H1B VISA DATA ANALYSIS

USING HADOOP FRAMEWORK

BY: PIYUSH KUMAR
ID: S180020200197

NIIT

SOME OF THE REASONS WHY BIG DATA IS GENERATING:

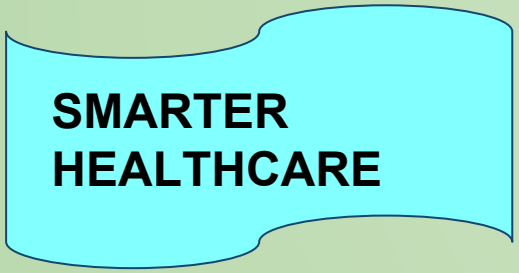
NEED OF MORE STORAGE CAPABILITIES

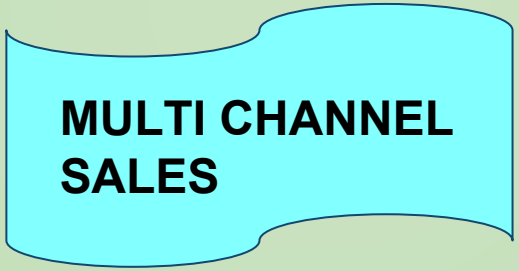
INCREASE OF PROCESSING POWER

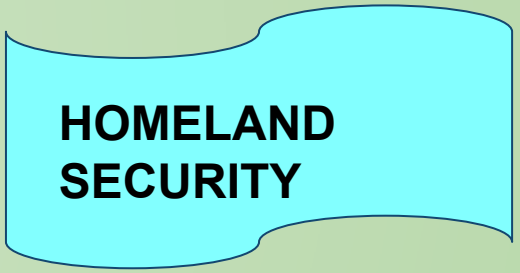
AVAILABILITY OF DATA (DIFFERENT DATA TYPES)

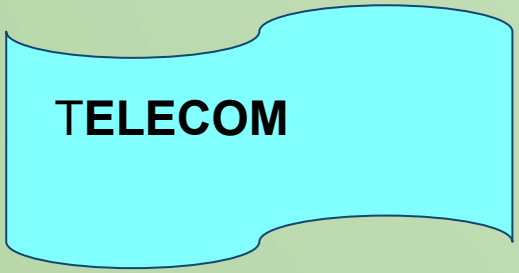
EVERYDAY WE GENERATE 2.5 QUINTILLIONS OF DATA; 90% OF THE DATA TODAY IN THE WORLD HAS BEEN CREATED IN THE LAST 2 YEARS.

