Università degli Studi di Catania
Corso di Laurea Magistrale in Informatica

Riccardo Raciti

Machine Learning in Cloud

Relazione Sistemi Centrali

Anno Accademico 2022–2023

Indice

Introduzione

L'oggetto di questa relazione saranno le tecnloige Cloud Services presenti sul mercato per l'ottimizzazione e la parallelizzazione di modelli di Machine Learning. La relazione sarà suddivia in capitoli, macroargomenti, e sottocapitoli, microargomenti.

Nel primo capitolo della relazione saranno fornite nozioni di base riguardanti il machine learning e i tipi di apprendimento che si possono intraprendere. Dopo di chè saranno presentate le due euristiche che guidano ad oggi il mercato e la scelta dell'architettura aziendale ovvero la scelta di una tecnologia on-premise o cloud, approfondendo i vantaggi e gli svantaggi di entrambe le soluzioni e si parlerà in maniera sommaria dei possibili costi di un'infrastruttura.

Il Cloud Services offre differenti tipi di servizi che sono: Saas, Paas e Iass. Essi verrano trattati e spiegati fornendone i vantaggi e svantaggi.

Dopo di ciò verranno presentate le aziende leader sul mercato AWS, Azure e GCP.

Per ognuna di esse verranno riportati i servizi disponibili per lo sviluppo di progetti di machine learning, spiegando il funzionamento di alcuni di essi e presentando dei WhitePaper, paper aziendali, delle aziende sopra citate per far vedere in concreto le possibilità nel'utilizzo di queste tecnologie. Non verranno trattati i prezzi e gli abbonamenti dei servizi poichè sarebbe solo una trascrizione delle tabelle dei servizi e non è lo scopo di questa relazione.

Alla fine della relazione verrà mostrato in concreto con una piccola guida su come poter parallelizzare il carico di lavoro tramite il servizio azure, parlando del parallelismo del modello e dei dati e del training distribuito. Di questo ultimo si vedranno le possibili implementazioni tramite le librerie supporta-

te ovvero PyTorch e TensorFlow, specificando alcune soluzioni possibili che offrono le librerie in collaborazione con Azure.

1 Machine Learning

Il Machine Learning (ML) è un sottoinsieme dell'intelligenza artificiale (AI) che si occupa di creare sistemi che apprendono o migliorano le performance in base ai dati che utilizzano. Intelligenza artificiale è un termine generico e si riferisce a sistemi o macchine che imitano l'intelligenza umana, simulando il funzionamento del cervello umano. I termini machine learning e AI vengono spesso utilizzati insieme e in modo interscambiabile, ma non hanno lo stesso significato. Un'importante distinzione è che sebbene tutto ciò che riguarda il machine learning rientra nell'intelligenza artificiale, l'intelligenza artificiale non include solo il machine learning.
Il machine learling si suddivide in due sottofamiglie in base al tipo di apprendimento che si effettua.

1.1 Appredinmento Supervisionato

Gli algoritmi di machine learning supervisionato sono i più utilizzati. Con questo modello, un data scientist agisce da guida e insegna all'algoritmo i risultati da generare. Esattamente come un bambino impara a identificare i frutti memorizzandoli in un libro illustrato, nel machine learning supervisionato l'algoritmo apprende da un set di dati già etichettato, chiamato Traning Set, e con un output predefinito.
La difficoltà in questo approccio sta nel etichettare correttamente ogni item con le rispettive label, si immagini un dataset formato da 1000 foto, che per i problemi di machine learning è un numero esiguo, se questo dataset fosse di automobili e l'obbietivo fosse quello di riconoscere il modello delle automobili data una foto il programmatore dovrebbe uno ad uno etichettare tutte le foto con l'opportuno modello dell'autovettura.

Gli algoritmi di regressione lineare e logistica, di classificazione multiclasse e le macchine a vettori di supporto sono alcuni esempi di machine learning supervisionato.

1.2 Apprendimento non Supervisioanto

Il machine learning non supervisionato utilizza un approccio più indipendente, in cui un computer impara a identificare processi e schemi complessi senza la guida attenta e costante di una persona. Il machine learning non supervisionato implica una formazione basata su dati privi di etichette e per i quali non è stato definito un output specifico.

Per continuare a utilizzare l'analogia precedente, il machine learning non supervisionato è simile a un bambino che impara a identificare le macchine osservando le forme e gli schemi, anziché memorizzando i nomi con l'aiuto di un insegnante. Il bambino cercherà le somiglianze tra le immagini e le suddividerà in gruppi, assegnando a ciascun gruppo la nuova etichetta corrispondente.

Gli algoritmi di clustering k-means, l'analisi di componenti principali e indipendenti e le regole di associazione sono esempi di machine learning non supervisionato.

1.3 Apprendimento Misto

Vi sono due metodi di Apprendimento che mischiano i due metodi sopra citati.

1.3.1 Apprendimento Semi Supervisionato

Ha le stesse applicazioni dell'apprendimento supervisionato. Ma per l'addestramento utilizza dati classificati e non: solitamente di un ridotto volume di dati classificati e un più ampio volume di dati non classificati (perchè acquisire questi ultimi è più economico e meno faticoso). Questo tipo di apprendimento può essere utilizzato con metodi di classificazione, regressione e previsione. L'apprendimento semi supervisionato è utile se la classificazione ha un costo troppo alto per permettere un processo di apprendimento completamente supervisionato. Un esempio recente sono le fotocamere capaci di identificare il volto delle persone.

1.3.2 Apprendimento per Rinforzo

Spesso viene usato in robotica, videogiochi e navigazione. Con l'apprendimento per rinforzo l'algoritmo scopre da quali azioni vengono generate le ricompense maggiori, passando per esperimenti ed errori. Questo tipo di apprendimento presenta tre componenti principali: l'agente (chi impara o prende decisioni), l'ambiente (tutto ciò con cui l'agente interagisce) e le azioni (cosa può fare l'agente). L'obiettivo dell'agente è scegliere quelle azioni che massimizzano la ricompensa prevista in un determinato lasso temporale. Scegliendo le azioni giuste, l'agente raggiungerà l'obiettivo più velocemente. Quindi l'obiettivo dell'apprendimento per rinforzo è quello di imparare quali sono le azioni migliori da attuare.

2 Infrastruttura

Tutto qeullo di cui si è discusso riguarda solo una parte del mondo del Machine Learning, un altro aspetto importante da tenere in considerazione è l'hardware necessario per poter esugire i task esposti precedentemente e il modo in cui li si può condividere tra più utilizzatori.

2.1 On-Premise

Come dice il termine stesso, sono soluzioni installate sui computer di ogni singolo utente o su un server aziendale. Il programma poi funziona a licenza e se ne può usufruire in base all'accordo stipulato.
Una volta che i software sono stati installati, è importante tenerli sempre aggiornati, infatti vengono rilasciati update frequenti fondamentali per migliorarne la stabilità e prevenire o affrontare minacce alla sicurezza.
Sfruttando una soluzione On-premise l'utente ha a disposizione tutti i dati, che vengono custoditi nel proprio server sotto il suo diretto controllo.

2.1.1 Vantaggi

- Una volta acquistata, una soluzione On-premise non prevede spese aggiuntive nel periodo in cui viene utilizzata.
- Si possono gestire gli aggiornamenti delle applicazioni.
- Per questioni legate alla privacy i dati saranno presenti nel sistema e sempre in possesso dell'utente, poichè non vi è il rischio che vengano divulgati e si può decidere quali sistemi difensivi adottare.

2.1.2 Svantaggi

- Nelle aziende più grandi potrebbero essere necessarie numerose licenze per rendere accessibile a tutti un determinato programma, diventando così una spesa non indifferente e piuttosto onerosa.

- L'aggionamento delle applicazioni può essere un'arma a doppio taglio poichè sarà svolta dagli impiegati dell'azienda stessa, quindi aumenta la loro mole di lavoro, oppure in aziende più grandi richiede una figura esperta.

- Spesso avviene che le aziende sviluppatrici piuttosto che fornire nuovi aggiornamenti, decidono di proporre direttamente un nuovo software, spingendo le aziende all'acquisto e a sostenere un'ulteriore spesa.

- Le aziende devono garantire la sicurezza dei dati storati nei loro server, questo avviene tramite acquisti, abbastanza onerosi, di licenze per la sicurezza, inoltre in caso di fuga di dati o di attacchi esterni ne risponderà l'azienda stessa.

2.2 Cloud

Le soluzioni in Cloud sono dei servizi che non prevedono alcuna installazione fisica nei dispositivi, ma sono gestiti da remoto e accessibili via Internet attraverso una piattaforma virtuale.
Sono quindi offerti da un provider e tenuti sotto controllo da un team tecnico esterno.
Con il Cloud il software è ospitato in un data center e tutto ciò che riguarda aggiornamenti, sviluppo e manutenzione è affidato al personale specializzato.
I dati, quindi, non sono fisicamente presenti in azienda, ma vi si può accedere in qualunque momento grazie a una connessione a Internet.

2.2.1 Vantaggi

- Non serve disporre di computer potenti o dispositivi all'avanguardia, il servizio funziona totalmente da remoto, garantendo sempre il massimo delle prestazioni grazie a server specializzati.

- La grossa mole di lavoro a cui l'utente può rinunciare, scaricando le mansioni sul personale tecnico, poichè l'host mette a disposizione un team specializzato incaricato di occuparsi di ogni dettaglio in modo competente e professionale, lasciando l'utente finale molto più libero e con meno responsabilità.

- L'azienda non è più responsabile dei dati, quindi gli eventuali provvedimenti ricadranno sul provider, anche se oggi in Europa si discute ancora molto su questa tematica.

2.2.2 Svantaggi

- Per fruire del servizio è fondamentale una rete Internet, e in caso di linea debole o interruzione della connessione, non è più possibile accedere al software.

