Бесплатная электронная книга

УЧУСЬ bigdata

Free unaffiliated eBook created from
Stack Overflow contributors.

#bigdata

Около

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

глава 1: Начало работы с bigdata

замечания

В этом разделе представлен обзор того, что такое bigdata, и почему разработчик может захотеть его использовать.

Большие данные - это данные, характеризуемые 4 В. Это объем, скорость, разнообразие и достоверность.

1. Объем - когда объем данных находится в огромном объеме, таком как терабайты или петабайты. Как говорится в отчете, за последние 2 или 3 года мы получили 90% -ные данные мира.
2. Velocity - скорость, с которой данные текут в системе. Например, миллионы пользователей, загружающих свой контент на сайтах социальных сетей, одновременно генерируют такие высокие данные, как диапазон терабайт / сек.
3. Разнообразие - различные типы данных, основанные на его характере. Он может быть структурирован (к которому относится большинство старых RDBMS), полуструктурированных (электронная почта, XML и т. Д.) И неструктурированных (видео, аудио, данные датчиков и т. Д.).
4. Veracity - это средство, с помощью которого мы получаем осмысленное представление о наших доступных данных. Это можно рассматривать как наиболее важный аспект данных, поскольку большинство бизнес-решений зависит от полезности данных.

Наиболее общей платформой, используемой для хранения и обработки больших данных, является **Hadoop Framework.** Он состоит из двух вещей:

1. Распределенная файловая система Hadoop (HDFS). Данные хранятся в распределенной файловой системе Hadoop (HDFS), которая на самом деле является кластером товарного оборудования, в отличие от примитивного способа хранения на серверах. Данные хранятся на HDFS и, возможно, обрабатываются для получения информации с использованием различных инструментов и рамки.
2. MapReduce (MR) - это платформа обработки по умолчанию для Hadoop. MapReduce (входит в состав Apache Hadoop)

С продвижением в Hadoop в Hadoop Community появились новые инструменты обработки. Многие из самых популярных инструментов / фреймворков:

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

И многое другое ..

Немногие из механизмов хранения, отличных от простой HDFS:

1. улей
2. HBase
3. Cassandra

И многое другое ..

Разработчик может быть заинтересован в возможностях обработки больших данных, чтобы он мог оказаться существенным различием в том, как мы смотрим на наши данные. В параллельной юниверсе мы также можем назвать большие данные как Rich-untamed-Data. Мы должны приручить эти огромные данные. С большими данными мы могли бы обработать скрытый потенциал уже существующих данных.

Лучший пример можно привести в поведении клиента по торговым сайтам, где их взгляды, клики и количество времени, проведенного на этом веб-сайте, сообщают онлайн-продавцу о приобретении продукта и отправке рекомендаций на основе поведения пользователя.

Examples

Что такое большие данные?

Большие данные в его самой базовой форме могут быть описаны как зонтичный термин, метризуемый различными аспектами данных. Эти различные аспекты

> Объем (огромное количество данных), скорость (большие скорости потока данных), разновидность (структурированные, неструктурированные и полуструктурированные данные) и достоверность (принятие правильных решений на основе данных).

Эти метрики трудно было позаботиться о реляционных базах данных по старости. Появилась необходимость в новой системе, и обработка Больших Данных оказалась на спасение. В то время как многие люди имеют разное понимание того, что такое «большие данные», вот несколько определений крупных данных, данных лидерами отрасли в секторе данных:

Определения:

- «Большие данные превышают охват широко используемых аппаратных сред и программных средств для их захвата, управления и обработки в течение приемлемого времени для его пользовательской популяции» (статья «Журнал Teradata Magazine», 2011 г.)
- «Большие данные относятся к наборам данных, размер которых превосходит возможности типичных инструментов программного обеспечения базы данных для сбора, хранения, управления и анализа». (The McKinsey Global Institute, 2012)
- «Большие данные - это набор наборов данных, настолько больших и сложных, что становится сложно обрабатывать с помощью инструментов управления базами данных на основе данных» (Wikipedia, 2014)
- «Большие данные - это высокопроизводительные, высокоскоростные и / или высокоуровневые информационные активы, для которых требуются новые формы обработки, чтобы обеспечить расширенное принятие решений, восстановление информации и оптимизацию процессов» (Gartner, 2012)

Когда данные становятся «большими»?

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

Пример с большими данными

Большие данные - это термин для наборов данных, которые являются настолько большими или сложными, что традиционные приложения обработки данных неадекватны для решения этих проблем. Задачи включают анализ, захват, обработку данных, поиск, обмен, хранение, передачу, визуализацию, запрос, обновление и конфиденциальность информации.

