BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Introduction to Big Data Analytics

Course Code: 23ST02VAC

Unit-I

BIG DATA - INTRODUCTION

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. S. Soundarya
Guest Faculty
Department of Statistics

UNIT – I

INTRODUCTION TO BIG DATA ANALYTICS

Data

Data is one of the prime factors of any business purpose. Business Enterprises are data-driven and without data, no one can have a competitive advantage. It has different definitions wherein the huge amount of data can be considered as Big Data. It is the most widely used technology these days in almost every business vertical.

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

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 also a data but with huge size.

Example of Big Data

There are some examples of Big Data Analytics in different areas such as retail, IT infrastructure, and social media.

Retail: As mentioned earlier, Big Data presents many opportunities to improve sales and marketing analytics.

An example of this is the U.S. retailer Target. After analyzing consumer purchasing behavior, Target's statisticians determined that the retailer made a great deal of money from three main life-event situations.

Definition of Big Data

Big data is high-velocity and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making. Big data refers to datasets whose size is typically beyond the storage capacity of and also complex for traditional database software tools Big data is anything beyond the human & technical infrastructure needed to support storage, processing and analysis.

History of Big Data

Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database. Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyzes big data sets) was developed that same year. NoSQL also began to gain popularity during this time. The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it. With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data. While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

Benefits of Big Data and Data Analytics

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

Types of Big Data

Following are the types of Big Data:

• Structured

• Unstructured

• Semi-structured

Structured data

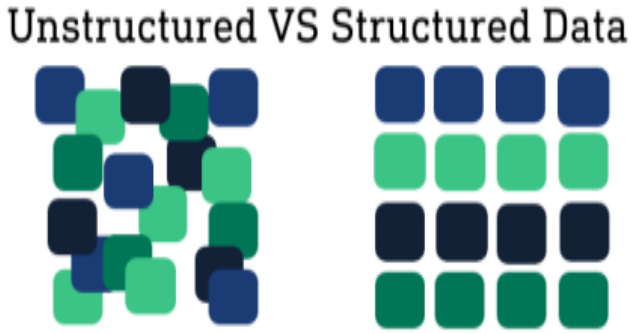
Structured data includes quantitative data that is stored in an organized manner. It consists of numerical and text data. It is easy to analyze and process structured data. It is generally stored in a relational database and can be queried using Structured Query Language (SQL).

Example of Structured data

Emp.ID	Emp.Name	Gender	Department	Salary(INR)
2383	ABC	Male	Finance	650,000
4623	XYZ	Male	Admin	5,000,000

Unstructured data

Unstructured data includes qualitative data that lacks any predefined structure and can come in a variety of formats (images, mp3 files, wav files, etc.). Unstructured data is said to lack “structure”. It is stored in a non-relational database and can be queried using NoSQL. There can be semi-structured data as well, which lies somewhat in between structured and unstructured data.



Semi-structured Data

Semi-structured data falls somewhere between structured data and unstructured data. It mostly translates to unstructured data that has metadata attached to it. Semi-structured data can be inherited such as location, time, email address or device ID stamp. It can even be a semantic tag attached to the data later. Consider the example of an email. The time an email was sent, the email addresses of the sender and the recipient, the IP address of the device that the email was sent from, and other relevant information are linked to the content of the email. While the actual content itself is not structured, these components enable the data to be grouped in a structured manner.

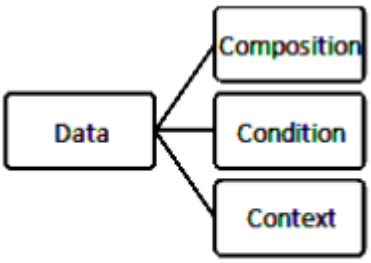
Characteristics of Data

Data has three key characteristics

1. Composition: The composition of data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it is static or real-time streaming.

2. Condition: The condition of data deals with the state of data, that is, "Can one use this data as is for analysis?" or "Does it require cleansing for further enhancement and enrichment?"

