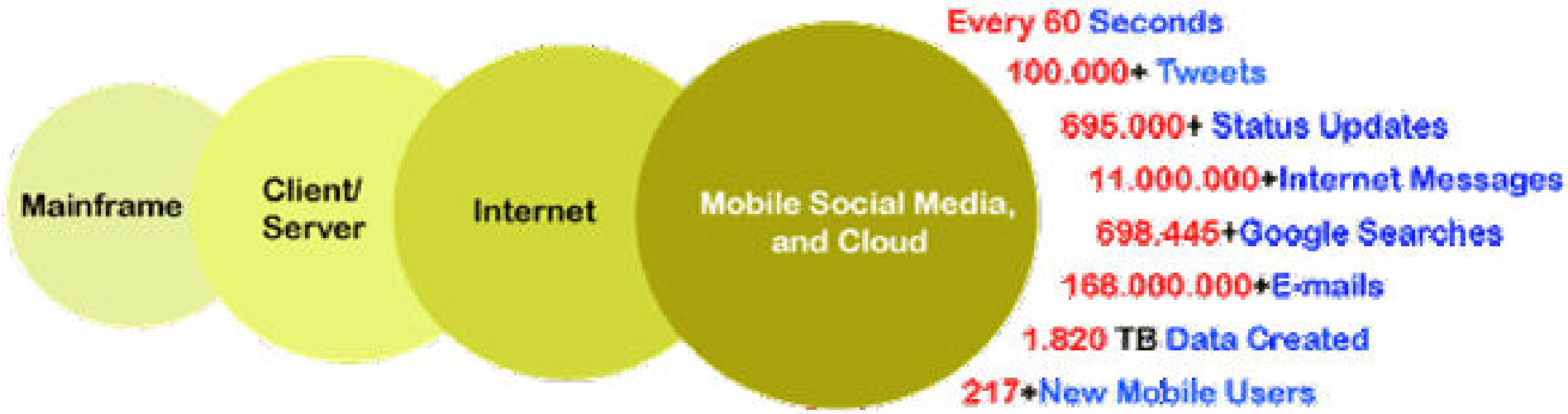
Big Data Concepts and Techniques

Lecturer: Narges Peyravi

What is Big Data?

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.

- **Big Data** is a term used for a collection of data sets that are large and complex, which is difficult to store and process using available database management tools or traditional data processing applications. The challenge includes capturing, curating, storing, searching, sharing, transferring, analyzing and visualization of this data.

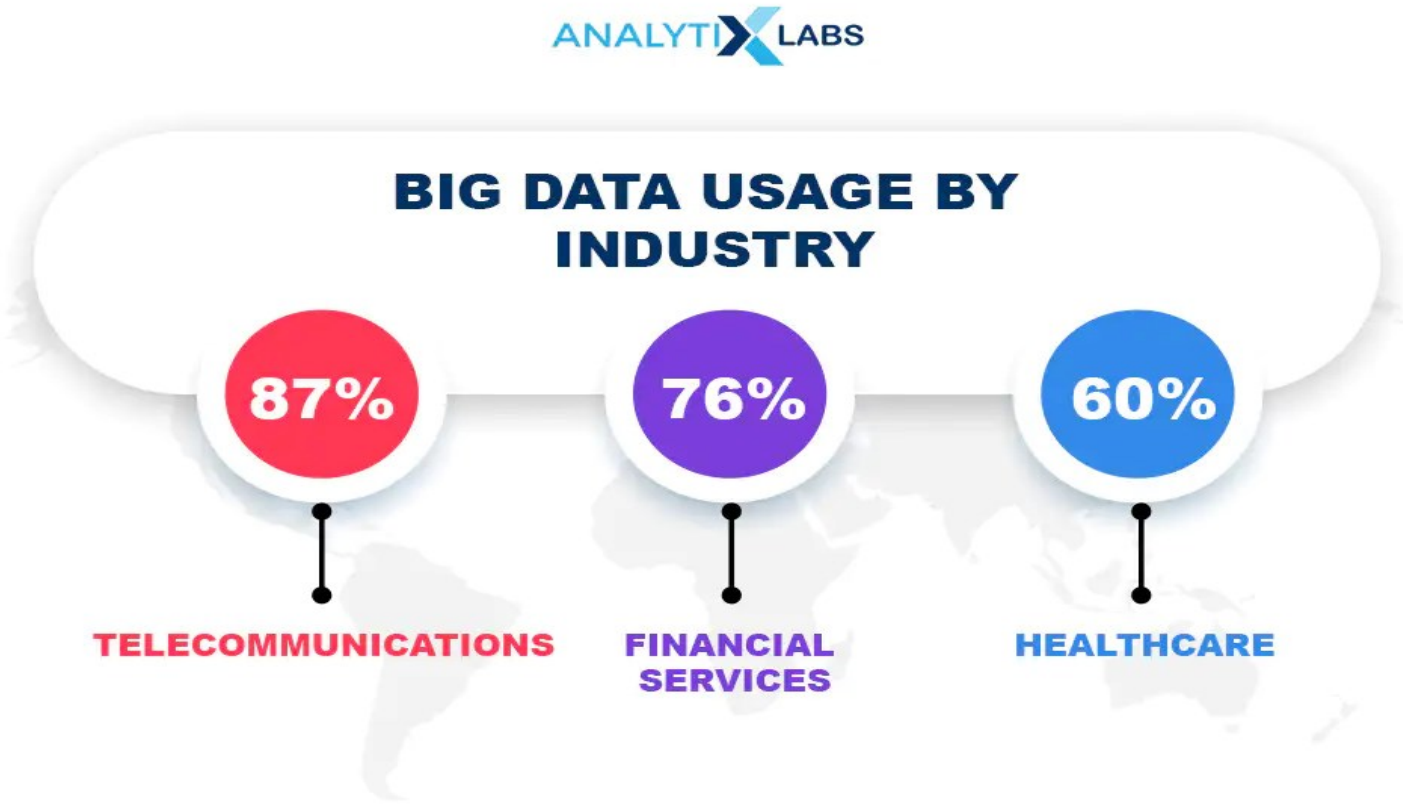


Examples of Big Data

- Daily we upload millions of bytes of data. 90 % of the world's data has been created in last two years.
- Walmart handles more than **1 million** customer transactions every hour.
- Facebook stores, accesses, and analyzes **30+ Petabytes** of user generated data.
- **230+ millions** of tweets are created every day.
- More than **5 billion** people are calling, texting, tweeting and browsing on mobile phones worldwide.
- YouTube users upload **48 hours** of new video every minute of the day.
- Amazon handles **15 million** customer click stream user data per day to recommend products.
- **294 billion** emails are sent every day. Services analyses this data to find the spams.
- Modern cars have close to **100 sensors** which monitors fuel level, tire pressure etc. , each vehicle generates a lot of sensor data.

Applications of Big Data

- Healthcare
- Telecom
- E-commerce
- Traffic control
- Manufacturing
- Search Quality
- media and entertainment
- IoT
- Education
- Banking
- Government
- …….



Main Components of Big Data

1. Ingestion

Ingestion refers to the process of gathering and preparing the data. You'd use the ETL (extract, transform, and load) process to prepare your data. In this phase, you have to identify your data sources, determine whether you'll gather the data in batches or stream it, and prepare it through cleansing, massaging, and organization. You perform the extract process in gathering the data and the transformation process in optimizing it.

2. Storage

Once you have gathered the necessary data, you'd need to store it. Here, you'll perform the final step of the ETL, the load process. You'd store your data in a data warehouse or a data lake, depending on your requirements. This is why it's crucial to understand your organization's goals while performing any big data process.

