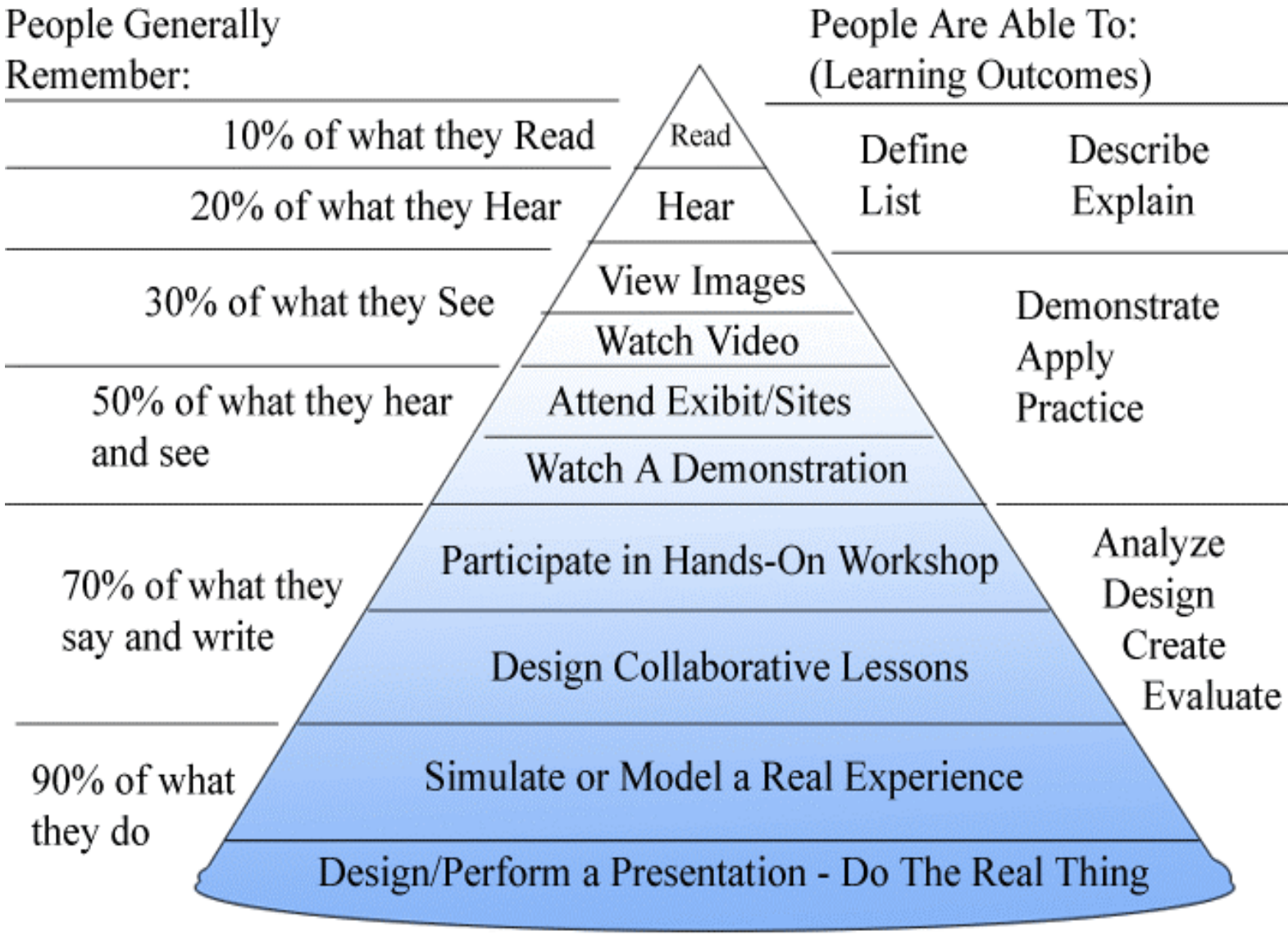
INTRODUCTION TO BIG DATA

Dr Thushari Silva
Department of Computational Mathematics
Faculty of Information Technology
University of Moratuwa

Why Attend the Lectures?



Year III - Semester 01			
Course Code:	**BSIT 31023**		
Course Name:	BIG DATA ANALYTICS		
Credit Value:	03		
Core/Optional	Core		
Hourly Breakdown	Theory	Tutorials/Practical	Independent Learning
	30	45	75

Course Aim/Intended Learning Outcomes:

At the end of the course unit students should be able to;

LO1: Describe the characteristics of big data and application perspective of big data.
LO2: Apply big data analytics frameworks including Hadoop, Yarn and SPARK and develop applications in big data.
LO3: Model schema-less databases and retrieve data using NoSQL.
LO4: Develop applications based on big data analytics framework.

Course Content:

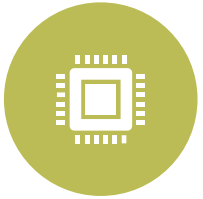
- Overview of Big data
  - Big data characteristics, application of big data, challengs in big data
- Technologies handling big data - Hadoop
  - Map-Reduce, HDFS, Hbase commands
- Hadoop YARN Architecture and Spark
  - YARN and its components, backward compatibility, log management in YARN, Spark, Data frames in Spark
- NoSQL Data Management
  - NoSQL features, document databases, graph databases, Database modeling with NoSQL, CAP Theorem

Ouline

Introduction to Big Data & its Application

Technologies Handling Big Data

Apache Hadoop & Map Reduce



Apache Spark

Spark SQL & NoSQL

Introduction to Big Data

What it is? and Why it Matters?

•01

Big Data, what is it?

traditional
computer science

data that will not fit in main memory.

data with a *large* number of observations and/or features.

statistics

other fields

non-traditional sample size (i.e. > 300 subjects); can't analyze in stats tools (Excel).

Big Data, What is it?

Big Data, What is it?

(ChatGPT, January 2023)

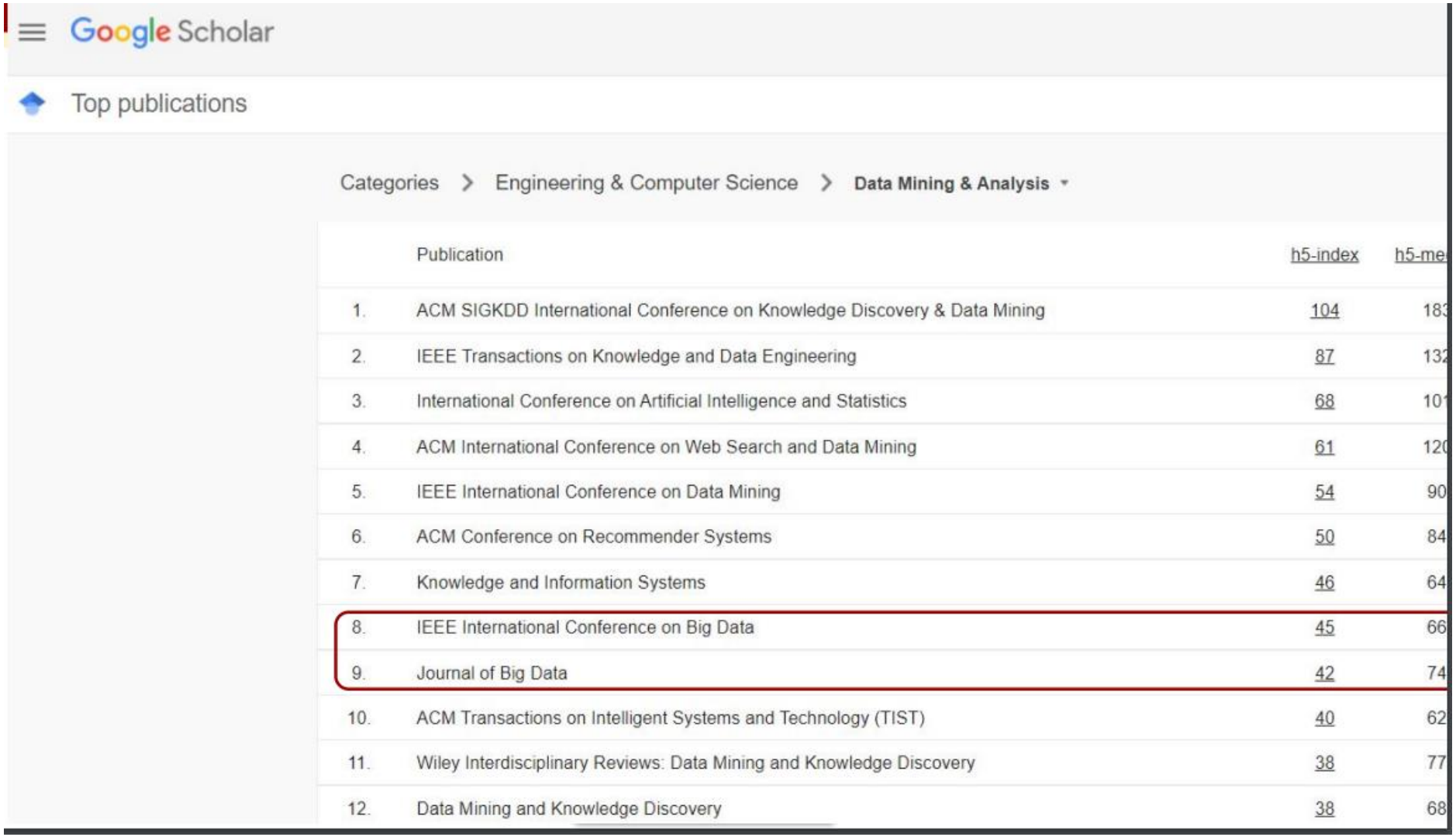
Big data refers to the large and complex sets of data that are generated and collected by organizations and individuals on a daily basis. These data sets can come from a variety of sources, such as social media, online transactions, and sensor data, and can be structured or unstructured.

One of the main challenges of big data is the ability to store, process, and analyze it effectively. Traditional data processing methods and technologies are often not able to handle the volume, velocity, and variety of big data. As a result, new technologies and approaches, such as distributed computing and machine learning, have been developed to help organizations make sense of their big data.

Big data can have a wide range of applications, from improving business operations and customer service to enabling new scientific discoveries and advancements in healthcare. For example, in business, big data can be used to gain insights into customer behavior, identify new market opportunities, and optimize supply chain operations. In healthcare, big data can be used to improve patient outcomes and develop personalized treatment plans.

Overall, big data is a rapidly growing field with many potential benefits for organizations and individuals, but also has the potential for privacy and security concerns. Therefore, it is important for organizations to have a robust data governance framework and for individuals to understand the implications of data collection and use.

Big Data, a buzz Word



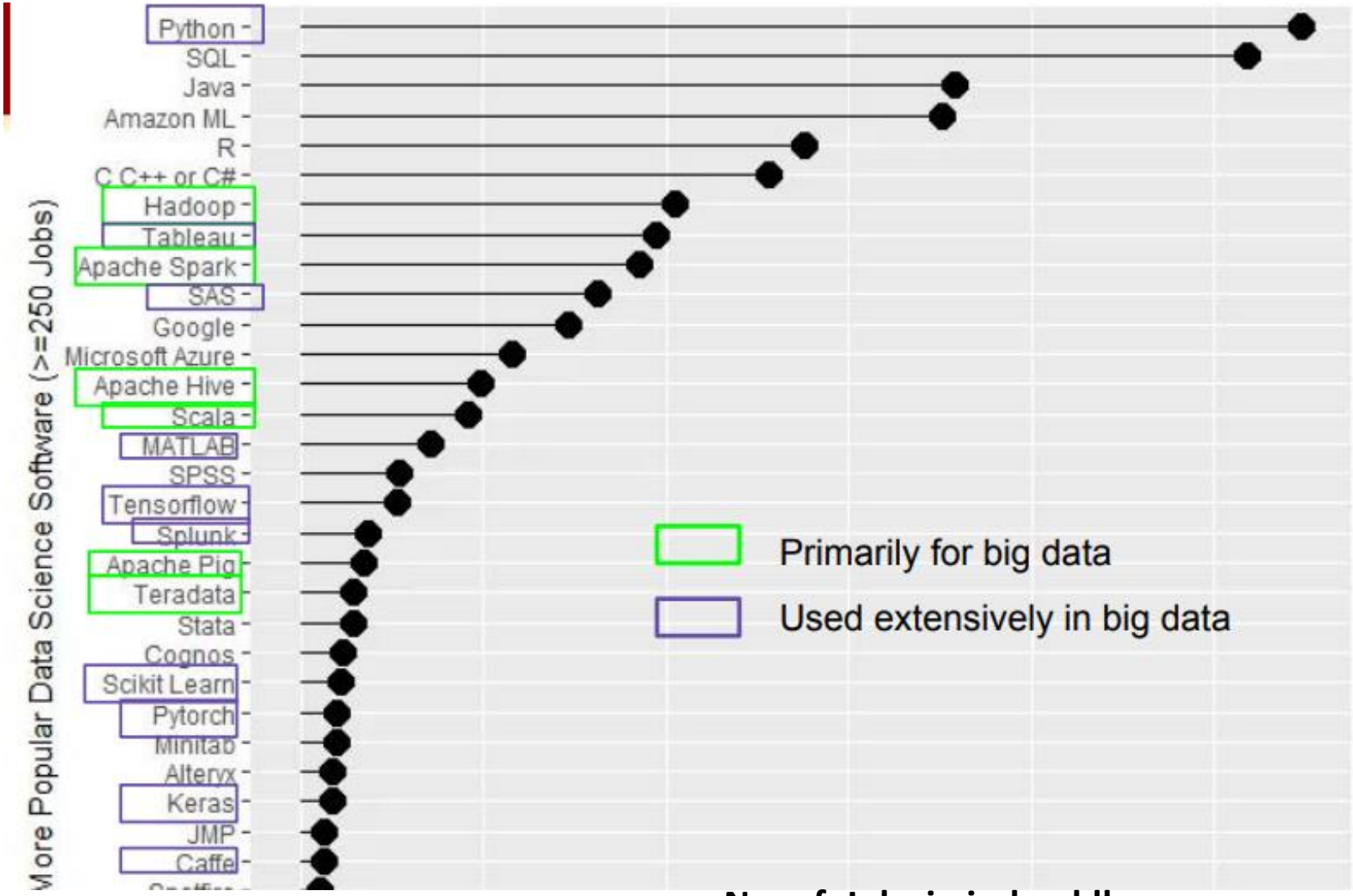
Google Scholar

Top publications

Categories > Engineering & Computer Science > **Data Mining & Analysis**

	Publication	h5-index	h5-me
1.	ACM SIGKDD International Conference on Knowledge Discovery & Data Mining	104	183
2.	IEEE Transactions on Knowledge and Data Engineering	87	132
3.	International Conference on Artificial Intelligence and Statistics	68	101
4.	ACM International Conference on Web Search and Data Mining	61	120
5.	IEEE International Conference on Data Mining	54	90
6.	ACM Conference on Recommender Systems	50	84
7.	Knowledge and Information Systems	46	64
8.	IEEE International Conference on Big Data	45	66
9.	Journal of Big Data	42	74
10.	ACM Transactions on Intelligent Systems and Technology (TIST)	40	62
11.	Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery	38	77
12.	Data Mining and Knowledge Discovery	38	68

Big Data, in demand?



No. of Jobs in indeed.lk

Big Data, What is it?

Goal: Generalizations
A ***model*** or ***summarization*** of the data.

How to analyze data that is mostly too large for main memory.

Analyses only possible with a *large* number of observations or features.

Big Data, What is it?

Goal: Generalizations
A *model* or *summarization* of the data.

E.g.

- Google's PageRank: *summarizes* web pages by a single number.
- Twitter financial market predictions: *Models* the stock market according to shifts in sentiment in Twitter.
- Distinguish tissue type in medical images: *Summarizes* millions of pixels into clusters.
- Mental health diagnosis in social media: *Models* presence of diagnosis as a distribution (a summary) of linguistic patterns.
- Frequent co-occurring purchases: *Summarize* billions of purchases as items that frequently are bought together.

Big Data, What is it?

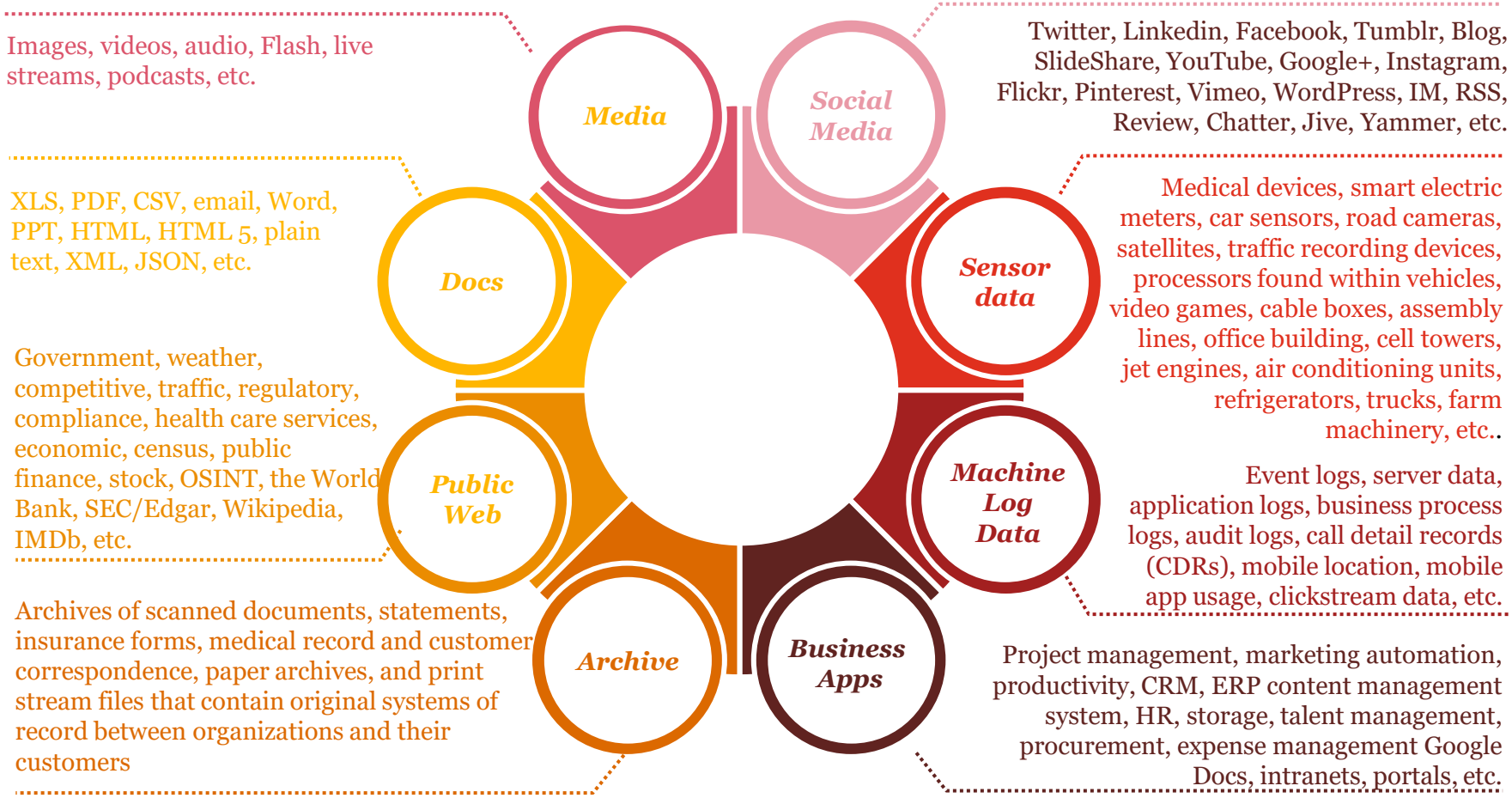
Goal: Generalizations
A *model* or *summarization* of the data.

1. Descriptive analytics
   Describe (*generalizes*) the data itself

2. Predictive analytics
   Create something *generalizeable* to new data

Types of Big Data

Variety is the most unique aspect of Big Data. New technologies and new types of data have driven much of the evolution around Big Data.



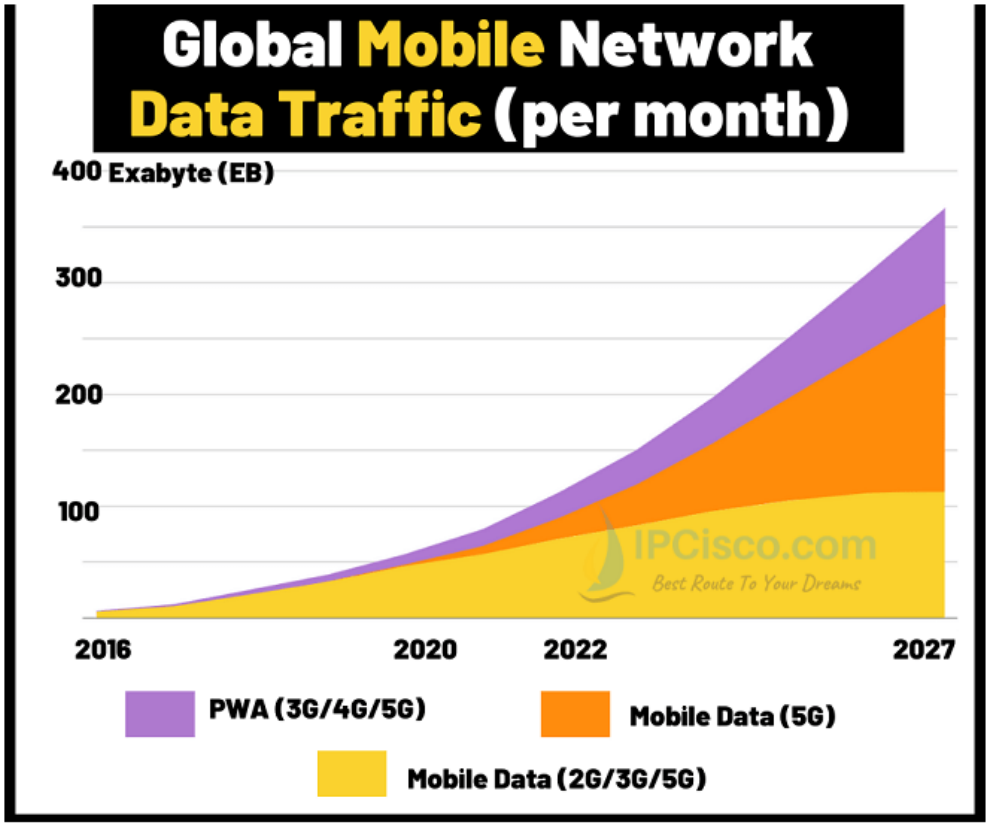
Why is Big Data valuable?

We have identified 5 key areas where Big Data is uniquely valuable:

Accessibility to Data	Enhanced visibility of relevant information and better transparency to massive amounts of data. Improved reporting to stakeholders.
Decision Making	Next generation analytics can enable automated decision making (inventory management, financial risk assessment, sensor data management, machinery tuning).
Marketing Trends	Segmentation of population to customize offerings and marketing campaigns (consumer goods, retail, social, clinical data, etc).
Performance Improvement	Exploration for, and discovery of, new needs, can drive organizations to fine tune for optimal performance and efficiency (employee data).
New Business Models/Services	Discovery of trends will lead organizations to form new business models to adapt by creating new service offerings for their customers. Intermediary companies with big data expertise will provide analytics to 3rd parties.

Big Data Growth Drivers

Global Mobile Data Traffic, 2015-2027

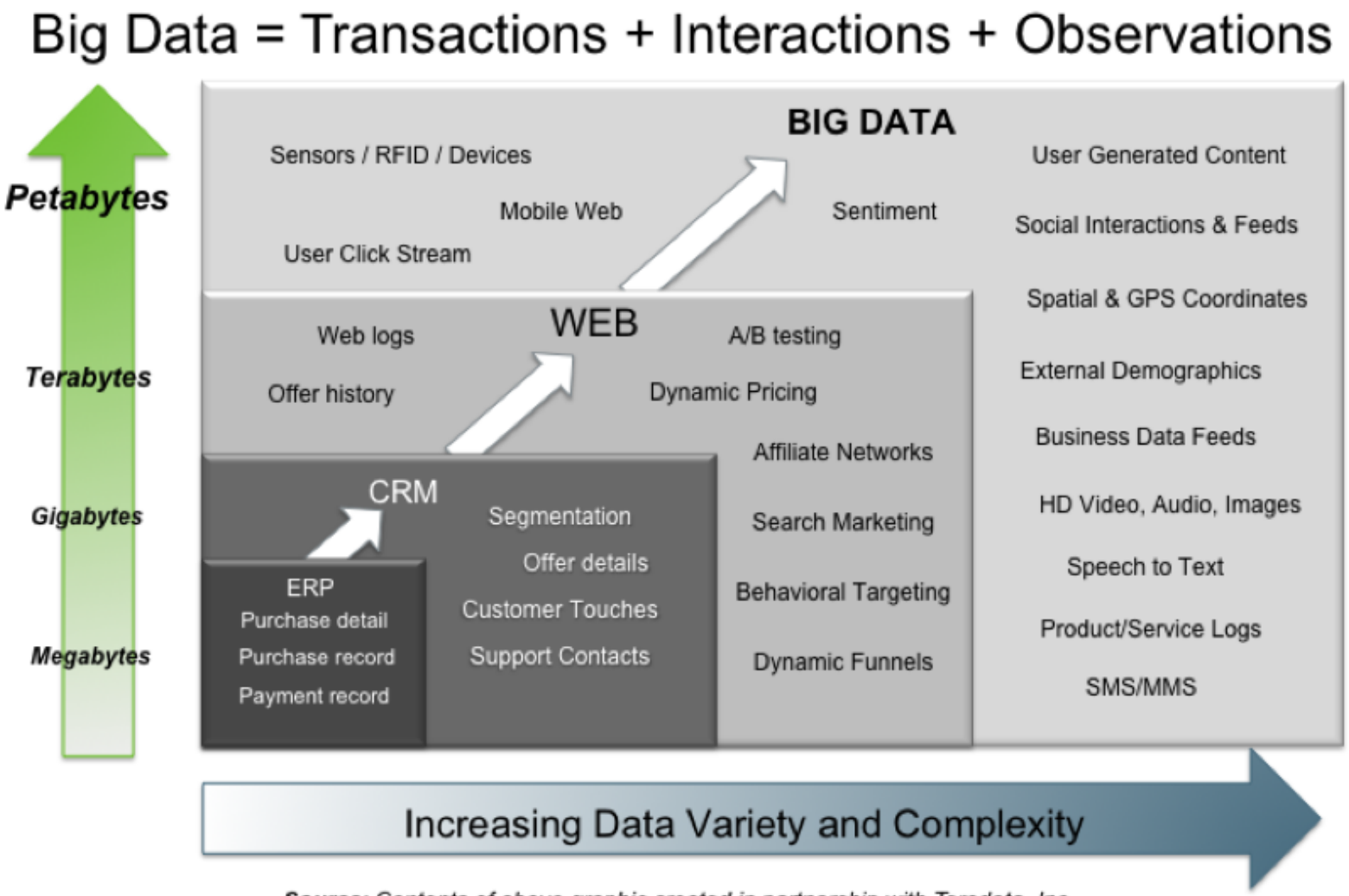


- **3-Major contributions**
  - Adapting to Smarter Mobile Devices
  - Defining Cell Network advances – 2G, 3G, 4G and 5G perspective
  - Reviewing Tiered Pricing – Unlimited data and Shared plans

What's Big Data?

