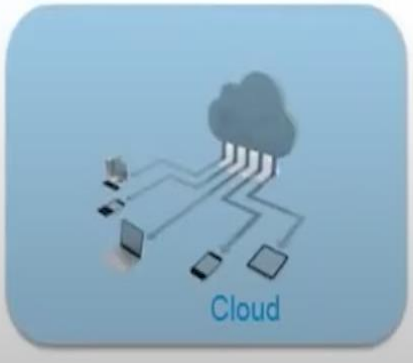
1. **Evolution of Technology**
2. IOT
3. Social Media
4. Data evolved to Big Data

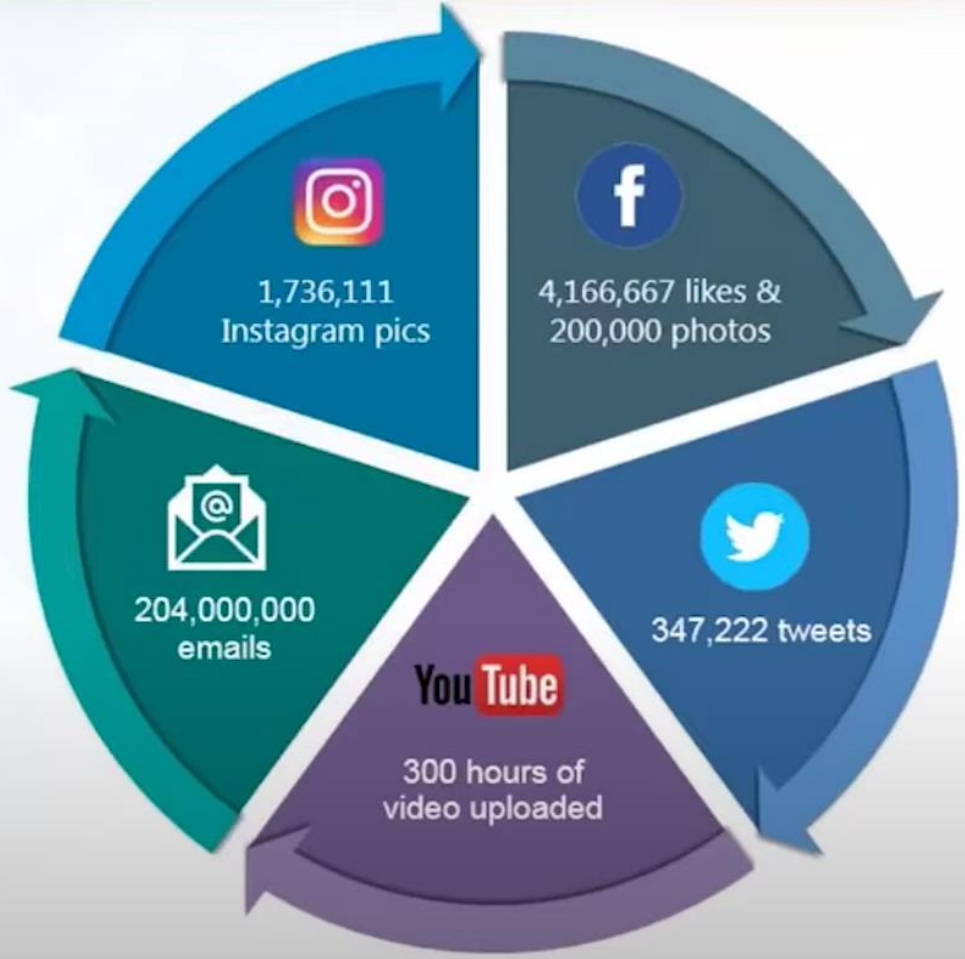


Ref:Edureka

1. Evolution of Technology
2. IOT
3. Social Media
4. Data evolved to Big Data

Ref:Edureka

1. Evolution of Technology
2. IOT
3. **Social Media**
4. Data evolved to Big Data

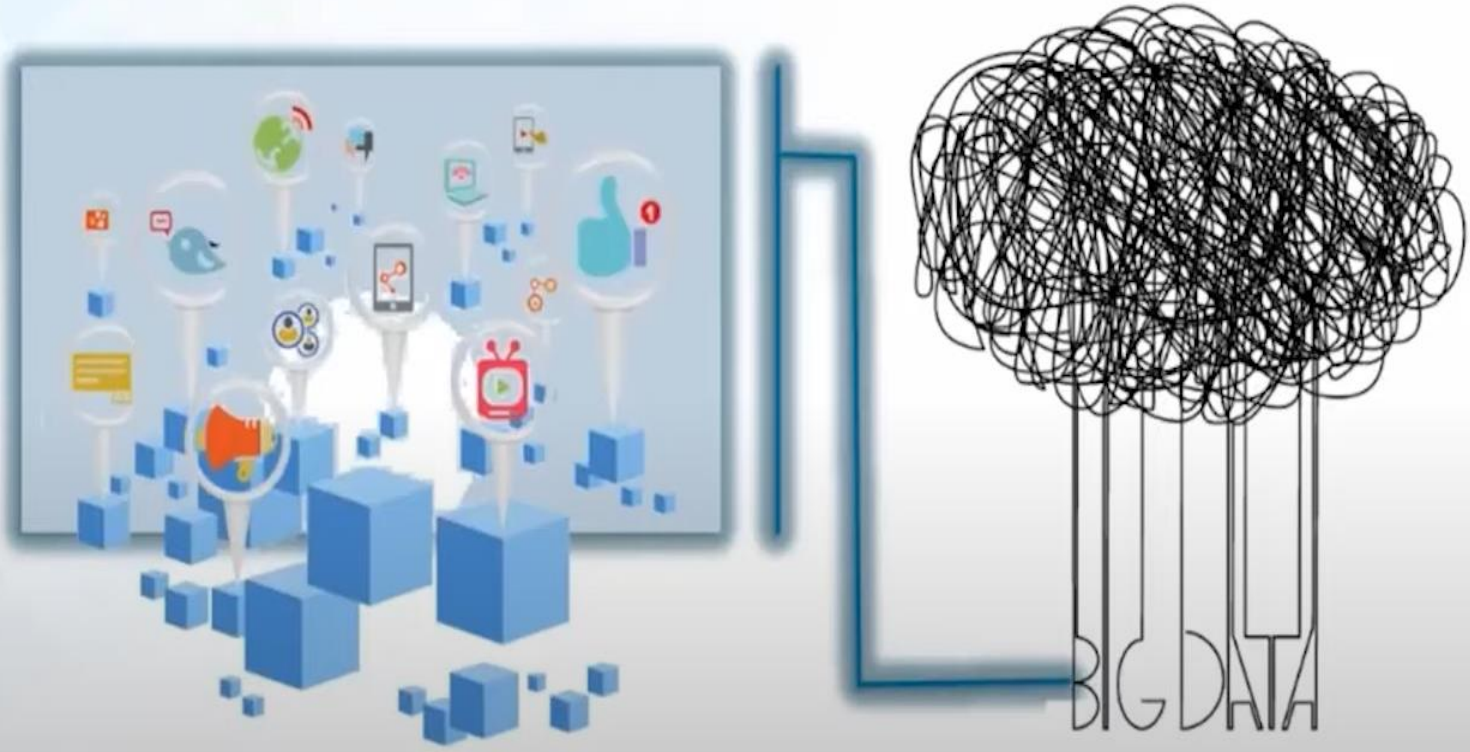


Social Media

1. Evolution of Technology
2. IOT
3. Social Media
4. Other Factors

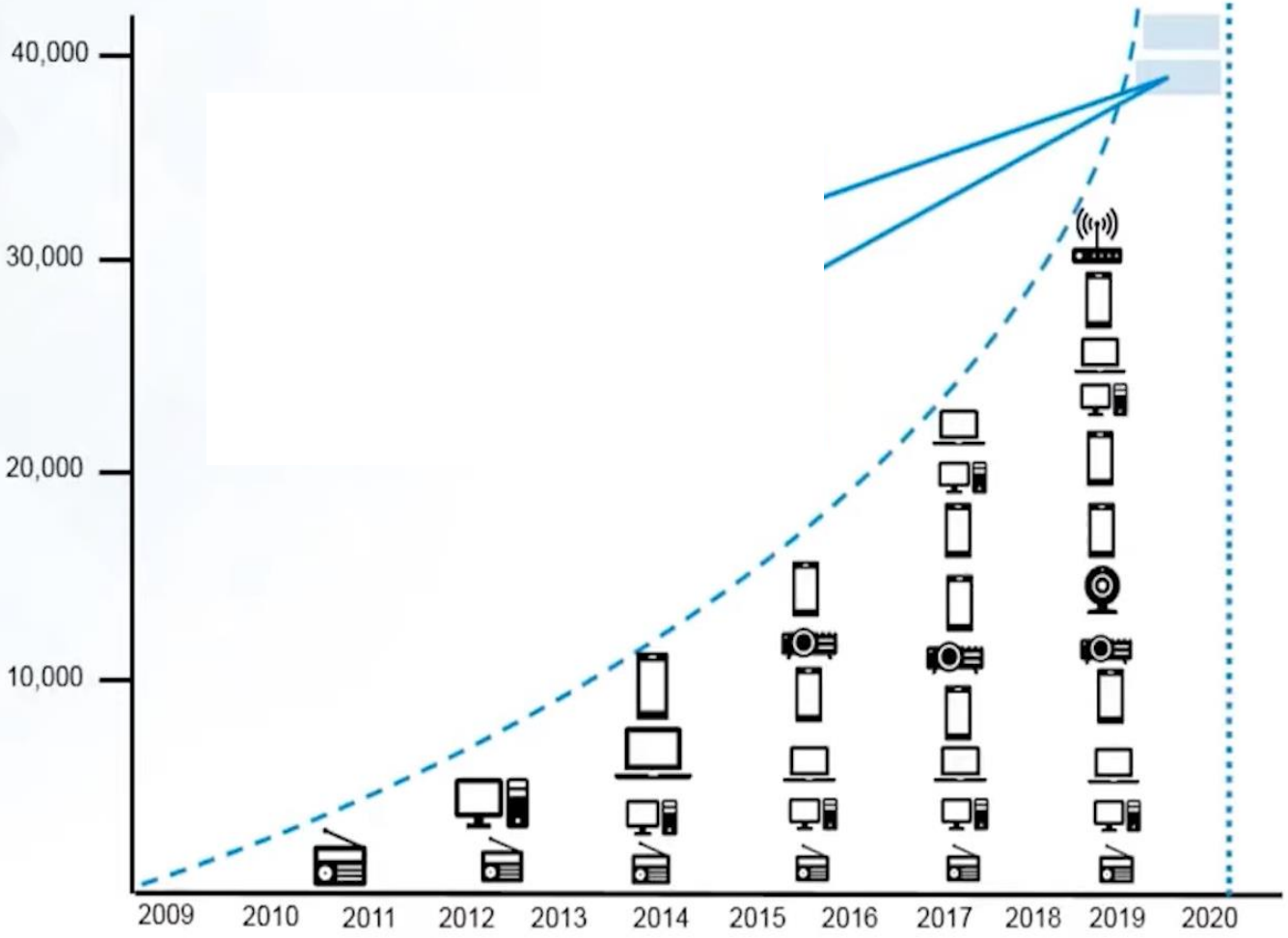
Ref:Edureka

Big data is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database system tools or traditional data processing applications



