Caccia ai dati: come ottenerli da servizi inaspettati

20 Giugno 2024 - 6 min. read

Quando interagiamo con qualcosa, generiamo dati.

Nel 2016, si prevedeva che un singolo utente internet avrebbe generato 1,7 MB di dati al secondo ogni giorno entro il 2020... Cifre decisamente sottostimate se guardiamo al mondo digitale di oggi.

I dati sono ovunque... basta solo sapere dove cercare!

Come piccoli tesori, a volte dobbiamo spingerci a cercarli nei "posti" più inaspettati, anche in quei servizi che avrebbero tutt'altro scopo rispetto alla generazione di informazioni. Così tutto diventa "dato", ciò che di più prezioso esista per i business di ogni settore.

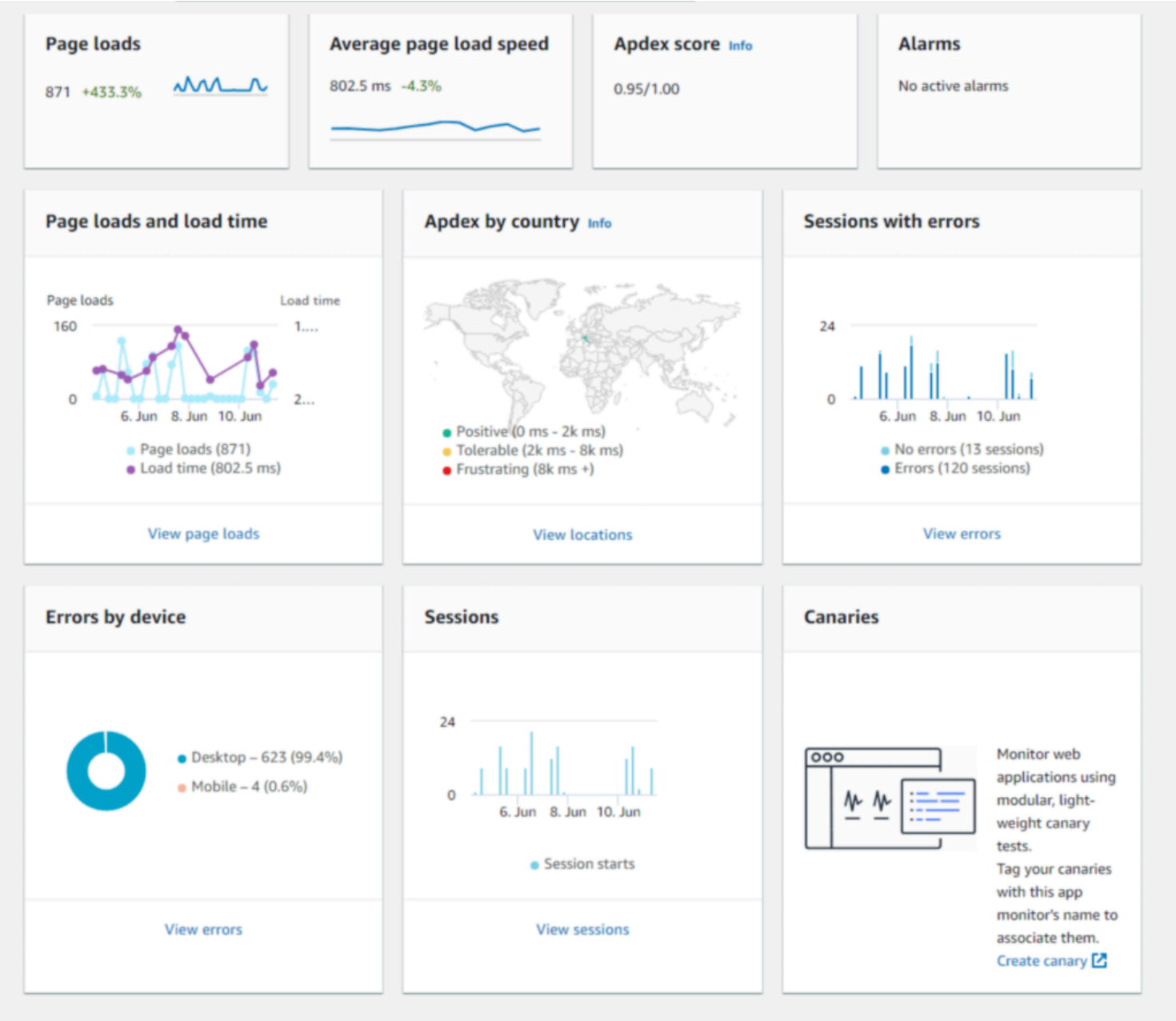
Questo fatto mi ricorda la storia dell'aragosta: storicamente, le aragoste erano considerate cibo per i poveri ed animali indesiderati. I primi americani le usavano addirittura come fertilizzante per il giardino o come esca per la pesca. Prigionieri e servi si lamentavano perché costretti a mangiare aragosta molto spesso... Dagli anni '20, improvvisamente la domanda di aragosta ha iniziato ad aumentare fino a diventare il cibo prezioso che conosciamo.

