PENSIERI E CONSIDERAZIONI DALL'AWS RE:INVENT 2019

re:Invent

beSharp | 17 Dicembre 2019

L'AWS re:Invent 2019 si è concluso.
È passata una settimana dalla fine dell'**evento Cloud più importante dell'anno**, il tempo necessario per riprendermi dal jet lag, per sistemare parzialmente lo spaventoso backlog di e-mail che solo una conferenza di una settimana può produrre e anche per divertirmi un po' sulla spiaggia dell'Oceano Pacifico insieme al team prima di tornare a casa. Quindi ora è il momento perfetto per riorganizzare i pensieri sull'evento e condividerli finalmente con voi.

È stato il mio **ottavo re:Invent di fila** (il primo nel 2012): sono abbastanza sicuro di essere l'unico in Italia – e, credo, una tra le pochissime persone al di fuori degli Stati Uniti – a detenere questo piccolo record.

Fun fact: sono italiano, ho 37 anni, e nella mia vita sono stato solo due volte a Venezia e ben otto volte nella "falsa" Venezia nel bel mezzo di Las Vegas; curioso, eh? E non solo per me!

Ho partecipato all'evento con lo stesso entusiasmo della prima volta, anche se dal 2012 molte cose sono cambiate relativamente sia ad AWS che al Cloud Computing in generale. Otto anni fa, i veri Cloud user erano organizzati in una piccola nicchia di pionieri alle prese con l'adozione precoce di un paradigma ancora relativamente nuovo. Oggi, al contrario, la corsa al Cloud ha ormai coinvolto tutti (o quasi). I siti dedicati all'argomento non si contano e migliaia sono i siti web, i blog e i post sui social che ci hanno già fornito carrellate più o meno dettagliate su tutti gli annunci della settimana a Las Vegas, alcuni aggiornamenti addirittura in real-time, mentre gli speaker erano ancora sul palco e parlarne!
Un paio di esempi? Ecco qui... e qui!

Per questo motivo, quest'anno, ho deciso di non esaminare ogni singolo annuncio, descrivendo caratteristiche e dettagli tecnici. Al contrario, mi concentrerò maggiormente su una **visione più a lungo termine di AWS** e dell'utilizzo di vecchi e nuovi servizi alla luce delle notizie a mio parere più dirompenti e rivoluzionarie da Las Vegas.
Quindi, cominciamo!

L'aspetto infrastrutturale è una delle più trascurate: la grande parte dell'hype su AWS è sempre più developer-driven e guidata da "quei tipi hipster che pensano che il mondo finisca ad un endpoint REST" J (cit.)
Ma, come si può facilmente intuire, l'innovazione in questo strato della torta è la chiave per poter rilasciare qualsiasi altro tipo di servizio di livello superiore.

Parlando di **computing**, due dei servizi più interessanti sono Local Zones (ultra-low-latency edge computing per le città metropolitane) e AWS Wavelength (ultra-low-latency edge computing all'interno delle reti 5G dei provider telco).
Entrambi i servizi, a mio avviso, sono basati su AWS Outpost, annunciato l'anno scorso al re:Invent, ma rilasciato in GA lo scorso 3 dicembre. Ho avuto l'opportunità di discutere di Outpost in modo estremamente approfondito (potendolo anche toccare fisicamente con mano per la prima volta all'Expo del re:Invent di quest'anno) con Anthony Liguori, uno dei supervisori del progetto. Outpost, mi ha detto, è un'esatta riproduzione (su scala ridotta) dell'infrastruttura EC2 ed EBS di ultima generazione effettivamente in produzione all'interno delle AZ pubbliche. Questo mi è sembrato piuttosto strano, dato che il rack Outpost è pieno di server 1U (AWS custom) e dispositivi di rete (AWS custom), ma non c'è traccia invece di dispositivi di storage.