3. Analysis

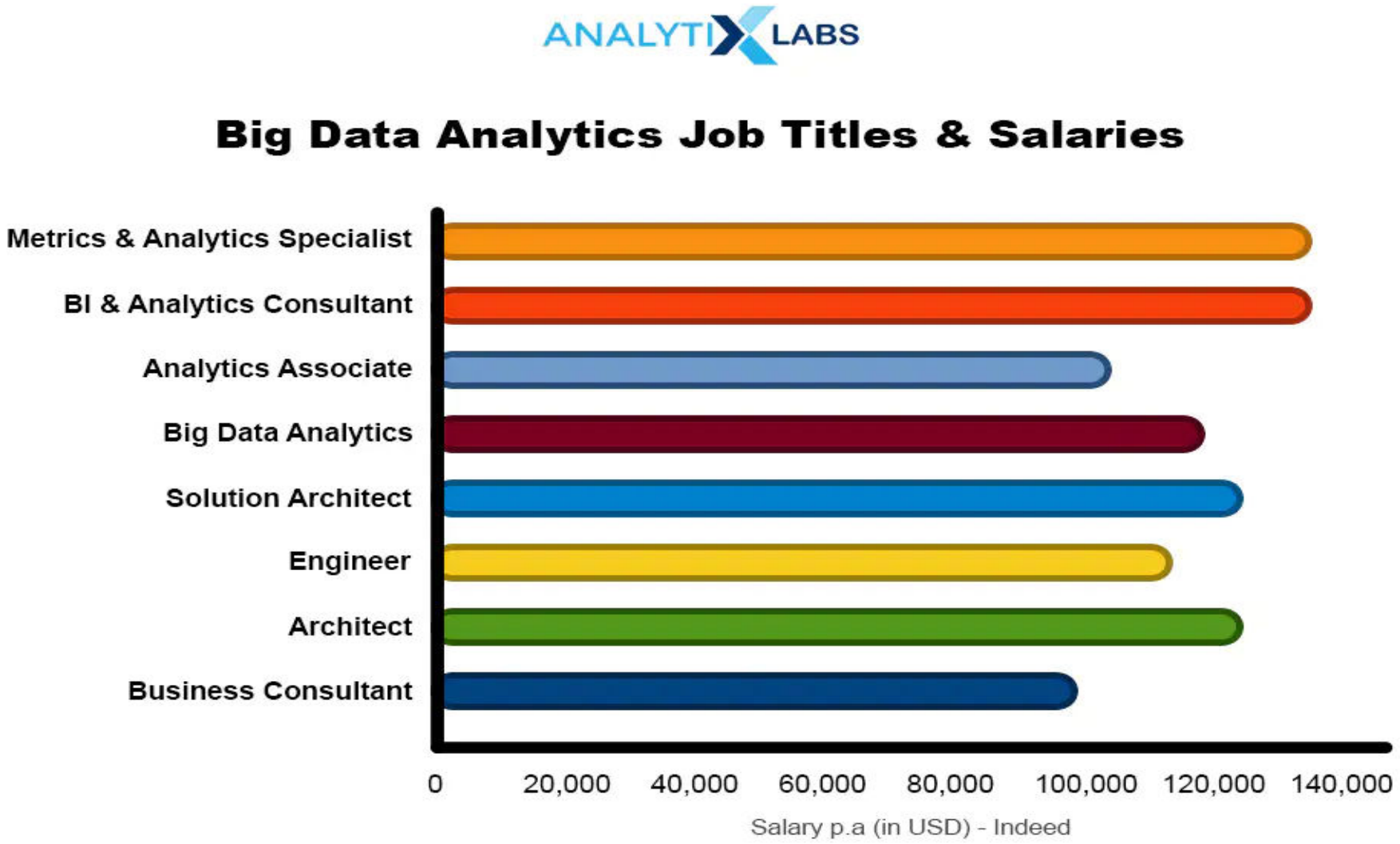
In this phase of your big data process, you'd analyze the data to generate valuable insights for your organization. There are four kinds of big data analytics: prescriptive, predictive, descriptive, and diagnostic. You'd use artificial intelligence and machine learning algorithms in this phase to analyze the data.

Main Components of Big Data

4. Consumption

This is the final phase of a big data process. Once you have analyzed the data and have found the insights, you have to share them with others. Here, you'd have to utilize data visualization and data storytelling to share your insights effectively with a non-technical audience such as stakeholders and project managers.

Big Data Analytics Job Title and Salaries



Advantages of Big Data

There are numerous advantages of Big Data for organizations. Some of the key ones are as follows:

. **Enhanced Decision-making**

Big data implementations can help businesses and organizations make better-informed decisions in less time. It allows them to use outside intelligence such as search engines and social media platforms to fine-tune their strategies. Big data can identify trends and patterns that would've been invisible otherwise, helping companies avoiding errors.

. **Data-driven Customer Service**

Another huge impact big data can have on all industries is in the customer service department. Companies are replacing the traditional customer feedback system with data-driven solutions. Such solutions can analyze customer feedback more efficiently and help them offer customer service to the consumers.

Advantages of Big Data

.Efficiency Optimization

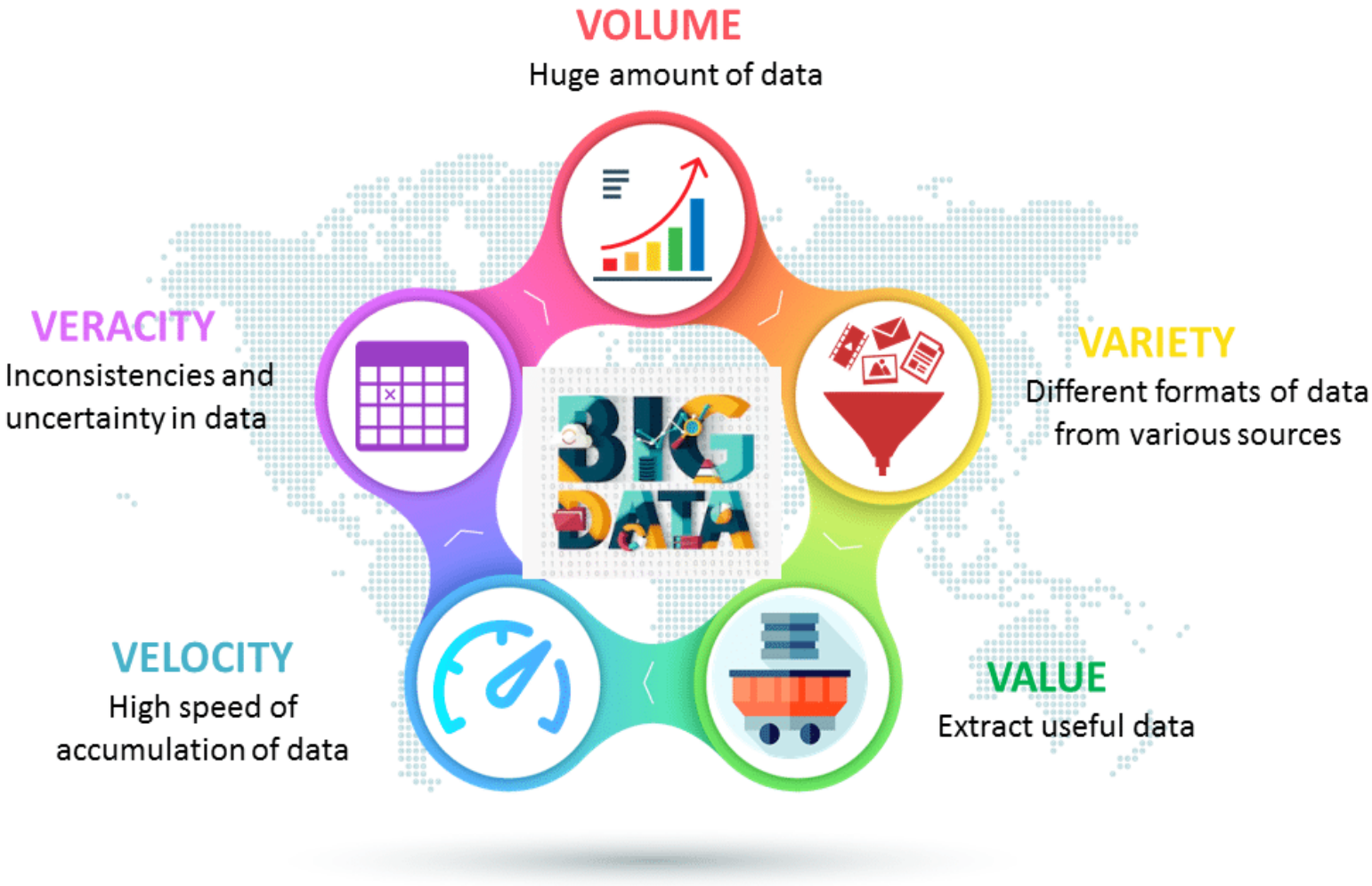
Organizations use big data to identify the weak areas present within them. Then, they use these findings to resolve those issues and enhance their operations substantially. For example, Big Data has substantially helped the manufacturing sector improve its efficiency through IoT and robotics.

