EBook Gratuito

APPRENDIMENTO bigdata

Free unaffiliated eBook created from
Stack Overflow contributors.

#bigdata

Sommario

Di ... 1

Capitolo 1: Iniziare con bigdata ... 2

Osservazioni ... 2

Uno sviluppatore potrebbe essere interessato alle capacità di elaborazione dei big data in ... 3

Examples ... 3

Che cosa sono i Big Data? ... 3

Esempio di Big Data ... 4

Cosa succede sotto i big data? ... 5

Capitolo 2: Iniziare con Big Data / Hadoop Security ... 7

introduzione ... 7

Osservazioni ... 7

Examples ... 8

Esempio per ACL ... 8

Titoli di coda ... 9

Di

Capitolo 1: Iniziare con bigdata

Osservazioni

Questa sezione fornisce una panoramica di cosa sono i bigdata e perché uno sviluppatore potrebbe volerlo utilizzare.

I big data sono i dati caratterizzati dalle 4 V. Questi sono Volume, Velocity, Variety e Veracity.

1. Volume - Quando la quantità di dati è in un volume enorme come Terabyte o Petabyte. Come afferma un rapporto, abbiamo generato il 90% di dati mondiali negli ultimi 2 o 3 anni.
2. Velocità: la velocità con cui i dati fluiscono nel sistema. Ad esempio, milioni di utenti che caricano i loro contenuti sui siti di social networking allo stesso tempo genera dati dell'ordine di un intervallo di Terabyte / sec.
3. Varietà: diversi tipi di dati in base alla loro natura. Può essere strutturato (con la maggior parte dei vecchi RDBMS), semi-strutturato (e-mail, XML ecc.) E non strutturato (video, audio, dati dei sensori ecc.).
4. Veracità: è il mezzo con cui otteniamo una visione significativa dei nostri dati disponibili. Questo può essere considerato l'aspetto più importante dei dati poiché la maggior parte delle decisioni aziendali dipende dall'utilità dei dati.

La piattaforma più generale utilizzata per archiviare ed elaborare i big data è il **framework Hadoop.** Consiste di 2 cose:

1. Hadoop Distributed File System (HDFS): i dati vengono archiviati su Hadoop Distributed File System (HDFS), che in realtà è un cluster di hardware di base diverso dal primitivo sistema di archiviazione sui server. I dati risiedono su HDFS e possono essere elaborati per ricavare informazioni utilizzando vari strumenti e quadri.
2. MapReduce (MR) - Questo è il framework di elaborazione predefinito per Hadoop. MapReduce (è una parte di Apache Hadoop)

Con un avanzamento in Hadoop, i nuovi strumenti di elaborazione hanno iniziato a emergere nella comunità Hadoop. Pochi degli strumenti / framework più popolari:

1. Apache Spark
2. Apache Storm
3. Apache Flink

   E tanti altri..

Pochi dei meccanismi di memorizzazione diversi dal semplice HDFS:

1. Alveare
2. HBase
3. cassandra

E tanti altri..

Uno sviluppatore potrebbe essere interessato alle capacità di elaborazione dei big data in modo che possa rivelarsi una grande differenza nel modo in cui guardiamo i nostri dati. In un universo parallelo, possiamo anche chiamare i big data come Rich-untamed-Data. Dobbiamo domare questi enormi dati. Con i grandi dati potremmo essere in grado di elaborare il potenziale nascosto di dati già esistenti.

Un esempio migliore può essere citato nel comportamento dei clic dei clienti sui siti web di shopping in cui le loro opinioni, i clic e la quantità di tempo speso su quel sito Web, indica al rivenditore online di procurarsi il prodotto e inviare raccomandazioni in base al comportamento dell'utente.

Examples

Che cosa sono i Big Data?

I Big Data, nella sua forma più elementare, possono essere descritti come il termine generico metricizzato da diversi aspetti dei dati. Questi diversi aspetti sono

> Volume (quantità enorme di dati), velocità (maggiori velocità del flusso di dati), varietà (dati strutturati, non strutturati e semi-strutturati) e veracità (prendere decisioni giuste in base ai dati).

