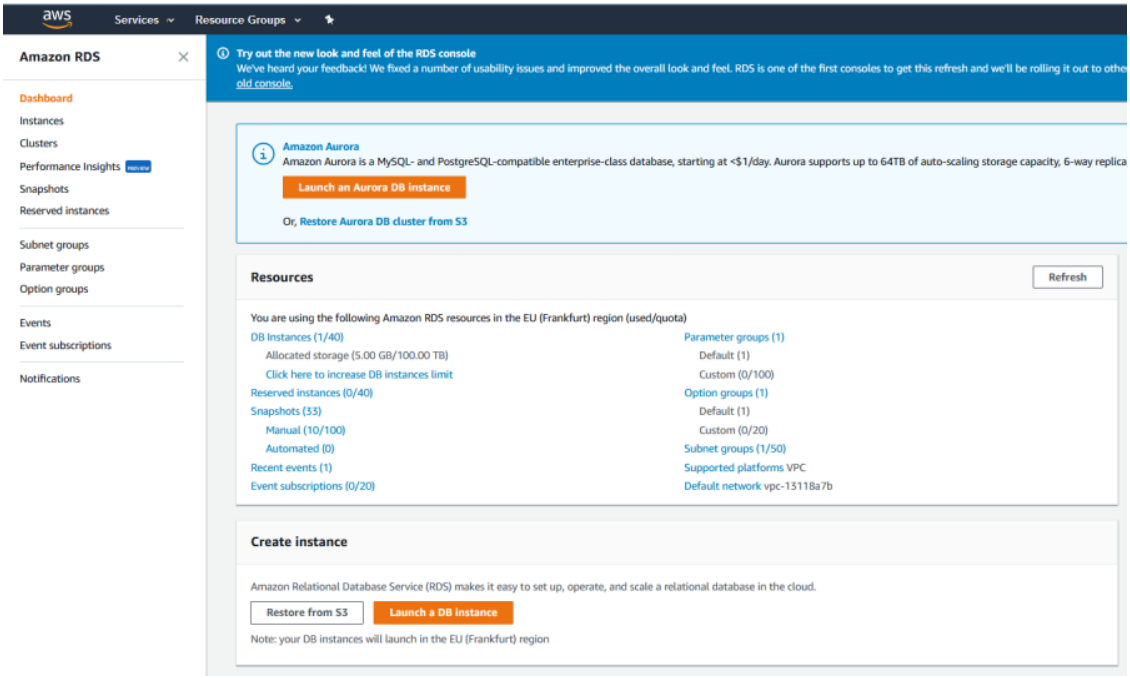
Verwendete Bausteine vom AWS System

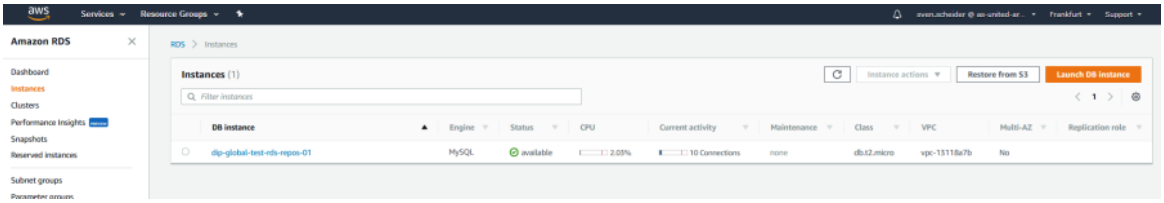
Verwendete Bausteine vom AWS System

Amazon-RDS

Amazon Relational Database Service (Amazon RDS) erleichtert Ihnen das Einrichten, Verwalten und Skalieren einer relationalen Datenbank in der Cloud. Dieser Service stellt kosteneffiziente und anpassbare Kapazitäten zur Verfügung und automatisiert zeitaufwendige Verwaltungsaufgaben wie Hardwarebereitstellung, Datenbankeinrichtung, Einlesen von Patches und Backups. Sie können Sich auf ihre Anwendungen konzentrieren und sich darum kümmern, dass sie die Vorgaben für Performance, Hochverfügbarkeit, Sicherheit und Kompatibilität erfüllen.
Amazon RDS ist für verschiedene Datenbank-Instance-Typen – optimiert für Arbeitsspeicher, Commons oder E/A – verfügbar und bietet Ihnen sechs vertraute Datenbank-Engines zur Auswahl, einschließlich Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle und Microsoft SQL Server. Sie können den AWS Database Migration Service nutzen, um Ihre bestehenden Datenbanken problemlos auf Amazon RDS zu migrieren oder zu replizieren.

