Industrial Machine Learning on Hadoop and Spark

Maks Nakhodnov, Bremen 2025

Distributed computing. Introduction

Plan:

What is big data, and where does it come from?

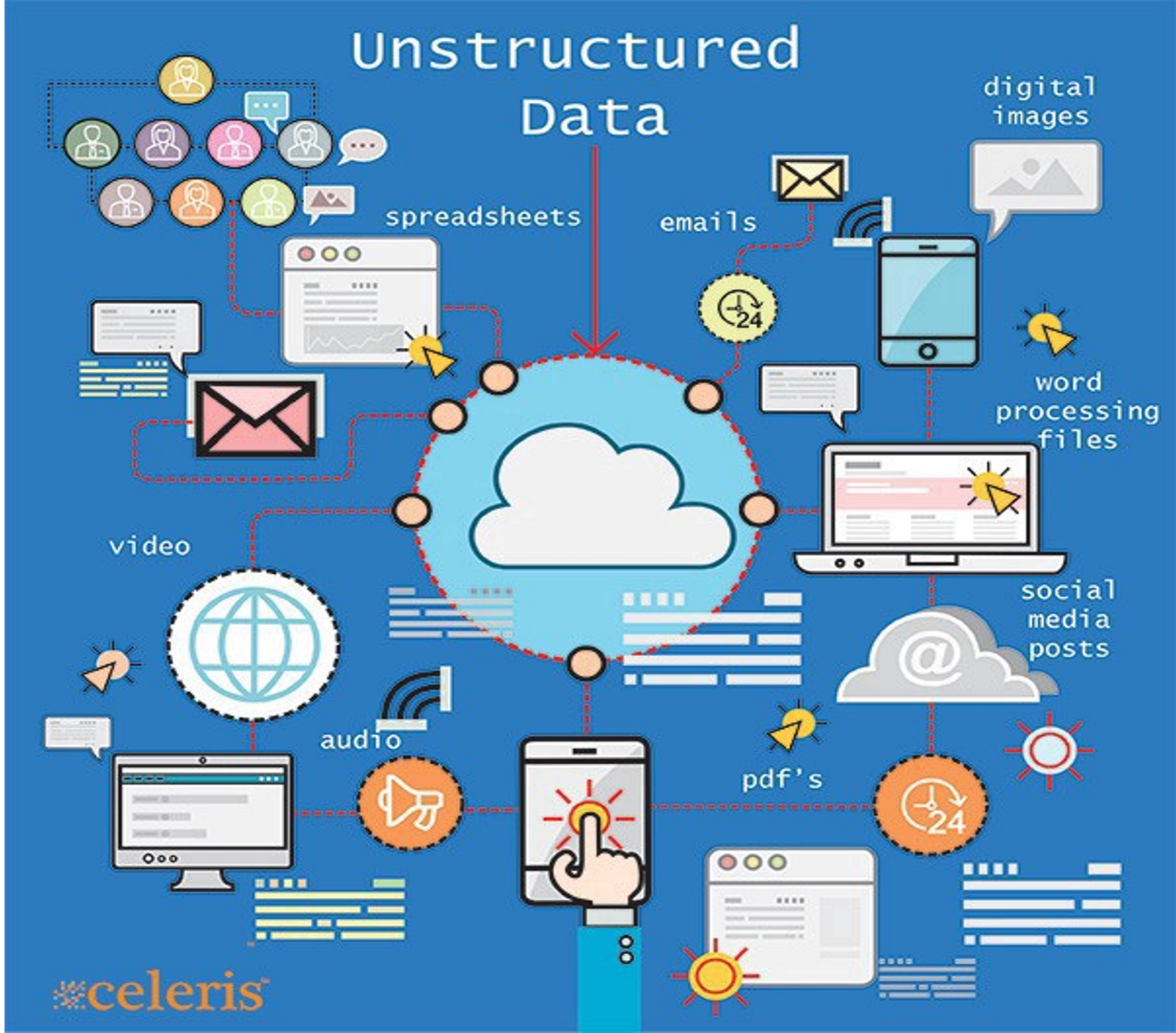
- Industries generating big data. Data explosion.
- Big data – the starting point. 5 main principles.
- How companies handle big data: IaaS / PaaS / SaaS.