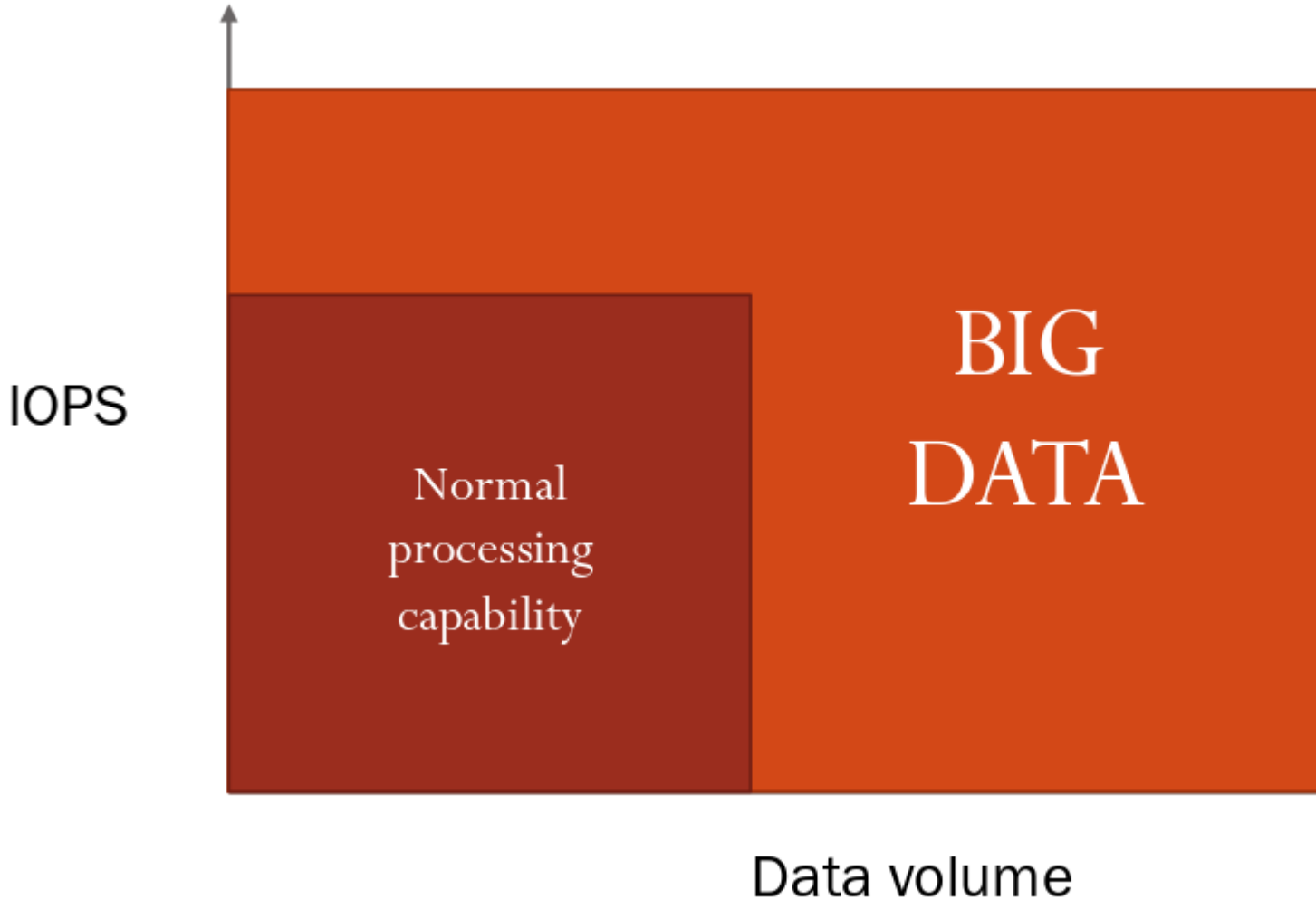
Queste metriche erano difficili da gestire con i database relazionali della vecchiaia. Nacque la necessità di un nuovo sistema e l'elaborazione dei Big Data venne in soccorso. Mentre molte persone hanno una comprensione diversa su cosa siano i Big Data, ecco alcune delle definizioni di Big Data fornite dai leader del settore Data nel settore:

definizioni:

- "I big data superano la portata degli ambienti hardware e degli strumenti software comunemente usati per catturarli, gestirli e processarli in un tempo intercambiabile tollerabile per la sua popolazione di utenti." (Articolo di Teradata Magazine, 2011)
- "I Big Data si riferiscono a set di dati la cui dimensione è al di là della capacità dei tipici strumenti software di database di acquisire, archiviare, gestire e analizzare." (The McKinsey Global Institute, 2012)
- "I big data sono una raccolta di set di dati così grandi e complessi che diventa difficile da elaborare utilizzando strumenti di gestione di database disponibili". (Wikipedia, 2014)
- "I Big Data sono risorse informative ad alto volume, ad alta velocità e / o ad alta varietà che richiedono nuove forme di elaborazione per consentire un processo decisionale potenziato,

recupero delle informazioni e ottimizzazione dei processi" (Gartner, 2012)

Quando i dati diventano "grandi"?




```
IOPS:Input/Output Operations Per Second
```

Esempio di Big Data

Big data è un termine per insiemi di dati così grandi o complessi che le tradizionali applicazioni di elaborazione dati sono inadeguate per gestirli. Le sfide includono analisi, acquisizione, gestione dei dati, ricerca, condivisione, archiviazione, trasferimento, visualizzazione, interrogazione, aggiornamento e riservatezza delle informazioni.

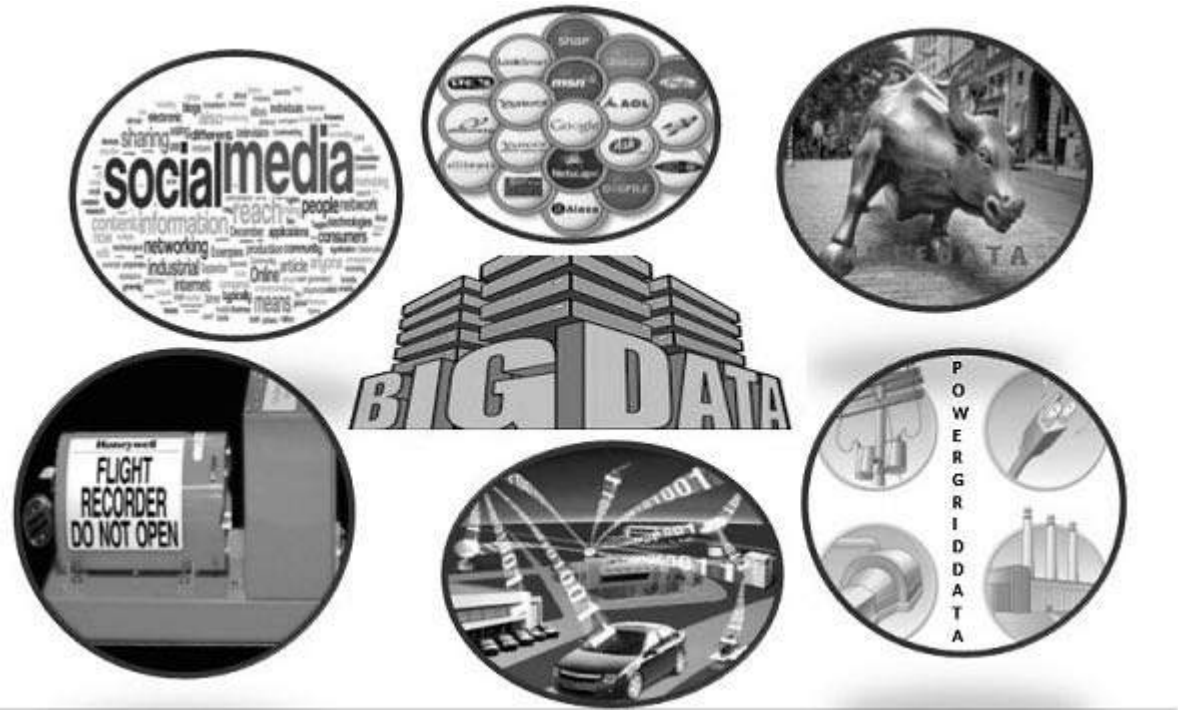
Un esempio generale di big data:

Dati raccolti dal social network facebook. Facebook raccoglie centinaia di terabyte (TB) di dati ogni giorno. I dati raccolti possono essere immagini, video, post, aggiornamenti, ecc. I dati variano da strutturati a non strutturati. Un like, share o reazione possono essere dati strutturati poiché ne conosciamo chiaramente la struttura. Mentre gli aggiornamenti o i post sono dati non strutturati che non seguono esattamente una struttura. Tutti questi dati formano insieme BigData!

Cosa succede sotto i big data?

