Introduction to Big Data
IN411 – M1S2 – 3 Credits

Course Grading

- Partial Exam
- Final Exam
- Second Session

Course Contents

- Chapter 1: Overview of Big Data
- Chapter 2: Hadoop Ecosystem
- Chapter 3: Apache Spark Basics
- Chapter 4: Introduction to NoSQL Databases
- Chapter 5: Big Data Storage & Data Lake Concepts

Course Contents

- **Chapter 1: Overview of Big Data**
- Chapter 2: Hadoop Ecosystem
- Chapter 3: Apache Spark Basics
- Chapter 4: Introduction to NoSQL Databases
- Chapter 5: Big Data Storage & Data Lake Concepts

Chapter 1:

Overview of Big Data

Content

- Big data definitions, characteristics (5 Vs), use cases
- Difference between traditional & big data approaches
- Big data ecosystem overview (Apache tools, Hadoop, Spark)

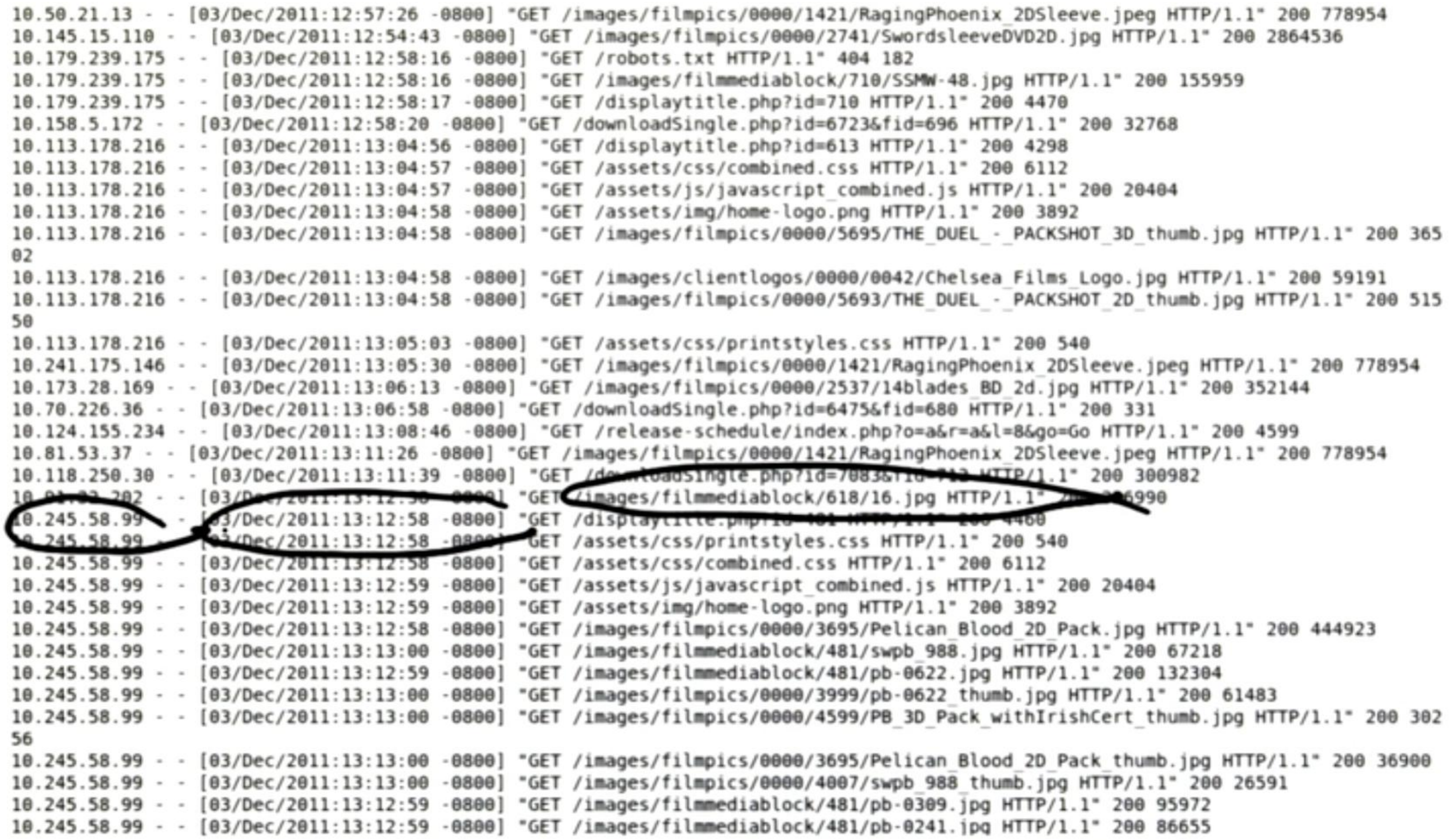
Big Data Introduction

Introduction

- Organizations have been **generating data** since way back, but as time goes on, more and more data is being generated.
- IBM estimates that, in this two years, the amount of generated data is more than 95% of word's data collected in all past years.
- For example, data about your mobile phone connected to the cell towers, are collected and logged in the phone company.
- Another example, medicine, research, search engine, AI tools, …

Introduction

- Another example, data collected during you visit a website like Amazon, Netflix, … **everything is logged**.



```
10.50.21.13 - - [03/Dec/2011:12:57:26 -0800] "GET /images/filmpics/0000/1421/RagingPhoenix_2DSleeve.jpeg HTTP/1.1" 200 778954
10.145.15.110 - - [03/Dec/2011:12:54:43 -0800] "GET /images/filmpics/0000/2741/SwordsleeveDVD2D.jpg HTTP/1.1" 200 2864536
10.179.239.175 - - [03/Dec/2011:12:58:16 -0800] "GET /robots.txt HTTP/1.1" 404 182
10.179.239.175 - - [03/Dec/2011:12:58:16 -0800] "GET /images/filmmediablock/710/SSMW-48.jpg HTTP/1.1" 200 155959
10.179.239.175 - - [03/Dec/2011:12:58:17 -0800] "GET /displaytitle.php?id=710 HTTP/1.1" 200 4470
10.158.5.172 - - [03/Dec/2011:12:58:20 -0800] "GET /downloadSingle.php?id=6723&fid=696 HTTP/1.1" 200 32768
10.113.178.216 - - [03/Dec/2011:13:04:56 -0800] "GET /displaytitle.php?id=613 HTTP/1.1" 200 4298
10.113.178.216 - - [03/Dec/2011:13:04:57 -0800] "GET /assets/css/combined.css HTTP/1.1" 200 6112
10.113.178.216 - - [03/Dec/2011:13:04:57 -0800] "GET /assets/js/javascript_combined.js HTTP/1.1" 200 20404
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /assets/img/home-logo.png HTTP/1.1" 200 3892
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /images/filmpics/0000/5695/THE_DUEL_-_PACKSHOT_3D_thumb.jpg HTTP/1.1" 200 365
02
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /images/clientlogos/0000/0042/Chelsea_Films_Logo.jpg HTTP/1.1" 200 59191
10.113.178.216 - - [03/Dec/2011:13:04:58 -0800] "GET /images/filmpics/0000/5693/THE_DUEL_-_PACKSHOT_2D_thumb.jpg HTTP/1.1" 200 515
50
10.113.178.216 - - [03/Dec/2011:13:05:03 -0800] "GET /assets/css/printstyles.css HTTP/1.1" 200 540
10.241.175.146 - - [03/Dec/2011:13:05:30 -0800] "GET /images/filmpics/0000/1421/RagingPhoenix_2DSleeve.jpeg HTTP/1.1" 200 778954
10.173.28.169 - - [03/Dec/2011:13:06:13 -0800] "GET /images/filmpics/0000/2537/14blades_BD_2d.jpg HTTP/1.1" 200 352144
10.70.226.36 - - [03/Dec/2011:13:06:58 -0800] "GET /downloadSingle.php?id=6475&fid=680 HTTP/1.1" 200 331
10.124.155.234 - - [03/Dec/2011:13:08:46 -0800] "GET /release-schedule/index.php?o=a&r=a&l=8&go=Go HTTP/1.1" 200 4599
10.81.53.37 - - [03/Dec/2011:13:11:26 -0800] "GET /images/filmpics/0000/1421/RagingPhoenix_2DSleeve.jpeg HTTP/1.1" 200 778954
10.118.250.30 - - [03/Dec/2011:13:11:39 -0800] "GET /downloadSingle.php?id=7083&fid=712 HTTP/1.1" 200 300982
10.[illegible] - - [03/Dec/2011:13:12:58 -0800] "GET /images/filmmediablock/618/16.jpg HTTP/1.1" 200 [illegible]6990
10.245.58.99 - - [03/Dec/2011:13:12:58 -0800] "GET /displaytitle.php?id=481 HTTP/1.1" 200 4460
10.245.58.99 - - [03/Dec/2011:13:12:58 -0800] "GET /assets/css/printstyles.css HTTP/1.1" 200 540
10.245.58.99 - - [03/Dec/2011:13:12:58 -0800] "GET /assets/css/combined.css HTTP/1.1" 200 6112
10.245.58.99 - - [03/Dec/2011:13:12:59 -0800] "GET /assets/js/javascript_combined.js HTTP/1.1" 200 20404
10.245.58.99 - - [03/Dec/2011:13:12:59 -0800] "GET /assets/img/home-logo.png HTTP/1.1" 200 3892
10.245.58.99 - - [03/Dec/2011:13:12:58 -0800] "GET /images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg HTTP/1.1" 200 444923
10.245.58.99 - - [03/Dec/2011:13:13:00 -0800] "GET /images/filmmediablock/481/swpb_988.jpg HTTP/1.1" 200 67218
10.245.58.99 - - [03/Dec/2011:13:12:59 -0800] "GET /images/filmmediablock/481/pb-0622.jpg HTTP/1.1" 200 132304
10.245.58.99 - - [03/Dec/2011:13:13:00 -0800] "GET /images/filmpics/0000/3999/pb-0622_thumb.jpg HTTP/1.1" 200 61483
10.245.58.99 - - [03/Dec/2011:13:13:00 -0800] "GET /images/filmpics/0000/4599/PB_3D_Pack_withIrishCert_thumb.jpg HTTP/1.1" 200 302
56
10.245.58.99 - - [03/Dec/2011:13:13:00 -0800] "GET /images/filmpics/0000/3695/Pelican_Blood_2D_Pack_thumb.jpg HTTP/1.1" 200 36900
10.245.58.99 - - [03/Dec/2011:13:13:00 -0800] "GET /images/filmpics/0000/4007/swpb_988_thumb.jpg HTTP/1.1" 200 26591
10.245.58.99 - - [03/Dec/2011:12:12:59 -0800] "GET /images/filmmediablock/481/pb-0309.jpg HTTP/1.1" 200 95972
10.245.58.99 - - [03/Dec/2011:13:12:59 -0800] "GET /images/filmmediablock/481/pb-0241.jpg HTTP/1.1" 200 86655
```

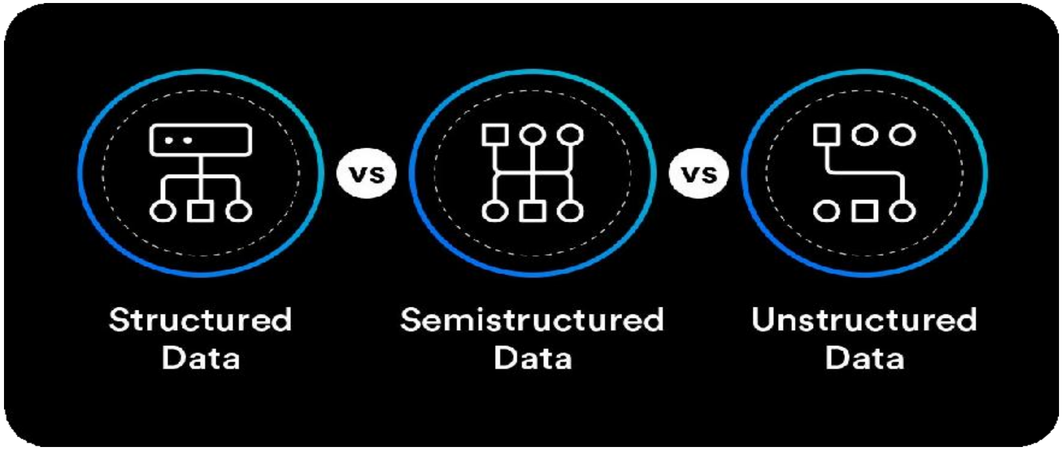
Introduction

Introduction

Definition

- Big Data refers to a collection of data sets so large, fast and complex, **impossible** to process them with the traditional data processing systems and usual databases and tools, requiring powerful technologies and advanced algorithms.

