Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges

by Marjani et.al

Presented by:
Ameya Samak,
Bhavishyya Muppalaneni,
Neeraj Bandi

INTRODUCTION

- The development of big data and IoT is rapidly accelerating
- These technologies have an impact on all aspects of business and society.
- The capacity to use and analyze enormous amounts of IoT data opens up a world of possibilities.
- Big data can be divided into three categories: quantity, diversity, and speed.

Benefits of IoT Big Data Analytics

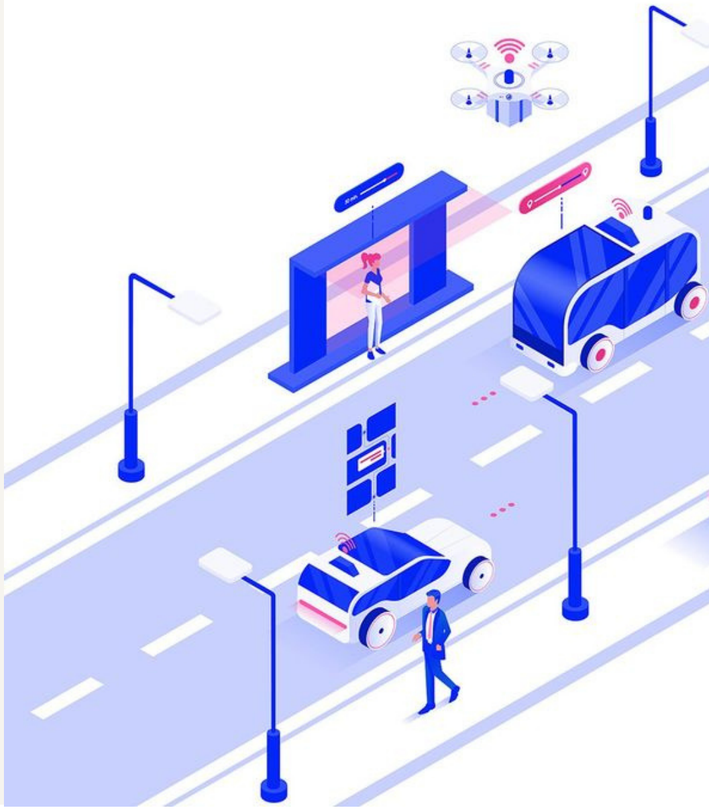
- The management of enormous volumes of information that may have an impact on organizations and the analysis of vast amounts of data can help both enterprises and individuals.

- IoT big data analytics attempts to help businesses better comprehend their data and make effective, informed decisions.

- Large amounts of unstructured data can be analyzed by data miners and scientists using conventional technologies.

IoT and Big Data Integration

- Solutions for integrating IoT and big data can help with challenges with storage, processing, data analytics, and visualization tools.
- It can help improve cooperation and communication amongst different things in a smart city.
- The integration can help application fields including smart ecological environments, smart traffic, smart grids, intelligent buildings, and logistic intelligent management.



The Advantages of IoT in Smart Environments

- IoT provides a seamless platform for devices to communicate and share information
- Adaptation of wireless technologies
- All surrounding electronic equipment can be connected to an IoT network
- Data Collection and Sharing
- Embedded Communication Devices
- IoT can be recognized in three paradigms: Internet-oriented, sensors, and knowledge

Big Data

- McKinsey Global Institute defines "big data" as the size of data sets that are better handled by tools for capturing, storing, processing, and analyzing such data.
- "The Digital Universe" study characterizes big data into three aspects: data sources, data analytics, and the presentation of the results of the analytics.
- The 3V's model, which highlights volume, variety, and velocity, is the most common description of the term "big data".
- Some researchers have introduced additional characteristics for big data, such as veracity, value, variability, and complexity.

Big Data Analytics

- Searching, mining, and analyzing large datasets to improve company performance
- Examining various data types to reveal patterns, correlations, trends, and other useful business information
- Main objective: better understanding of data for efficient and informed decisions
- Technologies and tools needed to transform structured, unstructured, and semi-structured data into understandable formats
- Algorithms must discover patterns, trends, and correlations over different time horizons
- Big data analytics is challenging due to data complexity and algorithm scalability
- Exploratory Data Analysis Environment is an example of a big data visual analytics system used for analyzing complex data sets

EXISTING ANALYTICS SYSTEMS

Real Time Analytics:

- Real-time analytics involves analyzing data as it is collected, which is useful for constantly changing data from sensors
- Two types of computer systems have been proposed for quick real-time analytics:

    1. Traditional databases that spread the work across multiple computers.
    2. Special types of fast computer memory.

- Examples of real-time analytics systems include Greenplum and Hana.

EXISTING ANALYTICS SYSTEMS

Off-line Analytics

- Offline analytics is a way of analyzing data when immediate results are not required.
- Useful when dealing with large volumes of data that require extensive processing time.
- Hadoop is a popular system used for offline analytics, saving costs on data processing.
- Improves the speed of data collection and analysis.
- Examples of systems that use offline analytics are SCRIBE, Kafka, Time-Tunnel, and Chukwa.

EXISTING ANALYTICS SYSTEMS

Memory-level Analytics

- Type of analysis for data that can fit in computer memory
- Enables real-time analysis using internal database technologies
- Ideal for when data size is not too big
- Example: MongoDB

EXISTING ANALYTICS SYSTEMS

BI analytics

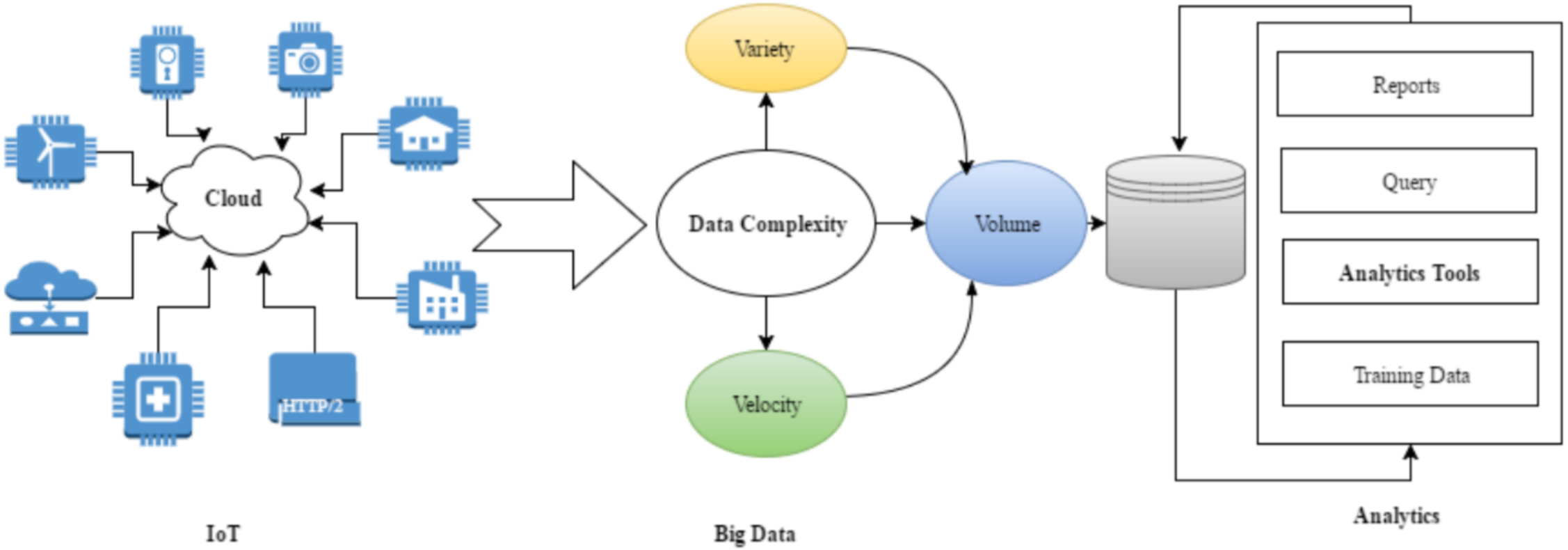
- BI analytics is used for data that is too big to fit into computer memory
- Data is imported into the BI analysis environment
- Supports data that is even bigger than a terabyte
- Helps to find new business opportunities
- Gives a competitive advantage and helps companies stay successful in the long term
- Easy to understand the meaning of the data with BI analytics

EXISTING ANALYTICS SYSTEMS

Massive analytics

- Used for handling extremely large datasets that can't be managed by traditional databases or BI analysis products
- Hadoop is the system used to store and analyze such data
- Enables businesses to get valuable insights from their data, make better decisions and reduce risks associated with decision-making
- Efficient and effective in providing services
- Crucial in helping businesses stay competitive and successful in the long term.