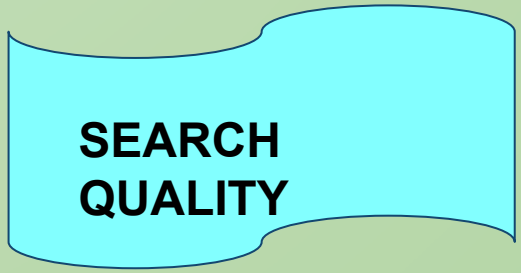
APPLICATION OF BIG DATA ANALYTICS:-

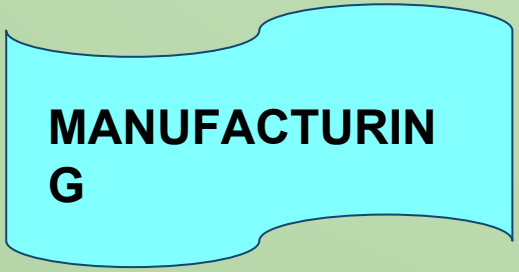




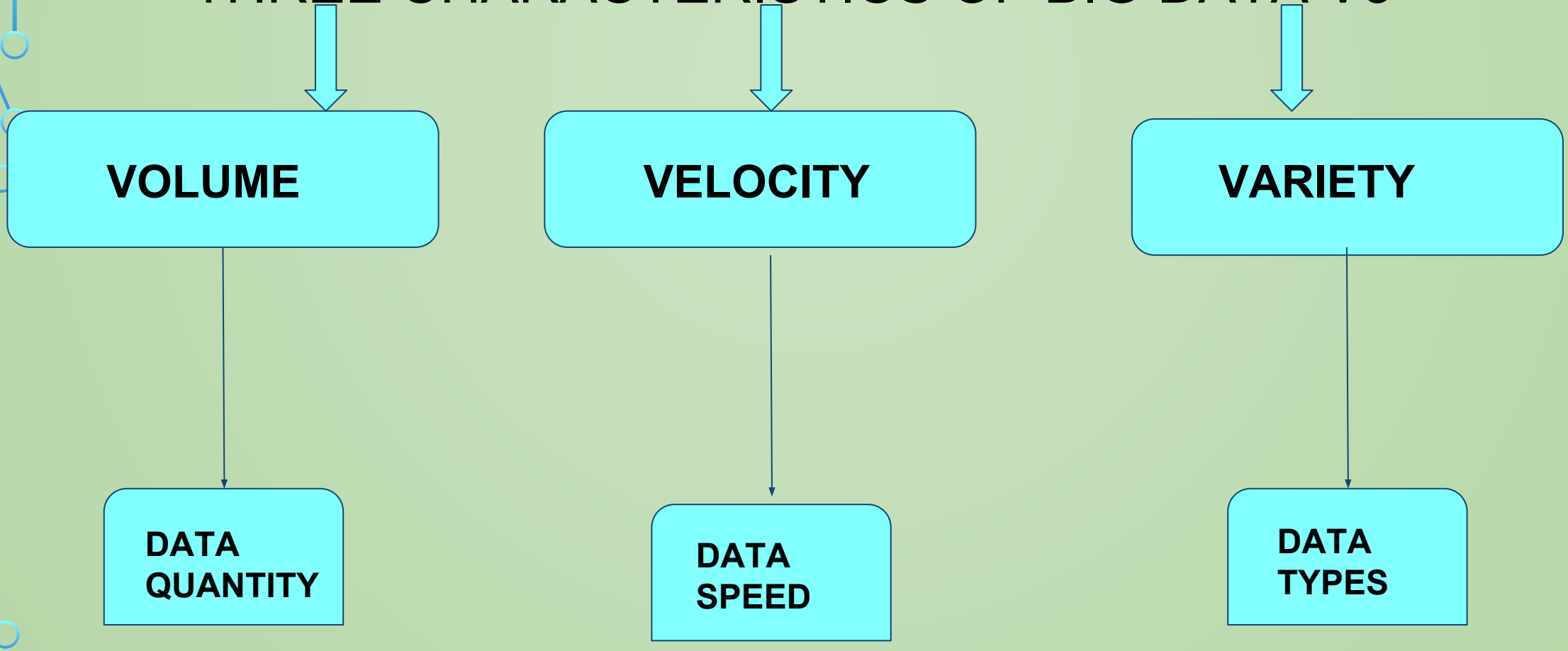








THREE CHARACTERISTICS OF BIG DATA V3



RISKS

1. Will be so overwhelmed:- need the right person and solve the correct problem.
2. Cost escalates too fast.
3. Many sources of big data is privacy:
   Self-regulation
   Legal- regulation

