Big Data, its challenges and large data storage

Presented By: Ijlal Ahmed Niazi

Ijlal.Niazi@student.lut.fi

Date: 7/6/2018

ITMO University, St.Petersburg

Course: Cluster, Grid, Clouds

Course Moderator: Andrey Shevel

Outline

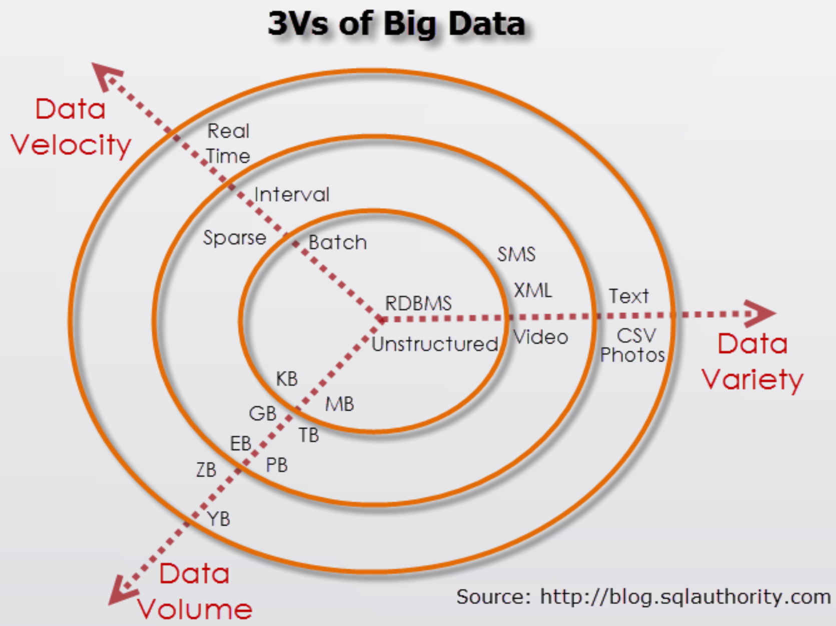
- Definition of big data
- Sources of big data
- Challenges in big data
- Data volume scale – Petabytes
- Storage of big data
- Challenges in storage

What is big data

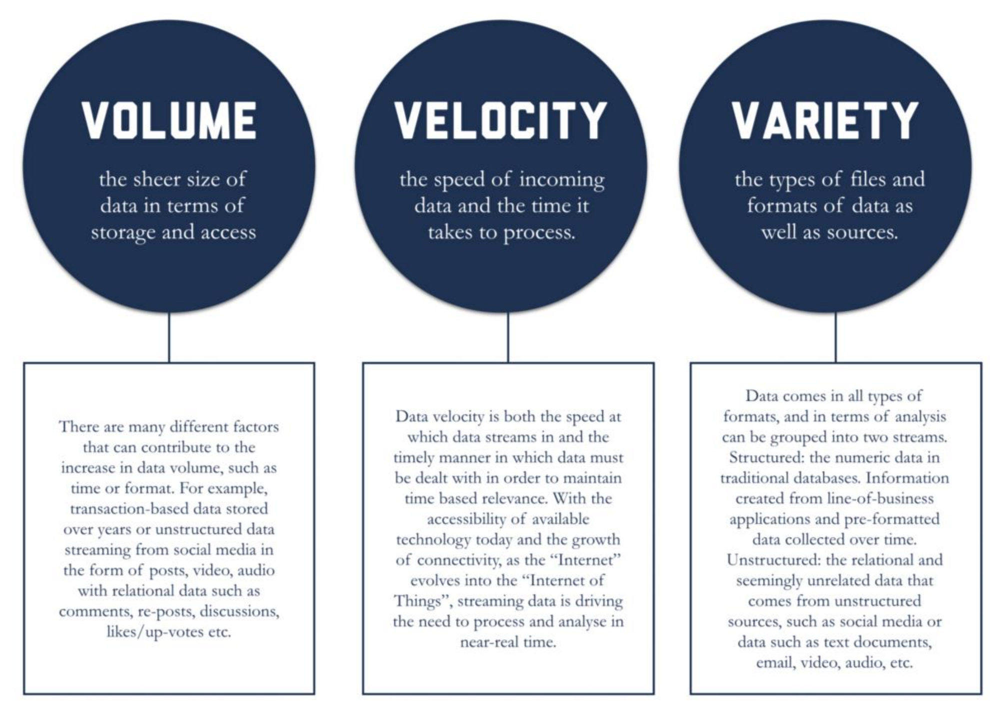
- Data of:
  - High magnitude
  - Advanced complexity
  - Non-conventional Analysis
- Data generated by devices would reach 507.5 zettabytes (ZB) per year (42.3 ZB per month) by 2019, 269 times higher than the amount of data being transmitted to datacenters from end-user devices and 49 times higher than total datacenter traffic

