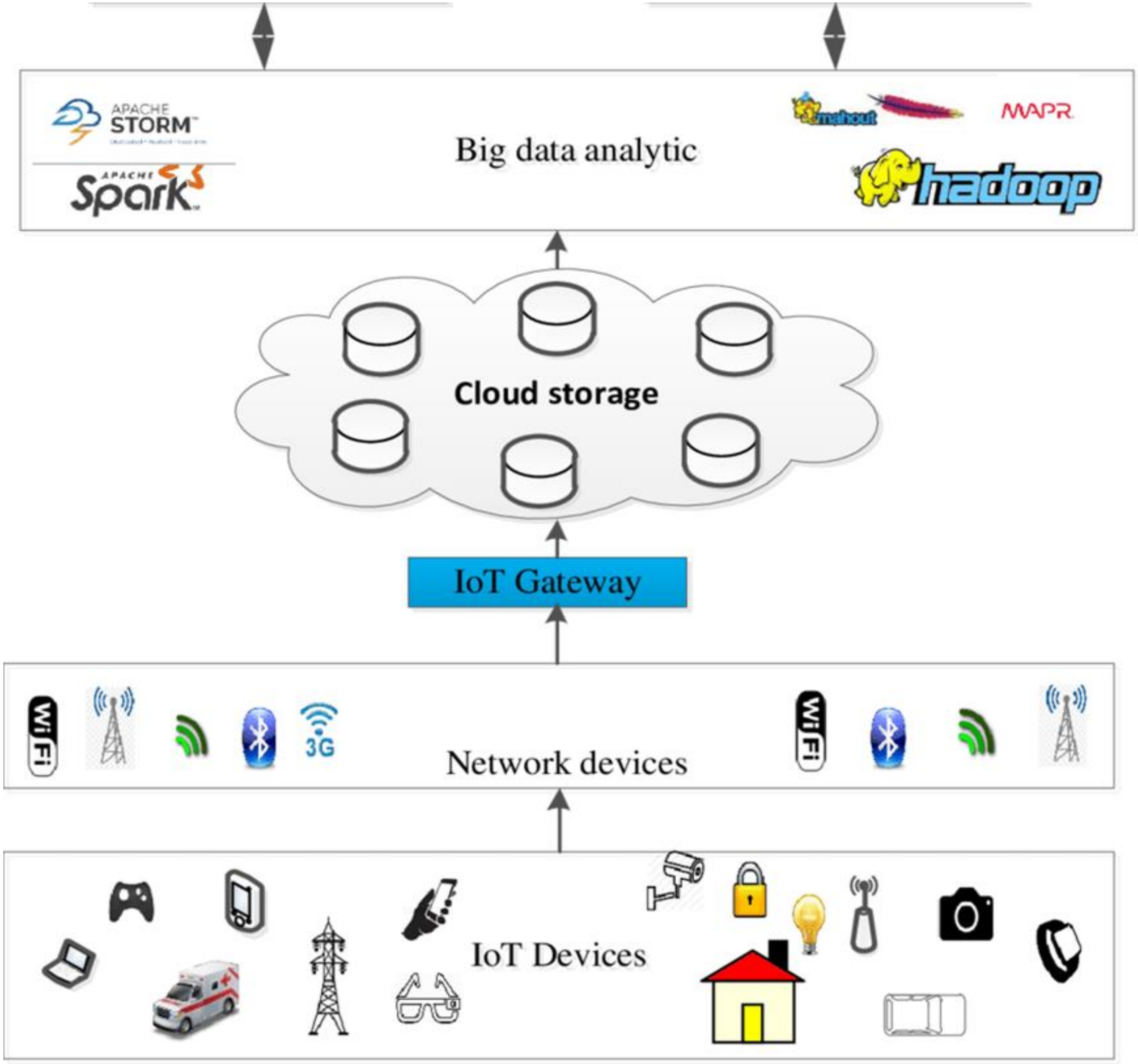
Dr. Ambarish G. Mohapatra

Big Data and Hadoop Cluster

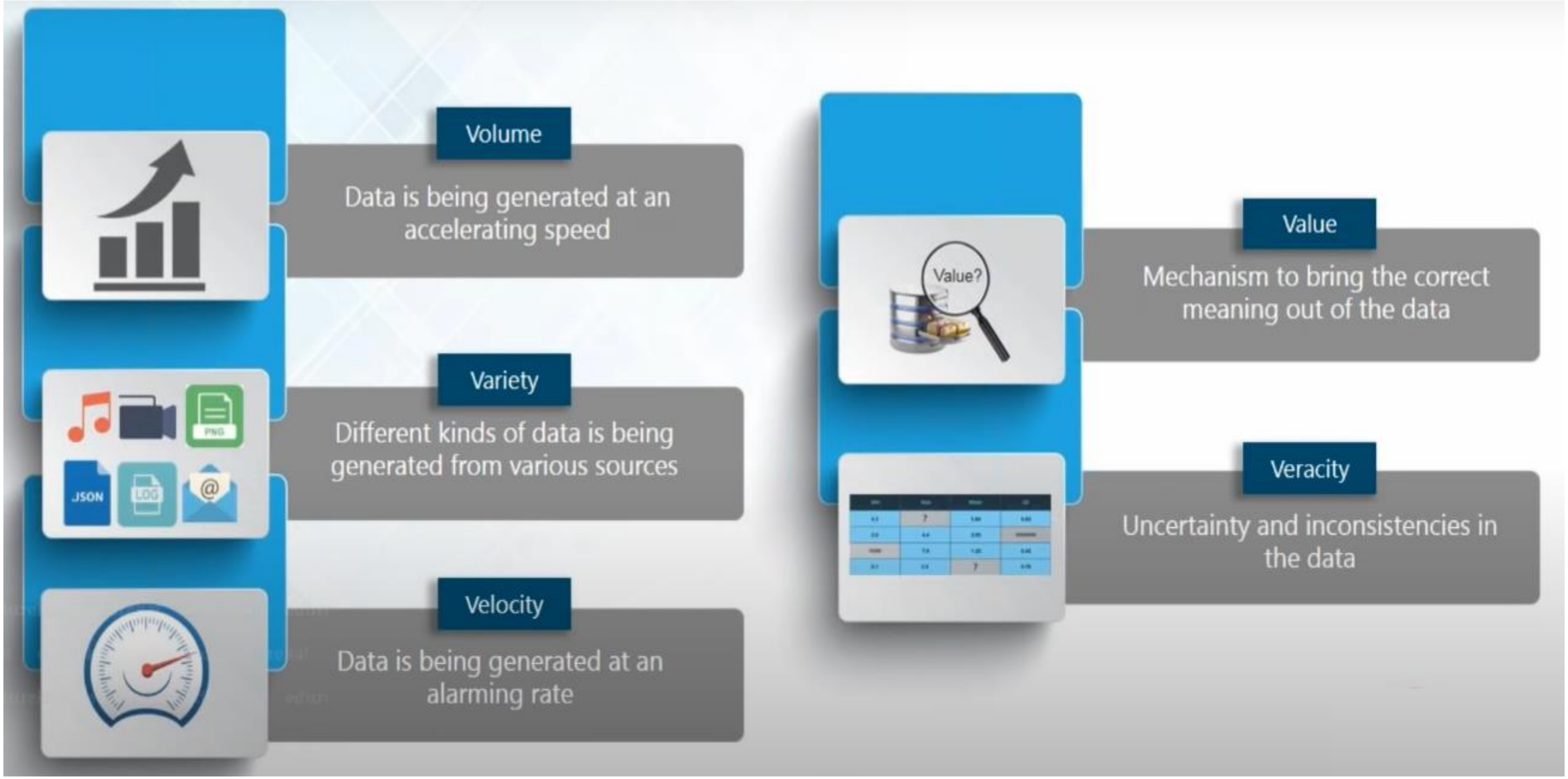
Silicon
...beyond teaching
Silicon Institute of Technology
Bhubaneswar