La risposta a questa mia obiezione è stata sorprendente... e anche relativamente semplice : tutti i server 1U sono full-length, quindi proprio accanto alla scheda madre c'è molto spazio per un sacco di SSD NVMe. Il chip Nitro gestisce direttamente tutte queste unità, quindi gli stessi dispositivi fisici possono agire sia come EBS che come Instance Storage (effimero), basato sulle diverse richieste degli utenti fatte alle API AWS. Nel caso dell'Instance Storage, il volume logico viene esposto "localmente" attraverso il bus PCIe, mentre l'EBS viene astratto attraverso diversi server e poi esposto attraverso Ethernet grazie all'enorme larghezza di banda di rete disponibile all'interno di Outpost. Immagino che una terza astrazione dello storage di Outpost basata sullo stesso hardware potrebbe essere S3 (senza la riconosciuta "11-nines durability" purtroppo!), che sarà disponibile nel 2020. Non mi sono stati forniti ulteriori dettagli sulla ridondanza dello storage e sulla tolleranza ai guasti (RAID? Erasure coding?), mi hanno solo informato che la magia di tutto questo avviene grazie – di nuovo – al chip Nitro. Quando uno specifico server 1U raggiunge una soglia predefinita di unità difettose, l'intero server viene contrassegnato come "da sostituire" e le procedure di controllo di Instance Storage ed EBS possono reagire di conseguenza. Una soluzione non convenzionale, ma a mio parere ingegnosa. Solo una piccola differenza con le AZ pubbliche: anche se i datacenter di AWS "girano" ufficialmente su un hardware di rete personalizzato (alimentato da ASIC di Annapurna Labs), all'interno di Outpost ci sono dispositivi di rete Juniper standard, ma estremamente potenti. Questo **massimizza la compatibilità** con i dispositivi di rete dei clienti.

Sempre parlando di computing: sembra che AWS stia facendo molto leva sull'acquisto degli Annapurna Labs, dato che gli altri due principali annunci, AWS Inferentia e AWS Graviton 2, riguardano, ancora una volta, il silicio personalizzato. Mentre Inferentia (che alimenta la nuova famiglia di instance Inf1) è semplicemente un chip ML personalizzato con un impressionante rapporto prestazioni/costo rispetto alle alternative standard basate su GPU. La novità rivoluzionaria

è probabilmente Graviton 2. E non per l'impressionante miglioramento delle specifiche rispetto al Graviton 1 dello scorso anno, la prima CPU personalizzata basata su ARM di AWS.

Tutto ciò è rivoluzionario: sembra che la sesta generazione di istanze EC2 si baserà (almeno per i primi tempi) SOLO sull'**architettura ARM**. Ad essere onesti, dubito fortemente che questa sia la fine definitiva dell'era x86 sul Public Cloud – esistono troppi carichi di lavoro ereditati che non saranno mai ricompilati per ARM – ma penso che sia comunque un grande segnale di cambiamento. L'architettura x86 ha 40 anni e nonostante le continue revisioni ed evoluzioni, porta con sè molti problemi e inefficienze, soprattutto a causa della parte "legacy" del set di istruzioni, mantenuta per ragioni di compatibilità. A queste considerazioni bisogna aggiungere che sono molte le voci (autorevoli) in circolazione riguardo al fatto che Apple stia cercando di abbandonare l'architettura Intel (e x86) per i suoi laptop (un altro campo, questo, dove il rapporto prestazioni/efficienza energetica è estremamente critico, esattamente come per il Cloud pubblico), sfruttando i suoi pluripremiati chip custom ARM (già in uso su iPhone e iPad). È una tendenza super-interessante, corroborata dalle scelte di due big della Silicon Valley.

