

# Big Data Analytics

Dr. Faisal Kamiran

# Classroom and TAs

- Course TAs:

1. Abdullah Zia

[msds19087@itu.edu.pk](mailto:msds19087@itu.edu.pk)

## Classroom

- Please join the classroom for this course using the following code.

Code:  
**dzbyqjf**

Link:

<https://classroom.google.com/c/N-Dc2MjU3OTEyOTA1?cjc=dzbyqjf>



# Tentative Course Modules

1. Data Modeling with Relational Databases
2. Data Modeling with NoSQL Databases
3. Data Warehousing
4. Cloud Computing
5. Big Data Tools (Spark, Hive, Hbase, Hadoop, Map-reduce)
6. Data Wrangling with Spark
7. Debugging and Optimization
8. Streaming Data
9. Data Pipelines with Apache Airflow

# Course Grading



**DATA SCIENCE  
LAB**



INFORMATION  
TECHNOLOGY  
UNIVERSITY

- Quizzes: 12%
- Assignments: 12%
- Final Project: 15%
- Mid-term Exam: 25%
- Final Exam: 35%

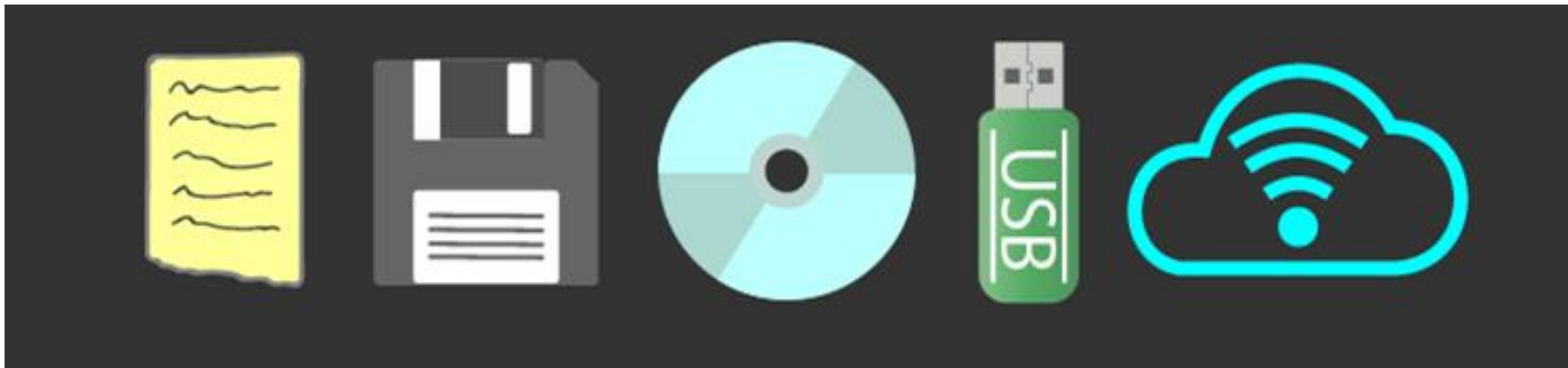


# What is Big Data Analytics?

- Big Data
  - Buzz Word
  - Datasets too large for modern relational databases.
  - Semi-Structured/ Unstructured Datasets
- Analytics
  - How to measure?
  - Identifying patterns in data.



