BIG
DATA
TOOLS
DISK
TARGET
APPLIED
SENSOR
SHARED
DEFINITION
CURRENT
MAY
MOVING
WITHIN
MANAGE
ARCHIVES
EXAMPLES
EVERY
PRACTITIONERS
CAPTURE
BUSINESS
TECHNOLOGIES
USING
INTERNET
DISTRIBUTED
SETS
GENOMICS
NEEDED
TIME
ONE
LARGE
MANAGEMENT
TERABYTES
CASE
ABILITY
SOFTWARE
INFORMATION
PROCESS
PETABYTES
INCLUDE
PERFORMANCE
STORE
ELAPSED
SYSTEMS
TOLERABLE
SIZE
RESEARCH
STORAGE
MPP
LOGS
COST
PARALLEL
SAN
MASSIVELY
GROW
SET
ORGANIZATIONS
RELATIONAL
USE
LARGER
CONTINUES
SOCIAL
COMPLEX
CITATION
TENS
ANALYTICS
BIOLOGICAL
PROCESSING
CAPACITY
HUNDREDS
NETWORKS
DATABASES
RECORDS
SEARCH
GARTNER
WORKING
DIFFICULTY

Evolution Of Data

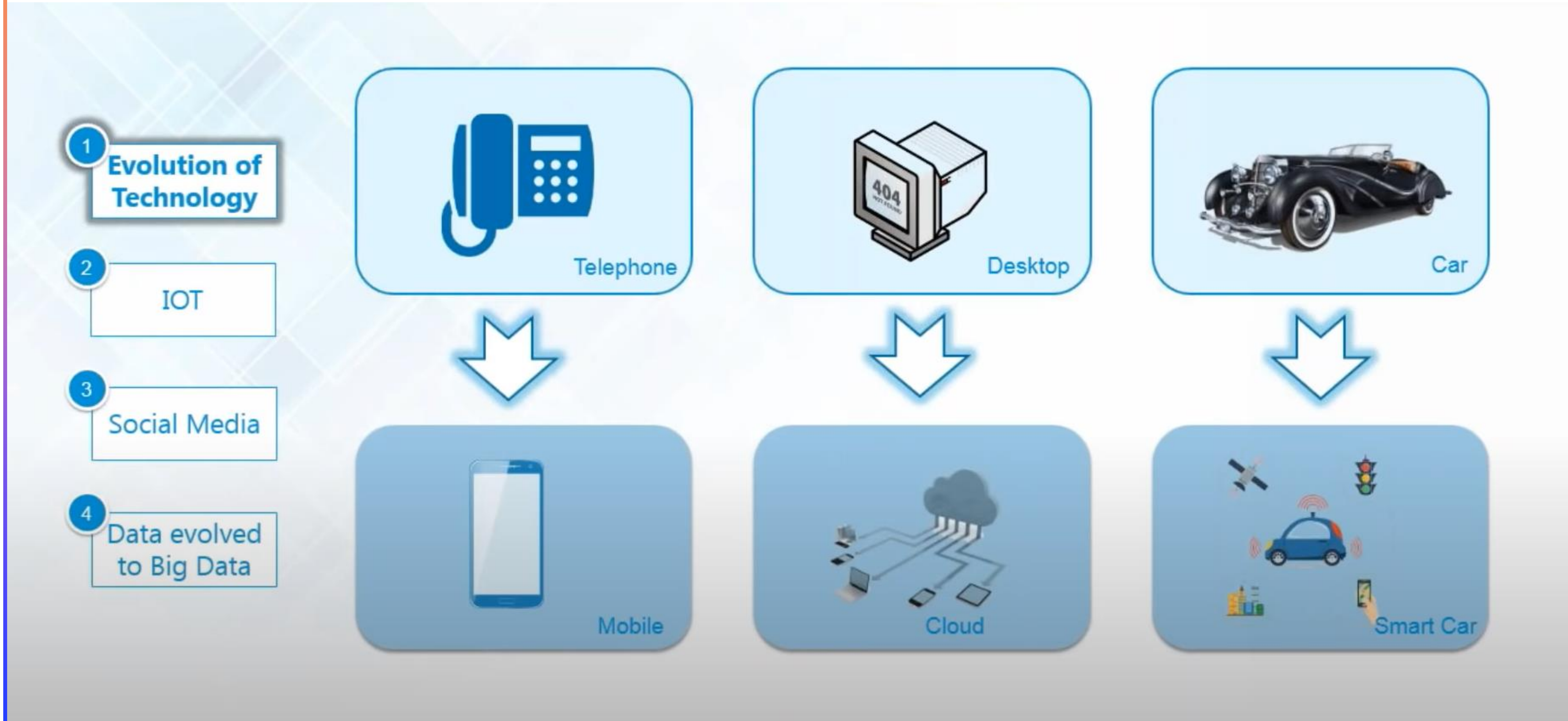
What is Big Data

Big Data as an Opportuniti

Problems in Encasing Opportunity

Haddop as a Solution

Evolution Of Data



1 Evolution of Technology

2 IOT

3 Social Media

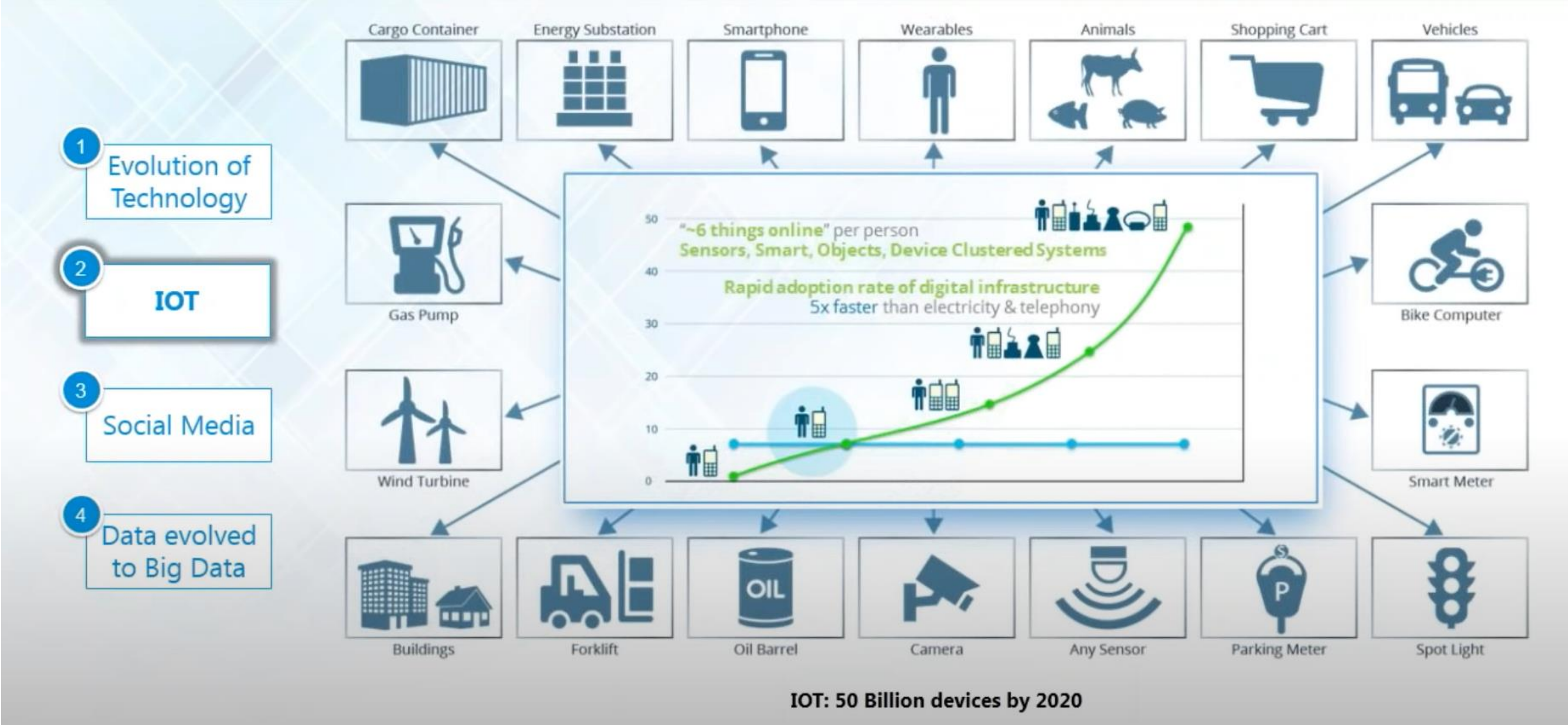
4 Data evolved to Big Data

Telephone → Mobile

Desktop → Cloud

Car → Smart Car

