



# BIG DATA CONCEPTS AND TOOLS

PERFORMED BY: BONDAREV PETR, 433

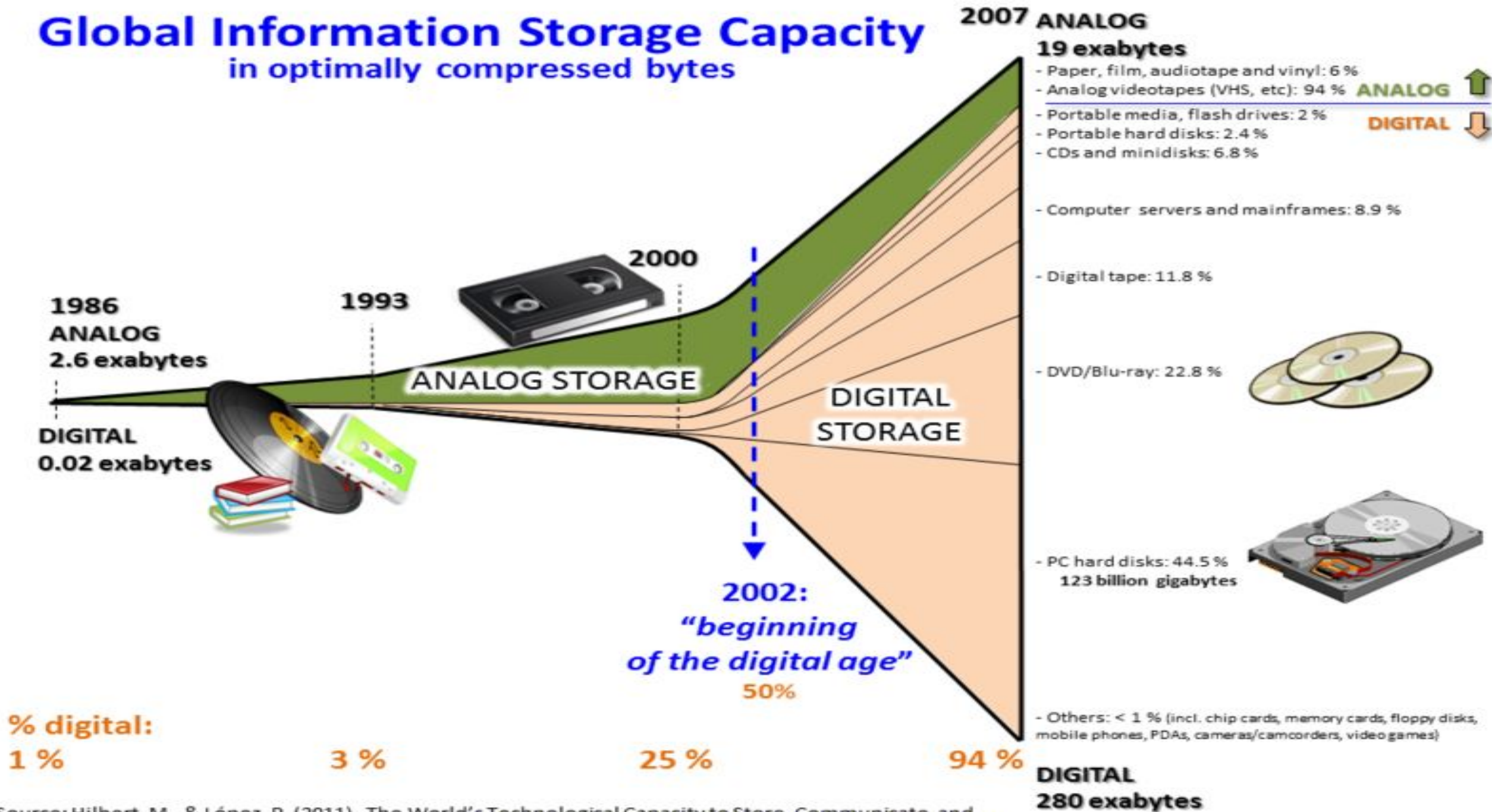
The term "Big Data" has launched a veritable industry of processes, personnel and technology to support what appears to be an exploding new field. Giant companies like Amazon and Wal-Mart as well as bodies such as the U.S. government and NASA are using Big Data to meet their business and/or strategic objectives. Big Data can also play a role for small or medium-sized companies and organizations that recognize the possibilities (which can be incredibly diverse) to capitalize upon the gains.

