



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

1	Machine Learning	3
1.1	Appredimento Supervisionato	3
1.2	Apprendimento non Supervisionato	4
1.3	Apprendimento Misto	5
1.3.1	Apprendimento Semi Supervisionato	5
1.3.2	Apprendimento per Rinforzo	5
2	Infrastruttura di calcolo per machine learning	6
2.1	On-Premise	6
2.1.1	Vantaggi	6
2.1.2	Svantaggi	7
2.2	Cloud	8
2.2.1	Vantaggi	8
2.2.2	Svantaggi	9
3	Architettura Necessaria	10
3.0.1	Memoria	10
3.0.2	RAM	11
3.0.3	Schede Video	11
4	Cloud Services	13
4.1	SaaS	13
4.1.1	Vantaggi del modello SaaS	14
4.1.2	Svantaggi del modello SaaS	16
4.2	PaaS	18
4.2.1	Vantaggi del modello PaaS	19
4.2.2	Svantaggi del modello PaaS	19
4.3	IaaS	21
4.3.1	Vantaggi del modello IaaS	21
4.3.2	Svantaggi del modello IaaS	23
5	Aziende sul mercato	24
5.1	AWS	24
5.1.1	SageMaker	25
5.1.2	Amazon SageMaker Autopilot	26
5.1.3	Amazon SageMaker Data Wrangler	27
5.1.4	Notebook Amazon SageMaker Studio	28
5.1.5	Amazon SageMaker Experiments	29
5.1.6	Debugger Amazon SageMaker	29
5.1.7	Informazioni Aggiuntive	31
5.1.8	AWS per Formula1	33
5.2	Azure	35

5.2.1	Azure Synapse Analytics	36
5.2.2	Azure Databrick	36
5.2.3	Analisi di flusso di Azure	37
5.2.4	Azure Machine Learning	38
5.2.5	Azure per Forza Horizon 5	41
5.3	Google Cloud Platform	44
5.3.1	Vertex AI	45
5.3.2	Vertex AI Workbench	46
5.3.3	Whitepaper GCP	47
6	Algoritmi di machine learning parallelizzabili	49
6.1	Parallelismo del modello	49
6.2	Parallelismo dei Dati	50
6.3	Traning Distribuito	51
6.3.1	MPI	51
6.3.2	Horovod	53
6.3.3	PyTorch	54
6.3.4	TensorFlow	59
6.4	Considerazioni	61
6.4.1	Vantaggi del Training Distribuito	61
6.4.2	Svantaggi del Training Distribuito	62
7	Conclusioni	63

Introduzione

Oggetto di questa relazione e' la trattazione delle tecnologie Cloud Services presenti sul mercato per l'ottimizzazione e la parallelizzazione di modelli di Machine Learning. La relazione sarà suddivisa 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 adottare. Successivamente 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 Computing offre differenti tipi di servizi che sono: SaaS (Software as a service), PaaS (Platform as a service) e IaaS (Infrastructure as a service). Essi verranno trattati e spiegati fornendone i vantaggi e svantaggi.

Dopo di ciò verranno presentate le aziende leader sul mercato AWS[1], Azure[2] e GCP[3].

Per ognuna di esse verranno riportati i servizi disponibili per lo sviluppo di progetti di machine learning, discutendo il contenuto di alcuni whitepaper, paper aziendali, delle aziende sopra citate per far vedere in concreto le possibilità nell'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.

Verrà mostrato in concreto con una piccola guida 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 supportate ovvero PyTorch[**26**] e TensorFlow[**27**], specificando alcune soluzioni possibili che offrono le librerie in collaborazione con Azure.

1 Machine Learning

Il Machine Learning (ML)[4] è un sottoinsieme dell'intelligenza artificiale (AI) che si occupa di creare sistemi in grado di apprendere o migliorare le prestazioni in base ai dati utilizzati. Il termine intelligenza artificiale si riferisce a sistemi o macchine che imitano l'intelligenza umana, simulando il funzionamento del cervello umano. Anche se i termini machine learning e AI sono spesso usati insieme e in modo intercambiabile, non hanno lo stesso significato. Una distinzione importante è che, mentre tutto ciò che riguarda il machine learning rientra nell'intelligenza artificiale, l'intelligenza artificiale non include solo il machine learning. Il machine learning è suddiviso in due sottocategorie in base al tipo di apprendimento che viene effettuato.

1.1 Apprendimento 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 Training Set, e con un output predefinito.

La difficoltà nel seguente approccio sta nell'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, un obiettivo potrebbe essere 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[5] e logistica[6], di classificazione multi-classe e le macchine a vettori[7] di supporto sono alcuni esempi di machine learning supervisionato.

1.2 Apprendimento non Supervisionato

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, potremmo affermare che 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[31], 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 [8].

1.3.1 Apprendimento Semi Supervisionato

Ha le stesse applicazioni dell'apprendimento supervisionato. Ma per l'addestramento utilizza sia dati classificati che non classificati. Solitamente si utilizza 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 di calcolo per machine learning

Tutto quello 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 eseguire 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 [9], 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, il che può comportare una spesa non indifferente e piuttosto onerosa.
- L'aggiornamento delle applicazioni può rappresentare un'arma a doppio taglio, poiché sarà svolto dagli impiegati dell'azienda stessa, aumentando la loro mole di lavoro, oppure richiedendo la figura di un esperto in aziende più grandi.
- Spesso le aziende sviluppatrici invece di fornire nuovi aggiornamenti, preferiscono proporre direttamente un nuovo software, spingendo le aziende all'acquisto e a sostenere un'ulteriore spesa.
- Le aziende devono garantire la sicurezza dei dati archiviati sui loro server, il che avviene attraverso l'acquisto di licenze per la sicurezza, che possono essere abbastanza costose. In caso di fuga di dati o di attacchi esterni, l'azienda stessa sarà ritenuta responsabile.

