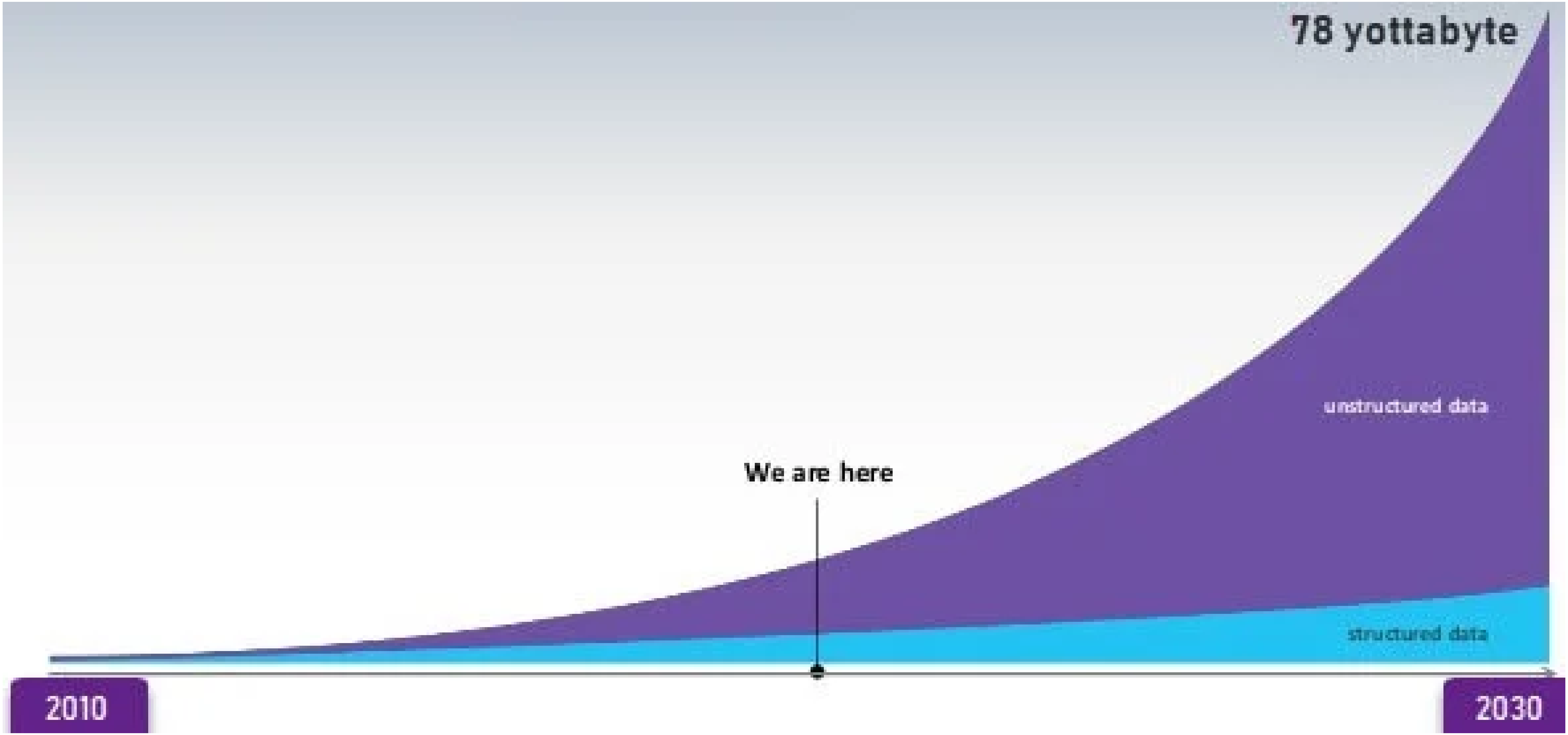
introduction to big data

DR. AMRITPAL SINGH

Data that is always increasing and cannot be processed and stored on a single machine is termed as Big Data.