- The term refers to large datasets that include a number of diverse formats: **unstructured; semi-structured; and structured data.**



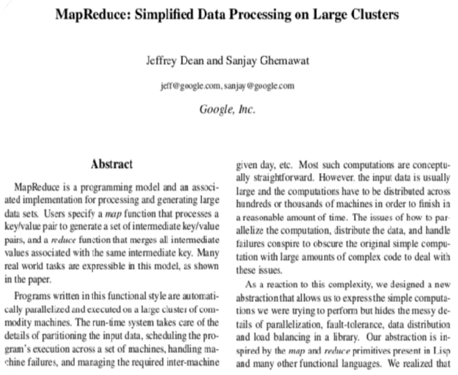
History of big data

History of big data

- The genesis of Big Data can be traced to the technological landscape of the 1960s and 1970s, a period marked by the introduction of computers for data processing.
- However, it wasn't until the 1990s that the term "Big Data" emerged.
- During the initial years of the 21st century, the advent of the internet and the widespread proliferation of digital devices triggered an unprecedented surge in the volume of data generated.

History of big data

- In the year 2004, Google does a groundbreaking technological innovation with the introduction of **MapReduce**.
- This transformative technology used in an era of large-scale data processing across distributed systems, leveraging the efficiency of commodity hardware.
- It laid the cornerstone for the development of **Hadoop**, an open-source platform that revolutionized the landscape of distributed data storage and processing.
- Hadoop, a manifestation of the principles encapsulated in MapReduce, was officially unveiled to the world in 2006.



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

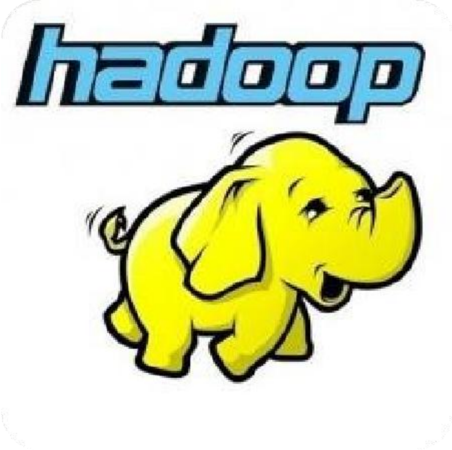
Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the *map* and *reduce* primitives present in Lisp and many other functional languages. We realized that



History of big data

- In the subsequent decade, 2010 up to now, the landscape of **Big Data technologies** underwent a profound metamorphosis, marked by the emergence and evolution of transformative components such as NoSQL databases, in-memory computing, and cloud computing, among a **spectrum** of other advancements.

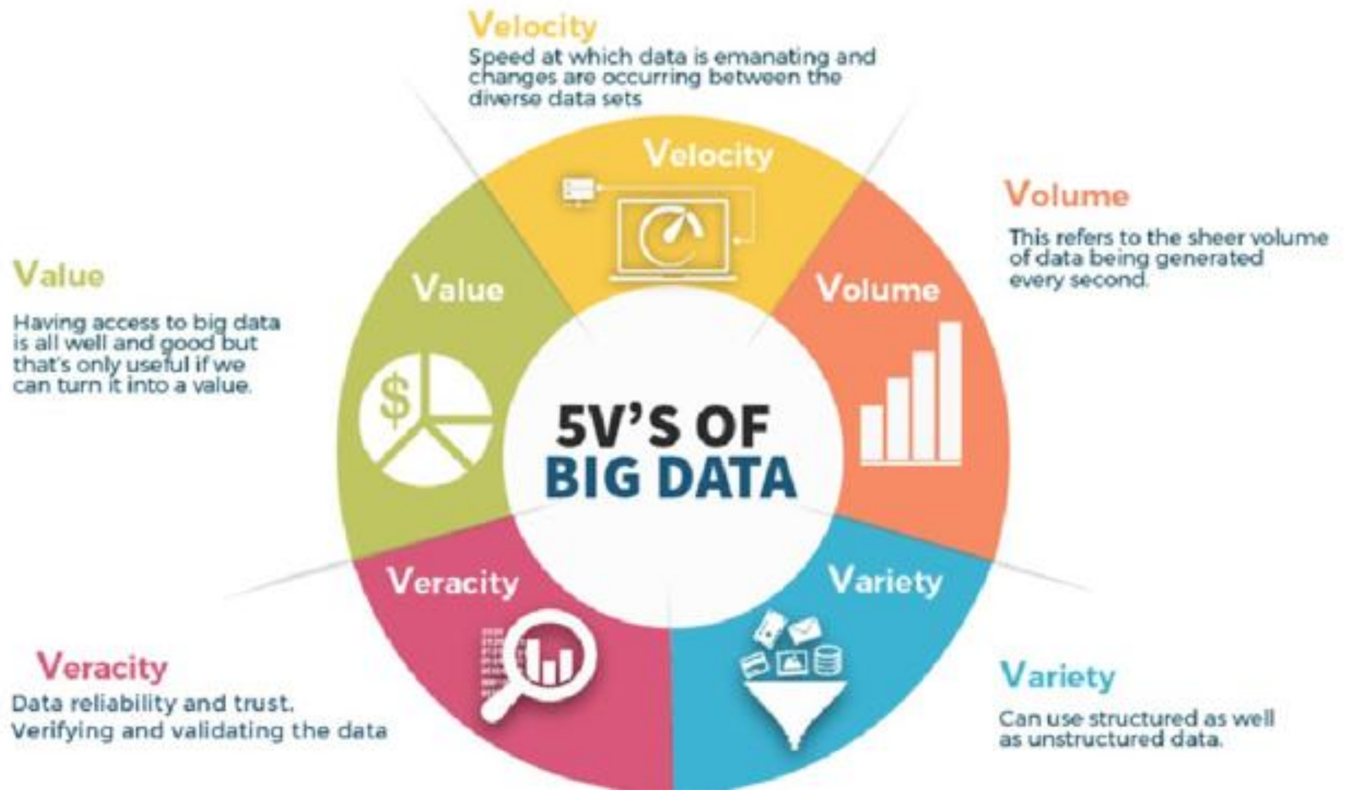
Key Drivers

- Explosion of digital content
- Growth of IoT and sensors
- Advances in distributed computing and cloud computing
- Cost reduction in storage and processing
- …

- **What should be the characteristics of big data ?**

Characteristics of Big Data (Vs)

- There are five main characteristics of Big Data, commonly known as the 5Vs of Big Data, which are:
  1. Volume
  2. Variety
  3. Velocity
  4. Veracity
  5. Value



Velocity
Speed at which data is emanating and changes are occurring between the diverse data sets

Volume
This refers to the sheer volume of data being generated every second.

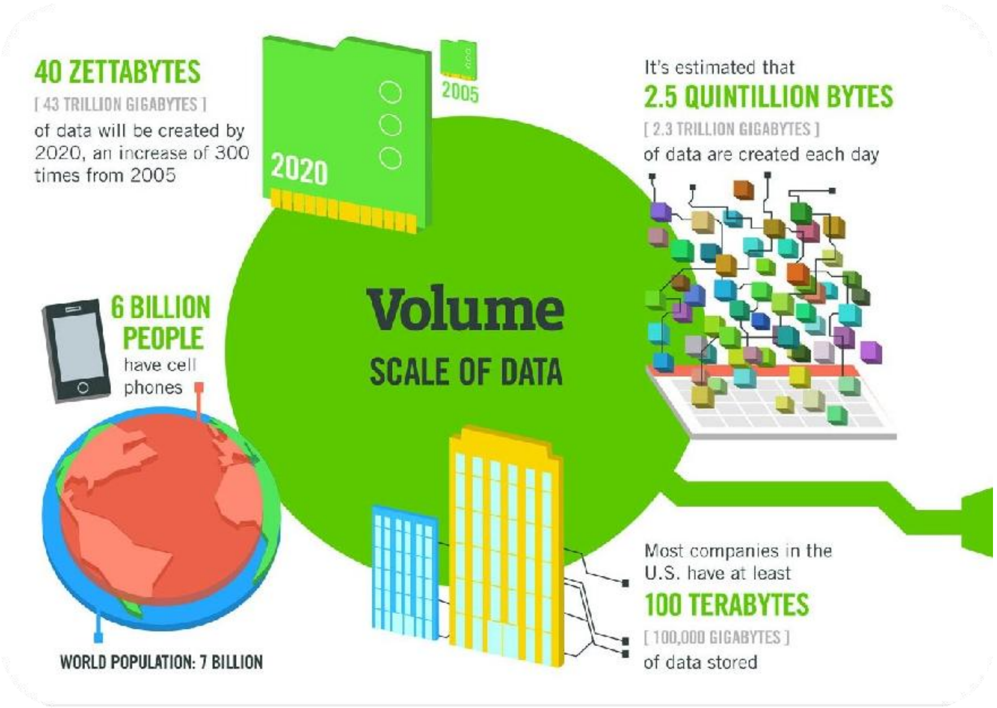
Value
Having access to big data is all well and good but that's only useful if we can turn it into a value.

Veracity
Data reliability and trust.
Verifying and validating the data

Variety
Can use structured as well as unstructured data.