Nutzung bei CDP: Zurzeit werden die RDS nur als Metadata Repository für die verschiedenen Hivemeta-Stores und für die Huemeta-Store der Units verwendet.

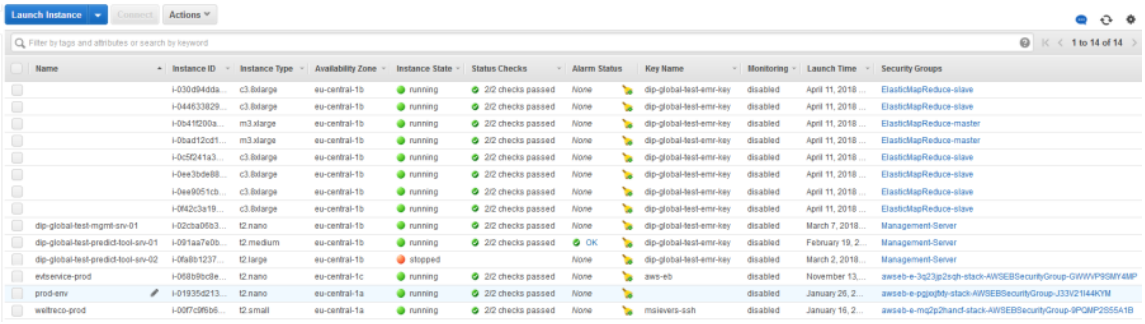




Amazon-EC2

Der Web-Service Amazon Elastic Compute Cloud (Amazon EC2) stellt sichere, skalierbare Rechenkapazitäten in der Cloud bereit. Der Service ist darauf ausgelegt, Cloud Computing für Entwickler zu erleichtern.
Mit der einfachen Web-Service-Oberfläche von Amazon EC2 können Sie mühelos Kapazität erhalten und konfigurieren. Sie ermöglicht Ihnen die vollständige Kontrolle über Ihre Rechenressourcen sowie die Ausführung in der bewährten Rechenumgebung von Amazon. Amazon EC2 verkürzt die zum Buchen und Hochfahren neuer Server-Instances benötigte Zeit auf wenige Minuten. So können Sie die Kapazität entsprechend den Änderungen Ihrer Datenverarbeitungsanforderungen schnell wie gewünscht anpassen. Indem Sie nur für die Kapazität zahlen, die Sie auch tatsächlich nutzen, verändert Amazon EC2 die wirtschaftlichen Rahmenbedingungen von Rechenoperationen. Amazon EC2 bietet Entwicklern die Tools, um ausfallsichere Anwendungen zu erstellen und diese von üblichen Fehlerszenarien zu isolieren.

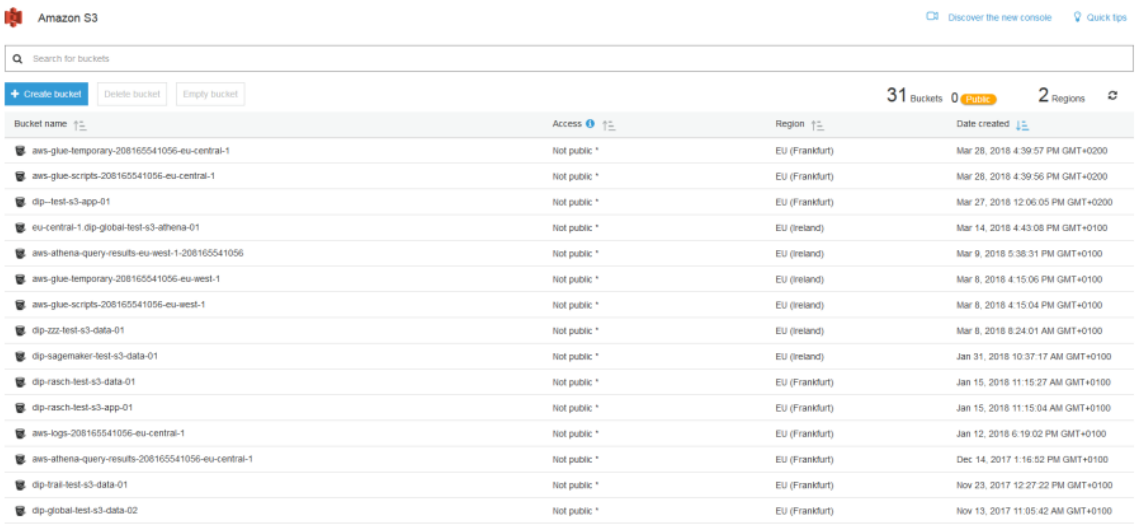
Nutzung bei CDP: Einmal werden die EC2-Instanzen im Rahmen der EMR-Cluster verwendet. Zudem gibt es einen Management-Server über den die ETL-Strecken über eine Crntab gesteuertwerden.



