Co-funded by the
Erasmus+ Programme
of the European Union

Master classes for Partner Universities
December 10, 2021

BIG DATA TECHNOLOGIES

Dr. Olha HALAN

Helga.halan@gmail.com

AGENDA

- **CONCEPT and PRINCIPLES** of **«BIG DATA»**
- Worldwide amount of the data
- The “3V“ Big Data principles
- Methods and technics of big data analysis
- **TECHNICAL TOOLS TO WORK WITH BIG DATA**
- NoSQL, MapReduce, Ecosystem Hadoop, Spark
- **CLOUD COMPUTING**
- **MACHINE LEARNING**
- **HARDWARE SOLUTIONS**
- How to work with Big Data
- Data-driven business

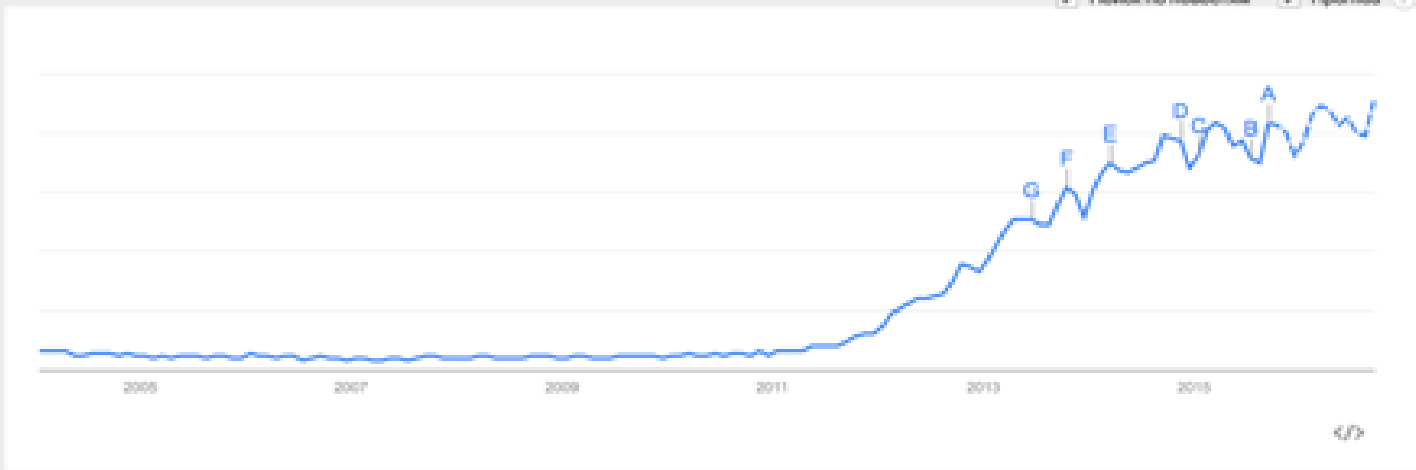
PART 1
«BIG DATA» DEFINITION

big data
Поисковый запрос

+Добавить запрос

Динамика популярности

Поиск по новостям Прогноз



Usually there are several definitions:

- Big Data – when data more than 100 Gb (500Gb, 1Tb, as you like)
- Big Data – it's data, which impossible to work with Excel
- Big Data – it's data, which is impossible to work with on one computer
- Big Data – it's any data
- Big Data is not present, it was established by marketers

The Wikipedia gives the definition of Big Data:

Big data - a series of approaches, tools and methods of processing structured and unstructured data of huge volumes and considerable variety to obtain human-perceived results, effective in continuous growth, distribution to numerous nodes of the computer network formed in the late 2000s, alternative to traditional database management systems and Business Intelligence class solutions.

THE WORLD AMOUNT OF THE DATA

- 2003 - 5 exabyte of the data (1 Eb = 1 mlrd Gb)
- 2008 - 0,18 zetabyte (1 Zb = 1024 Eb)
- 2015 - more than 6,5 Zb
- 2018 - 33 Zb
- 2020 - 40-45 Zb
- 2025 - 160-175 Zb (as expected)

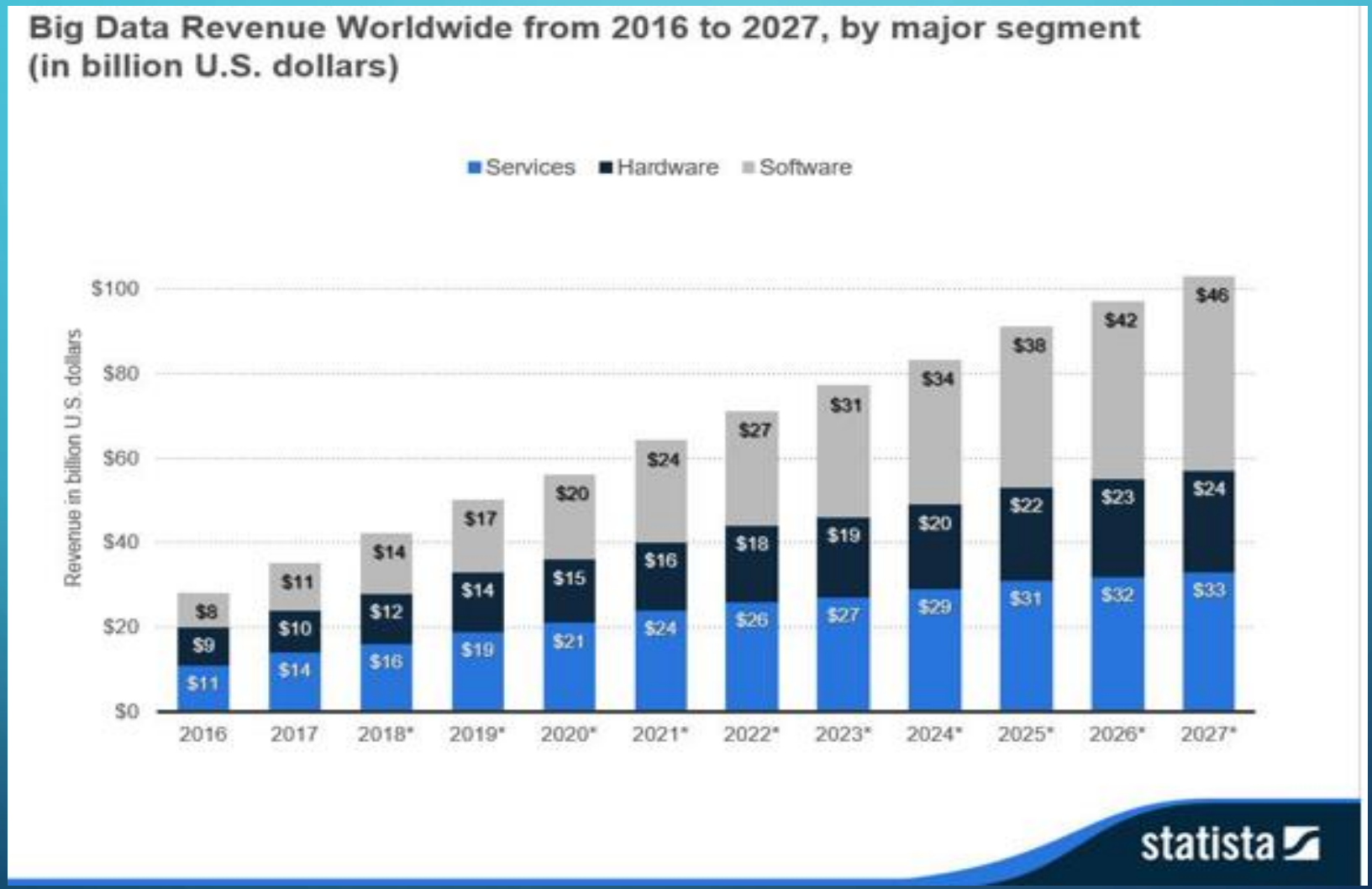
EXAMPLES OF DATA RESOURCES TO WORK WITH BIG DATA

THE MAIN ADVANTAGES OF «BIG DATA» USING:

- obtaining qualitatively new knowledge;
- expanding functionality;
- increase efficiency;
- ensuring a minimum cost;

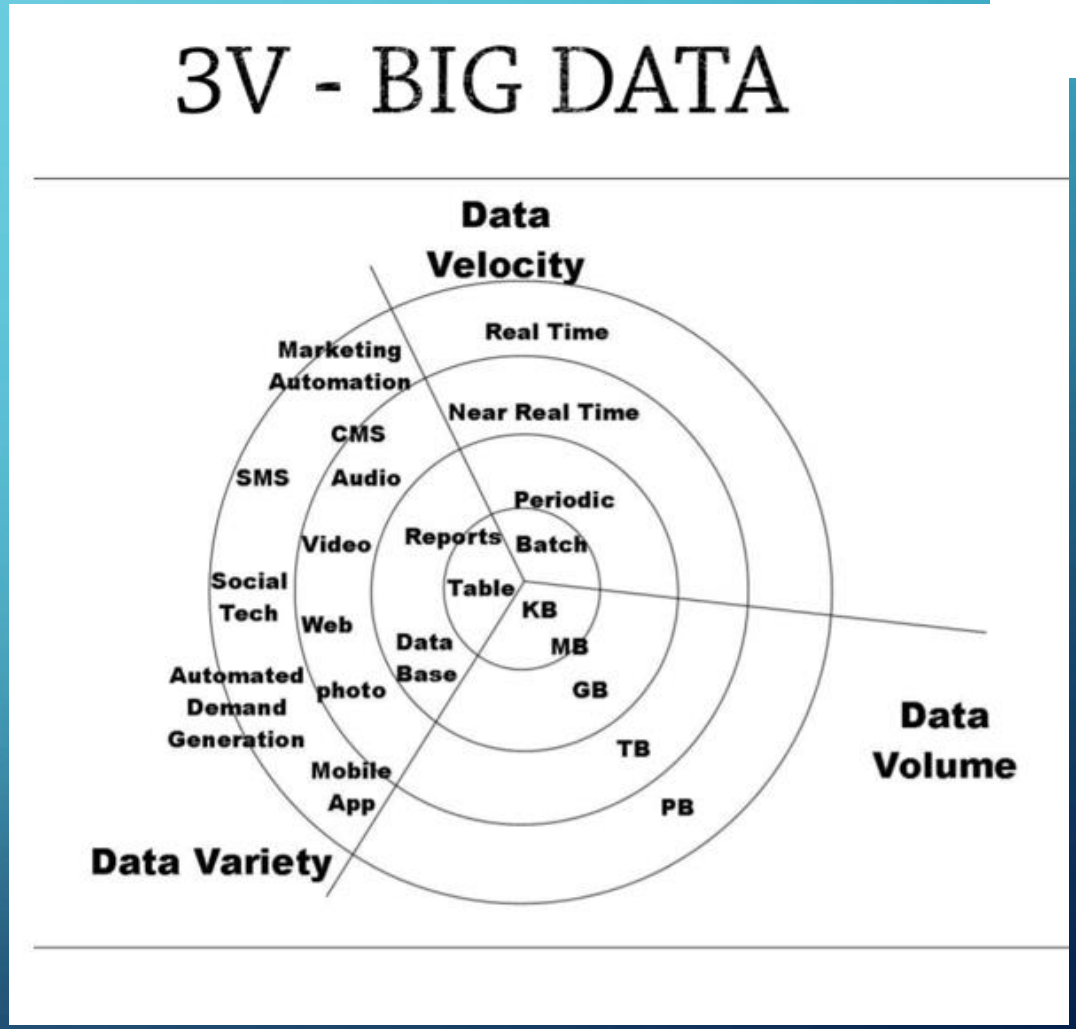
Possibilities of optimization of many spheres: public administration, medicine, telecommunications, finance, transport, production etc.

