Erfolgreich planen MLOps

AWS Präskriptive Leitlinien

AWS Präskriptive Leitlinien: Erfolgreich planen MLOps

Table of Contents

Erfolgreich planen MLOps

Bruno Klein, Amazon Web Services (AWS)

Dezember 2021 (Dokumentenverlauf)

Der Einsatz von Lösungen für maschinelles Lernen (ML) in der Produktion bringt viele Herausforderungen mit sich, die bei Standard-Softwareentwicklungsprojekten nicht auftreten. ML-Lösungen sind komplexer und schwieriger, sie von vornherein richtig zu machen. Sie kommen auch in normalerweise volatilen Umgebungen vor, in denen die Datenverteilung im Laufe der Zeit aus einer Vielzahl erwarteter und unerwarteter Gründe erheblich abweicht.

Diese Probleme werden noch dadurch verschärft, dass viele ML-Praktiker keinen Hintergrund in der Softwareentwicklung haben und daher möglicherweise nicht mit den Best Practices dieser Branche vertraut sind, wie dem Schreiben von testbarem Code, der Modularisierung von Komponenten und der effektiven Nutzung der Versionskontrolle. Diese Herausforderungen führen zu technischen Schulden, und die Lösungen werden im Laufe der Zeit immer komplexer und schwieriger zu warten, was auf einen sich verschärfenden Effekt für ML-Teams zurückzuführen ist.

In diesem Leitfaden werden bewährte Methoden für ML-Operationen (MLOps) aufgeführt, die dazu beitragen, diese Herausforderungen bei ML-Projekten und -Workloads zu bewältigen.

Da MLOps es sich um ein bereichsübergreifendes Problem handelt, betreffen diese Probleme nicht nur die Bereitstellungs- und Überwachungsprozesse, sondern den gesamten Modelllebenszyklus. In diesem Leitfaden sind MLOps bewährte Verfahren in vier Hauptbereiche unterteilt:

- Daten
- Training
- Bereitstellung
- Überwachung

Gezielte Geschäftsergebnisse

Der Einsatz von ML-Modellen in der Produktion ist eine Aufgabe, die kontinuierliche Anstrengungen und ein engagiertes Team erfordert, das diese Ressourcen während ihrer gesamten Lebensdauer (in einigen Fällen sogar Jahre) verwaltet. ML-Modelle können einen erheblichen Nutzen aus Geschäftsdaten ziehen, sind jedoch mit hohen Kosten verbunden. Um die Kosten zu

minimieren, sollten Unternehmen bewährte Verfahren in den Bereichen Softwareentwicklung und Datenwissenschaft anwenden. Sie sollten sich der Nuancen von ML-Systemen bewusst sein, z. B. der Datendrift, die dazu führt, dass Modelle nach einer Weile unerwartet funktionieren. Wenn sich Unternehmen dieser Bedenken bewusst sind, können sie ihre kurz- und langfristigen Geschäftsziele sicher und agil erreichen.

Es gibt verschiedene Arten von ML-Modellen, und die Branchen, auf die sie abzielen, haben unterschiedliche Arten von ML-Aufgaben und Geschäftsproblemen, sodass Sie für jedes Modell und jede Branche unterschiedliche Bedenken berücksichtigen müssen. Die in diesem Leitfaden dargelegten Verfahren sind nicht modell- oder geschäftsspezifisch, sondern gelten für ein breites Spektrum von Modellen und Branchen, um die Bereitstellungszeiten zu verbessern, die Produktivität zu steigern und für eine stärkere Unternehmensführung und Sicherheit zu sorgen.

Die Produktion von Modellen ist eine multidisziplinäre Aufgabe, für die Datenwissenschaftler, Ingenieure für maschinelles Lernen, Dateningenieure und Softwareingenieure erforderlich sind. Wir empfehlen Ihnen, beim Aufbau Ihres ML-Teams auf diese Fähigkeiten und Hintergründe zu achten.

Daten

DevOps ist ein Software-Engineering-Praxis, das sich mit der Operationalisierung von Software befasst. Zu den gemeinsamen Elementen DevOps gehören versionskontrollierter Code, CI/CD-Pipelines (Continuous Integration and Continuous Delivery), Komponententests sowie reproduzierbare Codeerstellungen und -bereitstellungen, die alle Code beinhalten. ML-Modelle sind ein Produkt aus Code und Daten, daher müssen Daten dieselben Standards erfüllen wie Code. MLOps muss sich mit datenbezogenen Fragen befassen, z. B. wie die Datenqualität aufrechterhalten werden kann, wie man Grenzfälle in Daten identifiziert, wie man Daten schützt und wie man Daten wartungsfreundlicher macht.

Themen

Labeling

Stellen Sie klare Anweisungen zur Kennzeichnung bereit

Ein Datensatz kann mehrdeutige Stichproben enthalten, die zu einer inkonsistenten Kennzeichnung des gesamten Datensatzes führen. Stellen Sie sich zum Beispiel die Aufgabe vor, Bilder zu kennzeichnen, auf denen ein Hund zu sehen ist. Einige Proben enthalten möglicherweise nur einen flüchtigen Blick auf das Tier. Sollten diese mit einem positiven oder negativen Etikett gekennzeichnet sein? Diese Art von Problem könnte gelöst werden, indem den Etikettierern klare und objektive Anweisungen gegeben werden.

Verwenden Sie Mehrheitsbeschlüsse

Betrachten wir nun das Problem der Kennzeichnung eines speech-to-text Datensatzes, der verrauschte Audiodateien enthält, mit Wörtern, die phonetisch ähnlich oder identisch mit anderen Wörtern sind, z. B. wissen und gehen, Schuh und zwei, Weinen und hoch oder richtig und schreiben. In diesem Fall könnten Labeler diese Samples uneinheitlich beschriften.

Um ein hohes Maß an Korrektheit bei der Kennzeichnung aufrechtzuerhalten, ist ein gängiger Ansatz die Mehrheitsabstimmung, bei der dieselbe Datenstichprobe mehreren Arbeitnehmern gegeben und

ihre Ergebnisse aggregiert werden. Diese Methode und ihre ausgefeilteren Varianten werden im Blogbeitrag Nutze die Weisheit der Massen mit Amazon SageMaker AI Ground Truth, um Daten genauer zu kommentieren, im Blog AWS Machine Learning beschrieben.

Spaltungen und Datenlecks

Datenlecks treten auf, wenn Ihr Modell während der Inferenz — also in dem Moment, in dem das Modell in Produktion ist und Prognoseanfragen empfängt — Daten erhält, auf die es keinen Zugriff haben sollte, z. B. Datenproben, die für das Training verwendet wurden, oder Informationen, die nicht verfügbar sind, wenn das Modell in der Produktion eingesetzt wird.

Wenn Ihr Modell versehentlich anhand von Trainingsdaten getestet wird, kann ein Datenverlust zu einer Überanpassung führen. Eine Überanpassung bedeutet, dass Ihr Modell nicht gut auf unsichtbare Daten übertragen werden kann. In diesem Abschnitt finden Sie bewährte Methoden zur Vermeidung von Datenlecks und Überanpassungen.

Teilen Sie Ihre Daten in mindestens drei Sätze auf

Eine häufige Ursache für Datenlecks ist die unsachgemäße Aufteilung (Aufteilung) Ihrer Daten während des Trainings. Beispielsweise könnte der Datenwissenschaftler das Modell wissentlich oder unwissentlich anhand der Daten trainiert haben, die für Tests verwendet wurden. In solchen Situationen können Sie sehr hohe Erfolgskennzahlen beobachten, die auf eine Überanpassung zurückzuführen sind. Um dieses Problem zu lösen, sollten Sie die Daten in mindestens drei Gruppen aufteilen: `training`, `validation` und `testing`.

Wenn Sie Ihre Daten auf diese Weise aufteilen, können Sie anhand des `validation` Satzes die Parameter auswählen und anpassen, die Sie zur Steuerung des Lernprozesses verwenden (Hyperparameter). Wenn Sie ein gewünschtes Ergebnis erzielt oder ein Verbesserungspotenzial erreicht haben, führen Sie eine Bewertung des Sets durch. `testing` Die Leistungskennzahlen für das `testing` Set sollten den Kennzahlen für die anderen Sets ähneln. Dies deutet darauf hin, dass kein Verteilungsgefälle zwischen den Sets besteht und dass Ihr Modell in der Produktion voraussichtlich gut generalisiert werden kann.

Verwenden Sie einen Algorithmus für die stratifizierte Aufteilung

Wenn Sie Ihre Daten in `training`, `validation` und `testing` für kleine Datensätze aufteilen oder wenn Sie mit stark unausgewogenen Daten arbeiten, stellen Sie sicher, dass Sie einen Algorithmus

für die stratifizierte Aufteilung verwenden. Durch die Stratifizierung wird gewährleistet, dass jede Aufteilung ungefähr die gleiche Anzahl oder Verteilung von Klassen für jede Aufteilung enthält. Die Scikit-Learn ML-Bibliothek implementiert bereits die Stratifizierung, ebenso wie Apache Spark.

Stellen Sie bei der Stichprobengröße sicher, dass die Validierungs- und Testsätze über genügend Daten für die Auswertung verfügen, damit Sie zu statistisch signifikanten Schlussfolgerungen gelangen können. Eine übliche Teilungsgröße für relativ kleine Datensätze (weniger als 1 Million Stichproben) beträgt beispielsweise 70%, 15% und 15% für `training`, `validation` und. `testing` Bei sehr großen Datensätzen (mehr als 1 Million Stichproben) können Sie 90%, 5% und 5% verwenden, um die verfügbaren Trainingsdaten zu maximieren.

In einigen Anwendungsfällen ist es sinnvoll, die Daten in weitere Datensätze aufzuteilen, da die Verteilung der Produktionsdaten während des Zeitraums, in dem sie erfasst wurden, möglicherweise radikale, plötzliche Veränderungen erfahren haben könnte. Stellen Sie sich zum Beispiel einen Datenerfassungsprozess zur Erstellung eines Modells zur Bedarfsprognose für Artikel aus Lebensmittelgeschäften vor. Wenn das Data-Science-Team die `training` Daten im Jahr 2019 und die `testing` Daten von Januar 2020 bis März 2020 sammeln würde, würde ein Modell am `testing` Set wahrscheinlich gut abschneiden. Wenn das Modell jedoch in der Produktion eingesetzt wird, hätte sich das Konsumverhalten bestimmter Artikel aufgrund der COVID-19-Pandemie bereits erheblich verändert, und das Modell würde zu schlechten Ergebnissen führen. In diesem Szenario wäre es sinnvoll, ein weiteres Set (zum Beispiel`recent_testing`) als zusätzliche Schutzmaßnahme für die Modellgenehmigung hinzuzufügen. Diese Ergänzung könnte Sie daran hindern, ein Modell für die Produktion zu genehmigen, das aufgrund von Vertriebsinkongruenzen sofort eine schlechte Leistung erbringen würde.

In einigen Fällen möchten Sie möglicherweise zusätzliche Stichproben `validation` oder `testing` Gruppen erstellen, die bestimmte Arten von Stichproben enthalten, z. B. Daten, die sich auf Bevölkerungsgruppen von Minderheiten beziehen. Es ist wichtig, diese Datenstichproben richtig zu machen, sie sind jedoch möglicherweise nicht gut im Gesamtdatensatz vertreten. Diese Datenteilmengen werden als Slices bezeichnet.

Nehmen wir als Beispiel ein ML-Modell für die Kreditanalyse, das anhand von Daten für ein ganzes Land trainiert und ausgewogen wurde, um den gesamten Bereich der Zielvariablen gleichermaßen zu berücksichtigen. Bedenken Sie außerdem, dass dieses Modell möglicherweise `City` über eine Funktion verfügt. Wenn die Bank, die dieses Modell verwendet, ihr Geschäft auf eine bestimmte Stadt ausdehnt, könnte sie daran interessiert sein, wie das Modell in dieser Region abschneidet. Eine Genehmigungspipeline sollte also nicht nur die Qualität des Modells auf der Grundlage der Testdaten für das gesamte Land bewerten, sondern auch Testdaten für ein bestimmtes Stadtviertel auswerten.

Wenn Datenwissenschaftler an einem neuen Modell arbeiten, können sie die Fähigkeiten des Modells leicht beurteilen und Grenzfälle berücksichtigen, indem sie unterrepräsentierte Bereiche in der Validierungsphase des Modells integrieren.

Ziehen Sie bei zufälligen Aufteilungen doppelte Stichproben in Betracht

Eine weitere, weniger häufige Ursache für Leckagen sind Datensätze, die möglicherweise zu viele doppelte Proben enthalten. In diesem Fall können verschiedene Teilmengen gemeinsame Stichproben haben, auch wenn Sie die Daten in Teilmengen aufteilen. Je nach Anzahl der Duplikate kann eine Überanpassung mit Generalisierung verwechselt werden.

Ziehen Sie Funktionen in Betracht, die beim Empfang von Schlussfolgerungen in der Produktion möglicherweise nicht verfügbar sind

Datenlecks treten auch auf, wenn Modelle mit Funktionen trainiert werden, die in der Produktion nicht verfügbar sind, und zwar in dem Moment, in dem die Schlussfolgerungen gezogen werden. Da Modelle häufig auf der Grundlage historischer Daten erstellt werden, können diese Daten mit zusätzlichen Spalten oder Werten angereichert werden, die zu einem bestimmten Zeitpunkt noch nicht vorhanden waren. Stellen Sie sich das Beispiel eines Kreditgenehmigungsmodells vor, das über eine Funktion verfügt, mit der nachverfolgt werden kann, wie viele Kredite ein Kunde in den letzten sechs Monaten bei der Bank aufgenommen hat. Es besteht die Gefahr von Datenlecks, wenn dieses Modell für die Kreditgenehmigung eines neuen Kunden eingesetzt und verwendet wird, der noch keine sechsmonatige Erfahrung mit der Bank hat.

Amazon SageMaker AI Feature Store hilft bei der Lösung dieses Problems. Mithilfe von Zeitreiseabfragen, mit denen Sie Daten zu bestimmten Zeitpunkten anzeigen können, können Sie Ihre Modelle genauer testen.

Feature-Shop

Die Verwendung von SageMaker AI Feature Store erhöht die Teamproduktivität, da dadurch die Grenzen zwischen Komponenten (z. B. Speicherplatz und Nutzung) entkoppelt werden. Es bietet auch die Wiederverwendbarkeit von Funktionen in verschiedenen Data-Science-Teams innerhalb Ihres Unternehmens.

Verwenden Sie Zeitreise-Abfragen

Die Funktionen für Zeitreisen im Feature Store helfen bei der Reproduktion von Modellerstellungen und unterstützen strengere Governance-Praktiken. Dies kann nützlich sein, wenn eine Organisation

die Datenherkunft bewerten möchte, ähnlich wie Versionskontrolltools wie Git Code bewerten. Zeitreiseabfragen helfen Unternehmen auch dabei, genaue Daten für Konformitätsprüfungen bereitzustellen. Weitere Informationen finden Sie unter Grundlegendes zu den wichtigsten Funktionen von Amazon SageMaker AI Feature Store im AWS Machine Learning Learning-Blog.

Verwenden von IAM-Rollen

Feature Store trägt auch zur Verbesserung der Sicherheit bei, ohne die Produktivität und Innovation des Teams zu beeinträchtigen. Sie können AWS Identity and Access Management (IAM-) Rollen verwenden, um bestimmten Benutzern oder Gruppen den detaillierten Zugriff auf bestimmte Funktionen zu gewähren oder einzuschränken.