IoT and Big Data

Dr. Ambarish G. Mohapatra



5’Vs of Big Data



Volume

Data is being generated at an accelerating speed

Variety

Different kinds of data is being generated from various sources

Velocity

Data is being generated at an alarming rate

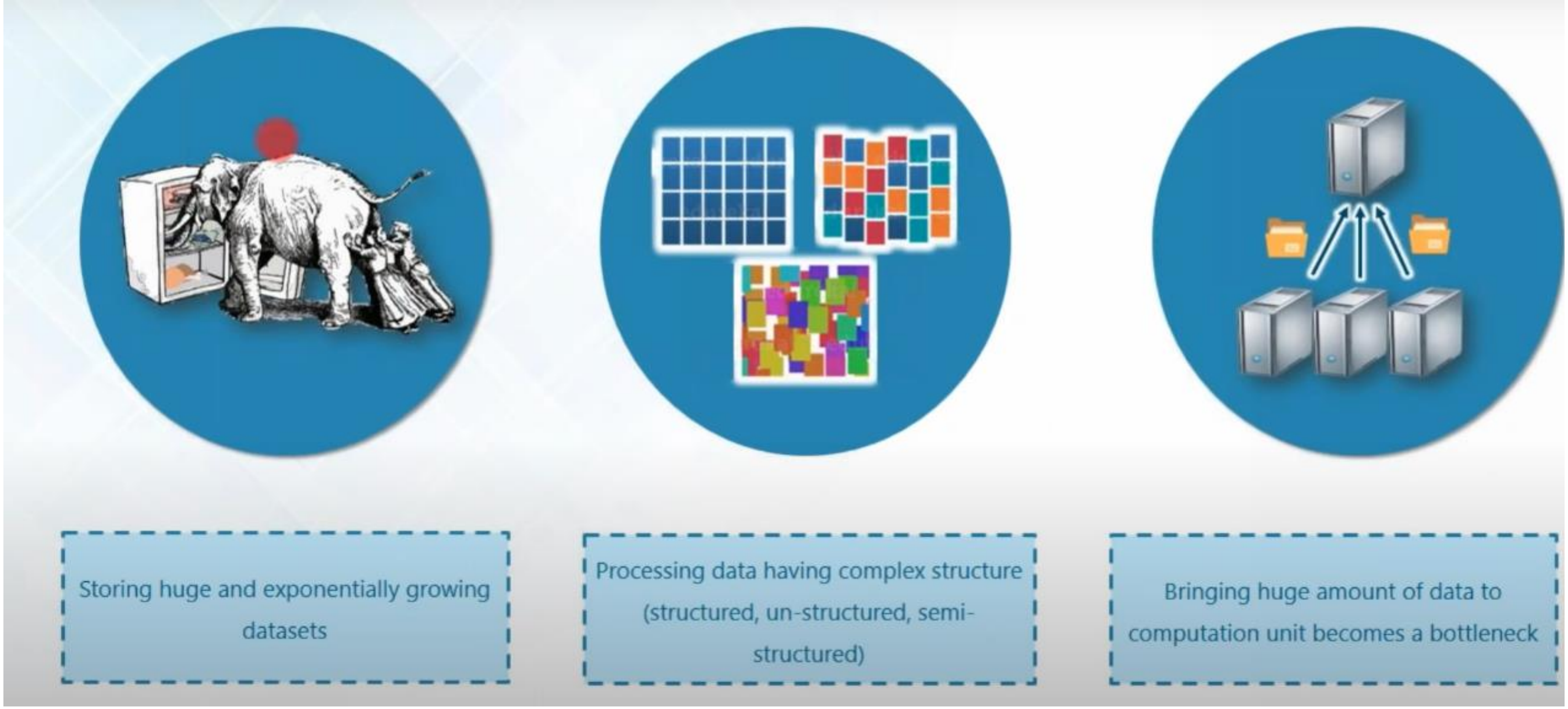
Value

Mechanism to bring the correct meaning out of the data

Veracity

Uncertainty and inconsistencies in the data

How BigData helps…



Storing huge and exponentially growing datasets

Processing data having complex structure (structured, un-structured, semi-structured)

Bringing huge amount of data to computation unit becomes a bottleneck

How Big Data Works

Big data gives you new insights that open up new opportunities and business models. Getting started involves three key actions:

1. **Integrate**

Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as extract, transform, and load (ETL) generally aren't up to the task. It requires new strategies and technologies to analyze big data sets at terabyte, or even petabyte, scale.

2. **Manage**

Big data requires storage. Your storage solution can be in the cloud, on premises, or both. You can store your data in any form you want and bring your desired processing requirements and necessary process engines to those data sets on an on-demand basis.