TRENDS AND PROSPECTS



THREE "V" AND WORKING PRINCIPLES WITH BIG DATA

- volume, physical volume
- velocity, the rate of data growth and the need for rapid processing
- variety - the ability to simultaneously process different types of data.



1. DATA VOLUME

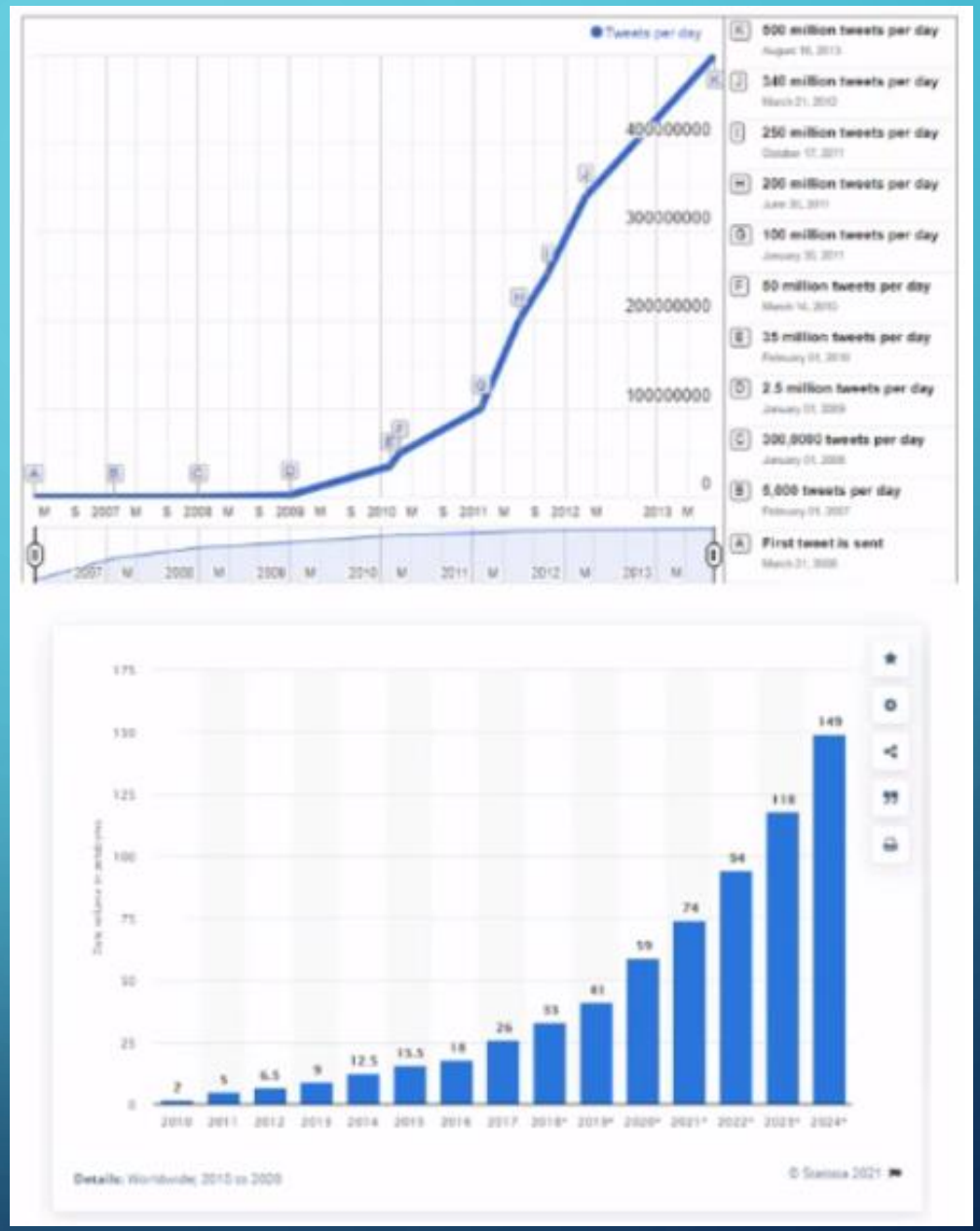
- Grows exponentially
- From 2 ZB in 2010

Till 59 in 2020

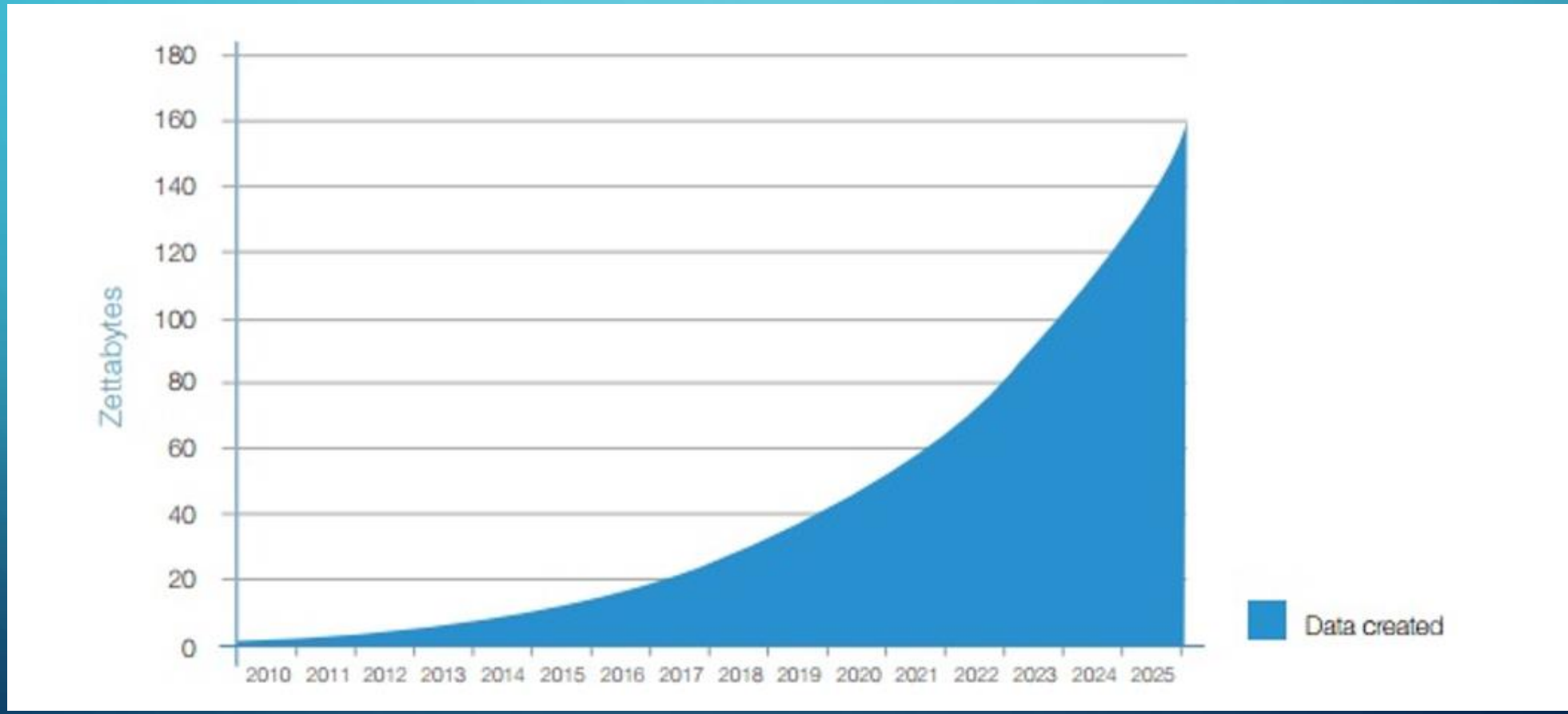
Doubling ~ each 2 years

($\log_2(29,5)=4,883$);