Industries generating big data.
Data explosion.

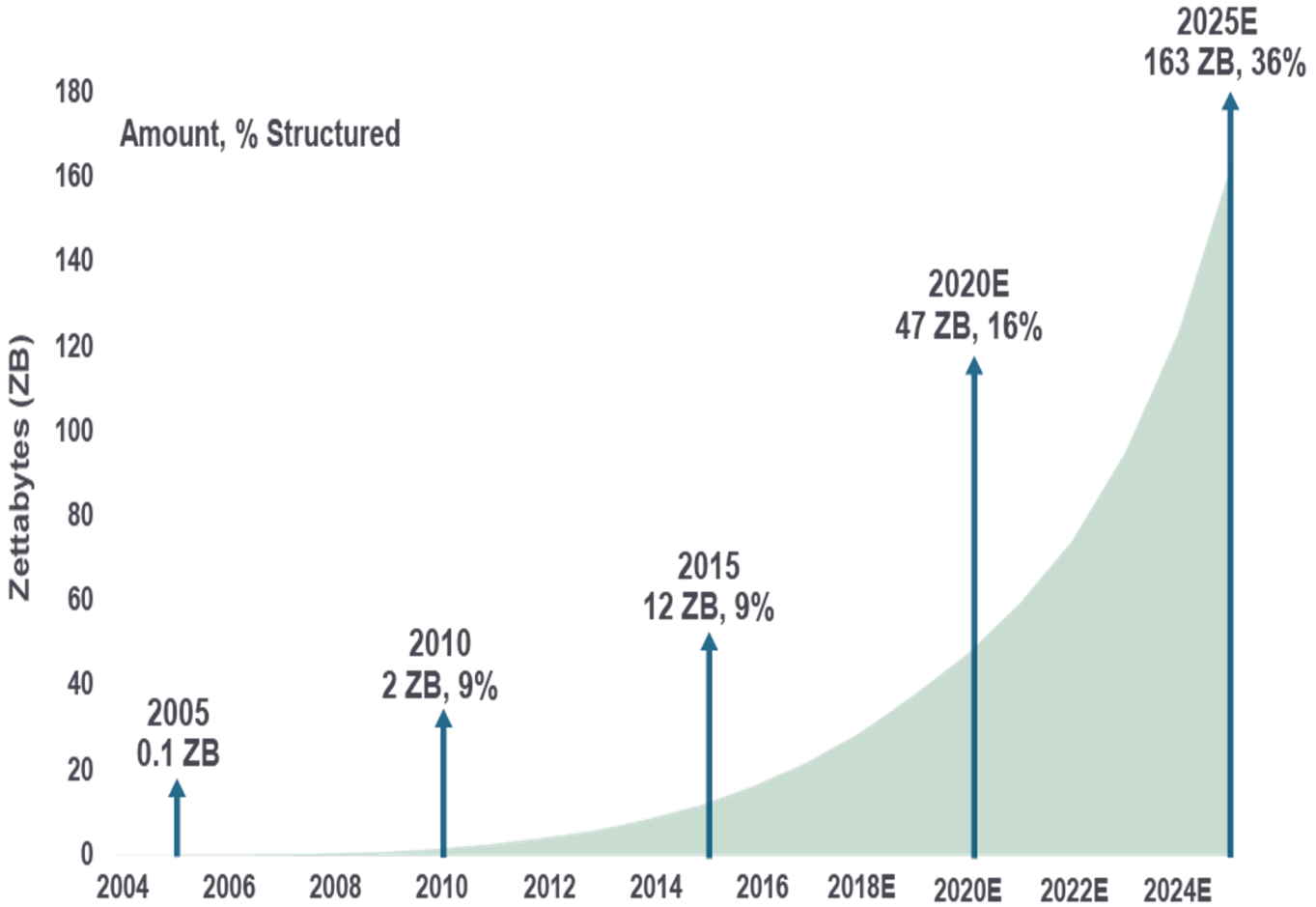
Origins of Big Data

Industries:

- Telecom
- Banking
- Social networks
- Media
- Industry
- Bioinformatics
- Internet of Things (IoT)



Data explosion



One zettabyte equals **one trillion gigabytes**.