Evolution Of Data



1 Evolution of Technology

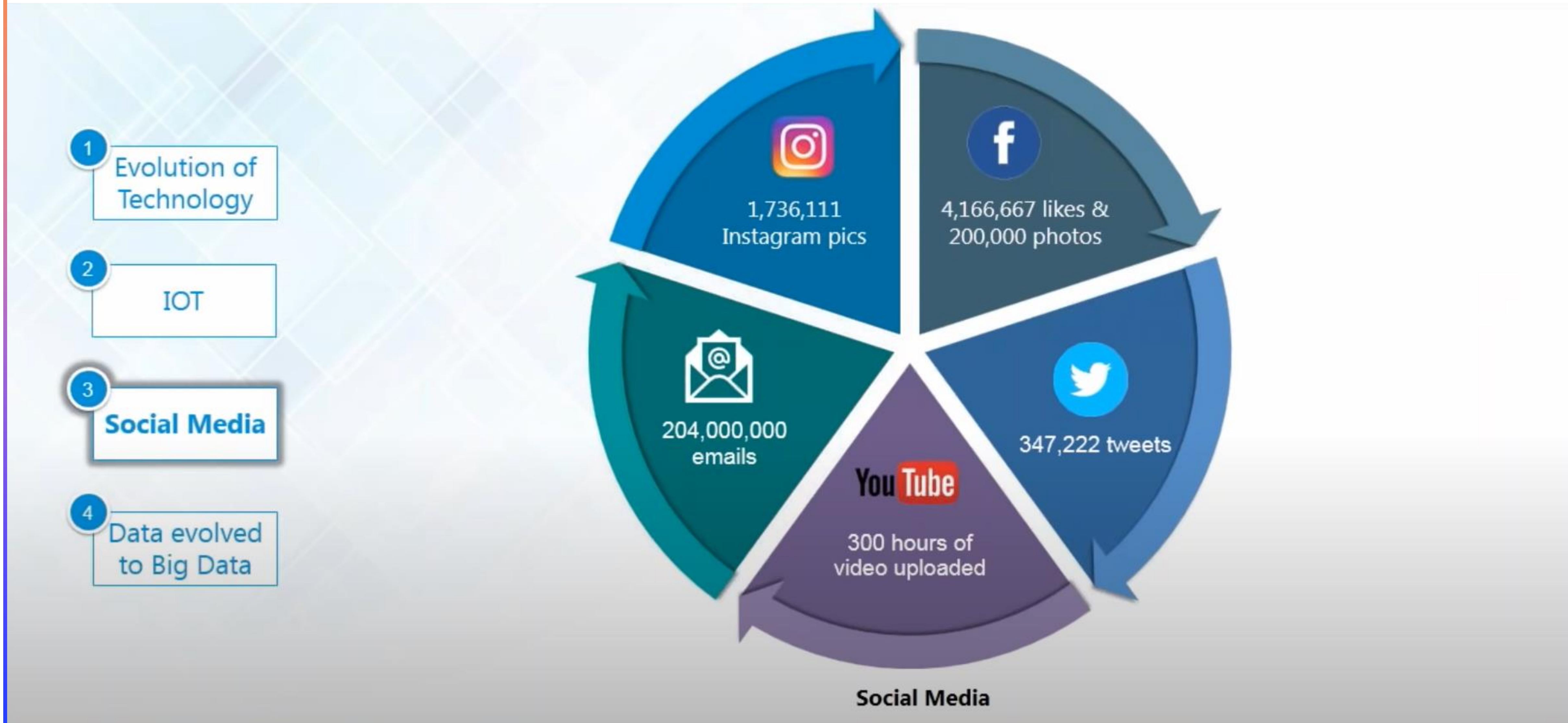
2 IOT

3 Social Media

4 Data evolved to Big Data

IOT: 50 Billion devices by 2020

Evolution Of Data



1. Evolution of Technology
2. IOT
3. **Social Media**
4. Data evolved to Big Data

1,736,111 Instagram pics

4,166,667 likes & 200,000 photos

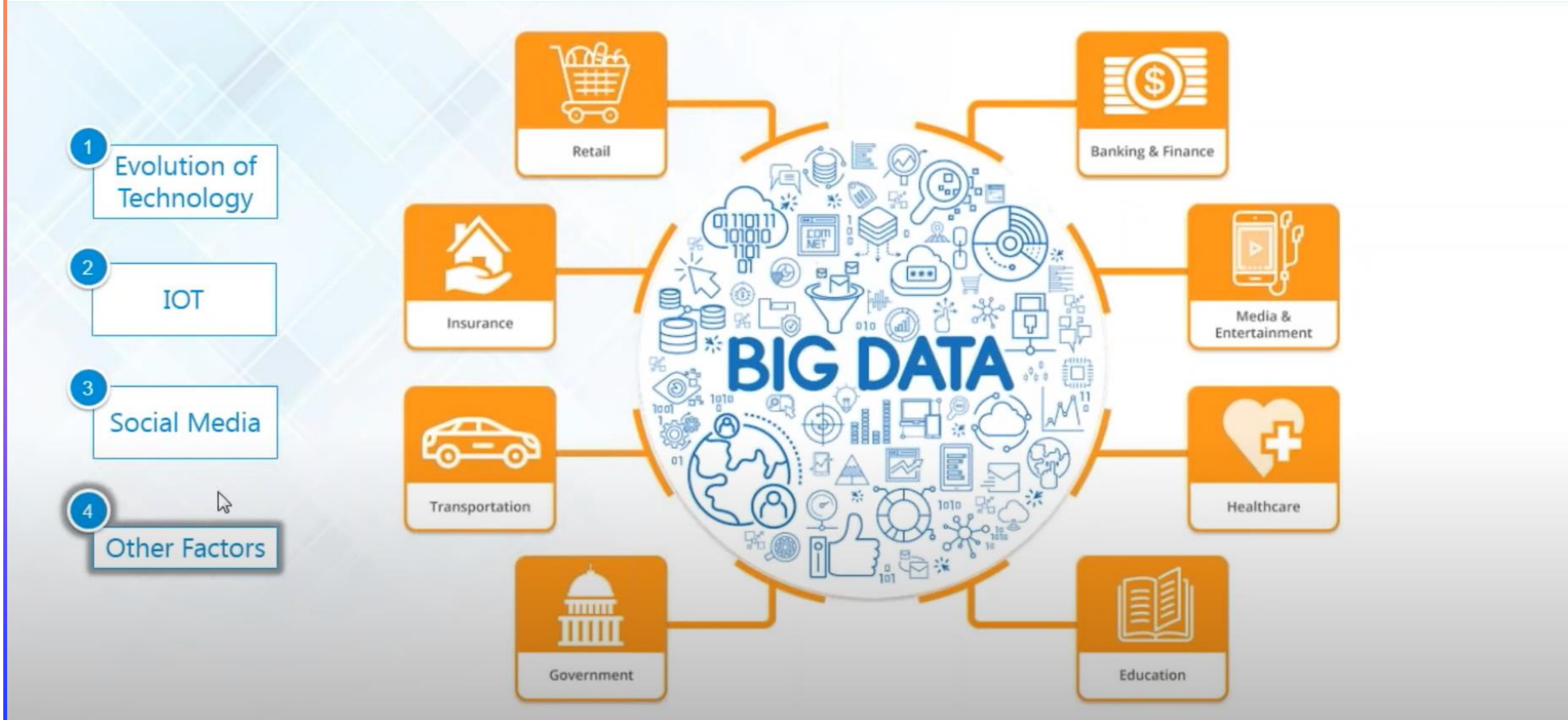
347,222 tweets

300 hours of video uploaded

204,000,000 emails

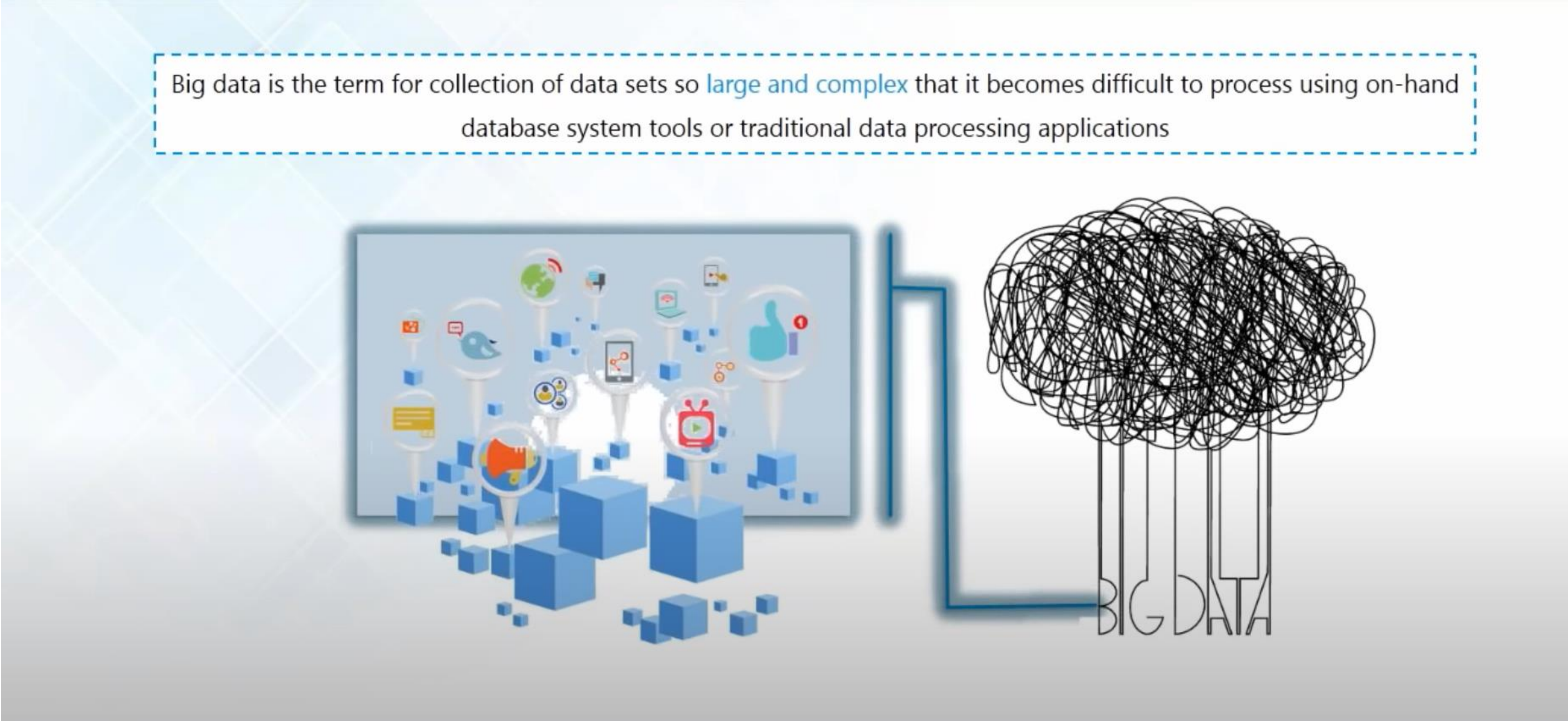
Social Media

Evolution Of Data



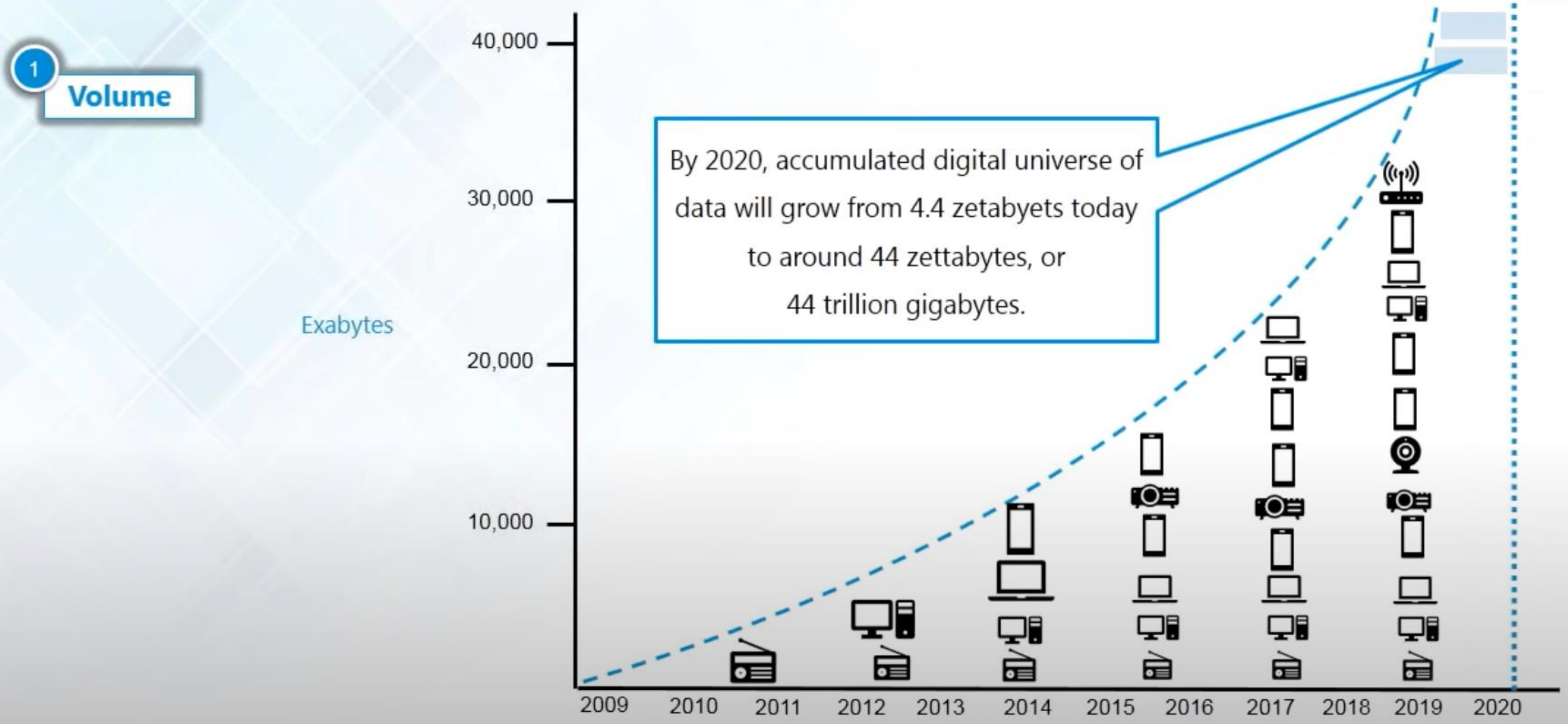
1. Evolution of Technology
2. IOT
3. Social Media
4. Other Factors

What is Big Data



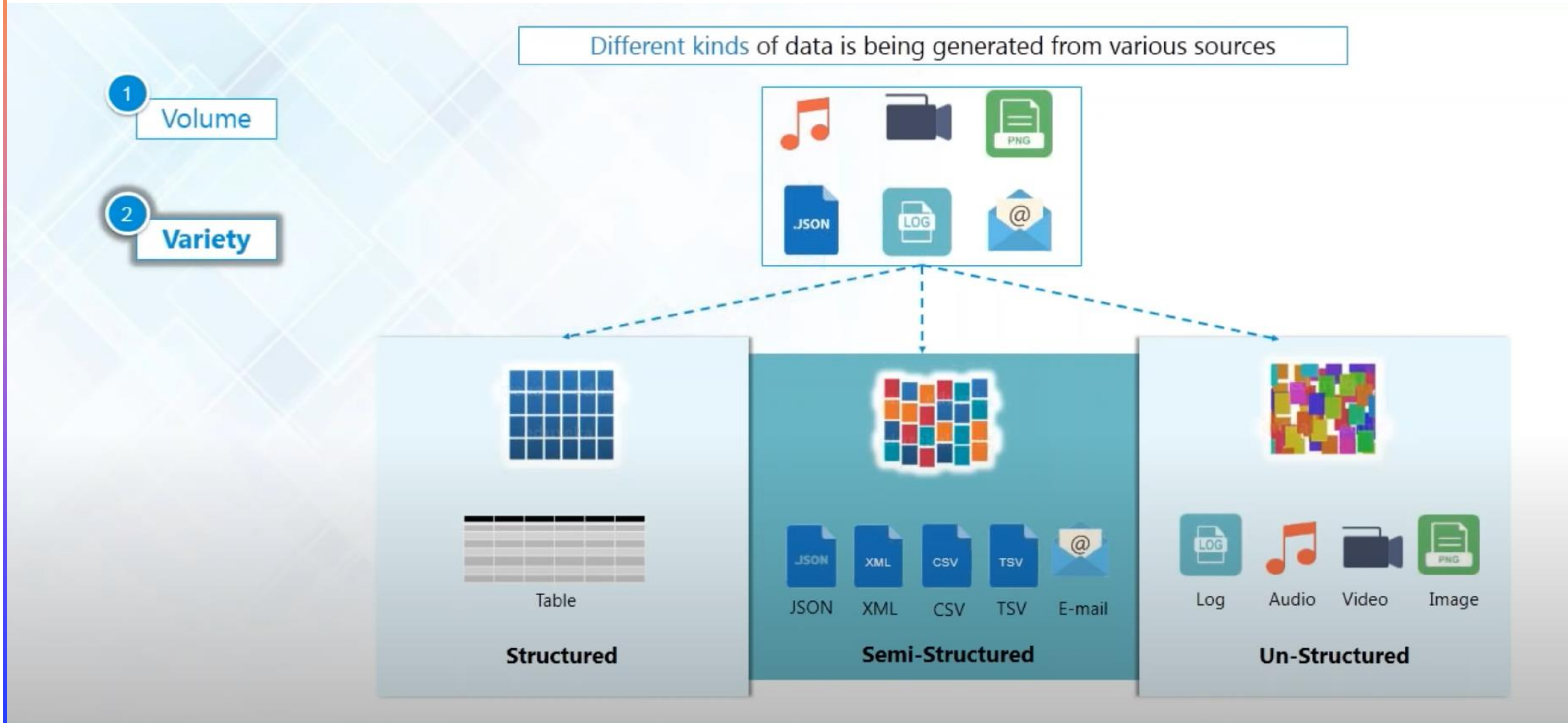
Big data is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database system tools or traditional data processing applications

What is Big Data



By 2020, accumulated digital universe of data will grow from 4.4 zetabyets today to around 44 zettabytes, or 44 trillion gigabytes.