The collage consists of several overlapping screenshots of news websites from February 2011. The primary headline across all sites is 'Global data storage calculated at 295 exabytes'. The websites include:

- BBC News:** Headline 'Global data storage calculated at 295 exabytes' by Jon Stewart, dated 11 February 2011.
- WELT ONLINE:** Headline 'Die Menschheit erstickt an ihren Daten' (Humanity suffocates on its data).
- NOW news:** Headline '人類資訊量 疊高可達月球' (Human information volume stacks up to the moon).
- Correio Braziliense:** Headline 'Necessidade de espaço para armazenar dados digitais é cada vez maior' (Need for space to store digital data is growing).
- la Repubblica.it:** Headline 'La memoria dell'umanità è alta più di 500mila km' (Humanity's memory is higher than 500,000 km).
- Público.es:** Headline '162 bibliotecas de Alejandría por cada ser humano' (162 libraries of Alexandria for every human).
- Scientific American:** Headline 'Speicherkapazität im Exabyte' (Storage capacity in the exabyte).




# Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

# WHY ARE BIG DATA SYSTEMS DIFFERENT?



An exact definition of "big data" is difficult to nail down because projects, vendors, practitioners, and business professionals view it quite differently. With that in mind, generally speaking, **big data** is:

- large datasets
- the category of computing strategies and technologies that are used to handle large datasets

# WHY ARE BIG DATA SYSTEMS DIFFERENT?

The basic requirements for working with big data are the same as the requirements for working with datasets of any size. However, the massive scale, the speed of ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions. The goal of most big data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.

In 2001, Gartner's Doug Laney first presented what became known as the "three Vs of big data" to describe some of the characteristics that make big data different from other data processing:

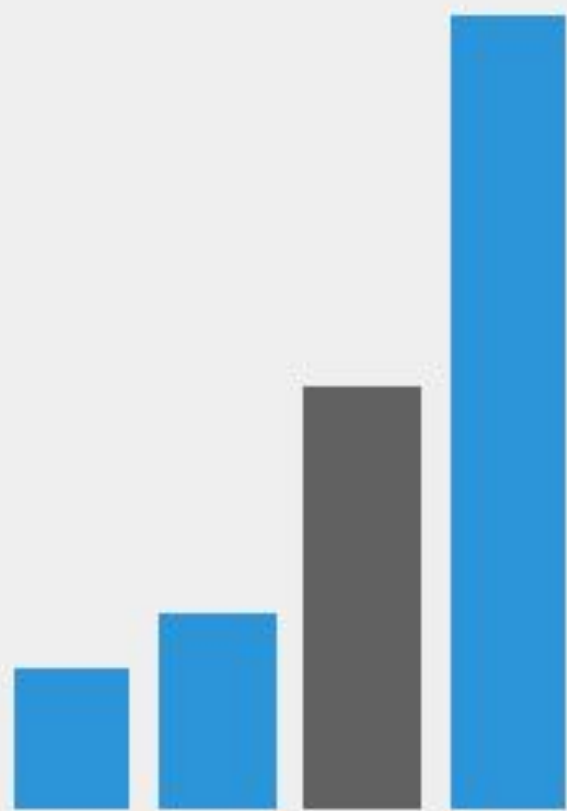


**BIG  
DATA**

**Douglas Laney**







Volume



Velocity



Variety

# OTHER CHARACTERISTICS

- **Veracity:** The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data (and consequently, the quality of the resulting analysis)
- **Variability:** Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process, or filter low quality data to make it more useful.
- **Value:** The ultimate challenge of big data is delivering value. Sometimes, the systems and processes in place are complex enough that using the data and extracting actual value can become difficult.