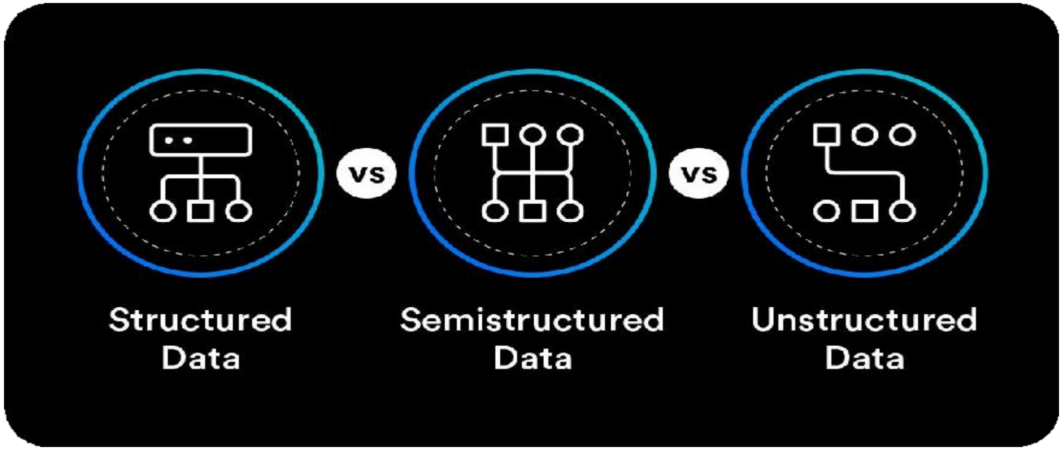
Characteristics of Big Data (Volume)

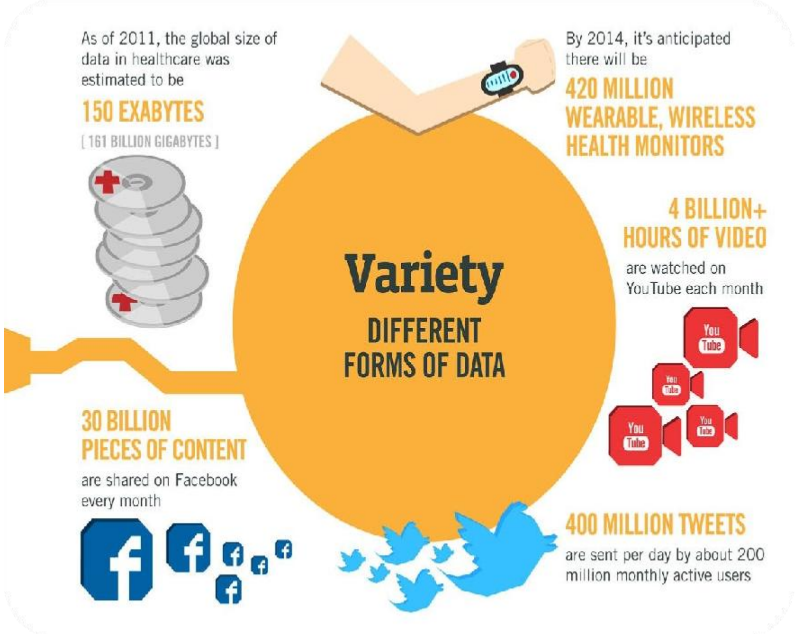
- **Massive data size**
- Many factors contribute to the increase in data volume:
- Transaction-based data stored through the years.
- Unstructured data streaming in from social media.
- Increasing amounts of sensor and machine-to-machine data being collected.


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Characteristics of Big Data (Variety)

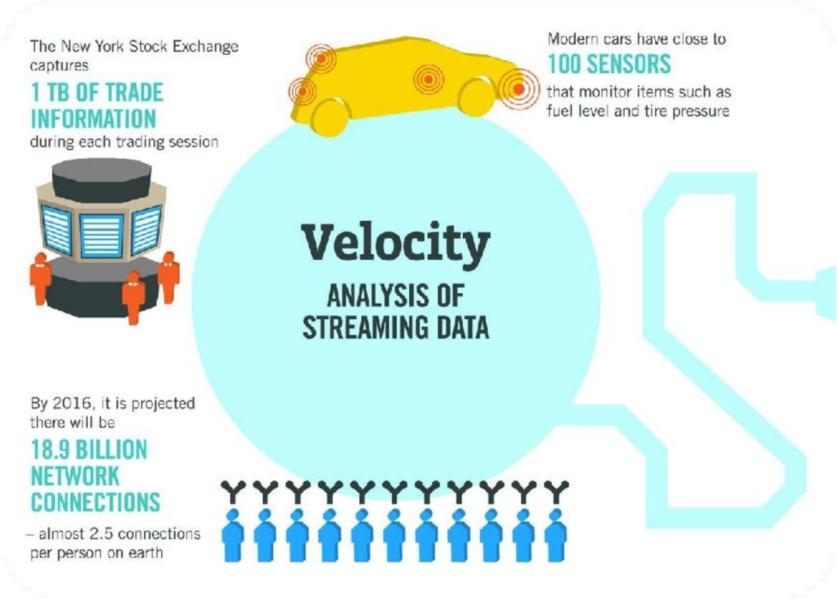
- Data today comes in all types of formats from different sources:
- **Structured**, numeric data in traditional databases.
- **Semi-structured** Information created from line-of-business applications.
- **Unstructured** text documents, email, video, audio, stock ticker data and financial transactions.




As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

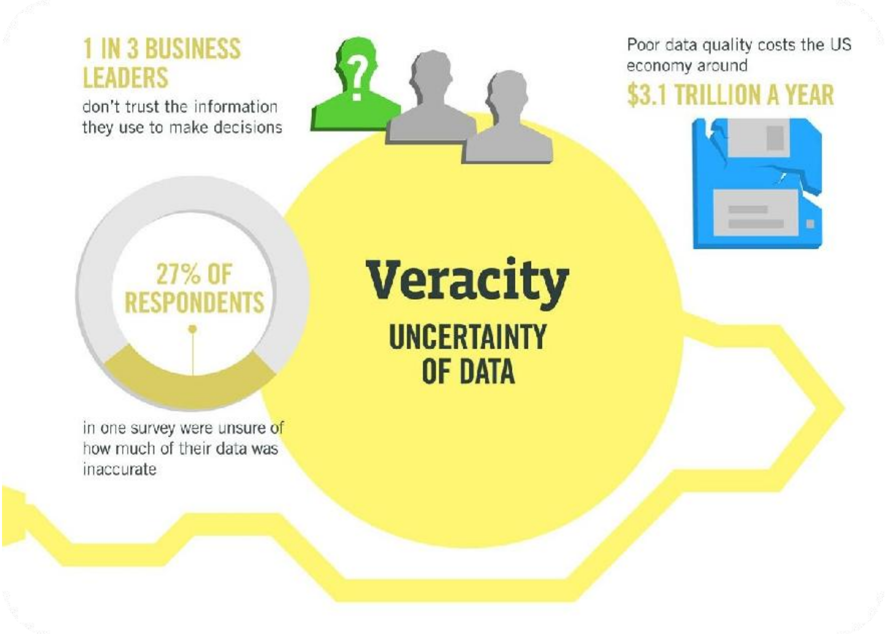
Characteristics of Big Data (Velocity)

- High-speed data generation and real-time processing.
- Data is streaming in at unprecedented speed and must be dealt with in a timely manner.
- RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.
- Reacting quickly enough to deal with data velocity is a challenge for most organizations.


The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

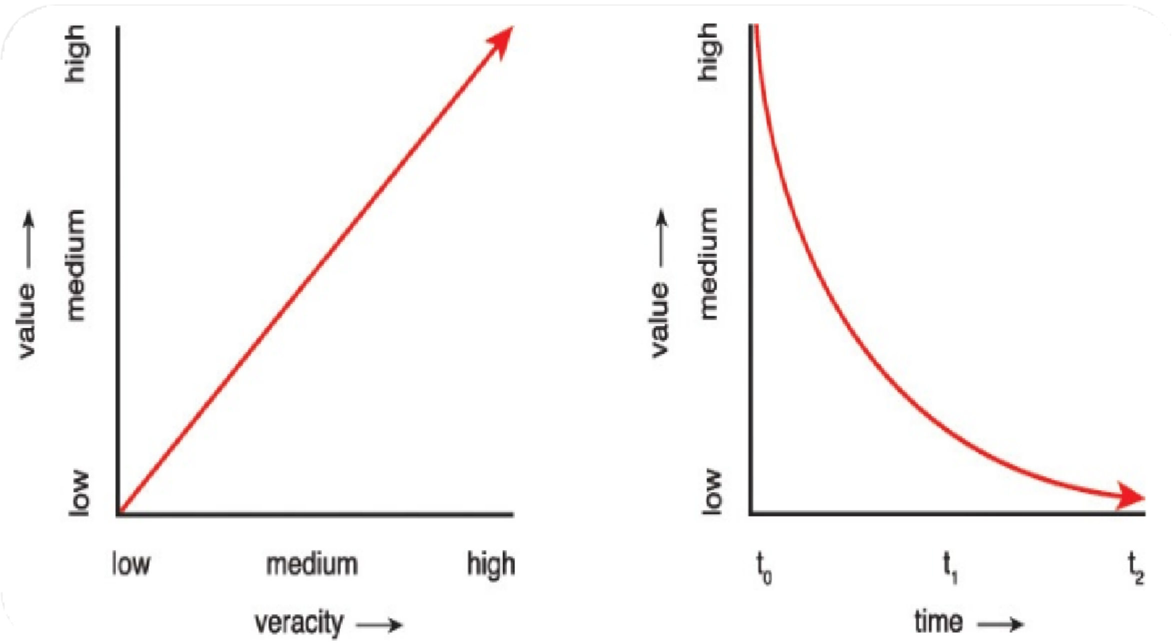
Characteristics of Big Data (Veracity)

- Big Data Veracity refers to the biases, noise and abnormality in data.
- In addition to data quality, uncertainty, and reliability challenges.
- Is the data that is being stored and mined meaningful to the problem being analyzed?.
- In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems.


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
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Characteristics of Big Data (Value)

- Value is defined as the usefulness of data for an enterprise.
- In addition to the ability to extract meaningful insights and business value.
- The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business;



Advantages

- **Enhanced decision-making capabilities**:
- Big Data grants organizations access to extensive datasets,
- enabling them to make decisions that are more informed and driven by data.

- **Increased efficiency and productivity**:
- Leveraging Big Data technologies empowers organizations to swiftly and precisely process and analyze data.
- This capability aids organizations in optimizing operations,
- minimizing waste and inefficiencies,
- and ultimately boosting productivity.

Advantages

- **Enhanced customer insights**:
- big data equips organizations with a detailed comprehension of their customers' behaviors, preferences, and requirements.
- This enables organizations to enhance their marketing and customer engagement strategies,
- ultimately resulting in increased customer satisfaction and loyalty.

- **Cost-effectiveness**:
- through enhanced efficiency and productivity, Big Data contributes to cost savings, thereby increasing organizational profitability.
- For instance, optimizing supply chain operations, lowering inventory costs, and enhancing resource allocation are areas where Big Data can be instrumental."

Challenges in Big Data

- **Technical Challenges**
- Scalability
- Storage management
- Real-time processing
- System reliability

- **Data Challenges**
- Quality and cleaning
- Integration from multiple sources
- Governance and lifecycle management

Challenges in Big Data

- **Security & Privacy**
- Data breaches
- Encryption
- Compliance (GDPR, data protection laws)

- **Ethical Challenges**
- Algorithmic bias
- Data ownership
- Responsible AI usage

Use cases and Emerging Technologies

Use cases – Streaming Apps

- This is probably the easiest to explain example of how Big Data and Data Science enhance a customer-focused and data-driven business.
- Data Science and Engineering people at Netflix are members of different business units, like content or product development, and they are responsible for implementing analytics at scale.
- They provide personalized movie and TV show recommendations, thumbnails, and trailers.
- Also, content popularity prediction before it goes live.