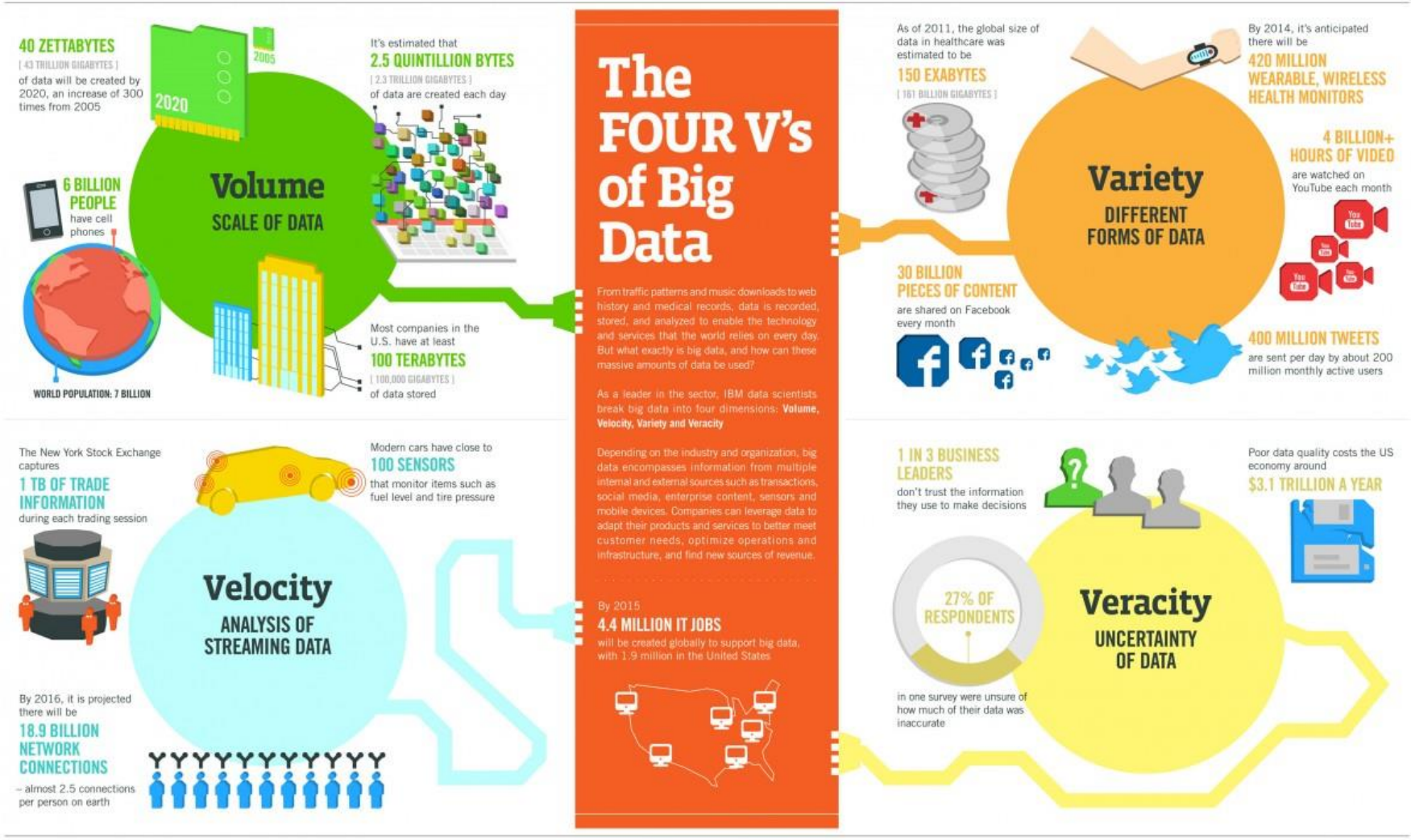
- **Big data**
  - a collection of data sets so large and complex,
  - difficult to process using on-hand database management tools or traditional data processing applications
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

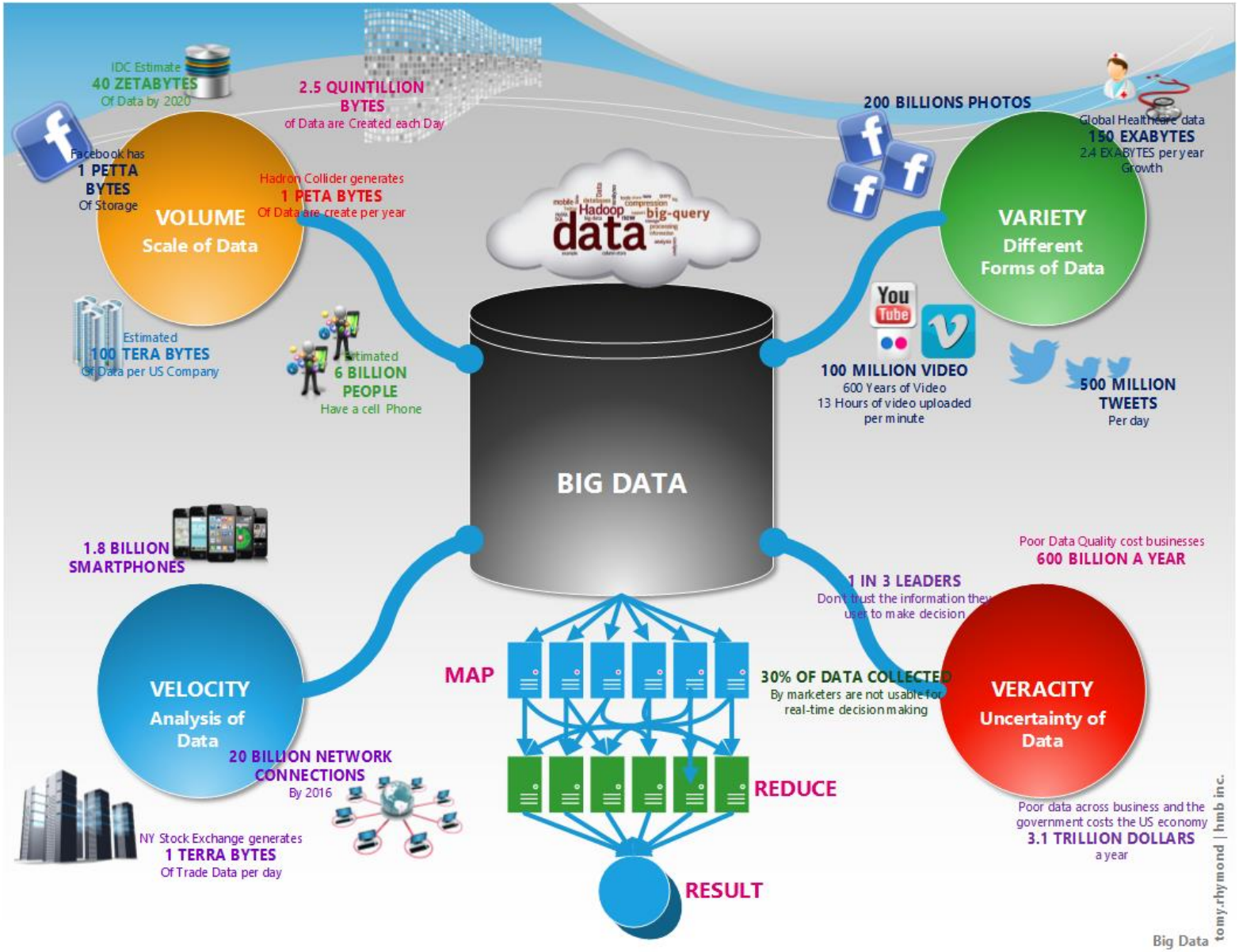
Big Data: 4V's



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

BigData Challenges & Characteristics




IDC Estimate
40 ZETABYTES
Of Data by 2020
2.5 QUINTILLION BYTES
of Data are Created each Day
Facebook has
1 PETTA BYTES
Of Storage
VOLUME
Scale of Data
Hadron Collider generates
1 PETA BYTES
Of Data are create per year
Estimated
100 TERA BYTES
Of Data per US Company
Estimated
6 BILLION PEOPLE
Have a cell Phone
Hadoop
data
big-query
BIG DATA
200 BILLIONS PHOTOS
Global Healthcare data
150 EXABYTES
2.4 EXABYTES per year
Growth
VARIETY
Different Forms of Data
100 MILLION VIDEO
600 Years of Video
13 Hours of video uploaded per minute
500 MILLION TWEETS
Per day
1.8 BILLION SMARTPHONES
VELOCITY
Analysis of Data
20 BILLION NETWORK CONNECTIONS
By 2016
NY Stock Exchange generates
1 TERRA BYTES
Of Trade Data per day
MAP
REDUCE
RESULT
Poor Data Quality cost businesses
600 BILLION A YEAR
1 IN 3 LEADERS
Don't trust the information they use to make decision
30% OF DATA COLLECTED
By marketers are not usable for real-time decision making
VERACITY
Uncertainty of Data
Poor data across business and the government costs the US economy
3.1 TRILLION DOLLARS
a year
Big Data
tomy.rhymond | hmb inc.

Volume

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

12+ TBs of tweet data every day

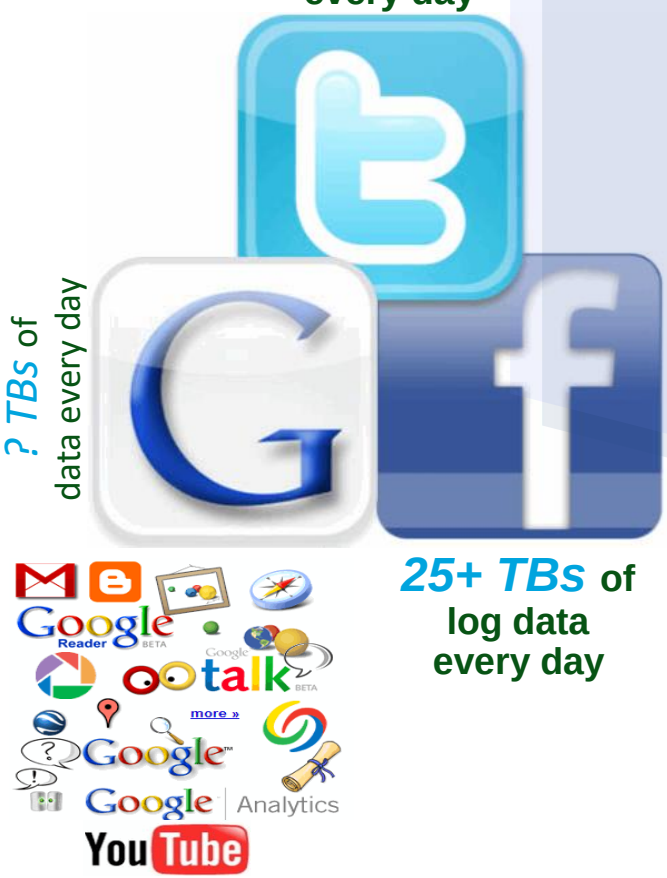
? TBs of data every day

25+ TBs of log data every day

100s of millions of GPS enabled devices sold annually

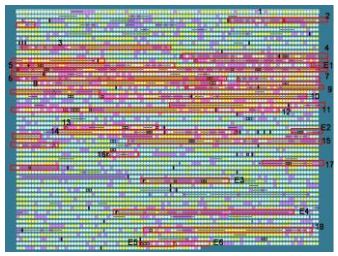
76 million smart meters in 2009… 200M by 2014

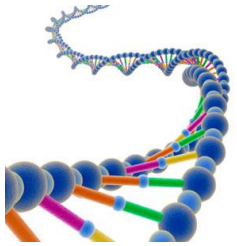
2+ billion people on the Web by end 2011

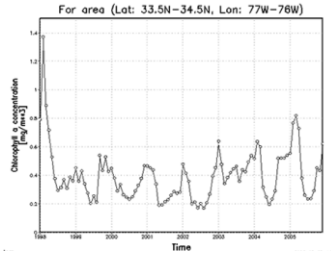


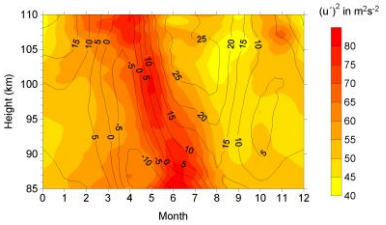
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)









To extract knowledge➔ all these types of data need to linked together

Velocity (Speed)

- Data generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

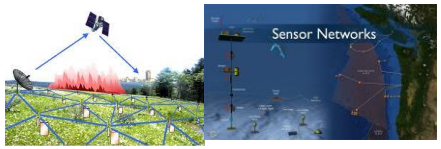
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

It's not just about the data…

It is important to understand the distinction between Big Data sets (large, unstructured, fast, and uncertain data) and 'Big Data Analytics'.

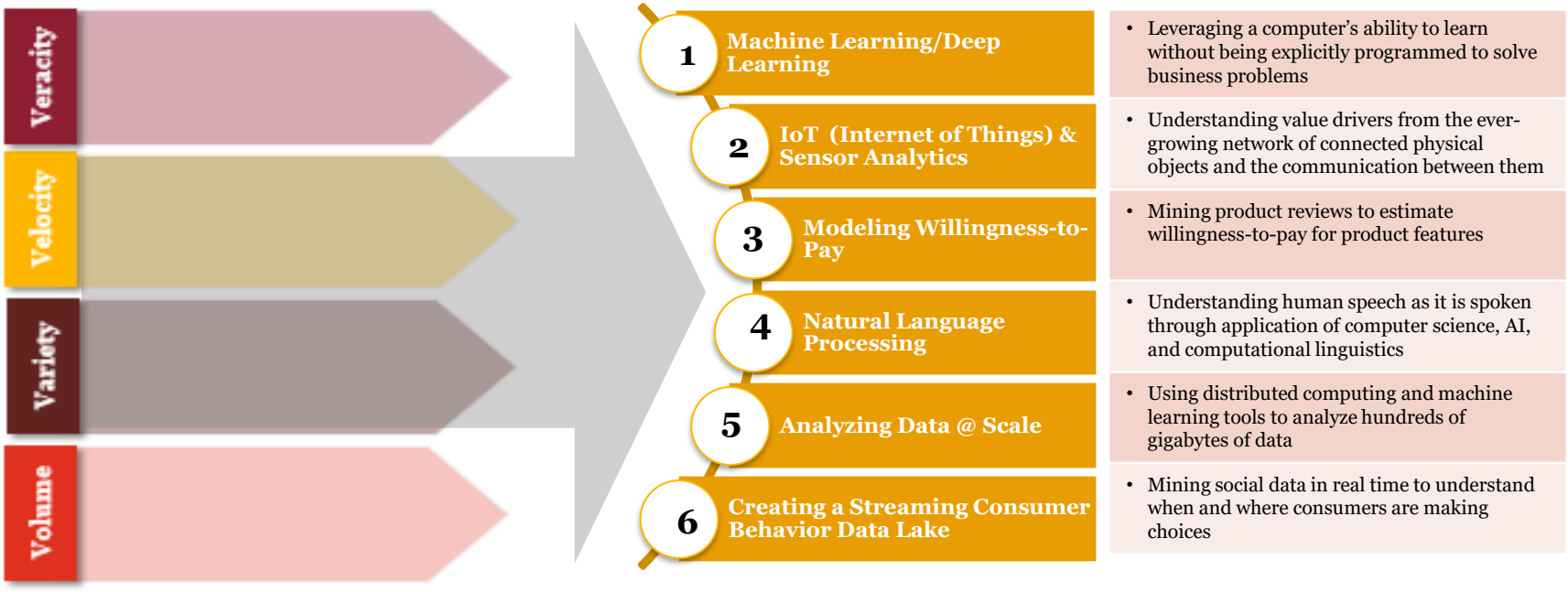
Big ***Data***

Refers to the DATA only

+

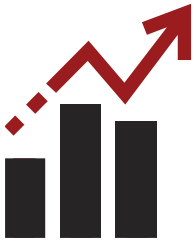
Big ***Data Analytics***

Methods of using Big Data to generate insight



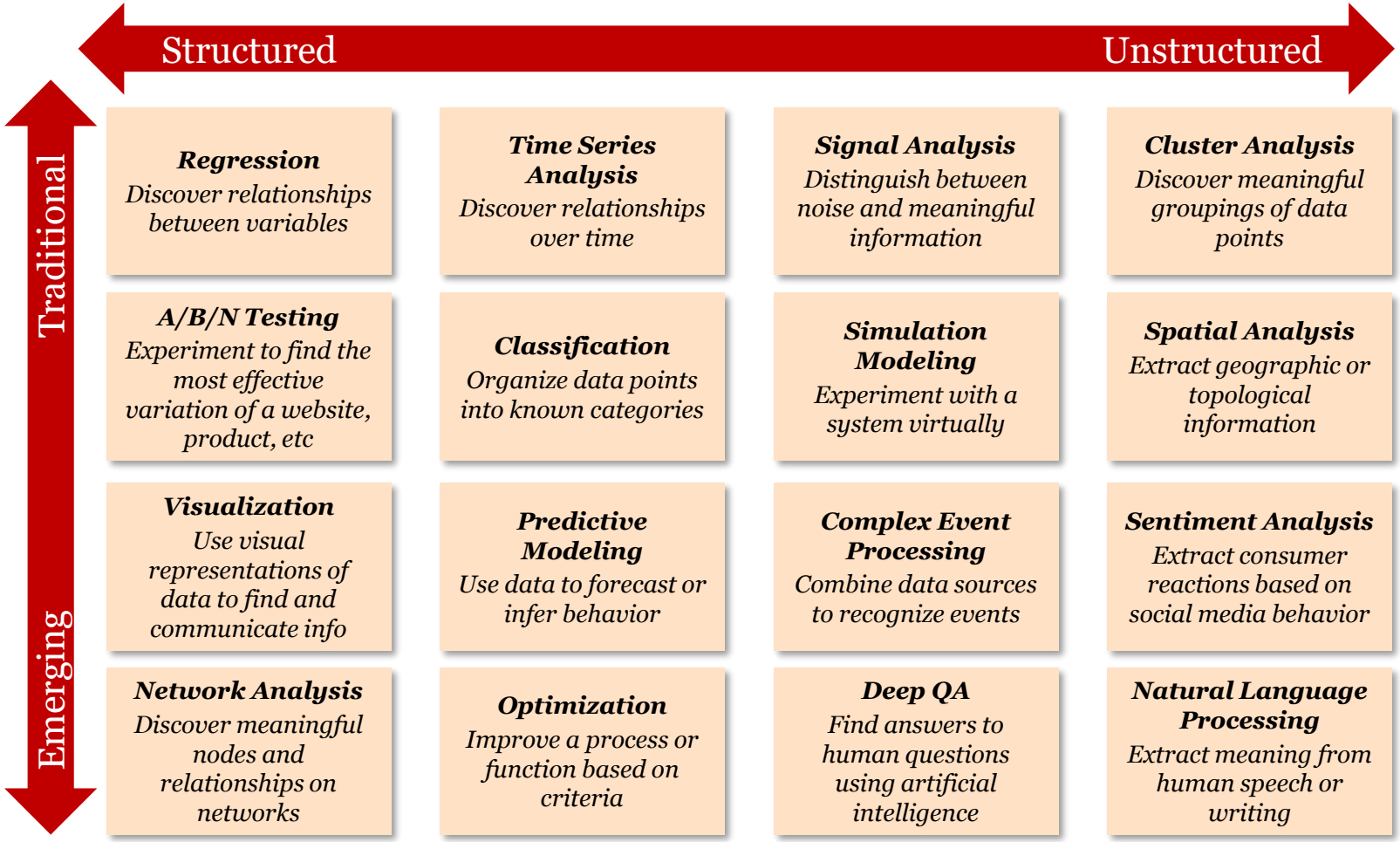
… It's also about what, how, and why you use it

Big Data Analytics – the process of harnessing Big Data to yield actionable insights – is a combination of five key elements:

Decisions	*Analytics*	*Data*	*Technology*	*Mindset & Skills*
The value of Big Data Analytics is driven by the unique decisions facing leaders, companies, and countries today. In turn, the type, frequency, speed, and complexity of decisions drive how Big Data Analytics is deployed.	To leverage the variety and volume of Big Data while managing its volatility, advanced analytical approaches are necessary, such as natural language processing, network analysis, simulative modeling, artificial intelligence, etc.	Big Data Analytics is about operationalizing new and more data, but it is also about data quality, data interoperability, data disaggregation, and the ability to modularize data structures to quickly absorb new data and new types of data.	To store, manage, and use Big Data often requires investments in new technologies and data processing methods, such as distributed processing (e.g., Hadoop), NoSQL storage, and Cloud computing.	Big Data Analytics requires firm commitment to using analytics in decision-making; a decisive mentality capable of employing in-the-moment intelligence; and investment in analytical technology, resources, and skills.
				

Big Data Analytical Capabilities

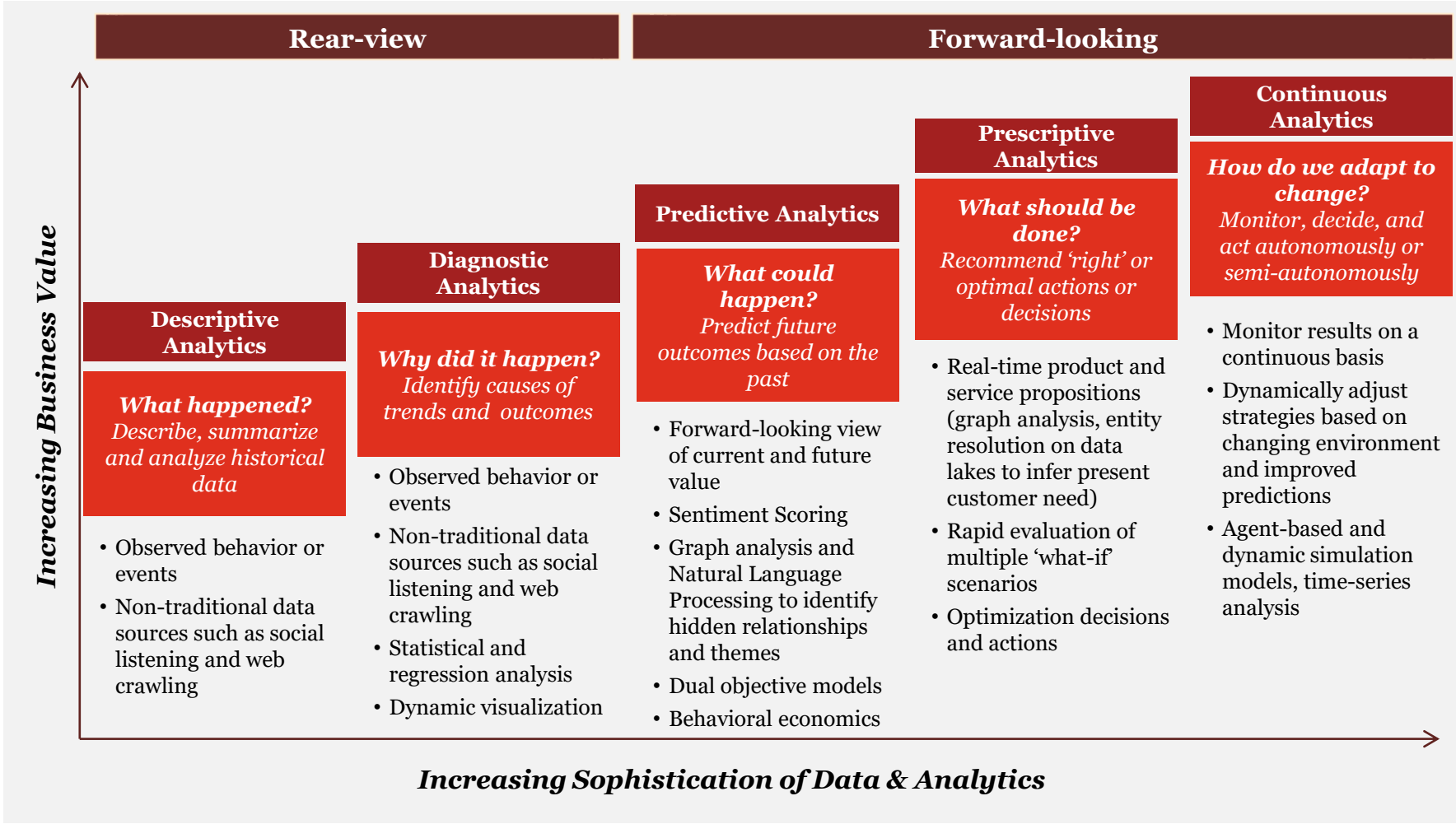
Continuing increases in processing capacity have opened the door to a range of advanced algorithms and modeling techniques that can produce valuable insights from Big Data.



