Introduction to Big Data

Original Slides by Dr Sandeep Deshmukh, SadePach Labs
Modifications by Dr Amey Karkare, IIT Kanpur

What is Big Data?

- Big data is data that exceeds the processing capacity of conventional database systems.
- The data is too big, moves too fast, or doesn’t fit the structures of your database architectures.
- To gain value from this data, you must choose an alternative way to process it.

https://www.oreilly.com/ideas/what-is-big-data

Definition

“Big data” is

high-volume, -velocity and -variety information assets

that demand cost-effective, innovative forms of information processing

for enhanced insight and decision making

By Gartner

Definition

“Big data” is

high-volume, -velocity and -variety information assets

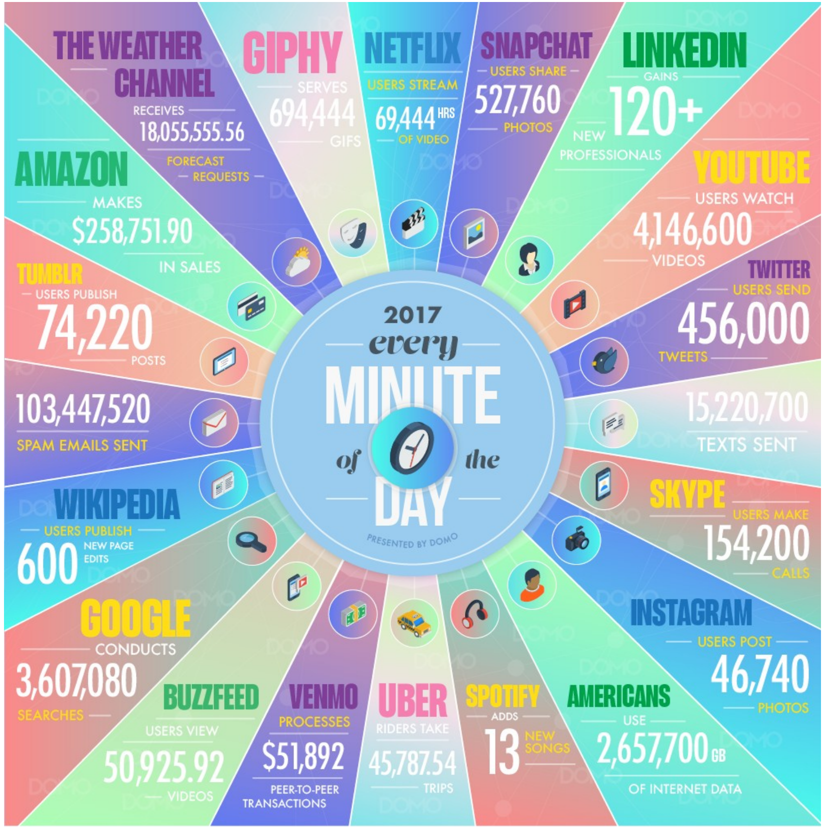
that demand cost-effective, innovative forms of information processing