Die folgende Richtlinie schränkt beispielsweise den Zugriff auf eine vertrauliche Funktion im Feature Store ein.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Deny",
            "Action": "*",
            "Resource": "arn:aws:s3:::amzn-s3-demo-bucket--usw2-az1--x-s3/12345678910/
sagemaker/us-east-2/offline-store/doctor-appointments"
        }
    ]
}
```

Weitere Informationen zur Datensicherheit und Verschlüsselung mit Feature Store finden Sie in der SageMaker KI-Dokumentation unter Sicherheit und Zugriffskontrolle.

Verwenden Sie Unit-Tests

Wenn Datenwissenschaftler Modelle auf der Grundlage bestimmter Daten erstellen, treffen sie häufig Annahmen über die Verteilung der Daten oder führen eine gründliche Analyse durch, um die Dateneigenschaften vollständig zu verstehen. Wenn diese Modelle eingesetzt werden, sind sie irgendwann veraltet. Wenn der Datensatz veraltet ist, trainieren Datenwissenschaftler, ML-Ingenieure und (in einigen Fällen) automatisierte Systeme das Modell mit neuen Daten, die aus einem Online- oder Offline-Speicher abgerufen werden, neu.

Die Verteilung dieser neuen Daten könnte sich jedoch geändert haben, was sich auf die Leistung des aktuellen Algorithmus auswirken könnte. Eine automatisierte Methode, um nach solchen Problemen zu suchen, besteht darin, das Konzept des Unit-Tests aus der Softwareentwicklung zu übernehmen. Üblicherweise werden anhand eines Frameworks wie Hypothesenteststatistiken (t-test) der Prozentsatz fehlender Werte, die Kardinalität kategorialer Variablen und die Frage, ob Spalten mit reellen Werten einer bestimmten erwarteten Verteilung entsprechen, getestet. Möglicherweise möchten Sie auch das Datenschema überprüfen, um sicherzustellen, dass es sich nicht geändert hat und nicht automatisch ungültige Eingabe-Features generiert.

Unit-Tests setzen voraus, dass Sie die Daten und ihre Domäne verstehen, damit Sie die genauen Assertions planen können, die im Rahmen des ML-Projekts ausgeführt werden sollen. Weitere Informationen finden Sie im Big Data-Blog unter Datenqualität im AWS großen Maßstab testen. PyDeequ

Training

MLOps befasst sich mit der Operationalisierung des ML-Lebenszyklus. Daher muss es Datenwissenschaftlern und Dateningenieuren die Arbeit erleichtern, pragmatische Modelle zu entwickeln, die den Geschäftsanforderungen entsprechen und auf lange Sicht gut funktionieren, ohne dass technische Schulden entstehen.

Folgen Sie den bewährten Methoden in diesem Abschnitt, um die Herausforderungen im Bereich Modelltraining zu bewältigen.

Themen

Erstellen Sie ein Basismodell

Wenn Praktiker mit einer ML-Lösung auf ein Geschäftsproblem stoßen, neigen sie in der Regel zunächst dazu, den state-of-the-art Algorithmus zu verwenden. Diese Vorgehensweise ist riskant, da der state-of-the-art Algorithmus wahrscheinlich nicht erprobt wurde. Darüber hinaus ist der state-of-the-art Algorithmus oft komplexer und nicht gut verstanden, sodass er möglicherweise nur zu geringfügigen Verbesserungen gegenüber einfacheren, alternativen Modellen führt. Eine bessere Vorgehensweise besteht darin, ein Basismodell zu erstellen, das relativ schnell validiert und implementiert werden kann und das Vertrauen der Projektbeteiligten gewinnen kann.

Wenn Sie einen Basisplan erstellen, empfehlen wir Ihnen, wann immer möglich, dessen metrische Leistung zu bewerten. Vergleichen Sie die Leistung des Basismodells mit anderen automatisierten oder manuellen Systemen, um dessen Erfolg sicherzustellen und sicherzustellen, dass die Modellimplementierung oder das Projekt mittel- und langfristig durchgeführt werden können.

Das Basismodell sollte mit ML-Technikern weiter validiert werden, um zu bestätigen, dass das Modell die für das Projekt festgelegten nichtfunktionalen Anforderungen erfüllen kann, wie z. B. die Inferenzzeit, wie oft sich die Verteilung der Daten voraussichtlich ändern wird, ob das Modell in

diesen Fällen leicht umtrainiert werden kann und wie es eingesetzt wird, was sich auf die Kosten der Lösung auswirken wird. Holen Sie sich multidisziplinäre Sichtweisen zu diesen Fragen ein, um die Wahrscheinlichkeit zu erhöhen, dass Sie ein erfolgreiches und langfristiges Modell entwickeln.

Datenwissenschaftler neigen möglicherweise dazu, einem Basismodell so viele Funktionen wie möglich hinzuzufügen. Dies erhöht zwar die Fähigkeit eines Modells, das gewünschte Ergebnis vorherzusagen, einige dieser Funktionen führen jedoch möglicherweise nur zu inkrementellen metrischen Verbesserungen. Viele Merkmale, insbesondere solche, die stark korreliert sind, sind möglicherweise überflüssig. Das Hinzufügen zu vieler Funktionen erhöht die Kosten, da mehr Rechenressourcen und Optimierungen erforderlich sind. Zu viele Funktionen wirken sich auch auf den day-to-day Betrieb des Modells aus, da Datendrift wahrscheinlicher wird oder schneller erfolgt.

Stellen Sie sich ein Modell vor, in dem zwei Eingabe-Features stark korreliert sind, aber nur ein Merkmal kausal ist. Beispielsweise könnte ein Modell, das vorhersagt, ob ein Kredit zahlungsunfähig sein wird, Eingabemerkmale wie Alter und Einkommen des Kunden haben, die stark korrelieren könnten, aber für die Gewährung oder Ablehnung eines Kredits sollte nur das Einkommen verwendet werden. Ein Modell, das anhand dieser beiden Merkmale trainiert wurde, könnte sich bei der Generierung der Prognoseausgabe auf das Merkmal stützen, das keine Kausalität aufweist, wie z. B. das Alter. Wenn das Modell nach der Serienproduktion Anfragen zu Inferenzen von Kunden erhält, die älter oder jünger sind als das im Trainingsset angegebene Durchschnittsalter, kann es zu schlechten Ergebnissen führen.

Darüber hinaus kann es bei jedem einzelnen Merkmal während der Produktion zu einer Verlagerung kommen, was zu einem unerwarteten Verhalten des Modells führen kann. Aus diesen Gründen gilt: Je mehr Merkmale ein Modell aufweist, desto anfälliger ist es in Bezug auf Drift und Veralterung.

Datenwissenschaftler sollten anhand von Korrelationsmaßen und Shapley-Werten beurteilen, welche Merkmale der Vorhersage einen ausreichenden Wert verleihen und beibehalten werden sollten. Solch komplexe Modelle erhöhen die Wahrscheinlichkeit einer Rückkopplungsschleife, in der das Modell die Umgebung verändert, für die es modelliert wurde. Ein Beispiel ist ein Empfehlungssystem, bei dem sich das Verbraucherverhalten aufgrund der Empfehlungen eines Modells ändern kann. Feedback-Schleifen, die modellübergreifend wirken, sind seltener. Stellen Sie sich zum Beispiel ein Empfehlungssystem vor, das Filme empfiehlt, und ein anderes System, das Bücher empfiehlt. Wenn beide Modelle auf dieselbe Gruppe von Verbrauchern abzielen, würden sie sich gegenseitig beeinflussen.

Überlegen Sie bei jedem Modell, das Sie entwickeln, welche Faktoren zu dieser Dynamik beitragen könnten, damit Sie wissen, welche Kennzahlen in der Produktion überwacht werden müssen.

Verwenden Sie einen datenzentrierten Ansatz und eine Fehleranalyse

Wenn Sie ein einfaches Modell verwenden, kann sich Ihr ML-Team darauf konzentrieren, die Daten selbst zu verbessern und statt eines modellzentrierten Ansatzes einen datenzentrierten Ansatz zu wählen. Wenn Ihr Projekt unstrukturierte Daten wie Bilder, Text, Audio und andere Formate verwendet, die von Menschen bewertet werden können (im Vergleich zu strukturierten Daten, die möglicherweise schwieriger sind, effizient einem Label zuzuordnen), ist die Durchführung einer Fehleranalyse eine bewährte Methode, um eine bessere Modellleistung zu erzielen.

Bei der Fehleranalyse wird ein Modell anhand eines Validierungssatzes bewertet und auf die häufigsten Fehler überprüft. Auf diese Weise können potenzielle Gruppen ähnlicher Datenstichproben identifiziert werden, bei denen das Modell möglicherweise Schwierigkeiten hat, sie richtig zu machen. Um eine Fehleranalyse durchzuführen, können Sie Schlussfolgerungen auflisten, die höhere Vorhersagefehler aufwiesen, oder Fehler einstufen, bei denen eine Stichprobe aus einer Klasse als aus einer anderen Klasse stammend vorhergesagt wurde.

Entwerfen Sie Ihr Modell für eine schnelle Iteration

Wenn Datenwissenschaftler sich an bewährte Verfahren halten, können sie während der Machbarkeitsstudie oder sogar bei einer Umschulung einfach und schnell mit einem neuen Algorithmus experimentieren oder verschiedene Funktionen kombinieren und anpassen. Dieses Experimentieren trägt zum Produktionserfolg bei. Eine bewährte Methode besteht darin, auf dem Basismodell aufzubauen, etwas komplexere Algorithmen zu verwenden und iterativ neue Funktionen hinzuzufügen, während gleichzeitig die Leistung im Trainings- und Validierungssatz überwacht wird, um das tatsächliche Verhalten mit dem erwarteten Verhalten zu vergleichen. Dieses Trainingsframework kann für ein optimales Gleichgewicht bei der Vorhersagekraft sorgen und dazu beitragen, dass Modelle so einfach wie möglich gehalten werden und weniger technische Schulden entstehen.

Für eine schnelle Iteration müssen Datenwissenschaftler verschiedene Modellimplementierungen austauschen, um das für bestimmte Daten am besten geeignete Modell zu ermitteln. Wenn Sie ein großes Team, eine kurze Frist und andere logistische Aufgaben im Zusammenhang mit dem Projektmanagement haben, kann eine schnelle Iteration ohne eine Methode schwierig sein.

In der Softwareentwicklung ist das Liskov-Substitutionsprinzip ein Mechanismus zur Gestaltung von Interaktionen zwischen Softwarekomponenten. Dieses Prinzip besagt, dass Sie in der Lage sein sollten, eine Implementierung einer Schnittstelle durch eine andere Implementierung zu ersetzen,

ohne die Client-Anwendung oder die Implementierung zu beschädigen. Wenn Sie Trainingscode für Ihr ML-System schreiben, können Sie dieses Prinzip anwenden, um Grenzen zu setzen und den Code zu kapseln, sodass Sie den Algorithmus einfach ersetzen und neue Algorithmen effektiver ausprobieren können.

Im folgenden Code können Sie beispielsweise neue Experimente hinzufügen, indem Sie einfach eine neue Klassenimplementierung hinzufügen.

```
from abc import ABC, abstractmethod

from pandas import DataFrame


class ExperimentRunner(object):

    def __init__(self, *experiments):
        self.experiments = experiments

    def run(self, df: DataFrame) -> None:
        for experiment in self.experiments:
            result = experiment.run(df)
            print(f'Experiment "{experiment.name}" gave result {result}')


class Experiment(ABC):

    @abstractmethod
    def run(self, df: DataFrame) -> float:
        pass

    @property
    @abstractmethod
    def name(self) -> str:
        pass


class Experiment1(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 1')
        return 0

    def name(self) -> str:
```

```
        return 'experiment 1'


class Experiment2(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 2')
        return 0

    def name(self) -> str:
        return 'experiment 2'


class Experiment3(Experiment):

    def run(self, df: DataFrame) -> float:
        print('performing experiment 3')
        return 0

    def name(self) -> str:
        return 'experiment 3'


