Module 1

Introduction to Big Data Technology and Hadoop

Thanachart Numnonda, Executive Director, IMC Institute
Mr. Aekanun Thongtae, Big Data Consultant, IMC Institute

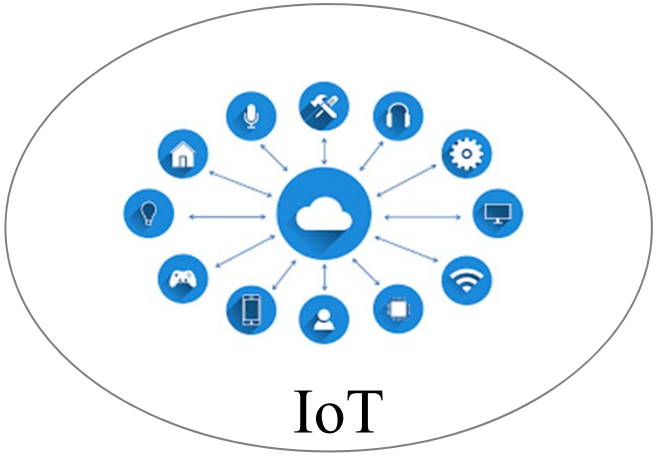
Thanisa Numnonda, Faculty of Information Technology,

King Mongkut's Institute of Technology Ladkrabang

Where do Data Come from?

Social Media

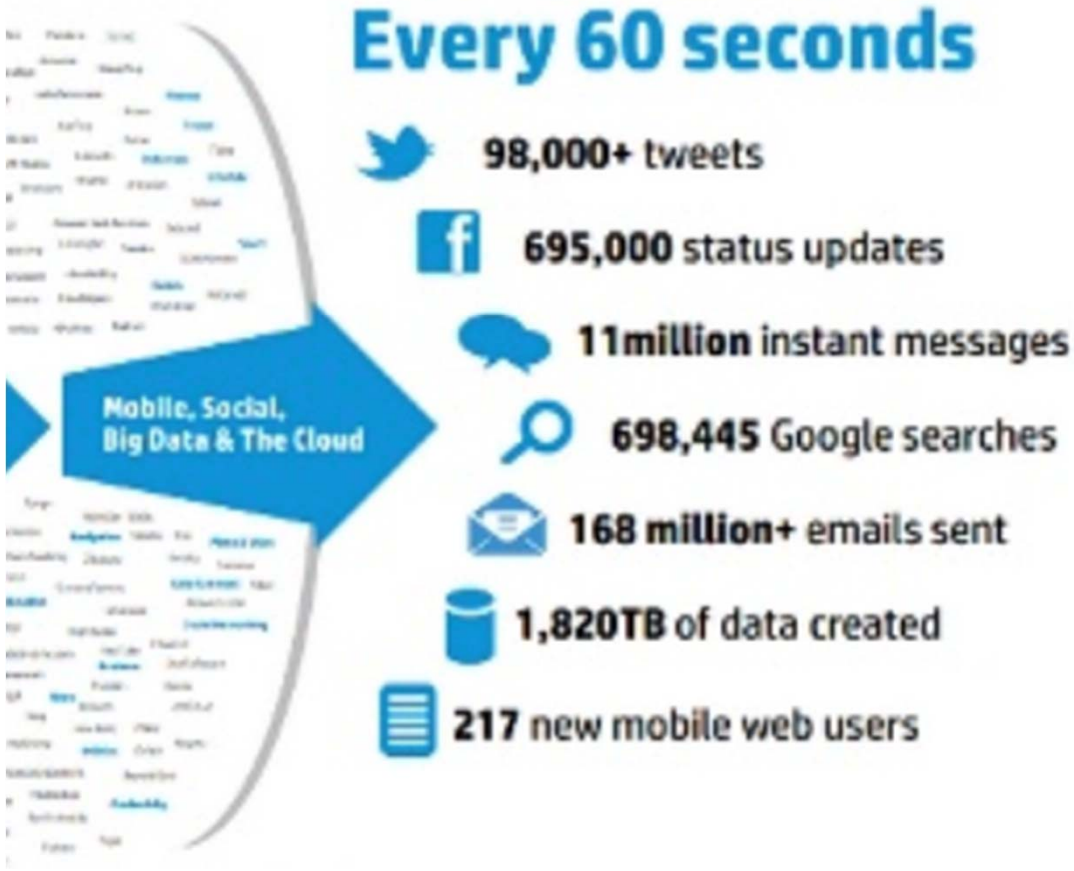
Mobile



IoT

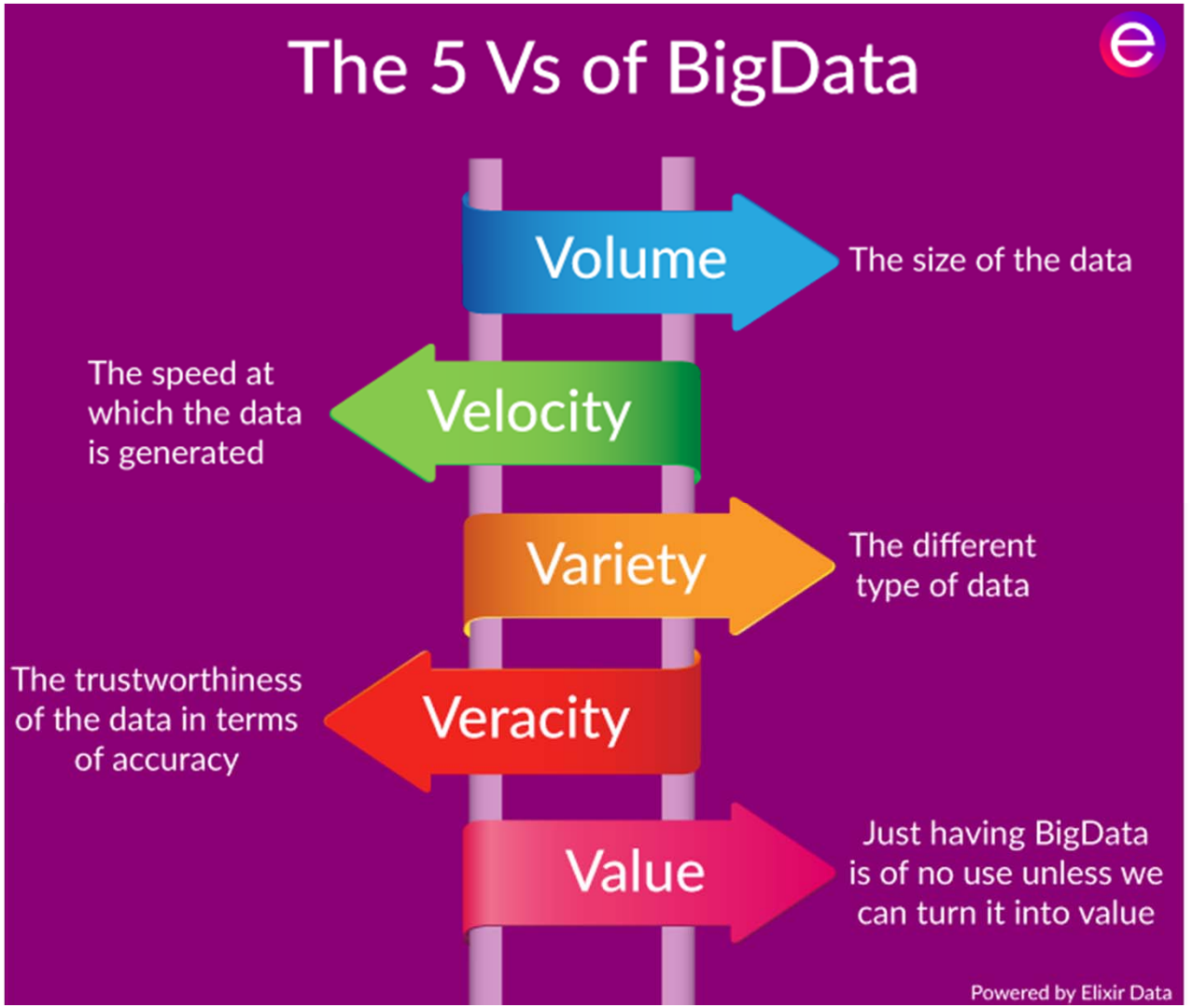
Transaction

HOW BIG IS BIG?

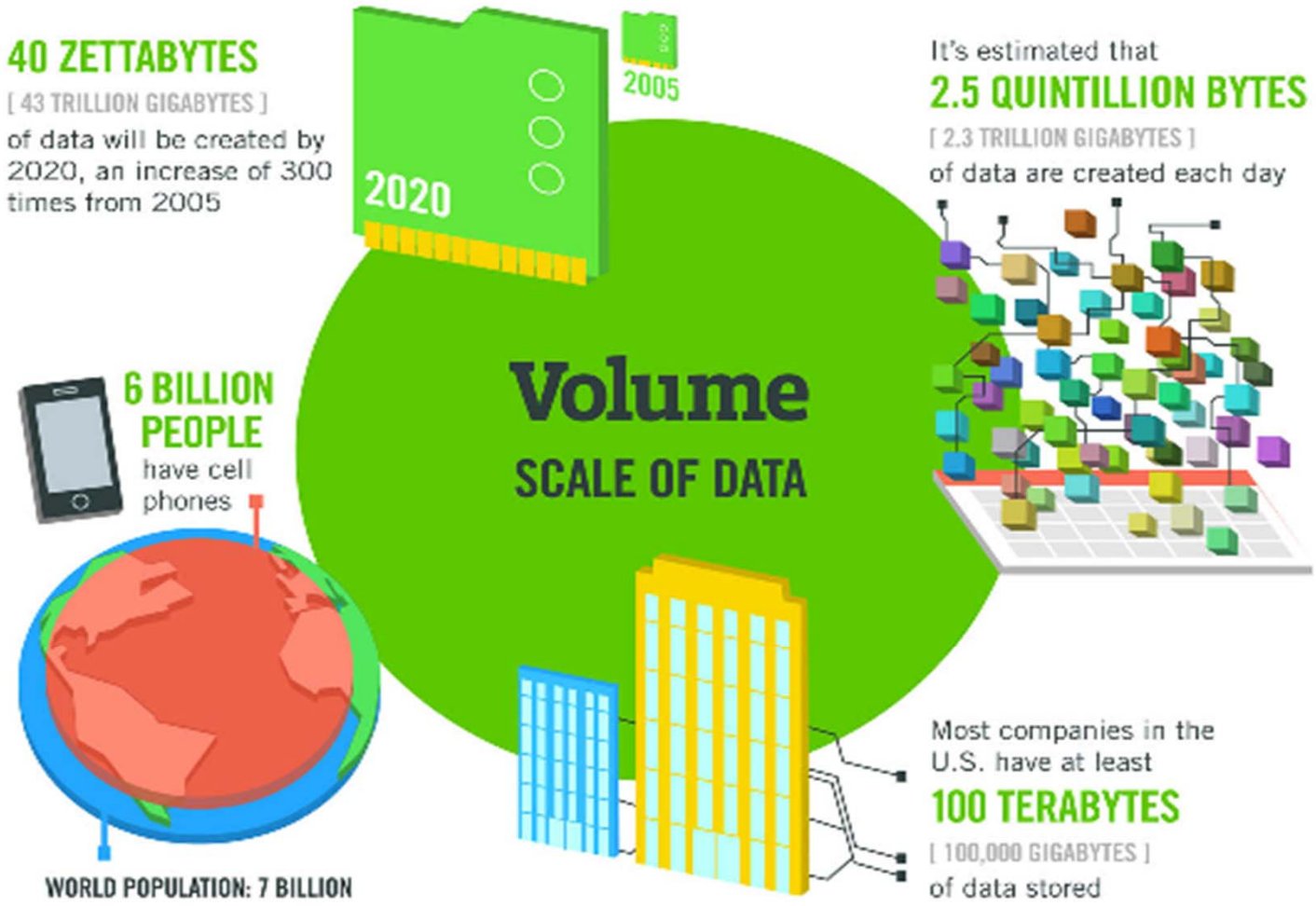


Source: https://www.slideshare.net/sfamilian/working-with-big-data-jan-2016-part-1/6-CONTEXT_WHATS_BIG_DATAHOW_BIG/

Big Data



Source: https://viblo.asia/p/why-we-need-modern-big-data-integration-platform-gAm5yx1Dldb/