for enhanced insight and decision making

By Gartner

The Three *V*-s

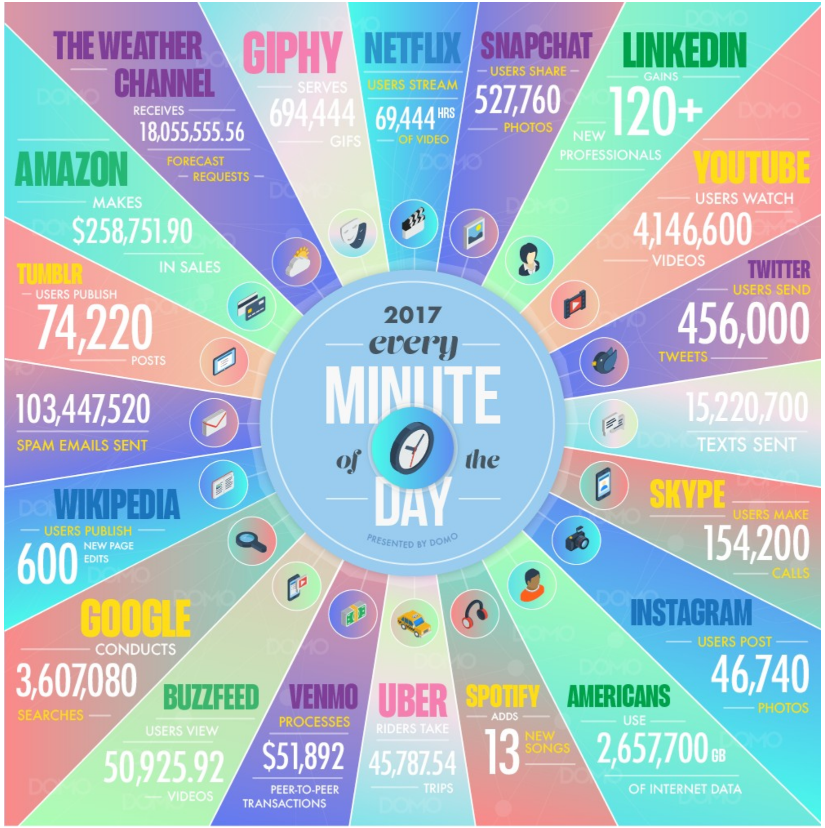
Volume



- Quantity of data
- Data sets too large to store and analyse using traditional databases

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

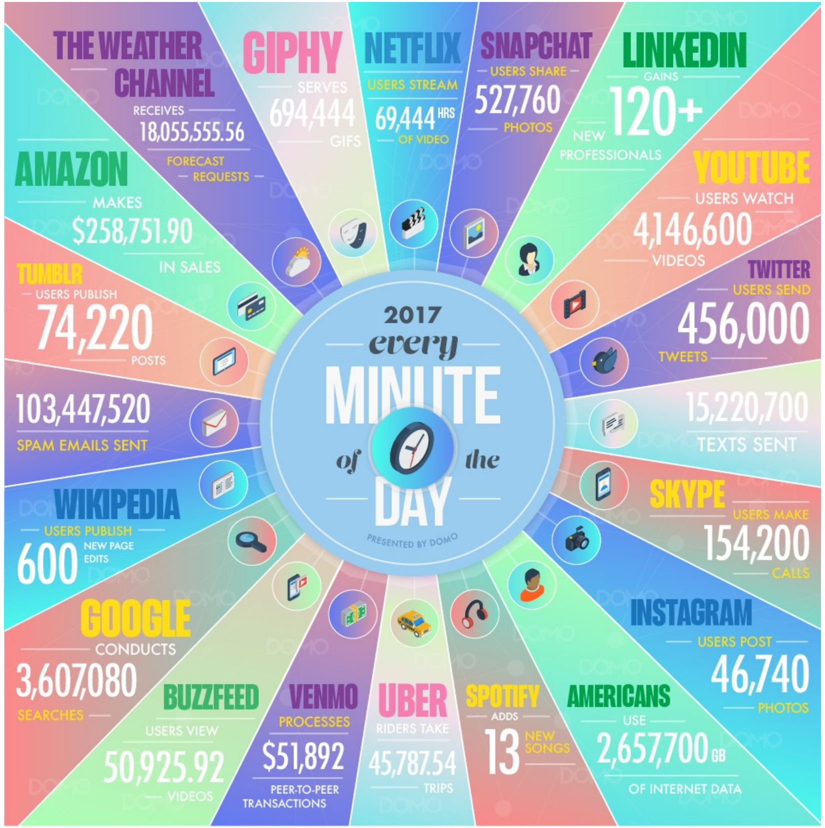
Velocity



- Speed at which data is generated
- Speed at which data is moving around and analysed
- Processing should be faster than generation
- Analyse data while it is being generated without even putting it into databases

