ADT302 - CONCEPTS IN BIG DATA ANALYTICS

Viju P Poonthottam
Asst. Professor
Dept. of AI & DS
MES CE Kuttippuram

February 27, 2023

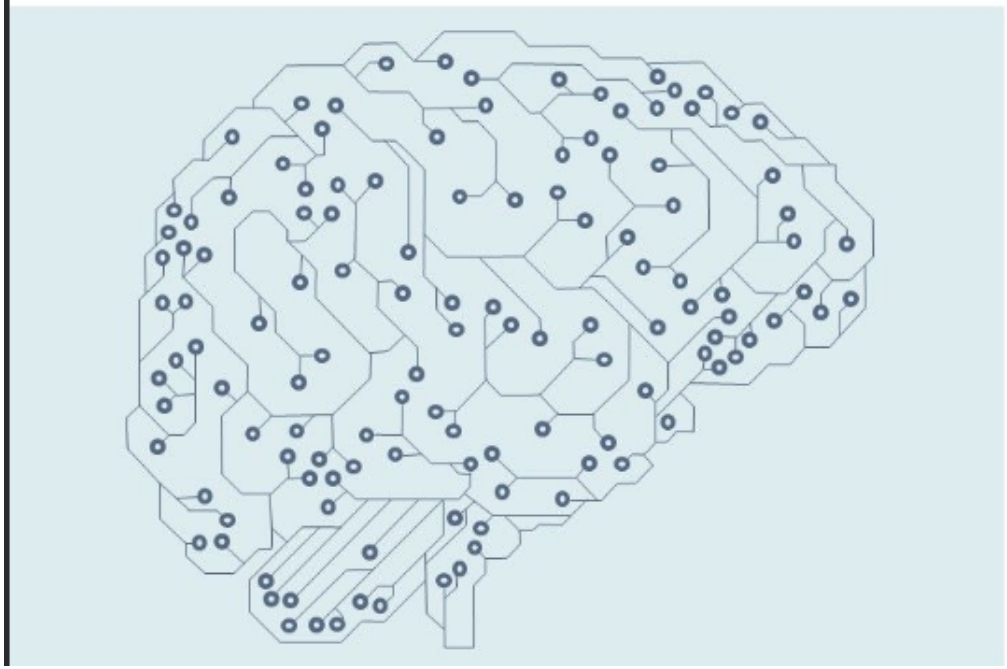
Table of contents

Module 1 -Introduction to Big Data(9 Hrs)

- Introduction to Big data, Conventional Data vs Big data,
- Big data architecture,
- Big data platforms,
- Nature of data,
- Analytic processes and tools,
- 5 V's of Big data,
- Big data analytical method,
- Intelligent data analysis,
- Big data analytics life cycle.

