



HP Z8 G4 mit NVIDIA- Grafikprozessoren oder Public Cloud für umfangreiche Data-Science-Aufgaben

Welche Lösung ist finanziell sinnvoller?

Powered by



Intellerts





Data Science ist eine rechenintensive Aufgabe. Mit der Verbesserung von Methoden und verstärktem Einsatz von Machine Learning stoßen neue Techniken an die Grenzen moderner Hardware. Es ist eine Herausforderung, mit den neuesten Errungenschaften Schritt zu halten, die neuesten Modelle zu implementieren und ihre Fähigkeiten zu testen.

Aus IT-Perspektive gibt es zwei wesentliche Ansätze zur Unterstützung Ihrer Data Scientists. Sie können entweder in lokale Hardware investieren oder die Rechenleistung der Cloud nutzen. In folgendem Geschäftsmodell vergleichen wir die Nutzung eines HP Z8 G4 Rechners mit zwei NVIDIA-GPUs mit einem Angebot ähnlicher Leistung eines großen Cloud-Anbieters. Dabei geht es darum, herauszufinden, ob der Einsatz eines HP Z8 G4 Rechners, der eine beträchtliche Vorabinvestition erfordert, im Vergleich zur Nutzung einer Cloud-Lösung sinnvoll ist.

Um unseren Fall greifbar zu machen, werden wir einige realistische datenwissenschaftliche Experimente als Ausgangspunkt für unseren Vergleich durchführen. Wir definieren eine realistische Arbeitslast und berechnen die Kosten für deren Ausführung im Zeitverlauf mit dem HP Z8 G4 Rechner als auch mit einer vergleichbaren Cloud-Instanz.

In diesem Whitepaper werden wir zunächst den Data Science-Workload erläutern, den wir als Grundlage für unsere Berechnungen verwenden. Anschließend werfen wir einen Blick auf die erforderliche Hardware. Wir werden zunächst beschreiben, wie die Arbeitslast auf einem lokalen Desktop-Rechner ausgeführt wird, und dann einen Blick auf eine vergleichbare Bare Metal Cloud-Instanz werfen. Abschließend berechnen wir die Gesamtkosten beider Optionen. Was rechnet sich am Ende finanziell besser: Eine Vorabinvestition in eine lokale Data Science Workstation wie die HP Z8 G4 oder monatliche Zahlungen an einen öffentlichen Cloud-Anbieter?



Ausgehend von einer realistischen Arbeitslast

Für unseren Vergleich benötigen wir einen realistischen Data Science-Workload. Wir haben uns dazu entschieden, mit Aufgaben der Namenserkennung (Named Entity Recognition, NER) zu experimentieren. NER ist ein fortgeschrittener Teilbereich der natürlichen Sprachverarbeitung (Natural Language Processing, NLP) mit einer Vielzahl von faszinierenden Anwendungen. Die Modelle, die der Namenserkennung zugrunde liegen, sind zwar komplex, jedoch nicht das Prinzip.

Bei NER zielt ein Algorithmus darauf ab, Informationen aus unstrukturiertem Text zu extrahieren, indem er sogenannte „benannte Entitäten“ aufspürt und klassifiziert. Betrachten wir ein vereinfachtes Beispiel. Sie könnten auf ein Dokument stoßen, das den folgenden Satz enthält:

“Big Business Corp. hat 10.000 Dollar für eine Rechnung der spanischen Abteilung von Desktop Inc. bezahlt.”

Mit Hilfe von NER könnte ein Algorithmus [Big Business Corp.] und [Desktop Inc.] als Firmennamen erkennen, [10.000 Dollar] als Betrag, [spanischen] als Adjektiv, das sich auf den Ort bezieht, und so weiter. Abb. 1 veranschaulicht dies anhand der Anwendung von NER auf Text von Wikipedia. Die Anwendung von NER auf eine große Menge unstrukturierter Daten könnte interessante Erkenntnisse für Analysten liefern, die Entscheidungen in den Bereichen Rechnungsprüfung, Investitionen, Finanzberichterstattung und Identifizierung verdächtiger Ereignisse treffen. Die weitere Analyse der strukturierten Daten, die durch NER generiert werden, könnte Erkenntnisse und Zusammenhänge liefern, die bisher nicht bekannt waren.