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

Variety

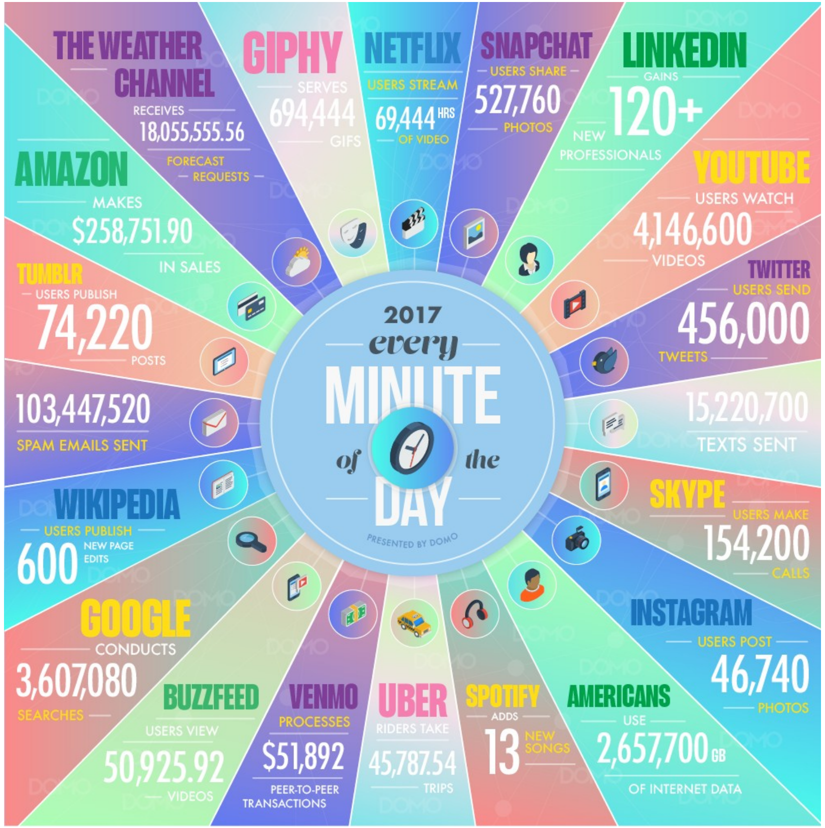


- Different types of data that we can use
- Generated by different entities
  - Humans
  - Machines (HW + SW)
  - Sensors