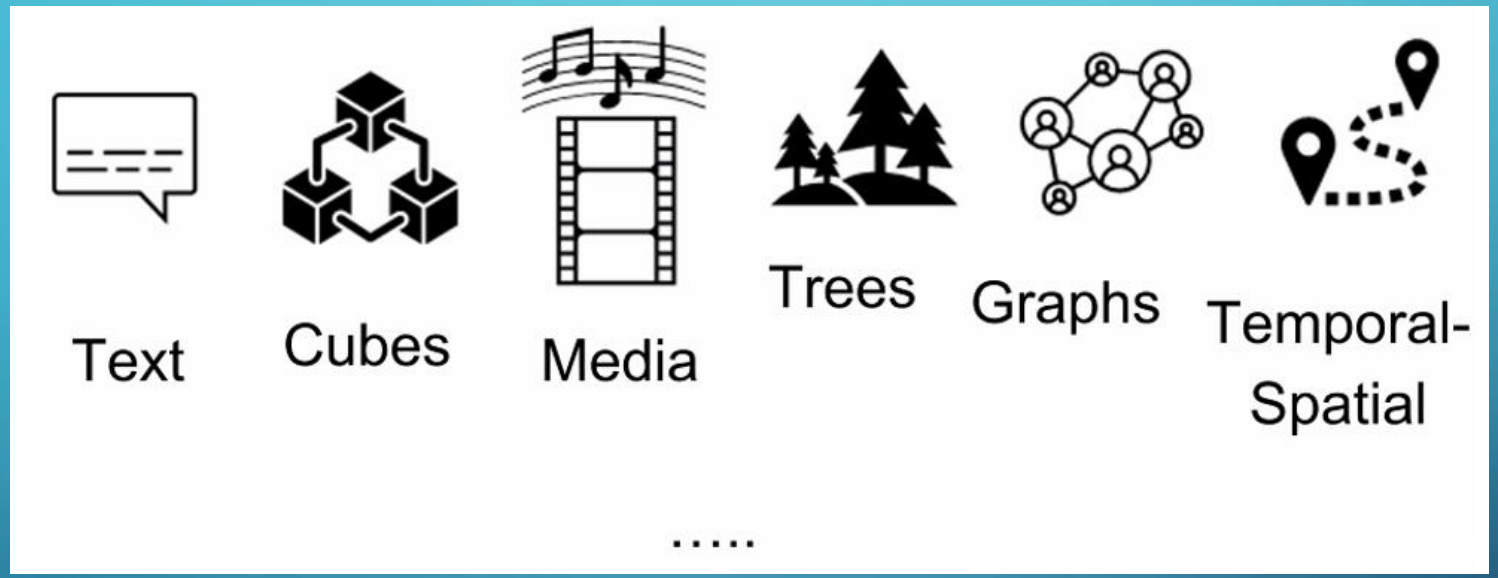
10p./4,883=~2p.



THE VOLUME OF DATA HAS BEEN GROWING RAPIDLY OVER THE LAST FEW YEARS AND PROMISES T
CONTINUE.
ONE OF THE MANY GRAPHS SHOWING THIS RAPID GROWTH :



2. DATA VARIETY



3. DATA VELOCITY

- They are quickly generated and require fast processing

https://www.internetlivestats.com/

- Delayed response, processing leads to failures:

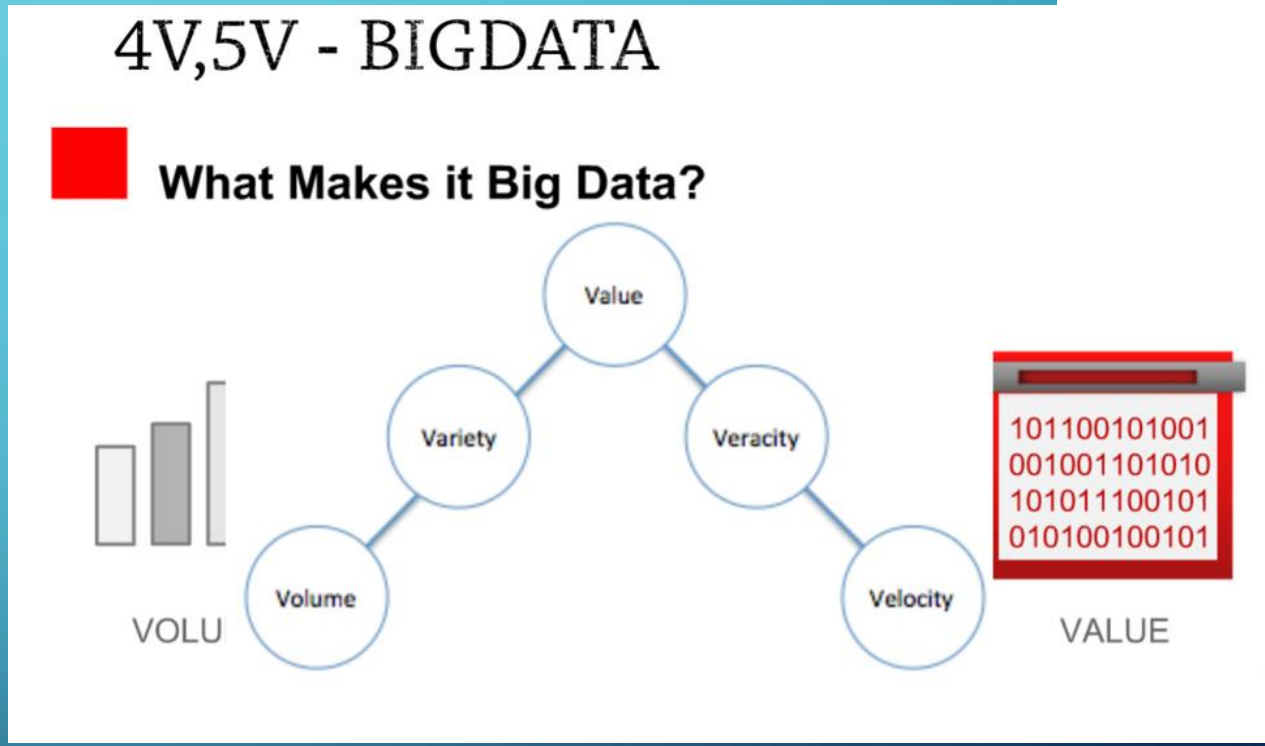
- Messenger
- Mobile marketing – location
- Medical support, telemedicine

4V (veracity)

5V (viability and value),

7V (variability and visualization).

- The IDC Company interprets the forth V as value



THE METHODS AND TECHNICS OF BIG DATA ANALISYS

McKinsey, an international consulting firm, provides 11 methods and techniques of analysis that can be applied to big data:

1. Methods Data Mining;
2. Crowdsourcing;
3. Data fusion and integration;
4. Machine Learning, Deep Learning;
5. Artificial neural networks;
6. Pattern recognition;
7. Forecasting analytics;
8. Simulation;
9. Spatial analysis;
10. Statistical analysis;
11. Visualization of analytical data

According to a report by McKinsey's "Global Institute, Big Data: The Next Frontier for Innovation, Competition, and Productivity", Big Data technologies can be useful in solving the following tasks :

- forecasting the market situation;
- marketing and sales optimization;
- product improvement;
- making managerial decisions;
- increase productivity;
- efficient logistics;
- monitoring the condition of fixed assets.

PART 2. TECHNICAL TOOLS TO WORK WITH BIG DATA TECHNOLOGIES

NoSQL

- Scalability
- Availability
- Atomicity
- Consistency

NoSQL - «Not Only SQL»

Google services:

GMail,

Google Maps,

Google Earth

Technologies:

MapReduce

Hadoop

NoSQL

- Amazon.com, IBM, Facebook, Netflix, EBay, Hulu, Yahoo!

ACID demands:

- atomicity,
- consistency,
- isolation,
- durability.

NoSQL instead of ACID - BASE:

- basic availability;
- soft state;
- eventual consistency.

NOSQL CHARACTERISTICS

- use of different types of storage;
- the ability to develop a database without setting the scheme;
- use of multiprocessor;
- linear scalability;
- innovation;
- reduction of development time;
- speed.

TYPES OF DATA STORAGES

1. **Storage «key- value»**

Berkeley DB,

MemcacheDB,

Redis, Riak,

Amazon DynamoDB.