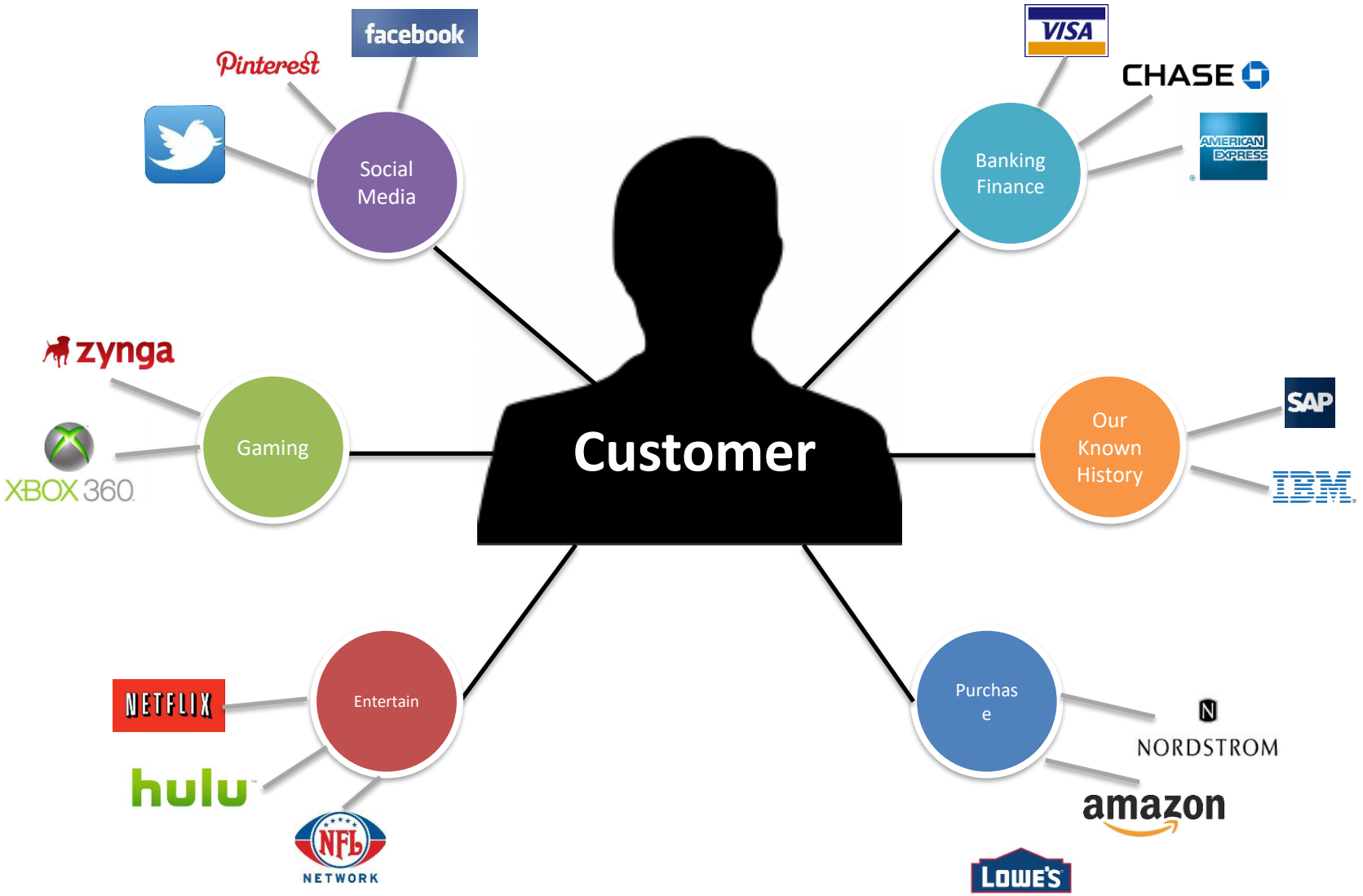
* *For more information on these analytic methods, see Appendix.*

Forward-Looking vs. Rear-View Analytics

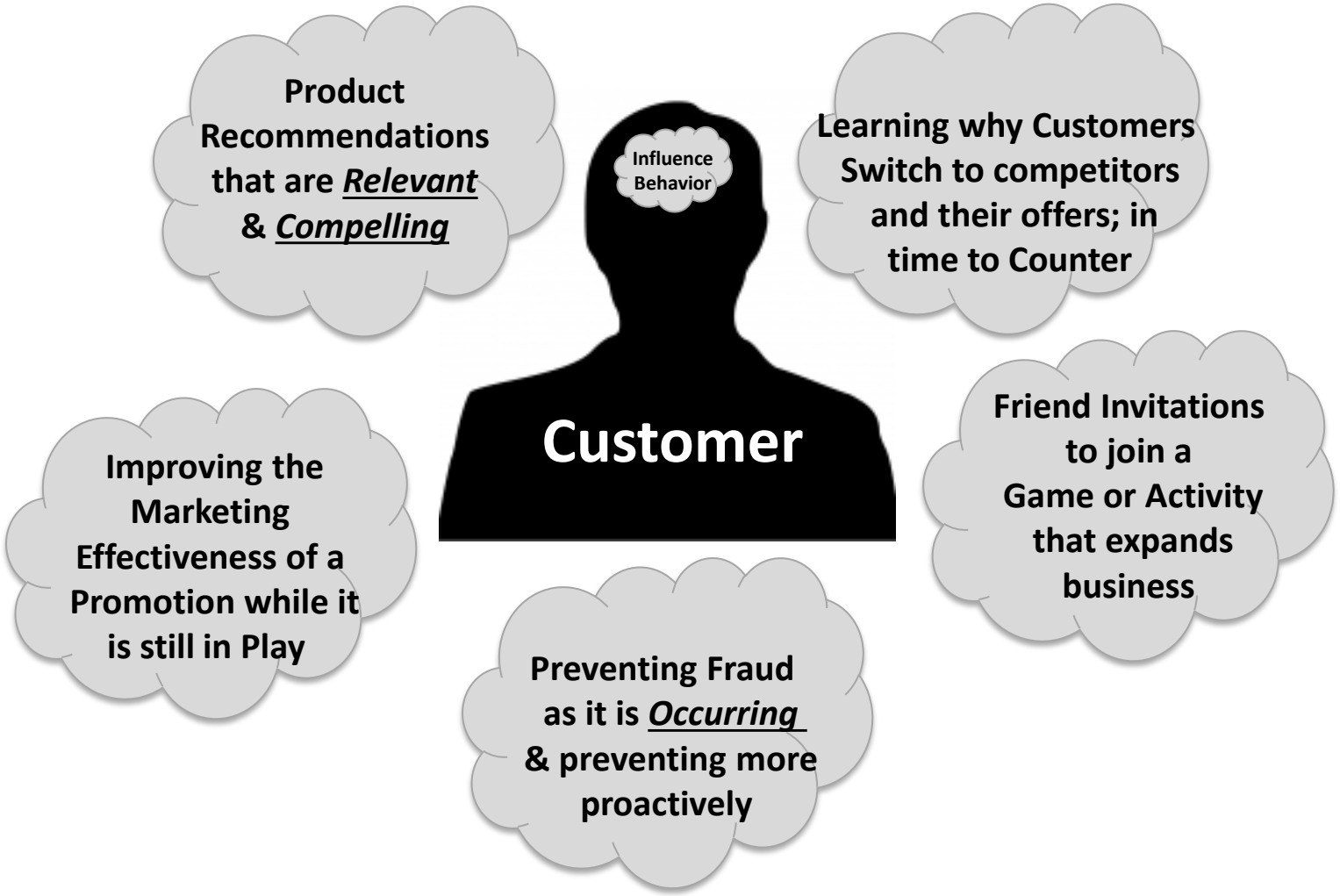
Big Data Analytics improves the speed and efficiency with which we understand the past, and opens up entirely new avenues for preparing for and adapting to the future.



A Single View to the Customer

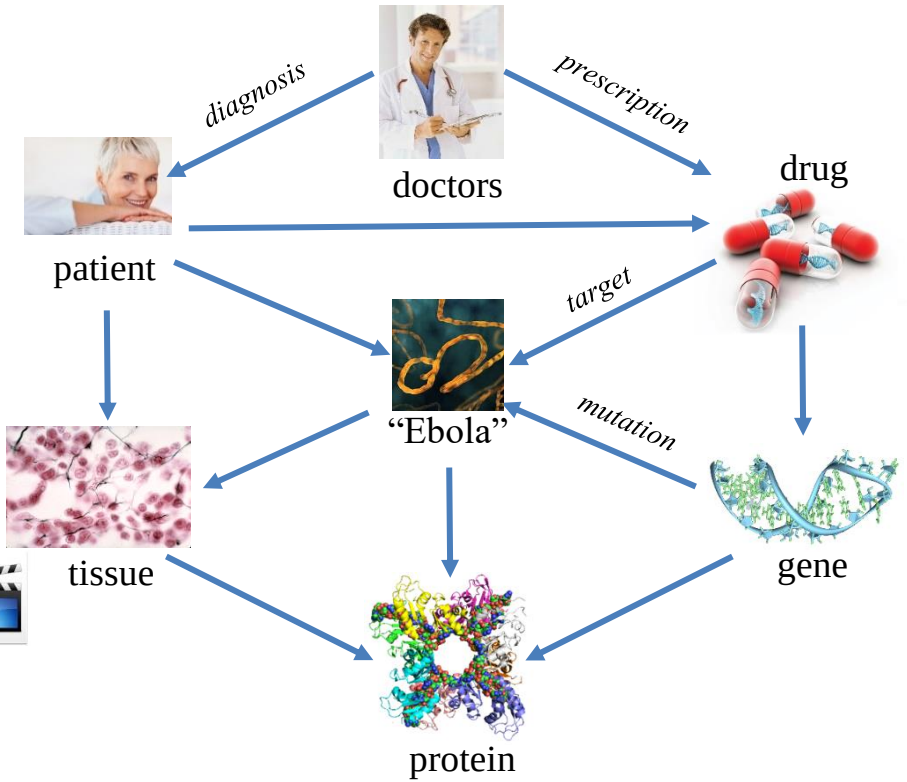


Real-Time Analytics/Decision Requirement



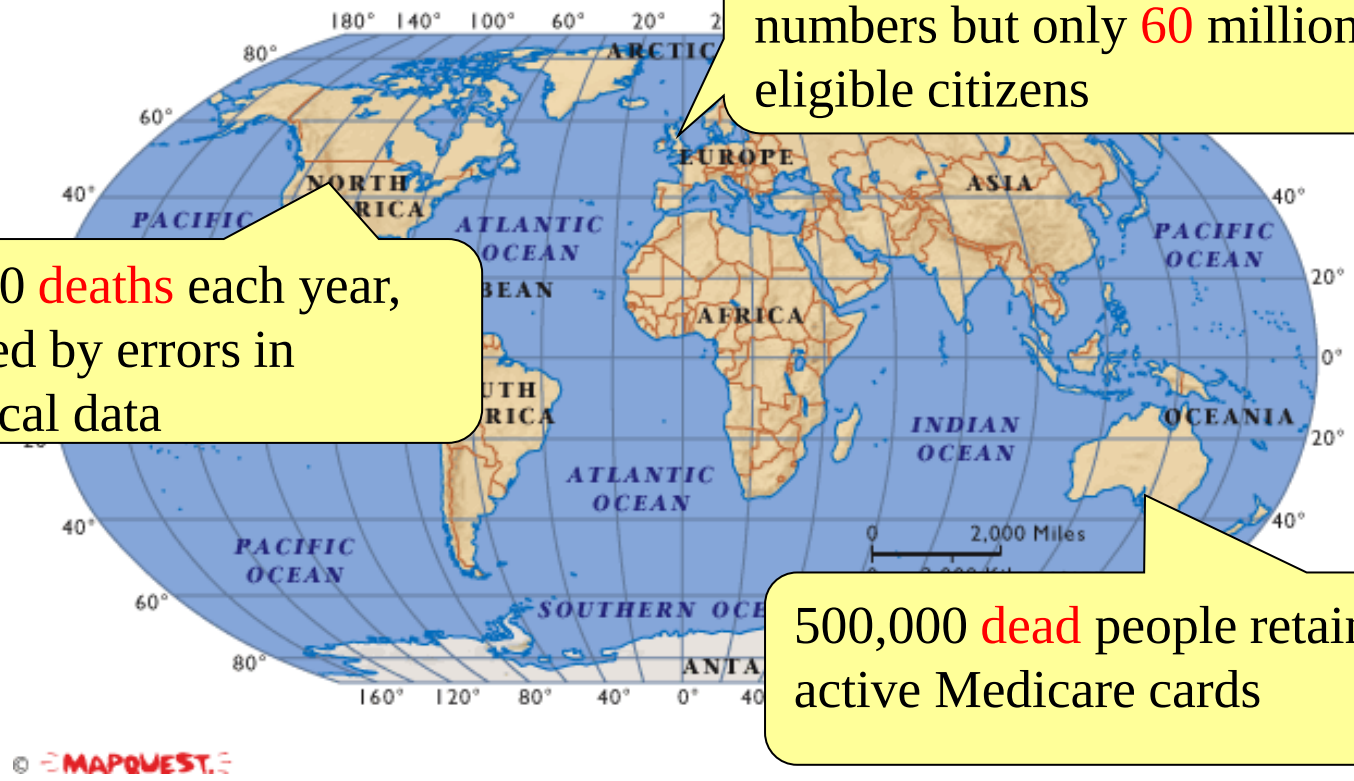
A Global View of Linked Big Data

Diversified social network



Heterogeneous information network

Data in real-life is often dirty



Visibility/Visualization

- Visible to the process of big data management
- Big Data – visibility = Black Hole?

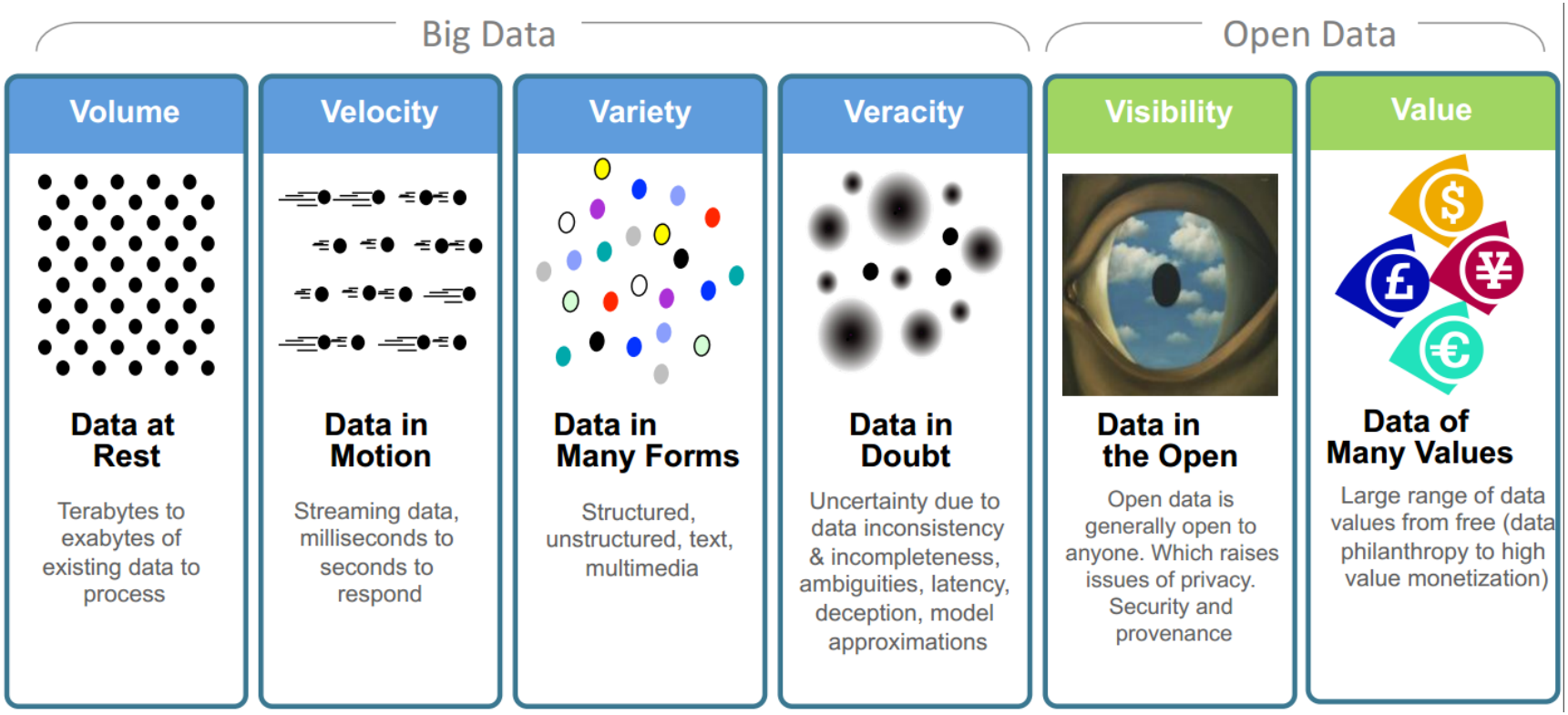
[A visualization of Divvy bike rides across Chicago]

- Big da

Value

- Big data is meaningless if it does not provide value toward some meaningful goal

Big Data: 6V in Summary



Transforming Energy and Utilities through Big Data & Analytics. By Anders Quitzau@IBM

8Vs of Big Data

Other V's

- Variability
  - Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing.
- Viscosity
  - This term is sometimes used to describe the latency or lag time in the data relative to the event being described. We found that this is just as easily understood as an element of Velocity.
- Virality
  - Defined by some users as the rate at which the data spreads; how often it is picked up and repeated by other users or events.
- Volatility
  - Big data volatility refers to how long is data valid and how long should it be stored. You need to determine at what point is data no longer relevant to the current analysis.
- More V's in the future …

Big Data Tag Cloud

Why Study Big Data Technologies?

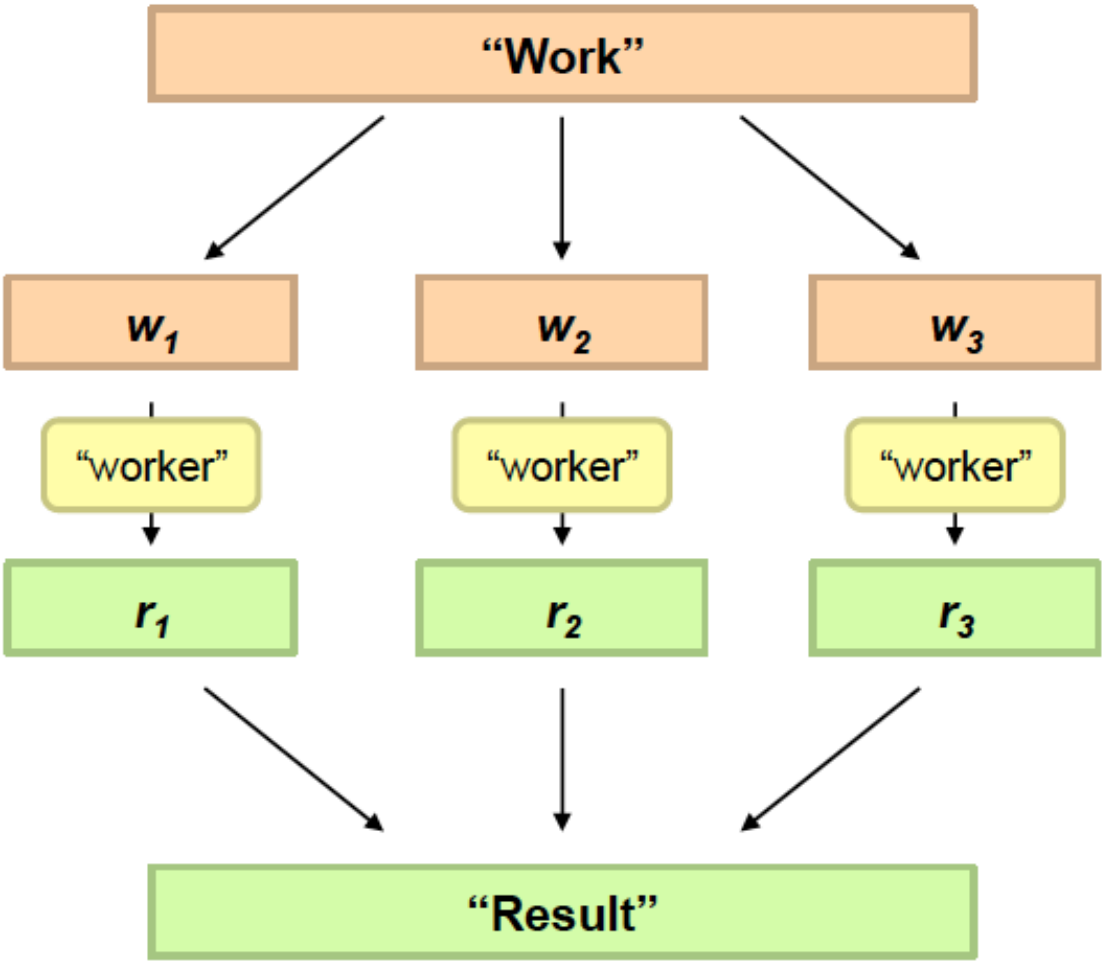
- The hottest topic in both research and industry
- Highly demanded in real world
- A promising future career
  - Research and development of big data systems:

    distributed systems (eg, Hadoop), visualization tools, data warehouse, OLAP, data integration, data quality control, …
  - Big data applications:

    social marketing, healthcare, …
  - Data analysis: to get values out of big data

    discovering and applying patterns, predicative analysis, business intelligence, privacy and security, …

Philosophy to Scale for Big Data Processing



Divide Work

Combine Results

Distributed processing is non-trivial

- How to assign tasks to different workers in an efficient way?
- What happens if tasks fail?
- How do workers exchange results?
- How to synchronize distributed tasks allocated to different workers?

Big data storage is challenging

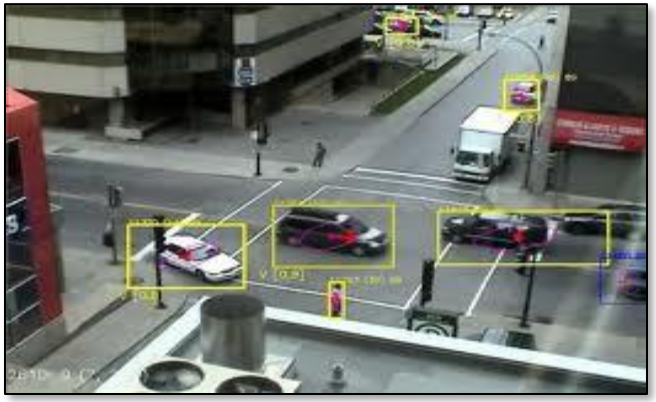
- Data Volumes are massive
- Reliability of Storing PBs of data is challenging
- All kinds of failures: Disk/Hardware/Network Failures
- Probability of failures simply increase with the number of machines …

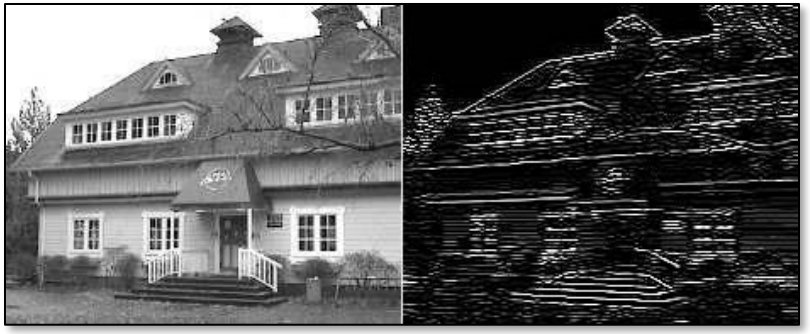
Image Analytics Overview

How can we extract insight from images and video?

Overview

- The process of pulling relevant information from an image or sets of images for advanced classification and traditional analysis
- Applies image capture, image processing, and machine learning techniques to extract, quantify, and structure, image information





Advantages

- Provides a method to structure, organize, and search information that is stored within images
- Offers an additional data set that can be applied to understanding consumer behavior, automating business processes, and discovering knowledge enterprise content

Image Analytics Tools

There are few standalone packages that are capable of performing robust image analysis; however, solutions can be developed using existing frameworks and analytics toolkits

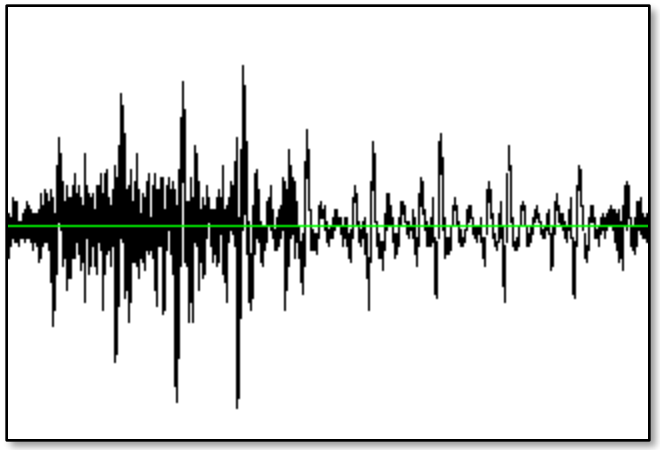
Tool	Overview	Image Processing	Computer Vision	Machine Learning
OpenCV	Open source library of computer vision functions that is accessible via C, Java, and Python	X	X	X
PAXit Image Analysis	Integrated image analysis platform that provides basic feature identification functions	X	X	
ImageJ	Java based image processing platform that can be accessed via an API and expanded with custom plugins	X		
PIL	Python image processing library	X		
PyBrain	A modular machine learning library for Python			X

Audio Analytics Overview

How can we extract insight from audio and voice media?