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

Additional *V*-s

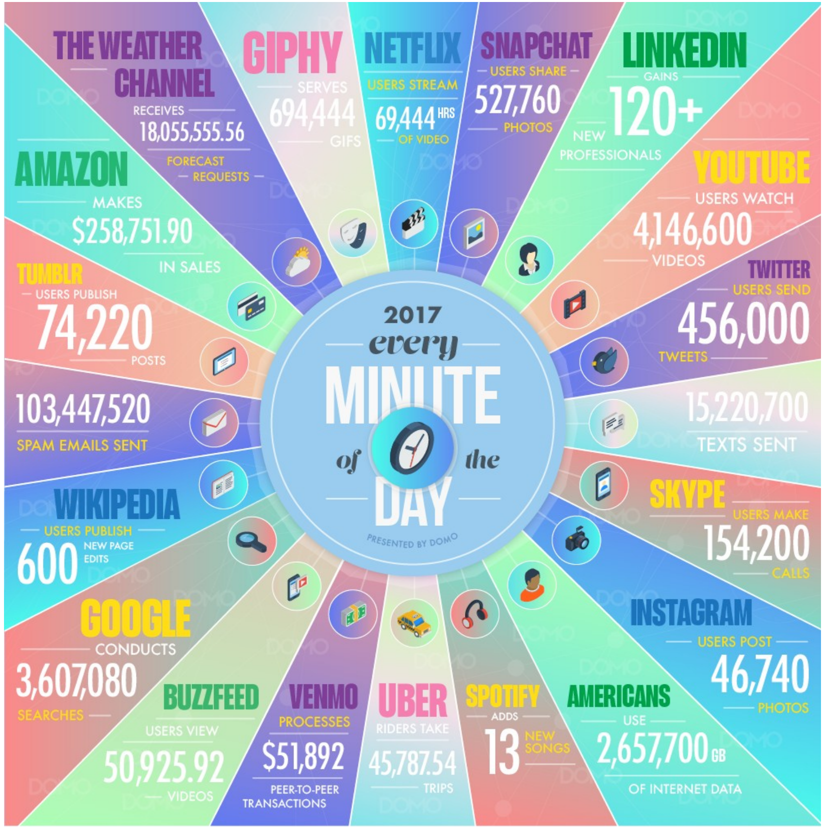
Veracity



- Messiness or trustworthiness of the data
- Volumes makes up for quality
  - Eg. Tweets with spelling mistakes, short words
  - u→you, thr→there, teh→the

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

Value



Getting value out of Big Data!!!

Image Source: https://www.domo.com/blog/data-never-sleeps-5/

Definition

“Big data” is

high-volume, -velocity and -variety information assets

that demand cost-effective, innovative forms of information processing

for enhanced insight and decision making

By Gartner

Definition

“Big data” is

high-volume, -velocity and -variety information assets

that demand cost-effective, innovative forms of information processing

for enhanced insight and decision making

By Gartner

Wikipedia Definition

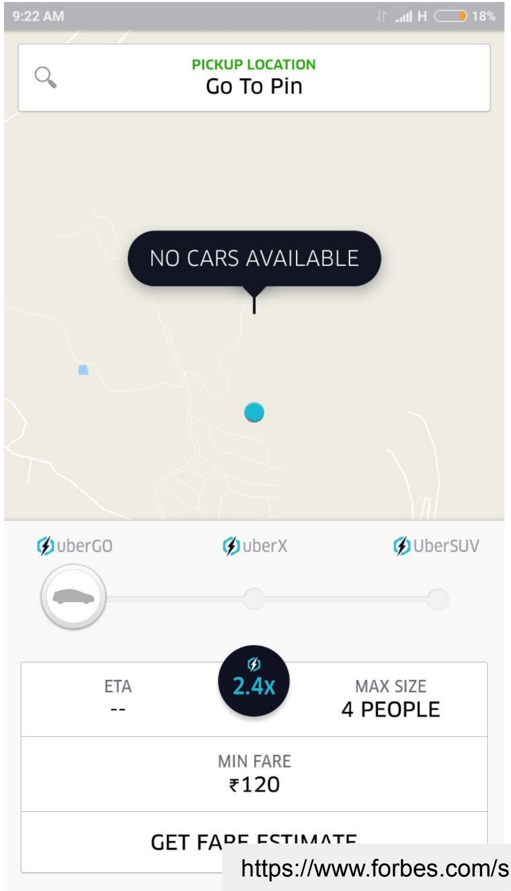
- Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate…
- Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. …
- The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. …
- Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Use cases

Use Case: Big Data in Oil & Gas Drilling

http://analytics-magazine.org/how-big-data-is-changing-the-oil-a-gas-industry/

Use Case: Uber - Pay Surge Pricing if Battery is Low



PICKUP LOCATION
Go To Pin

NO CARS AVAILABLE

ETA
--

2.4x

MAX SIZE
4 PEOPLE

MIN FARE
₹120

https://www.forbes.com/sites/amitchowdhry/2016/05/25/uber-low-battery/#19762c0474b3

Big Data Challenges

Big Data Challenges: Size does matter

1KB	Kilobyte
1MB	Megabyte
1GB	Gigabyte
1TB	Terabyte
1PB	Petabyte
1EB	Exabyte
1ZB	Zettabyte
1YB	Yottabyte