Key	Value
k1	value1
k2	value2
k3	value3

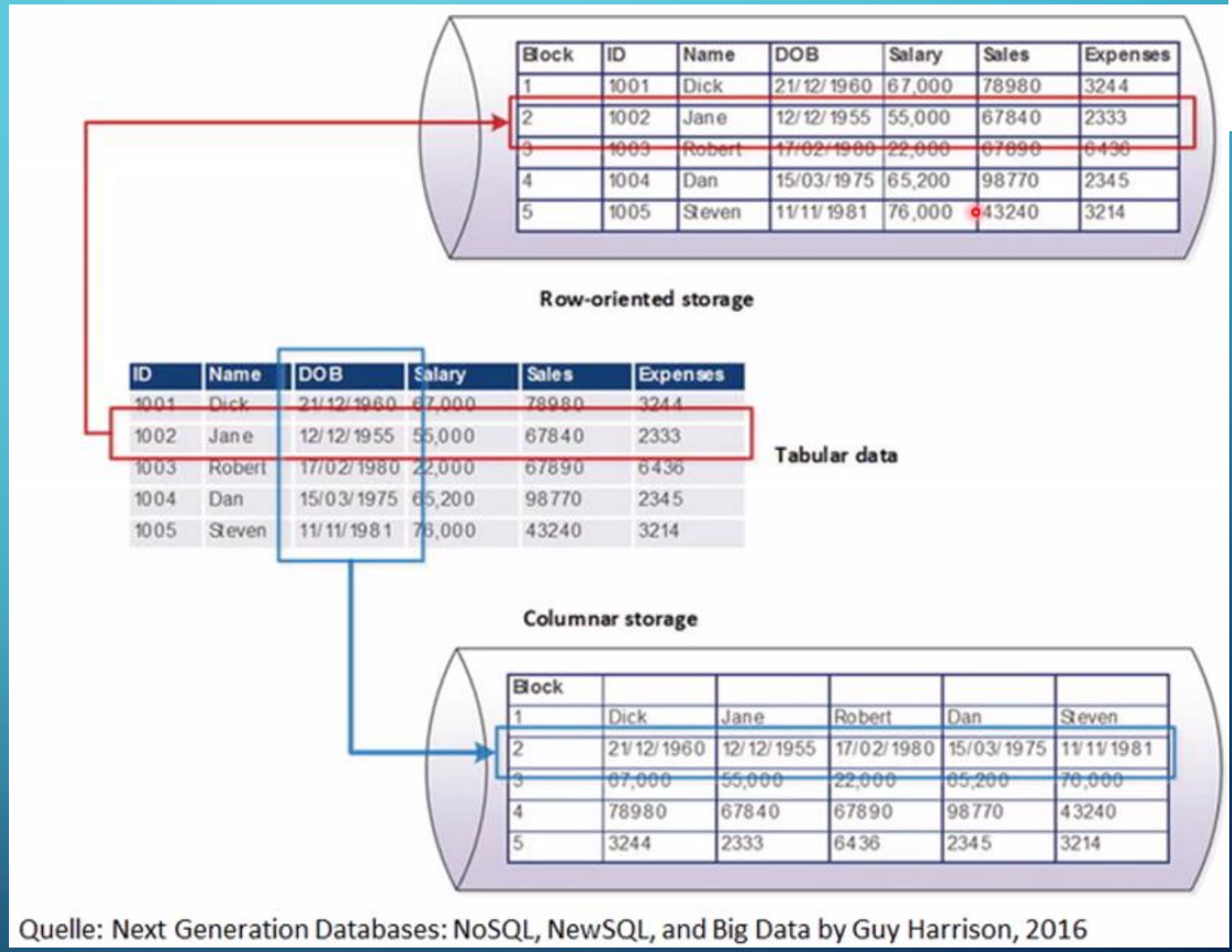
2. Repository of column families

Apache HBase,

Apache Cassandra,

Apache Accumulo,

Hypertable.



Quelle: Next Generation Databases: NoSQL, NewSQL, and Big Data by Guy Harrison, 2016

- **3.** Document-oriented

Format i.a. XML, JSON, BSON, YAML, RD

CouchDB, Couchbase, MarkLogic, MongoDB, eXist, Berkeley DB XML.

```
{
   }, "id": 1, "name":
   "football boot,
   "price": 199,
   "stock": {
      "warehouse": 120,
      "retail": 10
   }
}
```

4. Graph-based databases

Neo4j, AllegroGraph, Bigdata (RDF-storage), InfiniteGraph.

MAP REDUCE

Map Reduce - distributed computing model.

MapReduce assumes that the data is organized in the form of some records.

Data processing takes place in 3 stages :

- Stage **Map**.
- Stage **Shuffle**.
- Stage **Reduce**.

A FEW ADDITIONAL FACTS ABOUT MAP REDUCE:

1) the principle of horizontal scaling;
2) the principle of data locality;
3) full data scan.

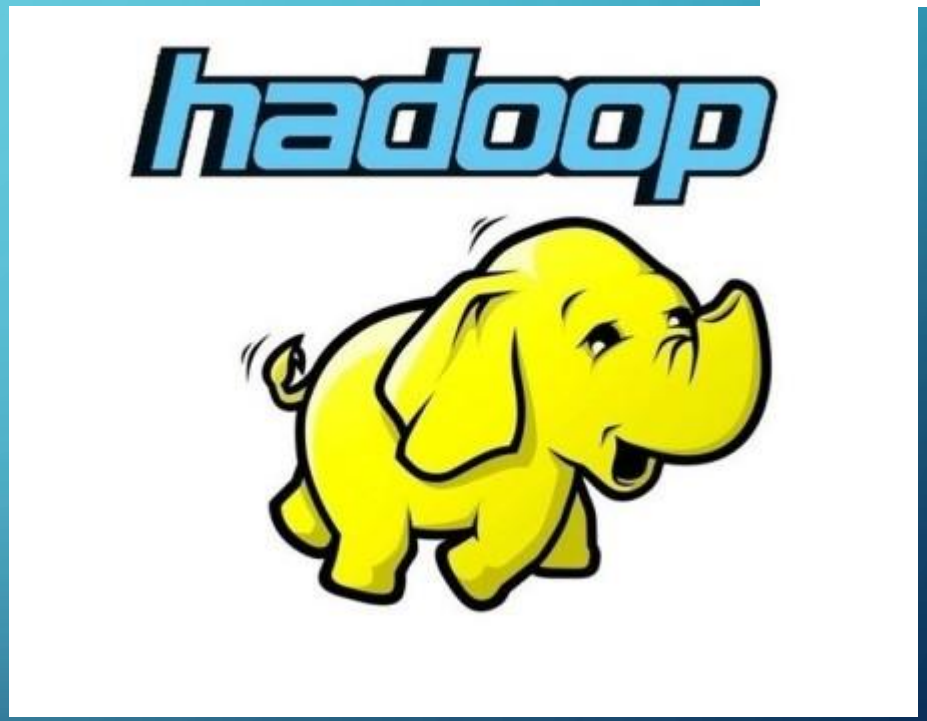
AVAILABLE IMPLEMENTATIONS OF THE MAPREDUCE :

- Google implemented MapReduce on C++ with interface on Python and Java;
- Greenplum - commercial implementation with language support with Python, Perl, SQL and others;
- GridGain - free open source implementation in the language Java;
- Apache Hadoop Project - free implementation MapReduce with open source on Java;
- Phoenix - realization MapReduce on C using allocated memory;
- MapReduce also realized on Cell Broadband Engine in C language;
- MapReduce implemented in Nvidia GPUs using CUDA;
- Qt Concurrent - a simplified version of the framework implemented by Qt in C ++, which is used to distribute a task between several cores of one computer;

- CouchDB uses MapReduce to define views on top of distributed documents;
- MongoDB allows you to use MapReduce to process requests in parallel on multiple servers;
- Skynet - open source implementation in Ruby;
- Disco - an implementation created by Nokia, its core is written in Erlang, and applications for it can be written in Python;
- Apache Hive - open source add-on from Facebook, which allows you to combine Hadoop and access to data in SQL - a similar language;
- Qizmt - an open source implementation from MySpace, written in C #;
- DryadLinq - implementation created by Microsoft Research based on a parallel version of Linq and Microsoft Dryad;
- YAMR (yet another mapreduce) implementation created by Microsoft Research Based on a parallel version of Lynch and Microsoft Dyad

HADOOP

- Hadoop — it is a framework that consists of a set of utilities for developing and executing distributed computing programs.
- Hadoop – powerful tools for working with big data
- Hadoop - Apache Software Foundation project
- Developed in Java within the MapReduce computing paradigm



WHO IS USED?

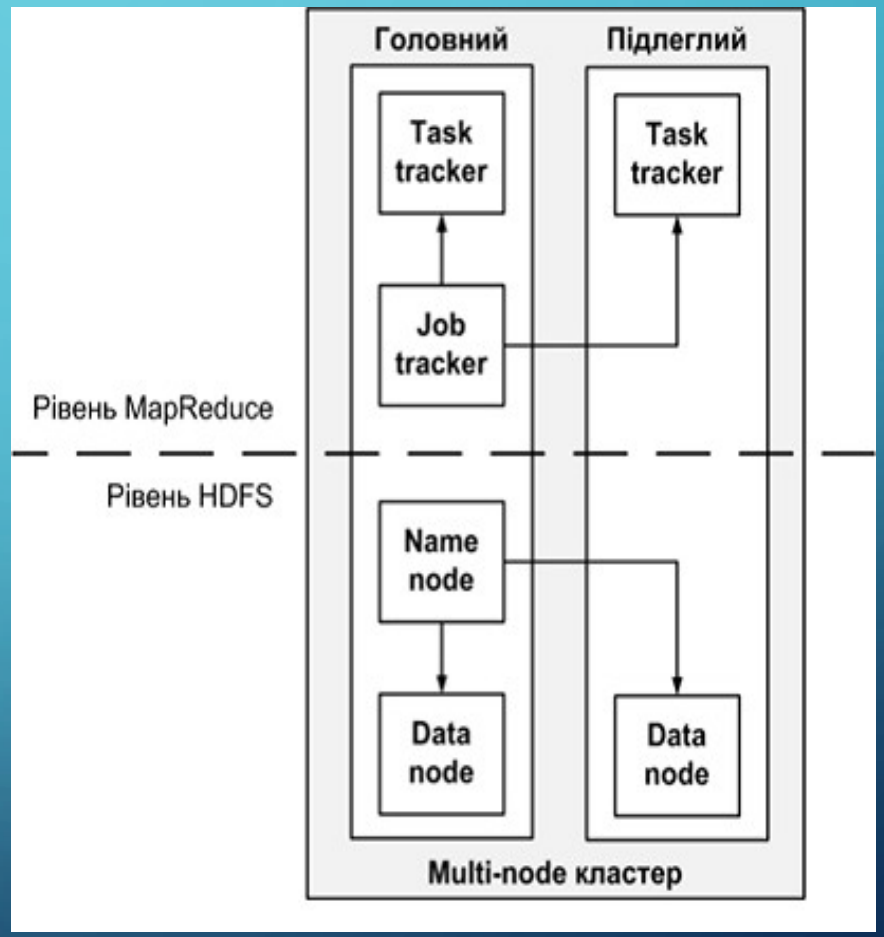
Hadoop - technology to work with BigData.

- eBay, Amazon, IBM, Facebook etc.
- Vendors of distributions : Cloudera, MapR, HDP.
- Hadoop complex and has such a large number of components.

CORE COMPONENTS OF THE HADOOP :

- Hadoop Common,
- HDFS,
- YARN,
- Hadoop MapReduce.

Multi-node Hadoop cluster



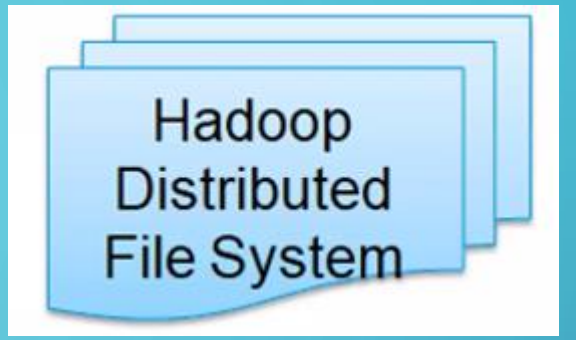
HADOOP COMMON

- hdfs dfs - command URI:

hdfs://example.com/file1 або file:///tmp/local/file2.

