

## DATENBLATT

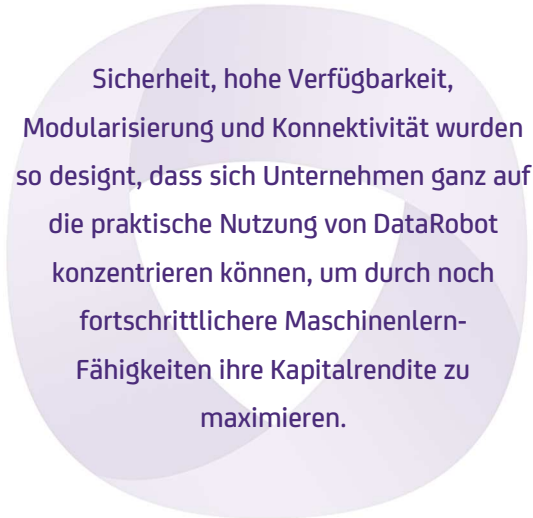
### DataRobot: Architektur

#### Überblick

DataRobot bietet eine führende Plattform für automatisiertes Maschinenlernen, die konsequent für den Einsatz im Unternehmen entwickelt wurde.

Die Plattform umfasst eine Reihe von Features und Funktionen, die dem IT-Team die sichere Bereitstellung von DataRobot in ganz unterschiedlichen Umgebungen ermöglicht.

Sicherheit, hohe Verfügbarkeit, Modularisierung und Konnektivität wurden so design't, dass sich Unternehmen ganz auf die praktische Nutzung von DataRobot konzentrieren können, um durch noch fortschrittlichere Maschinenlern-Fähigkeiten ihre Kapitalrendite zu maximieren.



Sicherheit, hohe Verfügbarkeit, Modularisierung und Konnektivität wurden so design't, dass sich Unternehmen ganz auf die praktische Nutzung von DataRobot konzentrieren können, um durch noch fortschrittlichere Maschinenlern-Fähigkeiten ihre Kapitalrendite zu maximieren.

#### Implementierungsoptionen

DataRobot bietet On-Premise-Unternehmenskunden zwei Bereitstellungsoptionen: als eigenständiges Linux-Deployment oder als Hadoop-Deployment. Via Linux-Deployment kann die Plattform in einer Vielzahl von Locations bereitgestellt werden:

von physischer Hardware über VMware Cluster bis hin zu Virtual-Private-Cloud(VPC)-Anbietern wie Amazon Web Services (AWS), Google Cloud und Microsoft Azure. Hadoop-Deployments ermöglichen die Installation von DataRobot in einem bereits bestehenden Hadoop-Cluster:

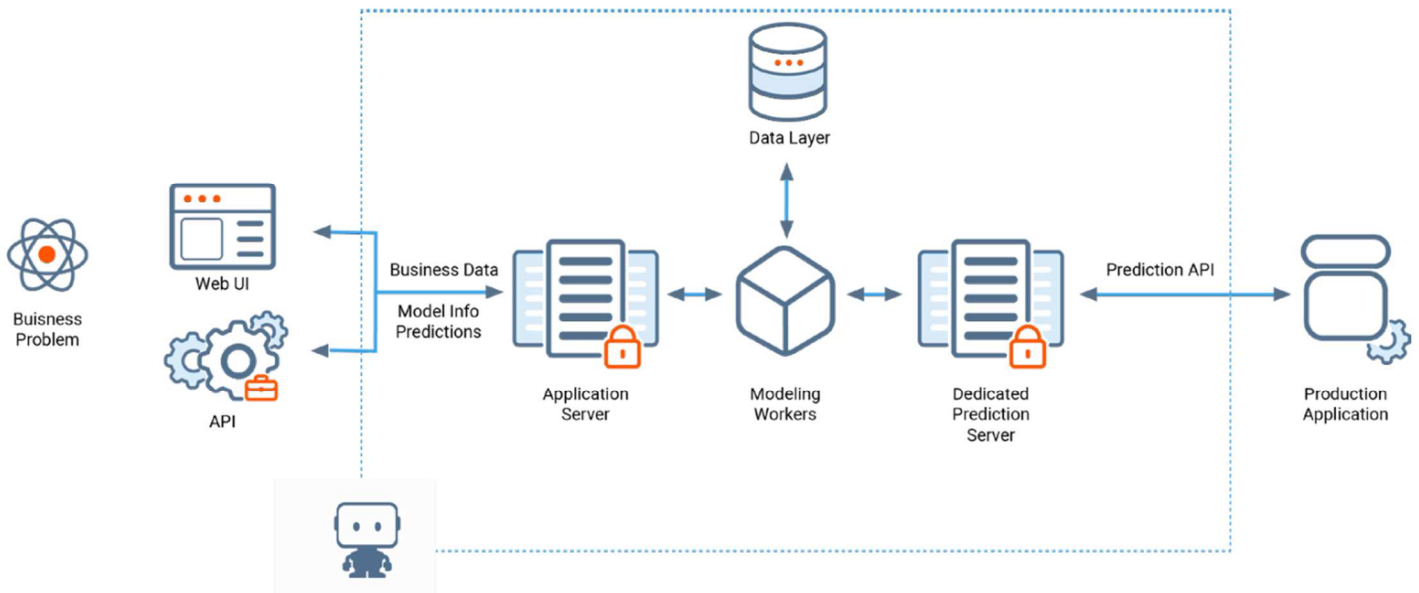
Das spart Hardware-Kosten und vereinfacht die Datenanbindung, die in Hadoop typischerweise schon vorhanden ist.

