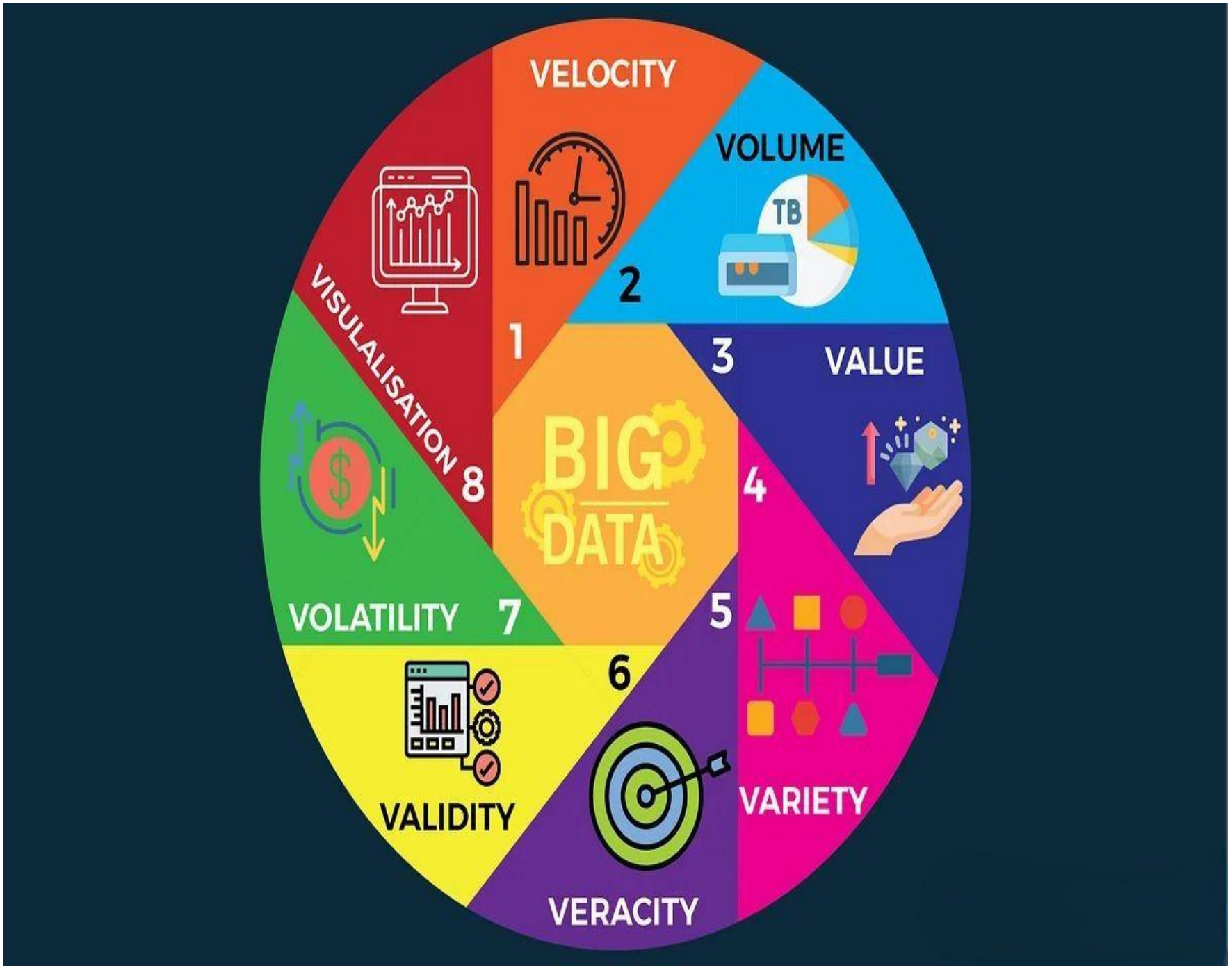
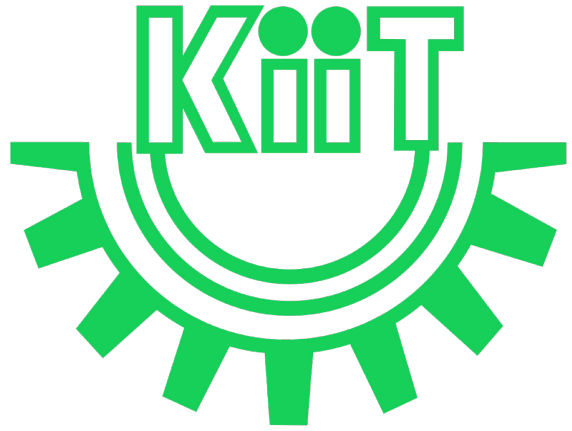


