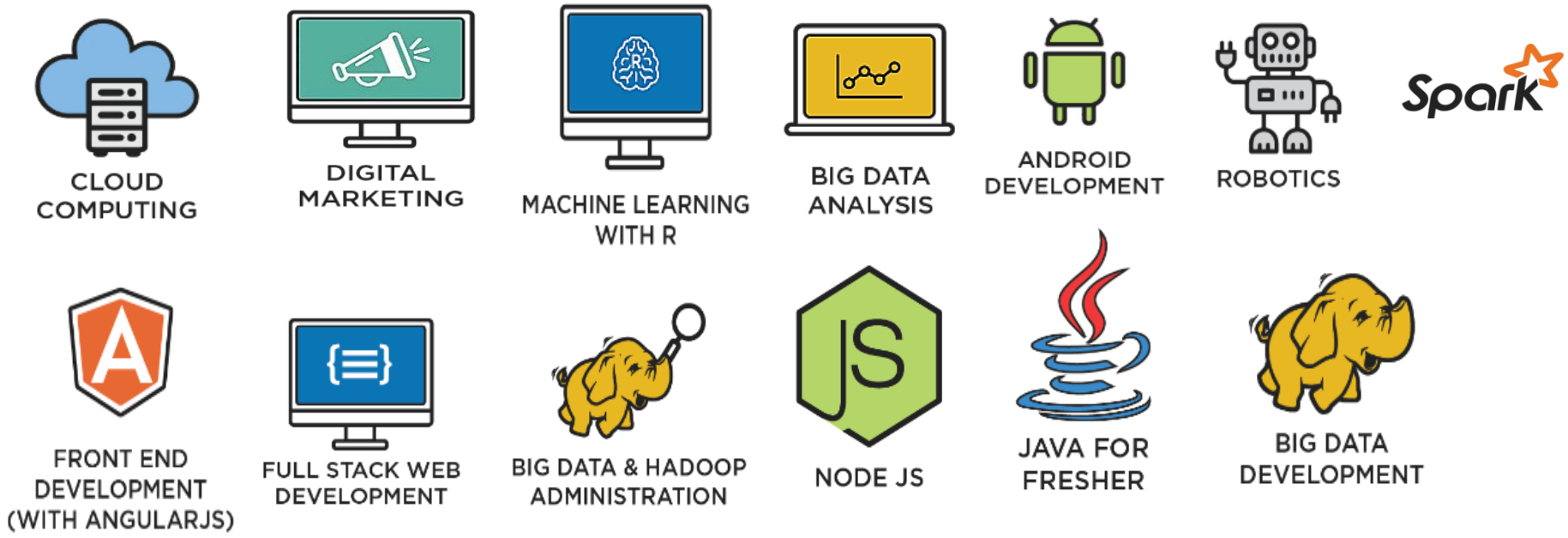

ACADGILD
LEARN. DO. EARN
BIG DATA
DEVELOPMENT

About AcadGild

ACADGILD

- AcadGild is a technology education start-up which provides online courses in the latest technologies.



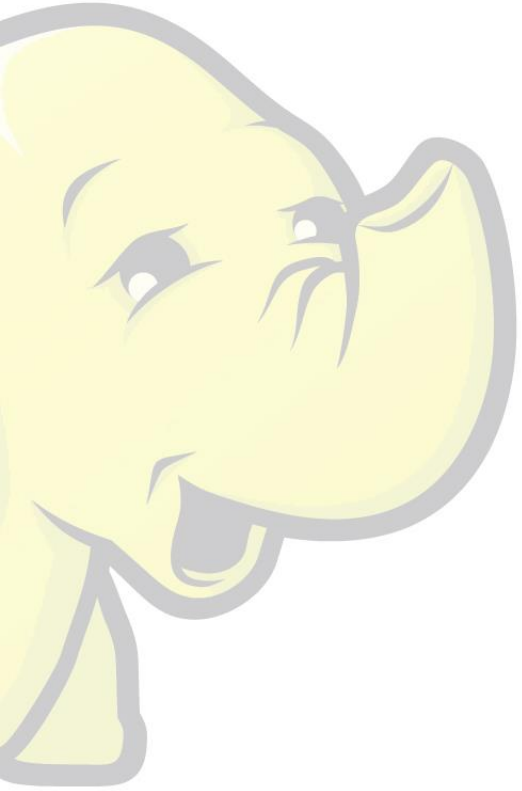
- AcadGild was founded by Mr. Vinod Dham, more popularly known as the “Father of the Pentium Chips.”