if __name__ == '__main__':
    runner = ExperimentRunner(*[
        Experiment1(),
        Experiment2(),
        Experiment3()
    ])
    df = ...
    runner.run(df)
```

Verfolgen Sie Ihre ML-Experimente

Wenn Sie mit einer großen Anzahl von Experimenten arbeiten, ist es wichtig abzuwägen, ob die Verbesserungen, die Sie beobachten, auf umgesetzte Änderungen oder auf Zufall zurückzuführen sind. Sie können Amazon SageMaker AI Experiments verwenden, um auf einfache Weise Experimente zu erstellen und ihnen Metadaten zur Nachverfolgung, zum Vergleich und zur Auswertung zuzuordnen.

Die Reduzierung der Zufälligkeit des Modellerstellungsprozesses ist nützlich für das Debuggen, die Fehlerbehebung und die Verbesserung der Steuerung, da Sie die Inferenz des Ausgabemodells bei gleichem Code und denselben Daten mit größerer Sicherheit vorhersagen können.

Aufgrund zufälliger Gewichtungsinitialisierung, parallel Rechensynchronität, innerer GPU-Komplexität und ähnlicher nichtdeterministischer Faktoren ist es oft nicht möglich, einen Trainingscode vollständig reproduzierbar zu machen. Die korrekte Festlegung von Zufallszahlen, um sicherzustellen, dass jeder Trainingslauf am selben Punkt beginnt und sich ähnlich verhält, verbessert jedoch die Vorhersagbarkeit der Ergebnisse erheblich.

Beheben Sie Fehler bei Trainingsaufträgen

In einigen Fällen kann es für Datenwissenschaftler schwierig sein, selbst ein sehr einfaches Basismodell anzupassen. In diesem Fall könnten sie entscheiden, dass sie einen Algorithmus benötigen, der besser für komplexe Funktionen geeignet ist. Ein guter Test besteht darin, die Basislinie eines sehr kleinen Teils des Datensatzes (z. B. etwa 10 Stichproben) zu verwenden, um sicherzustellen, dass der Algorithmus zu dieser Stichprobe passt. Auf diese Weise können Daten- oder Codeprobleme ausgeschlossen werden.

Ein weiteres hilfreiches Tool für das Debuggen komplexer Szenarien ist Amazon SageMaker AI Debugger, mit dem Probleme im Zusammenhang mit der algorithmischen Korrektheit und Infrastruktur, wie z. B. optimale Computernutzung, erfasst werden können.

Bereitstellung

In der Softwareentwicklung erfordert die Implementierung von Code in der Produktion die gebotene Sorgfalt, da sich Code möglicherweise unerwartet verhält, unvorhergesehenes Benutzerverhalten die Software beschädigen kann und unerwartete Grenzfälle gefunden werden können. Softwareingenieure und DevOps -ingenieure setzen in der Regel Komponententests und Rollback-Strategien ein, um diese Risiken zu minimieren. Mit ML erfordert die Einführung von Modellen in der Produktion noch mehr Planung, da davon ausgegangen wird, dass sich die reale Umgebung verändert. In vielen Fällen werden Modelle anhand von Kennzahlen validiert, die als Stellvertreter für die tatsächlichen Geschäftskennzahlen dienen, die sie verbessern möchten.

Folgen Sie den bewährten Methoden in diesem Abschnitt, um diese Herausforderungen zu bewältigen.

Themen

Automatisieren Sie den Bereitstellungszyklus

Der Schulungs- und Bereitstellungsprozess sollte vollständig automatisiert werden, um menschliche Fehler zu vermeiden und sicherzustellen, dass Build-Checks konsistent durchgeführt werden. Benutzer sollten keine Schreibzugriffsberechtigungen für die Produktionsumgebung haben.

Amazon SageMaker AI Pipelines und AWS CodePipelineHilfe bei der Erstellung von CI/CD pipelines for ML projects. One of the advantages of using a CI/CD Pipelines bestehen darin, dass der gesamte Code, der zum Erfassen von Daten, zum Trainieren eines Modells und zur Überwachung verwendet wird, mithilfe eines Tools wie Git versionskontrolliert werden kann. Manchmal müssen Sie ein Modell neu trainieren, indem Sie denselben Algorithmus und dieselben Hyperparameter, aber unterschiedliche Daten verwenden. Die einzige Möglichkeit, zu überprüfen, ob Sie die richtige Version des Algorithmus verwenden, ist die Verwendung von Quellcodeverwaltung und Tags. Sie können die von SageMaker AI bereitgestellten Standardprojektvorlagen als Ausgangspunkt für Ihre MLOps Praxis verwenden.

Wenn Sie CI/CD-Pipelines für die Bereitstellung Ihres Modells erstellen, achten Sie darauf, Ihre Build-Artefakte mit einer Build-ID, einer Codeversion oder einem Commit und einer Datenversion zu

kennzeichnen. Diese Vorgehensweise hilft Ihnen bei der Behebung von Bereitstellungsproblemen. Tagging ist manchmal auch für Modelle erforderlich, die Vorhersagen in stark regulierten Bereichen treffen. Die Fähigkeit, rückwärts zu arbeiten und die genauen Daten, Codes, Builds, Prüfungen und Genehmigungen zu identifizieren, die mit einem ML-Modell verknüpft sind, kann dazu beitragen, die Steuerung erheblich zu verbessern.

Ein Teil der Aufgabe der CI/CD-Pipeline besteht darin, Tests an dem durchzuführen, was sie gerade erstellt. Obwohl davon ausgegangen wird, dass Dateneinheitstests durchgeführt werden, bevor die Daten von einem feature store aufgenommen werden, ist die Pipeline dennoch dafür verantwortlich, Tests an der Eingabe und Ausgabe eines bestimmten Modells durchzuführen und wichtige Metriken zu überprüfen. Ein Beispiel für eine solche Prüfung besteht darin, ein neues Modell anhand eines festen Validierungssatzes zu validieren und anhand eines festgelegten Schwellenwerts zu bestätigen, dass seine Leistung mit dem vorherigen Modell vergleichbar ist. Wenn die Leistung deutlich unter den Erwartungen liegt, sollte der Bau fehlschlagen und das Modell sollte nicht in Produktion gehen.

Der umfangreiche Einsatz von CI/CD-Pipelines unterstützt auch Pull-Requests, wodurch menschliche Fehler vermieden werden können. Wenn Sie Pull-Requests verwenden, muss jede Codeänderung von mindestens einem anderen Teammitglied überprüft und genehmigt werden, bevor sie in Produktion gehen kann. Pull Requests sind auch nützlich, um Code zu identifizieren, der nicht den Geschäftsregeln entspricht, und um Wissen innerhalb des Teams zu verbreiten.

Wählen Sie eine Bereitstellungsstrategie

MLOps Zu den Bereitstellungsstrategien gehören blue/green, canary, shadow, and A/B Tests.

Blau/Grün

Blue/green deployments are very common in software development. In this mode, two systems are kept running during development: blue is the old environment (in this case, the model that is being replaced) and green is the newly released model that is going to production. Changes can easily be rolled back with minimum downtime, because the old system is kept alive. For more in-depth information about blue/greenBereitstellungen im Kontext von SageMaker, siehe den Blogbeitrag Sichere Bereitstellung und Überwachung von Amazon SageMaker KI-Endpunkten mit AWS CodePipeline und AWS CodeDeploy im AWS Machine Learning Learning-Blog.

Canary

Bereitstellungen auf Canary ähneln blue/green deployments in that both keep two models running together. However, in canary deployments, the new model is rolled out to users incrementally, until

all traffic eventually shifts over to the new model. As in blue/green Bereitstellungen. Das Risiko wird minimiert, da das neue (und potenziell fehlerhafte) Modell beim ersten Rollout genau überwacht wird und bei Problemen rückgängig gemacht werden kann. In SageMaker KI können Sie mithilfe der API die anfängliche Verteilung des Datenverkehrs festlegen. InitialVariantWeight

Shadow

Sie können Schattenbereitstellungen verwenden, um ein Modell sicher in die Produktion zu bringen. In diesem Modus arbeitet das neue Modell mit einem älteren Modell oder Geschäftsprozess zusammen und führt Schlussfolgerungen durch, ohne Entscheidungen zu beeinflussen. Dieser Modus kann als abschließende Prüfung oder als Experiment mit höherer Genauigkeit nützlich sein, bevor Sie das Modell zur Serienproduktion hochstufen.

Der Schattenmodus ist nützlich, wenn Sie kein Feedback zu Benutzerinferenzen benötigen. Sie können die Qualität der Vorhersagen beurteilen, indem Sie eine Fehleranalyse durchführen und das neue Modell mit dem alten Modell vergleichen, und Sie können die Ausgabeverteilung überwachen, um sicherzustellen, dass sie den Erwartungen entspricht. Informationen zur Schattenbereitstellung mit SageMaker KI finden Sie im Blogbeitrag Deploy Shadow ML-Modelle in Amazon SageMaker AI auf dem AWS Machine Learning Learning-Blog.

A/B-Tests

Wenn ML-Praktiker Modelle in ihren Umgebungen entwickeln, sind die Metriken, für die sie optimieren, oft Proxys für die Geschäftskennzahlen, die wirklich wichtig sind. Dies macht es schwierig, mit Sicherheit zu sagen, ob ein neues Modell die Geschäftsergebnisse wie Umsatz und Klickrate tatsächlich verbessert und die Anzahl der Benutzerbeschwerden reduziert.

Stellen Sie sich den Fall einer E-Commerce-Website vor, auf der das Geschäftsziel darin besteht, so viele Produkte wie möglich zu verkaufen. Das Bewertungsteam weiß, dass Umsatz und Kundenzufriedenheit direkt mit informativen und genauen Bewertungen korrelieren. Ein Teammitglied könnte einen neuen Algorithmus zur Bewertung von Bewertungen vorschlagen, um den Umsatz zu verbessern. Mithilfe von A/B-Tests könnten sie die alten und neuen Algorithmen auf verschiedene, aber ähnliche Benutzergruppen ausrollen und die Ergebnisse überwachen, um festzustellen, ob Nutzer, die Prognosen anhand des neueren Modells erhalten haben, eher Käufe tätigen.

A/B-Tests helfen auch dabei, die Auswirkungen von veralteten Modellen und Abweichungen auf das Geschäft einzuschätzen. Teams können neue Modelle mit einer gewissen Wiederholung in Produktion nehmen, A/B-Tests für jedes Modell durchführen und ein Diagramm zwischen Alter und

Leistung erstellen. Dies würde dem Team helfen, die schwankenden Datenunterschiede in ihren Produktionsdaten besser zu verstehen.

Weitere Informationen zur Durchführung von A/B-Tests mit SageMaker KI finden Sie im Blogbeitrag A/B-Tests von ML-Modellen in der Produktion mit Amazon SageMaker AI im AWS Machine Learning Learning-Blog.

Berücksichtigen Sie Ihre Anforderungen an die Inferenz

Mit SageMaker KI können Sie die zugrunde liegende Infrastruktur für die Bereitstellung Ihres Modells auf unterschiedliche Weise auswählen. Diese Funktionen zum Aufrufen von Inferenzen unterstützen unterschiedliche Anwendungsfälle und Kostenprofile. Zu Ihren Optionen gehören Inferenz in Echtzeit, asynchrone Inferenz und Batch-Transformation, wie in den folgenden Abschnitten beschrieben.

Echtzeit-Inferenz

Inferenz in Echtzeit ist ideal für Inferenz-Workloads, bei denen interaktive Echtzeitanforderungen mit niedriger Latenz erfüllt werden müssen. Sie können Ihr Modell für SageMaker KI-Hosting-Dienste bereitstellen und erhalten einen Endpunkt, der für Inferenzen verwendet werden kann. Diese Endgeräte werden vollständig verwaltet, unterstützen automatische Skalierung (siehe Automatische Skalierung von Amazon SageMaker KI-Modellen) und können in mehreren Availability Zones eingesetzt werden.

Wenn Sie ein Deep-Learning-Modell haben, das mit Apache MXNet PyTorch TensorFlow, oder erstellt wurde, können Sie auch Amazon SageMaker AI Elastic Inference (EI) verwenden. Mit EI können Sie jeder SageMaker AI-Instance einen Bruchteil zuordnen GPUs , um die Inferenz zu beschleunigen. Sie können die Client-Instanz auswählen, auf der Ihre Anwendung ausgeführt werden soll, und einen EI-Beschleuniger anhängen, um die richtige Menge an GPU-Beschleunigung für Ihre Inferenzanforderungen zu verwenden.

Eine weitere Option ist die Verwendung von Endpunkten mit mehreren Modellen, die eine skalierbare und kostengünstige Lösung für die Bereitstellung einer großen Anzahl von Modellen bieten. Diese Endgeräte verwenden einen gemeinsam genutzten Serving-Container, der mehrere Modelle hosten kann. Endgeräte mit mehreren Modellen reduzieren die Hosting-Kosten, da sie die Endpunktauslastung im Vergleich zur Verwendung von Endgeräten mit einem einzigen Modell verbessern. Sie reduzieren auch den Bereitstellungsaufwand, da SageMaker KI das Laden von Modellen im Speicher und deren Skalierung auf der Grundlage von Verkehrsmustern verwaltet.

Weitere bewährte Methoden für den Einsatz von ML-Modellen in SageMaker KI finden Sie in der SageMaker KI-Dokumentation unter Bewährte Methoden für die Bereitstellung.

Asynchrone Inferenz-Inferenz

Amazon SageMaker AI Asynchronous Inference ist eine SageMaker KI-Funktion, die eingehende Anfragen in eine Warteschlange stellt und sie asynchron verarbeitet. Diese Option ist ideal für Anfragen mit großen Nutzlasten von bis zu 1 GB, langen Verarbeitungszeiten und Latenzanforderungen nahezu in Echtzeit. Durch asynchrone Inferenz können Sie Kosten sparen, indem Sie die Anzahl der Instanzen automatisch auf Null skalieren, wenn keine Anfragen zu verarbeiten sind. Sie zahlen also nur, wenn Ihr Endpunkt Anfragen verarbeitet.

Batch-Transformation

Verwenden Sie die Batch-Transformation, wenn Sie Folgendes tun möchten:

- Vorverarbeitung von Datensätzen, um Rauschen oder Bias, das das Training oder Inferenz beeinträchtigt, aus Ihrem Datensatz zu entfernen.
- Abrufen von Inferenzen aus großen Datensätzen.
- Ausführen der Inferenz, wenn Sie keinen persistenten Endpunkt benötigen.
- Verknüpfen von Eingabedatensätzen mit Inferenzen, um die Interpretation der Ergebnisse zu unterstützen.

Überwachen

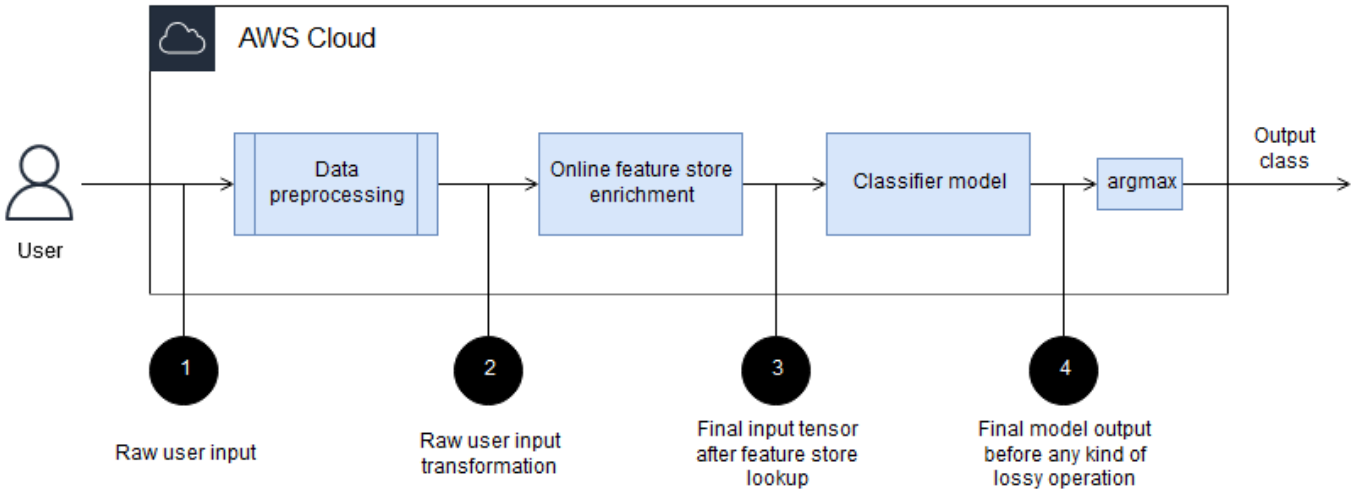
Wenn Modelle bereits in der Produktion sind und einen Mehrwert für Ihr Unternehmen bieten, sollten Sie kontinuierliche Prüfungen durchführen, um festzustellen, wann Modelle neu trainiert oder Maßnahmen ergriffen werden müssen.

Ihr Überwachungsteam sollte proaktiv und nicht reaktiv handeln, um das Datenverhalten der Umgebung besser zu verstehen und die Häufigkeit, Geschwindigkeit und Abruptheit von Datenverschiebungen zu ermitteln. Das Team sollte neue Grenzfälle in den Daten identifizieren, die im Trainingsset, im Validierungssatz und in anderen Grenzfällen möglicherweise unterrepräsentiert sind. Sie sollten QoS-Metriken (Quality of Service) speichern, Alarme verwenden, um sofort Maßnahmen zu ergreifen, wenn ein Problem auftritt, und eine Strategie zur Aufnahme und Änderung aktueller Datensätze definieren. Diese Verfahren beginnen mit der Protokollierung von Anfragen und Antworten für das Modell, um als Referenz für die Fehlerbehebung oder zusätzliche Erkenntnisse zu dienen.

Im Idealfall sollten Datentransformationen während der Verarbeitung in einigen wichtigen Phasen protokolliert werden:

- Vor jeder Art von Vorverarbeitung
- Nach jeder Art von Feature-Store-Anreicherung
- Nach allen Hauptphasen eines Modells
- Vor jeder Art von verlustbehafteter Funktion auf der Modellausgabe, wie `argmax`

Das folgende Diagramm veranschaulicht diese Phasen.



Sie können SageMaker AI Model Monitor verwenden, um Eingabe- und Ausgabedaten automatisch zu erfassen und in Amazon Simple Storage Service (Amazon S3) zu speichern. Sie können andere Arten der Zwischenprotokollierung implementieren, indem Sie Protokolle zu einem benutzerdefinierten Serving-Container hinzufügen.

Nachdem Sie die Daten aus den Modellen protokolliert haben, können Sie die Verteilungsabweichung überwachen. In einigen Fällen können Sie schon bald nach der Inferenz Ground Truth (Daten, die korrekt beschriftet sind) abrufen. Ein gängiges Beispiel hierfür ist ein Modell, das vorhersagt, welche Anzeigen für einen Nutzer am relevantesten sind. Sobald der Nutzer die Seite verlassen hat, können Sie feststellen, ob er auf die Anzeige geklickt hat. Wenn der Nutzer auf die Anzeige geklickt hat, können Sie diese Informationen protokollieren. In diesem einfachen Beispiel können Sie leicht quantifizieren, wie erfolgreich Ihr Modell ist, indem Sie eine Metrik wie Genauigkeit oder F1 verwenden, die sowohl im Training als auch im Einsatz gemessen werden kann. Weitere Informationen zu diesen Szenarien, in denen Sie Daten beschriftet haben, finden Sie in der SageMaker KI-Dokumentation unter Überwachen der Modellqualität. Diese einfachen Szenarien kommen jedoch selten vor, da Modelle häufig darauf ausgelegt sind, mathematisch sinnvolle Metriken zu optimieren, die nur stellvertretend für tatsächliche Geschäftsergebnisse sind. In solchen Fällen besteht die bewährte Methode darin, das Geschäftsergebnis zu überwachen, wenn ein Modell in der Produktion eingesetzt wird.

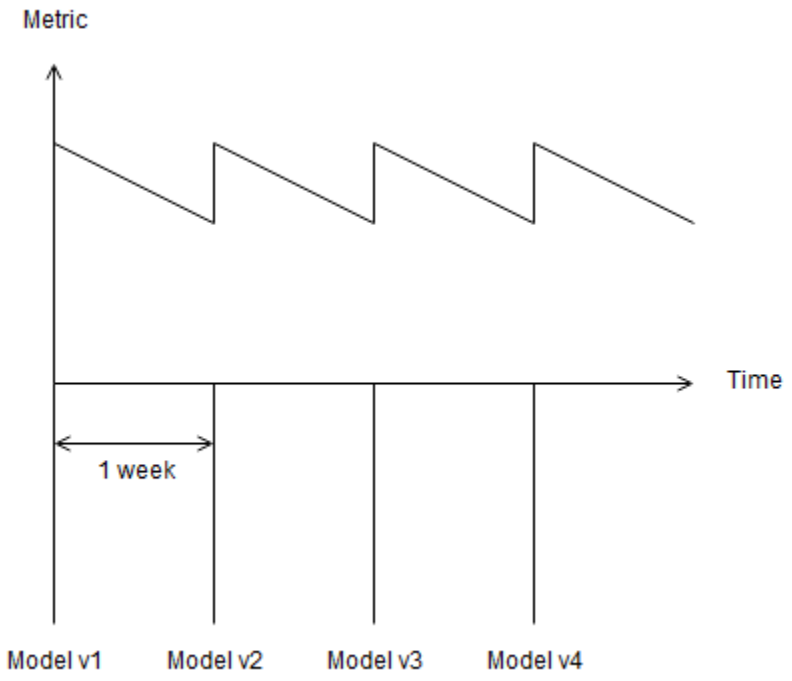
Stellen Sie sich das Beispiel eines Bewertungs-Ranking-Modells vor. Wenn das definierte Geschäftsergebnis des ML-Modells darin besteht, die relevantesten und nützlichsten Bewertungen oben auf der Webseite anzuzeigen, können Sie den Erfolg des Modells messen, indem Sie eine Schaltfläche wie „War das hilfreich?“ hinzufügen für jede Bewertung. Die Messung der Klickrate

dieser Schaltfläche könnte ein Maßstab für das Geschäftsergebnis sein, anhand dessen Sie messen können, wie gut Ihr Modell in der Produktion abschneidet.

Um die Abweichung der Eingabe- oder Ausgabebeschriftungen in SageMaker KI zu überwachen, können Sie die Datenqualitätsfunktionen von SageMaker AI Model Monitor verwenden, die sowohl die Eingabe als auch die Ausgabe überwachen. Sie können auch Ihre eigene Logik für SageMaker AI Model Monitor implementieren, indem Sie einen benutzerdefinierten Container erstellen.

Die Überwachung der Daten, die ein Modell sowohl während der Entwicklungszeit als auch während der Laufzeit empfängt, ist von entscheidender Bedeutung. Techniker sollten die Daten nicht nur auf Schemaänderungen, sondern auch auf Diskrepanzen bei der Verteilung überwachen. Das Erkennen von Schemaänderungen ist einfacher und kann durch eine Reihe von Regeln implementiert werden, aber Verteilungskonflikte sind oft schwieriger, vor allem, weil Sie dafür einen Schwellenwert definieren müssen, um zu quantifizieren, wann ein Alarm ausgelöst werden muss. In Fällen, in denen die überwachte Verteilung bekannt ist, ist es oft am einfachsten, die Parameter der Verteilung zu überwachen. Bei einer Normalverteilung wären das der Mittelwert und die Standardabweichung. Andere wichtige Kennzahlen, wie der Prozentsatz fehlender Werte, Maximal- und Minimalwerte, sind ebenfalls nützlich.