CS 3032: Big Data

Lec-7

In this Discussion . . .

- Exploring the Big Data Stack
  - Physical Infrastructure Layer
  - Platform Management Layer
  - Security Layer
  - Monitoring layer
  - Analytics Engine
  - Visualization Layer



Physical Infrastructure Layer

- Usually, Big Data analytics is based on the principles of:
  - **Performance**:
    - High-end infrastructure is required to deliver high performance with low **latency** (**the total time taken by a packet to travel from one node to another node**).
    - It is measured end-to-end, on the basis of a single transaction or query request.
    - Performance is rated high if the total time taken in processing a query request is low.
  - **Availability:** The infrastructure setup must be available at all times to ensure nearly a 100% uptime guarantee of service.

Physical Infrastructure Layer (Contd.)

- Usually, Big Data analytics is based on the principles of:
  - **Scalability:** The infrastructure must be **scalable** enough to **accommodate varying storage and computing requirements**. It must be capable enough to deal with any unexpected challenges.
  - **Flexibility:** Flexible infrastructures facilitate adding more resources to the setup and promote failure recovery.
  - **Cost**: Affordable infrastructure must be adopted including hardware, networking and storage requirements. Such parameters must be considered from the overall budget and trade-offs can be made, wherever necessary.

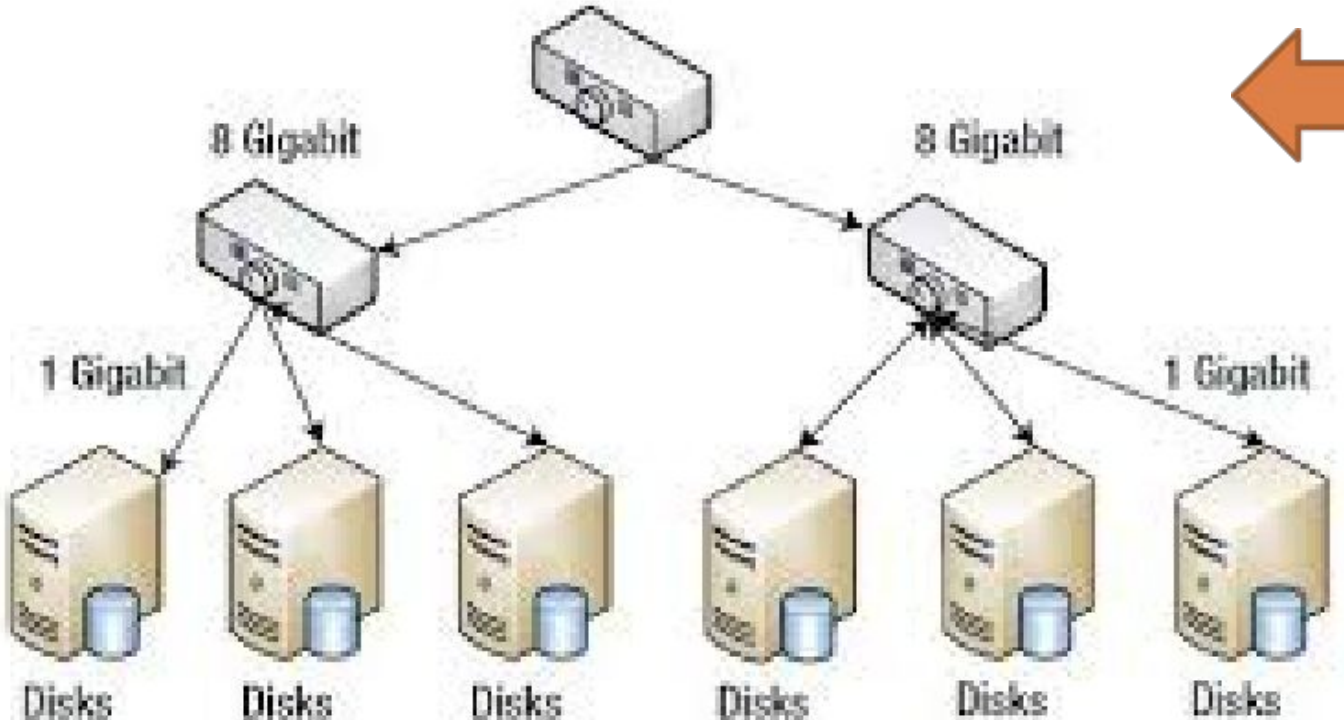
Physical Infrastructure Layer (Contd.)

