Kostenloses eBook

LERNEN bigdata

Free unaffiliated eBook created from
Stack Overflow contributors.

#bigdata

Inhaltsverzeichnis

Über

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: bigdata

It is an unofficial and free bigdata ebook created for educational purposes. All the content is extracted from Stack Overflow Documentation, which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official bigdata.

Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to info@zzzprojects.com

Kapitel 1: Erste Schritte mit Bigdata

Bemerkungen

In diesem Abschnitt erhalten Sie einen Überblick über Bigdata und warum ein Entwickler sie verwenden möchte.

Big Data sind die Daten, die durch die 4 V charakterisiert werden. Dies sind Volumen, Geschwindigkeit, Abwechslung und Wahrhaftigkeit.

1. Volume (Volumen) - Wenn sich die Datenmenge in einem großen Umfang wie Terabytes oder Petabytes befindet. Wie aus einem Bericht hervorgeht, haben wir in den letzten 2 oder 3 Jahren weltweit 90% Daten generiert.
2. Geschwindigkeit - Die Geschwindigkeit, mit der Daten im System fließen. Zum Beispiel erzeugen Millionen Benutzer, die ihre Inhalte gleichzeitig auf Social-Networking-Sites hochladen, Daten in der Größenordnung von Terabytes / Sek.
3. Vielfalt - Verschiedene Arten von Daten basierend auf ihrer Art. Es kann strukturiert sein (was die meisten alten RDBMS betreffen), halbstrukturiert (E-Mail, XML usw.) und unstrukturiert (Videos, Audios, Sensordaten usw.).
4. Wahrhaftigkeit - Es ist das Mittel, mit dem wir einen sinnvollen Einblick in unsere verfügbaren Daten erhalten. Dies kann als der wichtigste Aspekt von Daten betrachtet werden, da die meisten Geschäftsentscheidungen von der Nützlichkeit der Daten abhängen.

Die allgemeinste Plattform zum Speichern und Verarbeiten von Big Data ist das **Hadoop Framework.** Es besteht aus 2 Sachen:

1. Hadoop Distributed File System (HDFS) - Die Daten werden auf dem Hadoop Distributed File System (HDFS) gespeichert, bei dem es sich im Gegensatz zu der primitiven Art der Speicherung auf Servern um ein Cluster von Standardhardware handelt. Die Daten werden auf HDFS gespeichert und möglicherweise verarbeitet, um Erkenntnisse mithilfe verschiedener Tools abzuleiten und Rahmenbedingungen.
2. MapReduce (MR) - Dies ist das Standardverarbeitungsframework für Hadoop. MapReduce (ist ein Teil von Apache Hadoop)

Mit der Weiterentwicklung von Hadoop sind in der Hadoop-Community neue Verarbeitungs-Tools entstanden. Einige der beliebtesten Tools / Frameworks:

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

   Und viele mehr..

Einige Speichermechanismen außer normalem HDFS:

1. Bienenstock
2. HBase
3. Kassandra

Und viele mehr..

Ein Entwickler könnte an den Verarbeitungsmöglichkeiten von Big Data interessiert sein, so dass sich dies als ein wesentlicher Unterschied bei der Betrachtung unserer Daten herausstellen kann. In einem Paralleluniversum können wir Big Data auch als Rich-Unamed-Data bezeichnen. Wir müssen diese riesigen Daten zähmen. Mit Big Data können wir möglicherweise das verborgene Potenzial bereits vorhandener Daten verarbeiten.

Ein bestes Beispiel ist das Kundenklickverhalten auf den Shopping-Websites, in dem die Ansichten, Klicks und die auf dieser Website verbrachte Zeit den Online-Händler anweisen, ein Produkt zu beschaffen und Empfehlungen basierend auf dem Nutzerverhalten zu senden.

Examples

Was ist Big Data?