3. **Analyze**

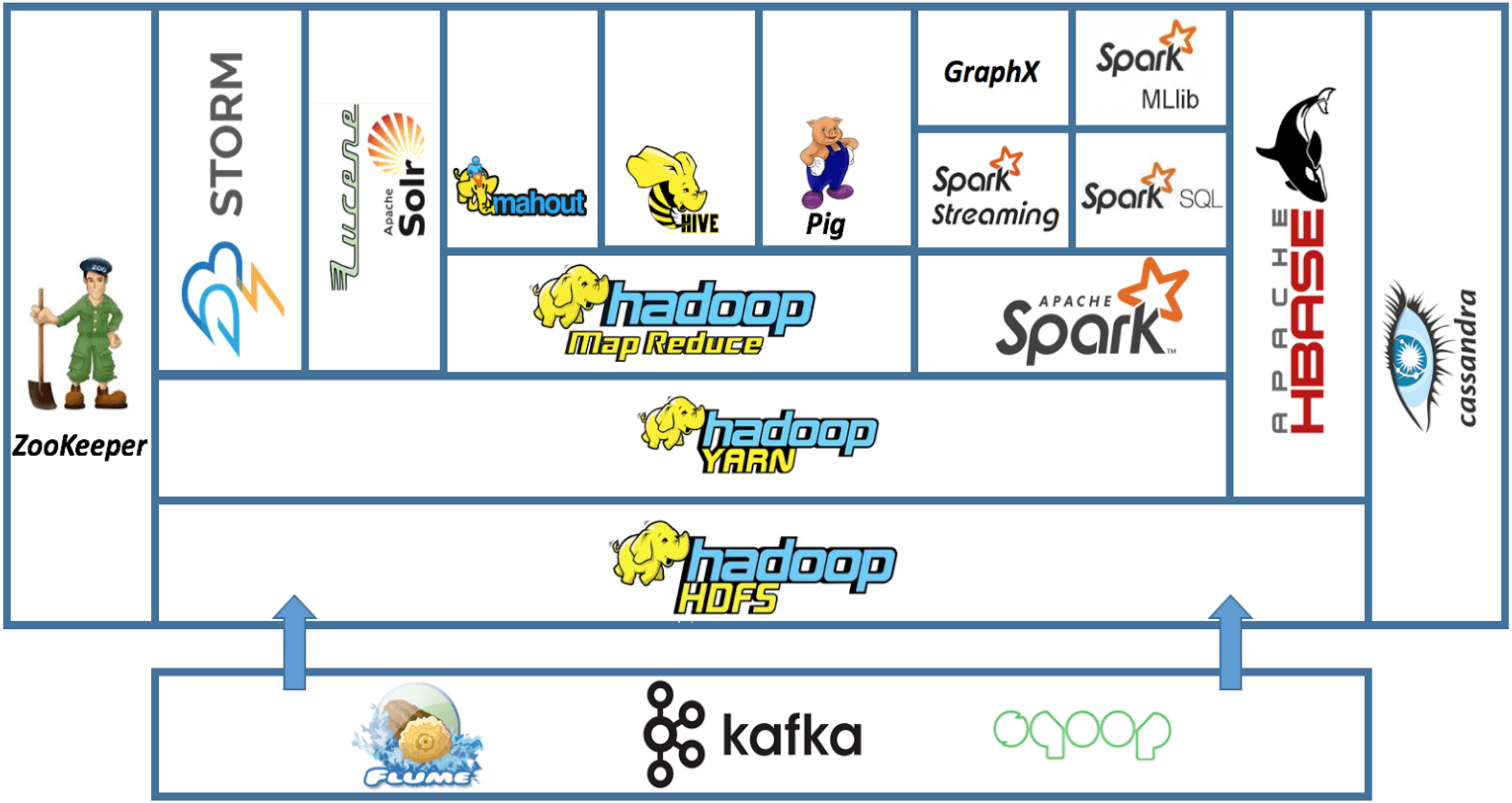
Your investment in big data pays off when you analyze and act on your data.

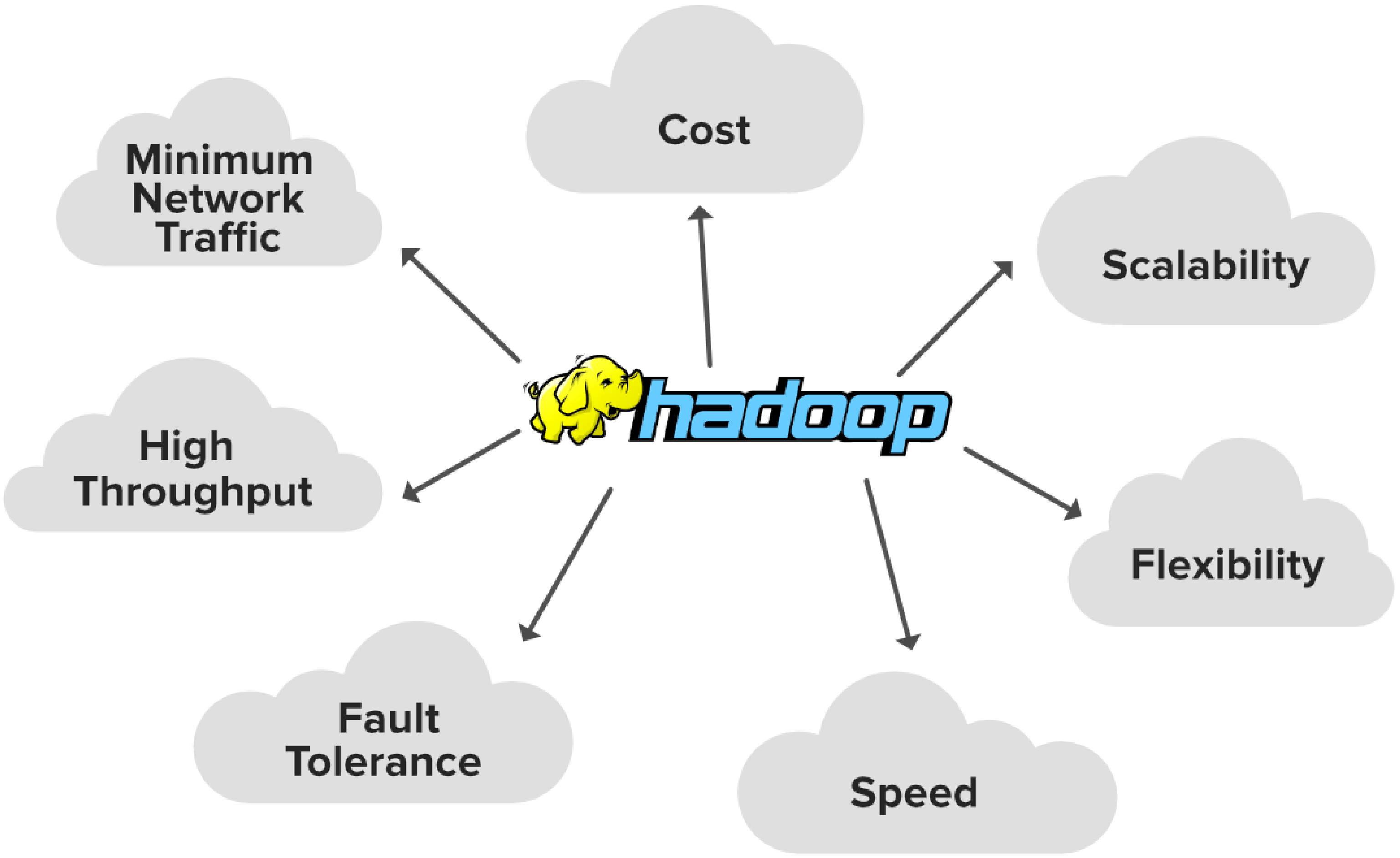
What is Hadoop?