In einem praktischen Szenario würden Data Scientists Modelle trainieren, ausführen und feinabstimmen, um verwertbare Ergebnisse für das Unternehmen zu erzielen. Wie wir im weiteren Verlauf des Whitepapers sehen werden, handelt es sich dabei um eine kontinuierliche, iterative Aufgabe.

Die **Niederlande** (GPE) (**Niederländisch** (NORP) : **Nederland** (GPE) [ˈneːdarlant] (About this soundlisten)); informell **Holland** (GPE) , ist ein Land, das hauptsächlich in **Westeuropa** (LOC) und teilweise in der **Karibik** (LOC) liegt. Es ist das größte der **vier** (KARDINALEN) Länder, die das **Königreich der Niederlande** (GPE) bilden. In **Europa** (LOC) bestehen **die Niederlande** (GPE) aus **zwölf** (KARDINALEN) Provinzen, die im Osten an **Deutschland** (GPE) , im Süden an **Belgien** (GPE) und im **Nordwesten an die Nordsee** (LOC) grenzen, mit Seegrenzen in der **Nordsee** (LOC) zu diesen Ländern und dem **Vereinigten Königreich** (GPE) . In der **Karibik** (LOC) besteht sie aus **drei** (KARDINALEN) besonderen Gemeinden: den Inseln **Bonaire** (GPE) , **Sint Eustatius** (PERSON) und **Saba** (GPE) . Die Amtssprache des Landes ist **Niederländisch** (NORP) , mit **Westfriesisch** (NORP) als zweiter Amtssprache in der Provinz **Friesland** (GPE) und **Englisch** (LANGUAGE) und **Papiamentu** (GPE) als zweiten Amtssprachen in den **Karibischen Niederlanden** (LOC) . **Niederländisch** (NORP) , **Niedersächsisch** (PERSON) und **Limburgisch** (NORP) sind anerkannte Regionalsprachen (im Osten bzw. Südosten gesprochen), während **Sinte Romani** (PERSON) und **Jiddisch** (GPE) anerkannte nicht-territoriale Sprachen sind.

- ✓ Person
- ✓ Norp
- ✓ Org
- ✓ Gpe
- ✓ Loc
- ✓ Produkt
- ✓ Event
- ✓ Kunstwerk
- ✓ Sprache
- ✓ Datum
- ✓ Zeit
- ✓ Prozent
- ✓ Geld
- ✓ Menge
- ✓ Ordinal
- ✓ Kardinal

Abb.1: NER extrahiert Informationen aus unstrukturiertem Text durch Klassifizierung von Entitäten.

Einrichten des Tests

Für unseren Test werden wir die Kosten für den Betrieb einiger beliebter NER-Modelle analysieren. Wir werden uns speziell mit Finanzinformationen befassen, da dies eine sehr realistische Anwendung mit NER ist. Die jüngsten Fortschritte bei der Verarbeitung natürlicher Sprache beruhen in hohem Maße auf großen vortrainierten Modellen, die mit modernsten Deep Learning-Techniken erstellt wurden. Darüber hinaus sind sie sehr ressourcenintensiv und erfordern mehrere Iterationen, um ein hervorragendes Ergebnis zu erzielen. Es sind mehrere fortschrittliche Modelle verfügbar. Für unser Experiment haben wir die folgenden Modelle ausgewählt:

- **Googles BERT** (Bidirectional Encoder Representations from Transformers) als das ursprüngliche Transformatormodell, das die ganze Revolution der Transformatoren ausgelöst hat;
- **Facebooks RoBERTa**, als eine robustere und optimierte Version des ursprünglichen BERT-Modells;
- **ELMo** Modell mit tiefen kontextualisierten Wortrepräsentationen, das vom Allen Institute for Artificial Intelligence entwickelt wurde;
- **Flair** kontextbezogene String-Einbettungen, erstellt von Zalando. Um eine bessere Leistung zu erzielen, haben wir sowohl Vorwärts- als auch Rückwärtseinbettungen verwendet;
- **XLNet**-Modell, das autoregressives Vortraining anwendet und BERT in einer Vielzahl von Aufgaben übertrifft;
- **XLNet-Roberta**, ein großes mehrsprachiges Modell auf der Grundlage von Facebooks RoBERTa.