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

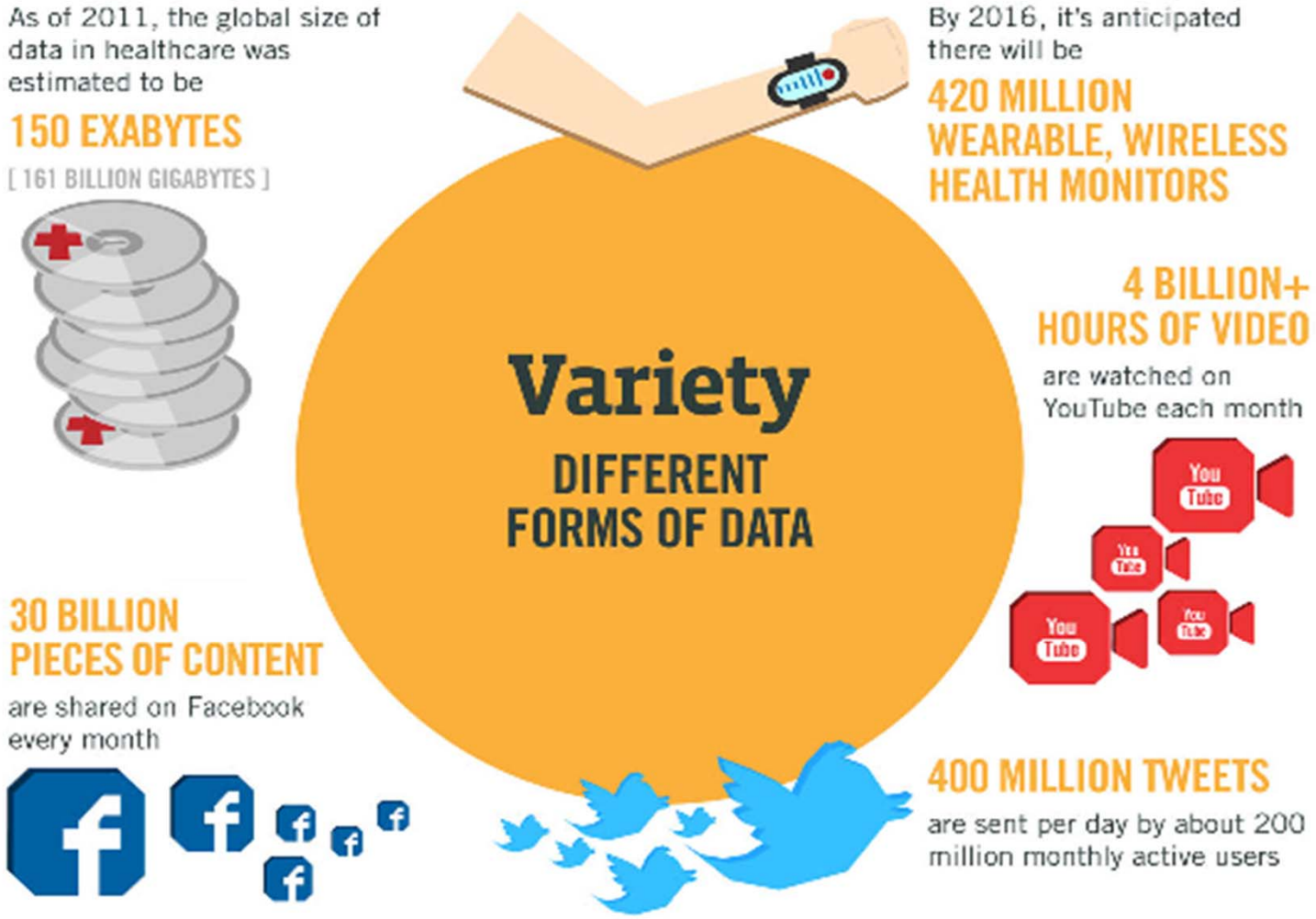
Velocity

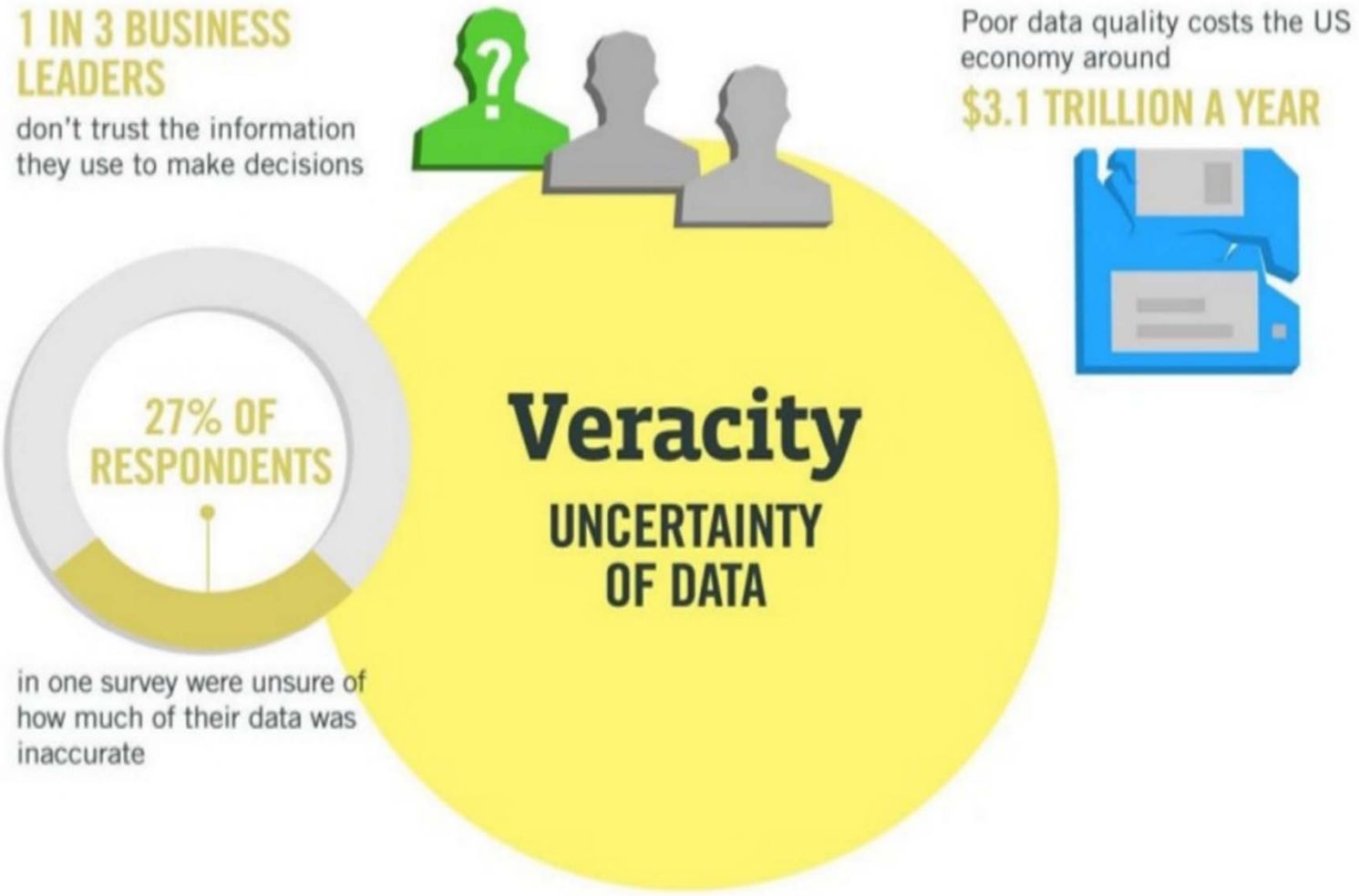
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth


As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2016, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

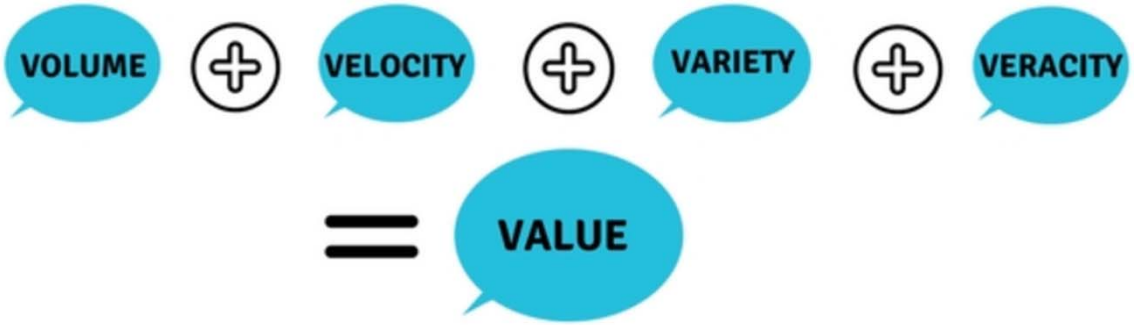
The Value of Big Data

Over 80% of organizations say:

Big Data is critical to meet strategic objectives.

Sharing insights is a must-have capability for businesses.

Big Data will amplify other technology innovations.



Why is Big Data Analytics Important?

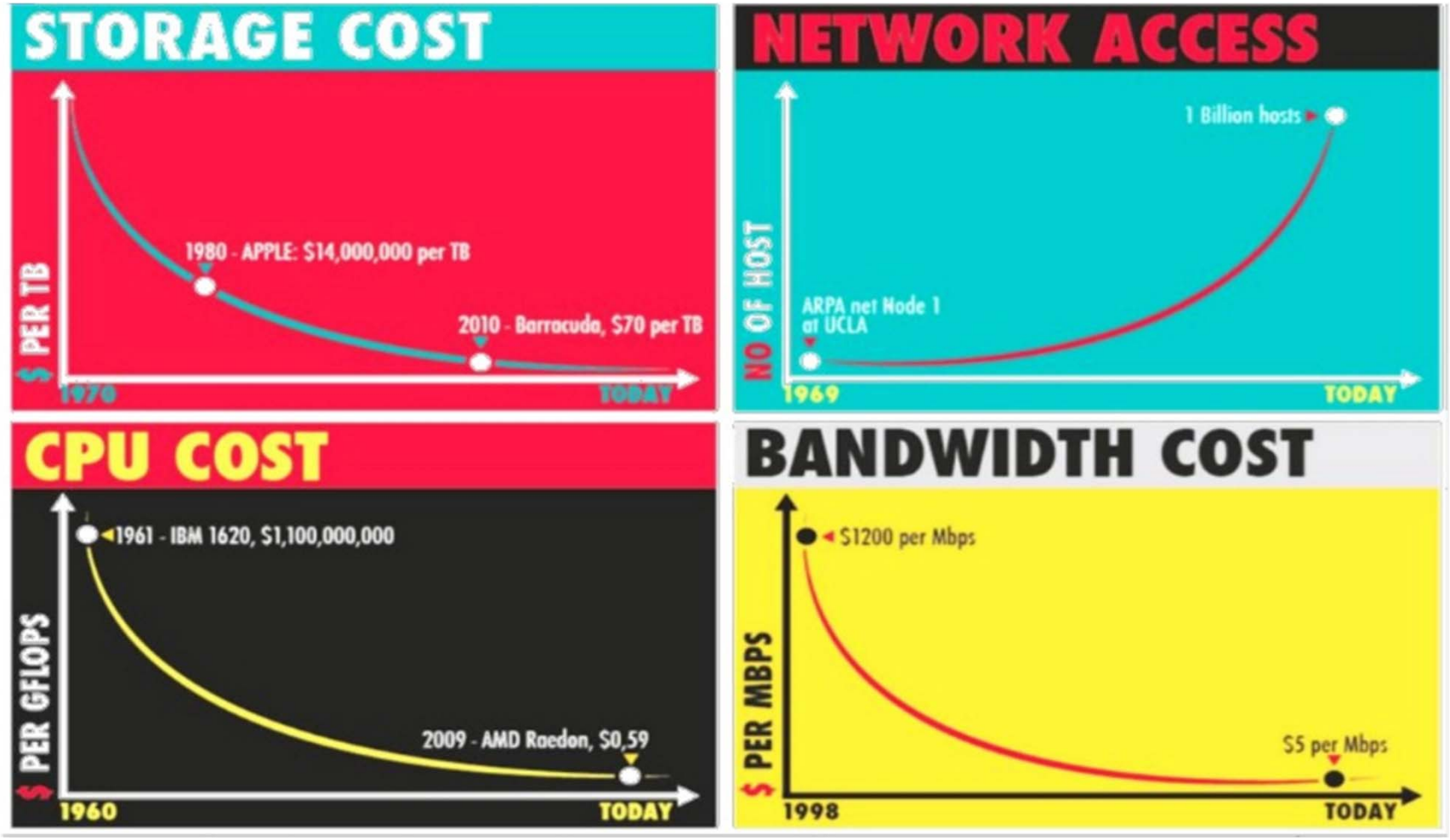
Cost reduction: identify more efficient ways of doing business.

Faster, better decision making: able to analyze information immediately – and make decisions based on what they've learned.