.Real-time Decision Making

Big Data has transformed several areas by enabling real-time trackings, such as inventory management, supply chain optimization, anti-money laundering, and fraud detection in banking & finance.

……..

Big Data Characteristics(5V)



Big Data Characteristics(5V)

- ***VOLUME***
- Volume refers to the ‘amount of data’, which is growing day by day at a very fast pace. The size of data generated by humans, machines and their interactions on social media itself is massive. Researchers have predicted that 40 Zettabytes (40,000 Exabytes) will be generated by 2020, which is an increase of 300 times from 2005.
- ***VELOCITY***
- Velocity is defined as the pace at which different sources generate the data every day. This flow of data is massive and continuous. There are 1.03 billion Daily Active Users on Mobile as of now, which is an increase of 22% year-over-year. This shows how fast the number of users are growing on social media and how fast the data is getting generated daily.
- ***VARIETY***
- As there are many sources which are contributing to Big Data, the type of data they are generating is different. Earlier, we used to get the data from excel and databases, now the data are coming in the form of images, audios, videos, sensor data etc. Hence, this variety of data creates problems in capturing, storage, mining and analyzing the data.

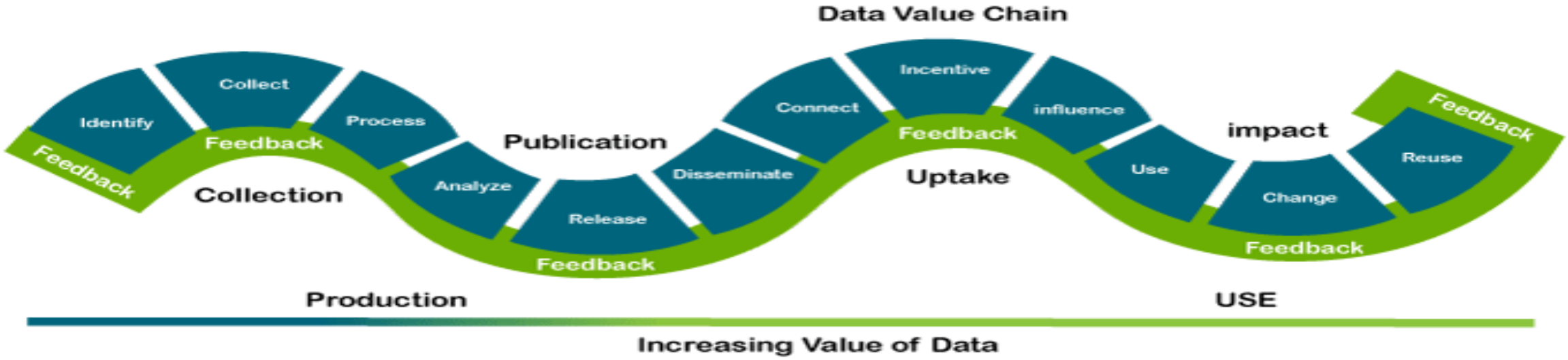
Big Data Characteristics(5V)

- **Veracity**

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development. For example, Facebook posts with hashtags.

- **Value**

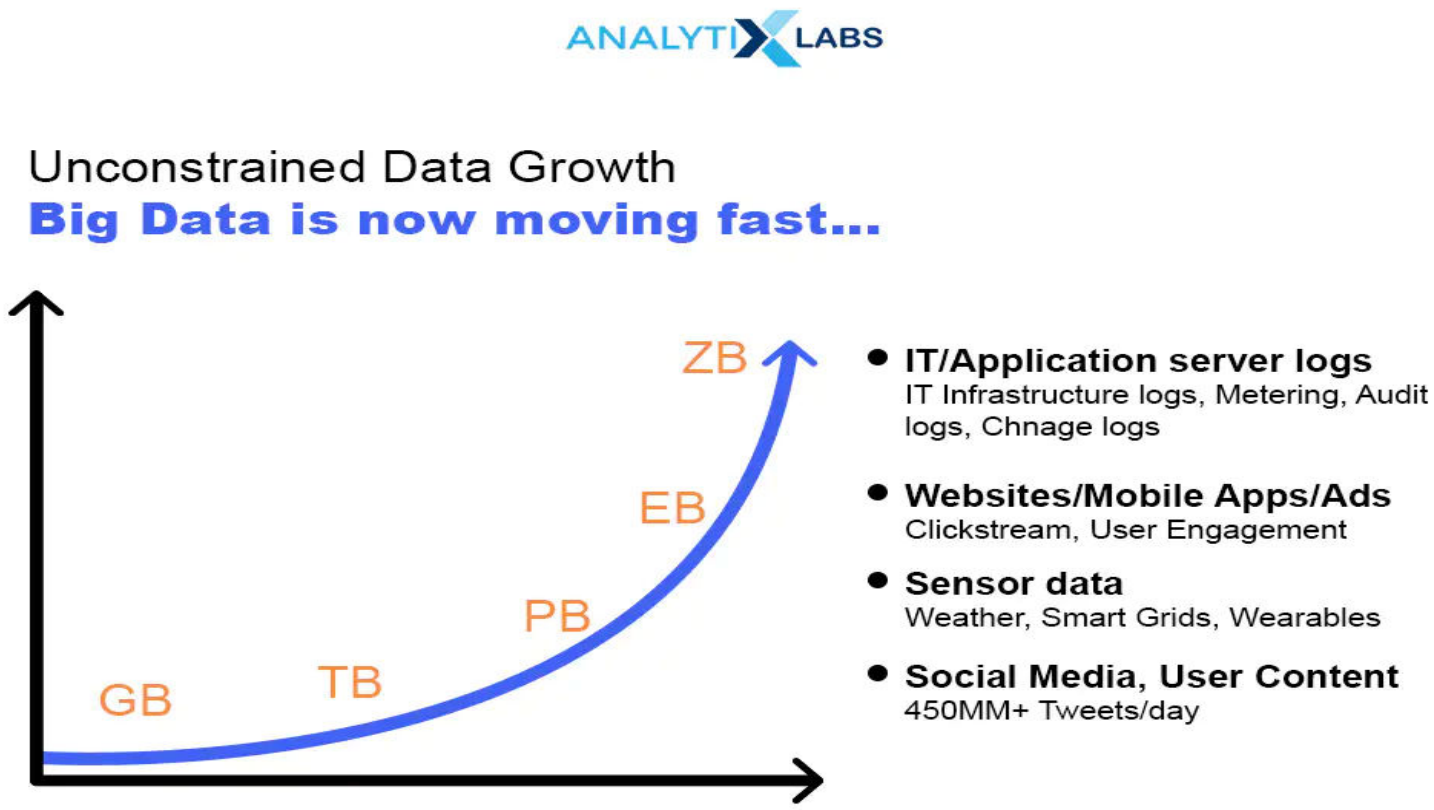
Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights.



