für die sehr preiswerte und flexible Speicherung beliebiger Data sets, also für strukturierte und unstrukturierte Daten an. Im Gegensatz zu den bekannten Dateisystemen enthalten die Objekte zwar die Daten, sind allerdings nicht in einer Hierarchie organisiert. Jedes Objekt befindet sich auf der gleichen Ebene eines Adressraums, wird mithilfe seiner erweiterten Metadaten charakterisiert und bekommt einen einzigartigen Identifikator

zugewiesen. Somit können Server oder Endanwender das Objekt beziehen und müssen den physischen Standort der Daten nicht kennen. Diese Herangehensweise ist für die Automatisierung und Rationalisierung der Datenspeicherung in Cloud-Computing-Umgebungen nützlich und darüberhinaus auch preisgünstiger.

Aus der Anwendersicht haben damit Entwickler, Data Engineers und Data Scientists bereits die relevanten Cloud Services, die sie zum Arbeiten und Analysieren benötigen. Diese Personengruppen werden dabei feststellen, dass Oracle auch bei der Ausgestaltung seiner Cloud Services eine lange Tradtion fortsetzt und weiterhin Open Source Technologien unterstützt bzw. ergänzend nutzt. In Abbildung 5 sind dafür stellvertretend einige Werkzeuge eingetragen: CloudBerry (ein komfortabler Windows-Client für Dateimanagement und -transfer); für Entwicklung und Analyse kommen webbasierte Notebooks (Jupyter, Zeppelin) mit einbindbaren Interpretern zum Einsatz, die zahlreiche Programmier- und Skriptsprachen unterstützen; RStudio für Arbeit mit der Statistik-Programmiersprache R. Statt leicht bedienbarer Benutzeroberflächen ist es für diese Klientel wichtiger, möglichst schnell neue Technolgieversionen (Spark, Hive usw.), Programmpakete (inbesondere für Python und R) oder populäre Frameworks (z.B. TensorFlow für Deep Learning inklusive Grafikprozessor- Support) auf der Big Data Plattform einsetzen zu können.

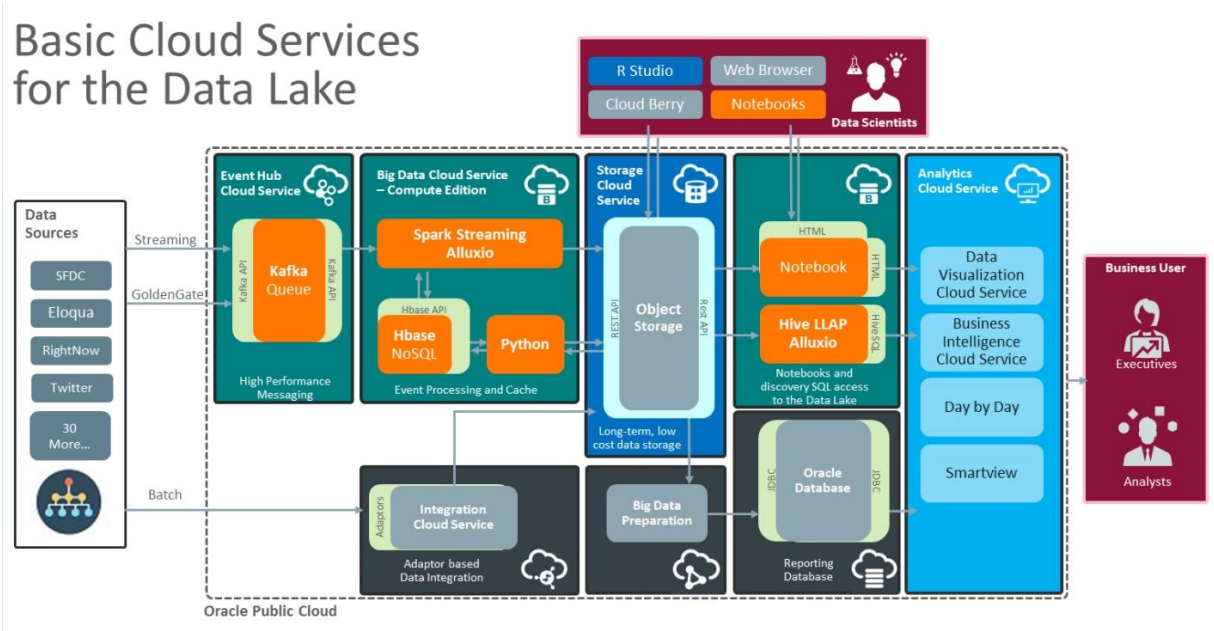

Abbildung 5: Event Hub, Big Data und Storage Cloud Services bilden die Basis für den Aufbau eines Cloud Data Lakes

So sind also die heutigen Anforderungen an die Data Lake Anwender durchaus höher im Vergleich zu den Business Intelligence Anwendern von „früher“. Oftmals sind Programmier- und Skriptingkenntnisse erforderlich, wenn bei Datenexperimenten im Discovery Lab oder bei der direkten Verwertung neuer Datenquellen nicht auf die formalen IT-Entwicklungsprozesse gewartet werden kann. Um dem zu begegnen ist für die Fachanwender erfreulicherweise ein Trend zu benutzerfreundlichen Werkzeugen für alle Aspekte der Wertschöpfungskette vom Laden der Daten bis zum Analysieren erkennbar. Entsprechend bietet auch Oracle für die Arbeit mit dem Data Lake u.a. Daten Integration und Big Data Preparation Cloud Services an und hat Reporting-, Data Discovery- sowie Analytics-Funktionen in einem Cloud Service gebündelt:

» **Analytics Cloud Service:** ein interaktiver Cloud Service für das Data Lake und die klassische rollenbasierte Informationsversorgung des Unternehmens mittels Business Intelligence Dashboards, Reports, proaktiven Alarmen usw. Für die ambitionierten Fachanwender ist besonders das Werkzeug Oracle Data Visualization interessant, mit dem sich mit Hilfe einer grafischen Oberfläche Daten aus

einer Vielzahl von Quellsystemen laden bzw. live abfragen lassen. Da die Ursprungsdaten nicht immer 100%ig perfekt für die geplanten Analysen sind, fällt den Anwendern damit nun die eigenständige Durchführung der Datenaufbereitung zu. Mit den integrierten „Lightweight-ETL"-Funktionen ist dies aber ohne Programmierkenntnisse möglich. Abbildung 6 gibt hierzu einen Eindruck, wie sich Data sets organisieren, vor der Analyse inspizieren (Lightwight Data Profiling) und in Data Flows kombinieren lassen. Durch die einfache Bedienung und enge Integration aller Funktionen können Fachbenutzer ihre Analysen innerhalb einer Anwendung von der Idee über mehrere Iterationen hinweg bis zum Endergebnis durchführen, Analyseschritte dokumentieren und für Live-Präsentationen aufbereiten.

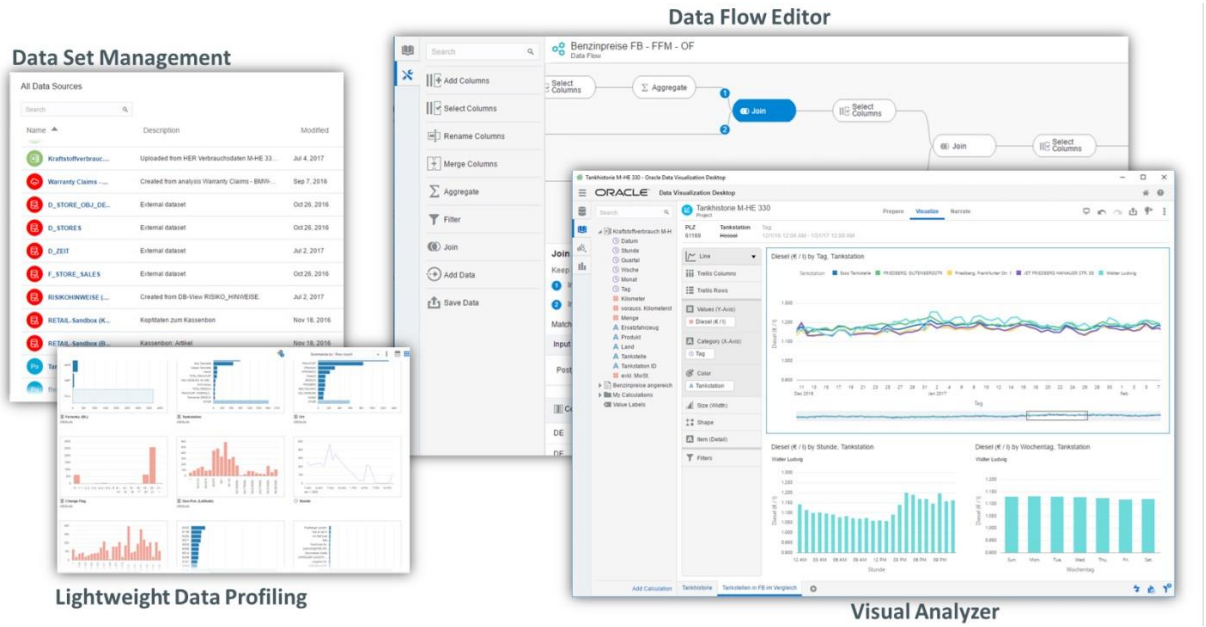

Abbildung 6: Datenexploration und visuelle Analyse in der Oracle Analytics Cloud

Fazit

Ein Data Lake ist mehr als nur Hadoop und ein spannendes Thema – wie die rasanten neuen Möglichkeiten und technologischen Entwicklungen zeigen. Es wird das Data Warehouse-Konzept nicht verdrängen, sondern in den meisten Fällen ergänzen. Mit Data Lakes können viele Anwendergruppen dynamisch und agil arbeiten, komplexe analytische Aufgabenstellungen sind mit dieser Infrastruktur lösbar. Oracle's Cloud-Umgebungen führen im Big Data Analytics Umfeld dazu alle benötigten Fähigkeiten zusammen und bieten beste Bedingungen für neue Entwicklungen bei gleichzeitiger Integration in die bestehende Systemlandschaft.

Kontakt:

Harald Erb
ORACLE Deutschland B.V. & Co. KG
Robert-Bosch-Straße 5
D-63303 Dreieich

Telefon: +49 (0) 6103-397 403
E-Mail harald.erb@oracle.com
Internet: www.oracle.com
LinkedIn: www.de.linkedin.com/in/haralderb