Overview

- The process of capturing audio and analyzing its features as to extract content and context of an event
- Applies speech analysis and signal processing principles to structure audio information for analysis via NLP or traditional analytics techniques



Advantages

- Provides a method for identifying events or common patterns within sound bytes
- Offers a way of capturing not only the content and topics within a conversation, but also the emotions and context

Audio Analytics Applications

	Analysis	Objectives
Voice Recognition	Analyze conversations to capture speech as text based dialog	• Capture and structure the content of conversations • Utilize structured speech as an input to text mining and natural language processing capabilities • Combine phone based conversations with other interaction data sets
Sound Matching	Analyze sound clips to identify specific events taking place	• Monitor customer interactions or business operations to capture events in real time • Use captured events for comparison, categorization and analysis with other data points
Sentiment Analysis	Monitor phone calls with customers to uncover sentiment towards the experience and/or products/services	• Capture the content of the conversation and conducting sentiment analysis based on word choice • Analyze the pitch, loudness, and rate of consumer speech to identify emotional state during the conversation and its cause
Employee /Customer Screening	Monitor customer and job candidate conversations to extract information from word usage and speech patterns that can inform or improve a screening process	• Analyze prescreen phone conversations to assess job candidate personality, interest in job, and fit to job requirements • Analyze customer conversations to assess level of risk and honestly when applying for a product or filing claim/complaints

Audio Analytics Tools

There are few tools on the market that provide a broad range of audio analysis capabilities. However, basic audio analysis and natural language tool kits can be combined for robust analytics

Tool	Overview	Audio Processing	Information Retrieval
Clam	A C++ library that provides varying level of audio processing and information retrieval capabilities	X	X
CallMiner	A tool that is capable of translating calls to a more structured text data set and combining with other communication forms		X
Nuance	Logs calls and structures audio for text based search and retrieval		X
yaafe	Aduio feature extraction toolkit with wrappers for several languages		X
PRAAT	Multiple platform audio analysis toolkit	X	

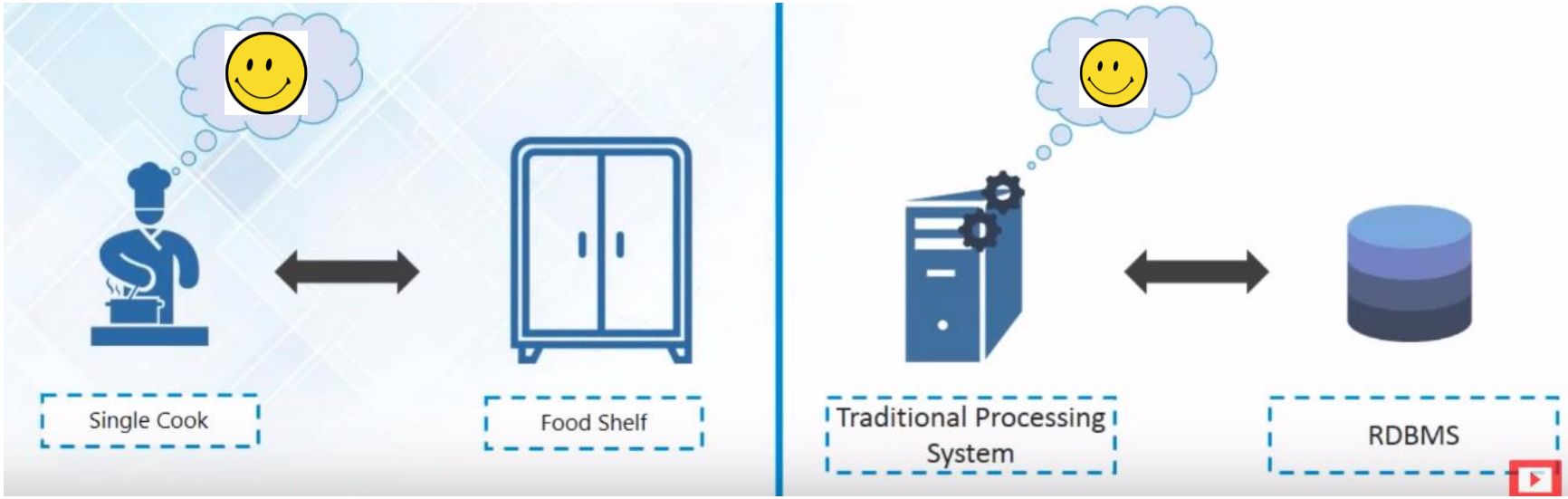
Story of Big Data and Traditional systems

Scenario:
Bob has opened a small restaurant in his city

Story of Big Data and Traditional systems

Traditional Scenario:
2 orders per hour

Traditional Scenario:
Data is generated at a steady rate and is structured in nature



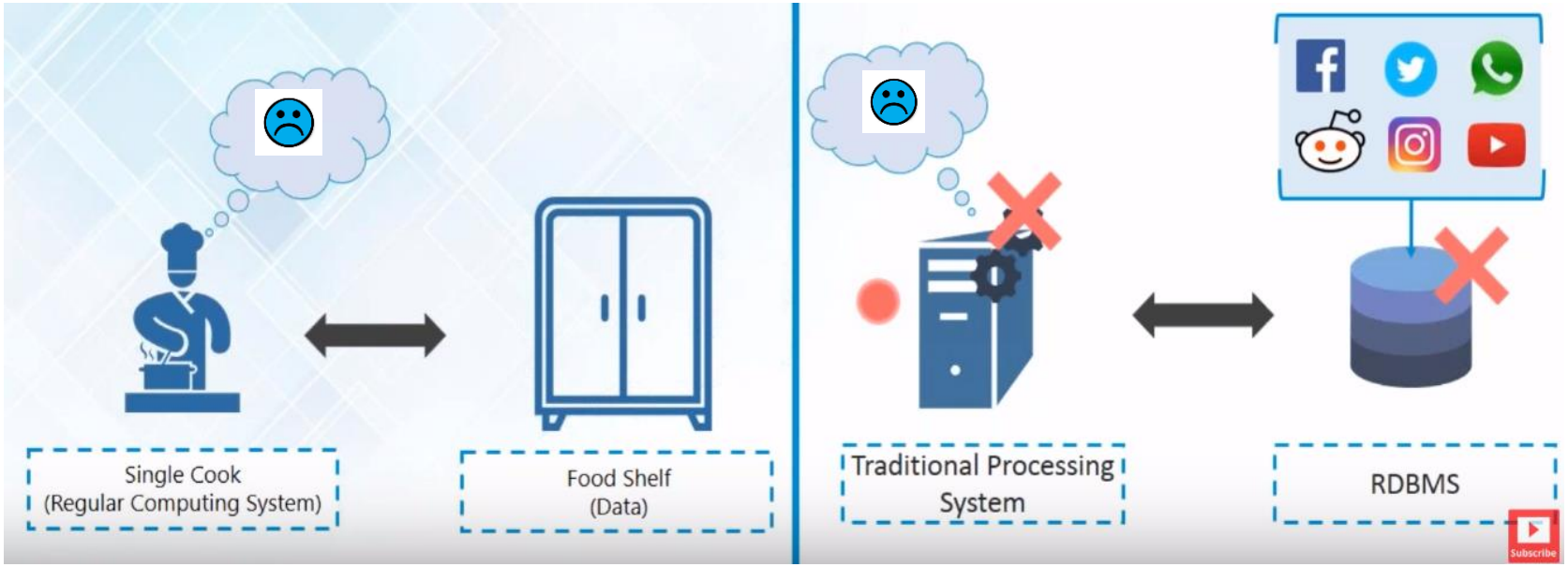
Failure of Traditional System

Scenario 2:

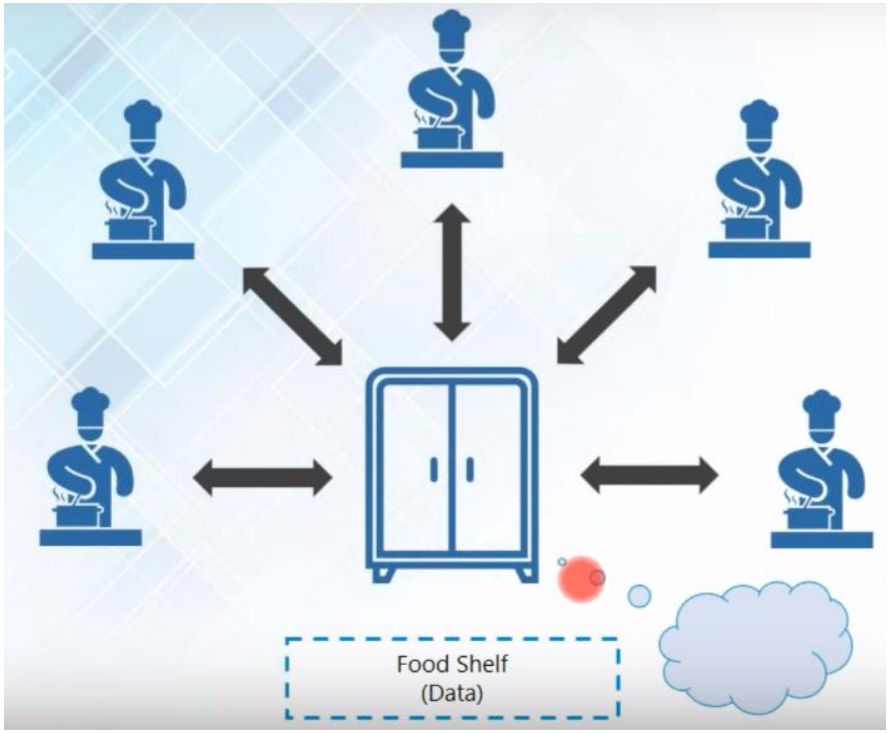
- They started taking online orders
- 10 orders per hour

Big Data Scenario :

- Heterogeneous data is being generated at an alarming rate by multiple sources



Need an effective Solution



Scenario 2:

- Multiple cook cooking food

Issues:

- A bottleneck at the food shell

Need an effective Solution

Scenario

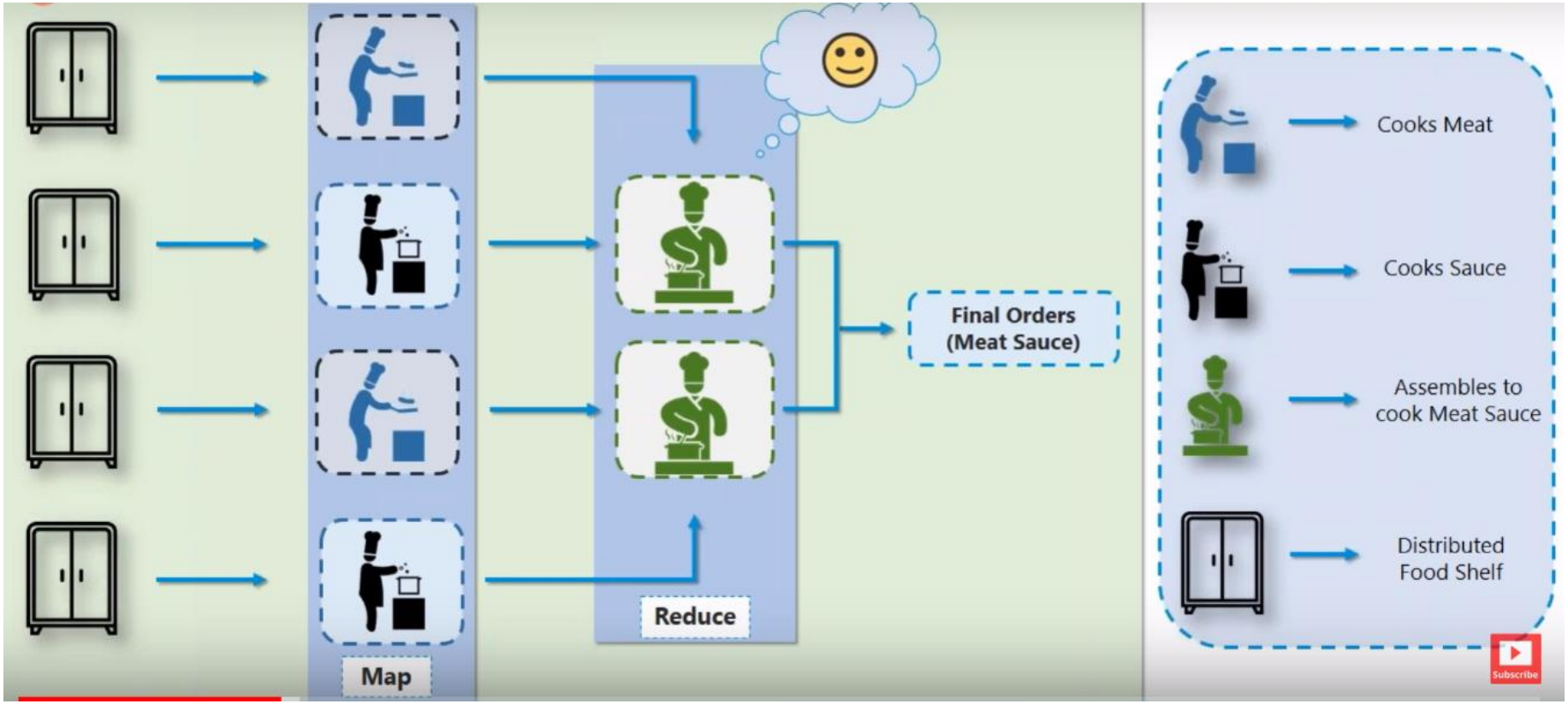
- Multiple Processing Unit for data processing

Issues:

- Bringing data to processing generated lots of Network overhead

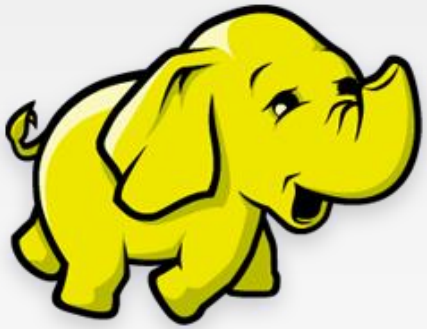
Solution : Distributed and Parallel approach

Effective solution



What is Hadoop

- Open-source data storage and processing platform
- Before the advent of Hadoop, storage and processing of big data was a big challenge
- Massively scalable, automatically parallelizable
  - Based on work from Google
    - Google: GFS + MapReduce + BigTable (Not open)
    - Hadoop: HDFS + Hadoop MapReduce + HBase（opensource)
- Named by Doug Cutting in 2006 (worked at Yahoo! at that time), after his son's toy elephant.



Hadoop offers

- Redundant, Fault-tolerant data storage
- Parallel computation framework
- Job coordination

Programmers

No longer need to worry about

Q: Where file is located?

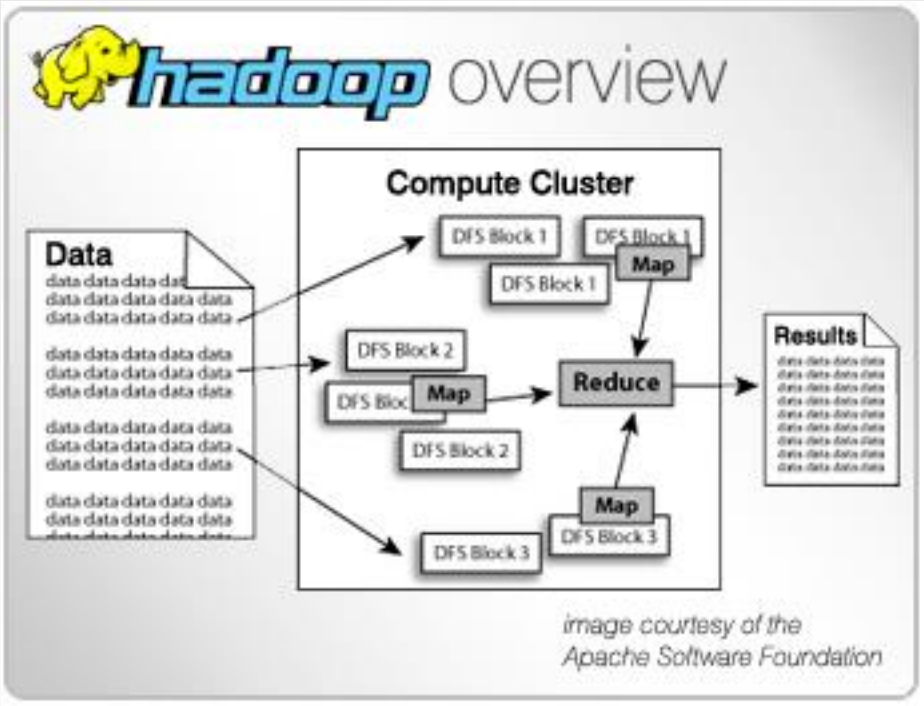
Q: How to handle failures & data lost?

Q: How to divide computation?

Q: How to program for scaling?

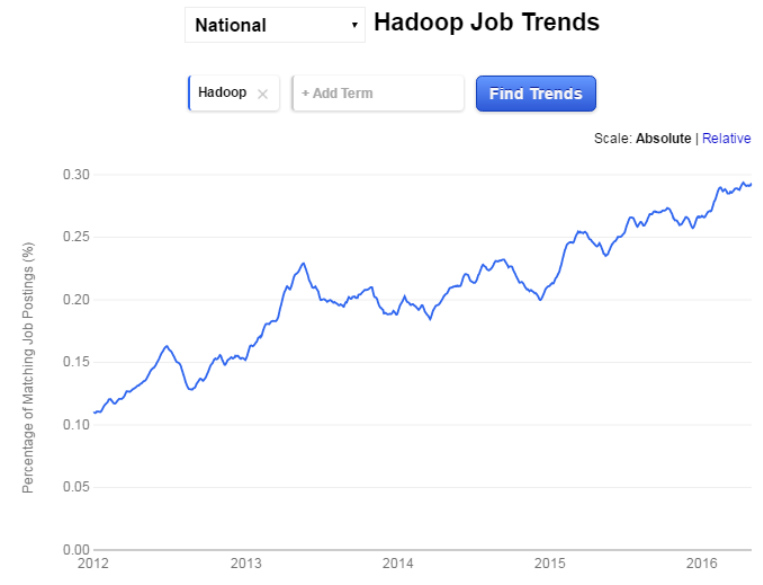
Why Use Hadoop?

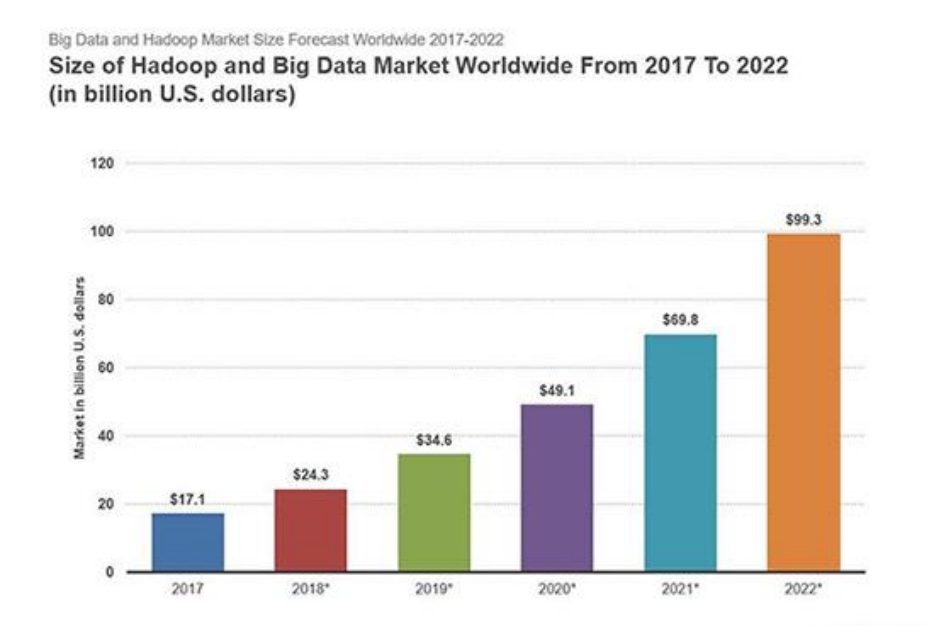
- Cheaper
  - Scales to Petabytes or more easily
- Faster
  - Parallel data processing
- Better
  - Suited for particular types of big data problems



Companies Using Hadoop

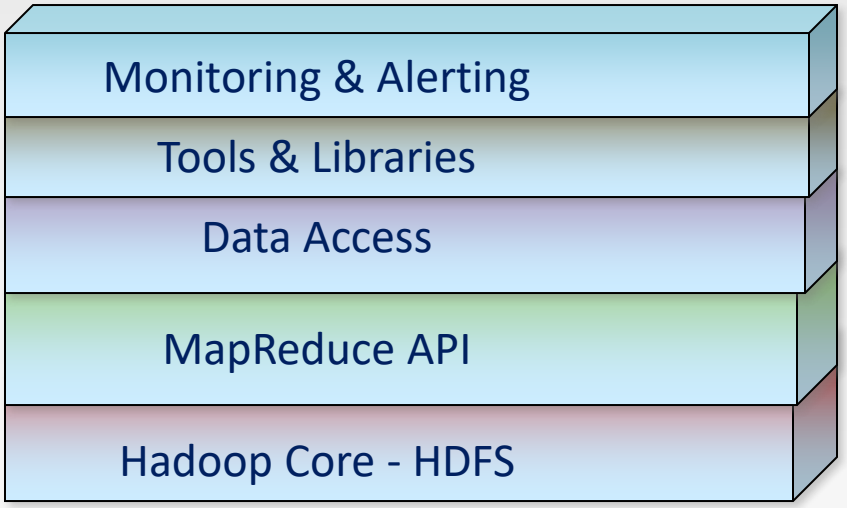
Forecast growth of Hadoop Job Market



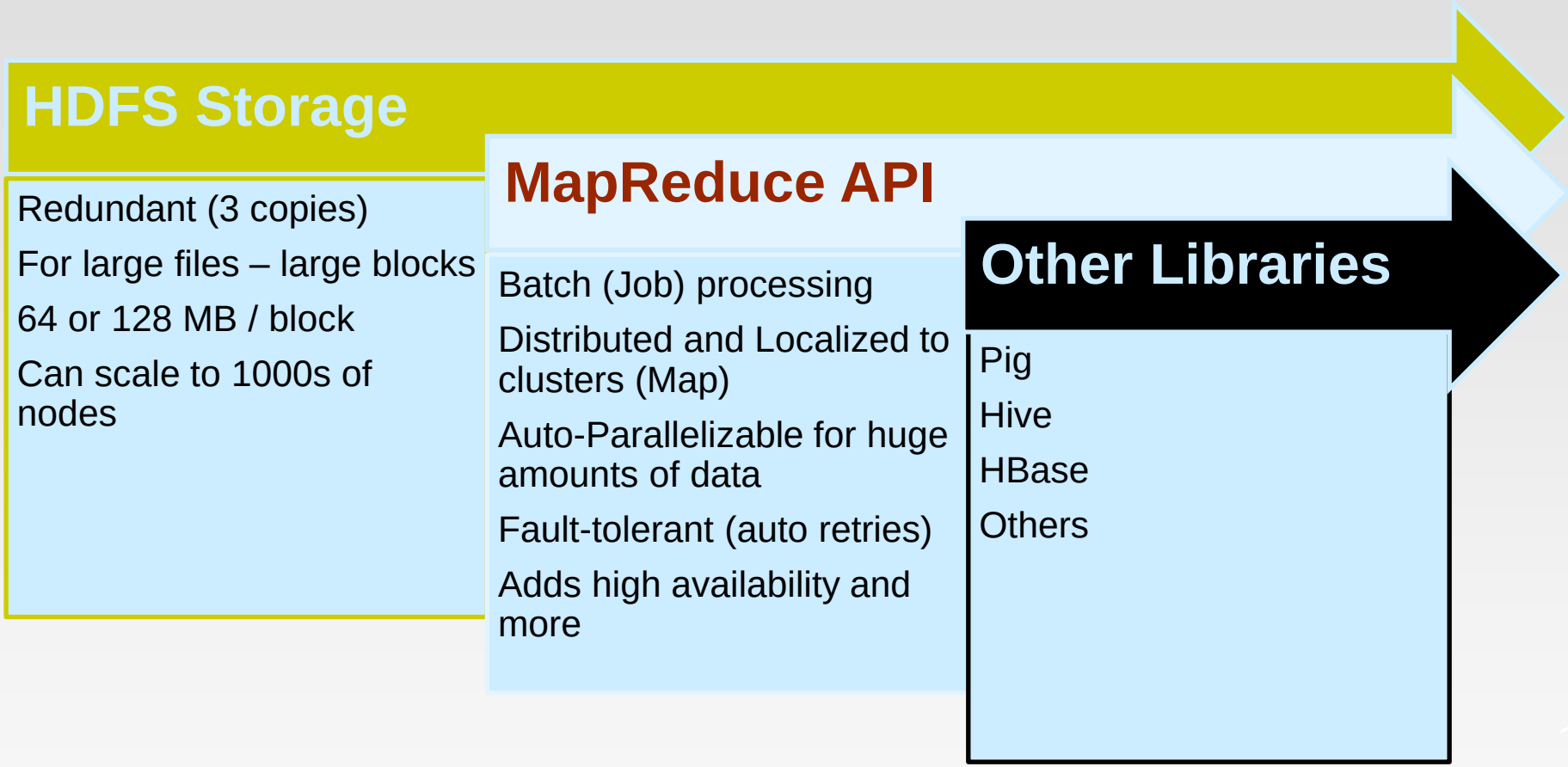


Hadoop is a set of Apache Frameworks and more…

- Data storage (HDFS)
  - Runs on commodity hardware (usually Linux)
  - Horizontally scalable
- Processing (MapReduce)
  - Parallelized (scalable) processing
  - Fault Tolerant
- Other Tools / Frameworks
  - Data Access
    - HBase, Hive, Pig, Mahout
  - Tools
    - Hue, Sqoop
  - Monitoring
    - Greenplum, Cloudera

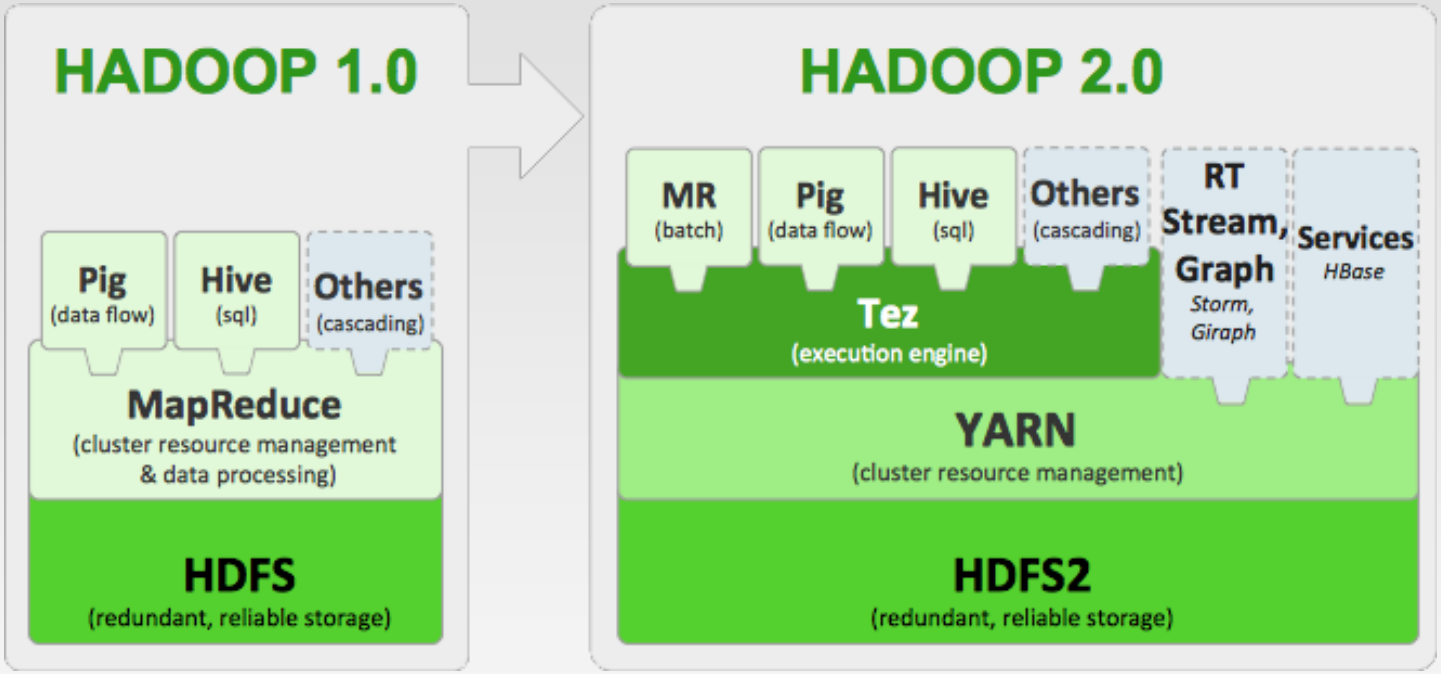


What are the core parts of a Hadoop distribution?