2.2 Cloud

Le soluzioni in Cloud [10] 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.
- Le mansioni non verranno più effettuate dal personale, ciò permetterà di risparmiare risorse e tempo, ma sarà l'host a mettere 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 [11] 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 machine learning si lavora con una grande mole di dati entrando in quello che viene chiamato Big Data [12].

Non vi è però solo la necessità di poter salvare questi dati, quindi infrastrutture capaci di contenere molti tera di dati, ma bisogna gestire anche la condivisione in tempo reale di questi dati per i vari sviluppatori che lavorano al progetto.

Il problema di come conservare i dati e di quanto spazio bisogna avere è un problema che ha bisogno di pianificazioni e studi appositi per ogni progetto, per esempio se si dovesse lavorare con dei video lo spazio richiesto sarebbe maggiore rispetto allo spazio necessario per conservare dati tabulati, come dati clinici.

Per questo motivo solitamente le aziende necessitano di grandi strutture interne che utilizzano come server privati.

3.0.2 RAM

Per poter trasferire 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 sì che il prezzo di tale hardware sia alto, considerando la quantità richiesta di RAM. In più nel caso in cui ad ogni sviluppatore si voglia garantire la stessa potenza bisogna considerare 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 NVIDIA per il machine learning. NVIDIA propone due schede video una nella forma base e un'altra nella forma pro.

La scheda video in figura è NVIDIA RTX 6000 [13], con la seguente scheda tecnica:

Caratteristiche delle GPU	NVIDIA RTX 6000
Memoria della GPU	GDDR6 da 48 GB con ECC
Display Port	4 DisplayPort 1.4*
Max consumo energetico	300 W
Bus grafico	PCIe Gen 4 x 16
Fattore di forma	Doppio slot 4,4" (H) x 10,5" (L)
Termica	Attiva
Supporto software vGPU**	NVIDIA vPC/vApps, NVIDIA RTX Virtual Workstation

Figura 1: 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 efficiente in termini di velocità e memoria è venduta a 6.800\$.

Questo è solo il prezzo di una singola scheda video, 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 Cloud Services

4.1 Saas

Il Software as a Service (SaaS)[14], ovvero il Software come un Servizio, consente agli utenti di accedere ad applicazioni basate sul cloud tramite Internet e di utilizzarle. Tra gli esempi più comuni di applicazioni SaaS ci sono la posta elettronica, i calendari e gli strumenti di produttività.

Il modello SaaS offre una soluzione software completa che può essere acquistata in base al consumo da un provider di servizi cloud. In pratica, si noleggia l'uso di un'app per l'organizzazione e gli utenti si connettono all'applicazione tramite Internet, solitamente utilizzando un web browser.

Tutta l'infrastruttura sottostante, il middleware, il software dell'applicazione e i dati dell'applicazione si trovano nel data center del provider di servizi. Il provider di servizi gestisce sia l'hardware che il software e, con il contratto di servizio appropriato, garantisce la disponibilità e la sicurezza dell'app e dei dati.

Il modello SaaS consente alle organizzazioni di essere operative in modo rapido ed efficiente con un'applicazione che richiede costi iniziali minimi.

4.1.1 Vantaggi del modello SaaS

I vantaggi derivanti dall'adozione di un modello SaaS[15]:

- Non è necessario acquistare, installare, aggiornare o gestire hardware, middleware o software.
- Applicazioni aziendali complesse come ERP e CRM diventano accessibili alle organizzazioni che non hanno le risorse per acquistare, distribuire e gestire l'infrastruttura e il software necessari.
- Il pagamento è proporzionato all'utilizzo, il che significa che le organizzazioni possono risparmiare grazie alla scalabilità verticale automatica in base al livello di utilizzo.
- I clienti possono utilizzare il software gratuitamente, eseguendo la maggior parte delle app SaaS direttamente dal Web browser senza la necessità di scaricare e installare alcun software.
- Il modello SaaS garantisce anche la mobilità della forza lavoro, in quanto gli utenti possono accedere alle applicazioni e ai dati SaaS da qualsiasi dispositivo mobile o computer connesso a Internet, inoltre non è necessario preoccuparsi di sviluppare app per eseguirle su diversi tipi di computer e dispositivi, poiché il provider di servizi SaaS gestisce tutto.
- Le organizzazioni non devono impiegare personale con un'esperienza speciale per gestire le problematiche di sicurezza correlate all'uso di dispositivi mobili, poiché il provider di servizi cloud si occupa di tutto.
- Una scelta attenta del provider di servizi garantisce la sicurezza dei dati, indipendentemente dal tipo di dispositivo utilizzato.

- Si può accedere ai dati delle applicazioni da qualsiasi luogo, poiché i dati sono archiviati nel cloud. Ciò significa che gli utenti possono accedere alle informazioni da qualsiasi dispositivo mobile o computer connesso a Internet e che i dati non andranno persi in caso di problemi al computer o al dispositivo di un utente.