Introduction to Big data

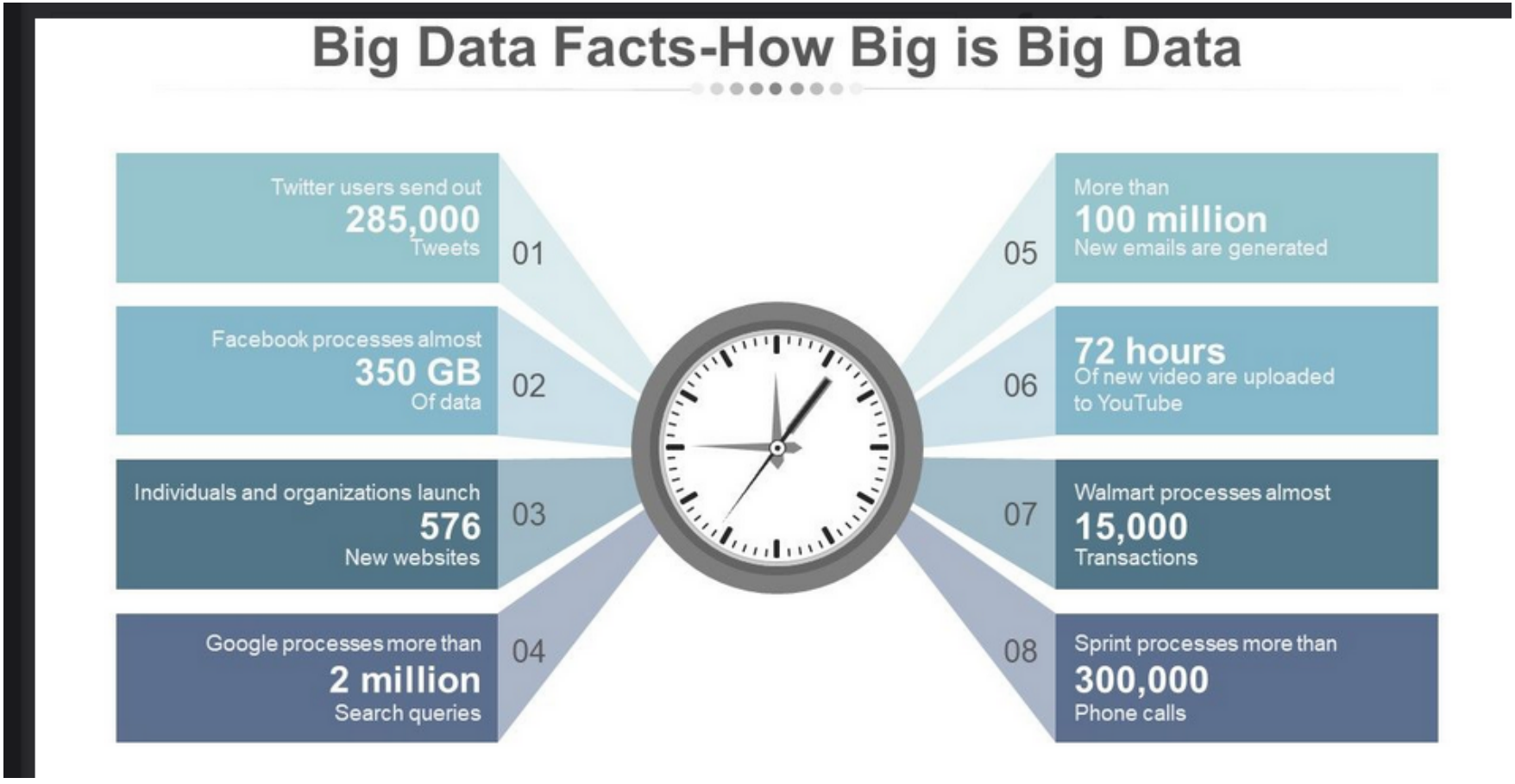
What is Big Data

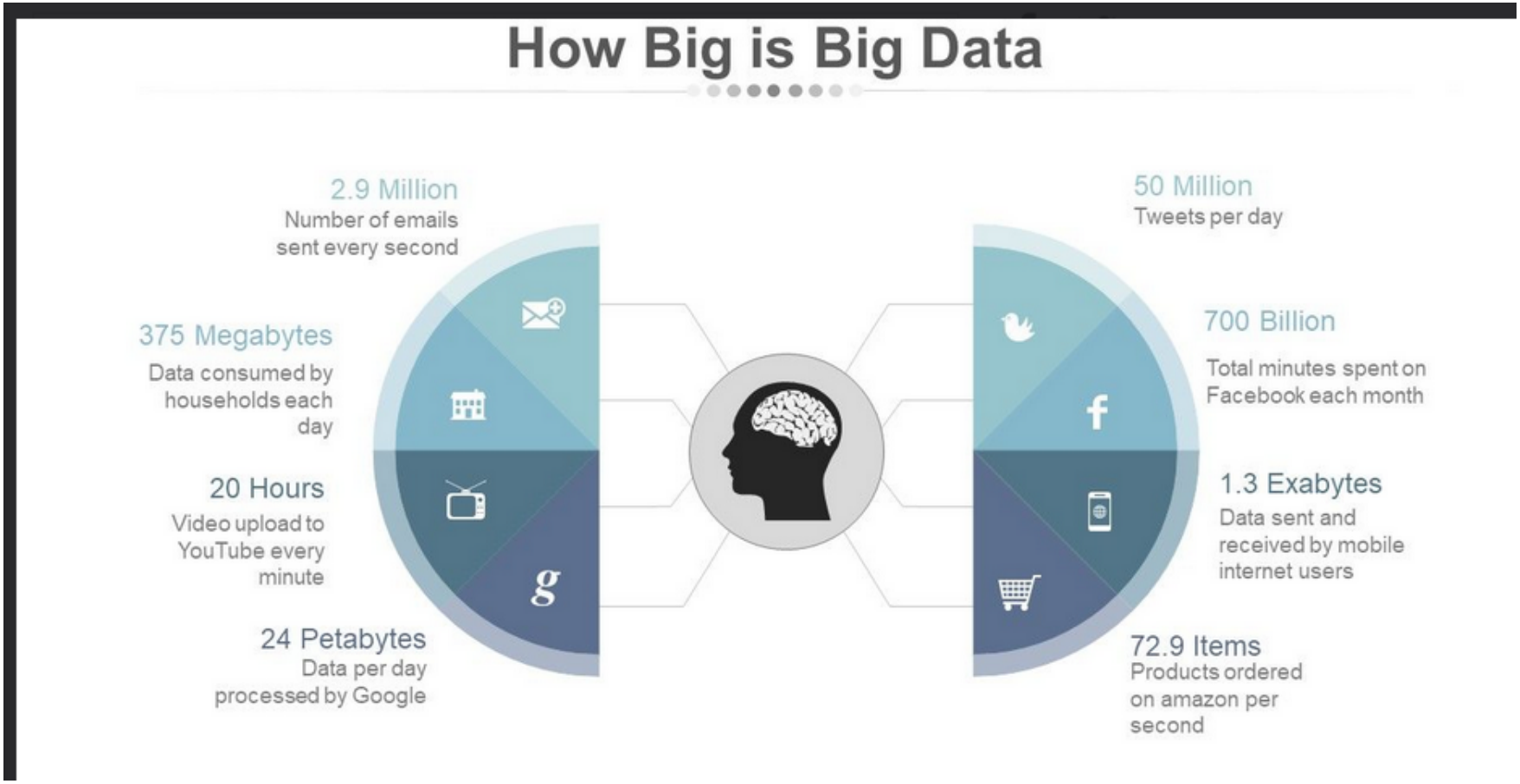


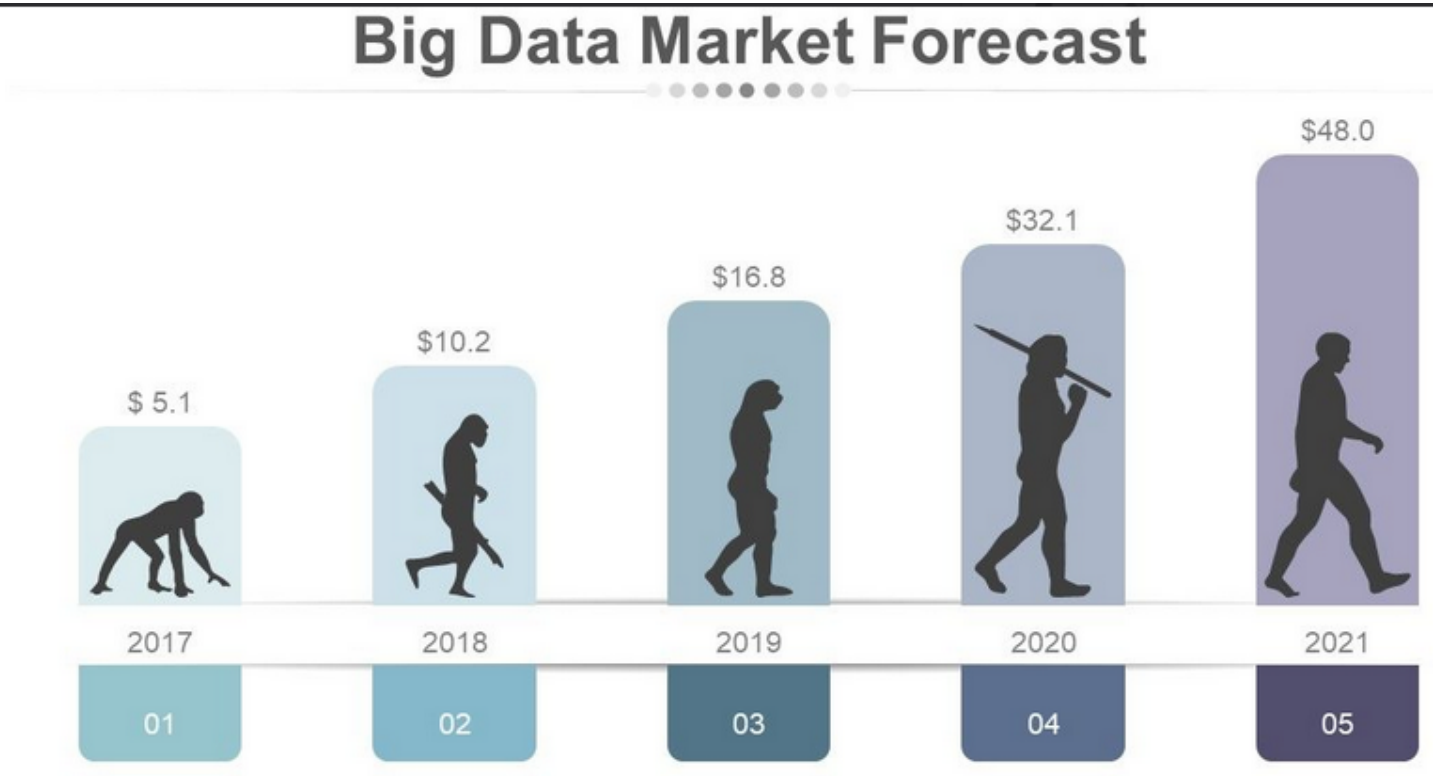
Big Data

Big data is characterized by large volumes of different types of data (e.g. Social, web, transaction, etc.) That builds very quickly.

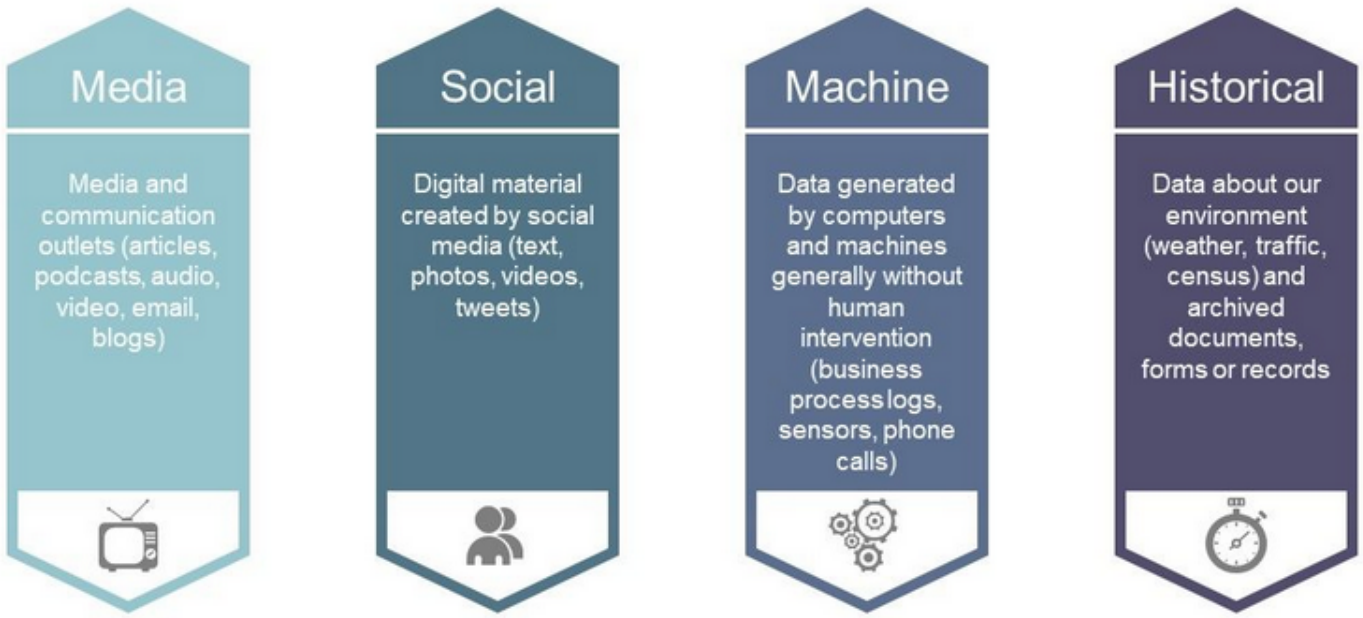
It exceeds the reach of commonly used hardware environments and software tools to capture, manage and process in a timely manner for its users.