RELATION BETWEEN IOT AND BIG DATA ANALYTICS



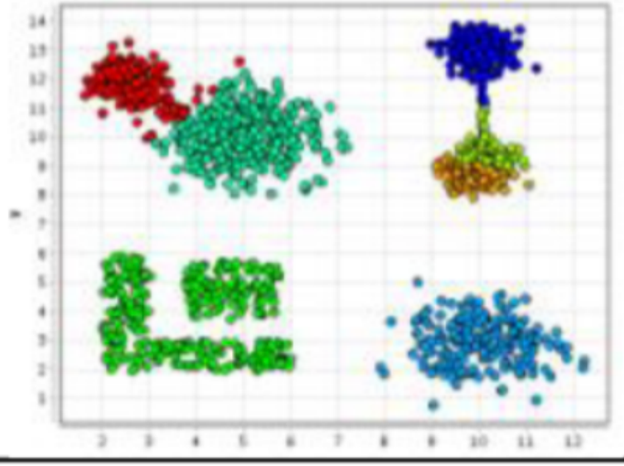
BIG DATA ANALYTICS METHODS

- Big data analytics aim to immediately extract knowledgeable information that helps in making predictions, identifying recent trends, finding hidden information, and ultimately, making decisions.
- Efficient and cost-effective big data analytics requires processing speeds equal to or faster than traditional data analytics for managing high-volume, high-velocity, and high-variety data.
- The notion that diverse and extensive data sets always lead to valuable insights in big data is not necessarily true, as larger datasets may contain more ambiguities and abnormalities.
- This outlines four main categories of big data analytics methods: classification, clustering, association rule mining, and prediction, and provides an overview of their various algorithms and subcategories, such as Bayesian network and partitioning, to extract and analyze information.