# Big Data Timeline



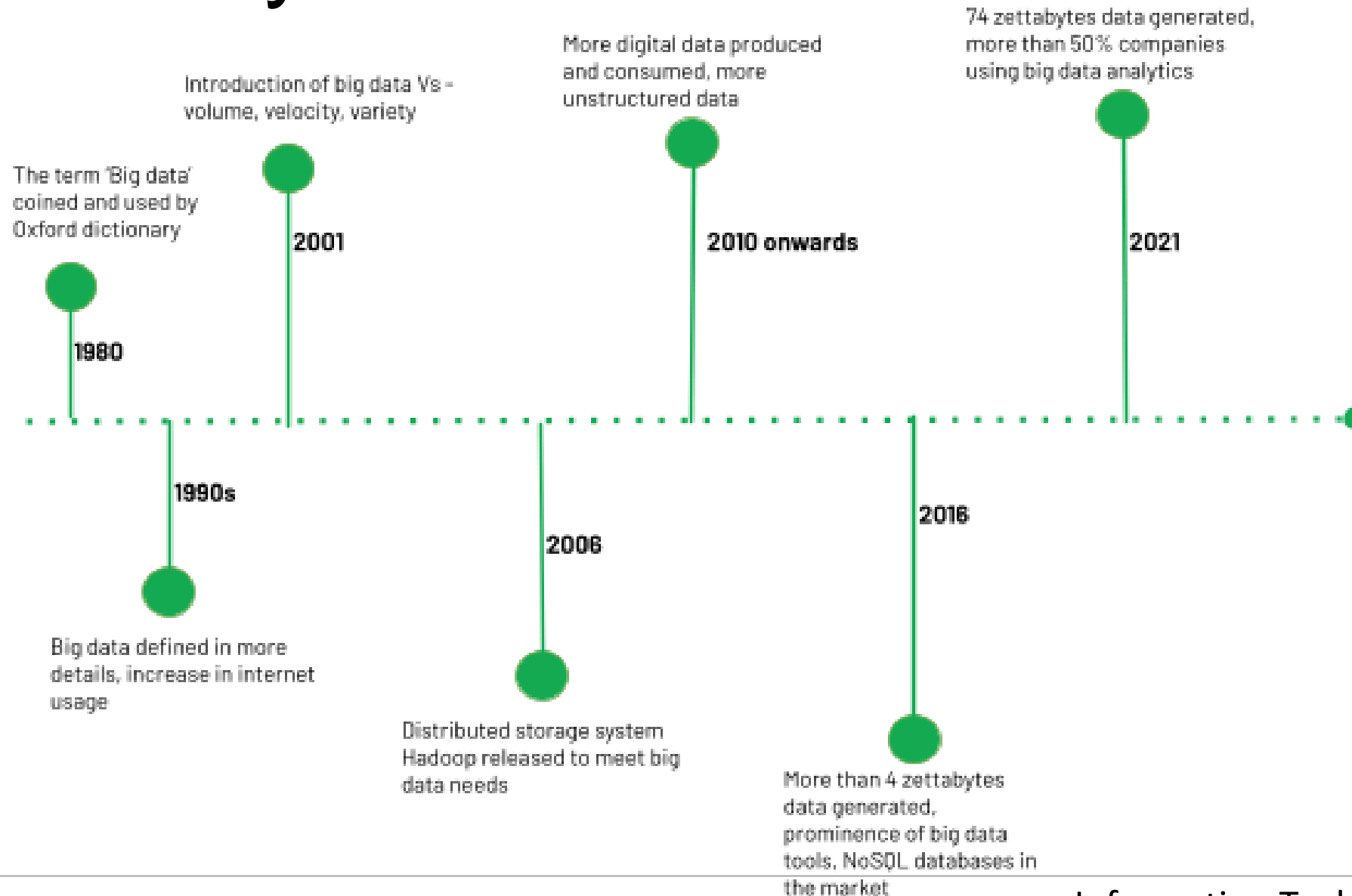
# Brief History



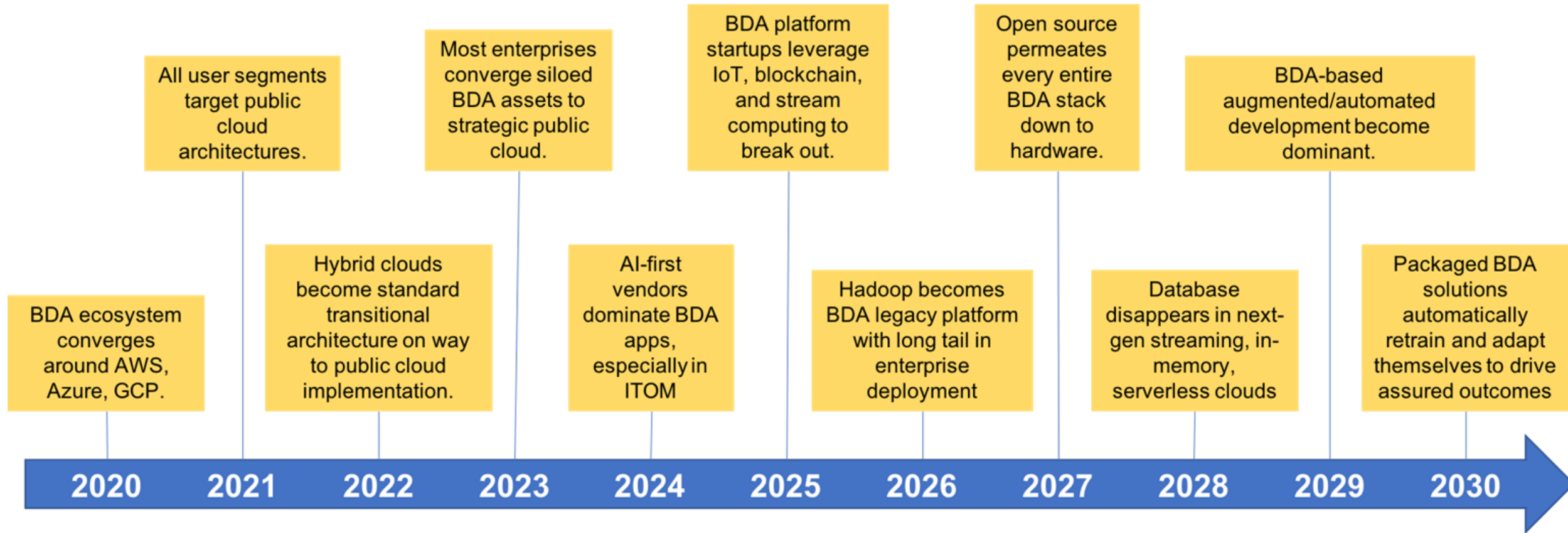
DATA SCIENCE  
LAB



INFORMATION  
TECHNOLOGY  
UNIVERSITY



# Scope of BDA in Future







# (Evolution of Unstructured Data)

- 1989: Birth of Internet
- 1998: Development of search engines
- Web 2.0 was introduced in 2004 that provided a base for unstructured and user-generated data that included blogs, Wikis, social media platforms, etc.





# (DATA VOLUMES INCREASED)

- There were 5 exabytes of information created by the entire world between the dawn of civilization and 2003 (Eric Schmidt, 2010)
- Now that same amount is created every two days.
- Increased demand of smartphone and smart IOT devices



