Table of Contents

[Unit of Data 1](#_Toc409657857)

[How Big are those numbers? 2](#_Toc409657858)

[Companies using Big Data 2](#_Toc409657859)

[Processing capabiltiy 3](#_Toc409657860)

[Some Real Time Hadoop Use cases 3](#_Toc409657861)

[China Mobil Guangdong 3](#_Toc409657862)

[Nokia 4](#_Toc409657863)

[7 V's 4](#_Toc409657864)

[Estimating Big Data 4](#_Toc409657865)

[DataSet Galleries 4](#_Toc409657866)

[Introduction to Hadoop 5](#_Toc409657867)

[Inspiration for Hadoop 6](#_Toc409657868)

[Configuration 1 6](#_Toc409657869)

[Configuration 2 6](#_Toc409657870)

# Unit of Data

smallest unit is bit

1 byte=8 bits

1 KB (Kilo Byte) = 1024 bytes =1024\*8 bits

1MB (Mega Byte) =1024 KB =(1024)^2 \* 8 bits

1 GB (Giga Byte) =1024 MB =(1024)^3 \* 8 bits

1 TB (Tera Byte) =1024GB =(1024)^4 \* 8 bits

1 PB (Peta Byte) =1024 TB =(1024)^5 \* 8 bits

1 EB (Exa Byte) =1024 PB =(1024)^6 \* 8 bits

1 ZB (Zetta Byte) =1024 EB =(1024)^7 \* 8 bits

1 YB (Yotta Byte) =1024 ZB =(1024)^8 \* 8 bits

1 XB (Xenotta Byte) =1024 YB =(1024)^9 \* 8 bits

# How Big are those numbers?

**Reference:-**

 **http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html**

1 byte =A single character

1 KB = A very short story

1 MB=A small novel (6 seconds of TV-quality video)

1 Gigabyte: A pickup truck filled with paper

1 Terabyte : 50000 trees made into paper

2 PB: All US academic research libraries

5 EB: All words ever spoken by human beings

# Companies using Big Data

**http://en.wikipedia.org/wiki/Petabyte**

[AT&T](http://en.wikipedia.org/wiki/AT%26T_Inc.) transfers about 30 petabytes of data through its networks each day

[Google](http://en.wikipedia.org/wiki/Google) processed about 24 petabytes of data per day in 2009

BitTorrent Sync has transferred over 30 petabytes of data since its pre-alpha release in January 2013

The 2009 movie [*Avatar*](http://en.wikipedia.org/wiki/Avatar_%282009_film%29) is reported to have taken over 1 petabyte of local storage at [Weta Digital](http://en.wikipedia.org/wiki/Weta_Digital) for the rendering of the 3D CGI effects

As of January 2013, [Facebook](http://en.wikipedia.org/wiki/Facebook) users had uploaded over 240 billion photos,[[19]](http://en.wikipedia.org/wiki/Petabyte#cite_note-19) with 350 million new photos every day. For each uploaded photo, Facebook generates and stores four images of different sizes, which translated to a total of 960 billion images and an estimated 357 petabytes of storage



**Some More interesting fact**

**Source:-** [**http://wikibon.org/blog/big-data-statistics/**](http://wikibon.org/blog/big-data-statistics/)

* customer spends USD 300,000 on shopping per minutes
* 600 new videos are uploaded per minutes
* Barack Obama used Big Data to win election
* Driver-less cars uses Big Data Processing for driving vehicles
* 100 terabytes of data are [uploaded daily to Facebook](http://wikibon.org/blog/big-data-statistics/);
* 2.7 Zetabytes of data exist in the digital universe today
* Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data
* Akamai analyzes 75 million events per day to better target advertisements
* Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.
* In 2008, Google was processing 20,000 terabytes of data (20 petabytes) a day
* The largest AT&T database boasts titles including the largest volume of data in one unique database (312 terabytes) and the second largest number of rows in a unique database (1.9 trillion), which comprises AT&T’s extensive calling records
* big data will be a $50 billion business by 2017



# Processing capability

Google process 20 PB a day

ebay 6.5 PB of data +50TB/day

#

# Some Real Time Hadoop Use cases

http://www.computerworld.com/article/2492877/government-it/barack-obama-s-big-data-won-the-us-election.html

### China Mobil Guangdong

**Problem:** Storing billions of mobile call records and providing real time access to the call records and billing information to customers.
Traditional storage/database systems couldn't scale to the loads and provide a cost effective solution

**Solution:** HBase is used to store billions of rows of call record details. 30TB of data is added monthly

**Hadoop vendor:** Intel

**Hadoop cluster size:** 100+ nodes

### Nokia

Nokia collects and analyzes vast amounts of data from mobile phones

**Problem:**
(1) Dealing with 100TB of structured data and 500TB+ of semi-structured data
(2) 10s of PB across Nokia, 1TB / day