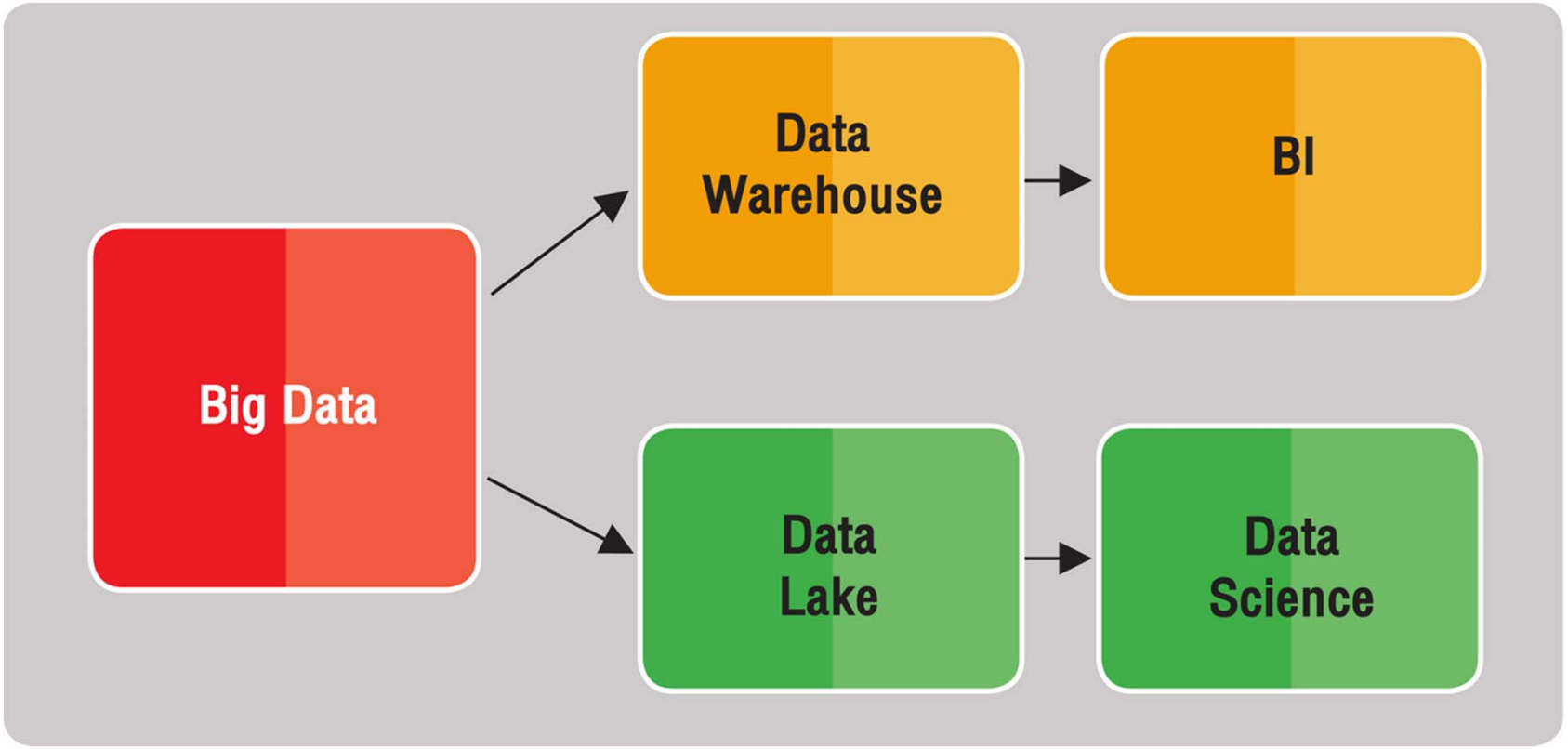
New products and services: give customers what they want - creating new products to meet customers' needs.

Source: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

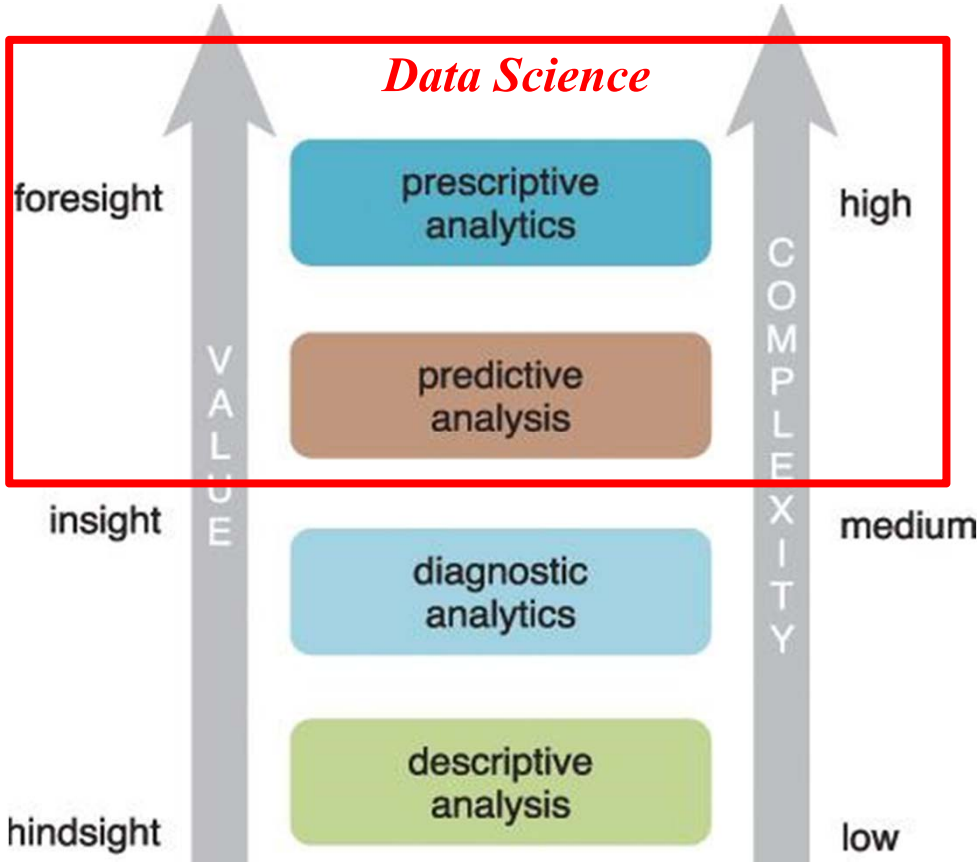
Big Data : Why Now?



Source: William EL KAIM, Enterprise Architecture and Technology Innovation


Big Data
Data Warehouse
BI
Data Lake
Data Science

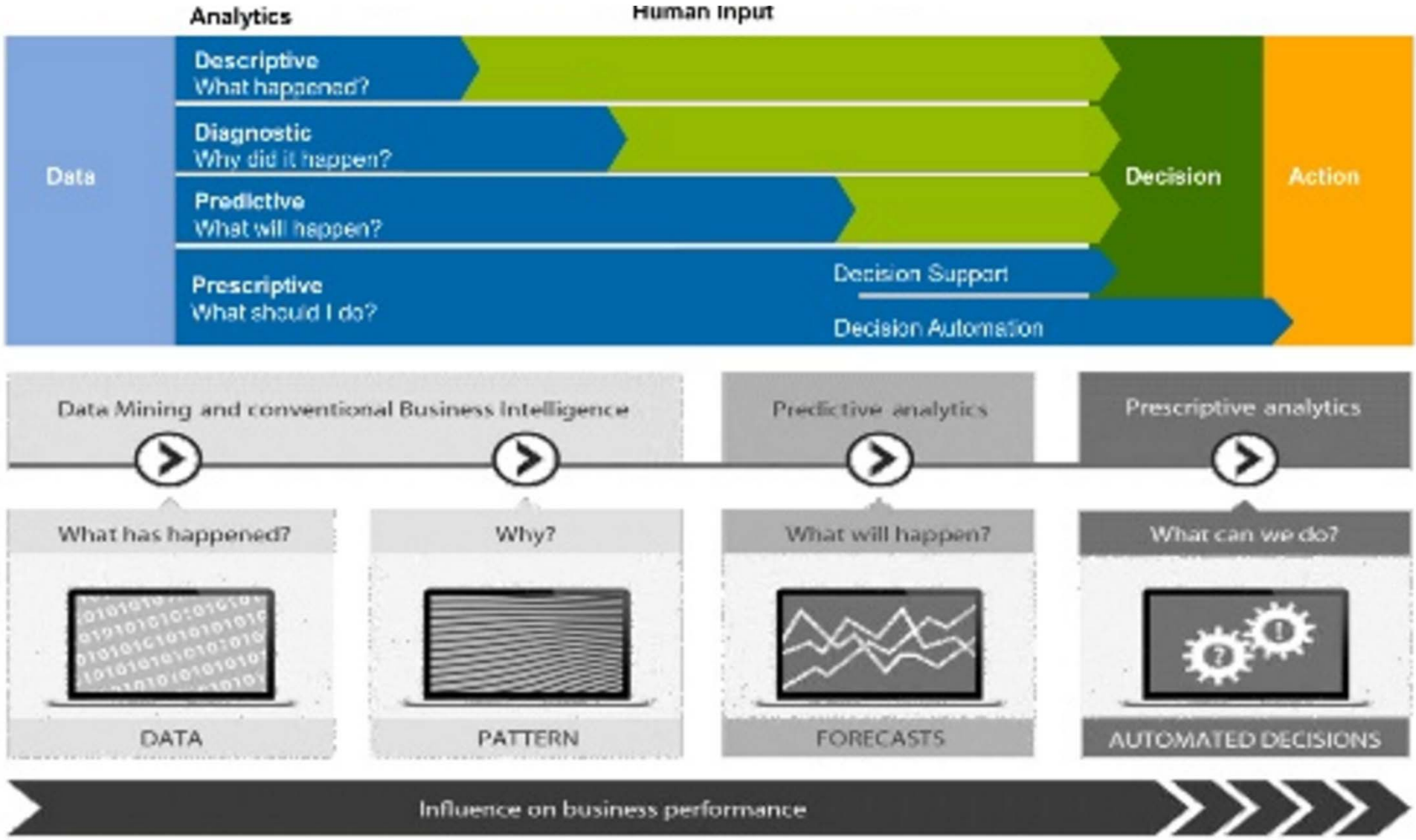
Data Analytics



- A discipline that includes the management of the complete data lifecycle

- Encompasses collecting, cleansing, organizing, storing, analyzing and governing data.

Adapted from: Thomas Eri et.al, "Big Data Fundamentals: Concepts, Drivers & Techniques," Prentice Hall, 2016