? TBs of  
data every day

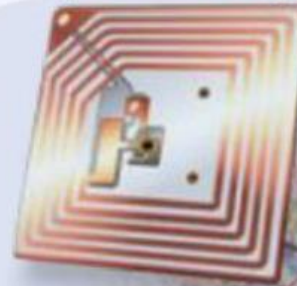


**12+ TBs**  
of tweet data  
every day



**25+ TBs** of  
log data  
every day

**30 billion** RFID  
tags today  
(1.3B in 2005)



**4.6 billion**  
camera  
phones  
world wide



**100s of millions**  
of GPS  
enabled  
devices sold  
annually



**76 million** smart meters  
in 2009...  
200M by 2014



**2+ billion**  
people on  
the Web  
by end  
2011



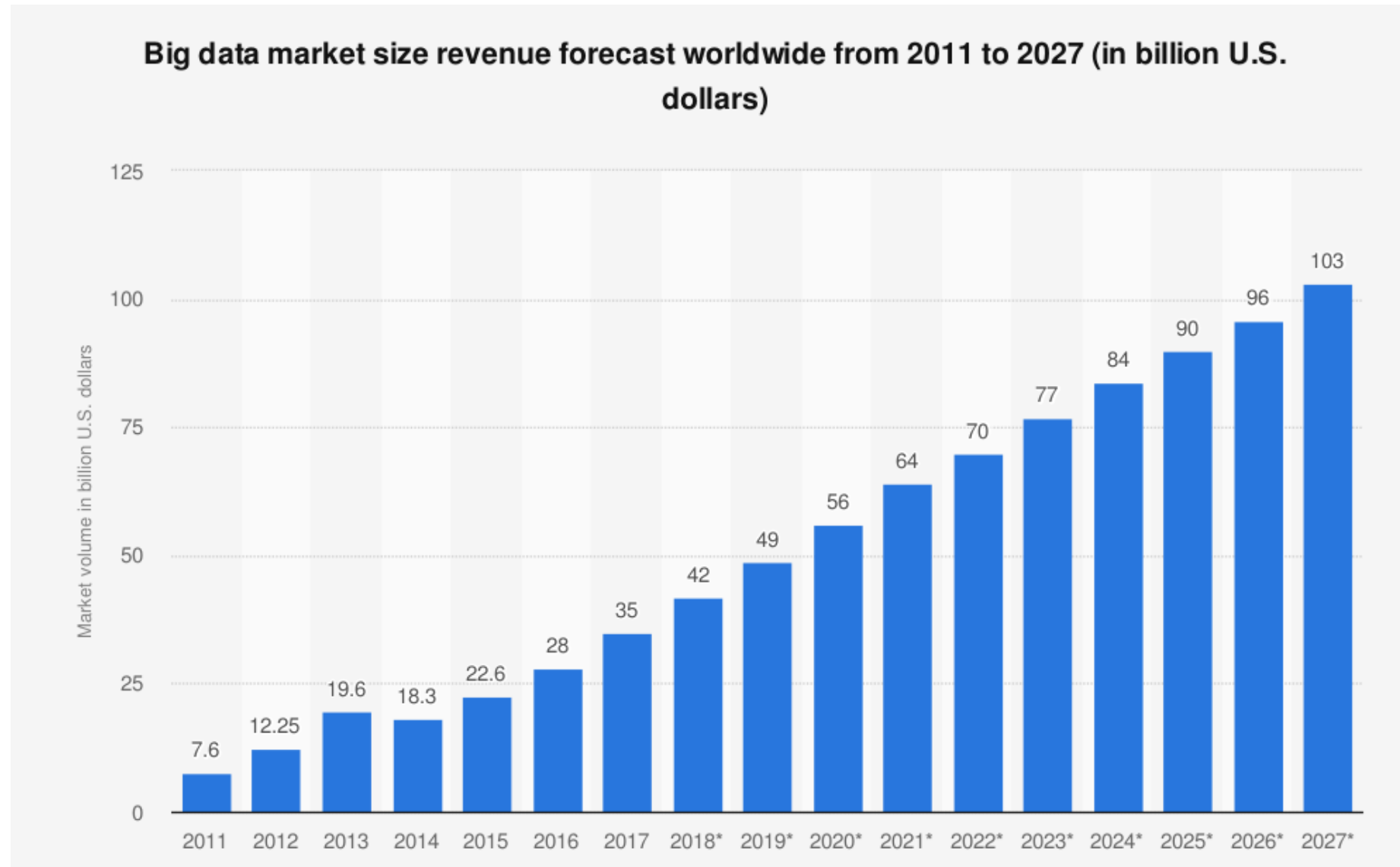




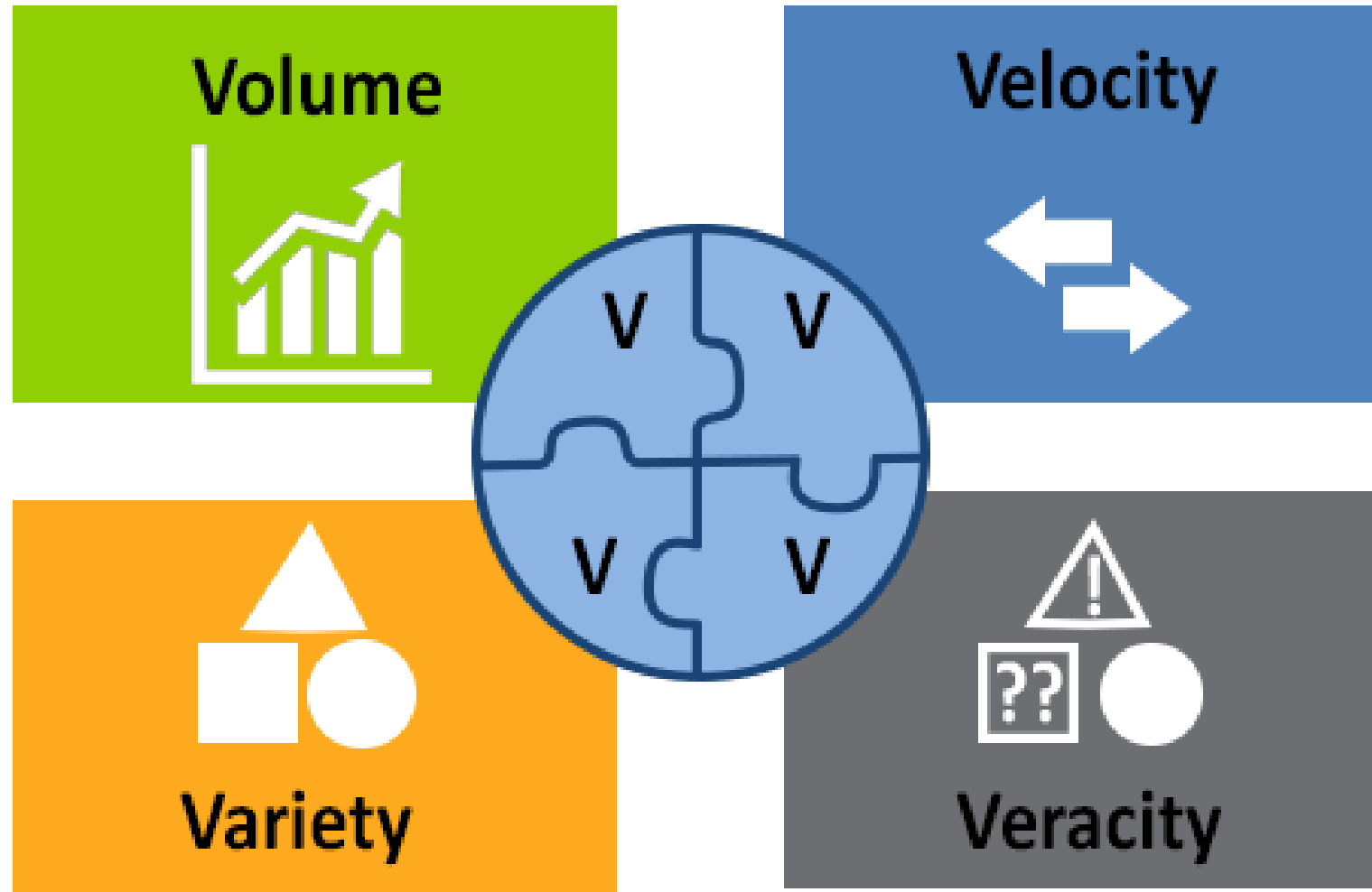
# GLOBAL MARKET - 2021

The global Big Data market was said to be worth **\$64bn** in 2021, and at that time, and predicted to rise to **\$103bn** by 2027 .





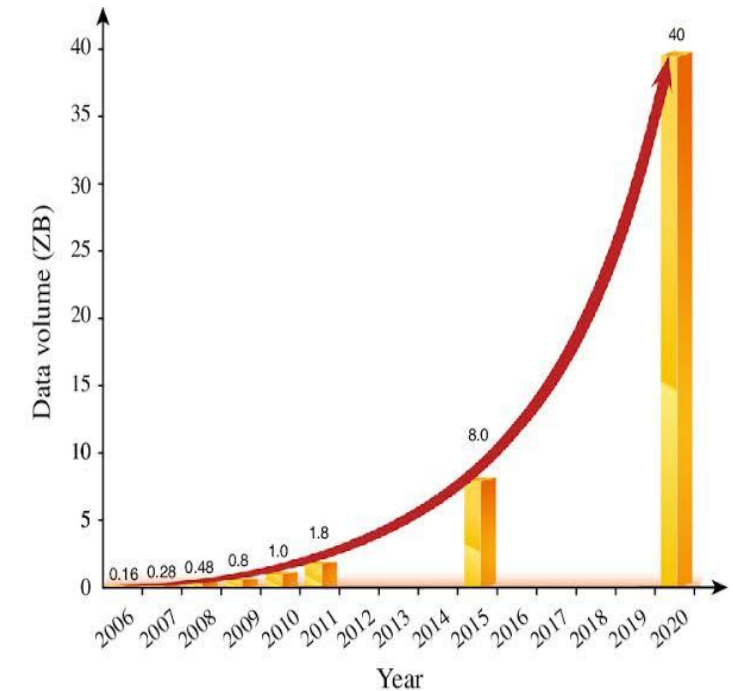
# Big Data Using 4 Vs



# Data Volume



- A typical PC might have had 10 gigabytes of storage in 2000.
- Today, WhatsApp users exchange up to 65 billion messages every day.
- More than 120 professionals join LinkedIn every minute.
- 88,000 YouTube videos are viewed every second.
- Twitter users send over 528,780 tweets every minute.
- Every minute there are 510,000 comments posted
- Internet users generate about **2.5 quintillion bytes** of data each day.



# Data Velocity



- **Velocity** is the rate at which the data is being generated
- High-frequency stock trading algorithms reflect market changes within microseconds.
- Machine to machine processes exchange data between billions of devices.
- Infrastructure and sensors generate massive log data in real-time.
- Online gaming systems support millions of concurrent users, each producing multiple inputs per second.