Hadoop 2.0

- Single Use System
  - Batch apps
- Multi-Purpose Platform
  - Batch, Interactive, Online, Streaming



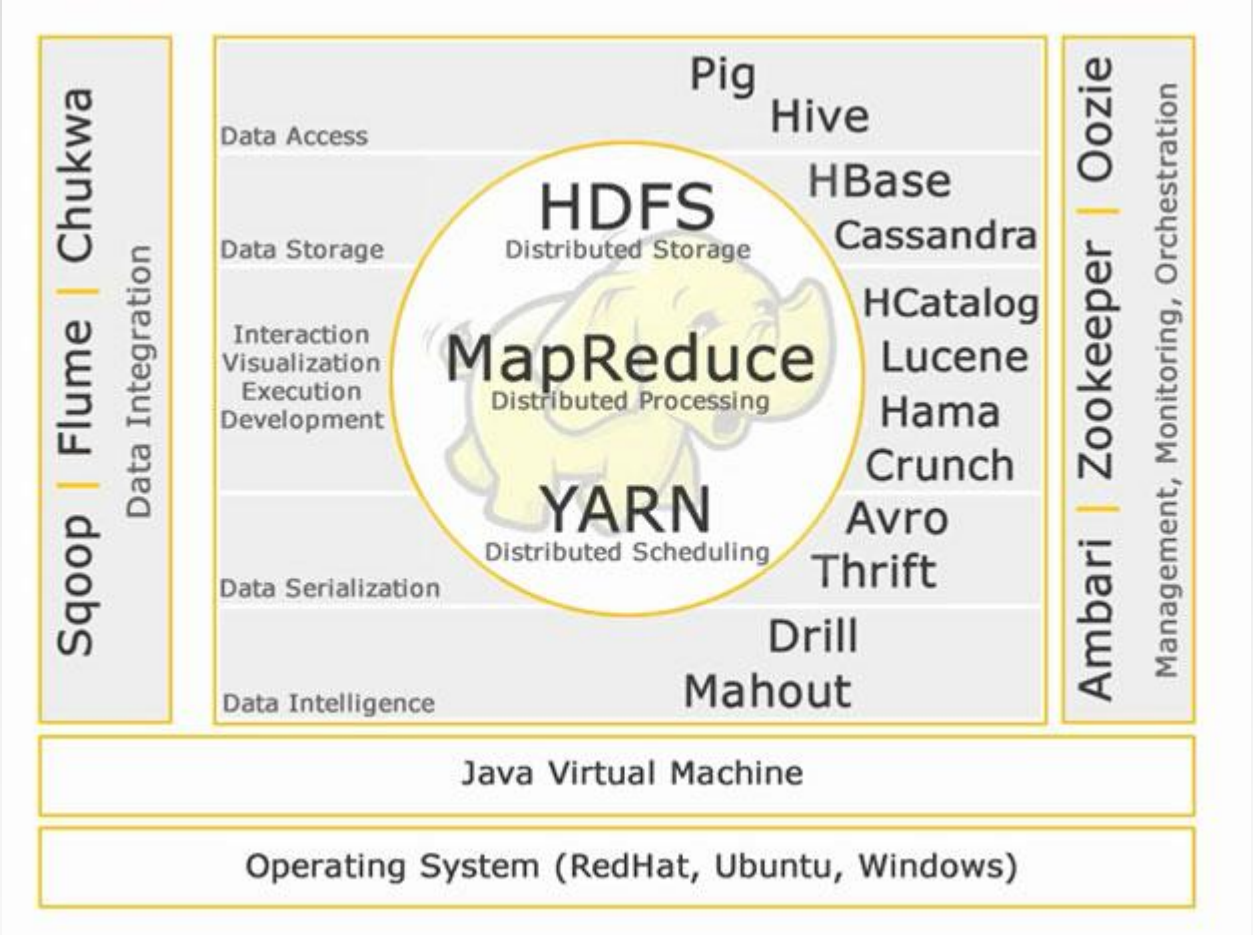
Hadoop YARN *(Yet Another Resource Negotiator)*: a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications

Comparison of Hadoop 1.x vs Hadoop 2.x

Hadoop 1.x	Hadoop 2.x
It supports only MapReduce tools	It supports more than the basic MapReduce tool like spark, Hbase etc.. i.e. other distributed computing models
Maximum limit is 4000 nodes per cluster	Scalable up to 10000 nodes per cluster
Works on slots. Each slot can run either Map or Reduce task	Works on Containers. In Containers they can do generic tasks.
Single Namenode is used	Multiple Namenodes may be used
Supports on Linux System	It supports Microsoft Windows
No additional files are needed to execute a MapReduce program	To run program of Hadoop 1.x in 2.x we need additional files
MapReduce does both processing and the cluster resource management	Cluster resource management is done by YARN and processing is done by another processing models
Failure of name node affects the stack	Hive, Pig, Hbase all are responsible for handling namenode failure.

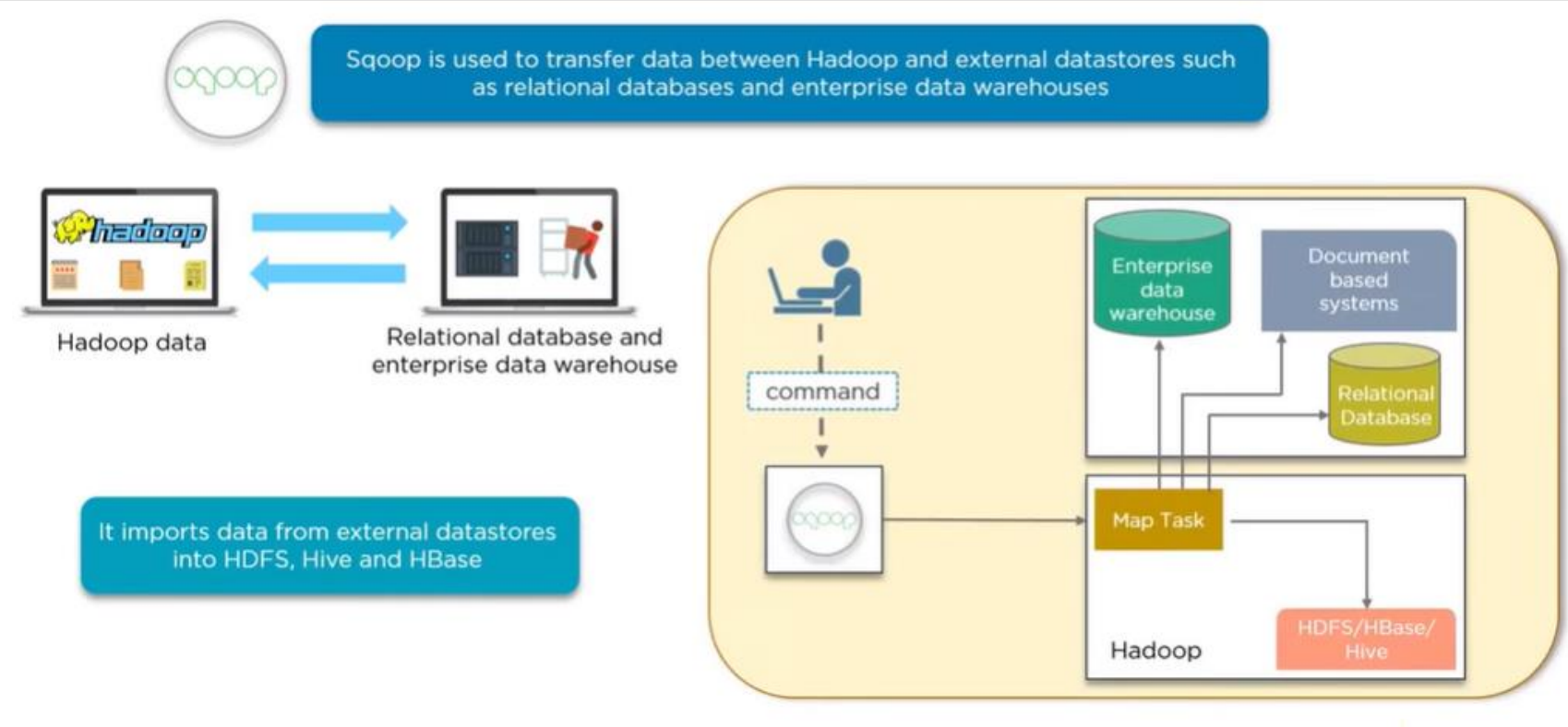
Hadoop Ecosystem

A combination of technologies which have proficient advantage in solving business problems.

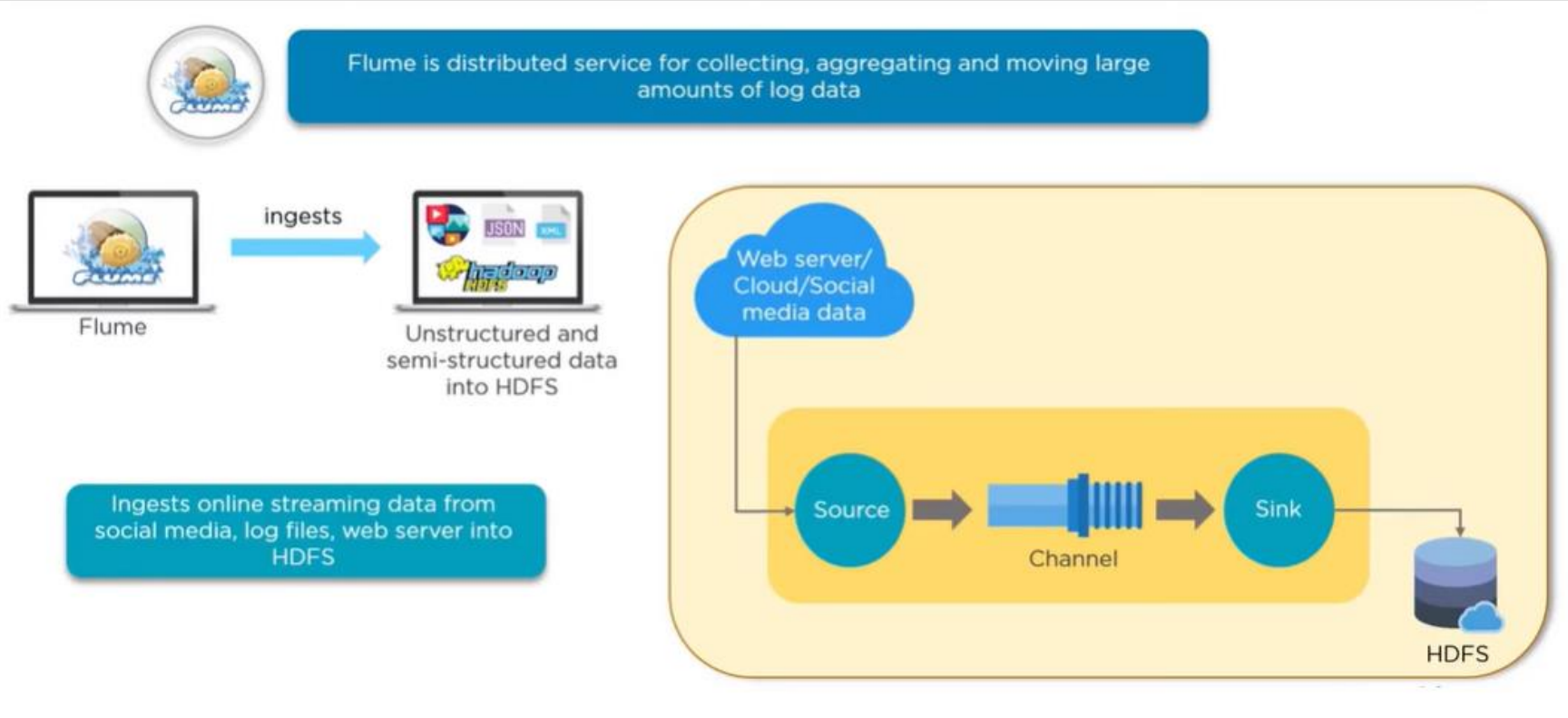


http://www.edupristine.com/blog/hadoop-ecosystem-and-components

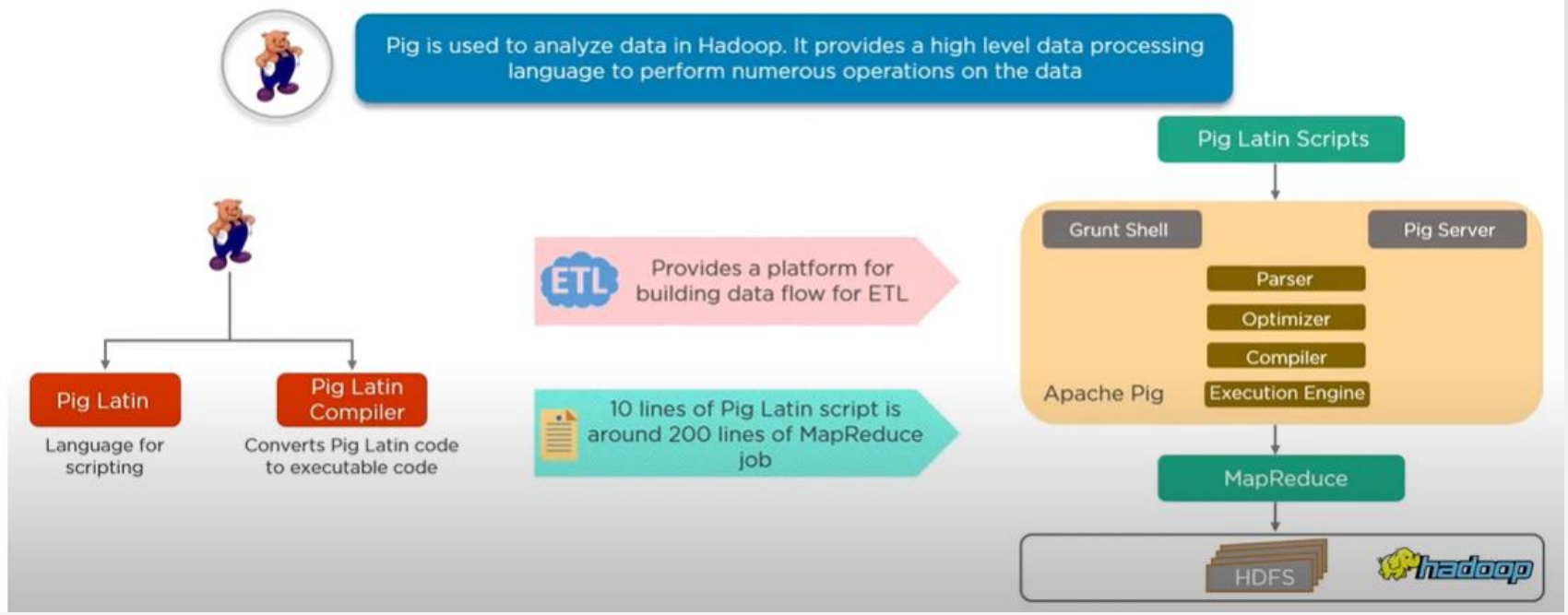
Sqoop


Sqoop is used to transfer data between Hadoop and external datastores such as relational databases and enterprise data warehouses
Hadoop data
Relational database and enterprise data warehouse
It imports data from external datastores into HDFS, Hive and HBase
command
Enterprise data warehouse
Document based systems
Relational Database
Map Task
HDFS/HBase/ Hive
Hadoop

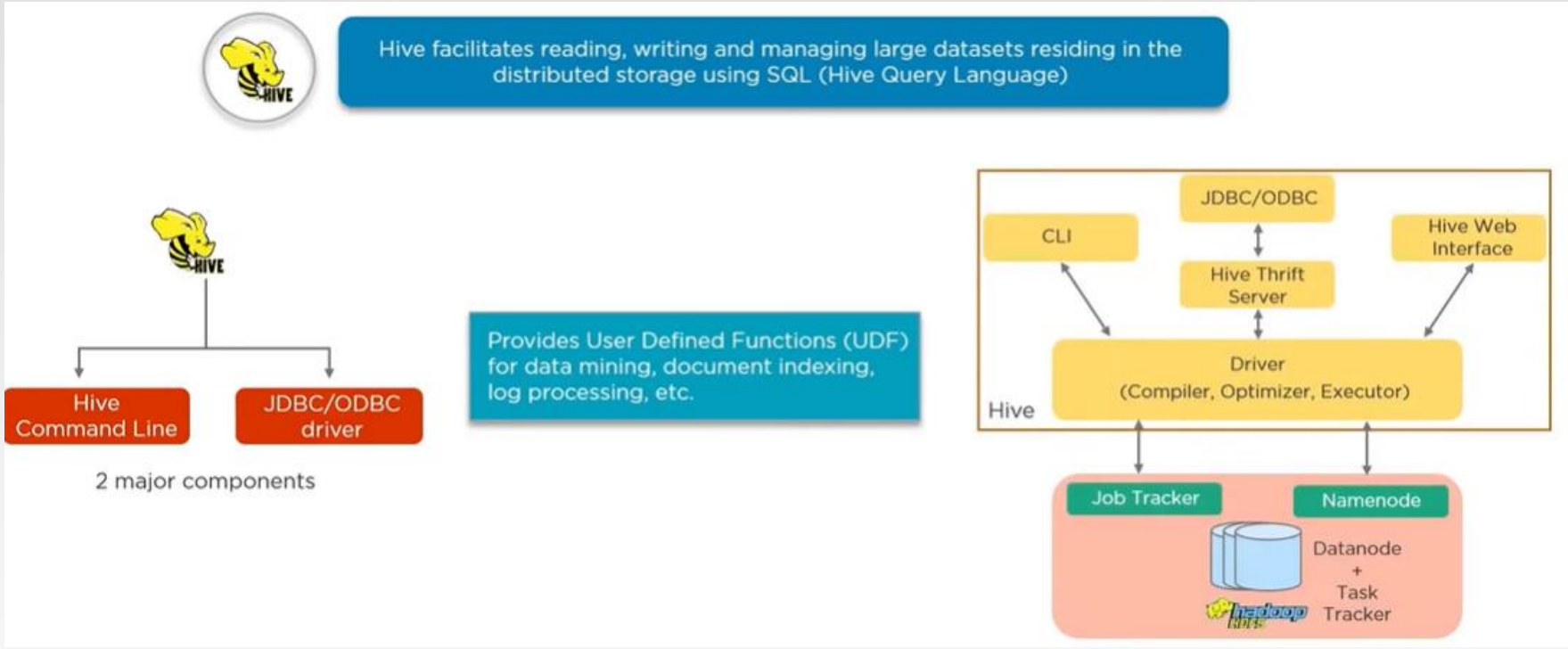
Flume


Flume is distributed service for collecting, aggregating and moving large amounts of log data
ingests
Flume
JSON
Unstructured and semi-structured data into HDFS
Ingests online streaming data from social media, log files, web server into HDFS
Web server/ Cloud/Social media data
Source
Channel
Sink
HDFS

Pig

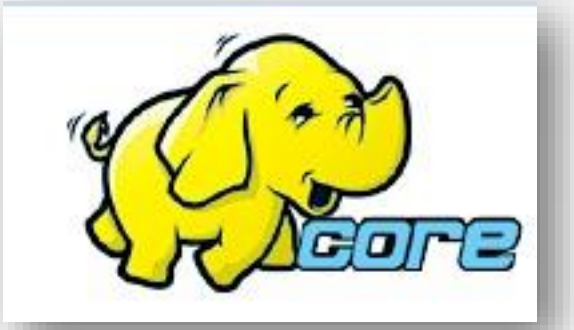

Pig is used to analyze data in Hadoop. It provides a high level data processing language to perform numerous operations on the data
Pig Latin
Language for scripting
Pig Latin Compiler
Converts Pig Latin code to executable code
ETL
Provides a platform for building data flow for ETL
10 lines of Pig Latin script is around 200 lines of MapReduce job
Pig Latin Scripts
Grunt Shell
Pig Server
Parser
Optimizer
Compiler
Apache Pig
Execution Engine
MapReduce
HDFS
hadoop

Hive


Hive facilitates reading, writing and managing large datasets residing in the distributed storage using SQL (Hive Query Language)
Hive
Command Line
JDBC/ODBC driver
2 major components
Provides User Defined Functions (UDF) for data mining, document indexing, log processing, etc.
JDBC/ODBC
CLI
Hive Web Interface
Hive Thrift Server
Driver
(Compiler, Optimizer, Executor)
Hive
Job Tracker
Namenode
Datanode
+
Task
Tracker

Common Hadoop Distributions

- Open Source
  - Apache
- Commercial
  - Cloudera
  - Hortonworks
  - MapR
  - AWS MapReduce
  - Microsoft Azure HDInsight (Beta)



Setting up Hadoop Development

Hadoop Binaries

Local install
- Linux
- Windows

Cloudera's Demo VM
- Need Virtualization software, i.e. VMware, etc…

Cloud
- AWS
- Microsoft (Beta)
- Others

Data Storage

Local
- File System
- HDFS Pseudo-distributed (single-node)