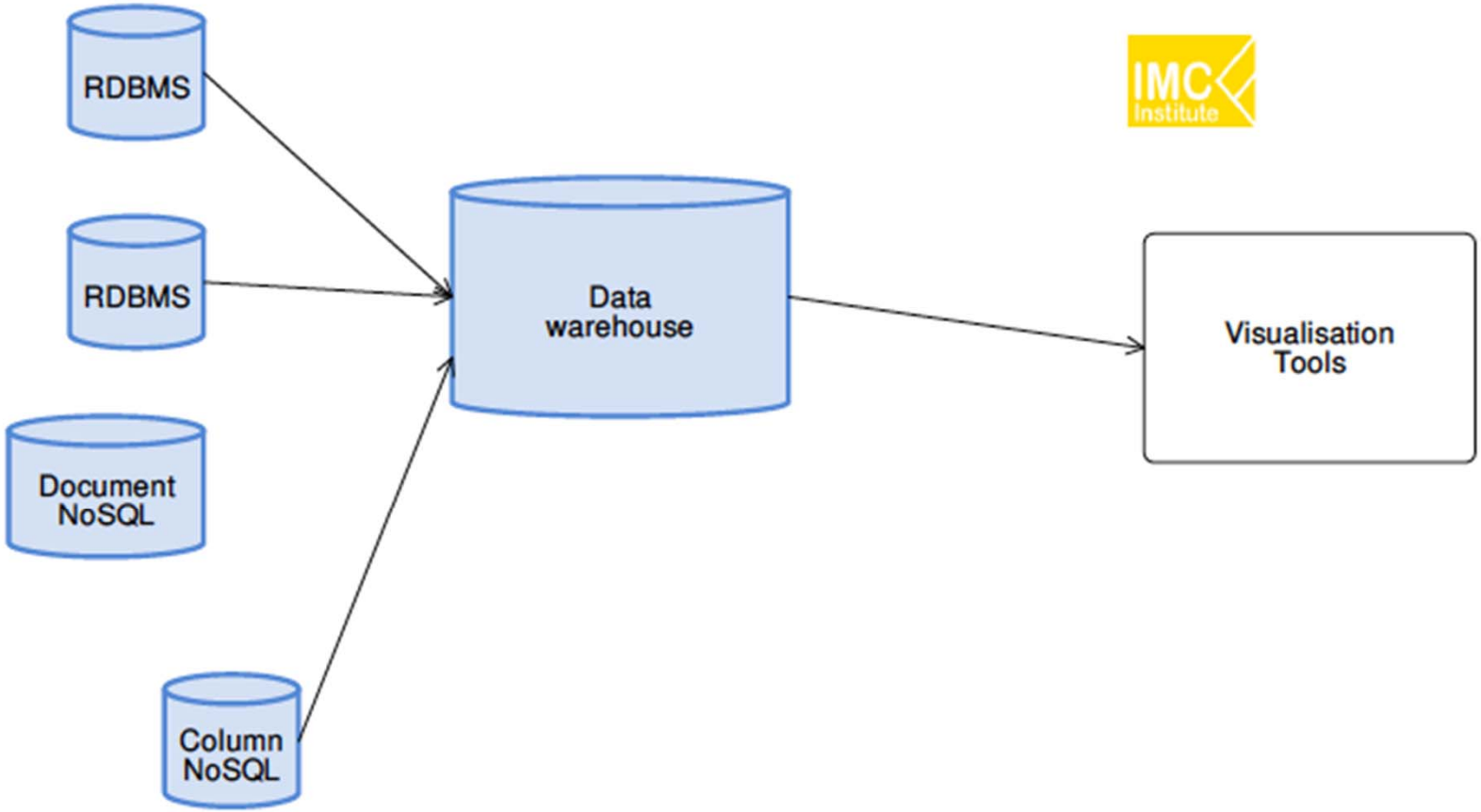


Source: William EL KAIM, Enterprise Architecture and Technology Innovation

We are forecasting the future based on the past

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

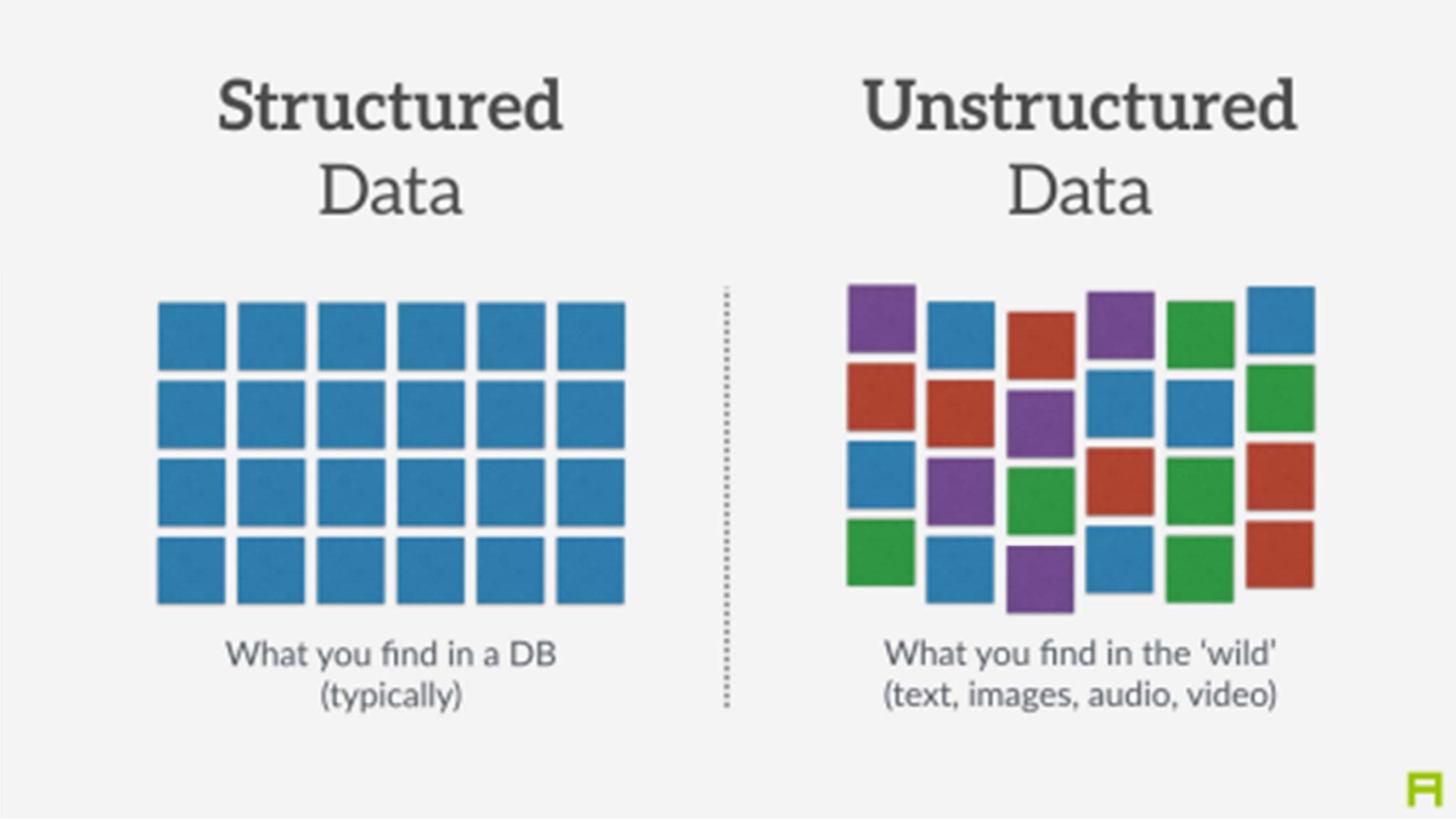
Big Data Architecture-Traditional



Semi-structured Data

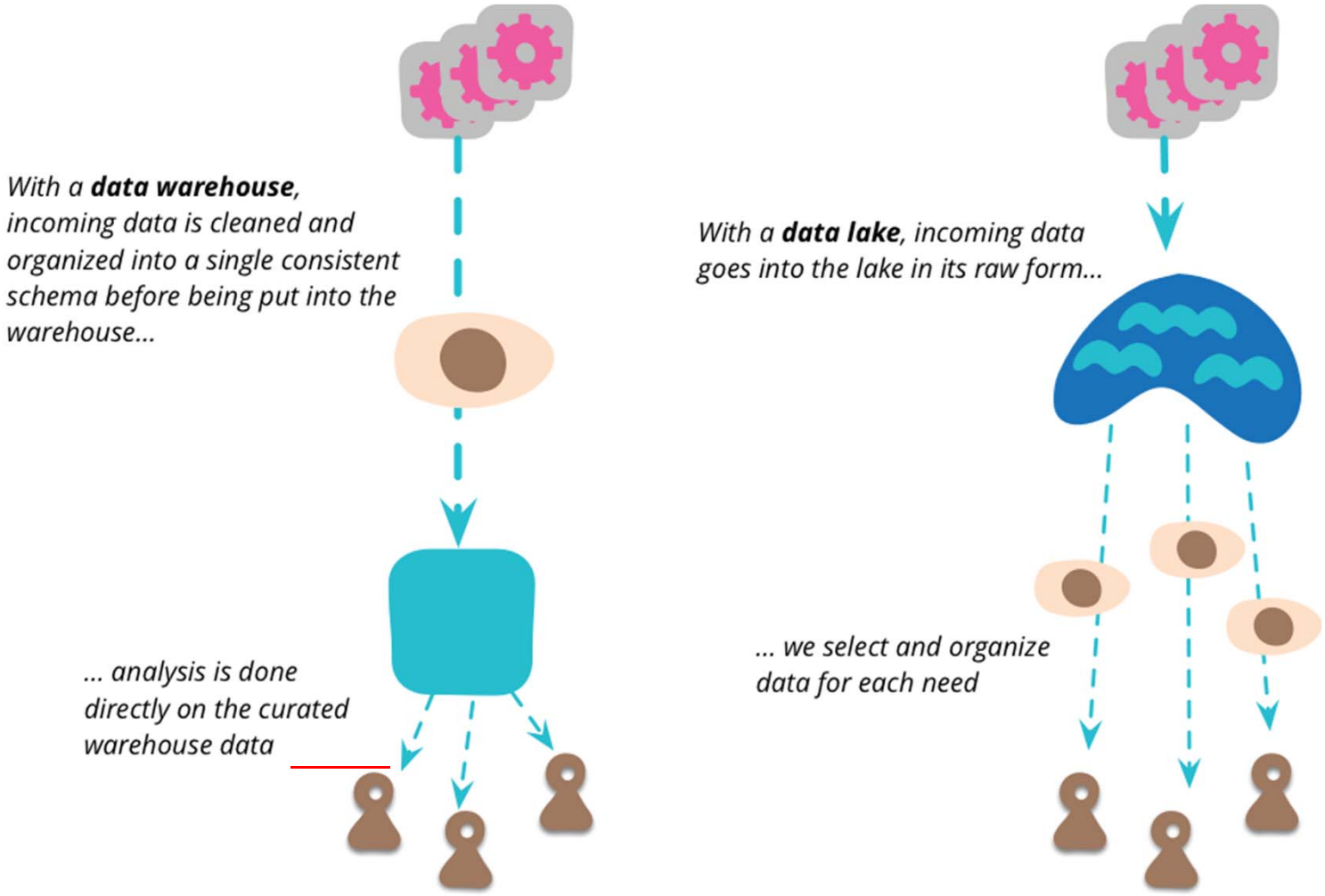
Source : Thomas Eri et.al, “Big Data Fundamentals: Concepts, Drivers & Techniques,” Prentice Hall, 2016

Data characteristics have been changed !



Source : rodneyrohrmann.blogspot.com

Differences between Data Lake and Data Warehouse



Source: martinfowler.com/bliki/DataLake.html

Data Warehouse vs. Data Lake

Data Warehouse	Issues	Data Lake
Structured Processed	DATA	Structured/Semi/Unstructured Raw
Schema-on-Write	DATA PROCESS	Schema-on-Read
Expensive for large data volumes	STORAGE	Designed for low cost storage
Less agile; fixed configuration	AGILITY	Highly agile; configure as required
Mature	SECURITY	Maturing
Business professionals	USER	Data scientists et al.

Adapted from Dave Kellermanns, "What is the Difference Between a Data Lake and a Data Warehouse?"

Major Industries Impacted with Big Data

Source: https://viblo.asia/p/why-we-need-modern-big-data-integration-platform-gAm5yx1Dldb/

Summary & Transactional Data

รอบค่าใช้บริการ (Bill Cycle)	ยอดค่าใช้บริการรอบปัจจุบัน (Current Charge)	ยอดค้างชำระ (Outstanding Balance)	รวมยอดที่ต้องชำระ (Total Outstanding Balance)	วันครบกำหนดชำระรอบปัจจุบัน (Current Due Date)
พฤศจิกายน 2560 (24/10/2560-23/11/2560)	2,204.20	0.00	2,204.20	19/12/2560

วันที่	เวลา	หมายเลขปลายทาง	เครือข่าย	ต้นทาง	ปลายทาง	บริการเสริม	หน่วย	ค่าบริการที่ใช้จริง (บาท)	ค่าบริการที่เรียกเก็บ (บาท)
โทรพื้นที่เดียวกัน (นาที:วินาที)							198:0	297.00	0.00
24/10/2017	12:23:43	0853624306	AIS	Bangkok	AIS		5:0	7.50	0.00
24/10/2017	14:52:50	0853624306	AIS	VoWiFi	AIS		3:0	4.50	0.00
24/10/2017	17:14:44	0853624306	AIS	Bangkok	AIS		16:0	24.00	0.00
24/10/2017	18:31:15	0853624306	AIS	Bangkok	AIS		6:0	9.00	0.00
24/10/2017	22:40:01	0853624306	AIS	Bangkok	AIS		5:0	7.50	0.00
25/10/2017	10:51:07	0969645614		Bangkok	CAT		2:0	3.00	0.00
25/10/2017	15:50:04	0819319162	AIS	Ratchaburi	AIS		1:0	1.50	0.00
27/10/2017	20:10:59	0815510308		Ratchaburi	DTN		2:0	3.00	0.00
28/10/2017	15:59:52	0955625050		VoWiFi	CAT		1:0	1.50	0.00
30/10/2017	18:11:05	0853624306	AIS	Ratchaburi	AIS		4:0	6.00	0.00
31/10/2017	17:25:39	032200654	Landline	Ratchaburi	Ratchaburi		1:0	1.50	0.00
01/11/2017	10:49:16	028888888	Landline	Bangkok	Bangkok		1:0	1.50	0.00
01/11/2017	10:50:32	028888888	Landline	Bangkok	Bangkok		5:0	7.50	0.00
01/11/2017	14:21:09	0824526464		Bangkok	DTN		2:0	3.00	0.00
01/11/2017	19:28:08	0922838959		Bangkok	DTN		1:0	1.50	0.00
02/11/2017	17:13:34	0955625050		Bangkok	CAT		2:0	3.00	0.00
03/11/2017	09:21:52	0853624306	AIS	Bangkok	AIS		1:0	1.50	0.00
03/11/2017	11:17:48	0918899077		Bangkok	CAT		4:0	6.00	0.00
03/11/2017	11:21:08	0952044317		Bangkok	CAT		4:0	6.00	0.00

Big Data Use Case

COMPANY
Starbucks Coffee

INDUSTRY
Food & Beverage

EMPLOYEES
160,000

TYPE
Behavioral Analytics

PURPOSE:

Starbucks collects data on its customers' purchasing habits in order to send personalized ads and coupon offers to the consumers' mobile phones. The company also identifies trends indicating whether customers are losing interest in their product and directs offers specifically to those customers in order to regenerate interest.