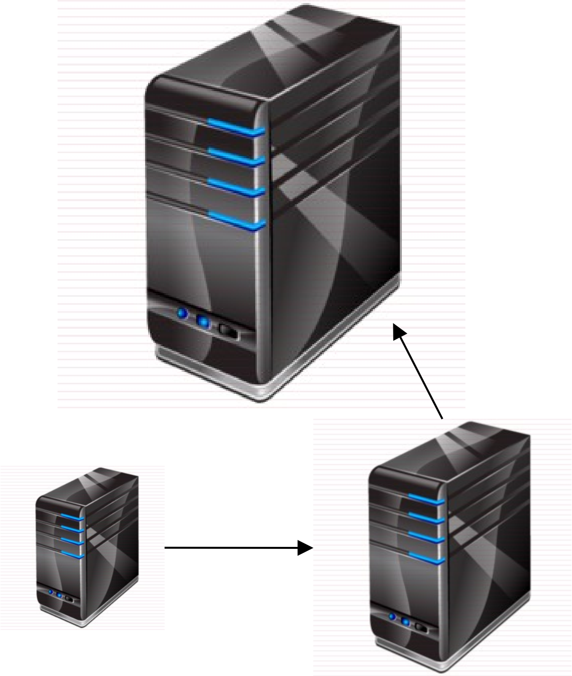
1 GB = 1 hr

1 TB = 1024 hrs = 102 days

1 PB = 286 yrs **> 1 lifetime**

1 EB = 293K yrs

Big Data Challenges: Vertical Vs Horizontal Scaling



Vertical Scaling

Horizontal Scaling

Big Data Challenges

Scaling

Source:
https://s-media-cache-ak0.pinimg.com/736x/10/0c/d0/100cd0da1c19e5d6f850ed23c3633714.jpg

Big Data Challenges: Scale of infrastructure

Image Source: https://datacenter.legrand.com

Further Reading

- A Brief History of Big Data Everyone Should Read
- Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity
- What is big data? - OpenSource.com & O'Reilly
- Uber Use Case
- 5 Big Data Use Cases To Watch
- Best Big Data Analytics Use Cases
- The 5 game changing big data use cases
- Big Data - The 5 Vs Everyone Must Know
- Top SlideShare Presentations on Big Data
- Google Data Center 360° Tour

Questions?

How to store *huge* files?

Requirements?

- Efficient Access
- Effective utilization of space
- Redundancy (Failsafe)
  - Given: probability of 1 disk failing is 1% per year
  - What are the chances that 1 out of 10^3 disk fails at a data center?

HDFS

Hadoop distributed File System

HDFS

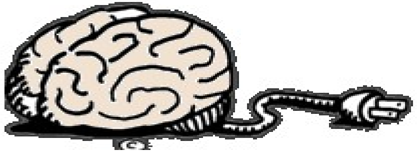
- Data storage system used by Hadoop
  - Hadoop: Project to develop open-source software for reliable, scalable, distributed computing★
  - Will discuss Hadoop later
- Components
- Architecture
- Tasks / Services