- commans: cat, chmod, chown, chgrp, cp, du, ls, mkdir, mv, rm, tail, recursive key -R for chmod, chown, chgrp)
- commands specific to Hadoop (for example, count, expunge, setrep).

HDFS



HDFS (Hadoop Distributed File System):

- All bloks in HDFS (except the last block of the file) have the same size
- Due to replication, the distributed system is resistant to failures of individual nodes.
- Files in HDFS can be written only once (modification is not supported).

YARN

YARN (Yet Another Resource Negotiator - «another resource mediocre»):

- Resource Manager
- як MapReduce - programs and any other distributed applications,
- the ability to perform several different tasks in parallel,
- Application Master.

HADOOP MAPREDUCE

- Hadoop MapReduce - software framework for programming distributed computing within the paradigm MapReduce.
- conversion of key-value source pairs into an intermediate set of key-value pairs, abbreviated set.
- The framework consists of three phases :

- shuffle
- sort
- reduce

3. Fault tolerance

Data replication

Restarting the taskbar

4. Encapsulation of implementation complexity.

- Who studies Hadoop?

- analysts and developers

- employees of banks, IT companies,

- large services with a large customer base.

SCALABILITY

- Yahoo (more than 4,000 nodes with 15 PB bytes each)
- Facebook (appr. 2000 nodes on 21 Pbt) and Ebay (700 nodes on 16 Pbt).
- Hadoop to version 2.0 maximum 4 thousand nodes when using 10 MapReduce - tasks per node.
- NameNode - 100 mln.
- Reduction of computing power– ARM, Intel Atom
- High-performance computing nodes - InfiniBand in Oracle Big Data Appliance, Fiber Channel and Ethernet with a bandwidth of 10 GB / s in FlexPod template configurations for "big data".
- Hadoop (before 2.0) - 10-100 base handlers on the cluster node,

  for tasks that do not require significant CPU time - up to 300.

In YARN uses configuration constants.

4. HIVE, PIG TA HBASE

- Hive - it's an add-on over Hadoop:

  - Hive can use SQL;

  - Hive creates tasks MapReduce;

  - tables in Hive superimposed on data in HDFS.

- Pig - written scripts are secretly transformed into MapReduce tasks that run on the Hadoop cluster.

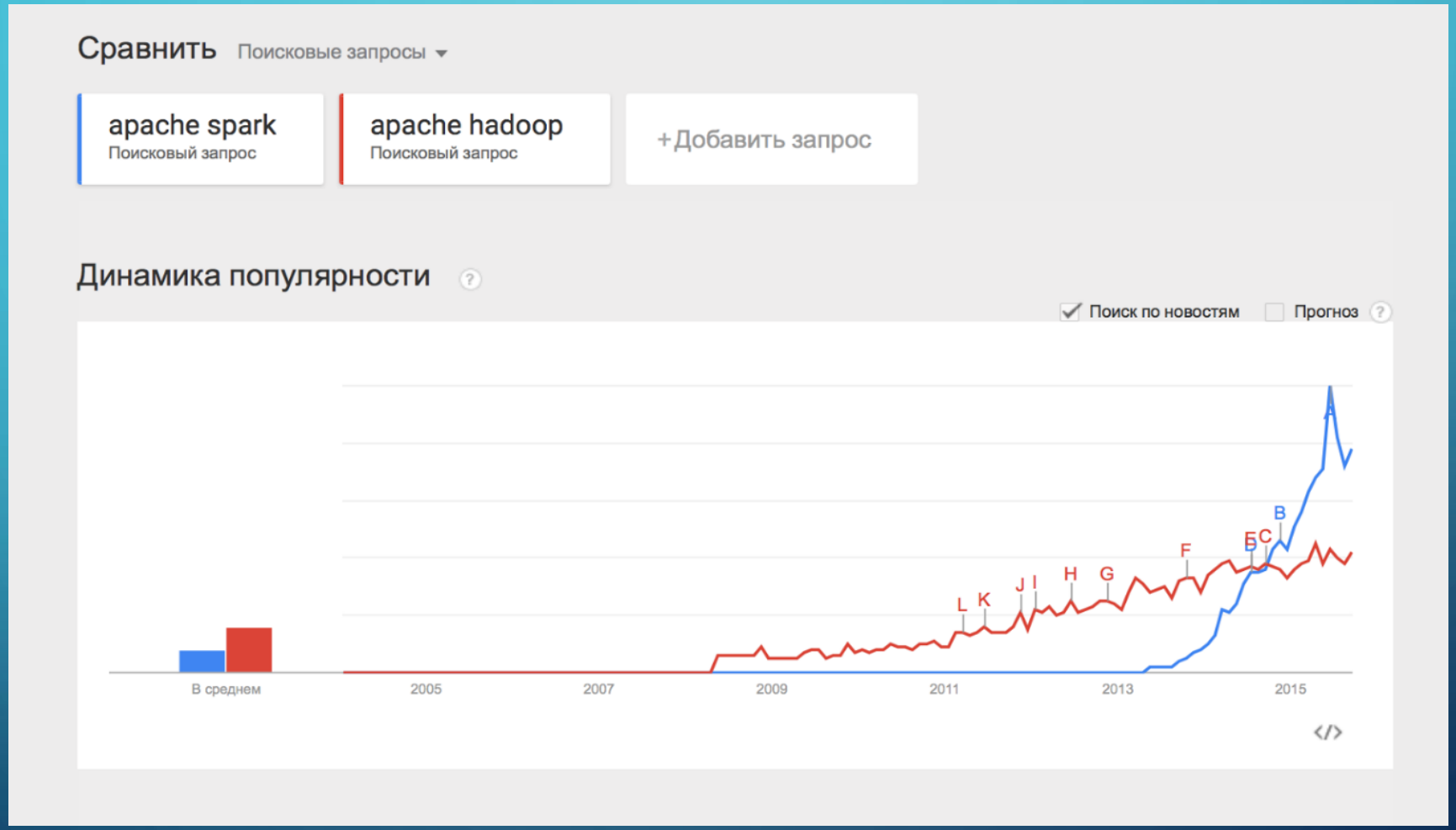
- HBase - this columnar database, located on top of HDFS, has the following limitations:

- search of a number on one key;

- transactions are not supported;

- only one-line operations are available.

Apache Spark

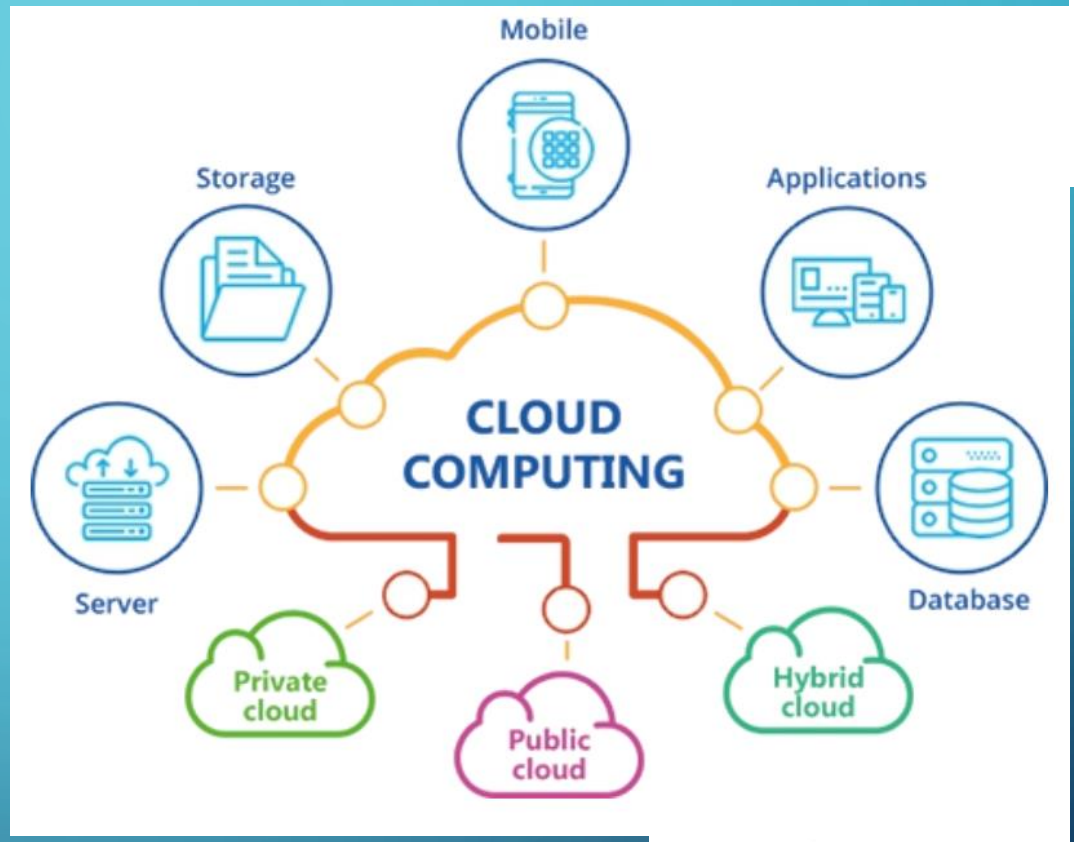


- Amazon, eBay и Yahoo!
- 100 times faster, on the disk - more than 10 times
- API for Scala, Java and Python
- Integrates with Hadoop and data sources (HDFS, Amazon S3, Hive, HBase, Cassandra, etc.)
- Hadoop YARN, Apache Mesos, can work offline

CLOUD COMPUTING

Properties of the cloud model of service use:

- mass ;
- homogeneity ;
- application virtualization;
- stability;
- cheap software;
- geographical unlimited use;
- service orientation;
- advanced security technologies.