Data Growth over the years



Big Data Examples

The New York Stock Exchange is an example of Big Data that generates about one terabyte of new trade data per day.

Big Data Examples

statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day.

Big Data Examples

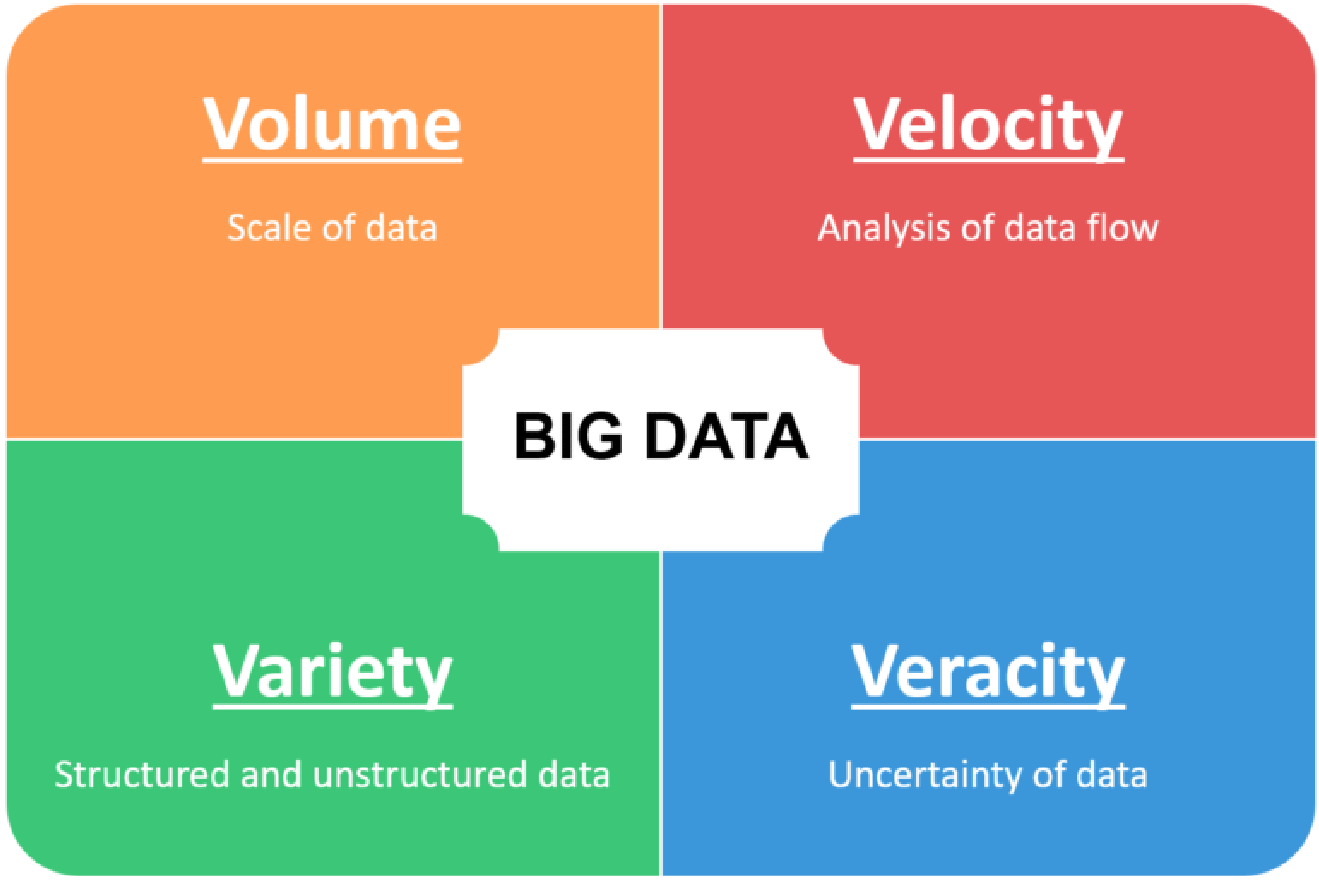
A single Jet engine can generate 10+terabytes of data in 30 minutes of flight time. With many thousand flights per day, generation of data reaches up to many Petabytes.

