

Verwendete Bausteine vom AWS System

- Verwendete Bausteine vom AWS System
 - Amazon-RDS
 - Amazon-EC2
 - Amazon-S3
 - Amazon-Athena
 - Amazon-EMR
 - AWS-Glue

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.

The screenshot displays the AWS Amazon RDS console. The top navigation bar includes the AWS logo, 'Services', and 'Resource Groups'. The left sidebar shows the 'Amazon RDS' dashboard with links to Dashboard, Instances, Clusters, Performance Insights, Snapshots, Reserved instances, Subnet groups, Parameter groups, Option groups, Events, Event subscriptions, and Notifications.

The main content area features a blue header with a message about the new RDS console look and feel. Below this, there's a section for 'Amazon Aurora' with a 'Launch an Aurora DB instance' button and a link to 'Restore Aurora DB cluster from S3'. The 'Resources' section lists various RDS resources in the EU (Frankfurt) region, including DB Instances (1/40), Allocated storage (5.00 GB/100.00 TB), Reserved instances (0/40), Snapshots (33), Manual (10/100), Automated (0), Recent events (1), Event subscriptions (0/20), Parameter groups (1), Custom (0/100), Option groups (1), Default (1), Custom (0/20), Subnet groups (1/50), Supported platforms VPC, and Default network vpc-13118a7b. A 'Refresh' button is present next to the resources list.

The 'Create instance' section provides information about Amazon RDS and includes buttons for 'Restore from S3' and 'Launch a DB instance'. A note states: 'Note: your DB instances will launch in the EU (Frankfurt) region'.

The bottom section shows the 'Instances (1)' table with a search bar and filters. The table lists one instance: 'db-global-test-rds-repo-01', which is a MySQL instance, currently available, with 2.03% CPU usage, 10 connections, and is running on the db.t2.micro class in the vpc-13118a7b VPC. The table also shows columns for Engine, Status, CPU, Current activity, Maintenance, Class, VPC, Multi-AZ, and Replication role.

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 gesteuert werden.

Launch Instance									
Filter by tags and attributes or search by keyword									
Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Key Name	Monitoring	Launch Time
	i-030d940da	c3.large	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
	i-044833629	c3.large	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
	i-0b410200a	m3.xlarge	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
	i-0ba012cd1	m3.xlarge	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
	i-0c50241a3	c3.large	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
	i-0ee3b0e88	c3.large	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
	i-0ee0051cb	c3.large	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
	i-042c3a18	c3.large	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	April 11, 2018
dp-global-test-mgmt-en-01	i-02ca03a18	t2.nano	eu-central-1b	running	2/2 checks passed	None	dp-global-test-emr-key	disabled	March 7, 2018
dp-global-test-predict-tool-en-01	i-091aa7a0b	t2.medium	eu-central-1b	running	2/2 checks passed	OK	dp-global-test-emr-key	disabled	February 19, 2018
dp-global-test-predict-tool-en-02	i-05ab1237	t2.large	eu-central-1b	stopped	2/2 checks passed	None	dp-global-test-emr-key	disabled	March 2, 2018
edvservice-prod	i-068080c8e	t2.nano	eu-central-1c	running	2/2 checks passed	None	aws-eb	disabled	November 13, 2017
prod-env	i-019350213	t2.nano	eu-central-1a	running	2/2 checks passed	None	aws-eb	disabled	January 26, 2018
weltrco-prod	i-0071d906	t2.small	eu-central-1a	running	2/2 checks passed	None	mailevers-ssh	disabled	January 16, 2018

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.