CATEGORIES (TYPES) OF "CLOUDS" BY FORM OF OWNERSHIP ARE :

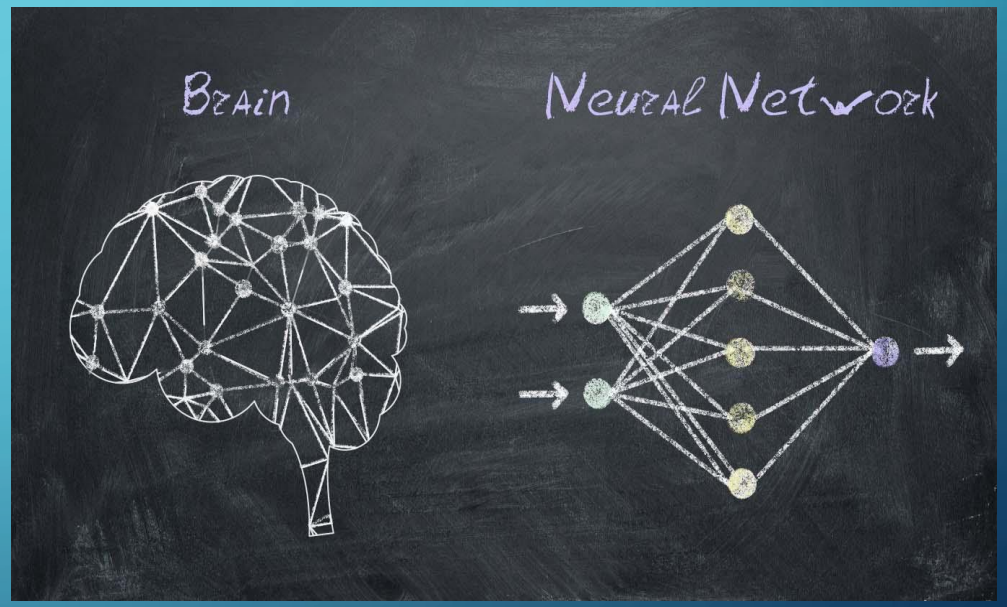
- **public cloud**

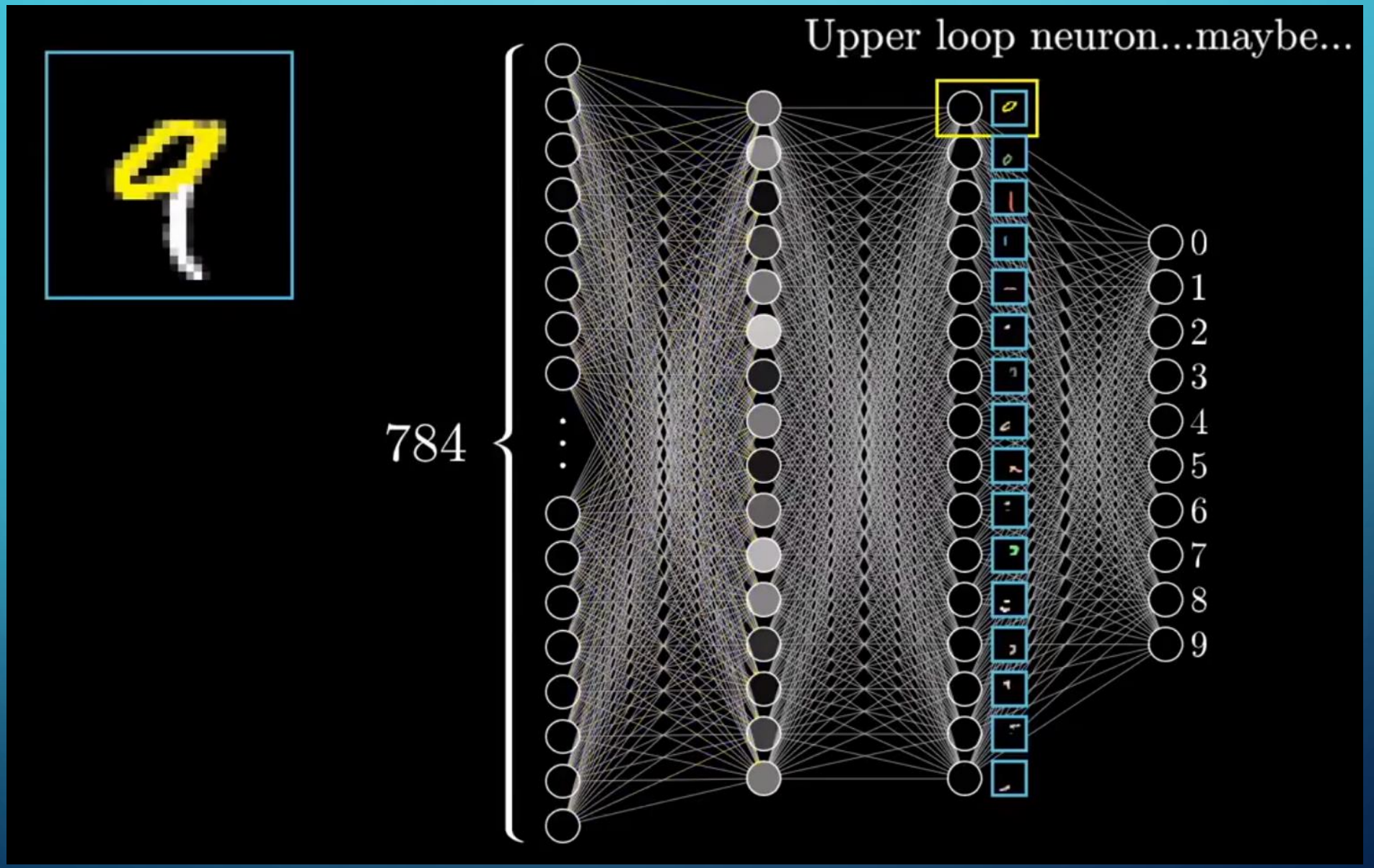
Amazon EC2, Google Apps / Docs, Microsoft Office Web.

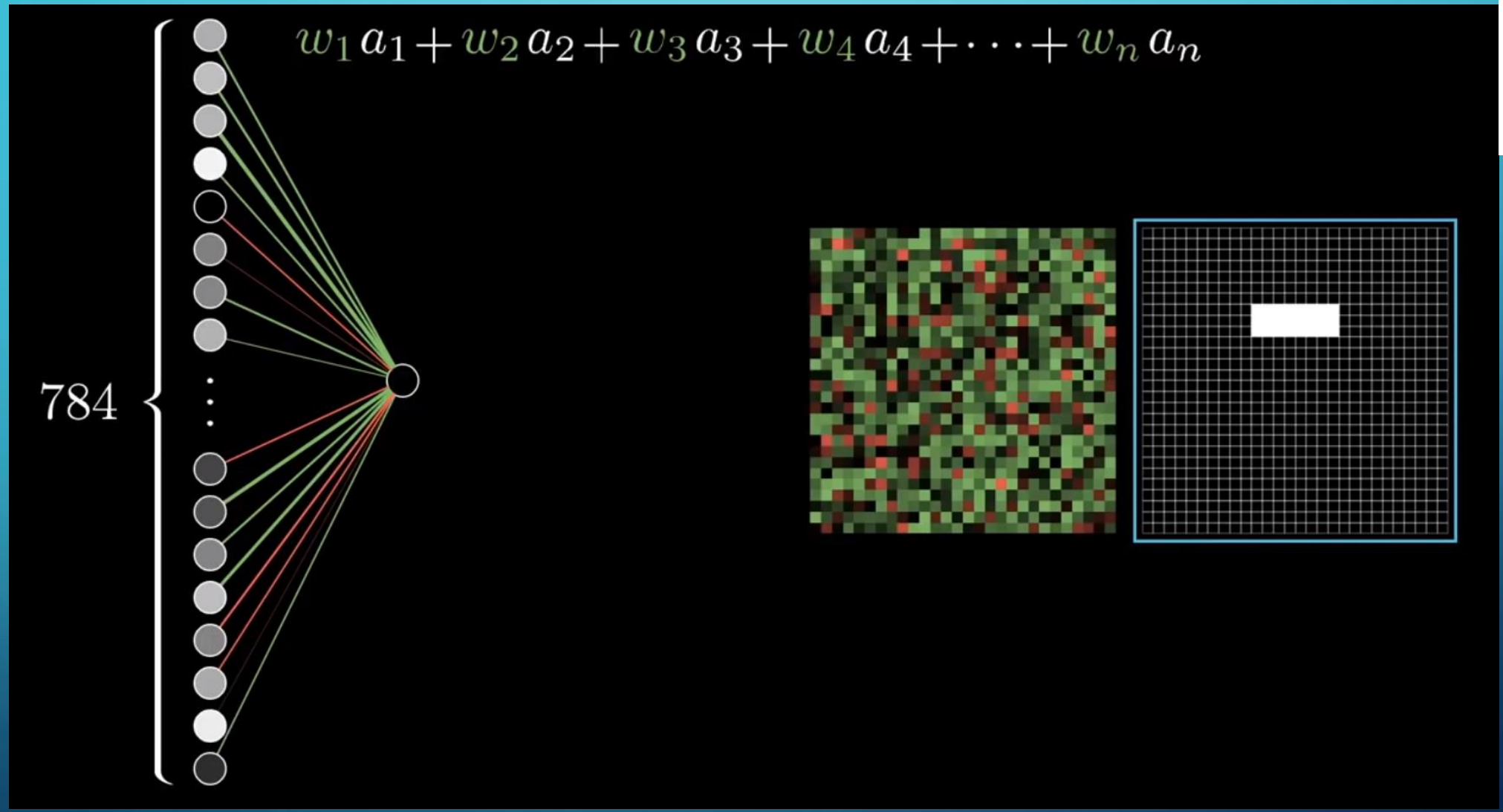
- **private cloud**
- **hybrid cloud**
- **community cloud**