Общий пример больших данных:

Данные, собранные сайтом социальной сети facebook. Каждый день Facebook собирает сотни терабайт (ТБ) данных. Собранные данные могут быть изображениями, видеороликами, сообщениями, обновлениями и т. Д. Данные варьируются от структурированных до неструктурированных. Подобная, доля или реакция, возможно, структурированные данные, поскольку мы четко знаем ее структуру. В то время как обновления или сообщения представляют собой неструктурированные данные, которые точно не соответствуют структуре. Все эти данные вместе образуют BigData!

Что входит в большие данные?

Большие данные включают данные, полученные различными устройствами и приложениями. Ниже приведены некоторые из полей, которые входят в сферу деятельности «Больших данных».

- Данные Black Box: это компонент вертолета, самолетов и самолетов и т. Д. Он захватывает голоса летного экипажа, записи микрофонов и наушников и информацию о производительности самолета.
- Социальные медиа-данные. Социальные медиа, такие как Facebook и Twitter, содержат информацию и мнения, публикуемые миллионами людей по всему миру.
- Данные фондовой биржи. Данные биржи содержат информацию о решениях «покупать» и «продавать», принимаемых на долю разных компаний, сделанных клиентами.
- Данные энергосети: данные энергосистемы содержат информацию, потребляемую конкретным узлом относительно базовой станции.
- Транспортные данные: данные о транспорте включают в себя модель, мощность, расстояние и доступность транспортного средства.
- Данные поисковой системы: поисковые системы извлекают множество данных из разных баз данных.

- Данные датчиков: данные с разных устройств, работающих с датчиками, например: метеорологические данные (погодные и климатические), данные сейсмического (землетрясения), данные океанических (приливы, цунами и т. Д.).

Таким образом, Big Data включает в себя огромный объем, высокую скорость и расширяемость разнообразных данных. Данные в нем будут трех типов.

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

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

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

Прочитайте Начало работы с bigdata онлайн: https://riptutorial.com/ru/bigdata/topic/6890/начало-работы-с-bigdata

глава 2: Начало работы с Big Data / Hadoop Security

Вступление

Мы можем защитить данные в Hadoop различными способами. Каждый метод имеет свои преимущества. Мы также можем комбинировать более одного метода для лучшего результата. В этом разделе рассматриваются преимущества и ограничения каждого метода

замечания

1. Kerberos - это протокол сетевой аутентификации:

а. Преимущество: Аутентификация пользователей на начальном уровне.

б. Ограничение: Kerberos предотвращает несанкционированный доступ пользователей к среде. Но после входа в систему он не будет предоставлять подробные проверки уровня, такие как таблица, столбец, папка, уровень файла и т. Д.

2. Apache Sentry - это система для принудительного выполнения мелкозернистых ролей

а. Преимущество: аутентификации на уровне приложений, такие как Hive, Impala, Solr и т. Д. Он может контролировать доступ к DB, таблице, уровню столбца для конкретного пользователя / группы.

б. Ограничение: он не может контролировать папки HDFS, которые подчеркиваются за такими приложениями, как Hive, Impala и т. Д. Ex: таблица hive prod.table1, хранящаяся в / user/hive/warehouse/prod.db/table1. Настройка часовой роли в Hue может контролировать только доступ к таблице / столбцу в Hue, но возможно, что пользователь может управлять доступом к папкам непосредственно в HDFS

с. Ограничение: папки HDFS, которые не связаны с Hive, Impala и т. Д., Не будут контролироваться

3. Список контроля доступа (ACL) - это список записей управления доступом (ACE). Каждый ACE в ACL идентифицирует доверительного управляющего и указывает права доступа, запрещенные или проверенные для этого доверительного управляющего

а. Преимущество: доступ к папке возможен пользователями, использующими

4. HDFS Encryption реализует прозрачное сквозное шифрование данных, считываемых и записываемых в HDFS

a. Преимущество: Шифрование данных обеспечит дополнительную безопасность уровня. В целом шифрование данных требуется нескольким государственным, финансовым и регулирующим органам

Examples

Пример для ACL

```
hadoop fs -setfacl
```

Прочитайте Начало работы с Big Data / Hadoop Security онлайн:
https://riptutorial.com/ru/bigdata/topic/9869/начало-работы-с-big-data---hadoop-security

кредиты

S. No	Главы	Contributors
1	Начало работы с bigdata	Ani Menon, Community, Mr. P, NeoWelkin, Sayali Sonawane
2	Начало работы с Big Data / Hadoop Security	saranvisa