L'annuncio più "visionario" di tutto il re:Invent – a chiusura del nostro capitolo sulla potenza di calcolo – è senza dubbio quello di Amazon Braket (nella speranza che non sia solo un annuncio "marketing-driven" per aggiudicarsi la Quantum Supremacy contro Google e IBM). Il quantum computing è tanto complesso da comprendere in teoria, quanto da implementare nel mondo reale. Per questo motivo, l'arrivo di **infrastrutture quantistiche gestite** (anche se estremamente semplici e allo stadio di prototipo) è una grande opportunità per gli scienziati, fino ad ora impegnati nella difficile ricerca che di un "hardware" adeguato per condurre le loro ricerche in questo campo.

Una menzione d'onore va a Lambda Provisioned Concurrency, che ha finalmente risolto il famigerato problema del cold-start per le applicazioni serverless (pagando una piccola tassa uscendo da un modello di puro on-demand), e a EKS Fargate, annunciato 2 anni fa, ma ancora, secondo me, l'unico modo ragionevole per far funzionare Kubernetes su AWS.

Sul fronte networking nessun annuncio particolarmente eclatante, ma comunque molti annunci interessanti che vanno ad aggiungere nuove funzionalità ai servizi già esistenti, rendendoli più pronti per l'impresa in termini di supporto di topologie di rete complesse (supporto multicast, inter-region peering e Network Manager per Transit Gateway), fornendo prestazioni adeguate (VPN site-

to-site accelerate), e supportando integrazioni di terze parti (VPC ingress routing). Con quest'ultimo, le aziende molto grandi saranno in grado di portare nelle loro VPC appliance di sicurezza e UTM distribuiti tra ogni livello applicativo e ogni subnet per motivi di conformità, non avendo principalmente idea di cosa farne.

Non provateci a casa.

Molti annunci intriganti per **Database, DWH, Data Lakes** aggiungono funzionalità (integrazione di Aurora Machine Learning, Redshift Managed Storage), prestazioni (RDS Proxy, **AQUA per Redshift**) e governance (Athena e Redshift Federated Query). Redshift, in particolare, sta ottenendo un numero impressionante di nuove funzionalità e prestazioni migliorate. Un altro messaggio chiaro e diretto ai clienti Oracle e un altro episodio della battaglia tra le diverse filosofie dei due giganti dell'informatica: **database specializzati vs. database general-purpose**. Chi vincerà?

Il piatto forte dell'intero evento è stato come previsto l'argomento **AI/ML,** il topic di tendenza per eccellenza degli ultimi tre anni nel panorama Cloud, se non addirittura in tutto il mondo IT (chiedo scusa agli amanti delle blockchain!).

Per quanto riguarda i servizi di IA, sembra che gli ingegneri di Seattle stiano mettendo l'esperienza AI di AWS (e di Amazon.com) in servizi e campi sempre più diversi, **dal rilevamento delle frodi al customer care, dal riconoscimento delle immagini, all'analisi del codice**... tra tutte le applicazioni possibili, ecco infatti l'idea di **Amazon CodeGuru** di sfruttare i modelli ML per rivedere e profilare il codice. Ma il servizio è in fase iniziale (per ora solo supporto Java), e discretamente costoso. Sono molto curioso di vedere se le ottimizzazioni delle prestazioni e dei costi derivanti dai **processi di revisione e di profilazione** riusciranno a bilanciare il costo del servizio. Devo solo resuscitare alcuni vecchi pezzi di codice Java per provare 🙂

Lato ML, fondamentalmente, si tratta di **mettere SageMaker sotto steroidi**, aggiungere un IDE dedicato, un ambiente di collaborazione, effettuare esperimenti, introdurre automazione dell'elaborazione, monitoraggio dei modelli e debugging... addirittura si parla di generazione automatica dei modelli. Su quest'ultimo aspetto ho alcuni dubbi: quale sarà il compromesso tra facilità d'uso, precisione e flessibilità? Forse si può chiedere ad alcuni umani...

