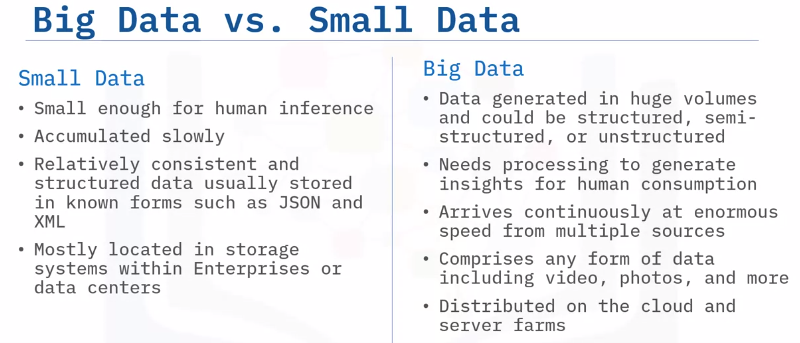
# Introduction to Big Data

**Learning Objectives**

* Define what Big Data is and identify key Big Data characteristics.
* Provide examples of Big Data related technologies, describe the impact of Big Data on businesses and people, and define the relationship between the Internet of Things (IoT) and Big Data.
* Compare linear to parallel processing and explain why Big Data requires parallel processing.
* Describe scalability, including horizontal scaling, embarrassing parallel calculations, and fault tolerance as they relate to Big Data.
* List the key Big Data ecosystem tooling categories and their associated major tools and vendors.
* Explain the role of open-source in Big Data and list related platforms and open-source frameworks.
* Describe the key sources and different types of Big Data.
* Provide examples of real-world Big Data use cases.

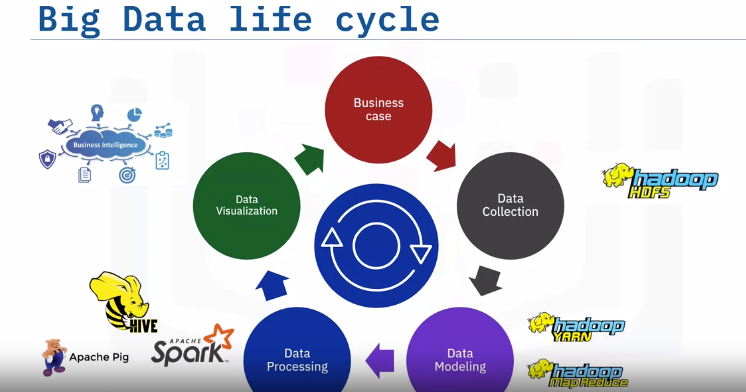
Introduction:

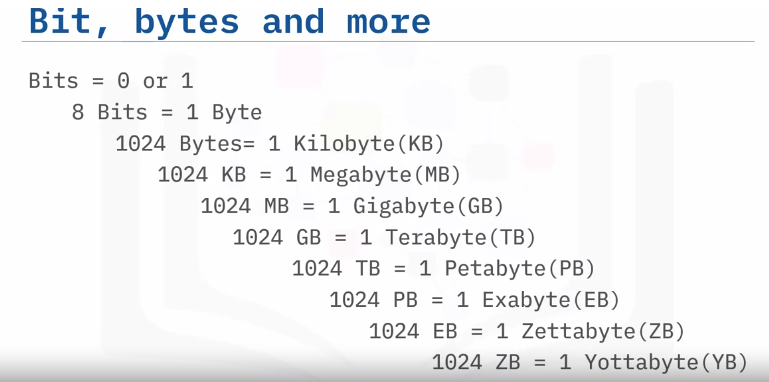
The latest statistics report that the accumulated world’s data will grow from 4.4 zettabytes to 44 zettabytes, with much of that data classified as Big Data. Revenues based on Big Data analytics are projected to increase to $103 billion by 2027. Understandably, organizations across industries want to harness the competitive advantages of Big Data analytics. This course provides us with the foundational knowledge and hands-on lab experience we need to understand what Big Data is and learn how organizations use Apache Hadoop, Apache Spark, including Apache Spark SQL, and Kubernetes to expedite and optimize Big Data processing.



It is a common misconception that Big Data refers to just large volumes of data. In reality, Big Data is the entire life cycle of working with large volumes of data. Let’s take a look at each phase in the Big Data life cycle. Big Data collection is initiated as a result of a business problem or requirement. As data is collected, it gets stored using a framework for distributed storage such as Hadoop HDFS.

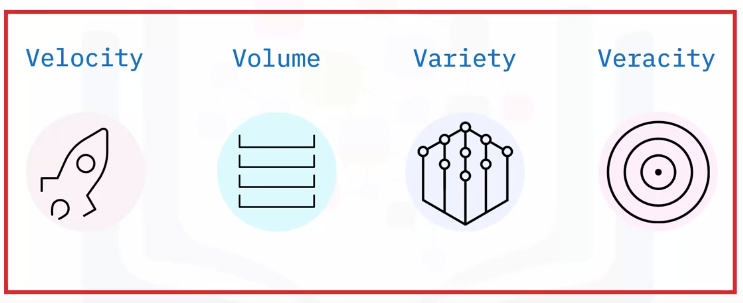
To make sense of all the data collected, Map and Reduce tasks and scripts create a data model to store it in a database. This data model includes the various data entities (or objects), and the relationship and rules between these entities. After modeling, data is ready to be processed. Tools such as Apache Spark are used to produce meaningful information from the modeled data. Finally, the processed data is visualized and presented in a graphical format such as charts and graphs. This visualized data is then used for making meaningful business decisions and lead to new business cases, thereby creating a continuous life cycle.