Sie können auch laufende Monitoring-Jobs erstellen, bei denen Trainingsdaten und Inferenzdaten ausgewählt und deren Verteilungen verglichen werden. Sie können diese Jobs sowohl für die Modelleingabe als auch für die Modellausgabe erstellen und die Daten im Zeitverlauf grafisch darstellen, um plötzliche oder allmähliche Abweichungen zu visualisieren. Dies wird in der folgenden Tabelle veranschaulicht.



Um das Driftprofil der Daten besser zu verstehen, z. B. wie oft, mit welcher Geschwindigkeit oder wie plötzlich sich die Datenverteilung signifikant ändert, empfehlen wir, kontinuierlich neue Modellversionen bereitzustellen und deren Leistung zu überwachen. Wenn Ihr Team beispielsweise jede Woche ein neues Modell einführt und feststellt, dass sich die Modellleistung jedes Mal erheblich verbessert, kann es festlegen, dass es neue Modelle mindestens in weniger als einer Woche liefern sollte.

Nächste Schritte und Ressourcen

Dieser Leitfaden führt Sie durch einige Überlegungen bei der Planung des Lebenszyklus der Machine-Learning-Modelle, die Sie in die Produktion bringen möchten. Es behandelt Herausforderungen und bewährte Verfahren in vier Bereichen — Daten, Schulung, Bereitstellung und Überwachung — und enthält zusätzliche relevante Ressourcen.

AWS bietet das Well-Architected Framework, das Cloud-Architekten dabei unterstützt, sichere, leistungsstarke, belastbare und effiziente Infrastrukturen für eine Vielzahl von Anwendungen, Workloads und Technologiedomänen aufzubauen. Weitere Informationen finden Sie in der von AWS Well-Architected angebotenen Machine Learning Lens.

Ressourcen

Dokumentation zu Amazon SageMaker AI

- Amazon SageMaker AI Feature Store
- Sicherheit und Zugriffskontrolle im Feature Store
- Shapley-Werte
- Amazon SageMaker KI-Debugger
- Amazon SageMaker KI-Pipelines
- Amazon SageMaker AI-Standardprojektvorlagen
- SageMaker KI-Inferenz in Echtzeit
- Automatisches Skalieren von Amazon SageMaker AI-Modellen
- Asynchrone Amazon SageMaker AI-Inferenz
- SageMaker KI-Modellmonitor

AWS Tools für Entwickler

- AWS CodePipeline

AWS Blog-Beiträge

- Die wichtigsten Funktionen von Amazon SageMaker AI Feature Store verstehen

- Testen Sie die Datenqualität im großen Maßstab mit PyDeequ
- SageMaker KI-Experimente mit Amazon
- Sichere Bereitstellung und Überwachung von SageMaker Amazon-Endpunkten mit CodePipeline und AWS CodeDeploy
- Stellen Sie Schatten-ML-Modelle in Amazon SageMaker AI bereit
- A/B-Tests von ML-Modellen in der Produktion mit Amazon AI SageMaker

Dokumentverlauf

In der folgenden Tabelle werden wichtige Änderungen in diesem Leitfaden beschrieben. Um Benachrichtigungen über zukünftige Aktualisierungen zu erhalten, können Sie einen RSS-Feed abonnieren.

Änderung	Beschreibung	Datum
Erste Veröffentlichung	—	20. Dezember 2021

AWS Glossar zu präskriptiven Leitlinien

Die folgenden Begriffe werden häufig in Strategien, Leitfäden und Mustern von AWS Prescriptive Guidance verwendet. Um Einträge vorzuschlagen, verwenden Sie bitte den Link Feedback geben am Ende des Glossars.

Zahlen

7 Rs

Sieben gängige Migrationsstrategien für die Verlagerung von Anwendungen in die Cloud. Diese Strategien bauen auf den 5 Rs auf, die Gartner 2011 identifiziert hat, und bestehen aus folgenden Elementen:

- Faktorwechsel/Architekturwechsel – Verschieben Sie eine Anwendung und ändern Sie ihre Architektur, indem Sie alle Vorteile cloudnativer Feature nutzen, um Agilität, Leistung und Skalierbarkeit zu verbessern. Dies beinhaltet in der Regel die Portierung des Betriebssystems und der Datenbank. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank auf die Amazon Aurora PostgreSQL-kompatible Edition.
- Plattformwechsel (Lift and Reshape) – Verschieben Sie eine Anwendung in die Cloud und führen Sie ein gewisses Maß an Optimierung ein, um die Cloud-Funktionen zu nutzen. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank zu Amazon Relational Database Service (Amazon RDS) für Oracle in der. AWS Cloud
- Neukauf (Drop and Shop) – Wechseln Sie zu einem anderen Produkt, indem Sie typischerweise von einer herkömmlichen Lizenz zu einem SaaS-Modell wechseln. Beispiel: Migrieren Sie Ihr CRM-System (Customer Relationship Management) zu Salesforce.com.
- Hostwechsel (Lift and Shift) – Verschieben Sie eine Anwendung in die Cloud, ohne Änderungen vorzunehmen, um die Cloud-Funktionen zu nutzen. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank zu Oracle auf einer EC2-Instanz in der. AWS Cloud
- Verschieben (Lift and Shift auf Hypervisor-Ebene) – Verlagern Sie die Infrastruktur in die Cloud, ohne neue Hardware kaufen, Anwendungen umschreiben oder Ihre bestehenden Abläufe ändern zu müssen. Sie migrieren Server von einer lokalen Plattform zu einem Cloud-Dienst für dieselbe Plattform. Beispiel: Migrieren Sie eine Microsoft Hyper-V Anwendung zu AWS.
- Beibehaltung (Wiederaufgreifen) – Bewahren Sie Anwendungen in Ihrer Quellumgebung auf. Dazu können Anwendungen gehören, die einen umfangreichen Faktorwechsel erfordern und

die Sie auf einen späteren Zeitpunkt verschieben möchten, sowie ältere Anwendungen, die Sie beibehalten möchten, da es keine geschäftliche Rechtfertigung für ihre Migration gibt.

- Außerbetriebnahme – Dekommissionierung oder Entfernung von Anwendungen, die in Ihrer Quellumgebung nicht mehr benötigt werden.

A

ABAC

Siehe attributbasierte Zugriffskontrolle.

abstrahierte Dienste

Siehe Managed Services.

ACID

Siehe Atomarität, Konsistenz, Isolierung und Haltbarkeit.

Aktiv-Aktiv-Migration

Eine Datenbankmigrationsmethode, bei der die Quell- und Zieldatenbanken synchron gehalten werden (mithilfe eines bidirektionalen Replikationstools oder dualer Schreibvorgänge) und beide Datenbanken Transaktionen von miteinander verbundenen Anwendungen während der Migration verarbeiten. Diese Methode unterstützt die Migration in kleinen, kontrollierten Batches, anstatt einen einmaligen Cutover zu erfordern. Es ist flexibler, erfordert aber mehr Arbeit als eine aktiv-passive Migration.

Aktiv-Passiv-Migration

Eine Datenbankmigrationsmethode, bei der die Quell- und Zieldatenbanken synchron gehalten werden, aber nur die Quelldatenbank verarbeitet Transaktionen von verbindenden Anwendungen, während Daten in die Zieldatenbank repliziert werden. Die Zieldatenbank akzeptiert während der Migration keine Transaktionen.

Aggregatfunktion

Eine SQL-Funktion, die mit einer Gruppe von Zeilen arbeitet und einen einzelnen Rückgabewert für die Gruppe berechnet. Beispiele für Aggregatfunktionen sind `SUM` und`MAX`.

AI

Siehe künstliche Intelligenz.

AIOps

Siehe Operationen im Bereich künstliche Intelligenz.

Anonymisierung

Der Prozess des dauerhaften Löschens personenbezogener Daten in einem Datensatz. Anonymisierung kann zum Schutz der Privatsphäre beitragen. Anonymisierte Daten gelten nicht mehr als personenbezogene Daten.

Anti-Muster

Eine häufig verwendete Lösung für ein wiederkehrendes Problem, bei dem die Lösung kontraproduktiv, ineffektiv oder weniger wirksam als eine Alternative ist.

Anwendungssteuerung

Ein Sicherheitsansatz, bei dem nur zugelassene Anwendungen verwendet werden können, um ein System vor Schadsoftware zu schützen.

Anwendungsportfolio

Eine Sammlung detaillierter Informationen zu jeder Anwendung, die von einer Organisation verwendet wird, einschließlich der Kosten für die Erstellung und Wartung der Anwendung und ihres Geschäftswerts. Diese Informationen sind entscheidend für den Prozess der Portfoliofindung und -analyse und hilft bei der Identifizierung und Priorisierung der Anwendungen, die migriert, modernisiert und optimiert werden sollen.

künstliche Intelligenz (KI)

Das Gebiet der Datenverarbeitungswissenschaft, das sich der Nutzung von Computertechnologien zur Ausführung kognitiver Funktionen widmet, die typischerweise mit Menschen in Verbindung gebracht werden, wie Lernen, Problemlösen und Erkennen von Mustern. Weitere Informationen finden Sie unter Was ist künstliche Intelligenz?

Operationen mit künstlicher Intelligenz (AIOps)

Der Prozess des Einsatzes von Techniken des Machine Learning zur Lösung betrieblicher Probleme, zur Reduzierung betrieblicher Zwischenfälle und menschlicher Eingriffe sowie zur Steigerung der Servicequalität. Weitere Informationen zur Verwendung in der AWS Migrationsstrategie finden Sie im Operations Integration Guide. AIOps

Asymmetrische Verschlüsselung

Ein Verschlüsselungsalgorithmus, der ein Schlüsselpaar, einen öffentlichen Schlüssel für die Verschlüsselung und einen privaten Schlüssel für die Entschlüsselung verwendet. Sie können den

öffentlichen Schlüssel teilen, da er nicht für die Entschlüsselung verwendet wird. Der Zugriff auf den privaten Schlüssel sollte jedoch stark eingeschränkt sein.

Atomizität, Konsistenz, Isolierung, Haltbarkeit (ACID)

Eine Reihe von Softwareeigenschaften, die die Datenvalidität und betriebliche Zuverlässigkeit einer Datenbank auch bei Fehlern, Stromausfällen oder anderen Problemen gewährleisten.

Attributbasierte Zugriffskontrolle (ABAC)

Die Praxis, detaillierte Berechtigungen auf der Grundlage von Benutzerattributen wie Abteilung, Aufgabenrolle und Teamname zu erstellen. Weitere Informationen finden Sie unter ABAC AWS in der AWS Identity and Access Management (IAM-) Dokumentation.

maßgebliche Datenquelle

Ein Ort, an dem Sie die primäre Version der Daten speichern, die als die zuverlässigste Informationsquelle angesehen wird. Sie können Daten aus der maßgeblichen Datenquelle an andere Speicherorte kopieren, um die Daten zu verarbeiten oder zu ändern, z. B. zu anonymisieren, zu redigieren oder zu pseudonymisieren.

Availability Zone

Ein bestimmter Standort innerhalb einer AWS-Region , der vor Ausfällen in anderen Availability Zones geschützt ist und kostengünstige Netzwerkkonnektivität mit niedriger Latenz zu anderen Availability Zones in derselben Region bietet.

AWS Framework für die Einführung der Cloud (AWS CAF)

Ein Framework mit Richtlinien und bewährten Verfahren, das Unternehmen bei der Entwicklung eines effizienten und effektiven Plans für die erfolgreiche Umstellung auf die Cloud unterstützt. AWS AWS CAF unterteilt die Leitlinien in sechs Schwerpunktbereiche, die als Perspektiven bezeichnet werden: Unternehmen, Mitarbeiter, Unternehmensführung, Plattform, Sicherheit und Betrieb. Die Perspektiven Geschäft, Mitarbeiter und Unternehmensführung konzentrieren sich auf Geschäftskompetenzen und -prozesse, während sich die Perspektiven Plattform, Sicherheit und Betriebsabläufe auf technische Fähigkeiten und Prozesse konzentrieren. Die Personalperspektive zielt beispielsweise auf Stakeholder ab, die sich mit Personalwesen (HR), Personalfunktionen und Personalmanagement befassen. Aus dieser Perspektive bietet AWS CAF Leitlinien für Personalentwicklung, Schulung und Kommunikation, um das Unternehmen auf eine erfolgreiche Cloud-Einführung vorzubereiten. Weitere Informationen finden Sie auf der AWS -CAF-Webseite und dem AWS -CAF-Whitepaper.

AWS Workload-Qualifizierungsrahmen (AWS WQF)

Ein Tool, das Workloads bei der Datenbankmigration bewertet, Migrationsstrategien empfiehlt und Arbeitsschätzungen bereitstellt. AWS WQF ist in () enthalten. AWS Schema Conversion Tool AWS SCT Es analysiert Datenbankschemas und Codeobjekte, Anwendungscode, Abhängigkeiten und Leistungsmerkmale und stellt Bewertungsberichte bereit.

B

schlechter Bot

Ein Bot, der Einzelpersonen oder Organisationen stören oder ihnen Schaden zufügen soll.

BCP

Siehe Planung der Geschäftskontinuität.

Verhaltensdiagramm

Eine einheitliche, interaktive Ansicht des Ressourcenverhaltens und der Interaktionen im Laufe der Zeit. Sie können ein Verhaltensdiagramm mit Amazon Detective verwenden, um fehlgeschlagene Anmeldeversuche, verdächtige API-Aufrufe und ähnliche Vorgänge zu untersuchen. Weitere Informationen finden Sie unter Daten in einem Verhaltensdiagramm in der Detective-Dokumentation.

Big-Endian-System

Ein System, welches das höchstwertige Byte zuerst speichert. Siehe auch Endianness.

Binäre Klassifikation

Ein Prozess, der ein binäres Ergebnis vorhersagt (eine von zwei möglichen Klassen). Beispielsweise könnte Ihr ML-Modell möglicherweise Probleme wie „Handelt es sich bei dieser E-Mail um Spam oder nicht?“ vorhersagen müssen oder „Ist dieses Produkt ein Buch oder ein Auto?“

Bloom-Filter

Eine probabilistische, speichereffiziente Datenstruktur, mit der getestet wird, ob ein Element Teil einer Menge ist.

Blau/Grün-Bereitstellung

Eine Bereitstellungsstrategie, bei der Sie zwei separate, aber identische Umgebungen erstellen. Sie führen die aktuelle Anwendungsversion in einer Umgebung (blau) und die neue

Anwendungsversion in der anderen Umgebung (grün) aus. Mit dieser Strategie können Sie schnell und mit minimalen Auswirkungen ein Rollback durchführen.

Bot

Eine Softwareanwendung, die automatisierte Aufgaben über das Internet ausführt und menschliche Aktivitäten oder Interaktionen simuliert. Manche Bots sind nützlich oder nützlich, wie z. B. Webcrawler, die Informationen im Internet indexieren. Einige andere Bots, sogenannte bösartige Bots, sollen Einzelpersonen oder Organisationen stören oder ihnen Schaden zufügen.

Botnetz

Netzwerke von Bots, die mit Malware infiziert sind und unter der Kontrolle einer einzigen Partei stehen, die als Bot-Herder oder Bot-Operator bezeichnet wird. Botnetze sind der bekannteste Mechanismus zur Skalierung von Bots und ihrer Wirkung.

branch

Ein containerisierter Bereich eines Code-Repositorys. Der erste Zweig, der in einem Repository erstellt wurde, ist der Hauptzweig. Sie können einen neuen Zweig aus einem vorhandenen Zweig erstellen und dann Feature entwickeln oder Fehler in dem neuen Zweig beheben. Ein Zweig, den Sie erstellen, um ein Feature zu erstellen, wird allgemein als Feature-Zweig bezeichnet. Wenn das Feature zur Veröffentlichung bereit ist, führen Sie den Feature-Zweig wieder mit dem Hauptzweig zusammen. Weitere Informationen finden Sie unter Über Branches (GitHub Dokumentation).

Zugang durch Glasbruch

Unter außergewöhnlichen Umständen und im Rahmen eines genehmigten Verfahrens ist dies eine schnelle Methode für einen Benutzer, auf einen Bereich zuzugreifen AWS-Konto , für den er normalerweise keine Zugriffsrechte besitzt. Weitere Informationen finden Sie unter dem Indikator Implementation break-glass procedures in den AWS Well-Architected-Leitlinien.

Brownfield-Strategie

Die bestehende Infrastruktur in Ihrer Umgebung. Wenn Sie eine Brownfield-Strategie für eine Systemarchitektur anwenden, richten Sie sich bei der Gestaltung der Architektur nach den Einschränkungen der aktuellen Systeme und Infrastruktur. Wenn Sie die bestehende Infrastruktur erweitern, könnten Sie Brownfield- und Greenfield-Strategien mischen.

Puffer-Cache

Der Speicherbereich, in dem die am häufigsten abgerufenen Daten gespeichert werden.

Geschäftsfähigkeit

Was ein Unternehmen tut, um Wert zu generieren (z. B. Vertrieb, Kundenservice oder Marketing). Microservices-Architekturen und Entwicklungsentscheidungen können von den Geschäftskapazitäten beeinflusst werden. Weitere Informationen finden Sie im Abschnitt Organisiert nach Geschäftskapazitäten des Whitepapers Ausführen von containerisierten Microservices in AWS.

Planung der Geschäftskontinuität (BCP)

Ein Plan, der die potenziellen Auswirkungen eines störenden Ereignisses, wie z. B. einer groß angelegten Migration, auf den Betrieb berücksichtigt und es einem Unternehmen ermöglicht, den Betrieb schnell wieder aufzunehmen.

C

CAF

Weitere Informationen finden Sie unter Framework AWS für die Cloud-Einführung.

Bereitstellung auf Kanaren

Die langsame und schrittweise Veröffentlichung einer Version für Endbenutzer. Wenn Sie sich sicher sind, stellen Sie die neue Version bereit und ersetzen die aktuelle Version vollständig.

CCoE

Weitere Informationen finden Sie im Cloud Center of Excellence.

CDC

Siehe Erfassung von Änderungsdaten.

Erfassung von Datenänderungen (CDC)

Der Prozess der Nachverfolgung von Änderungen an einer Datenquelle, z. B. einer Datenbanktabelle, und der Aufzeichnung von Metadaten zu der Änderung. Sie können CDC für verschiedene Zwecke verwenden, z. B. für die Prüfung oder Replikation von Änderungen in einem Zielsystem, um die Synchronisation aufrechtzuerhalten.

Chaos-Technik

Absichtliches Einführen von Ausfällen oder Störungsereignissen, um die Widerstandsfähigkeit eines Systems zu testen. Sie können AWS Fault Injection Service (AWS FIS) verwenden, um Experimente durchzuführen, die Ihre AWS Workloads stress, und deren Reaktion zu bewerten.

CI/CD

Siehe Continuous Integration und Continuous Delivery.

Klassifizierung

Ein Kategorisierungsprozess, der bei der Erstellung von Vorhersagen hilft. ML-Modelle für Klassifikationsprobleme sagen einen diskreten Wert voraus. Diskrete Werte unterscheiden sich immer voneinander. Beispielsweise muss ein Modell möglicherweise auswerten, ob auf einem Bild ein Auto zu sehen ist oder nicht.

clientseitige Verschlüsselung

Lokale Verschlüsselung von Daten, bevor das Ziel sie AWS-Service empfängt.

Cloud-Exzellenzzentrum (CCoE)

Ein multidisziplinäres Team, das die Cloud-Einführung in der gesamten Organisation vorantreibt, einschließlich der Entwicklung bewährter Cloud-Methoden, der Mobilisierung von Ressourcen, der Festlegung von Migrationszeitplänen und der Begleitung der Organisation durch groß angelegte Transformationen. Weitere Informationen finden Sie in den CCoE-Beiträgen im AWS Cloud Enterprise Strategy Blog.