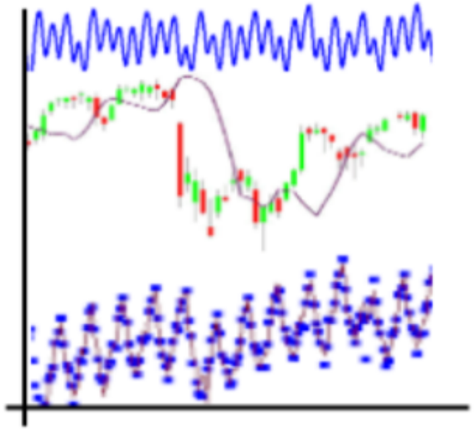
Big Data Analytics

Classification

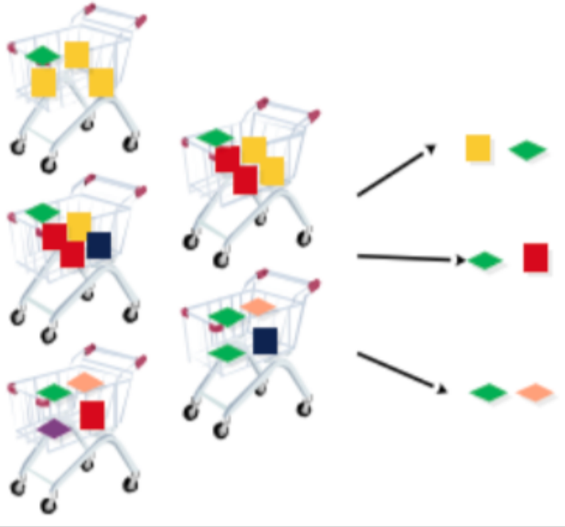
Clustering



Prediction

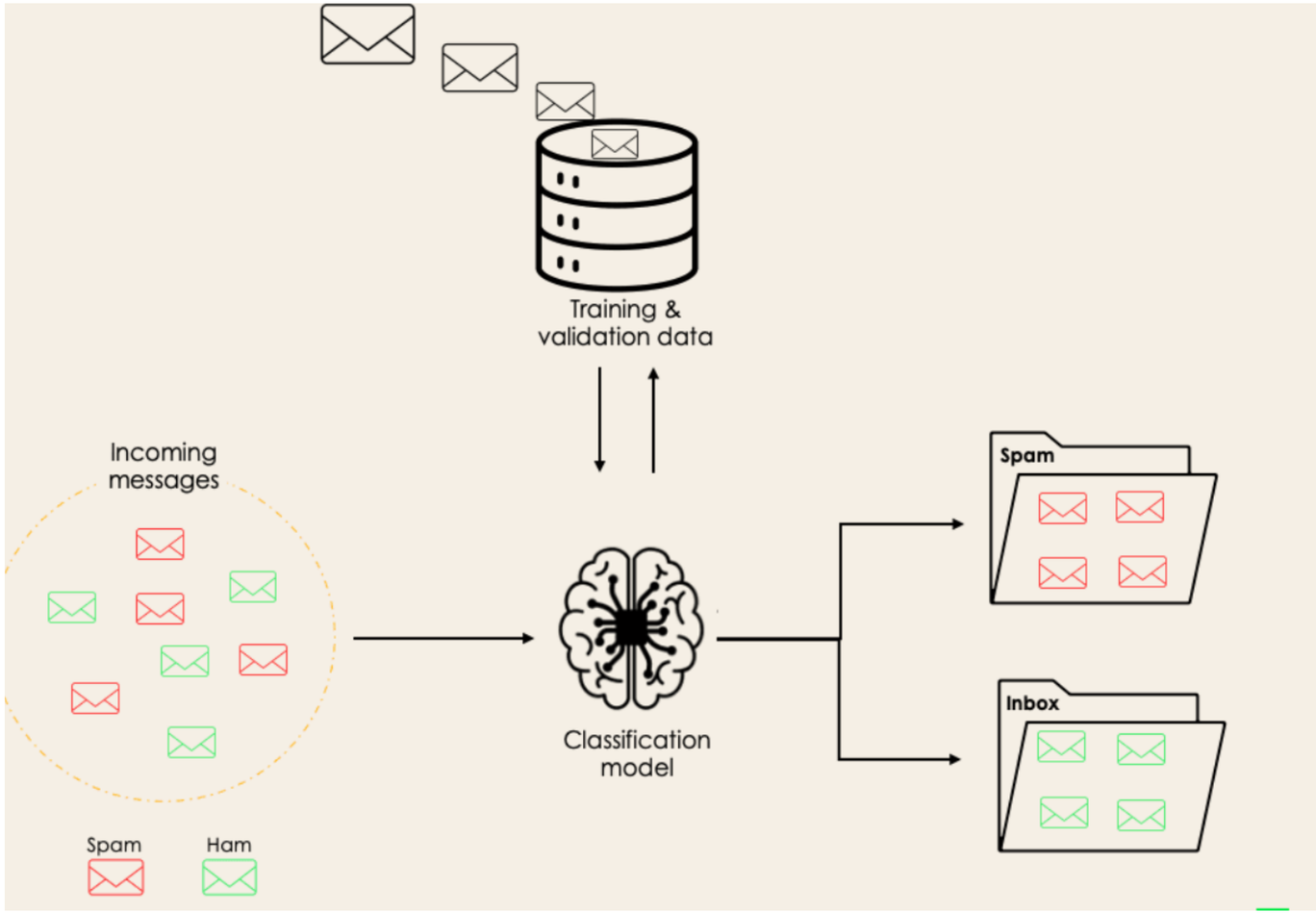


Association Rule



CLASSIFICATION

- Classification is a supervised learning approach that uses prior knowledge as training data to classify data objects into groups
- Bayesian networks are useful for analyzing complex data structures in big data and describes as directed acyclic graphs with nodes representing random variables and edges.
- SVM is a classification technique that analyses data patterns and organizes them using statistical learning theory. Text classification, pattern matching, health diagnostics, and commerce are some of the applications of SVM classification in big data analytics.
- KNN is a valuable tool for discovering hidden patterns in big data and retrieving objects similar to a predefined category, with extensions for anomaly detection, high-dimensional data, and scientific experiments, making it one of the popular data mining techniques for big data analytics.

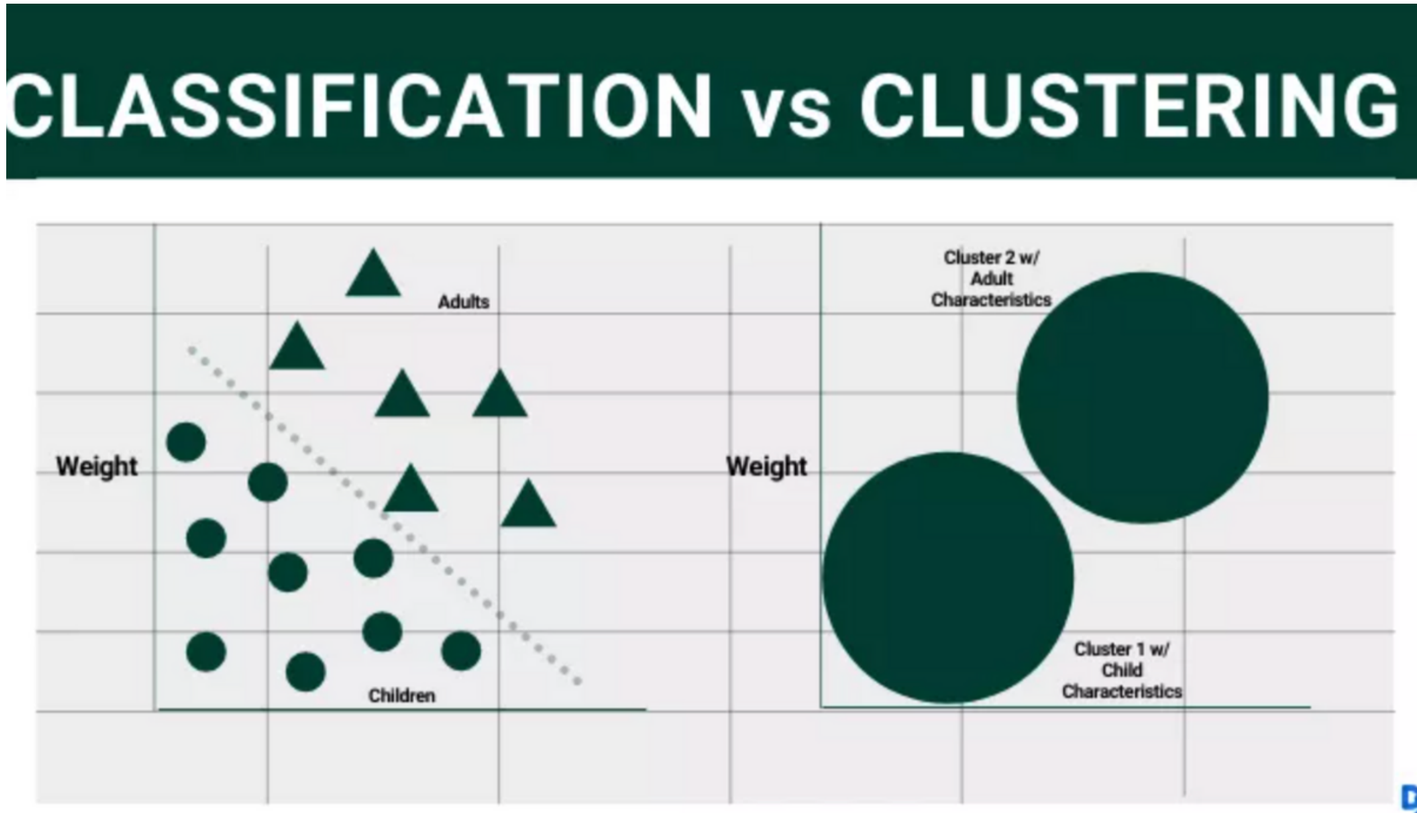


CLUSTERING

- Clustering is an unsupervised learning approach in big data analytics, which groups objects based on their meaningful features.
- Grouping objects into clusters simplifies data manipulation.
- The two well-known clustering methods are hierarchical clustering and partitioning.

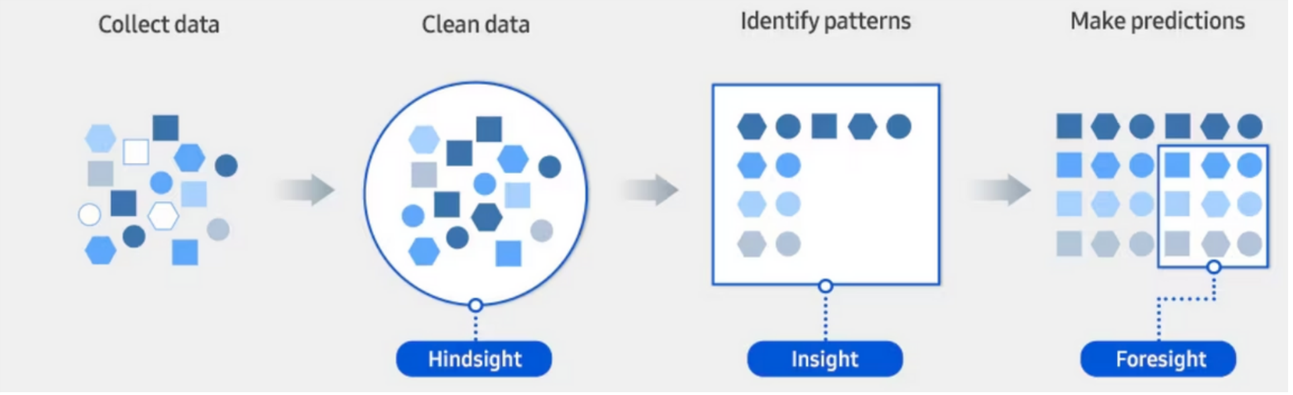
.

- Hierarchical clustering creates agglomerative clusters by combining small clusters into a hierarchical tree, while divisive clusters are formed by dividing a single cluster of all data objects into smaller clusters.
- Partition method partitions data into a predetermined number of non-overlapping clusters. Popular partitioning algorithms include k-means, k-means, and fuzzy c-means.