- Software platform that lets one easily write and run applications that process vast amounts of data. It includes:
  - HDFS – Hadoop distributed file system
  - MapReduce – offline computing engine

- Yahoo! is the biggest contributor
- Here's what makes it especially useful:
  - **Scalable:** It can reliably store and process petabytes.
  - **Economical:** It distributes the data and processing across clusters of commonly available computers (in thousands).
  - **Efficient:** By distributing the data, it can process it in parallel **on the nodes where the data is located.**
  - **Reliable:** It automatically maintains multiple copies of data and automatically redeploys computing tasks based on failures.

Hadoop Ecosystem





Cost

Scalability

Flexibility

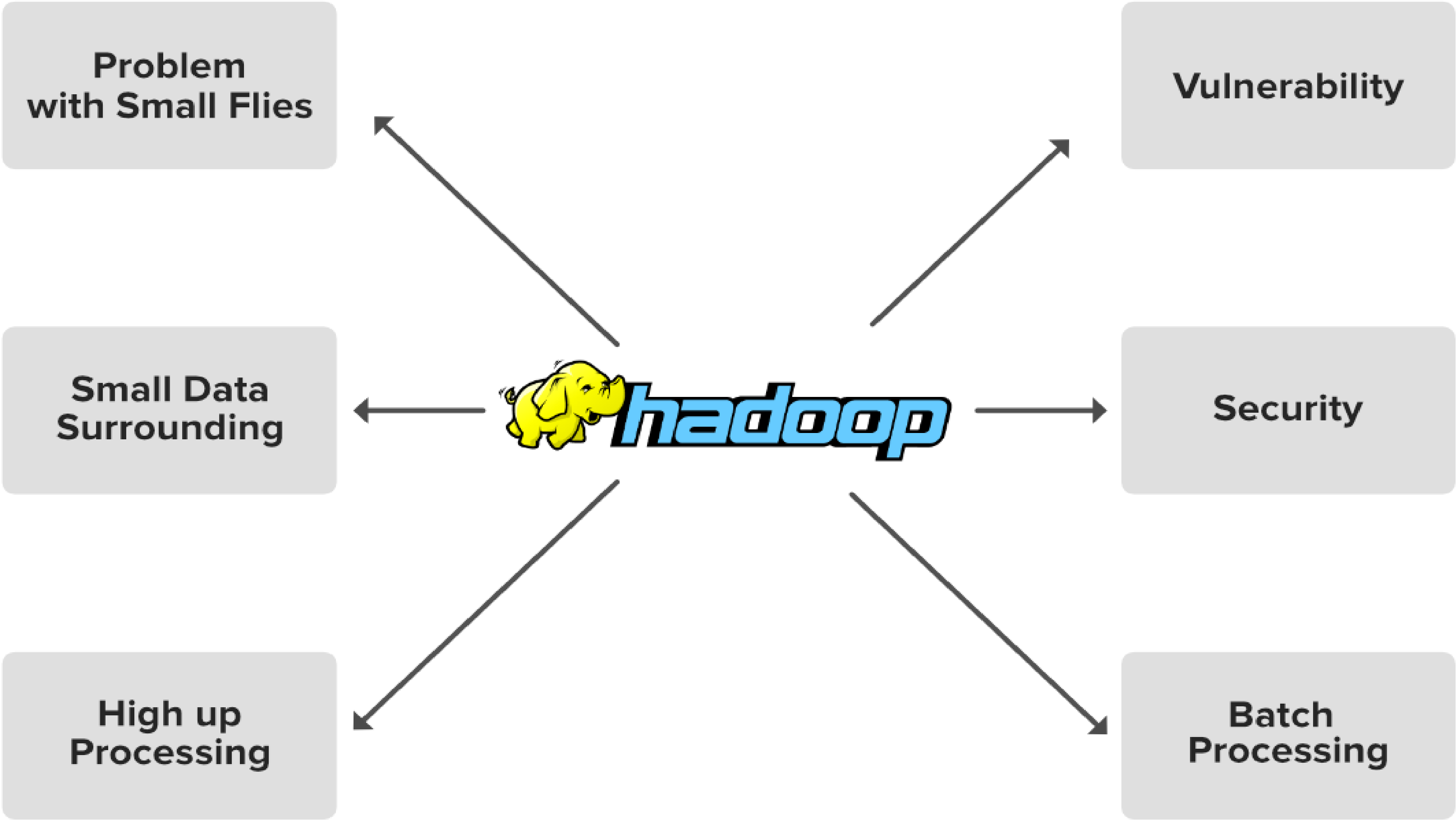
Speed

Fault Tolerance

High Throughput

Minimum Network Traffic

Hadoop Advantages



Problem with Small Flies

Vulnerability

Small Data Surrounding

Security

High up Processing

Batch Processing

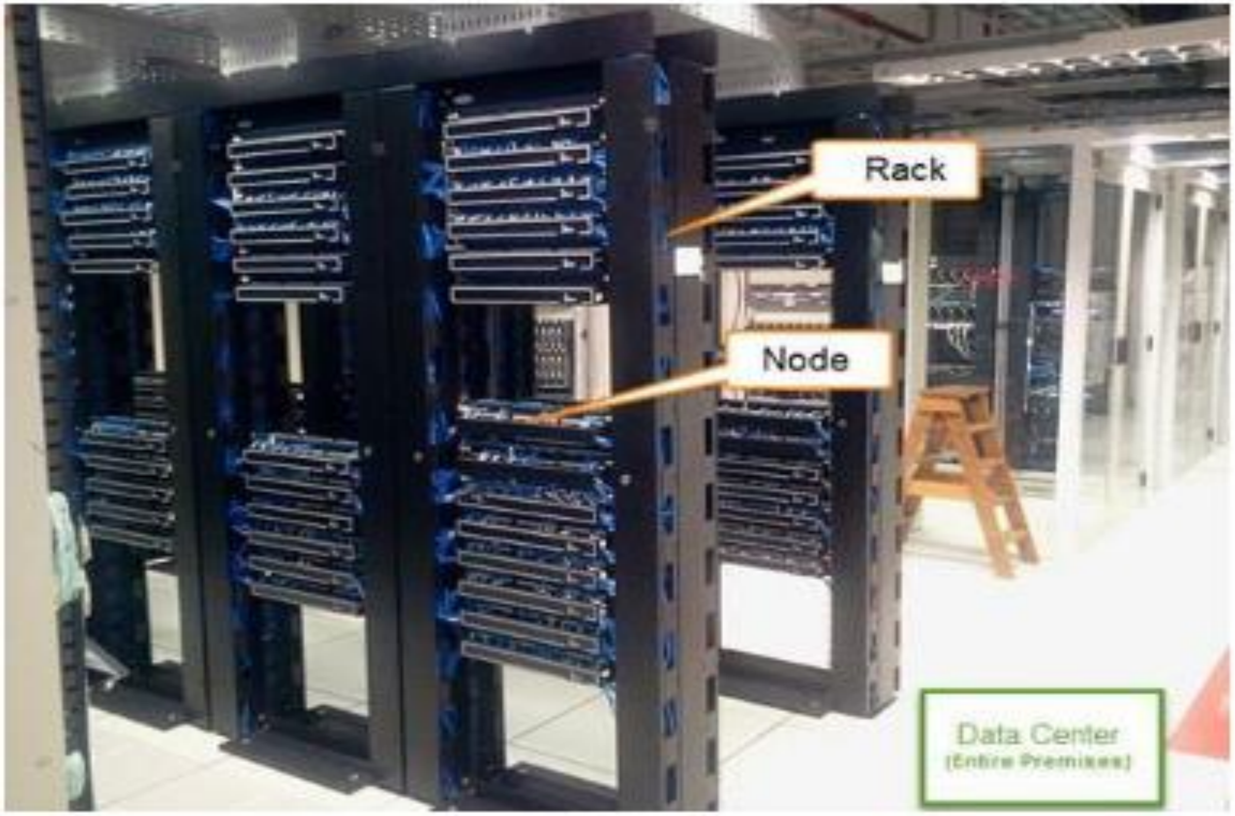
Hadoop Disadvantages

Hadoop Cluster

- In general, a computer cluster is a collection of various computers that work collectively as a single system.
- “**A hadoop cluster is a collection of independent components connected through a dedicated network to work as a single centralized data processing resource**”.
- “A hadoop cluster can be referred to as a computational computer cluster for storing and analyzing big data (structured, semi-structured and unstructured) in a distributed environment”.
- “A computational computer cluster that distributes data analysis workload across various cluster nodes that work collectively to process the data in parallel”.

Hadoop Cluster Architecture