Amazon-S3

Heutzutage müssen Unternehmen ihre Daten in massivem Umfang einfach und sicher erfassen, speichern und analysieren können. Amazon S3 ist ein Objektspeicher zum Speichern und Abrufen beliebiger Datenmengen aus allen Speicherorten: von Websites und mobilen Apps, Unternehmensanwendungen sowie Daten von IoT-Sensoren oder Geräten. Es ist auf eine 99,999999999%ige Haltbarkeit ausgelegt und speichert Daten für Millionen von Anwendungen, die von Marktführern aus allen Branchen verwendet werden. S3 bietet umfassende Sicherheits und Compliance-Funktionen, die selbst die strengsten rechtlichen Anforderungen erfüllen. Es verleiht Kunden Flexibilität bei der Verwaltung von Daten zur Kostenoptimierung, Zugriffssteuerung und Compliance. S3 bietet Funktionen für direkte Abfrage, mit denen Sie leistungsstarke Analysen unmittelbar in Ihren in S3 gespeicherten Daten ausführen können. Und Amazon S3 ist die am häufigsten unterstützte, verfügbare Speicherplattform mit dem größten Ökosystem an ISV-Lösungen und Systemintegratorpartnern.

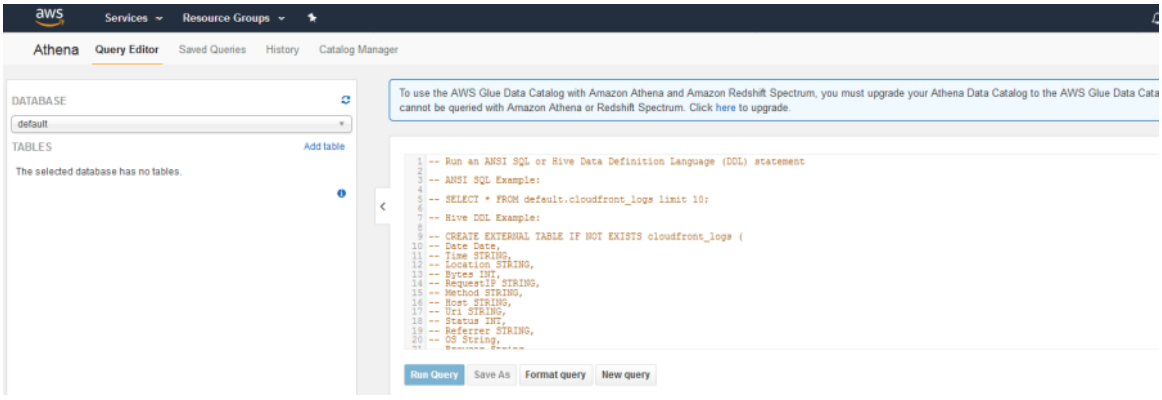
Nutzung bei CDP: Der CDP Data Lake basiert auf S3. Für jede Unit gibt es jeweils einen Bucket, der die Daten enthält ("_data_"), sowie einen Bucket in dem alle Applikations-Skripte verwaltet werden ("_app_"). Bei diesem Bucket ist eine Versionierung der Dateien aktiviert. Zusätzlich sind auf beiden Buckets sogenannte Lifecycle-Rules hinterlegt, welche die Daten nach vorgegebenen Zeiträumen löschen.