Für das Training der Modelle verwenden wir Daten von Ontonotes. Ontonotes enthält eine riesige Menge an Textdaten aus Telefongesprächen, Newswire, Newsgroups, Rundfunknachrichten, Rundfunkgesprächen und Weblogs. Wir haben die verfügbaren religiösen Texte ausgeschlossen, da sie für das Training unserer Benchmark-Modelle irrelevant sind.

Darüber hinaus haben wir Daten der US-Börsenaufsichtsbehörde (SEC) verwendet. Die Dokumente in diesem Datensatz sind für das Training unseres Finanz-NER-Modells außerordentlich relevant. Leider enthalten die SEC-Daten nur grundlegende Entitätstypen wie Person, Ort, Organisation oder Verschiedenes, während die Ontonotes-Daten verwendet werden können, um ein Modell zur Erkennung von 20 Entitätstypen zu trainieren. Wir haben ebenfalls FinBERT verwendet, ein sogenanntes Transformatormodell, das von der Hong Kong University of Science and Technology anhand einer großen Menge von Finanzdokumenten trainiert wurde.

Die Besonderheiten der einzelnen Modelle sind für Data Scientists relevant. Aus geschäftlicher Sicht ähneln sie sich jedoch alle. Sie müssen trainiert werden und führen NER auf einem Satz unstrukturierter Daten durch. Das Training, die Ausführung und die Feinabstimmung der Modelle verbrauchen eine Menge Systemressourcen. Wir haben das Experiment mit verschiedenen Modellen durchgeführt, um sicherzustellen, dass es keine Verzerrungen in Bezug auf die zugrunde liegende Hardware gibt.



Die Hardware

Zunächst werden wir unsere Experimente auf lokaler Hardware durchführen. Da NER eine ressourcenintensive Aufgabe ist, benötigen wir eine leistungsfähige Maschine. In diesem Geschäftsfall verwenden wir eine HP Z8 G4 Data Science Workstation mit zwei hochmodernen NVIDIA RTX8000-Grafikprozessoren, die jeweils mit 48 GB VRAM ausgestattet sind. Das System selbst verfügt über 376 GB RAM. Damit stehen uns eine Fülle von Systemressourcen zur Verfügung, die es uns ermöglichen, die größten Versionen aller oben genannten Sprachmodelle zu verwenden. Darüber hinaus ist das System mit zwei Intel(R) Xeon(R) Gold 6242R CPUs (je 20 Kerne), einem 4 TB HP Z Turbo-Datenlaufwerk, einem 1.450-Watt-Netzteil und Ubuntu 20.04 ausgestattet.