1 kilobyte	**1 000**
1 megabyte	**1 000 000**
1 gigabyte	**1 000 000 000**
1 terabyte	**1 000 000 000 000**
1 petabyte	**1 000 000 000 000 000**
1 exabyte	**1 000 000 000 000 000 000**
1 zettabyte	**1 000 000 000 000 000 000 000**

Processes 20Pb per day (2008)
Downloads 20B web pages per day (2012)

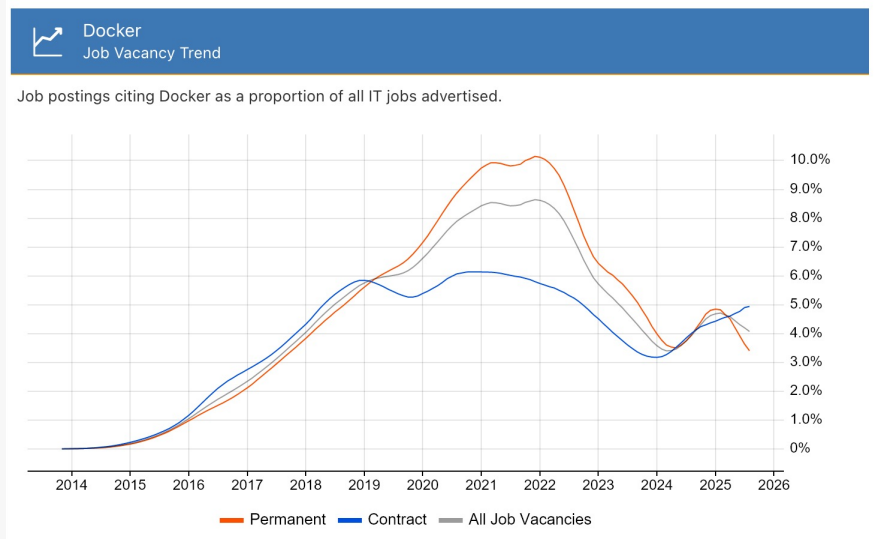
>10 PB data,
75B DB (6/2012)

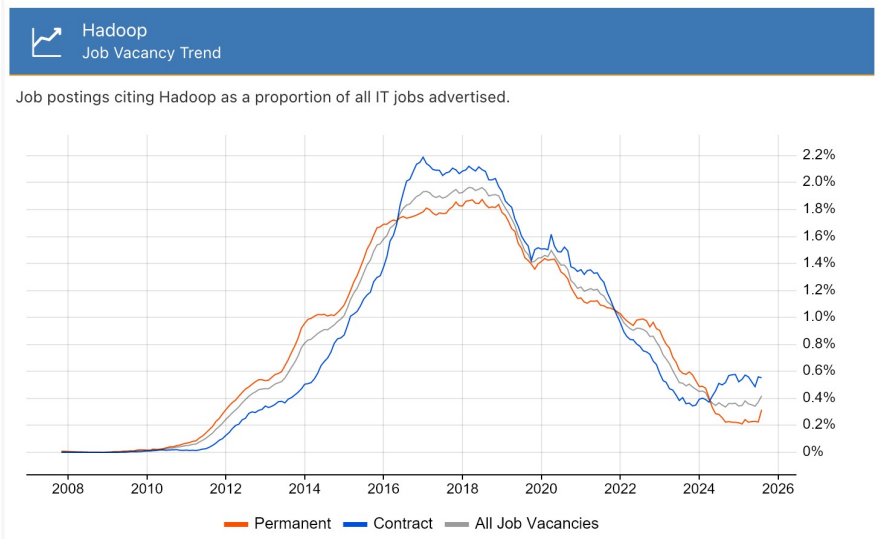
>100 Pb user data +
500 TB/day (8/2012)