- A hadoop cluster architecture consists of a data centre, rack and the node that actually executes the jobs.
- **Data centre consists of the racks and racks consists of nodes.**



- A medium to large cluster consists of a two or three level hadoop cluster architecture that is built with **rack mounted servers**.
- **Every rack of servers is interconnected through 1 gigabyte of Ethernet (1 GigE).**
- Each **rack level switch** in a hadoop cluster is connected to a cluster level switch which are in turn connected to other cluster level switches or they uplink to other switching infrastructure.

Components of a Hadoop Cluster

Hadoop cluster consists of three components -

Master Node –

- Master node in a hadoop cluster is responsible for storing data in HDFS and executing parallel computation the stored data using MapReduce.
- Master Node has 3 nodes –
  - – NameNode,
  - – Secondary NameNode and
  - – JobTracker.

Components of a Hadoop Cluster

- JobTracker monitors the parallel processing of data using MapReduce while the NameNode handles the data storage function with HDFS.
- NameNode keeps a track of all the information on files (i.e. the metadata on files) such as the access time of the file, which user is accessing a file on current time and which file is saved in which hadoop cluster.
- The secondary NameNode keeps a backup of the NameNode data.

Components of a Hadoop Cluster

Slave/Worker Node-

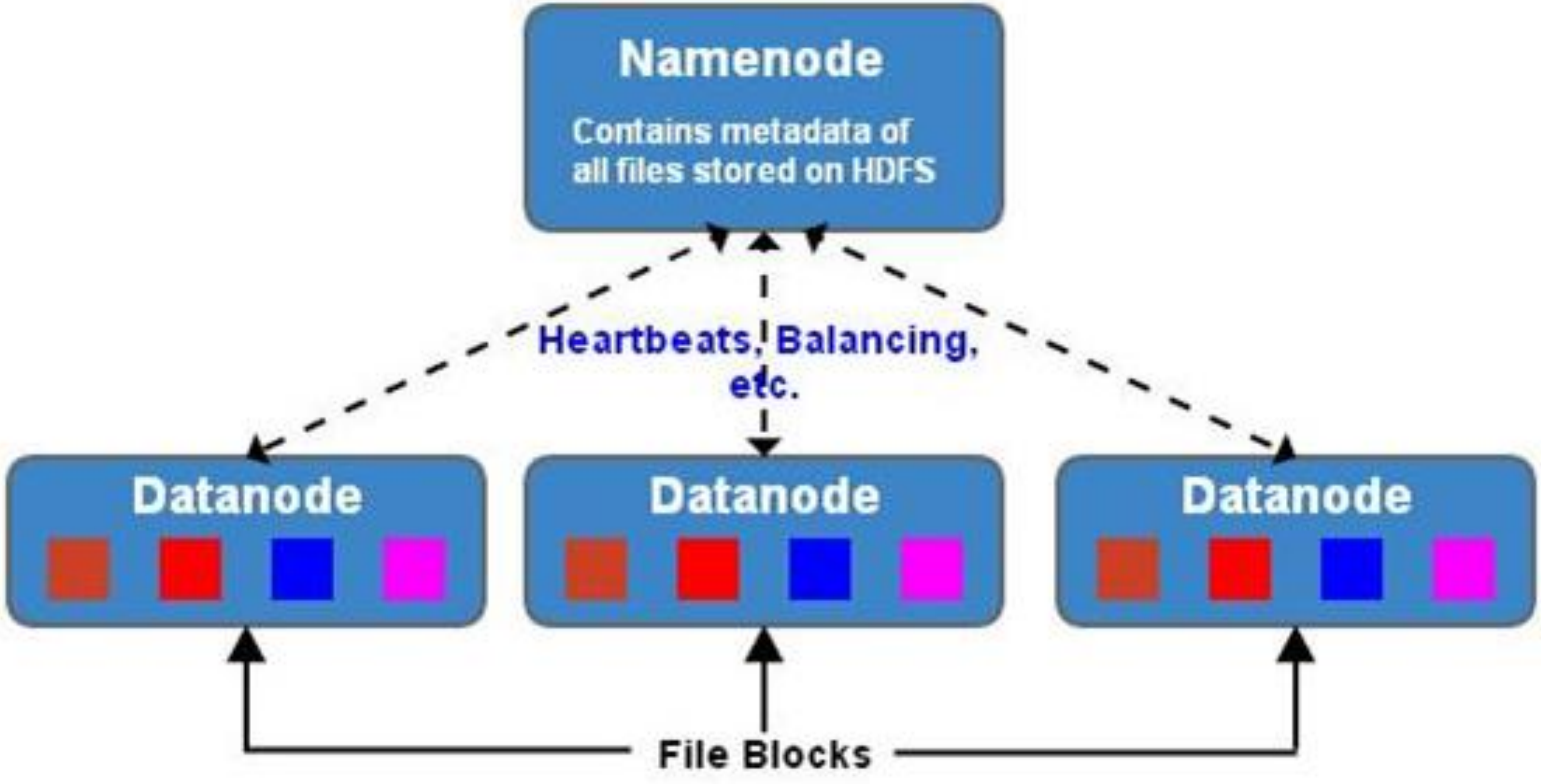
- This component in a hadoop cluster is responsible for storing the data and performing computations.
- Every slave/worker node runs both a TaskTracker and a DataNode service to communicate with the Master node in the cluster.
- The DataNode service is secondary to the NameNode and the TaskTracker service is secondary to the JobTracker.

