

Reference System Architecture for a Machine Learning powered Dashboard

Themenbereiche:	Business Process Modelling, Artificial intelligence & machine learning, Data science
Studierender:	Livio Weber
Betreuungsperson:	Dr. Ossmane Krini
Experte:	Dr. Jeremy Callner
Auftraggebender:	Roche Diagnostics International AG
Keywords:	AWS, CDK, Cloud Computing

1. Aufgabenstellung

Um neue Testmethoden zu realisieren, werden Chemikalien entwickelt, die zu einem bestimmten Virus passen. Dieser Entwicklungsprozess ist forschungsintensiv und erfordert mehrere Zyklen. Das Gesamtprojekt der Arbeit zielt darauf ab, diesen Prozess zu verbessern. Zur Unterstützung dieser Modifikation soll, ein besserer Einblick in die gesammelten Daten gewährt werden.

Die Zielsetzung der vorliegenden Diplomarbeit liegt in der Entwicklung einer Referenz System Architektur für künftige Machine Learning Applikationen. Die implementierten Ansätze sollen zukünftig als Basis sowohl für die Wahl der Architektur sowie auch als Grundlage für neue Anwendungen dienen. Der Fokus der vorliegenden Arbeit liegt im Designen der Architektur. Der Prozessaufbau und getroffene Entscheidungen werden nachvollziehbar aufgezeigt und begründet.

2. Lösungskonzept

Im Rahmen der vorliegenden Bachelordiplomarbeit wird eine Referenz System Architektur designt. Dabei werden verschiedene Architekturkonzepte aufgezeigt und verglichen. Das final erarbeitete Konzept wird als Proof-of-Concept Implementierung realisiert. Die Architektur wird mit Amazon Web Services (AWS) umgesetzt. Konkret wird dabei mit Python und AWS Cloud Development Kit (CDK) gearbeitet. Die Architektur umfasst die entsprechende Quelldatenbank mit den zu berechnenden Daten bis hin zum Dashboard, in welchem die Resultate angezeigt werden.

3. Spezielle Herausforderungen

Vor Beginn der Arbeiten gab es aufseiten des Autors so gut wie kein Know-how in den Bereichen AWS, CDK, Docker und Cloud Computing. Daher war vor allem die Implementierung anspruchsvoll. Zudem erwies es sich als Herausforderung, für spezifische Probleme die geeigneten Antworten im Internet zu finden. Viele der AWS-spezifischen Tutorials und Informationen verweisen auf die AWS-Homepage. Diese ist wie ein Assistent aufgebaut und ermöglicht es dem Benutzer, Dienste und Funktionen zu erstellen und zu kombinieren. Da das CDK für verschiedene Sprachen angeboten wird, mussten einige der gefundenen Lösungen erst in Python

übersetzt werden, damit sie funktionieren. Zusätzlich gibt es bei AWS mittlerweile eine grosse Anzahl verschiedener Services. Den richtigen Dienst zu finden war auch deswegen diffizil.

4. Ergebnisse

Im Rahmen der Bachelordiplomarbeit wurde eine Architektur erarbeitet, welche sich von der Source Datenbank bis hin zum visualisierten Resultat im Dashboard erstreckt.