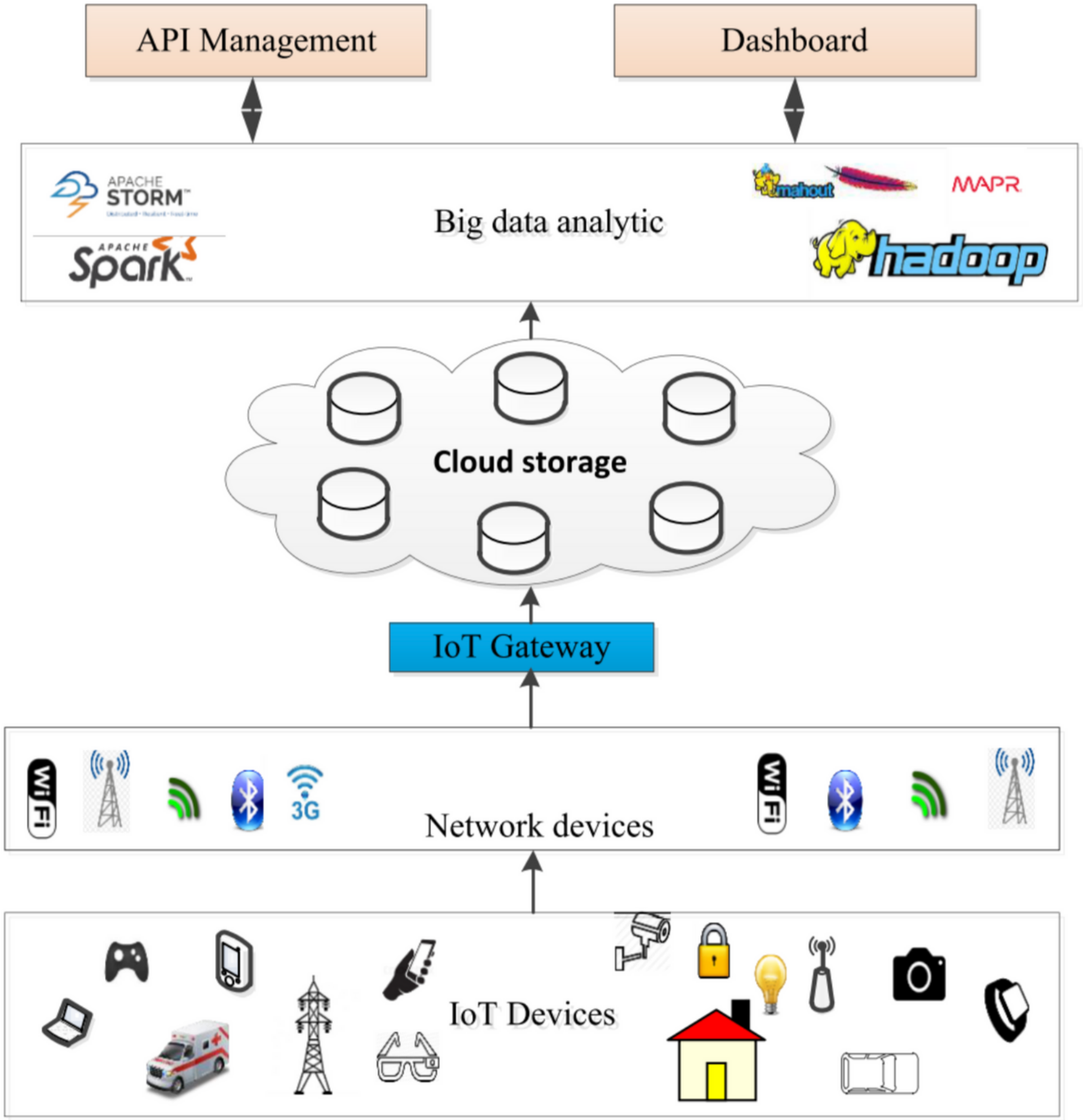
PREDICTIVE ANALYTICS

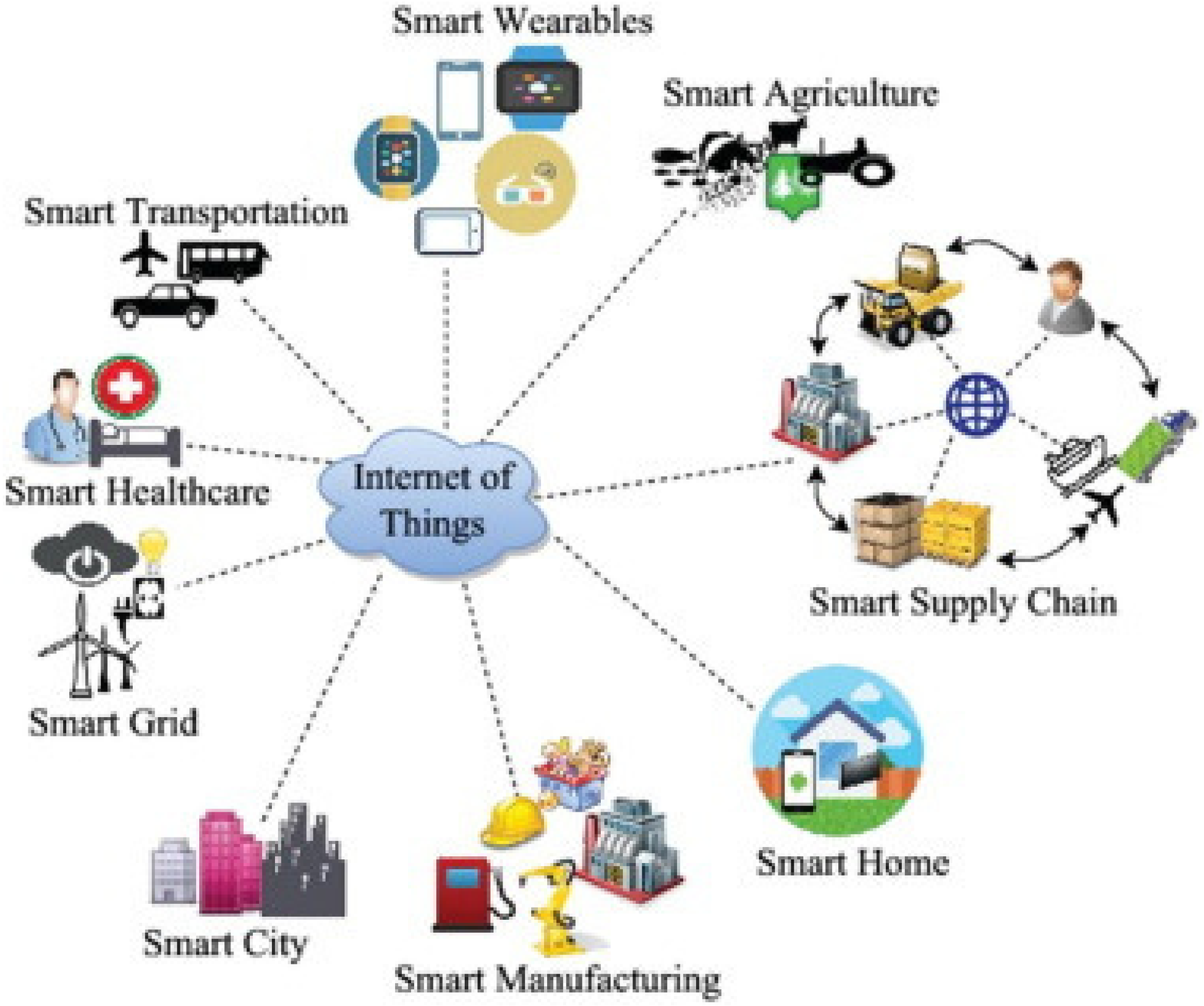
- Predictive analytics uses historical data to make predictions about future events or trends. It involves statistical and machine learning techniques to identify patterns and relationships in the data.
- SVM and fuzzy logic algorithms are commonly used to obtain regression curves for predictions, such as for natural disasters or customer buying behavior.
- Time series analysis is a popular technique for data representation in big data analytics. It reduces high dimensionality associated with big data and offers improved decision making.
- ARMA, bitmaps, and wavelet functions are popular time series representation techniques.
- Predictive analytics is used in various industries, including healthcare, finance, marketing, and manufacturing.
- Predictive analytics provides insights into customer behavior, market trends, and business performance, helping businesses make better decisions.