Types of Data

Structured Data
Unstructured Data
Semi-Structured Data

V's of Big Data

Volume
Variety
Velocity
Variability
Veracity
Visualization
Value



Volume

Scale of data

Velocity

Analysis of data flow

BIG DATA

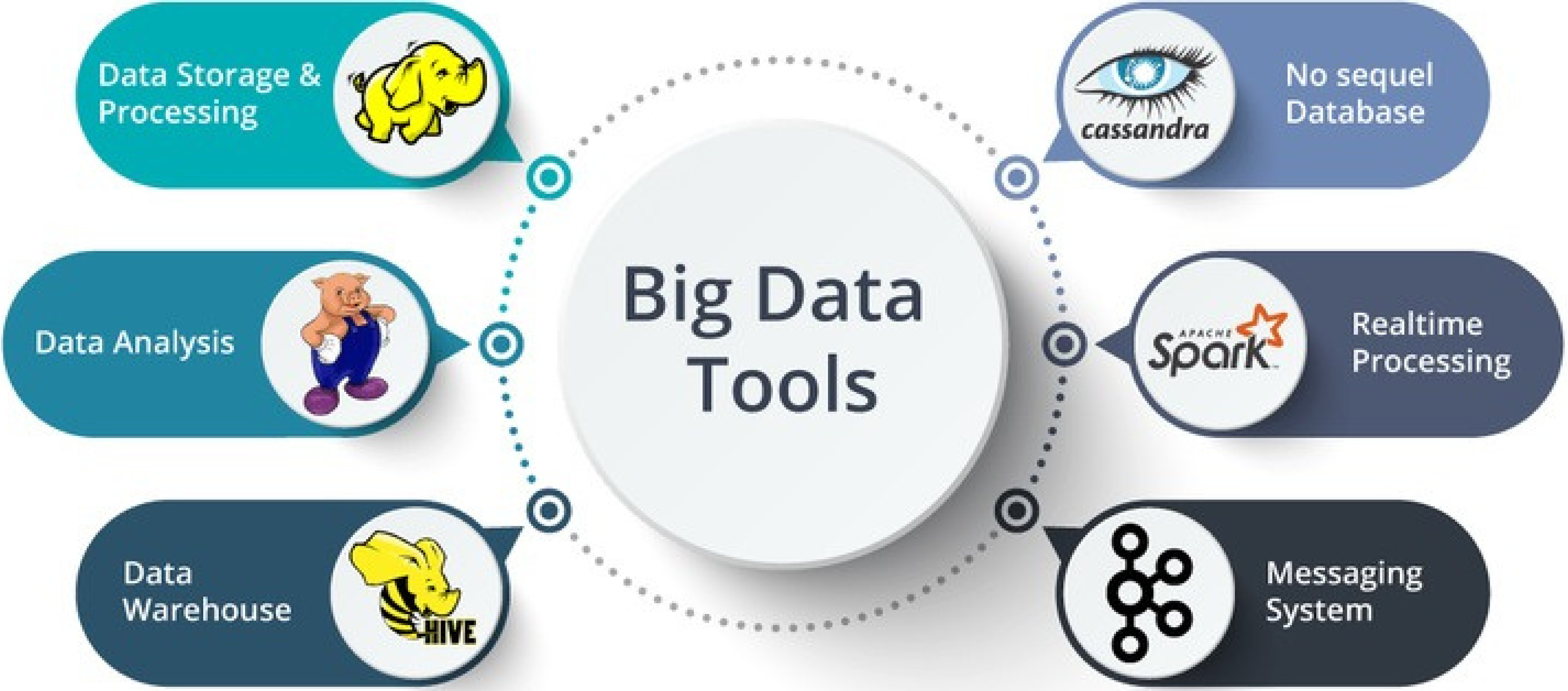
Variety

Structured and unstructured data

Veracity

Uncertainty of data

Big Data Tools



Big Data Tools

Data Storage & Processing