# Data Variety



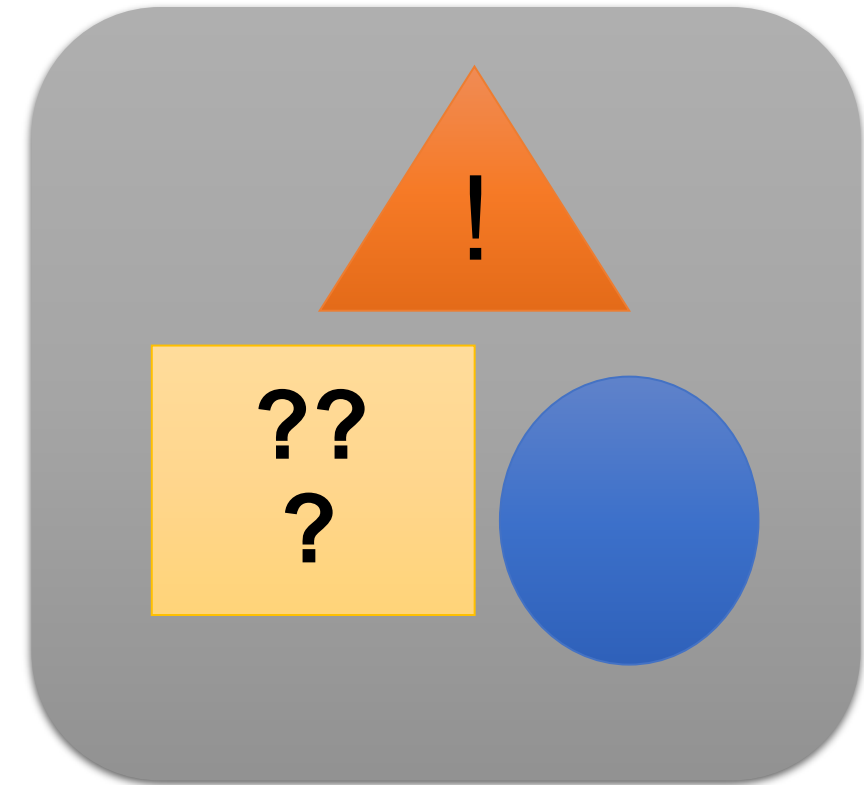
- Big Data isn't just numbers, dates, and strings. Big Data is also
  - Geospatial data
  - 3D data
  - Audio and video
  - unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data.





# Data Veracity







- Biases, noise and abnormality in data
- Is the data that is being stored, and mined meaningful to the problem being analyzed?
- Need to have your team and partners work to help keep your data clean





# The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume*, *variety* and *velocity*. Over time, other Vs have been added to descriptions of big data:

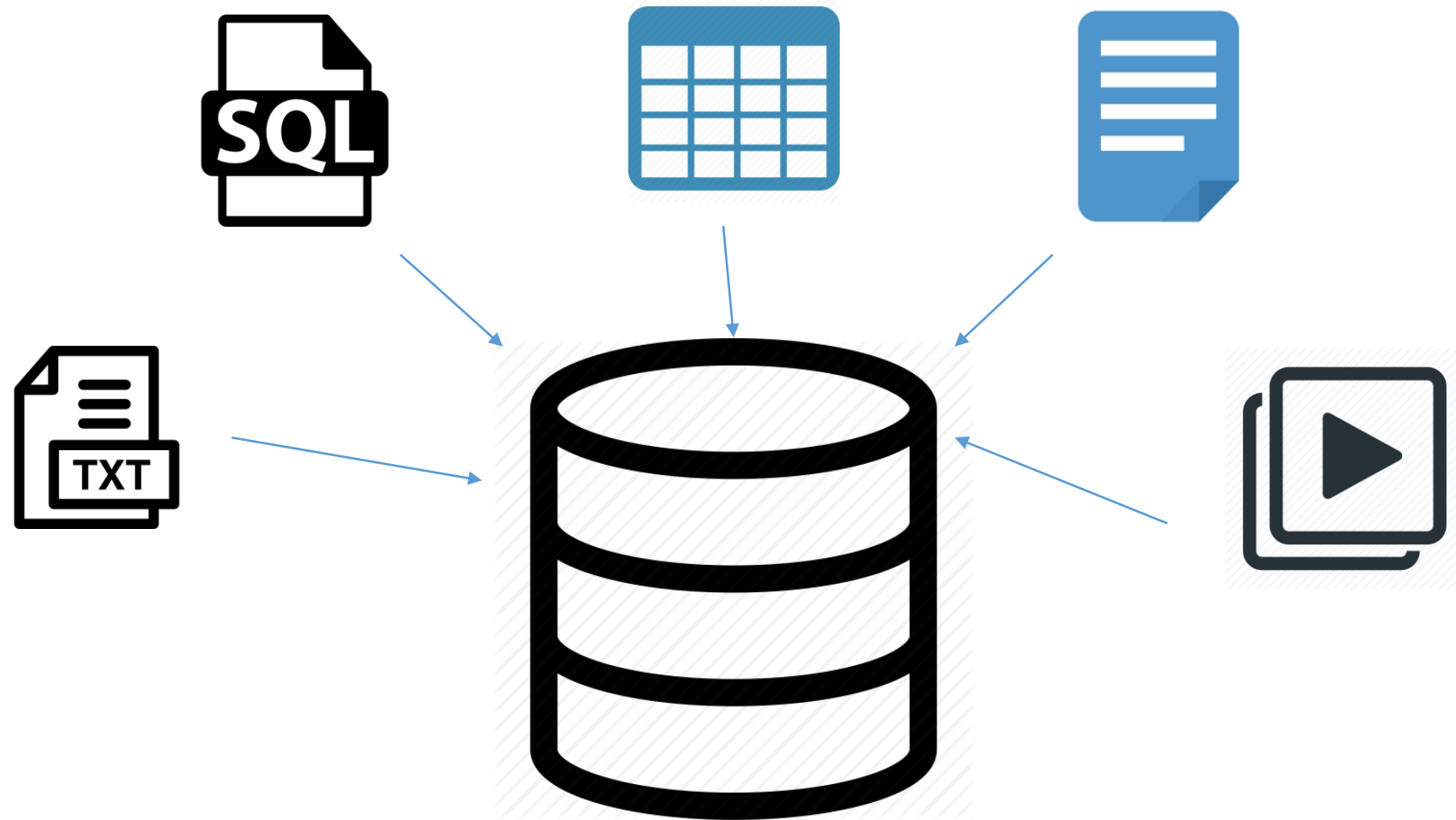
VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					