Ref:Edureka

1 **Volume**

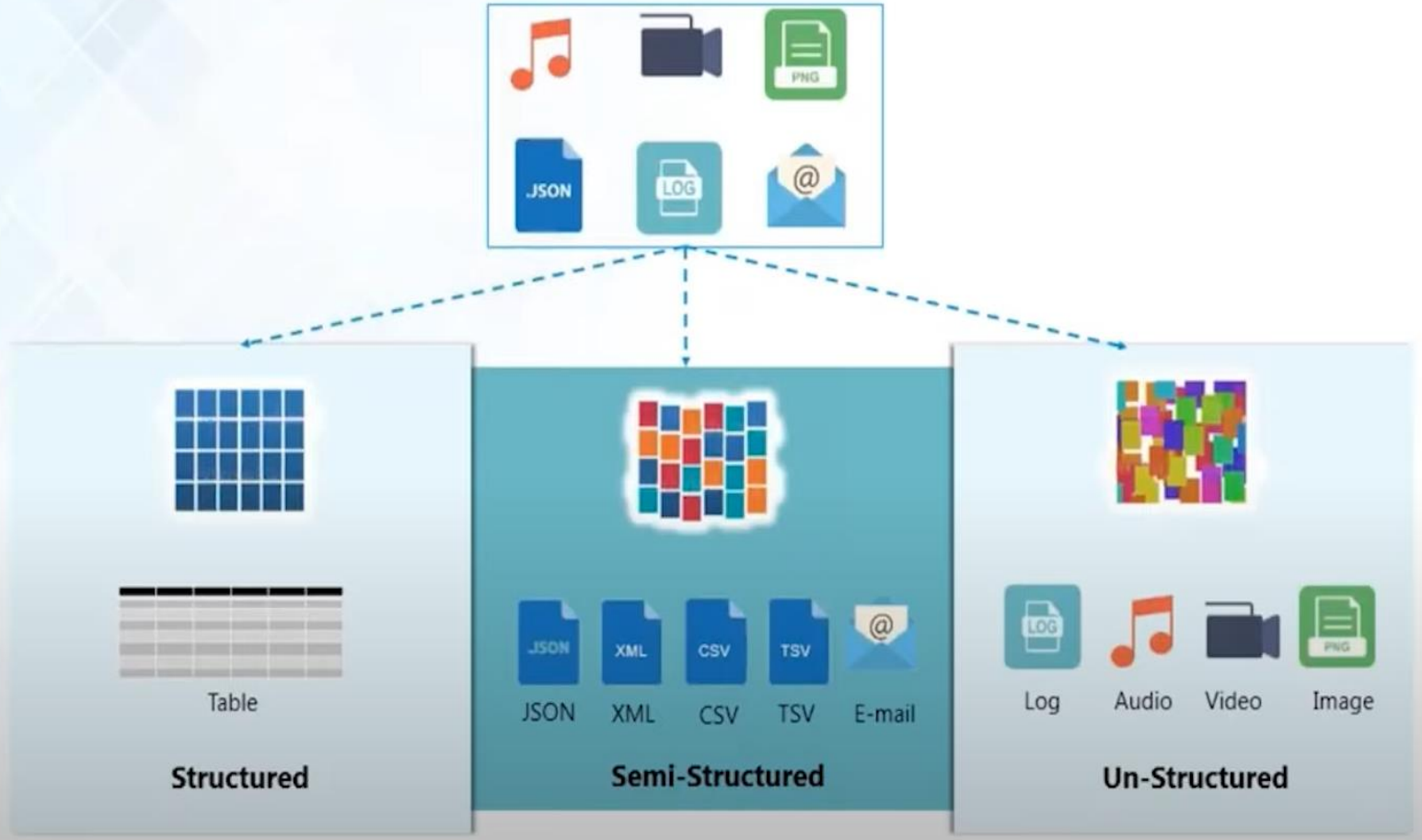


Ref:Edureka

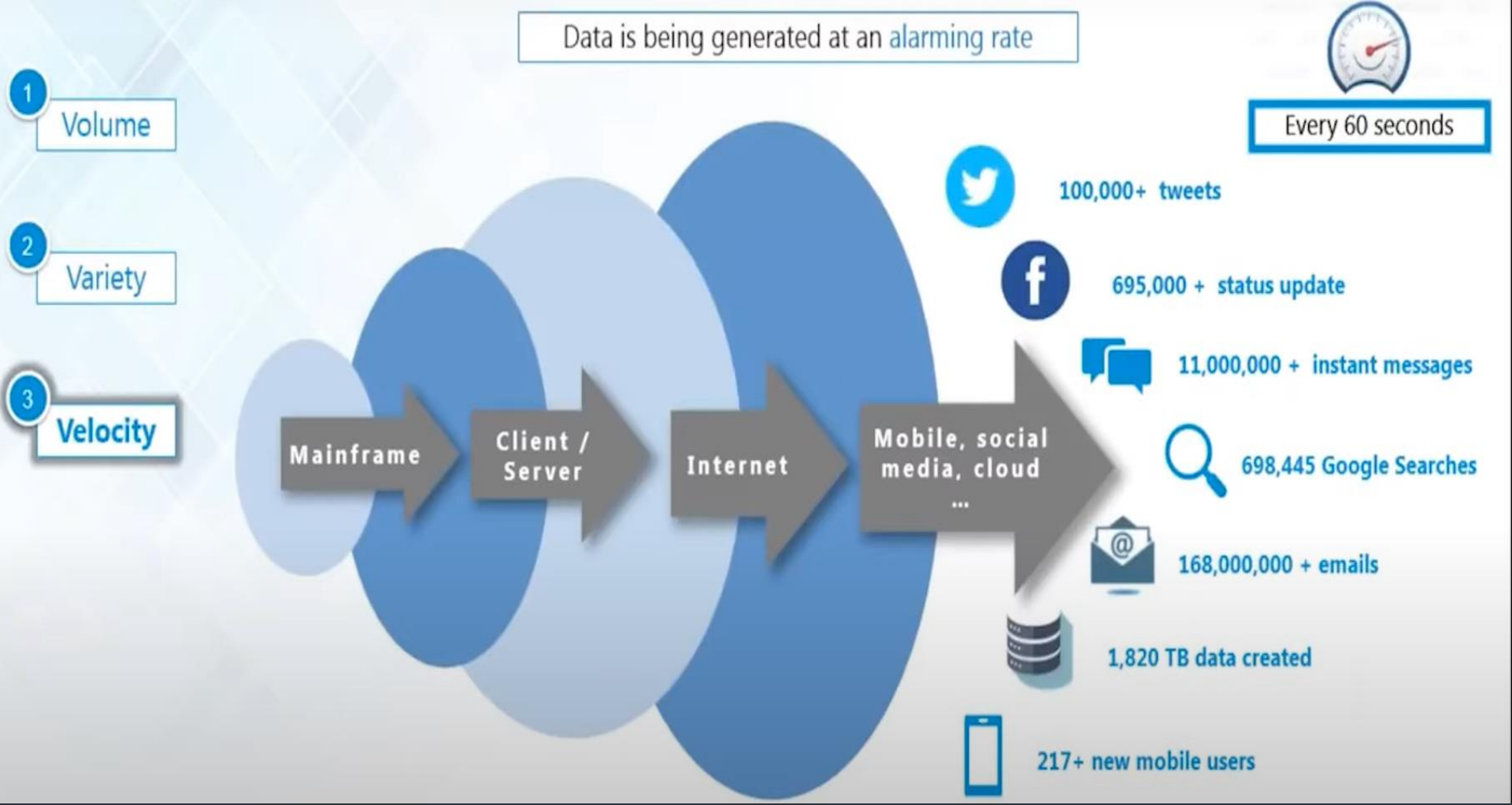
Different kinds of data is being generated from various sources

1. Volume

2. **Variety**



Ref:Edureka


Data is being generated at an alarming rate
1
Volume
2
Variety
3
Velocity
Mainframe
Client / Server
Internet
Mobile, social media, cloud ...
Every 60 seconds
100,000+ tweets
695,000 + status update
11,000,000 + instant messages
698,445 Google Searches
168,000,000 + emails
1,820 TB data created
217+ new mobile users

Ref:Edureka

Mechanism to bring the correct meaning out of the data

1. Volume
2. Variety
3. Velocity
4. **Value**

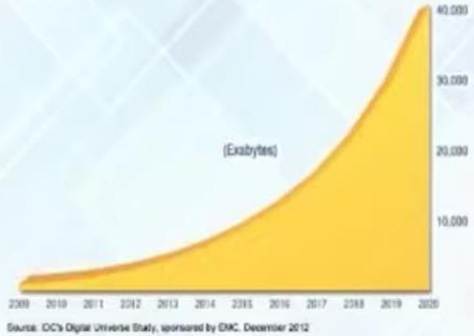
Ref:Edureka

1. Volume
2. Variety
3. Velocity
4. Value
5. **Veracity**

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

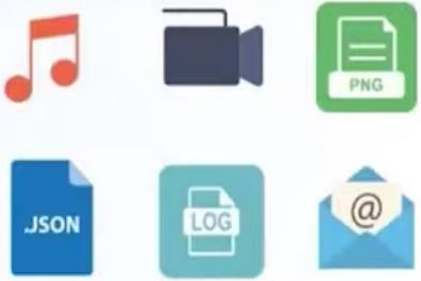
Uncertainty and inconsistencies in the data

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Volume



Different kinds of data is being generated from various sources

Variety

Data is being generated at an alarming rate

Velocity

Mechanism to bring the correct meaning out of the data

Value

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

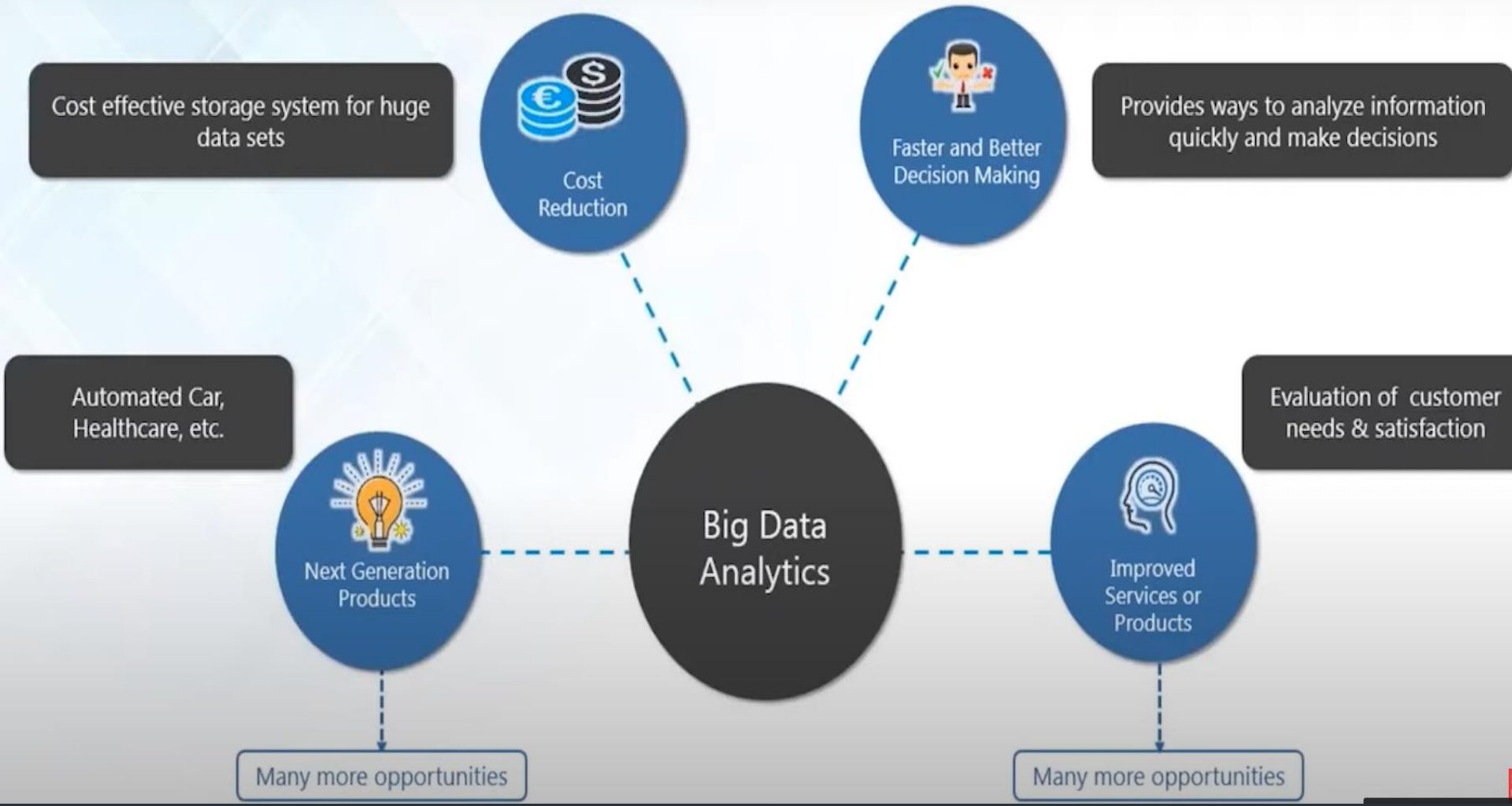
Uncertainty and inconsistencies in the data

Veracity

. . . . . .

V's associated with Big Data may grow with time

Ref:Edureka


Cost effective storage system for huge data sets
Cost Reduction
Faster and Better Decision Making
Provides ways to analyze information quickly and make decisions
Automated Car, Healthcare, etc.
Next Generation Products
Big Data Analytics
Improved Services or Products
Evaluation of customer needs & satisfaction
Many more opportunities
Many more opportunities

Ref:Edureka

Data was collected in 1 Month

Earlier

Data is collected in 15 Minutes

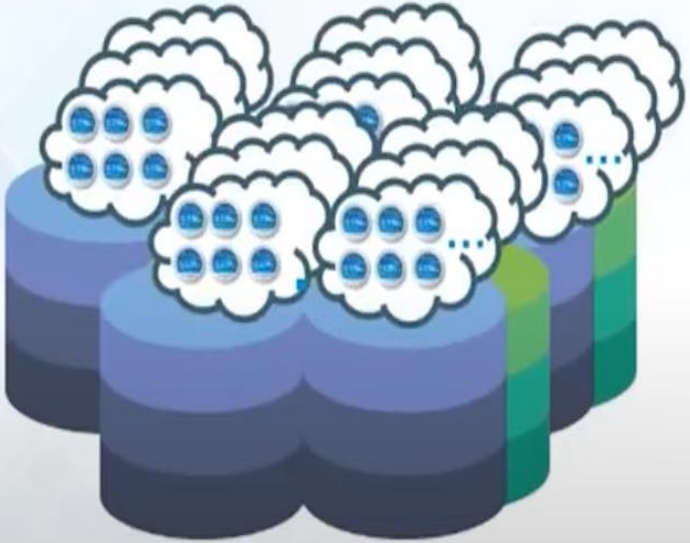
Now

Managing the large volume and velocity of information generated by short-interval reads of smart meter data can overwhelm existing IT resources

96 million reads per day for every million meters

Big Data generated by Smart Meter

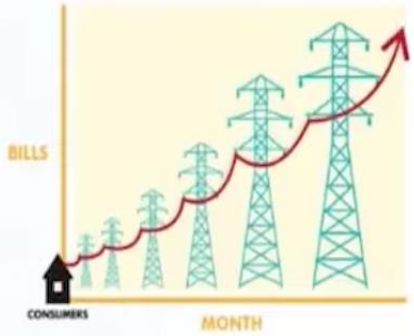
To manage and use this information to gain insight, utility companies must be capable of high-volume data management and advanced analytics designed to transform data into actionable insights.



Store

Analyze

Before analyzing Big Data



Energy utilization and billing has increased

After analyzing Big Data

During peak-load the users require more energy

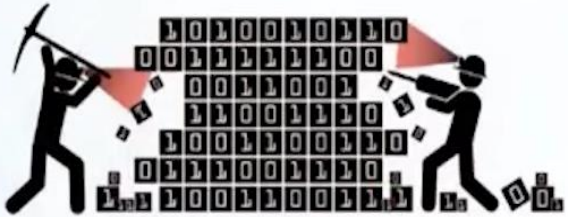
During off-peak times the users required less energy

Time-of-use pricing encourages cost-savvy retail like industrial heavy machines to be used at off-peak times

Ref:Edureka

IBM offers an integrated suite of products designed to enable IT to leverage big data in a variety of ways that can contribute to the success of energy companies

Data Analysis



Data Mining

Data Warehousing

User Data Security

Reporting

IBM Solution

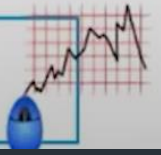
1. Managing smart meter data

2. Monitoring the distribution grid

3. Optimizing unit commitment

4. Optimizing energy trading

5. Forecasting and scheduling loads



Ref:Edureka

1 Making Smarter and More Efficient Organisations

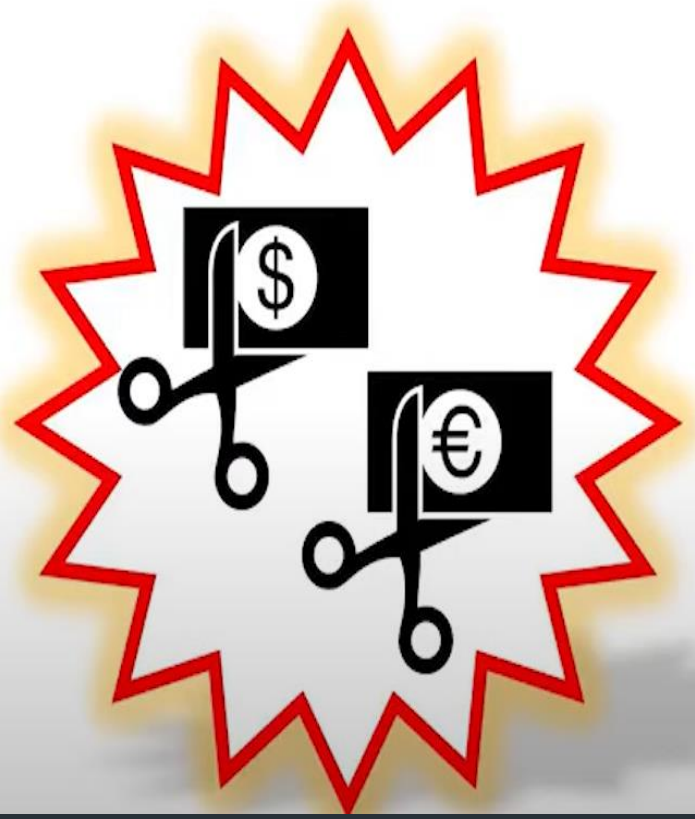
New York Police Department is utilizing data patterns, scientific analysis, and technological tools to prevent the occurrence of crime

2 Optimize Business Operations by analysing customer behaviour

- Analysing all the clicks of every visitor on a website
- Studying the paths leading them to buy products
- Customer Satisfaction

Amazon uses customer click-stream data and historical purchase data of more than 300 million customers and each user is shown customized results on customized web pages.

3 Cost Reduction



Parkland Hospital uses analytics and predictive modelling to identify high-risk patients and predict likely outcomes once patients are sent home. As a result, Parkland reduced 30-day readmissions for patients with heart failure, by 31 percent, saving $500,000 annually.

Ref:Edureka

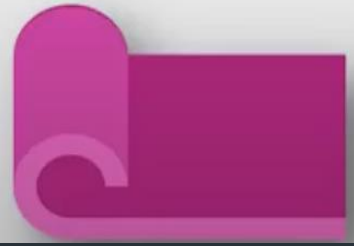
4 Next Generation Products

Big Data tools are used to operate Google's Self Driving Cars. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings.

Netflix launched the seasons of its TV show House of Cards based on the user reviews, ratings and viewership.