Source: Big Data

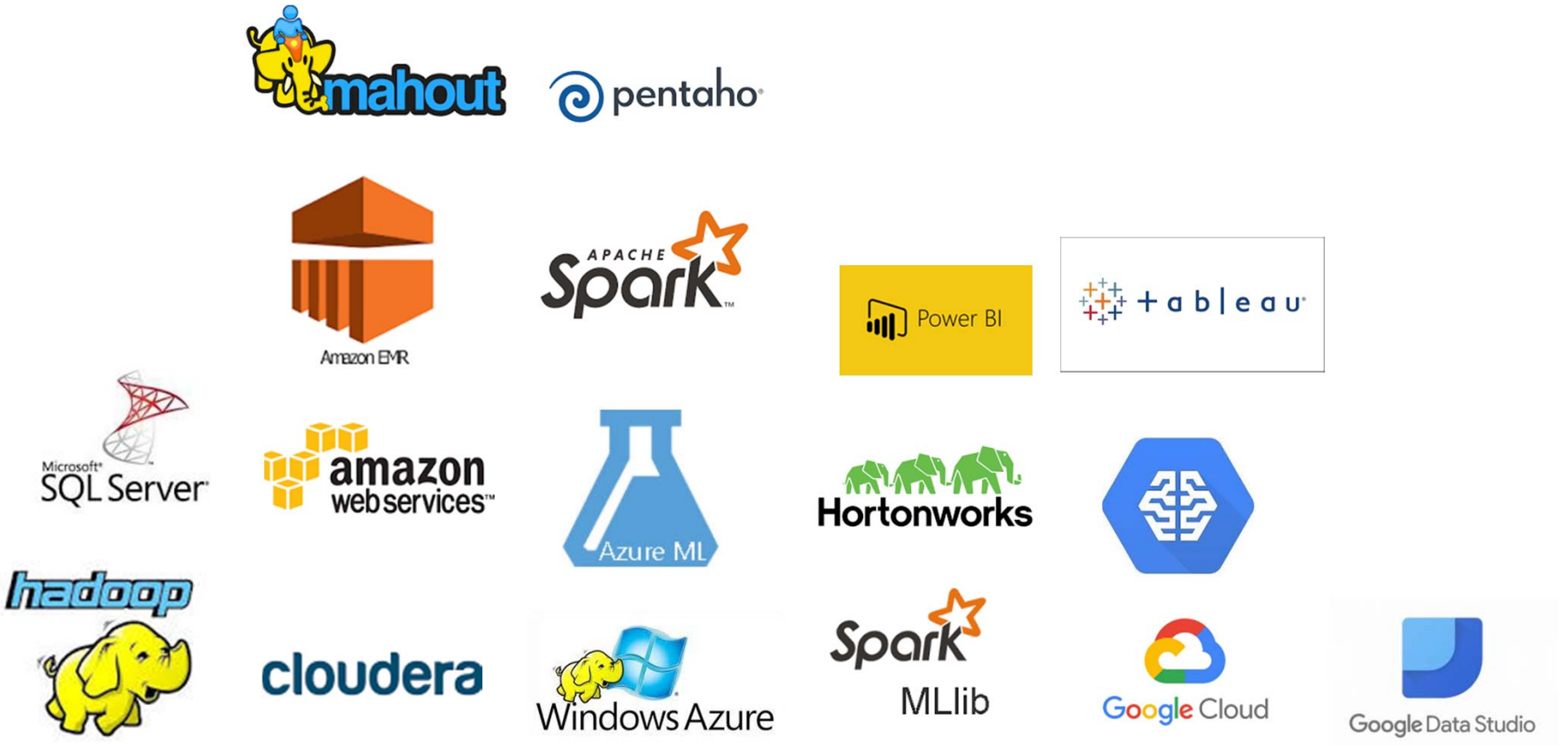
Big Data Use Case

PURPOSE:

Kayak uses big data analytics to create a predictive model that tells users if the price for a particular flight will go up or down within the next week. The system uses one billion search queries to find the cheapest flights, as well as popular destinations and the busiest airports. The algorithm is constantly improved by tracking the flights to see if its predictions are correct.

Source: Big Data

History and Evolution of Big Data Analytics



Technology

APACHE FLUME

Data Collection/ Ingestion

Google BigQuery

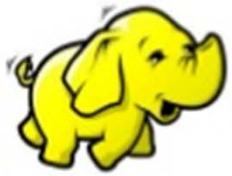
Google Cloud Storage

Data Storage

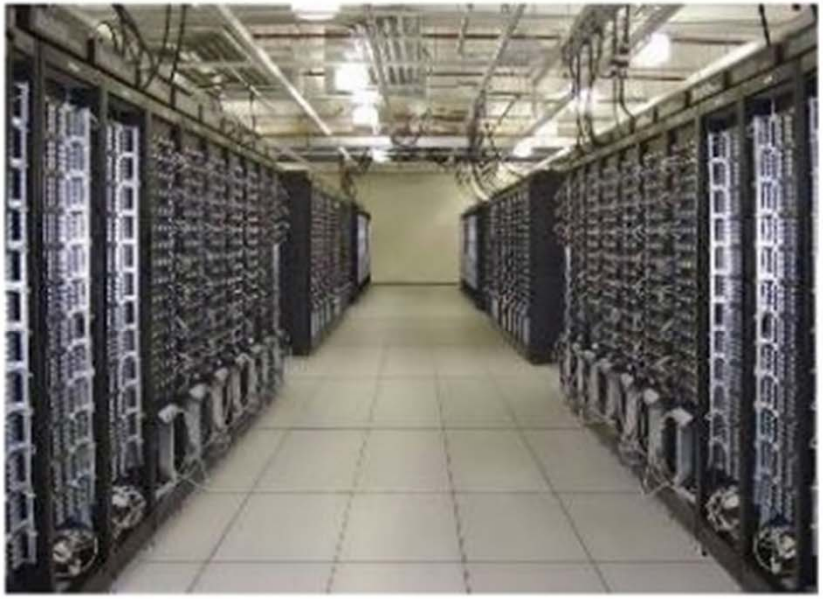
MLlib

Data Analysis/ Processing

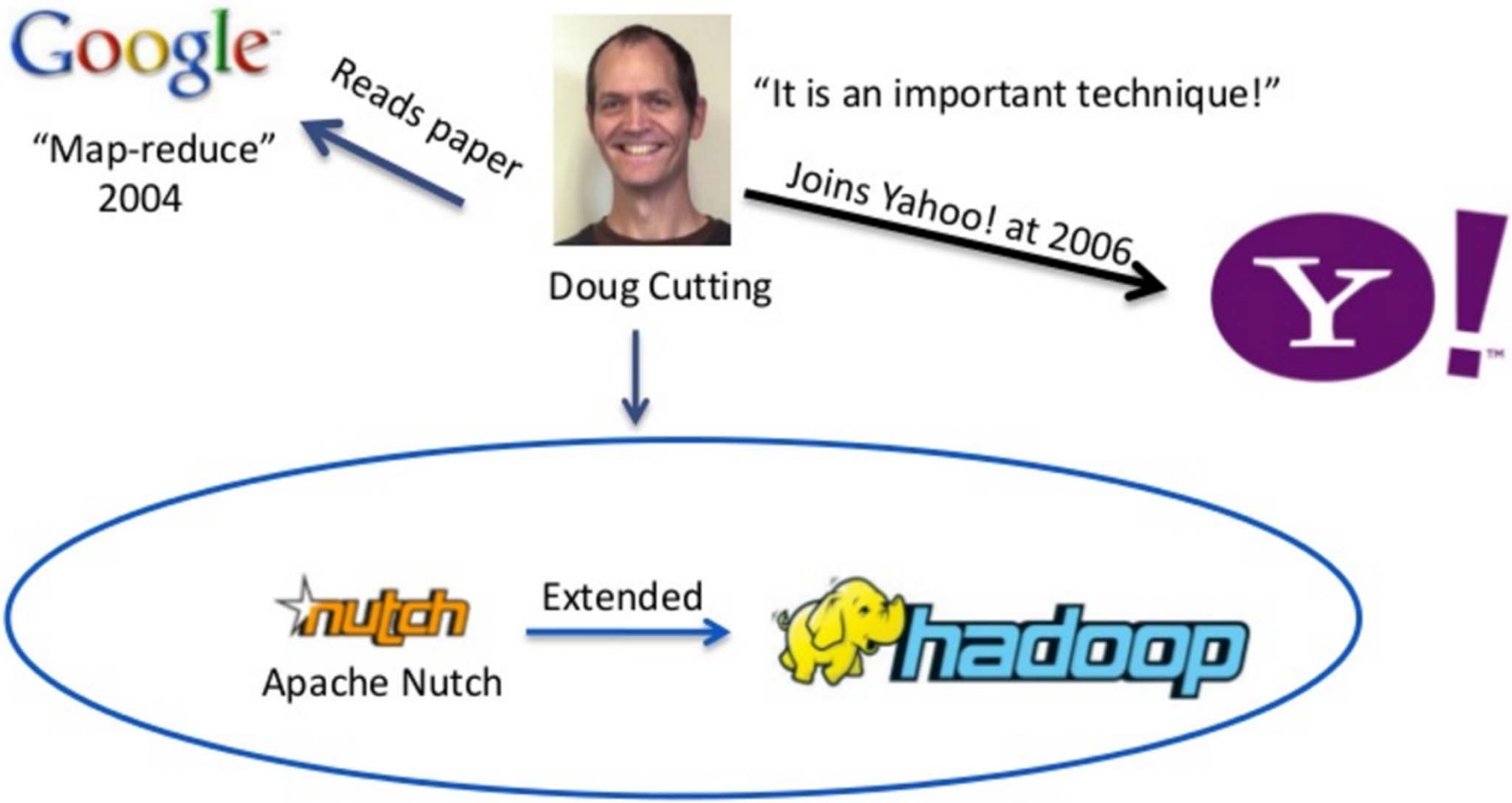
Data visualisation



What is HADOOP?

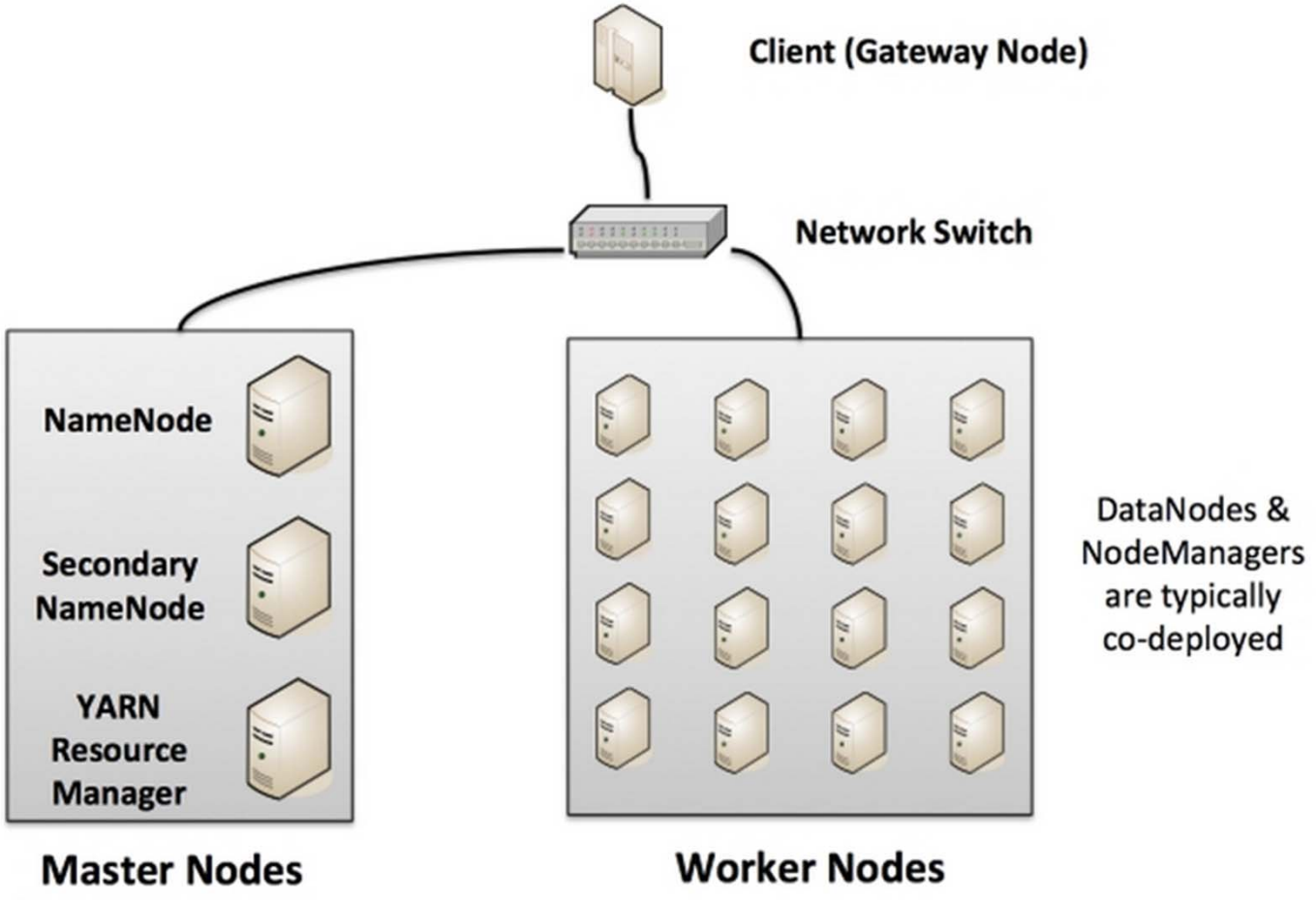


- Hadoop is open source framework for big data. Both distributed storage and processing.
- Hadoop is reliable and fault tolerant with no rely on hardware for these properties.
- Hadoop has unique horisontal scalability. Currently — from single computer up to thousands of cluster nodes.