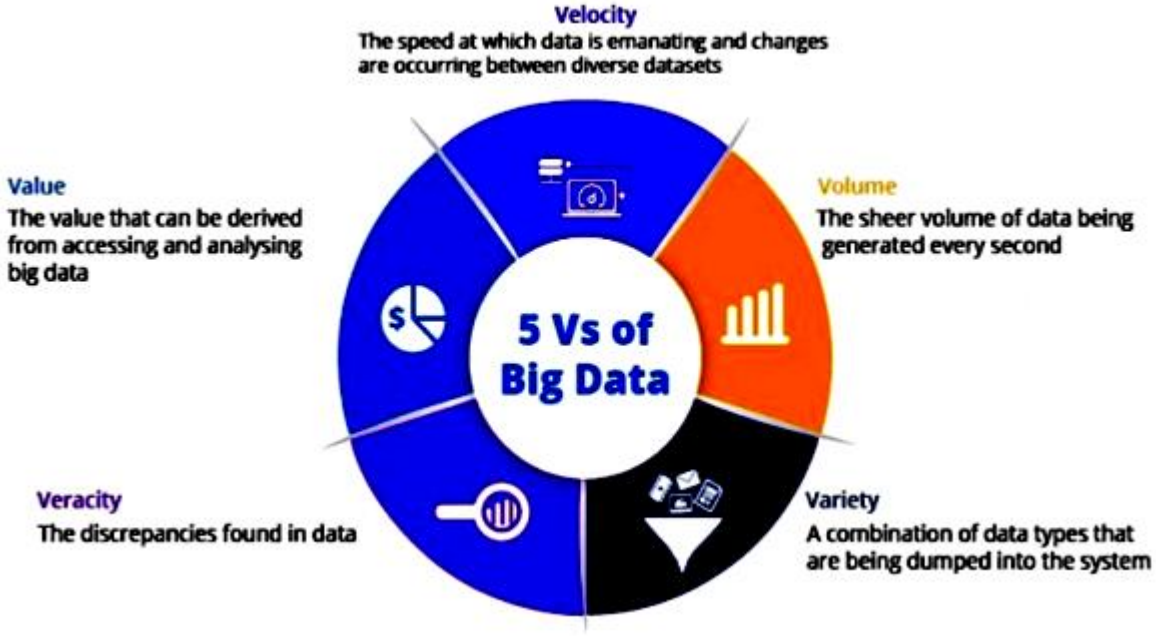
3. Context: The context of data deals with "Where has this data been generated?" "Why was this data generated?" How sensitive is this data?" "What are the events associated with this data?" and so on. Small data (data as it existed prior to the big data revolution) is about certainty. It is about known data sources; it is about no major changes to the composition or context of data.



Characteristics of data

The V's of Big Data (or) Characteristics of Big Data

Big Data has the following distinct characteristics:



1. Volume: The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with Big Data solutions.

2. Variety: Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis

applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

3. Velocity: The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

4. Value: Benefit generated by using the information contained in the data to improve to outcomes of actions. e.g. profit, medical or social benefits, customer, employee, or personal satisfaction.

5. Veracity: Since packages get lost during execution, we need to start again from the stage of mining raw data to convert it into valuable data. And this process goes on. There will also be uncertainties and inconsistencies in the data that can be overcome by veracity. Veracity means the trustworthiness and quality of data. The veracity of data must be maintained. For example, think about Facebook posts, hashtags, abbreviations, images, videos, etc., which make the posts unreliable and hamper the quality of their content. Collecting loads and loads of data is of no use if the quality and trustworthiness of the data are not up to the mark.

Important Big Data

The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyze it to find answers which will enable:

1. Cost Savings: Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.

2. Time Reductions: The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learning.

3. Understand the market conditions: By analyzing big data you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.

4. Control online reputation: Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.

5. Using Big Data Analytics to Boost Customer Acquisition and Retention : The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business cannot afford to disregard the high competition it faces. If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end, loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behavior is important to trigger loyalty.

6. Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights : Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company's product line and of course ensuring that the marketing campaigns are powerful.

7. Big Data Analytics As a Driver of Innovations and Product Development : Another huge advantage of big data is the ability to help companies innovate and redevelop their products.