Use cases – Social Networks

- Have you ever seen one of the videos on Facebook that shows a “flashback” of posts, likes, or images - like the ones you might see on your birthday or on the anniversary of becoming friends with someone? If so, you have seen examples of how Facebook uses Big Data.

Use cases – Flight Companies

- A commercial flight can generate about 10 Terabytes of operational information every 30-minute interval of work.
- About 22,000 daily flights are operated in a given day, worldwide.
- This might give us a glimpse of the data deluge generated by machines and sensor networks regularly (Internet of Things).
- So, Smarter maintenance and Safer flights.

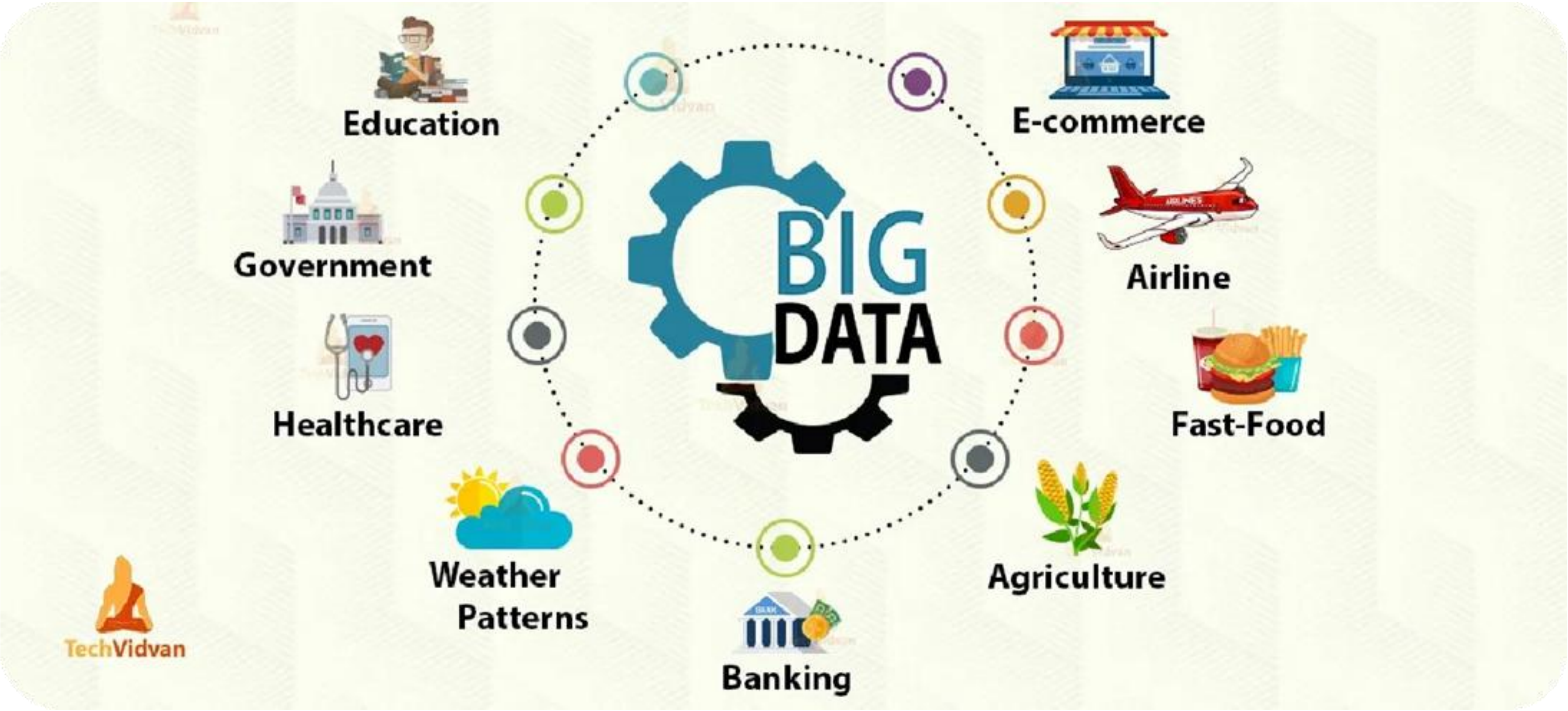
Use cases – Smart Cities

- Smart-city projects integrate real-time data from many different data sources into a single data hub.
- Some smart-city projects involve building brand-new cities that are smart from the ground up. However, most smart-city projects involve the retrofitting of existing cities with new sensor networks and data- processing centers.
- For example, in the Smart Santander project in Spain, more than 12,000 networked sensors have been installed across the city to measure temperature, noise, ambient lighting and parking.

Use cases

- **Business & Finance:** Fraud detection, customer analytics, algorithmic trading.
- **Healthcare:** Genomics, disease prediction, personalized medicine.
- **Smart Cities:** Traffic optimization, public safety, energy efficiency.
- **Cybersecurity:** Threat detection, anomaly detection, log analytics.
- **Scientific Research:** Climate modeling, astronomy, particle physics.
- **Marketing:** Behavior analytics, recommendation engines.

Use cases


Education
E-commerce
Government
BIG DATA
Airline
Healthcare
Fast-Food
Weather Patterns
Agriculture
Banking
TechVidvan

Emerging Technologies

- **Artificial Intelligence (AI) & Machine Learning (ML)**: Extracts insights, builds predictive models, automates decisions.
- **Cloud Computing & Cloud-Native Technologies**: Provides scalable storage, compute elasticity, and distributed processing.
- **Internet of Things (IoT) & Edge Computing**: Generates high-velocity real-time data streams.
- **Blockchain & Distributed Ledger Technologies**: Ensures data integrity, immutability, and secure sharing.
- **Cybersecurity & Privacy-Enhancing Technologies**: Protects large-scale data assets and ensures regulatory compliance.

Emerging Technologies

- **Data Engineering & Analytics Platforms**: Manages data pipelines, lifecycle, and quality.
- **Digital Twin & Simulation Technologies**: Models real-world systems using live data.
- **Extended Reality (XR**): AR, VR, MR. Uses analytics for immersive data visualization and user behavior modeling.
- **Quantum Computing**: Potential acceleration of complex data processing problems.
- **5G / 6G & High-Speed Connectivity**: Enables real-time, large-scale data transfer.
- **Robotic Process Automation (RPA):** Automates data ingestion, processing, and workflows.

Emerging Technologies

- **Industry-Specific Emerging Big Data Applications:**
- **Healthcare:** Precision Medicine, Medical Imaging AI, Genomics Analytics.
- **Finance:** Algorithmic Trading, Fraud Detection AI.
- **Smart Cities:** Traffic Analytics, Energy Optimization.
- **Manufacturing:** Industry 4.0 Analytics, Predictive Maintenance.

Emerging Technologies

- **Big Data Storage & Processing Frameworks:** Distributed storage, batch processing, real-time analytics.
- **Example:**
- Apache Hadoop
- Apache Spark
- Apache Flink
- Apache Kafka (Streaming & Event Platforms)
- Apache Storm
- Apache HBase
- NoSQL Databases (MongoDB, Cassandra)
- Object Storage (Amazon S3, Azure Blob)

Difference Between Traditional Data and Big Data

- **Traditional Data Approach** focuses on managing structured, moderate-size datasets using centralized relational databases and predefined schemas.
- It is optimized for business reporting, transaction processing, and structured analytics.

- **Big Data Approach** is designed to handle massive, fast-moving, and diverse datasets using distributed storage, parallel processing, and flexible data models.
- It supports advanced analytics, machine learning, and real-time decision-making.

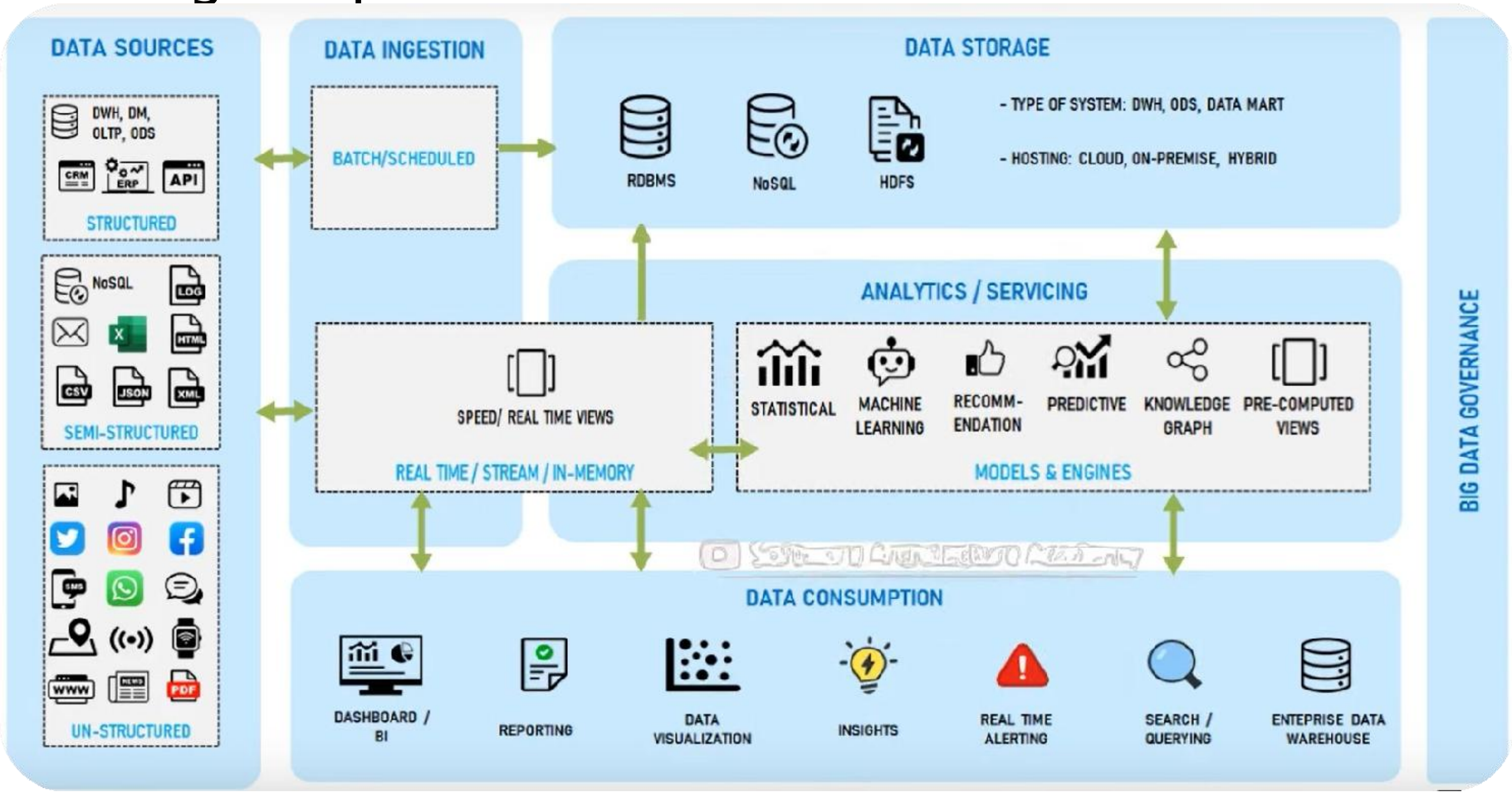
Difference Between Traditional Data and Big Data