Data Analysis

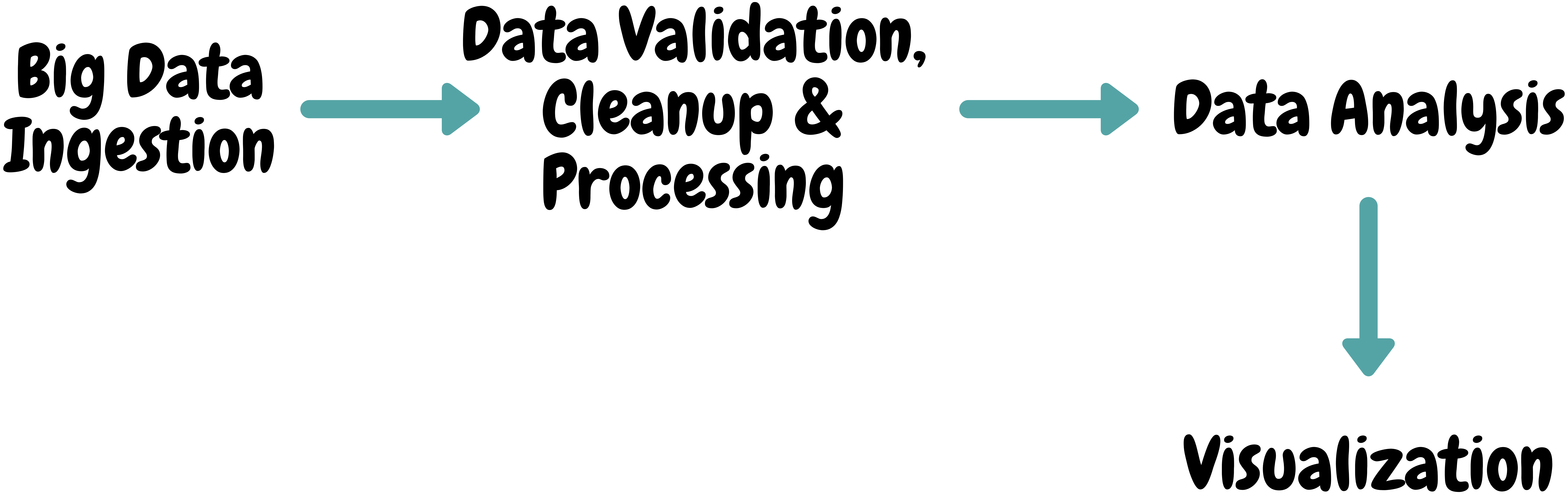
Data Warehouse

No sequel Database

Realtime Processing

Messaging System

Big Data Pipeline



Big Data Ingestion

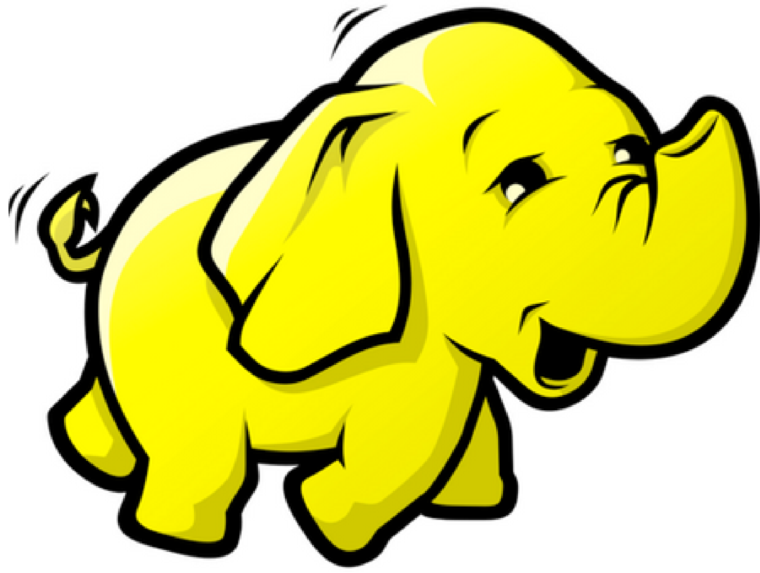
Data Validation, Cleanup & Processing

Data Analysis

Visualization

Intro to Hadoop

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging from GBs to PBs.