Cloud Computing

Die Cloud-Technologie, die typischerweise für die Ferndatenspeicherung und das IoT-Gerätemanagement verwendet wird. Cloud Computing ist häufig mit Edge-Computing-Technologie verbunden.

Cloud-Betriebsmodell

In einer IT-Organisation das Betriebsmodell, das zum Aufbau, zur Weiterentwicklung und Optimierung einer oder mehrerer Cloud-Umgebungen verwendet wird. Weitere Informationen finden Sie unter Aufbau Ihres Cloud-Betriebsmodells.

Phasen der Einführung der Cloud

Die vier Phasen, die Unternehmen bei der Migration in der Regel durchlaufen AWS Cloud:

- Projekt – Durchführung einiger Cloud-bezogener Projekte zu Machbarkeitsnachweisen und zu Lernzwecken
- Fundament — Tätigen Sie grundlegende Investitionen, um Ihre Cloud-Einführung zu skalieren (z. B. Einrichtung einer landing zone, Definition eines CCo E, Einrichtung eines Betriebsmodells)

- Migration – Migrieren einzelner Anwendungen
- Neuentwicklung – Optimierung von Produkten und Services und Innovation in der Cloud

Diese Phasen wurden von Stephen Orban im Blogbeitrag The Journey Toward Cloud-First & the Stages of Adoption im AWS Cloud Enterprise Strategy-Blog definiert. Informationen darüber, wie sie mit der AWS Migrationsstrategie zusammenhängen, finden Sie im Leitfaden zur Vorbereitung der Migration.

CMDB

Siehe Datenbank für das Konfigurationsmanagement.

Code-Repository

Ein Ort, an dem Quellcode und andere Komponenten wie Dokumentation, Beispiele und Skripts gespeichert und im Rahmen von Versionskontrollprozessen aktualisiert werden. Zu den gängigen Cloud-Repositorys gehören GitHub oderBitbucket Cloud. Jede Version des Codes wird Zweig genannt. In einer Microservice-Struktur ist jedes Repository einer einzelnen Funktionalität gewidmet. Eine einzelne CI/CD-Pipeline kann mehrere Repositorien verwenden.

Kalter Cache

Ein Puffer-Cache, der leer oder nicht gut gefüllt ist oder veraltete oder irrelevante Daten enthält. Dies beeinträchtigt die Leistung, da die Datenbank-Instance aus dem Hauptspeicher oder der Festplatte lesen muss, was langsamer ist als das Lesen aus dem Puffercache.

Kalte Daten

Daten, auf die selten zugegriffen wird und die in der Regel historisch sind. Bei der Abfrage dieser Art von Daten sind langsame Abfragen in der Regel akzeptabel. Durch die Verlagerung dieser Daten auf leistungsschwächere und kostengünstigere Speicherstufen oder -klassen können Kosten gesenkt werden.

Computer Vision (CV)

Ein Bereich der KI, der maschinelles Lernen nutzt, um Informationen aus visuellen Formaten wie digitalen Bildern und Videos zu analysieren und zu extrahieren. Amazon SageMaker AI bietet beispielsweise Bildverarbeitungsalgorithmen für CV.

Drift in der Konfiguration

Bei einer Arbeitslast eine Änderung der Konfiguration gegenüber dem erwarteten Zustand. Dies kann dazu führen, dass der Workload nicht mehr richtlinienkonform wird, und zwar in der Regel schrittweise und unbeabsichtigt.

Verwaltung der Datenbankkonfiguration (CMDB)

Ein Repository, das Informationen über eine Datenbank und ihre IT-Umgebung speichert und verwaltet, inklusive Hardware- und Softwarekomponenten und deren Konfigurationen. In der Regel verwenden Sie Daten aus einer CMDB in der Phase der Portfolioerkennung und -analyse der Migration.

Konformitätspaket

Eine Sammlung von AWS Config Regeln und Abhilfemaßnahmen, die Sie zusammenstellen können, um Ihre Konformitäts- und Sicherheitsprüfungen individuell anzupassen. Mithilfe einer YAML-Vorlage können Sie ein Conformance Pack als einzelne Entität in einer AWS-Konto AND-Region oder unternehmensweit bereitstellen. Weitere Informationen finden Sie in der Dokumentation unter Conformance Packs. AWS Config

Kontinuierliche Bereitstellung und kontinuierliche Integration (CI/CD)

Der Prozess der Automatisierung der Quell-, Build-, Test-, Staging- und Produktionsphasen des Softwareveröffentlichungsprozesses. CI/CD wird allgemein als Pipeline beschrieben. CI/CD kann Ihnen helfen, Prozesse zu automatisieren, die Produktivität zu steigern, die Codequalität zu verbessern und schneller zu liefern. Weitere Informationen finden Sie unter Vorteile der kontinuierlichen Auslieferung. CD kann auch für kontinuierliche Bereitstellung stehen. Weitere Informationen finden Sie unterKontinuierliche Auslieferung im Vergleich zu kontinuierlicher Bereitstellung.

CV

Siehe Computer Vision.

D

Daten im Ruhezustand

Daten, die in Ihrem Netzwerk stationär sind, z. B. Daten, die sich im Speicher befinden.

Datenklassifizierung

Ein Prozess zur Identifizierung und Kategorisierung der Daten in Ihrem Netzwerk auf der Grundlage ihrer Kritikalität und Sensitivität. Sie ist eine wichtige Komponente jeder Strategie für das Management von Cybersecurity-Risiken, da sie Ihnen hilft, die geeigneten Schutz- und Aufbewahrungskontrollen für die Daten zu bestimmen. Die Datenklassifizierung ist ein Bestandteil

der Sicherheitssäule im AWS Well-Architected Framework. Weitere Informationen finden Sie unter Datenklassifizierung.

Datendrift

Eine signifikante Abweichung zwischen den Produktionsdaten und den Daten, die zum Trainieren eines ML-Modells verwendet wurden, oder eine signifikante Änderung der Eingabedaten im Laufe der Zeit. Datendrift kann die Gesamtqualität, Genauigkeit und Fairness von ML-Modellvorhersagen beeinträchtigen.

Daten während der Übertragung

Daten, die sich aktiv durch Ihr Netzwerk bewegen, z. B. zwischen Netzwerkressourcen.

Datennetz

Ein architektonisches Framework, das verteilte, dezentrale Dateneigentum mit zentraler Verwaltung und Steuerung ermöglicht.

Datenminimierung

Das Prinzip, nur die Daten zu sammeln und zu verarbeiten, die unbedingt erforderlich sind. Durch Datenminimierung im AWS Cloud können Datenschutzrisiken, Kosten und der CO2-Fußabdruck Ihrer Analysen reduziert werden.

Datenperimeter

Eine Reihe präventiver Schutzmaßnahmen in Ihrer AWS Umgebung, die sicherstellen, dass nur vertrauenswürdige Identitäten auf vertrauenswürdige Ressourcen von erwarteten Netzwerken zugreifen. Weitere Informationen finden Sie unter Aufbau eines Datenperimeters auf. AWS

Vorverarbeitung der Daten

Rohdaten in ein Format umzuwandeln, das von Ihrem ML-Modell problemlos verarbeitet werden kann. Die Vorverarbeitung von Daten kann bedeuten, dass bestimmte Spalten oder Zeilen entfernt und fehlende, inkonsistente oder doppelte Werte behoben werden.

Herkunft der Daten

Der Prozess der Nachverfolgung des Ursprungs und der Geschichte von Daten während ihres gesamten Lebenszyklus, z. B. wie die Daten generiert, übertragen und gespeichert wurden.

betroffene Person

Eine Person, deren Daten gesammelt und verarbeitet werden.

Data Warehouse

Ein Datenverwaltungssystem, das Business Intelligence wie Analysen unterstützt. Data Warehouses enthalten in der Regel große Mengen historischer Daten und werden in der Regel für Abfragen und Analysen verwendet.

Datenbankdefinitionssprache (DDL)

Anweisungen oder Befehle zum Erstellen oder Ändern der Struktur von Tabellen und Objekten in einer Datenbank.

Datenbankmanipulationssprache (DML)

Anweisungen oder Befehle zum Ändern (Einfügen, Aktualisieren und Löschen) von Informationen in einer Datenbank.

DDL

Siehe Datenbankdefinitionssprache.

Deep-Ensemble

Mehrere Deep-Learning-Modelle zur Vorhersage kombinieren. Sie können Deep-Ensembles verwenden, um eine genauere Vorhersage zu erhalten oder um die Unsicherheit von Vorhersagen abzuschätzen.

Deep Learning

Ein ML-Teilbereich, der mehrere Schichten künstlicher neuronaler Netzwerke verwendet, um die Zuordnung zwischen Eingabedaten und Zielvariablen von Interesse zu ermitteln.

defense-in-depth

Ein Ansatz zur Informationssicherheit, bei dem eine Reihe von Sicherheitsmechanismen und -kontrollen sorgfältig in einem Computernetzwerk verteilt werden, um die Vertraulichkeit, Integrität und Verfügbarkeit des Netzwerks und der darin enthaltenen Daten zu schützen. Wenn Sie diese Strategie anwenden AWS, fügen Sie mehrere Steuerelemente auf verschiedenen Ebenen der AWS Organizations Struktur hinzu, um die Ressourcen zu schützen. Ein defense-in-depth Ansatz könnte beispielsweise Multi-Faktor-Authentifizierung, Netzwerksegmentierung und Verschlüsselung kombinieren.

delegierter Administrator

In AWS Organizations kann ein kompatibler Dienst ein AWS Mitgliedskonto registrieren, um die Konten der Organisation und die Berechtigungen für diesen Dienst zu verwalten. Dieses Konto

wird als delegierter Administrator für diesen Service bezeichnet. Weitere Informationen und eine Liste kompatibler Services finden Sie unter Services, die mit AWS Organizations funktionieren in der AWS Organizations -Dokumentation.

Einsatz

Der Prozess, bei dem eine Anwendung, neue Feature oder Codekorrekturen in der Zielumgebung verfügbar gemacht werden. Die Bereitstellung umfasst das Implementieren von Änderungen an einer Codebasis und das anschließende Erstellen und Ausführen dieser Codebasis in den Anwendungsumgebungen.

Entwicklungsumgebung

Siehe Umgebung.

Detektivische Kontrolle

Eine Sicherheitskontrolle, die darauf ausgelegt ist, ein Ereignis zu erkennen, zu protokollieren und zu warnen, nachdem ein Ereignis eingetreten ist. Diese Kontrollen stellen eine zweite Verteidigungslinie dar und warnen Sie vor Sicherheitsereignissen, bei denen die vorhandenen präventiven Kontrollen umgangen wurden. Weitere Informationen finden Sie unter Detektivische Kontrolle in Implementierung von Sicherheitskontrollen in AWS.

Abbildung des Wertstroms in der Entwicklung (DVSM)

Ein Prozess zur Identifizierung und Priorisierung von Einschränkungen, die sich negativ auf Geschwindigkeit und Qualität im Lebenszyklus der Softwareentwicklung auswirken. DVSM erweitert den Prozess der Wertstromanalyse, der ursprünglich für Lean-Manufacturing-Praktiken konzipiert wurde. Es konzentriert sich auf die Schritte und Teams, die erforderlich sind, um durch den Softwareentwicklungsprozess Mehrwert zu schaffen und zu steigern.

digitaler Zwilling

Eine virtuelle Darstellung eines realen Systems, z. B. eines Gebäudes, einer Fabrik, einer Industrieanlage oder einer Produktionslinie. Digitale Zwillinge unterstützen vorausschauende Wartung, Fernüberwachung und Produktionsoptimierung.

Maßtabelle

In einem Sternschema eine kleinere Tabelle, die Datenattribute zu quantitativen Daten in einer Faktentabelle enthält. Bei Attributen von Dimensionstabellen handelt es sich in der Regel um Textfelder oder diskrete Zahlen, die sich wie Text verhalten. Diese Attribute werden häufig zum Einschränken von Abfragen, zum Filtern und zur Kennzeichnung von Ergebnismengen verwendet.

Katastrophe

Ein Ereignis, das verhindert, dass ein Workload oder ein System seine Geschäftsziele an seinem primären Einsatzort erfüllt. Diese Ereignisse können Naturkatastrophen, technische Ausfälle oder das Ergebnis menschlichen Handelns sein, z. B. unbeabsichtigte Fehlkonfigurationen oder ein Malware-Angriff.

Notfallwiederherstellung (DR)

Die Strategie und der Prozess, mit denen Sie Ausfallzeiten und Datenverluste aufgrund einer Katastrophe minimieren. Weitere Informationen finden Sie unter Disaster Recovery von Workloads unter AWS: Wiederherstellung in der Cloud im AWS Well-Architected Framework.

DML

Siehe Sprache zur Datenbankmanipulation.

Domainorientiertes Design

Ein Ansatz zur Entwicklung eines komplexen Softwaresystems, bei dem seine Komponenten mit sich entwickelnden Domains oder Kerngeschäftszielen verknüpft werden, denen jede Komponente dient. Dieses Konzept wurde von Eric Evans in seinem Buch Domaingesteuertes Design: Bewältigen der Komplexität im Herzen der Software (Boston: Addison-Wesley Professional, 2003) vorgestellt. Informationen darüber, wie Sie domaingesteuertes Design mit dem Strangler-Fig-Muster verwenden können, finden Sie unter Schrittweises Modernisieren älterer Microsoft ASP.NET (ASMX)-Webservices mithilfe von Containern und Amazon API Gateway.

DR

Siehe Disaster Recovery.

Erkennung von Driften

Verfolgung von Abweichungen von einer Basiskonfiguration. Sie können es beispielsweise verwenden, AWS CloudFormation um Abweichungen bei den Systemressourcen zu erkennen, oder Sie können AWS Control Tower damit Änderungen in Ihrer landing zone erkennen, die sich auf die Einhaltung von Governance-Anforderungen auswirken könnten.

DVSM

Siehe Abbildung des Wertstroms in der Entwicklung.

E

EDA

Siehe explorative Datenanalyse.

EDI

Siehe elektronischer Datenaustausch.

Edge-Computing

Die Technologie, die die Rechenleistung für intelligente Geräte an den Rändern eines IoT-Netzwerks erhöht. Im Vergleich zu Cloud Computing kann Edge Computing die Kommunikationslatenz reduzieren und die Reaktionszeit verbessern.

elektronischer Datenaustausch (EDI)

Der automatisierte Austausch von Geschäftsdokumenten zwischen Organisationen. Weitere Informationen finden Sie unter Was ist elektronischer Datenaustausch.

Verschlüsselung

Ein Rechenprozess, der Klartextdaten, die für Menschen lesbar sind, in Chiffretext umwandelt.

Verschlüsselungsschlüssel

Eine kryptografische Zeichenfolge aus zufälligen Bits, die von einem Verschlüsselungsalgorithmus generiert wird. Schlüssel können unterschiedlich lang sein, und jeder Schlüssel ist so konzipiert, dass er unvorhersehbar und einzigartig ist.

Endianismus

Die Reihenfolge, in der Bytes im Computerspeicher gespeichert werden. Big-Endian-Systeme speichern das höchstwertige Byte zuerst. Little-Endian-Systeme speichern das niedrigwertigste Byte zuerst.

Endpunkt

Siehe Service-Endpunkt.

Endpunkt-Services

Ein Service, den Sie in einer Virtual Private Cloud (VPC) hosten können, um ihn mit anderen Benutzern zu teilen. Sie können einen Endpunktdienst mit anderen AWS-Konten oder AWS Identity and Access Management (IAM AWS PrivateLink -) Prinzipalen erstellen und diesen

Berechtigungen gewähren. Diese Konten oder Prinzipale können sich privat mit Ihrem Endpunktservice verbinden, indem sie Schnittstellen-VPC-Endpunkte erstellen. Weitere Informationen finden Sie unter Einen Endpunkt-Service erstellen in der Amazon Virtual Private Cloud (Amazon VPC)-Dokumentation.

Unternehmensressourcenplanung (ERP)

Ein System, das wichtige Geschäftsprozesse (wie Buchhaltung, MES und Projektmanagement) für ein Unternehmen automatisiert und verwaltet.

Envelope-Verschlüsselung

Der Prozess der Verschlüsselung eines Verschlüsselungsschlüssels mit einem anderen Verschlüsselungsschlüssel. Weitere Informationen finden Sie unter Envelope-Verschlüsselung in der AWS Key Management Service (AWS KMS) -Dokumentation.

Umgebung

Eine Instance einer laufenden Anwendung. Die folgenden Arten von Umgebungen sind beim Cloud-Computing üblich:

- Entwicklungsumgebung – Eine Instance einer laufenden Anwendung, die nur dem Kernteam zur Verfügung steht, das für die Wartung der Anwendung verantwortlich ist. Entwicklungsumgebungen werden verwendet, um Änderungen zu testen, bevor sie in höhere Umgebungen übertragen werden. Diese Art von Umgebung wird manchmal als Testumgebung bezeichnet.
- Niedrigere Umgebungen – Alle Entwicklungsumgebungen für eine Anwendung, z. B. solche, die für erste Builds und Tests verwendet wurden.
- Produktionsumgebung – Eine Instance einer laufenden Anwendung, auf die Endbenutzer zugreifen können. In einer CI/CD Pipeline ist die Produktionsumgebung die letzte Bereitstellungsumgebung.
- Höhere Umgebungen – Alle Umgebungen, auf die auch andere Benutzer als das Kernentwicklungsteam zugreifen können. Dies kann eine Produktionsumgebung, Vorproduktionsumgebungen und Umgebungen für Benutzerakzeptanztests umfassen.

Epics

In der agilen Methodik sind dies funktionale Kategorien, die Ihnen helfen, Ihre Arbeit zu organisieren und zu priorisieren. Epics bieten eine allgemeine Beschreibung der Anforderungen und Implementierungsaufgaben. Zu den Sicherheitsepen AWS von CAF gehören beispielsweise Identitäts- und Zugriffsmanagement, Detektivkontrollen, Infrastruktursicherheit, Datenschutz und

Reaktion auf Vorfälle. Weitere Informationen zu Epics in der AWS -Migrationsstrategie finden Sie im Leitfaden zur Programm-Implementierung.

ERP

Siehe Enterprise Resource Planning.

Explorative Datenanalyse (EDA)

Der Prozess der Analyse eines Datensatzes, um seine Hauptmerkmale zu verstehen. Sie sammeln oder aggregieren Daten und führen dann erste Untersuchungen durch, um Muster zu finden, Anomalien zu erkennen und Annahmen zu überprüfen. EDA wird durchgeführt, indem zusammenfassende Statistiken berechnet und Datenvisualisierungen erstellt werden.

F

Faktentabelle

Die zentrale Tabelle in einem Sternschema. Sie speichert quantitative Daten über den Geschäftsbetrieb. In der Regel enthält eine Faktentabelle zwei Arten von Spalten: Spalten, die Kennzahlen enthalten, und Spalten, die einen Fremdschlüssel für eine Dimensionstabelle enthalten.

schnell scheitern

Eine Philosophie, die häufige und inkrementelle Tests verwendet, um den Entwicklungslebenszyklus zu verkürzen. Dies ist ein wichtiger Bestandteil eines agilen Ansatzes.

Grenze zur Fehlerisolierung

Dabei handelt es sich um eine Grenze AWS Cloud, z. B. eine Availability Zone AWS-Region, eine Steuerungsebene oder eine Datenebene, die die Auswirkungen eines Fehlers begrenzt und die Widerstandsfähigkeit von Workloads verbessert. Weitere Informationen finden Sie unter Grenzen zur AWS Fehlerisolierung.

Feature-Zweig

Siehe Zweig.

Features

Die Eingabedaten, die Sie verwenden, um eine Vorhersage zu treffen. In einem Fertigungskontext könnten Feature beispielsweise Bilder sein, die regelmäßig von der Fertigungslinie aus aufgenommen werden.

Bedeutung der Feature