Components of a Hadoop Cluster

Client Nodes –

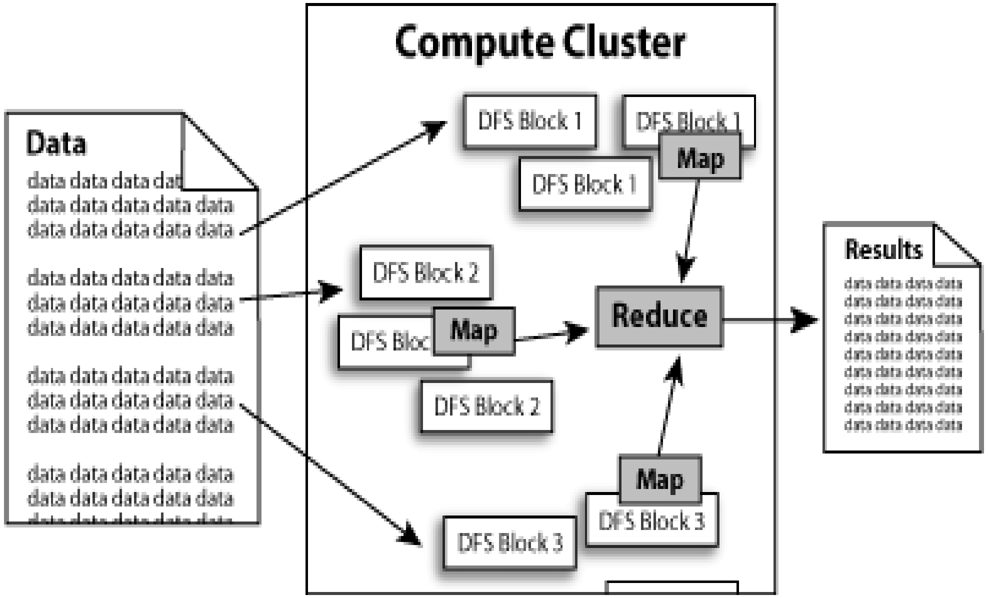
- Client node has hadoop installed with all the required cluster configuration settings and is responsible for loading all the data into the hadoop cluster.
- Client node submits mapreduce jobs describing on how data needs to be processed and then the output is retrieved by the client node once the job processing is completed.

Components of a Hadoop Cluster


Namenode
Contains metadata of all files stored on HDFS
Heartbeats, Balancing, etc.
Datanode
Datanode
Datanode
File Blocks

Applications of Hadoop

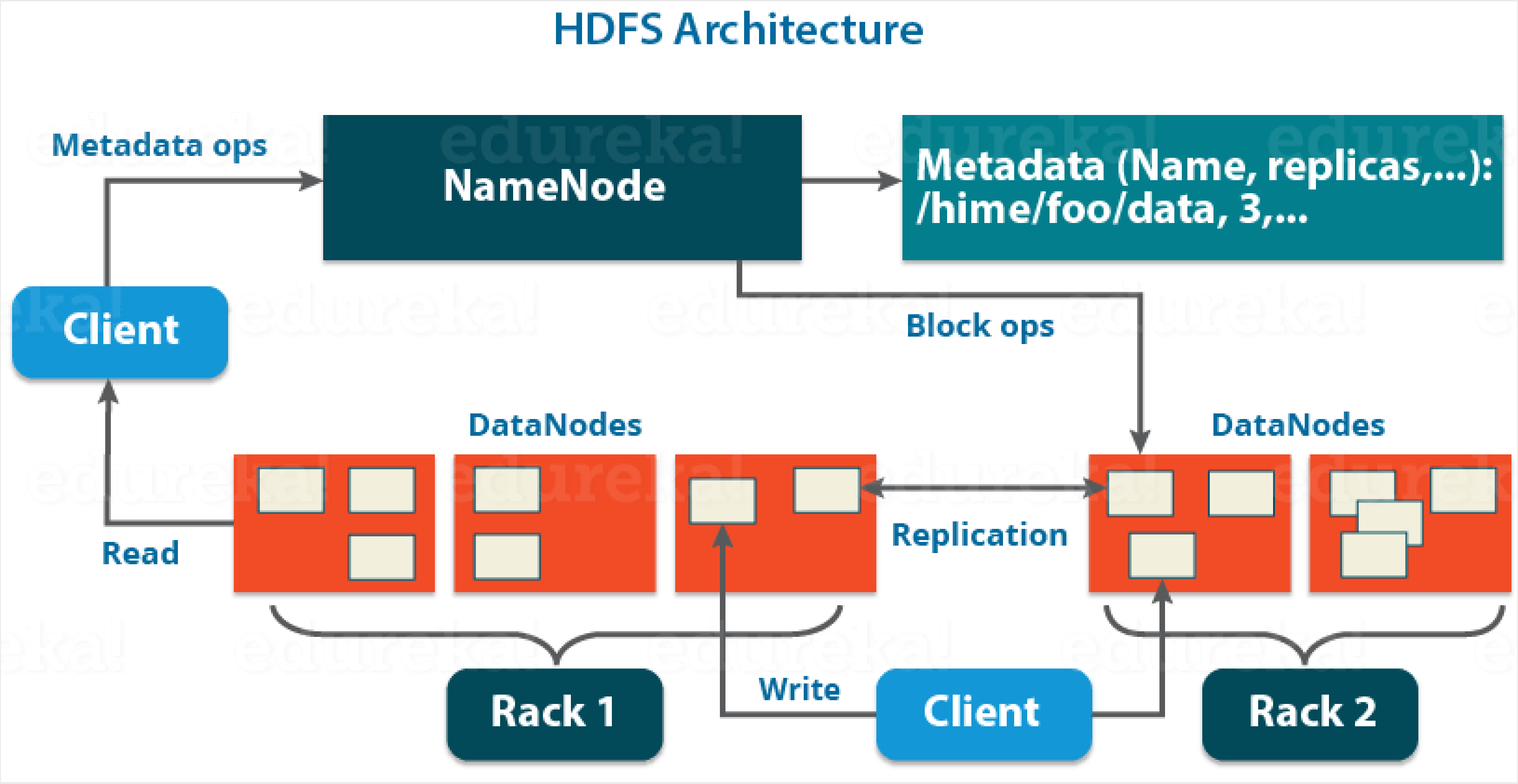
- Hadoop implements Google's MapReduce, using HDFS
- MapReduce divides applications into many small blocks of work.
- HDFS creates multiple replicas of data blocks for reliability, placing them on compute nodes around the cluster.
- MapReduce can then process the data where it is located.
- Hadoop 's target is to run on clusters of the order of 10,000-nodes.