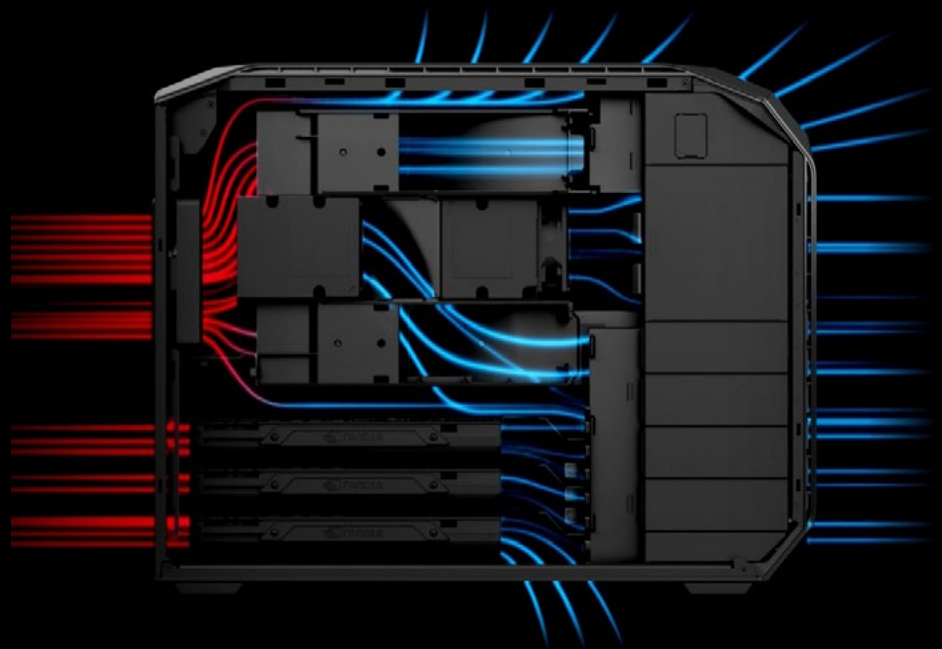


Abb.2: Die in unserem Experiment verwendete HP Z8 G4-Konfiguration.



NER arbeitet genau wie andere natürlichsprachliche Anwendungen und Deep Learnings im Allgemeinen am effizientesten, wenn die Softwaremodelle über Zugang zu Hardware-Beschleunigern wie den NVIDIA GPUs verfügen. Dies begründet sich darin, dass Training und Inferenz an sich nicht unbedingt komplexe Aufgaben sind. Oder genauer gesagt: Eine einzige Berechnung, die für einen einzigen Schritt im Prozess erforderlich ist, ist nicht so rechenintensiv. Um ein komplexes Modell in einem akzeptablen Zeitrahmen ausführen zu können, muss die Hardware in der Lage sein, sehr viele kleine Berechnungen gleichzeitig auszuführen.

CPUs sind sehr gut darin, weniger schwere Operationen auszuführen, während sich GPUs als Beschleuniger durch die parallele Ausführung kleinerer Anweisungen auszeichnen – genau das, was wir jetzt brauchen. Darüber hinaus bestimmt die Speichergröße die Größe des Modells und die Datenmenge, die ein System verarbeiten kann. Wie Sie sehen können, ist unser HP Z8 G4 mit einer ordentlichen Menge an Arbeitsspeicher und zwei leistungsstarken GPUs gut für die Aufgabe gerüstet.

Test und Abstimmung

Während unseres Experiments haben wir versucht, unsere HP-Maschine so weit wie möglich auszunutzen. Das bedeutet, dass wir die Größe der Datenstapel für das Training der Modelle schrittweise erhöht haben, bis einige Modelle anfangen, die verfügbaren Ressourcen zu erschöpfen. Wir versuchten, die Modelle so weit wie möglich zu optimieren.

Data Science ist keine einfache Aufgabe. Die Erstellung eines genauen und effizienten Modells, das relevante Ergebnisse liefert, kann Hunderte von Versuchen erfordern. In der Data Science im Allgemeinen und in unserem NER-Experiment im Besonderen gibt es keinen endlichen Zustand oder eine perfekte Antwort. Als Data Scientist müssen Sie an dem Modell arbeiten und es immer weiter verbessern, damit es immer relevanter und präziser wird. Das bedeutet, dass Sie eine Menge Experimente durchführen, einige Parameter für jedes Modell einstellen und diese alle noch einmal durchführen müssen.

Im Durchschnitt dauerte die Durchführung eines Experiments mit einem einzigen Modell 3 Stunden und 46 Minuten. Bei unserer ersten Iteration dauerte die Ausführung aller Modelle 23 Stunden. Von diesem Ausgangspunkt aus haben wir die Modelle in zwanzig verschiedenen Iterationen verfeinert. Nach jedem Durchlauf bewerteten wir die Leistung und die Laufzeit des Modells und führten weitere Iterationen durch. Letztendlich dauerte es achteinhalb Stunden, bis alle Modelle mit der vollen GPU-Kapazität des HP Z8 G4 liefen.