- So it can be concluded that a **robust and inexpensive physical infrastructures needs to be implemented for Big Data**. *This requirement is handled by the* **Hadoop physical infrastructure layer**.

  - The **Hadoop physical infrastructure layer** is based on distributed computing model, which allows the physical storage of data in many different locations be linked with each other through networks & distributed file systems.
  - It also supports data redundancy.

Physical Infrastructure Layer (Contd.)

- This layer takes care of the hardware and network requirements and can provide a virtualized cloud environment or a distributed grid of commodity servers over a fast gigabit network.
- An example scenario of a hardware topology used for Big Data Implementation is as follows: [The rack is a physical collection of nodes in our Hadoop cluster].



The main components of a Hadoop infrastructure:

1. *n commodity servers (8-corem 24GBs RAM, 4 to 12 TBs)*
2. *2-level network (20 to 40 nodes per rack)*

Physical Infrastructure Design Considerations

- **Physical Redundant Networks:** In the Big data environment, networks should be redundant and capable of accommodating the anticipated volume and velocity of the inbound and outbound data in case of heavy network traffic.

    - The strategy must be prepared for improving the network performance to handle the increase in the volume, velocity, and variety of data.

Physical Infrastructure Design Considerations (Contd.)

- **Physical Redundant Networks:** In the Big data environment, networks should be redundant and capable of accommodating the anticipated volume and velocity of the inbound and outbound data in case of heavy network traffic.

  - *Network redundancy* is a process through which additional or alternate instances of network devices, equipment and communication mediums are installed within network infrastructure.

Physical Infrastructure Design Considerations (Contd.)

- **Physical Redundant Networks:** In the Big data environment, networks should be redundant and capable of accommodating the anticipated volume and velocity of the inbound and outbound data in case of heavy network traffic.

  - It is a method for ensuring network availability in case of a network device or path failure and unavailability.

Physical Infrastructure Design Considerations

- **Managing Hardware: Storage and Servers** – Hardware resources for storage and servers must have sufficient speed and capacity to handle all expected types of Big Data. If slow servers are connected to high-speed networks, the slow performance of the servers will be little use and can at times also become a bottleneck.
- **Infrastructure Operations -** Proper management of data handling operations provides a well-managed environment, which in turn gives the greatest levels of performance and flexibility.

Platform Management Layer

- This layer provides tools and programming languages for NoSQL databases.
- This layer uses HDFS on top of Hadoop Physical infrastructure layer.
- Hadoop contains various tools to help store, access and analyse large volumes of streaming data using real time analysis tools.