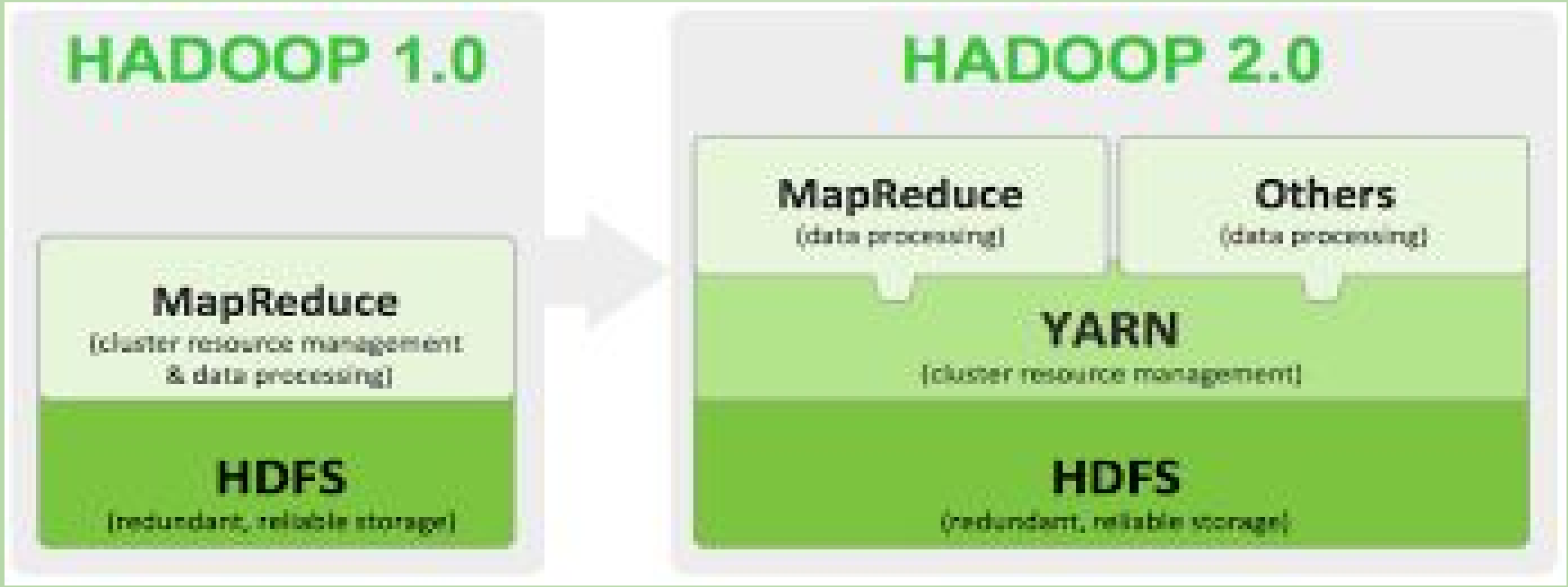
BENEFITS

1. Cost reductions and Time reduction
2. New product development and optimized offerings
3. Smart decision making.
4. Determining root causes of failures, issues and defects in near-real time.
5. Generating coupons at the point of sale, based on the customer's buying habits.
6. Detecting fraudulent behaviour before it affects the organization.

What is Apache Hadoop?

- Open source software framework from apache designed for storage and processing of large data in volume.
- Written in java.
- Not OLAP(online analytical processing) but used for offline processing.
- Cutting named the program after his son’s toy elephant.
- Used by yahoo, twitter, facebook etc.

THE HADOOP MODULES



HDFS(Hadoop distributed file system)