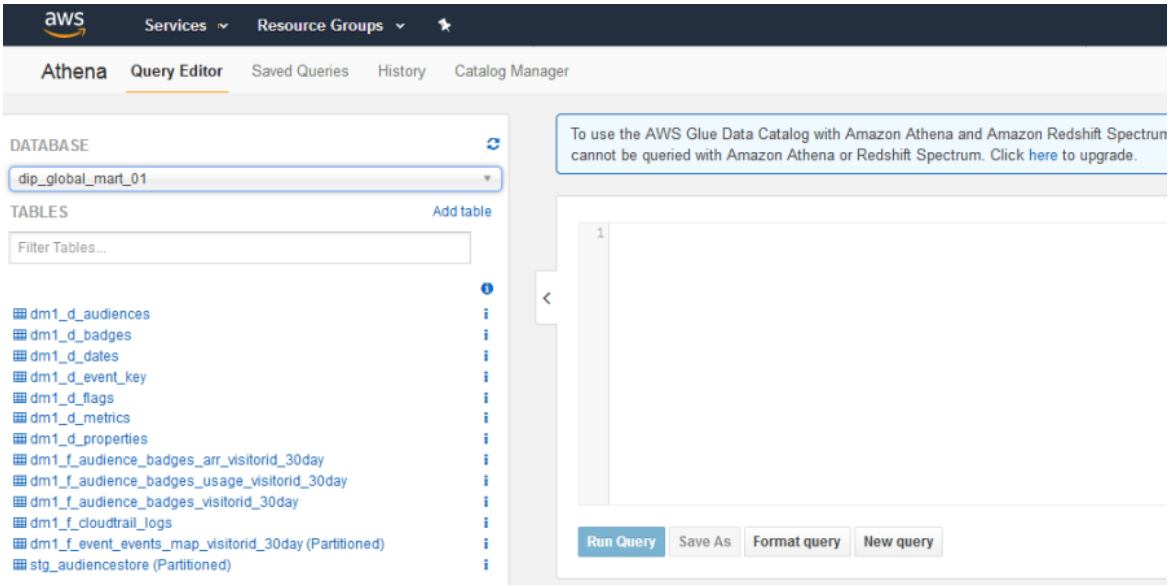


Amazon-Athena

Amazon Athena ist ein interaktiver Abfrageservice, der die Analyse von Daten in Amazon S3 mit Standard-SQL erleichtert. Athena kommt ohne Server aus, deshalb gibt es auch keine Infrastruktur zu verwalten und Sie zahlen nur für die Abfragen, die Sie auch ausführen.
Athena ist benutzerfreundlich. Verweisen Sie einfach auf Ihre Daten in Amazon S3, definieren Sie das Schema und starten Sie die Abfrage mit Standard-SQL. Die meisten Ergebnisse erhalten Sie in Sekundenschnell. Mit Athena sind keine komplexen ETL-Aufträge zur Vorbereitung der Daten für die Analyse erforderlich. Dadurch kann jeder mit SQL-Kenntnissen schnell große Datensätze analysieren.
Athena ist für die Nutzung mit dem AWS Glue-Datenkatalog vorkonfiguriert. Sie können damit ein einheitliches Metadaten-Repository für verschiedene Services erstellen, Datenquellen nach Schemata durchsuchen, den Katalog mit neuen und geänderten Tabellen- und Partitionsdefinitionen füllen und die Schemaversionierung aufrechterhalten. Die vollständig verwalteten ETL-Funktionen von Glue ermöglichen es Ihnen, Daten zu transformieren oder in Spaltenformate zu konvertieren. Sie senken dadurch die Kosten und steigern die Leistung.

Nutzung bei CDP: Wird für die Auswertung des Data Lakes verwendet. Dazu muss ein eigener Metadata-Stores angelegt werden. Zukünftig muss auch der Metadata-Store der EMR-Cluster hier überführt werden. Zurzeit laufen diese noch parallel. Athena kann dann mit verschieden SQL Clients bzw. BI Tools verbunden werden.