I big data riguardano i dati prodotti da diversi dispositivi e applicazioni. Di seguito sono riportati alcuni dei campi che rientrano nell'ombrello dei Big Data.

- Black Box Data: è un componente di elicotteri, aerei e jet, ecc. Cattura le voci dell'equipaggio di condotta, le registrazioni di microfoni e auricolari e le informazioni sulle prestazioni dell'aeromobile.
- Dati sui social media: i social media come Facebook e Twitter contengono le informazioni e le opinioni pubblicate da milioni di persone in tutto il mondo.
- Dati di borsa: i dati di borsa contengono informazioni sulle decisioni di acquisto e vendita di una quota di diverse società effettuate dai clienti.
- Dati Power Grid: i dati della griglia di alimentazione contengono informazioni consumate da un nodo particolare rispetto a una stazione base.
- Dati di trasporto: i dati di trasporto includono il modello, la capacità, la distanza e la disponibilità di un veicolo.
- Dati del motore di ricerca: i motori di ricerca recuperano molti dati da diversi database.
- Dati sensore: dati provenienti da diversi dispositivi che lavorano su sensori, ad esempio: dati meteorologici (meteo e climatici), dati sismici (terremoti), dati oceanici (maree, tsunami ecc.).



Pertanto, i Big Data includono enormi volumi, alta velocità ed estensibile varietà di dati. I dati in esso saranno di tre tipi.

```
1. Structured data : Mostly data from Relational Databases.
```

```
2. Semi Structured data : XML data, email data.

3. Unstructured data : Word, PDF, Text, Media Logs.
```

Leggi Iniziare con bigdata online: https://riptutorial.com/it/bigdata/topic/6890/iniziare-con-bigdata

Capitolo 2: Iniziare con Big Data / Hadoop Security

introduzione

Siamo in grado di proteggere i dati in Hadoop utilizzando diversi metodi. Ogni metodo ha i suoi vantaggi. Possiamo anche combinare più di un metodo per ottenere risultati migliori. Questo argomento copre i vantaggi e le limitazioni di ciascun metodo

Osservazioni

1. Kerberos è un protocollo di autenticazione di rete:

un. Vantaggio: autentica gli utenti a livello di entrata.

b. Limitazione: Kerberos impedisce agli utenti non autorizzati di accedere all'ambiente. Ma dopo il login, non fornirà autenticazioni di livello dettagliate come tabella, colonna, cartella, livello di file, ecc

2. Apache Sentry è un sistema per il rafforzamento del ruolo di base

un. Vantaggio: autenticazioni a livello di applicazione come Hive, Impala, Solr, ecc. Può controllare l'accesso su DB, tabella, livello di colonna per un particolare utente / gruppo.

b. Limitazione: non può controllare le cartelle HDFS che sono sottolineate dietro applicazioni come Hive, Impala, ecc. Ex: Hive table prod.table1 memorizzato in /user/hive/warehouse/prod.db/table1. L'impostazione del ruolo di sentinella in Hue può controllare solo l'accesso a tabelle / colonne in Hue, ma è possibile che l'utente possa accedere alle cartelle direttamente in HDFS

c. Limitazione: le cartelle HDFS che non sono correlate a Hive, Impala, ecc. Non saranno controllate

3. Un elenco di controllo di accesso (ACL) è un elenco di voci di controllo di accesso (ACE). Ogni ACE in un ACL identifica un trustee e specifica i diritti di accesso consentiti, negati o controllati per tale trustee

un. Vantaggio: l' accesso al livello della cartella è possibile tramite gli utenti che utilizzano

4. La crittografia HDFS implementa la crittografia trasparente, end-to-end dei dati letti da e scritti su HDFS

un. Vantaggio: crittografare i dati fornirà ulteriore sicurezza di livello. In generale, la crittografia dei dati è richiesta da numerose entità governative, finanziarie e regolatorie

Examples

Esempio per ACL

```
hadoop fs -setfacl
```

Leggi Iniziare con Big Data / Hadoop Security online:
https://riptutorial.com/it/bigdata/topic/9869/iniziare-con-big-data---hadoop-security

Titoli di coda

S. No	Capitoli	Contributors
1	Iniziare con bigdata	Ani Menon, Community, Mr. P, NeoWelkin, Sayali Sonawane
2	Iniziare con Big Data / Hadoop Security	saranvisa