The V’s model

- Three V’s of big data to describe characteristics



Source: http://blog.sqlauthority.com



VOLUME

the sheer size of data in terms of storage and access

There are many different factors that can contribute to the increase in data volume, such as time or format. For example, transaction-based data stored over years or unstructured data streaming from social media in the form of posts, video, audio with relational data such as comments, re-posts, discussions, likes/up-votes etc.

VELOCITY

the speed of incoming data and the time it takes to process.

Data velocity is both the speed at which data streams in and the timely manner in which data must be dealt with in order to maintain time based relevance. With the accessibility of available technology today and the growth of connectivity, as the “Internet” evolves into the “Internet of Things”, streaming data is driving the need to process and analyse in near-real time.

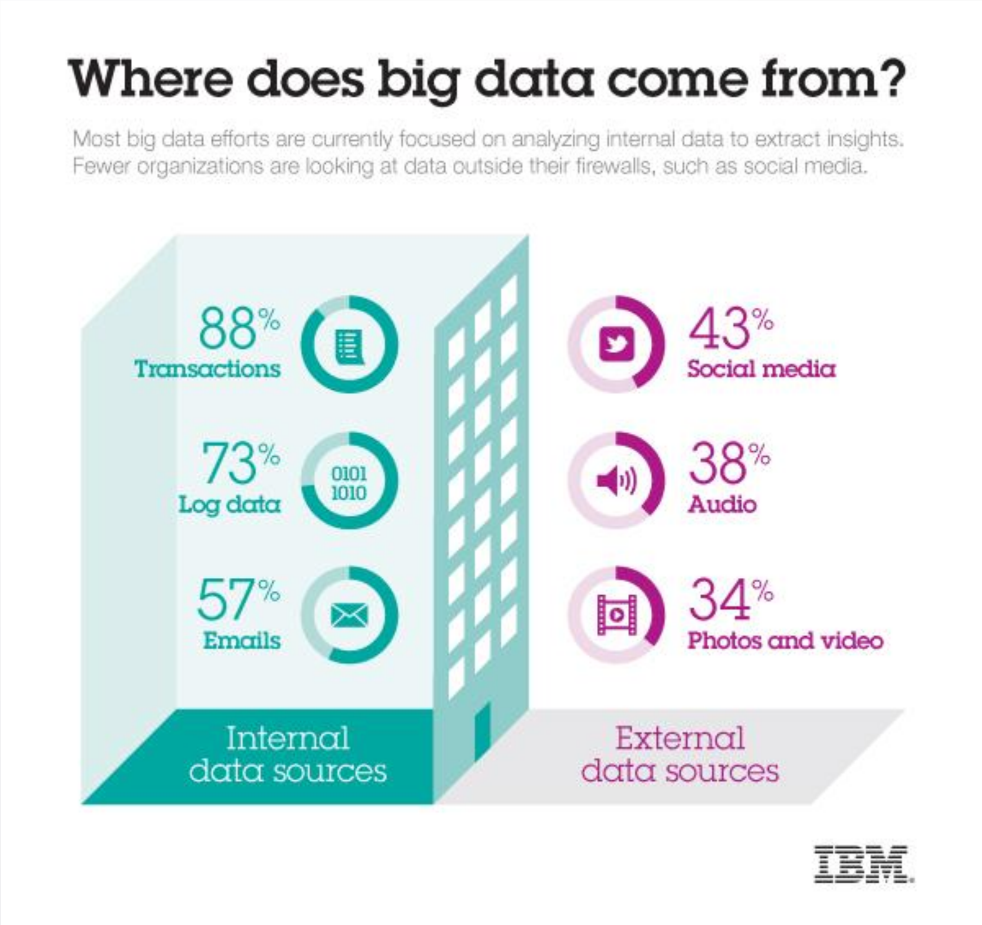
VARIETY

the types of files and formats of data as well as sources.

Data comes in all types of formats, and in terms of analysis can be grouped into two streams. Structured: the numeric data in traditional databases. Information created from line-of-business applications and pre-formatted data collected over time. Unstructured: the relational and seemingly unrelated data that comes from unstructured sources, such as social media or data such as text documents, email, video, audio, etc.

Sources of big data

- Internet of things
- Research
- Economics
- Social Media
- Telecommunication



Where does big data come from?

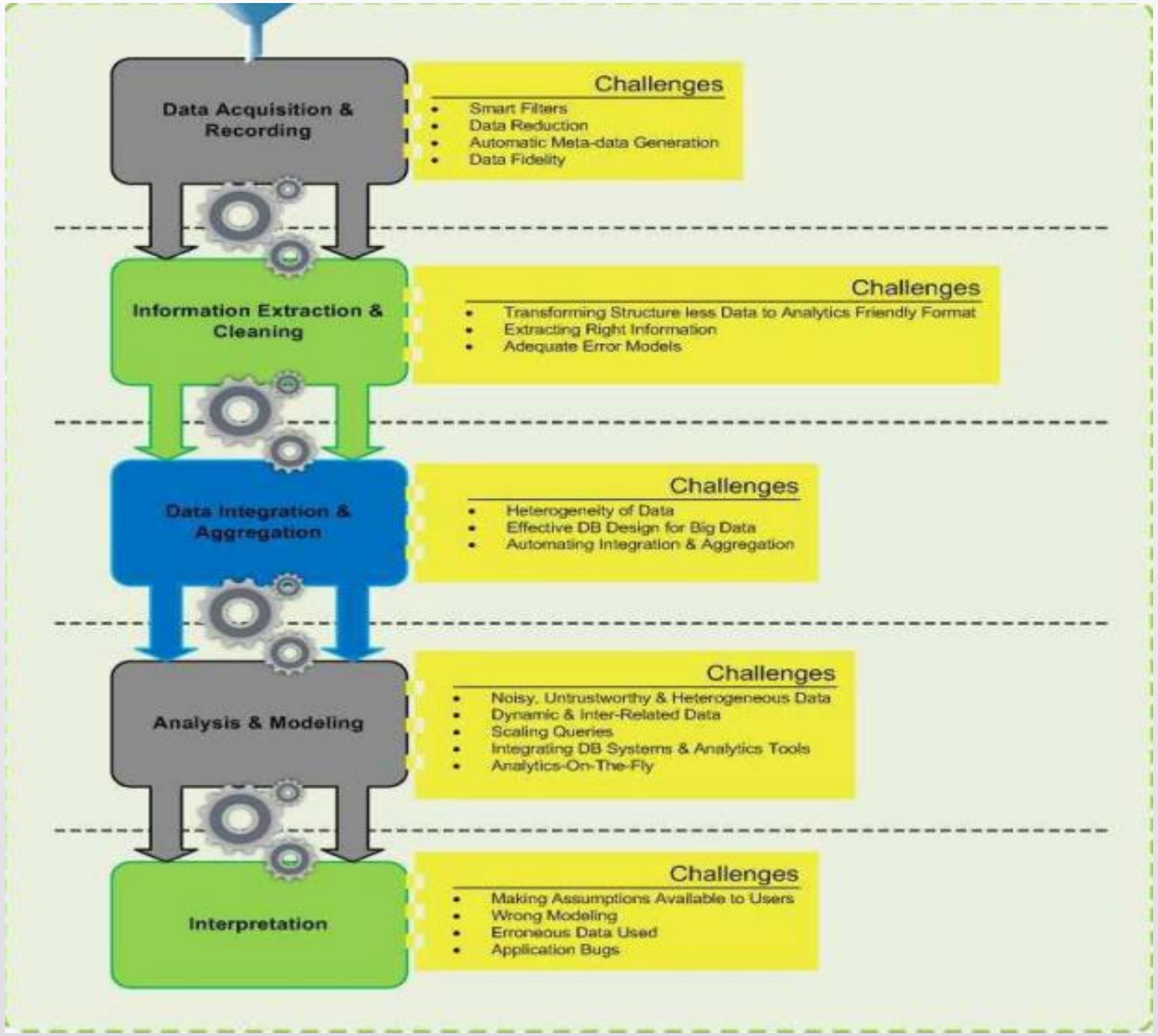
Most big data efforts are currently focused on analyzing internal data to extract insights. Fewer organizations are looking at data outside their firewalls, such as social media.