Vediamo come possiamo trovare dati... nell'aragosta di oggi!

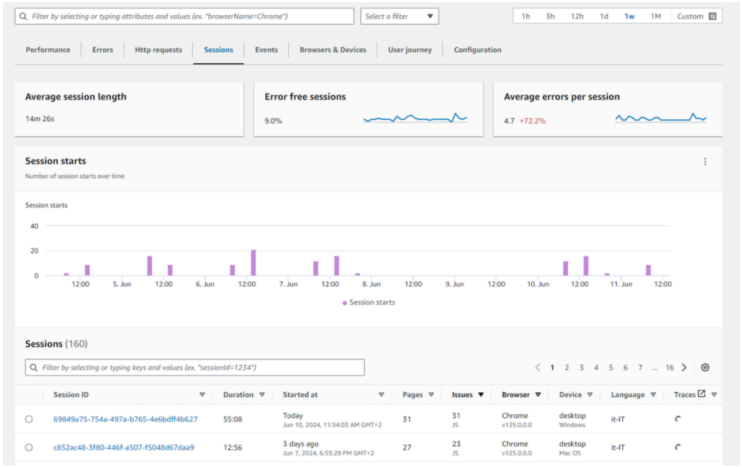
CloudWatch RUM

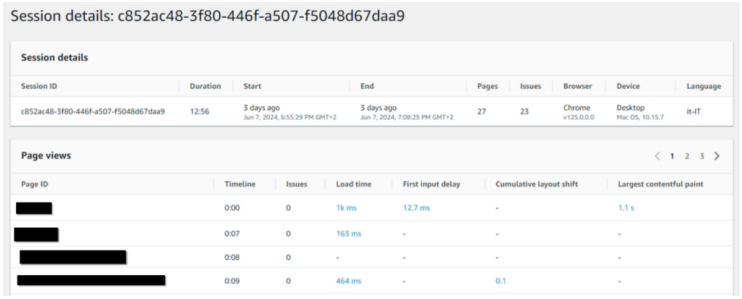
CloudWatch RUM è un'aggiunta relativamente nuova ai servizi Amazon CloudWatch. Monitora l'attività degli utenti reali per tracciare le performance di un sito web (web vitals) e identificare problemi nelle sessioni utente. Il suo vero potere è l'integrazione con X-ray per le applicazioni serverless; permette infatti di tracciare un errore che si è verificato in una lambda e risalire all'attività dell'utente che lo ha generato. Se volete approfondire questo aspetto, i miei colleghi Alessio e Daniele hanno già parlato di X-Ray qui.

CloudWatch RUM mette a disposizione uno snippet da inserire nel sito web che, una volta inserito, invia metriche ed eventi agli endpoint del servizio, permettendo la visualizzazione dei dati in una dashboard. Questo è un esempio:



Si possono anche vedere le sessioni utente, il percorso nel sito e le statistiche sugli errori (vedendone anche il dettaglio).





CloudWatch RUM, oltre ad essere uno strumento utile per il monitoring, può permetterci realmente di prendere **decisioni di business data-driven**.

Sebbene venga sempre utilizzato come strumento di monitoraggio e reporting, può anche essere una fonte di dati che permettono di rivelare informazioni altrimenti difficilemente ottenibili: infatti, il suo vero valore è dato dagli eventi e le metriche catturate dagli utenti reali.

Non ci resta quindi che svelare quali informazioni cela ed il loro potenziale.

Il primo tipo di dati (facilmente disponibili e accessibili) sono le metriche come visualizzazioni di pagina, conteggio delle sessioni, e così via. A questo indirizzo è disponibile l'elenco completo.