4.1.2 Svantaggi del modllo SaaS

Alcuni svantaggi possono essere[16]:

- L'utilizzo dei Software as a Service (SaaS) comporta il trasferimento dei dati dell'azienda a un fornitore di servizi esterno, il che può comportare alcune criticità.
 - In primo luogo, il provider potrebbe non adottare una policy di riservatezza adeguata per proteggere i dati aziendali caricati sui suoi server. Pertanto, è di fondamentale importanza scegliere un fornitore di servizi trasparente e affidabile, che garantisca la sicurezza e la riservatezza dei dati;
 - In secondo luogo, il provider potrebbe non avere un'infrastruttura sicura per ospitare i dati dell'azienda, esponendoli al rischio di furto o accessi non autorizzati.
- In questo caso, anche la connessione a internet deve essere sempre attiva, in quanto i SaaS richiedono una connessione stabile e veloce per funzionare correttamente.
- Alcuni fornitori di SaaS, come ad esempio Google, consentono di utilizzare i loro software anche in modalità offline, sincronizzando i dati non appena la connessione sarà nuovamente disponibile. Questa opzione è particolarmente utile in situazioni di emergenza, ma 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. In caso di fallimento o cessazione dell'attività del provider, l'applicazione diventerebbe inaccessibile, causando

potenziali problemi aziendali. Pertanto, è importante scegliere un fornitore di servizi affidabile e di lunga durata per minimizzare questo rischio.

- Il trasferimento dei dati dell'azienda a un fornitore di servizi esterno comporta alcune criticità, tra cui la protezione dei dati, la sicurezza dell'infrastruttura e la cessazione dei servizi. Nonostante ciò scegliendo un fornitore di servizi affidabile e attento a questi problemi, è possibile sfruttare i vantaggi dei SaaS, come l'accesso remoto ai dati, la flessibilità e la riduzione dei costi di gestione IT.

4.2 PaaS

Il modello PaaS (Platform as a Service)[17] è stato sviluppato per fornire un'alternativa ai problemi che sorgono quando le aziende cercano di sviluppare e distribuire le proprie applicazioni in modo efficiente nel cloud. La piattaforma PaaS fornisce un ambiente completo per lo sviluppo e la distribuzione di applicazioni, con risorse che consentono di distribuire qualsiasi cosa, da semplici app basate sul cloud ad applicazioni aziendali sofisticate abilitate per il cloud.

In pratica, l'utente può acquistare le risorse necessarie da un provider di servizi cloud e accedervi tramite una connessione Internet sicura, senza la necessità di acquistare e gestire licenze software, middleware, agenti di orchestrazione di contenitori come Kubernetes o strumenti di sviluppo e altre risorse. Questo significa che l'utente può concentrarsi sulla gestione delle applicazioni e dei servizi che sviluppa, mentre il provider di servizi cloud gestisce tutto il resto.

Il modello PaaS fornisce molti vantaggi, tra cui la riduzione delle spese e della complessità legate all'acquisto e alla gestione di risorse, la possibilità di pagare solo per il consumo effettivo di risorse e la disponibilità di un ambiente completo per lo sviluppo e la distribuzione di applicazioni. Inoltre, il modello PaaS consente di aumentare la produttività degli sviluppatori, ridurre i tempi di rilascio delle applicazioni e migliorare l'affidabilità e la scalabilità delle applicazioni distribuite nel cloud.

4.2.1 Vantaggi del modllo PaaS

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 poiché non subbentreranno i costi per la realizzazione dell'infrastruttura, per le manutenzioni, per gli aggiornamenti e/o per l'acquisto di nuove licenze software.

4.2.2 Svantaggi del modllo PaaS

- Poichè è il fornitore a occuparsi della fornitura non si ha alcun controllo sull'infrastruttura e non si può implementare da soli le feature.
- 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, nonostante 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

Il servizio di cloud computing IaaS (Infrastructure as a service)[18] fornisce risorse di calcolo, archiviazione e rete on-demand e con pagamento in base al consumo. La migrazione dell'infrastruttura dell'organizzazione a una soluzione IaaS può ridurre la manutenzione dei data center locali, risparmiare sui costi dell'hardware e fornire informazioni dettagliate in tempo reale. IaaS offre la flessibilità necessaria per dimensionare le risorse IT in base alla domanda, permette di effettuare rapidamente il provisioning di nuove applicazioni e aumentare l'affidabilità dell'infrastruttura. L'utilizzo di IaaS consente di evitare il costo e la complessità dell'acquisto e della gestione di server fisici e dell'infrastruttura del data center, e ogni risorsa viene offerta come componente di servizio distinto con pagamento solo per il periodo in cui è necessaria.

4.3.1 Vantaggi del modello IaaS

- 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 nel-

le 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 del modllo IaaS

- 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 effettuare 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 usufruire del servizio.

5 Aziende sul mercato

5.1 AWS

Amazon Web Services, AWS [1], 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.

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

Amazon SageMaker[19] è un servizio di machine learning offerto da AWS che offre un ambiente completamente gestito per preparare i dati e costruire, addestrare e implementare modelli di machine learning. Grazie all'utilizzo di infrastrutture, strumenti e flussi di lavoro completamente gestiti, Amazon SageMaker consente di offrire disponibilità elevata senza previsione di 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. Inoltre, Amazon SageMaker crittografa gli artefatti dei modelli di machine learning e di altri elementi del sistema, sia inattivi sia in transito, e le richieste all'API della console di SageMaker vengono inoltrate tramite una connessione sicura (SSL).

Per quanto riguarda l'accesso alle risorse, si può avvalere del 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 anche 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 machine learning collegato.