Cloud
- AWS
- Azure
- Others

MapReduce

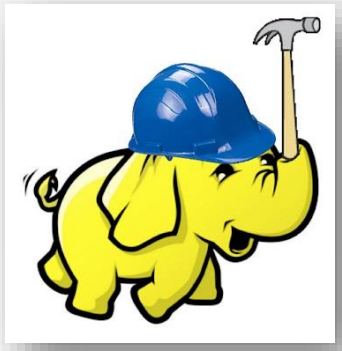
Local

Cloud

Other Libraries & Tools

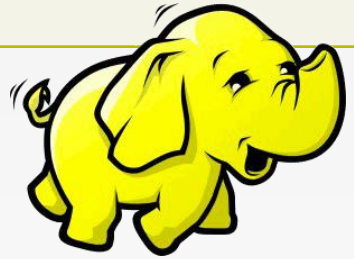
Vendor Tools

Libraries

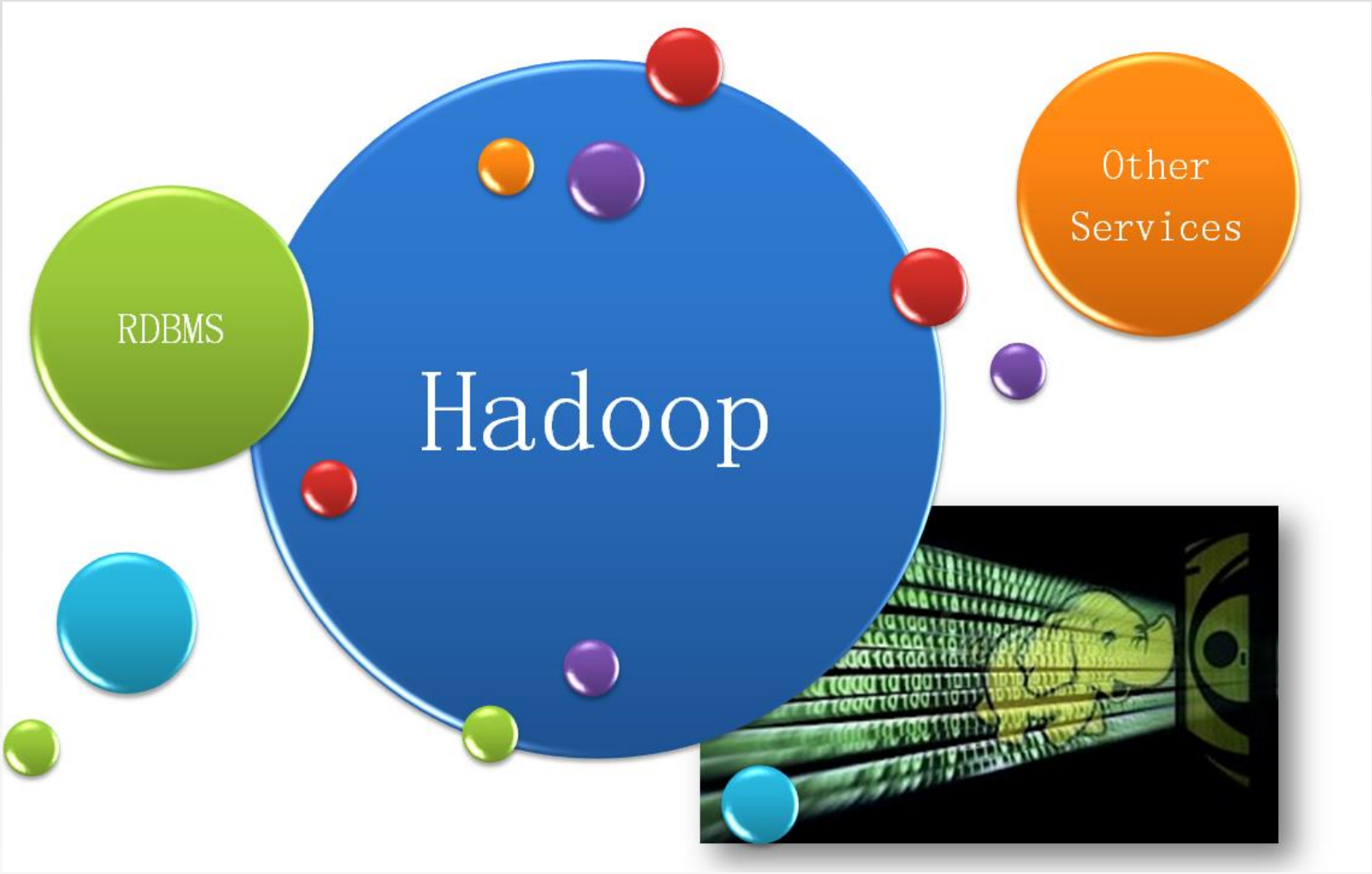


Comparing: RDBMS vs. Hadoop

	Traditional RDBMS	Hadoop / MapReduce
Data Size	Gigabytes (Terabytes)	Petabytes (Hexabytes)
Access	Interactive and Batch	Batch – NOT Interactive
Updates	Read / Write many times	Write once, Read many times
Structure	Static Schema	Dynamic Schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear
Query Response Time	Can be near immediate	Has latency (due to batch processing)

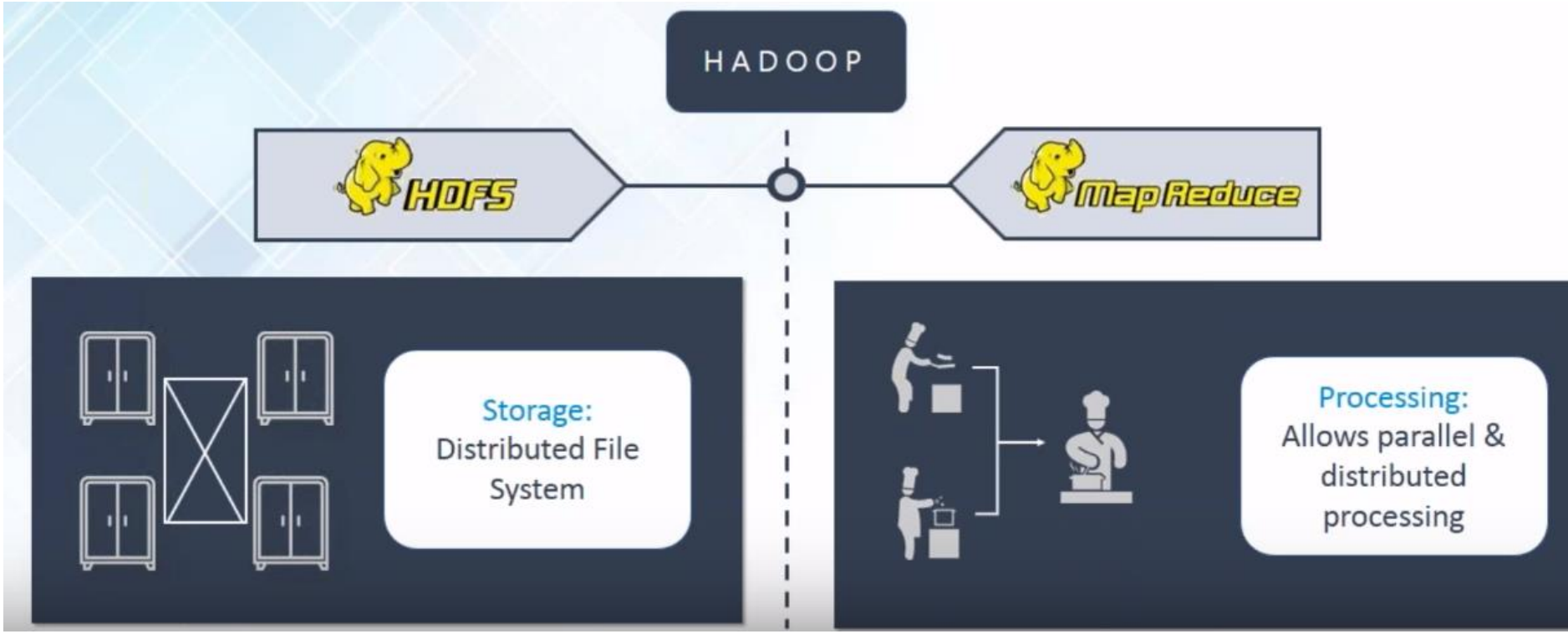


The Changing Data Management Landscape



Apache Hadoop: Framework to process big data

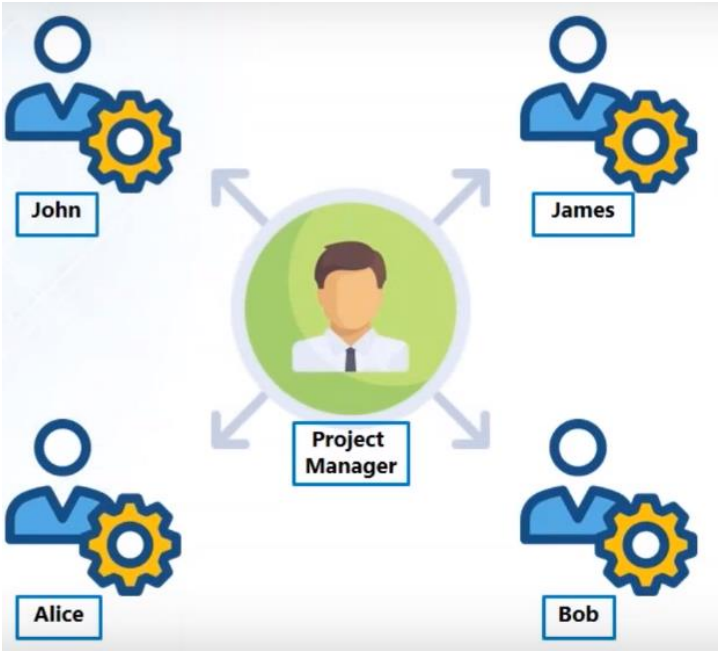
- Hadoop is a framework that allows us to store and process large data sets in parallel and distributed fashion



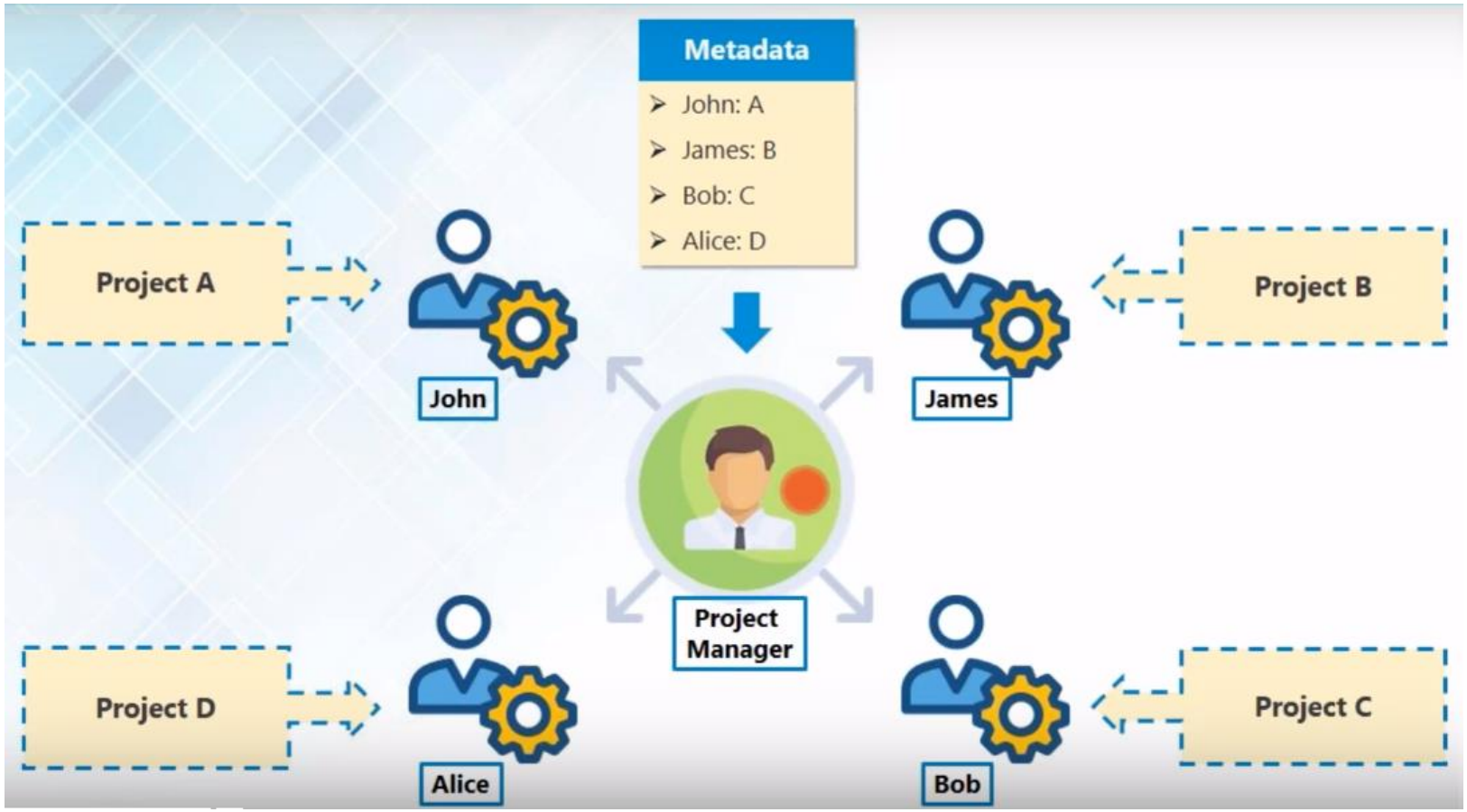
Hadoop: Master/Slave Architecture

Scenario

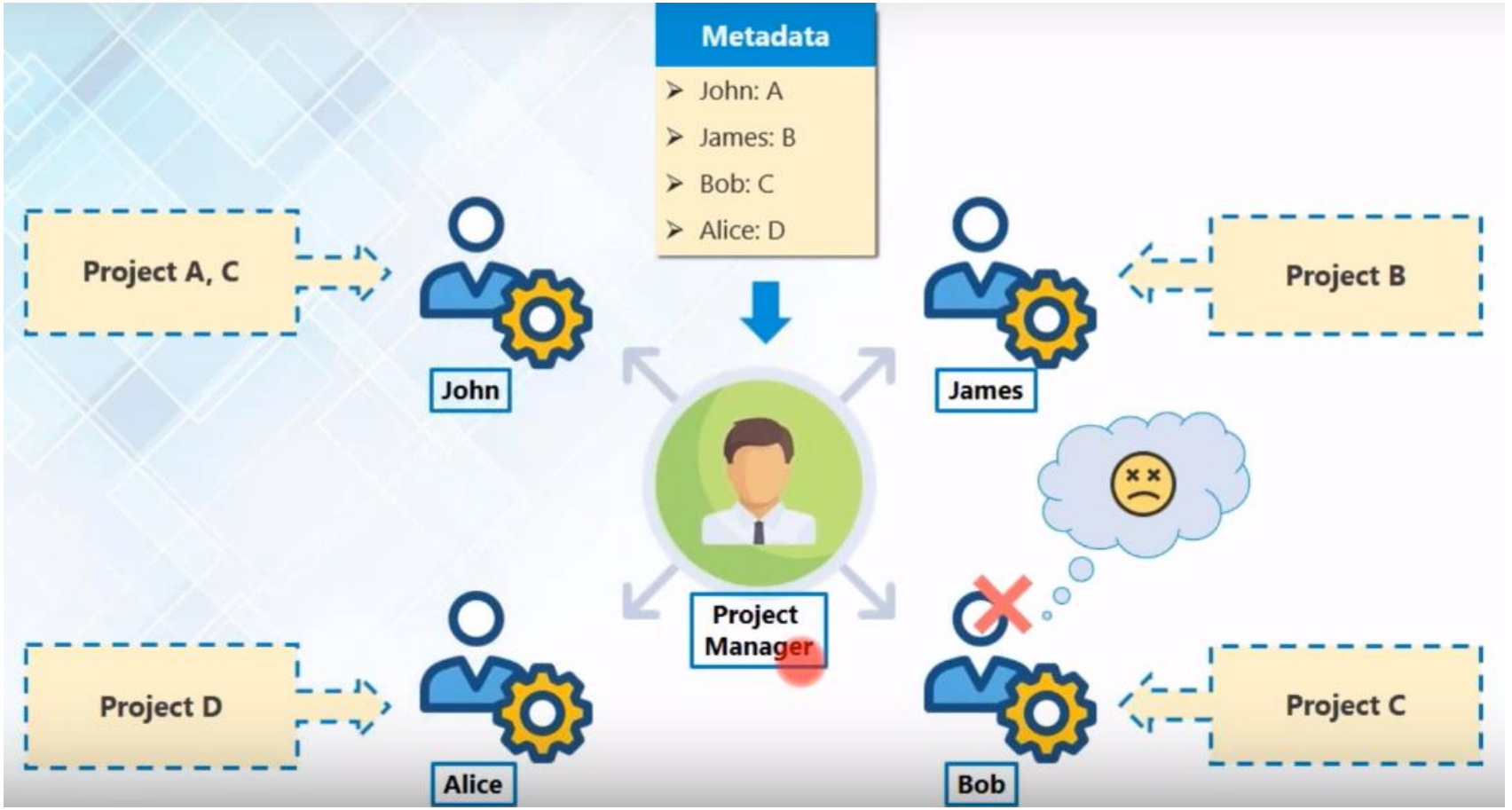
- A project manager managing a team of four employees.
- He assigns project to each of them and tracks the progress



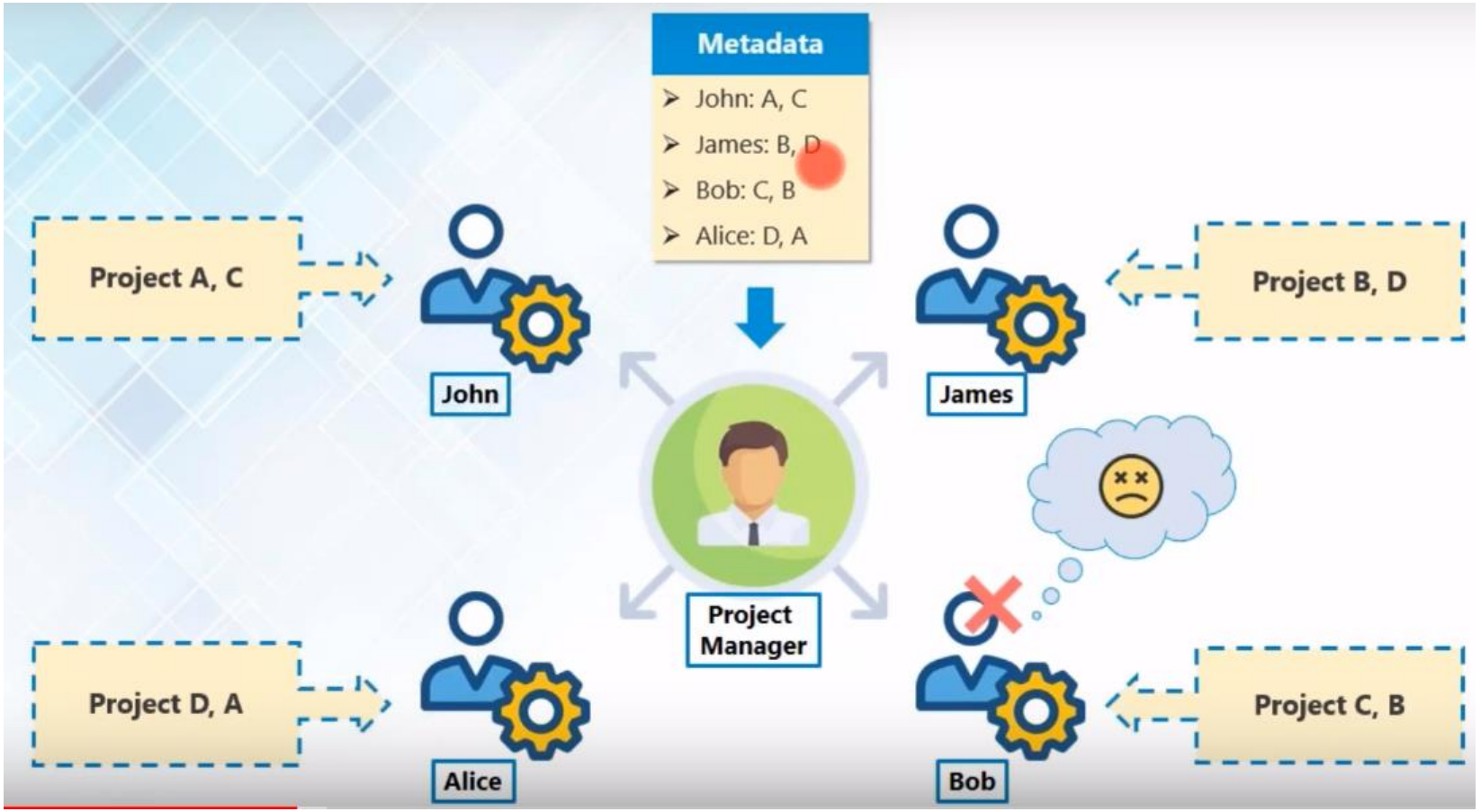
Scenario


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

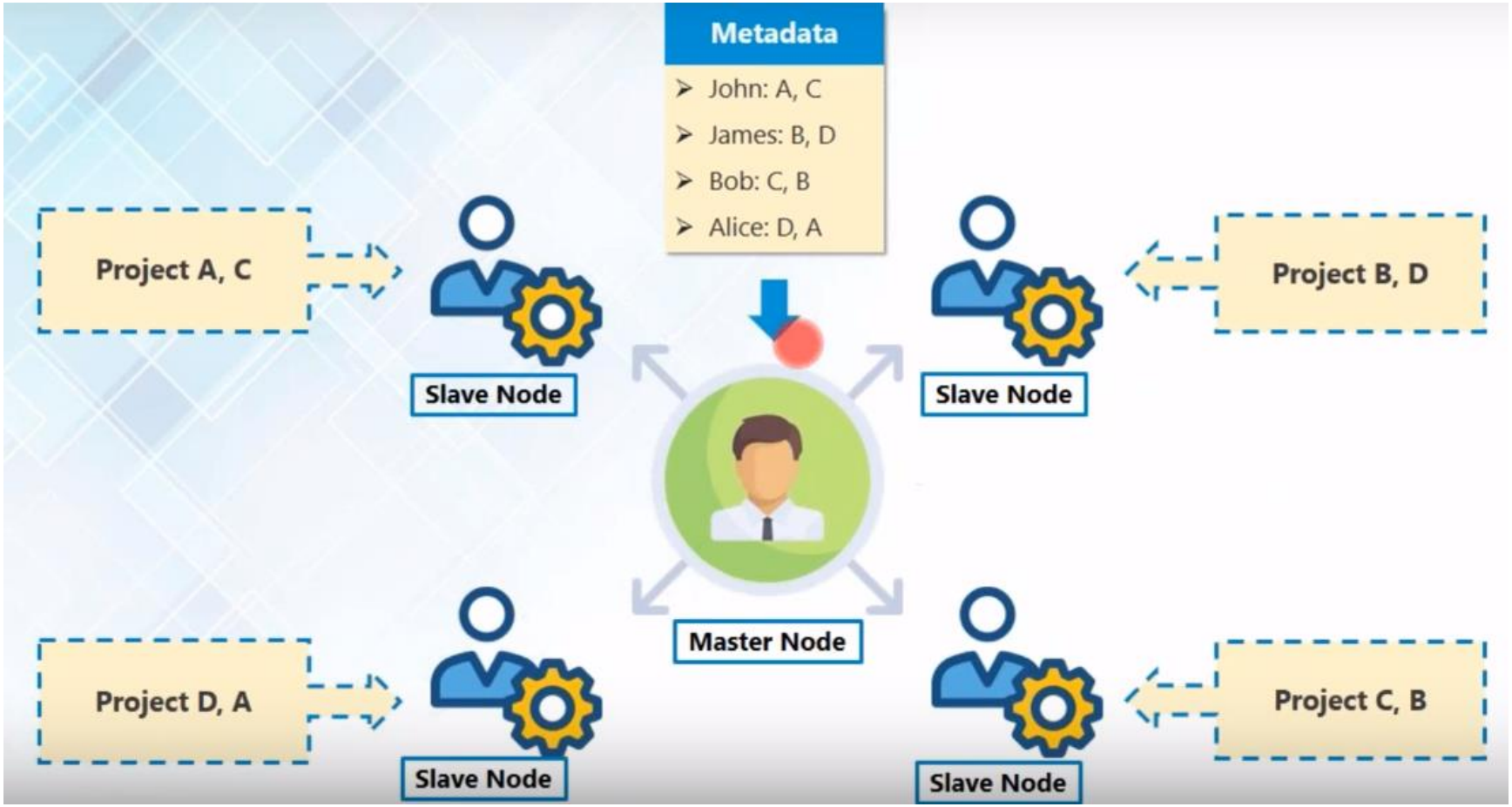
Scenario


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

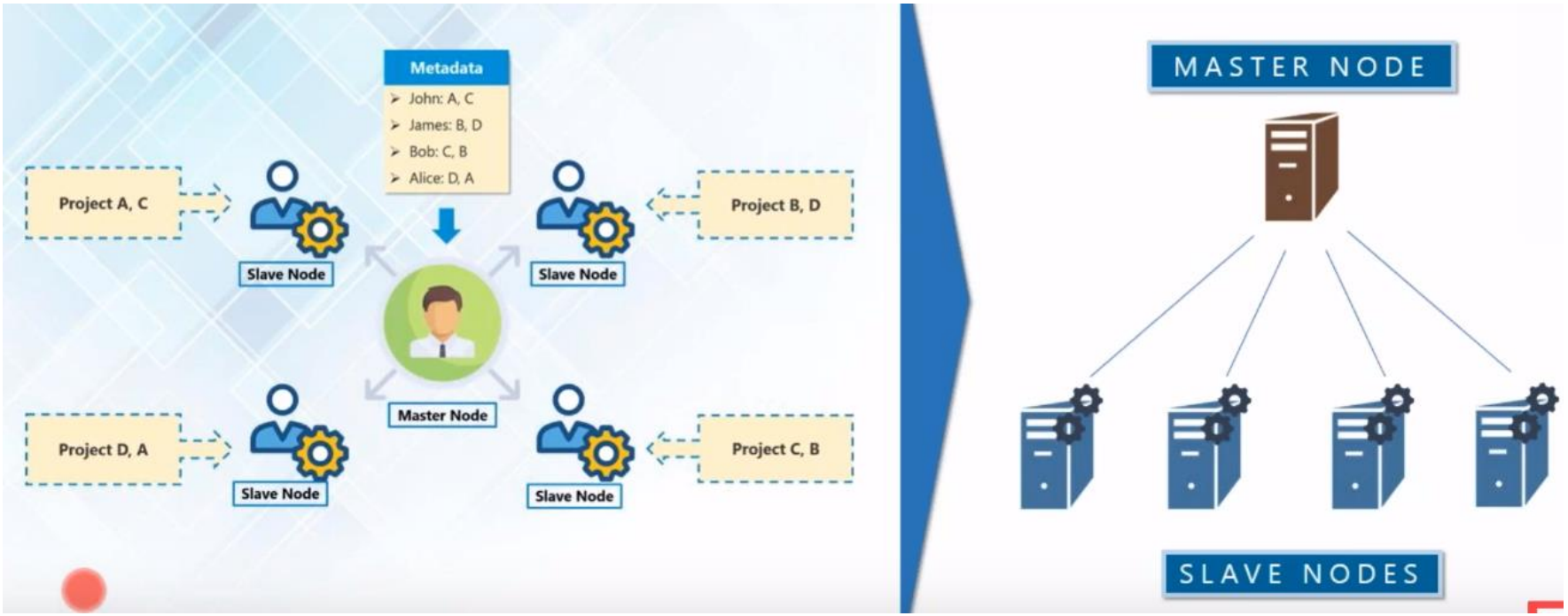
Scenario


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

Scenario

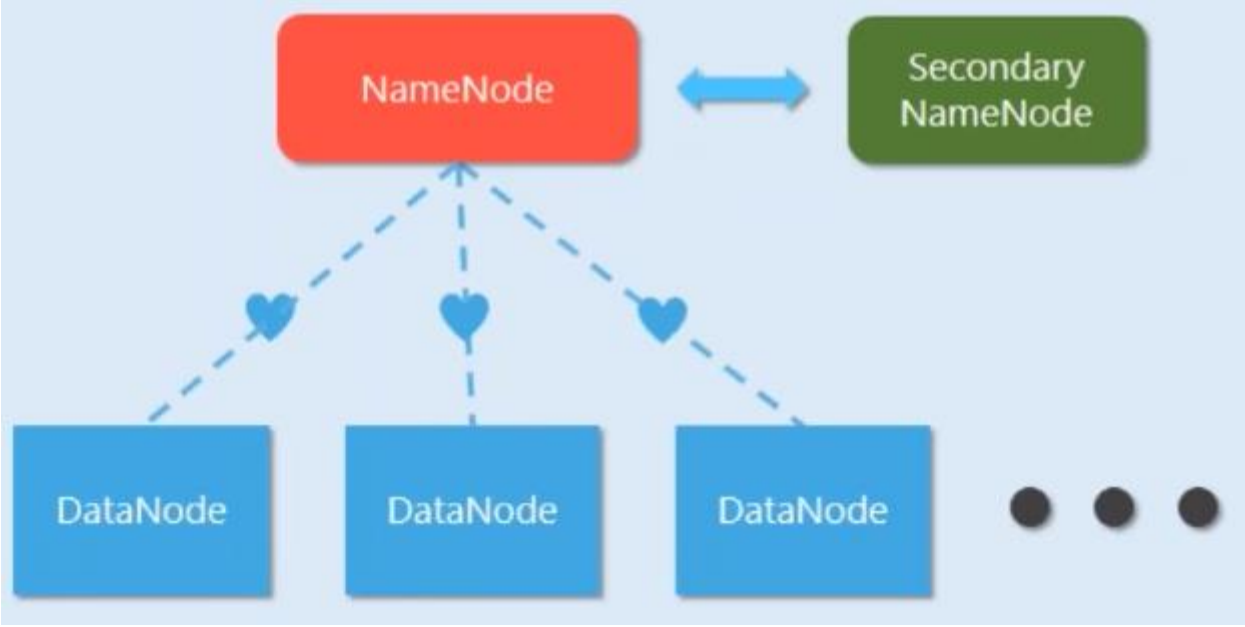

Metadata
John: A, C
James: B, D
Bob: C, B
Alice: D, A
Project A, C
Slave Node
Project B, D
Slave Node
Master Node
Project D, A
Slave Node
Project C, B
Slave Node