What are the various domains in which big data arise?

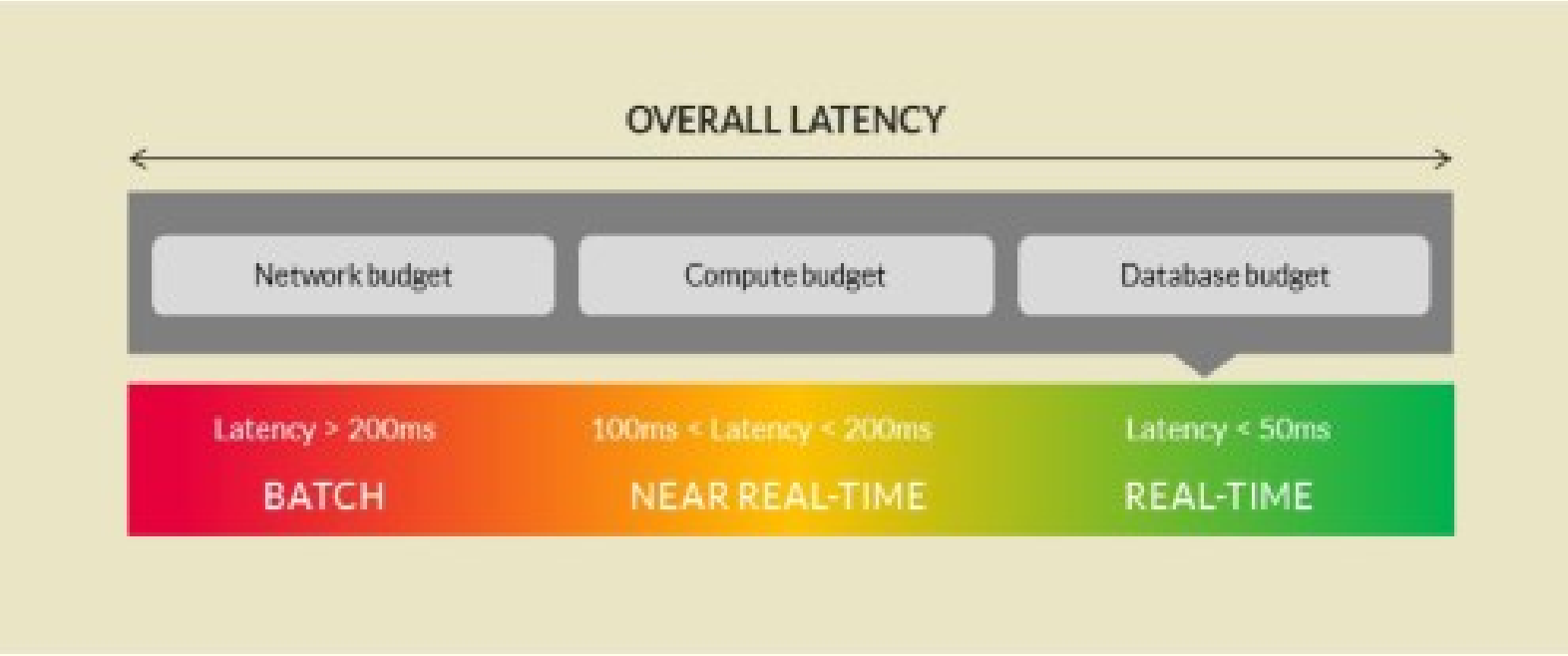
The particular domain in which data arises will determine the types of architecture that will be required to store it, process it, and perform analytics on it.

- **Latency**
- **Structure**
- **Domain**
- **Infrastructure**
- **Storage**
- **Analytics**



Latency

- Real-time (financial streams, complex event processing (CEP), intrusion detection, fraud detection)
- Near real-time (ad placement)
- Batch (retail, forensics, bioinformatics, geodata, historical data of various types)



Structure

Big Data could be of three types

Structured

Semi-Structured

Unstructured

Structure

- ***Structured***

The data that can be stored and processed in a fixed format is called as Structured Data. Data stored in a relational database management system (RDBMS) is one example of 'structured' data. It is easy to process structured data as it has a fixed schema. Structured Query Language (SQL) is often used to manage such kind of Data.

- ***Semi-Structured***
- Semi-Structured Data is a type of data which does not have a formal structure of a data model, i.e. XML files or JSON documents are examples of semi-structured data. Examples Of Semi-structured Data

Personal data stored in an XML file:

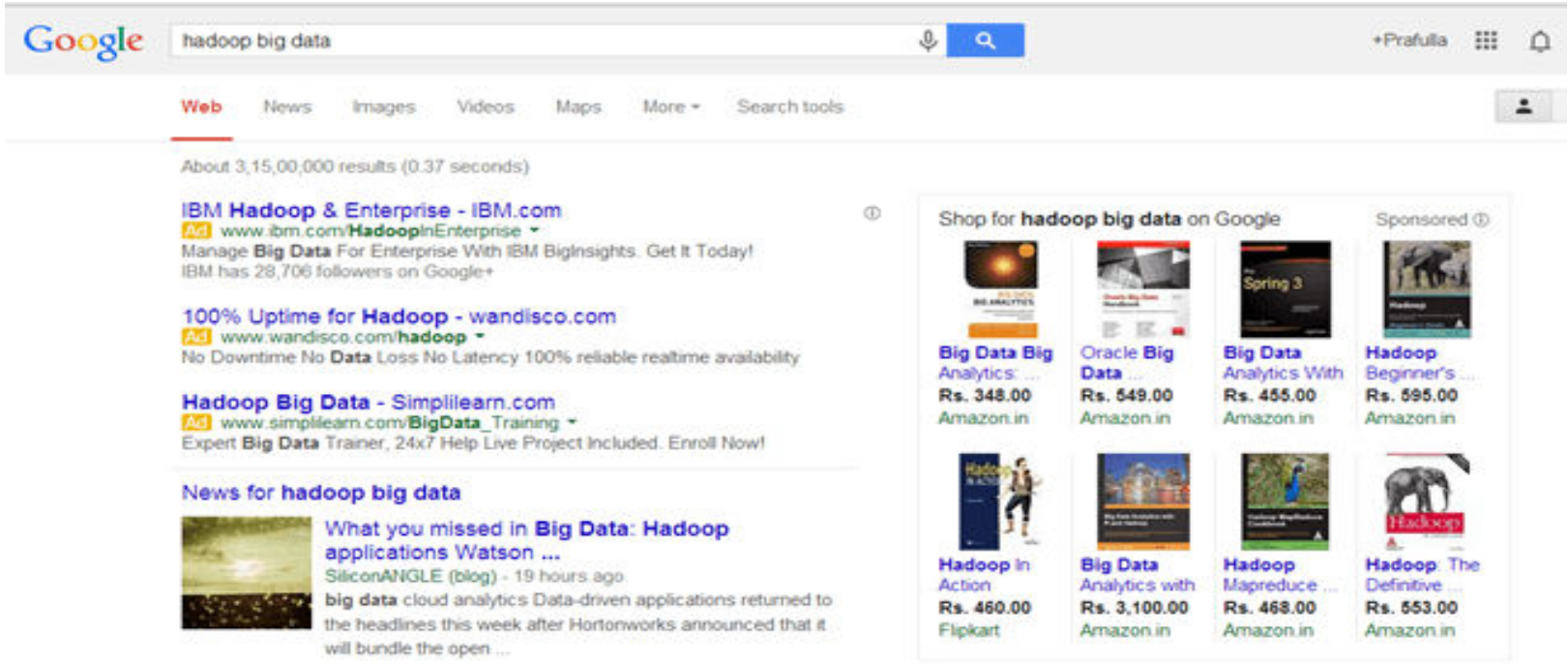
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