Infine, Amazon SageMaker supporta anche il cloud privato virtuale (VPC)

di Amazon e AWS PrivateLink, offrendo così un ulteriore livello di sicurezza per le applicazioni di machine learning che richiedono di operare in un ambiente privato e isolato.

5.1.2 Amazon SageMaker Autopilot

Amazon SageMaker Autopilot rappresenta una grande innovazione nel settore del machine learning automatico, offrendo una totale visibilità e controllo sui modelli di ML. Questa funzionalità è in grado di analizzare automaticamente i dati grezzi, applicare i processor di caratteristiche, selezionare il miglior set di algoritmi, addestrare e ottimizzare diversi modelli, controllarne le prestazioni e stilare una classifica dei modelli migliori, il tutto in pochi clic. Il risultato è un modello dalle migliori prestazioni possibili, che può essere distribuito in una frazione del tempo normalmente richiesto per l'addestramento del modello. Con SageMaker Autopilot è possibile avere una completa trasparenza su come il modello è stato creato e quali componenti contiene, grazie alla sua integrazione con Amazon SageMaker Studio. Inoltre, è possibile selezionare fino a 50 modelli diversi generati da SageMaker Autopilot all'interno di SageMaker Studio, semplificando la scelta del modello migliore per il caso d'uso. SageMaker Autopilot può essere utilizzato da chiunque, anche da coloro che non hanno esperienza di machine learning, ma anche da sviluppatori esperti per creare un modello di base per iterazioni successive. Amazon SageMaker JumpStart offre oltre 150 modelli open source pre-caricati da PyTorch Hub e TensorFlow Hub. Per le attività visive come la classificazione di immagini e il rilevamento di oggetti, sono disponibili modelli come ResNet, MobileNet e Single-Shot Detector (SSD). Per le attività testuali come la classificazione di frasi, la classificazione dei testi e le risposte alle domande, invece, è possibile utilizzare modelli come BERT, RoBERTa e

DistilBERT. SageMaker JumpStart offre soluzioni preconfigurate con tutti i servizi AWS necessari per avviare rapidamente una soluzione in produzione, totalmente personalizzabili per adattarsi alle esigenze specifiche di ciascun caso d'uso e set di dati.

5.1.3 Amazon SageMaker Data Wrangler

Amazon SageMaker Data Wrangler[20] è uno strumento che semplifica l'aggregazione e la preparazione dei dati per l'apprendimento automatico, riducendo il tempo necessario per tali attività. Grazie ad una singola interfaccia all'interno di Amazon SageMaker Studio, gli utenti possono facilmente cercare e importare dati da diverse fonti, come Amazon S3, Amazon Athena, Amazon Redshift, AWS Lake Formation, Amazon SageMaker Feature Store e Snowflake. Possono essere interrogati e importati dati da oltre 40 origini diverse, registrati nel Catalogo dati AWS Glue attraverso Amazon AppFlow. In questo modo, gli utenti hanno accesso ad una vasta gamma di dati e possono manipolare facilmente i dati in modo da renderli pronti per l'analisi e l'elaborazione.

5.1.3.1 Visualizzazione dei dati

SageMaker Data Wrangler[20] 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. Permette 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 risultati

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 memo-

ria 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 di machine learning nel Cloud

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. 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 disposizione 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 AWS per Formula1

La Formula 1 (F1) ha scelto AWS per le seguenti ragioni[21]:

"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.*

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 piloti. 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.

La F1 sta studiando l'utilizzo del Machine Learning nelle proprie simulazio-

ni, fornendo nuove informazioni utili all'organizzazione con oltre 550 milioni di punti dati 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.

Uno dei vantaggi più importanti dato dall'utilizzo di questo servizio è stata la possibilità di poter prelevare ed elaborare i dati in tempo reale, in modo da fornire ai fan quasi simultaneamente informazioni utili per seguire la gara durante la gara stessa.

Un ulteriore vantaggio, non ai fini del marketing, è stato l'utilizzo di AWS High Performance Computing per poter, tramite delle simulazioni, produrre autovetture più performanti sia a livello di consumo che di velocità. Questo è un vantaggio notevole poiché sono stati utilizzati dati di cui già erano muniti e non hanno dovuto progettare da zero nuovi veicoli con costi e tempi elevati, ma tramite l'utilizzo di simulazioni hanno potuto effettuare più prove per poi trovare la soluzione migliore per poi progettare e produrre solo l'auto migliore.

5.2 Azure

Azure è una piattaforma cloud pubblica di Microsoft che offre una vasta gamma di servizi, tra cui funzionalità PaaS (Platform as a Service), IaaS (Infrastructure as a Service) e servizi di database gestiti. Questa piattaforma consente ai clienti di creare, distribuire e gestire le loro applicazioni e servizi su una vasta gamma di infrastrutture distribuite in tutto il mondo. Azure è progettata per essere altamente scalabile, sicura e affidabile, con funzionalità avanzate di sicurezza, monitoraggio e gestione. La piattaforma offre un'ampia scelta di strumenti di sviluppo e integrazione, supportando diversi linguaggi di programmazione e tecnologie, tra cui: .NET, Java, Python e Node.js.

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

- Azure Synapse Analytics;
- Azure Databricks;
- Microsoft Purview;
- Data Factory di Azure;
- HDInsight;
- 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 [22] è un servizio di analisi che offre la possibilità di eseguire query sui dati in base alle esigenze, utilizzando opzioni serverless o dedicate su larga scala. Combina integrazione dei dati, funzionalità aziendali di data warehousing e analisi dei Big Data in un'unica esperienza unificata per l'inserimento, l'esplorazione, la preparazione, la trasformazione, la gestione e la distribuzione dei dati. Grazie a questo servizio, è possibile fornire informazioni dettagliate da tutti i dati, in diversi data warehouse e sistemi di analisi dei Big Data, in modo veloce ed efficiente, e applicare modelli di Machine Learning a tutte le app intelligenti.

