Arvato Systems Whitepaper

Modern Data Management mit Amazon Web Services

ARVATO SYSTEMS

Inhaltsverzeichnis

Einleitung

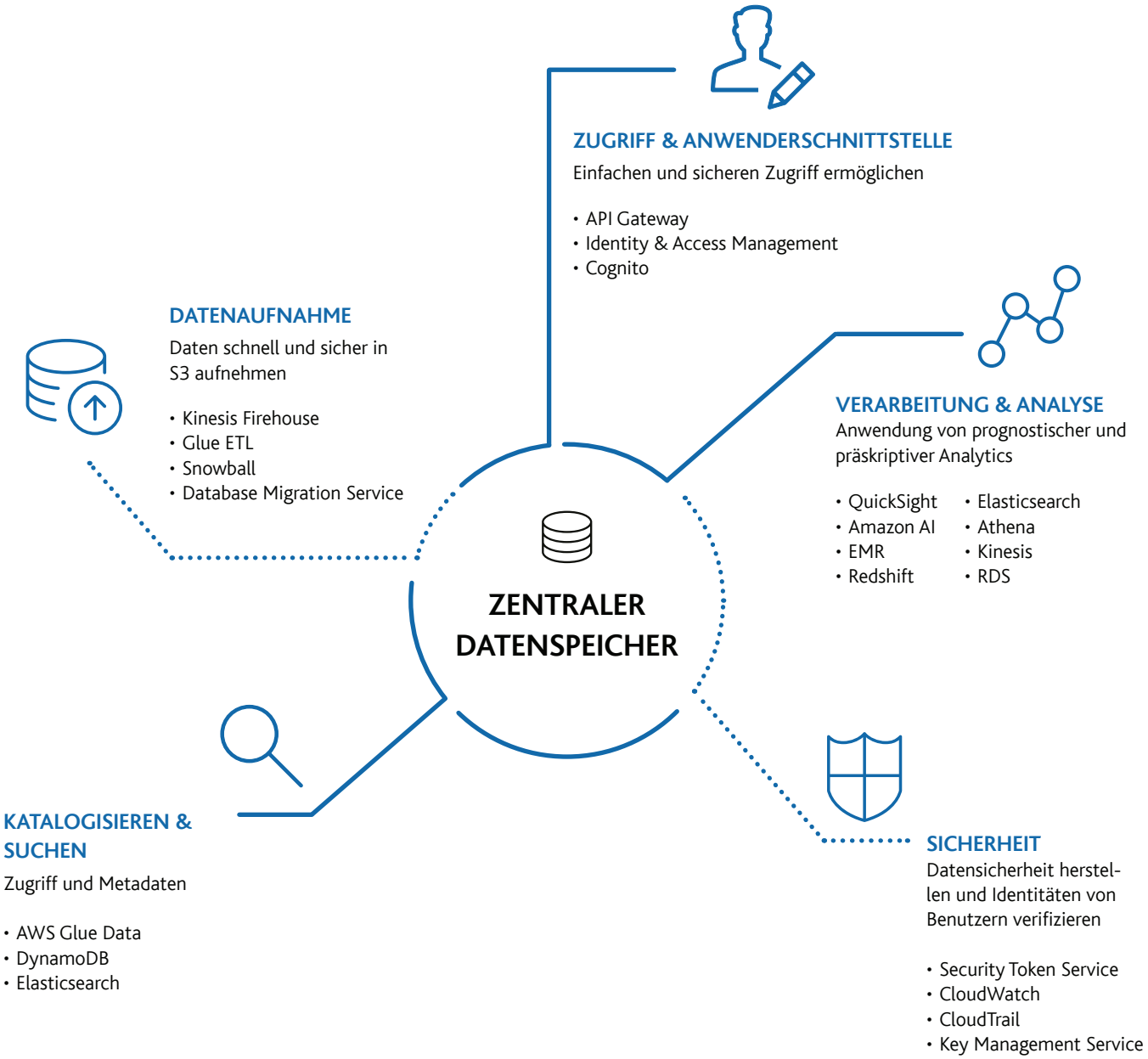
Im Whitepaper „Die 5 Herausforderungen der Datenverarbeitung und des Datenmanagements“ wurden bereits die Probleme entlang der Datenwertschöpfungskette sichtbar gemacht und aufgezeigt, dass klassische, silierte On-Premises Lösungen den heutigen Ansprüchen des modernen Datenmanagements nicht mehr gewachsen sind.