Challenges in Big data

- In General:
  - Volume, Variety and Velocity
- Data Transfer
  - Conventional methods of data transfer do not work efficiently
  - Networks not designed to cater for such massive volumes
  - One solution: Process at the “edge”

Challenges in Big data

- Data Processing
  - Quality and relevance are important characteristics
  - General Statistical analysis tools not applicable
  - OLAP Tools (On-line Analytical Processing Tools), HPC and Machine Learning



Data Acquisition & Recording

Challenges
- Smart Filters
- Data Reduction
- Automatic Meta-data Generation
- Data Fidelity

Information Extraction & Cleaning

Challenges
- Transforming Structure less Data to Analytics Friendly Format
- Extracting Right Information
- Adequate Error Models

Data Integration & Aggregation

Challenges
- Heterogeneity of Data
- Effective DB Design for Big Data
- Automating Integration & Aggregation

Analysis & Modeling

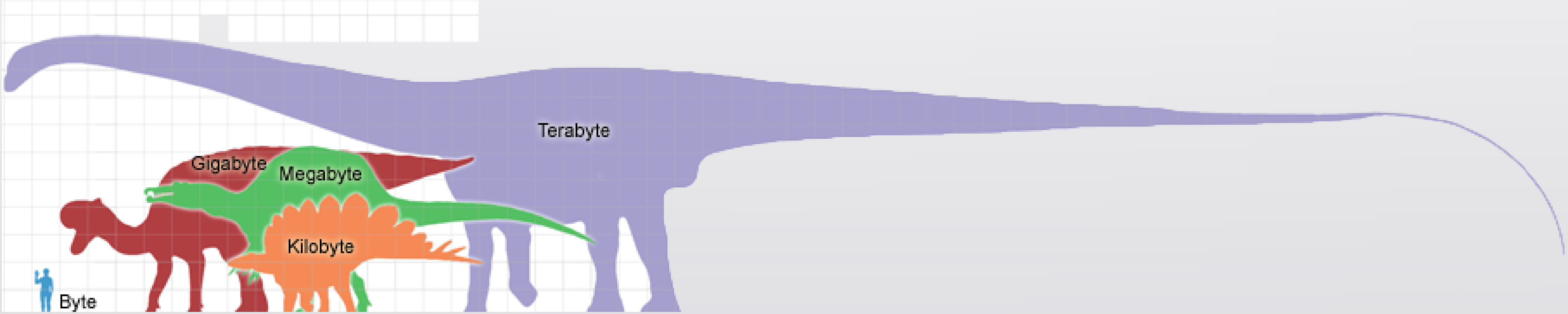
Challenges
- Noisy, Untrustworthy & Heterogeneous Data
- Dynamic & Inter-Related Data
- Scaling Queries
- Integrating DB Systems & Analytics Tools
- Analytics-On-The-Fly