Structure

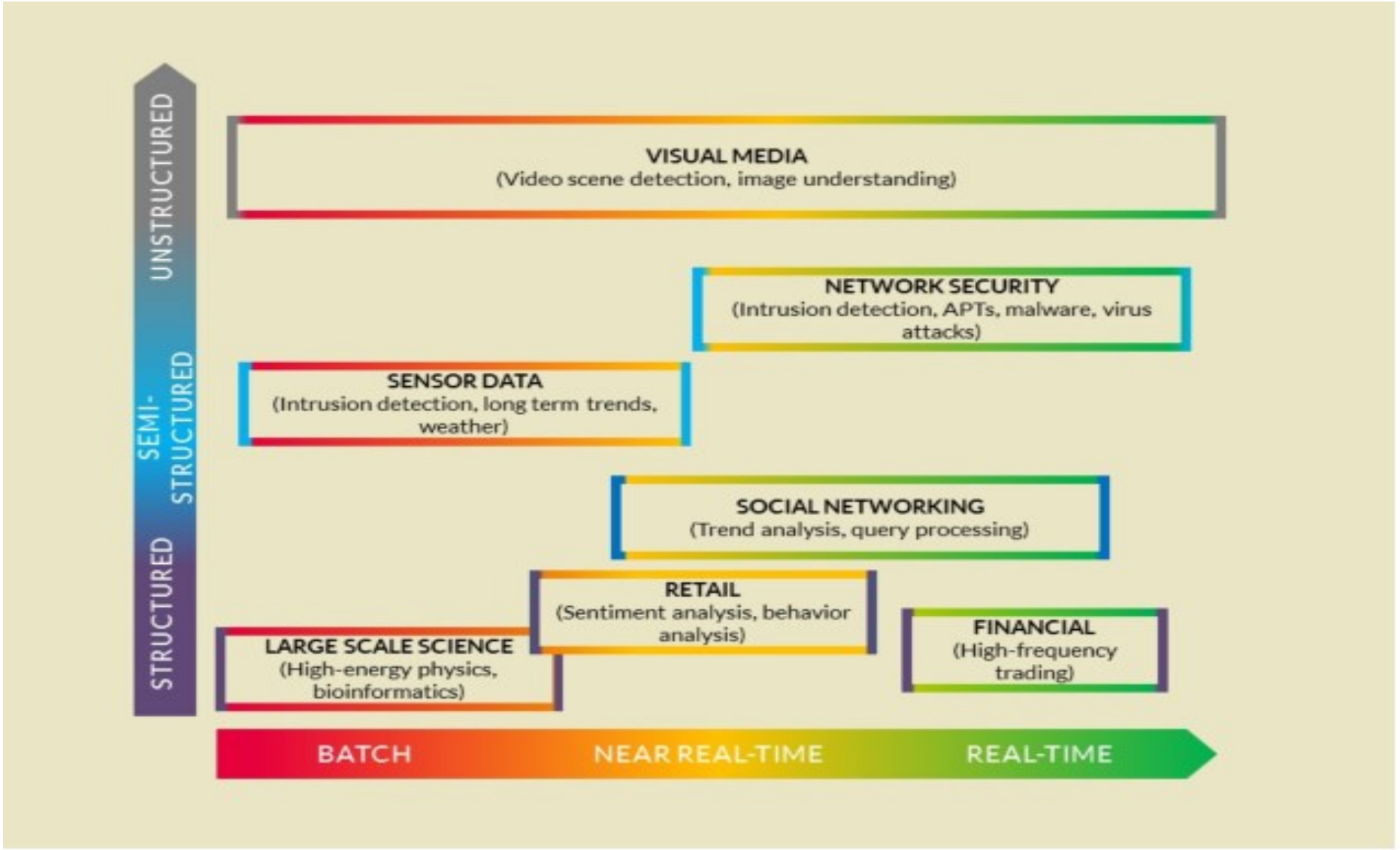
- ***Unstructured***

The data which have unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data. Text Files and multimedia contents like images, audios, videos are example of unstructured data.

Examples Of Un-structured Data: The output returned by ‘Google Search’:



Mapping the big data verticals

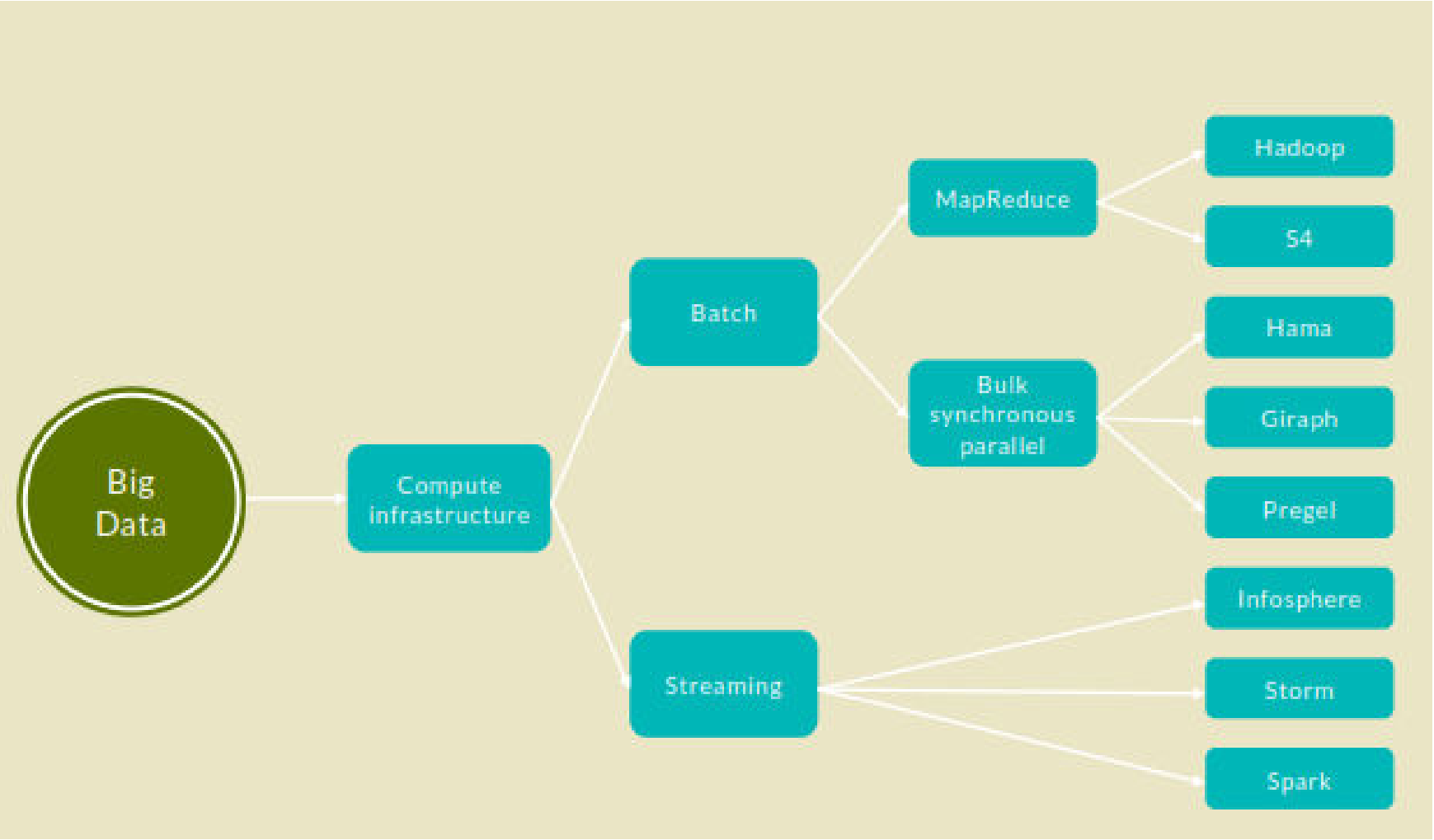


Domain

Figure illustrates the various domains and specific subdomains in which big data processing issues arise.