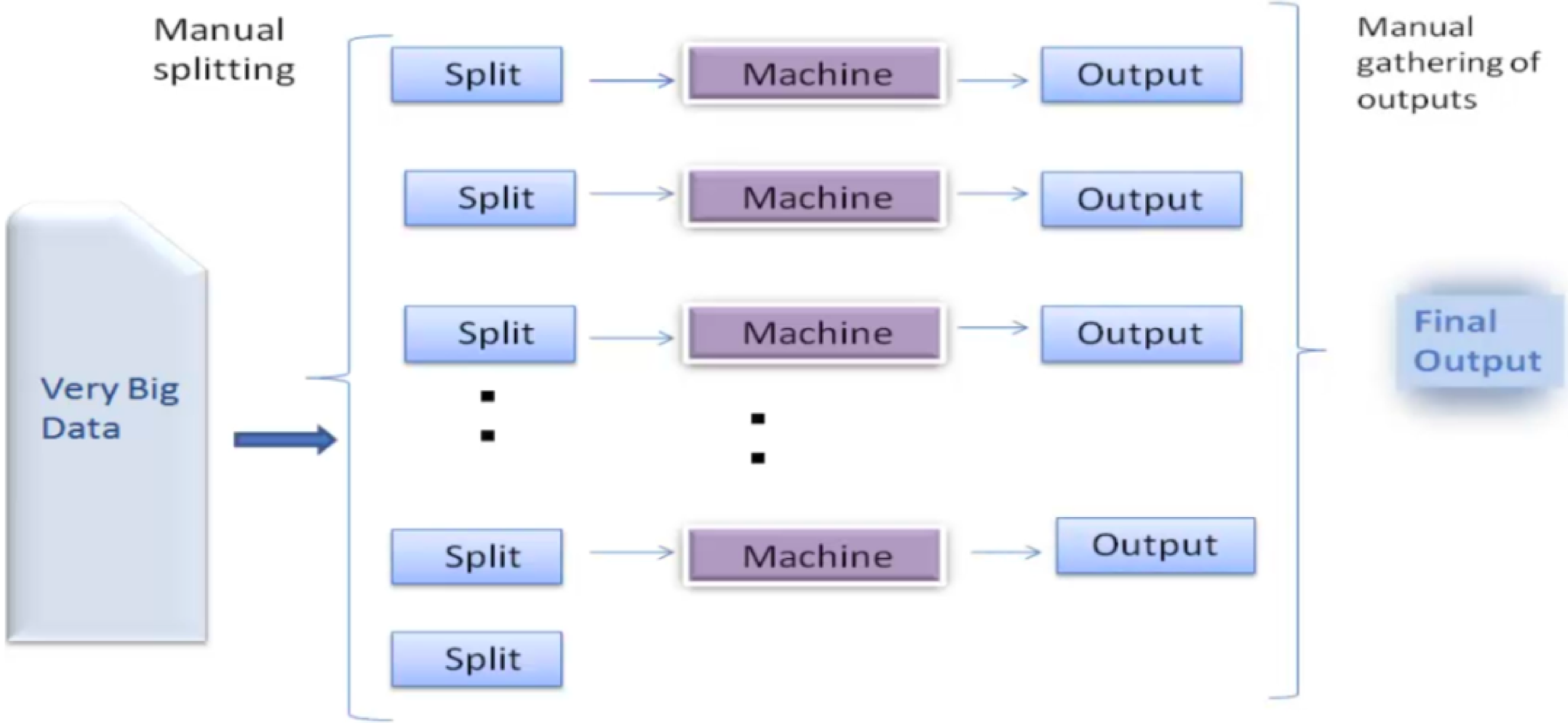


WHAT IS MAP REDUCE?