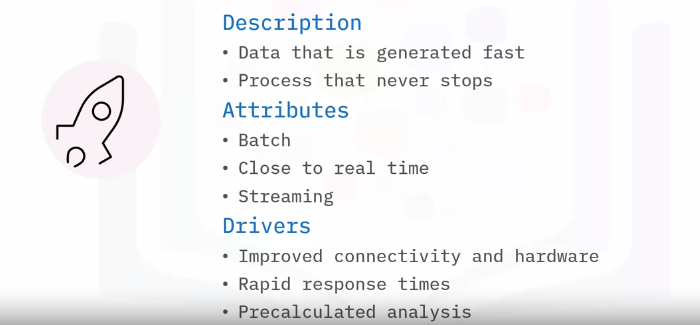
# Four V’s:

When we talk about Big Data, we traditionally talk about the four V’s of Big Data. These are: Velocity, Volume, Variety, and Veracity. Velocity is the speed at which data arrives. Volume is the increase in the amount of data stored over time. Variety is the diversity of data. Many forms of data exist and need to be stored. Veracity is the certainty of data. With a large amount of data available, how will we know if the data collected is accurate or inaccurate? These four main components are used to describe the dimensions of Big Data.



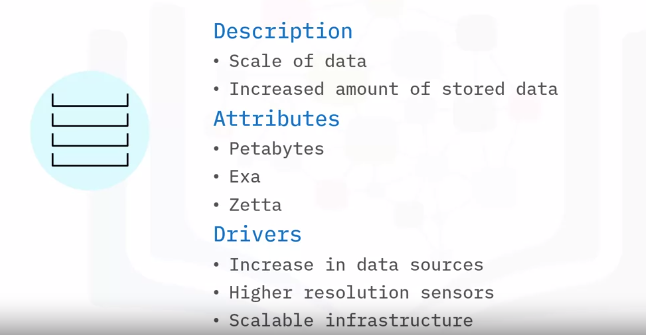
## Velocity:

Velocity signifies that data is being generated extremely fast, and the process never stops. Data must be processed quickly so that decisions can be made at the speed with which the data arrives. Velocity’s main attributes are: batch close to real-time, and streaming. What are the drivers? Definitely improving connectivity and hardware. Just think about all the devices that are connected through the Internet today and all the super-fast response times. Big Data also supports upscaling of pre-calculated analysis.



## Volume:

Volume is the increase in the amount of data stored. The amount of Big Data generated is vast compared to traditional data sources. Volume attributes of Big Data are: Petabytes, Exa, and Zetta, to name just a few. Typical drivers of volume in Big Data are: the increase in data sources, higher resolution sensors, and scalable hardware infrastructure.

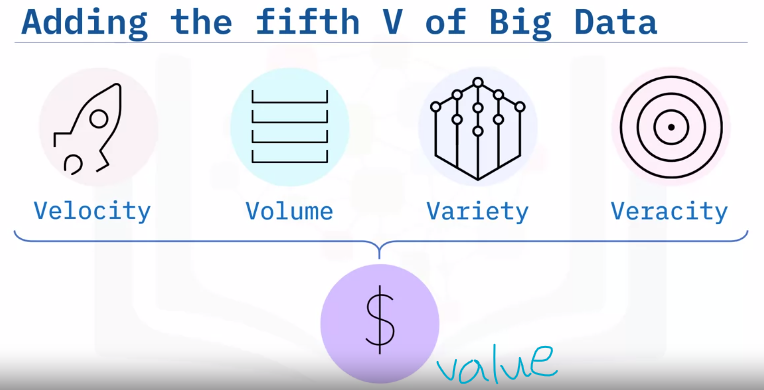


## Variety:

Variety is the diversity of the data. Data is generated by people and processes through the use of machines, from both inside and outside an organization. Some of the data is structured and semi-structured, but most is unstructured. The main attributes are structure, complexity, and origin. Drivers of Variety in Big Data can be: mobile technologies, scalable hardware infrastructure, resilience, fault recovery, and efficient storage and retrieval.

## Veracity:

Veracity is the quality, origin, and conformity to facts and accuracy of the data. This is because data comes from both within and outside an organization. Attributes include consistency and completeness, integrity, and ambiguity. Drivers of Veracity in Big Data are: cost and the need for traceability, robust ingestion, and extract, transform, load (ETL) mechanisms.



Big Data has another V that must be considered. The fifth V of Big Data is Value. It is the outcome of making intelligent business decisions from leveraging the previous four V’s. The ultimate goal of an organization is to: produce value in the form of faster and smarter business decisions, increase efficient use of resources, and discover of new market opportunities. Big Data supports innovation and thus creates value.

**Summary:**

Big Data is the digital trace that gets generated in this digital era. Big Data is a high-volume, high-velocity, and/or high-variety information asset that demands cost-effective and innovative tools for processing. The core features of Big Data are the 4 V’s: Velocity, Volume, Variety, and Veracity. Big Data creates a fifth V, Value, when collected, stored, and processed correctly.