Der Vergleich dieser beiden On-Premise-Deployment-Optionen zeigt, dass die Features größtenteils identisch sind. Der Unterschied von Hadoop- und Linux-Deployments liegt in der Größe des von DataRobot nutzbaren Datensets:

Hadoop, das sich als Branchenstandard für Big-Data-Plattformen etabliert hat, ermöglicht DataRobot die Aufnahme (Ingest) von bis zu 100 GB Daten für das Training der ML-Modelle. Standalone-Linux-Deployments sind aktuell auf 10 GB begrenzt.

## Schlüsselmodule

Die Schlüsselmodule der DataRobot-Umgebung sind: Anwendungsserver, Datenschicht, Modellierer und Vorhersage-Server. Als konzeptueller Flowchart lässt sich das wie folgt darstellen:



Der Anwendungsserver beinhaltet alle primären Verwaltungskomponenten. Er übernimmt Authentifizierung, Projekt- und Nutzerverwaltung, und bietet einen API-Endpunkt.

Er verwaltet auch die Queue der Modellierungs-Anfragen aus verschiedenen Projekte, die dann von den Modellierern übernommen werden, d. h. einer Rechenressource, die DataRobot-Anwendern das parallele Training von ML-Modellen und ggf. auch die Generierung von Vorhersagen erlaubt.

Die Modellierer sind zustandslos und so konfigurierbar, dass sie je nach Bedarf in die Umgebung integriert oder herausgenommen werden können. So können bei einer VPC-Konfiguration Hardware-Kosten eingespart werden. Innerhalb eines Hadoop-Clusters sind die Modellierer YARN-Container.

Trainierte ML-Modelle werden in die Datenschicht zurückgeschrieben; das Modell-Leaderboard auf dem Anwendungsserver reflektiert ihre Genauigkeit. Trainierte Modelle können auch auf den Prediction-Serverumgebungen bereitgestellt werden.

Dedizierte Vorhersage-Server sind der wichtigste Teil jeder auf Analytics basierenden Unternehmung. Sie ermöglichen Echtzeit-

Dank des dedizierten Vorhersage-Servers können trainierte Modelle innerhalb der DataRobot-Plattform schnell in einem dedizierten System zur Vorhersage-Erstellung bereitgestellt werden.

Entscheidungen: schnell und zuverlässig, ohne Angst vorm Scheitern oder Zuspätkommen.

Zudem werden die wichtigsten Statistiken zu diesen Vorhersagen und die gelieferten Daten zurück an den Anwendungsserver gegeben und den Nutzern angezeigt, um die Gesundheit der Modelle zu überwachen. Der Vorhersage-Server kann auch in einer von der DataRobot-Plattform getrennten Umgebung bereitgestellt werden, was Unternehmen das Deployment von Modellen in segregierten Netzwerken ermöglicht.

## Speicherung und Konnektivität

Ohne Daten gibt es weder Data Science noch Maschinenlernen. DataRobot bietet mehrere Möglichkeiten zur Anbindung von Unternehmensdaten, die zur Erstellung von ML-Modellen genutzt werden sollen. DataRobot liest Daten aus URLs (inkl. Amazon S3 buckets), HDFS (für Hadoop-Installationen), Datenbanken (via JDBC-Driver) und direkt aus hochgeladenen Flat Files (csv, xlsx, tsv, etc.). Alle Daten werden bei der Übertragung verschlüsselt.

Die Datenspeicherung in DataRobot hängt von der gewählten Deployment-Methode ab. Hadoop-Deployments speichern alle relevanten Daten in HDFS.

Bei regulären Linux-Deployments werden die Daten in einem Gluster-Dateisystem gespeichert, das mit Mehrfachkopien repliziert werden kann. Bei Nutzung einer AWS VPC kann DataRobot direkt in einen privaten Amazon S3 bucket schreiben. Im Interesse maximaler Sicherheit können alle in DataRobot gespeicherten Daten verschlüsselt werden.

## Innere / Interne Komponenten

Alle DataRobot-Komponenten sind modular im Design und können problemlos über mehrere Maschinen verteilt werden. Eben dieses Design ermöglicht DataRobot die horizontale Skalierung zur Anpassung an veränderte Geschäftsanforderungen.

Alle Dienste laufen in Docker-Containern. Dadurch können mehrere Instanzen bestimmter Dienste auf mehreren Maschinen laufen, was hohe Verfügbarkeit und Resilienz in Disaster-Recovery-Situationen gewährleistet.

Alternativ kann das Unternehmen alle Prozesse auf einem Server laufen lassen – sowohl bei Umgebungen mit weniger stringenten Verfügbarkeitsanforderungen oder zur Evaluierung der Plattform als Teil einer Machbarkeitsstudie (Proof of Concept).

## Ihr Kontakt:

Heyde AG, Fuchsiastr. 10, 8024 Zürich || [datarobot@heyde.ch](mailto:datarobot@heyde.ch) || +41 44 405 60 50