What is Big Data



1 Volume

2 Variety

Different kinds of data is being generated from various sources

Table

Structured

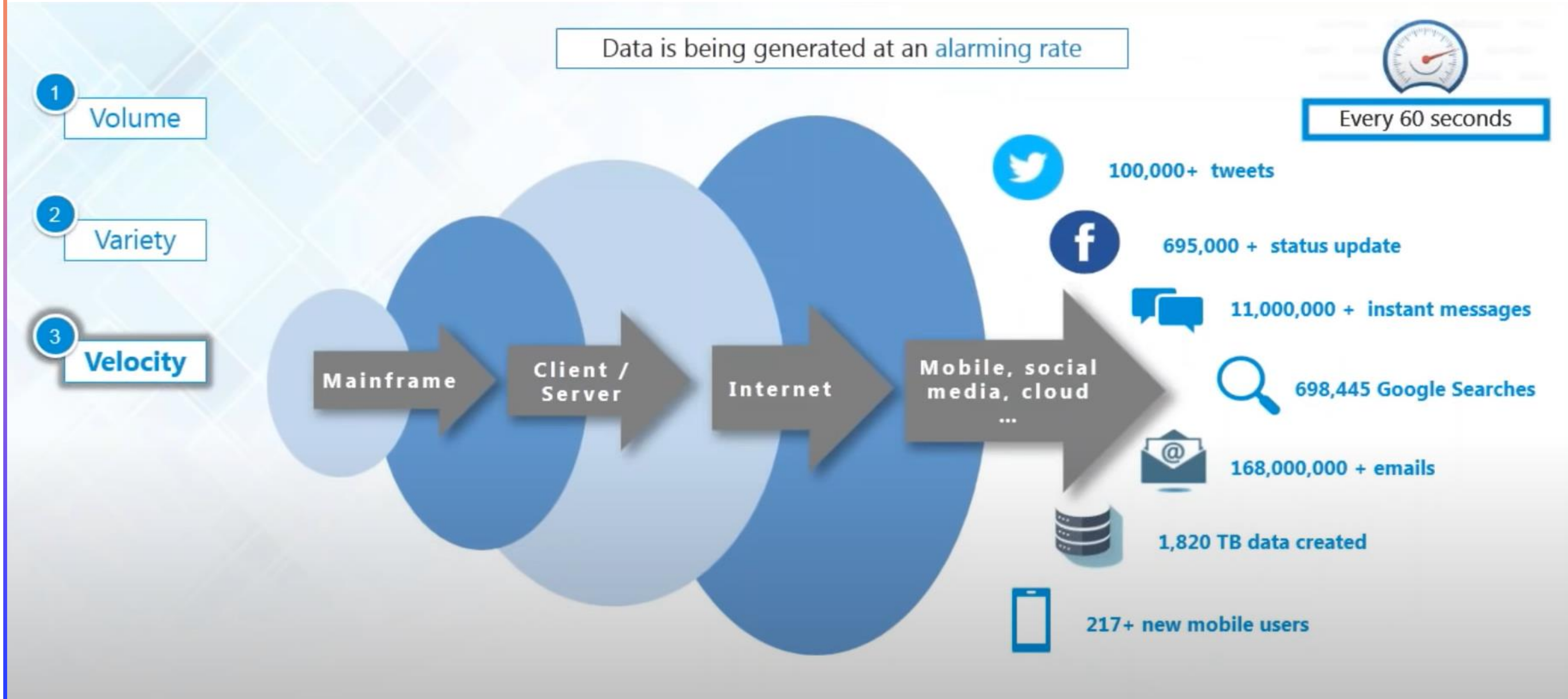
JSON XML CSV TSV E-mail

Semi-Structured

Log Audio Video Image

Un-Structured

What is Big Data



Data is being generated at an alarming rate

1 Volume

2 Variety

3 Velocity

Mainframe → Client / Server → Internet → Mobile, social media, cloud ...

Every 60 seconds

100,000+ tweets

695,000 + status update

11,000,000 + instant messages

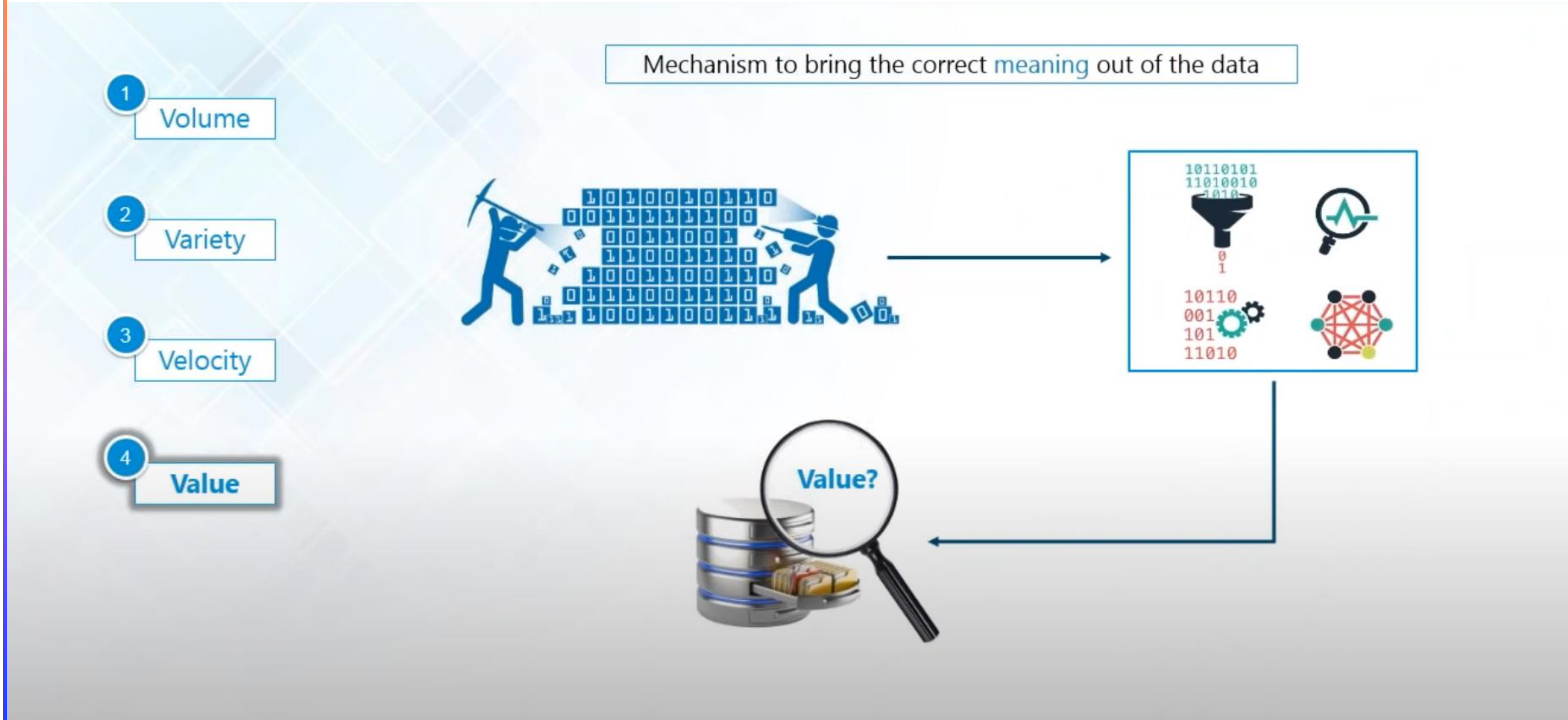
698,445 Google Searches

168,000,000 + emails

1,820 TB data created

217+ new mobile users

What is Big Data



1 Volume

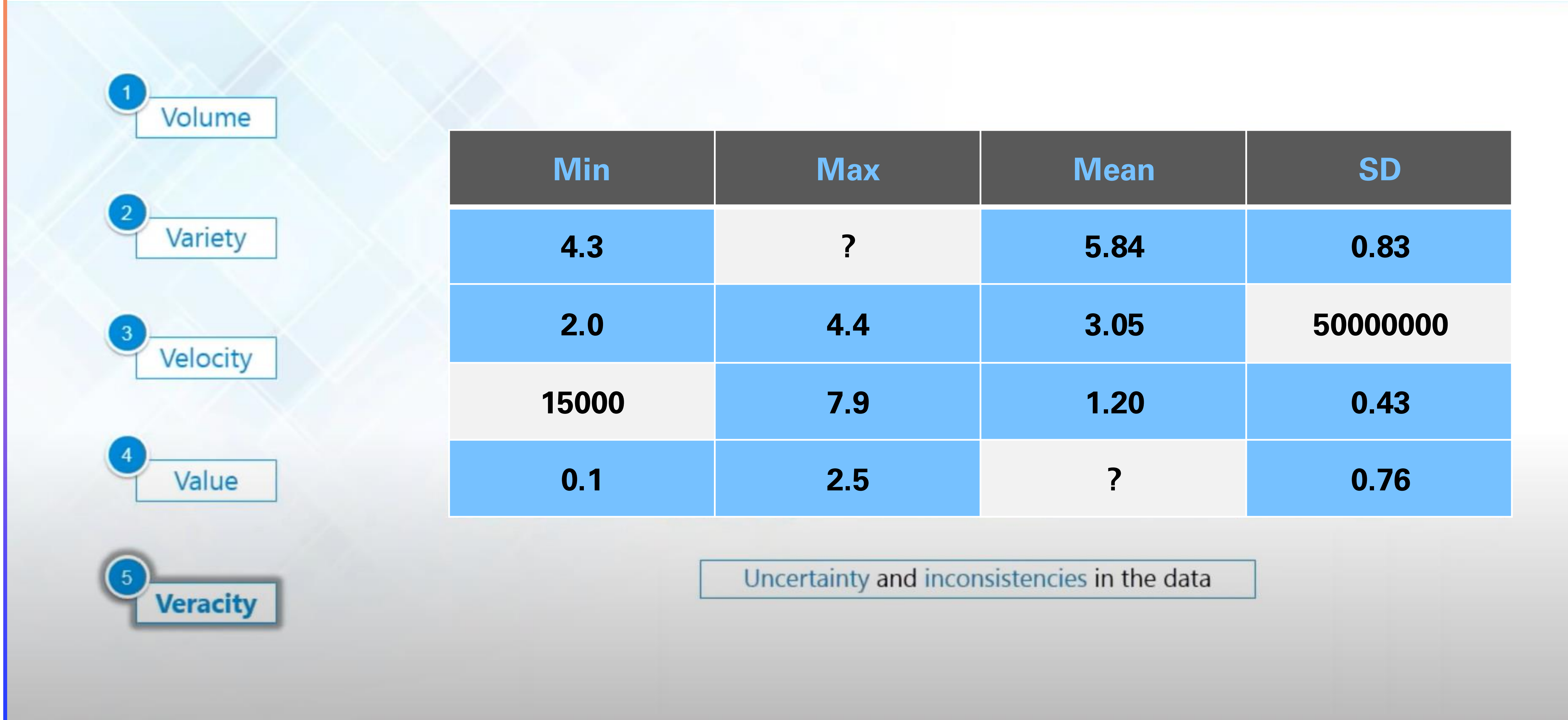
2 Variety

3 Velocity

4 **Value**

Mechanism to bring the correct meaning out of the data

What is Big Data



1 Volume

2 Variety

3 Velocity

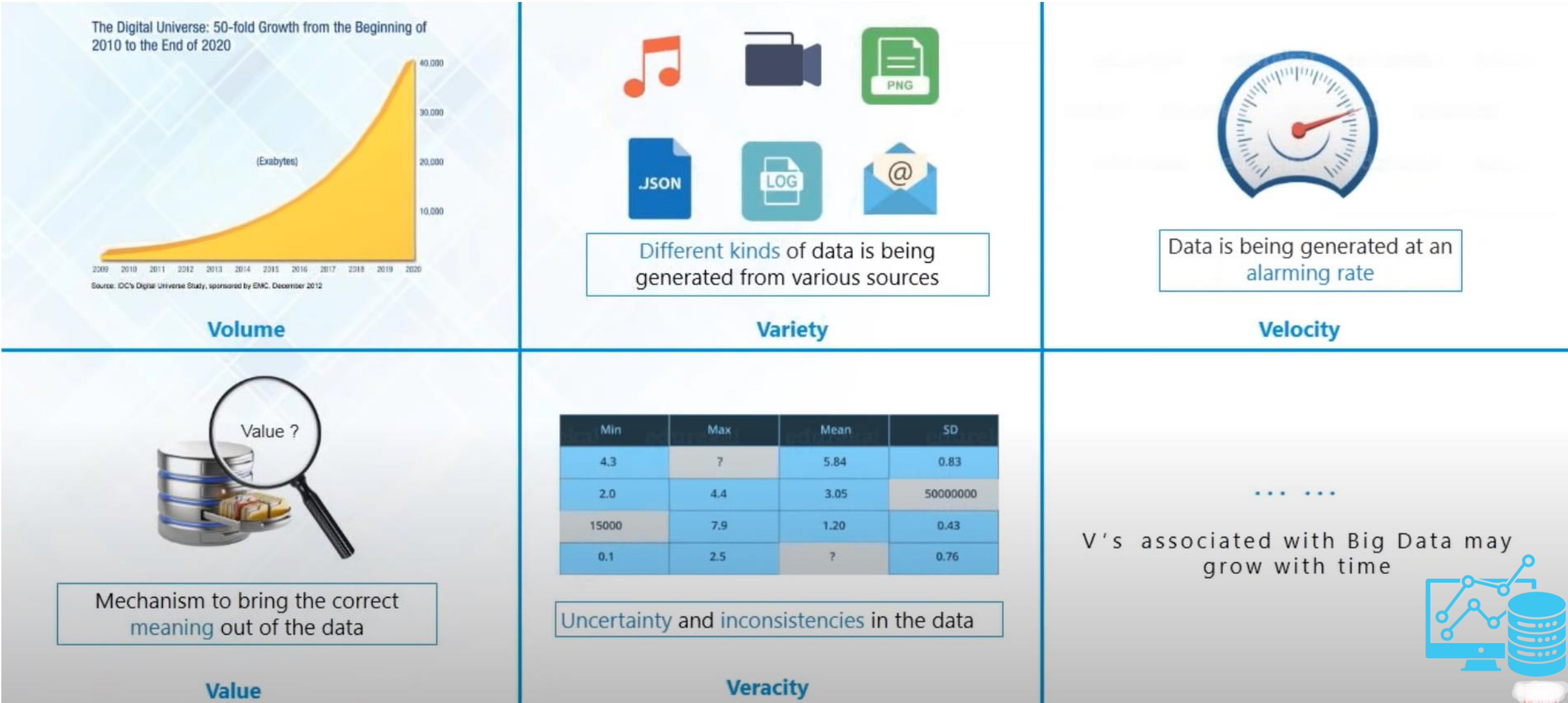
4 Value

5 **Veracity**

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Uncertainty and inconsistencies in the data