Where to use:-

Very largefiles

Streaming data access

Commodity hardware

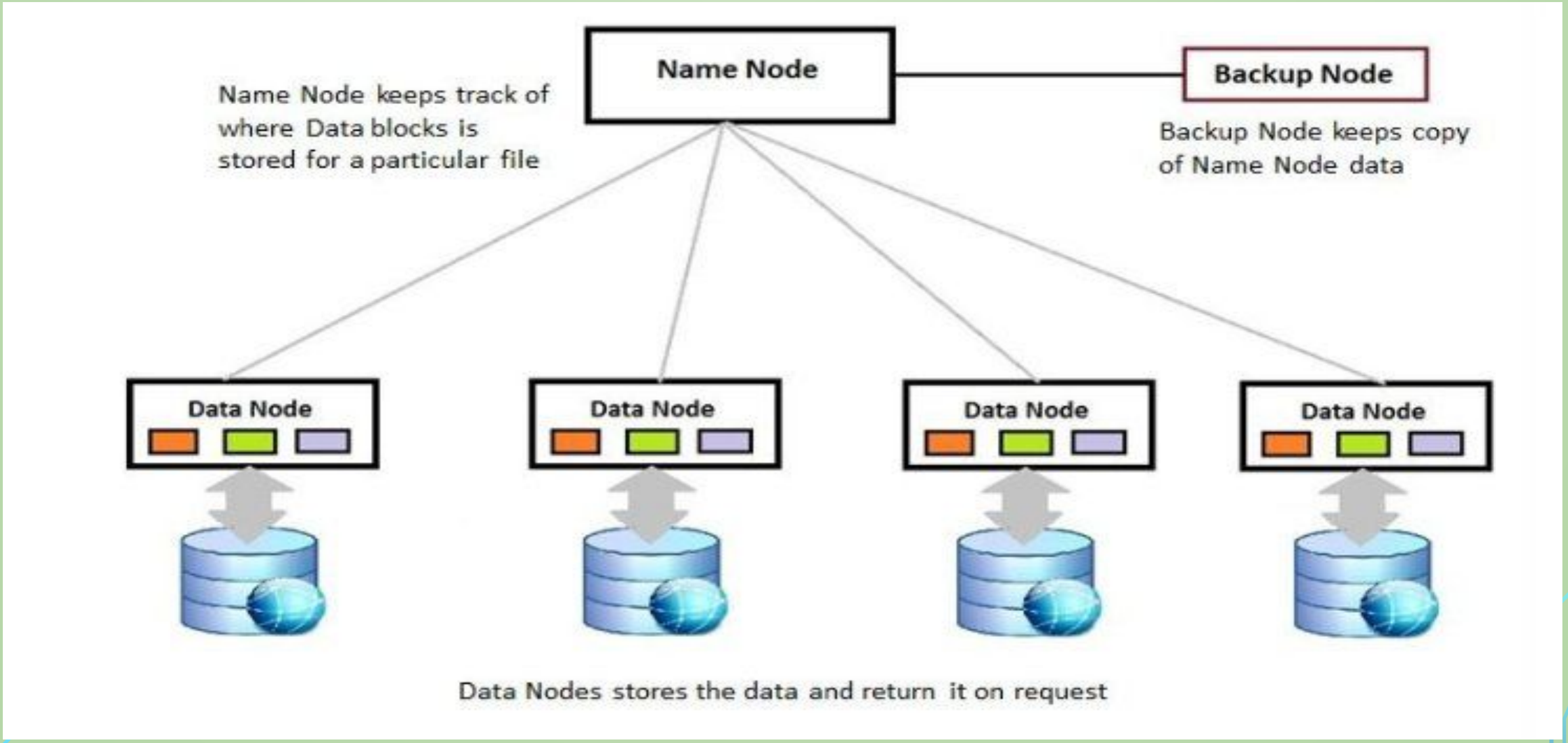
Where not to use:-

Low latency data access

Lots of small files

Multiple writes

HDFS DATANODE AND NAMENODE



YARN

COMPONENTS:-

Client
Resource manager
Node manager
Mapreduce application master

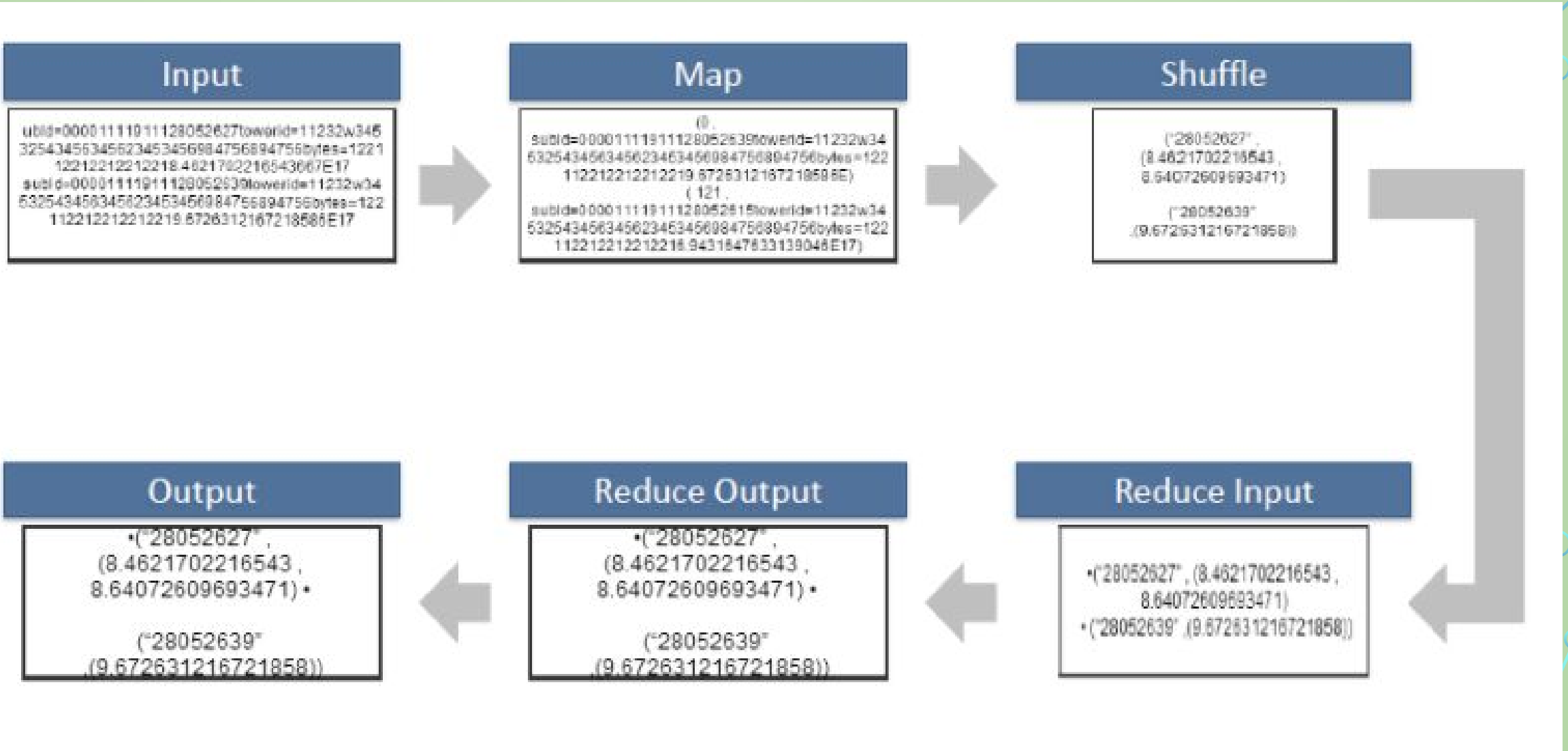
BENEFITS:-

Scalability
Utiliazation
Multitenancy

MAPREDUCE OVERVIEW

- A method for distributing computation across multiple nodes
- Each node processes the data that is stored at that node
- Consists of two main phases
  1. Map
  2. Reduce

Steps in MAPREDUCE



THE MAPPER

- Reads data as key/value pairs
  - The key is often discarded
- Outputs zero or more key/value pairs

SORT AND SHUFFLE

- Output from the mapper is sorted by key
- All values with the same key are guaranteed to go to the same machine

THE REDUCER

- The Reducer code reads the outputs generated by the different mappers as pairs and emits key value pairs.
- The reducer outputs zero or more final key/value pair.

OTHER TOOLS IN HADOOP FRAMEWORK

- Pig: Hadoop processing with scripting
- Hive: Hadoop processing with SQL
- HBase: Database model built on top of Hadoop
- Sqoop: For importing and exporting data from RDBMS to HDFS and vice versa.
- Flume: Designed for large scale data movement
- Oozie: Scheduler system to run and manage Hadoop jobs.
- Zookeeper: Co-ordinate and manage service in a distributed environment.

THANK YOU :)