FREE eBook

LEARNING bigdata

Free unaffiliated eBook created from
Stack Overflow contributors.

#bigdata

Table of Contents

About

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

Chapter 1: Getting started with bigdata

Remarks

This section provides an overview of what bigdata is, and why a developer might want to use it.

Big data is the data characterized by the 4 V's. These are Volume, Velocity, Variety and Veracity.

1. Volume - When the amount of data is in huge volume like Terabytes or Petabytes. As a report says, we have generated world's 90 % data over the last 2 or 3 years.
2. Velocity - The speed at which data is flowing in the system. For instance, millions of users uploading their content on Social networking Sites at the same time generates data as high as in the range of Terabytes/sec.
3. Variety - Different types of data based on its nature. It can be Structured(which most of the old RDBMS deals with), Semi-Structured(email, XML etc.) and Unstructured(Videos, Audios, Sensor Data etc.).
4. Veracity - It is the means with which we get a meaningful insight in our available data. This can be considered as the most important aspect of data as most of the business decision depends on the usefulness of data.

The most general platform used to store and process big data is the **Hadoop Framework.** It consists of 2 things:

1. Hadoop Distributed File System(HDFS) - Data is stored on Hadoop Distributed File System(HDFS) which is actually a cluster of commodity hardware unlike the primitive way of storing on servers.The data resides on HDFS and maybe processed to derive insights using various tools and frameworks.
2. MapReduce(MR) - This is the default processing framework for Hadoop.MapReduce (is a part of Apache Hadoop)

With an advancement in Hadoop , new processing tools started emerging in the Hadoop Community.Few of the most popular tools/frameworks:

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

   And many more..

Few of the storage mechanisms other than plain HDFS:

1. Hive
2. HBase
3. Cassandra

And many more..

A developer might be interested in the processing capabilities of big data so that it can prove to be a major difference in how we look at our data. In a parallel universe, we can also call big data as Rich-untamed-Data. We have to tame this huge data.With big data we might be able to process the hidden potential of already existing data.

A best example can be cited in the customer click behavior over the shopping websites wherein their views, clicks and the amount of time spent on that website, tells the online retailer to procure product and send recommendations based on user behavior.

Examples

What is Big Data?

Big Data, in its most basic form, can be described as the umbrella term metricized by different aspects of data. These different aspects are