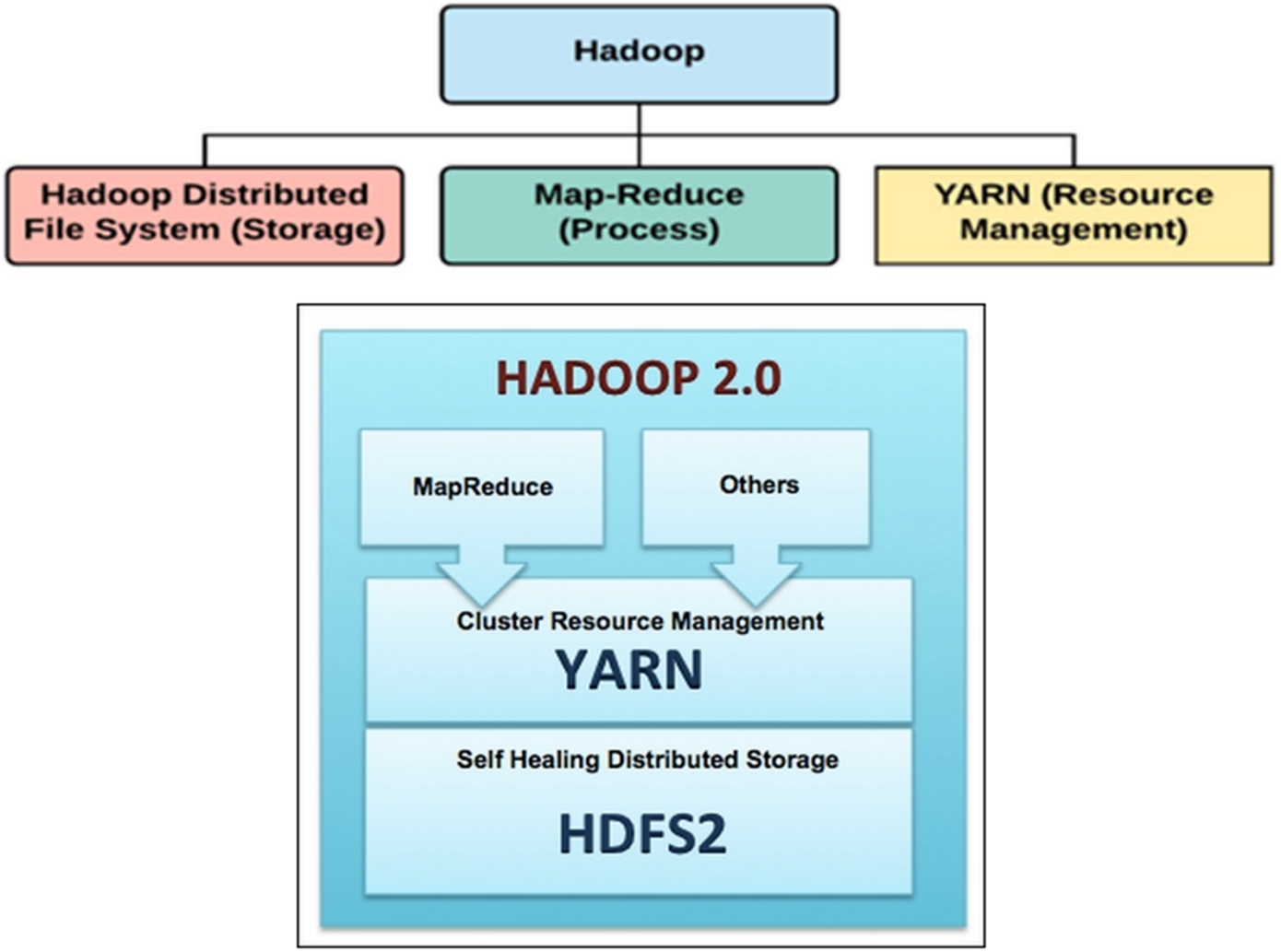
Source: Xiaoxiao Shi

Hadoop Cluster



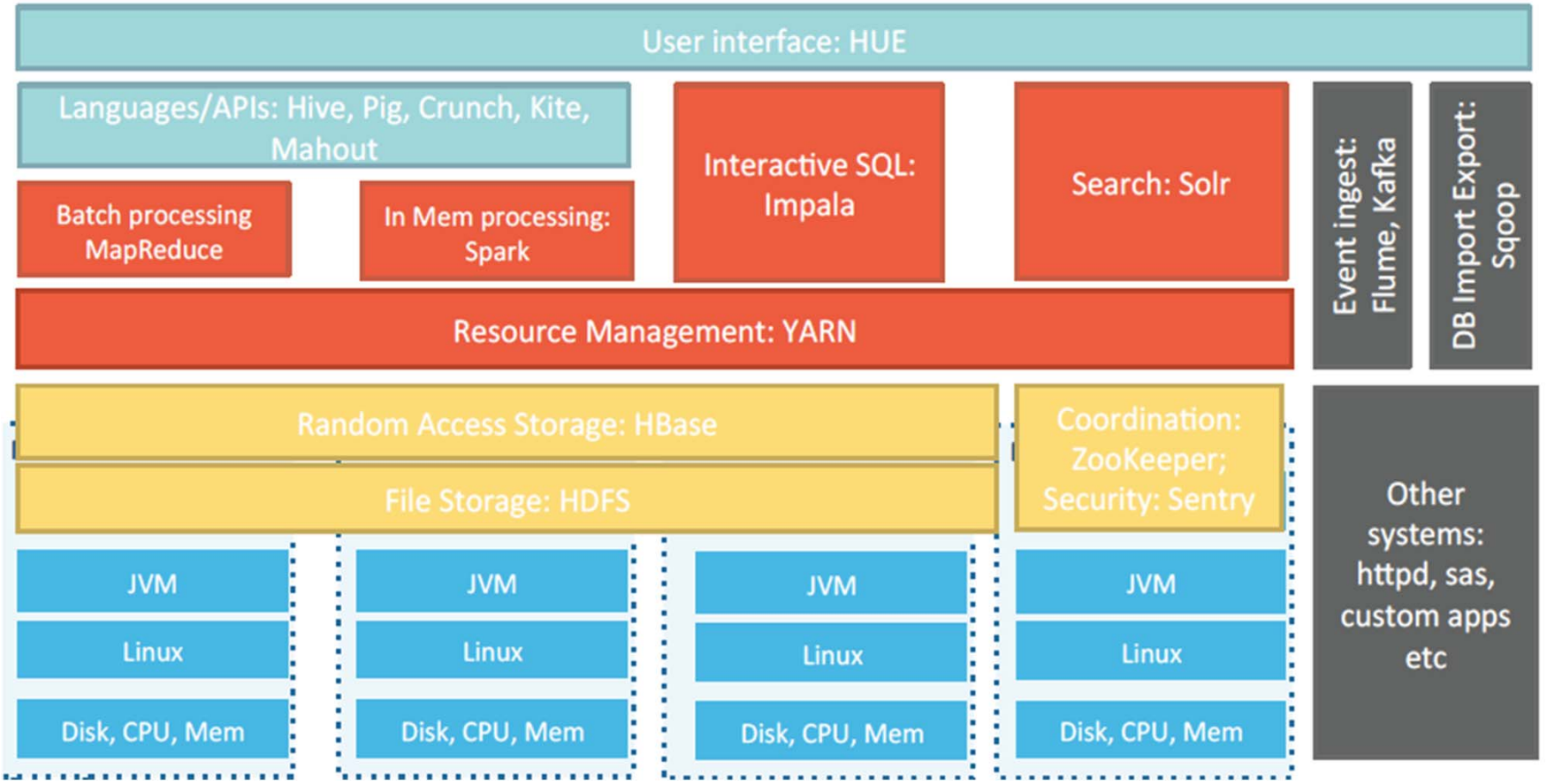
Source: HDInsight Essentials - Second Edition

Hadoop 2.0



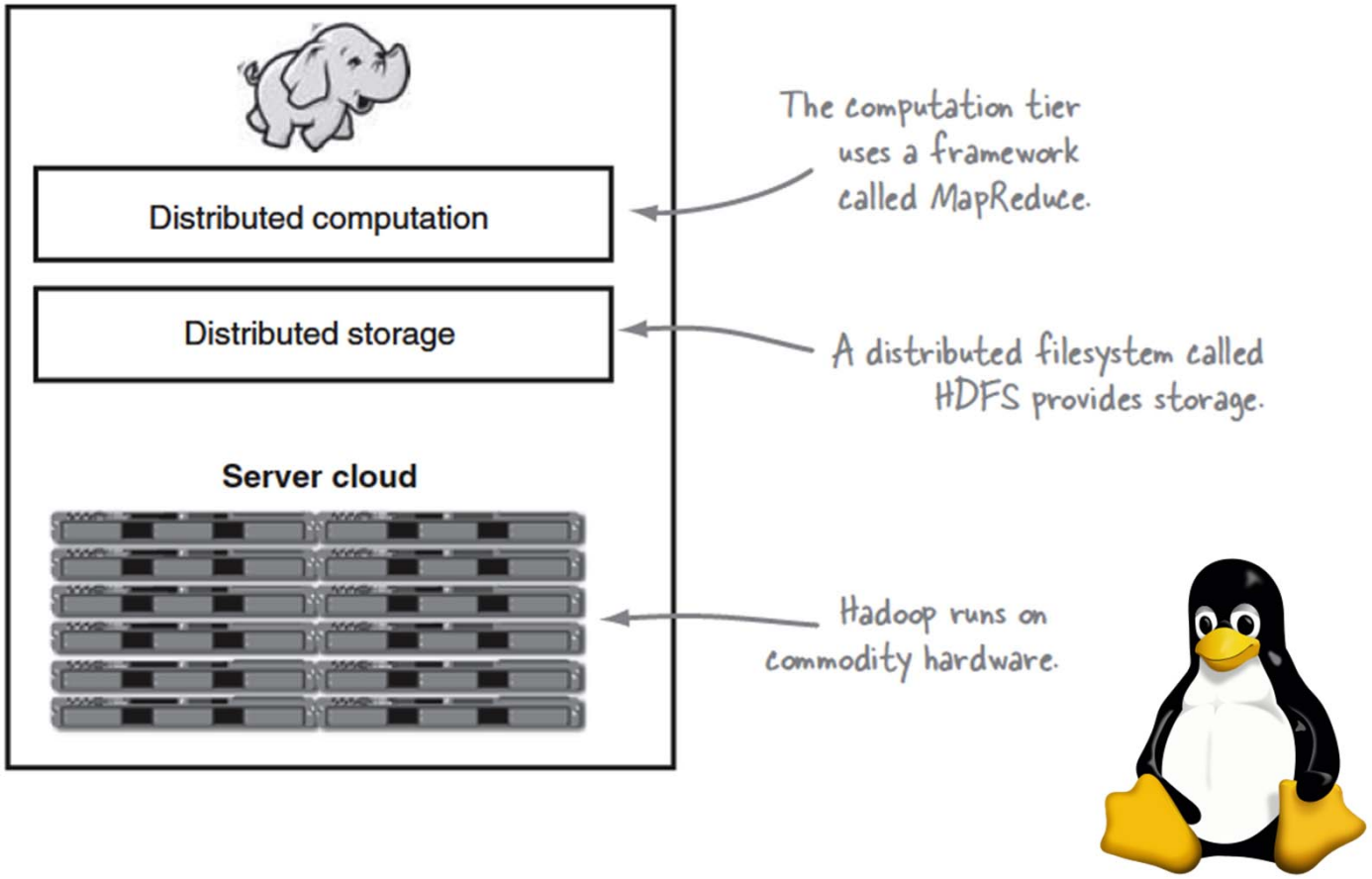
Source: HDInsight Essentials - Second Edition