Das Kernziel von Datenmanagement muss darin bestehen, Unternehmen erfolgreicher zu machen. Es geht heutzutage nicht mehr darum regelmäßig die Vergangenheit in Berichten abzubilden, sondern man benötigt Einblicke in Echtzeit. Es geht nicht mehr darum Kennzahlen miteinander abzugleichen, sondern darum, neues Wissen und Prognosen für die Zukunft zu erlangen. Kunden wollen nicht mehr sortiert nach Gruppen angesprochen werden, sondern sie erwarten möglichst individualisierte und passende Angebote. Es sind nicht mehr nur vereinzelte Spezialisten, die Mehrwerte aus Daten generieren, sondern unterschiedlichste Fachbereiche werden zur treibenden Kraft.

Um diesen Anforderungen zu entsprechen, hat Amazon Web Services sein breites Portfolio an Einzel-Lösungen unter dem Dach einer modernen Datenplattform gebündelt.

Durch das geschickte Zusammenspiel von Plattformdiensten und Mechanismen, ergibt sich ein extrem zukunftsfähiges Fundament, welches klassische Ansätze, genauso wie State-of-the-Art Lösungen, kosteneffizient abbilden kann.

Im Rahmen dieses Whitepapers soll ein Überblick über die grundlegenden Kernfunktionen und die von „klassischen“ Lösungen abweichenden Zusammenhänge gegeben werden.


ZUGRIFF & ANWENDERSCHNITTSTELLE
Einfachen und sicheren Zugriff ermöglichen
• API Gateway
• Identity & Access Management
• Cognito
DATENAUFNAHME
Daten schnell und sicher in S3 aufnehmen
• Kinesis Firehouse
• Glue ETL
• Snowball
• Database Migration Service
VERARBEITUNG & ANALYSE
Anwendung von prognostischer und präskriptiver Analytics
• QuickSight
• Amazon AI
• EMR
• Redshift
• Elasticsearch
• Athena
• Kinesis
• RDS
ZENTRALER DATENSPEICHER
KATALOGISIEREN & SUCHEN
Zugriff und Metadaten
• AWS Glue Data
• DynamoDB
• Elasticsearch
SICHERHEIT
Datensicherheit herstellen und Identitäten von Benutzern verifizieren
• Security Token Service
• CloudWatch
• CloudTrail
• Key Management Service

Überblick: AWS in der Datenwertschöpfungskette:

Datenaufnahme

AWS bietet eine Vielzahl an innovativen und Use-Case-spezifischen Integrations- und Transport-Möglichkeiten für Daten.

Datenspeicherung

Datensilos können zusammengeführt werden und durch automatisierte Katalog-Dienste verwaltet und zugänglich gemacht werden. Eine breite Palette an Werkzeugen ermöglicht eine situative und angepasste Transformation zur Weiterverarbeitung.

Datenanalyse

Eine der wichtigsten Funktionen, eines auf AWS basierenden Data Lakes, ist die Möglichkeit, Datenbestände direkt zu transformieren und abzufragen, ohne Cluster bereitstellen und verwalten zu müssen. Auf diese Weise können anspruchsvolle analytische Abfragen direkt auf Datenbeständen ausgeführt werden, ohne Daten kopieren und in separate Analyseplattformen oder Data Warehouses laden zu müssen. Bestehende Analysesysteme können weiter genutzt werden und durch innovative Plattformdienste ergänzt werden.

Data Governance

Dank der API-basierten Datenspeicherung und Plattformsteuerung bleiben keine manuellen oder maschinellen Interaktionen mehr ungesehen. Kombiniert mit im Kern der Plattform verbauten Identity-, Logging- und Compliance-Lösungen, können Datenflüsse und Zugriffe feingranular gesteuert und protokolliert werden. Darüber hinaus können KI-Dienste aktiv und passiv bei der Einhaltung von Vorgaben unterstützen.

Datenschutz

AWS schützt die Cloud an sich. Der Nutzer behält die 100%ige Hoheit über seine Daten und ist dafür verantwortlich, die durch den Cloudanbieter zur Verfügung gestellten Mechanismen einzurichten und durchzusetzen. Das Zusammenspiel aus Plattform-Zertifizierung, verfügbaren Verschlüsselungstechnologien und der API-basierten Governance ermöglichen den Aufbau von „Security by Design“-Lösungen.

Amazon S3 – das Herzstück der modernen Datenplattform auf AWS

Amazon Simple Storage Service (S3) agiert als Fundament jeglicher auf AWS-Technologien basierender Data Lake-Lösungen. S3 stellt einen unlimitierten Datenspeicher zur Verfügung, welcher seine Leistungsfähigkeit anpassen kann und dank mehrfacher automatischer Replikation innerhalb einer Region (z.B. Frankfurt) für eine Haltbarkeit von 99,99999999999% konzipiert wurde. Seine nativen Verschlüsselungs- und Kontroll-Funktionen ermöglichen es, beliebige Dateiformate zentral, sicher und kosteneffizient zu speichern.