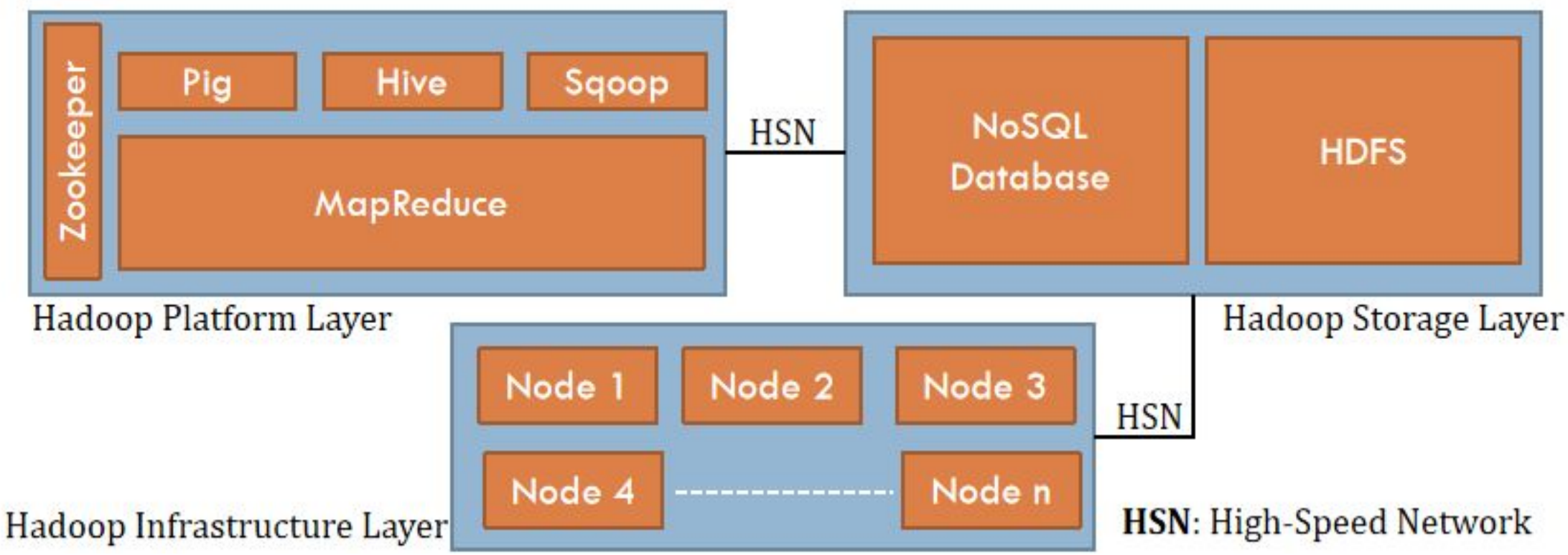
Key elements of Platform Management Layer are:

- Pig
- ZooKeeper
- Hive
- Sqoop
- MapReduce

Platform Management Layer (Contd.)

- Redundancy is built into this infrastructure for the very simple reason that we are dealing with large volume of data from different sources.
- The key building blocks of the Hadoop platform management layer is **MapReduce** programming which executes set of functions against a large amount of data in batch mode.
- The map function does the distributed computation task while the reduce function combines all the elements back together to provide a result.

Platform Management Layer (Contd.)



Security Layer

- It handles all the security measures that must be included in Big Data model and Big Data Architecture.
- **Big Data uses distributed systems in security layer**. Some security checks that are a prerequisite for security in big data are:
  - It must authenticate nodes by using protocols such as **Kerberos** (protocol for authenticating service requests between trusted hosts across an untrusted network, such as the internet).
  - It must enable file-layer encryption.
  - It must subscribe a key management service for trusted keys and certificates.

Security Layer (Contd.)

- Big Data uses distributed systems in security layer. Some security checks that are a prerequisite for security in big data are:
  - It must maintain logs of the communication that occurs between nodes and trace any anomalies across layers by using distributed logging mechanisms.

Security Layer (Contd.)