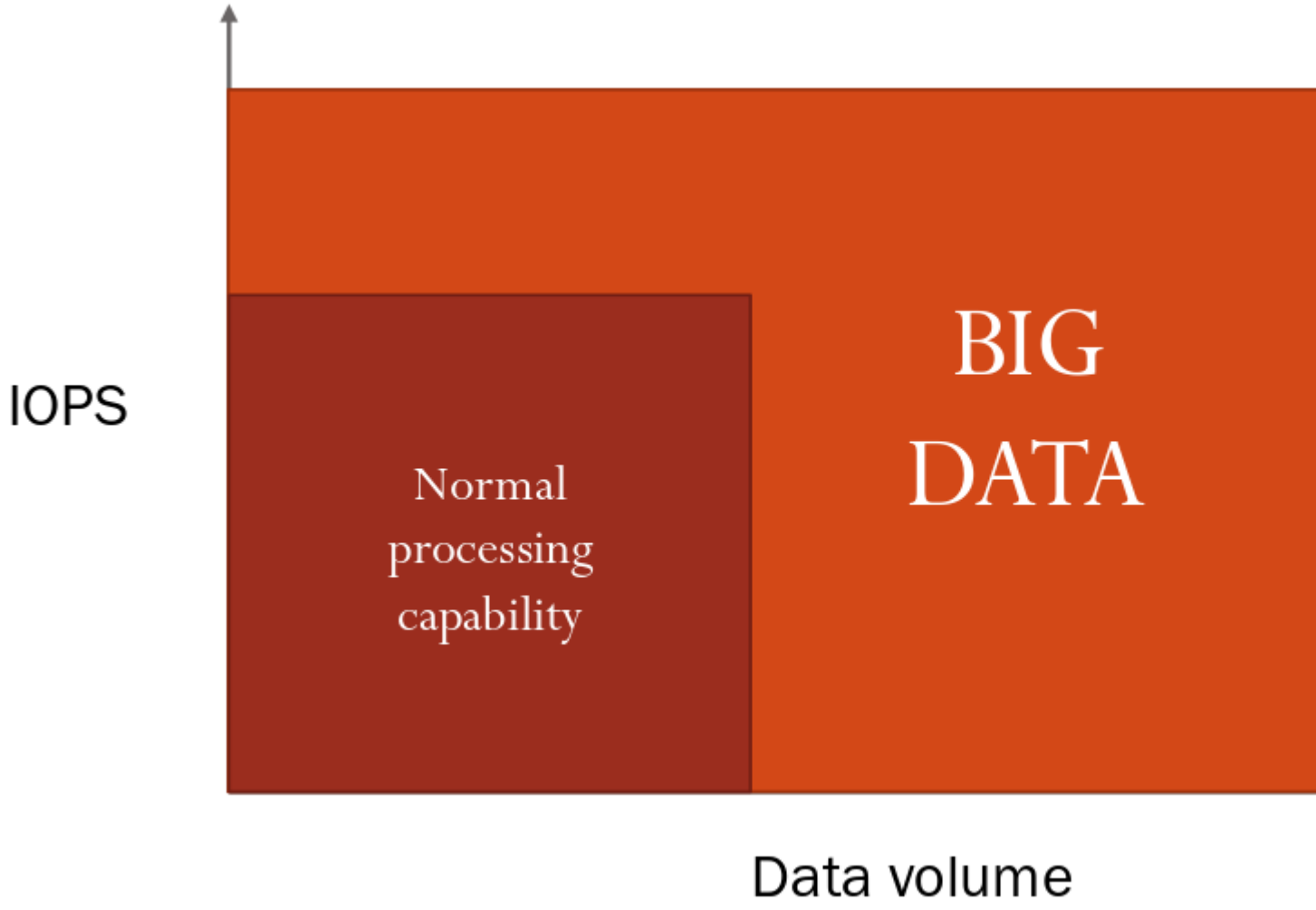
> Volume(Huge quantity of Data), Velocity(Greater dataflow speeds), Variety(Structured, Unstructured and Semi-structured Data) and Veracity(Making right decisions based on data).

These metrics were hard to be taken care of by old age relational databases. A need for a new system arose and Big Data processing came to the rescue. While many people have different understanding on what Big Data is, here are few of the definitions of Big Data given by industry leaders in Data sector:

Definitions:

- “Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it with in a tolerable elapsed time for its user population.” (Teradata Magazine article, 2011)
- “Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.” (The McKinsey Global Institute, 2012)
- “Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools.” (Wikipedia, 2014)
- "Big Data are high--volume,high--velocity,and/or high--variety information assets that require new forms of processing to enable enhanced decision making,insight recovery and process optimization" (Gartner,2012)

When data become “Big”?




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

Big Data example

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

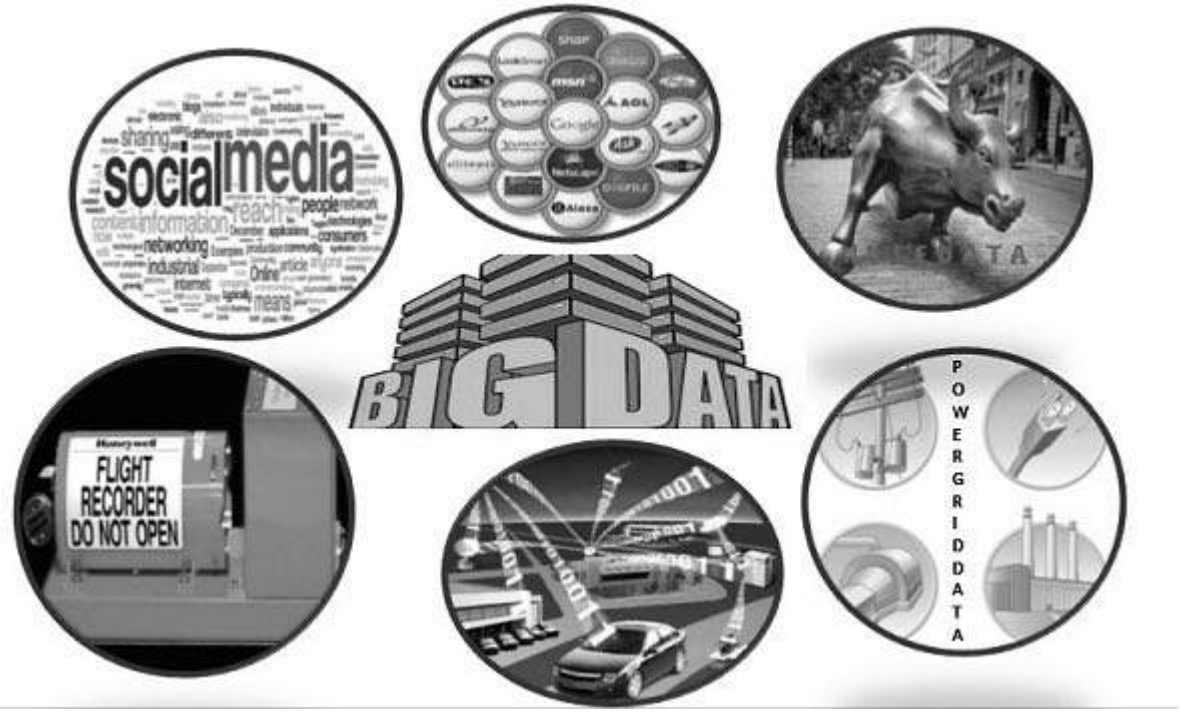
A general example of big data:

Data collected by social networking site facebook. Facebook collects hundreds of terabytes(TB) of data every day. Data collected may be images, videos, posts, updates, etc. The data varies from structured to unstructured. A like, share or reaction maybe structured data as we clearly know the structure of it. Whereas updates or posts are unstructured data which don't exactly follow a structure. All this data together forms BigData!

What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

- Black Box Data : It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- Social Media Data : Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- Stock Exchange Data : The stock exchange data holds information about the ‘buy’ and ‘sell’ decisions made on a share of different companies made by the customers.
- Power Grid Data : The power grid data holds information consumed by a particular node with respect to a base station.
- Transport Data : Transport data includes model, capacity, distance and availability of a vehicle.
- Search Engine Data : Search engines retrieve lots of data from different databases.
- Sensor Data : Data from different devices working on sensors, example: Meteorological (weather and climate) data, Seismic (earthquake) data, Oceanic (Tides, Tsunami etc.) data.



Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

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

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

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

Read Getting started with bigdata online: https://riptutorial.com/bigdata/topic/6890/getting-started-with-bigdata

Chapter 2: Getting started with Big Data/Hadoop Security

Introduction

We can secure the data in Hadoop using different methods. Each method has its own advantages. We can also combine more than one method for better result. This topic covers the advantages & limitations of each method

Remarks

1. Kerberos is a network authentication protocol:

a. Advantage: Authenticate users at the entry level.

b. Limitation: Kerberos prevents unauthorized user access to the environment. But after login, it will not provide detailed level authentications like table, column, folder, file level, etc

2. Apache Sentry is a system for enforcing fine grained role bas

a. Advantage: Application level authentications like Hive, Impala, Solr, etc. It can control access on DB, table, column level for a particular user/group.

b. Limitation: It cannot control the HDFS folders which are underlined behind applications like Hive, Impala, etc. Ex: Hive table prod.table1 stored in /user/hive/warehouse/prod.db/table1. The sentry role setup in Hue can control only table/column access in Hue but It is possible that user can manage to access folders directly in HDFS

c. Limitation: HDFS folders which are not related to Hive, Impala, etc will not be controlled

3. An access control list (ACL) is a list of access control entries (ACE). Each ACE in an ACL identifies a trustee and specifies the access rights allowed, denied, or audited for that trustee

a. Advantage: Folder level access is possible by users using

4. HDFS Encryption implements transparent, end-to-end encryption of data read from and written to HDFS

a. Advantage: Encrypt the data will provide additional level security. In General, Data encryption is required by a number of different government, financial, and regulatory entities

Examples

Example for ACL

```
hadoop fs -setfacl
```

Read Getting started with Big Data/Hadoop Security online:
https://riptutorial.com/bigdata/topic/9869/getting-started-with-big-data-hadoop-security

Credits

S. No	Chapters	Contributors
1	Getting started with bigdata	Ani Menon, Community, Mr. P, NeoWelkin, Sayali Sonawane
2	Getting started with Big Data/Hadoop Security	saranvisa