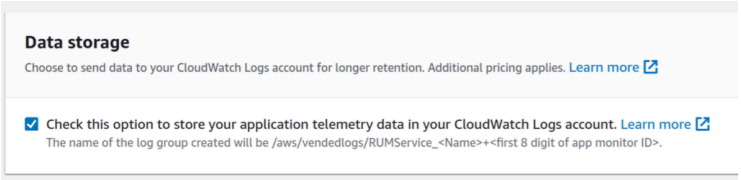
È possibile anche definire metriche personalizzate basate sugli eventi, come ad esempio una metrica personalizzata che conta solo gli accessi da Chrome e Safari dagli Stati Uniti.

Come per ogni metrica, sono impostabili allarmi ed è possibile calcolare metriche derivate, ma questi sono argomenti all'ordine del giorno nel campo monitoraggio.

La seconda tipologia di dati è meno evidente, ma più preziosa: si tratta **eventi raw generati da utenti reali**.

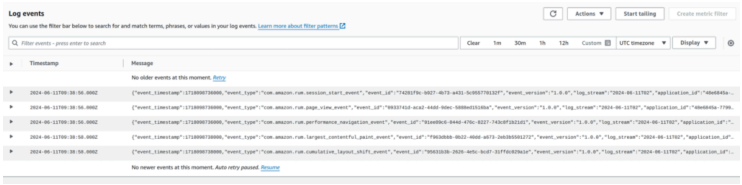
Nella configurazione di RUM è possibile abilitare l'integrazione con CloudWatch Logs. In questo modo, tutti gli eventi relativi alle attività generati dagli utenti vengono inviati quindi

ad un log group in formato JSON.



Attenzione: i log group, per default, non impostano un periodo di retention dei dati e RUM ne genera tanti. Per non avere sorprese nella bolletta è meglio valutare una retention economicamente sostenibile.

Questo è un esempio di log generati da una visualizzazione di pagina:



Abbiamo quindi finalmente i dati su navigazione e performance a disposizione.

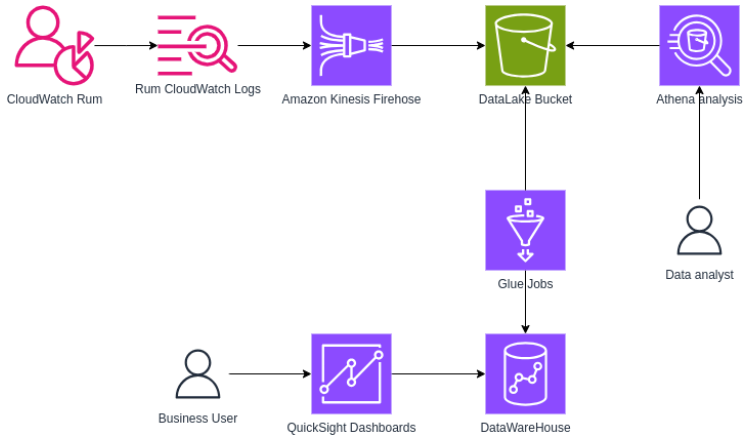
Questo ne è un estratto:


```
{
    "event_timestamp": 1718098736000,
    "event_type": "com.amazon.rum.performance_navigation_event",
    "event_id": "91ee89c6-044d-476c-8227-743c8f1b21d1",
    "event_version": "1.0.0",
    "log_stream": "2024-06-11T02",
    "application_id": "48e6845a-7799-4e76-90f6-a3602dac887a",
    "application_version": "1.0.0",
    "metadata": {
        "version": "1.0.0",
        "browserLanguage": "en-US",
        "browserName": "Edge",
        "browserVersion": "125.0.0.0",
        "osName": "Linux",
        "osVersion": "x86_64",
        "deviceType": "desktop",
        "platformType": "web",
        "pageId": "/",
        "interaction": 0,
        "title": "test",
        "domain": "blog.besharp.it",
```

```
        "aws:client": "arw-script",
        "aws:clientVersion": "1.16.1",
        "countryCode": "IT",
        "subdivisionCode": "PV"
    },
    "user_details": {
        "userId": "ac3f587b-3887-4661-8200-52d8ce5768f7",
        "sessionId": "13df94e4-7076-44a2-8b49-5dbb49e55e59"
    },
    "event_details": {
        "version": "1.0.0",
        "initiatorType": "navigation",
        "navigationType": "navigate",
        "startTime": 0,
        "unloadEventStart": 0,
        "promptForUnload": 0,
        "redirectCount": 0,
        "redirectStart": 0,
        "redirectTime": 0,
        "workerStart": 0,
        "workerTime": 0,
        "fetchStart": 4.7999999998137355,
        "domainLookupStart": 34,
        "dns": 0,
        "nextHopProtocol": "h2",
        "connectStart": 34,
        "connect": 15.799999999813735,
        "secureConnectionStart": 37.90000000037253,
        "tlsTime": 11.899999999441206,
        "requestStart": 50.10000000055879,
        "timeToFirstByte": 151.8999999994412,
        "responseStart": 202,
        "responseTime": 1.1000000005587935,
        "domInteractive": 276.20000000018626,
        "domContentLoadedEventStart": 276.20000000018626,
        "domContentLoaded": 0,
        "domComplete": 452.1000000005588,
        "domProcessingTime": 249,
        "loadEventStart": 452.1000000005588,
        "loadEventTime": 0.09999999962747097,
```

```
        "duration": 452.20000000018626,
        "headerSize": 300,
        "transferSize": 944,
        "compressionRatio": 2.0683229813664594,
        "navigationTimingLevel": 2
    }
}
```