Azure Synapse Analytics riduce notevolmente il tempo necessario per lo sviluppo dei progetti, eliminando gli ostacoli nei dati ed eseguendo analisi sui dati operativi e di app aziendali con collegamento ad Azure Synapse senza spostare i dati, garantendo la massima sicurezza e privacy con funzionalità avanzate come la sicurezza a livello di colonna e di riga e la 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 SparkTM[32] 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 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 [23] descritto è un'offerta di Azure, la piattaforma cloud di Microsoft, che consente di elaborare flussi di dati in tempo reale senza dover gestire l'infrastruttura sottostante. La pipeline di streaming serverless end-to-end permette di configurare rapidamente una pipeline di elaborazione dei dati che preleva i dati da una o più origini, li

elabora utilizzando SQL o codice personalizzato in JavaScript o C#, e li invia ai sink desiderati.

Il servizio offre funzionalità avanzate per l'elaborazione di flussi di dati in tempo reale, come la possibilità di garantire l'elaborazione di eventi "Exactly Once", una disponibilità del 99,9% e funzionalità di ripristino predefinite. Il servizio consente di acquisire informazioni dettagliate e di analisi in tempo reale più vicine alla posizione di origine dei dati, il che rende possibili nuovi scenari con vere architetture ibride per l'elaborazione di flussi.

Il servizio di analisi in tempo reale di Azure offre anche 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. La configurazione avviene facilmente tramite una pipeline di integrazione continua e recapito continuo (CI-CD) con latenze inferiori al secondo per i carichi di lavoro più complessi.

In sintesi, il servizio di analisi in tempo reale di Azure è una soluzione potente e facile da usare per l'elaborazione di flussi di dati in tempo reale, che offre funzionalità avanzate e modelli predefiniti di machine learning per aiutare a ottenere informazioni dettagliate in modo efficiente e rapido.

5.2.4 Azure Machine Learning

Azure Machine Learning[24] è una piattaforma completa che supporta il ciclo di vita di Machine Learning in modo rapido e sicuro. Questa piattaforma offre funzionalità avanzate di MLOps, che consentono di accelerare la creazione, la distribuzione e la gestione dei modelli di Machine Learning. Offre una vasta gamma di strumenti integrati e di interoperabilità open source che consentono di aumentare la produttività del team di sviluppo.

Azure Machine Learning supporta lo sviluppo di modelli di Machine Learning attraverso la funzionalità di studio, che offre un'esperienza di sviluppo completa per creare, eseguire il training e distribuire i modelli. È possibile collaborare con Jupyter Notebook utilizzando i framework e le librerie open source più diffusi. Permette creazione di modelli accurati utilizzando il Machine Learning automatizzato per modelli tabulari, di testo e di immagini mediante la progettazione delle funzionalità e lo sweep degli iperparametri, in maniera rapida e accurata. Azure Machine Learning offre anche la possibilità di passare dal training da locale al cloud utilizzando Visual Studio Code con scalabilità automatica con potenti cluster di CPU e GPU basati sul cloud con tecnologia della rete NVIDIA Quantum InfiniBand.

Azure Machine Learning semplifica la distribuzione e la gestione di migliaia di modelli in più ambienti utilizzando MLOps. È possibile automatizzare i flussi di lavoro utilizzando pipeline ripetibili per l'integrazione continua e il recapito continuo (CI/CD). Azure Machine Learning offre endpoint completamente gestiti per previsioni in batch e in tempo reale, che consentono di accelerare la distribuzione e l'assegnazione di un punteggio ai modelli. È possibile condividere e personalizzare gli artefatti di Machine Learning tra più team per la collaborazione tra aree di lavoro utilizzando i registri. Permette anche il monitoraggio continuo delle metriche relative alle prestazioni dei modelli, rilevando la derivata dei dati e attivando la ripetizione del training per migliorare le prestazioni dei modelli. Azure Machine Learning consente anche la valutazione dei 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.

Azure Machine Learning aumenta la sicurezza nel ciclo di vita di Machine

Learning offrendo funzionalità complete che includono identità, dati, rete, monitoraggio e conformità. È possibile proteggere le soluzioni utilizzando il controllo degli accessi in base al ruolo personalizzato, le reti virtuali, la crittografia dei dati, gli endpoint privati e gli indirizzi IP privati. Permette l'esecuzione del training e la distribuzione dei modelli in locale per soddisfare i requisiti di sovranità dei dati. Queste funzionalità offrono maggiore sicurezza e riducono il rischio di fuga di dati e di violazione della sicurezza durante il ciclo di vita di Machine Learning.

5.2.5 Azure per Forza Horizon 5

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**[25] ha usufruito dei servizi di Azure per i suoi scopi.

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 esperienze 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. Grazie al servizio 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 mi-

nimo 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.

5.3 Google Cloud Platform

Google Cloud Platform (GCP) è una piattaforma di cloud computing che offre una vasta gamma di servizi, tra cui l'archiviazione, l'elaborazione e l'analisi dei dati, la gestione delle macchine virtuali, l'intelligenza artificiale, la sicurezza e molto altro. La piattaforma è costruita sulla stessa infrastruttura che Google utilizza per i propri servizi, il che significa che gli utenti di GCP beneficiano di una tecnologia avanzata e scalabile che è stata testata in modo estensivo a livello globale.