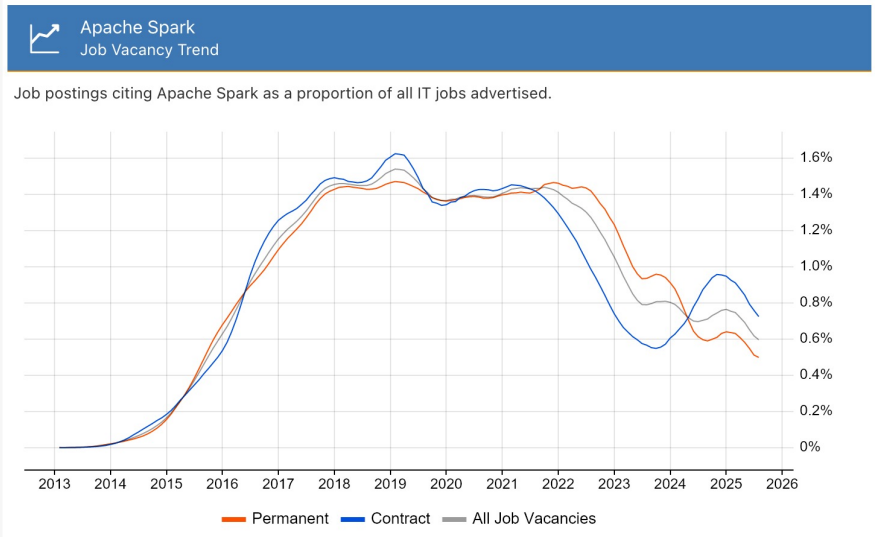
S3: 449B objects (7/2011)

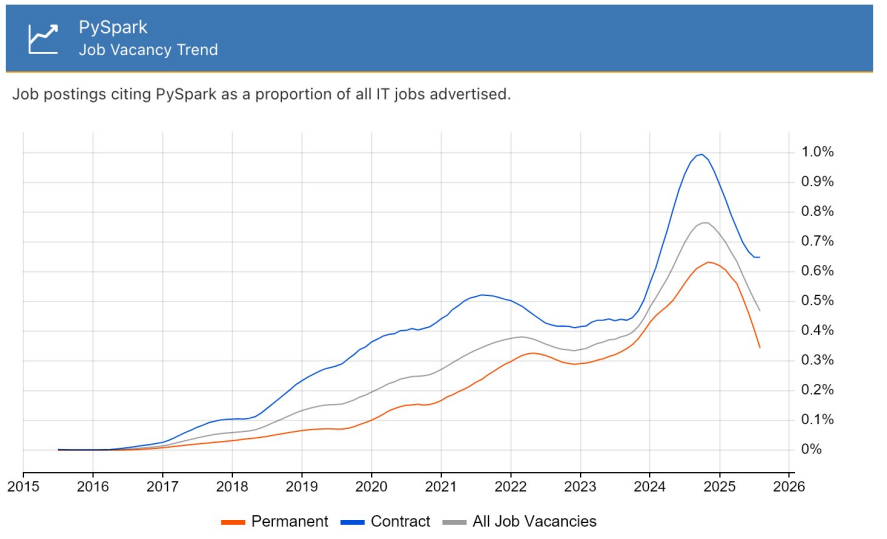
How much is “big data”?

Job trends









Source: https://www.itjobswatch.co.uk/jobs/uk/

Big data – the starting point.
5 main principles.