## 40 ZETTABYTES

[ 43 TRILLION GIGABYTES ]

of data will be created by 2020, an increase of 300 times from 2005



## Volume SCALE OF DATA

### It's estimated that 2.5 QUINTILLION BYTES

[ 2.3 TRILLION GIGABYTES ]  
of data are created each day

Most companies in the U.S. have at least  
**100 TERABYTES**  
[ 100,000 GIGABYTES ]  
of data stored

The New York Stock Exchange captures  
**1 TB OF TRADE INFORMATION**  
during each trading session



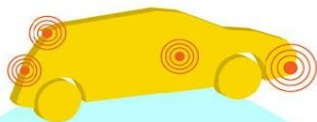
By 2016, it is projected there will be  
**18.9 BILLION NETWORK CONNECTIONS**

— almost 2.5 connections per person on earth



## Velocity ANALYSIS OF STREAMING DATA

Modern cars have close to  
**100 SENSORS**  
that monitor items such as fuel level and tire pressure



# The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015  
**4.4 MILLION IT JOBS**  
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**  
[ 161 BILLION GIGABYTES ]



**30 BILLION  
PIECES OF CONTENT**  
are shared on Facebook every month



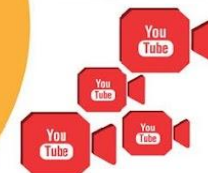
## Variety DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be

**420 MILLION  
WEARABLE, WIRELESS  
HEALTH MONITORS**



**4 BILLION+  
HOURS OF VIDEO**  
are watched on YouTube each month



**400 MILLION TWEETS**  
are sent per day by about 200 million monthly active users



**1 IN 3 BUSINESS  
LEADERS**

don't trust the information they use to make decisions



Poor data quality costs the US economy around

**\$3.1 TRILLION A YEAR**



in one survey were unsure of how much of their data was inaccurate

## Veracity UNCERTAINTY OF DATA



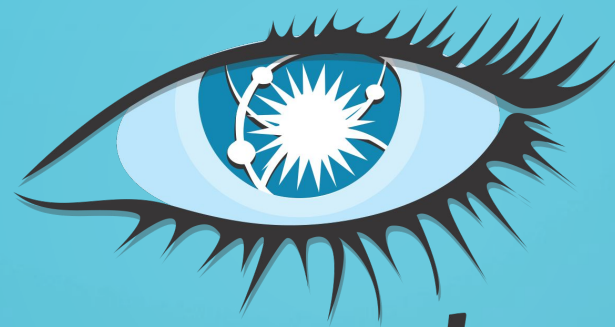
# TOOLS

There are thousands of Big Data tools out there for data analysis today. Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making.





Great product from Apache that has been used by many large corporations. Among the most important features of this advanced software library is superior processing of voluminous data sets in clusters of computers using effective programming models. Corporations choose Hadoop because of its great processing capabilities plus developer provides regular updates and improvements to the product.



# *cassandra*

This tool is widely used today because it provides an effective management of large amounts of data. It is a database that offers high availability and scalability without compromising the performance of commodity hardware and cloud infrastructure. Among the main advantages of Cassandra highlighted by the development are fault tolerance, performance, decentralization, professional support, durability, elasticity, and scalability. Indeed, such users of Cassandra as eBay and Netflix may prove them.

**eBay**

**NETFLIX**





This tool makes the list because of its superior streaming data processing capabilities in real time. It also integrates with many other tools such as Apache Slider to manage and secure the data. The use cases of Storm include data monetization, real time customer management, cyber security analytics, operational dashboards, and threat detection. These functions provide awesome business opportunities.



The HPCC platform combines a range of big data analysis tools. It is a package solution with tools for data profiling, cleansing, job scheduling and automation. Like Hadoop, it also leverages commodity computing clusters to provide high-performance, parallel data processing for big data applications.

It uses ECL (a language specially designed to work with big data) as the scripting language for ETL engine. The HPCC platform supports both parallel batch data processing (Thor) and real-time query applications using indexed data files (Roxie).



# elasticsearch

Elasticsearch is a dependable and safe open source platform where you can take any data from any source, in any format and search, analyze it and envision it in real time. Elasticsearch is designed for horizontal scalability, reliability, and ease of management. All of this achieved while combining the speed of search with the potential of analytics. It is based on Lucene a retrieval software library originally compiled in Java. It uses a developer-friendly, JSON-style, query language that works well for structured, unstructured and time-series data.





THANKS FOR YOUR ATTENTION!

# SOURCES

- <https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology>
- <https://www.techrepublic.com/blog/big-data-analytics/big-data-basic-concepts-and-benefits-explained/>
- <https://bluewavebuzzblog.wordpress.com/2014/03/20/what-is-big-data-and-how-can-it-help-marketers/>
- <http://bigdata-madesimple.com/top-10-tools-for-working-with-big-data-for-successful-analytics-developers-2/>
- <https://www.newgenapps.com/blog/top-best-open-source-big-data-tools>