Weitere fundamentale Eigenschaften für eine moderne Datenplattform sind:

Zentralisierte Datenarchitektur
AWS S3 steuert Datenzugriff auf Basis von Mandanten. Dies ermöglicht es, verschiedenen Benutzern einer Organisation mit unterschiedlichen Analyse Tools, auf denselben Datensatz zuzugreifen. Somit werden Speicher- und Verwaltungskosten gesenkt, da seltener zusätzliche Kopien von Daten benötigt werden.

Entkoppeln von Speicher und Rechenleistung
Traditionelle Hadoop- und Datawarehouse-Lösungen koppeln Speicher und Rechenleistung relational aneinander. S3 ermöglicht es, Rechenleistung (EC2) unabhängig von der zu analysierenden Datenmenge, dem realen Bedarf entsprechend, zu allokieren. Dies führt zu einer erhöhten Effizienz und ermöglicht sehr kosteneffiziente Architekturen.

Integration mit serverlosen AWS-Diensten
PaaS (Plattform-as-a-Service) Dienste wie Amazon Athena, Amazon Redshift Spectrum, Amazon Recognition und AWS Glue können automatisiert und ohne zusätzliche Aufwände direkt Abfragen auf S3 durchführen. In Kombination mit AWS Lambda (Serverless-Funktionen) können somit Automatismen gestartet werden, welche komplette Arbeitsketten abbilden können. Da die Plattformdienste nach Nutzung abgerechnet werden, ermöglicht dieses Zusammenspiel von Diensten nicht nur signifikante Produktivitätssteigerung, sondern auch ein schnelles und risikofreies Experimentieren.

Standardisierter API-Zugriff
Die AWS S3 RESTful API gilt als de facto Industriestandard und ermöglicht es Organisation somit in den meisten Fällen bestehende Werkzeuge weiter zu nutzen und externe Lösungen zu integrieren, ohne Anpassung vornehmen zu müssen.

Datenaufnahme

Eine der Kernfunktionen einer Data Lake-Architektur ist die Fähigkeit, schnell und einfach jegliche Art von Datentypen zu erfassen. Echtzeit-Streaming-Daten und Massendatenbestände von lokalen Speicherplattformen müssen ständig mit legacy-generierten Datenbeständen aus lokalen Plattformen wie Mainframes oder On-Premises Data Warehouses verbunden werden. AWS bietet Services und Funktionen für alle diese Szenarien.

AWS Kinesis Firehose
Wie im vorigen Whitepaper beschrieben, scheitern viele bestehende Lösungen daran, moderne und unstrukturierte Daten in Echtzeit aufzunehmen. AWS bietet mit Amazon Kinesis Firehose einen komplett gemanagten PaaS-Dienst, welcher Echtzeitdatenströme direkt an S3 sendet. Kinesis Firehose skaliert automatisch bedarfsgerecht und benötigt keine weitere Administration. Darüber hinaus kann Kinesis Firehose Daten schon während des Streaming Prozesses beispielsweise komprimieren, verschlüsseln oder manipulieren. Es ist somit möglich komplexe Datenzulieferprozesse und Verarbeitungen zu automatisieren.

AWS Snowball
Um bestehende Daten aus On-Premises Plattformen oder Hadoop Clustern in der Cloud zusammenzuführen, bietet AWS mit AWS Snowball eine Möglichkeit große Datenmengen per Logistik zu versenden. Größte Datenmengen können damit schnell, kosteneffizient und verschlüsselt ausgetauscht und transportiert werden.

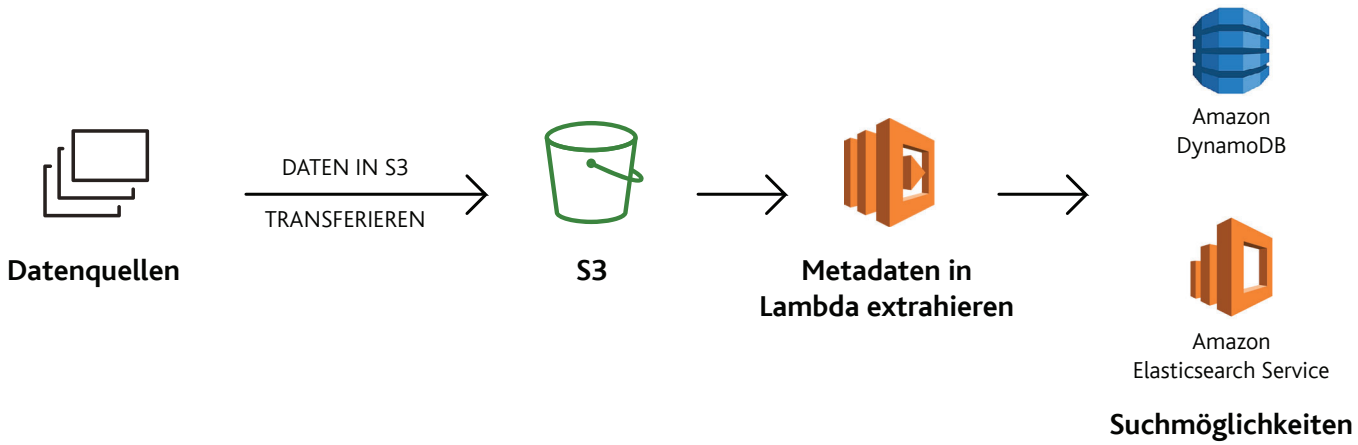
AWS Storage Gateway
Mit AWS Storage Gateway kann ein S3 Data Lake mit einer Legacy Datenplattform verbunden werden. File Gateway verbindet ein klassisches On-Premises Netzwerk Laufwerk mit dem S3 Data Lake über eine verschlüsselte Leitung und ermöglicht es so, auch Anwendungen und Plattformen, welche nativ nicht für den Cloud-Betrieb vorgesehen sind, zu integrieren.