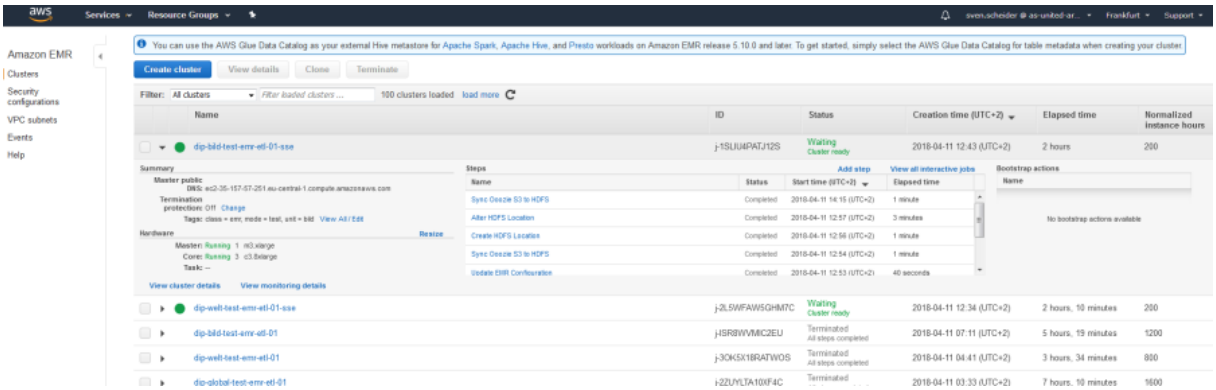




Amazon-EMR

Einfache Ausführung und Skalierung von Apache Hadoop, Spark, HBase, Presto, Hive und anderen Big Data-Frameworks. Amazon EMR bietet ein verwaltetes Hadoop-Framework, mit dem Sie umfangreiche Datenmengen einfach, schnell und kosteneffektiv in dynamisch skalierbaren Amazon EC2 Instances verarbeiten können. Sie können in Amazon EMR auch andere beliebte verteilte Frameworks wie Apache Spark, HBase, Presto und Flink ausführen. Darüber hinaus haben Sie die Möglichkeit, mit Daten in anderen AWS-Datenspeichern wie Amazon S3 und Amazon DynamoDB zu interagieren.
Amazon EMR verarbeitet sicher und zuverlässig eine breite Palette von Big Data-Anwendungsfällen. Hierzu zählen unter anderem Protokollanalysen, Web-Indizierungen, Datentransformationen (ETL), maschinelles Lernen, Finanzanalysen, wissenschaftliche Simulationen und Bioinformatik.

Nutzung bei CDP: Der EMR Cluster dient dazu die Rohdaten mit ETL und Hive zu transformieren, so dass die Daten einfach und schnell auswertbaren Strukturen vorliegen.



AWS-Glue

Einfacher, flexibler und kosteneffektiver ETL-Service.
AWS Glue ist ein vollständig verwalteter ETL-Service (Extrahieren, Transformieren und Laden), der Kunden das Vorbereiten und Laden ihrer Daten zu Analysezwecken erleichtert. Sie können ETL-Aufträge mit wenigen Mausklicks in der AWS Management Console erstellen und ausführen. Indem Sie AWS Glue einfach auf Ihre in AWS gespeicherten Daten verweisen, werden Ihre Daten von der Anwendung erkannt und die zugehörigen Metadaten (z. B. Tabellendefinitionen und Schemata) im AWS Glue-Datenkatalog gespeichert. Sie können Ihre Daten unmittelbar nach der Katalogisierung durchsuchen, abfragen und für ETL nutzen. AWS Glue generiert den Code zum Ausführen von Datentransformationen sowie zum Laden Ihrer Daten.
Sie haben die Möglichkeit, den von AWS Glue generierten Code anzupassen, wiederzuverwenden und zu portieren. Sobald Ihr ETL-Auftrag bereit ist, können Sie dessen Ausführung innerhalb der vollständig verwalteten, skalierten Apache Spark-Umgebung von AWS Glue planen. AWS Glue bietet eine flexibles Planungstool mit Abhängigkeitsauflösung, Auftragsüberwachung und Benachrichtigungsfunktion.
Da AWS Glue serverlos ist, erübrigt sich das Anschaffen, Einrichten und Verwalten von Infrastruktur. AWS Glue stellt automatisch die für den Auftrag erforderliche Umgebung bereit. Kunden zahlen nur für die zum Ausführen von ETL-Aufträgen verbrauchten Rechenressourcen. Mit AWS Glue lassen sich Daten binnen weniger Minuten zu Analysezwecken bereitstellen.

Nutzung bei CDP: Da der DIenst erst seit Sommer 2018 zur Verfügung steht, befindet er sich derzeit in einer Testphase und soll später den EMR Cluster ablösen. Damit hätte man dann sowohl für die Analyse als auch für die ETL Strecken eine Serverless Variante implementiert.