Advantages of Big Data Processing

Ability to process Big Data in DBMS brings in multiple benefits, such as-

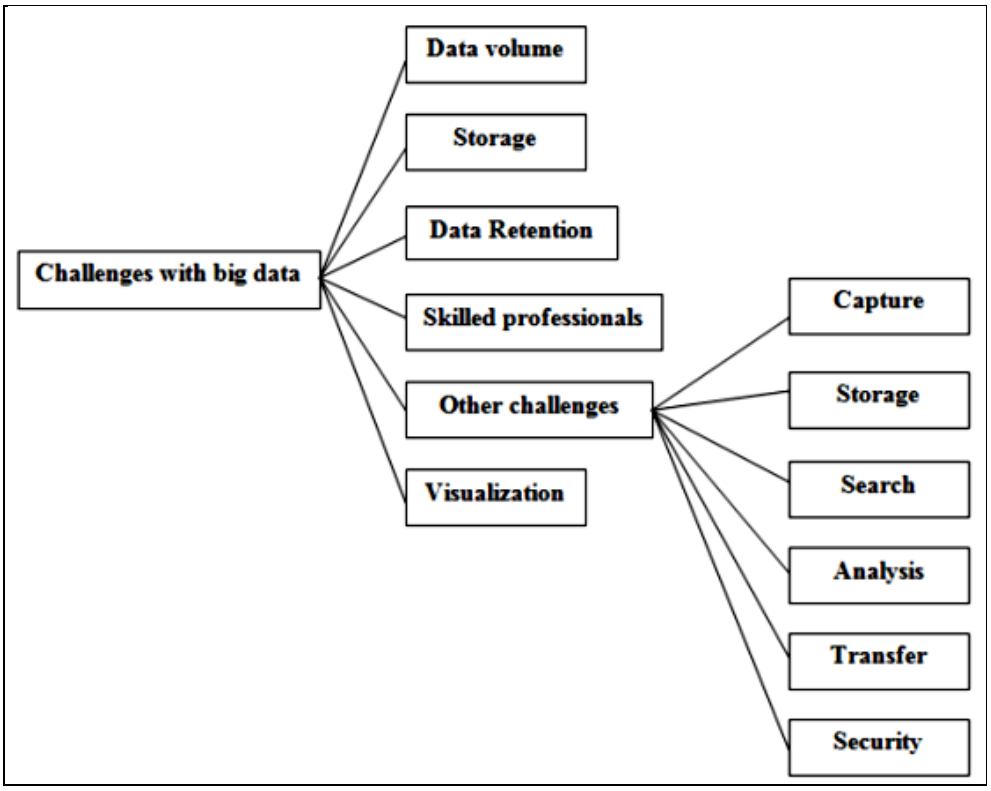
- Businesses can utilize outside intelligence while taking decisions Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.
- Improved customer service - Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.
- Early identification of risk to the product/services, if any
- Better operational efficiency - Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

Evolution of Big Data

1970s and before was the era of mainframes. The data was essentially primitive and structured. Relational databases evolved in 1980s and 1990s. The era was of data intensive applications. The World Wide Web (WWW) and the Internet of Things (IOT) have led to an onslaught of structured, unstructured, and multimedia data.

	Data Generation and Storage	**Data Utilization**	**Data Driven**
Complex and Unstructured	-	-	Structured data, Unstructured data, Multimedia data
Complex and Relational	-	Relational databases: Data-Intensive applications	-
Primitive And Structured	Mainframes: Basic data storage.	-	-
	1970 and before	**Relational (1980 and 1990s)**	**2000 and beyond**

Challenges with big data



Data volume: Data today is growing at an exponential rate. This high tide of data will continue to rise continuously. The key questions are –

- "will all this data be useful for analysis?",
- "Do we work with all this data or subset of it?",
- "How will we separate the knowledge from the noise?" etc

Storage: Cloud computing is the answer to managing infrastructure for big data as far as cost- efficiency, elasticity and easy upgrading / downgrading is concerned. This further complicates the decision to host big data solutions outside the enterprise.

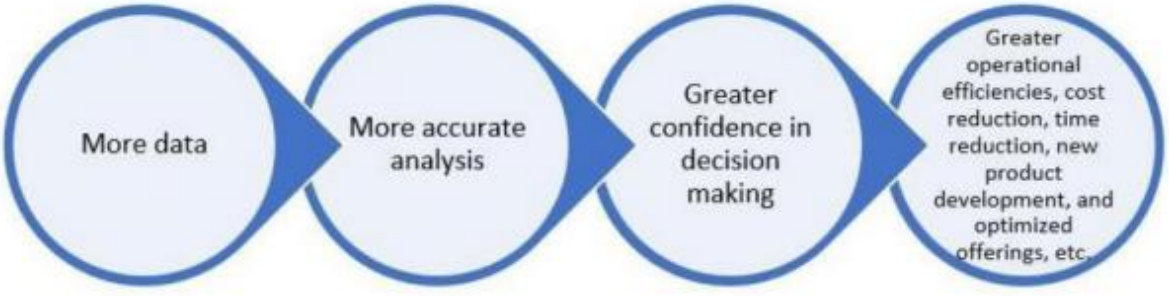
Data retention: How long should one retain this data? Some data may require for log-term decision, but some data may quickly become irrelevant and obsolete.