Dieser Prozess der Ausführung aller Modelle, der Feinabstimmung und der Iteration wird unser Benchmark für die Berechnung der Kosten von Data Science sowohl auf dem HP Z8 G4 als auch in der Cloud sein.

In die Cloud

Die öffentliche Cloud soll eine interessante Alternative zu unserer High-End-Data-Science-Maschine bilden. Man könnte versucht sein, die Experimente in der Cloud durchzuführen. Die Anfangskosten sind zugegebenermaßen deutlich geringer. Wie wir im vorigen Kapitel erörtert haben, ist die Data Science jedoch kein einmaliges Projekt, sondern erfordert kontinuierliche Iteration und Verbesserung. Wie wird die Cloud im Vergleich zum Desktop abschneiden, wenn wir den gesamten Lebenszyklus eines NER-Projekts berücksichtigen, einschließlich der kontinuierlichen Verbesserung?

Zunächst einmal benötigen wir eine Instanz, die mit unserer HP Z8 G4 Workstation vergleichbar ist. AWS bietet die Instanz g4db.metal an. Wir wollen eine Bare Metal-Instanz, da wir beabsichtigen, dieselbe Software darauf laufen zu lassen, und wir möchten nicht auf einen bestimmten Data Science-Service eines Cloud-Anbieters beschränkt sein.

Die Bare Metal-Instanz wird von bis zu acht NVIDIA T4 Tensor Core-GPUs mit 320 Turing Tensor Cores und 2.560 CUDA Cores angetrieben. Die GPUs verfügen jeweils über 16 Gb Speicher. Wir entscheiden uns für eine Instanz mit 384 Gb RAM. Die CPU ist ein benutzerdefinierter Xeon Scalable-Prozessor mit bis zu 64 vCPUs. Das ist weniger als bei unserer HP-Workstation, dies sollte sich jedoch bei unserer experimentellen Arbeitslast nicht signifikant auswirken.



Im Vergleich zum HP Z8 ist die AWS Bare Metal-Instanz in einigen Bereichen deutlich leistungsfähiger. Sie verfügt ungefähr über die doppelte Anzahl an GPU-Kernen. Die Kerne im HP Z8 G4 arbeiten jedoch mit fast der doppelten Taktrate. Das bedeutet, dass die AWS-Instanz im Vergleich zum Desktop doppelt so viele parallele Berechnungen durchführen kann, aber jede Berechnung doppelt so lange dauert. Theoretisch sollten sich diese Unterschiede mehr oder weniger gegenseitig aufheben. Wir haben den genauen Leistungsunterschied in der Praxis nicht überprüft, sind aber zuversichtlich, dass die Auswirkungen auf unsere Ergebnisse minimal sein werden. Andere Ressourcen sind ähnlich. Im Allgemeinen liegen das HP Z8 G4 und das AWS g4dn.metal auf dem gleichen Niveau. Der HP Z8 G4 ist mit flexiblen Konfigurationen erhältlich, die AWS-Instanz hingegen nicht: Sie müssen sich an das Angebot anpassen, das Ihren Anforderungen am besten entspricht.



Technisches Leistungsvermögen	Typ	HP Z8 G4 Konfiguration mit 2 X Quadro RTX 8000	AWS g4dn.metal Konfiguration mit 8 X Tesla T4
CUDA Parallel-Prozessorkerne	Insgesamt	4608 x 2 = 9216	2560 x 8 = 20480
NVIDIA Tensor-Kerne	Insgesamt	576 x 2 = 1152	320 x 8 = 2560
Kerntaktgeschwindigkeit		1395 MHz	585 MHz
Boost-Taktfrequenz		1770 MHz	1590 MHz
Anzahl der Transistoren		18,600 Millionen	13,600 Millionen
Thermische Entwurfsleistung (TDP)		260 Watt	70 Watt
Schnittstelle		PCIe 3.0 x16	PCIe 3.0 x16
Speicherart		GDDR6	GDDR6
Maximale RAM-Menge	Insgesamt	48GB x 2 = 96GB	16GB x 8 = 128GB
Speicherbusbreite		384 Bit	256 Bit
Speichertaktfrequenz		14000 MHz	10000 MHz
Speicher-Bandbreite		672.0 GB/s	320.0 GB/s