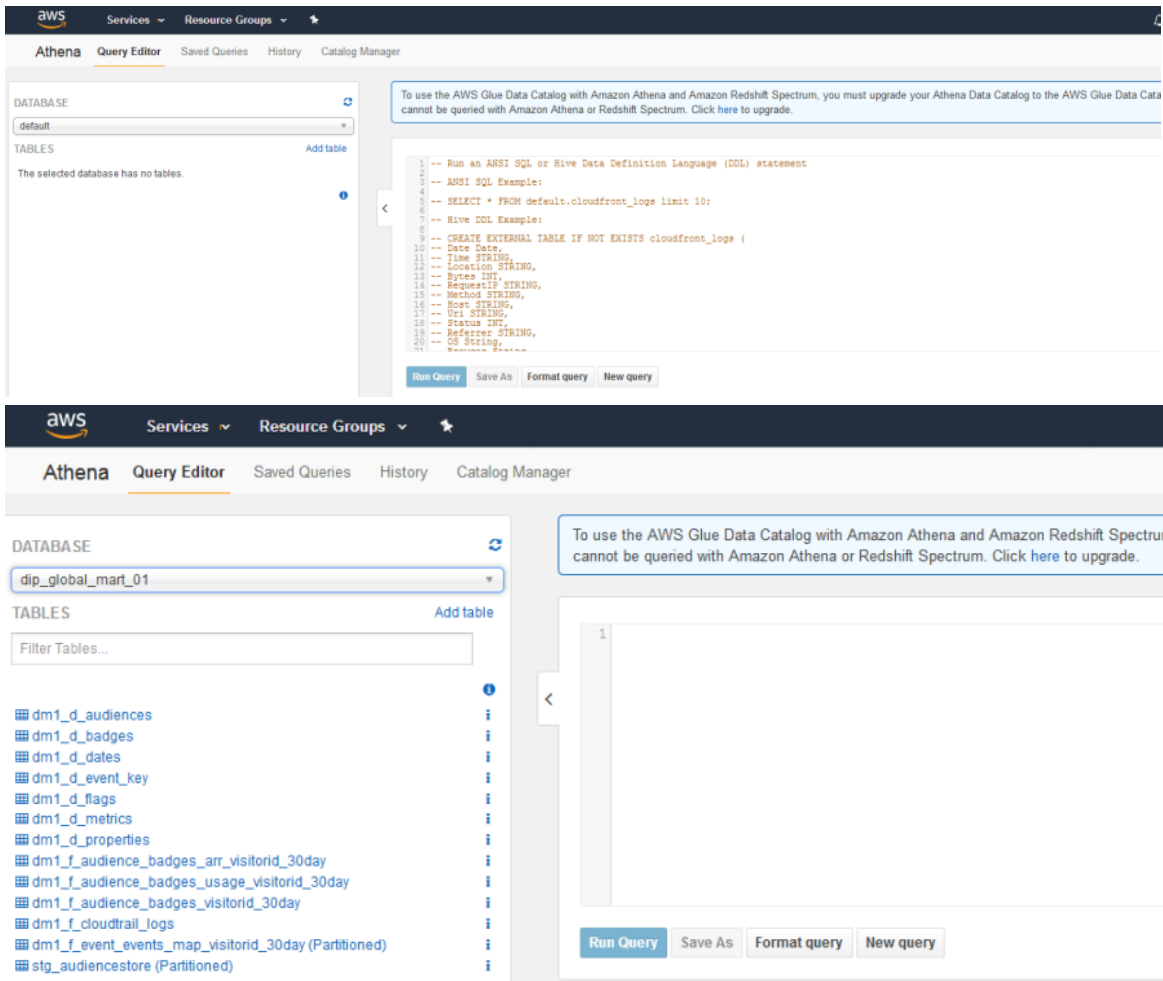
Nutzung bei CDP: Der CDP Data Lake basiert auf S3. Für jede Unit gibt es jeweils einen Bucket, der die Daten enthält ("_data_"), sowie einen Bucket in dem alle Applikations-Skripte verwaltet werden ("_app_"). Bei diesem Bucket ist eine Versionierung der Dateien aktiviert. Zusätzlich sind auf beiden Buckets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