5 V's of Big Data



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Volume

Different kinds of data is being generated from various sources

Variety

Data is being generated at an alarming rate

Velocity

Mechanism to bring the correct meaning out of the data

Value

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

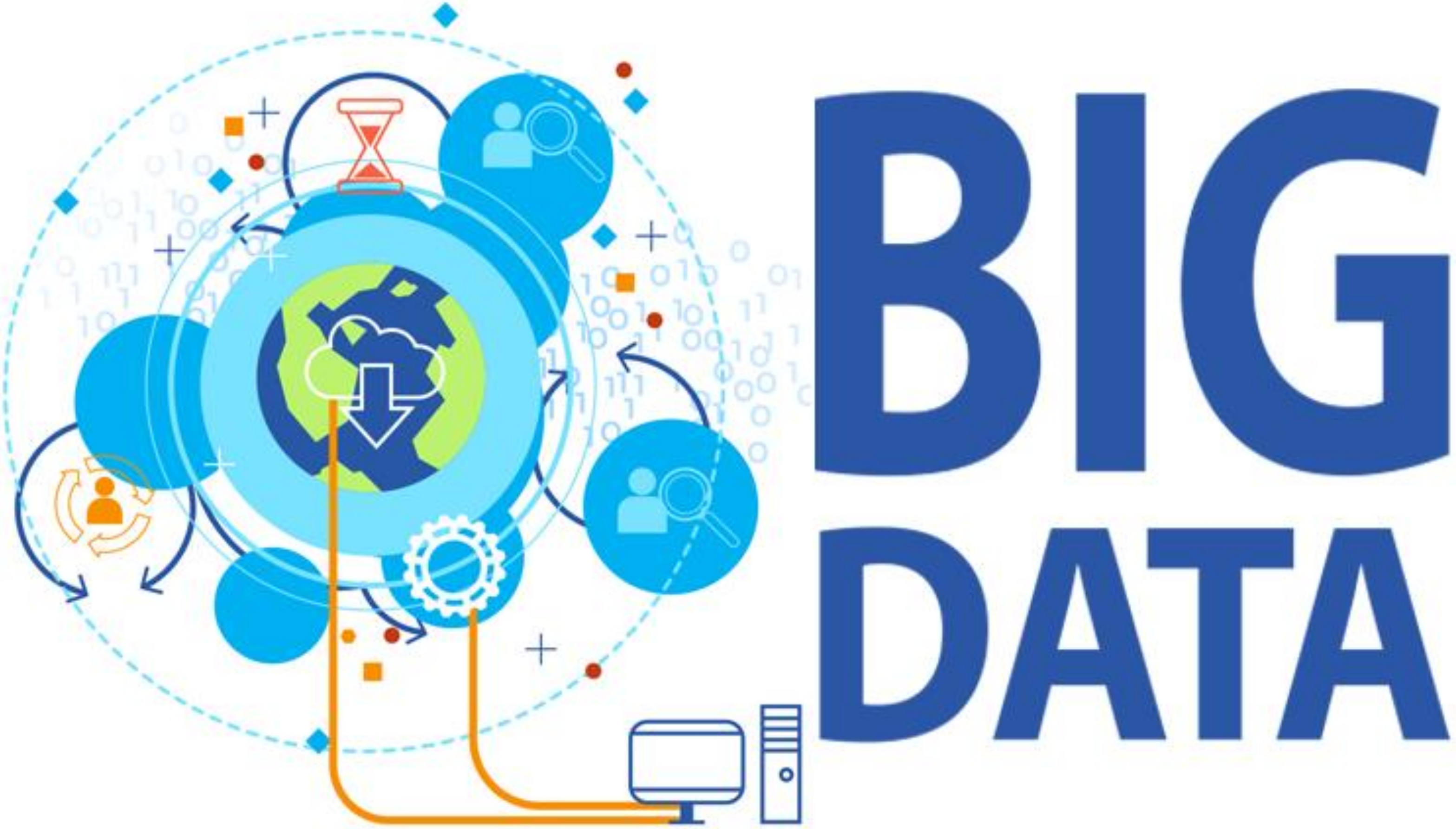
Uncertainty and inconsistencies in the data

Veracity

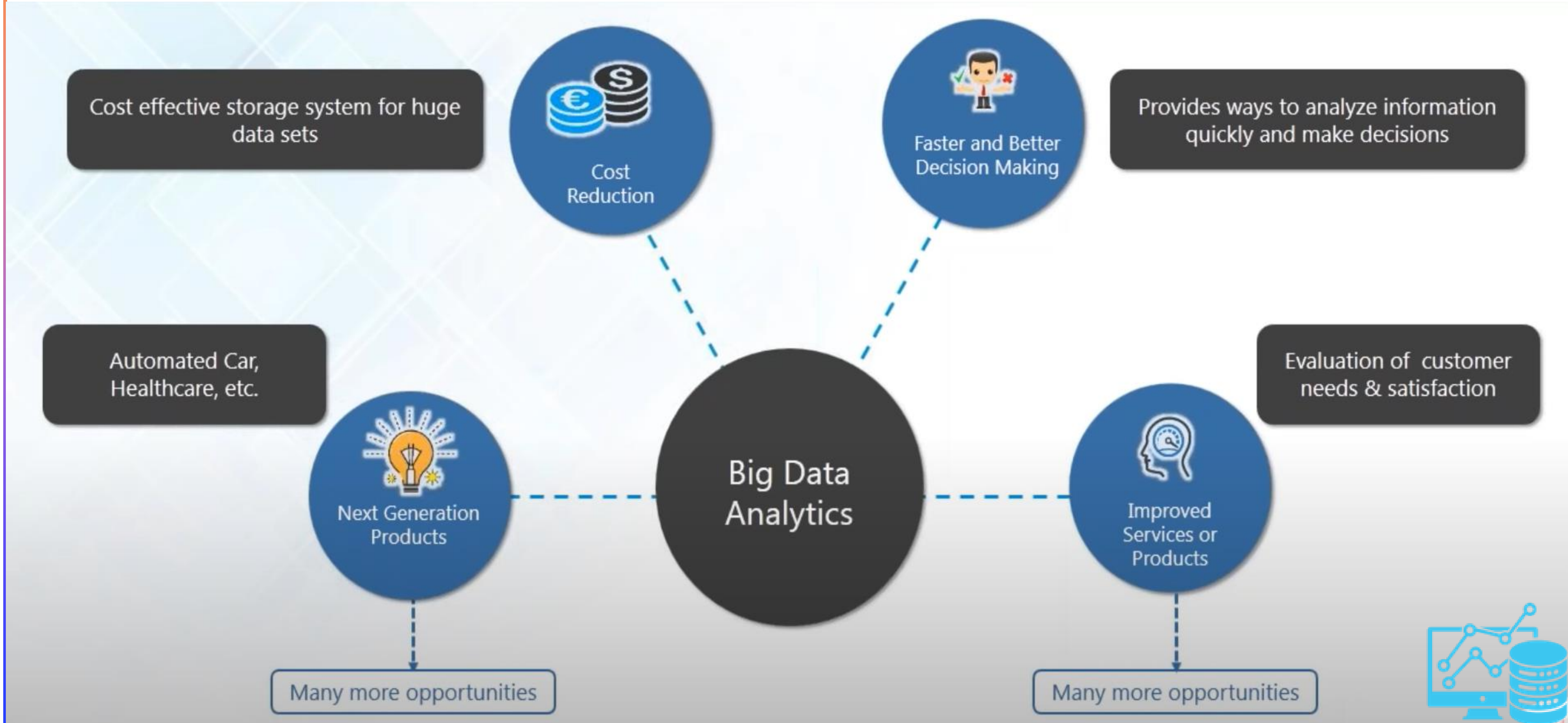
... ...

V's associated with Big Data may grow with time

Big Data as an Oppotunity



Big Data as an Oppotunity



Big Data Analytics

Cost Reduction

Cost effective storage system for huge data sets

Faster and Better Decision Making

Provides ways to analyze information quickly and make decisions

Next Generation Products

Automated Car, Healthcare, etc.

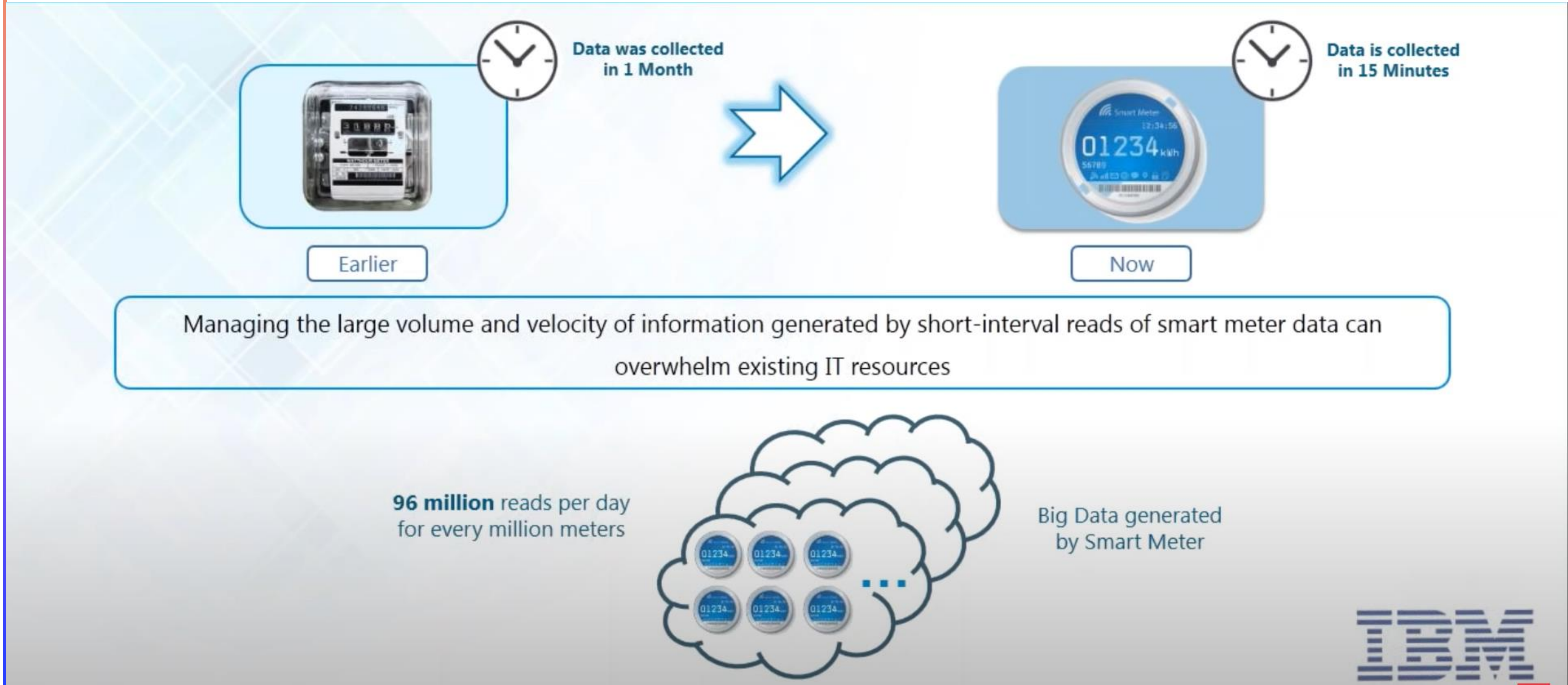
Many more opportunities

Improved Services or Products

Evaluation of customer needs & satisfaction

Many more opportunities

Big Data Collected by Smart Meter



Data was collected in 1 Month

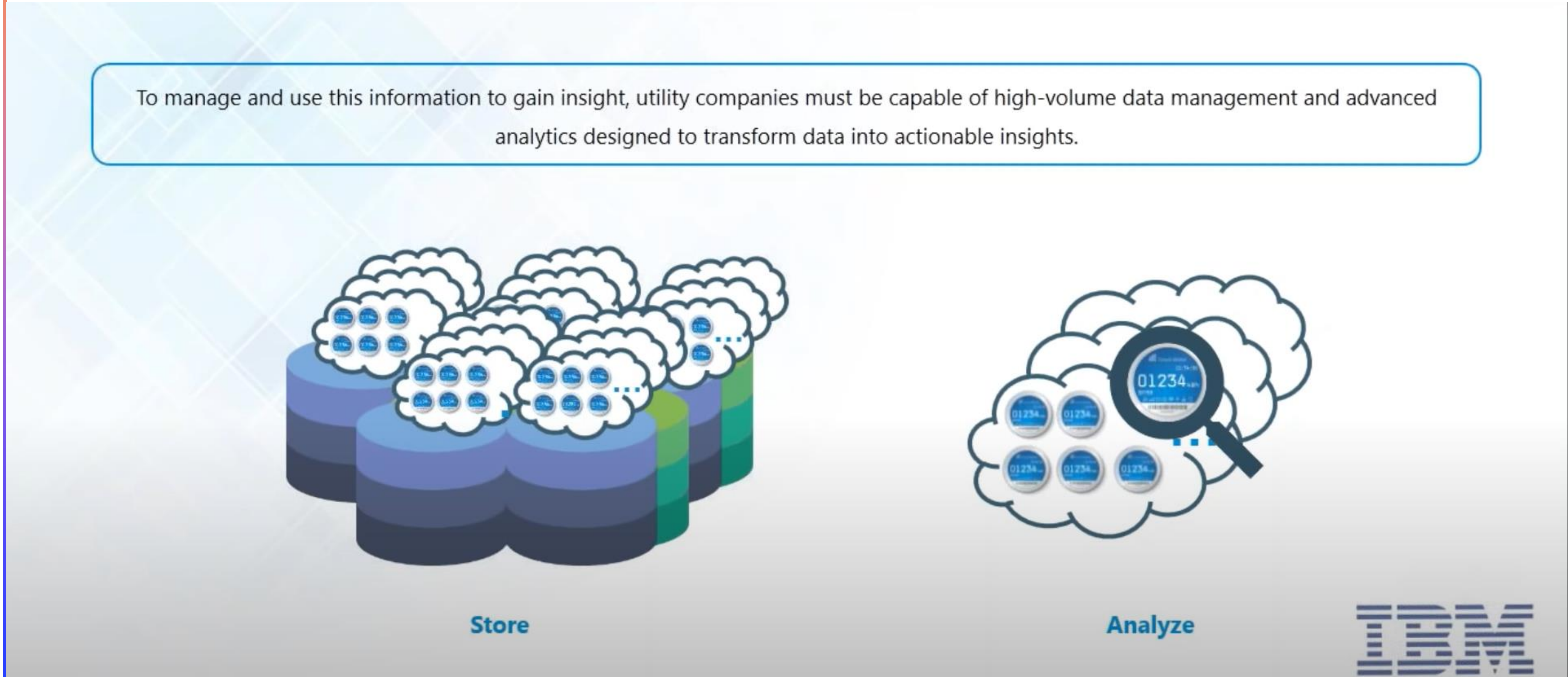
Data is collected in 15 Minutes

Managing the large volume and velocity of information generated by short-interval reads of smart meter data can overwhelm existing IT resources