NETFLIX

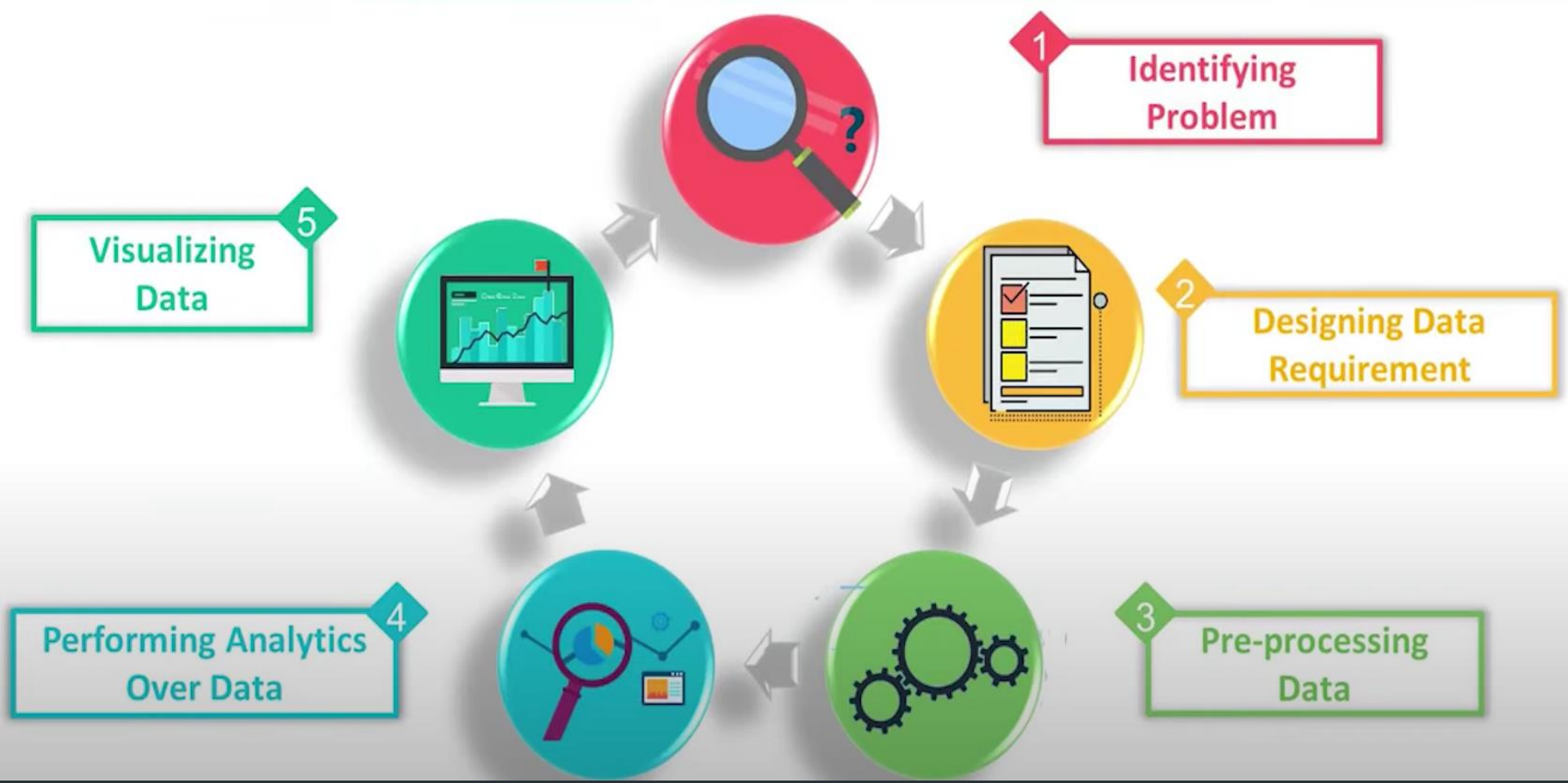
A smart yoga mat has sensors embedded in the mat will be able to provide feedback on your postures, score your practice, and even guide you through an at-home practice.



Ref:Edureka

"Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights"

Ref:Edureka


1
Identifying Problem
2
Designing Data Requirement
3
Pre-processing Data
4
Performing Analytics Over Data
5
Visualizing Data

Ref:Edureka

Types of Big Data Analytics

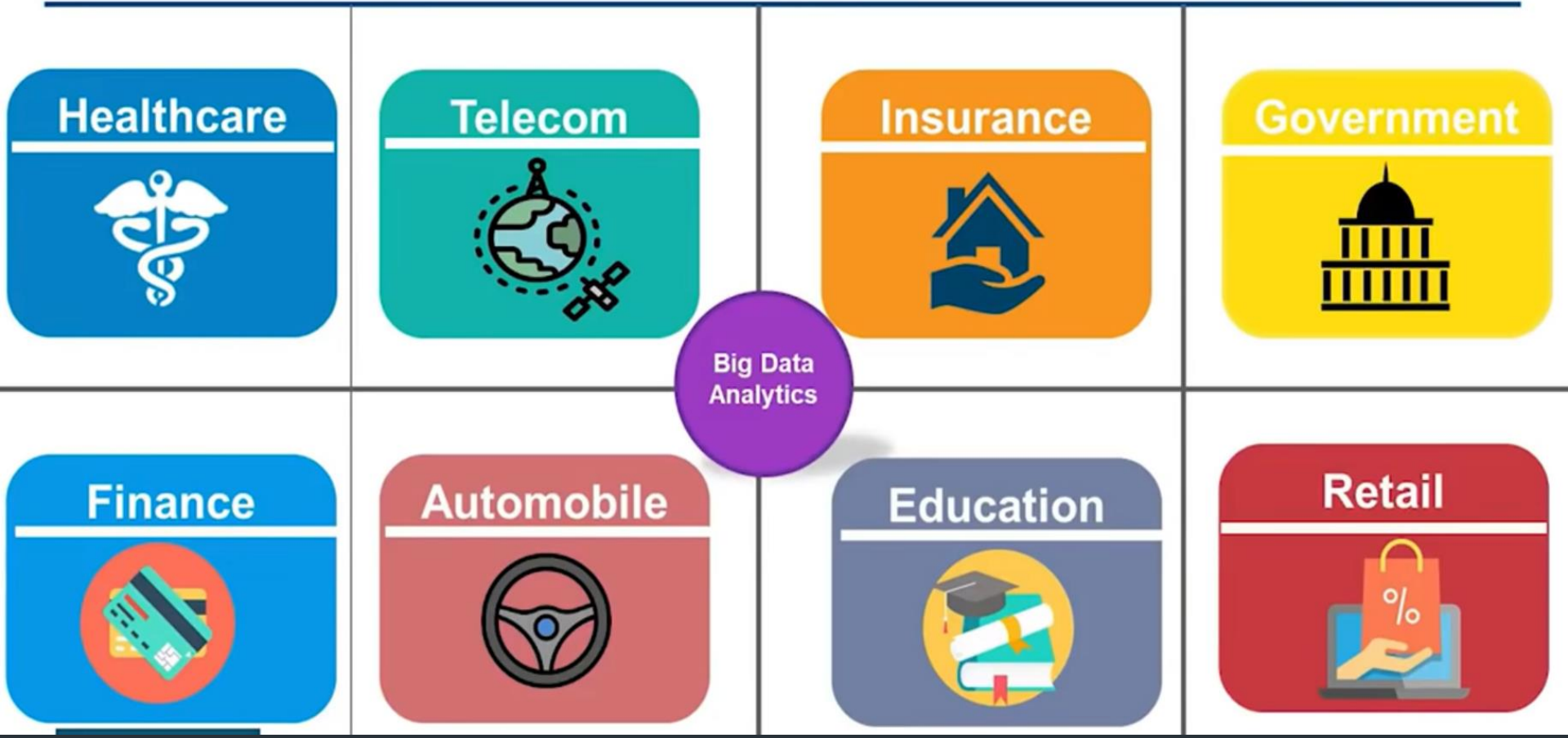
1. Descriptive Analysis
2. Predictive Analysis
3. Prescriptive Analysis
4. Diagnostic Analytics

What is happening now based on incoming data.

Google Analytics Tool is the best example for descriptive analysis. A business gets result from the web server through the tool which help understand what actually happened in the past and validate if a promotional campaign was successful or not based on basic parameters like page views.

Ref:Edureka

Domains using Big Data Analytics



Ref:Edureka

Use Case 1 - Starbucks

Starbucks uses behavioural analytics to cater to its customers

Starbucks gather a lot of info about their customers' coffee-buying habits from their preferred drinks to what time of day they're usually ordering

The company directs exciting offers and coupons to their customers and ensures to maintain their interest

Use Case 2 – Procter & Gamble

P&G uses Market Basket Analysis and price optimization to optimize their products

Procter&Gamble

Market Basket Analysis, analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets"

P&G

The company uses simulation models and predictive analysis in order to create the best design for its products.

Restaurant Scenario

Scenario:
Bob has opened a small restaurant in his city

Traditional Scenario:

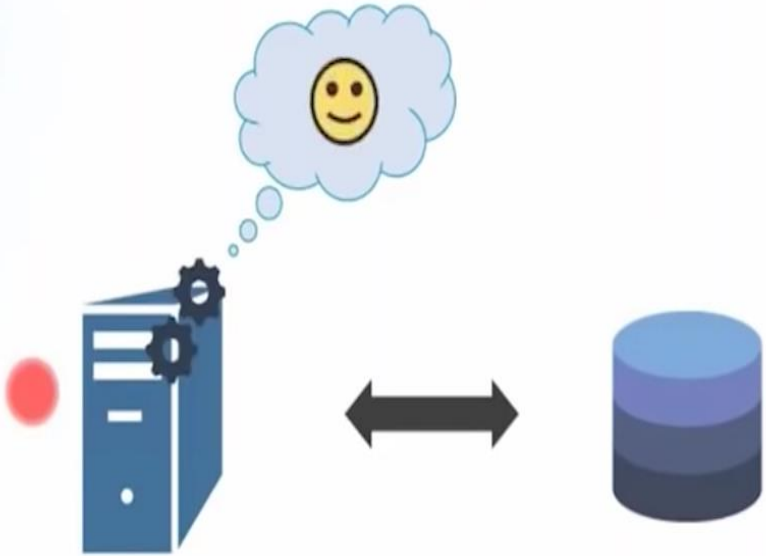
2 orders per hour

Single Cook

Food Shelf

Traditional Scenario:

Data is generated at a steady rate and is structured in nature



Traditional Processing System

RDBMS

Scenario 2:

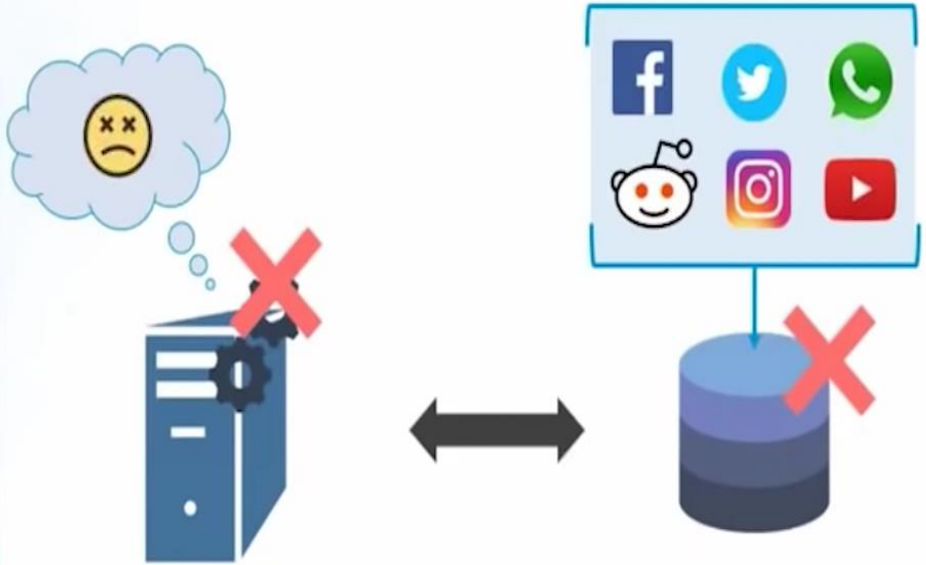
- They started taking Online orders
- 10 orders per hour

Single Cook
(Regular Computing System)

Food Shelf
(Data)

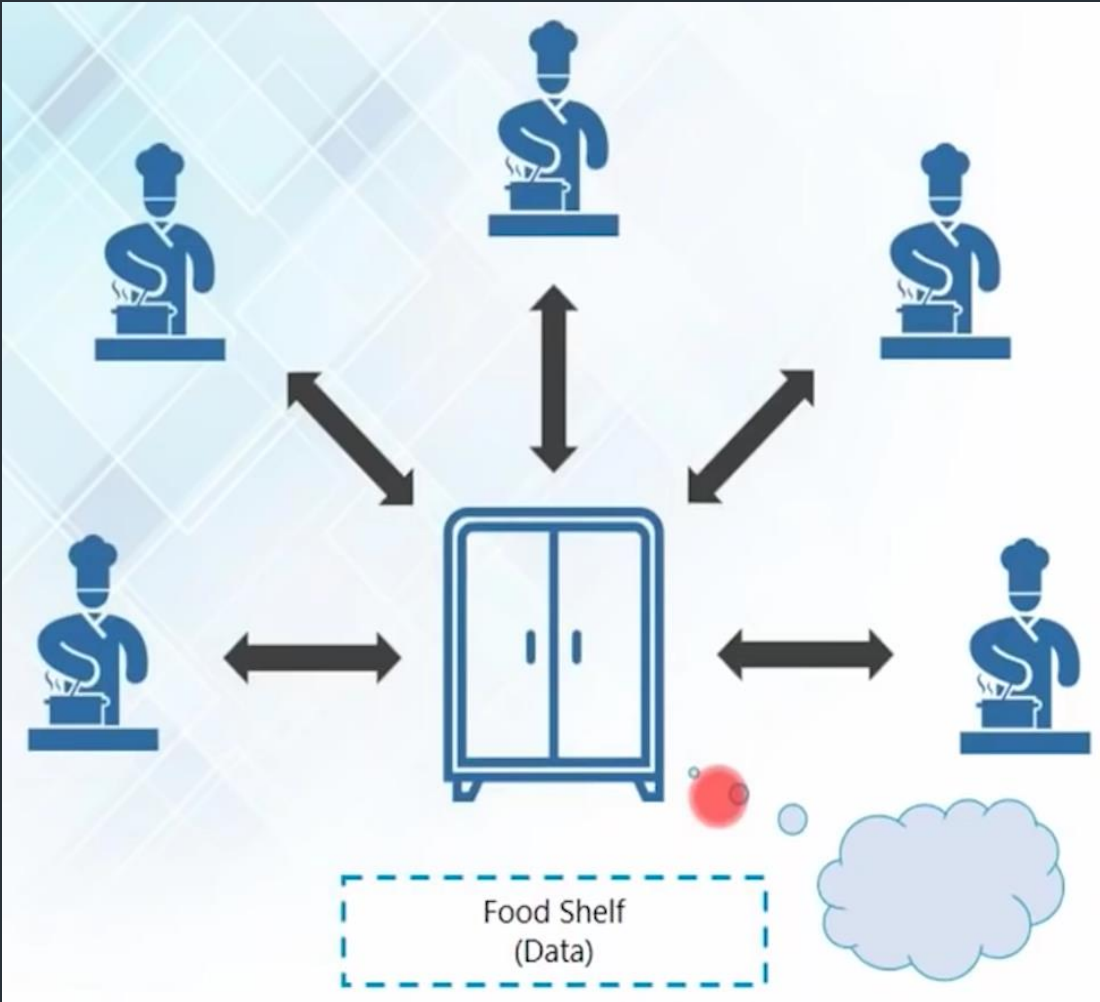
Big Data Scenario:

Heterogenous data is being generated at an alarming rate by multiple sources



Traditional Processing System

RDBMS

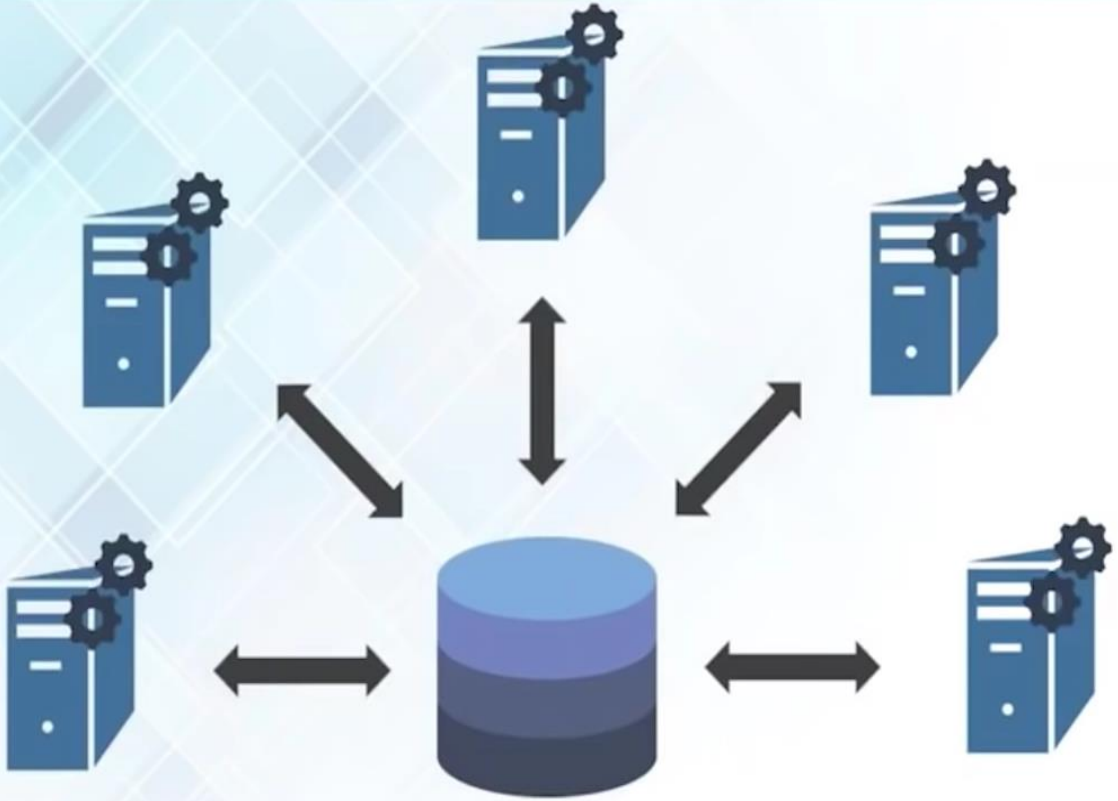


Scenario:

Multiple Cook cooking food

Issue:

Food Shelf becomes the BOTTLENECK



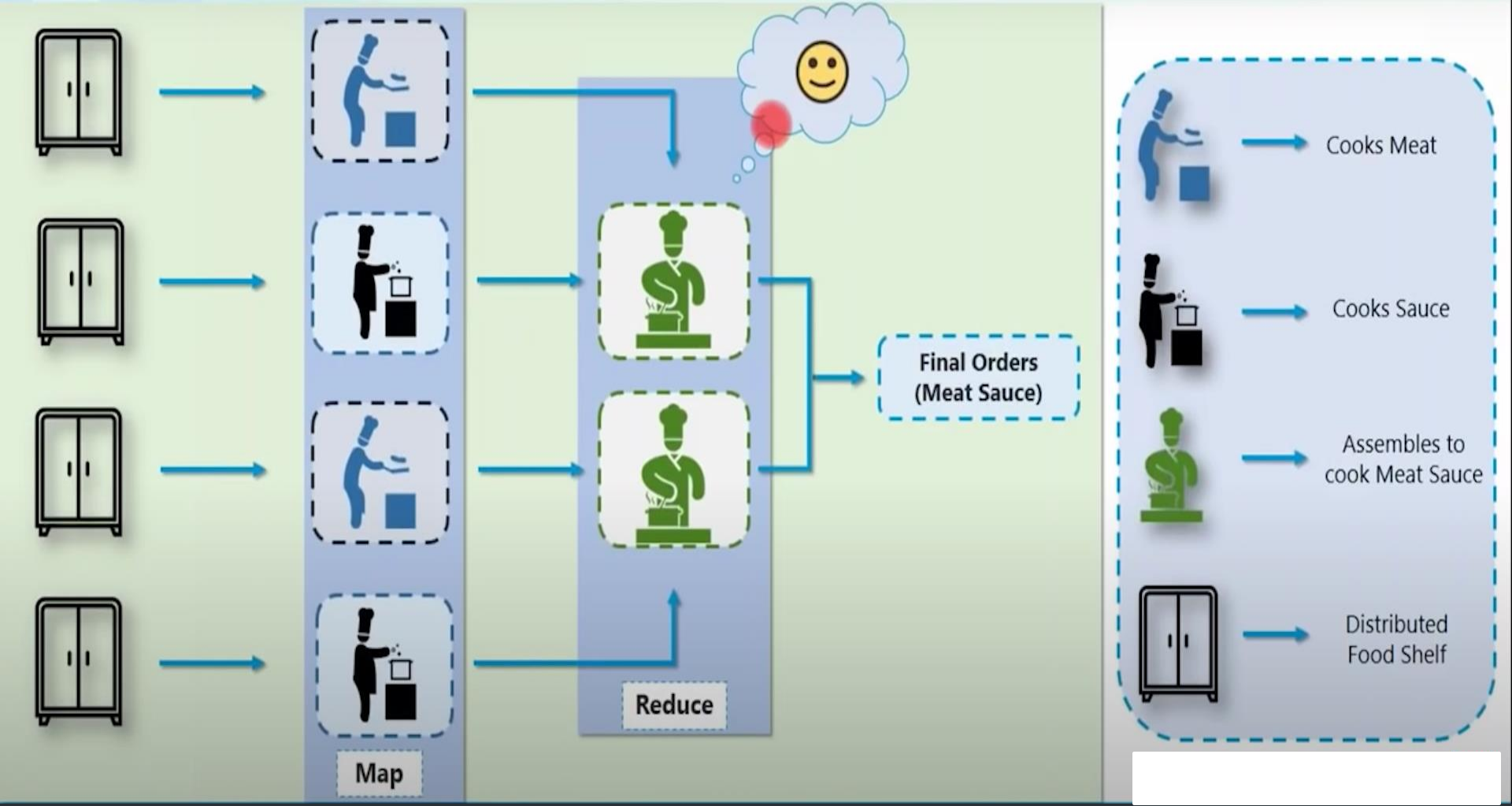
Scenario:

Multiple Processing Unit for data processing

Issue:

Bringing data to processing generated lots of Network overhead

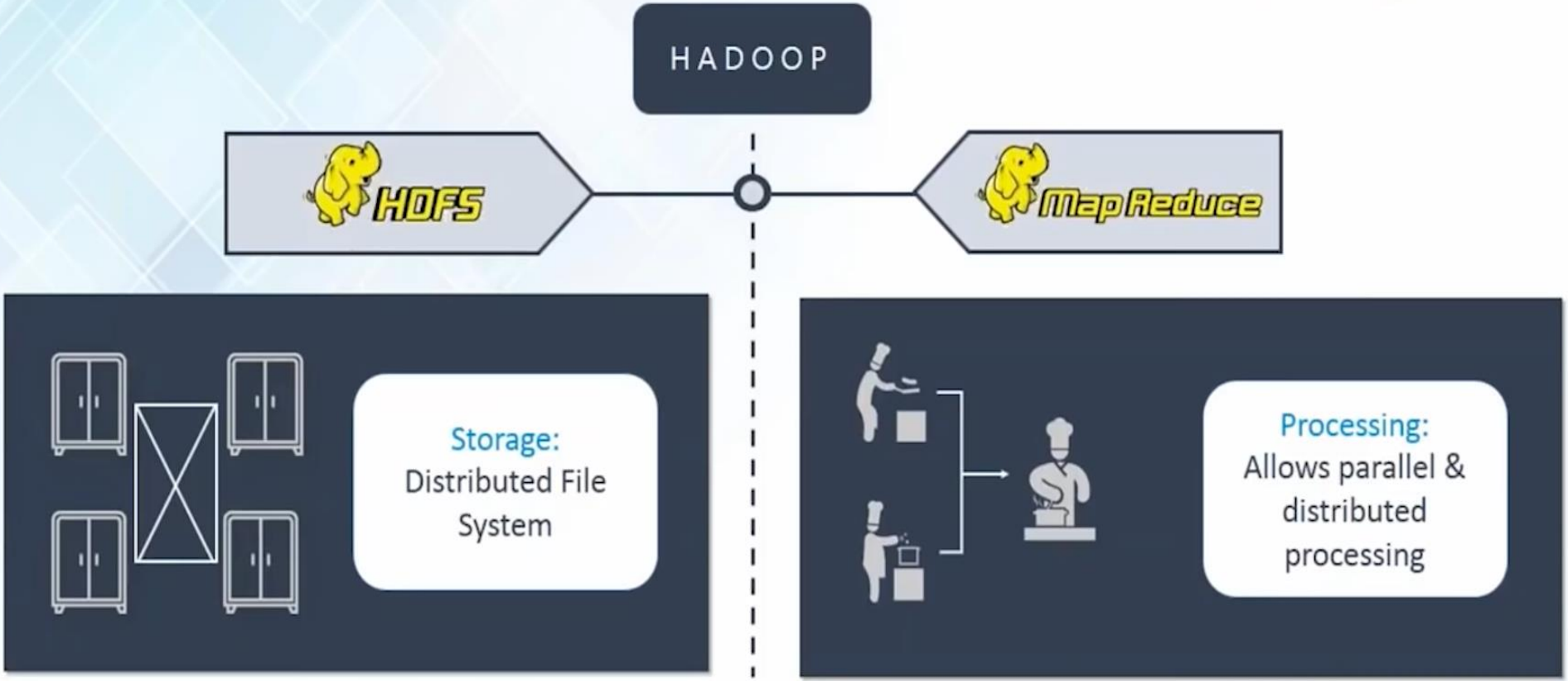
Data Warehouse


Map
Reduce
Final Orders
(Meat Sauce)
Cooks Meat
Cooks Sauce
Assembles to
cook Meat Sauce
Distributed
Food Shelf

Ref:Edureka

Apache Hadoop

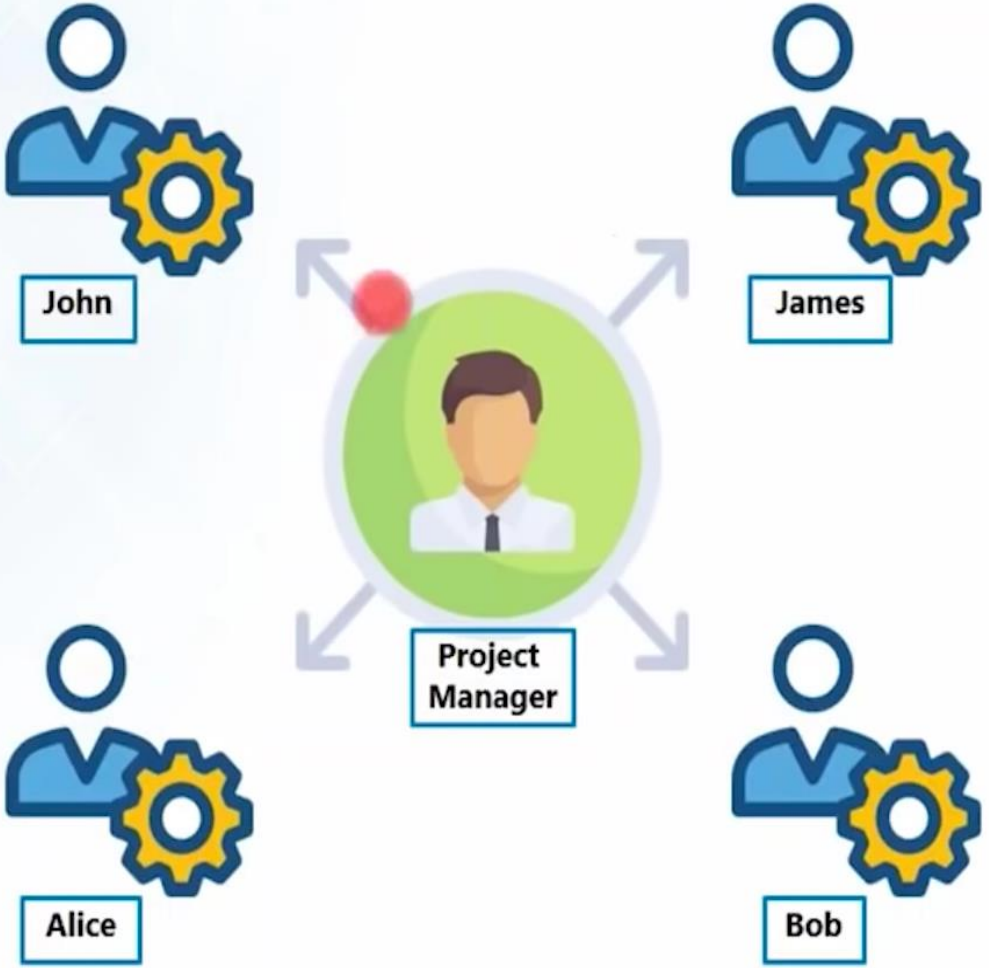
Hadoop is a framework that allows us to store and process large data sets in parallel and distributed fashion



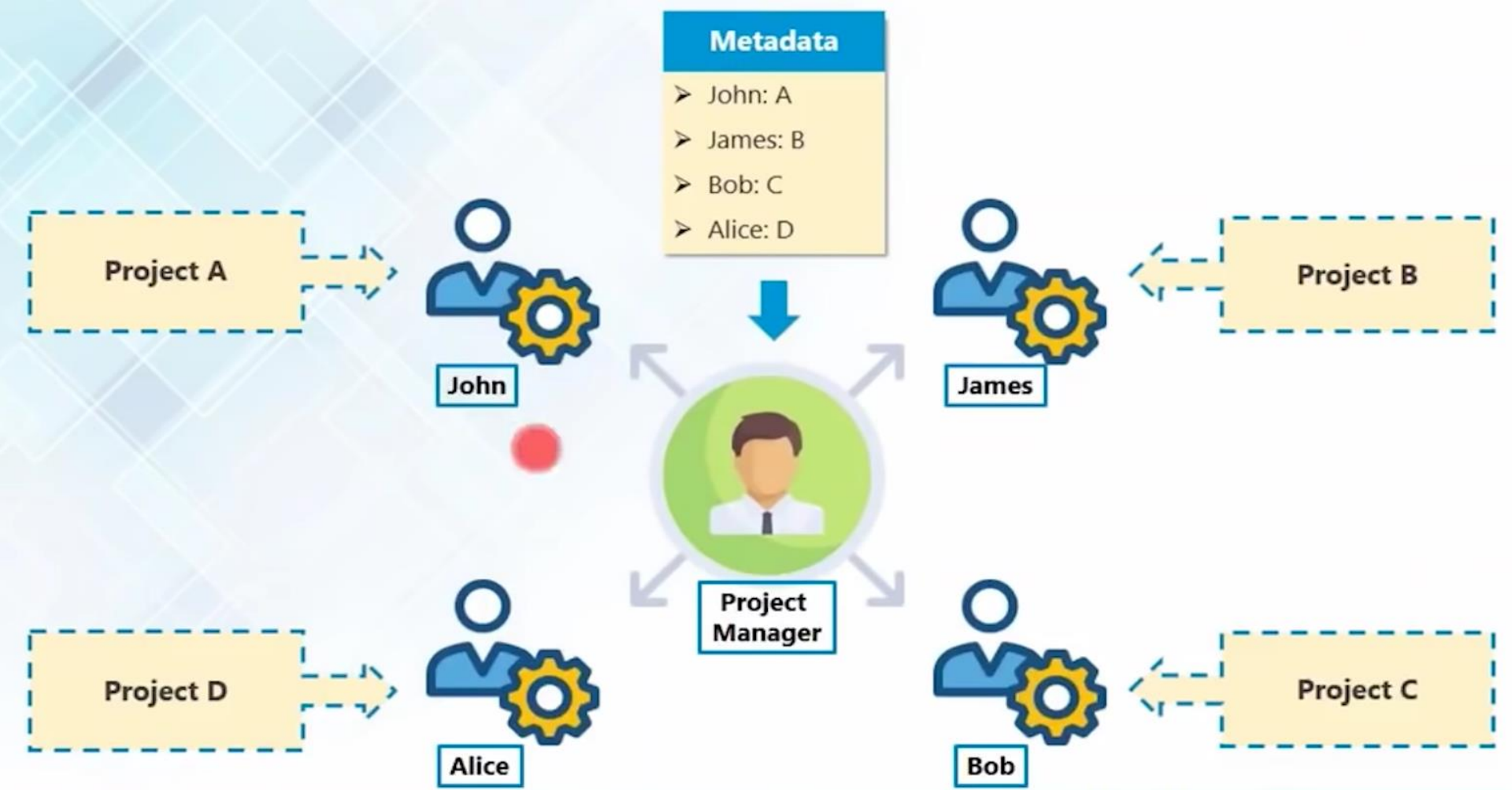
Ref:Edureka

Scenario:

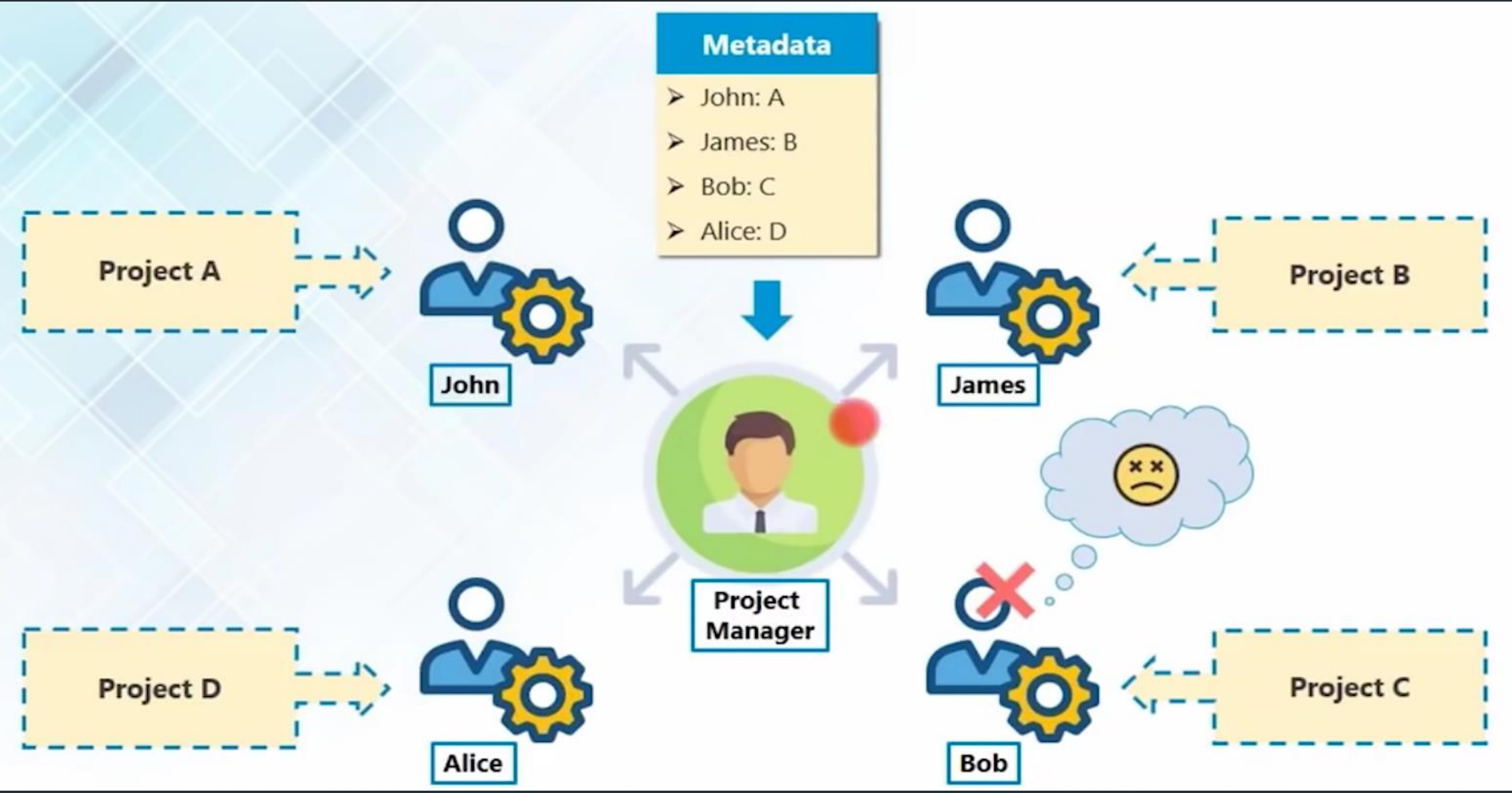
A project Manager managing a team of four employees. He assigns project to each of them and tracks the progress



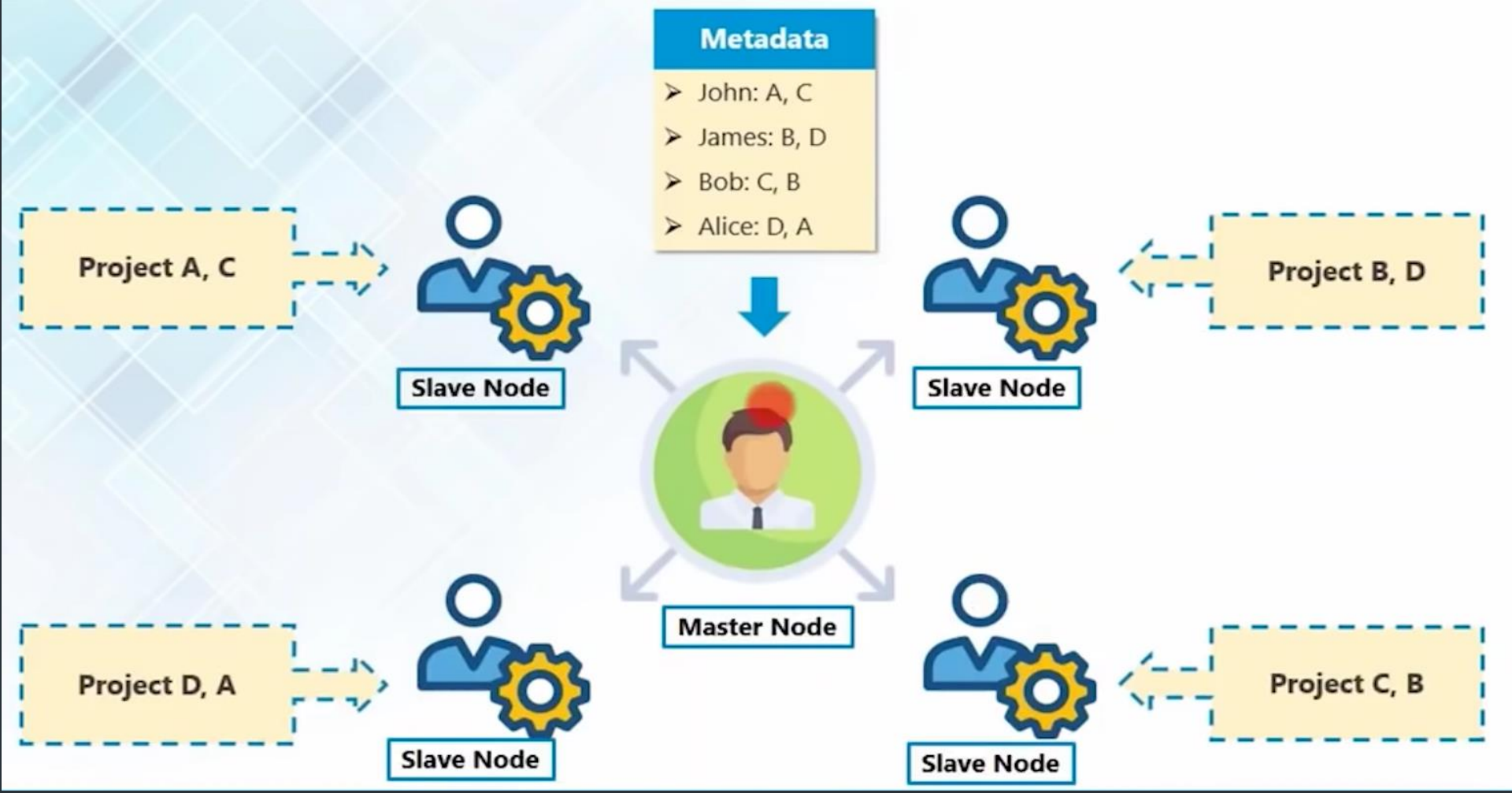
Ref:Edureka


Metadata
John: A
James: B
Bob: C
Alice: D
Project A
Project B
Project D
Project C
John
James
Alice
Bob
Project Manager

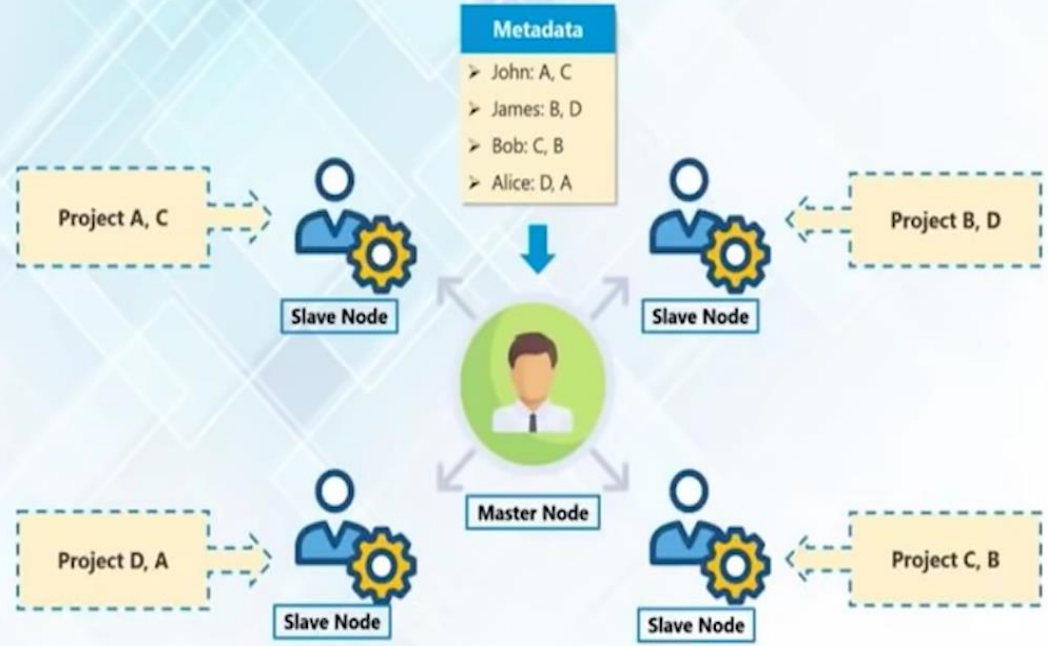
Ref:Edureka

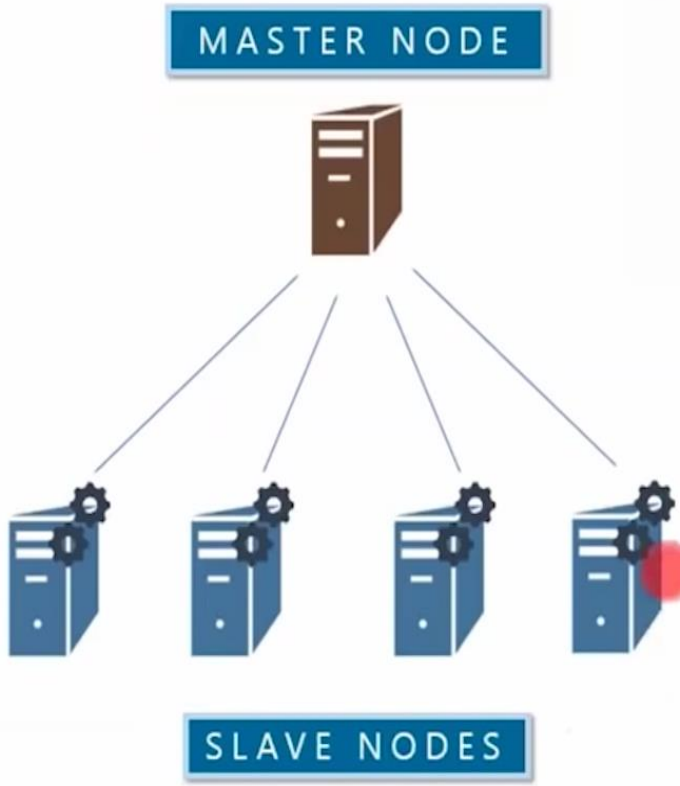

Metadata
John: A
James: B
Bob: C
Alice: D
Project A
John
James
Project B
Project Manager
Project D
Alice
Bob
Project C

Ref:Edureka


Metadata
John: A, C
James: B, D
Bob: C, B
Alice: D, A
Project A, C
Slave Node
Project B, D
Slave Node
Master Node
Project D, A
Slave Node
Project C, B
Slave Node

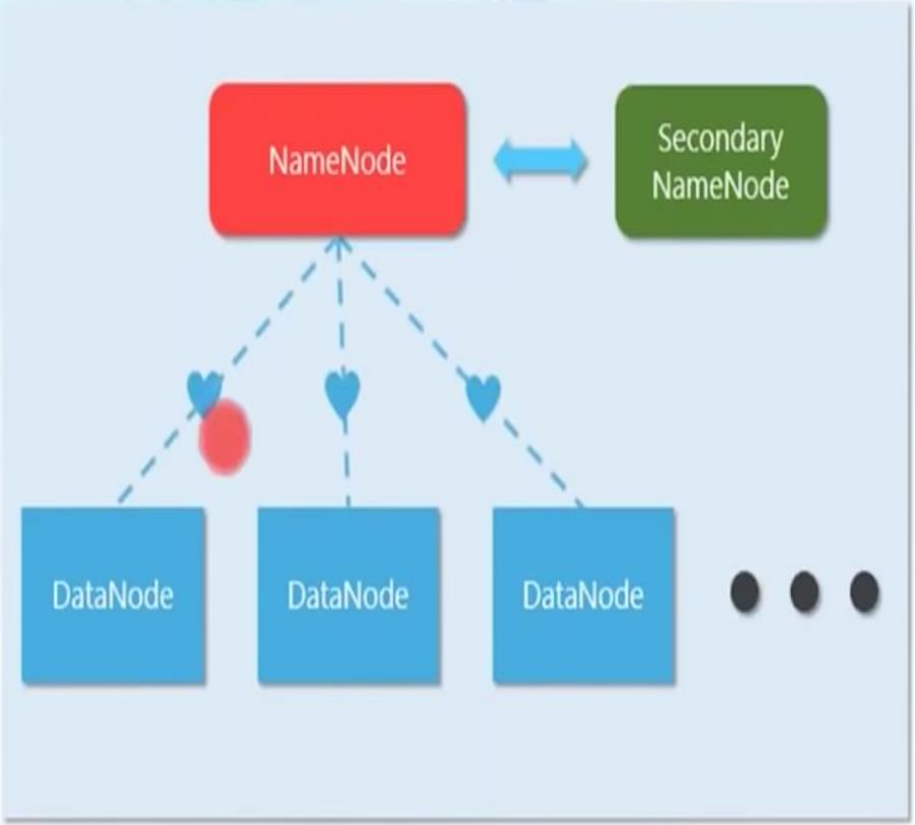
Ref:Edureka


Metadata
John: A, C
James: B, D
Bob: C, B
Alice: D, A
Project A, C
Slave Node
Project B, D
Slave Node
Master Node
Project D, A
Slave Node
Project C, B
Slave Node


MASTER NODE
SLAVE NODES

Ref:Edureka

HDFS



NameNode:

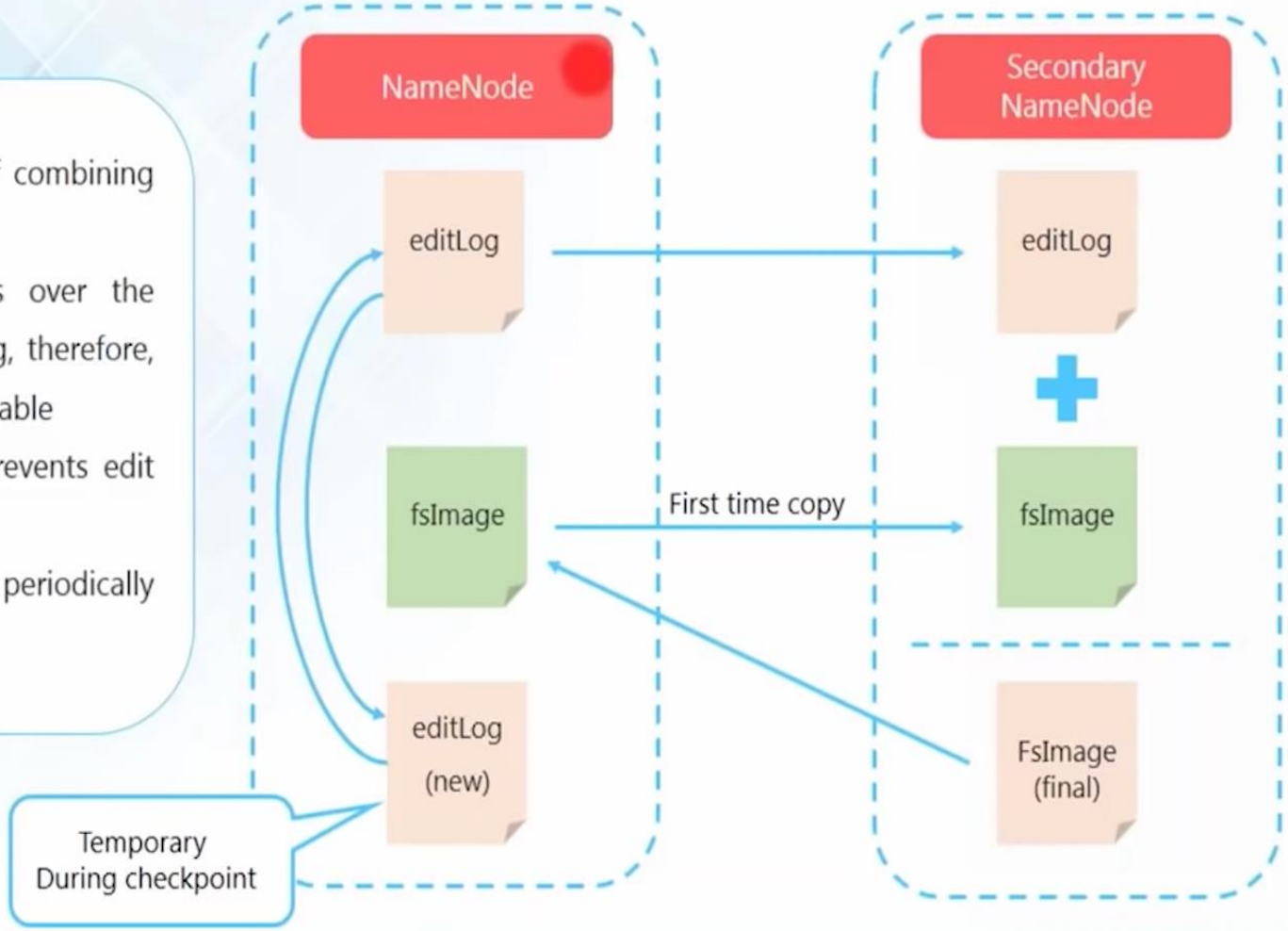
- Maintains and Manages DataNodes
- Records metadata i.e. information about data blocks e.g. location of blocks stored, the size of the files, permissions, hierarchy, etc.
- Receives heartbeat and block report from all the DataNodes

DataNode:

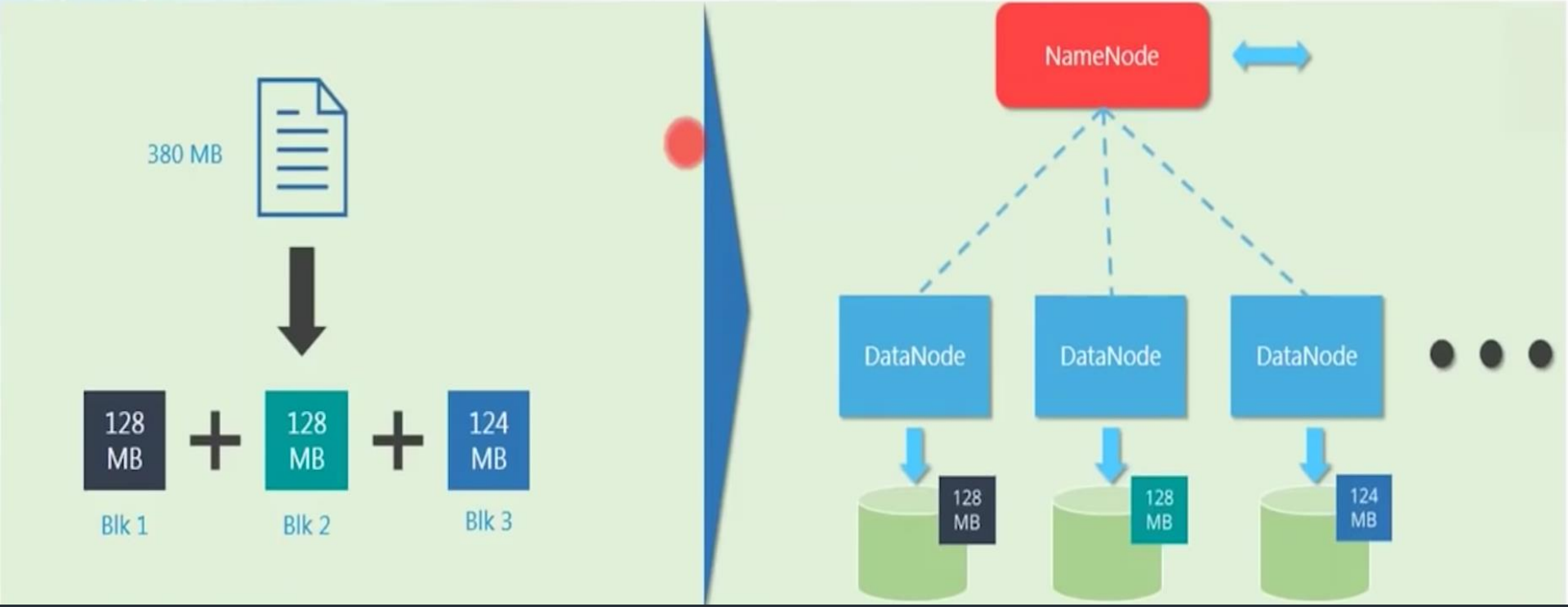
- Slave daemons
- Stores actual data
- Serves read and write requests from the clients

Ref:Edureka

- Checkpointing is a process of combining edit logs with FsImage
- Secondary NameNode takes over the responsibility of checkpointing, therefore, making NameNode more available
- Allows faster Failover as it prevents edit logs from getting too huge
- Checkpointing happens periodically (default: 1 hour)



- Each file is stored on HDFS as blocks
- The default size of each block is 128 MB in Apache Hadoop 2.x (64 MB in Apache Hadoop 1.x)

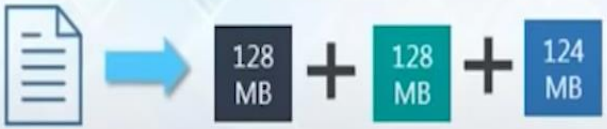


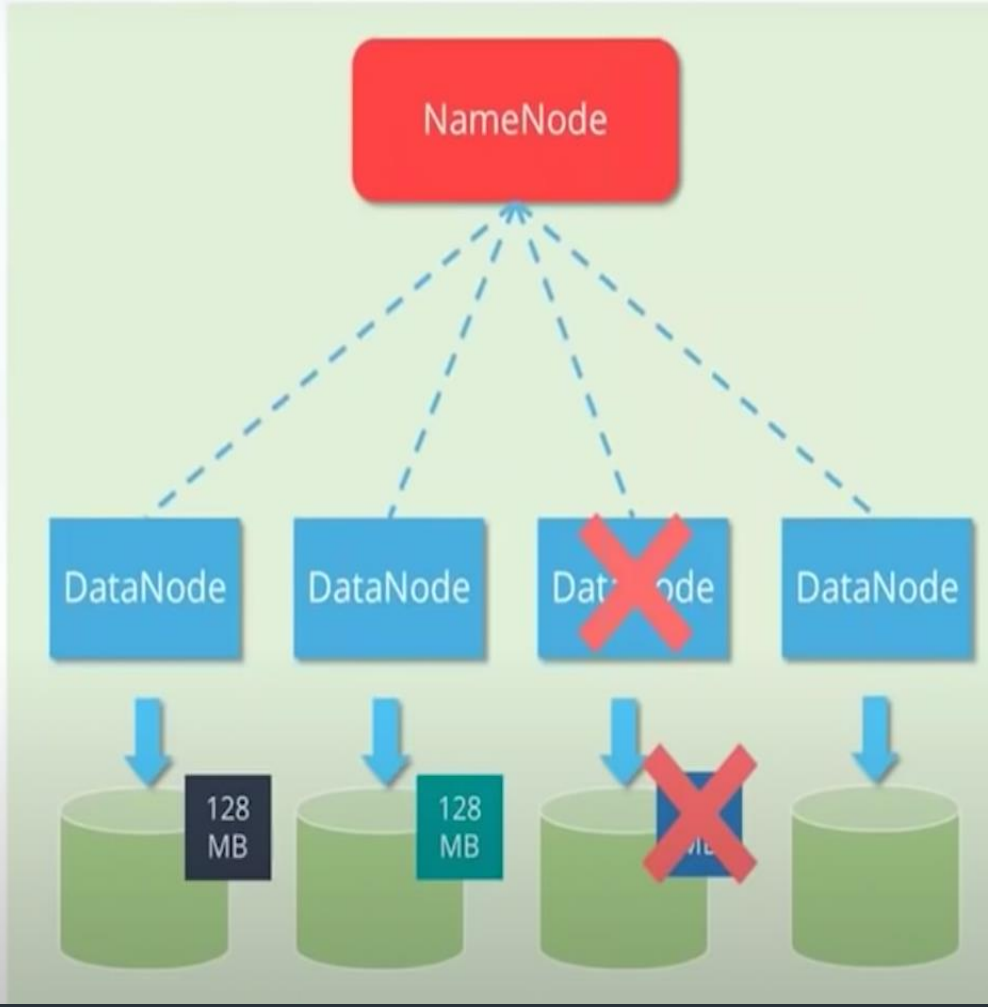
Ref:Edureka

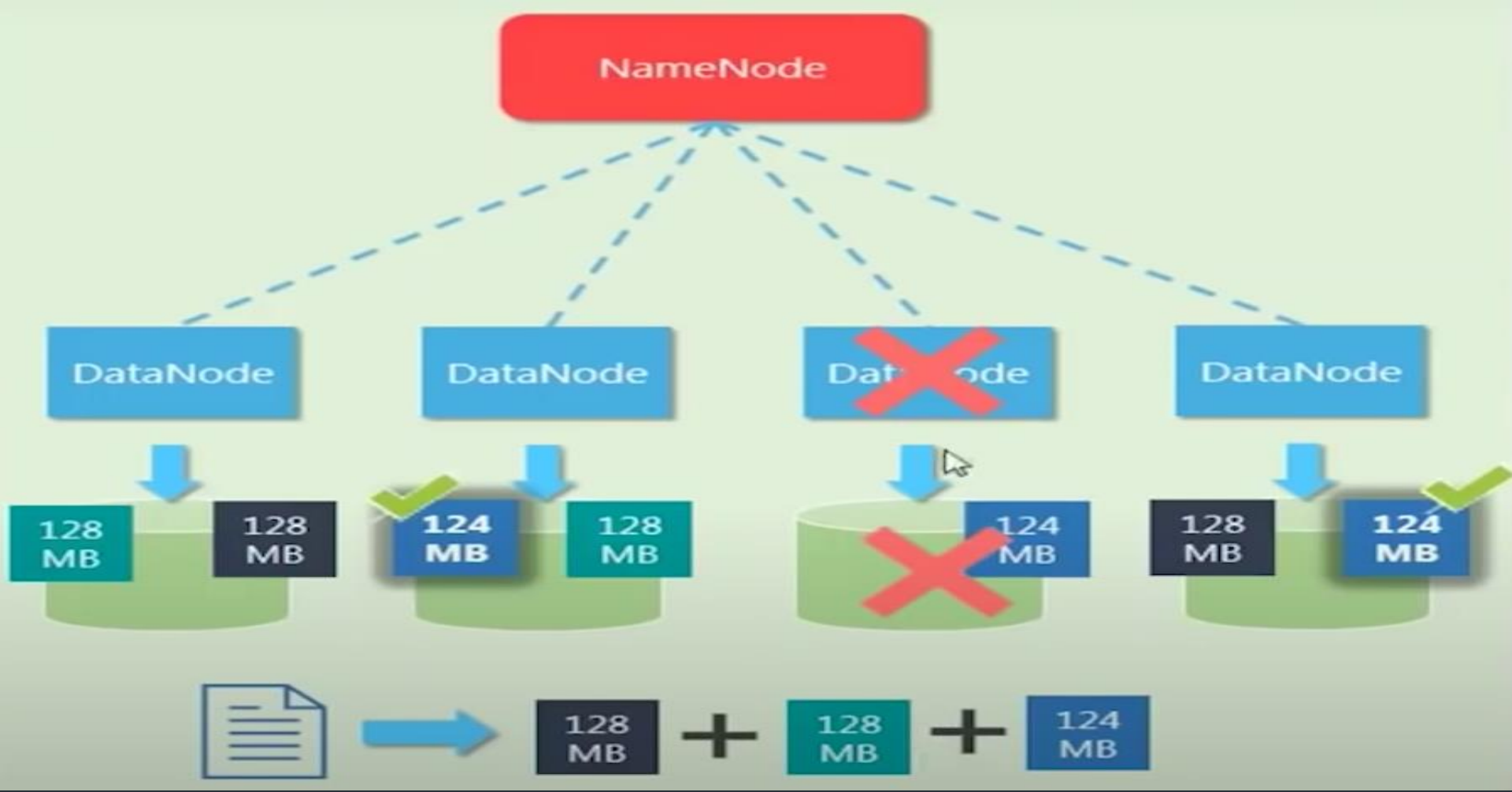
Fault Tolerance

Scenario:

One of the DataNodes crashed containing the data blocks

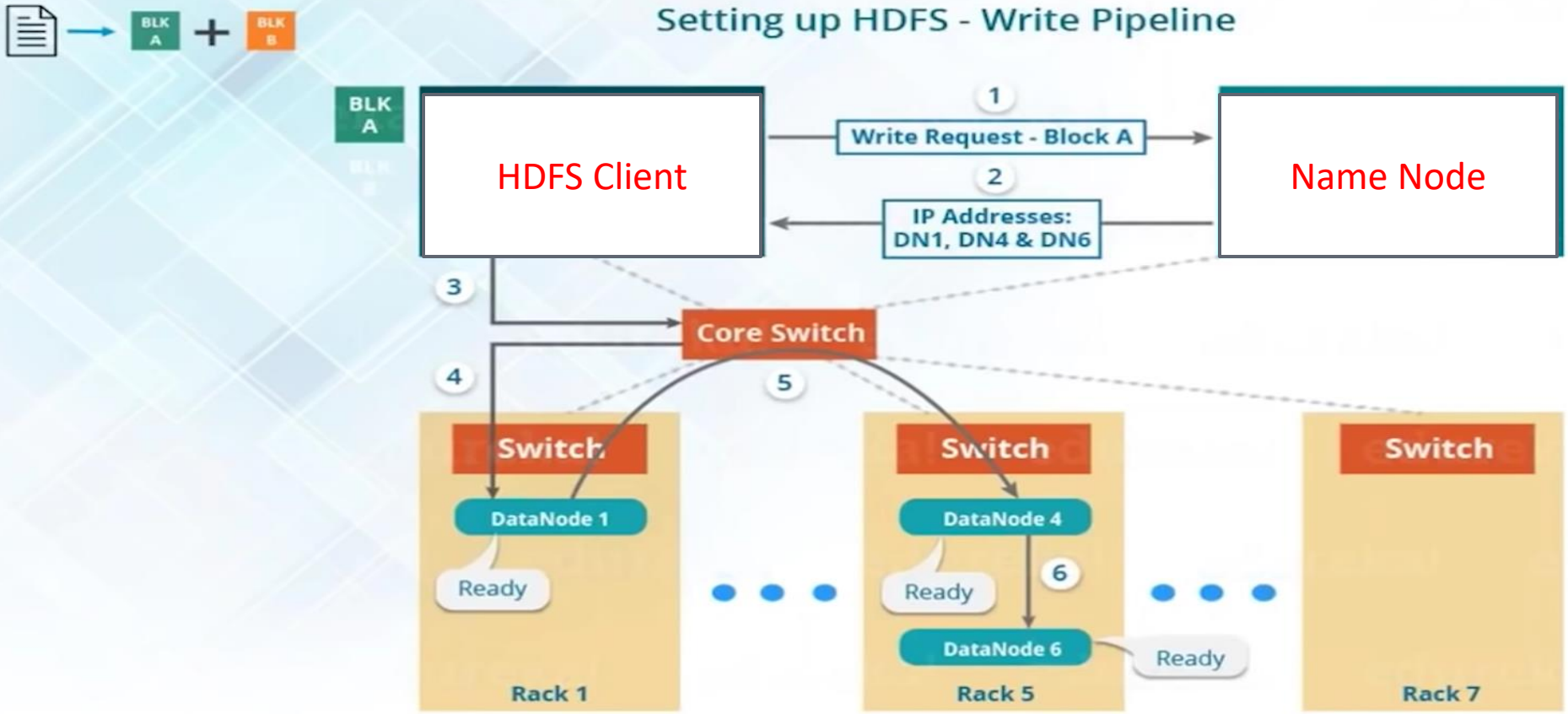





NameNode
DataNode
DataNode
DataNode
128 MB
128 MB
124 MB
128 MB
124 MB
128 MB
124 MB
128 MB
+
128 MB
+
124 MB