Amazon S3				Discover the new console	Quick tips
Search for buckets					
+ Create bucket Delete bucket Empty bucket				31 Buckets	0 Quotas
				2 Regions	
Bucket name	Access	Region	Date created		
aws-glue-temporary-208165541056-eu-central-1	Not public *	EU (Frankfurt)	Mar 28, 2018 4:39:57 PM GMT+0200		
aws-glue-scripts-208165541056-eu-central-1	Not public *	EU (Frankfurt)	Mar 28, 2018 4:39:56 PM GMT+0200		
dp-test-s3-app-01	Not public *	EU (Frankfurt)	Mar 27, 2018 12:06:05 PM GMT+0200		
eu-central-1-dp-global-test-s3-athena-01	Not public *	EU (Ireland)	Mar 14, 2018 4:43:08 PM GMT+0100		
aws-athena-query-results-eu-west-1-208165541056	Not public *	EU (Ireland)	Mar 9, 2018 5:38:31 PM GMT+0100		
aws-glue-temporary-208165541056-eu-west-1	Not public *	EU (Ireland)	Mar 8, 2018 4:15:06 PM GMT+0100		
aws-glue-scripts-208165541056-eu-west-1	Not public *	EU (Ireland)	Mar 8, 2018 4:15:04 PM GMT+0100		
dp-zzz-test-s3-data-01	Not public *	EU (Ireland)	Mar 8, 2018 8:24:01 AM GMT+0100		
dp-sagemaker-test-s3-data-01	Not public *	EU (Ireland)	Jan 31, 2018 10:37:17 AM GMT+0100		
dp-rasch-test-s3-data-01	Not public *	EU (Frankfurt)	Jan 15, 2018 11:15:27 AM GMT+0100		
dp-rasch-test-s3-app-01	Not public *	EU (Frankfurt)	Jan 15, 2018 11:15:04 AM GMT+0100		
aws-logs-208165541056-eu-central-1	Not public *	EU (Frankfurt)	Jan 12, 2018 6:19:02 PM GMT+0100		
aws-athena-query-results-208165541056-eu-central-1	Not public *	EU (Frankfurt)	Dec 14, 2017 1:16:52 PM GMT+0100		
dp-trai-test-s3-data-01	Not public *	EU (Frankfurt)	Nov 23, 2017 12:27:22 PM GMT+0100		
dp-global-test-s3-data-02	Not public *	EU (Frankfurt)	Nov 13, 2017 11:05:42 AM GMT+0100		

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 Sekundenbruchteilen. 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-Store angelegt werden. Zukünftig muss auch der Metadata-Store der EMR-Cluster hier überführt werden. Zurzeit laufen diese noch parallel. Athena kann dann mit verschiedenen 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 Services ▾ Resource Groups ▾ aws-scheduler @ ap-united-ar... Frankfurt Support

Amazon EMR ←

- Clusters
- Security configurations
- VPC subnets
- Events
- Help

● You can use the AWS Glue Data Catalog as your external Hive metastore for Apache Spark, Apache Hive, and Presto workloads on Amazon EMR release 5.10.0 and later. To get started, simply select the AWS Glue Data Catalog for table metadata when creating your cluster.

Create cluster
View details
Close
Terminate

Filter:

All clusters

▾ Filter loaded clusters ...
100 clusters loaded
load more

Name	ID	Status	Creation time (UTC+2)	Elapsed time	Normalized instance hours
<input type="checkbox"/> ● dip-bdd-test-emr-etl-01-us-east-1	j-LSLUPKATJ2S	Waiting <small>Cluster ready</small>	2018-04-11 12:43 (UTC+2)	2 hours	200

Summary

Master public: DNS: ec2-35-157-57-251.eu-central-1.compute.amazonaws.com

Termination protection: Off Change

Tags: class = ext; mode = test; env = bdd View All/Edit

Hardware [Resize](#)

Master: Running 1 x c3.xlarge

Core: Running 3 x c3.xlarge

Task: --

[View cluster details](#) [View monitoring details](#)

Steps

Name	Status	Start time (UTC+2)	Elapsed time	Bootstrap actions
Sysnc Ozone S3 to HDFS	Completed	2018-04-11 14:15 (UTC+2)	1 minute	No bootstrap actions available
Alter HDFS Location	Completed	2018-04-11 12:57 (UTC+2)	3 minutes	
Create HDFS Location	Completed	2018-04-11 12:58 (UTC+2)	1 minute	
Sysnc Ozone S3 to HDFS	Completed	2018-04-11 12:54 (UTC+2)	1 minute	
Update EMR Configuration	Completed	2018-04-11 12:53 (UTC+2)	40 seconds	

<input type="checkbox"/>	● dip-web-test-emr-etl-01-us-east-1	j-LSHFAWGSQHMTC	Waiting <small>Cluster ready</small>	2018-04-11 12:34 (UTC+2)	2 hours, 10 minutes	200
<input type="checkbox"/>	dip-bdd-test-emr-etl-01	j-HSRBWMCZEU	Terminated <small>All steps completed</small>	2018-04-11 07:11 (UTC+2)	5 hours, 19 minutes	1200
<input type="checkbox"/>	dip-web-test-emr-etl-01	j-3OKSXIBRWIWS	Terminated <small>All steps completed</small>	2018-04-11 04:41 (UTC+2)	3 hours, 34 minutes	800
<input type="checkbox"/>	dip-global-test-emr-etl-01	i-ZU0YLYTA10MFC	Terminated	2018-04-11 03:33 (UTC+2)	7 hours, 10 minutes	1600