Big data – 5 main principles

Veracity

Volume

- Terabytes
- Records
- Transactions
- Tables, files

Value

- **Insight**
- **Accuracy**
- **Impact**

Velocity

- Tweets
- Facebook posts
- Device sensors

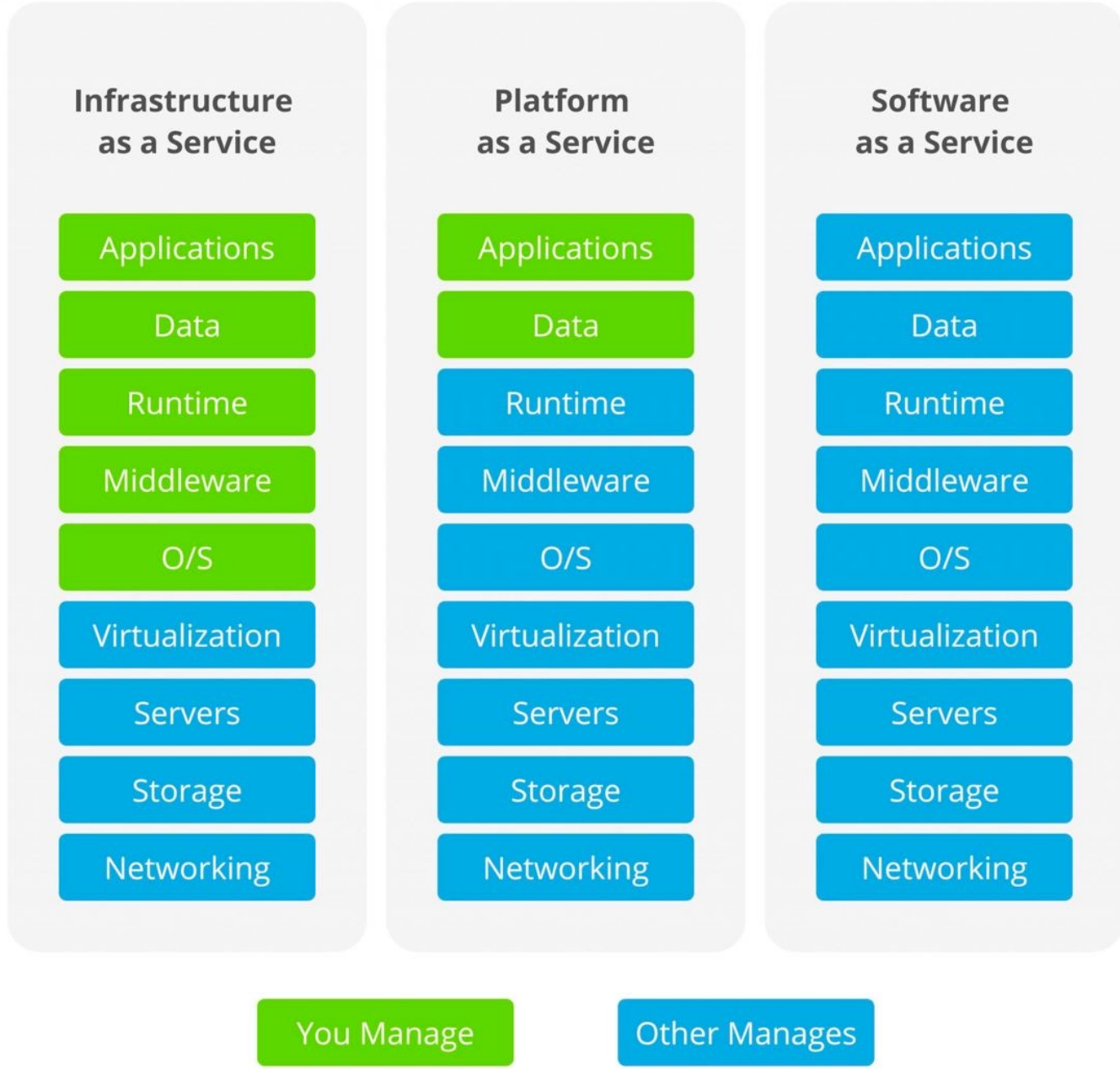
- **Data generation speed**
- **Generated in real-time**
- **Online and offline data**
- **In streams, batches, or bits**

Variety

- Video
- Audio
- Text data
- Time series

- **Structured**
- **Unstructured**
- **Semi-structured**

How do companies handle big data?