★ http://hadoop.apache.org/

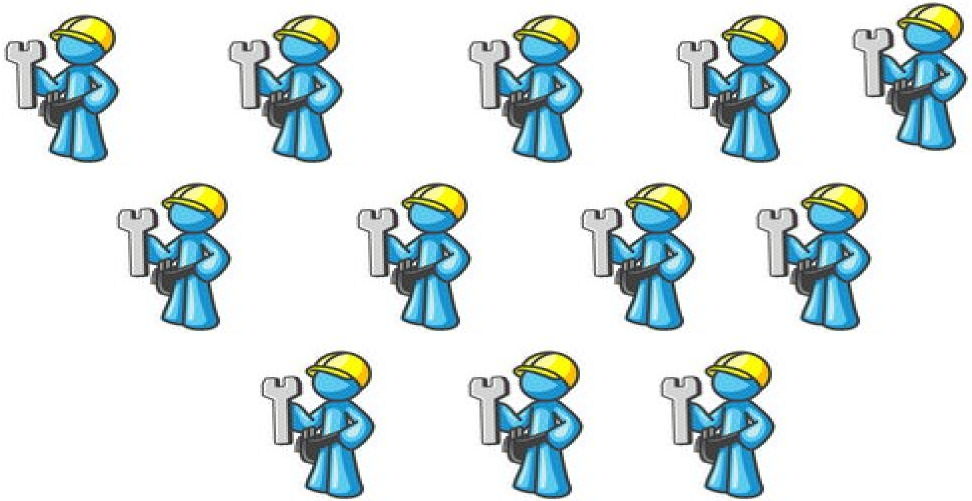
Components of HDFS

Secondary NameNode

Active NameNode



Standby NameNode



DataNodes

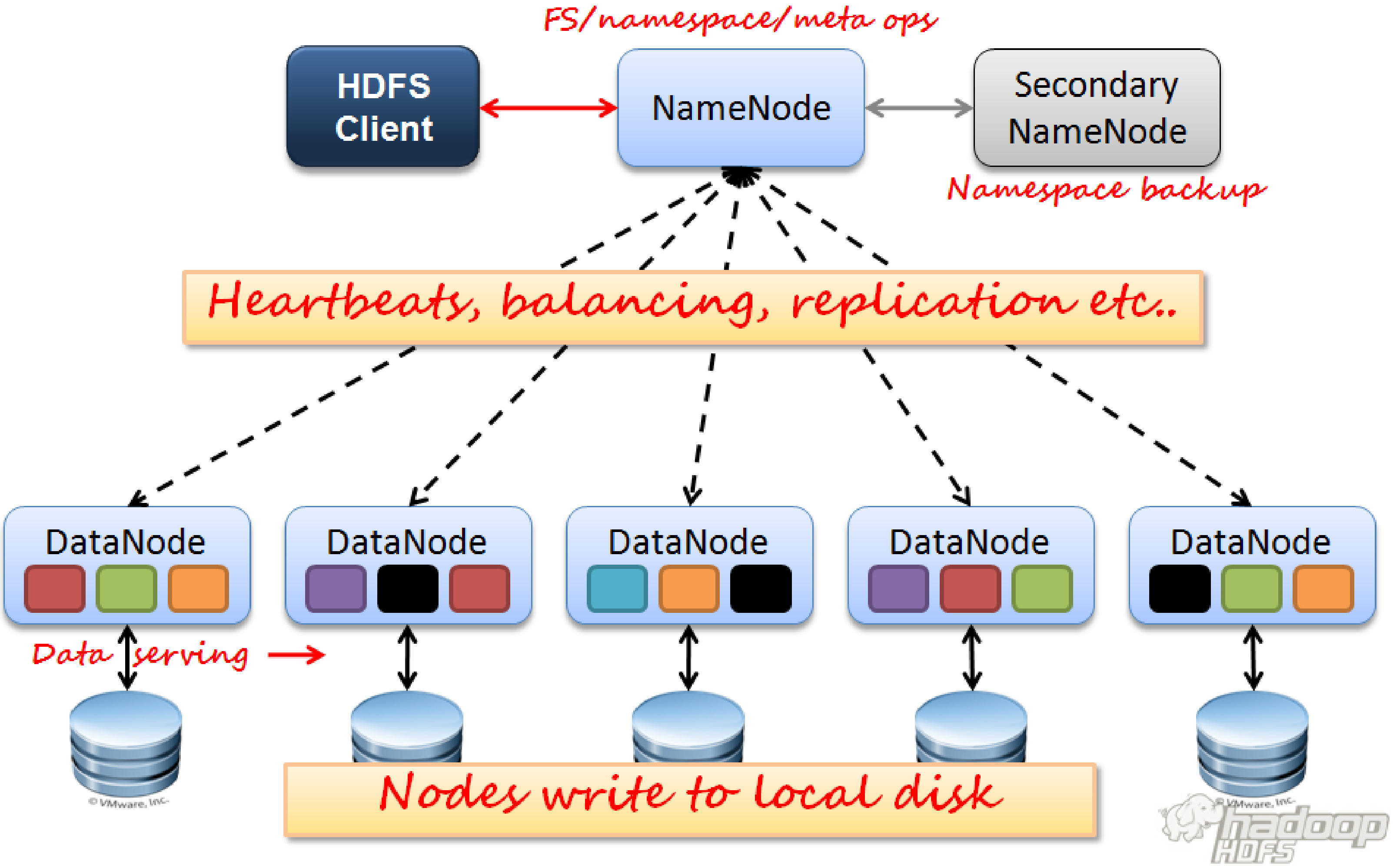
Terminology

- **HDFS**: Hadoop Distributed File System
- **Datanode**: A DataNode stores data in HDFS.
- **Namenode**: The centerpiece of an HDFS file system.
  - Keeps the directory tree of all files in the file system
  - Tracks where across the cluster the file data is kept.
    - Does not store the data of these files itself.
  - Active : Actively serving request
  - Standby: Becomes Active if the current Active node fails

Terminology