Hadoop: Master/Slave Architecture



HDFS Core Components

- NameNode
- DataNode
- Secondary NameNode



NameNode & DataNode

- **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

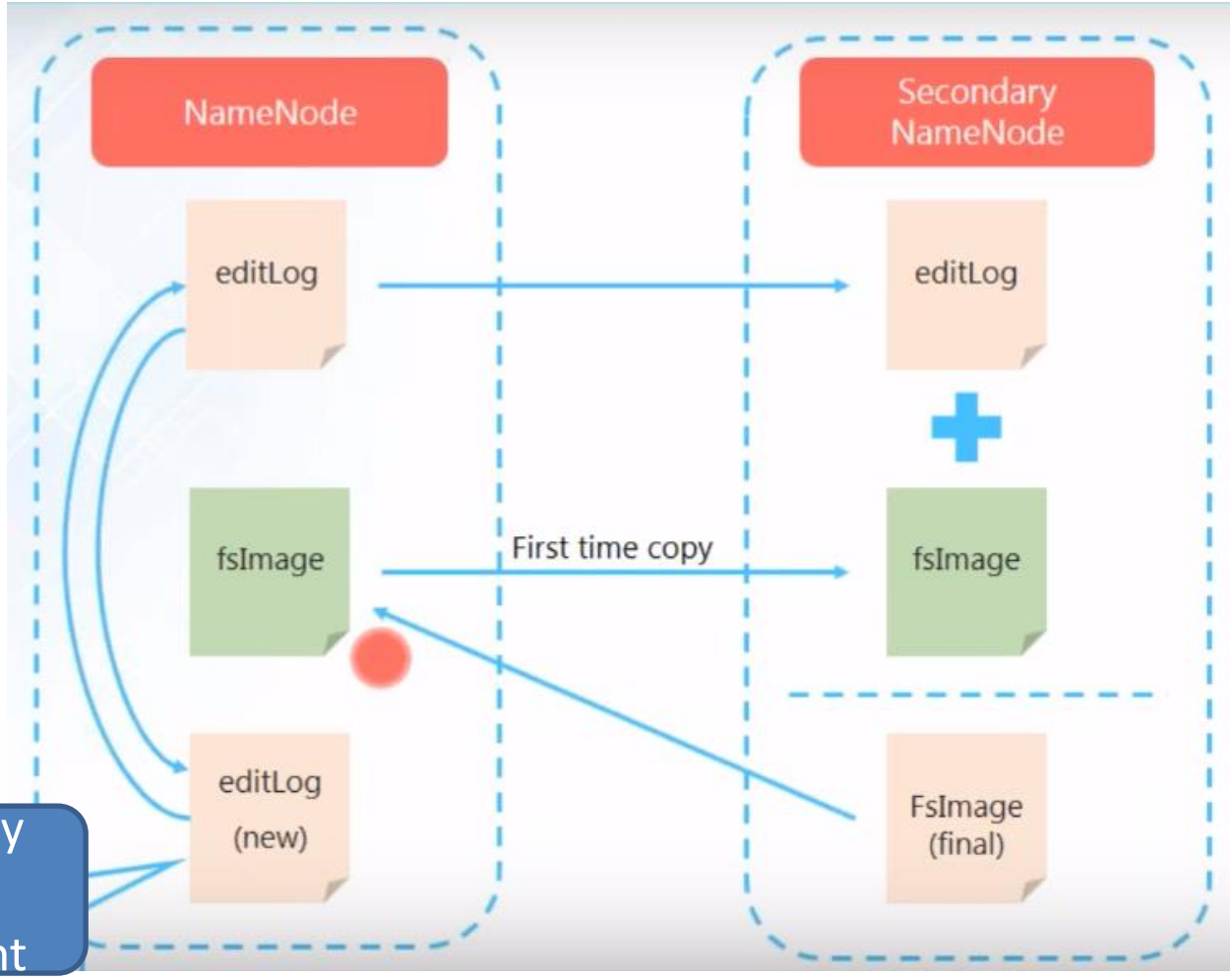
NameNode & DataNode

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

Secondary NameNode & Checkpointing

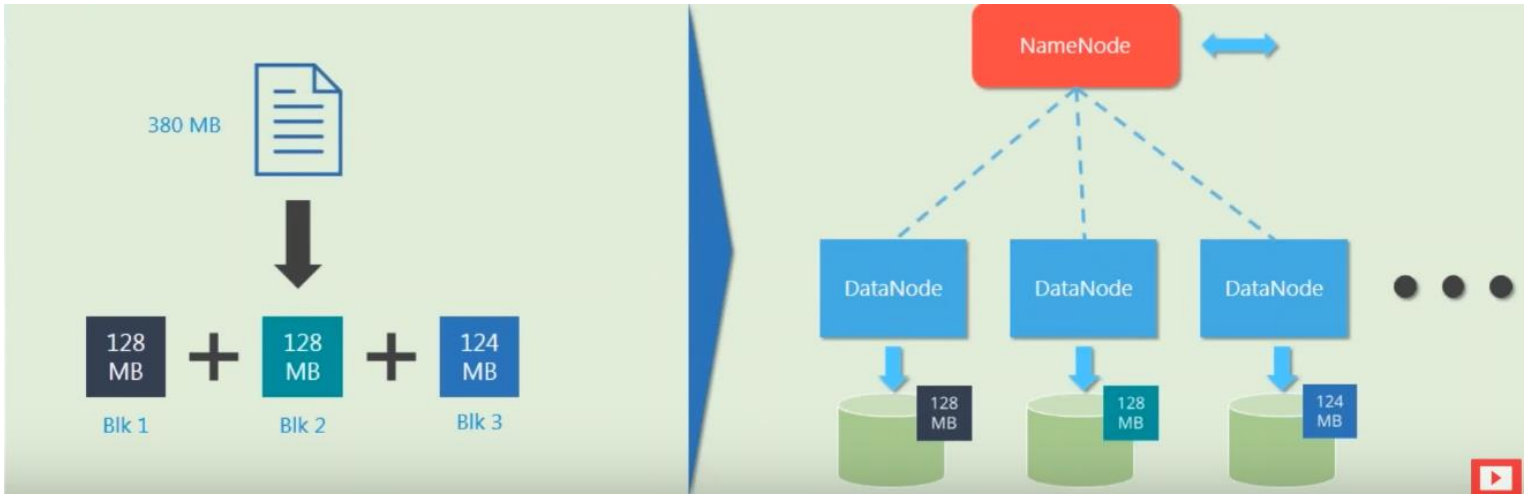
- Checkpointing is a process of combining edit logs with fsImage
- Secondary NameNode takes over the responsibility of checkpointing, therefore making NameNode more available
- Allows faster failover as it prevents edit logs from becoming too huge
- Checkpointing happens periodically (Default 1 hour)

Secondary NameNode & Checkpointing



HDFS Data Blocks

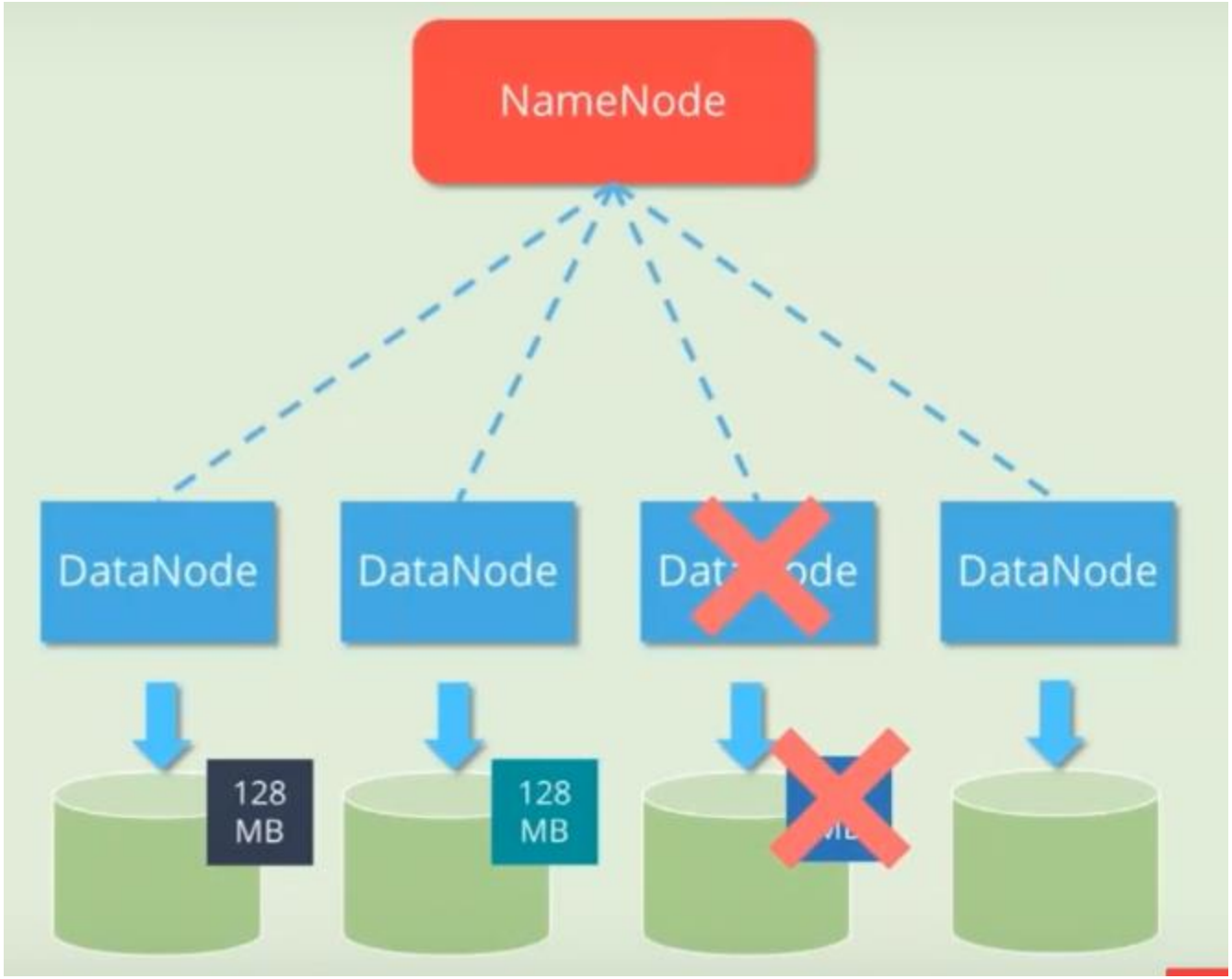
- Each file is stored on HDFS as blocks
- The default size of each block is 128MB in Apache Hadoop 2.x (64MB in Apache Hadoop 1.x)



HDFS Data Blocks

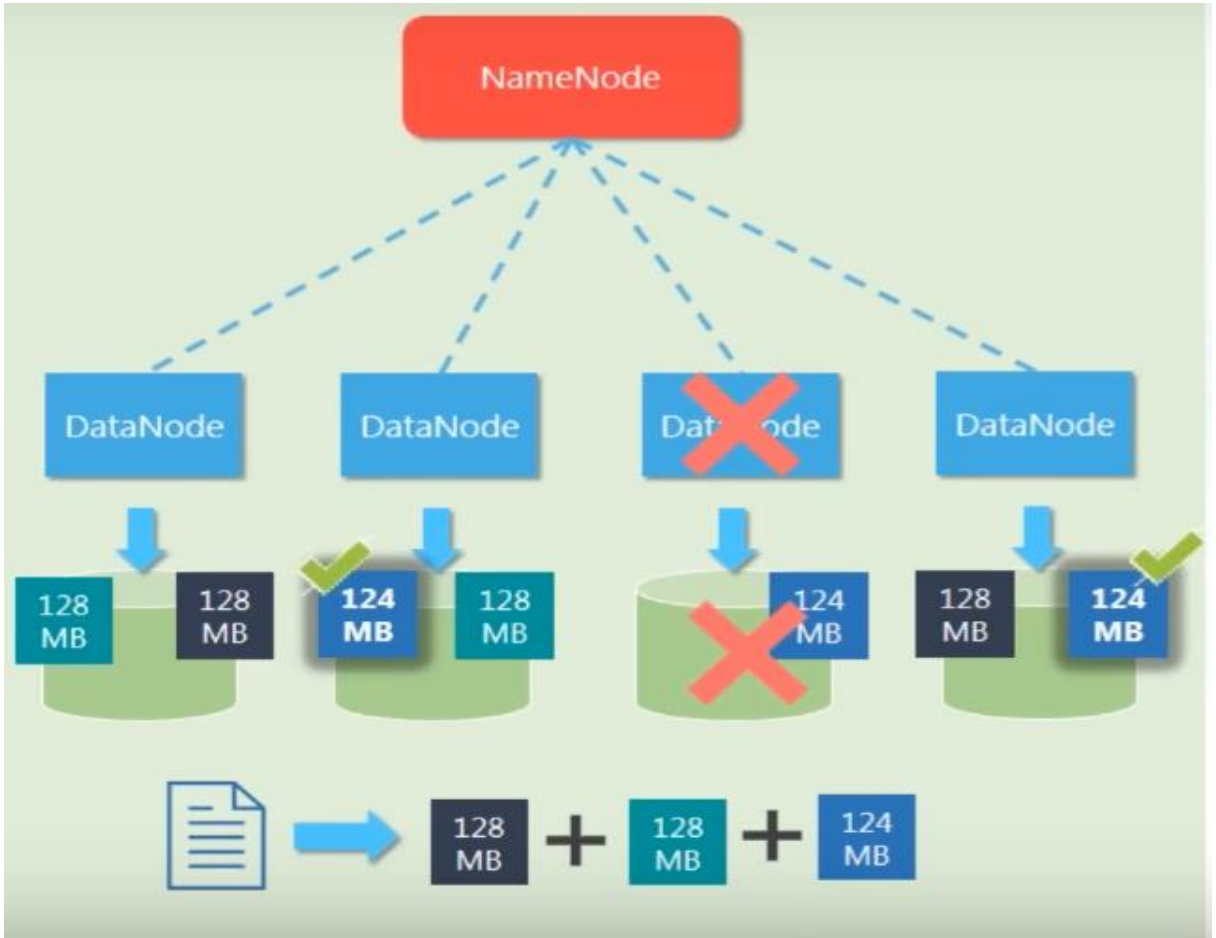
- **Advantages**
  - Save disk space
  - Flexibility – can expanded by adding required number of dataNodes
  - Runtime efficiency – associated with parallel distributed computing

Fault Tolerance

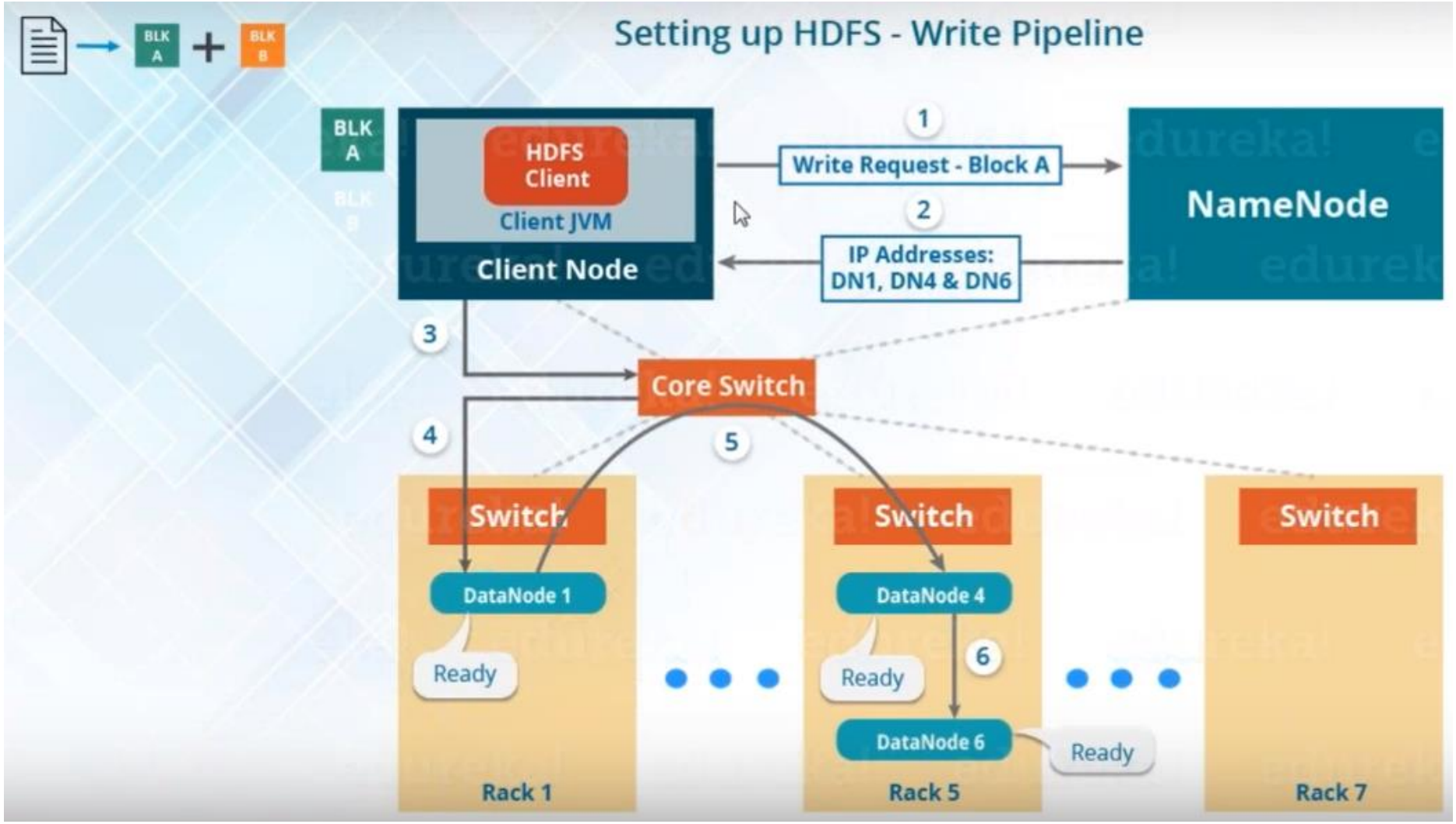


Fault Tolerance : Replication Factor

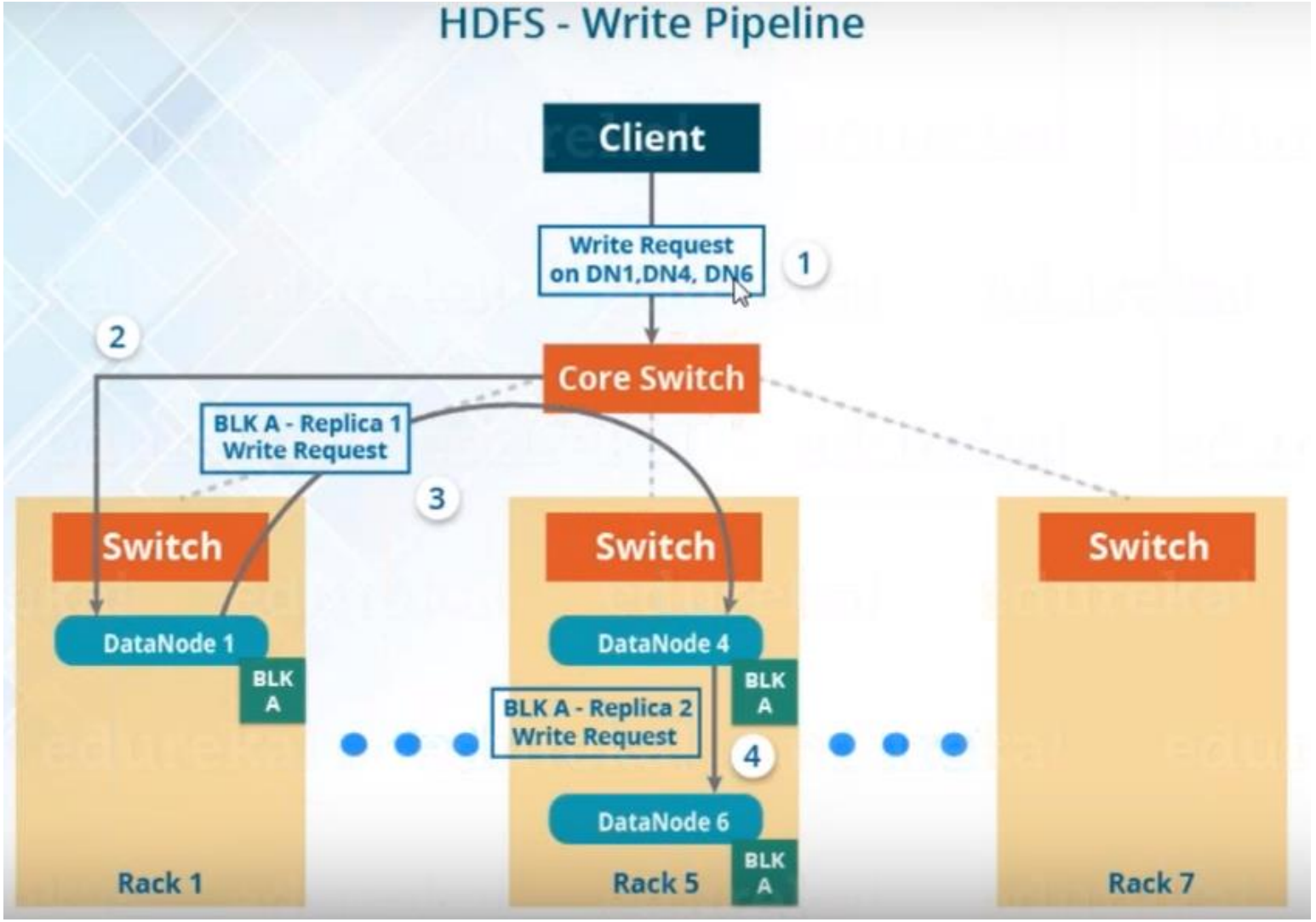
- Each data block is replicated and are distributed across different DataNode