Datenspeicherung

In den vorangestellten Abschnitten wurde bereits dargelegt, dass S3 es ermöglicht, alle Unternehmensdaten in Ihren jeweiligen Rohformen aufzunehmen und diverse Dienste eine konstante Datenaufnahme sicherstellen können.
Eine der größten Herausforderung jeder Data Lake-Architektur besteht jedoch darin, die Übersicht zu behalten. Daten werden ständig zur Weiterverarbeitung verändert, dupliziert und manipuliert, weshalb ein essentieller Bestandteil jeder zentralen Speicherlösung ein Datenkatalog als sogenannte Single-Source-of-Truth darstellt. AWS bietet hier zwei Lösungsansätze die abhängig von den Anforderungen coexistieren können. Diverse Plattformdienste ermöglichen es, einen umfassenden Datenkatalog auf Basis nicht-relationaler Datenbanktechnologien zu erstellen, und ein Hive Metastore Catalog (HCatalog) macht Transformationsverläufe für weitere Analysewerkzeuge zugänglich:

Umfassender Datenkatalog

Die Erstellung eines Umfassenden Datenkatalogs ist ein gutes Beispiel für die Reife der AWS als Plattform. Durch die Kombination mehrerer individuell skalierender Dienste kann ein kosteneffizienter höchst performanter Katalog erstellt werden. Das Konzept sieht wie folgt aus: Immer wenn Daten bei S3 ankommen, wird eine Serverless Compute Funktion (Lambda) ausgelöst, welche Objekt- und Metadaten in eine gemanagte nicht-relationale-Datenbank (Amazon Dynamo DB) schreibt. Ein, ebenfalls gemanagter Elasticsearch Service (Amazon ES) kann anschließend für hoch performante Suchen eingesetzt werden.



Catalog mit AWS Glue

AWS Glue kann einen Apache Hive kompatiblen Metastore Catalog erstellen, welcher für spätere Analysetätigkeiten wichtig wird und Analyseprojekte signifikant erleichtern kann. AWS Glue kann von Haus aus mit weit verbreiteten Datentypen, wie z.B. JSON, CSV und Parquet umgehen und lässt sich um weitere Klassifizierungsfunktionalitäten erweitern. Der generierte Katalog kann von jedem Standard Hive Metastore kompatiblen Analyze Tool genutzt werden und ist kompatibel mit den vielfältigen Werkzeugen welche nächsten Kapitel detaillierter beleuchtet werden.

ETL (Extract, Transform Load)

Sind die Rohdaten gespeichert und katalogisiert, müssen sie auch heute noch oft in verarbeitbare Formate umgewandelt werden, damit sie ihr volles Potenzial entfalten können und von möglich vielen unterschiedlichen Endanwendern im Unternehmen konsumiert und weiterverarbeitet werden können. AWS empfiehlt hier Parquet als demokratisierendes Format, da es u.a. effiziente ad-hoc SQL Abfragen ermöglicht. Zur Umwandlung stellt AWS wieder eine Vielzahl an Möglichkeiten zur Verfügung. Ob sich eine Organisation dazu entscheidet bedarfsgerechte Compute Cluster (Amazon EMR) zur Umwandlung zu nutzen oder den ETL Prozess mit AWS Glue zu automatisieren, ist meist eher eine Frage der vorhandenen Erfahrungen und Präferenzen. In beiden Fällen können die für die Umwandlung der Daten benötigten Ressourcen genau an den Bedarf angepasst werden. Dies ermöglicht es Organisationen Kosten und Geschwindigkeit stets der realen Anforderung entsprechend anzupassen und bestehende ETL Prozesse zu optimieren.