Method	Applications										
	Disaster management	Healthcare	Medical Imaging	Human Genetics	Market Analysis	Industry	Speech Recognition	Bioinformatics	NLP	Social Network Analysis	e-governance
Classification [46]	-	-	✓	-	-	✓	✓	-	✓	-	✓
Clustering [57]	-	✓	✓	✓	✓	✓	-	✓	-	✓	✓
Association rule[58, 65]	-	✓	-	-	✓	✓	-	✓	-	-	✓
Prediction [61]	✓	-	-	-	✓	-	-	-	-	✓	-
Time Series [62] [63] [64]	✓	-	✓	-	-	-	✓	-	-	✓	✓

IoT ARCHITECTURE FOR BIG DATA ANALYTICS




Smart Wearables
Smart Agriculture
Smart Transportation
Smart Healthcare
Internet of Things
Smart Supply Chain
Smart Grid
Smart Home
Smart City
Smart Manufacturing

USE CASES

Smart Metering

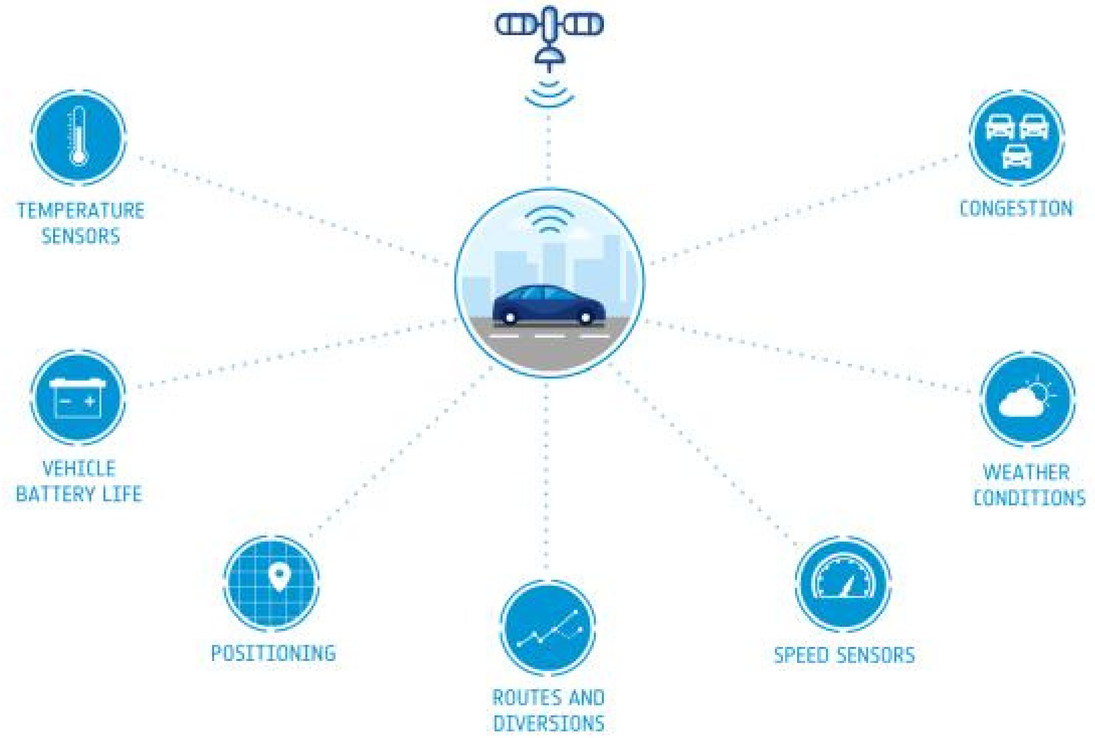
- Smart metering creates a large amount of data from different sources, such as smart grids, tank levels, and water flows. Collecting and analyzing smart meter data in an IoT environment assists decision makers in predicting electricity consumption.
- The analytics of a smart meter can be used to forecast demands, prevent problems, and satisfy strategic goals through specific pricing plans.
- Utility companies must be capable of high-volume data management and advanced analytics designed to convert data into actionable insights.

Smart Transportation

- A smart transportation system is an IoT-based use case that supports the smart city concept. Introducing IoT into vehicular technologies enables traffic congestion management to exhibit significantly better performance than the existing infrastructure.

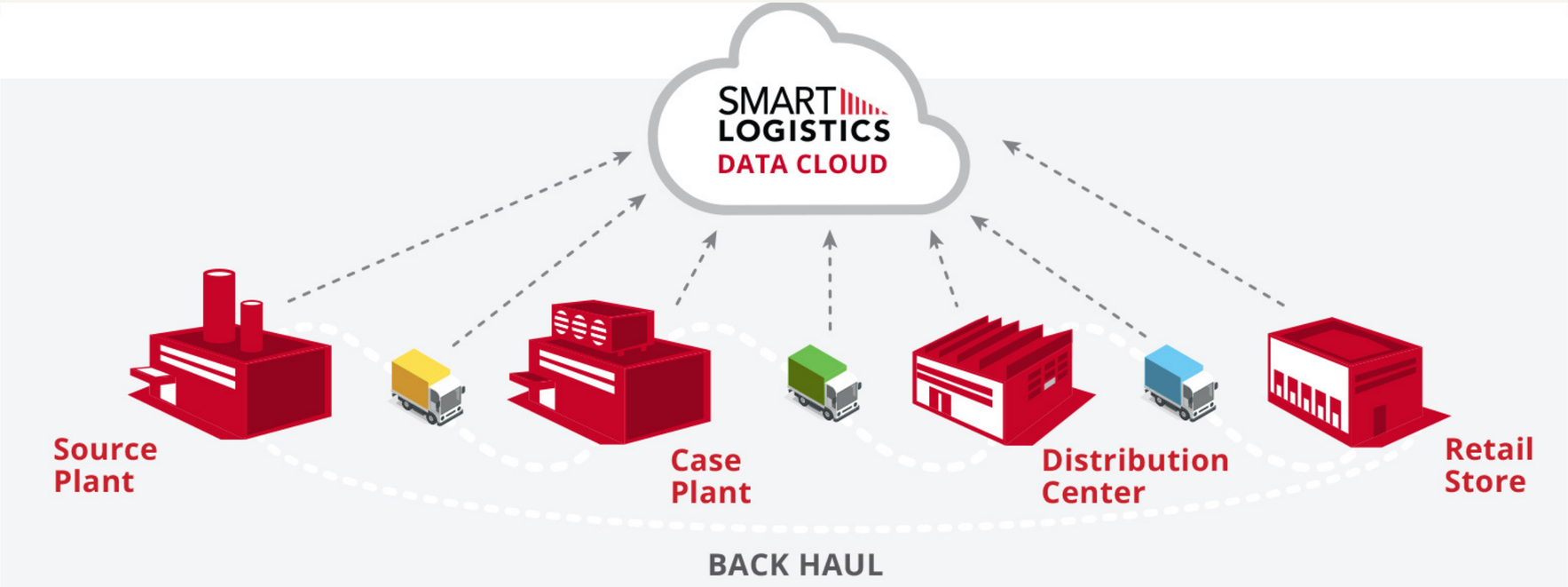
- Satellite navigation systems and sensors can be applied in real time to optimize the routing of vehicles. IoT technologies can improve existing traffic systems in which vehicles can communicate with one another in a systematic manner without human intervention.