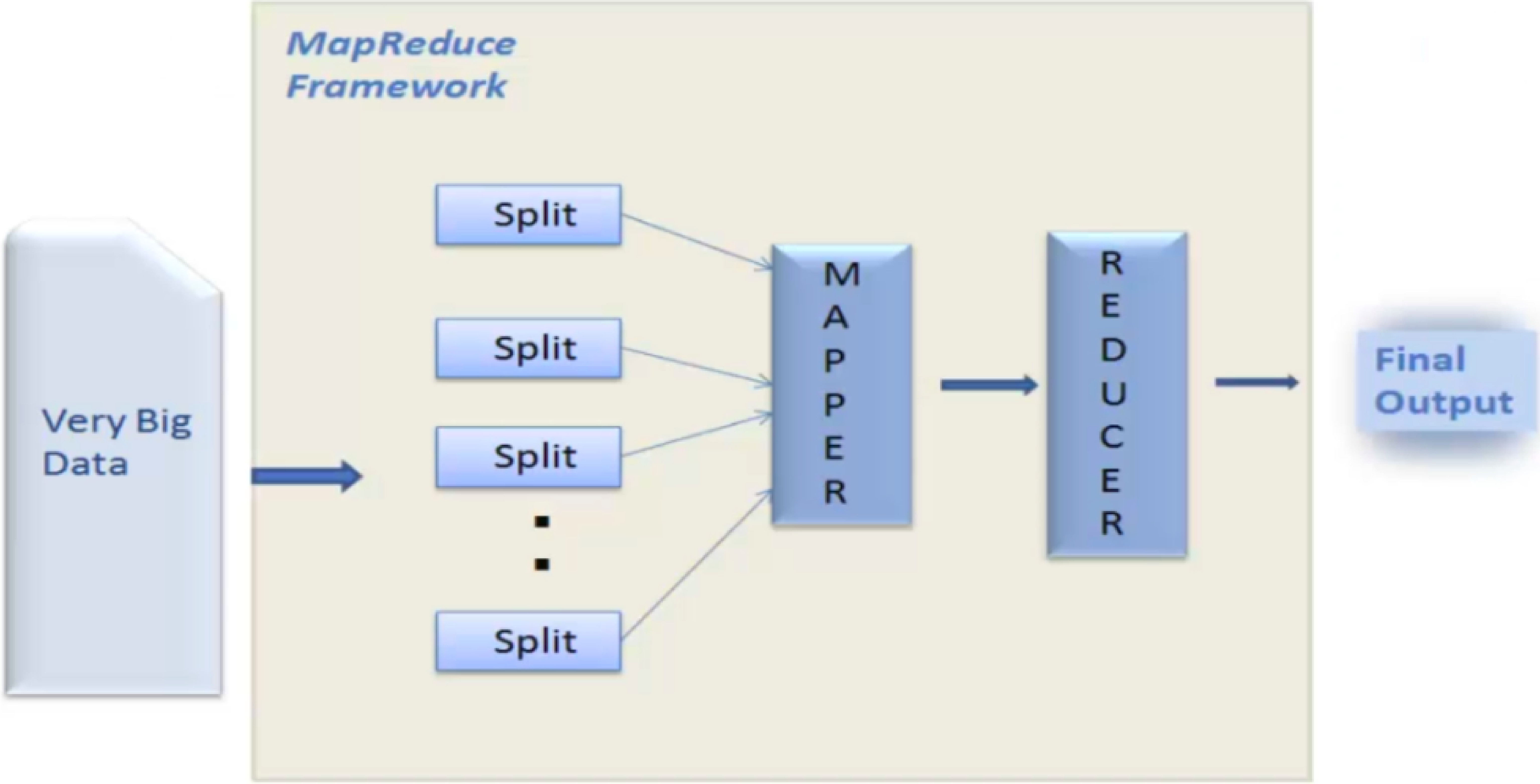
MapReduce is the core component for data processing in Hadoop framework.