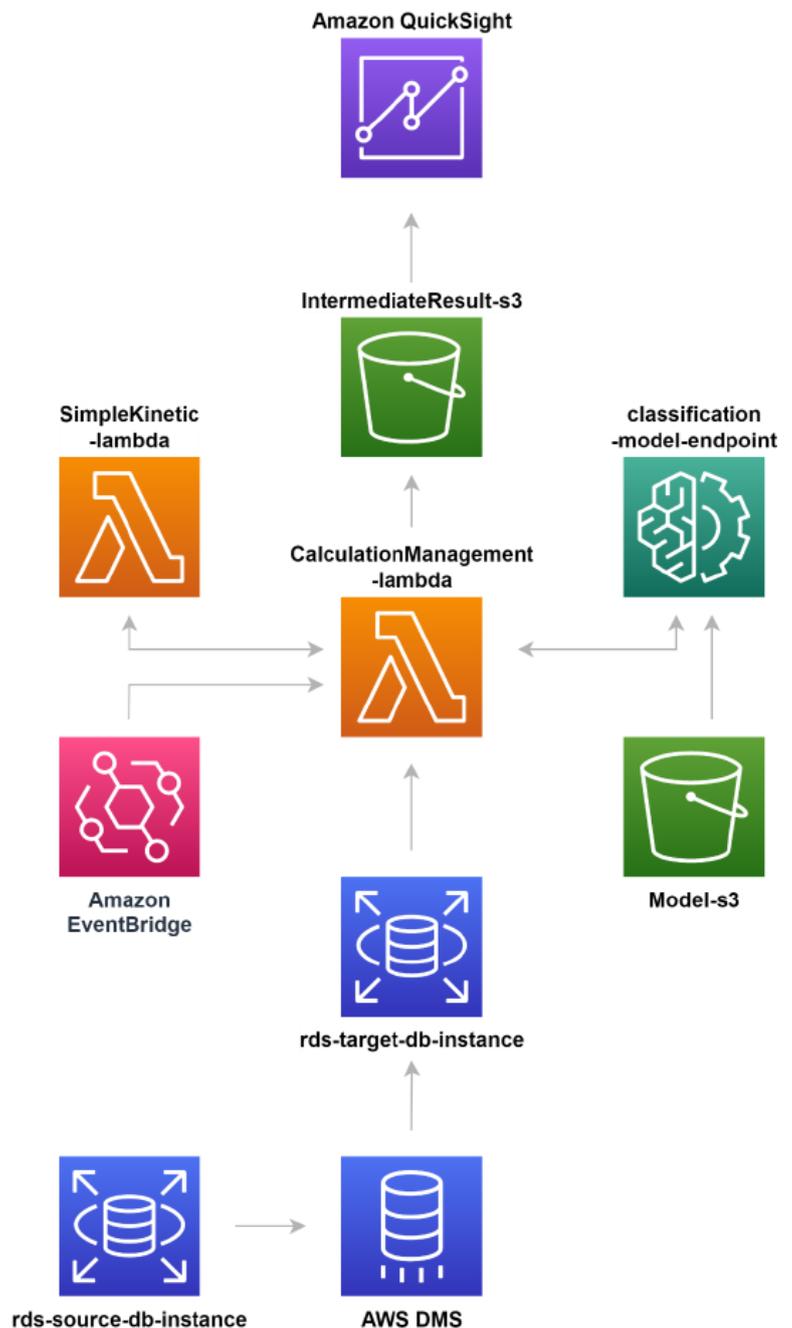
Die Quelldatenbank wird dabei repliziert und es wird jeweils mit den replizierten Daten weitergearbeitet. Eine zentrale AWS Lambda-Funktion managt das Ausführen des konventionellen Algorithmus (SimpleKinetic-lambda) sowie des Machine Learning Modells. (classification-model-endpoint). Diese Berechnung wird in regelmässigen Intervallen durchgeführt, so dass dadurch wiederholt neue Datensatz-Einträge berücksichtigt werden.

Die Berechnungsergebnisse werden auf einem Amazon S3-Bucket zwischengespeichert und durch Amazon QuickSight zur Verfügung gestellt. Mithilfe des Amazons QuickSight-Editors werden beispielhafte Analysen zusammengebaut. Währenddessen bietet der Editor viele verschiedene Diagramm und Filter-Möglichkeiten.

5. Ausblick

Die Referenzarchitektur hat mit der Fertigstellung der Bachelorarbeit einen stabilen Zustand erreicht. Ausführungen zu weiteren Konzepten und Ideen mussten im Rahmen der Arbeit eingegrenzt werden und wurden daher in den Ausführungen nicht im Detail dargestellt oder umgesetzt.

Beispielsweise wird aktuell ein bereits trainiertes maschinelles Lernmodell verwendet. Um das Modell auf dem neuesten Stand zu halten, muss es manuell bereitgestellt und ausgetauscht werden. Bei einer Weiterführung und Überarbeitung des Projekts könnte der Trainings- und Validierungsprozess in die Architektur eingebaut werden.



Des Weiteren wurde im Rahmen der Referenzarchitektur mit einer praktikablen und daher übersichtlichen Menge an Daten gearbeitet. In einer professionellen und produktiven Umgebung wäre das anders. Daher scheint es sinnvoll, die aktuelle Berechnung künftig ebenfalls anzupassen und entsprechend zu überarbeiten. So könnte beispielsweise eine Batch-Berechnung implementiert werden. Der Prozess könnte aber auch so angepasst werden, dass jeweils nur die neuen Datensätze berechnet werden. Durch die aktuelle Referenzarchitektur werden jeweils alle Datensätze an das Machine Learning Modell und an den Algorithmus weitergeleitet. In einer professionellen Umgebung und der entsprechend grösseren Auslastung könnte das System bei zu kleinen Berechnungsintervallen ansonsten an seine Grenzen stossen.

Die Grundlage ist somit geschaffen und diese wird innerhalb der Firma Roche Diagnostics International AG zur Weiterarbeit und Weiterentwicklung bereitgestellt.