Abb.3 Obwohl es Unterschiede in der Konfiguration der AWS Bare Metal-Instanz und der HP Z8 G4-Workstation gibt, zeigen sie eine ähnliche Leistung.

Die Ausführung der letzten Iteration der Experimente auf der AWS-Instanz und dem HP G8 Z4 dauert ähnlich lange: 8 Stunden und 30 Minuten. Werfen wir einen Blick auf die AWS-Preise. Für unseren Vergleich wählen wir die On-Demand-Preise, da diese Strategie die Flexibilität der Cloud unterstreicht.

Der Preis für eine AWS g4dn.metal Instanz für eine einzige Stunde beträgt 9,78 Dollar. Die Durchführung einer Iteration des Experiments dauerte mehr als 8 Stunden. Es lässt sich leicht nachweisen, dass der Endpreis für die Entwicklung unseres NER-Modells 106,5 Dollar betragen hätte, d. h. 9 Stunden multipliziert mit 9,78 Dollar pro Stunde, zuzüglich der üblichen Mehrwertsteuer von 21 %.

8 Stunden und 30 Minuten

Betrachten wir nun den gesamten Workflow. Unser abschließender Durchlauf dauert 8 Stunden und 30 Minuten, erforderte aber eine Menge Optimierungsarbeit. Für unsere allererste Iteration benötigten wir 23 Stunden. Im Laufe von 20 Iterationen gelang es uns, die Laufzeit auf 8 Stunden und 30 Minuten zu verkürzen. Insgesamt haben wir 320 Stunden ((23h +9h)/2 * 20 Iterationen) an Berechnungen mit dem HP Z8 G4 Rechner durchgeführt. Bei Verwendung der oben erwähnten AWS g4db.metal Instanz hätte uns das 3.129,6 Dollar ohne Mehrwertsteuer und 3.789,82 Dollar inklusive Mehrwertsteuer gekostet; außerdem sind darin keine zusätzlichen Kosten für Netzwerke, Speicher oder zusätzliche Sicherheitsebenen enthalten.

Vergleich

Der HP Z8 G4 ist nicht günstig. Ohne nennenswerte Rabatte seitens HP kostet die Konfiguration der verwendeten Data Science Workstation rund 30.000 Dollar. Auch der Betrieb der Workstation ist nicht kostenlos. Der HP Z8 G4 verbraucht 1.450 Watt pro Stunde. Legt man die Stromkosten in den Niederlanden zum Zeitpunkt des Experiments zugrunde, ergibt sich für die gesamte Laufzeit von 320 Stunden eine Stromrechnung von 53,36 Dollar. Im Großen und Ganzen ist dies jedoch zu vernachlässigen. Schließlich erfordert ein eigener Desktop eine gewisse Wartung. Nehmen wir an, die Wartungskosten betragen 300 Dollar pro Monat. Die Anfangsinvestition für den Betrieb des Experiments vor Ort beträgt 30.000 Dollar für die Maschine plus 300 Dollar für die Wartung und 53 Dollar für Strom, insgesamt also 30.353 Dollar.

Die Nutzung von AWS für einen 320-Stunden-Lauf kostet 3.790 Dollar. Einschließlich der notwendigen Kosten für andere Dienste wie Speicherplatz und Vernetzung errechnen wir einen monatlichen Endpreis von 4,927 Dollar. Es ist wichtig zu erkennen, dass Data Scientists in einem realen Szenario nicht einfach nach einem Monat aufhören würden, Experimente durchzuführen. Sie würden die Modelle weiter verfeinern und neue einführen. Das würde bedeuten, dass sie unser Experiment Monat für Monat durchführen könnten und jedes Mal 320 Stunden an Berechnungen benötigen würden.