Datenanalyse

Sind die Daten erstmal, wie in den vorherigen Abschnitten beschrieben, im Data Lake angekommen und vorbereitet ist es an der Zeit ihr echtes Potenzial zu entfalten. Grundsätzlich lassen sich in AWS so gut wie alle im Markt verfügbaren Analyse- und Auswertungs-Lösungen einsetzen. Organisationen können weiterhin ihre Daten in Data Warehouse-Lösungen wie Amazon Redshift laden und bestehende Prozesse somit zusätzlich weiter nutzen. Da alle Lösungen on-demand zur Verfügung gestellt werden und Ihre Leistungsfähigkeit nur noch durch den Preis limitiert wird, kann jedoch sehr viel schneller und situativer vorgegangen werden. Organisationen müssen sich nicht mehr um begrenzte On-Premises Ressourcen streiten, was allen datengetriebenen Initiativen eines Unternehmens zu Gute kommt.

Amazon Redshift

Amazon Redshift ist, laut AWS, das zur Zeit schnellste und günstigste Data Warehouse in der Cloud. Es ist tief verknüpft mit einem AWS Data Lake, was Lade- und Speicherprozesse erleichtert. Da Redshift Cluster on-demand zur Verfügung stehen und die Leistungsfähigkeit in den meisten Szenarien nur noch durch den Preis limitiert ist, müssen sich Organisationen nicht mehr um intern begrenzte DW-Ressourcen „streiten". Dank der automatisierten Bereitstellung, eingebauter Fehlertoleranz und automatisierter Backups wird aus dem Data Warehouse ein bedarfsgerechtet Werkzeug. Redshift kann auf ungeladene Daten im Data Lake direkt zugreifen und seit neusten können sogar Abfragen auf relationale Datenbanken in RDS (Relational Database Service) durchgeführt werden.

In-Place-Abfragen

Die Fähigkeit, SQL Abfragen direkt auf Rohdaten (strukturiert sowie unstrukturiert) ausführen zu können, ist aus unternehmerischer Sicht der vielleicht wichtigste Vorteil einer auf AWS basierenden Datenplattform. Sie senkt die technologische, fachliche und organisatorische Hürde innerhalb eines Unternehmens immens und ermöglicht somit eine langfristige Transformation hin zum Datengetriebenen Unternehmen. Dies bedeutetet im Detail: Jeder Mitarbeiter mit SQL-Kenntnissen kann anspruchsvolle analytische Abfragen direkt auf die in S3 gespeicherten Datenbestände ausführen. Es ist wenig Verständnis der komplexen Data Lake Architektur von Nöten und zeitaufwändige Lade- oder ETL-Prozesse entfallen. Endanwender (Fachabteilungen, Data Scientists, Analysten, etc.), welche von zeitraubenden und teuren ETL-Prozesse befreit sind, können schneller, eigenständiger und somit kreativer Mehrwerte aus Daten generieren. Da Kosten nur entsprechend der Analysetätigkeiten anfallen, werden Risiken minimiert und Fachabteilungen können eigenverantwortlicher aktiv werden und somit schneller und häufiger Ergebnisse erzielen.

Technologisch werden diese In-Place-Abfragen erneut durch eine Zusammenspiel mehrerer Dienste ermöglicht. AWS Glue bietet, wie in den vorherigen Abschnitten beschrieben, die Datenerkennungs- und ETL-Funktionen, und Amazon Athena und Amazon Redshift Spectrum bieten die direkten Abfrage selbst:

Amazon Athena

Amazon Athena ist ein interaktiver Abfragedienst, mit dem Daten direkt in Amazon S3 mit Standard-SQL analysiert werden können. Mit ein paar Aktionen in der AWS Management Console können auch Nicht-Experten Athena direkt für Datenbestände verwenden die im Data Lake gespeichert sind und Standard-SQL verwenden, um Ad-hoc-Abfragen auszuführen und innerhalb von Sekunden Ergebnisse zu erhalten. Athena ist serverlos, daher muss keine Infrastruktur eingerichtet oder verwaltet werden, und Unternehmen zahlen nur für das Volumen der Datenbestände, die während der ausgeführten Abfragen gescannt wurden. Athena skaliert automatisch und führt Abfragen parallel aus, sodass die Ergebnisse auch bei großen Datenmengen und komplexen Abfragen schnell sind. Athena kann unstrukturierte, halbstrukturierte und strukturierte Datensätze verarbeiten. Zu den unterstützten Datenbestandsformaten gehören CSV-, JSON- oder spaltenweise Datenformate wie Apache Parquet und Apache ORC. Athena lässt sich zur einfachen Visualisierung in Amazon QuickSight (BI-Tool) integrieren. Es kann auch mit Berichts- und Business Intelligence-Tools von Drittanbietern verwendet werden, indem diese Tools mit einem JDBC-Treiber mit Athena verbunden werden.