Skilled professionals: In order to develop, manage and run those applications that generate insights, organizations need professionals who possess a high-level proficiency in data sciences.

Other challenges: Other challenges of big data are with respect to capture, storage, search, analysis, transfer and security of big data.

Visualization: Big data refers to datasets whose size is typically beyond the storage capacity of traditional database software tools. There is no explicit definition of how big the data set should be for it to be considered big data. Data visualization (computer graphics) is becoming popular as a separate discipline. There are very few data visualization experts.

Big data : The more data we have for analysis, the greater will be the analytical accuracy and the greater would be the confidence in our decisions based on these analytical findings. The analytical accuracy will lead a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and originating new products, new services, and optimizing existing services (refer bellow figure).



Terminologies used in Big Data Environments

As-a-service infrastructure: Data-as-a-service, software-as-a-service, platform-as-a-service – all refer to the idea that rather than selling data, licences to use data, or platforms for running Big Data technology, it can be provided "as a service", rather than as a product. This reduces the upfront capital investment necessary for customers to begin putting their data, or platforms, to work for them, as the provider bears all of the costs of setting up and hosting the infrastructure. As a customer, as-a-service infrastructure can greatly reduce the initial cost and setup time of getting Big Data initiatives up and running.

Data science: Data science is the professional field that deals with turning data into value such as new insights or predictive models. It brings together expertise from fields including statistics, mathematics, computer science, communication as well as domain expertise such as business knowledge. Data scientist has recently been voted the No 1 job in the U.S., based on current demand and salary and career opportunities.

Data mining: Data mining is the process of discovering insights from data. In terms of Big Data, because it is so large, this is generally done by computational methods in an automated way using methods such as decision trees, clustering analysis and, most recently, machine learning. This can be thought of as using the brute mathematical power of computers to spot patterns in data which would not be visible to the human eye due to the complexity of the dataset.

Hadoop : Hadoop is a framework for Big Data computing which has been released into the public domain as open source software, and so can freely be used by anyone. It consists of a number of modules all tailored for a different vital step of the Big Data process – from file storage (Hadoop File System – HDFS) to database (HBase) to carrying out data operations (Hadoop MapReduce). It has become so popular due to its power and flexibility that it has developed its own industry of retailers (selling tailored versions), support service providers and consultants.

Predictive modelling : At its simplest, this is predicting what will happen next based on data about what has happened previously. In the Big Data age, because there is more data around than ever before, predictions are becoming more and more accurate. Predictive modelling is a core component of most Big Data initiatives, which are formulated to help us choose the course of action which will lead to the most desirable outcome. The speed of modern computers and the volume of data available means that predictions can be made based on a huge number of variables, allowing an ever-increasing number of variables to be assessed for the probability that it will lead to success.

Map Reduce : Map Reduce is a computing procedure for working with large datasets, which was devised due to difficulty of reading and analyzing really Big Data using conventional computing methodologies. As its name suggest, it consists of two procedures – mapping (sorting information into the format needed for analysis – i.e. sorting a list of people according to their age) and reducing.

NoSQL: NoSQL refers to a database format designed to hold more than data which is simply arranged into tables, rows, and columns, as is the case in a conventional relational database. This database format has proven very popular in Big Data applications because Big Data is often messy, unstructured and does not easily fit into traditional database frameworks.

Python: Python is a programming language which has become very popular in the Big Data space due to its ability to work very well with large, unstructured datasets (see Part II for the difference between structured and unstructured data). It is considered to be easier to learn for a data science beginner than other languages such as R (see also Part II) and more flexible.