- **Secondary Namenode:**
  - helper node for namenode
  - Puts a checkpoint in filesystem which will help Namenode to function better

Architecture



FS/namespace/meta ops

HDFS Client

NameNode

Secondary NameNode

Namespace backup

Heartbeats, balancing, replication etc..

DataNode

DataNode

DataNode

DataNode

DataNode

Data serving

Nodes write to local disk

Storing file on HDFS

Motivation: Reliability, Availability , Network Bandwidth

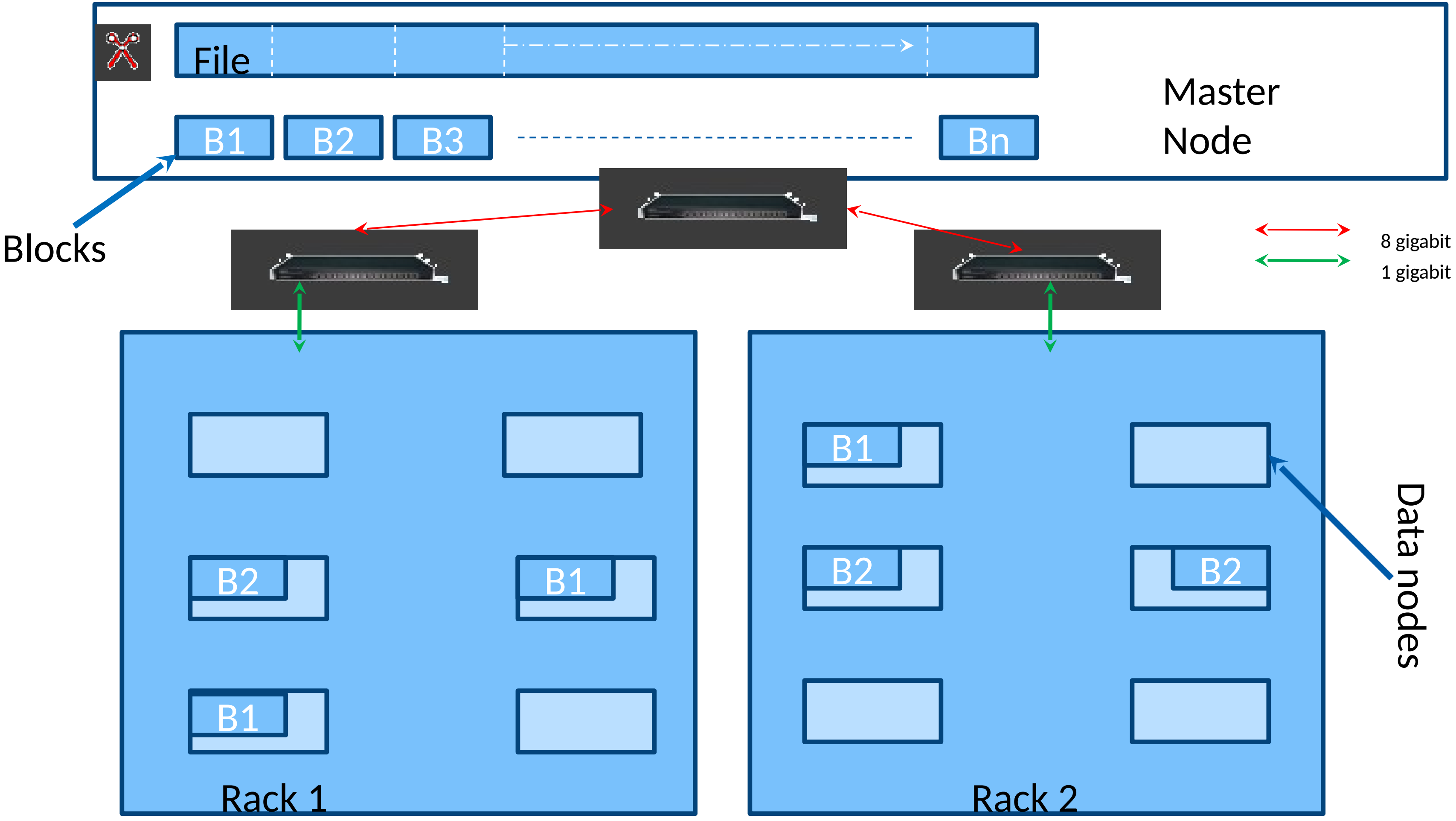
- The input file (say 1 TB) is split into smaller chunks/blocks of 128 MB
- The chunks are stored on multiple nodes as independent files on data nodes