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 utiliz-

zare 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.

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**[26] e **TensorFlow**[27].

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[28] del **modello** e dei **dati**.

6.1 Parallelismo del modello

Il parallelismo del modello è un'approccio al parallelismo nell'elaborazione dei dati in cui un modello viene suddiviso in parti e ciascuna parte viene eseguita contemporaneamente in diversi nodi di elaborazione. Questo approccio è anche noto come parallelismo di rete.

Il grado di parallelismo dipende dalla complessità dell'algoritmo e dal numero di parti in cui il modello viene suddiviso. A differenza del parallelismo dei dati, in cui i dati vengono suddivisi tra i nodi di lavoro, nel parallelismo del modello ogni nodo di lavoro opera sugli stessi dati di training.

Il parallelismo del modello richiede una maggiore complessità nell'implementazione rispetto al parallelismo dei dati, ma presenta alcuni vantaggi. Ad esempio, i modelli di dimensioni maggiori non rappresentano un problema in quanto ogni nodo opera solo su una sottosezione del modello.

Nel parallelismo del modello i nodi di lavoro devono sincronizzare solo i

parametri condivisi, solitamente una volta per ogni passaggio di propagazione in avanti o all'indietro. Ciò significa che il parallelismo del modello è più efficiente dal punto di vista computazionale rispetto ad altri approcci di parallelismo.

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.

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 Training Distribuito

6.3.1 MPI

Azure Machine Learning offre un processo MPI[29] per il training distribuito di modelli di deep learning. Il processo MPI di Azure Machine Learning consente di avviare un determinato numero di processi in ogni nodo, con un comando di avvio MPI completo (mpirun) costruito automaticamente dietro le quinte.

Tuttavia, non è possibile fornire comandi head-node-launcher completi come mpirun o DeepSpeed launcher.

Per eseguire il training distribuito tramite MPI, è necessario seguire alcuni passaggi:

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 oggetto **command** con *instance_count*. L'*instance_count* deve essere uguale al numero di GPU per nodo per l'avvio per processo oppure impostato 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 parametro di distribuzione di command per specificare le impostazioni per MpiDistribution. La distribuzione MPI può essere configurata per utilizzare diverse strategie, come la replicazione del modello e la suddivisione dei dati. È possibile specificare il numero di processi MPI per nodo e il numero totale di processi da avviare.

È possibile definire un oggetto `command` e specificare le impostazioni per `MpiDistribution` per avviare il training distribuito.

```
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"
)
```


6.3.2 Horovod

Horovod[30] è un framework di deep learning distribuito che supporta diverse configurazioni di processo per eseguire il training di modelli su cluster di GPU. MPI (Message Passing Interface) è uno dei protocolli di comunicazione supportati da Horovod per la gestione della comunicazione tra i processi.

Quando si utilizza Horovod con il framework di deep learning su Azure Machine Learning, è consigliabile utilizzare la configurazione del processo MPI per garantire che il codice segua le migliori pratiche. Ciò include l'istru-mentazione corretta del codice di training con Horovod prima di aggiungere parti di Azure Machine Learning, l'uso di ambienti di Azure Machine Learning che contengono Horovod e MPI, e la creazione di un oggetto command con la distribuzione desiderata.

Gli ambienti GPU curati di PyTorch e TensorFlow sono preconfigurati con Horovod e le relative dipendenze, semplificando notevolmente l'uso di Horovod per il training distribuito di modelli di deep learning su cluster di GPU. Quindi l'uso della configurazione del processo MPI con Horovod per il training distribuito di modelli di deep learning su Azure Machine Learning può aiutare a garantire una migliore gestione della comunicazione tra i processi e una maggiore efficienza complessiva del training.

6.3.3 PyTorch

Azure Machine Learning (AML) supporta l'esecuzione di processi distribuiti utilizzando le funzionalità di training distribuito native di PyTorch (`torch.distributed`).

Il training distribuito è una tecnica che consente di distribuire il carico di lavoro del training su più nodi di calcolo, accelerando il processo di training e consentendo di gestire set di dati più grandi. Il training distribuito in PyTorch si basa su un gruppo di processi che si conoscono tra loro e possono comunicare tra loro utilizzando un back-end.

Per creare il gruppo di processi in PyTorch, tutti i processi distribuiti devono chiamare la funzione `torch.distributed.init_process_group`. Questa funzione permette ai processi di unirsi collettivamente per formare un gruppo di processi che collaboreranno durante il training.

Durante l'esecuzione di un training distribuito, PyTorch utilizza una serie di tecniche per suddividere il modello, i dati e i gradienti su più nodi di calcolo. In questo modo, il carico di lavoro viene distribuito in modo equo tra i nodi, consentendo di ottenere un'accelerazione significativa del processo di training.

AML semplifica l'implementazione di training distribuito in PyTorch fornendo un'interfaccia intuitiva e facile da usare che consente di gestire il training distribuito in modo trasparente, senza dover preoccuparsi di gestire l'infrastruttura sottostante.