R Programming: R is another programming language commonly used in Big Data, and can be thought of as more specialised than Python, being geared towards statistics. Its strength lies in its powerful handling of structured data. Like Python, it has an active community of users who are constantly expanding and adding to its capabilities by creating new libraries and extensions.

Recommendation Engine: A recommendation engine is basically an algorithm, or collection of algorithms, designed to match an entity (for example, a customer) with something they are looking for. Recommendation engines used by the likes of Netflix or Amazon heavily rely on Big Data technology to gain an overview of their customers and, using predictive modelling, match them with products to buy or content to consume. The economic incentives offered by recommendation engines has been a driving force behind a lot of commercial Big Data initiatives and developments over the last decade.

Real-time: Real-time means “as it happens” and in Big Data refers to a system or process which is able to give data-driven insights based on what is happening at the present moment. Recent years have seen a large push for the development of systems capable of processing and offering insights in real-time (or near-real-time), and advances in computing power as well as development of techniques such as machine learning have made it a reality in many applications today.

Reporting: The crucial “last step” of many Big Data initiative involves getting the right information to the people who need it to make decisions, at the right time. When this step is automated, analytics is applied to the insights themselves to ensure that they are communicated

in a way that they will be understood and easy to act on. This will usually involve creating multiple reports based on the same data or insights but each intended for a different audience (for example, in-depth technical analysis for engineers, and an overview of the impact on the bottom line for c-level executives).

Spark: Spark is another open source framework like Hadoop but more recently developed and more suited to handling cutting-edge Big Data tasks involving real time analytics and machine learning. Unlike Hadoop it does not include its own filesystem, though it is designed to work with Hadoop's HDFS or a number of other options. However, for certain data related processes it is able to calculate at over 100 times the speed of Hadoop, thanks to its in-memory processing capability. This means it is becoming an increasingly popular choice for projects involving deep learning, neural networks and other compute-intensive tasks.

Structured Data: Structured data is simply data that can be arranged neatly into charts and tables consisting of rows, columns or multi-dimensioned matrixes. This is traditionally the way that computers have stored data, and information in this format can easily and simply be processed and mined for insights. Data gathered from machines is often a good example of structured data, where various data points – speed, temperature, rate of failure, RPM etc. – can be neatly recorded and tabulated for analysis.

Unstructured Data: Unstructured data is any data which cannot easily be put into conventional charts and tables. This can include video data, pictures, recorded sounds, text written in human languages and a great deal more. This data has traditionally been far harder to draw insight from using computers which were generally designed to read and analyze structured information. However, since it has become apparent that a huge amount of value can be locked away in this unstructured data, great efforts have been made to create applications which are capable of understanding unstructured data – for example visual recognition and natural language processing.

Visualization : Humans find it very hard to understand and draw insights from large amounts of text or numerical data – we can do it, but it takes time, and our concentration and attention is limited. For this reason effort has been made to develop computer applications capable of rendering information in a visual form – charts and graphics which highlight the most important insights which have resulted from our Big Data projects. A subfield of reporting (see above), visualizing is now often an automated process, with visualizations customized by algorithm to be understandable to the people who need to act or take decisions based on them. Basically Available, Soft State and Eventual Consistency Basically Available: The system is guaranteed to be available in event of failure.

- **Soft State**: The state of the data could change without application interactions due to eventual consistency.
- **Eventual Consistency:** The system will be eventually consistent after the application input. The data will be replicated to different nodes and will eventually reach a consistent state. But the consistency is not guaranteed at a transaction level.

Other Definitions of BASE Include:

- "A system allowing horizontal scaling, fault tolerance and high availability at the cost of consistency," (Akshay Pore)
- An alternative to the ACID data processing model. (Dan Pritchett)
- A consistency model that values availability with less strict assurance of consistency than in an ACID database model. (Neo4j)
- An acronym used to describe the properties of certain databases, usually NoSql. (Stackoverflow)

BASE Case Examples:

- Creating a value-based model for a health organization from disparate informational sources
- Using shopping cart applications on a website
- Uncovering fraud rings and scams
- Monitoring network and IT infrastructure security
- Managing and reusing document content

Businesses use BASE Database Types to:

- Use the object-oriented architecture native to the cloud
- Process large amounts of unstructured data very quickly
- Support data research
- Find patterns, such as fraud
- Gain business insights
- Manage data from the Internet of Things (IoT)