- Big Data uses distributed systems in security layer. Some security checks that are a prerequisite for security in big data are:
  - It must ensure a secure communication between nodes by using the Secure Sockets Layer (SSL)
  - It must validate data during the deployments of datasets or while applying service patches on virtual nodes

Monitoring Layer

- Monitoring Layer uses monitoring systems that provide machine communication and monitoring.
- These monitoring systems remain aware of all the configurations and functions of the OS as well as the hardware.
- The machine communication is provided with the help of high level protocols like XML. Monitoring systems also provide tools for data storage and visualization.
- Some tools for monitoring big data are Ganglia and Nagios.

Analytics Engine

- Analytics layer contains analytics engine that is used to analyse huge amount of data (usually unstructured). The analysis can be text analysis, statistical analysis etc.
- Big data Analytics Engines are classified into 2 types:

Search Engines	Real-Time Engines
• It requires very fast search engines and cognitive data discovery system to analyse tremendous volumes of data. • The data must be indexed and searched for analytical processing.	• Real time applications generate high volumes of data at a very fast speed. • Real time engines are required to perform analysis for big data environment for this type of processing.

Analytics Engine (Contd.)

- Some statistical and numerical methods used for analyzing various unstructured data sources are:
  - Natural Language Processing
  - Text Mining
  - Machine Learning
  - Linguistic Computation
  - Search and Sort Algorithms
  - Syntax and Lexical Analysis

Analytics Engine (Contd.)

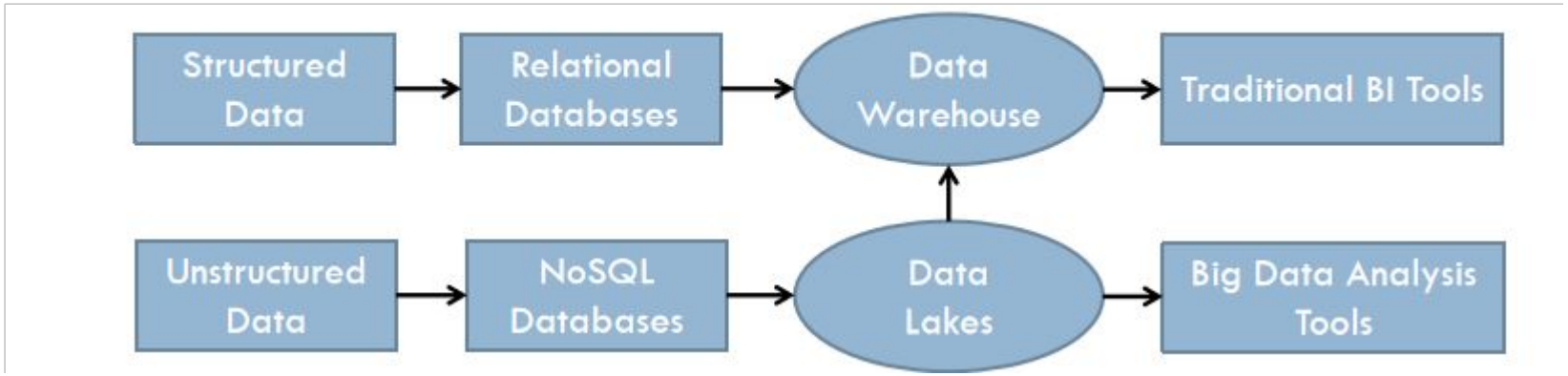
- Some examples of different types of unstructured data that are available as large dataset include the following:
  - Machine generated data such as RFID feeds and weather data
  - Documents containing textual patterns
  - Data generated from application logs about upcoming or down time details or about maintenance and upgrade details.

Visualization Layer

- Visualization layer is involved with the visualizing and interpreting of big data.
- It is very crucial as it gives a great deal of information about the data at a glance.
- Data Visualization has many techniques and methods which can be used for simulations and also deriving conclusions of the big data.

Visualization Layer (Contd.)

- It works on top of data warehouses and Operational Data Stores (ODS). Some examples of popular visualization and dashboard tools are **Tableau**, **Spotfire**, **D3**, **DataWrapper** etc.
- These tools work on the top of the traditional components such as reports, dashboards, scorecards, and queries.