- Our aim is to provide millions of high school graduates, college graduates, and working professionals, skills to make them ready for jobs.

Course Objectives

ACADGILD

- Define and describe big data with examples
- Understand the solution of Big Data-Hadoop Technology
- Learn about the Hadoop architecture
- Create your own single node and multi-node cluster set-up on a Linux platform
- Learn about the key components of the Hadoop ecosystem in depth
    - MapReduce, Yarn, HDFS, Pig, Hive, Hbase, and Oozie
- Learn data loading techniques of Flume and Sqoop
- Assignments/hands-on lab practices will be imparted with step-by-step guidance on tasks; this will help you become an expert in writing your own programs on Hadoop2.x
    - Toward the end of the course, you will apply your learnings by working on two end-to-end real life projects on Hadoop!



Session 1

How to Solve the Big Data Problem

Agenda

ACADGILD

S. No.	Agenda Title
1.	Why is Data So Important?
2.	Pre-requisite: Data Scale
3.	What is Big Data?
4.	Big Bank: Big Challenge
5.	Customer Churn Analysis
6.	Point-of-Sale Transaction Analysis
7.	Common Problems
8.	3 Vs of Big Data
9.	Defining Big Data
10.	Sources of Data Flood
11.	Exploding Data Problem
12.	Redefining the Challenges of Big Data
13.	Possible Solutions

S. No.	Agenda Title
14.	Scaling Up Vs. Scaling Out
15.	Challenges of Scaling Out
16.	Solution for Data Explosion-Hadoop
17.	Hadoop: Introduction

Why is Data So Important?

ACADGILD

Every day challenge:

- Score Card
- Mobile Bills
- Movie Ratings
- Monthly Expenditure

It is data and its analysis which helps us **take better decisions** or **make better choices**

Pre-requisite – Data Scale

ACADGILD

- 8 bits = 1 byte
- 1024 bytes = 1 kilobyte (KB)
- 1024 KB = 1 megabyte (MB)
- 1024 MB = 1 gigabyte (GB)
- 1024 GB = 1 Terabyte (TB)

- **1 TB ==> 1,610 CDs worth of data**

- 1 TB is not the hard-drive capacity of a majority of commodity/personal machines even today
- 1024 TB = 1 Petabyte (PB)

- **1 PB of data ==> 2,23,100 DVDs**

- **Facebook processes ~300 PB everyday**

- An Exabyte (EB) = 1,024 PB
- A Zettabyte (ZB) = 1,024 EB
- A Yottabyte (YB) = 1,024 ZB

What is Big Data?

ACADGILD

Big Bank: Big Challenge

ACADGILD

- A large bank has to take separate data warehouses from multiple departments and combine them into a single global repository for analysis

- That very large bank with several consumer lines of business need to analyze customer activity across multiple products to predict credit risk with greater accuracy

Big Bank: Big Challenge (Contd.)