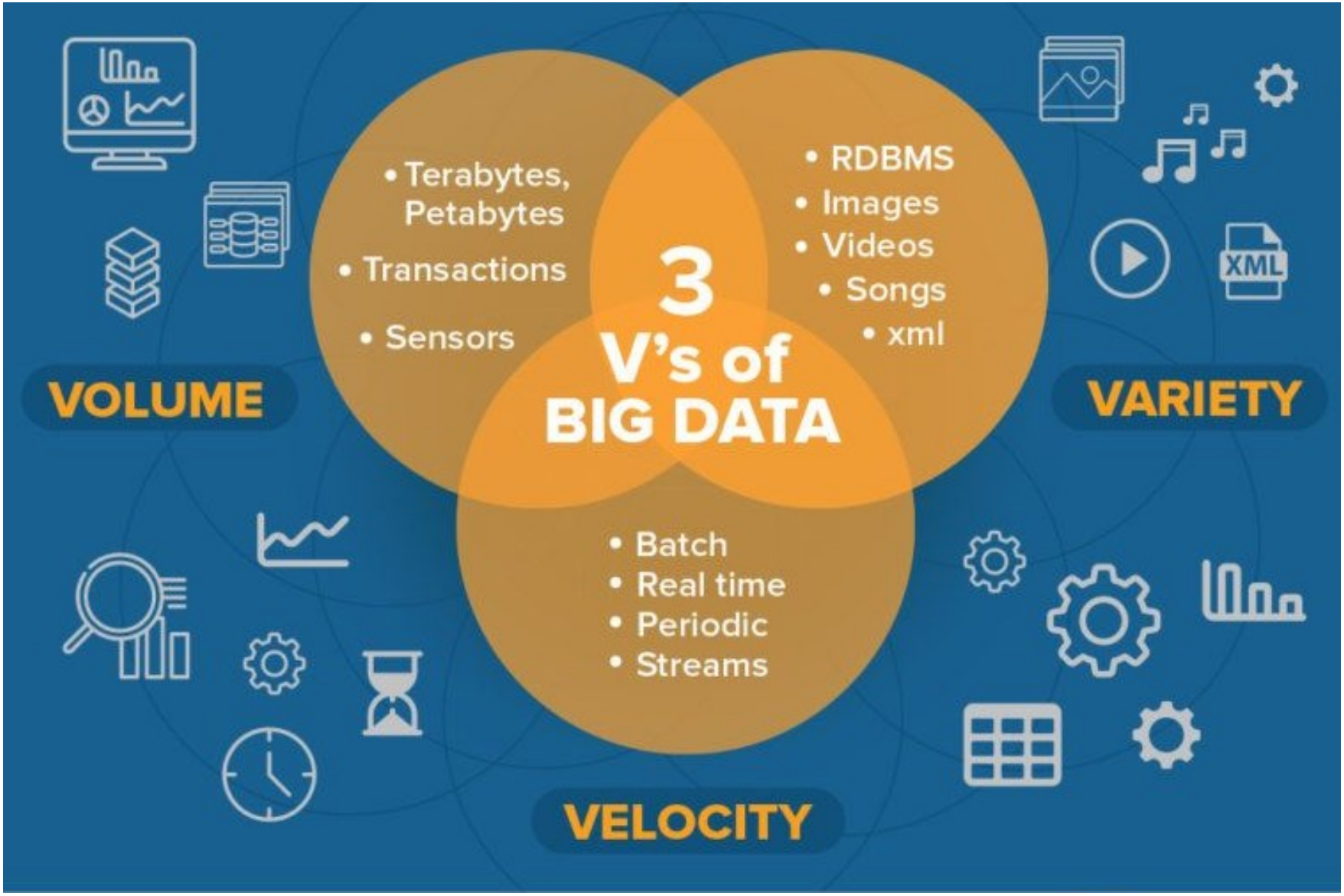

Big Data Facts-How Big is Big Data
Twitter users send out
285,000
Tweets
01
Facebook processes almost
350 GB
Of data
02
Individuals and organizations launch
576
New websites
03
Google processes more than
2 million
Search queries
04
More than
100 million
New emails are generated
05
72 hours
Of new video are uploaded
to YouTube
06
Walmart processes almost
15,000
Transactions
07
Sprint processes more than
300,000
Phone calls
08