Infrastructure as a Service
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
Platform as a Service
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
Software as a Service
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
You Manage
Other Manages

- **IaaS — Infrastructure as a Service.** Provides infrastructure resources such as virtual servers, storage, and networks (Google Compute Engine, DigitalOcean, Amazon Web Services (AWS), Cisco Metacloud).
  - Migrating IT systems to the cloud
  - Saving on infrastructure costs
  - Rapid business launch
  - Scaling infrastructure
  - Infrastructure for companies with demand spikes
  - Development and testing

- **PaaS — Platform as a Service.** Provides a platform for building and deploying applications (Windows Azure, OpenShift, Heroku, Google App Engine).
  - Databases
  - Application development in containers
  - Big data analytics
  - Machine learning

- **SaaS — Software as a Service.** Provides software applications as a service (Google App Engine, Dropbox, JIRA).
  - Email
  - CRM systems
  - Task planners
  - Website builders

The first key question: how to store big data?

Filesystem

The main functions of a filesystem are:

- Storing and organizing data on a storage device as files
- Determining the maximum supported data volume on the storage device
- Creating, reading, and deleting files
- Assigning and modifying file attributes (size, creation and modification time, owner and creator, read-only, hidden, temporary, archive, executable, maximum filename length, etc.)
- Defining the structure of a file
- Searching for files
- Organizing directories for logical file organization
- Protecting files in case of system failure
- Protecting files from unauthorized access and modification

Linux File System Directories

/bin: Where Linux core commands reside like ls, mv.

/boot: Where boot loader and boot files are located.

/dev: Where all physical drives are mounted like USBs DVDs.

/etc: Contains configurations for the installed packages.

/home: Where every user will have a personal folder to put his folders with his name like /home/likegeeks.

/lib: Where the libraries of the installed packages located since libraries shared among all packages, unlike Windows, you may find duplicates in different folders.

/media: Here are the external devices like DVDs and USB sticks that are mounted, and you can access their files from here.

/mnt: Where you mount other things Network locations and some distros, you may find your mounted USB or DVD.

/opt: Some optional packages are located here and managed by the package manager.

/proc: Because everything on Linux is a file, this folder for processes running on the system, and you can access them and see much info about the current processes.

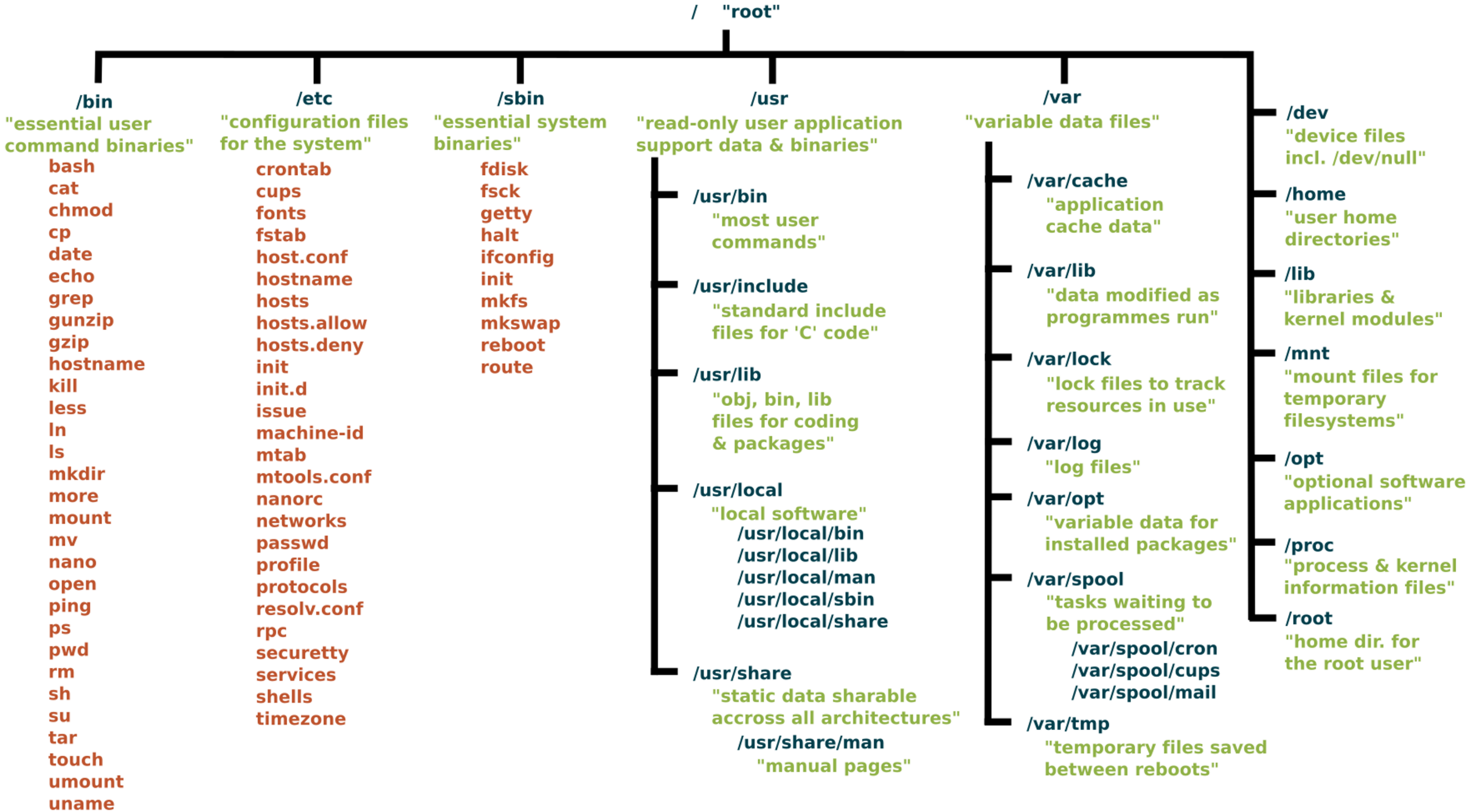
/root: The home folder for the root user.