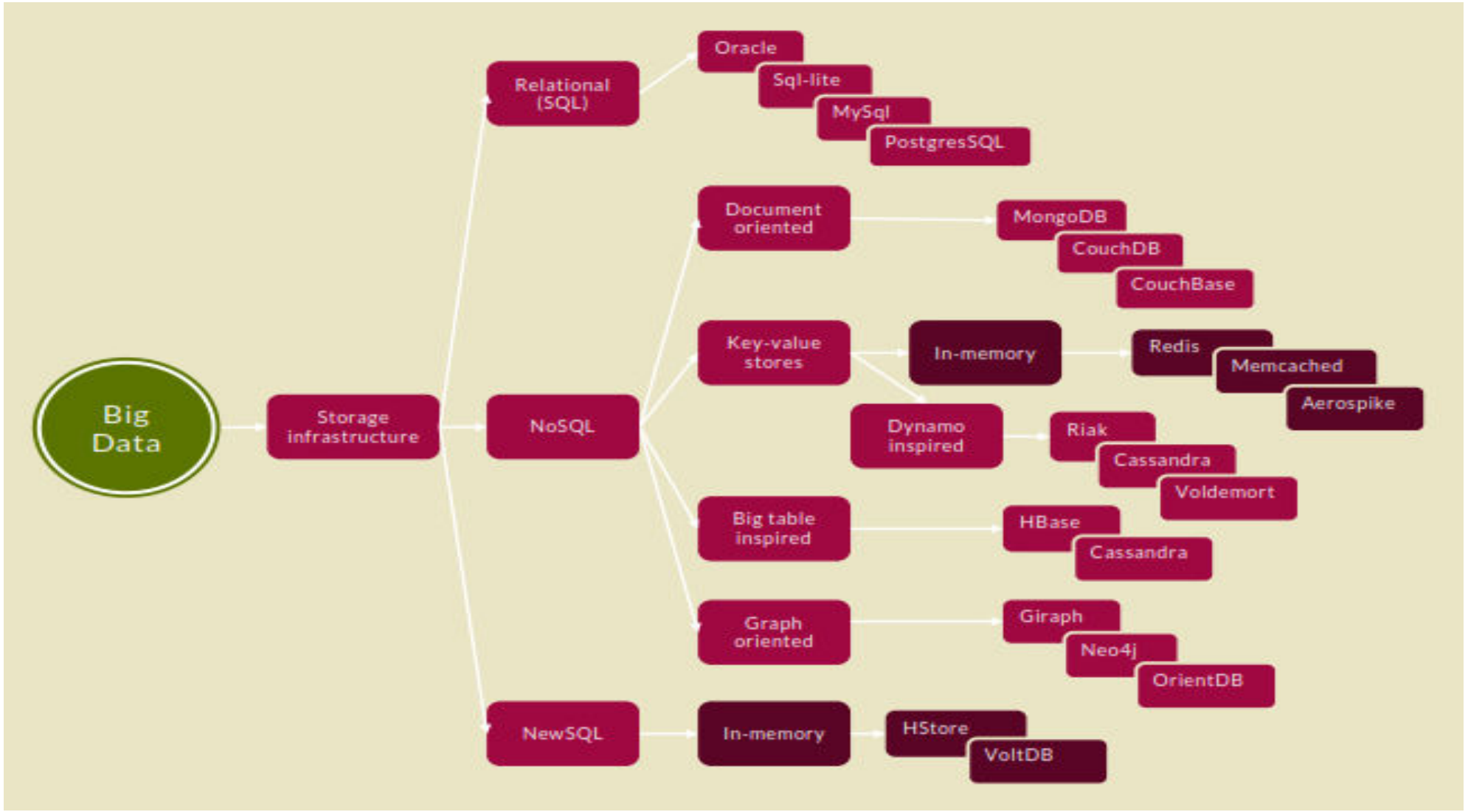
Infrastructure

Figure shows the taxonomy for the various styles of processing architectures.