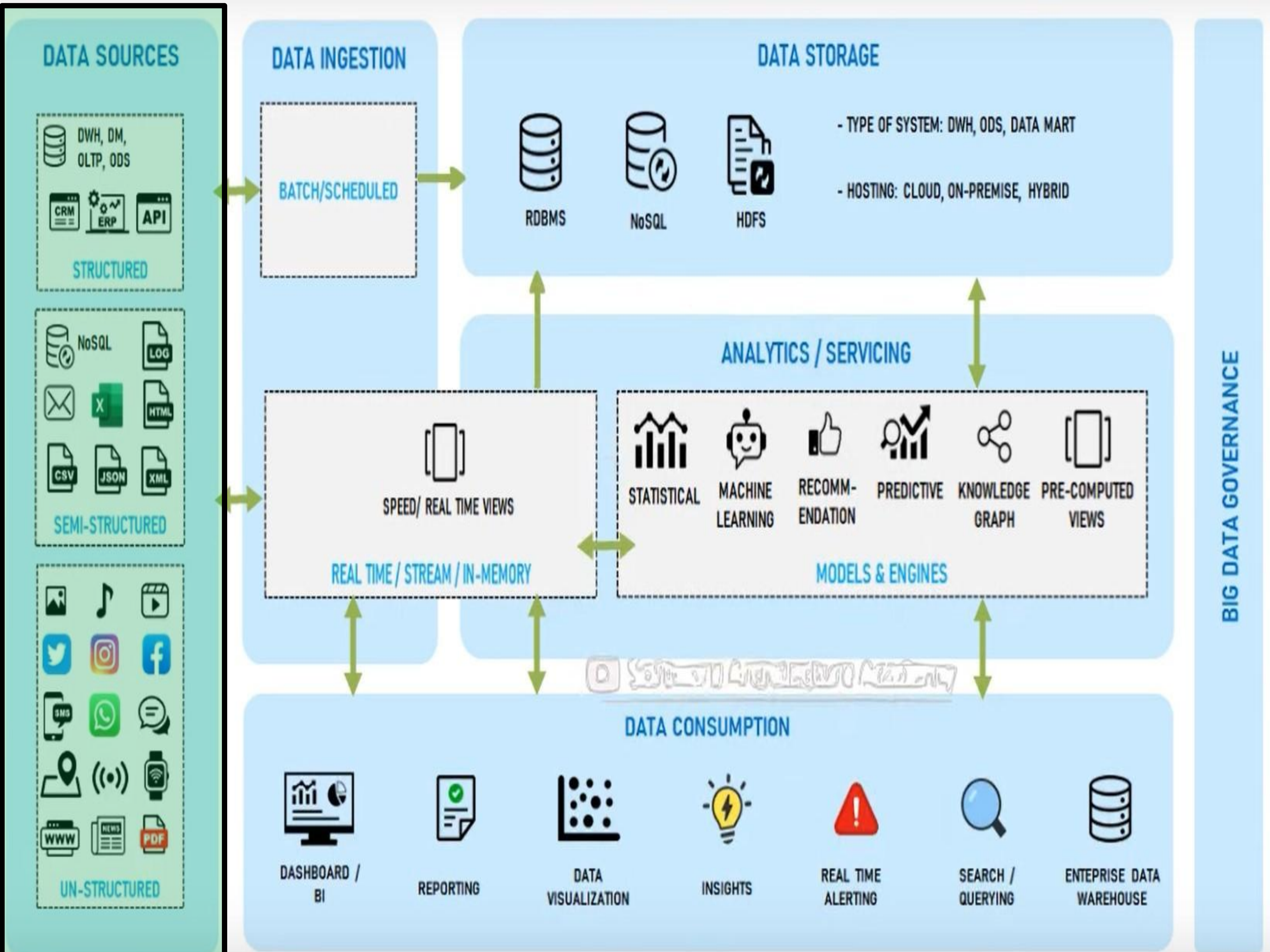
Dimension	Traditional Data	Big Data
Data Size	Small to medium (MB–TB)	Massive (TB–PB–EB+)
Data Types	Structured only	Structured, semi-structured, unstructured
Schema	Schema-on-write (fixed)	Schema-on-read (flexible)
Storage	Centralized databases	Distributed file systems & object stores
Processing Model	Batch processing	Batch + real-time + streaming
Scalability	Vertical (scale up)	Horizontal (scale out)
Architecture	Monolithic	Distributed & cluster-based
Fault Tolerance	Limited	Built-in redundancy & replication
Query Model	SQL	SQL + NoSQL + MapReduce + Spark
Analytics	BI, reporting	AI, ML, predictive analytics
Cost Model	High-cost enterprise hardware	Commodity hardware / cloud-based
Performance Goal	Transaction efficiency	Large-scale throughput & speed

Big Data Architecture and Ecosystem Overview

Big Data Architecture

- A big data architecture is designed to handle the **ingestion**, **processing**, and **analysis** of data that is too large or complex for traditional database systems.
- Most big data architectures include some or all of the following components:





DATA SOURCES

DWH, DM,
OLTP, ODS

STRUCTURED

NoSQL

SEMI-STRUCTURED

UN-STRUCTURED

DATA INGESTION

BATCH/SCHEDULED

SPEED/ REAL TIME VIEWS

REAL TIME / STREAM / IN-MEMORY

DATA STORAGE

RDBMS

NoSQL

HDFS

- TYPE OF SYSTEM: DWH, ODS, DATA MART
- HOSTING: CLOUD, ON-PREMISE, HYBRID

ANALYTICS / SERVICING

STATISTICAL

MACHINE LEARNING

RECOMM-ENDATION

PREDICTIVE

KNOWLEDGE GRAPH

PRE-COMPUTED VIEWS

MODELS & ENGINES

DATA CONSUMPTION

DASHBOARD / BI

REPORTING

DATA VISUALIZATION

INSIGHTS

REAL TIME ALERTING

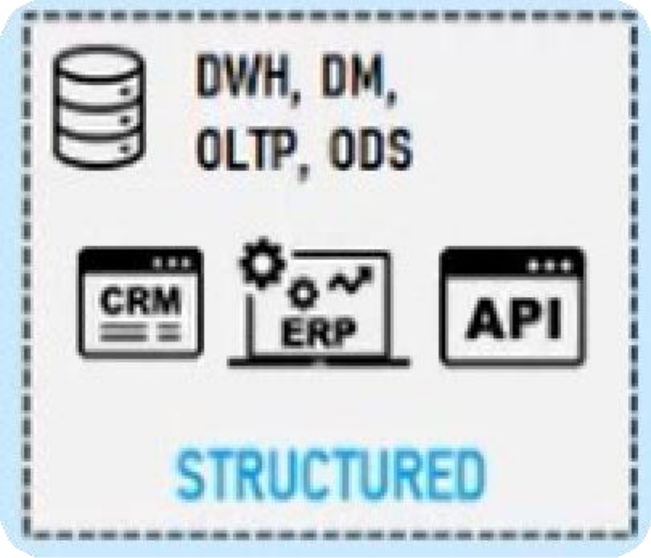
SEARCH / QUERYING

ENTEPRISE DATA WAREHOUSE

BIG DATA GOVERNANCE

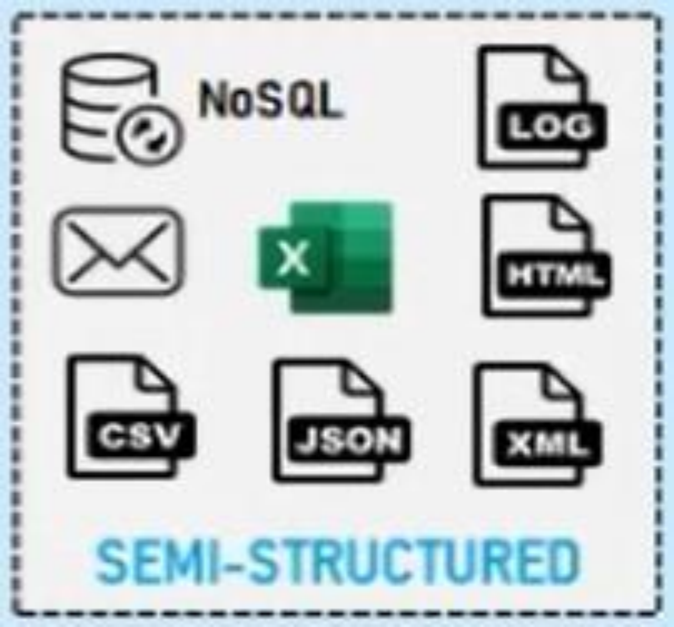
Data Sources

- All big data solutions start with one or more data sources. They can be broadly classified into **three categories**.
- **Structured data** sources are the most organized forms of data, frequently originating from <u>relational databases and tables</u> where the structure is clearly defined.
- Common examples of structured data sources include SQL databases like MySQL, Oracle, and Microsoft SQL Server.