Visualization Layer (Contd.) : Data Lake

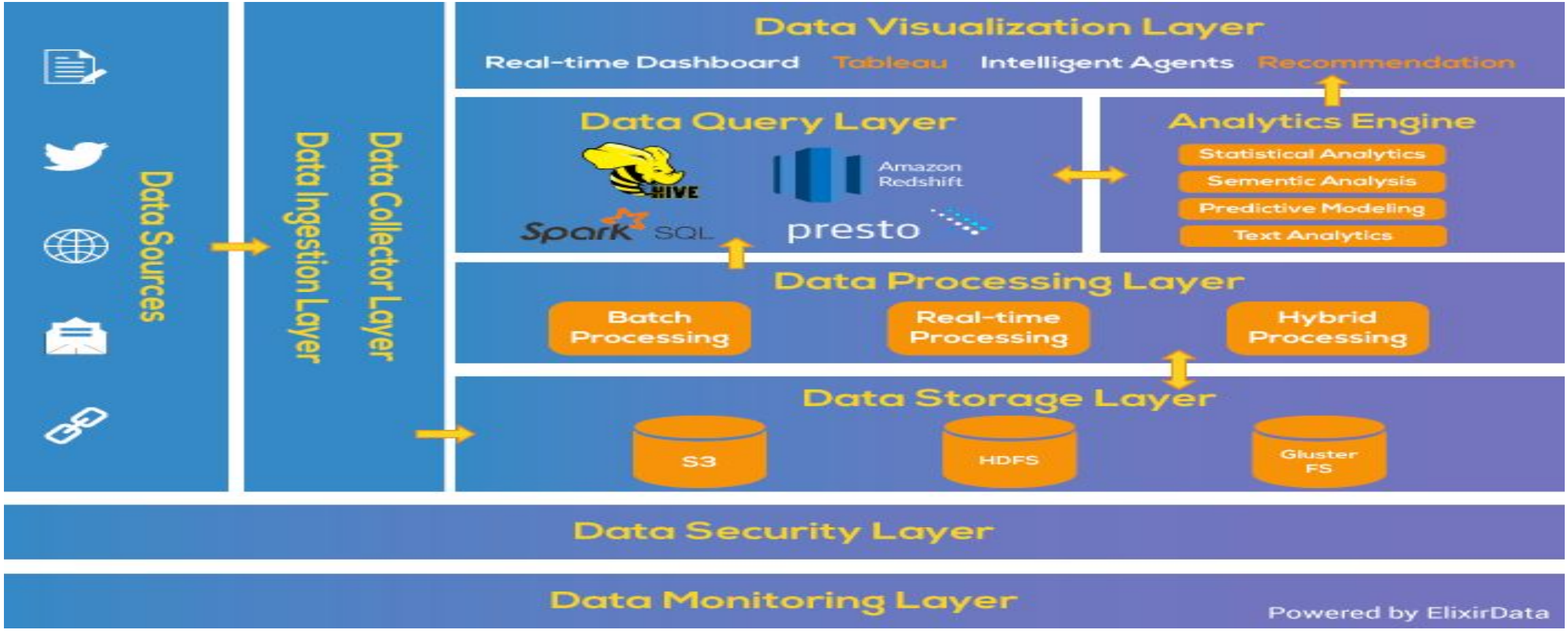
- A data lake is a centralized repository that allows to store all your structured and unstructured data at any scale.
- It can store data as-is, without having to first structure the data, and run different types of analytics from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	schema-on-write	schema-on-read
Users	Business analysts	Data scientists, and Business analysts
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, and data discovery

Visualization Layer (Contd.)

- Visualization in Visualization Layer can be carried out with the help of the following approaches: **Server Visualization**, **Network Visualization**, **Data** and **Storage Visualization**, **Application Visualization**.

Big Data Architecture Layers



References

1. https://csveda.com/big-data-architecture-layers/
2. https://www.rcvacademy.com/big-data-layers/
3.