- I dati non saranno più presenti fisicamente in azienda e non si avrà più il controllo diretto sulle informazioni sensibili, è importante notare però che il provider fornirà tutte le garanzie, quinti attestati e licenze, che garantiscono la sicurezza dei dati.

3 Architettura Necessaria

Un altro importante fattore da analizzare quando si parla di Machine Learning riguarda i costi per effettuare il training dei modelli.
Per training si intende quella fase in cui il modello viene allenato, ottimizzato, per il suo task.
In questa fase, durante gli anni, ci si è resi conto che il miglior modo per effettuare il training è spostare questa particolare fase sulla GPU (Graphics Processing Unit), si consideri che non si parla di schede video per i consumatori comuni ma è Hardware ottimizzato per il machine learning, il cui costo è elevato.

3.0.1 Memoria

Nei contesti del machin learning si lavora con una grande mole di dati entrando in quello che viene chiamato Big Data.
Non vi è però solo la necessità di poter storare questi dati, quindi infrastrutture capaci di contenere tera e tera di dati, ma bisogna gestire anche la condivisione in tempo reale di questi dati per i vari sviluppatori che lavorano al progetto.

3.0.2 RAM

Per poter trasfetire i dati velocemente dalla memoria centrale, server, al computer per poi passarlo alla GPU, vi è la necessità di molta RAM con prestazioni performanti. Queste caratteristiche richieste fanno si che il prezzo di tale hardware sia alto, considerando la quantità richiesta di RAM. In più nel caso in cui ad ogni sviluppatore si vuole garantire quella potenza bisogna

tener conto che la spesa dovrà essere moltiplicata per tutti gli sviluppatori non essendo una spesa globale come quella dello storage nel server centrale.

3.0.3 Schede Video

Per portare un esempio analizziamo le schede video proposte da INVIDIA per il machine learning. INVIDIA propone due schede video una nella forma base e un'altra nella forma pro.

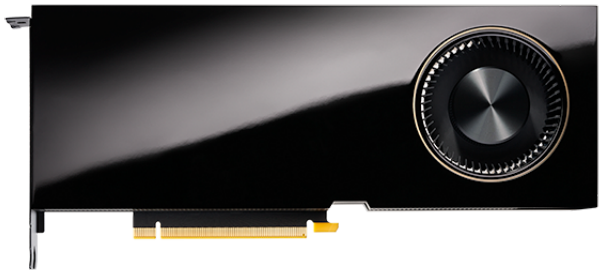

Figura 1: NVIDIA RTX 6000

La scheda video in figura è NVIDIA RTX 6000, con la seguente scheda tecnica:

GPU Features	NVIDIA RTX™ A6000
GPU Memory	48 GB GDDR6 with error-correcting code (ECC)
Display Ports	4x DisplayPort 1.4a*
Max Power Consumption	300 W
Graphics Bus	PCI Express Gen 4 x 16
Form Factor	4.4" (H) x 10.5" (L) dual slot
Thermal	Active
NVLink	2-way low profile (2-slot and 3-slot bridges) Connect 2 RTX A6000
vGPU Software Support	NVIDIA vPC/vApps, NVIDIA RTX Virtual Workstation, NVIDIA Virtual Compute Server
vGPU Profiles Supported	**See the Virtual GPU Licensing Guide**
VR Ready	Yes

Figura 2: Scheda tecnica NVIDIA RTX 6000

Questa scheda video offerta da Invidia è la versione base, ed è venduta ad un prezzo di 4.650$, mentre la versione pro che è ancora più efficente in termini di velocità e memoria è venduta a 6.800$.

Questo è solo il prezzo di una singola schedavideo, nella realtà di un'azienda che decide di occuparsi di machine learning si capisce da subito che i costi lievitano, è importante considerare pure i costi dei server dove memorizzare i dati e cosa non meno importante i costi di manutenzione e mantenimento di suddetti dispositivi, i quali consumano non poca corrente elettrica.

4 Clous Services

4.1 Saas

SaaS, Software as a Service (Il Software come un Servizio), consente agli utenti di connettersi ad app basate sul cloud tramite Internet e usare tali app. Esempi comuni sono la posta elettronica, i calendari e gli strumenti di produttività.

Il modello SaaS offre una soluzione software completa che si può acquistare con pagamento in base al consumo da un provider di servizi cloud.

Si noleggia l'uso di un'app per l'organizzazione e gli utenti si connetteranno all'app tramite Internet, in genere con un Web browser.

L'infrastruttura sottostante, il middleware, il software delle app e i dati delle app si trovano tutti nel data center del provider di servizi.

Il provider di servizi gestisce l'hardware e il software e, con il contratto di servizio appropriato, garantisce la disponibilità e la sicurezza dell'app e dei dati. Il modello SaaS consente all'organizzazione di essere rapidamente operativa con un'app con costi iniziali minimi.

4.1.1 Vantaggi

- Non si devono acquistare, installare, aggiornare o gestire hardware, middleware o software.
- Le applicazioni aziendali più sofisticate, come ERP e CRM, diventano abbordabili per le organizzazioni che non dispongono delle risorse per acquistare, distribuire e gestire l'infrastruttura e il software necessari.
- Il pagamento è proporzionato all'utilizzo, si può anche risparmiare in quanto il servizio SaaS offre scalabilità verticale automatica in base al livello di utilizzo.
- L'uso del software da parte del client è gratuito, gli utenti potranno eseguire la maggior parte delle app SaaS direttamente dal Web browser senza che sia necessario scaricare e installare alcun software, sebbene alcune app richiedano plug-in. Questo significa che si devono acquistare e installare software speciale per gli utenti.
- Il modello SaaS permette di garantire in modo semplice la mobilità alla forza lavoro perché gli utenti possono accedere a dati e app SaaS da qualsiasi dispositivo mobile o computer connesso a Internet, non ci si deve preoccupare di sviluppare app per l'esecuzione in tipi di computer e dispositivi diversi, poichè predispone tutto il provider di servizi.
- Non serve personale con un'esperienza speciale per gestire le problematiche di sicurezza correlate all'uso di dispositivi mobili.
- Una scelta attenta del provider di servizi garantisce la sicurezza dei dati, indipendentemente dal tipo di dispositivo in cui vengono utilizzati.

- Si può accedere ai dati delle app da qualunque luogo, con i dati archiviati nel cloud, gli utenti possono accedere alle informazioni da qualsiasi dispositivo mobile o computer connesso a Internet e quando i dati di un'app vengono archiviati nel cloud, non vanno persi in caso di problemi al computer o al dispositivo di un utente.

4.1.2 Svantaggi

- L'utilizzo dei SaaS implica il trasferimento dei dati dell' azienda a un provider di servizi.
- L'upload dei file aziendali sui server di un'altra società può comportare due criticità:
  - Il provider potrebbe non adottare una policy di riservatezza adeguata;
  - Potrebbe non avere un'infrastruttura sicura. Per questi motivi è di fondamentale importanza scegliere un fornitore di servizi trasparente e affidabile, per evitare furti o accessi non autorizzati ai dati.
- Anche in questo scenario la connessione ad internet deve essere sempre attiva, i Software as a Service necessitano di una connessione ad internet veloce e sempre attiva.
- Vi sono alcuni fornitori, come ad esempio Google, che permettono di utilizzare i software anche in modalità offline, sincronizzando i dati non appena la connessione sarà nuovamente attiva. Si tratta di una soluzione che deve essere presa in considerazione in casi di emergenza, poiché spesso le versioni offline dei software hanno meno funzionalità.
- La cessazione dei servizi da parte del provider è uno dei principali svantaggi delle soluzioni cloud.
- I Software as a Service necessitano di un'infrastruttura informatica sempre attiva per essere utilizzati. Nel caso in cui il provider cessi la sua attività, l'applicazione non sarebbe più accessibile. Per evitare possibili problemi di questo tipo, è necessario affidarsi a partner affidabili.

4.2 PaaS

Il modello PaaS, Platform as a Service (piattaforma distribuita come servizio), viene usato per ovviare a questi problemi.
Una piattaforma distribuita come servizio (PaaS, Platform as a Service) è un ambiente di sviluppo e distribuzione completo nel cloud, con risorse che consentono di distribuire qualsiasi cosa, da semplici app basate sul cloud ad applicazioni aziendali sofisticate abilitate per il cloud. Si possono acquistare le risorse necessarie da un provider di servizi cloud con pagamento in base al consumo e accedervi tramite una connessione Internet sicura.
Il modello PaaS consente di evitare le spese e la complessità legate all'acquisto e alla gestione di licenze software, middleware, agenti di orchestrazione di contenitori come Kubernetes o strumenti di sviluppo e altre risorse. L'utente gestisce le applicazioni e i servizi che sviluppa e in genere il provider di servizi cloud gestisce tutto il resto.

4.2.1 Vantaggi

L'utilizzo della piattaforma PaaS offre innumerevoli vantaggi agli sviluppatori.

- Lo sviluppo di applicazioni è più facile e veloce, poiché non serve realizzare e gestire una propria infrastruttura, ciò permette di abbassare i costi, permettendo di far arrivare i prodotti prima sul mercato.
- Le prestazioni sono scalabili, cioè è possibile ampliare o diminuire in modo flessibile le capacità richieste a seconda delle proprie esigenze.
- Il fatto che sia il fornitore a occuparsi dell'infrastruttura è un vantaggio poichè non subbentreranno i costi per la realizzazione dell'infrastrut-

tura, per le manutenzioni, per gli aggiornamenti e/o per l'acquisto di nuove licenze software.

4.2.2 Svantaggi

- Poichè è il fornitore a occuparsi della fornitura non si ha alcun controllo sull'infrastruttura e non si può implementare da soli le feature. Inoltre sono utilizzabili solo i linguaggi di programmazione e i tool messi a disposizione dal fornitore.