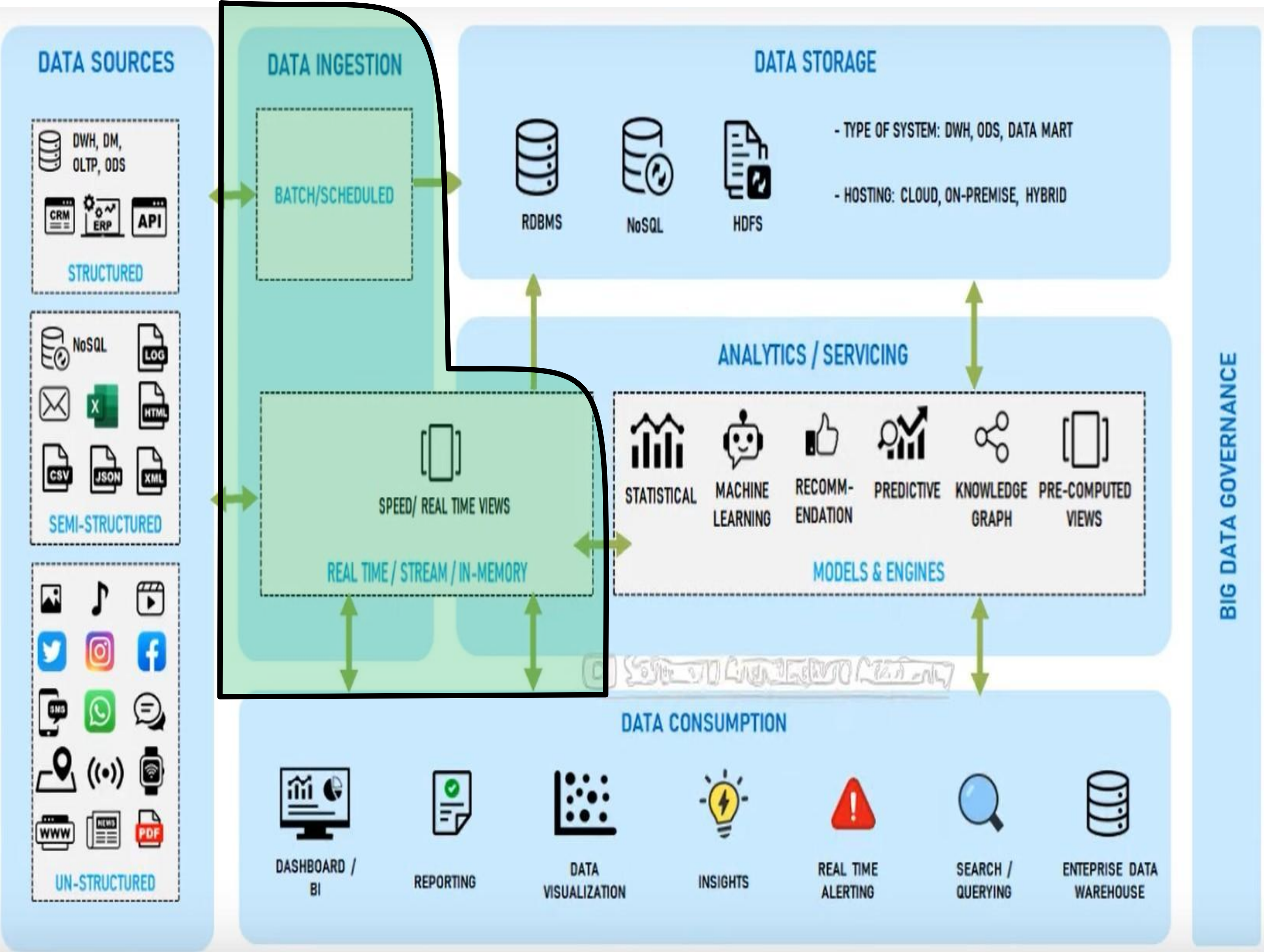
Data Sources

- **Semi-structured data** sources exhibit a certain level of organization but do not neatly fit into tabular structures.
- Examples encompass data formats such as HTML, XML, and JSON files.
- Although these formats may possess hierarchical or tagged structures, additional processing is required to render them fully structured.



Data Sources

- **Unstructured data** sources, on the other hand, encompass a diverse array of data types lacking a predefined structure.
- Examples of unstructured data include sensor data in industrial Internet of Things (IoT) applications, videos and audio streams, images, and content from social media platforms such as tweets or Facebook posts.



DATA SOURCES

DWH, DM, OLTP, ODS

STRUCTURED

NoSQL

SEMI-STRUCTURED

UN-STRUCTURED

DATA INGESTION

BATCH/SCHEDULED

SPEED/ REAL TIME VIEWS

REAL TIME / STREAM / IN-MEMORY

DATA STORAGE

RDBMS

NoSQL

HDFS

- TYPE OF SYSTEM: DWH, ODS, DATA MART
- HOSTING: CLOUD, ON-PREMISE, HYBRID

ANALYTICS / SERVICING

STATISTICAL

MACHINE LEARNING

RECOMM-ENDATION

PREDICTIVE

KNOWLEDGE GRAPH

PRE-COMPUTED VIEWS

MODELS & ENGINES

DATA CONSUMPTION

DASHBOARD / BI

REPORTING

DATA VISUALIZATION

INSIGHTS

REAL TIME ALERTING

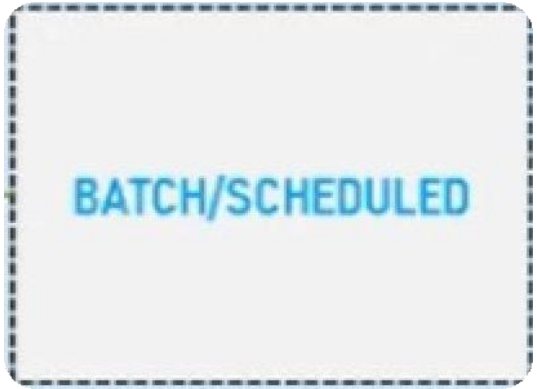
SEARCH / QUERYING

ENTEPRISE DATA WAREHOUSE

BIG DATA GOVERNANCE

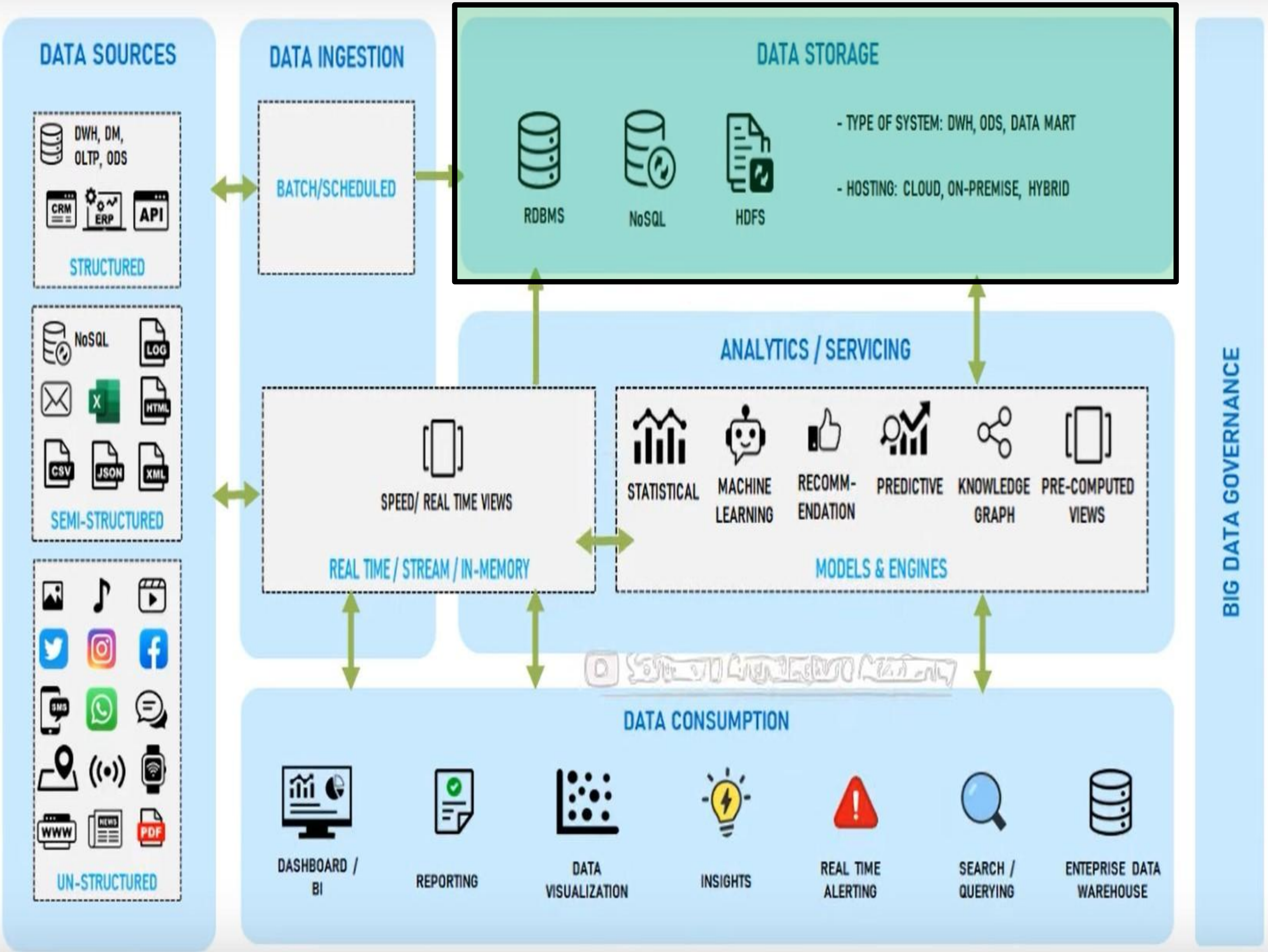
Data ingestion

- **Data ingestion** is the process of importing (load) data.
- It serves as the gateway through which data enters the big data architecture, either in **batch** or **real-time modes**, before undergoing further processing.
- **Batch** ingestion is a **scheduled, interval-based** approach to data importation, commonly set to run on a regular basis, such as nightly or weekly, to transfer large chunks of data at once.
- Apache NiFi is often used as a tool for batch ingestion.



Data ingestion

- **Real-time** ingestion involves the **immediate entry** of data into the big data architecture as it is generated.
- This is particularly crucial for **time-sensitive applications** such as fraud detection or real- time analytics.
- Popular tools for handling real-time data ingestion include Apache Kafka.