/sbin: Like /bin, but binaries here are for root user only.

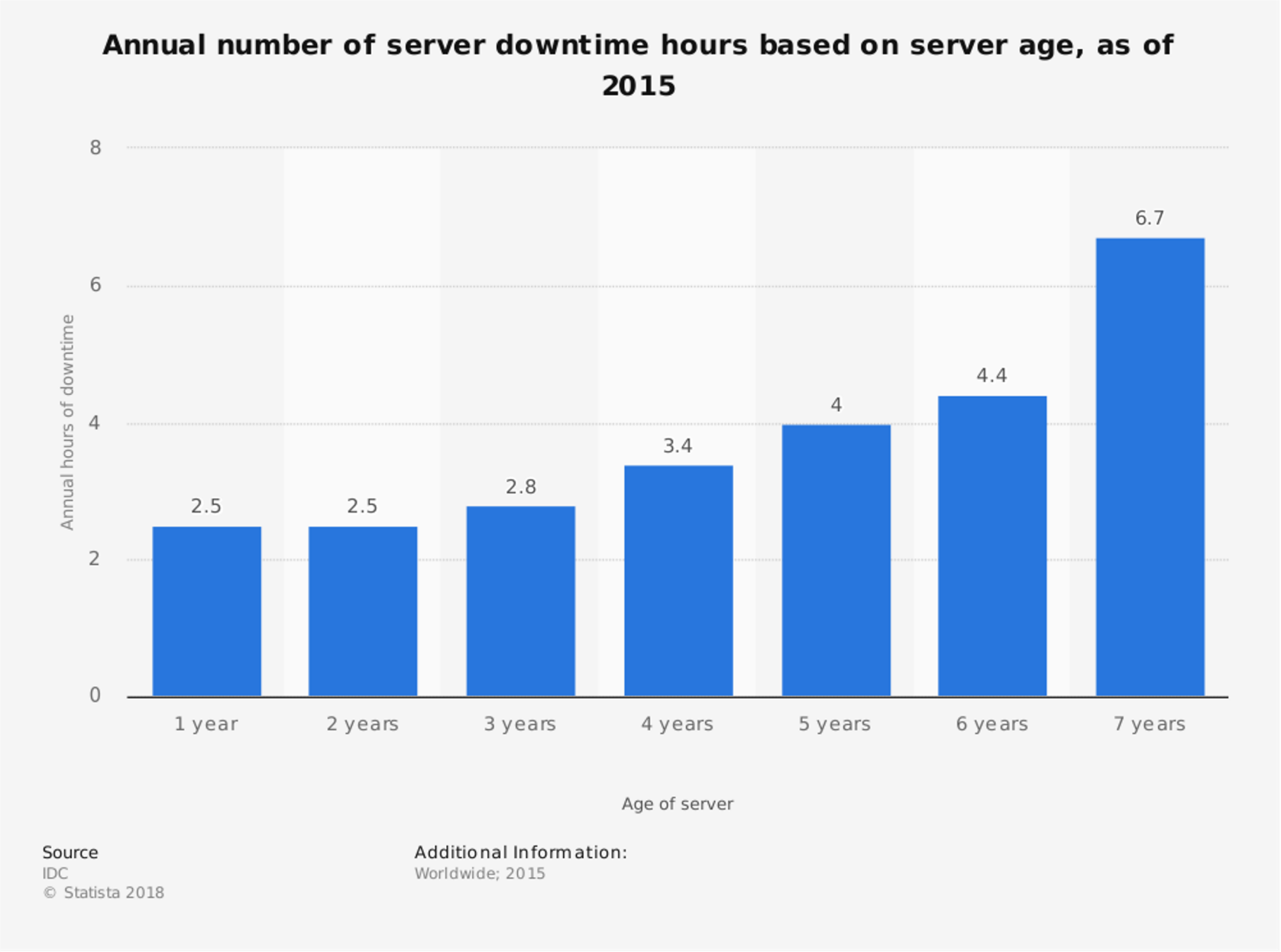
/tmp: Contains the temporary files.