- Il progetto è più o meno vincolato all'ambiente di sviluppo scelto, nonostane sia possibile far migrare un progetto piccolo. Nel caso di applicazioni più grandi il codice non si può sempre riportare per intero durante il trasferimento su un'altra piattaforma e in questi casi deve essere invece riscritto daccapo.

- Anche se ad oggi potrebbe essere considerato un problema banale, bisogna considerare che è provvidenziale affidarsi ad aziende già presenti sul mercato e affermate poichè in caso di fallimento di esse tutto il lavoro verrebbe perso.

4.3 Iaas

IaaS (Infrastructure as a service) è un tipo di servizio di cloud computing che offre risorse di calcolo, archiviazione e rete essenziali on demand e con pagamento in base al consumo.
La migrazione dell'infrastruttura dell'organizzazione a una soluzione IaaS contribuisce alla riduzione della manutenzione dei data center locali, al risparmio di denaro sui costi dell'hardware e al recupero di informazioni dettagliate aziendali in tempo reale. Le soluzioni offrono la flessibilità necessaria per dimensionare le risorse IT in base alla domanda.
Consentono anche di effettuare rapidamente il provisioning di nuove applicazioni e aumentare l'affidabilità dell'infrastruttura sottostante.
IaaS permette di evitare il costo e la complessità dell'acquisto e della gestione di server fisici e dell'infrastruttura del data center. Ogni risorsa viene offerta come componente di servizio distinto e si deve pagare per una risorsa specifica solo per il periodo di tempo in cui è necessaria.

4.3.1 Vantaggi

- Le risorse dei server cloud nel data center sono condivise da tutti gli utenti, il prezzo è molto più basso per ogni singola azienda che utilizza tali risorse.

- Non è necessario scaricare e installare i prodotti software sui singoli dispositivi, ne consegue che il personale tecnico non deve essere impegnato su compiti relativi all'installazione o all'aggiornamento del software.

- Altro vantaggio è l'affidabilità, poiché le richieste di rete sono distribuite su numerosi server e, in alcuni casi, su diverse sedi fisiche, i fornitori

di servizi possono garantire tempi di attività maggiori e ridondanze nelle prestazioni assicurando minori possibilità di interruzioni del servizio per l'azienda, ma anche per i clienti finali.

- Le risorse di calcolo, storage e rete sono distribuite su richiesta, questo è un pregio in ottica di scalabilità. Gli utenti pagano solo per le risorse di cui hanno bisogno in quel tal momento, potendo aumentarle o diminuirle in modo rapido e mirato.

- L'accessibilità può essere aumentata e ridotta, le applicazioni e i software possono essere aggiunti e rimossi e la capacità dell'infrastruttura può crescere secondo le necessità. Questo è particolarmente prezioso per le aziende che devono affrontare fluttuazioni stagionali della domanda.

- La manutenzione dell'infrastruttura IT in sede richiede competenze interne e costi di assistenza continui, le risorse virtuali possono, invece, essere fornite ovunque e in tempi rapidi, garantendo così doti di flessibilità significative.

- Le risorse basate sul cloud sono protette dalle interruzioni di servizio e dai guasti che possono colpire l'hardware fisico presente in sede, un vantaggio in termini di stabilità e sicurezza, dato che dei termini di servizio fanno parte anche la continuità operativa e il disaster recovery.

4.3.2 Svantaggi

- Sebbene la mancanza di responsabilità relative alla presenza fisica o meno degli hardware necessari permetta di risparmiare sui costi e sull'impegno, questa caratteristica presenta anche uno svantaggio, come utenti non si ha infatti alcun influsso né sulla disponibilità del servizio, né sulla funzionalità delle singole componenti.

- Anche in materia di sicurezza e protezione dei dati non si avrà la situazione in pugno.

- Un ulteriore svantaggio è che, sebbene con IaaS sia possibile passare da un provider a un altro in qualsiasi momento, la mancanza di standard e interfacce unitarie rendono il passaggio molto scomodo e difficoltoso.

- Vi sono possibili problemi legati alle linee guide sulla protezione dei dati per via del luogo dove si trovano i server del provider.
  Nel caso in cui si volesse effetturare un cambio di provider non è detto che lo si possa fare con facilità, potendo trasferire tutti i dati e task.

- Essenziale la connessione ad internet, senza essa non si usuffruire del servizio.

4.4 Conclusioni sul Cloud Services

Il funzionamento nonchè i vantaggi e gli svantaggi delle tre componenti del Cloud Services sono molto simili tra loro, poichè le tecnologie presentate che formano una struttura piramidale, come mostrato in figura.

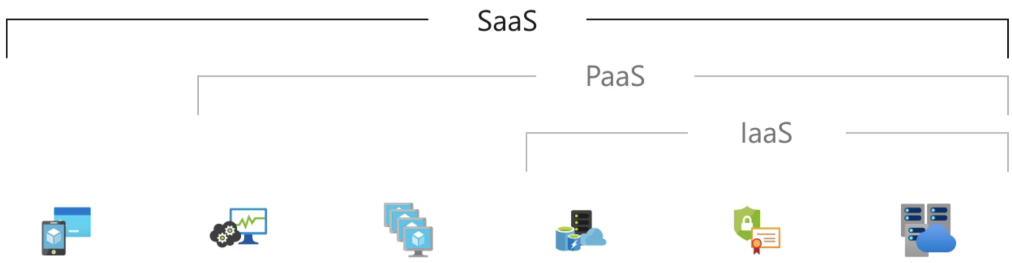

Figura 3: Struttura Cloud Services

5 Aziende sul mercato

5.1 AWS

Amazon Web Services, AWS, è la piattaforma cloud più utilizzata nel mondo, offre più di 200 servizi completi da data center a livello globale.
Milioni di sviluppatori, start-up in rapida crescita, le grandi aziende e le agenzie governative leader di settore, utilizzano AWS per diminuire i costi, diventare più agili e innovarsi in modo più rapido.

Figura 4: Logo AWS

AWS offre tantissimi servizi tra i quali:

- Amazon SageMaker;
- Amazon Augmented AI;
- Amazon CodeGuru;
- Amazon CodeWhisperer;
- Amazon Comprehend;
- Amazon Comprehend Medical;
- Amazon DevOps Guru;

- Amazon Elastic Inference;
- Amazon Forecast;
- Amazon Fraud Detector;
- Amazon HealthLake;
- **...** .

In realtà la lista non si fermerebbe qui ma vi sono altri 25 servizi, oltre quelli nominati sopra, per il machine learning messi a disposizione da AWS.

5.1.1 SageMaker

Uno dei servizi più utilizzati nel machine learning con AWS è **SageMaker**. Amazon SageMaker è un servizio completamente gestito per preparare i dati e costruire, addestrare e implementare modelli di machine learning (ML) impiegando infrastrutture, strumenti e flussi di lavoro completamente gestiti.
Amazon SageMaker è progettato per offrire disponibilità elevata. Non sono previsti finestre di manutenzione o tempi di inattività pianificati. Le API SageMaker vengono eseguite nei data center sicuri e ad elevata disponibilità di Amazon, con replica dello stack di servizi configurata fra tre strutture in ogni regione AWS per offrire tolleranza ai guasti in caso di inattività di un server o di interruzione del servizio nella zona di disponibilità.
Amazon SageMaker crittografa gli artefatti dei modello di ML e di altri elementi del sistema, sia inattivi sia in transito. Le richieste all'API e alla console di SageMaker vengono inoltrate tramite una connessione sicura (SSL). Ci si può avvalere deel ruolo di AWS Identity and Access Management in SageMaker per fornire le autorizzazioni per accedere alle risorse per l'addestramento e l'implementazione manualmente. È possibile utilizzare bucket

Amazon Simple Storage Service (Amazon S3) crittografati per dati e artefatti di modello, nonché applicare una chiave del Servizio di gestione delle chiavi AWS (AWS KMS) a notebook di SageMaker, processi di addestramento ed endpoint per crittografare il volume di archiviazione dedicato al ML collegato. Amazon SageMaker supporta inoltre il cloud privato virtuale (VPC) di Amazon e AWS PrivateLink.

5.1.2 Amazon SageMaker Autopilot

Amazon SageMaker Autopilot è la prima funzionalità automatica di machine learning del settore che offre controllo e visibilità totali sui modelli di ML. SageMaker Autopilot esamina automaticamente i dati non elaborati, applica i processori di caratteristiche, sceglie il miglior set di algoritmi, addestra e ottimizza diversi modelli, controlla le loro prestazioni e in base a queste stila una classifica dei modelli, il tutto in pochi clic. Il risultato è un modello con le migliori prestazioni possibili, che si può distribuire in una frazione del tempo normalmente richiesto per la formazione del modello. Si ha piena visibilità su come il modello è stato creato, cosa c'è al suo interno e SageMaker Autopilot si integra con Amazon SageMaker Studio. Si possono ricercare fino a 50 modelli diversi generati da SageMaker Autopilot all'interno di SageMaker Studio, facilitando la scelta del miglior modello per il caso d'uso. SageMaker Autopilot può essere utilizzato per produrre facilmente un modello anche da chi non possiede esperienza di machine learning e da sviluppatori esperti per sviluppare un modello di base per le successive iterazioni del team.
Amazon SageMaker JumpStart include oltre 150 modelli open source precedentemente addestrati da PyTorch Hub e TensorFlow Hub. Per attività visive come classificazioni di immagini e rilevamento di oggetti, è possibile usare modelli come ResNet, MobileNet e Single-Shot Detector (SSD).

Per attività testuali come classificazioni di frasi, classificazione di testi e risposte a domande è possibile utilizzare modelli come BERT, RoBERTa e DistilBERT.

SageMaker JumpStart offre soluzioni preconfigurate con tutti i servizi AWS necessari per avviare una soluzione in produzione. Le soluzioni sono totalmente personalizzabili, in modo da poterle modificare in base al proprio specifico caso d'uso e set di dati.