In layman's term Mapreduce helps to split the input data set into a number of parts and run a program on all data parts parallel at once. The term MapReduce refers to two separate and distinct tasks.

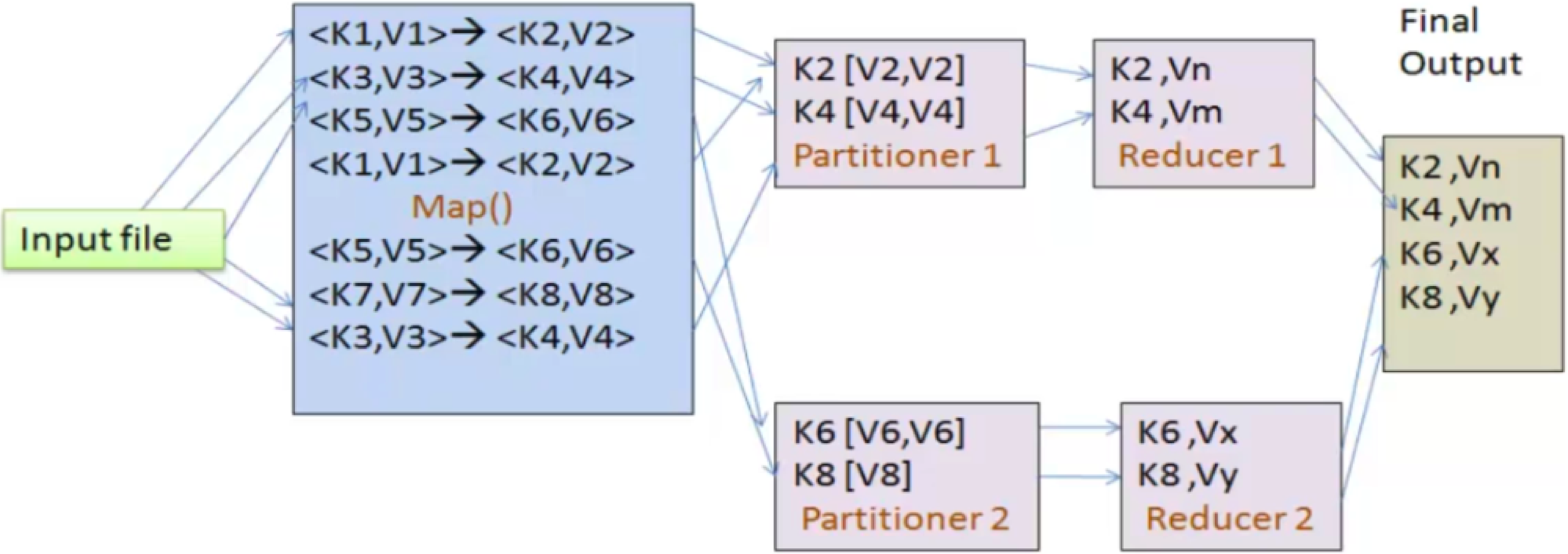
TRADITIONAL APPROACH



MAP REDUCE APPROACH



MAP REDUCE APPROACH



Input file
<K1,V1>→ <K2,V2>
<K3,V3>→ <K4,V4>
<K5,V5>→ <K6,V6>
<K1,V1>→ <K2,V2>
Map()
<K5,V5>→ <K6,V6>
<K7,V7>→ <K8,V8>
<K3,V3>→ <K4,V4>
K2 [V2,V2]
K4 [V4,V4]
Partitioner 1
K6 [V6,V6]
K8 [V8]
Partitioner 2
K2 ,Vn
K4 ,Vm
Reducer 1
K6 ,Vx
K8 ,Vy
Reducer 2
Final Output
K2 ,Vn
K4 ,Vm
K6 ,Vx
K8 ,Vy