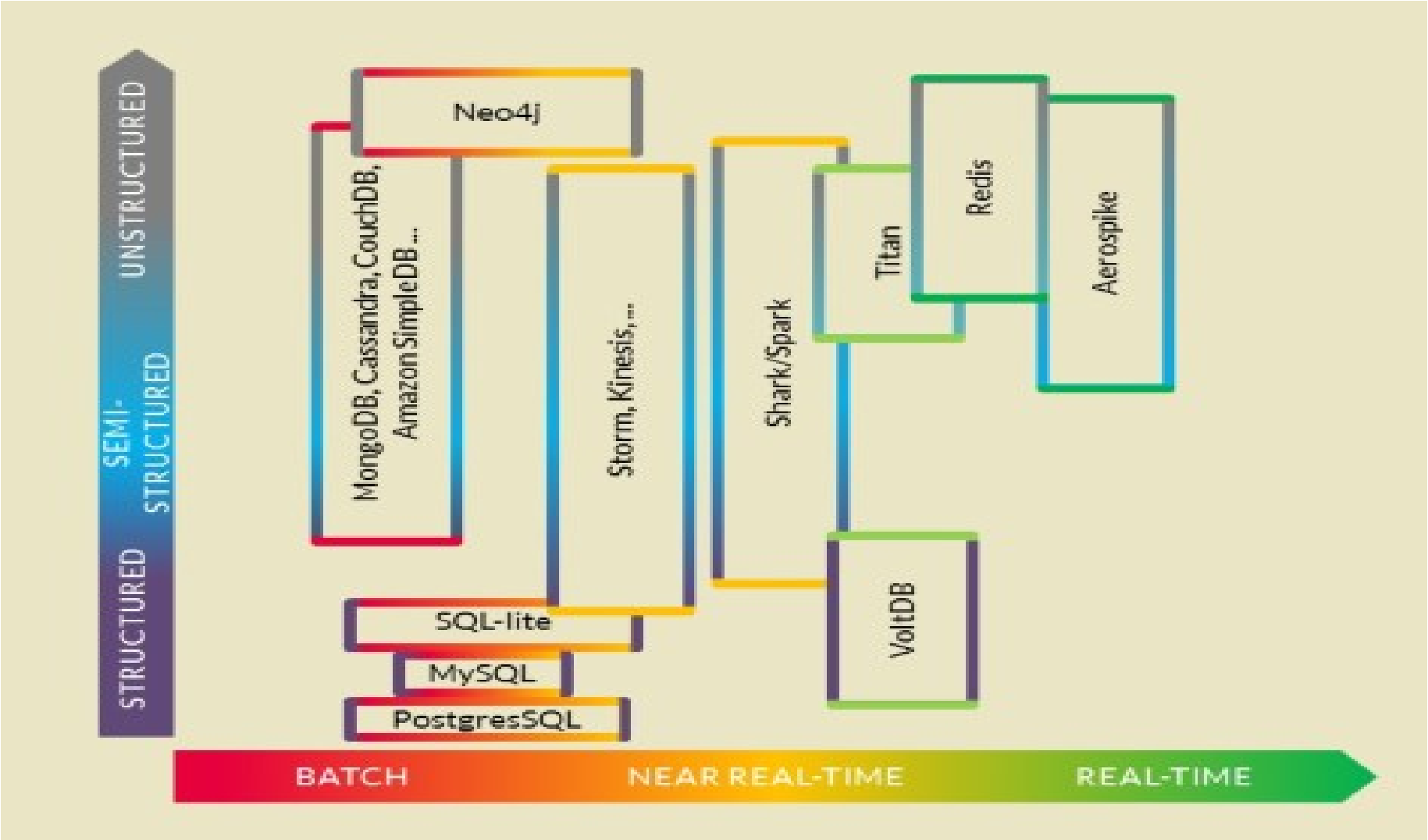


storage

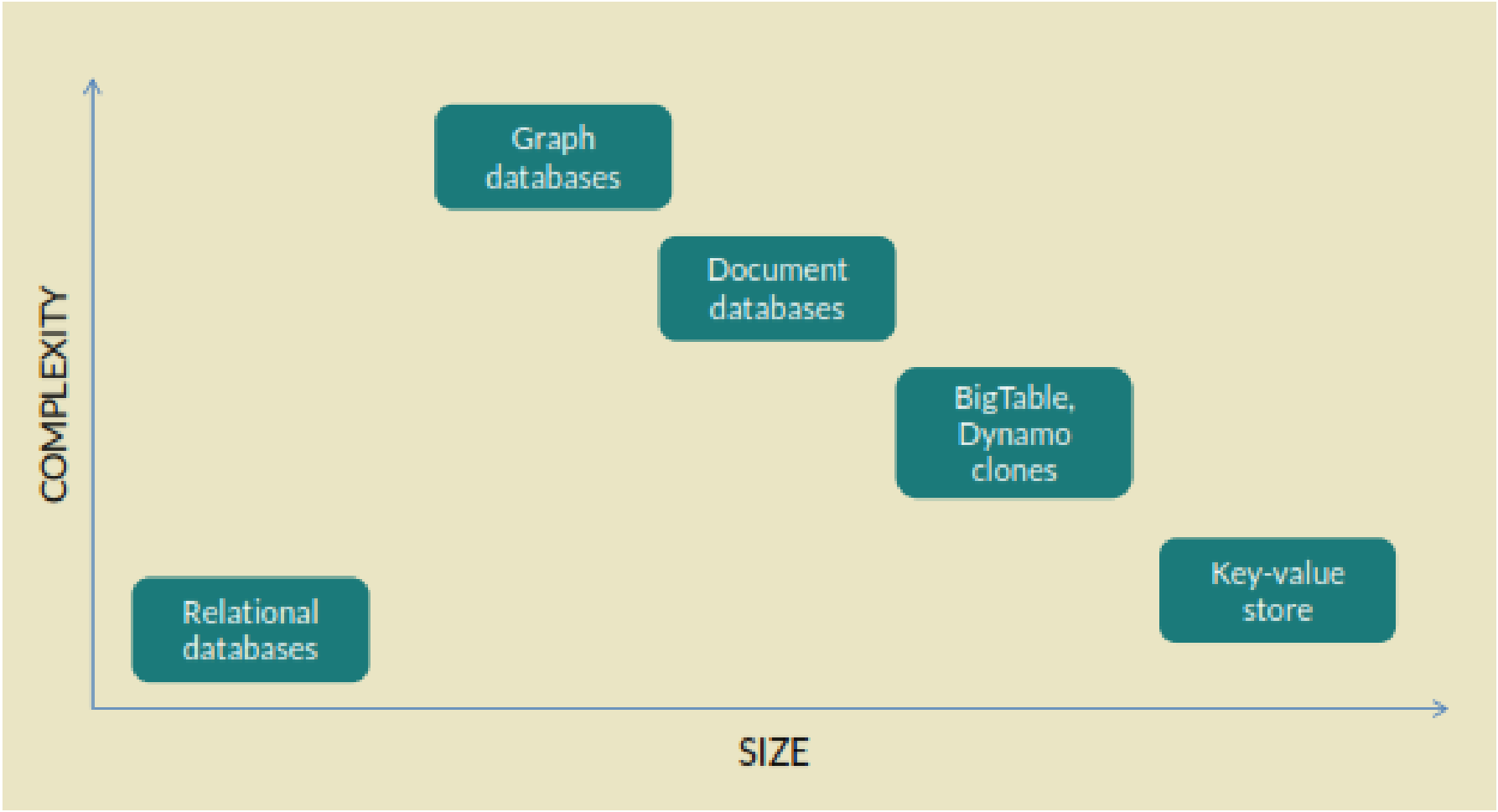
Figure shows the taxonomy for the various types of databases that are used for big data storage.



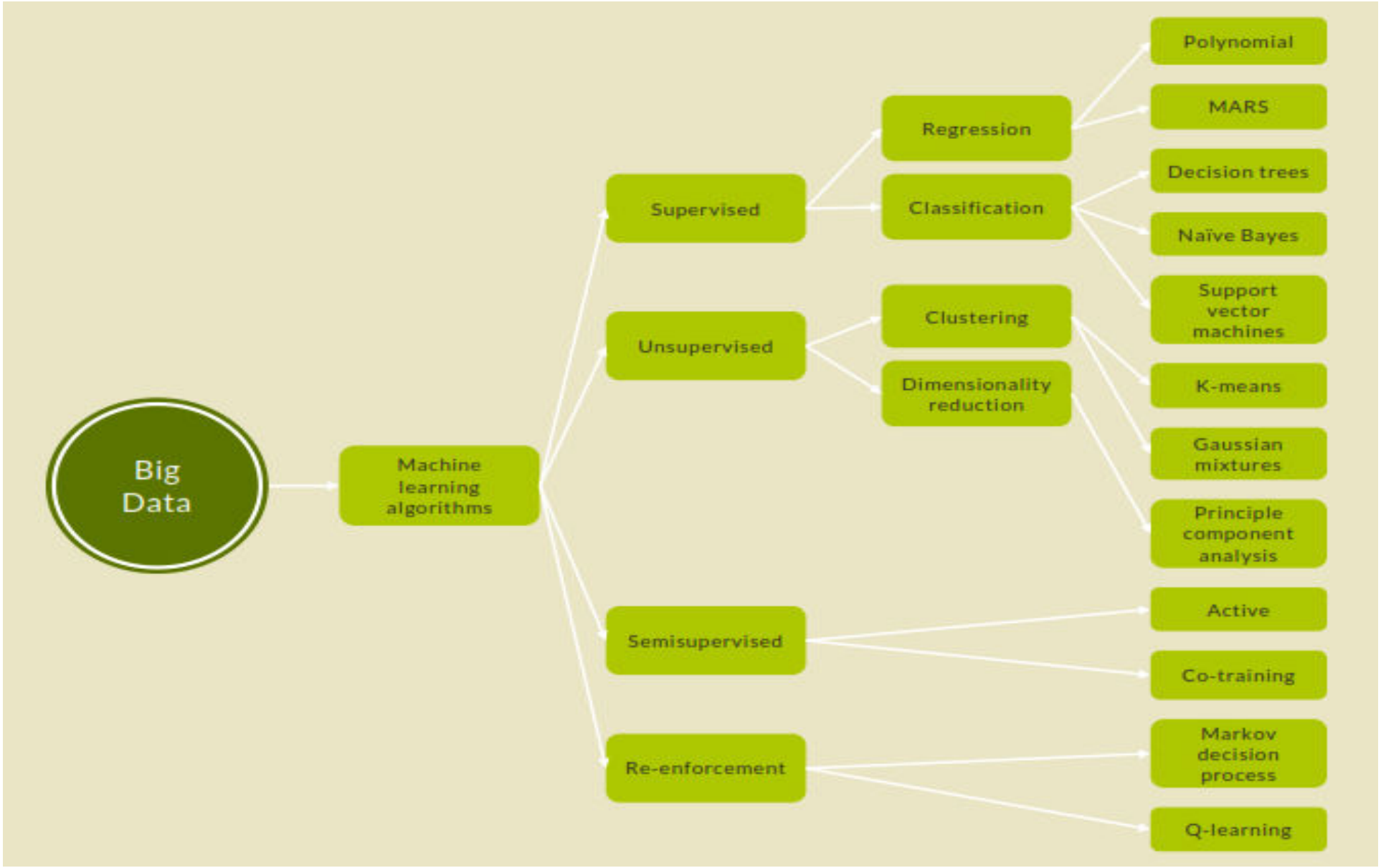
Storage technologies map



various categories of databases regarding the complexity of the data items



Analytics


Big Data
Machine learning algorithms
Supervised
Unsupervised
Semisupervised
Re-enforcement
Regression
Classification
Clustering
Dimensionality reduction
Polynomial
MARS
Decision trees
Naïve Bayes
Support vector machines
K-means
Gaussian mixtures
Principle component analysis
Active
Co-training
Markov decision process
Q-learning