Wie wichtig ein Feature für die Vorhersagen eines Modells ist. Dies wird in der Regel als numerischer Wert ausgedrückt, der mit verschiedenen Techniken wie Shapley Additive Explanations (SHAP) und integrierten Gradienten berechnet werden kann. Weitere Informationen finden Sie unter Interpretierbarkeit von Modellen für maschinelles Lernen mit AWS.

Featuretransformation

Daten für den ML-Prozess optimieren, einschließlich der Anreicherung von Daten mit zusätzlichen Quellen, der Skalierung von Werten oder der Extraktion mehrerer Informationssätze aus einem einzigen Datenfeld. Das ermöglicht dem ML-Modell, von den Daten profitieren. Wenn Sie beispielsweise das Datum „27.05.2021 00:15:37“ in „2021“, „Mai“, „Donnerstag“ und „15“ aufschlüsseln, können Sie dem Lernalgorithmus helfen, nuancierte Muster zu erlernen, die mit verschiedenen Datenkomponenten verknüpft sind.

Eingabeaufforderung mit wenigen Klicks

Bereitstellung einer kleinen Anzahl von Beispielen, die die Aufgabe und das gewünschte Ergebnis veranschaulichen, bevor das LLM aufgefordert wird, eine ähnliche Aufgabe auszuführen. Bei dieser Technik handelt es sich um eine Anwendung des kontextbezogenen Lernens, bei der Modelle anhand von Beispielen (Aufnahmen) lernen, die in Eingabeaufforderungen eingebettet sind. Bei Aufgaben, die spezifische Formatierungs-, Argumentations- oder Fachkenntnisse erfordern, kann die Eingabeaufforderung mit wenigen Handgriffen effektiv sein. Siehe auch Zero-Shot Prompting.

FGAC

Siehe detaillierte Zugriffskontrolle.

Feinkörnige Zugriffskontrolle (FGAC)

Die Verwendung mehrerer Bedingungen, um eine Zugriffsanfrage zuzulassen oder abzulehnen.

Flash-Cut-Migration

Eine Datenbankmigrationsmethode, bei der eine kontinuierliche Datenreplikation durch Erfassung von Änderungsdaten verwendet wird, um Daten in kürzester Zeit zu migrieren, anstatt einen schrittweisen Ansatz zu verwenden. Ziel ist es, Ausfallzeiten auf ein Minimum zu beschränken.

FM

Siehe Fundamentmodell.

Fundamentmodell (FM)

Ein großes neuronales Deep-Learning-Netzwerk, das mit riesigen Datensätzen generalisierter und unbeschrifteter Daten trainiert wurde. FMs sind in der Lage, eine Vielzahl allgemeiner Aufgaben zu erfüllen, z. B. Sprache zu verstehen, Text und Bilder zu generieren und Konversationen in natürlicher Sprache zu führen. Weitere Informationen finden Sie unter Was sind Foundation-Modelle.

G

Generative KI

Eine Untergruppe von KI-Modellen, die mit großen Datenmengen trainiert wurden und mit einer einfachen Textaufforderung neue Inhalte und Artefakte wie Bilder, Videos, Text und Audio erstellen können. Weitere Informationen finden Sie unter Was ist Generative KI.

Geoblocking

Siehe geografische Einschränkungen.

Geografische Einschränkungen (Geoblocking)

Bei Amazon eine Option CloudFront, um zu verhindern, dass Benutzer in bestimmten Ländern auf Inhaltsverteilungen zugreifen. Sie können eine Zulassungsliste oder eine Sperrliste verwenden, um zugelassene und gesperrte Länder anzugeben. Weitere Informationen finden Sie in der Dokumentation unter Beschränkung der geografischen Verteilung Ihrer Inhalte. CloudFront

Gitflow-Workflow

Ein Ansatz, bei dem niedrigere und höhere Umgebungen unterschiedliche Zweige in einem Quellcode-Repository verwenden. Der Gitflow-Workflow gilt als veraltet, und der Trunk-basierte Workflow ist der moderne, bevorzugte Ansatz.

goldenes Bild

Ein Snapshot eines Systems oder einer Software, der als Vorlage für die Bereitstellung neuer Instanzen dieses Systems oder dieser Software verwendet wird. In der Fertigung kann ein Golden Image beispielsweise zur Bereitstellung von Software auf mehreren Geräten verwendet werden und trägt zur Verbesserung der Geschwindigkeit, Skalierbarkeit und Produktivität bei der Geräteherstellung bei.

Greenfield-Strategie

Das Fehlen vorhandener Infrastruktur in einer neuen Umgebung. Bei der Einführung einer Neuausrichtung einer Systemarchitektur können Sie alle neuen Technologien ohne Einschränkung der Kompatibilität mit der vorhandenen Infrastruktur auswählen, auch bekannt als Brownfield. Wenn Sie die bestehende Infrastruktur erweitern, könnten Sie Brownfield- und Greenfield-Strategien mischen.

Integritätsschutz

Eine allgemeine Regel, die dazu beiträgt, Ressourcen, Richtlinien und die Einhaltung von Vorschriften in allen Unternehmenseinheiten zu regeln (OUs). Präventiver Integritätsschutz setzt Richtlinien durch, um die Einhaltung von Standards zu gewährleisten. Sie werden mithilfe von Service-Kontrollrichtlinien und IAM-Berechtigungsgrenzen implementiert. Detektivischer Integritätsschutz erkennt Richtlinienverstöße und Compliance-Probleme und generiert Warnmeldungen zur Abhilfe. Sie werden mithilfe von AWS Config, AWS Security Hub CSPM, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector und benutzerdefinierten AWS Lambda Prüfungen implementiert.

H

HEKTAR

Siehe Hochverfügbarkeit.

Heterogene Datenbankmigration

Migrieren Sie Ihre Quelldatenbank in eine Zieldatenbank, die eine andere Datenbank-Engine verwendet (z. B. Oracle zu Amazon Aurora). Eine heterogene Migration ist in der Regel Teil einer Neuarchitektur, und die Konvertierung des Schemas kann eine komplexe Aufgabe sein. AWS bietet AWS SCT, welches bei Schemakonvertierungen hilft.

hohe Verfügbarkeit (HA)

Die Fähigkeit eines Workloads, im Falle von Herausforderungen oder Katastrophen kontinuierlich und ohne Eingreifen zu arbeiten. HA-Systeme sind so konzipiert, dass sie automatisch ein Failover durchführen, gleichbleibend hohe Leistung bieten und unterschiedliche Lasten und Ausfälle mit minimalen Leistungseinbußen bewältigen.

historische Modernisierung

Ein Ansatz zur Modernisierung und Aufrüstung von Betriebstechnologiesystemen (OT), um den Bedürfnissen der Fertigungsindustrie besser gerecht zu werden. Ein Historian ist eine Art von Datenbank, die verwendet wird, um Daten aus verschiedenen Quellen in einer Fabrik zu sammeln und zu speichern.

Daten zurückhalten

Ein Teil historischer, beschrifteter Daten, der aus einem Datensatz zurückgehalten wird, der zum Trainieren eines Modells für maschinelles Lernen verwendet wird. Sie können Holdout-Daten verwenden, um die Modellleistung zu bewerten, indem Sie die Modellvorhersagen mit den Holdout-Daten vergleichen.

Homogene Datenbankmigration

Migrieren Sie Ihre Quelldatenbank zu einer Zieldatenbank, die dieselbe Datenbank-Engine verwendet (z. B. Microsoft SQL Server zu Amazon RDS für SQL Server). Eine homogene Migration ist in der Regel Teil eines Hostwechsels oder eines Plattformwechsels. Sie können native Datenbankserviceprogramme verwenden, um das Schema zu migrieren.

heiße Daten

Daten, auf die häufig zugegriffen wird, z. B. Echtzeitdaten oder aktuelle Translationsdaten. Für diese Daten ist in der Regel eine leistungsstarke Speicherebene oder -klasse erforderlich, um schnelle Abfrageantworten zu ermöglichen.

Hotfix

Eine dringende Lösung für ein kritisches Problem in einer Produktionsumgebung. Aufgrund seiner Dringlichkeit wird ein Hotfix normalerweise außerhalb des typischen DevOps Release-Workflows erstellt.

Hypercare-Phase

Unmittelbar nach dem Cutover, der Zeitraum, in dem ein Migrationsteam die migrierten Anwendungen in der Cloud verwaltet und überwacht, um etwaige Probleme zu beheben. In der Regel dauert dieser Zeitraum 1–4 Tage. Am Ende der Hypercare-Phase überträgt das Migrationsteam in der Regel die Verantwortung für die Anwendungen an das Cloud-Betriebsteam.

I

IaC

Sehen Sie Infrastruktur als Code.

Identitätsbasierte Richtlinie

Eine Richtlinie, die einem oder mehreren IAM-Prinzipalen zugeordnet ist und deren Berechtigungen innerhalb der AWS Cloud Umgebung definiert.

Leerlaufanwendung

Eine Anwendung mit einer durchschnittlichen CPU- und Arbeitsspeicherauslastung zwischen 5 und 20 Prozent über einen Zeitraum von 90 Tagen. In einem Migrationsprojekt ist es üblich, diese Anwendungen außer Betrieb zu nehmen oder sie On-Premises beizubehalten.

IIoT

Siehe Industrielles Internet der Dinge.

unveränderliche Infrastruktur

Ein Modell, das eine neue Infrastruktur für Produktionsworkloads bereitstellt, anstatt die bestehende Infrastruktur zu aktualisieren, zu patchen oder zu modifizieren. Unveränderliche Infrastrukturen sind von Natur aus konsistenter, zuverlässiger und vorhersehbarer als veränderliche Infrastrukturen. Weitere Informationen finden Sie in der Best Practice Deploy using immutable infrastructure im AWS Well-Architected Framework.

Eingehende (ingress) VPC

In einer Architektur AWS mit mehreren Konten ist dies eine VPC, die Netzwerkverbindungen von außerhalb einer Anwendung akzeptiert, überprüft und weiterleitet. Die AWS Security Reference Architecture empfiehlt, Ihr Netzwerkkonto mit eingehendem und ausgehendem Datenverkehr und Inspektion einzurichten, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

Inkrementelle Migration

Eine Cutover-Strategie, bei der Sie Ihre Anwendung in kleinen Teilen migrieren, anstatt eine einziges vollständiges Cutover durchzuführen. Beispielsweise könnten Sie zunächst nur einige Microservices oder Benutzer auf das neue System umstellen. Nachdem Sie sich vergewissert haben, dass alles ordnungsgemäß funktioniert, können Sie weitere Microservices oder Benutzer

schrittweise verschieben, bis Sie Ihr Legacy-System außer Betrieb nehmen können. Diese Strategie reduziert die mit großen Migrationen verbundenen Risiken.

Industrie 4.0

Ein Begriff, der 2016 von Klaus Schwab eingeführt wurde und sich auf die Modernisierung von Fertigungsprozessen durch Fortschritte in den Bereichen Konnektivität, Echtzeitdaten, Automatisierung, Analytik und KI/ML bezieht.

Infrastruktur

Alle Ressourcen und Komponenten, die in der Umgebung einer Anwendung enthalten sind.

Infrastructure as Code (IaC)

Der Prozess der Bereitstellung und Verwaltung der Infrastruktur einer Anwendung mithilfe einer Reihe von Konfigurationsdateien. IaC soll Ihnen helfen, das Infrastrukturmanagement zu zentralisieren, Ressourcen zu standardisieren und schnell zu skalieren, sodass neue Umgebungen wiederholbar, zuverlässig und konsistent sind.

industrielles Internet der Dinge (T) IIo

Einsatz von mit dem Internet verbundenen Sensoren und Geräten in Industriesektoren wie Fertigung, Energie, Automobilindustrie, Gesundheitswesen, Biowissenschaften und Landwirtschaft. Weitere Informationen finden Sie unter Aufbau einer digitalen Transformationsstrategie für das industrielle Internet der Dinge (IIoT).

Inspektions-VPC

In einer Architektur AWS mit mehreren Konten eine zentralisierte VPC, die Inspektionen des Netzwerkverkehrs zwischen VPCs (in demselben oder unterschiedlichen AWS-Regionen), dem Internet und lokalen Netzwerken verwaltet. In der AWS Security Reference Architecture wird empfohlen, Ihr Netzwerkkonto mit eingehendem und ausgehendem Datenverkehr sowie Inspektionen einzurichten, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

Internet of Things (IoT)

Das Netzwerk verbundener physischer Objekte mit eingebetteten Sensoren oder Prozessoren, das über das Internet oder über ein lokales Kommunikationsnetzwerk mit anderen Geräten und Systemen kommuniziert. Weitere Informationen finden Sie unter Was ist IoT?

Interpretierbarkeit

Ein Merkmal eines Modells für Machine Learning, das beschreibt, inwieweit ein Mensch verstehen kann, wie die Vorhersagen des Modells von seinen Eingaben abhängen. Weitere Informationen finden Sie unter Interpretierbarkeit des Modells für maschinelles Lernen mit. AWS

IoT

Siehe Internet der Dinge.

IT information library (ITIL, IT-Informationsbibliothek)

Eine Reihe von bewährten Methoden für die Bereitstellung von IT-Services und die Abstimmung dieser Services auf die Geschäftsanforderungen. ITIL bietet die Grundlage für ITSM.

T service management (ITSM, IT-Servicemanagement)

Aktivitäten im Zusammenhang mit der Gestaltung, Implementierung, Verwaltung und Unterstützung von IT-Services für eine Organisation. Informationen zur Integration von Cloud-Vorgängen mit ITSM-Tools finden Sie im Leitfaden zur Betriebsintegration.

BIS

Siehe IT-Informationsbibliothek.

ITSM

Siehe IT-Servicemanagement.

L

Labelbasierte Zugangskontrolle (LBAC)

Eine Implementierung der Mandatory Access Control (MAC), bei der den Benutzern und den Daten selbst jeweils explizit ein Sicherheitslabelwert zugewiesen wird. Die Schnittmenge zwischen der Benutzersicherheitsbeschriftung und der Datensicherheitsbeschriftung bestimmt, welche Zeilen und Spalten für den Benutzer sichtbar sind.

Landing Zone

Eine landing zone ist eine gut strukturierte AWS Umgebung mit mehreren Konten, die skalierbar und sicher ist. Dies ist ein Ausgangspunkt, von dem aus Ihre Organisationen Workloads und Anwendungen schnell und mit Vertrauen in ihre Sicherheits- und Infrastrukturumgebung starten

und bereitstellen können. Weitere Informationen zu Landing Zones finden Sie unter Einrichtung einer sicheren und skalierbaren AWS -Umgebung mit mehreren Konten..

großes Sprachmodell (LLM)

Ein Deep-Learning-KI-Modell, das anhand einer riesigen Datenmenge vorab trainiert wurde. Ein LLM kann mehrere Aufgaben ausführen, z. B. Fragen beantworten, Dokumente zusammenfassen, Text in andere Sprachen übersetzen und Sätze vervollständigen. Weitere Informationen finden Sie unter Was sind. LLMs

Große Migration

Eine Migration von 300 oder mehr Servern.

SCHWARZ

Siehe Labelbasierte Zugriffskontrolle.

Geringste Berechtigung

Die bewährte Sicherheitsmethode, bei der nur die für die Durchführung einer Aufgabe erforderlichen Mindestberechtigungen erteilt werden. Weitere Informationen finden Sie unter Geringste Berechtigungen anwenden in der IAM-Dokumentation.

Lift and Shift

Siehe 7 Rs.

Little-Endian-System

Ein System, welches das niedrigwertigste Byte zuerst speichert. Siehe auch Endianness.

LLM

Siehe großes Sprachmodell.

Niedrigere Umgebungen

Siehe Umgebung.

M

Machine Learning (ML)

Eine Art künstlicher Intelligenz, die Algorithmen und Techniken zur Mustererkennung und zum Lernen verwendet. ML analysiert aufgezeichnete Daten, wie z. B. Daten aus dem Internet der

Dinge (IoT), und lernt daraus, um ein statistisches Modell auf der Grundlage von Mustern zu erstellen. Weitere Informationen finden Sie unter Machine Learning.

Hauptzweig

Siehe Filiale.

Malware

Software, die entwickelt wurde, um die Computersicherheit oder den Datenschutz zu gefährden. Malware kann Computersysteme stören, vertrauliche Informationen durchsickern lassen oder sich unbefugten Zugriff verschaffen. Beispiele für Malware sind Viren, Würmer, Ransomware, Trojaner, Spyware und Keylogger.

verwaltete Dienste

AWS-Services für die die Infrastrukturebene, das Betriebssystem und die Plattformen AWS betrieben werden, und Sie greifen auf die Endgeräte zu, um Daten zu speichern und abzurufen. Amazon Simple Storage Service (Amazon S3) und Amazon DynamoDB sind Beispiele für Managed Services. Diese werden auch als abstrakte Dienste bezeichnet.

Manufacturing Execution System (MES)

Ein Softwaresystem zur Verfolgung, Überwachung, Dokumentation und Steuerung von Produktionsprozessen, bei denen Rohstoffe in der Fertigung zu fertigen Produkten umgewandelt werden.

MAP

Siehe Migration Acceleration Program.

Mechanismus

Ein vollständiger Prozess, bei dem Sie ein Tool erstellen, die Akzeptanz des Tools vorantreiben und anschließend die Ergebnisse überprüfen, um Anpassungen vorzunehmen. Ein Mechanismus ist ein Zyklus, der sich im Laufe seiner Tätigkeit selbst verstärkt und verbessert. Weitere Informationen finden Sie unter Aufbau von Mechanismen im AWS Well-Architected Framework.

Mitgliedskonto

Alle AWS-Konten außer dem Verwaltungskonto, die Teil einer Organisation in sind. AWS Organizations Ein Konto kann jeweils nur Mitglied einer Organisation sein.

MES

Siehe Manufacturing Execution System.

Message Queuing-Telemetrietransport (MQTT)

Ein leichtes machine-to-machine (M2M) -Kommunikationsprotokoll, das auf dem Publish/Subscribe-Muster für IoT-Geräte mit beschränkten Ressourcen basiert.

Microservice

Ein kleiner, unabhängiger Dienst, der über genau definierte Kanäle kommuniziert APIs und in der Regel kleinen, eigenständigen Teams gehört. Ein Versicherungssystem kann beispielsweise Microservices beinhalten, die Geschäftsfunktionen wie Vertrieb oder Marketing oder Subdomains wie Einkauf, Schadenersatz oder Analytik zugeordnet sind. Zu den Vorteilen von Microservices gehören Agilität, flexible Skalierung, einfache Bereitstellung, wiederverwendbarer Code und Ausfallsicherheit. Weitere Informationen finden Sie unter Integration von Microservices mithilfe serverloser Dienste. AWS

Microservices-Architekturen

Ein Ansatz zur Erstellung einer Anwendung mit unabhängigen Komponenten, die jeden Anwendungsprozess als Microservice ausführen. Diese Microservices kommunizieren mithilfe von Lightweight über eine klar definierte Schnittstelle. APIs Jeder Microservice in dieser Architektur kann aktualisiert, bereitgestellt und skaliert werden, um den Bedarf an bestimmten Funktionen einer Anwendung zu decken. Weitere Informationen finden Sie unter Implementierung von Microservices auf. AWS

Migration Acceleration Program (MAP)

Ein AWS Programm, das Beratung, Unterstützung, Schulungen und Services bietet, um Unternehmen dabei zu unterstützen, eine solide betriebliche Grundlage für die Umstellung auf die Cloud zu schaffen und die anfänglichen Kosten von Migrationen auszugleichen. MAP umfasst eine Migrationsmethode für die methodische Durchführung von Legacy-Migrationen sowie eine Reihe von Tools zur Automatisierung und Beschleunigung gängiger Migrationsszenarien.

Migration in großem Maßstab