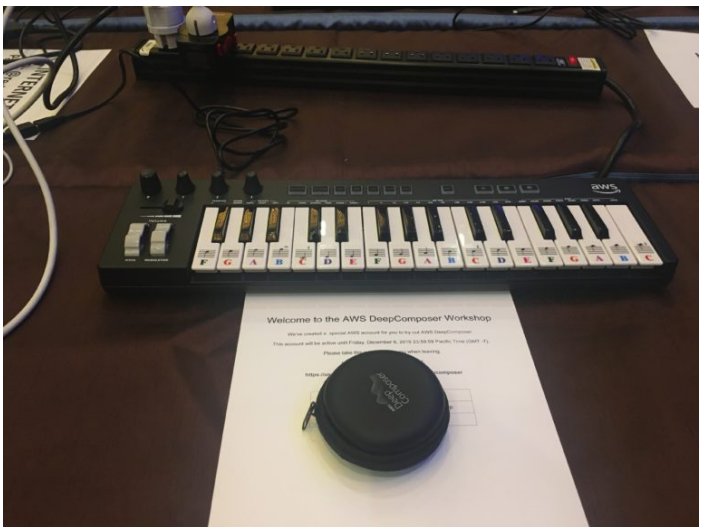
DeepComposer merita una discussione a parte (ma certamente non meritava la mia coda di 6 ore per entrare nell'ultimo re:Invent workshop disponibile...). Sono un tastierista, e quando Matt Wood ha annunciato DeepComposer, ero così eccitato al pensiero di mostrare la potenza del Machine Learning ai nostri clienti mentre suonavo assoli di synth in stile Dream-Theater davanti a loro... Ma poi, ecco il meme "aspettative contro realtà": io che suono una versione con un solo dito di "Twinkle twinkle little star" senza successo cercando di istruire il servizio DeepComposer su come armonizzarlo e organizzarlo... Scherzi a parte, credo che ci siano stati molti fraintendimenti su questo servizio.

Facciamo chiarezza:

- la tastiera non ha nessun inference hardware proprio al suo interno. E' solamente uno standard MIDI controller estremamente semplice e basilare. Il servizio può essere utilizzato con un **MIDI**

input device.

- DeepComposer è in preview privata e, per il momento, ha molte limitazioni: tempo fisso a 100 bpm, una lunghezza di registrazione massima a 4/4.: 2 bars. E' complicato armonizzare melodie monofoniche e arrangiare armonie complesse. Ad oggi, è possibile raggiungere risultati estremamente migliori con una tastiera digitale auto-rythm degli anno '80.
- Se non siete soddisfatti con nessuno dei 5 modelli pronti disponibili (rock, pop, classical, jack, country), potrete (o meglio, in futuro sarà possibile farlo...) creare il vostro modello DeepComposer. Ma questa è probabilmente la parte più complicata dell'intero progetto.



In realtà penso che DeepComposer sia stato lanciato non tanto per essere, esso stesso, un servizio pronto per la produzione, ma per mostrare meglio la potenza delle Reti Generative Avversarie (GAN) – la stessa tecnica ML usata per i Deep Fakes. Una nota molto curiosa a questo proposito è che **nei modelli di DeepComposer disponibili durante l'anteprima, le due reti non vengono addestrate utilizzando file audio, ma utilizzando immagini!** Nello specifico le immagini "piano roll", **rappresentazioni 2D di una melodia suonata nel tempo** . In questo modo è più facile trattare i dati audio come se si trattasse di un'immagine, potendo riutilizzare molte ben note tecniche di estrazione di caratteristiche legate all'immagine. In conclusione: **soluzione brillante; esecuzione da verificare.**

Questo è tutto per ora da Las Vegas! (in realtà da Pavia)
Se vuoi dare un seguito alla discussione sul re:Invent 2019, condividi i tuoi commenti qui sotto o contattaci!

beSharp

Dal 2011 beSharp guida le aziende italiane sul Cloud. Dalla piccola impresa alla grande multinazionale, dal manifatturiero al terziario avanzato, aiutiamo le realtà più all'avanguardia a realizzare progetti innovativi in campo IT.

Get in touch

beSharp.it
proud2becloud@besharp.it