5.1.3 Amazon SageMaker Data Wrangler

Amazon SageMaker Data Wrangler riduce il tempo richiesto per l'aggregazione e la preparazione dei dati per il ML.

Fornisce una singola interfaccia in Amazon SageMaker Studio, con essa è possibile cercare e importare dati da Amazon S3, Amazon Athena, Amazon Redshift, AWS Lake Formation, Amazon SageMaker Feature Store e Snowflake facilmente.

Inoltre, è possibile interrogare e importare dati trasferiti da oltre 40 origini dei dati e registrati nel Catalogo dati AWS Glue da Amazon AppFlow.

5.1.3.1 Visualizzazione dei dati

SageMaker Data Wrangler carica, aggrega e visualizza automaticamente i dati grezzi. Dopo avere importato i dati in SageMaker Data Wrangler, è possibile visualizzare istogrammi e riepiloghi colonnari generati automaticamente.

Per analizzare i dati in maggiore dettaglio e identificare errori potenziali, è possibile utilizzare il report sulla qualità dei dati e relativi approfondimenti di SageMaker Data Wrangler, che fornisce statistiche di riepilogo e segnalazioni relative alla qualità dei dati. Inoltre, è possibile eseguire l'analisi

delle distorsioni supportata da Amazon SageMaker Clarify direttamente da SageMaker Data Wrangler per rilevare potenziali distorsioni durante la preparazione dei dati. Da qui, è possibile utilizzare le trasformazioni predefinite di SageMaker Data Wrangler per preparare i dati. Una volta che i dati sono pronti, è possibile costruire flussi di lavoro di ML completamente automatizzati con Pipeline Amazon SageMaker o importare i dati in Amazon SageMaker Feature Store.

5.1.4 Notebook Amazon SageMaker Studio

I notebook Amazon SageMaker Studio sono notebook Jupyter gestiti, facili da utilizzare e ideali per la collaborazione. I notebook Amazon SageMaker Studio si integrano con gli strumenti progettati ad hoc per il ML in SageMaker e in altri servizi AWS per fornire lo sviluppo del ML end-to-end in Amazon SageMaker Studio, l'ambiente di sviluppo integrato(IDE) completo per il ML.
Ciascun utente avrà una home directory isolata e indipendente da una determinata istanza. Questa directory viene automaticamente montata su tutti i server e kernel dei notebook all'avvio, in modo da poter accedere ai notebook e ad altri file anche quando si cambia istanza per visualizzarli ed eseguirli.

5.1.4.1 Condivisione dei risukltati

I professionisti del machine learning possono creare uno spazio di lavoro condiviso all'interno dei quali i membri del team possono leggere e modificare in maniera collaborativa i notebook Amazon SageMaker Studio. Utilizzando gli spazi condivisi, i colleghi possono modificare a più mani lo stesso file del notebook, eseguire simultaneamente il codice del notebook e rivedere insieme i risultati, eliminando le sequenze di passaggi e ottimizzando la collaborazione.

Negli spazi condivisi, i team del ML dispongono di un supporto incorporato per servizi come BitBucket e AWS CodeCommit, potendo così gestire con facilità versioni differenti dei propri notebook e confrontare le modifiche nel corso del tempo. Le risorse create all'interno dei notebook, come esperimenti e modelli di ML, vengono automaticamente salvati e associati allo specifico spazio di lavoro in cui sono stati creati affinché i team possano coordinarsi e organizzarsi con maggiore facilità e accelerare lo sviluppo dei modelli di ML.

5.1.5 Amazon SageMaker Experiments

Amazon SageMaker Experiments aiuta a organizzare e controllare le iterazioni sui modelli di ML. Esperiment SageMaker aiuta a gestire le iterazioni con l'acquisizione automatica di parametri di input, configurazioni e risultati salvandoli come "esperimenti". Si può lavorare all'interno dell'interfaccia visiva di Amazon SageMaker Studio, in cui si possono sfogliare gli esperimenti attivi, cercare gli esperimenti precedenti in base alle loro caratteristiche, rivederne i risultati e confrontare i risultati degli esperimenti in modo visivo.

5.1.6 Debugger Amazon SageMaker

Debugger Amazon SageMaker acquisisce automaticamente metriche in tempo reale durante l'addestramento, come matrici di confusione e gradienti di apprendimento, per contribuire a migliorare la precisione del modello. Le metriche di Debugger SageMaker possono essere visualizzate in Amazon SageMaker Studio per una facile comprensione. Debugger SageMaker può anche generare segnalazioni e avvisi di correzione quando vengono rilevati comuni problemi di addestramento. SageMaker Debugger inoltre monitora e

profila automaticamente le risorse di sistema come CPU, GPU, rete e memoria in tempo reale, e fornisce consigli sulla loro riassegnazione. Ciò ti abilita ad utilizzare le tue risorse in modo efficiente durante l'addestramento e aiuta a ridurre i costi e le risorse.

5.1.7 Informazioni Aggiuntive

5.1.7.1 Distribuzione Modelli

Amazon SageMaker è in grado di distribuire automaticamente modelli di deep learning e grandi set di addestramento fra istanze AWS GPU in una frazione del tempo necessario per costruire e ottimizzare queste strategie di distribuzione manualmente. Le due tecniche di addestramento distribuite che SageMaker applica sono il parallelismo dei dati e il parallelismo dei modelli. Il parallelismo dei dati viene applicato per migliorare la velocità di addestramento dividendo i dati equamente fra più istanze della GPU, permettendo a ciascuna istanza di addestrarsi contemporaneamente. Il parallelismo del modello è utile per i modelli troppo grandi per essere memorizzati su una singola GPU e richiedono che il modello sia partizionato in parti più piccole prima di essere distribuito su più GPU. Con solo poche righe di codice aggiuntivo negli script di addestramento PyTorch e TensorFlow, SageMaker applicherà automaticamente il parallelismo dei dati o il parallelismo dei modelli, consentendo di sviluppare e distribuire i modelli più velocemente. SageMaker determinerà il miglior approccio per dividere il modello usando algoritmi di partizionamento dei grafici per bilanciare il calcolo di ciascuna GPU e minimizzando la comunicazione tra istanze GPU. Inoltre, SageMaker ottimizza i processi di addestramento distribuiti tramite algoritmi che sfruttano appieno le capacità di calcolo e la rete di AWS per raggiungere un'efficienza di scalabilità quasi lineare. In questo modo, è possibile completare l'addestramento più velocemente rispetto alle implementazioni open source manuali.

Una volta creati e formati dei modelli, Amazon SageMaker fornisce tre opzioni per implementarli, permettendo l'inizio di previsioni.

1. **L'inferenza in tempo reale** è adeguata per carichi di lavoro con requisiti di latenza di millisecondi, dimensioni di payload fino a 6MB e tempi di elaborazione fino a 60 secondi;

2. **La trasformazione in batch** è ideale per previsioni offline su grandi batch di dati disponibili in anticipo;

3. **L'inferenza asincrona** è progettata per carichi di lavoro che non richiedono latenza inferiore al secondo, dimensioni di payload fino a 1 GB e tempi di elaborazione fino a 15 minuti.

5.1.7.2 Costi

I servizi appena elencati sono solo alcuni dei servizi che mette a disposizione AWS per il machine learning. Ogni servizio elencato è un servizio a pagamento e Amazon mette a disposizone una funzione per ridurre i costi. I **Savings Plans** di Amazon SageMaker contribuiscono a ridurre i costi fino al 64%. I piani si applicano automaticamente all'uso idoneo delle istanze ML SageMaker, inclusi SageMaker Studio Notebooks, istanze Notebook SageMaker, SageMaker Processing, SageMaker Data Wrangler, SageMaker Training, SageMaker Real-Time Inference e SageMaker Batch Transform, indipendentemente dalla famiglia di istanze, dalla dimensione o dalla regione.
Se non si volesse considerare questa soluzione e si volesse optare per il costo effettivo in base al consumo basta andare sul sito di AWS e vedere il catalogo.

5.1.8 Whitepaper AWS

La Formula 1 (F1) ha scelto AWS per le seguenti ragioni:

"Avevamo bisogno di un fornitore di tecnologia che ci aiutasse a innovare più velocemente e a proiettare la nostra organizzazione verso il futuro, quindi AWS era il partner ideale. Approfittando delle straordinarie risorse di AWS e delle sue innovative tecnologie cloud, siamo stati in grado di far comprendere meglio ai fan le decisioni prese in frazioni di secondo sulla pista, di riprogettare le auto di F1 del futuro, di capire la grande quantità di dati della F1, di effettuare analisi e applicare il machine learning per sfruttare al meglio le potenzialità dei dati e tanto altro. Siamo entusiasti di ciò che abbiamo realizzato ed emozionati di vedere cos'altro possiamo fare insieme".

Queste sono le motivazioni citate da:

Ross Brawn, *Managing Director of Motor Sports, F1.*

Figura 5: Logo F1

Le funzionalità più ampie e approfondite di AWS hanno cambiato il modo in cui la F1 raccoglie, analizza e sfrutta dati e contenuti per prendere decisioni. Con 300 sensori su ogni auto da corsa di F1 in grado di generare oltre 1,1 milioni di punti dati al secondo trasmessi dalle auto ai box, la F1 è uno sport fortemente basato sui dati.

La F1 e AWS utilizzano i dati per migliorare le prestazioni di auto e pilo-

ti. Grazie all'utilizzo di AWS High Performance Computing, la F1 è stata in grado di effettuare simulazioni aerodinamiche per sviluppare la sua auto del futuro più veloce del 70% con la riduzione della perdita di portanza dal 50% al 15%. Questa notevole riduzione permette al pilota più indietro di avere maggiori possibilità di sorpasso e, allo stesso tempo, rende la corsa più entusiasmante per i fan. L'auto è stata presentata in occasione della stagione 2022. Inoltre, la F1 sta studiando l'utilizzo del Machine Learning nelle proprie simulazioni, fornendo nuove informazioni utili all'organizzazione con oltre 550 milioni di punti data raccolti attraverso più di 5.000 simulazioni con una o più auto.