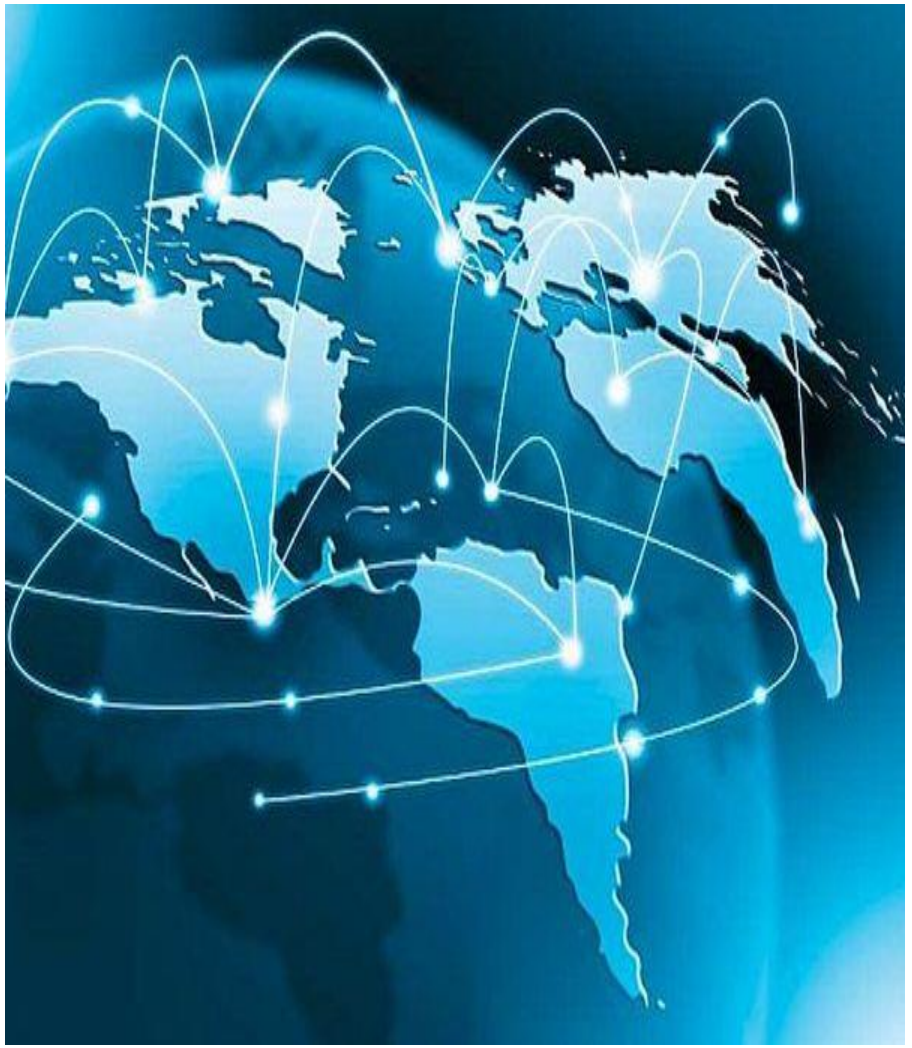
ACADGILD

- With the economic downturn of 2008, the bank had significant exposure in its mortgage business to defaults by its borrowers
- Understanding that the risk required by banks to build a comprehensive picture of its customers is challenging

Customer Churn Analysis

ACADGILD

- A large telecommunications provider analyzed call logs and complex data from multiple sources
- A large mobile carrier needed to analyze multiple data sources to understand how and why customers decided to terminate their service contracts
- Were customers actually leaving or were they merely trading one service plan for another?
- Were they leaving the company entirely and moving to a competitor?
- Were pricing, coverage gaps, or device issues a factor?
- What other issues were important, and how could the provider improve satisfaction and retain customers?



Point-of-Sale Transaction Analysis

ACADGILD

- A large retailer doing Point-of-Sale transactional analysis needs to combine larger quantities of PoS transaction analysis to forecast demand and improve the return.
- It wants to combine this new information with recent and historical sales data from PoS systems to increase sales and improve them.
- Traditional data warehousing systems are an expensive place to store complex data from new sources.
- They do not, generally, support the kind of sophisticated analyses—sentiment, language processing and others—that apply to this new data.