Der Prozess, bei dem der Großteil des Anwendungsportfolios in Wellen in die Cloud verlagert wird, wobei in jeder Welle mehr Anwendungen schneller migriert werden. In dieser Phase werden die bewährten Verfahren und Erkenntnisse aus den früheren Phasen zur Implementierung einer Migrationsfabrik von Teams, Tools und Prozessen zur Optimierung der Migration von Workloads durch Automatisierung und agile Bereitstellung verwendet. Dies ist die dritte Phase der AWS -Migrationsstrategie.

Migrationsfabrik

Funktionsübergreifende Teams, die die Migration von Workloads durch automatisierte, agile Ansätze optimieren. Zu den Teams in der Migrationsabteilung gehören in der Regel Betriebsabläufe, Geschäftsanalysten und Eigentümer, Migrationsingenieure, Entwickler und DevOps Experten, die in Sprints arbeiten. Zwischen 20 und 50 Prozent eines Unternehmensanwendungsportfolios bestehen aus sich wiederholenden Mustern, die durch einen Fabrik-Ansatz optimiert werden können. Weitere Informationen finden Sie in Diskussion über Migrationsfabriken und den Leitfaden zur Cloud-Migration-Fabrik in diesem Inhaltssatz.

Migrationsmetadaten

Die Informationen über die Anwendung und den Server, die für den Abschluss der Migration benötigt werden. Für jedes Migrationsmuster ist ein anderer Satz von Migrationsmetadaten erforderlich. Beispiele für Migrationsmetadaten sind das Zielsubnetz, die Sicherheitsgruppe und AWS das Konto.

Migrationsmuster

Eine wiederholbare Migrationsaufgabe, in der die Migrationsstrategie, das Migrationsziel und die verwendete Migrationsanwendung oder der verwendete Migrationsservice detailliert beschrieben werden. Beispiel: Rehost-Migration zu Amazon EC2 mit AWS Application Migration Service.

Migration Portfolio Assessment (MPA)

Ein Online-Tool, das Informationen zur Validierung des Geschäftsszenarios für die Migration auf das bereitstellt. AWS Cloud MPA bietet eine detaillierte Portfoliobewertung (richtige Servergröße, Preisgestaltung, Gesamtbetriebskostenanalyse, Migrationskostenanalyse) sowie Migrationsplanung (Anwendungsdatenanalyse und Datenerfassung, Anwendungsgruppierung, Migrationspriorisierung und Wellenplanung). Das MPA-Tool (Anmeldung erforderlich) steht allen AWS Beratern und APN-Partnerberatern kostenlos zur Verfügung.

Migration Readiness Assessment (MRA)

Der Prozess, bei dem mithilfe des AWS CAF Erkenntnisse über den Cloud-Bereitschaftsstatus eines Unternehmens gewonnen, Stärken und Schwächen identifiziert und ein Aktionsplan zur Schließung festgestellter Lücken erstellt wird. Weitere Informationen finden Sie im Benutzerhandbuch für Migration Readiness. MRA ist die erste Phase der AWS - Migrationsstrategie.

Migrationsstrategie

Der Ansatz, der verwendet wurde, um einen Workload auf den AWS Cloud zu migrieren. Weitere Informationen finden Sie im Eintrag 7 Rs in diesem Glossar und unter Mobilisieren Sie Ihr Unternehmen, um umfangreiche Migrationen zu beschleunigen.

ML

Siehe maschinelles Lernen.

Modernisierung

Umwandlung einer veralteten (veralteten oder monolithischen) Anwendung und ihrer Infrastruktur in ein agiles, elastisches und hochverfügbares System in der Cloud, um Kosten zu senken, die Effizienz zu steigern und Innovationen zu nutzen. Weitere Informationen finden Sie unter Strategie zur Modernisierung von Anwendungen in der AWS Cloud.

Bewertung der Modernisierungsfähigkeit

Eine Bewertung, anhand derer festgestellt werden kann, ob die Anwendungen einer Organisation für die Modernisierung bereit sind, Vorteile, Risiken und Abhängigkeiten identifiziert und ermittelt wird, wie gut die Organisation den zukünftigen Status dieser Anwendungen unterstützen kann. Das Ergebnis der Bewertung ist eine Vorlage der Zielarchitektur, eine Roadmap, in der die Entwicklungsphasen und Meilensteine des Modernisierungsprozesses detailliert beschrieben werden, sowie ein Aktionsplan zur Behebung festgestellter Lücken. Weitere Informationen finden Sie unter Evaluierung der Modernisierungsbereitschaft von Anwendungen in der AWS Cloud.

Monolithische Anwendungen (Monolithen)

Anwendungen, die als ein einziger Service mit eng gekoppelten Prozessen ausgeführt werden. Monolithische Anwendungen haben verschiedene Nachteile. Wenn ein Anwendungs-Feature stark nachgefragt wird, muss die gesamte Architektur skaliert werden. Das Hinzufügen oder Verbessern der Feature einer monolithischen Anwendung wird ebenfalls komplexer, wenn die Codebasis wächst. Um diese Probleme zu beheben, können Sie eine Microservices-Architektur verwenden. Weitere Informationen finden Sie unter Zerlegen von Monolithen in Microservices.

MPA

Siehe Bewertung des Migrationsportfolios.

MQTT

Siehe Message Queuing-Telemetrietransport.

Mehrklassen-Klassifizierung

Ein Prozess, der dabei hilft, Vorhersagen für mehrere Klassen zu generieren (wobei eines von mehr als zwei Ergebnissen vorhergesagt wird). Ein ML-Modell könnte beispielsweise fragen: „Ist dieses Produkt ein Buch, ein Auto oder ein Telefon?“ oder „Welche Kategorie von Produkten ist für diesen Kunden am interessantesten?“

veränderbare Infrastruktur

Ein Modell, das die bestehende Infrastruktur für Produktionsworkloads aktualisiert und modifiziert. Für eine verbesserte Konsistenz, Zuverlässigkeit und Vorhersagbarkeit empfiehlt das AWS Well-Architected Framework die Verwendung einer unveränderlichen Infrastruktur als bewährte Methode.

O

OAC

Siehe Origin Access Control.

EICHE

Siehe Zugriffsidentität von Origin.

COM

Siehe organisatorisches Change-Management.

Offline-Migration

Eine Migrationsmethode, bei der der Quell-Workload während des Migrationsprozesses heruntergefahren wird. Diese Methode ist mit längeren Ausfallzeiten verbunden und wird in der Regel für kleine, unkritische Workloads verwendet.

OI

Siehe Betriebsintegration.

OLA

Siehe Vereinbarung auf operativer Ebene.

Online-Migration

Eine Migrationsmethode, bei der der Quell-Workload auf das Zielsystem kopiert wird, ohne offline genommen zu werden. Anwendungen, die mit dem Workload verbunden sind, können während

der Migration weiterhin funktionieren. Diese Methode beinhaltet keine bis minimale Ausfallzeit und wird in der Regel für kritische Produktionsworkloads verwendet.

OPC-UA

Siehe Open Process Communications — Unified Architecture.

Offene Prozesskommunikation — Einheitliche Architektur (OPC-UA)

Ein machine-to-machine (M2M) -Kommunikationsprotokoll für die industrielle Automatisierung. OPC-UA bietet einen Interoperabilitätsstandard mit Datenverschlüsselungs-, Authentifizierungs- und Autorisierungsschemata.

Vereinbarung auf Betriebsebene (OLA)

Eine Vereinbarung, in der klargestellt wird, welche funktionalen IT-Gruppen sich gegenseitig versprechen zu liefern, um ein Service Level Agreement (SLA) zu unterstützen.

Überprüfung der Betriebsbereitschaft (ORR)

Eine Checkliste mit Fragen und zugehörigen bewährten Methoden, die Ihnen helfen, Vorfälle und mögliche Ausfälle zu verstehen, zu bewerten, zu verhindern oder deren Umfang zu reduzieren. Weitere Informationen finden Sie unter Operational Readiness Reviews (ORR) im AWS Well-Architected Framework.

Betriebstechnologie (OT)

Hardware- und Softwaresysteme, die mit der physischen Umgebung zusammenarbeiten, um industrielle Abläufe, Ausrüstung und Infrastruktur zu steuern. In der Fertigung ist die Integration von OT- und Informationstechnologie (IT) -Systemen ein zentraler Schwerpunkt der Industrie 4.0-Transformationen.

Betriebsintegration (OI)

Der Prozess der Modernisierung von Abläufen in der Cloud, der Bereitschaftsplanung, Automatisierung und Integration umfasst. Weitere Informationen finden Sie im Leitfaden zur Betriebsintegration.

Organisationspfad

Ein Pfad, der von erstellt wird und in AWS CloudTrail dem alle Ereignisse für alle AWS-Konten in einer Organisation protokolliert werden. AWS Organizations Diese Spur wird in jedem AWS-Konto , der Teil der Organisation ist, erstellt und verfolgt die Aktivität in jedem Konto. Weitere Informationen finden Sie in der CloudTrail Dokumentation unter Einen Trail für eine Organisation erstellen.

Organisatorisches Veränderungsmanagement (OCM)

Ein Framework für das Management wichtiger, disruptiver Geschäftstransformationen aus Sicht der Mitarbeiter, der Kultur und der Führung. OCM hilft Organisationen dabei, sich auf neue Systeme und Strategien vorzubereiten und auf diese umzustellen, indem es die Akzeptanz von Veränderungen beschleunigt, Übergangsprobleme angeht und kulturelle und organisatorische Veränderungen vorantreibt. In der AWS Migrationsstrategie wird dieses Framework aufgrund der Geschwindigkeit des Wandels, der bei Projekten zur Cloud-Einführung erforderlich ist, als Mitarbeiterbeschleunigung bezeichnet. Weitere Informationen finden Sie im OCM-Handbuch.

Ursprungszugriffskontrolle (OAC)

In CloudFront, eine erweiterte Option zur Zugriffsbeschränkung, um Ihre Amazon Simple Storage Service (Amazon S3) -Inhalte zu sichern. OAC unterstützt alle S3-Buckets insgesamt AWS-Regionen, serverseitige Verschlüsselung mit AWS KMS (SSE-KMS) sowie dynamische PUT und `DELETE` Anfragen an den S3-Bucket.

Ursprungszugriffsidentität (OAI)

In CloudFront, eine Option zur Zugriffsbeschränkung, um Ihre Amazon S3 S3-Inhalte zu sichern. Wenn Sie OAI verwenden, CloudFront erstellt es einen Principal, mit dem sich Amazon S3 authentifizieren kann. Authentifizierte Principals können nur über eine bestimmte Distribution auf Inhalte in einem S3-Bucket zugreifen. CloudFront Siehe auch OAC, das eine detailliertere und verbesserte Zugriffskontrolle bietet.

ORR

Weitere Informationen finden Sie unter Überprüfung der Betriebsbereitschaft.

NICHT

Siehe Betriebstechnologie.

Ausgehende (egress) VPC

In einer Architektur AWS mit mehreren Konten eine VPC, die Netzwerkverbindungen verarbeitet, die von einer Anwendung aus initiiert werden. Die AWS Security Reference Architecture empfiehlt die Einrichtung Ihres Netzwerkkontos mit eingehendem und ausgehendem Datenverkehr sowie Inspektion, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

P

Berechtigungsgrenze

Eine IAM-Verwaltungsrichtlinie, die den IAM-Prinzipalen zugeordnet ist, um die maximalen Berechtigungen festzulegen, die der Benutzer oder die Rolle haben kann. Weitere Informationen finden Sie unter Berechtigungsgrenzen für IAM-Entitys in der IAM-Dokumentation.

persönlich identifizierbare Informationen (PII)

Informationen, die, wenn sie direkt betrachtet oder mit anderen verwandten Daten kombiniert werden, verwendet werden können, um vernünftige Rückschlüsse auf die Identität einer Person zu ziehen. Beispiele für personenbezogene Daten sind Namen, Adressen und Kontaktinformationen.

Personenbezogene Daten

Siehe persönlich identifizierbare Informationen.

Playbook

Eine Reihe vordefinierter Schritte, die die mit Migrationen verbundenen Aufgaben erfassen, z. B. die Bereitstellung zentraler Betriebsfunktionen in der Cloud. Ein Playbook kann die Form von Skripten, automatisierten Runbooks oder einer Zusammenfassung der Prozesse oder Schritte annehmen, die für den Betrieb Ihrer modernisierten Umgebung erforderlich sind.

PLC

Siehe programmierbare Logiksteuerung.

PLM

Siehe Produktlebenszyklusmanagement.

policy

Ein Objekt, das Berechtigungen definieren (siehe identitätsbasierte Richtlinie), Zugriffsbedingungen spezifizieren (siehe ressourcenbasierte Richtlinie) oder die maximalen Berechtigungen für alle Konten in einer Organisation definieren kann AWS Organizations (siehe Dienststeuerungsrichtlinie).

Polyglotte Beharrlichkeit

Unabhängige Auswahl der Datenspeichertechnologie eines Microservices auf der Grundlage von Datenzugriffsmustern und anderen Anforderungen. Wenn Ihre Microservices über dieselbe Datenspeichertechnologie verfügen, kann dies zu Implementierungsproblemen oder zu

Leistungseinbußen führen. Microservices lassen sich leichter implementieren und erzielen eine bessere Leistung und Skalierbarkeit, wenn sie den Datenspeicher verwenden, der ihren Anforderungen am besten entspricht.

Portfoliobewertung

Ein Prozess, bei dem das Anwendungsportfolio ermittelt, analysiert und priorisiert wird, um die Migration zu planen. Weitere Informationen finden Sie in Bewerten der Migrationsbereitschaft.

predicate

Eine Abfragebedingung, die `true` oder zurückgibt`false`, was üblicherweise in einer Klausel vorkommt. `WHERE`

Prädikat Pushdown

Eine Technik zur Optimierung von Datenbankabfragen, bei der die Daten in der Abfrage vor der Übertragung gefiltert werden. Dadurch wird die Datenmenge reduziert, die aus der relationalen Datenbank abgerufen und verarbeitet werden muss, und die Abfrageleistung wird verbessert.

Präventive Kontrolle

Eine Sicherheitskontrolle, die verhindern soll, dass ein Ereignis eintritt. Diese Kontrollen stellen eine erste Verteidigungslinie dar, um unbefugten Zugriff oder unerwünschte Änderungen an Ihrem Netzwerk zu verhindern. Weitere Informationen finden Sie unter Präventive Kontrolle in Implementierung von Sicherheitskontrollen in AWS.

Prinzipal

Eine Entität AWS , die Aktionen ausführen und auf Ressourcen zugreifen kann. Diese Entität ist in der Regel ein Root-Benutzer für eine AWS-Konto, eine IAM-Rolle oder einen Benutzer. Weitere Informationen finden Sie unter Prinzipal in Rollenbegriffe und -konzepte in der IAM-Dokumentation.

Datenschutz von Natur aus

Ein systemtechnischer Ansatz, der den Datenschutz während des gesamten Entwicklungsprozesses berücksichtigt.

Privat gehostete Zonen

Ein Container, der Informationen darüber enthält, wie Amazon Route 53 auf DNS-Abfragen für eine Domain und deren Subdomains innerhalb einer oder mehrerer VPCs Domains antworten soll. Weitere Informationen finden Sie unter Arbeiten mit privat gehosteten Zonen in der Route-53-Dokumentation.

proaktive Steuerung

Eine Sicherheitskontrolle, die den Einsatz nicht richtlinienkonformer Ressourcen verhindern soll. Diese Steuerelemente scannen Ressourcen, bevor sie bereitgestellt werden. Wenn die Ressource nicht der Kontrolle entspricht, wird sie nicht bereitgestellt. Weitere Informationen finden Sie im Referenzhandbuch zu Kontrollen in der AWS Control Tower Dokumentation und unter Proaktive Kontrollen unter Implementierung von Sicherheitskontrollen am AWS.

Produktlebenszyklusmanagement (PLM)

Das Management von Daten und Prozessen für ein Produkt während seines gesamten Lebenszyklus, vom Design, der Entwicklung und Markteinführung über Wachstum und Reife bis hin zur Markteinführung und Markteinführung.

Produktionsumgebung

Siehe Umgebung.

Speicherprogrammierbare Steuerung (SPS)

In der Fertigung ein äußerst zuverlässiger, anpassungsfähiger Computer, der Maschinen überwacht und Fertigungsprozesse automatisiert.

schnelle Verkettung

Verwendung der Ausgabe einer LLM-Eingabeaufforderung als Eingabe für die nächste Aufforderung, um bessere Antworten zu generieren. Diese Technik wird verwendet, um eine komplexe Aufgabe in Unteraufgaben zu unterteilen oder um eine vorläufige Antwort iterativ zu verfeinern oder zu erweitern. Sie trägt dazu bei, die Genauigkeit und Relevanz der Antworten eines Modells zu verbessern und ermöglicht detailliertere, personalisierte Ergebnisse.

Pseudonymisierung

Der Prozess, bei dem persönliche Identifikatoren in einem Datensatz durch Platzhalterwerte ersetzt werden. Pseudonymisierung kann zum Schutz der Privatsphäre beitragen. Pseudonymisierte Daten gelten weiterhin als personenbezogene Daten.

publish/subscribe (pub/sub)

Ein Muster, das asynchrone Kommunikation zwischen Microservices ermöglicht, um die Skalierbarkeit und Reaktionsfähigkeit zu verbessern. In einem auf Microservices basierenden MES kann ein Microservice beispielsweise Ereignismeldungen in einem Kanal veröffentlichen, den andere Microservices abonnieren können. Das System kann neue Microservices hinzufügen, ohne den Veröffentlichungsservice zu ändern.

Q

Abfrageplan

Eine Reihe von Schritten, wie Anweisungen, die für den Zugriff auf die Daten in einem relationalen SQL-Datenbanksystem verwendet werden.

Abfrageplanregression

Wenn ein Datenbankserviceoptimierer einen weniger optimalen Plan wählt als vor einer bestimmten Änderung der Datenbankumgebung. Dies kann durch Änderungen an Statistiken, Beschränkungen, Umgebungseinstellungen, Abfrageparameter-Bindungen und Aktualisierungen der Datenbank-Engine verursacht werden.

R

RACI-Matrix

Siehe verantwortlich, rechenschaftspflichtig, konsultiert, informiert (RACI).

RAG

Siehe Erweiterte Generierung beim Abrufen.

Ransomware

Eine bösartige Software, die entwickelt wurde, um den Zugriff auf ein Computersystem oder Daten zu blockieren, bis eine Zahlung erfolgt ist.

RASCI-Matrix

Siehe verantwortlich, rechenschaftspflichtig, konsultiert, informiert (RACI).

RCAC

Siehe Zugriffskontrolle für Zeilen und Spalten.

Read Replica

Eine Kopie einer Datenbank, die nur für Lesezwecke verwendet wird. Sie können Abfragen an das Lesereplikat weiterleiten, um die Belastung auf Ihrer Primärdatenbank zu reduzieren.

neu strukturieren

Siehe 7 Rs.

Recovery Point Objective (RPO)

: Die maximal zulässige Zeitspanne seit dem letzten Datenwiederherstellungspunkt. Damit wird festgelegt, was als akzeptabler Datenverlust zwischen dem letzten Wiederherstellungspunkt und der Serviceunterbrechung gilt.

Wiederherstellungszeitziel (RTO)

: Die maximal zulässige Verzögerung zwischen der Betriebsunterbrechung und der Wiederherstellung des Dienstes.

Refaktorierung

: Siehe 7 Rs.

Region

: Eine Sammlung von AWS Ressourcen in einem geografischen Gebiet. Jeder AWS-Region ist isoliert und unabhängig von den anderen, um Fehlertoleranz, Stabilität und Belastbarkeit zu gewährleisten. Weitere Informationen finden Sie unter Geben Sie an, was AWS-Regionen Ihr Konto verwenden kann.

Regression

: Eine ML-Technik, die einen numerischen Wert vorhersagt. Zum Beispiel, um das Problem „Zu welchem Preis wird dieses Haus verkauft werden?“ zu lösen Ein ML-Modell könnte ein lineares Regressionsmodell verwenden, um den Verkaufspreis eines Hauses auf der Grundlage bekannter Fakten über das Haus (z. B. die Quadratmeterzahl) vorherzusagen.