DATA SOURCES

DWH, DM,
OLTP, ODS

CRM ERP API

STRUCTURED

NoSQL LOG

X HTML

CSV JSON XML

SEMI-STRUCTURED

SMS

WWW NEWS PDF

UN-STRUCTURED

DATA INGESTION

BATCH/SCHEDULED

SPEED/ REAL TIME VIEWS

REAL TIME / STREAM / IN-MEMORY

DATA STORAGE

RDBMS

NoSQL

HDFS

- TYPE OF SYSTEM: DWH, ODS, DATA MART

- HOSTING: CLOUD, ON-PREMISE, HYBRID

ANALYTICS / SERVICING

STATISTICAL

MACHINE LEARNING

RECOMM-ENDATION

PREDICTIVE

KNOWLEDGE GRAPH

PRE-COMPUTED VIEWS

MODELS & ENGINES

DATA CONSUMPTION

DASHBOARD / BI

REPORTING

DATA VISUALIZATION

INSIGHTS

REAL TIME ALERTING

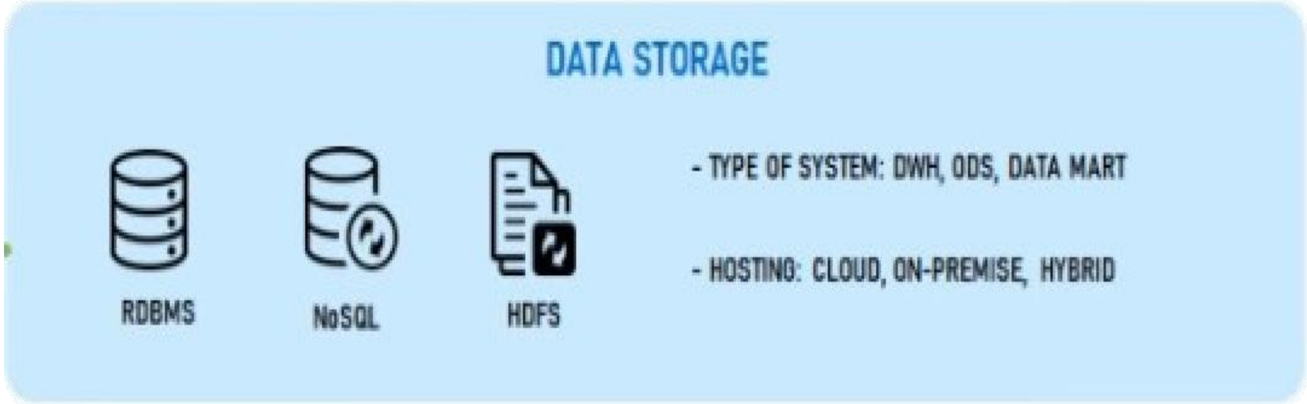
SEARCH / QUERYING

ENTEPRISE DATA WAREHOUSE

BIG DATA GOVERNANCE

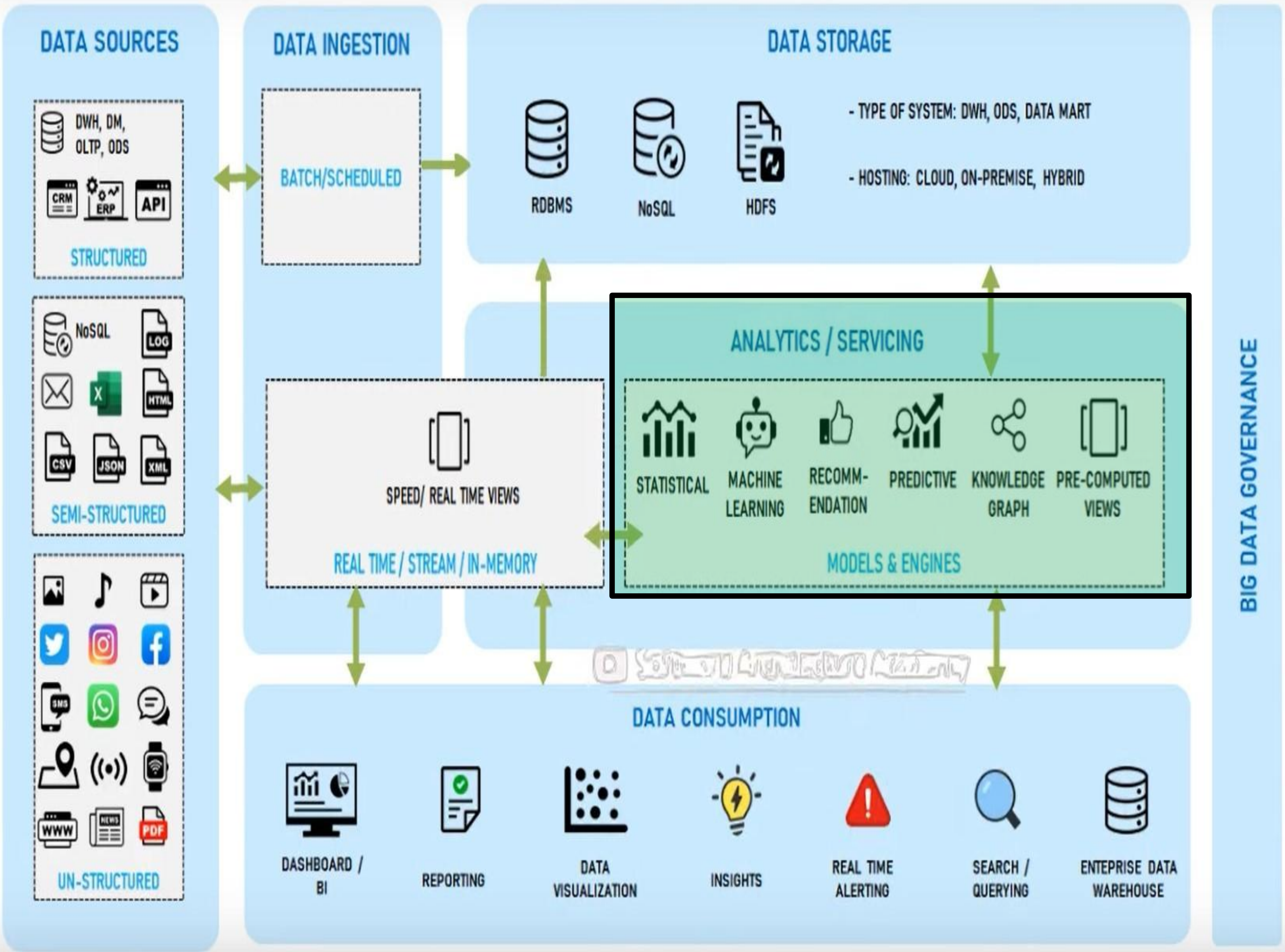
Data Storage and Processing

- The data storage and processing layer is where the ingested data **resides** and undergoes transformations to make it more **accessible and valuable for analysis**.
- Distributed file systems like Hadoop Distributed File System (**HDFS**) or Amazon S3 enable storing and processing large volumes of data across multiple nodes.
- Distributed file systems can significantly improve performance and fault tolerance by spreading data across multiple machines.


DATA STORAGE
RDBMS
NoSQL
HDFS
- TYPE OF SYSTEM: DWH, ODS, DATA MART
- HOSTING: CLOUD, ON-PREMISE, HYBRID

Data Storage and Processing

- Here are a few transformation processes that happen at this layer:
- **The data cleaning process** revolves around the removal or correction of inaccurate records, discrepancies, or inconsistencies present in the data.
- **Data enrichment** enhances the original data set by introducing additional information or context, thereby adding value.
- **Normalization** is a process that transforms the data into a standardized format, ensuring uniformity and consistency.
- **Structuring** frequently entails breaking down unstructured or semi-structured data into a structured form suitable for analysis.



DATA SOURCES

DWH, DM, OLTP, ODS

CRM ERP API

STRUCTURED

NoSQL LOG HTML CSV JSON XML

SEMI-STRUCTURED

WWW PDF

UN-STRUCTURED

DATA INGESTION

BATCH/SCHEDULED

SPEED/ REAL TIME VIEWS

REAL TIME / STREAM / IN-MEMORY

DATA STORAGE

RDBMS

NoSQL

HDFS

- TYPE OF SYSTEM: DWH, ODS, DATA MART
- HOSTING: CLOUD, ON-PREMISE, HYBRID

ANALYTICS / SERVICING

STATISTICAL

MACHINE LEARNING

RECOMM-ENDATION

PREDICTIVE

KNOWLEDGE GRAPH

PRE-COMPUTED VIEWS

MODELS & ENGINES

DATA CONSUMPTION

DASHBOARD / BI

REPORTING

DATA VISUALIZATION

INSIGHTS

REAL TIME ALERTING

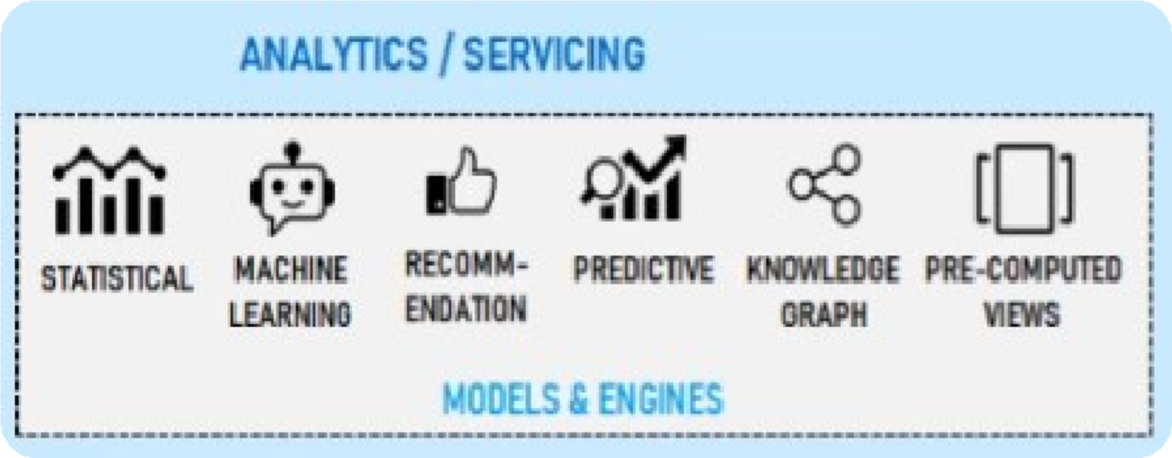
SEARCH / QUERYING

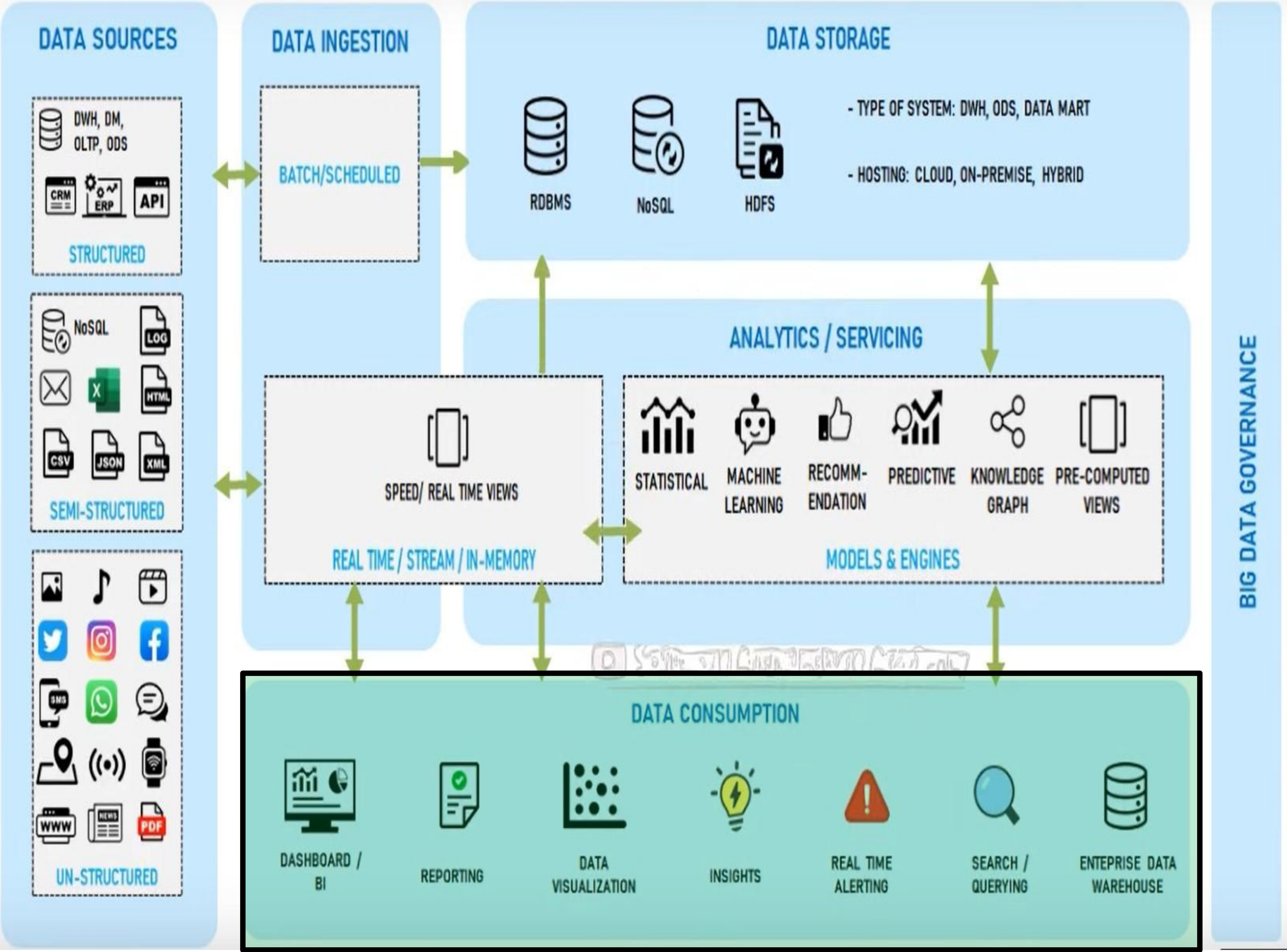
ENTEPRISE DATA WAREHOUSE

BIG DATA GOVERNANCE

Analytics and Servicing

- Analytical sandboxes function as isolated environments dedicated to **data exploration**, enabling various activities such as discovery, machine learning, predictive modeling, and exploratory data analysis.
- These environments provide a **secure space** for users to experiment with and analyze data without impacting the integrity of the overall data infrastructure.