# Why Big Data Analytics

- Better Approach than Traditional Approaches as it deals with:
  - Unstructured Data
  - Real-time Data
- Business Organization
  - Optimizing Flow of Operations
  - Increase quality and profit
  - Use data to understand customers
  - Create new products and services
- Government
  - Smart City Initiatives
  - Crime Control

# New Approach to Data





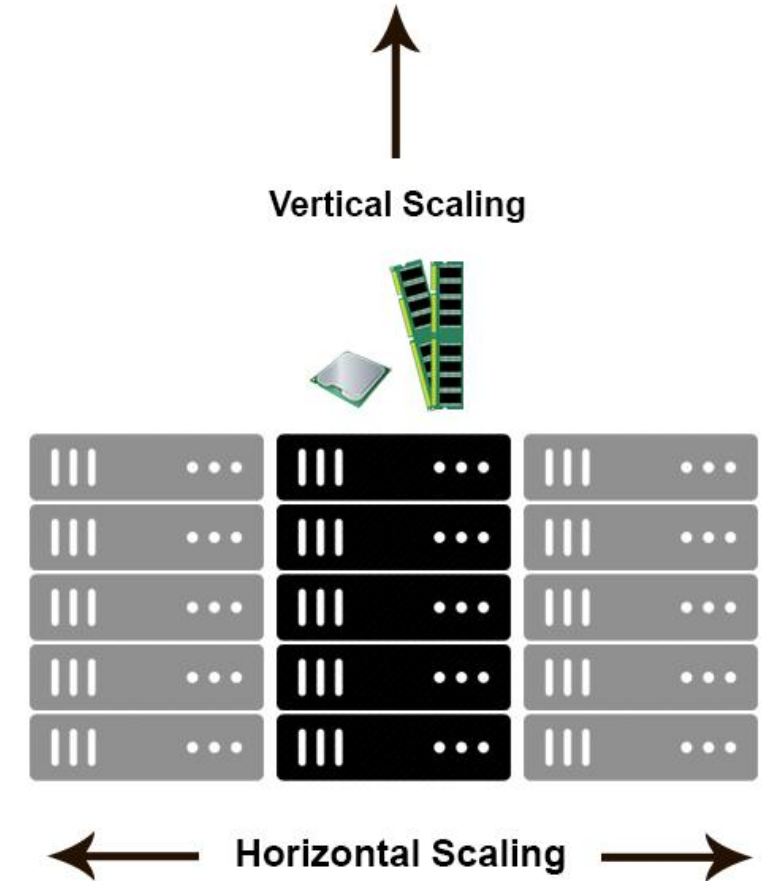
# Big Data Analytics

- Where processing is hosted?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is stored?
  - Distributed Storage (e.g. Amazon S3)
- What is the programming model?
  - Distributed Processing (e.g. MapReduce)
- How data is stored & indexed?
  - High-performance schema-free databases (e.g., MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing



# Vertical Scaling Vs. Horizontal Scaling?

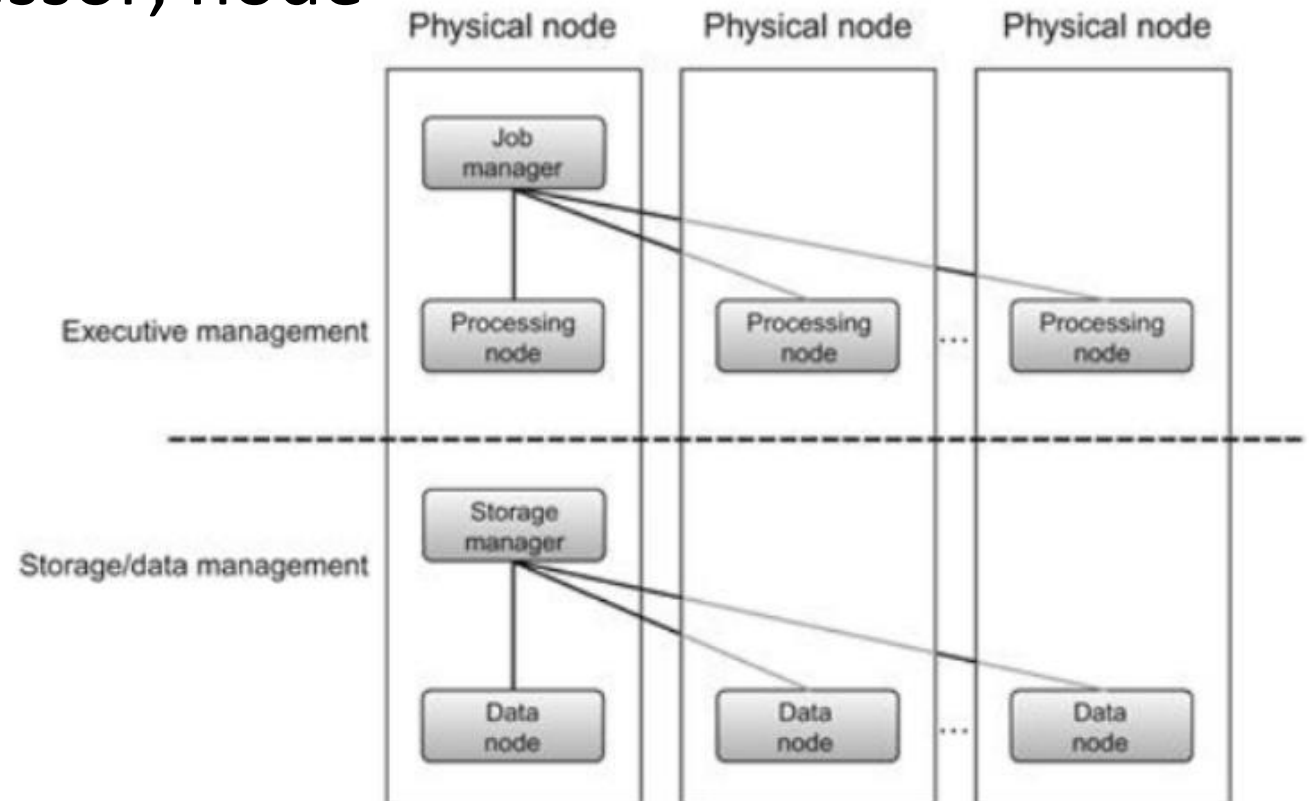
- **Horizontal scaling** means that you scale by adding more machines into your pool of resources
- **Vertical scaling** means that you scale by adding more power (CPU, RAM) to an existing machine.