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

I back-end di comunicazione più comuni usati sono `mpi`, `nccl` e `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 corrispondente `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_SIZE` e `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

# === Note on path ===
# can be can be a local path or a cloud path. AzureML supports https://`, `abfss://`, `wasbs://` and
# `azureml://` URIs.
# Local paths are automatically uploaded to the default datastore in the cloud.
# More details on supported paths: https://docs.microsoft.com/azure/machine-learning/how-to-read-write-data-v2#supported-paths

inputs = {
    "cifar": Input(
        type=AssetTypes.URI_FOLDER, path=returned_job.outputs.cifar.path
    ), # path="azureml:azureml_stoic_cartoon_wgb3lgvgky_output_data_cifar:1",
    #path="azureml://datastores/workspaceblobstore/paths/azureml/stoic_cartoon_wgb3lgvgky/cifar/"),
    "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 ${inputs.epoch} --batch-size\n${inputs.batchsize} --workers ${inputs.workers} --learning-rate ${inputs.lr} --momentum\n${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", # Change the name to the gpu cluster of your workspace.
    instance_count=2, # In this, only 2 node cluster was created.
    distribution={
        "type": "PyTorch",
        # set process count to the number of gpus per node
        # NV6 has only 1 GPU
        "process_count_per_instance": 1,
    },
)

```

Figura 2: Codice implementativo

6.3.3.1 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.

6.3.4 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 parameters* o l'oggetto *TensorFlowDistribution*.

```
# create the command
job = command(
    code="./src", # local path where the code is stored
    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
    # description: Train a basic neural network with TensorFlow on the MNIST dataset, distributed via
    TensorFlow.
)