Hadoop Ecosystem



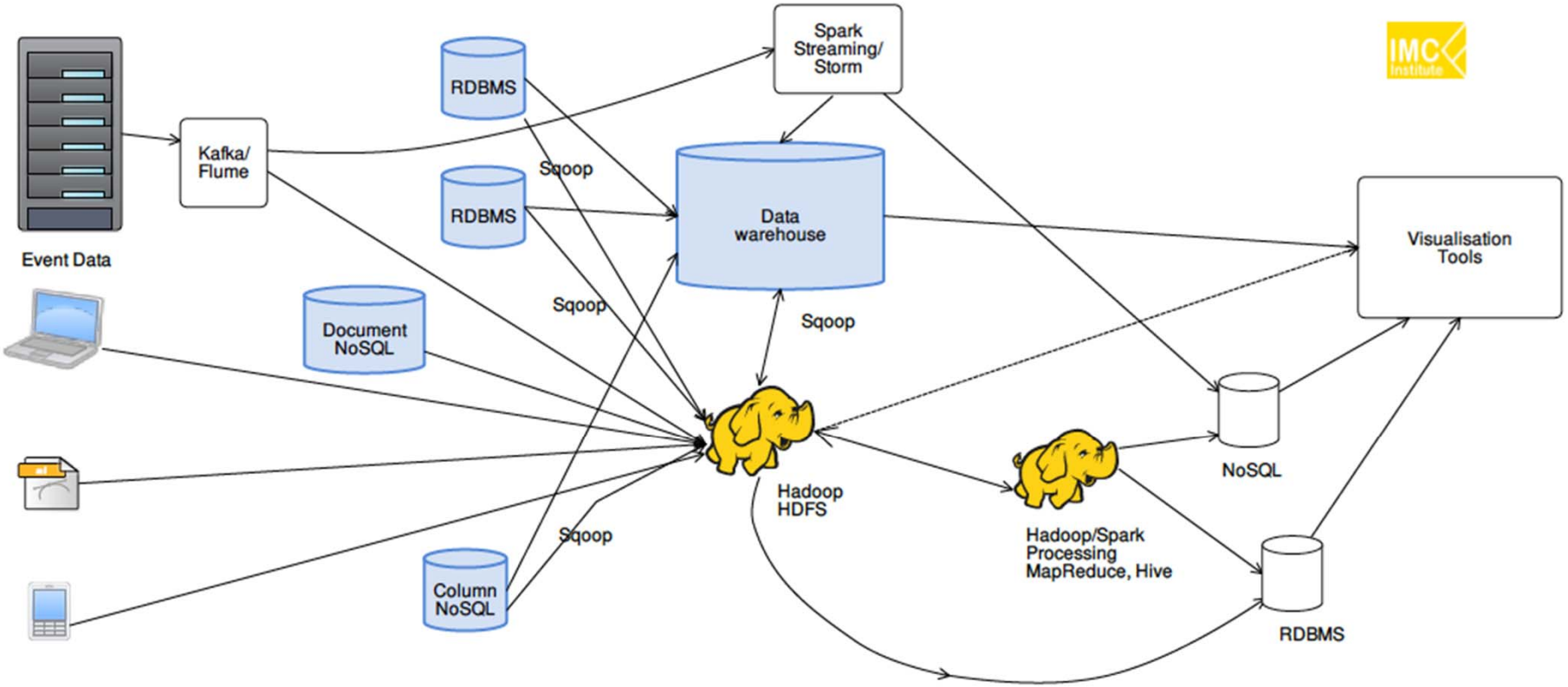
Source: Apache Hadoop Operations for Production Systems, Cloudera

Hadoop Environment

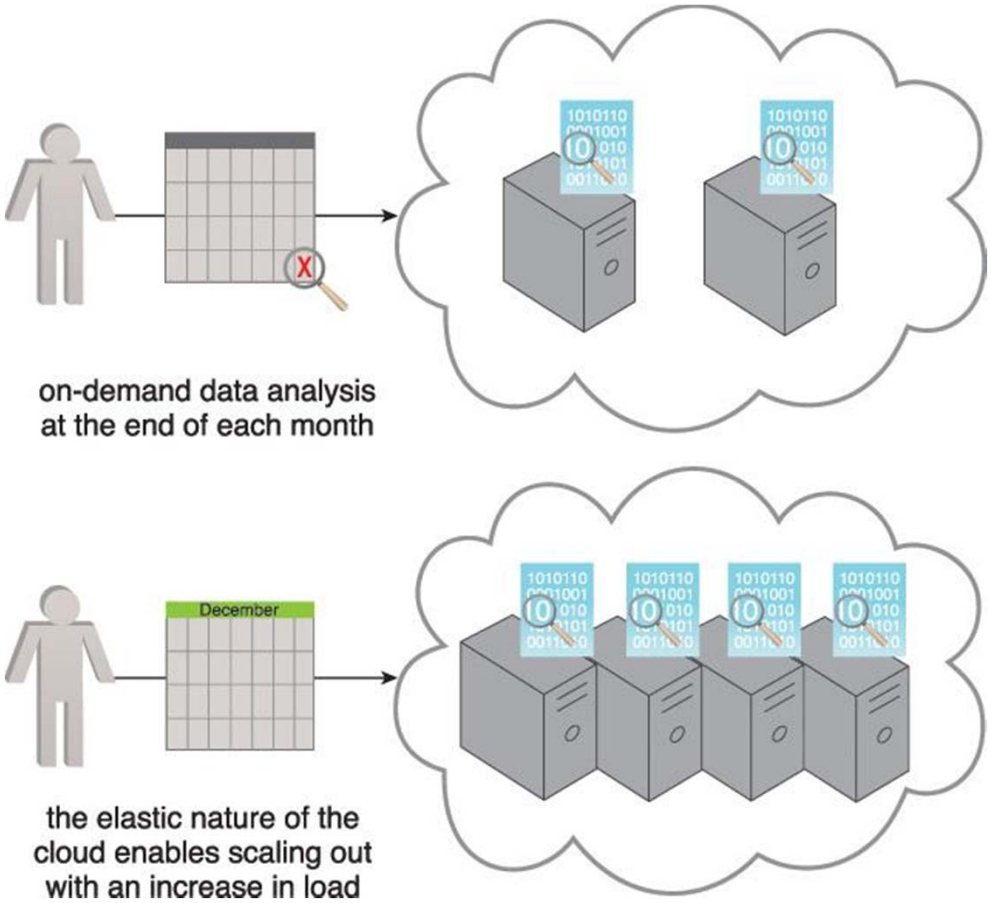


Source: Hadoop in Practice; Alex Holmes

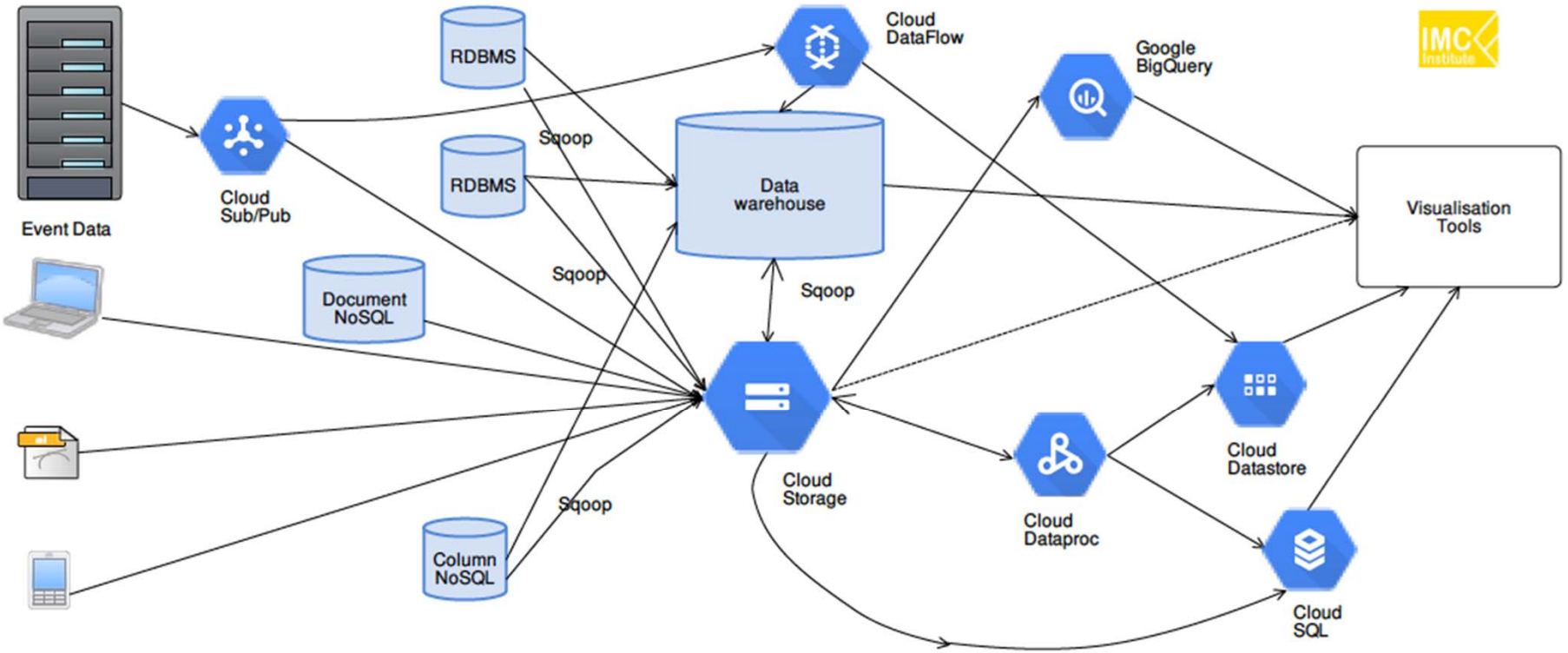
Big Data Architecture-Hadoop



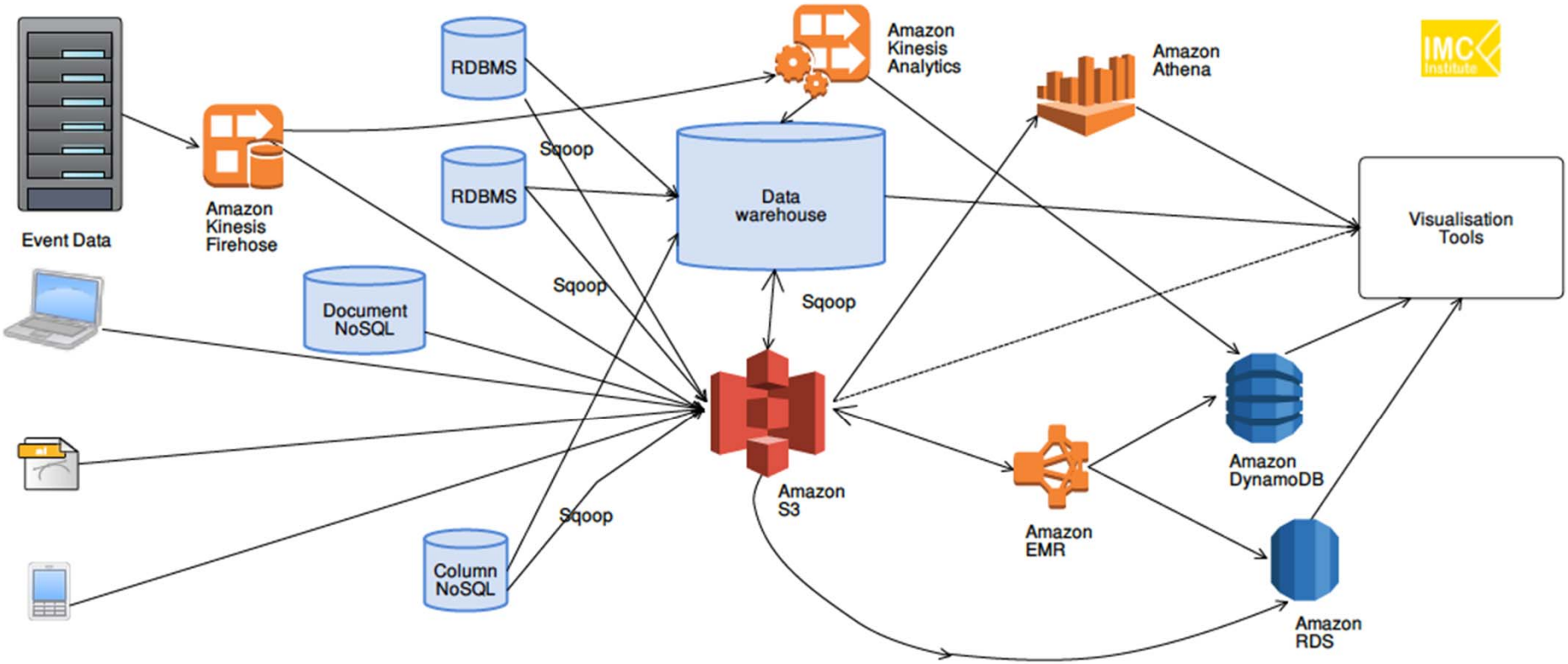
Why Cloud Computing ?