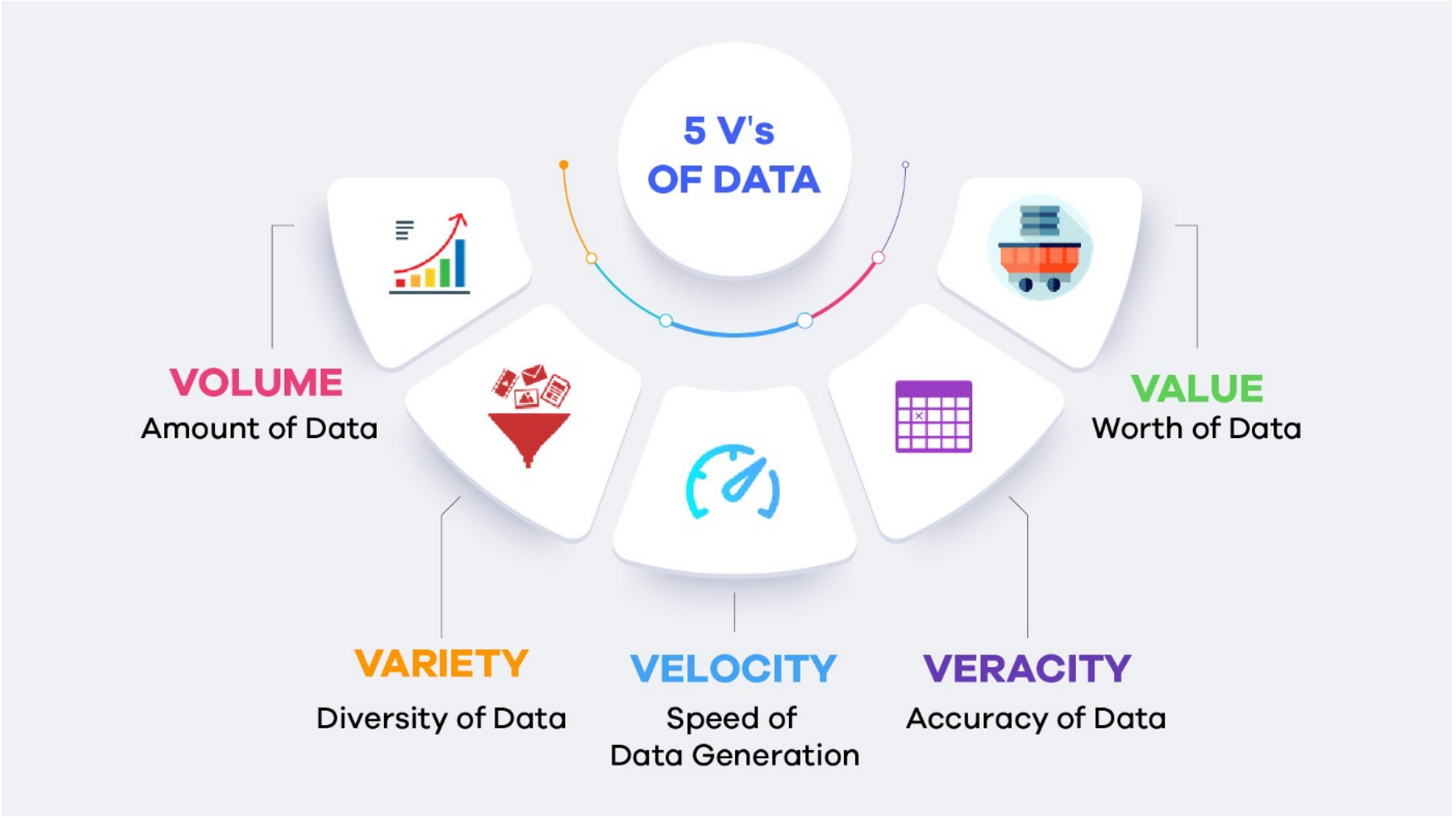

How Big is Big Data
2.9 Million
Number of emails sent every second
375 Megabytes
Data consumed by households each day
20 Hours
Video upload to YouTube every minute
24 Petabytes
Data per day processed by Google
50 Million
Tweets per day
700 Billion
Total minutes spent on Facebook each month
1.3 Exabytes
Data sent and received by mobile internet users
72.9 Items
Products ordered on amazon per second


Big Data Market Forecast
$ 5.1
$10.2
$16.8
$32.1
$48.0
2017
2018
2019
2020
2021
01
02
03
04
05

Sources of Big Data




• Terabytes, Petabytes
• Transactions
• Sensors
VOLUME
3 V's of BIG DATA
• RDBMS
• Images
• Videos
• Songs
• xml
VARIETY
XML
• Batch
• Real time
• Periodic
• Streams
VELOCITY


5 V's
OF DATA
VOLUME
Amount of Data
VALUE
Worth of Data
VARIETY
Diversity of Data
VELOCITY
Speed of
Data Generation
VERACITY
Accuracy of Data

Small Data vs Big Data

Small Data

- Low Volumes
- Batch Velocities
- Structured Varieties

Vs

Big Data

- Into Petabyte Volumes
- Real-time Velocities
- Multistructured Varieties

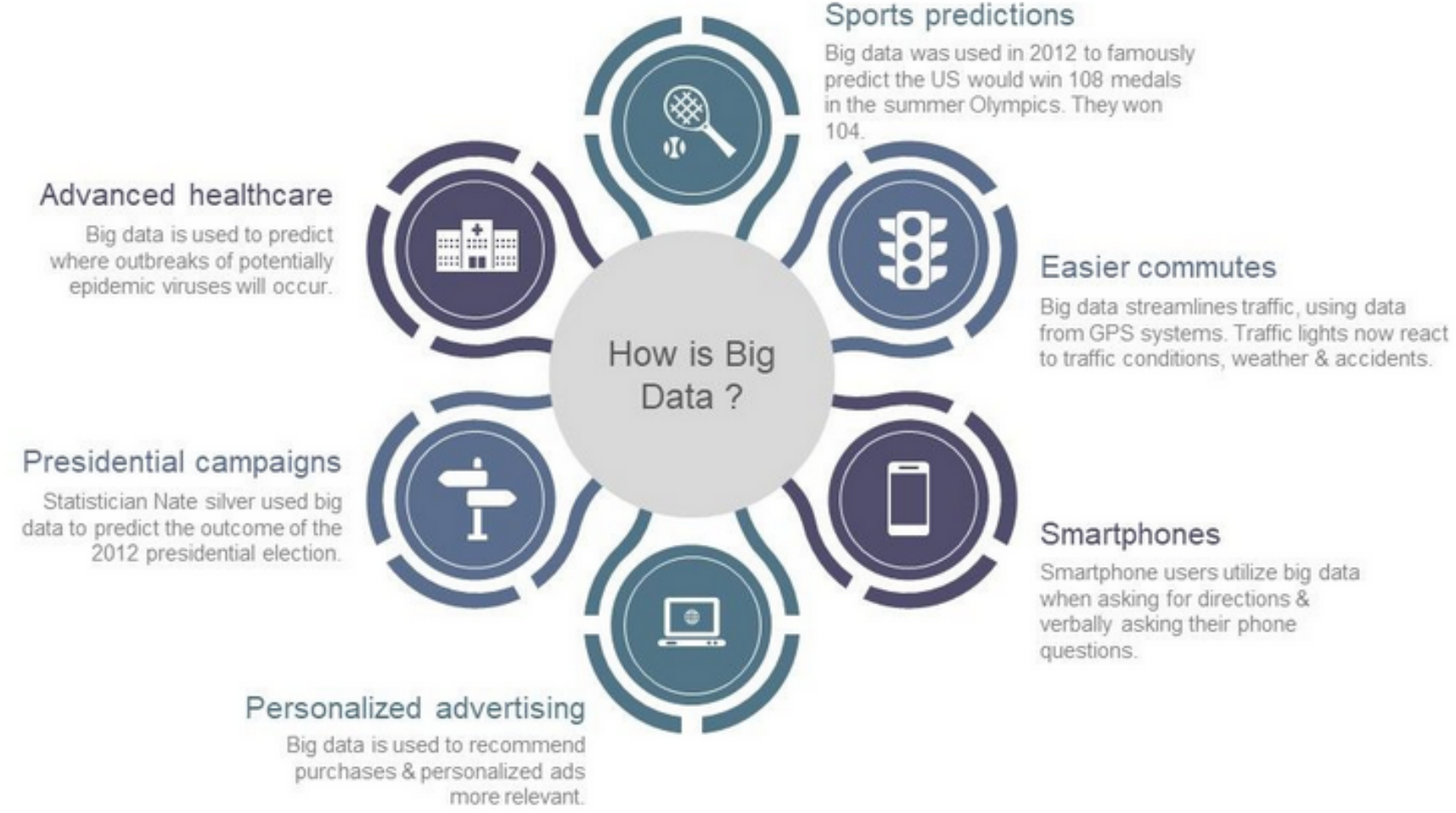
Forms/ Type of Big Data

Data that does not reside in fixed locations generally refers to free-form text, which is ubiquitous.