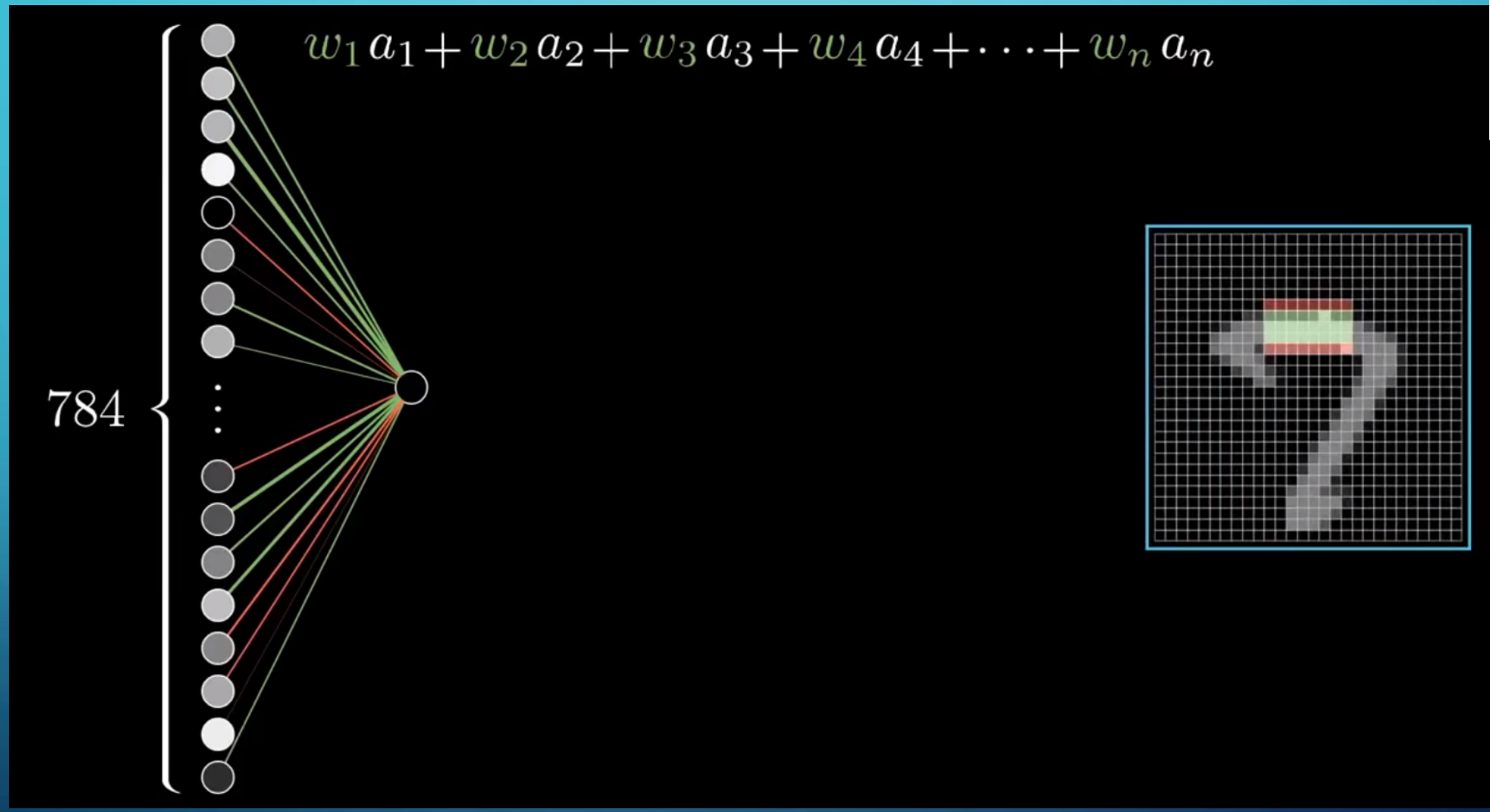
MACHINE LEARNING

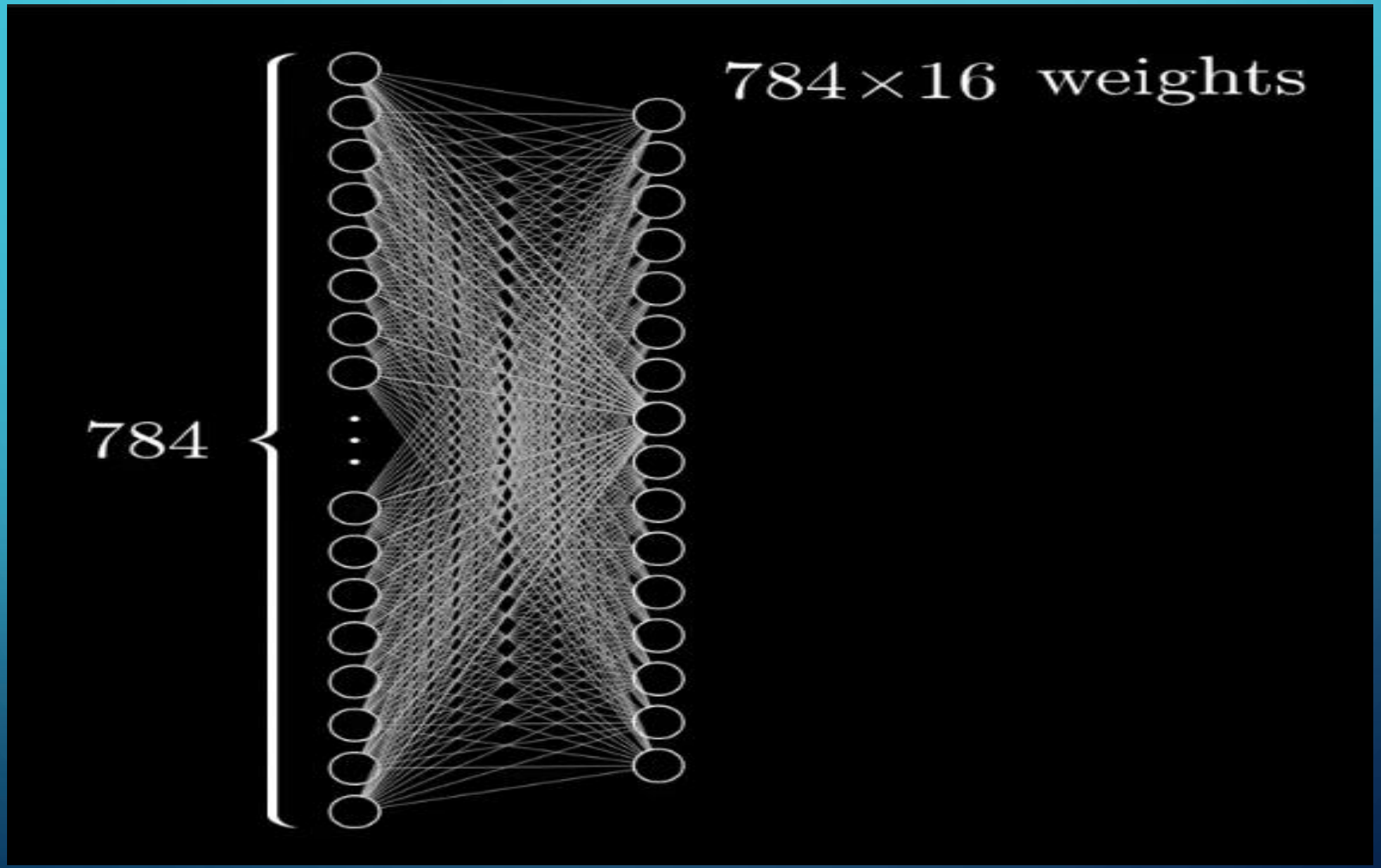
- machine learning(**Machine learning, ML)**
- **Data science -** (Cognitive Science, Big Data)
- **Neural Network, Artificial intelligence, AI)**
- **Deep Learning, DL)**