/usr: Where the utilities and files shared between users on Linux.

/var: Contains system logs and other variable data.


/ "root"
/bin
"essential user command binaries"
bash
cat
chmod
cp
date
echo
grep
gunzip
gzip
hostname
kill
less
ln
ls
mkdir
more
mount
mv
nano
open
ping
ps
pwd
rm
sh
su
tar
touch
umount
uname
/etc
"configuration files for the system"
crontab
cups
fonts
fstab
host.conf
hostname
hosts
hosts.allow
hosts.deny
init
init.d
issue
machine-id
mtab
mtools.conf
nanorc
networks
passwd
profile
protocols
resolv.conf
rpc
securetty
services
shells
timezone
/sbin
"essential system binaries"
fdisk
fsck
getty
halt
ifconfig
init
mkfs
mkswap
reboot
route
/usr
"read-only user application support data & binaries"
/usr/bin
"most user commands"
/usr/include
"standard include files for 'C' code"
/usr/lib
"obj, bin, lib files for coding & packages"
/usr/local
"local software"
/usr/local/bin
/usr/local/lib
/usr/local/man
/usr/local/sbin
/usr/local/share
/usr/share
"static data sharable accross all architectures"
/usr/share/man
"manual pages"
/var
"variable data files"
/var/cache
"application cache data"
/var/lib
"data modified as programmes run"
/var/lock
"lock files to track resources in use"
/var/log
"log files"
/var/opt
"variable data for installed packages"
/var/spool
"tasks waiting to be processed"
/var/spool/cron
/var/spool/cups
/var/spool/mail
/var/tmp
"temporary files saved between reboots"
/dev
"device files incl. /dev/null"
/home
"user home directories"
/lib
"libraries & kernel modules"
/mnt
"mount files for temporary filesystems"
/opt
"optional software applications"
/proc
"process & kernel information files"
/root
"home dir. for the root user"

Fault tolerance in data storage and computation



The probability that a failure will occur in the next hour:

$$P = 2.5 / (24 * 365) = 0.00028$$

$$P(\text{will not fail}) = (1 - P) = 0.9997$$

A cluster of 1000 machines

The probability that one of the servers will fail in the next hour:

$$1 - 0.9997^{1000} = 0.25$$