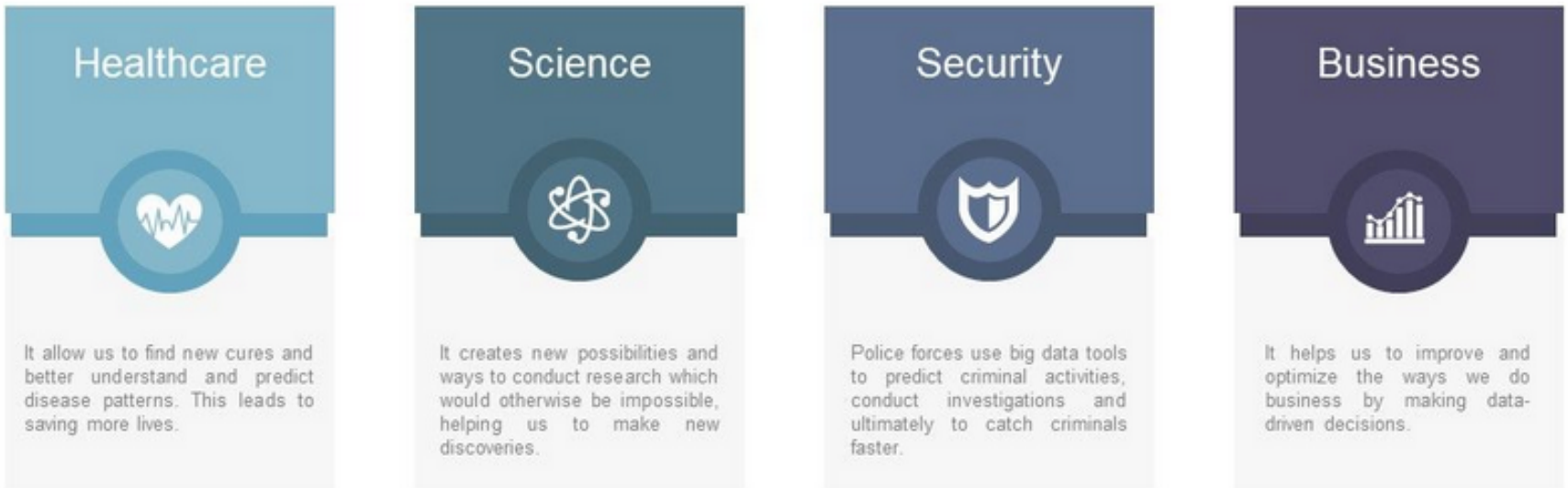
Data that resides in fixed fields within a record or file.

Between the tow forms where "tags" or "structure" are associated or embedded within unstructured data.