L'esperienza dei fan durante i week-end di gara è cambiata. Con AWS, la F1 è stata in grado di trasformare milioni di punti dati trasmessi dalle auto ai box in un'esperienza entusiasmante per i fan grazie a F1 Insights. La F1 si avvale dei dati raccolti durante le gare degli ultimi 70 anni, archiviati in Amazon S3, analizzati da complessi modelli e condivisi con i fan sotto forma di informazioni utili per rivelare le sfumature delle decisioni prese in frazioni di secondo e per evidenziare le prestazioni grazie a queste statistiche avanzate.

L'**architettura** utilizzata è la seguente

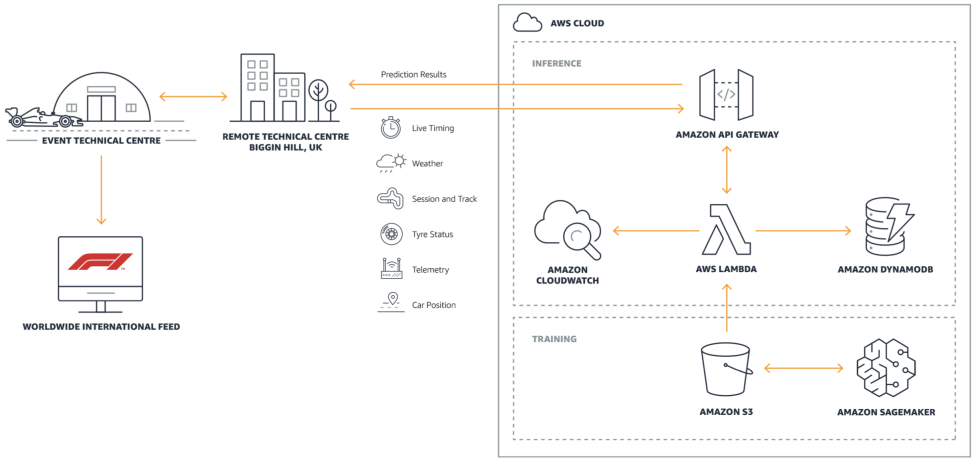

Figura 6: Architettura F1

5.2 Azure

Azure è una piattaforma cloud pubblica di Microsoft, offre un'ampia raccolta di servizi, che include funzionalità PaaS, IaaS e di servizio di database gestito.

Figura 7: Logo Azure

I servizi volti al macine learning che mette a disposizione Azure sono i seguenti

- Azure Synapse Analytics;
- Azure Databricks;
- Microsoft Purview;
- Data Factory di Azure;
- HDIsight;
- Analisi di flusso di Azure;
- Machine Learning;
- Azure Analysis Services;
- Azure Data Lake Storage;
- Esplora dati di Azure;

5.2.1 Azure Synapse Analytics

Azure Synapse Analytics è un servizio di analisi senza limiti che riunisce integrazione dei dati, funzionalità aziendali di data warehousing e analisi dei Big Data. Offre la libertà di eseguire query sui dati in base alle esigenze, usando opzioni serverless o dedicate, su larga scala. Azure Synapse combina questi mondi grazie a un'esperienza unificata per l'inserimento, l'esplorazione, la preparazione, la trasformazione, la gestione e la distribuzione dei dati per esigenze immediate di business intelligence e apprendimento automatico. Ciò che permette Azure Synapse Analytics è di fornire informazioni dettagliate da tutti i dati, in diversi data warehouse e sistemi di analisi dei Big Data, con velocità elevata, di espandere l'individuazione di informazioni dettagliate da tutti i dati e applicare modelli di Machine Learning a tutte le app intelligenti, riducendo notevolmente il tempo necessario per lo sviluppo dei progetti grazie a un'esperienza unificata per lo sviluppo di soluzioni di analisi end-to-end, elimina inoltre gli ostacoli nei dati ed esegue analisi sui dati operativi e di app aziendali con Collegamento ad Azure Synapse senza spostare i dati proteggendo i dati con le funzionalità più avanzate per la sicurezza e la privacy disponibili sul mercato, ad esempio la sicurezza a livello di colonna e di riga e Dynamic Data Masking.

5.2.2 Azure Databrick

Permette di acquisire informazioni dettagliate da tutti i dati e creare soluzioni di intelligenza artificiale con Azure Databricks, configurando l'ambiente Apache Spark™ in pochi minuti, ridimensionandolo automaticamente e di collaborare a progetti condivisi in un'area di lavoro interattiva.
Azure Databricks supporta Python, Scala, R, Java e SQL, oltre ai framework e le librerie di data science, ad esempio TensorFlow, PyTorch e scikit-learn.

Azure Databricks offre le versioni più recenti di Apache Spark e permette di integrarle facilmente con le librerie open source. Permette la configurazione di cluster e la creazione rapida in un ambiente Apache Spark completamente gestito che offre la disponibilità e la scala globale di Azure. I cluster vengono installati, configurati e ottimizzati per garantire affidabilità e prestazioni senza necessità di monitoraggio. Sfrutta le funzionalità di scalabilità automatica e terminazione automatica per migliorare il costo totale di proprietà. La collaborazione è permessa da una piattaforma aperta e unificata per eseguire tutti i tipi di carichi di lavoro di analisi, indipendentemente dal tipo di sviluppatore che ci stia lavorando, che esso sia un data scientist, un data engineer o un business analyst. Vi è la versione semplificata dei notebook con GitHub e Azure DevOps.

Accedendo alle funzionalità avanzate e automatizzate di Machine Learning usando Azure Machine Learning integrato si potranno identificare rapidamente gli algoritmi e gli iperparametri idonei, semplificando la gestione, il monitoraggio e l'aggiornamento dei modelli di Machine Learning distribuiti dal cloud alla rete perimetrale. Azure Machine Learning fornisce inoltre un registro centralizzato per esperimenti, pipeline di Machine Learning e modelli. Combina i dati su qualsiasi scala e fornisce informazioni dettagliate tramite dashboard di analisi e report operativi. Automatizza lo spostamento dei dati con Azure Data Factory, caricando i dati in Azure Data Lake Storage, trasformandoli e pulendoli con Azure Databricks e rendendoli disponibili per l'analisi con Azure Synapse Analytics.

5.2.3 Analisi di flusso di Azure

Il servizio di analisi in tempo reale, facile da usare, progettato per carichi di lavoro cruciali. Permette la creazione di una pipeline di streaming serverless

end-to-end con pochi clic.

Esegue analisi complesse senza che sia necessario imparare a usare nuovi framework di elaborazione o effettuare il provisioning di macchine virtuali o cluster. Usa un linguaggio SQL familiare con possibilità di estensione con codice personalizzato per JavaScript e C# per casi d'uso più avanzati. Abilita con facilità scenari come la creazione di dashboard a bassa latenza, ETL di streaming e avvisi in tempo reale con l'integrazione con un solo clic tra origini e sink.

Consente un'elaborazione di eventi garantita di tipo "Exactly Once" con una disponibilità al 99,9% e funzionalità di ripristino predefinite. La configurazione avviene con facilità tramite una pipeline di integrazione continua e recapito continuo (CI-CD) con latenze inferiori al secondo per i carichi di lavoro più complessi.

Si acquisice la funzionalità di informazioni dettagliata e di analisi in tempo reale più vicine alla posizione di origine dei dati, rendendo possibili nuovi scenari con vere architetture ibride per l'elaborazione di flussi eseguendo la stessa query sul cloud o nei dispositivi perimetrali.

Vi sono molti modelli predefiniti di Machine Learning per ridurre il tempo necessario all'ottenimento di informazioni dettagliate, con funzionalità basate su Machine Learning per eseguire il rilevamento delle anomalie nei processi di streaming con Analisi di flusso di Azure.

5.2.4 Azure Machine Learning

Azure Machine Learning consente ai data scientist e sviluppatori di creare, distribuire e gestire modelli in modo più rapido e sicuro. Accelera il time-to-value con le operazioni di Machine Learning (MLOps), l'interoperabilità open source e gli strumenti integrati.

Migliora la produttività con la funzionalità studio, un'esperienza di sviluppo che supporta tutte le attività di Machine Learning, per creare, eseguire il training e distribuire modelli. Collabora con Jupyter Notebook usando il supporto predefinito per i framework e le librerie open source più diffusi. Creando rapidamente modelli accurati conMachine Learning automatizzato per modelli tabulari, di testo e di immagini usando la progettazione delle funzionalità e lo sweep degli iperparametri. Utilizzando Visual Studio Code per passare senza problemi al training da locale a cloud con scalabilità automatica con potenti cluster di CPU e GPU basati sul cloud con tecnologia della rete NVIDIA Quantum InfiniBand.

Semplifica la distribuzione e la gestione di migliaia di modelli in più ambienti usando MLOps. Accelerando la distribuzione e l'assegnazione di un punteggio ai modelli grazie a endpoint completamente gestiti per previsioni in batch e in tempo reale. Usa pipeline ripetibili per automatizzare i flussi di lavoro per integrazione continua e recapito continuo (CI/CD). Permettendo la condivisione e l'individualizzazione di artefatti di Machine Learning tra più team per la collaborazione tra aree di lavoro usando i registri. Monitoraggio continuo delle metriche relative alle prestazioni dei modelli, rilevando la derivata dei dati e attiva la ripetizione del training per migliorare le prestazioni dei modelli. Consente la valutazione di modelli di Machine Learning con flussi di lavoro riproducibili e automatizzati per valutare l'equità dei modelli, la spiegabilità, l'analisi degli errori, l'analisi causale, le prestazioni del modello e l'analisi esplorativa dei dati. Effettuando interventi reali con l'analisi causale nel dashboard di intelligenza artificiale responsabile e genera una scorecard in fase di distribuzione. Contestualizza le metriche di intelligenza artificiale responsabili per i destinatari tecnici e non tecnici per coinvolgere gli stakeholder e semplificare la revisione della conformità.