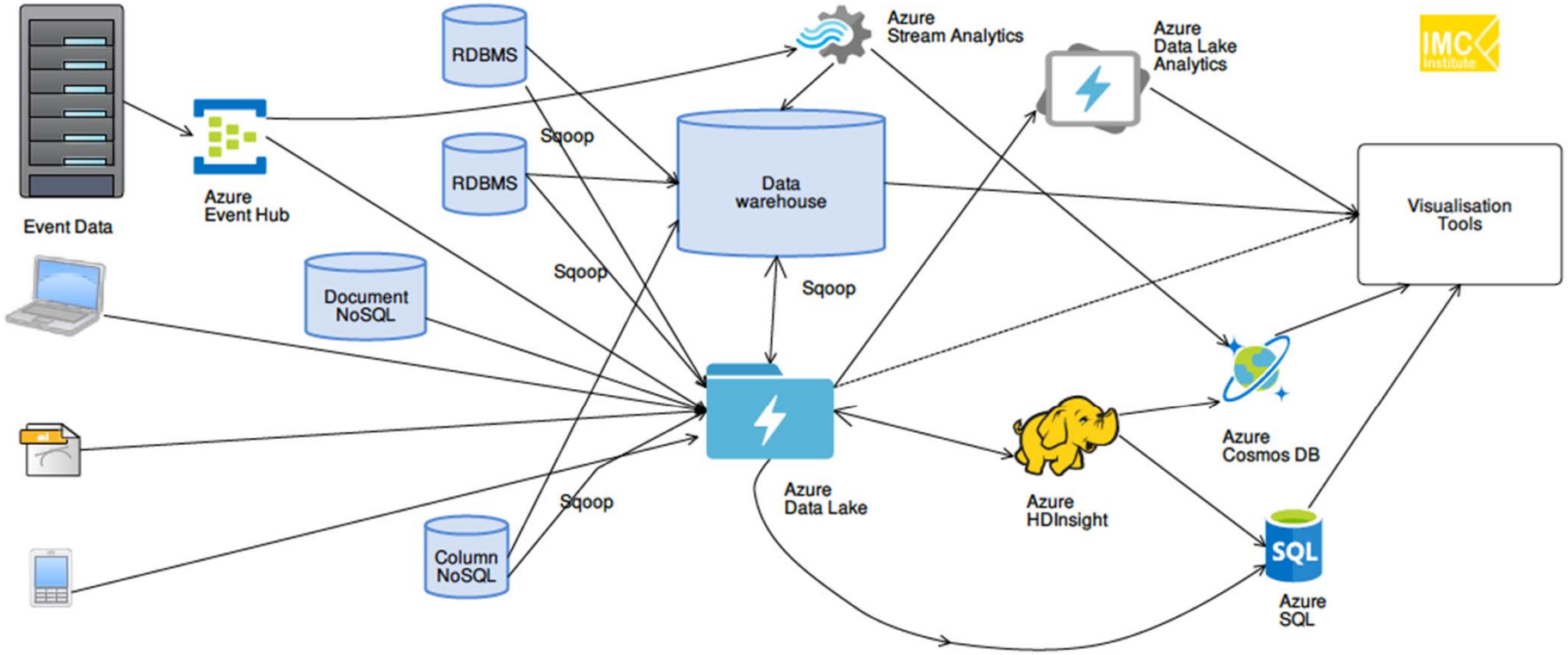
Big Data Architecture-GCP



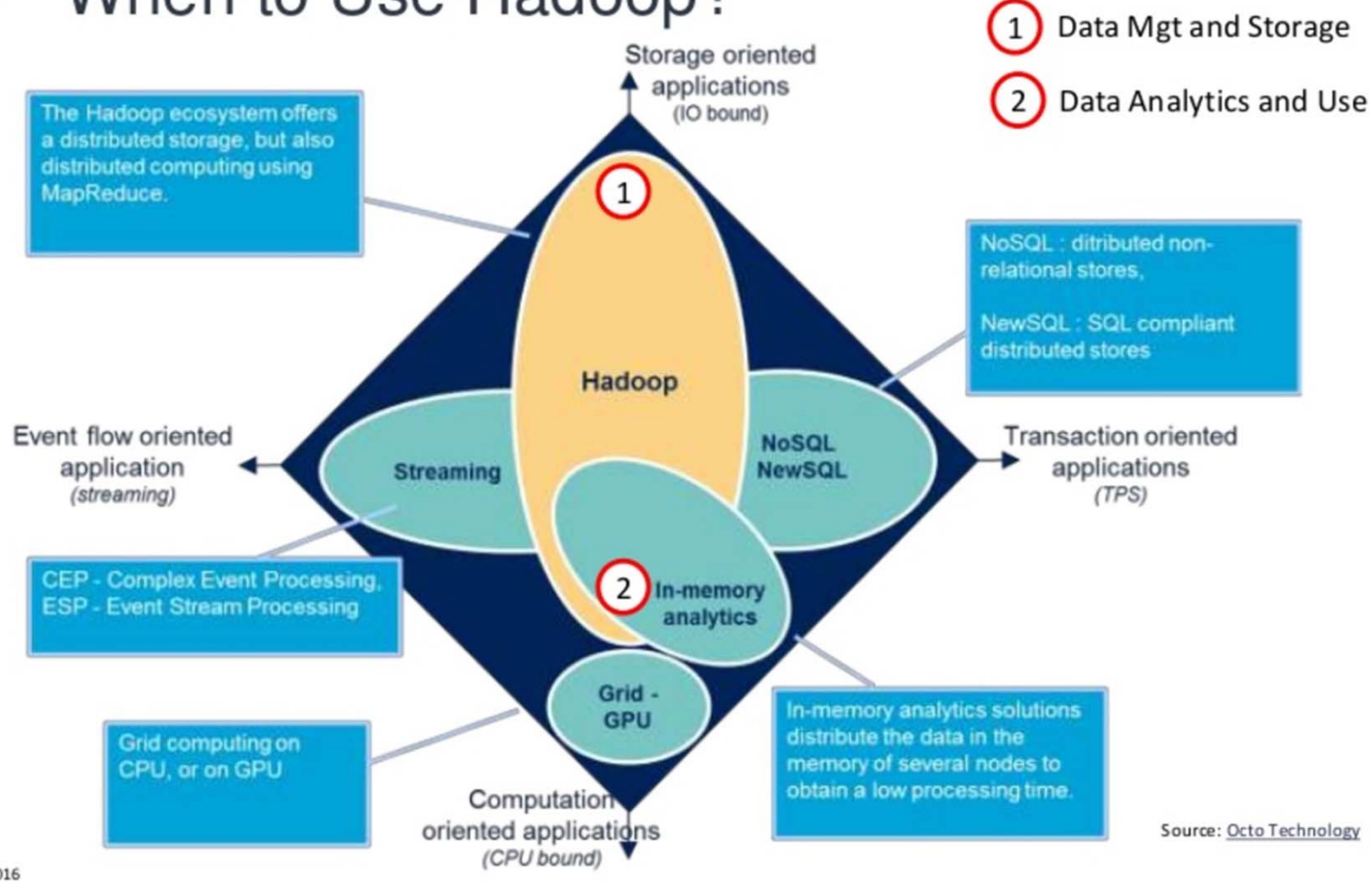
Big Data Architecture-AWS



Big Data Architecture-Azure



When to Use Hadoop?



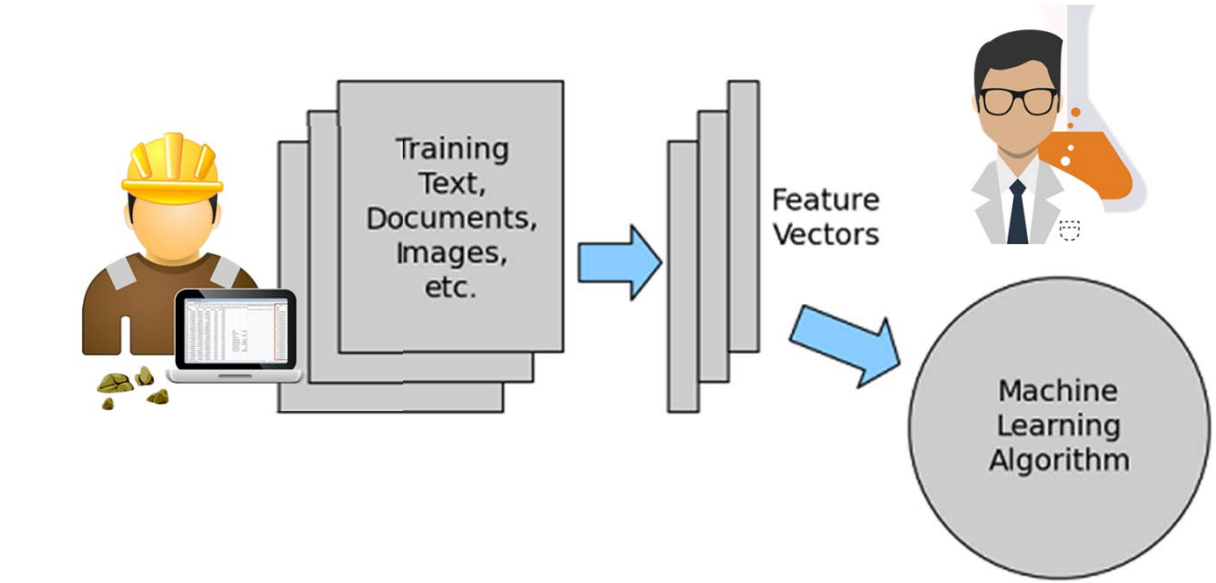
35

When Not to Use Hadoop

- Real Time Analytics (Solved in Hadoop V2)
  - Since Hadoop V1 cannot be used for real time analytics, people explored and developed a new way in which they can use the strength of Hadoop (HDFS) and make the processing real time. So, the industry accepted way is to store the Big Data in HDFS and mount Spark over it. By using spark the processing can be done in real time and in a flash (real quick). Apache Kudu is also a complementary solution to use.
- To Replace Existing Infrastructure
  - All the historical big data can be stored in Hadoop HDFS and it can be processed and transformed into a structured manageable data. After processing the data in Hadoop you often need to send the output to other database technologies for BI, decision support, reporting etc.
- Small Datasets
  - Hadoop framework is not recommended for small-structured datasets as you have other tools available in market which can do this work quite easily and at a fast pace than Hadoop. For a small data analytics, Hadoop can be costlier than other tools.

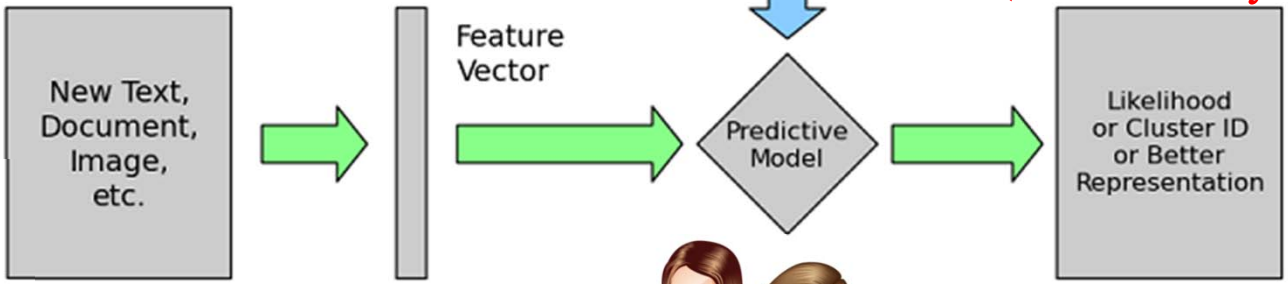
Source: Edureka

Platforms & Cost for Modeling & Prediction



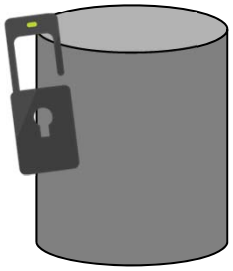
Dataproc (Big Data as a Service)

(5-6 mins/day * 30 days)



Real-Time Analytics Platform using Hadoop

Data Source



Authentication

Logs

Semi-

Structure

Ingestion

kafka

Real-Time Processing

Spark Streaming

Storage

HIVE

APACHE HBASE

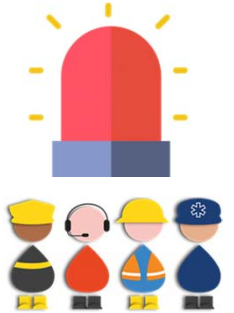
hadoop HDFS

Batch Processing

cloudera IMPALA

APACHE Spark

Presentation



Power BI

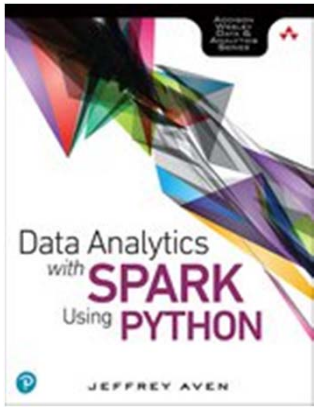
tableau

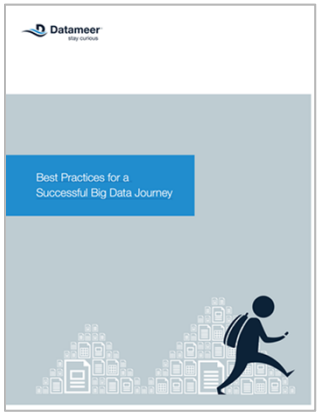
Why are they mostly being talked together ?

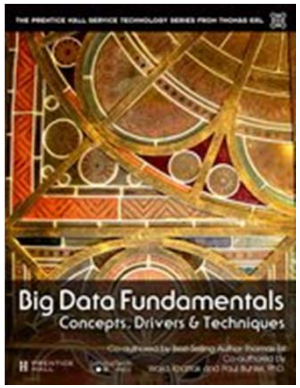
Big Data Analytics

Data Science

Strategic Value







Data Analytics & Big Data Technology

are together things to create values from

Big Data.