96 million reads per day for every million meters

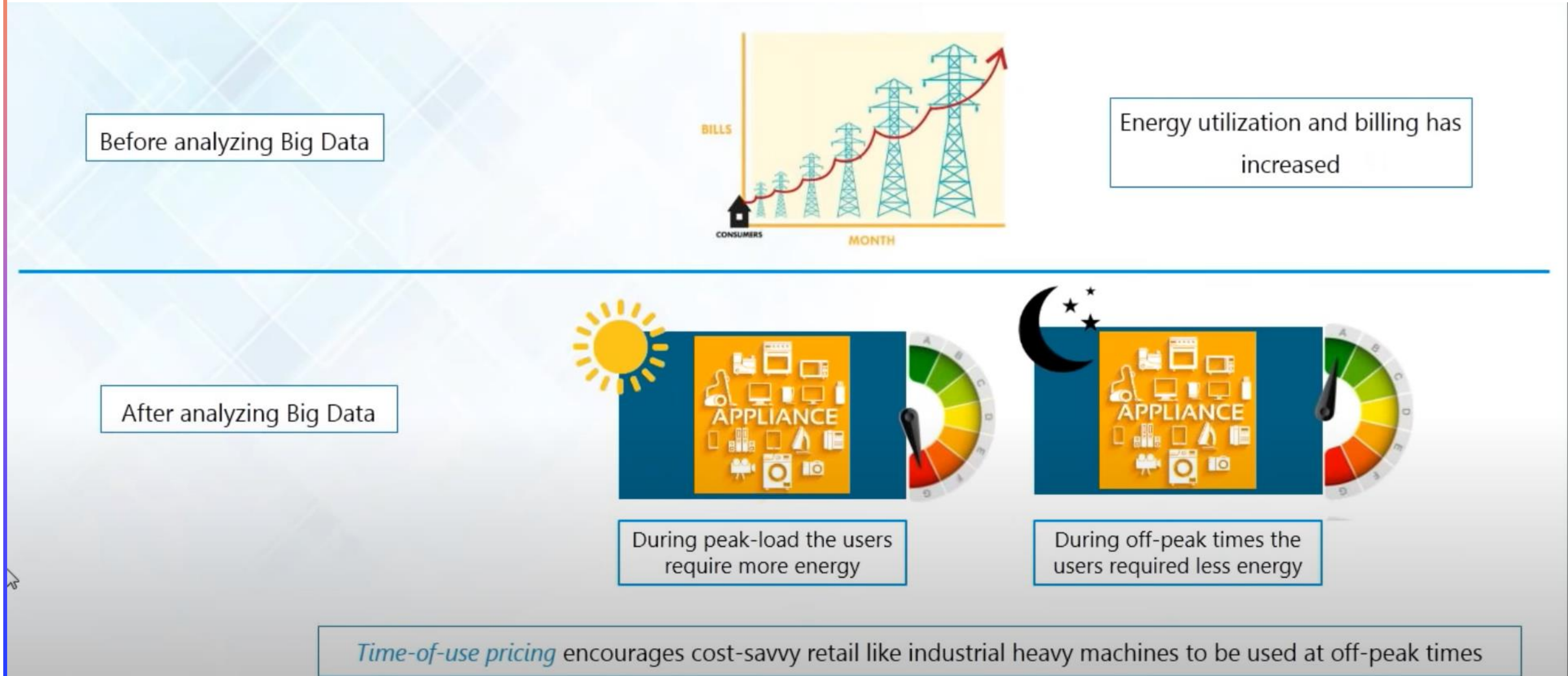
Big Data generated by Smart Meter

Problem with Smart Meter Big Data



To manage and use this information to gain insight, utility companies must be capable of high-volume data management and advanced analytics designed to transform data into actionable insights.

How Smart Meter Big Data Is Analysed



Before analyzing Big Data

Energy utilization and billing has increased

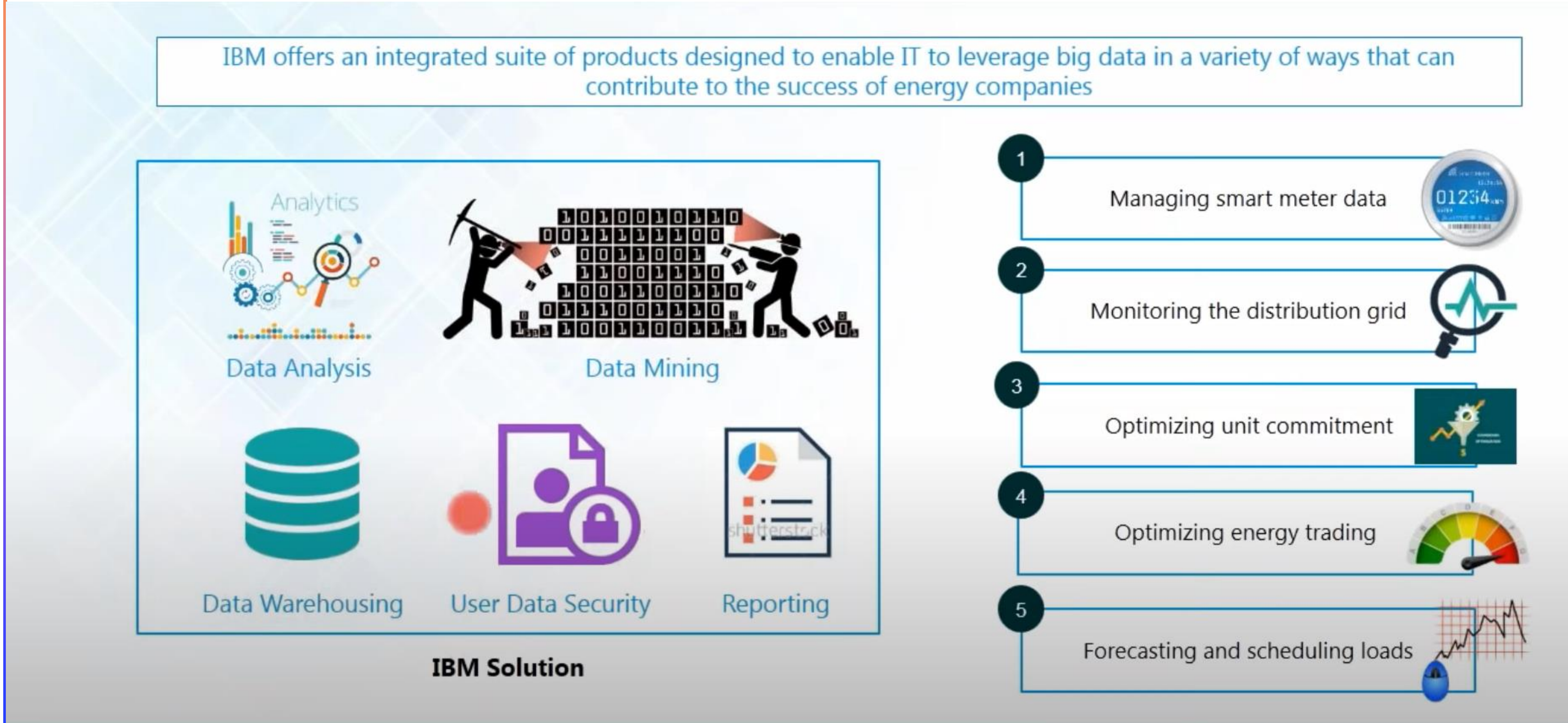
After analyzing Big Data

During peak-load the users require more energy

During off-peak times the users required less energy

Time-of-use pricing encourages cost-savvy retail like industrial heavy machines to be used at off-peak times

IBM Smart Meter Solution



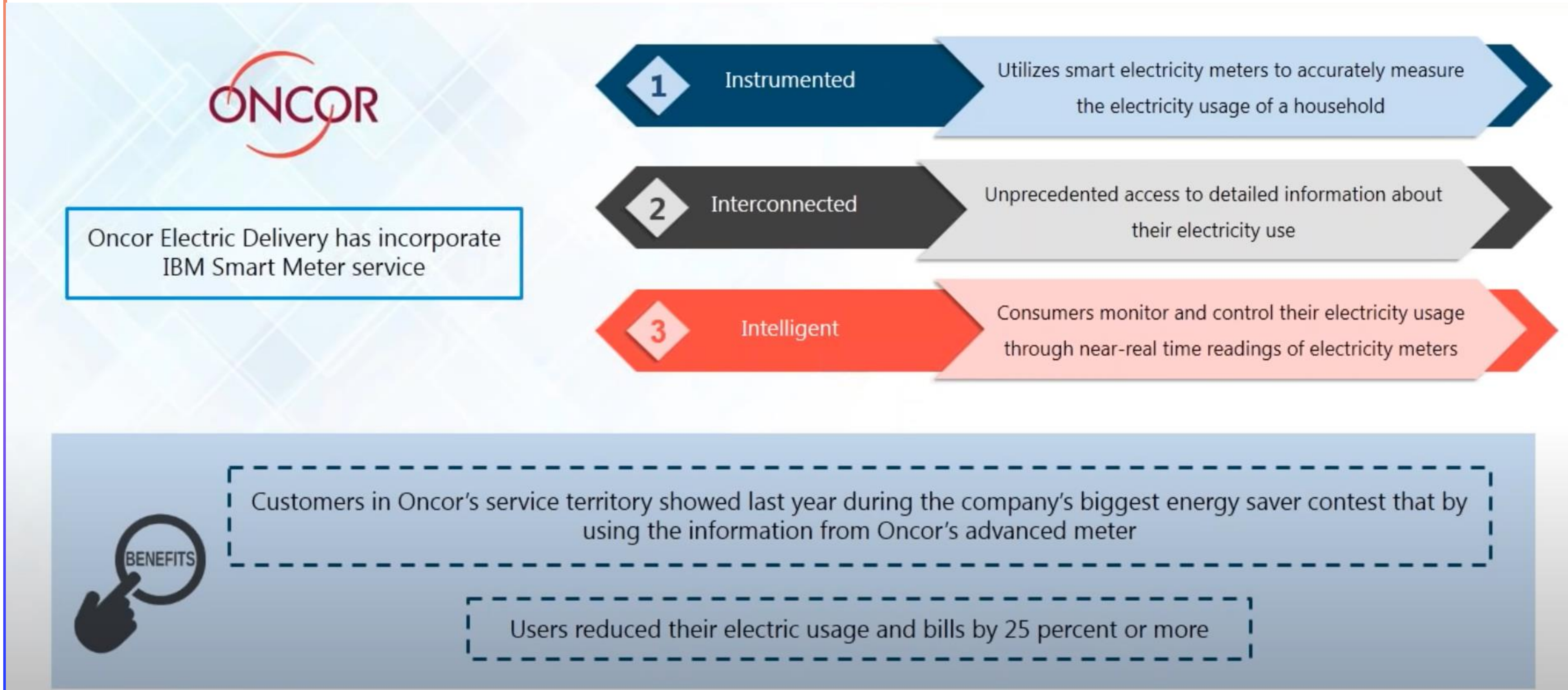
IBM offers an integrated suite of products designed to enable IT to leverage big data in a variety of ways that can contribute to the success of energy companies

- Data Analysis
- Data Mining
- Data Warehousing
- User Data Security
- Reporting

IBM Solution

1. Managing smart meter data
2. Monitoring the distribution grid
3. Optimizing unit commitment
4. Optimizing energy trading
5. Forecasting and scheduling loads