rehosten

: Siehe 7 Rs.

Veröffentlichung

: In einem Bereitstellungsprozess der Akt der Förderung von Änderungen an einer Produktionsumgebung.

umziehen

: Siehe 7 Rs.

neue Plattform

: Siehe 7 Rs.

Rückkauf

Siehe 7 Rs.

Ausfallsicherheit

Die Fähigkeit einer Anwendung, Störungen zu widerstehen oder sich von ihnen zu erholen. Hochverfügbarkeit und Notfallwiederherstellung sind häufig Überlegungen bei der Planung der Ausfallsicherheit in der. AWS Cloud Weitere Informationen finden Sie unter AWS Cloud Resilienz.

Ressourcenbasierte Richtlinie

Eine mit einer Ressource verknüpfte Richtlinie, z. B. ein Amazon-S3-Bucket, ein Endpunkt oder ein Verschlüsselungsschlüssel. Diese Art von Richtlinie legt fest, welchen Prinzipalen der Zugriff gewährt wird, welche Aktionen unterstützt werden und welche anderen Bedingungen erfüllt sein müssen.

RACI-Matrix (verantwortlich, rechenschaftspflichtig, konsultiert, informiert)

Eine Matrix, die die Rollen und Verantwortlichkeiten aller an Migrationsaktivitäten und Cloud-Operationen beteiligten Parteien definiert. Der Matrixname leitet sich von den in der Matrix definierten Zuständigkeitstypen ab: verantwortlich (R), rechenschaftspflichtig (A), konsultiert (C) und informiert (I). Der Unterstützungstyp (S) ist optional. Wenn Sie Unterstützung einbeziehen, wird die Matrix als RASCI-Matrix bezeichnet, und wenn Sie sie ausschließen, wird sie als RACI-Matrix bezeichnet.

Reaktive Kontrolle

Eine Sicherheitskontrolle, die darauf ausgelegt ist, die Behebung unerwünschter Ereignisse oder Abweichungen von Ihren Sicherheitsstandards voranzutreiben. Weitere Informationen finden Sie unter Reaktive Kontrolle in Implementieren von Sicherheitskontrollen in AWS.

Beibehaltung

Siehe 7 Rs.

zurückziehen

Siehe 7 Rs.

Retrieval Augmented Generation (RAG)

Eine generative KI-Technologie, bei der ein LLM auf eine maßgebliche Datenquelle verweist, die sich außerhalb seiner Trainingsdatenquellen befindet, bevor eine Antwort generiert wird. Ein

RAG-Modell könnte beispielsweise eine semantische Suche in der Wissensdatenbank oder in benutzerdefinierten Daten einer Organisation durchführen. Weitere Informationen finden Sie unter Was ist RAG.

Drehung

Der Vorgang, bei dem ein Geheimnis regelmäßig aktualisiert wird, um es einem Angreifer zu erschweren, auf die Anmeldeinformationen zuzugreifen.

Zugriffskontrolle für Zeilen und Spalten (RCAC)

Die Verwendung einfacher, flexibler SQL-Ausdrücke mit definierten Zugriffsregeln. RCAC besteht aus Zeilenberechtigungen und Spaltenmasken.

RPO

Siehe Recovery Point Objective.

RTO

Siehe Ziel für die Erholungszeit.

Runbook

Eine Reihe manueller oder automatisierter Verfahren, die zur Ausführung einer bestimmten Aufgabe erforderlich sind. Diese sind in der Regel darauf ausgelegt, sich wiederholende Operationen oder Verfahren mit hohen Fehlerquoten zu rationalisieren.

S

SAML 2.0

Ein offener Standard, den viele Identitätsanbieter (IdPs) verwenden. Diese Funktion ermöglicht föderiertes Single Sign-On (SSO), sodass sich Benutzer bei den API-Vorgängen anmelden AWS-Managementkonsole oder die AWS API-Operationen aufrufen können, ohne dass Sie einen Benutzer in IAM für alle in Ihrer Organisation erstellen müssen. Weitere Informationen zum SAML-2.0.-basierten Verbund finden Sie unter Über den SAML-2.0-basierten Verbund in der IAM-Dokumentation.

SCADA

Siehe Aufsichtskontrolle und Datenerfassung.

SCP

Siehe Richtlinie zur Dienstkontrolle.

Secret

Interne AWS Secrets Manager, vertrauliche oder eingeschränkte Informationen, wie z. B. ein Passwort oder Benutzeranmeldeinformationen, die Sie in verschlüsselter Form speichern. Es besteht aus dem geheimen Wert und seinen Metadaten. Der geheime Wert kann binär, eine einzelne Zeichenfolge oder mehrere Zeichenketten sein. Weitere Informationen finden Sie unter Was ist in einem Secrets Manager Manager-Geheimnis? in der Secrets Manager Manager-Dokumentation.

Sicherheit durch Design

Ein systemtechnischer Ansatz, der die Sicherheit während des gesamten Entwicklungsprozesses berücksichtigt.

Sicherheitskontrolle

Ein technischer oder administrativer Integritätsschutz, der die Fähigkeit eines Bedrohungsakteurs, eine Schwachstelle auszunutzen, verhindert, erkennt oder einschränkt. Es gibt vier Haupttypen von Sicherheitskontrollen: präventiv, detektiv, reaktionsschnell und proaktiv.

Härtung der Sicherheit

Der Prozess, bei dem die Angriffsfläche reduziert wird, um sie widerstandsfähiger gegen Angriffe zu machen. Dies kann Aktionen wie das Entfernen von Ressourcen, die nicht mehr benötigt werden, die Implementierung der bewährten Sicherheitsmethode der Gewährung geringster Berechtigungen oder die Deaktivierung unnötiger Feature in Konfigurationsdateien umfassen.

System zur Verwaltung von Sicherheitsinformationen und Ereignissen (security information and event management – SIEM)

Tools und Services, die Systeme für das Sicherheitsinformationsmanagement (SIM) und das Management von Sicherheitsereignissen (SEM) kombinieren. Ein SIEM-System sammelt, überwacht und analysiert Daten von Servern, Netzwerken, Geräten und anderen Quellen, um Bedrohungen und Sicherheitsverletzungen zu erkennen und Warnmeldungen zu generieren.

Automatisierung von Sicherheitsreaktionen

Eine vordefinierte und programmierte Aktion, die darauf ausgelegt ist, automatisch auf ein Sicherheitsereignis zu reagieren oder es zu beheben. Diese Automatisierungen dienen als detektive oder reaktionsschnelle Sicherheitskontrollen, die Sie bei der Implementierung bewährter

AWS Sicherheitsmethoden unterstützen. Beispiele für automatisierte Antwortaktionen sind das Ändern einer VPC-Sicherheitsgruppe, das Patchen einer Amazon EC2 EC2-Instance oder das Rotieren von Anmeldeinformationen.

Serverseitige Verschlüsselung

Verschlüsselung von Daten am Zielort durch denjenigen AWS-Service , der sie empfängt.

Service-Kontrollrichtlinie (SCP)

Eine Richtlinie, die eine zentrale Steuerung der Berechtigungen für alle Konten in einer Organisation in ermöglicht AWS Organizations. SCPs Definieren Sie Leitplanken oder legen Sie Grenzwerte für Aktionen fest, die ein Administrator an Benutzer oder Rollen delegieren kann. Sie können sie SCPs als Zulassungs- oder Ablehnungslisten verwenden, um festzulegen, welche Dienste oder Aktionen zulässig oder verboten sind. Weitere Informationen finden Sie in der AWS Organizations Dokumentation unter Richtlinien zur Dienststeuerung.

Service-Endpunkt

Die URL des Einstiegspunkts für einen AWS-Service. Sie können den Endpunkt verwenden, um programmgesteuert eine Verbindung zum Zielservice herzustellen. Weitere Informationen finden Sie unter AWS-Service -Endpunkte in der Allgemeine AWS-Referenz.

Service Level Agreement (SLA)

Eine Vereinbarung, in der klargestellt wird, was ein IT-Team seinen Kunden zu bieten verspricht, z. B. in Bezug auf Verfügbarkeit und Leistung der Services.

Service-Level-Indikator (SLI)

Eine Messung eines Leistungsaspekts eines Dienstes, z. B. seiner Fehlerrate, Verfügbarkeit oder Durchsatz.

Service-Level-Ziel (SLO)

Eine Zielkennzahl, die den Zustand eines Dienstes darstellt, gemessen anhand eines Service-Level-Indikators.

Modell der geteilten Verantwortung

Ein Modell, das die Verantwortung beschreibt, mit der Sie gemeinsam AWS für Cloud-Sicherheit und Compliance verantwortlich sind. AWS ist für die Sicherheit der Cloud verantwortlich, während Sie für die Sicherheit in der Cloud verantwortlich sind. Weitere Informationen finden Sie unter Modell der geteilten Verantwortung.

SIEM

Siehe Sicherheitsinformations- und Event-Management-System.

Single Point of Failure (SPOF)

Ein Fehler in einer einzelnen, kritischen Komponente einer Anwendung, der das System stören kann.

SLA

Siehe Service Level Agreement.

SLI

Siehe Service-Level-Indikator.

ALSO

Siehe Service-Level-Ziel.

split-and-seed Modell

Ein Muster für die Skalierung und Beschleunigung von Modernisierungsprojekten. Sobald neue Features und Produktversionen definiert werden, teilt sich das Kernteam auf, um neue Produktteams zu bilden. Dies trägt zur Skalierung der Fähigkeiten und Services Ihrer Organisation bei, verbessert die Produktivität der Entwickler und unterstützt schnelle Innovationen. Weitere Informationen finden Sie unter Schrittweiser Ansatz zur Modernisierung von Anwendungen in der AWS Cloud.

SPOTTEN

Siehe Single Point of Failure.

Sternschema

Eine Datenbank-Organisationsstruktur, die eine große Faktentabelle zum Speichern von Transaktions- oder Messdaten und eine oder mehrere kleinere dimensionale Tabellen zum Speichern von Datenattributen verwendet. Diese Struktur ist für die Verwendung in einem Data Warehouse oder für Business Intelligence-Zwecke konzipiert.

Strangler-Fig-Muster

Ein Ansatz zur Modernisierung monolithischer Systeme, bei dem die Systemfunktionen schrittweise umgeschrieben und ersetzt werden, bis das Legacy-System außer Betrieb

genommen werden kann. Dieses Muster verwendet die Analogie einer Feigenrebe, die zu einem etablierten Baum heranwächst und schließlich ihren Wirt überwindet und ersetzt. Das Muster wurde eingeführt von Martin Fowler als Möglichkeit, Risiken beim Umschreiben monolithischer Systeme zu managen. Ein Beispiel für die Anwendung dieses Musters finden Sie unter Schrittweises Modernisieren älterer Microsoft ASP.NET (ASMX)-Webservices mithilfe von Containern und Amazon API Gateway.

Subnetz

Ein Bereich von IP-Adressen in Ihrer VPC. Ein Subnetz muss sich in einer einzigen Availability Zone befinden.

Aufsichtskontrolle und Datenerfassung (SCADA)

In der Fertigung ein System, das Hardware und Software zur Überwachung von Sachanlagen und Produktionsabläufen verwendet.

Symmetrische Verschlüsselung

Ein Verschlüsselungsalgorithmus, der denselben Schlüssel zum Verschlüsseln und Entschlüsseln der Daten verwendet.

synthetisches Testen

Testen eines Systems auf eine Weise, die Benutzerinteraktionen simuliert, um potenzielle Probleme zu erkennen oder die Leistung zu überwachen. Sie können Amazon CloudWatch Synthetics verwenden, um diese Tests zu erstellen.

Systemaufforderung

Eine Technik, mit der einem LLM Kontext, Anweisungen oder Richtlinien zur Verfügung gestellt werden, um sein Verhalten zu steuern. Systemaufforderungen helfen dabei, den Kontext festzulegen und Regeln für Interaktionen mit Benutzern festzulegen.

T

tags

Schlüssel-Wert-Paare, die als Metadaten für die Organisation Ihrer Ressourcen dienen. AWS Mit Tags können Sie Ressourcen verwalten, identifizieren, organisieren, suchen und filtern. Weitere Informationen finden Sie unter Markieren Ihrer AWS -Ressourcen.

Zielvariable

Der Wert, den Sie in überwachtem ML vorhersagen möchten. Dies wird auch als Ergebnisvariable bezeichnet. In einer Fertigungsumgebung könnte die Zielvariable beispielsweise ein Produktfehler sein.

Aufgabenliste

Ein Tool, das verwendet wird, um den Fortschritt anhand eines Runbooks zu verfolgen. Eine Aufgabenliste enthält eine Übersicht über das Runbook und eine Liste mit allgemeinen Aufgaben, die erledigt werden müssen. Für jede allgemeine Aufgabe werden der geschätzte Zeitaufwand, der Eigentümer und der Fortschritt angegeben.

Testumgebungen

Siehe Umgebung.

Training

Daten für Ihr ML-Modell bereitstellen, aus denen es lernen kann. Die Trainingsdaten müssen die richtige Antwort enthalten. Der Lernalgorithmus findet Muster in den Trainingsdaten, die die Attribute der Input-Daten dem Ziel (die Antwort, die Sie voraussagen möchten) zuordnen. Es gibt ein ML-Modell aus, das diese Muster erfasst. Sie können dann das ML-Modell verwenden, um Voraussagen für neue Daten zu erhalten, bei denen Sie das Ziel nicht kennen.

Transit-Gateway

Ein Netzwerk-Transit-Hub, über den Sie Ihre Netzwerke VPCs und Ihre lokalen Netzwerke miteinander verbinden können. Weitere Informationen finden Sie in der Dokumentation unter Was ist ein Transit-Gateway. AWS Transit Gateway

Stammbasierter Workflow

Ein Ansatz, bei dem Entwickler Feature lokal in einem Feature-Zweig erstellen und testen und diese Änderungen dann im Hauptzweig zusammenführen. Der Hauptzweig wird dann sequentiell für die Entwicklungs-, Vorproduktions- und Produktionsumgebungen erstellt.

Vertrauenswürdiger Zugriff

Gewährung von Berechtigungen für einen Dienst, den Sie angeben, um Aufgaben in Ihrer Organisation AWS Organizations und in deren Konten in Ihrem Namen auszuführen. Der vertrauenswürdige Service erstellt in jedem Konto eine mit dem Service verknüpfte Rolle, wenn diese Rolle benötigt wird, um Verwaltungsaufgaben für Sie auszuführen. Weitere Informationen finden Sie in der AWS Organizations Dokumentation unter Verwendung AWS Organizations mit anderen AWS Diensten.

Optimieren

Aspekte Ihres Trainingsprozesses ändern, um die Genauigkeit des ML-Modells zu verbessern. Sie können das ML-Modell z. B. trainieren, indem Sie einen Beschriftungssatz generieren, Beschriftungen hinzufügen und diese Schritte dann mehrmals unter verschiedenen Einstellungen wiederholen, um das Modell zu optimieren.

Zwei-Pizzen-Team

Ein kleines DevOps Team, das Sie mit zwei Pizzen ernähren können. Eine Teamgröße von zwei Pizzen gewährleistet die bestmögliche Gelegenheit zur Zusammenarbeit bei der Softwareentwicklung.

U

Unsicherheit

Ein Konzept, das sich auf ungenaue, unvollständige oder unbekannte Informationen bezieht, die die Zuverlässigkeit von prädiktiven ML-Modellen untergraben können. Es gibt zwei Arten von Unsicherheit: Epistemische Unsicherheit wird durch begrenzte, unvollständige Daten verursacht, wohingegen aleatorische Unsicherheit durch Rauschen und Randomisierung verursacht wird, die in den Daten liegt.

undifferenzierte Aufgaben

Diese Arbeit wird auch als Schwerstarbeit bezeichnet. Dabei handelt es sich um Arbeiten, die zwar für die Erstellung und den Betrieb einer Anwendung erforderlich sind, aber dem Endbenutzer keinen direkten Mehrwert bieten oder keinen Wettbewerbsvorteil bieten. Beispiele für undifferenzierte Aufgaben sind Beschaffung, Wartung und Kapazitätsplanung.

höhere Umgebungen

Siehe Umgebung.

V

Vacuuming

Ein Vorgang zur Datenbankwartung, bei dem die Datenbank nach inkrementellen Aktualisierungen bereinigt wird, um Speicherplatz zurückzugewinnen und die Leistung zu verbessern.

Versionskontrolle

Prozesse und Tools zur Nachverfolgung von Änderungen, z. B. Änderungen am Quellcode in einem Repository.

VPC-Peering

Eine Verbindung zwischen zwei VPCs , die es Ihnen ermöglicht, den Verkehr mithilfe privater IP-Adressen weiterzuleiten. Weitere Informationen finden Sie unter Was ist VPC-Peering? in der Amazon-VPC-Dokumentation.

Schwachstelle

Ein Software- oder Hardwarefehler, der die Sicherheit des Systems beeinträchtigt.

W

Warmer Cache

Ein Puffer-Cache, der aktuelle, relevante Daten enthält, auf die häufig zugegriffen wird. Die Datenbank-Instance kann aus dem Puffer-Cache lesen, was schneller ist als das Lesen aus dem Hauptspeicher oder von der Festplatte.

warme Daten

Daten, auf die selten zugegriffen wird. Bei der Abfrage dieser Art von Daten sind mäßig langsame Abfragen in der Regel akzeptabel.

Fensterfunktion

Eine SQL-Funktion, die eine Berechnung für eine Gruppe von Zeilen durchführt, die sich in irgendeiner Weise auf den aktuellen Datensatz beziehen. Fensterfunktionen sind nützlich für die Verarbeitung von Aufgaben wie die Berechnung eines gleitenden Durchschnitts oder für den Zugriff auf den Wert von Zeilen auf der Grundlage der relativen Position der aktuellen Zeile.

Workload

Ein Workload ist eine Sammlung von Ressourcen und Code, die einen Unternehmenswert bietet, wie z. B. eine kundenorientierte Anwendung oder ein Backend-Prozess.

Workstream

Funktionsgruppen in einem Migrationsprojekt, die für eine bestimmte Reihe von Aufgaben verantwortlich sind. Jeder Workstream ist unabhängig, unterstützt aber die anderen Workstreams

im Projekt. Der Portfolio-Workstream ist beispielsweise für die Priorisierung von Anwendungen, die Wellenplanung und die Erfassung von Migrationsmetadaten verantwortlich. Der Portfolio-Workstream liefert diese Komponenten an den Migrations-Workstream, der dann die Server und Anwendungen migriert.

WURM

Sehen Sie einmal schreiben, viele lesen.

WQF

Siehe AWS Workload-Qualifizierungsrahmen.

einmal schreiben, viele lesen (WORM)

Ein Speichermodell, das Daten ein einziges Mal schreibt und verhindert, dass die Daten gelöscht oder geändert werden. Autorisierte Benutzer können die Daten so oft wie nötig lesen, aber sie können sie nicht ändern. Diese Datenspeicherinfrastruktur gilt als unveränderlich.

Z

Zero-Day-Exploit

Ein Angriff, in der Regel Malware, der eine Zero-Day-Sicherheitslücke ausnutzt.

Zero-Day-Sicherheitslücke

Ein unfehlbarer Fehler oder eine Sicherheitslücke in einem Produktionssystem. Bedrohungsakteure können diese Art von Sicherheitslücke nutzen, um das System anzugreifen. Entwickler werden aufgrund des Angriffs häufig auf die Sicherheitsanfälligkeit aufmerksam.

Eingabeaufforderung ohne Zwischenfälle

Bereitstellung von Anweisungen für die Ausführung einer Aufgabe an einen LLM, jedoch ohne Beispiele (Schnappschüsse), die ihm als Orientierungshilfe dienen könnten. Der LLM muss sein vortrainiertes Wissen einsetzen, um die Aufgabe zu bewältigen. Die Effektivität von Zero-Shot Prompting hängt von der Komplexität der Aufgabe und der Qualität der Aufforderung ab. Siehe auch Few-Shot-Prompting.

Zombie-Anwendung

Eine Anwendung, deren durchschnittliche CPU- und Arbeitsspeichernutzung unter 5 Prozent liegt. In einem Migrationsprojekt ist es üblich, diese Anwendungen außer Betrieb zu nehmen.

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.