Sapendo dove si trovano i dati, diventa facile esportarli ed integrarli in un Data Lake. Ad esempio, i log di CloudWatch possono essere inviati Kinesis Firehose e scritti quindi in un bucket S3.



Una volta esportati i dati, è possibile raffinarli, estrarre le informazioni e integrarle negli strumenti di analisi interni. A questo punto le possibilità sono infinite.

Ad esempio, si potrebbe pensare di correlare le vendite di un e-commerce con la posizione dell'utente, il tempo di risposta della pagina o i dati di navigazione della sessione.

Potremmo scoprire risultati inaspettati, come ad esempio che le condizioni meteorologiche di una particolare area influenzano le vendite. Per aggiungere altre informazioni ci viene in aiuto la possibilità di generare eventi personalizzati in RUM, che possono fungere da segnaposto e facilitare il lavoro di correlazione dei comportamenti.

Con tutte queste informazioni è possibile capire meglio il comportamento degli utenti, il posizionamento SEO, semplificare i processi interni o, semplicemente, avere un sistema di monitoraggio migliore per sapere chi e quando allertare in caso di anomalie e outlier nei dati. Senza contare la possibilità di creare dashboard per i clienti interni ed esterni; a chi non piace una bella dashboard :)

Una volta arricchito il Data Lake esistente, anche con queste informazioni originariamente destinate al monitoraggio è possibile realmente prendere decisioni accurate basate sui dati.

Questa soluzione non mira a sostituire Google Analytics, ma a facilitare l'integrazione e l'analisi dei dati: sappiamo che è possibile esportare i dati da GA-4 a BigQuery, ma se esiste già un DataWareHouse basato su AWS, l'integrazione può richiedere tempo. Il nostro sforzo è mirato infatti a supportare le decisioni, e non a mantenere integrazioni tra diversi sistemi.

Per Concludere

Scegliere la fonte di dati giusta invece di lavorare sulle integrazioni può essere vantaggioso e a volte basta scavare nei servizi per scoprire che le informazioni che desideriamo sono già presenti (spesso molte di più di quelle di cui avremmo bisogno!).

Se già sfruttate CloudWatch RUM, potete facilmente sfruttare l'integrazione con altri servizi AWS con poco sforzo ed utilizzando solo servizi gestiti, anche se, a prima vista, può sembrare complesso.

Avete mai pensato a come ottenere dati da fonti... insolite?

Dopo aver conosciuto questo tipo di integrazione vi sono venute in mente correlazioni inusuali?

Fatecelo sapere nei commenti!

About Proud2beCloud

Proud2beCloud è il blog di beSharp, APN Premier Consulting Partner italiano esperto nella progettazione, implementazione e gestione di infrastrutture Cloud complesse e servizi AWS avanzati. Prima di essere scrittori, siamo Solutions Architect che, dal 2007, lavorano quotidianamente con i servizi AWS. Siamo innovatori alla costante ricerca della soluzione più all'avanguardia per noi e per i nostri clienti. Su Proud2beCloud condividiamo regolarmente i nostri migliori spunti con chi come noi, per lavoro o per passione, lavora con il Cloud di AWS. Partecipa alla discussione!

Damiano Giorgi

Ex sistemista on-prem, pigro e incline all'automazione di task noiosi. Alla ricerca costante di novità tecnologiche e quindi passato al cloud per trovare nuovi stimoli.L'unico hardware a cui mi

dedico ora è quello del mio basso; se non mi trovate in ufficio o in sala prove provate al pub o in qualche aeroporto!