- Sensors incorporated into vehicles can provide real-time information to measure engine health, determine whether equipment requires maintenance, and predict errors.



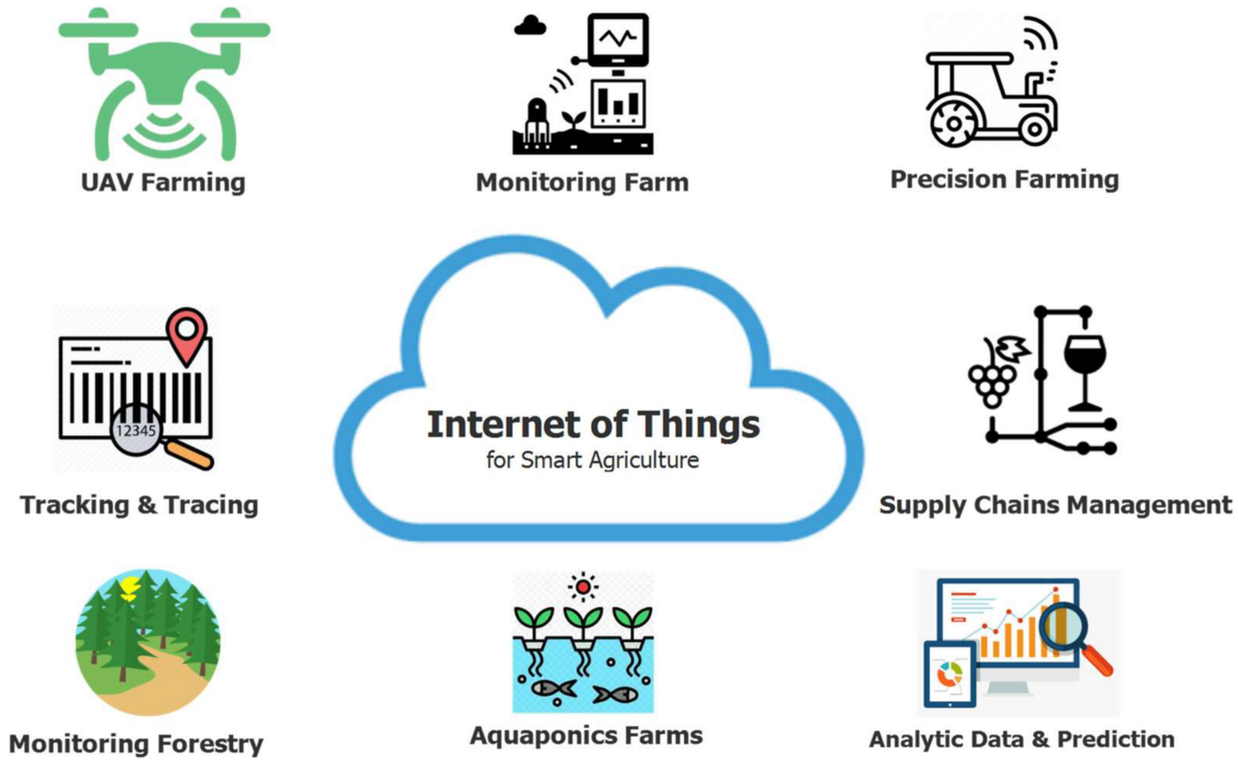
Smart Supply Chain

- More than 1 million elevators worldwide have remote accessibility thanks to embedded sensor technology that can broadcast and receive information. Future supply chains using IoT technology will benefit from in-transit visibility, a crucial use case.
- Position, identification, and other tracking information provided by RFIDs and cloud-based GPS are crucial for in-transit visibility.
- Factory equipment that is connected to the Internet of Things will be able to exchange data within predefined limits and change equipment settings or process workflow to improve performance.
- A supply chain may make judgments and manage the environment with the aid of massive IoT data analytics.



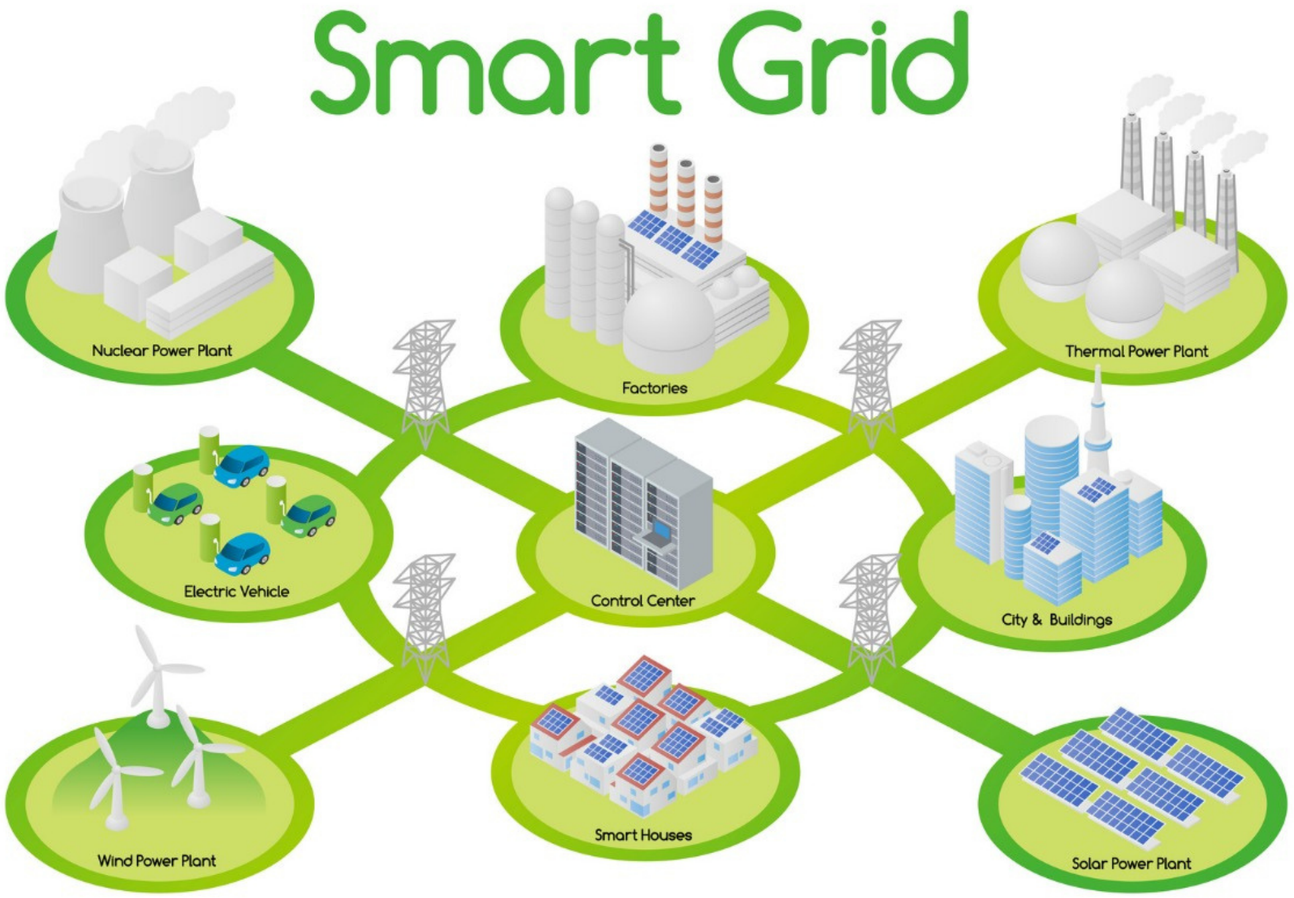
Smart Agriculture

- Smart agriculture is a beneficial use case in big IoT data analytics. Sensors are installed in fields to obtain data on moisture level of soil, trunk diameter of plants, microclimate condition, and humidity level, as well as to forecast weather.
- Sensors transmit obtained data using network and communication devices.
- The analytics layer processes the data obtained from the sensor network to issue commands.
- Automatic climate control according to harvesting requirements, timely and controlled irrigation, and humidity control for fungus prevention are some benefits of smart agriculture.
- Smart agriculture can also enhance food safety by enabling real-time monitoring of crops for contamination or disease