HDFS Write Mechanism – Pipeline Setup

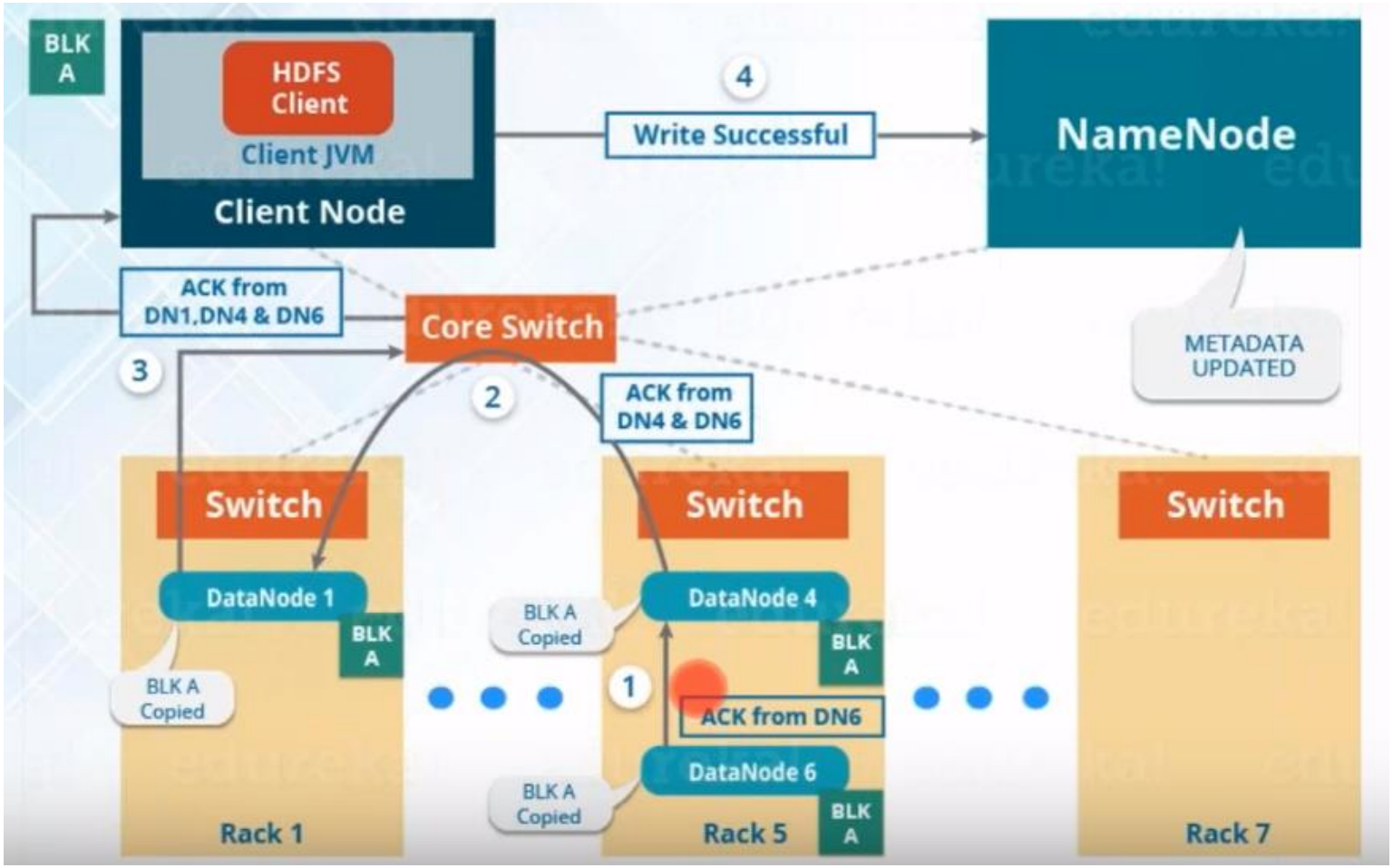


HDFS Write Mechanism – Writing a Block

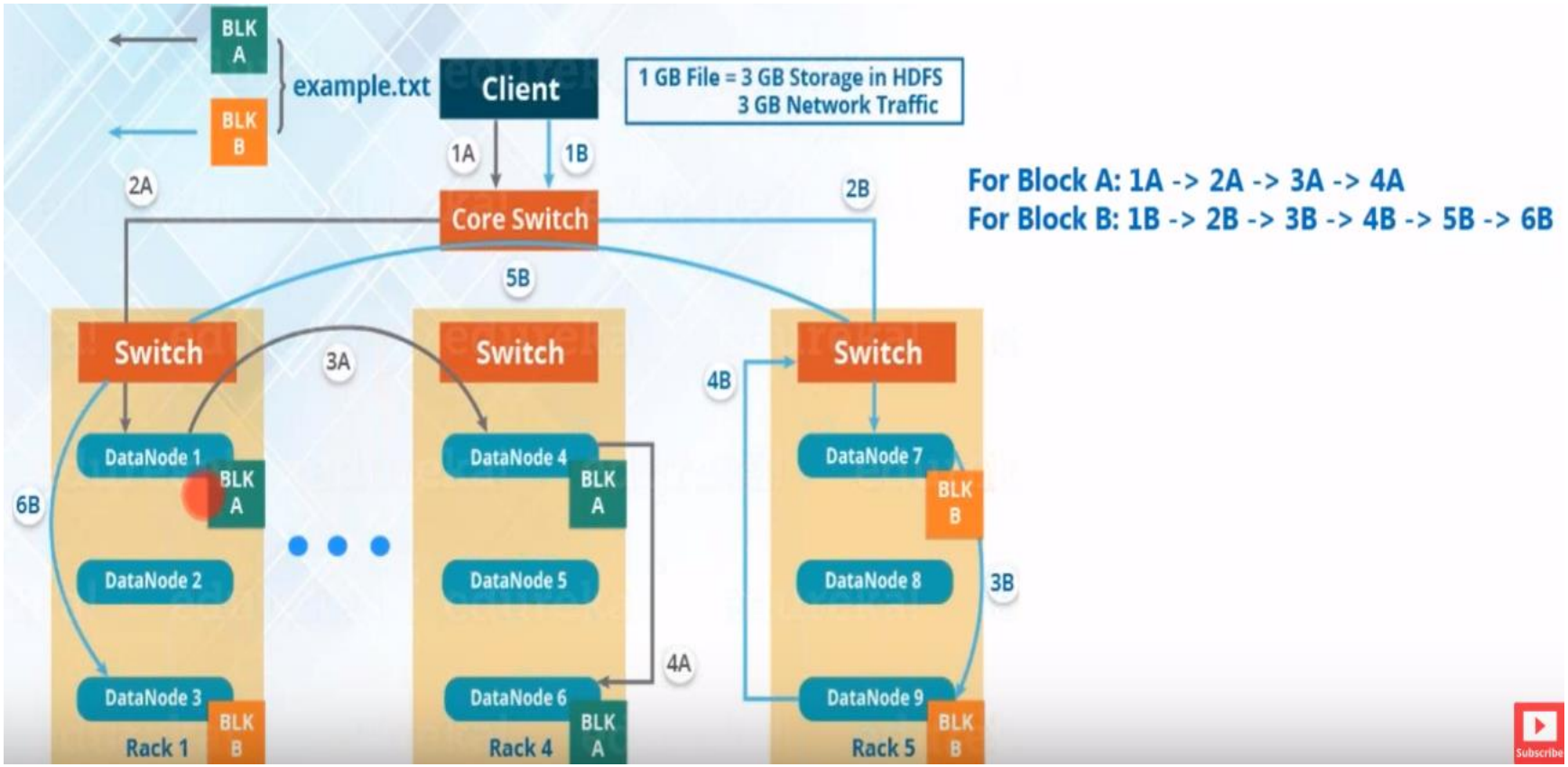


HDFS Write Mechanism – Acknowledgement

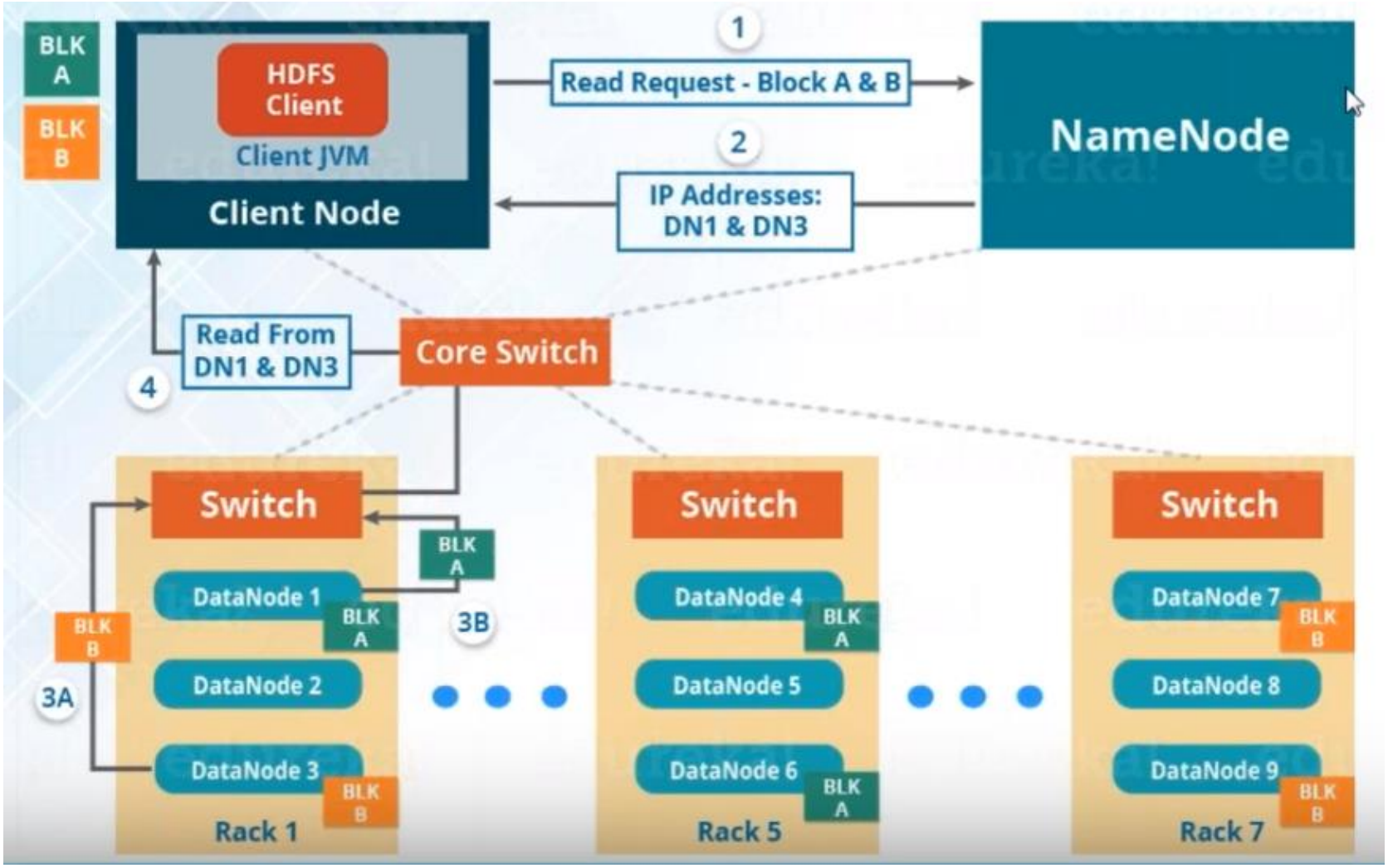
Acknowledgement in HDFS - Write



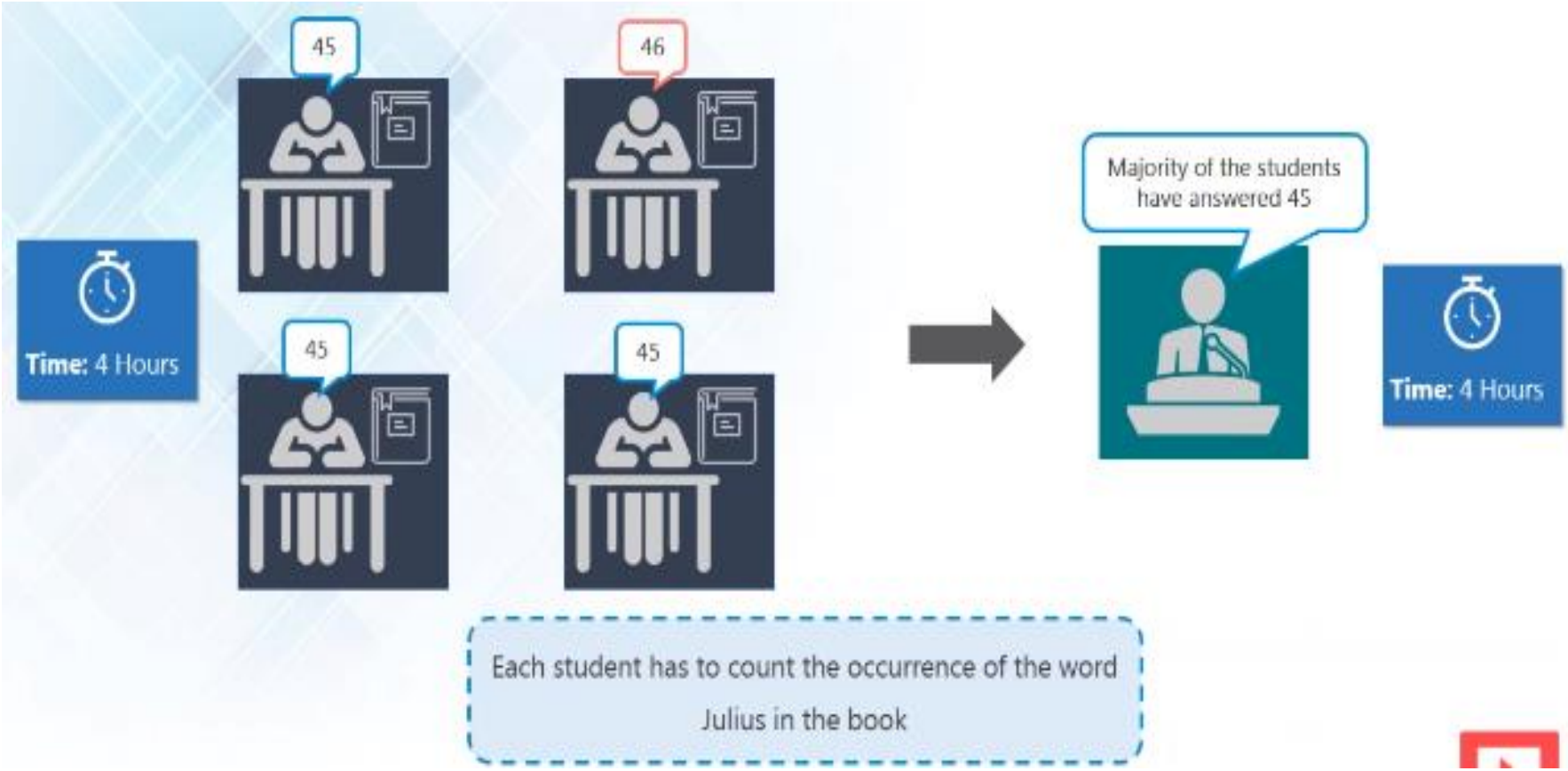
HDFS Multi-block write Mechanism



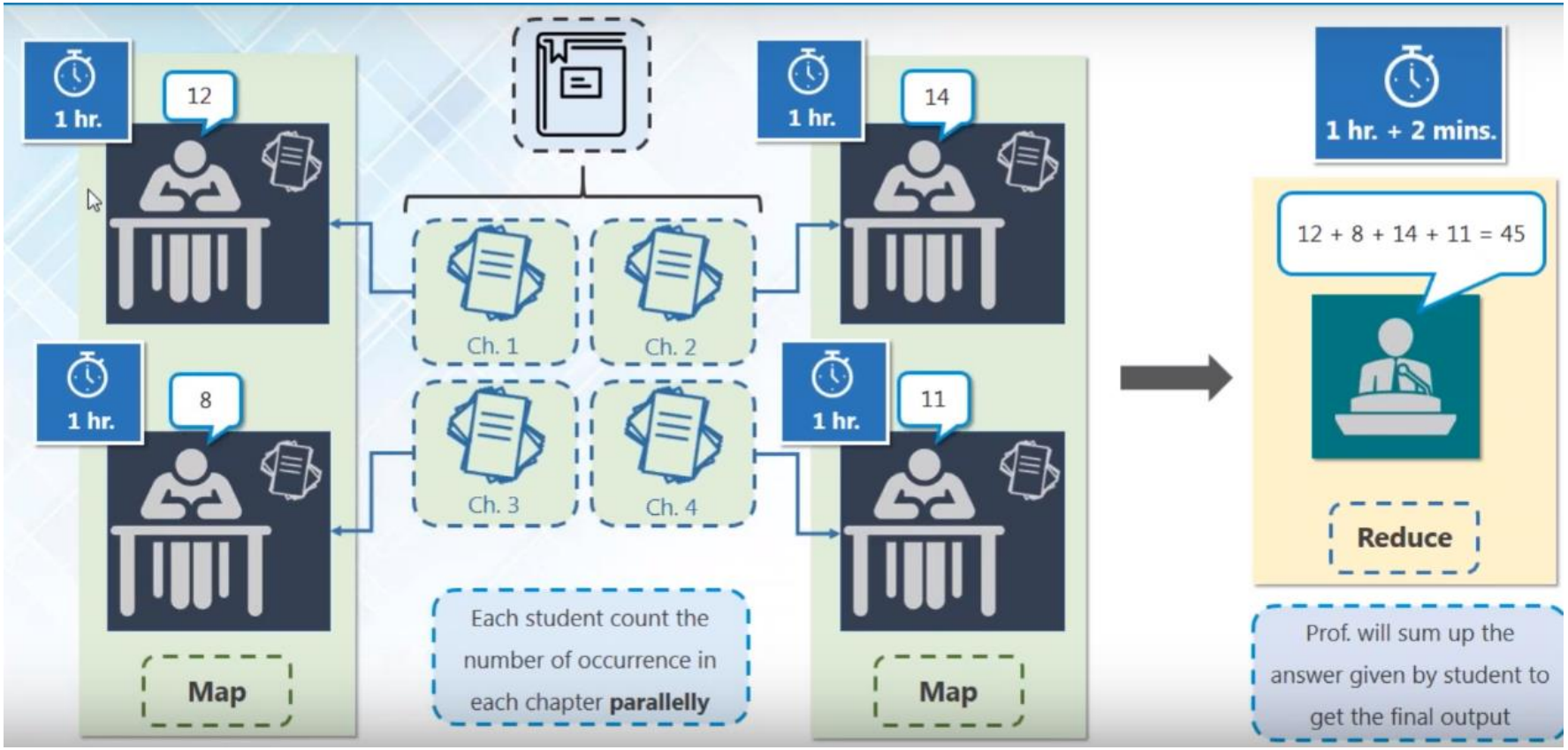
HDFS Read Mechanism



Story of Map Reduce

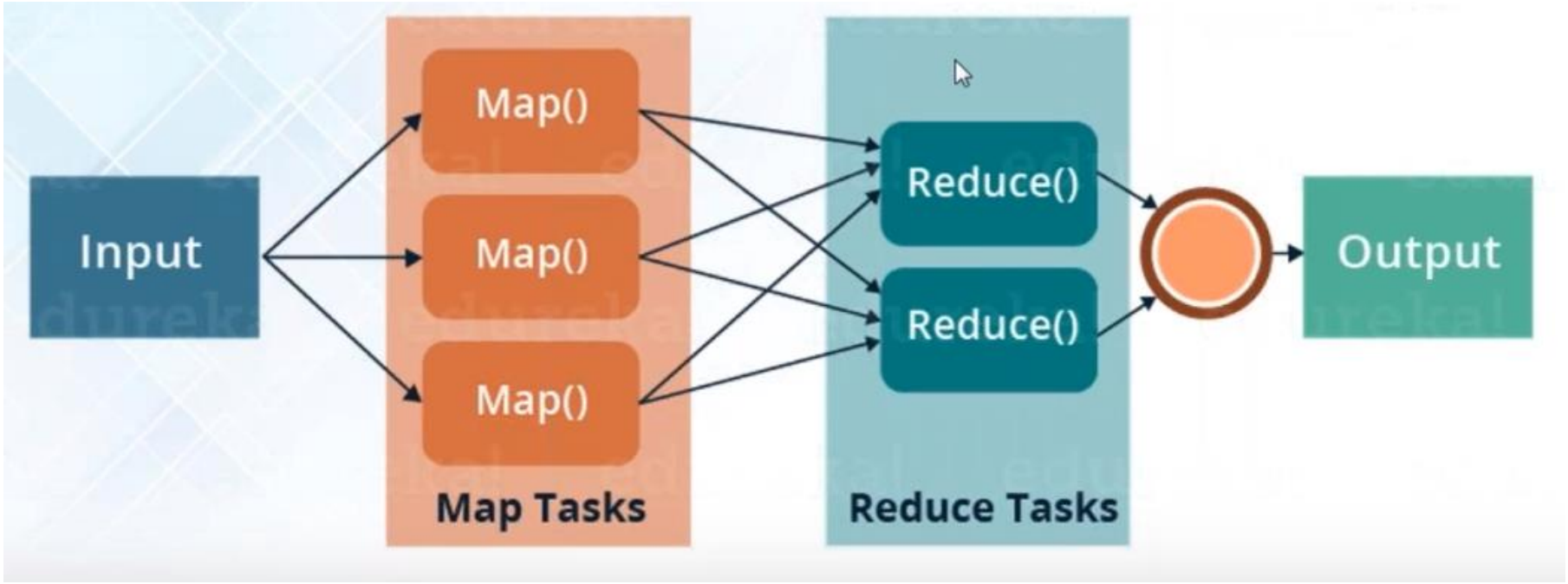


Story of Map Reduce



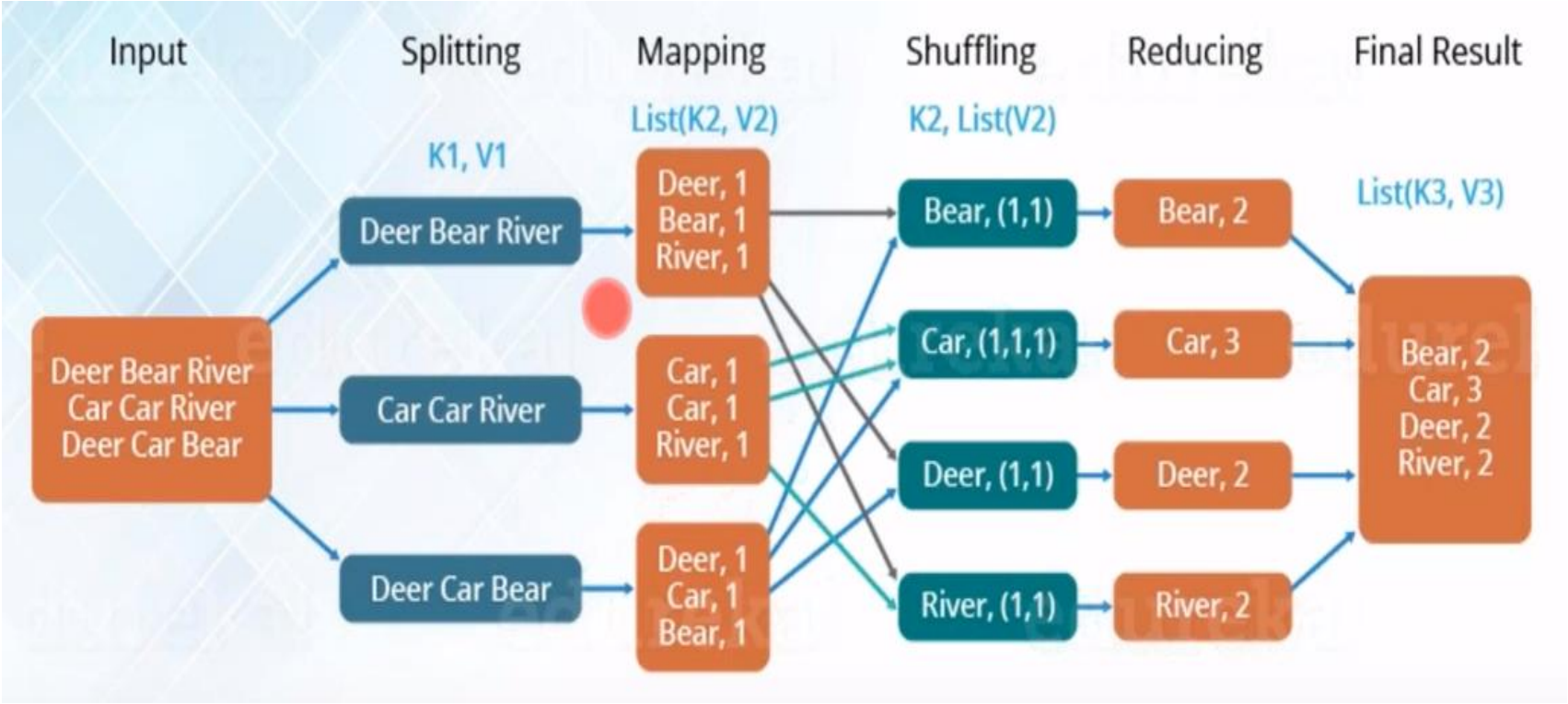
What is MapReduce?

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



MapReduce Word Count Program

The overall MapReduce word Count process



Map Reduce Word Count Program

- Three major parts of MapReduce Program:

Mapper Code:

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

Reducer Code

- Write reducer logic here. It combines the intermediate key-value pair generated by Mapper to get final aggregated output

Driver Code

- Specify all the job configurations over here including job name, input path, output path etc….

Limitations of Hadoop

- Hadoop is NOT good at
  - Processing small files
  - Interactive queries & Real time processing
  - Updates
  - For random access data retrieval
  - Does not support ACID properties

  (No transactions can be performed in Hadoop system)

Questions

- What is Big data?
- What solutions will be used to handle big data?
- Why it is useful to develop big data-based application compared to traditional web-based system?
- What factors should be considered when identifying big data?

Question

- “Map-Reduce is not suitable for applications that re-use a working data set across multiple parallel operations” Do you agree with this statement. Justify your answer.