Aumenta la sicurezza nel ciclo di vita di Machine Learning con funzionalità complete che includono identità, dati, rete, monitoraggio e conformità.
Permette la protezionde delle soluzioni usando il controllo degli accessi in base al ruolo personalizzato, le reti virtuali, la crittografia dei dati, gli endpoint privati e gli indirizzi IP privati. Si può eseguire il training e distribuire i modelli in locale per soddisfare i requisiti di sovranità dei dati.

5.2.5 Whitepaper Azure

Per rimanere in tema sportivo si potrebbe parlare di come l'NBA, campionato americano di basket, abbia utilizzato Azure per gestire i propri dati, ma per non risultare ripetitivo con gli esempi vedremo invece come **Forza Horizon 5** ha usufruito dei servizi di Azure per i suoi scopi.

Figura 8: Logo Forza Horizon 5

Quando è stato lanciato, Forza Horizon 5 ha visto più di 10 milioni di giocatori simultanei, la più grande prima settimana nella storia di Xbox Game Studios. Forza Horizon 5 è anche stata una straordinaria vittoria dietro le quinte, per la piattaforma Azure come servizio (PaaS) fornendo un equipaggio virtuale e un'architettura container basata su Windows ha permesso di adattare alle mutevoli richieste a velocità di curvatura. Gli sviluppatori di giochi si sono affidati alla scalabilità automatica dei cluster Azure Kubernetes Service (AKS) per soddisfare le esigenze di prestazioni più impegnative, mentre i servizi Azure completamente gestiti hanno liberato il team da attività di gestione dell'infrastruttura dispendiose in termini di tempo.
"Con i nostri stack di elaborazione, archiviazione e dati tutti su Azure, i nostri team di ingegneri possono ora dedicare più tempo a creare nuove espe-

rienze per i nostri giocatori piuttosto che gestire la nostra infrastruttura".

Queste sono state le parole di:

Daniel Adent: *General Manager of ForzaTech.*

AKS ha anche permesso al team di iterare molto più velocemente durante lo stress test, un passo mission-critical nella preparazione per il giorno del lancio. Prima di passare ad AKS, gli sviluppatori ForzaTech impiegavano mezz'ora per scambiare vecchie immagini sulle loro VM e prepararle per il caricamento. In AKS, questo processo ha richiesto secondi. La velocità extra ha consentito al team di effettuare riparazioni immediate, superare i blocchi stradali e scalare secondo necessità. Inoltre, il team ha potuto eseguire gli aggiornamenti del cluster in anticipo, aggiungendo flessibilità e agilità agli stress test impegnativi.

Ben prima del lancio, gli stress test hanno dimostrato che i servizi di ForzaTech potevano facilmente scalare da 600.000 a 3 milioni di utenti simultanei.

"È stata una migrazione facile per passare ad AKS e rimanere su Windows... Nel giro di un mese, in pratica avevamo convertito i nostri servizi in AKS e avevamo tutto in esecuzione".

Notevole dichiarazione di:

Tyler Hennessy: *Principal Software Engineering Lead.*

L'architettura Forza Horizon 5 include 17 servizi principali che vengono eseguiti in AKS. Tra questi, il servizio di classifica dice agli utenti dove si collocano, mentre il servizio di contenuti generati dall'utente (UGC) supporta la possibilità per i giocatori di caricare foto e altri contenuti. Il servizio di casa d'aste consente ai giocatori di vendersi auto l'un l'altro e il servizio di analisi raccoglie la telemetria del gioco e del giocatore utilizzata nei rapporti.

Per semplificare la migrazione ai contenitori Windows, ForzaTech ha creato un playbook basato sull'esperienza di spostamento del servizio UGC. La

porta d'ingresso per tutte le richieste dei client è il servizio aggregatore, una singola distribuzione di scalabilità automatica che esegue più pod in AKS. Tutti i pod interagiscono con le stesse configurazioni e segreti AKS. Il servizio aggregatore interagisce con gli altri microservizi in esecuzione in vari container. Le loro architetture simili hanno reso il playbook facile da seguire. Sono strumentati con sonde di salute di vitalità e prontezza di Kubernetes, in modo che il team possa identificare rapidamente eventuali problemi e apportare modifiche, come la regolazione della scala.

Per la continuità aziendale, due istanze di ogni servizio vengono eseguite su sole quattro VM. Prima del passaggio ad AKS, ForzaTech eseguiva un minimo di 40 VM in ogni momento. Questa configurazione ha ridotto il costo dell'ambiente di test del 90%.

Per aumentare ulteriormente la velocità, le VM nei cluster AKS consentono reti accelerate, una latenza ultra-bassa e un percorso ad alte prestazioni che riduce il jitter e diminuisce l'utilizzo della CPU. Ottenendo una riduzione della latenza del 50% su alcune chiamate di rete back-end.

Il livello di storage svolge anche un ruolo nel raggiungimento degli obiettivi di scalabilità e prestazioni. Il servizio aggregatore interagisce con Azure Cache for Redis, una soluzione di memorizzazione nella cache altamente scalabile. Quando il throughput ha bisogno di un aumento, i dati possono essere frammentati su più account di archiviazione di Azure, con un impatto diretto sulla capacità del team di scalare.

L'architettura utilizzata è la seguente:

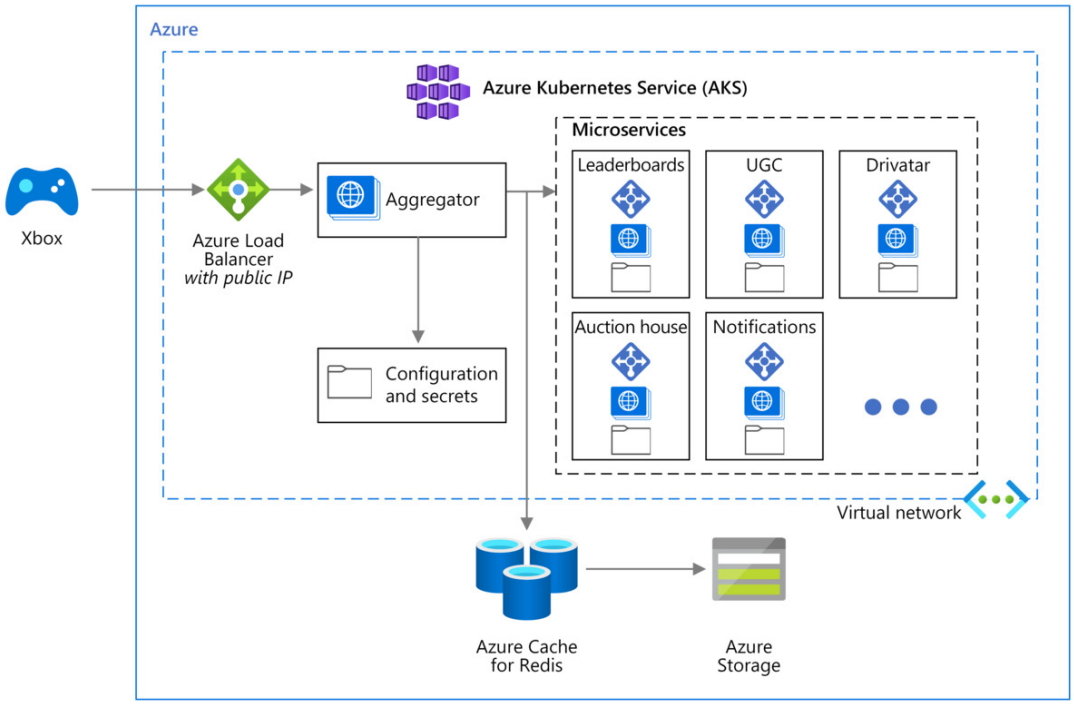

Figura 9: Architettura Forza Horizon 5

5.3 Google Cloud Platform

Google Cloud Platform(GCP), offerto da Google, è una suite di servizi di cloud computing che funziona sulla stessa infrastruttura che Google utilizza internamente per i suoi prodotti per gli utenti finali, come Ricerca Google, Gmail, Google Drive e YouTube.

Figura 10: Logo Google Cloud Platform

I servizi che offre google riguardanti il machine learning sono i seguenti:

- Vertex AI;
- Vertex AI Workbench;
- AI Infrastructure;
- AutoML;
- Natural Language AI;
- Speech-to-Text;
- Text-to-Speech;

- Translation AI;
- Video AI;
- Vision AI;
- Dialogflow;
- ecc.

5.3.1 Vertex AI

Un'interfaccia utente unificata per l'intero flusso di lavoro di ML, Vertex AI unisce i servizi Google Cloud per creare ML in un'unica interfaccia utente e API unificata. In Vertex AI, si può addestrare e confrontare facilmente i modelli utilizzando AutoML o l'addestramento con codice personalizzato e tutti i modelli vengono archiviati in un unico repository di modelli centrale. Si possono eseguire il deployment dei modelli sugli stessi endpoint su Vertex AI.

Fornisce API preaddestrate per visione artificiale, video, linguaggio naturale e molto altro.