Smart Grid

- The smart grid is a new generation of power grid that uses two-way communication technologies and computing capabilities to improve reliability, safety, and efficiency with real-time control and monitoring.
- Integrating renewable and distributed energy is a major challenge in a power system that requires a smart grid to manage the volatile behavior of distributed energy resources.
- Most energy systems have to follow government laws and regulations, as well as consider business analysis and potential legal constraints.
- Grid sensors and devices continuously and rapidly generate data related to control loops and protection and require real-time processing and analytics along with machine-to-machine communication.
- Benefits of smart grids include reduced energy costs, increased renewable energy integration, improved grid reliability, and increased grid resiliency in the face of disruptions such as natural disasters or cyber attacks.



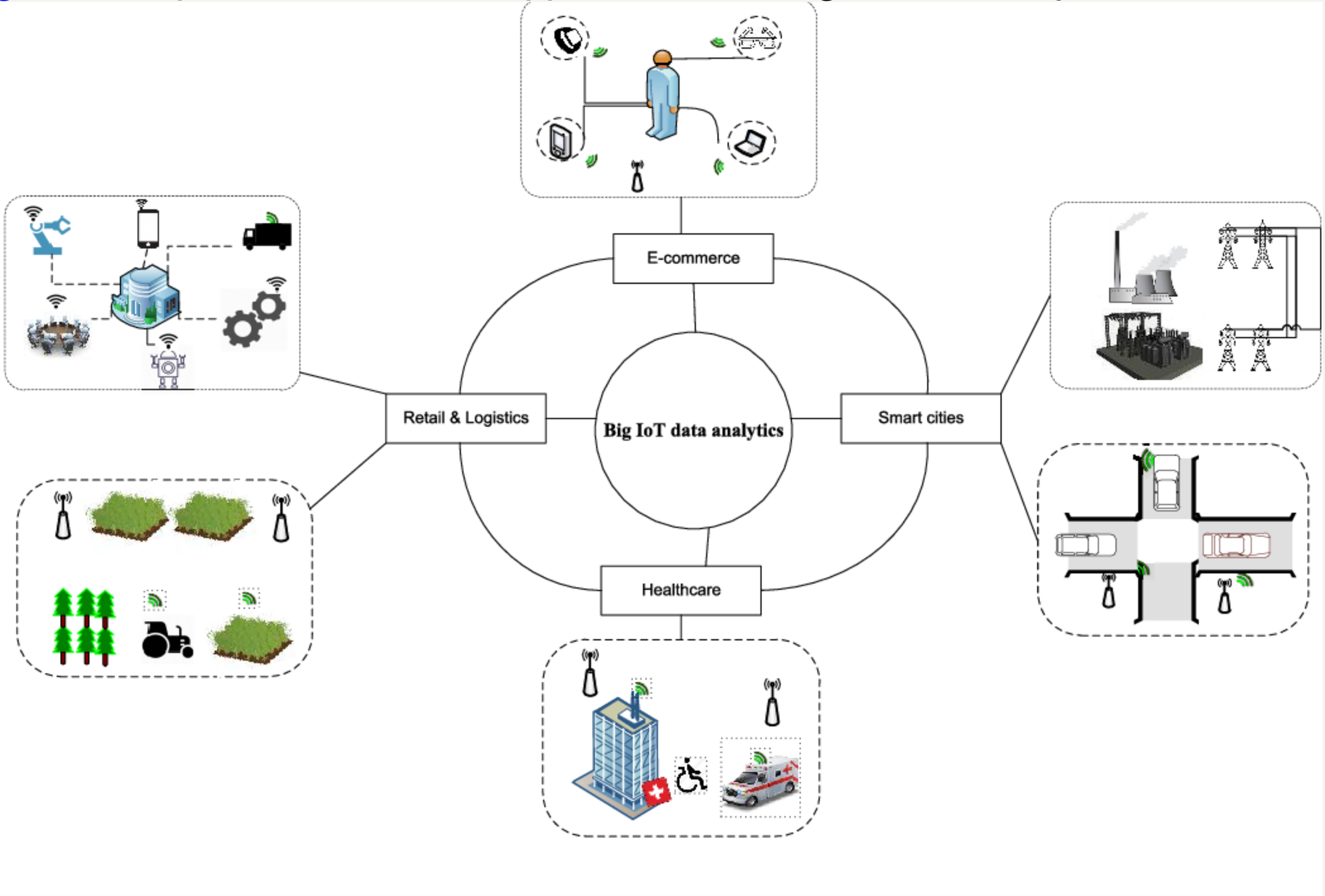
Opportunities

IoT is currently considered one of the most profound transitions in technology. It provides numerous opportunities for big data analytics.

For Instance,
In the healthcare industry, IoT data analytics can be used to monitor vital signs and detect early signs of health problems

In the agriculture industry, it is used to monitor crops, soil conditions, and weather patterns, helping farmers to optimize their yields and reduce waste.

Figure. Example of use cases and opportunities for big loT data analytics architecture.



E-COMMERCE

Big IoT data analytics provides organizations, well-designed tools to process real-time big data.

Big IoT data exhibit heterogeneity, increasing volume, and real-time data processing features.

The main success areas in e-commerce:

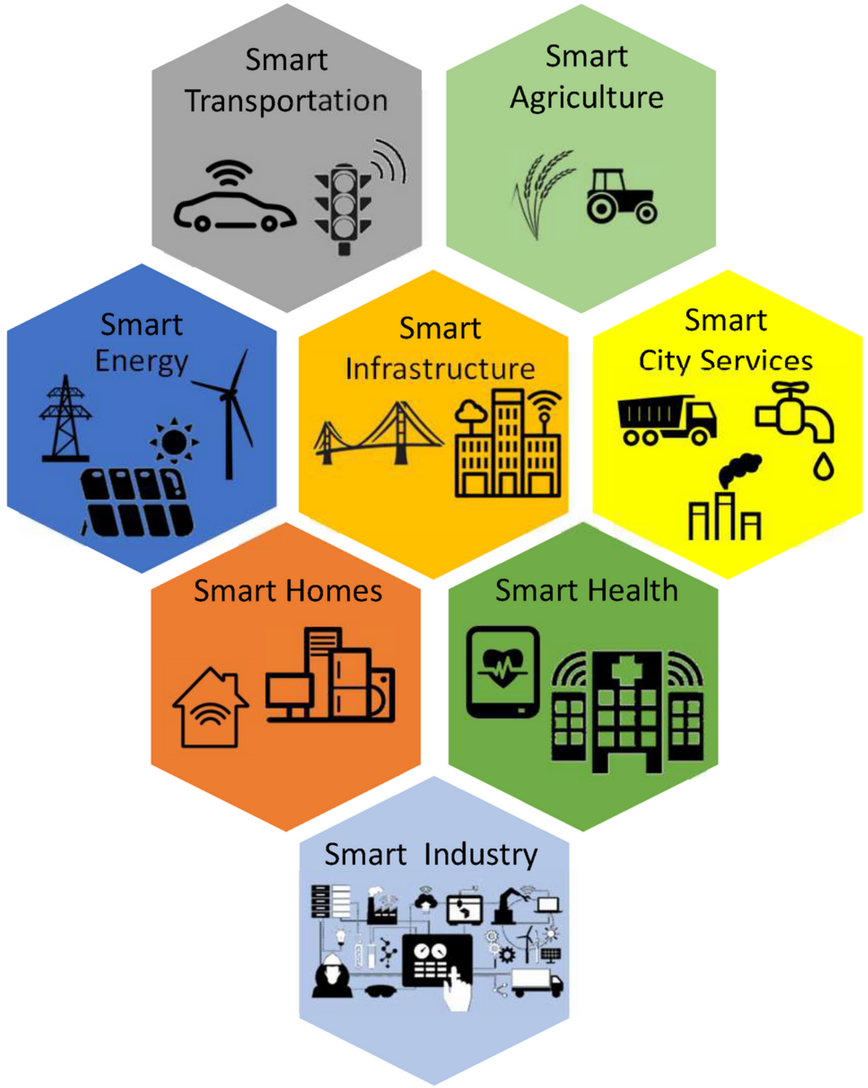
a. Customer Segmentation and Personalization
b. Product Optimization
c. Revenue Growth
d. Inventory management
e. Risk management

SMART CITIES

Smart cities generate vast amounts of data that can be leveraged to improve efficiency and optimize services.

Cloud computing technology has dramatically reduced the cost of storing and processing data, making it more feasible to analyze large volumes of data.

Hadoop with YARN resource manager has offered recent advancement in big data technology to support and handle numerous workloads, real-time processing, and streaming data ingestion.



RETAIL AND LOGISTICS

In logistics, RFID keeps track of containers, pallets, and crates, providing real-time tracking and monitoring capabilities.

Applying data analytics to logistic data sets can improve the shipment experience of customers, reduce delays, and increase efficiency.

Retail companies can leverage customer data to predict trends and demands, optimize pricing plans, and plan seasonal promotions efficiently to maximize profit.

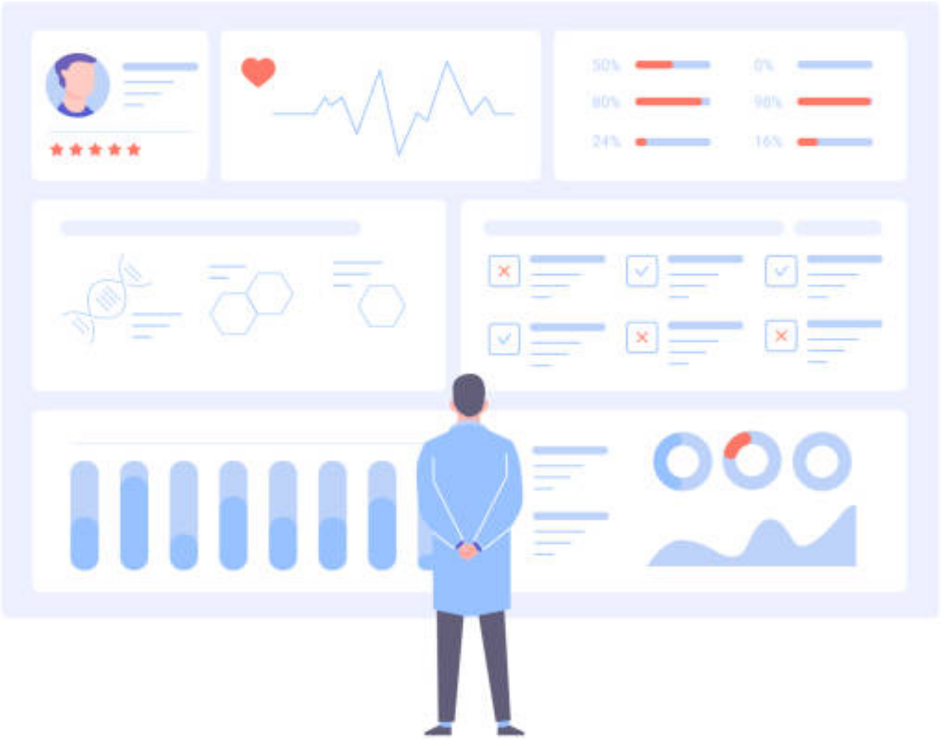
HEALTHCARE

Smart health monitoring devices generate massive amounts of patient data.
Examples: Fetal monitors, temperature and blood glucose monitors.

Data analytics can efficiently assess patient physical conditions, diagnose diseases in early stages, improve clinical quality of care, and ensure patient safety.

Physician profile can be improved by analyzing patient data, leading to higher customer satisfaction, acquisition, and retention.

Integration of data analytics with smart health monitoring devices has the potential to significantly improve healthcare outcomes and save lives.



Open Challenges and Future Directions

PRIVACY

Big data analytics tools pose privacy risks as personal information can be inferred or restored even from anonymous data.

Heterogeneity of devices and generated data types increases security risks and requires non-repudiable identification system and meta-repository for auditing.

Existing security solutions not applicable to dynamic data. Security problems can arise from timely updates, incident management, interoperability, and protocol convergence.

Legislative and regulatory issues need to be considered while signing SLAs.

Guidelines to overcome these challenges include open ecosystem with standard APIs, device protection while communicating, and hardcoded best security practices.

DATA MINING

The high-volume, high-velocity, and high-variety qualities of big IoT data make exploration, integration, and extraction processes difficult.

Accurate information extraction from diverse and complex data requires analyzing data properties and finding associations among different data points.

Researchers have proposed parallel and sequential programming models and different algorithms to minimize query response time and study dynamic data mining methods and stream data.

For example, parallel k-means algorithm and parallel association rule mining.

There is a need to devise algorithms compatible with the latest parallel architectures, and synchronization issues may occur in parallel computing.

VISUALIZATION

Big data analytics and visualization are essential for obtaining insights from IoT data, but it is challenging due to large size and high dimensionality of data.

Response time is a desirable factor in big IoT data analytics, and cloud computing architectures supported with rich GUI facilities can help to achieve better insights.

Fine-grained dimensions are necessary for identifying observable correlations, patterns, and outliers.

Several important issues are addressed in big data visualization, such as visual noise, information loss, large image observation, frequently changing image, and high performance requirements.

Visualization should adhere to guidelines such as providing special attention to metadata, being interactive and requiring maximum user involvement, and being built based on the dynamic nature of the generated data.

INTEGRATION

Data integration involves collecting, storing, and providing data with a unified view.

Integrating diverse data types is a complex task in merging different systems or applications.

Challenges of data integration include overlapping data, increasing performance and scalability, and enabling real-time data access. Reshaping semi-structured and unstructured data structures before integrating and analyzing also becomes difficult.

Technologies such as text mining, machine learning, natural language processing, and information extraction can be used to extract entities and relationships from textual data.

However, developing new technologies is necessary to extract information from non-text formats of unstructured data, such as images and videos.

Conclusion

The growth of IoT has led to a high frequency of data production, making big data analytics crucial. Existing big IoT data analytics solutions are in their early stages of development and the real-time analytics solutions will be necessary for quick insights in the future.

To fully realize the potential of big IoT data analytics, it is necessary to develop new technologies, algorithms, and architectures that can effectively handle the complexities of IoT data.

QUESTIONS