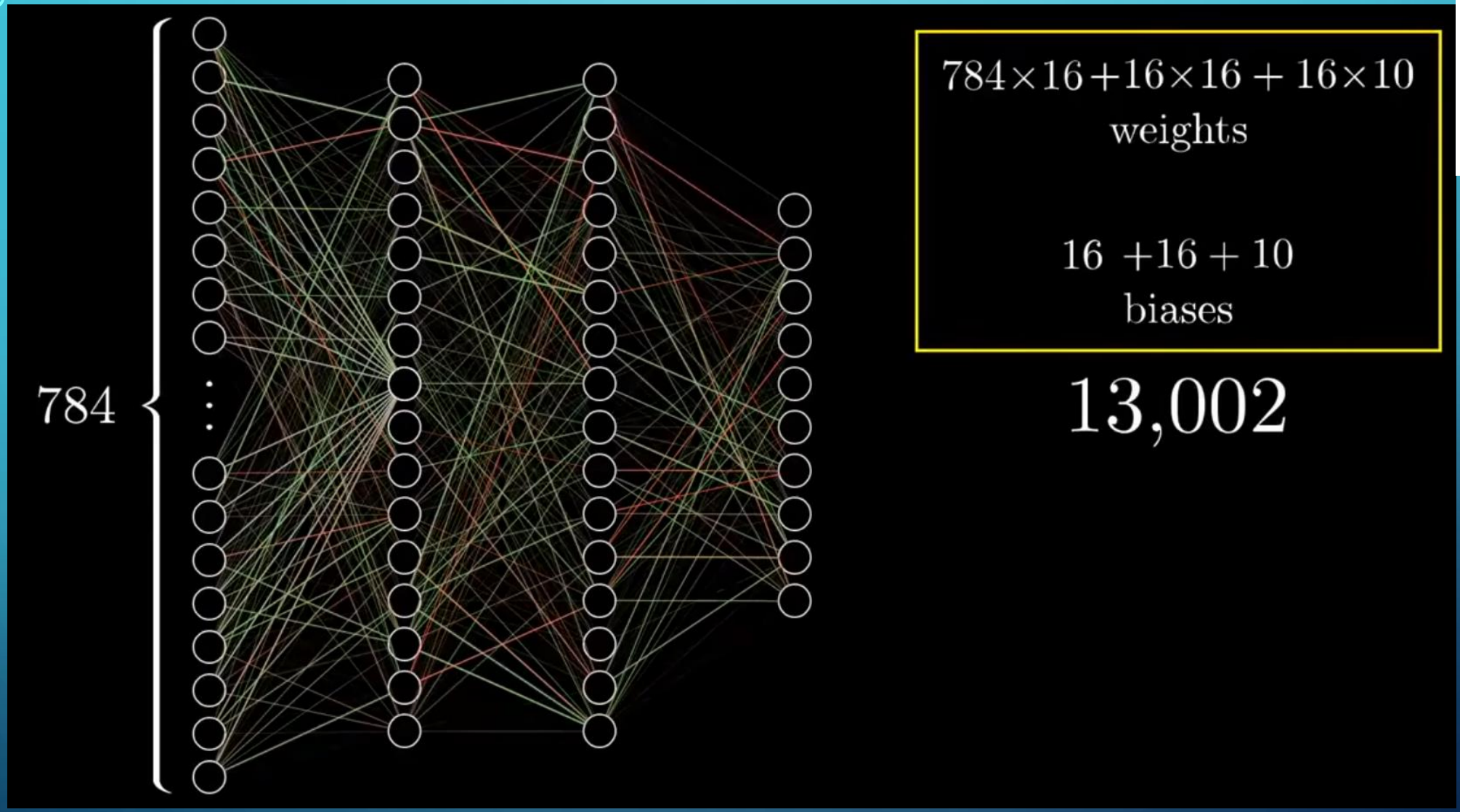


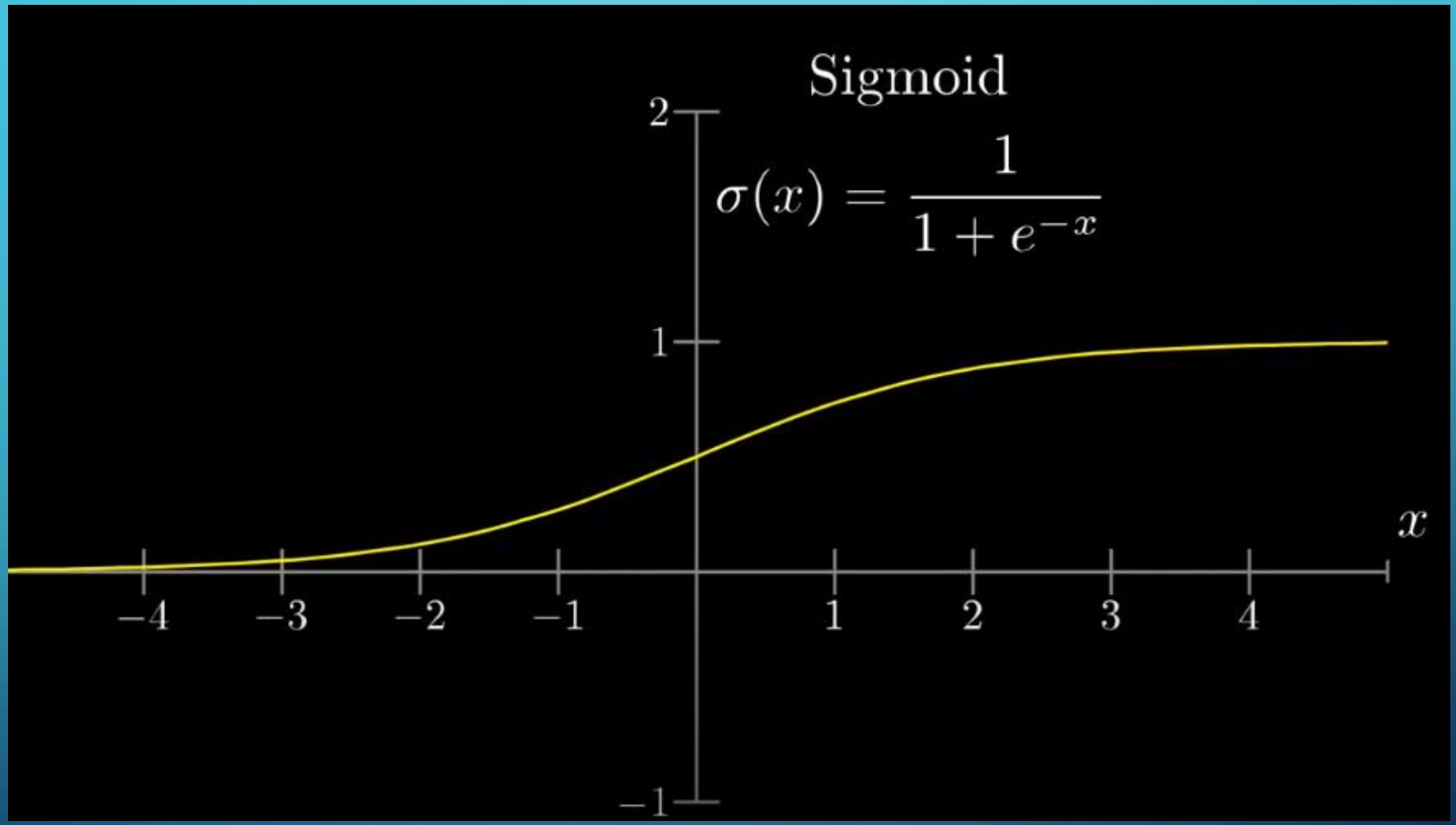

Upper loop neuron...maybe...
784
0
1
2
3
4
5
6
7
8
9


$w_1 a_1 + w_2 a_2 + w_3 a_3 + w_4 a_4 + \cdots + w_n a_n$
784

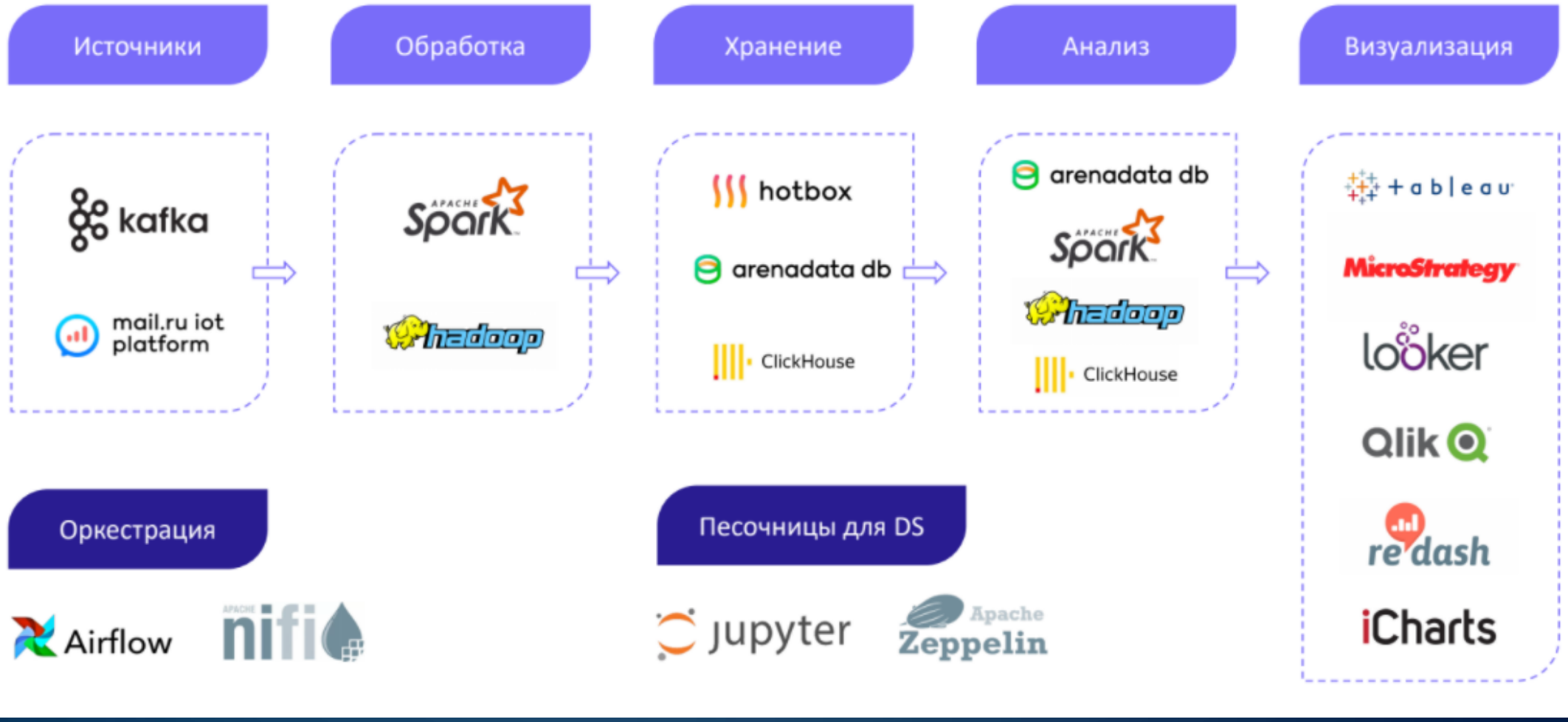

$w_1 a_1 + w_2 a_2 + w_3 a_3 + w_4 a_4 + \cdots + w_n a_n$
784


784×16 weights
784


784
$784 \times 16 + 16 \times 16 + 16 \times 10$
weights
$16 + 16 + 10$
biases
13,002


Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$
2
1
−1
x
−4
−3
−2
−1
1
2
3
4

HOW TO WORK WITH BIG DATA



GREAT INSIGHTS

- Car sharing – hour-coordinate
- Insight: traffic jams
- Insight: you need to set a stop

GREAT CONCEPTS

- ETL – Extract Transform Load
- Lake of data
- DWH – Data Werehouse
- ESB – data bus
- Enrichment
- CDC – Change Data Capture
- MDM – Master Data Management

PLATFORMS


kafka
Prometheus
Grafana
Vault
Jenkins
GitLab
sonarqube
Unit
docker
Sonatype
Nexus
nifi
Spark
Streaming
APACHE
Spark
HELM
kubernetes
Istio
redis
APPLICATIONS
hadoop
Greenplum
PostgreSQL
jupyter

Thank you!

HELGA.HALAN@GMAIL.COM