Integra in modo semplice il machine learning per visione artificiale, video, traduzione e linguaggio naturale nelle applicazioni esistenti. AutoML consente agli sviluppatori di addestrare modelli di specifici per le loro esigenze aziendali con un'esperienza minima nell'ambito del machine learning e con il minimo impegno. Con un registro gestito centralmente per tutti i set di dati in tutti i tipi di dati (visione artificiale, linguaggio naturale e dati tabulari). Tramite Vertex AI Workbench, Vertex AI è integrata in modo nativo con BigQuery, Dataproc e Spark, si può utilizzare BigQuery ML per creare ed eseguire modelli di machine learning in BigQuery utilizzando query SQL

standard su strumenti e fogli di lavoro di business intelligence esistenti, oppure si possono esportare set di dati da BigQuery direttamente in Vertex AI Workbench ed eseguire i modelli da lì, per poi utilizzare Vertex Data Labeling per generare etichette ad alta precisione per la raccolta dei dati.
Vertex AI si integra con framework open source ampiamente utilizzati come TensorFlow, PyTorch e scikit-learn, oltre al supporto di tutti i framework ML e i rami dell'intelligenza artificiale tramite container personalizzati per l'addestramento e la previsione.
Vertex AI Vision riduce il tempo necessario per creare applicazioni di visione artificiale da settimane a ore, abbassando in questo modo i costi di produzione. Offre un'interfaccia con trascinamento semplice da utilizzare e una libreria di modelli di machine learning preaddestrati per attività comuni come il conteggio del numero di partecipanti, il riconoscimento dei prodotti e il rilevamento di oggetti. Offre inoltre la possibilità di importare i tuoi modelli AutoML o ML personalizzati esistenti da Vertex AI.

5.3.2 Vertex AI Workbench

Le estensioni si connettono rapidamente all'intera infrastruttura dati, tra cui BigQuery, Data Lake, Dataproc e Spark, effettuando lo scale up o lo scale out rapidamente a seconda delle esigenze di analisi e AI.
Permette la scrittura di query SQL e Spark da una cella di blocco note sensibile alla sintassi e abilitata al completamento automatico.
Tutti gli aspetti delle operazioni di calcolo sono autogestiti. Il timeout di inattività e l'arresto automatico ottimizzano il costo totale di proprietà.
Controlli di sicurezza pronti all'uso di Google Cloud. Single Sign-On e autenticazione semplice ad altri servizi Google Cloud.
Data lake e Spark in un unico posto Indipendentemente dal fatto che si

utilizzi TensorFlow, PyTorch o Spark, si può eseguire qualsiasi motore da Vertex AI Workbench.

Con pochi clic, si possono collegare i blocchi note ai flussi di lavoro operativi consolidati, utilizzando blocchi note per l'addestramento distribuito, l'ottimizzazione degli iperparametri o l'addestramento continuo pianificato o attivato. La profonda integrazione con i servizi di Vertex AI inserisce MLOps nel blocco note senza la necessità di riscrivere il codice o nuovi flussi di lavoro. CI/CD semplificato Integrazione di Kubeflow Pipelines per utilizzare Notebooks come destinazione di deployment ideale, testata e verificata. Permette la condivisione dell'output delle celle del blocco note aggiornate periodicamente per la creazione di rapporti e per la contabilità.

5.3.3 Whitepaper GCP

A differenza di AWS e Azure parleremo di un problema risolto da GCP senza entrare nello specifico dell'architettura poichè google non mette a disposizione del pubblico le soluzioni adottate.

In questa sezione parleremo di come **PayPal** utilizza i servizi di Google Cloud Platform per adempiere correttamente al suo funzionamento.

Figura 11: Logo PayPal

PayPal sta abilitando nuovi mercati con lo scopo di aiutare le persone di tutto il mondo a entrare nell'economia globale con servizi finanziari semplici, convenienti e sicuri.

Sfruttando la potenza di Google Cloud, servono più di 300 milioni di clienti e sviluppano servizi online, mobili e in-store in 200 mercati, in 100 valute,

con l'obiettivo di servire un miliardo di utenti ogni giorno.

"I dati stanno esplodendo. Le aspettative dei clienti stanno cambiando. A meno che tu non cambi il modo di pensare, non sarai in grado di passare attraverso la trasformazione. Con Google Cloud, hai sistemi più affidabili, sempre e sicuri".

Queste sono state le motivazioni di:

Sri Shivananda *CTO di PayPal.*

Configurazione più rapida dell'ambiente per gli sviluppatori, carichi di lavoro basati su container distribuiti in pochi secondi, transazioni vicino alla fonte che si allineano con le normative locali, sicurezza di rete insuperabile, con livelli di crittografia e rilevamento delle frodi di nuova generazione e analisi dei dati che aggiunge valore e guida la creazione del prodotto.

Tutto questo è quello che ha offerto Google Cloud a PayPal.

Negli ultimi tre anni, PayPal ha spostato i carichi di lavoro mission-critical su Google Cloud. Il nuovo PayPal sarà ibrido e multi-cloud, unificando fornitori, posizioni e funzionalità disparate per un successo futuro condiviso.

6 Algoritmi di machine learning parallelizzabili

In questa sezione vedremo com'è possibile effettuare training distribuito con Azure utilizzando le librerie **PyTorch** e **TensorFlow**.
Nel training distribuito il carico di lavoro per il training di un modello viene suddiviso e condiviso tra più mini processori, denominati **nodi di lavoro**. Questi nodi di lavoro funzionano in parallelo per velocizzare il training del modello. Il training distribuito può essere usato per i modelli di Machine Learning tradizionali, ma è più adatto per le attività di calcolo e con un utilizzo intensivo del tempo, ad esempio l'apprendimento avanzato per il training di modelli di Deep Learning.
Azure offre due tipi di parallelizzazione del **modello** e dei **dati**.

6.1 Parallelismo del modello

Nel parallelismo del modello, noto anche come parallelismo di rete, il modello viene segmentato in parti diverse che possono essere eseguite simultaneamente in nodi diversi e ognuno verrà eseguito sugli stessi dati. La scalabilità di questo metodo dipende dal grado di parallelizzazione delle attività dell'algoritmo ed è più complesso implementarlo rispetto al parallelismo dei dati.
Nel parallelismo del modello, i nodi di lavoro devono sincronizzare solo i parametri condivisi, in genere una volta per ogni passaggio di propagazione avanti o indietro. Inoltre, i modelli di dimensioni maggiori non sono un problema perché ogni nodo opera su una sottosezione del modello sugli stessi dati di training.

6.2 Parallelismo dei Dati

Il parallelismo dei dati è il più semplice da implementare dei due approcci di training distribuiti ed è sufficiente per la maggior parte dei casi d'uso. In questo approccio, i dati vengono suddivisi in partizioni, in cui il numero di partizioni è uguale al numero totale di nodi disponibili, nel cluster di calcolo. Il modello viene copiato in ognuno dei nodi di lavoro e ogni ruolo di lavoro opera sul proprio subset dei dati. Ogni nodo deve avere la capacità di supportare il modello sottoposto a training, ovvero il modello deve adattarsi interamente a ogni nodo. Il diagramma seguente fornisce una dimostrazione visiva di questo approccio.

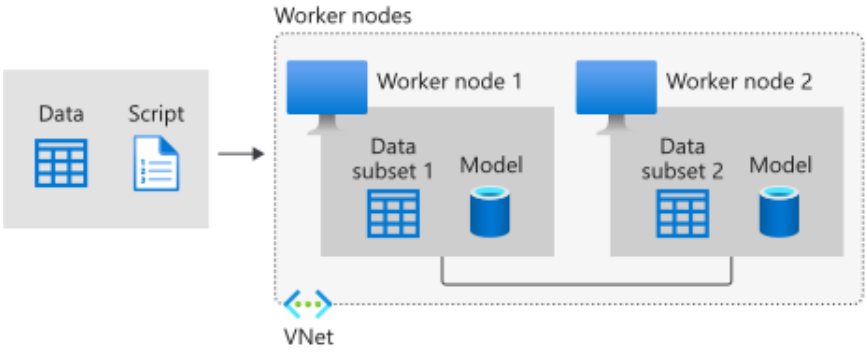

Figura 12: Traning Distribuito

Ogni nodo calcola in modo indipendente gli errori tra le stime per i relativi campioni di training e gli output etichettati. A sua volta, ogni nodo aggiorna il modello in base agli errori e deve comunicare tutte le modifiche apportate agli altri nodi per aggiornare i modelli corrispondenti. Ciò significa che i nodi di lavoro devono sincronizzare i parametri del modello o le sfumature alla fine del calcolo batch per assicurarsi che eseguano il training di un modello coerente.

6.3 Traning Distribuito

6.3.0.1 MPI

Azure Machine Learning offre un processo MPI per avviare un determinato numero di processi in ogni nodo. Azure Machine Learning costruisce il comando di avvio MPI completo(mpirun) dietro le quinte. Non è possibile fornire comandi head-node-launcher completi come mpirun o DeepSpeed launcher.

Per eseguire il training distribuito tramite MPI, seguire questa procedura:

1. Usare un ambiente di Azure Machine Learning con il framework di Deep Learning preferito e MPI. Azure Machine Learning offre un ambiente curato per i framework più diffusi.

2. Definire un **command** oggetto con *instance_ count*. L'instance_count deve essere uguale al numero di GPU per nodo per l'avvio per processo oppure lo si può impostare su 1(impostazione predefinita) per ogni nodo-launch, in questo caso lo script utente sarà responsabile dell'avvio dei processi per ogni nodo.

3. Usare il distribution parametro di command per specificare le impostazioni per MpiDistribution.

```
from azure.ai.ml import command, MpiDistribution
job = command(
    code="./src",  # local path where the code is stored
    command="python train.py --epochs ${{inputs.epochs}}",
    inputs={"epochs": 1},
    environment="AzureML-tensorflow-2.7-ubuntu20.04-py38-cuda11-gpu@latest",
    compute="gpu-cluster",
    instance_count=2,
    distribution=MpiDistribution(process_count_per_instance=2),
    display_name="tensorflow-mnist-distributed-horovod-example"
)
```

6.3.0.2 Horovod

Usare la configurazione del processo MPI quando si usa Horovod per il training distribuito con il framework di Deep Learning, permette di assicurarsi che il codice segua questi suggerimenti:

- Il codice di training viene instrumentato correttamente con Horovod prima di aggiungere le parti di Azure Machine Learning;
- L'ambiente di Azure Machine Learning contiene Horovod e MPI. Gli ambienti GPU curati di PyTorch e TensorFlow sono preconfigurati con Horovod e le relative dipendenze;
- Creare un oggetto command con la distribuzione desiderata.