Challenges with Big Data

- **Data Quality** –The data here is very messy, inconsistent and incomplete. Dirty data cost $600 billion to the companies every year in the United States.
- **Discovery** – Finding insights on Big Data is like finding a needle in a haystack. Analyzing petabytes of data using extremely powerful algorithms to find patterns and insights are very difficult.
- **Storage** – The more data an organization has, the more complex the problems of managing it can become. The question that arises here is “Where to store it?”. We need a storage system which can easily scale up or down on-demand.
- **Analytics** – In the case of Big Data, most of the time we are unaware of the kind of data we are dealing with, so analyzing that data is even more difficult.
- **Security** – Since the data is huge in size, keeping it secure is another challenge. It includes user authentication, restricting access based on a user, recording data access histories, proper use of data encryption etc.
- **Lack of Talent –** There are a lot of Big Data projects in major organizations, but a sophisticated team of developers, data scientists and analysts who also have sufficient amount of domain knowledge is still a challenge.

Top 10 Big Data Tools

- Apache Hadoop
- Apache Spark
- Apache Flink
- Apache Storm
- Apache Cassandra
- MongoDB
- Kafka
- CouchDB
- RapidMiner
- R Programming

Apache Hadoop

- Hadoop is an open-source framework from Apache and runs on commodity hardware. It is used to store process and analyze Big Data.
- Apache Hadoop enables parallel processing of data as it works on multiple machines simultaneously.
- It consists of 3 parts-
  - **Hadoop Distributed File System (HDFS) –** It is the storage layer of Hadoop.
  - **Map-Reduce –** It is the data processing layer of Hadoop.
  - **YARN –** It is the resource management layer of Hadoop.
- Hadoop does not support real-time processing. It only supports batch processing.
- Hadoop cannot do in-memory calculations.
- Hadoop is written in Java.

Apache Spark

- Apache Spark is an open source framework for data analytics, machine learning algorithms, and fast cluster computing.
- Apache Spark™ is a unified analytics engine for large-scale data processing.
- Spark, unlike Hadoop, supports both real-time as well as batch processing. It is a general-purpose clustering system.
- It was built on top of Hadoop MapReduce and it extends the MapReduce model to efficiently use more types of computations which includes Interactive Queries and Stream Processing.
- The main feature of Spark is its **in-memory cluster computing** that increases the processing speed of an application.
- Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud.

Apache Flink

- Apache Flink is an open source platform which is a streaming data flow engine that provides communication, fault-tolerance, and data-distribution for distributed computations over data streams.
- Flink is a top-level project of Apache.
- It is written in Java and Scala. It provides high accuracy results even for late-arriving data.
- Apache Flink provides APIs for creating several applications which use flink engine.
- It is a scalable **data analytics** framework that is fully compatible with **Hadoop**.
- Flink can execute both stream processing and batch processing easily.
- Apache Flink excels at processing unbounded and bounded data sets.

Apache Storm

- Apache Storm is a cross-platform, distributed stream processing, and fault-tolerant real-time computational framework.
- It is free and open-source.
- It is written in Clojure and Java.
- Its architecture is based on describe sources of information and manipulations in order to permit batch, distributed processing of unbounded streams of data.
- Among many, Groupon, Yahoo, Alibaba, and The Weather Channel are some of the famous organizations that use Apache Storm.
- It has very low latency.

Apache Cassandra

- Cassandra is a distributed database management system developed by Apache Software Foundation in 2008. It uses techniques based on NoSQL and is an open source software.
- Cassandra has mainly been used in big data applications which use real-time data such as those from sensor components or from social networking websites.

The key features of Apache Cassandra include:

- Ability to function on less powerful hardware.
- Cassandra architecture implements a key-value database system.
- Cassandra Query Language
- Distributed deployment and high application scalability
- Fault tolerance and decentralized system
- Apache Cassandra performs fast read/write functions.
- Tunable consistency and MapReduce Support

MongoDB

- **MongoDB** is a document-oriented NoSQL database used for high volume data storage. Instead of using tables and rows as in the traditional relational databases, MongoDB makes use of collections and documents.
- Documents consist of key-value pairs which are the basic unit of data in MongoDB. Collections contain sets of documents and function which is the equivalent of relational database tables.
- MongoDB stores data in a JSON document format. The fact that it is a JSON store also means that information can be retrieved very quickly.
- MongoDB addresses the “Variety” aspect of Big Data. It deals with the ways to represent different data types efficiently with colossal read/write scalability and huge availability of transactional systems in real time.
- MongoDB blends seamlessly with programming languages like JavaScript, Ruby and Python; this seamless blending conveys high coding velocity.

Apache Kafka

- Apache Kafka is a distributed data store optimized for ingesting and processing streaming data in real-time.
- Kafka is primarily used to build real-time streaming data pipelines and applications that adapt to the data streams. It combines messaging, storage, and stream processing to allow storage and analysis of both historical and real-time data.
- Kafka is written in Scala and Java.
- Kafka supports low latency message delivery and gives guarantee for fault tolerance in the presence of machine failures.
- Kafka was originally created at LinkedIn, where it played a part in analysing the connections between their millions of professional users in order to build networks between people.
- Kafka was originally designed to track the behaviour of visitors to large, busy websites (such as LinkedIn). By analysing the clickstream data (how the user navigates the site and what functionality they use) of every session, a greater understanding of user behaviour is achievable. This makes it possible to predict which news articles, or products for sale, a visitor might be interested in.

Apache CouchDB

- **Apache CouchDB** is an open-source document-oriented NoSQL database
- CouchDB is a document-oriented database and within each document fields are stored as key-value maps.
- CouchDB uses multiple formats and protocols to store, transfer, and process its data.
- It uses JSON to store data, JavaScript as its query language using MapReduce, and HTTP for an API.
- Data in CouchDB is stored in semi-structured documents that are flexible with individual implicit structures, but it is a simple document model for data storage and sharing.
- Document updates (add, edit, delete) follow Atomicity, i.e., they will be saved completely or not saved at all. The database will not have any partially saved or edited documents.

RapidMiner

- RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics.

- RapidMiner offers dozens of different operators or ways to connect to data. The data can be stored in a flat file such as a comma-separated values (CSV) file or spreadsheet, in a database such as a Microsoft SQLServer table, or it can be stored in other proprietary formats such as SAS or Stata or SPSS, etc.

- RapidMiner is a free of charge, open source software tool for data and text mining.

- Our end-to-end data science platform offers all of the data preparation and machine learning capabilities needed to drive real impact across your organization.

- Radoop: Analyzing Big Data with RapidMiner and Hadoop.

R Programming

- R is a potent language used broadly for statistical computing and data analysis.
- It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- R used in data science focuses on the language’s statistical and graphical uses.
- R for data science is used in industries such as telecommunications, banking, and media.
- R is a suite of operators for calculations on arrays, in particular matrices.