Interpretation

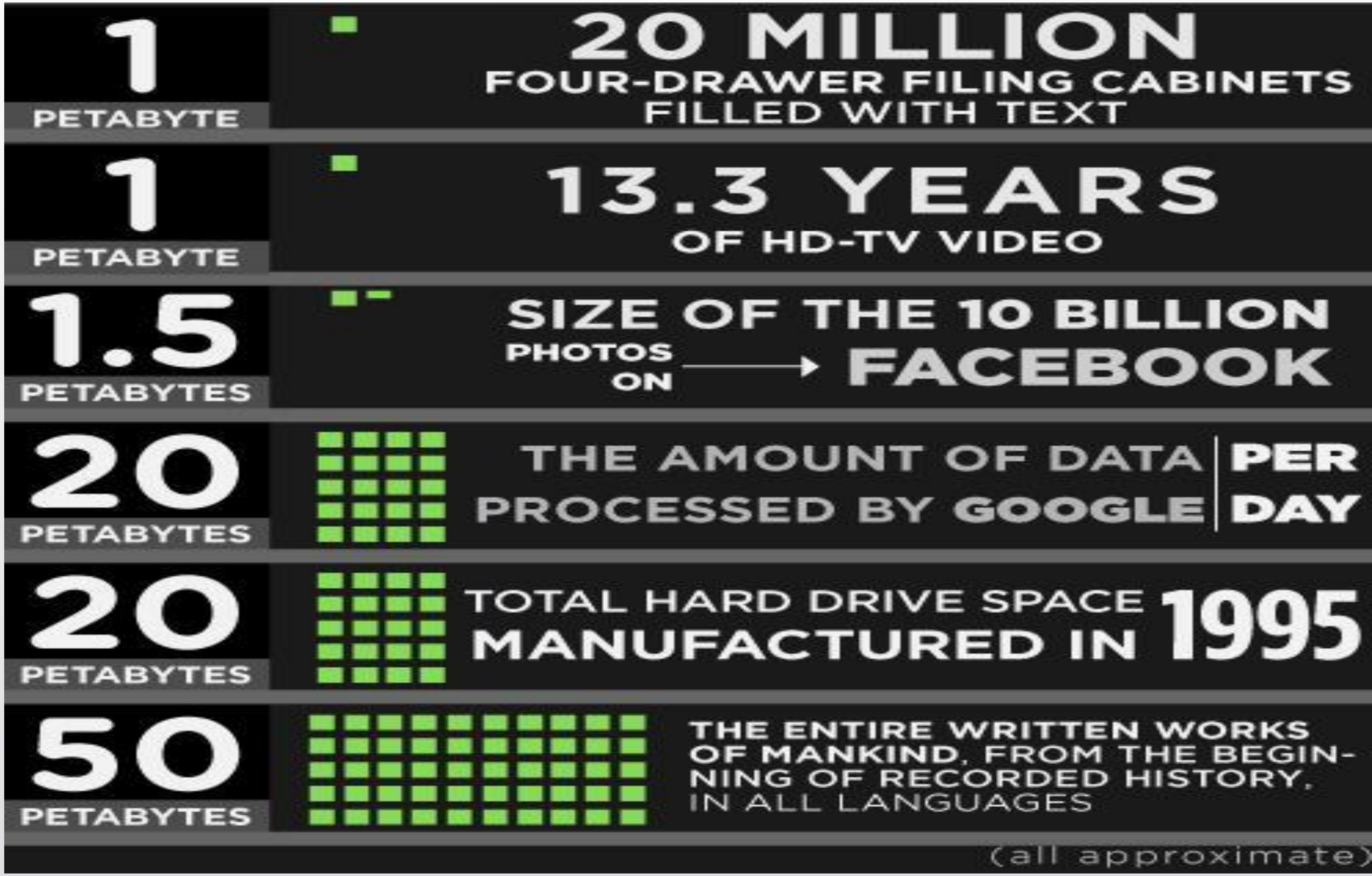
Challenges
- Making Assumptions Available to Users
- Wrong Modeling
- Erroneous Data Used
- Application Bugs