# Computing Resources for Big Data

- Processing capability: Processor, node
- Memory
- Storage
- Network

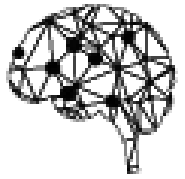






# Techniques towards Big Data

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High speed networks
- Task and thread management
- Data Retrieval
- Machine Learning

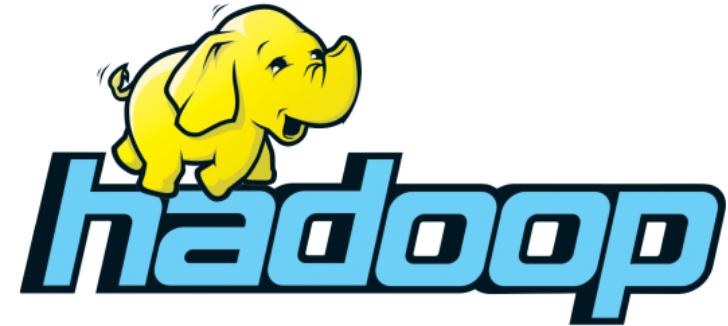


DATA SCIENCE  
LAB

# Big Data Platforms



# Big Data Platforms



- Hadoop
  - HDFS
  - Map-reduce
- Spark
  - RDD



# What is Hadoop?

Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware.



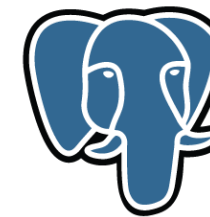


# RELATIONAL DATABASES

In general, RDBMS systems have been considered as the **one-size-fits-all** data retrieval and persistence solution for decades



PostgreSQL



Microsoft®  
SQL Server®



# NOSQL DATABASES

NoSQL stands for:

- No Relational
- No RDBMS
- **Not Only SQL**
  - Allows SQL-like query languages to be used.



NoSQL is an umbrella term for all databases and data stores that do not follow the RDBMS principles