Redshift Spectrum

Eine zweite Möglichkeit zur direkten Abfrage ist die Verwendung von Amazon Redshift Spectrum. verwandt mit Amazons gemanagtem Data Warehouse Redshift, ermöglicht Spectum es SQL-Abfragen direkt auf große Datenmengen (bis zu Exabyte) ausführen, die in einem Amazon S3-basierten Data Lake gespeichert sind. Amazon Redshift Spectrum wendet eine ausgefeilte Abfrageoptimierung an und skaliert die Verarbeitung über Tausende von Knoten, sodass die Ergebnisse, auch bei großen Datenmengen und komplex Anfragen schnell sind. Redshift Spectrum kann eine Vielzahl von im Data Lake gespeicherten Datenbeständen direkt abfragen, einschließlich CSV, TSV, Parkett, Sequenz und RCFile. Da Redshift Spectrum die SQL-Syntax von Amazon Redshift unterstützt, können Unternehmen anspruchsvolle Abfragen mit denselben BI-Tools ausführen, welches sie aktuell verwenden. Darüber hinaus ist es möglich, geladene Daten aus dem Redshift Warehouse mit Roh-Daten aus dem Data Lake zu kombinieren was sehr flexible Anwendungsszenarien ermöglicht. Da Amazon Athena und Amazon Redshift einen gemeinsamen Datenkatalog und gemeinsame Datenformate verwenden, können Organisationen sowohl Athena als auch Redshift Spectrum für dieselben Datenbestände verwenden. Normalerweise verwendet man Athena für die Ad-hoc-Datenerkennung und SQL-Abfrage und anschließend Redshift Spectrum für komplexere Abfragen und Szenarien, in denen eine große Anzahl von Data Lake-Benutzern gleichzeitig BI- und Berichts-Workloads ausführen möchten.

Amazon EMR (Hadoop)

Amazon EMR ist ein Compute-Framework, mit dem Daten schnell, einfach und kostengünstig verarbeitet werden können. Amazon EMR verwendet Apache Hadoop, ein Open-Source-Framework, um Daten und Verarbeitung auf einen elastisch veränderbaren Cluster von EC2-Instanzen zu verteilen. Außerdem können alle gängigen Hadoop-Tools wie Hive, Pig, Spark und HBase verwendet werden. Amazon EMR erledigt alle Aufgaben, die mit der Bereitstellung, Verwaltung und Wartung der Infrastruktur und Software eines Hadoop-Clusters verbunden sind und ist direkt in Amazon S3 integriert. Mit Amazon EMR kann ein dauerhafter Cluster gestartet werden, welcher auf unbestimmte Zeit erhalten bleibt oder es kann ein temporärer Cluster gestartet werden, der nach Abschluss der Analyse beendet wird. In beiden Szenarien zahlen Kunden nur für die Stunden, in denen der Cluster aktiv ist.

Amazon Sagemaker und KI-Dienste
Dieser Abschnitt kann stellvertretend für eine ganze Reihe eigenständiger zukünftiger Publikationen gesehen werden.
Um die Komplexität im Rahmen zu halten, soll sich auf die konzeptionelle Ausrichtung von AWS konzentriert werden.
Dass man in der Cloud alle Arten von Maschine Learning Technologien anwenden kann, diese stark von Skalierung profitieren und dass ein zentralisierter Zugang zu Daten, wirtschaftlich sinnvolle Machine Learning-Projekte, wird an diesem Punkt des Paper als allgemein anerkannt vorausgesetzt.

Was AWS im Machine Learning (ML) Kontext wirklich ausmacht, liegt in der Logik von Plattformen und der Art wie AWS Maschine Learning zugänglich macht. Generell fühlen sich ML-Experten auf Grund der Vielzahl an technologischen Möglichkeiten bei AWS sehr wohl, was erklärt, warum nirgends auf der Welt so viele ML-Lösungen in Produktion laufen wie bei AWS. Viel wichtiger für den deutschen Mittelstand ist jedoch, dass AWS mit Lösungen wie Amazon Sagemaker und der stetig wachsenden Zahl an, als API-Call abrufbaren, KI-Diensten, das Thema Maschine Learning in die Hände von Analysten, Data-Scientists und Entwicklern gelegt hat und somit die Einstiegshürde für Unternehmen signifikant gesenkt hat.