ONCOR using IBM Smart Meter Solution



ONCOR

Oncor Electric Delivery has incorporate IBM Smart Meter service

1 Instrumented — Utilizes smart electricity meters to accurately measure the electricity usage of a household

2 Interconnected — Unprecedented access to detailed information about their electricity use

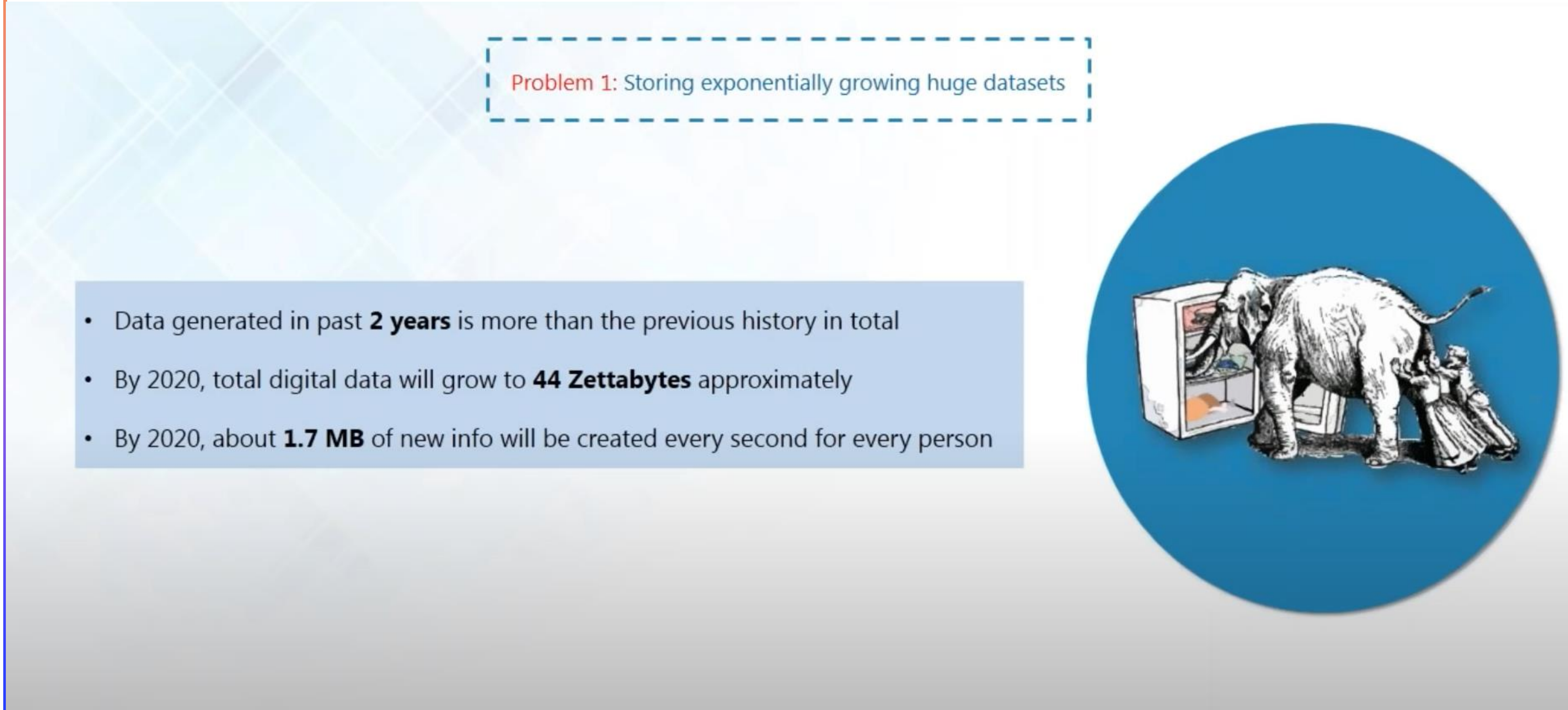
3 Intelligent — Consumers monitor and control their electricity usage through near-real time readings of electricity meters

BENEFITS

Customers in Oncor's service territory showed last year during the company's biggest energy saver contest that by using the information from Oncor's advanced meter

Users reduced their electric usage and bills by 25 percent or more

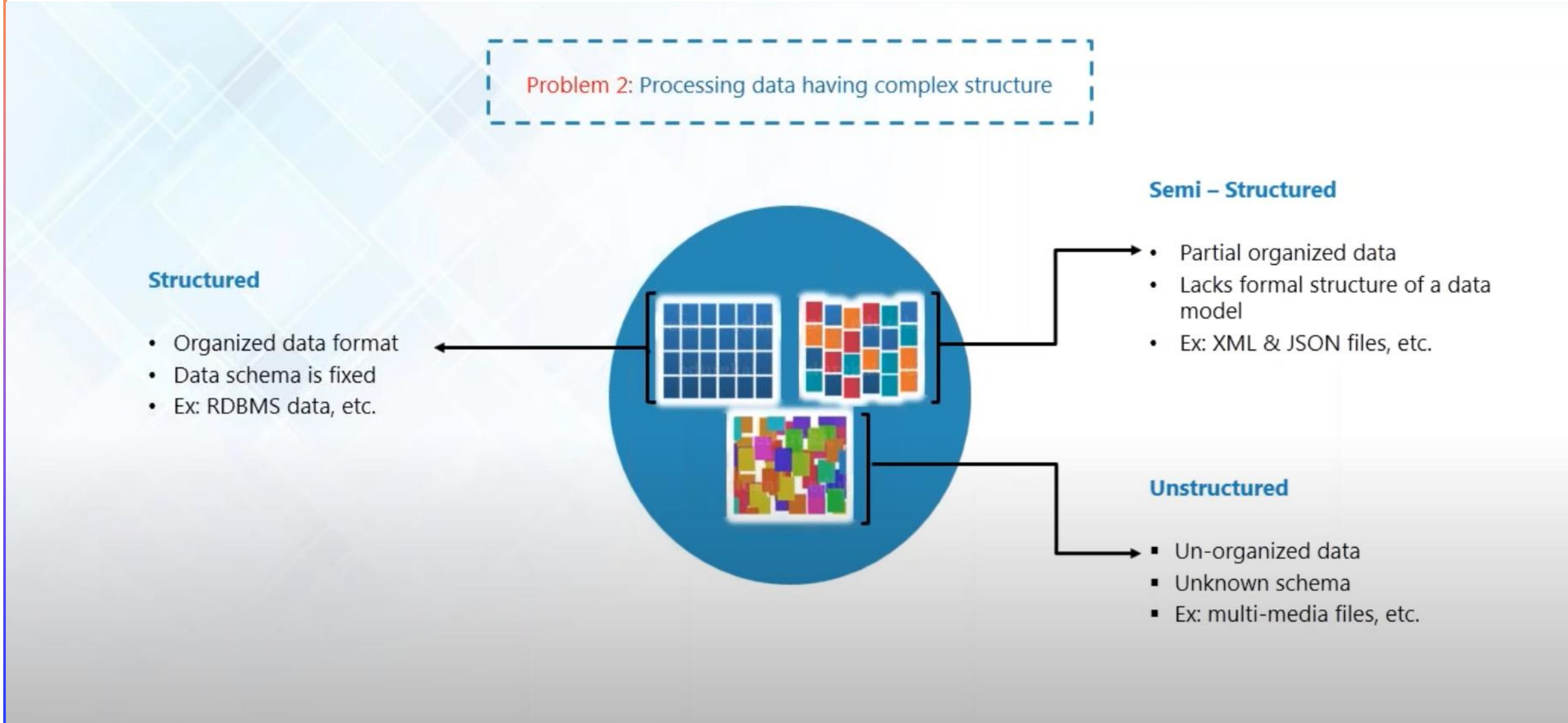
Propblems with Big Data



Problem 1: Storing exponentially growing huge datasets

- Data generated in past **2 years** is more than the previous history in total
- By 2020, total digital data will grow to **44 Zettabytes** approximately
- By 2020, about **1.7 MB** of new info will be created every second for every person

Propblems with Big Data



Problem 2: Processing data having complex structure

Structured

- Organized data format
- Data schema is fixed
- Ex: RDBMS data, etc.

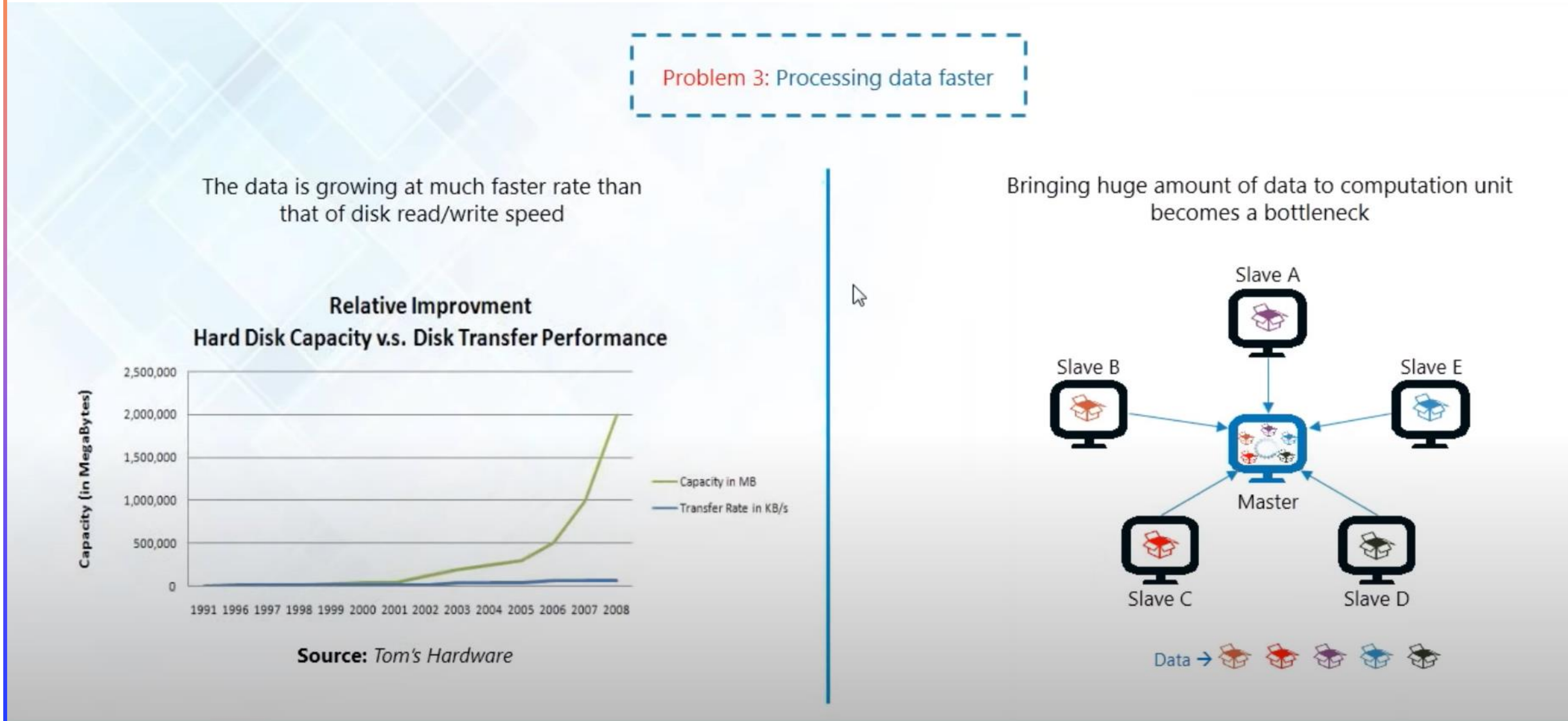
Semi – Structured

- Partial organized data
- Lacks formal structure of a data model
- Ex: XML & JSON files, etc.

Unstructured

- Un-organized data
- Unknown schema
- Ex: multi-media files, etc.

Propblems with Big Data



Problem 3: Processing data faster

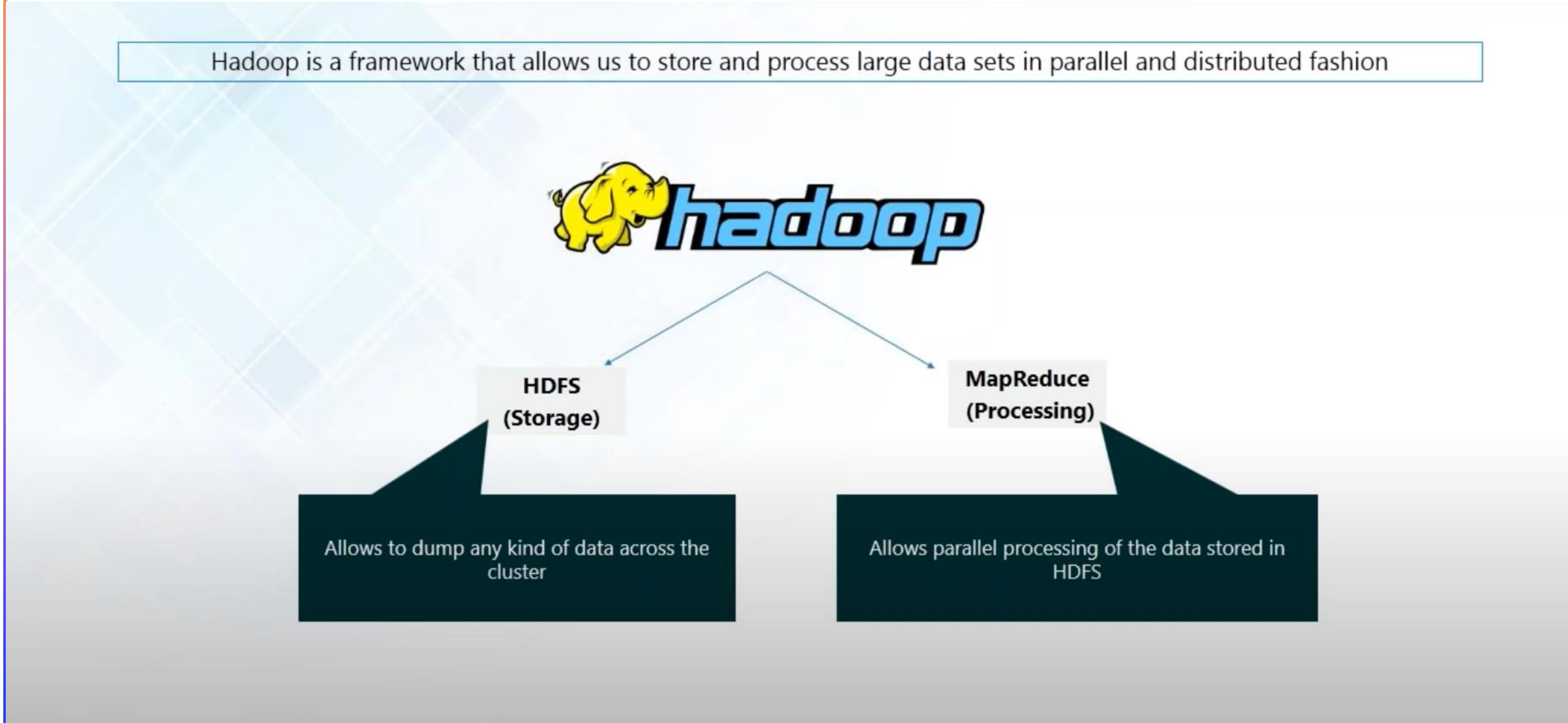
The data is growing at much faster rate than that of disk read/write speed