Big Data kann in seiner grundlegendsten Form als Oberbegriff bezeichnet werden, der anhand verschiedener Datenaspekte gemessen wird. Diese verschiedenen Aspekte sind

> Volumen (riesige Datenmenge), Geschwindigkeit (höhere Datenflussgeschwindigkeiten), Vielfalt (strukturierte, unstrukturierte und halbstrukturierte Daten) und Richtigkeit (richtige, auf Daten basierende Entscheidungen).

Diese Metriken waren nur schwer durch relationale Datenbanken im Alter zu bewältigen. Es bestand ein Bedarf an einem neuen System, und die Verarbeitung von Big Data kam zur Rettung. Während viele Menschen ein unterschiedliches Verständnis für Big Data haben, sind hier einige Definitionen von Big Data, die von Branchenführern im Datensektor vorgegeben werden:

Definitionen:

- "Große Datenmengen überschreiten die Reichweite von häufig verwendeten Hardwareumgebungen und Softwaretools, um diese innerhalb einer für die Benutzer akzeptablen Zeit zu erfassen, zu verwalten und zu verarbeiten." (Artikel des Teradata Magazine, 2011)
- "Big Data" bezieht sich auf Datensätze, deren Größe nicht die Fähigkeit typischer

Datenbanksoftware-Tools zum Erfassen, Speichern, Verwalten und Analysieren hat. "(The McKinsey Global Institute, 2012)

- „Big Data ist eine Sammlung von Datensätzen, die so groß und komplex sind, dass die Verarbeitung mit handelsüblichen Datenbankverwaltungstools schwierig wird.“ (Wikipedia, 2014)
- "Big Data sind Datenbestände mit hohem Volumen, hoher Geschwindigkeit und / oder vielfältigen Informationen, für die neue Verarbeitungsformen erforderlich sind, um eine bessere Entscheidungsfindung, Wiederherstellung von Informationen und Prozessoptimierung zu ermöglichen" (Gartner, 2012)

Wenn Daten zu „groß“ werden?


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

Big Data-Beispiel

Big Data ist ein Begriff für Datensätze, die so groß oder komplex sind, dass herkömmliche Datenverarbeitungsanwendungen nicht ausreichen, um damit umzugehen. Zu den Herausforderungen gehören Analyse, Erfassung, Datenaufbereitung, Suche, gemeinsame

Nutzung, Speicherung, Übertragung, Visualisierung, Abfragen, Aktualisieren und Datenschutz.

Ein allgemeines Beispiel für Big Data:

Daten, die von der Social-Networking-Site Facebook gesammelt wurden. Facebook sammelt täglich Hunderte von Terabyte (TB) Daten. Die erfassten Daten können Bilder, Videos, Beiträge, Updates usw. sein. Die Daten variieren von strukturiert bis unstrukturiert. Eine Like-, Share- oder Reaction-Struktur kann strukturierte Daten enthalten, da wir die Struktur klar kennen. Updates oder Posts sind dagegen unstrukturierte Daten, die nicht genau einer Struktur folgen. All diese Daten bilden zusammen BigData!

Was kommt unter Big Data?

Bei Big Data handelt es sich um Daten, die von verschiedenen Geräten und Anwendungen erzeugt werden. Im Folgenden sind einige Felder aufgeführt, die unter das Dach von Big Data fallen.

- Black Box Data: Es ist eine Komponente von Hubschraubern, Flugzeugen und Jets usw. Es erfasst Stimmen der Flugbesatzung, Aufzeichnungen von Mikrofonen und Ohrhörern sowie die Leistungsinformationen des Flugzeugs.
- Social Media-Daten: Social Media wie Facebook und Twitter enthalten Informationen und Ansichten, die von Millionen Menschen auf der ganzen Welt gepostet werden.
- Börsendaten: Die Börsendaten enthalten Informationen zu den von den Kunden getroffenen Anteilen verschiedener Unternehmen bezüglich der Kauf- und Verkaufsentscheidungen.
- Stromnetzdaten: Die Stromnetzdaten enthalten Informationen, die von einem bestimmten Knoten in Bezug auf eine Basisstation verbraucht werden.
- Transportdaten: Die Transportdaten umfassen Modell, Kapazität, Entfernung und Verfügbarkeit eines Fahrzeugs.
- Suchmaschinendaten: Suchmaschinen rufen viele Daten aus verschiedenen Datenbanken ab.
- Sensordaten: Daten von verschiedenen an Sensoren arbeitenden Geräten, Beispiel: Wetterdaten (Wetter und Klima), seismische Daten (Erdbeben), ozeanische Daten (Gezeiten, Tsunami usw.).