Die Plattformlogik, welche auf strategischer Ebene bei der Auswahl eines Cloudanbieters nicht unterschätzt werden darf, ist mit Blick auf die KI-Dienste einfach erklärt: Je mehr Daten ein Algorithmus verarbeitet umso schlauer wird er. Je mehr Menschen z.B. mit Alexa sprechen, umso besser werden die dahinterliegenden Dienste Lex und Polly. Möchte ein Unternehmen seiner Anwendung das Sprechen beibringen, muss ein Entwickler nur noch Polly ansteuern, um automatisch von dieser Entwicklung zu profitieren.

Data-Scientists und Analysten profitieren eher von Amazon Sagemaker.
Dieser Dienst macht Maschine Learning zugänglich, indem er die komplizierte, für „echtes" Maschine Learning aber notwendige, Infrastrukturen extrahiert und bei Bedarf verfügbar macht. Sind genug Daten verfügbar, können Experten somit ohne große Hindernisse sich auf den explorativen Prozess der Mehrwertgewinnung durch Maschine Learning konzentrieren. Da Sagemaker ebenfalls API-fähig ist, können entwickelte Intelligenzen und Lernprozesse automatisiert mit dem Data Lake verbunden werden und kontinuierliche Wissensvorsprünge und Datenwertschöpfungs-Loops generieren.

Data Governance-Vorgaben und Datenschutz

Immer wenn mit Daten gearbeitet wird, müssen strenge und differenzierte Sicherheits- und Zugriffskontrollen sowie Methoden zum Schutz und zur Verwaltung der Datenbestände implementiert werden. Da Data Governance und Datenschutz eng miteinander verknüpft sind, werden beide Aspekte in der Datenwertschöpfungskette in diesem Abschnitt gemeinsam behandelt. Der Fokus liegt hier darauf konzeptionelle Mechanismen zu erklären.

Eine Data Lake-Lösung unter AWS - mit Amazon S3 als Kern - bietet eine Reihe robuster Funktionen und Dienste, mit denen Daten auch in ‚großen Umgebungen mit mehreren Mandanten vor internen und externen Bedrohungen geschützt werden können. Darüber hinaus ermöglichen innovative Amazon S3-Datenverwaltungsfunktionen die Automatisierung und Skalierung des Data Lake-Speichermanagements, selbst wenn es Milliarden von Objekten und Petabyte an Datenbeständen enthält.

Security by Design
Jegliche Interaktion mit der AWS stellt einen API-Call dar. Ob in der Konsole manuelle Einstellungen vorgenommen werden, oder ob Daten durch einzelne Programme gelesen oder geschrieben werden, am Ende läuft alles auf standardisierte Maschinen-Kommunikation hinaus. Diese werden durch AWS CloudTrail, einem gemanagten Logging-Dienst, aufgezeichnet und können in nahezu Echtzeit ausgewertet werden. Diese Transparenz, stellt ein wichtiges konzeptionelles Fundament der Cloud-Security dar.

Zugriffskontrolle und Identitäts-Management
AWS IAM (Identity und Access Management) ist ein weiterer auf alle AWS-Dienste anwendbarer Kernservice, welcher eine Zugriffskontrolle auf Basis von Richtlinien und rollenbasierten Zugriffskontrollen ermöglicht. Da auch Systeme und Dienste eine Identität besitzen und im Standard jegliche Interaktionen durch IAM unterbunden werden (Least-Privilege-Modell), stellt IAM ein weiteres wichtiges Element der Cloud-Security dar. Im Kontext eines Data Lakes ist die Möglichkeit der Vergabe von Benutzer-Rollen erwähnenswert, da diese die spätere kontrollierbare Demokratisierung der Daten immens erleichtert indem sichergestellt wird, dass Systeme und Nutzer nur die Daten im Data Lake sehen, die Ihnen zugeordnet sind.

Objekt Tagging
Da Data Lake-Lösungen von Natur aus mandantenfähig sind und viele Organisationen, Geschäftsbereiche, Benutzer und Anwendungen Datenbestände verwenden und verarbeiten, ist es sehr wichtig, Datenbestände all diesen Entitäten zuzuordnen und Richtlinien für die kohärente Verwaltung dieser Bestände festzulegen. Das Objekt-Tagging wird nicht nur zur Datenklassifizierung verwendet, sondern bietet auch andere wichtige Funktionen. Objekt-Tags können in Verbindung mit IAM verwendet werden, um eine genaue Steuerung der Zugriffsberechtigungen zu ermöglichen. Beispielsweise kann einem bestimmten Data Lake-Benutzer die Berechtigung erteilt werden, nur Objekte mit bestimmten Tags zu lesen.