More Hadoop Applications

- Adknowledge - to build the recommender system for behavioral targeting, plus other clickstream analytics; clusters vary from 50 to 200 nodes, mostly on EC2.
- Contextweb - to store ad serving log and use it as a source for Ad optimizations/ Analytics/reporting/machine learning; 23 machine cluster with 184 cores and about 35TB raw storage. Each (commodity) node has 8 cores, 8GB RAM and 1.7 TB of storage.
- Cornell University Web Lab: Generating web graphs on 100 nodes (dual 2.4GHz Xeon Processor, 2 GB RAM, 72GB Hard Drive)
- NetSeer - Up to 1000 instances on Amazon EC2 ; Data storage in Amazon S3; Used for crawling, processing, serving and log analysis
- The New York Times : Large scale image conversions ; EC2 to run hadoop on a large virtual cluster
- Powerset / Microsoft - Natural Language Search; up to 400 instances on Amazon EC2 ; data storage in Amazon S3

HDFS Architecture



- HDFS follows the master-slave architecture and it has the following elements.

Features of HDFS

- It is suitable for the distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of namenode and datanode help users to easily check the status of cluster.
- Streaming access to file system data.
- HDFS provides file permissions and authentication.
- Single Namespace for entire cluster
- Data Coherency
  - Write-once-read-many access model
  - Client can only append to existing files
- **Files are broken up into blocks**
  - **Typically 64MB block size**
  - **Each block replicated on multiple DataNodes**
- Intelligent Client
  - Client can find location of blocks
  - Client accesses data directly from DataNode

What is cluster size in Hadoop?

A Hadoop cluster size is a set of metrics that defines storage and compute capabilities to run Hadoop workloads, namely :

- **Number of nodes**: number of Master nodes, number of Edge Nodes, number of Worker Nodes.
- **Configuration of each type node**: number of cores per node, RAM and Disk Volume.

Advantages of a Hadoop Cluster Setup

- As big data grows exponentially, parallel processing capabilities of a Hadoop cluster help in increasing the speed of analysis process. However, the processing power of a hadoop cluster might become inadequate with increasing volume of data.
- In such scenarios, hadoop clusters can scaled out easily to keep up with speed of analysis by adding extra cluster nodes without having to make modifications to the application logic.

Advantages of a Hadoop Cluster Setup

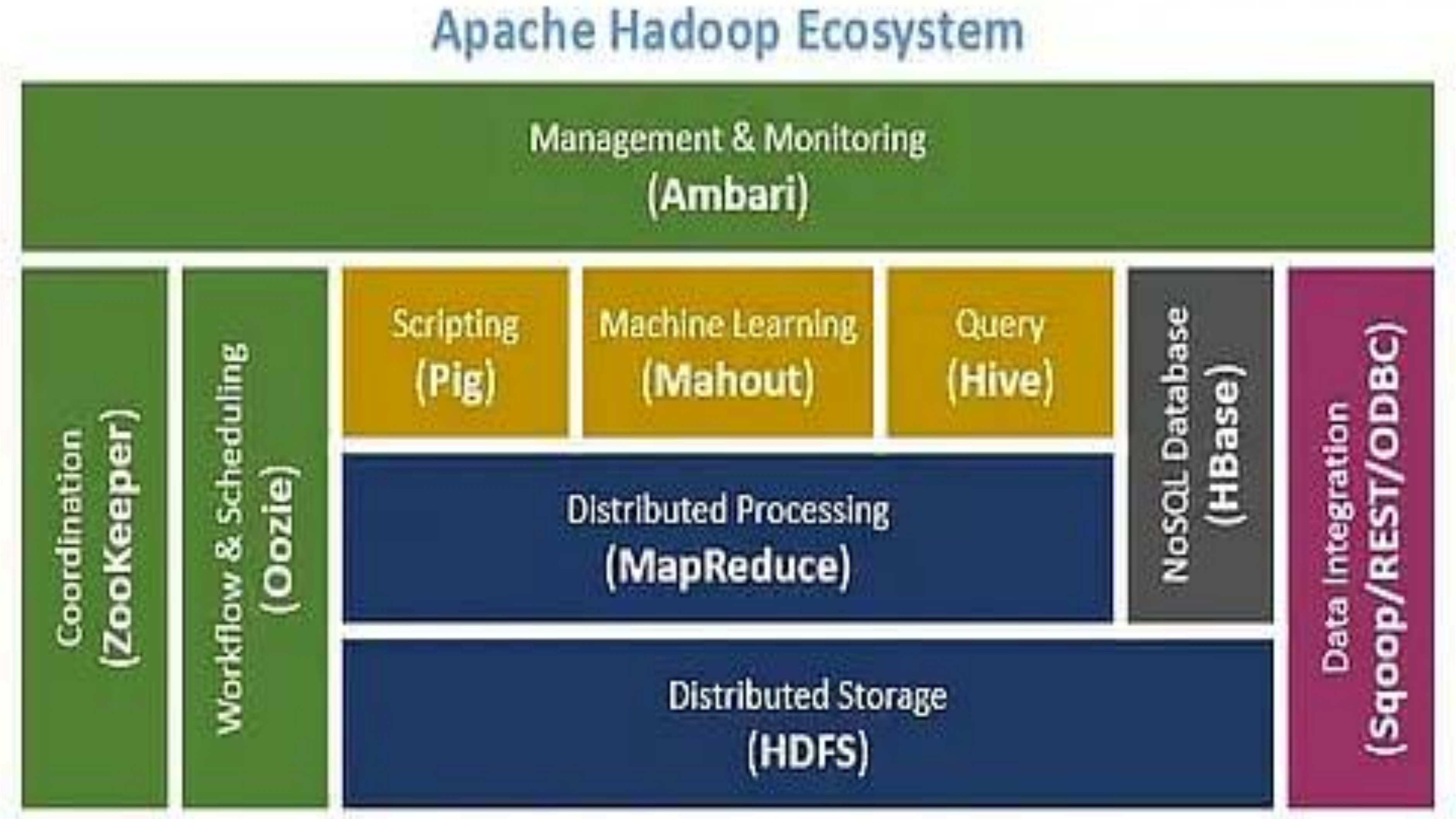
- Hadoop cluster setup is inexpensive as they are held down by cheap commodity hardware. Any organization can setup a powerful hadoop cluster without having to spend on expensive server hardware.
- Hadoop clusters are resilient to failure, meaning whenever data is sent to a particular node for analysis, it is also replicated to other nodes on the hadoop cluster. If the node fails then the replicated copy of the data present on the other node in the cluster can be used for analysis.