Petabytes of Data

- A single PB is 1,024 TB
- 1 TB = 1,024 GB = 1,048,576 MB = 1,073,741,824 KB = 1,099,511,627,776 B
- …… a lot



A bit of Context

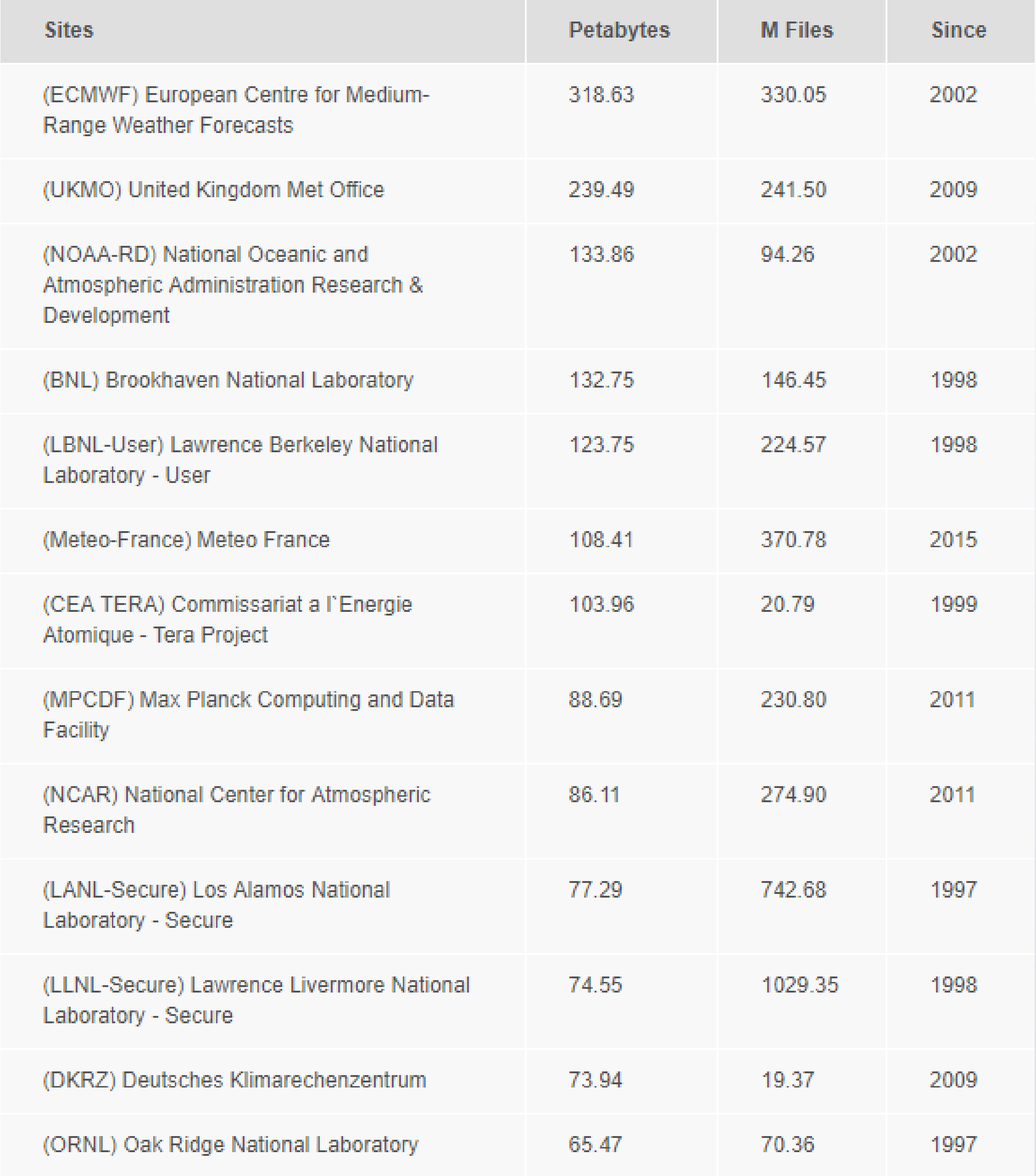


1 PETABYTE	20 MILLION FOUR-DRAWER FILING CABINETS FILLED WITH TEXT
1 PETABYTE	13.3 YEARS OF HD-TV VIDEO
1.5 PETABYTES	SIZE OF THE 10 BILLION PHOTOS ON → FACEBOOK
20 PETABYTES	THE AMOUNT OF DATA PROCESSED BY GOOGLE PER DAY
20 PETABYTES	TOTAL HARD DRIVE SPACE MANUFACTURED IN 1995
50 PETABYTES	THE ENTIRE WRITTEN WORKS OF MANKIND, FROM THE BEGINNING OF RECORDED HISTORY, IN ALL LANGUAGES

(all approximate)

Who Generates so much?

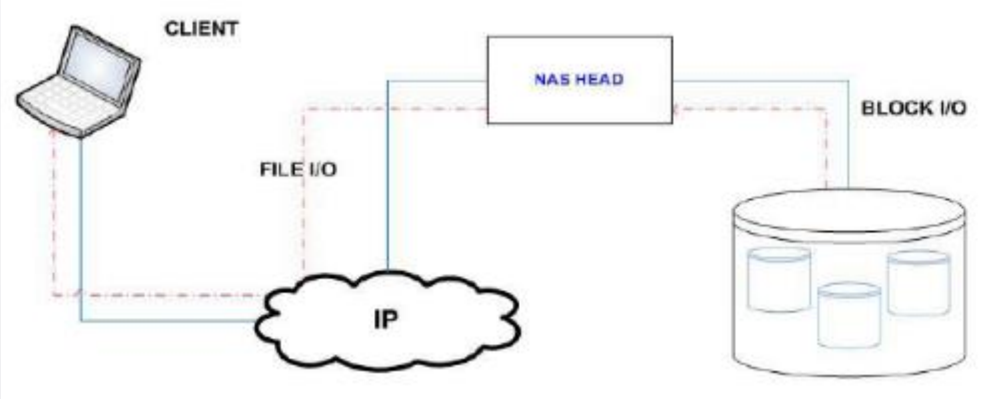
- AT&T transfers about 30 petabytes of data through its networks each day.
- The 2009 movie Avatar is reported to have taken over 1 petabyte of local storage at Weta Digital for the rendering of the 3D CGI effects.



Sites	Petabytes	M Files	Since
(ECMWF) European Centre for Medium-Range Weather Forecasts	318.63	330.05	2002
(UKMO) United Kingdom Met Office	239.49	241.50	2009
(NOAA-RD) National Oceanic and Atmospheric Administration Research & Development	133.86	94.26	2002
(BNL) Brookhaven National Laboratory	132.75	146.45	1998
(LBNL-User) Lawrence Berkeley National Laboratory - User	123.75	224.57	1998
(Meteo-France) Meteo France	108.41	370.78	2015
(CEA TERA) Commissariat a l`Energie Atomique - Tera Project	103.96	20.79	1999
(MPCDF) Max Planck Computing and Data Facility	88.69	230.80	2011
(NCAR) National Center for Atmospheric Research	86.11	274.90	2011
(LANL-Secure) Los Alamos National Laboratory - Secure	77.29	742.68	1997
(LLNL-Secure) Lawrence Livermore National Laboratory - Secure	74.55	1029.35	1998
(DKRZ) Deutsches Klimarechenzentrum	73.94	19.37	2009
(ORNL) Oak Ridge National Laboratory	65.47	70.36	1997

How is it stored?