Datenverschlüsselung
Obwohl Nutzer-Richtlinien und IAM im Fundament kontrollieren, wer Zugriff auf welche Daten in S3 hat, muss trotzdem sichergestellt werden, dass Daten auch im Falle eines unerlaubten oder böswilligen Zugriffs gesichert sind. AWS S3 verschlüsselt Daten standardmäßig serverseitig. Kombiniert man diese Funktionalität mit der zusätzlich angebotenen Kundenseitigen-Verschlüsselung kann sichergestellt werden, dass selbst theoretische Datenzugriffe durch den Cloudanbieter oder Behörden unmöglich gemacht werden. Der AWS Key-Management-Service ermöglicht es Kunden ihre Schlüssel zu verwalten und ist in CloudTrail integriert, was Revisionssicherheit und Auditierbarkeit sicherstellt. Nimmt man noch Dienste wie das AWS API Gateway und den Authentifizierungsdienst Amazon Cognito hinzu, kann der Zugriff einzelner Datenkonsumenten um eine weitere Sicherheitshürde ergänzt werden.

Datensicherheit und Compliance Vorgaben
Wie bereits in vorgestellten Abschnitten erwähnt ist S3 für eine Haltbarkeit von 99,99999999999% konzipiert. Dies bedeutet mathematisch, dass bei 10.000.000 Dateien in 10.000 Jahren keine Datei verloren geht. Um dem Ausfall einer ganzen Region (z.B. Frankfurt) vorzubeugen, können alle Daten mit allen ihren Eigenschaften automatisiert in eine weitere Region (z.B. Paris) repliziert werden. Gegen menschliche Fehler, schützt automatische Versionierung und Langzeitarchivierungen in AWS Glacier (Cold Storage).

Die Struktur von Amazon S3, in Kombination mit anderen AWS-Diensten wie IAM, AWS KMS, Amazon Cognito und Amazon API Gateway stellt somit sicher, dass ein S3 Data Lake die strengsten Anforderungen an Datensicherheit, Compliance, Datenschutz und Verfügbarkeit erfüllt. Amazon S3 umfasst eine breite Palette von Zertifizierungen, darunter PCI-DSS, HIPAA / HITECH, FedRAMP, SEC-Regel 17-a-4, FISMA, EU-Datenschutzrichtlinie und viele weiter andere Zertifizierungen globaler Agenturen. Diese Compliance- und Schutzstufen ermöglichen es Unternehmen, auf AWS einen Data Lake zu erstellen, der sicherer und risikoärmer arbeitet als einer, der in ihren On-Premises-Rechenzentren eingerichtet wurde.

Schlusswort

Wenn Sie bis hier hin gelesen haben, haben Sie bereits einen guten Einblick in die Komplexität auf der einen Seite aber auch die Reife der AWS Modern Data Platform auf der anderen Seite erhalten. Es sollte deutlich geworden sein, dass das Zusammenspiel grundlegender Eigenschaften, wie die API-Steuerung, und die plattformübergreifenden Sicherheits- und Managementdienste zukunftssicheres und agiles Arbeiten mit Daten erst ermöglichen. Und darüber hinaus die Plattform selbst ein Teil der Lösung darstellt.
Wenn Ihr Ziel darin besteht Ihrem Unternehmen ein langfristiges Datenfundament zur Verfügung zu stellen, welches es in die strategische Lage versetzt, auf Veränderungen zu reagieren und Innovation zu fördern während bewährte Prozessketten weiter verfügbar bleiben, empfiehlt es sich, sich tiefergehend mit Amazon Web Services zu beschäftigen.
Damit Sie und Ihre Experten die „AWS-Spache" lernen, ist es wichtig frühzeitig von der Theorie in die Praxis zu wechseln. Erfahrungen aus der Praxis haben gezeigt, dass wenn man die Grundgrammatik erstmal verstanden hat, alles weitere nur noch Vokabeln sind. Unsere Experten aus der AWS Business Group stehen jeder Zeit für vertiefende und erhellende Gespräch zur Verfügung und helfen Ihnen und Ihrer Organisation auch gerne auf Ihren ersten Metern in Richtung eines datengetriebenen Geschäftsmodels.

Schnell ins moderne Datenmanagement einsteigen mit AWS Lake Formation:
AWS Lake Formation automatisiert die Erstellung, Sicherung und Verwaltung eines Data Lakes unter Nutzung der in diesem Whitepaper beschriebenen Dienste und Funktionen.
Unsere Experten können somit innerhalb weniger Tage mit Ihnen einen AWS Data Lake aufsetzten.
Sprechen Sie uns an! aws@bertelsmann.de

AWS Lake Formation | Quelle: Amazon Web Services (2020) | https://aws.amazon.com/de/lake-formation/

ARVATO SYSTEMS

Arvato Systems, Reinhard-Mohn-Straße 18, D-33333 Gütersloh
cloud@bertelsmann.de | arvato-systems.de/cloud