Bei Verwendung der HP Z8 G4 Workstation belaufen sich die Gesamtkosten für die Durchführung des Experiments über einen Zeitraum von sieben Monaten einschließlich der anfänglichen Hardware-Investitionen, der Wartung und der Stromkosten auf 32.471 Dollar. Die Nutzung der AWS g4dn.metal Instanz für sieben Monate würde 34.489 Dollar kosten. Anders ausgedrückt: Für ein dediziertes Data Science-Projekt würde es nur 7 Monate dauern, bis sich die Investition in einen HP Z8 G4 Desktop auszahlt. Ab diesem Zeitpunkt werden die Gewinne nur noch größer. Die Durchführung von Experimenten für ein ganzes Jahr hätte 34.236 Dollar gekostet. Hätten Sie sich für die Nutzung der Cloud anstelle einer lokalen Workstation entschieden, könnten die Kosten leicht auf fast 60.000 Dollar ansteigen.

Wenn das Data Science-Team die Workstation zwei Jahre lang bis zur maximalen Auslastung nutzt, würden die TCO immer noch nicht über 40.000 Dollar liegen (38.472 Dollar, um genau zu sein). Die Nutzung der Cloud für eine ähnliche Zeitspanne kostet 118.248 Dollar.

Monat	Einsatz einer HP Z8 G4 Maschine				Einsatz der AWS g4dn.metal On-Demand-Instanz		Einsatz der AWS g4dn.metal On-Demand-Instanz (+ andere Dienste)	
	Kosten der Maschine (HPZ8)	Wartung (HPZ8)	Stromkosten (HPZ8)	Kumulativ (HPZ8)	Monatlicher Tarif (AWS)	Kumulativ (AWS)	Monatlicher Preis (AWS+30 %)	Kumulativ (AWS+30 %)
1	30,000	300	53	30,353	3,790	3,790	4,927	4,927
2		300	53	30,706	3,790	7,580	4,927	9,854
3		300	53	31,059	3,790	11,370	4,927	14,781
4		300	53	31,412	3,790	15,160	4,927	19,708
5		300	53	31,765	3,790	18,950	4,927	24,635
6		300	53	32,118	3,790	22,740	4,927	29,562
7		300	53	32,471	3,790	26,530	4,927	34,489
8		300	53	32,824	3,790	30,320	4,927	39,416
9		300	53	33,177	3,790	34,110	4,927	44,343
10		300	53	33,530	3,790	37,900	4,927	49,270
11		300	53	33,883	3,790	41,690	4,927	54,197
12		300	53	34,236	3,790	45,480	4,927	59,124
13		300	53	34,589	3,790	49,270	4,927	64,051
14		300	53	34,942	3,790	53,060	4,927	68,978
15		300	53	35,295	3,790	56,850	4,927	73,905
16		300	53	35,648	3,790	60,640	4,927	78,832
17		300	53	36,001	3,790	64,430	4,927	83,759
18		300	53	36,354	3,790	68,220	4,927	88,686
19		300	53	36,707	3,790	72,010	4,927	93,613
20		300	53	37,060	3,790	75,800	4,927	98,540
21		300	53	37,413	3,790	79,590	4,927	103,467
22		300	53	37,766	3,790	83,380	4,927	108,394
23		300	53	38,119	3,790	87,170	4,927	113,321
24		300	53	38,472	3,790	90,960	4,927	118,248