Daher umfasst Big Data ein enormes Volumen, hohe Geschwindigkeit und erweiterbare Datenvielfalt. Die Daten werden drei Arten haben.

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

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

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

Erste Schritte mit Bigdata online lesen: https://riptutorial.com/de/bigdata/topic/6890/erste-schritte-mit-bigdata

Kapitel 2: Erste Schritte mit Big Data / Hadoop Security

Einführung

Wir können die Daten in Hadoop mit verschiedenen Methoden sichern. Jede Methode hat ihre eigenen Vorteile. Wir können auch mehrere Methoden für ein besseres Ergebnis kombinieren. Dieses Thema behandelt die Vorteile und Einschränkungen jeder Methode

Bemerkungen

1. Kerberos ist ein Netzwerkauthentifizierungsprotokoll:

ein. Vorteil: Authentifizieren Sie Benutzer auf der Einstiegsebene.

b. Einschränkung: Kerberos verhindert den Zugriff unberechtigter Benutzer auf die Umgebung. Nach der Anmeldung werden jedoch keine detaillierten Authentifizierungen wie Tabellen, Spalten, Ordner, Dateiebene usw. bereitgestellt

2. Apache Sentry ist ein System zur Durchsetzung feinkörniger Rollenbasis

ein. Vorteil: Authentifizierungen auf Anwendungsebene wie Hive, Impala, Solr usw. Sie können den Zugriff auf DB-, Tabellen- und Spaltenebene für einen bestimmten Benutzer / eine bestimmte Gruppe steuern.

b. Einschränkung: Die HDFS-Ordner, die hinter Anwendungen wie Hive, Impala usw. unterstrichen werden, können nicht gesteuert werden. Das Sentry-Rollen-Setup in Hue kann nur den Zugriff auf Tabellen / Spalten in Hue steuern. Es ist jedoch möglich, dass der Benutzer den direkten Zugriff auf Ordner in HDFS verwaltet

c. Einschränkung: HDFS-Ordner, die sich nicht auf Hive, Impala usw. beziehen, werden nicht gesteuert

3. Eine Zugriffssteuerungsliste (ACL) ist eine Liste von Zugriffssteuerungseinträgen (ACE). Jeder ACE in einer ACL identifiziert einen Trustee und gibt die für diesen Trustee zugelassenen, abgelehnten oder geprüften Zugriffsrechte an

ein. Vorteil: Zugriff auf Ordnerebene ist für Benutzer möglich

4. HDFS-Verschlüsselung implementiert eine transparente Ende-zu-Ende-Verschlüsselung von Daten, die aus HDFS gelesen und in diese geschrieben werden

ein. Vorteil: Durch die Verschlüsselung der Daten wird eine zusätzliche Sicherheitsstufe geschaffen. Im Allgemeinen ist die Datenverschlüsselung von verschiedenen Regierungs-, Finanz- und Aufsichtsbehörden erforderlich

Examples

Beispiel für ACL

```
hadoop fs -setfacl
```

Erste Schritte mit Big Data / Hadoop Security online lesen:
https://riptutorial.com/de/bigdata/topic/9869/erste-schritte-mit-big-data---hadoop-security

Credits

S. No	Kapitel	Contributors
1	Erste Schritte mit Bigdata	Ani Menon, Community, Mr. P, NeoWelkin, Sayali Sonawane
2	Erste Schritte mit Big Data / Hadoop Security	saranvisa