Impact of Big Data



Impact of Big Data



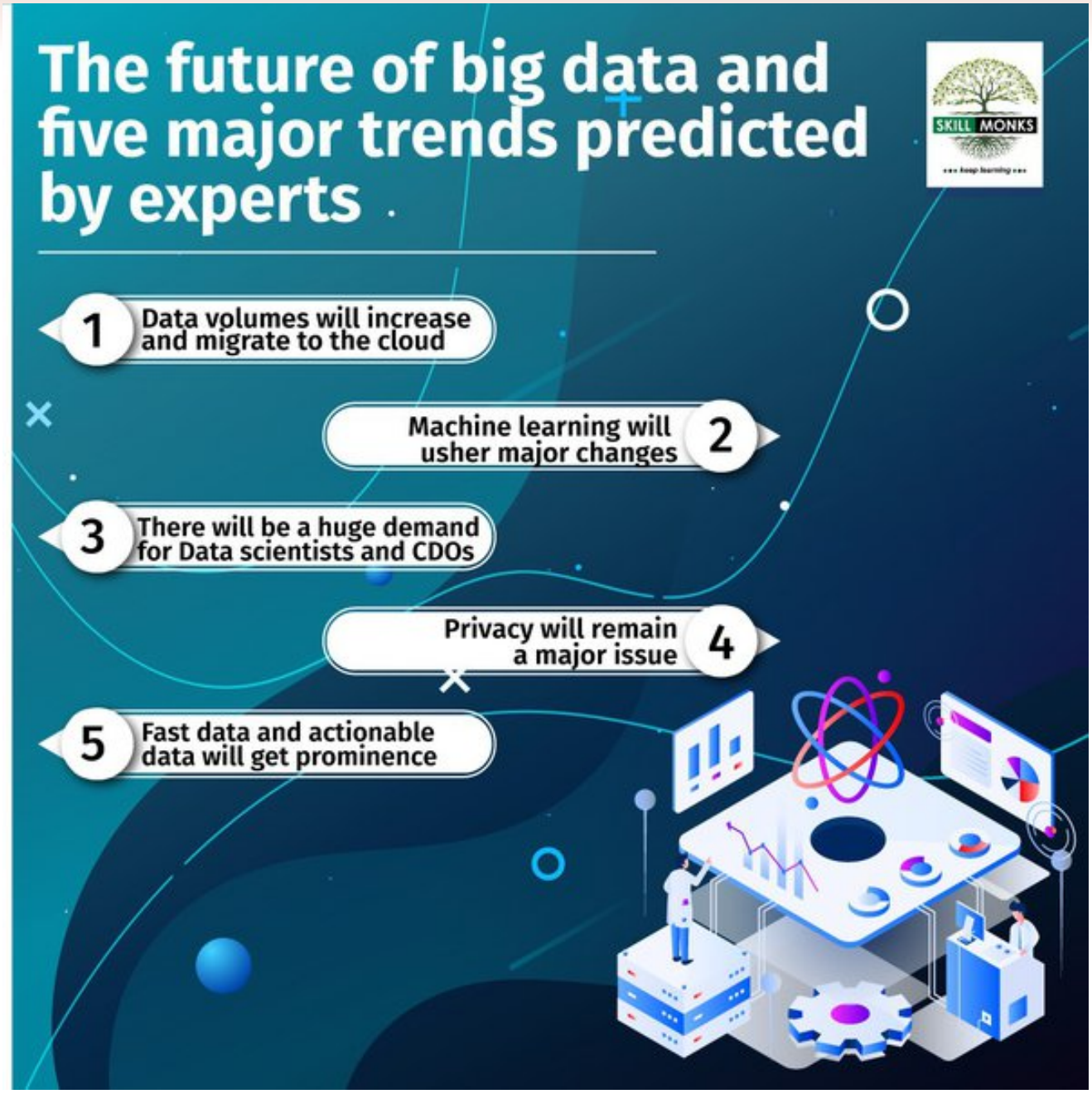
Benefits of Big Data

Increased efficiency

Better business decision making

Improved customer experience and engagement

Achieved financial savings


The future of big data and five major trends predicted by experts
SKILL MONKS
1 Data volumes will increase and migrate to the cloud
2 Machine learning will usher major changes
3 There will be a huge demand for Data scientists and CDOs
4 Privacy will remain a major issue
5 Fast data and actionable data will get prominence

BigData Opportunities and Challenges

- Lack of sufficiently skilled IT staff & Cost of Technology
- Managing Data quality
- Data Integration

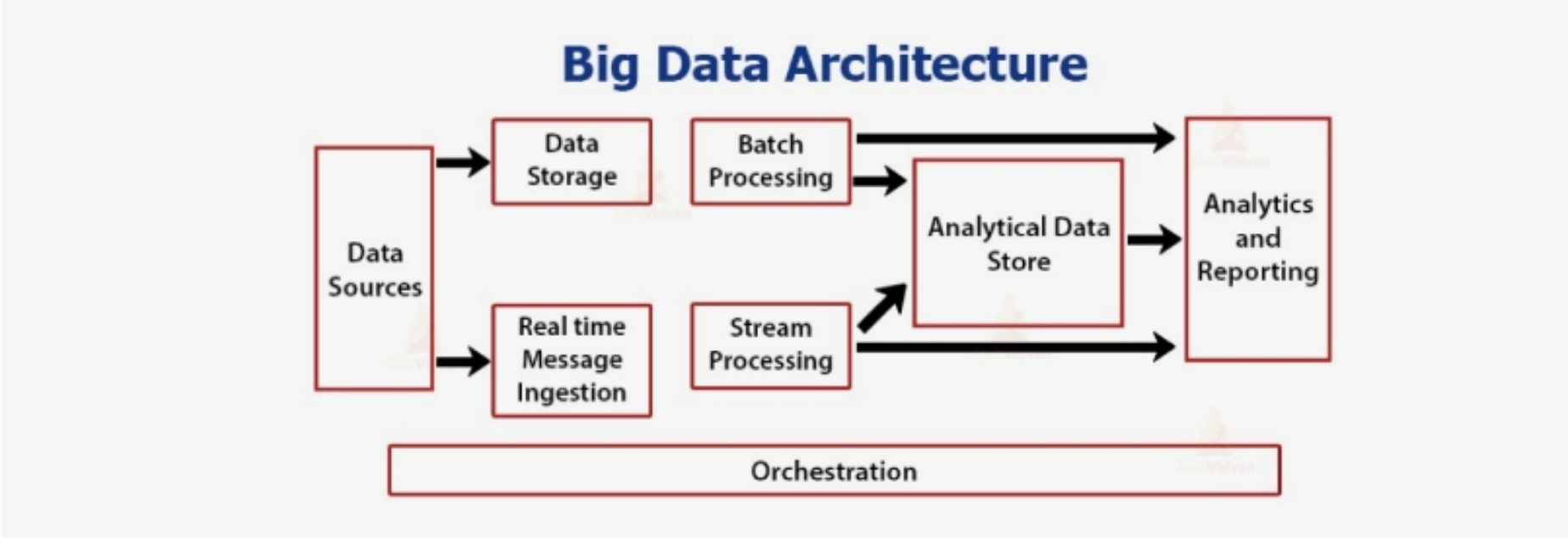

Why Big Data is right choice for your Career?
Exponential Rise of Data
Great Demand for Data Analysts
Huge Skill Gap among People
Unstructured Data Analytics
Lucrative Job Profile

Conventional Data vs Big data

- Size
- Flexibility - fixed schema V/s dynamic schema
- Architecture
- Real-time analytics
- Multitude of sources
- Enables exploratory analysis

Big Data Architecture

- Data sources
- Data storage.
- Batch processing
- Real-time message ingestion.
- Analytical data store.
- Analysis and reporting.
- Orchestration


Big Data Architecture
Data Sources
Data Storage
Real time Message Ingestion
Batch Processing
Stream Processing
Analytical Data Store
Analytics and Reporting
Orchestration

Data sources

- All big data solutions start with one or more data sources. Examples include:
- Application data stores, such as relational databases.
- Static files produced by applications, such as web server log files.
- Real-time data sources, such as IoT devices.

Data storage

- Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats.
- This kind of store is often called a data lake.
- Options for implementing this storage include Azure Data Lake Store or blob containers in Azure Storage.

Batch processing

- Because the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis.

Real-time message ingestion

- If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing.
- This might be a simple data store, where incoming messages are dropped into a folder for processing.
- many solutions need a message ingestion store to act as a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics.

Analytical data store

- Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools.