Challenges of a Hadoop Cluster

- **Issue with small files -** Hadoop struggles with large volumes of small files - smaller than the Hadoop block size of 128MB or 256MB by default. It wasn't designed to support big data in a scalable way. Instead, Hadoop works well when there are a small number of large files. Ultimately when you increase the volume of small files, it overloads the Namenode as it stores namespace for the system.
- **High processing overhead -** reading and writing operations in Hadoop can get very expensive quickly especially when processing large amounts of data. This all comes down to Hadoop's inability to do in-memory processing and instead data is read and written from and to the disk.

Challenges of a Hadoop Cluster

- **Only batch processing is supported -** Hadoop is built for small volumes of large files in batches. This goes back to the way data is collected and stored which all has to be done before processing starts. What this ultimately means is that streaming data is not supported and it cannot do real-time processing with low latency.
- **Iterative Processing -** Hadoop has a data flow structure is set-up in sequential stages which makes it impossible to do iterative processing or use for ML.


Apache Hadoop Ecosystem
Management & Monitoring
(Ambari)
Coordination
(ZooKeeper)
Workflow & Scheduling
(Oozie)
Scripting
(Pig)
Machine Learning
(Mahout)
Query
(Hive)
NoSQL Database
(HBase)
Data Integration
(Sqoop/REST/ODBC)
Distributed Processing
(MapReduce)
Distributed Storage
(HDFS)

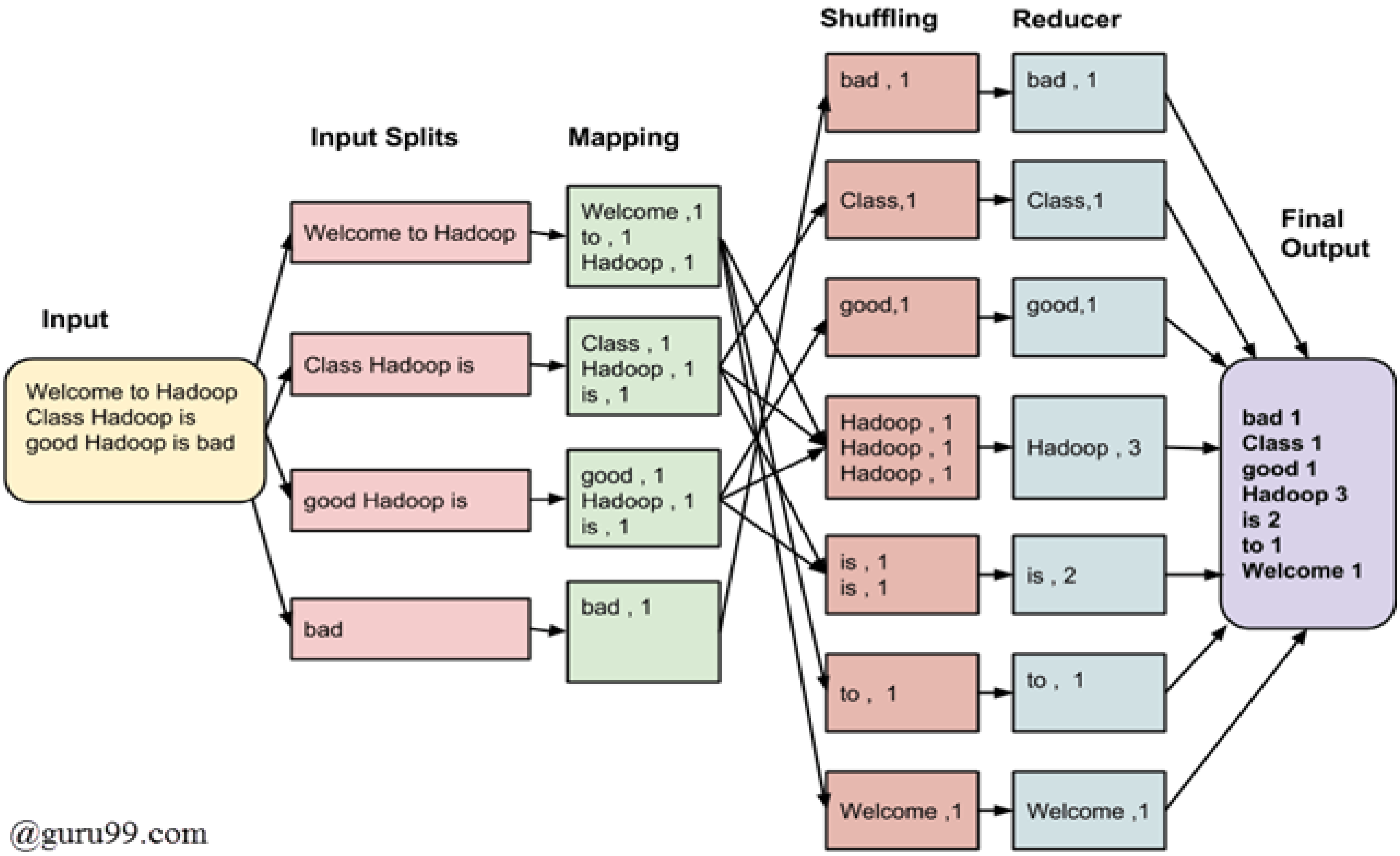
MapReduce Paradigm

- **Programming model developed at Google**
- Sort/merge based distributed computing
- Initially, it was intended for their internal search/indexing application, but now used extensively by more organizations **(e.g., Yahoo, Amazon.com, IBM, etc.)**
- It is functional style programming (e.g., LISP) that is naturally parallelizable across a large cluster of workstations or PCS.
- **The underlying system takes care of the partitioning of the input data, scheduling the program's execution across several machines, handling machine failures, and managing required inter-machine communication. (This is the key for Hadoop's success)**

How does MapReduce work?

- The run time partitions the input and provides it to different Map instances;
- Map (key, value) → (key’, value’)
- The run time collects the (key’, value’) pairs and distributes them to several Reduce functions so that each Reduce function gets the pairs with the same key’.
- **Each Reduce produces a single (or zero) file output.**
- **Map and Reduce are user written functions**

Example of word count problem using MapReduce



@guru99.com

Thanks.