# Important Big Data Tools



DATA SCIENCE  
LAB

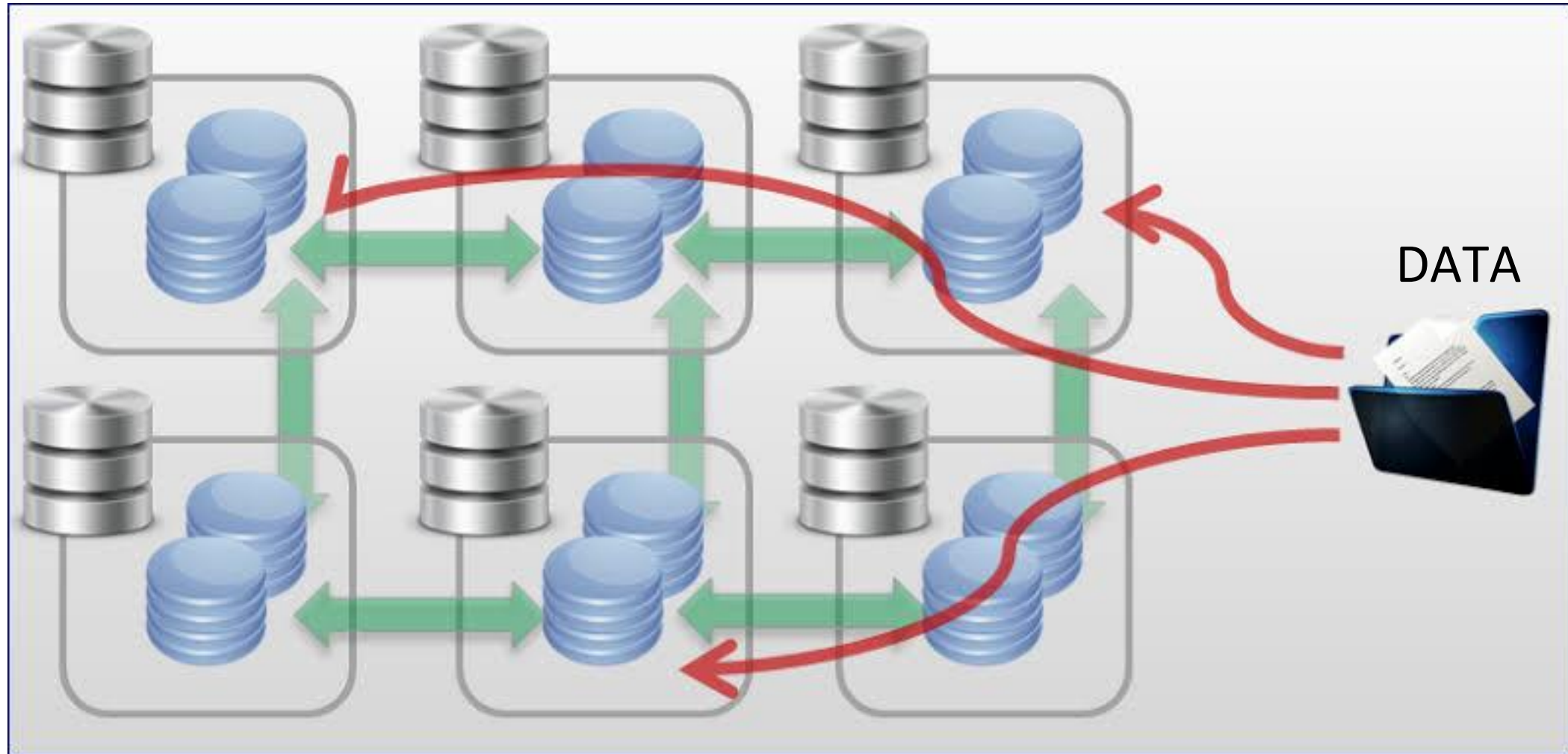


INFORMATION  
TECHNOLOGY  
UNIVERSITY



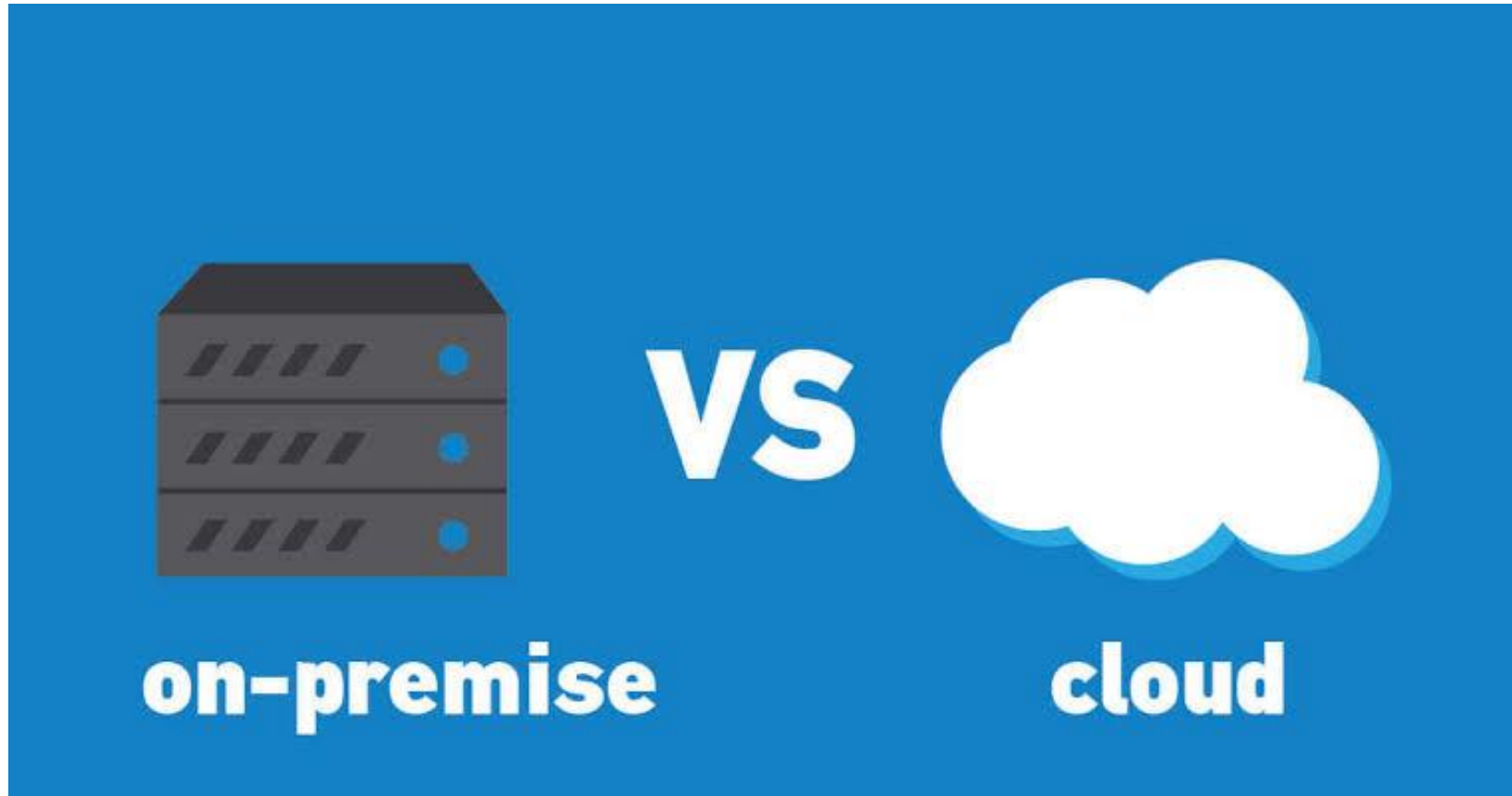


# Moving Computation to Data



Source: UC San Diego, Big Data







# On Premises

- Software and technology
- located within the physical confines of an enterprise
- Often in the company's data centers – as opposed to running remotely on hosted servers or in the cloud.



On Premise



# Cloud Computing

- Differs from on-premises software in one critical way.
- Third-party provider hosts all that for you.
- Allows companies to pay on an as-needed basis.
- Allows companies to effectively scale up or down depending on overall usage, user requirements, and the growth of a company.



# Questions?