6.3.1 PyTorch

Azure Machine Learning supporta l'esecuzione di processi distribuiti usando le funzionalità di training distribuite native di PyTorch (torch.distributed). Il backbone di qualsiasi training distribuito si basa su un gruppo di processi che si conoscono tra loro e possono comunicare tra loro usando un back-end. Per PyTorch, il gruppo di processi viene creato chiamando torch.distributed.init_process_group in tutti i processi distribuiti per formare collettivamente un gruppo di processi.

```
torch.distributed.init_process_group(backend='nccl',
                                     init_method='env://', ...)
```

I back-end di comunicazione più comuni usati sono mpi, nccle gloo.
L'*init_method* indica in che modo ogni processo può individuarsi tra loro, come inizializzare e verificare il gruppo di processi usando il back-end di comunicazione. Per impostazione predefinita, se init_method non è specificato PyTorch userà il metodo di inizializzazione della variabile di ambiente (env://). L'init_method è il metodo di inizializzazione consigliato da usare nel codice di training per eseguire PyTorch distribuito in Azure Machine Learning.

PyTorch cercherà le variabili di ambiente seguenti per l'inizializzazione:

- **MASTER_ADDR** - Indirizzo IP del computer che ospiterà il processo con classificazione 0.
- **MASTER_PORT** - Porta libera nel computer che ospiterà il processo con classificazione 0.
- **WORLD_SIZE** - Numero totale di processi. Deve essere uguale al numero totale di dispositivi (GPU) usati per il training distribuito.
- **RANK** - Classificazione (globale) del processo corrente. I valori possibili sono da 0 a (dimensione globale - 1).

Oltre a questi, molte applicazioni richiederanno anche le variabili di ambiente seguenti:

- **LOCAL_RANK** - Classificazione locale (relativa) del processo all'interno del nodo. I valori possibili sono da 0 a (# di processi nel nodo - 1). Queste informazioni sono utili perché molte operazioni, ad esempio la preparazione dei dati, devono essere eseguite una sola volta per ogni nodo in genere in local_rank = 0.
- **NODE_RANK** - Classificazione del nodo per il training multinodo. I valori possibili sono da 0 a (numero totale di nodi - 1).

Non è necessario usare un'utilità di avvio come torch.distributed.launch.

Per eseguire un processo PyTorch distribuito:

1. Specificare lo script di training e gli argomenti;

2. Creare un oggetto command e specificare il tipo come PyTorch e l'oggetto process_count_per_instancedistribution nel parametro corrisponde process_count_per_instance al numero totale di processi da eseguire per il processo. process_count_per_instance deve in genere essere uguale a # GPUs per node x # nodes. Se process_count_per_instance non viene specificato, Azure Machine Learning avvierà per impostazione predefinita un processo per nodo.

Azure Machine Learning imposterà le MASTER_ADDR variabili di ambiente, MASTER_PORT, WORLD_SIZEe NODE_RANK in ogni nodo e imposterà le variabili di ambiente e LOCAL_RANK a livello RANK di processo.

```
from azure.ai.ml import command
from azure.ai.ml.entities import Data
from azure.ai.ml import Input
from azure.ai.ml import Output
from azure.ai.ml.constants import AssetTypes

inputs = {
    "cifar": Input(
        type=AssetTypes.URI_FOLDER, path=returned_job.outputs.cifar.path
    ),
    "epoch": 10,
    "batchsize": 64,
    "workers": 2,
    "lr": 0.01,
    "momen": 0.9,
    "prtfreq": 200,
    "output": "./outputs",
}

job = command(
    code="./src",  # local path where the code is stored
    command="python train.py --data-dir ${{inputs.cifar}} --
    epochs ${{inputs.epoch}} --batch-size ${{inputs.batchsize}}
    --workers ${{inputs.workers}} --learning-rate ${{inputs.lr}}
    --momentum ${{inputs.momen}} --print-freq ${{inputs.prtfreq}}
    --model-dir ${{inputs.output}}",
    inputs=inputs,
```

```
    environment="azureml:AzureML-pytorch-1.9-ubuntu18.04-py37-cuda11-gpu:6",
    compute="gpu-cluster",
    instance_count=2,
    distribution={
        "type": "PyTorch",
        "process_count_per_instance": 1,
    },
)
```

6.3.1.1 DeepSpeed

DeepSpeed è supportato come tool di prima classe in Azure Machine Learning per eseguire processi distribuiti con scalabilità quasi lineare in termini di:

1. Aumento delle dimensioni del modello;
2. Aumento del numero di GPU.

DeepSpeed può essere abilitato usando la distribuzione di Pytorch o MPI per l'esecuzione del training distribuito. Azure Machine Learning supporta l'utilità di avvio per avviare il DeepSpeed training distribuito e l'ottimizzazione automatica per ottenere una configurazione ottimale ds. È possibile usare un ambiente curato per un ambiente predefinito con le tecnologie più recenti all'avanguardia, tra cui DeepSpeed, ORT, MSSCCLe Pytorch per i processi di training DeepSpeed.

6.3.2 TensorFlow

Se si usa TensorFlow distribuito nativo nel codice di training, ad esempio l'API di *tf.distribute.Strategy TensorFlow 2.x*, è possibile avviare il processo distribuito tramite Azure Machine Learning usando distribution parametres o l'oggetto TensorFlowDistribution.

```
job = command(
    code="./src",
    command="python main.py --epochs
    ${{inputs.epochs}} --model-dir
    ${{inputs.model_dir}}",
    inputs={"epochs": 1, "model_dir":
    "outputs/keras-model"},
    environment="AzureML-tensorflow-2.4-ubuntu18.04-
    py37-cuda11-gpu@latest",
    compute="cpu-cluster",
    instance_count=2,
    # distribution = {"type": "mpi", "process_count_per_instance": 1},
    distribution={
        "type": "tensorflow",
        "parameter_server_count": 1,
        "worker_count": 2,
        "added_property": 7,
    },
    # distribution = {
    #        "type": "pytorch",
    #        "process_count_per_instance": 4,
```

```
    #        "additional_prop": {"nested_prop": 3},
    #    },
    display_name="tensorflow-mnist-distributed-example"
    # experiment_name: tensorflow-mnist-distributed-example
)

# Può anche impostare la distribuzione in un passaggio
# separato e utilizzando gli oggetti digitati
# Invece di un dict
job.distribution = TensorFlowDistribution(parameter_server_count=1,
worker_count=2)
```

Se lo script di training usa la strategia del server dei parametri per il training distribuito, ad esempio per tensorFlow 1.x legacy, sarà necessario specificare anche il numero di server di parametri da usare nel processo, all'interno del distribution parametro di command. Nell'esempio "parameter_server_count" : 1 precedente, e ""worker_count": 2,

6.3.2.1 TF_Config

In TensorFlow la variabile di ambiente **TF_CONFIG** è necessaria per il training su più computer. Per i processi TensorFlow, Azure Machine Learning configurerà e imposterà la variabile TF_CONFIG in modo appropriato per ogni ruolo di lavoro prima di eseguire lo script di training. È possibile accedere TF_CONFIG dallo script di training se è necessario:
os.environ['TF_CONFIG'].

Esempio di TF_CONFIG impostato in un nodo chief worker:

```
TF_CONFIG='{
    "cluster": {
        "worker": ["host0:2222", "host1:2222"]
    },
    "task": {"type": "worker", "index": 0},
    "environment": "cloud"
}'
```

6.4 InfiniBand

Man mano che aumenta il numero di macchine virtuali che eseguono il training di un modello, il tempo necessario per eseguire il training del modello deve diminuire. La diminuzione del tempo, idealmente, deve essere linearemente proporzionale al numero di macchine virtuali di training. Ad esempio, se il training di un modello in una macchina virtuale richiede 100 secondi, il training dello stesso modello su due macchine virtuali dovrebbe richiedere idealmente 50 secondi, il training del modello su quattro macchine virtuali richiederà 25 secondi e così via.

InfiniBand può essere un fattore importante per ottenere questa scalabilità lineare. InfiniBand consente la comunicazione da GPU a GPU a bassa latenza tra nodi in un cluster. InfiniBand richiede hardware specializzato per funzionare. Alcune serie di macchine virtuali di Azure, in particolare nc, ND e serie H, includono macchine virtuali con supporto per RDMA con SR-IOV e InfiniBand. Queste macchine virtuali comunicano sulla rete InfiniBand a bassa latenza e larghezza di banda elevata, che è molto più efficiente rispetto alla connettività basata su Ethernet. SR-IOV per InfiniBand consente prestazioni quasi bare metal per qualsiasi libreria MPI . Questi SKU sono progettati per soddisfare le esigenze di carichi di lavoro di Machine Learning a elevato utilizzo di calcolo e accelerati dalla GPU. In genere, gli SKU delle macchine virtuali con "r" nel nome contengono l'hardware InfiniBand richiesto e quelli senza "r" in genere non lo contengono.

'r' è un riferimento a RDMA, che indica "accesso diretto alla memoria remota".

Ad esempio, lo SKU Standard_NC24rs_v3 della macchina virtuale è abilitato per InfiniBand, ma lo SKU Standard_NC24s_v3 non lo consente. A parte le funzionalità InfiniBand, le specifiche tra questi due SKU sono in

gran parte uguali, entrambe hanno 24 core, 448 GB di RAM, 4 GPU dello stesso SKU e via dicendo.

Se si crea un AmlCompute cluster di una di queste dimensioni abilitate per RDMA, le dimensioni abilitate per InfiniBand, l'immagine del sistema operativo verrà fornita con il driver MEllanox OFED necessario per abilitare InfiniBand preinstallato e preconfigurato.