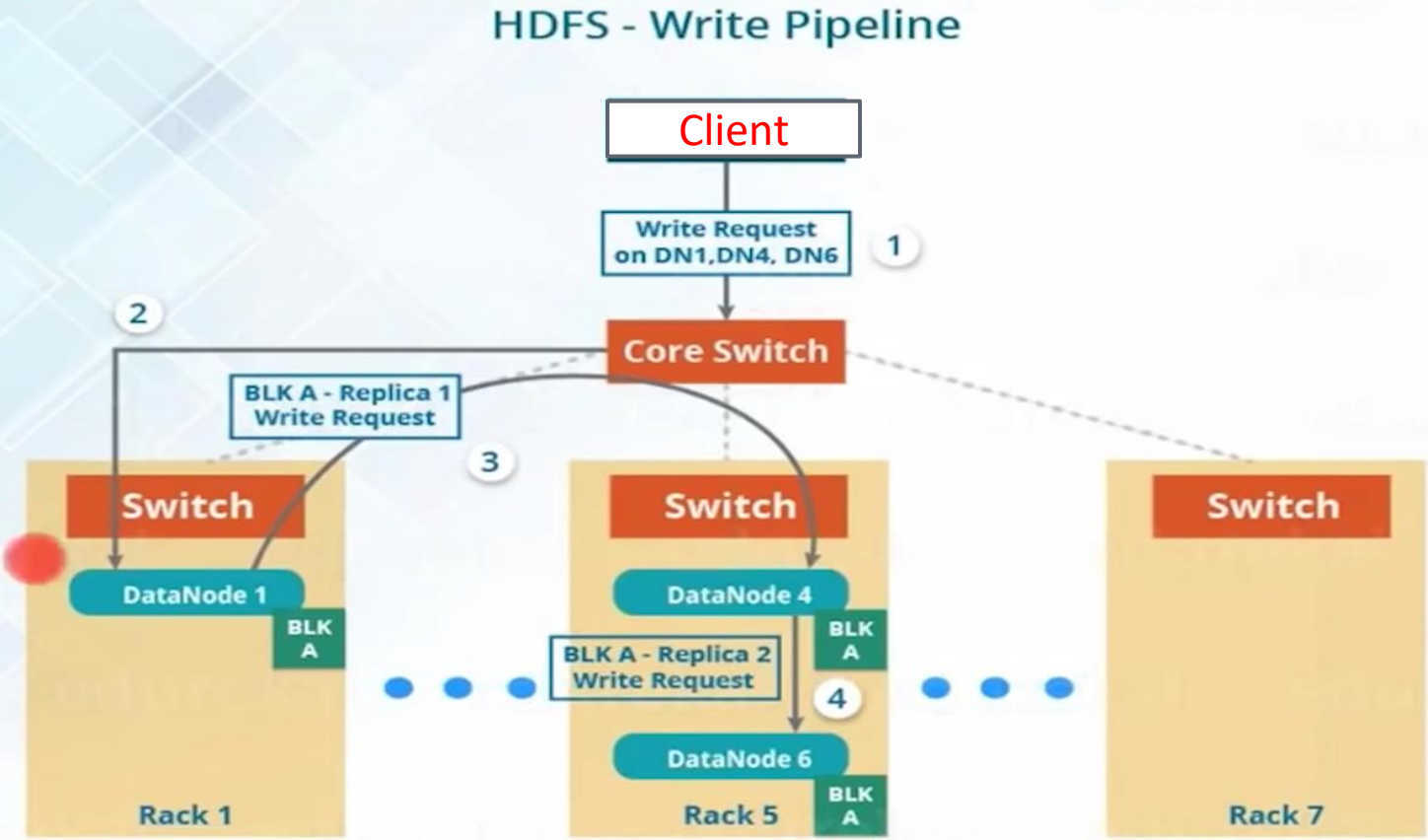
Ref:Edureka

HDFS Write Mechanism – Pipeline Setup



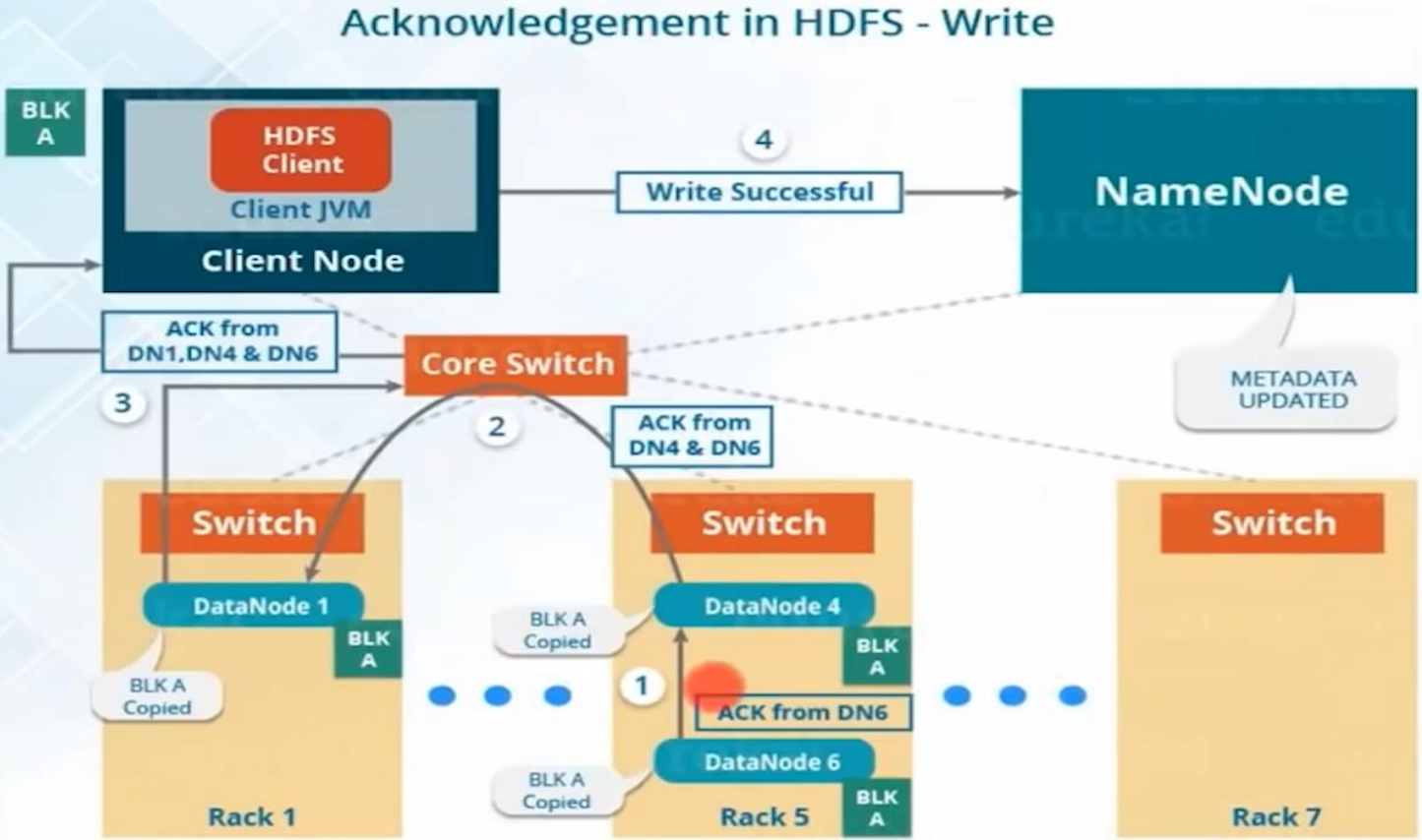
Ref:Edureka

HDFS Write Mechanism – Writing a Block



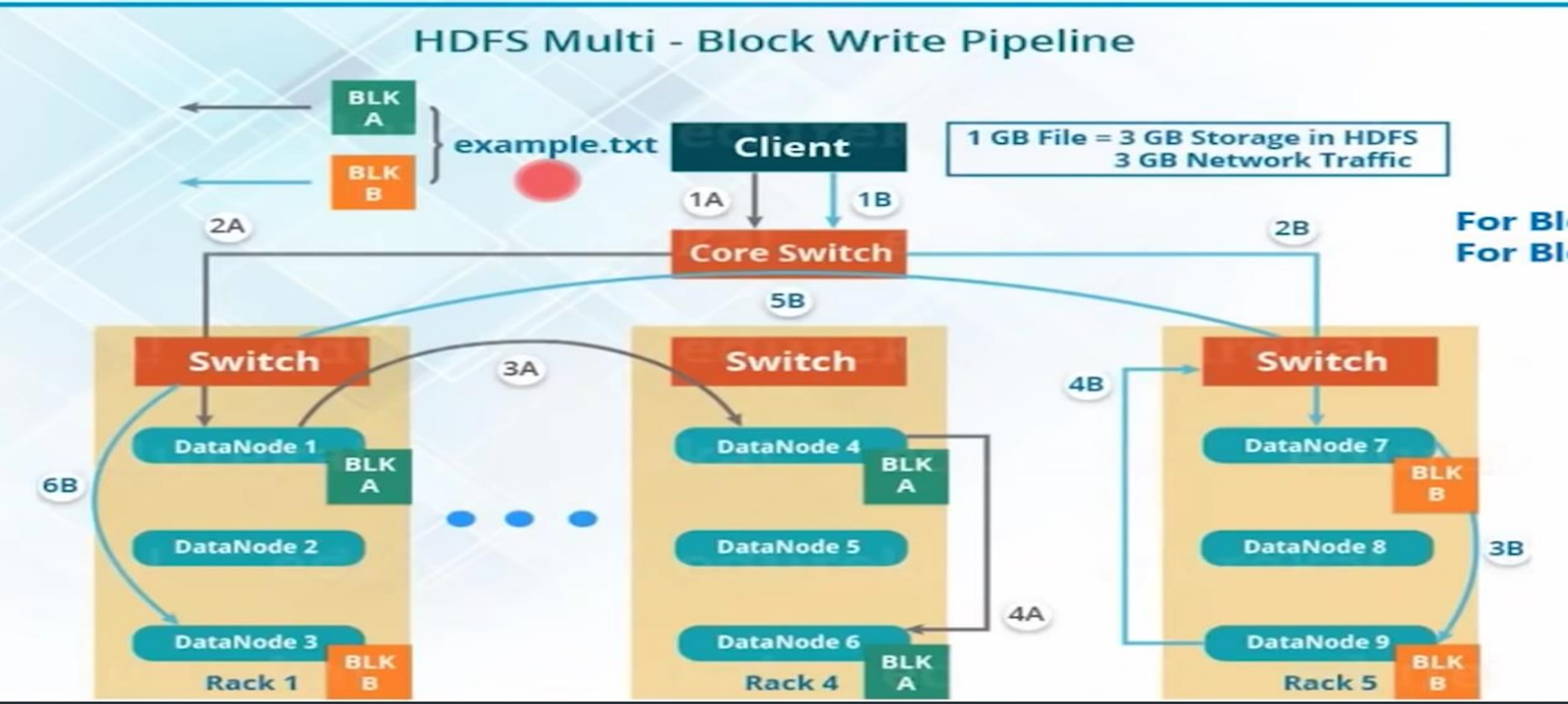
Ref:Edureka

HDFS Write Mechanism - Acknowledgement



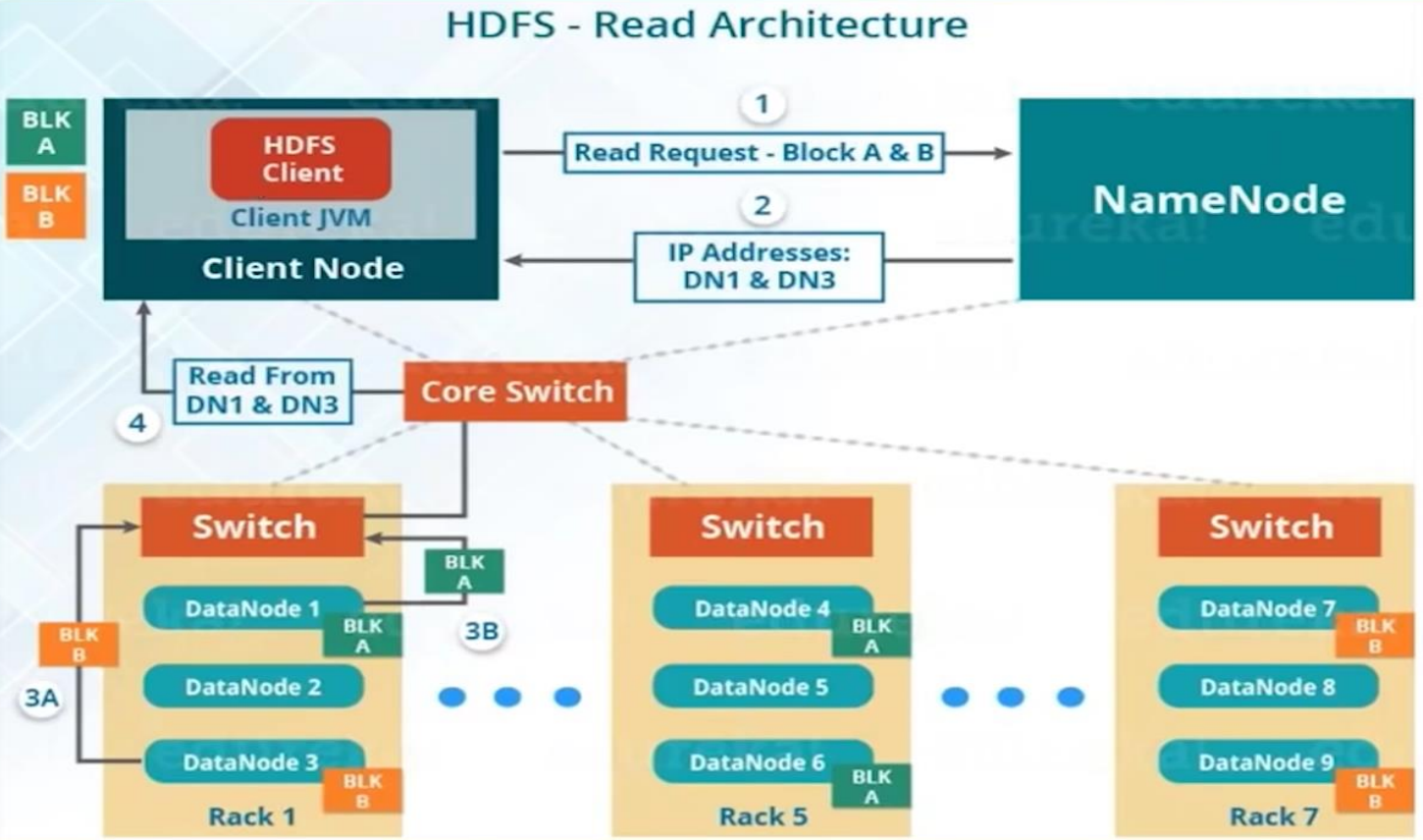
Ref:Edureka

HDFS Multi-Block Write Mechanism

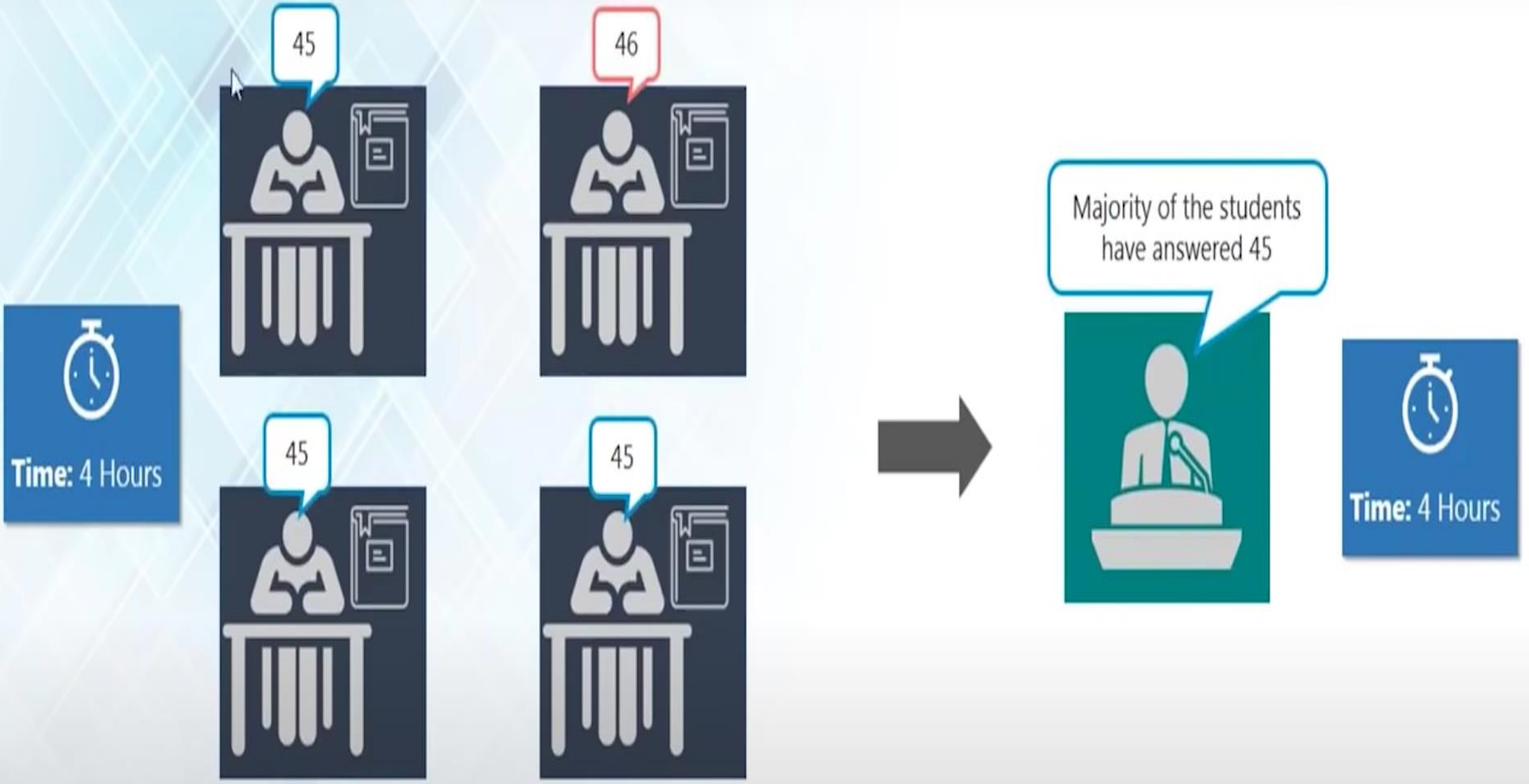


Ref:Edureka

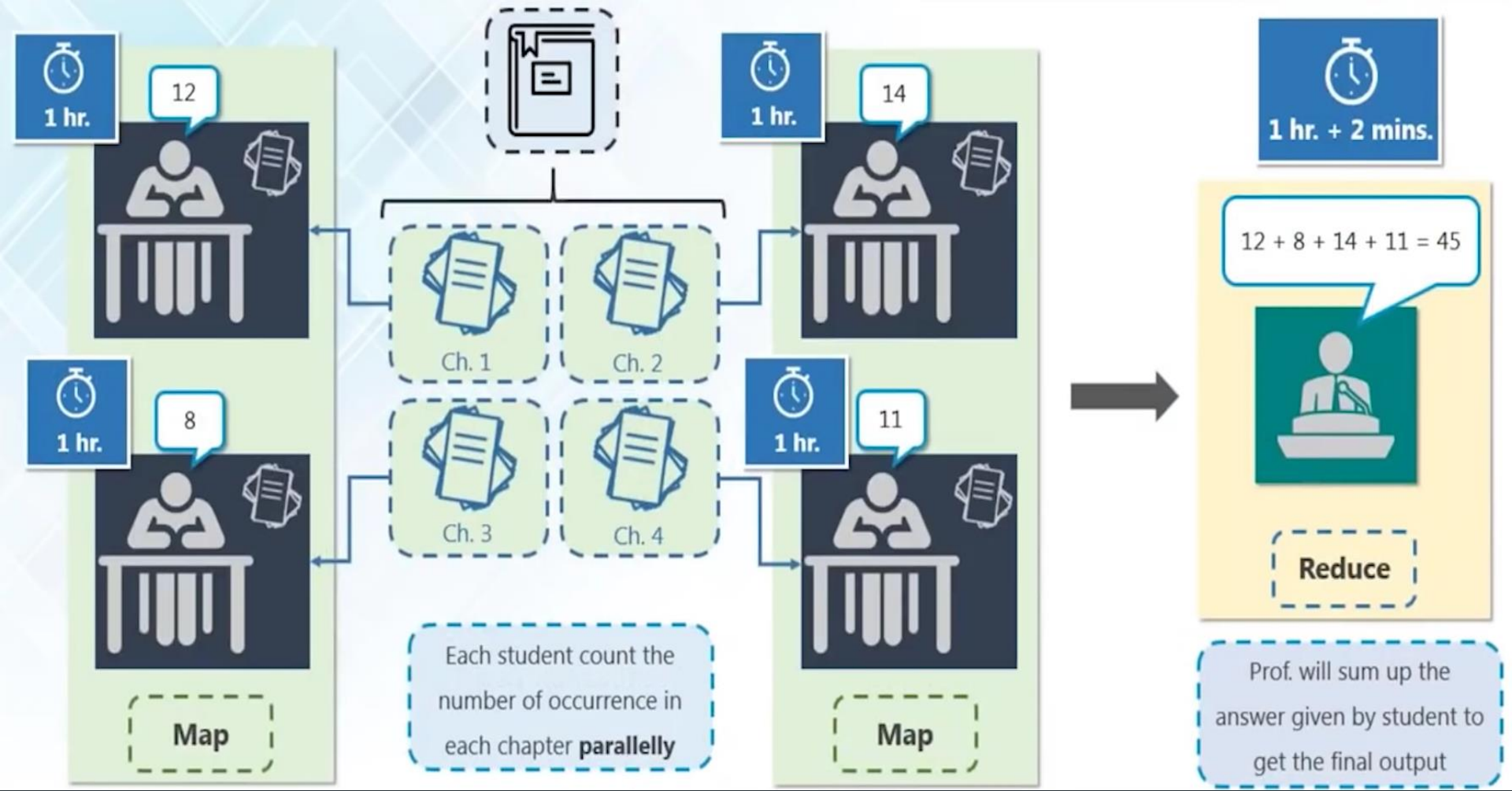
HDFS Read Mechanism



Ref:Edureka

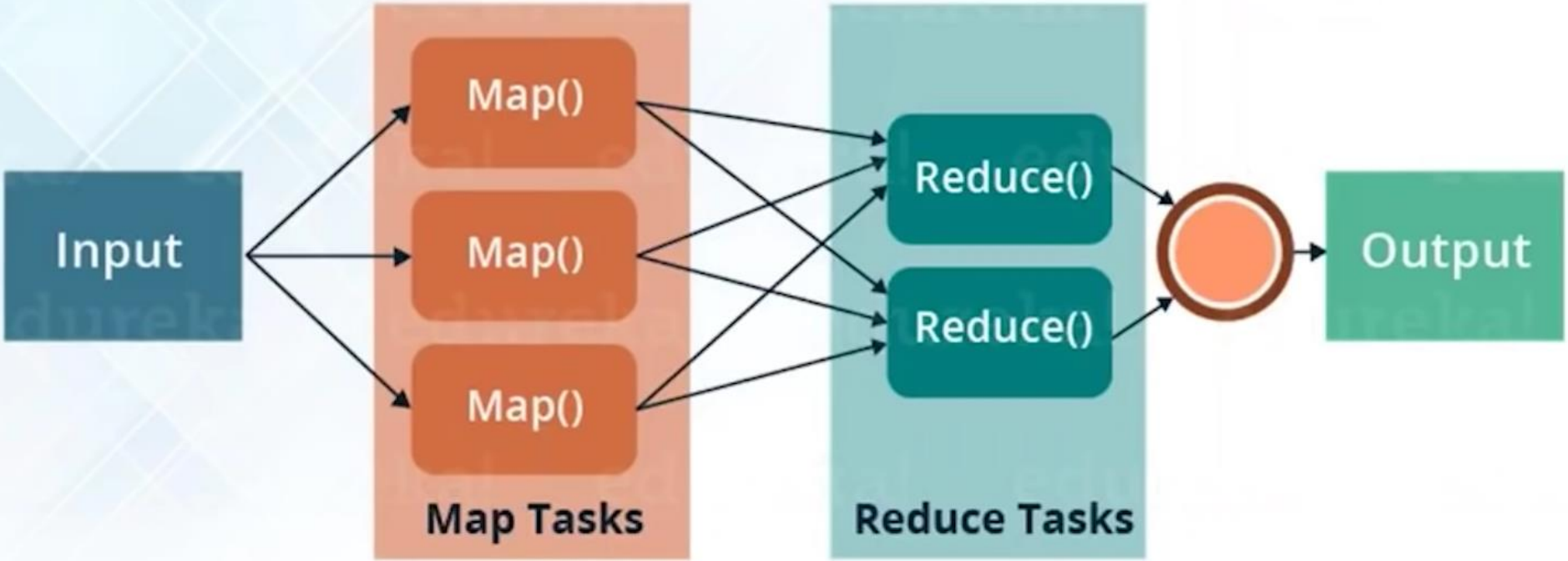

45
46
45
45
Time: 4 Hours
Majority of the students have answered 45
Time: 4 Hours

Ref:Edureka


1 hr.
12
1 hr.
14
1 hr. + 2 mins.
Ch. 1
Ch. 2
12 + 8 + 14 + 11 = 45
1 hr.
8
1 hr.
11
Ch. 3
Ch. 4
Reduce
Each student count the number of occurrence in each chapter parallelly
Map
Map
Prof. will sum up the answer given by student to get the final output

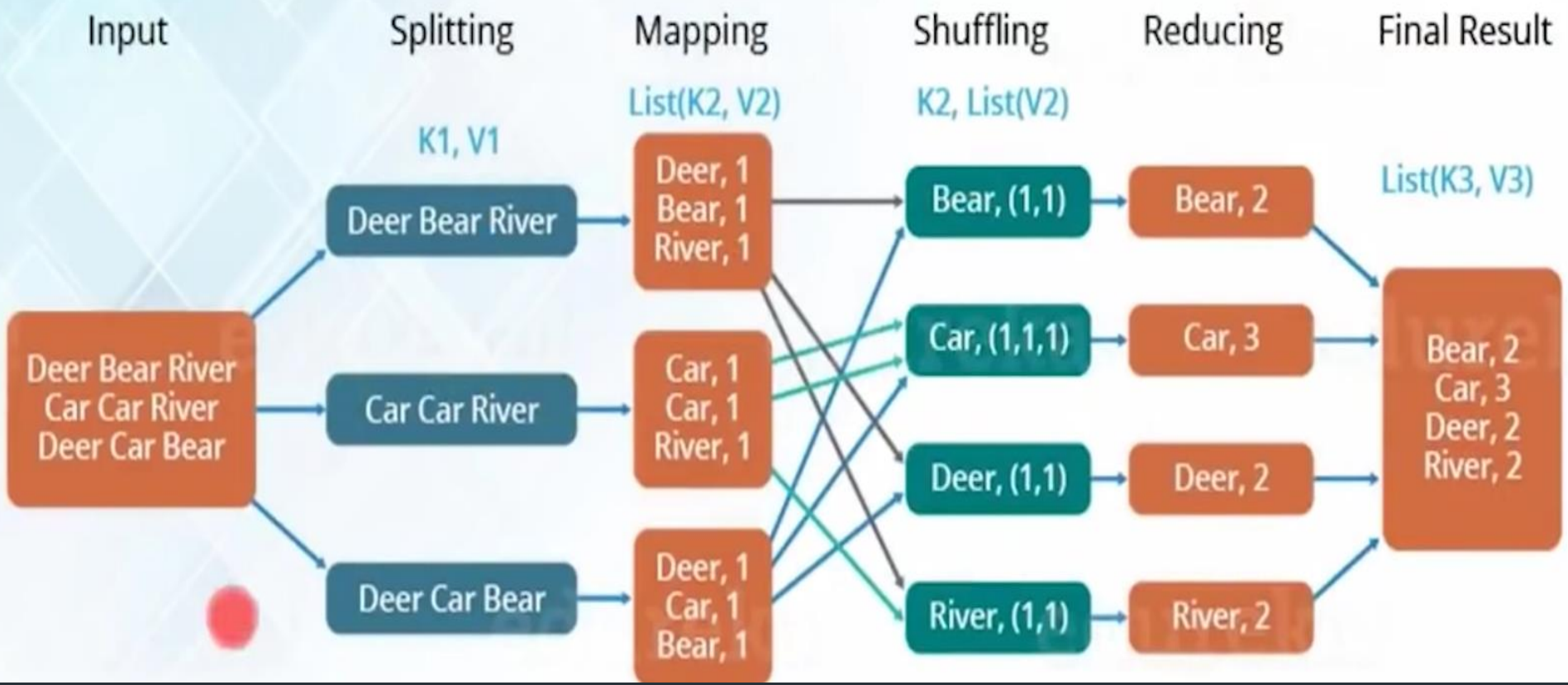
Ref:Edureka

MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment



Ref:Edureka

The Overall MapReduce Word Count Process



Three Major Parts of MapReduce Program:

1

Mapper Code:

You write the mapper logic over here i.e. how map task will process the data to produce the key-value pair to be aggregated

2

Reducer Code:

You write reducer logic here which combines the intermediate key-value pair generated by Mapper to give the final aggregated output

3

Driver Code

You specify all the job configurations over here like job name, Input path, output path, etc.

HADOOP Installation