**Solution:** HDFS data warehouse allows storing all the semi/multi structured data and offers processing data at peta byte scale

**Hadoop Vendor:** Cloudera

**Cluster/Data size:**
(1) 500TB of data
(2) 10s of PB across Nokia, 1TB / day

https://gigaom.com/2012/06/05/10-ways-companies-are-using-hadoop-to-do-more-than-serve-ads/

# 7 V's

1. **Volume:** The amount of data generated
	* By 2020,we will have 50 times the data wrt to 2011 data.
	* Aeroplane generates 2.5 billion TB of data each year through sensor.
	* Self-driving car will generate 2 PB of data each year.
2. **Variety:** Range of data types and sources. structured data, unstructured data. Earlier data was structured but these days 90% of the data are not structured.
3. **Velocity:** The speed at which data is generated
	* 100 hours of video on Youtube are uploaded every minute
	* 200 million emails are sent every minute
	* 2.5 million queries on Google every minute.
4. **Veracity:** Data should be correct and complete without error. data should not be lost due to transmission of data. Say you have are log files in four formats from six systems, some incomplete, with noise and errors. These inaccurate data has to be corrected or remove to get more effective result.
5. **Visualization:** Once the data is processed we need some way to present the data in a manner that's readable and accessible. The data should be processed to visualize it in terms of business through different graphs charts etc.
6. **Value**: Last year US Health care system saved around $450 billion in Health care spending. The cost of poor data is also huge. Data on its own is worthless. The value lies in rigorous analysis of accurate data, and the information and insights this provides.
7. **Variability**: A single word can have multiple meaning. While processing data we should be clear about its context. The below 3 lines has different meaning for the word **"Great"**
	* “Delicious muesli from the @imaginarycafe- what a great way to start the day!”
	* “Greatly disappointed that my local Imaginary Cafe have stopped stocking BLTs.”
	* “Had to wait in line for 45 minutes at the Imaginary Cafe today. Great, well there’s my lunchbreak gone…”

**Conclusion:** Well, it’s massive and rapidly-expanding, but it’s also noisy, messy, constantly-changing, in hundreds of formats and virtually worthless without analysis and visualisation.

# Estimating Big Data

When size is atleast 1 TB(1 PB)

40TB 100 TB

# DataSet Galleries

http://www.infochimps.com/

http://datamob.org/datasets

http://aws.amazon.com/public-data-sets/

http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free

http://datacatalog.worldbank.org/

http://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public

**Evolution of Hadoop:**

* 2003- Google published GFS Papers
* 2004- Google published Map-Reduce Papers
* 2004- Doug Cutting designing Web Crawler called Apache Nutch.
* Doug cutting & Michael Cafarella implemented GFS to come out with Hadoop.

# Introduction to Hadoop

A Java Distributed Framework for processing big data

Scientist: Doug cutting & Michael Cafarella

Google: published a paper called GFS(Google File System)

Implementation: Lucene, (search engine to search Document), to get all the document available in the internet,store them on some file system,and then start processing something on it.

[Hadoop](http://hadoop.apache.org/) is a framework written in Java for running applications on large clusters of commodity hardware and incorporates features similar to those of the [Google File System](http://en.wikipedia.org/wiki/Google_File_System) and of [MapReduce](http://en.wikipedia.org/wiki/MapReduce). [HDFS](http://hadoop.apache.org/hdfs/docs/current/hdfs_design.html) is a highly fault-tolerant distributed file system and like Hadoop designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications that have large data sets.

Consider a text file of 10 GB containing employee data in the below format and lets say you have 100 such files,summing the data upto 1 TB.

**employee.txt(**empId,empName,empAge,empSal,empDes**)**

101,prasad,t20,1000,lead

Assume you have around 100,00,000,00000,0000000000 records and you would like to find out all the employees above 60 years of age.How do you program them traditionally.



10 GB= 10 min

1 TB= 1000 minutes =16 hours

Google process 20 PB of data per day

To process 20 PB it will take 3200 hours = 133 days

# Inspiration for Hadoop

1. To store huge data(unlimited)
2. To process huge data

## Configuration 1

* 2 TB HDD
* 24 GB\*N(unlimited)
* 8 core-processor\* Unlimited

## Configuration 2

* 4 TB HDD
* 48 GB RAM
* 8 core Processor

http://hortonworks.com/blog/best-practices-for-selecting-apache-hadoop-hardware/

**Note:**

* Dont use a single supercomputer to process the data
* Use N- simple computers to process the data
* Follow the concept of A farmer->Ox for field(2)