# can also set the distribution in a separate step and using the typed objects instead of a dict
job.distribution = TensorFlowDistribution(parameter_server_count=1, worker_count=2)
```

Figura 3: Codice implementativo

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,

6.3.4.1 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 Considerazioni

La sezione precedente spiega in dettaglio i concetti di base del training distribuito, le due forme di parallelismo offerte da Azure (parallelismo del modello e parallelismo dei dati) e come utilizzare il processo MPI per il training distribuito. Inoltre, vengono menzionati strumenti come Horovod e DeepSpeed che semplificano l'implementazione del training distribuito. Nei paragrafi successivi verranno evidenziati i vantaggi e gli svantaggi del training distribuito con Azure e le librerie sopra menzionate.

6.4.1 Vantaggi del Training Distribuito

- Il training distribuito suddivide il carico di lavoro tra più nodi di lavoro, consentendo una maggiore velocità di training del modello.
- Il parallelismo del modello permette di suddividere il modello stesso in parti eseguite contemporaneamente in diversi nodi di elaborazione, offrendo maggiore efficienza computazionale.
- Il parallelismo del modello è adatto per modelli di dimensioni maggiori, in quanto ogni nodo di lavoro opera solo su una sottosezione del modello.
- Azure Machine Learning offre un processo MPI per il training distribuito di modelli di deep learning, semplificando l'avvio e la gestione dei processi distribuiti.

6.4.2 Svantaggi del Training Distribuito

- L'implementazione del parallelismo del modello richiede maggiore complessità rispetto al parallelismo dei dati.
- Nel parallelismo del modello, i nodi di lavoro devono sincronizzare i parametri condivisi, aggiungendo un'ulteriore complessità all'implementazione.
- Nel parallelismo dei dati, ogni nodo di lavoro deve adattare interamente il modello, quindi è necessaria una maggiore capacità di supporto del modello per ogni nodo.
- L'uso di processi MPI per il training distribuito richiede la configurazione adeguata e l'uso di strumenti come Horovod per garantire una comunicazione efficiente tra i processi.

7 Conclusioni

Il machine learning è diventato una tecnologia sempre più importante per molte aziende e organizzazioni che cercano di ottenere vantaggi competitivi. Con l'avvento dell'era del cloud computing, le soluzioni di machine learning basate sul cloud stanno diventando sempre più popolari, offrendo vantaggi come la scalabilità, la flessibilità e la riduzione dei costi operativi.

Esiste ancora una differenza significativa tra le soluzioni di machine learning on-premise e cloud computing. Le soluzioni on-premise offrono maggiore controllo e personalizzazione ma richiedono una maggiore infrastruttura e competenze tecniche. D'altra parte, le soluzioni di machine learning basate sul cloud sono più facili da implementare e gestire ma possono essere limitate in termini di personalizzazione e controllo.

La scelta tra una soluzione di machine learning on-premise e una basata sul cloud dipende dalle esigenze specifiche dell'organizzazione e dalle risorse disponibili. Le soluzioni basate sul cloud possono essere una scelta ideale per le organizzazioni che cercano di implementare il machine learning in modo rapido ed economico, mentre le soluzioni on-premise possono essere la scelta migliore per organizzazioni con requisiti specifici di sicurezza, personalizzazione e controllo.

Bisogna considerare anche le differenze tra le soluzioni di machine learning on-premise e cloud computing anche in termini di prestazioni e scalabilità. Le soluzioni basate sul cloud, come ad esempio Azure Machine Learning o AWS SageMaker, offrono un'ampia gamma di risorse di calcolo e di archiviazione scalabili che possono essere facilmente adattate alle esigenze dell'organizzazione. Queste soluzioni sono dotate di algoritmi di ottimizzazione automatica che consentono di ottimizzare le prestazioni del modello in modo

efficiente.

D'altra parte, le soluzioni di machine learning on-premise richiedono una maggiore attenzione alla configurazione dell'infrastruttura, alla gestione della sicurezza e alla manutenzione. La scalabilità delle soluzioni on-premise può essere limitata dalle risorse hardware disponibili e dalla complessità dell'infrastruttura.

Infine, le soluzioni di machine learning basate sul cloud offrono anche vantaggi in termini di collaborazione e condivisione dei dati. Grazie alla possibilità di archiviare e condividere i dati in modo sicuro e centralizzato, le organizzazioni possono collaborare facilmente con team di data scientist e condividere facilmente i dati e i modelli di machine learning in modo da migliorare le prestazioni del modello e accelerare il processo di sviluppo.

In conclusione, la scelta tra una soluzione di machine learning on-premise e una basata sul cloud dipende dalle specifiche esigenze dell'organizzazione in termini di prestazioni, scalabilità, personalizzazione e sicurezza. Le soluzioni basate sul cloud offrono vantaggi significativi in termini di flessibilità, scalabilità e collaborazione, ma è importante considerare anche le limitazioni e le sfide legate alla gestione dell'infrastruttura cloud e alla protezione dei dati.

Riferimenti bibliografici

AWS

[1] [1] <https://aws.amazon.com/it/>.

Azure

[2] <https://azure.microsoft.com/it-it/>.

Google Cloud Platform

[3] <https://cloud.google.com/?hl=it>

Machine learning

[4] <https://www.oracle.com/it/artificial-intelligence/machine-learning/what-is-machine-learning/>

Regressore Lineare

[5] <https://paperswithcode.com/method/linear-regression>

Regressore Logistico

[6] <https://paperswithcode.com/method/logistic-regression>

Support Vector Machine

[7] <https://paperswithcode.com/paper/deep-learningusing-linear-support-vector>

Appendimento Misto

[8] https://www.sas.com/it_it/insights/analytics/machine-learning.html

On-Premise

[9] <https://www.it-impresa.it/blog/software-on-premise/>
<https://www.it-impresa.it/blog/on-premise-cloud/>

Cloud

[10] <https://www.it-impresa.it/blog/on-premise-cloud/>

TrainingCompare

[11] Title:Acceleratingrecurrentneuralnetworksinanalyticsservers:
ComparisonofFPGA,CPU,GPU,andASIC;URL:<https://ieeexplore.ieee.org/abstract/document/7577314/metrics#metrics>

Big Data

[12] https://www.researchgate.net/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper

RTX

[13] <https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/rtx-6000/proviz-print-rtx6000-datasheet-web-2504660.pdf>

Saas

[14] [https://azure.microsoft.com/it-it/resources/
cloud-computing-dictionary/what-is-saas](https://azure.microsoft.com/it-it/resources/cloud-computing-dictionary/what-is-saas)

Vantaggi Saas

[15] [https://azure.microsoft.com/it-it/resources/
cloud-computing-dictionary/what-is-saas](https://azure.microsoft.com/it-it/resources/cloud-computing-dictionary/what-is-saas)

Svantaggi di Saas

[16] <https://timeflow.it/blog/ict-tips/cose-il-saas-vantaggi-e-svantaggi/>

Paas

[17] [https://azure.microsoft.com/it-it/resources/
cloud-computing-dictionary/what-is-paas](https://azure.microsoft.com/it-it/resources/cloud-computing-dictionary/what-is-paas)

Iaas

[18] [https://azure.microsoft.com/it-it/resources/
cloud-computing-dictionary/what-is-iaas](https://azure.microsoft.com/it-it/resources/cloud-computing-dictionary/what-is-iaas)

SageMaker

[19] <https://aws.amazon.com/it/sagemaker/faqs/>

SageMaker Data Wrangler

[20] [https://aws.amazon.com/it/sagemaker/
data-wrangler/?sagemaker-data-wrangler-whats-new.
sort-by=item.additionalFields.postDateTime&](https://aws.amazon.com/it/sagemaker/data-wrangler/?sagemaker-data-wrangler-whats-new.sort-by=item.additionalFields.postDateTime&)

sagemaker-data-wrangler-whats-new.sort-order=desc

Formula 1

[21] <https://aws.amazon.com/it/sports/f1/>

Azure Synapse Analytics

[22] https://azure.microsoft.com/it-it/products/synapse-analytics/?ef_id=_k_4e753241266b1c7ccbdd5677f37e4414_k_&OCID=AIDcmmy6frl1tq_SEM_k_4e753241266b1c7ccbdd5677f37e4414_k_&msclkid=4e753241266b1c7ccbdd5677f37e4414

Azure Flusso

[23] <https://azure.microsoft.com/it-it/products/stream-analytics>

Azure Machine Learning

[24] <https://azure.microsoft.com/it-it/products/machine-learning>

Forza Horizon 5

[25] <https://customers.microsoft.com/en-us/story/1498781140435260527-forza-horizon-5-crosses-finish-line-fueled-by-azure>

Pytorch

[26] <https://pytorch.org>

TensorFlow

[27] <https://www.tensorflow.org/?hl=it>

Parallel

[28] [https://learn.microsoft.com/it-it/azure/machine-learning/
concept-distributed-training](https://learn.microsoft.com/it-it/azure/machine-learning/concept-distributed-training)

MPI

[29] [https://learn.microsoft.com/it-it/azure/machine-learning/
how-to-train-distributed-gpu?view=azureml-api-2](https://learn.microsoft.com/it-it/azure/machine-learning/how-to-train-distributed-gpu?view=azureml-api-2)

Horovod

[30] [https://learn.microsoft.com/it-it/azure/machine-learning/
how-to-train-distributed-gpu?view=azureml-api-2](https://learn.microsoft.com/it-it/azure/machine-learning/how-to-train-distributed-gpu?view=azureml-api-2)

K-means

[31] <https://paperswithcode.com/paper/explainable-k-means-and-k-medians-clustering>

Apache Spark

[32] <https://spark.apache.org>