DATA SOURCES

DWH, DM, OLTP, ODS

STRUCTURED

NoSQL

SEMI-STRUCTURED

UN-STRUCTURED

DATA INGESTION

BATCH/SCHEDULED

SPEED/ REAL TIME VIEWS

REAL TIME / STREAM / IN-MEMORY

DATA STORAGE

RDBMS

NoSQL

HDFS

- TYPE OF SYSTEM: DWH, ODS, DATA MART
- HOSTING: CLOUD, ON-PREMISE, HYBRID

ANALYTICS / SERVICING

STATISTICAL

MACHINE LEARNING

RECOMM-ENDATION

PREDICTIVE

KNOWLEDGE GRAPH

PRE-COMPUTED VIEWS

MODELS & ENGINES

DATA CONSUMPTION

DASHBOARD / BI

REPORTING

DATA VISUALIZATION

INSIGHTS

REAL TIME ALERTING

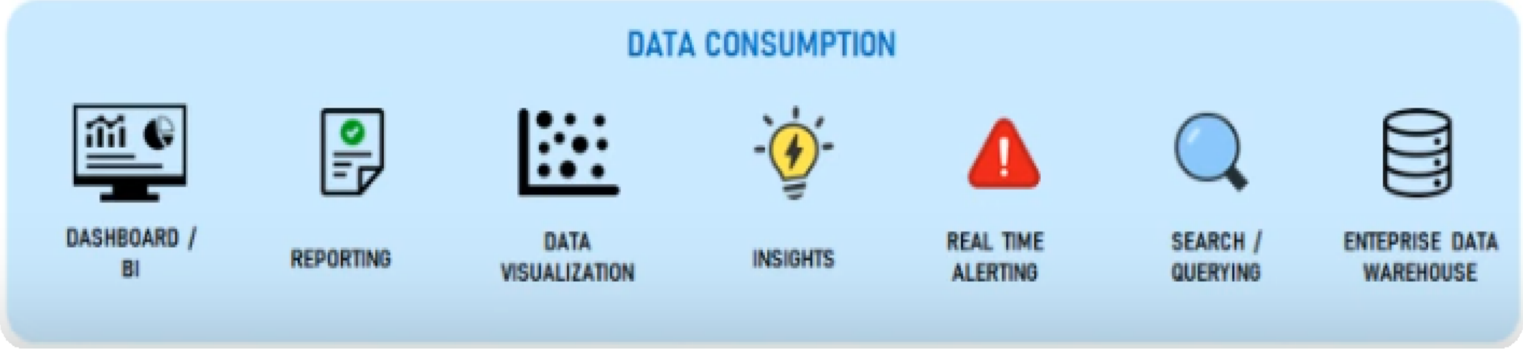
SEARCH / QUERYING

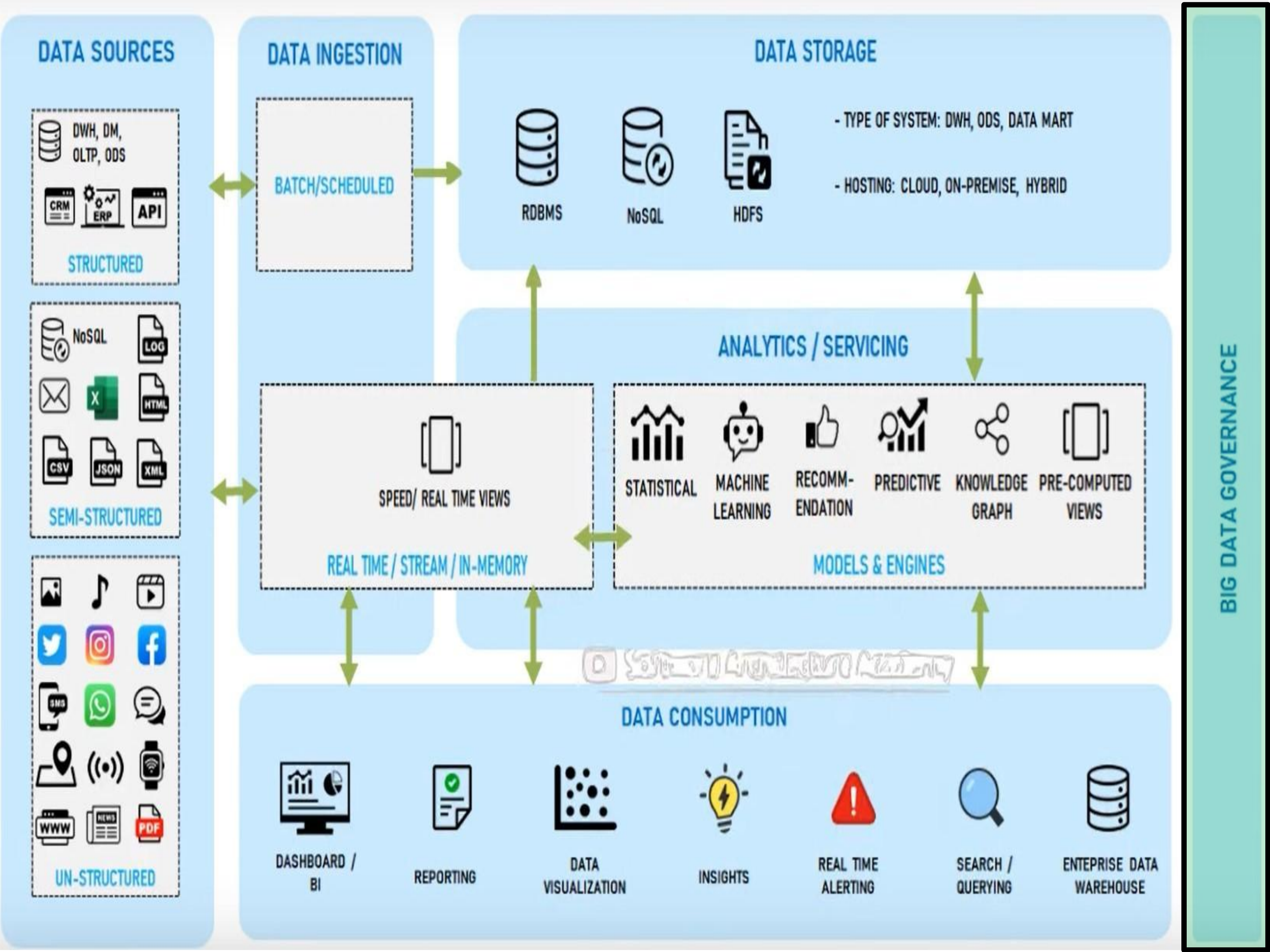
ENTEPRISE DATA WAREHOUSE

BIG DATA GOVERNANCE

Big Data Architecture

- The goal of most big data solutions is to provide insights into the data through analysis and reporting.
- It might support self-service BI, using the modeling and visualization technologies in Microsoft Power BI or Microsoft Excel.
- Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts.
- This layer is instrumental for specialized roles like **data analysts, business analysts, and decision-makers**, who utilize the processed data to inform and drive business decisions.





DATA SOURCES

DWH, DM, OLTP, ODS

CRM ERP API

STRUCTURED

NoSQL LOG HTML CSV JSON XML

SEMI-STRUCTURED

SMS WWW NEWS PDF

UN-STRUCTURED

DATA INGESTION

BATCH/SCHEDULED

SPEED/ REAL TIME VIEWS

REAL TIME / STREAM / IN-MEMORY

DATA STORAGE

RDBMS

NoSQL

HDFS

- TYPE OF SYSTEM: DWH, ODS, DATA MART
- HOSTING: CLOUD, ON-PREMISE, HYBRID

ANALYTICS / SERVICING

STATISTICAL

MACHINE LEARNING

RECOMM-ENDATION

PREDICTIVE

KNOWLEDGE GRAPH

PRE-COMPUTED VIEWS

MODELS & ENGINES

DATA CONSUMPTION

DASHBOARD / BI

REPORTING

DATA VISUALIZATION

INSIGHTS

REAL TIME ALERTING

SEARCH / QUERYING

ENTEPRISE DATA WAREHOUSE

BIG DATA GOVERNANCE

Big Data Governance

- An overarching layer of **governance, security and monitoring** is integral to the entire data flow.
- Governance plays a critical role in establishing and enforcing rules, policies, and procedures governing data access, quality, and usability.
- This ensures consistency in information and responsible data use.
- Tools such as Apache Atlas can be implemented to add this governance layer.
- Security protocols are in place to safeguard against unauthorized data access and ensure compliance with data protection regulations.
- These measures play a crucial role in protecting sensitive information and maintaining the integrity of data assets.

Big Data Tools

Open Source

- Processing Big Data involves ingesting, cleaning, and organizing the collected data to extract meaningful insights.
- Various tools and frameworks have been developed to assist in this process, offering different ways to handle and analyze large datasets:
- **Apache Hadoop**: Hadoop is an open-source framework designed to **process** and **store** Big Data across distributed clusters of computers.
- It comprises several components, including the Hadoop Distributed File System (HDFS) for **data storage** and MapReduce for **parallel data processing**.
- Hadoop enables fault-tolerance and horizontal scalability, making it an ideal solution for large-scale data processing tasks.

Open Source

- **Apache Spark**: Spark is another powerful open-source Big Data processing engine capable of handling batch and streaming data.
- It supports in-memory processing, which accelerates data processing tasks compared to Hadoop's MapReduce.
- Spark can be integrated with Hadoop and other storage systems, making it a versatile choice for various Big Data processing tasks, including machine learning and graph processing.

Open Source

- **NoSQL databases**: NoSQL databases are a category of databases that are designed for handling unstructured and semi-structured data.
- They provide a flexible and scalable system for storing and retrieving data, and they include several popular databases such as MongoDB, Couchbase, and Apache CouchDB.

Commercial

- The biggest public cloud providers at the moment are:
- **Amazon**: Recognized for its world-leading Amazon Web Services infrastructure, Amazon provides a comprehensive range of services, from backing up Kindle libraries to running advanced deep-learning models for platforms like Netflix.
- **Microsoft**: Azure, Microsoft's cloud platform, leverages the familiarity users have developed over decades with its software and standards, making it a preferred choice for many businesses.
- **Google**: In recent years, Google has rapidly expanded its cloud platform, Google Cloud Services, in an effort to compete with market leaders Amazon and Microsoft.

Thank you