Storing file on HDFS

- To ensure that data is not lost, data can typically be replicated on:
  - local rack
  - remote rack (in case local rack fails)
  - remote node (in case local node fails)
  - randomly
- Default replication factor is 3

Storing file on HDFS

- Default replication factor is 3
  - first replica of a block will be stored on a local rack
  - the next replica will be stored on a remote rack
  - the third replica will be stored on the same remote rack but on a different Datanode
  - Why?
- More replicas?
  - the rest will be placed on random Datanodes
  - As far as possible, no more than two replicas are kept on the same rack


File
Master
Node
B1
B2
B3
Bn
Blocks
8 gigabit
1 gigabit
B2
B1
B1
B1
B2
B2
Data nodes
Rack 1
Rack 2

Tasks of NameNode

- Manages File System
  - mapping files to blocks and blocks to data nodes
- Maintaining status of data nodes
  - Heartbeat
    - Datanode sends heartbeat at regular intervals
    - If heartbeat is not received, datanode is declared dead
  - Blockreport
    - DataNode sends list of blocks on it
    - Used to check health of HDFS

NameNode Functions

- ❑ Replication
  - ➢ On Datanode failure
  - ➢ On Disk failure
  - ➢ On Block corruption
- ❑ Data integrity
  - ➢ Checksum for each block
  - ➢ Stored in hidden file

- ❑ Rebalancing - balancer tool
  - ➢ Addition of new nodes
  - ➢ Decommissioning
  - ➢ Deletion of some files

HDFS Robustness

- Safemode
  - At startup: No replication possible
  - Receives Heartbeats and Blockreports from Datanodes
  - Only a percentage of blocks are checked for defined replication factor
- Replicate blocks wherever necessary

All is well ⍰ → Exit Safemode

HDFS Summary

- Fault tolerant
- Scalable
- Reliable
- File are distributed in large blocks for
  - Efficient reads
  - Parallel access

Questions?