Relative Improvment
Hard Disk Capacity v.s. Disk Transfer Performance

Source: *Tom's Hardware*

Bringing huge amount of data to computation unit becomes a bottleneck

Hadoop



Hadoop is a framework that allows us to store and process large data sets in parallel and distributed fashion

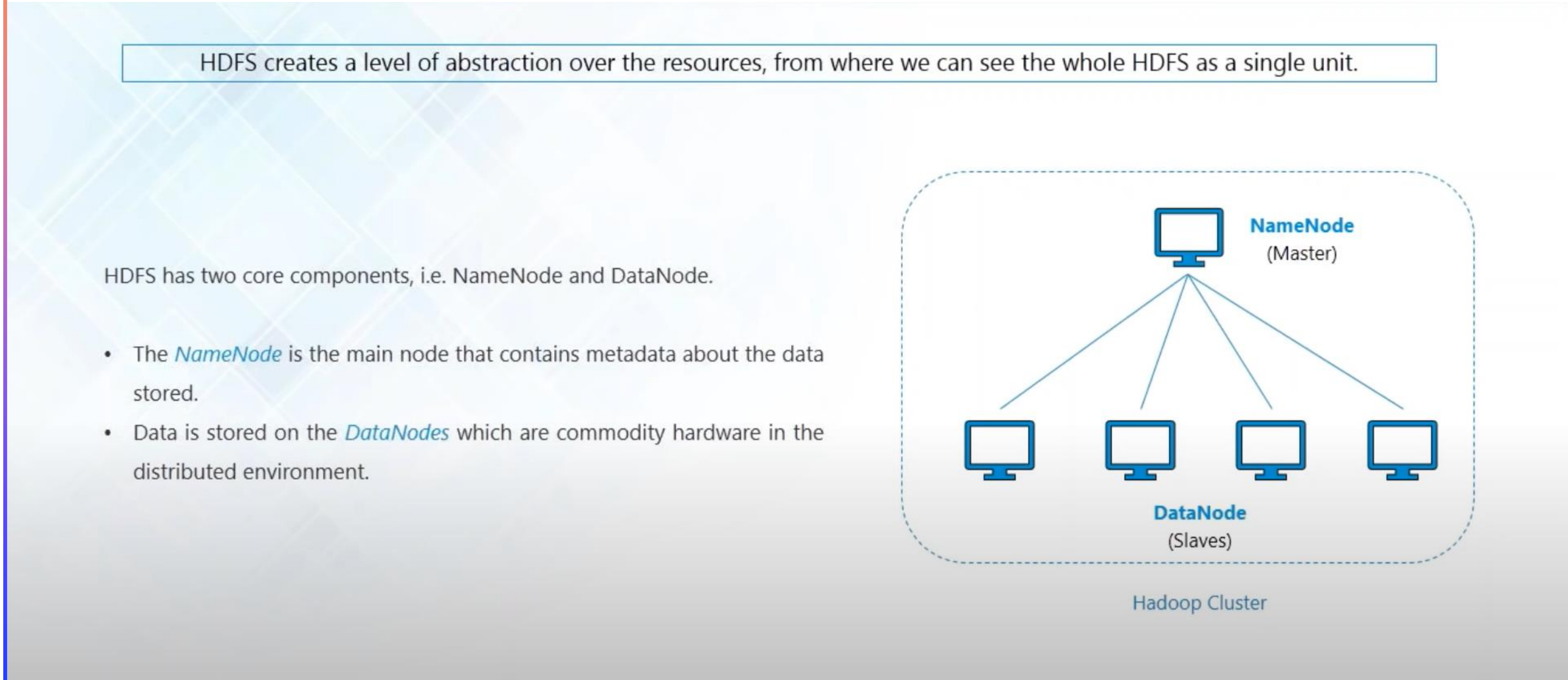
HDFS
(Storage)

Allows to dump any kind of data across the cluster

MapReduce
(Processing)

Allows parallel processing of the data stored in HDFS

Hadoop Distributed File System

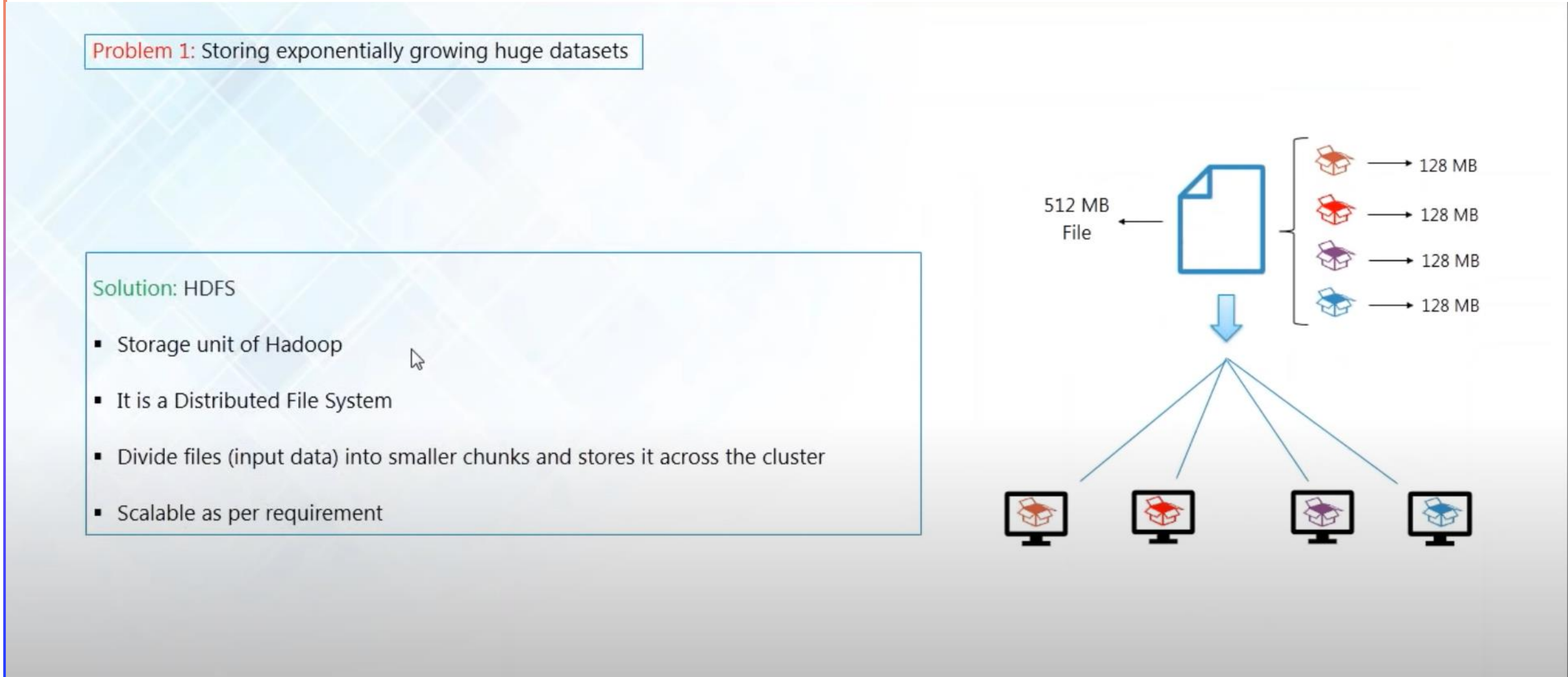


HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit.

HDFS has two core components, i.e. NameNode and DataNode.

- The *NameNode* is the main node that contains metadata about the data stored.
- Data is stored on the *DataNodes* which are commodity hardware in the distributed environment.

Storing Data (Solution)

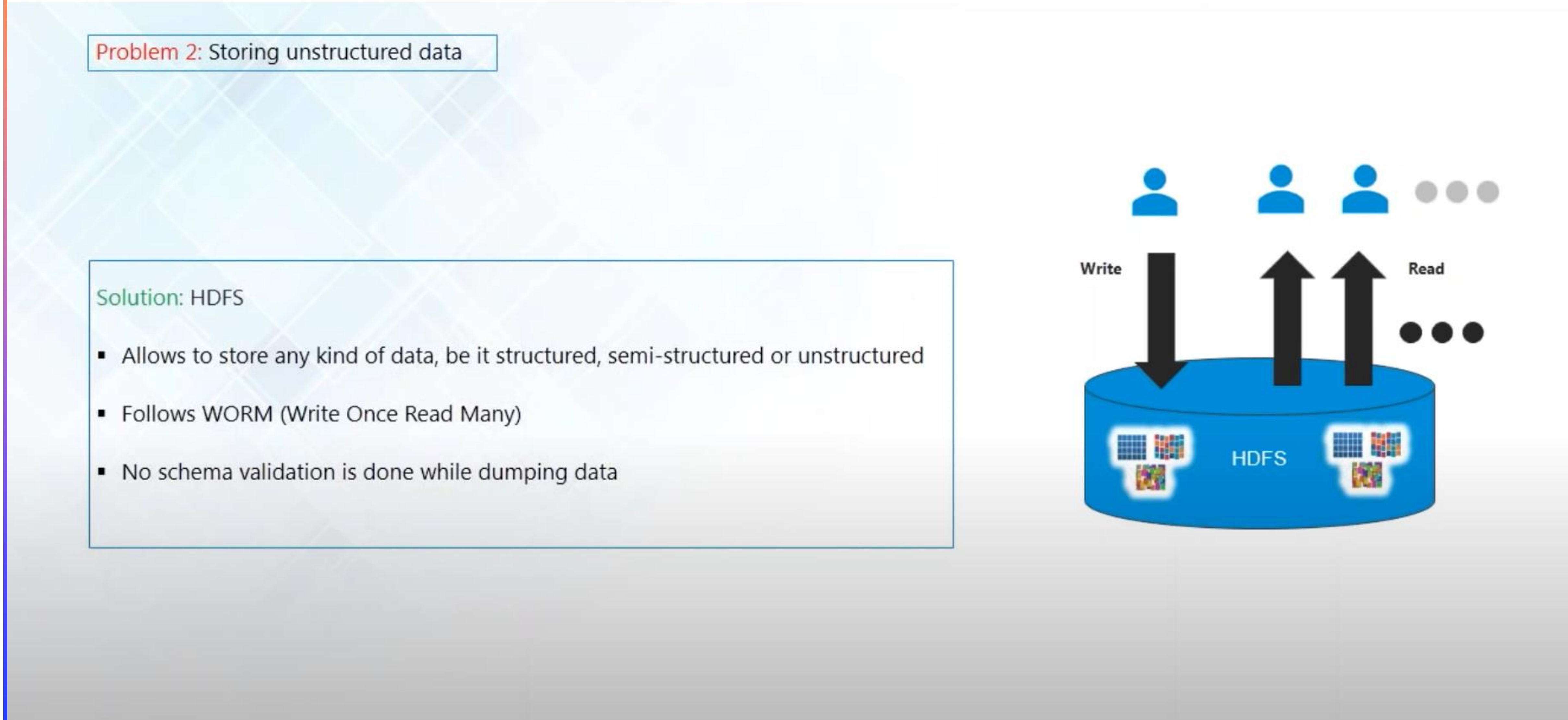


Problem 1: Storing exponentially growing huge datasets

Solution: HDFS

- Storage unit of Hadoop
- It is a Distributed File System
- Divide files (input data) into smaller chunks and stores it across the cluster
- Scalable as per requirement

Store Different Kinds Of Data (Solution)

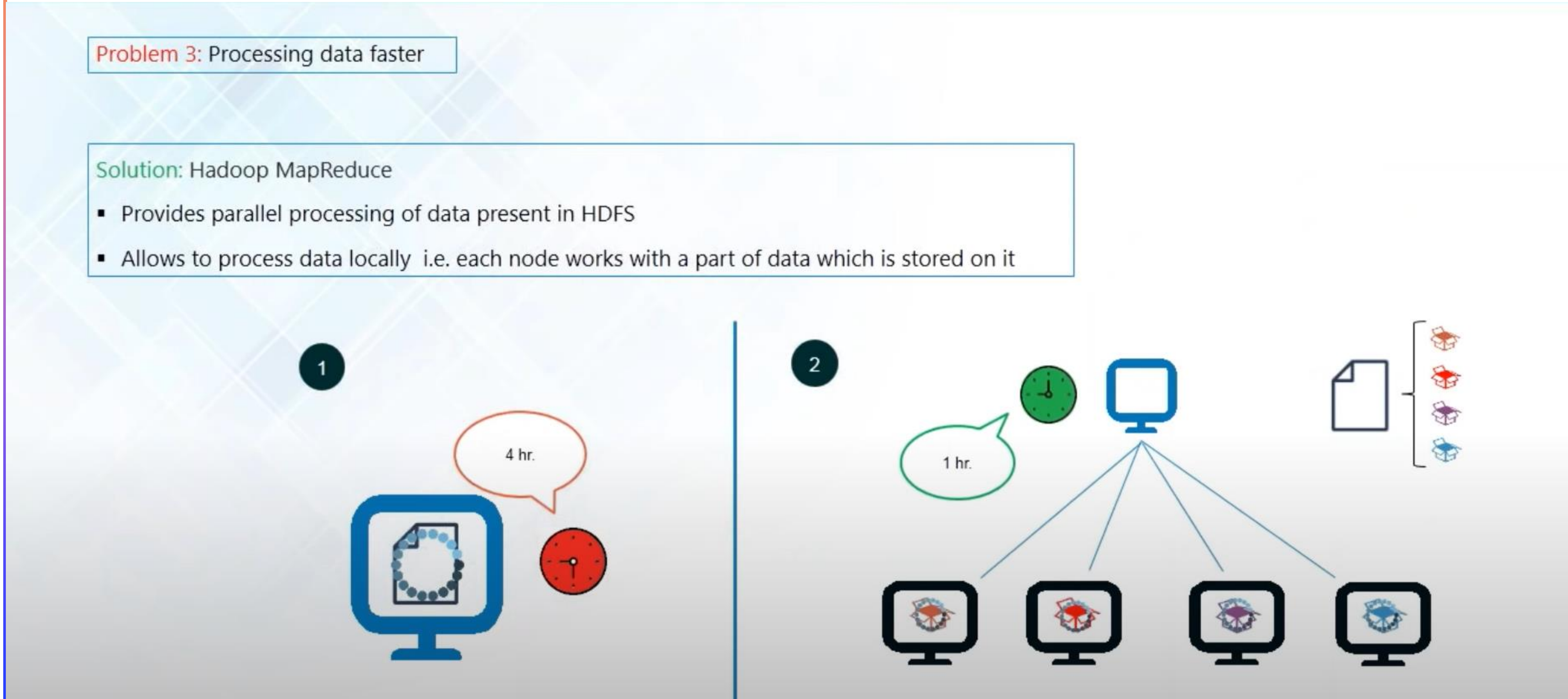


Problem 2: Storing unstructured data

Solution: HDFS

- Allows to store any kind of data, be it structured, semi-structured or unstructured
- Follows WORM (Write Once Read Many)
- No schema validation is done while dumping data