Kumulative Kosten, USD

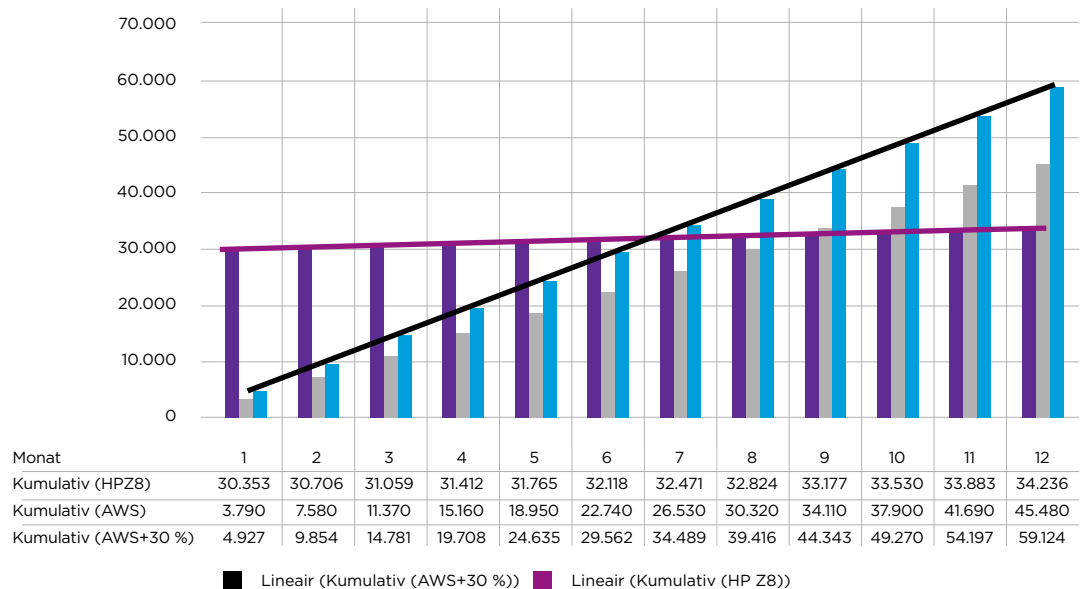


Abb.5: Der Kostenanstieg bei der Nutzung der Cloud im Vergleich zur Nutzung eines dedizierten Arbeitsplatzes ist offensichtlich. Die Erstinvestition in eine Data-Science-Maschine macht sich je nach Nutzung nach etwa 7 Monaten bezahlt.

Schlussfolgerung

Es ist offensichtlich, dass die öffentliche Cloud einen Platz in der Data Science hat. Wäre unser Experiment einmalig durchgeführt worden, wäre die AWS-Instanz eine interessante Lösung. Wenn Sie jedoch planen, regelmäßig Data Science-Aufgaben auszuführen, sollte sich der HP Z8 G4 nach etwa 7 Monaten auszahlen. Je intensiver die Nutzung, desto schneller werden Sie im Vergleich zur Cloud Geld sparen. In unserem Experiment wurde eine monatliche Laufzeit von 320 Stunden als Basiswert verwendet, basierend auf realem Training und unter Verwendung von NER-Modellen. Es ist sicherlich möglich, den Einsatz der Maschine im täglichen Betrieb weiter zu maximieren.

Unternehmen, die Data Science ernst nehmen, sollten sich von den anfänglichen Investitionen in eine dedizierte Workstation nicht abschrecken lassen. Wie dieses Experiment demonstriert, amortisiert sich die Investition in einen HP Z8 G4 im Vergleich zur Cloud bereits nach 7 Monaten. In zwei Jahren werden die Gesamtkosten der Cloud fast 120.000 Dollar erreichen. Für diesen Betrag können Sie vier HP Z8 G4 erwerben, die die vierfache Leistung erbringen, und Sie würden immer noch einen Gewinn erzielen.

Kontakt

Weitere Informationen zu HP, HP Produkten und Services sowie zum Support finden Sie auf unserer Website:

Workstationspecialist.de





Jetzt beraten lassen!



Ihren persönlichen Ansprechpartner

Markus Hömig-Heidrich

erreichen Sie telefonisch unter:

+49 221.588 320 31

E-Mail:

hoemig@workstation-pro.com

WORKSTATIONPRO
Eine gute Entscheidung.

WORKSTATIONPRO GmbH
Vorgebirgstraße 59
50677 Köln
www.workstation-pro.com