- Data Centers
- Cloud providers employ a huge number of commodity disk, across several data centers.
- Network attached storage (NAS)

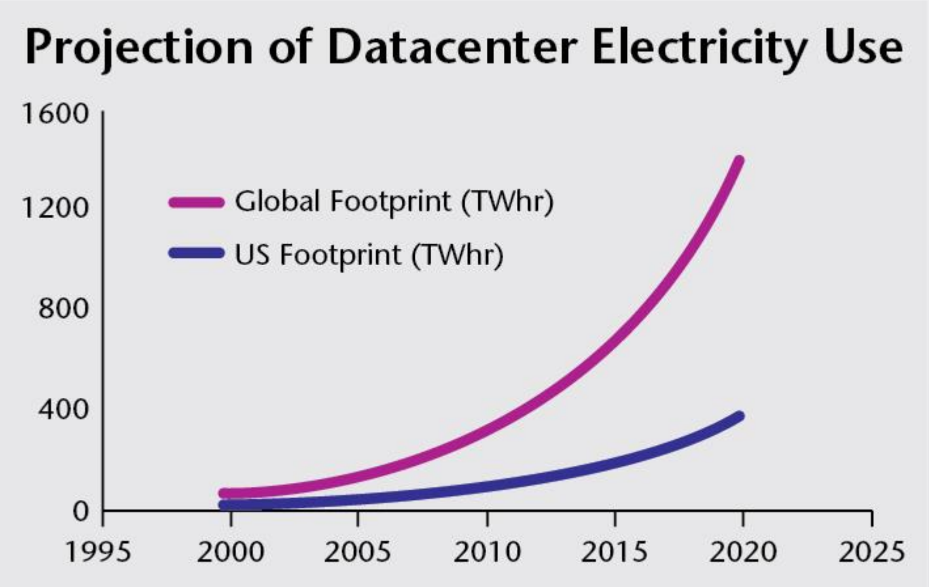


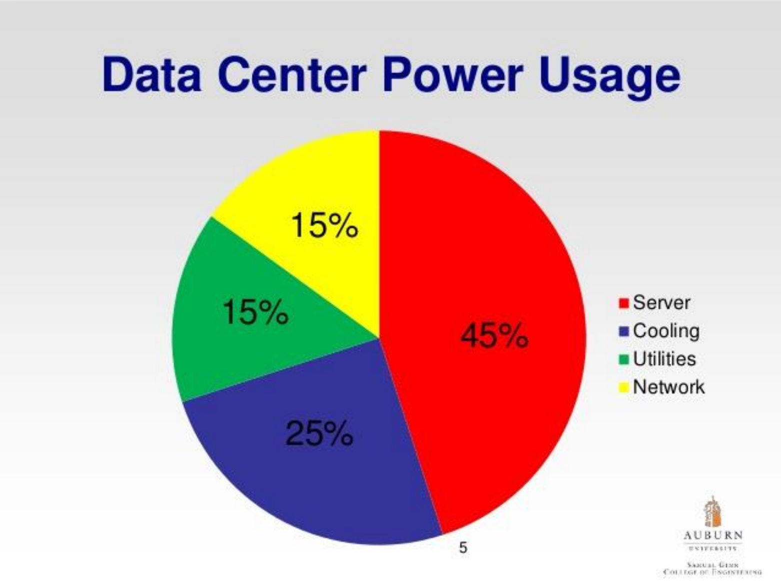
Challenges in Storage

- Must be scalable
- Must provide necessary rate of input/output operations to deliver data to analytics tools
- Data backup and archiving
- Distributed File systems such as HDFS (Hadoop Distributed File systems) and GFS (Google File System) and cloud computing platforms provide solutions

The Sustainability Aspect

- Storage systems and data centers consume huge amounts of power.





The Sustainability Aspect

- Energy efficient distributed file systems e.g. GreenHDFS
- Power-aware storage management algorithms
- Energy aware scheduling

Thank you for your attention!