Processing Data Faster (Solution)

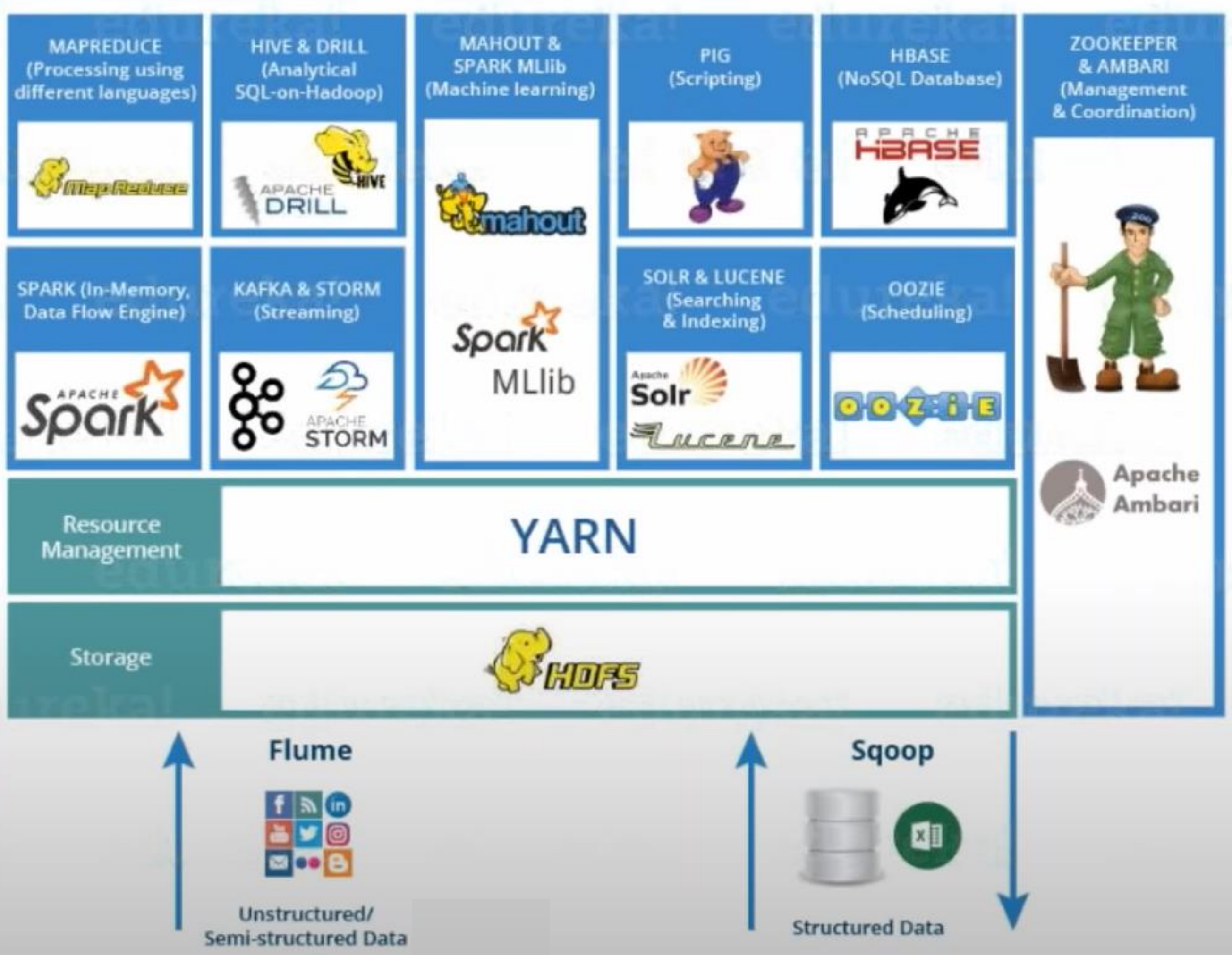


Problem 3: Processing data faster

Solution: Hadoop MapReduce

- Provides parallel processing of data present in HDFS
- Allows to process data locally i.e. each node works with a part of data which is stored on it

Hadoop Ecosystem



MAPREDUCE (Processing using different languages)

HIVE & DRILL (Analytical SQL-on-Hadoop)

MAHOUT & SPARK MLlib (Machine learning)

PIG (Scripting)

HBASE (NoSQL Database)

ZOOKEEPER & AMBARI (Management & Coordination)

SPARK (In-Memory, Data Flow Engine)

KAFKA & STORM (Streaming)

SOLR & LUCENE (Searching & Indexing)

OOZIE (Scheduling)

Resource Management

YARN

Storage

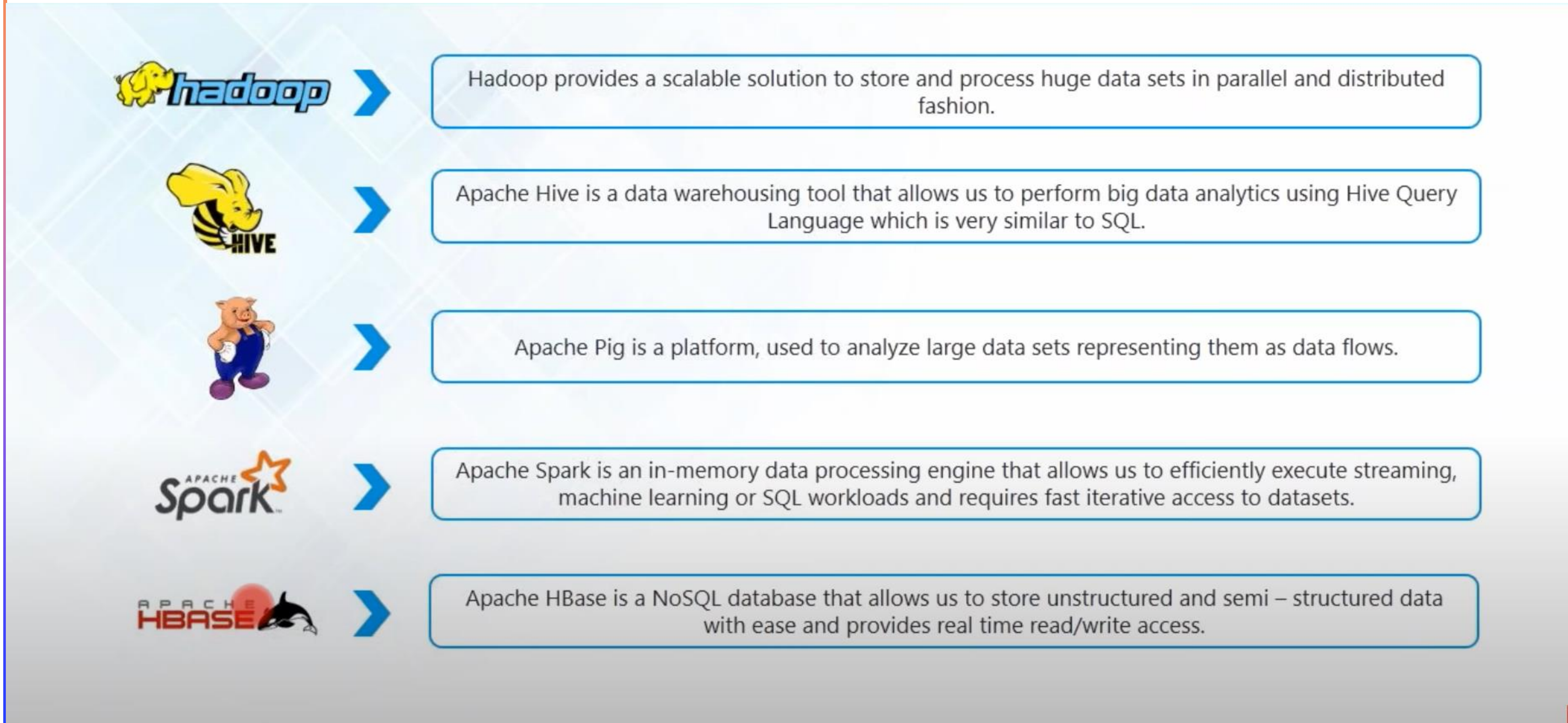
Flume

Unstructured/ Semi-structured Data

Sqoop

Structured Data

Hadoop Ecosystem



Hadoop provides a scalable solution to store and process huge data sets in parallel and distributed fashion.

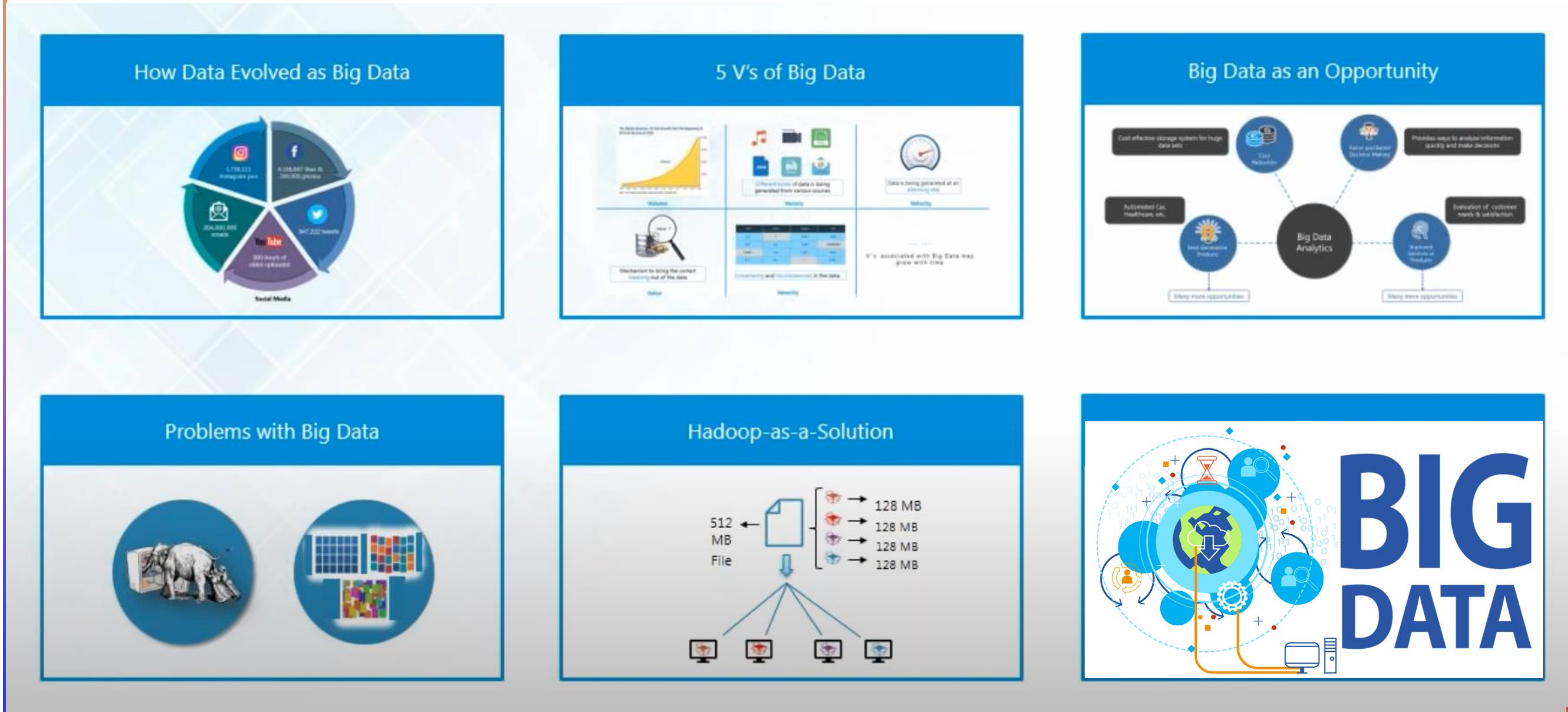
Apache Hive is a data warehousing tool that allows us to perform big data analytics using Hive Query Language which is very similar to SQL.

Apache Pig is a platform, used to analyze large data sets representing them as data flows.

Apache Spark is an in-memory data processing engine that allows us to efficiently execute streaming, machine learning or SQL workloads and requires fast iterative access to datasets.

Apache HBase is a NoSQL database that allows us to store unstructured and semi – structured data with ease and provides real time read/write access.

Session In A Minute



How Data Evolved as Big Data

1,738,111 Instagram pics
4,166,667 likes & 200,000 photos
347,222 tweets
300 hours of video uploaded
204,000,000 emails
Social Media

5 V's of Big Data

Volume
Different kinds of data is being generated from various sources
Variety
Data is being generated at an alarming rate
Velocity
Mechanism to bring the correct meaning out of the data
Value
Uncertainty and inconsistencies in the data
Veracity
V's associated with Big Data may grow with time

Big Data as an Opportunity

Cost effective storage system for huge data sets
Cost Reduction
Faster and Better Decision Making
Provides ways to analyze information quickly and make decisions
Automated Car, Healthcare, etc.
Next Generation Products
Big Data Analytics
Improved Services or Products
Evaluation of customer needs & satisfaction
Many more opportunities
Many more opportunities

Problems with Big Data

Hadoop-as-a-Solution

512 MB File
128 MB
128 MB
128 MB
128 MB