Orchestration

- Most big data solutions consist of repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report or dashboard.
- To automate these workflows, you can use an orchestration technology such Azure Data Factory or Apache Oozie and Sqoop.

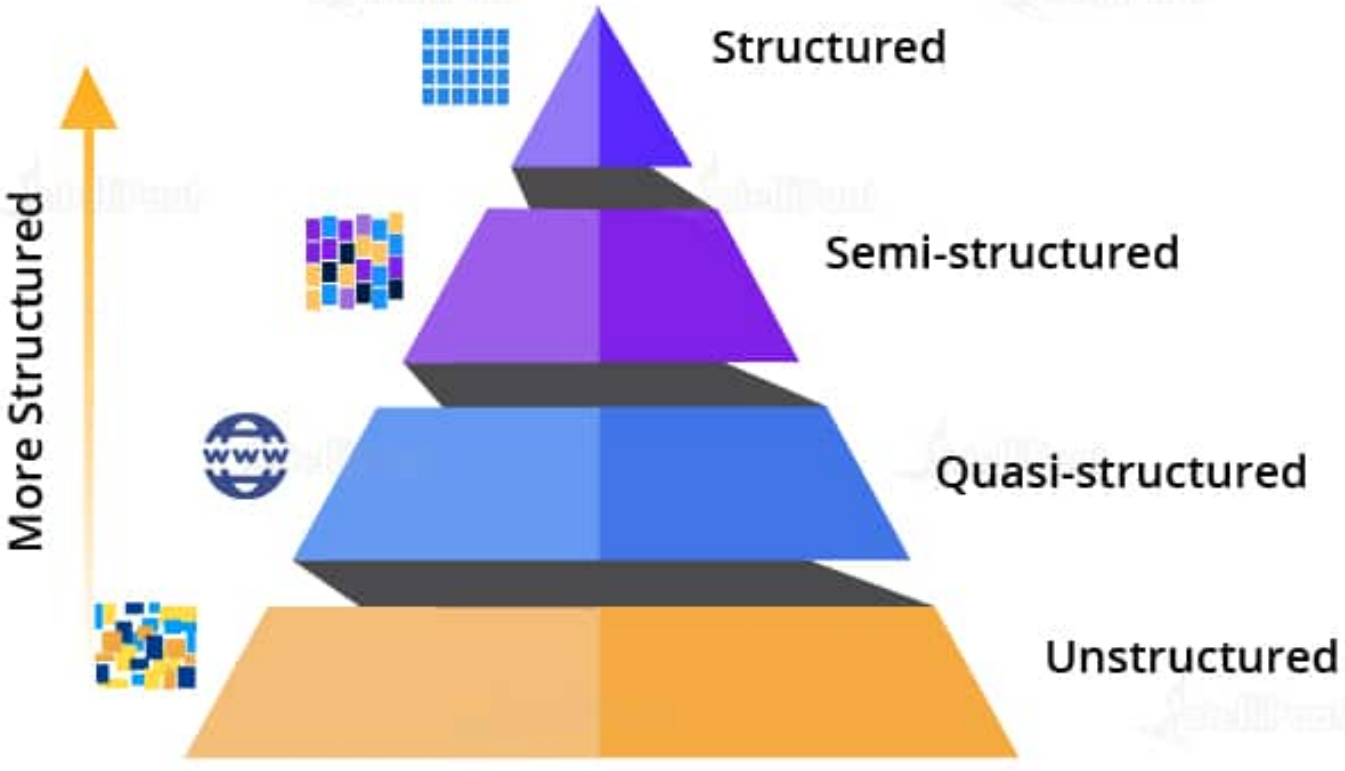
Big Data Platform

- Apache Hadoop
- Cloudera
- Databricks
- Hortonworks
- SAP HANA
- and many more

Nature of data

- Structured Data
- Semi structured Data
- Quasi-structured Data
- Unstructured Data

Nature of data



Analytic processes

- Deployment
- Business Understanding
- Data Exploration
- Data Preparation
- Data Modeling
- Data Evaluation

Nature of data

Analytic processes-Deployment

- plan the deployment and monitoring and maintenance,
- we need to produce a final report and review the project.

Analytic processes-Business Understanding

- The very first step consists of business understanding.
- Whenever any requirement occurs, firstly we need to determine the business objective,
- assess the situation,
- determine data mining goals and then produce the project plan as per the requirement.

Analytic processes- Data Exploration

- For the further process, we need to gather initial data, describe and explore the data and verify data quality to ensure it contains the data we require.
- Data collected from the various sources is described in terms of its application and the need for the project in this phase.
- This is necessary to verify the quality of data collected.

Analytic processes- Data Preparation

- we need to select data as per the need, clean it, construct it to get useful information and then integrate it all.
- Finally, we need to format the data to get the appropriate data.
- Data is selected, cleaned, and integrated into the format finalized for the analysis in this phase.

Analytic processes- Data Modeling

- select a modeling technique, generate test design, build a model and assess the model built.
- The data model is build to
  – analyze relationships between various selected objects in the data,
  – test cases are built for assessing the model and model is tested and implemented on the data in this phase.
- 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)

Analytic processes- Data Evaluation

- The process of exploring data and reports
  – in order to extract meaningful insights,
  – which can be used to better understand and improve business performance.

Modern Analytic Tools

- Batch processing tools -Apache Hadoop
- Stream Processing tools -Storm,Apache flink,Kinesis,
- Interactive Analysis tools.- Google's Dremel,Apache drill,

Categories of Modern Analytic Tools

- Big data tools for HPC and supercomputing
- Collective communication operations
- Big data tools on clouds
-
-

****

Analytic processes- TOOLS

Big Data Tools Based on Batch Processing

-
-

Analytic processes- TOOLS

Business Understanding

-
-

Analytic processes- TOOLS

Data Exploration

-
-

Analytic processes- TOOLS

Data Preparation

-
-

Analytic processes- TOOLS

Data Modeling

-
-

Analytic processes- TOOLS

Data Evaluation

-
-