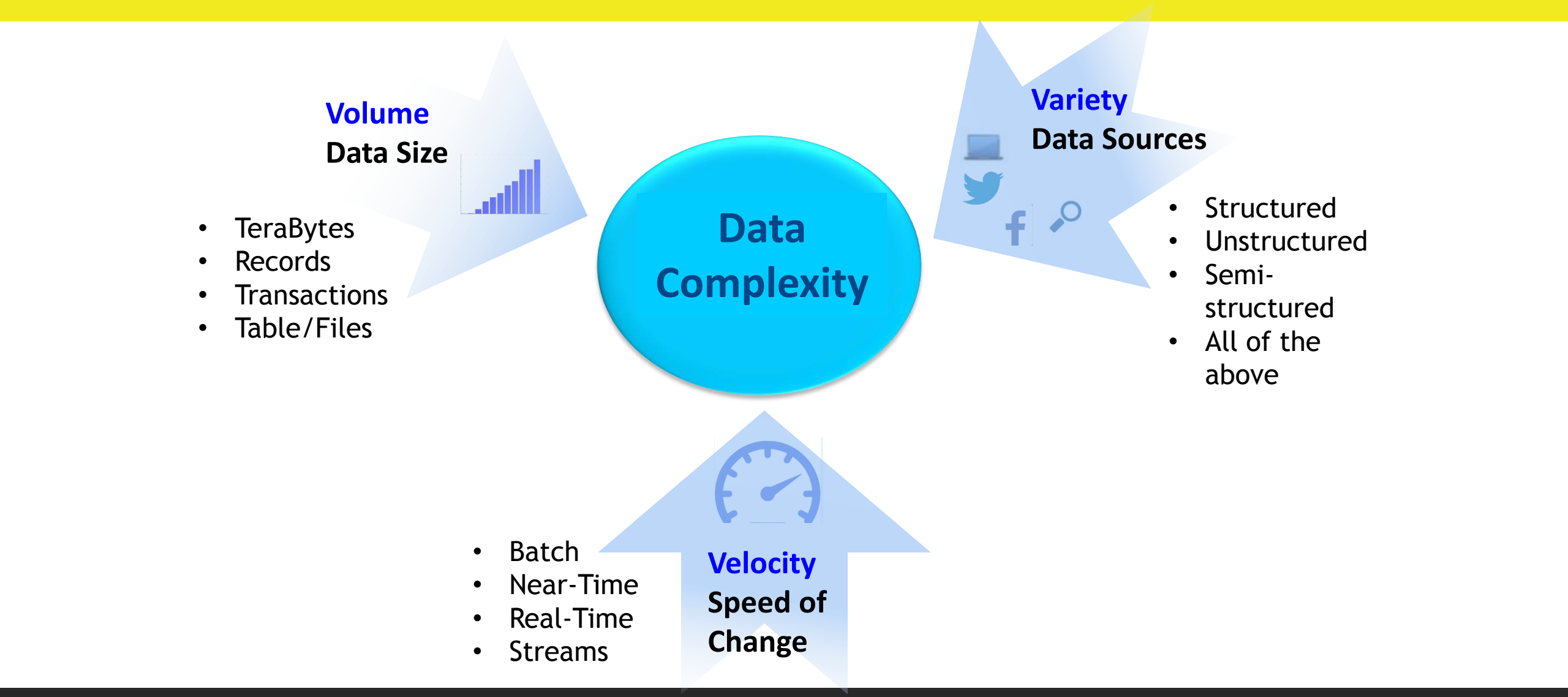
Common Problems

ACADGILD

- Unimaginable size of data
- Heterogeneous systems
- Traditional systems do not scale up
- RDBMS is costly
- Building single system is complex and not cost effective
- Data Analysis, Machine Learning, and Predictive Analysis do not have a common infrastructure

3 Vs of Big Data

ACADGILD



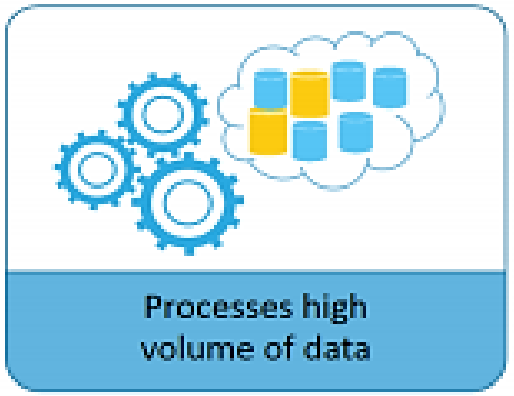
Defining Big Data

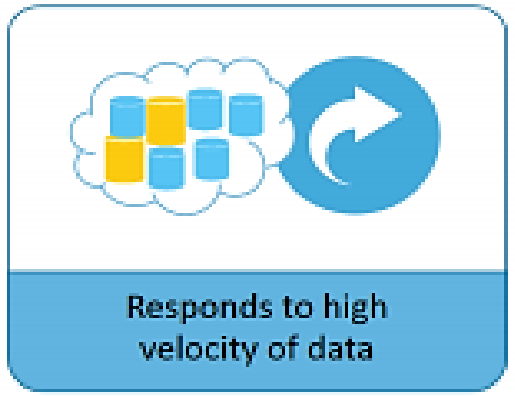
ACADGILD

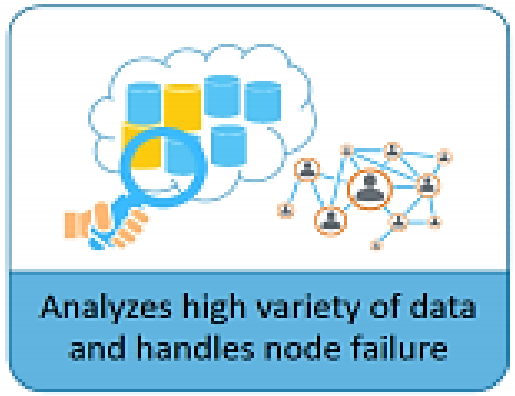
- According to Gartner, Big Data is a **high-volume**, **high-velocity**, and **high-variety** information asset that demands cost-effective, innovative forms of information processing for enhanced insight and decision making

Defining Big Data (Contd.)

ACADGILD





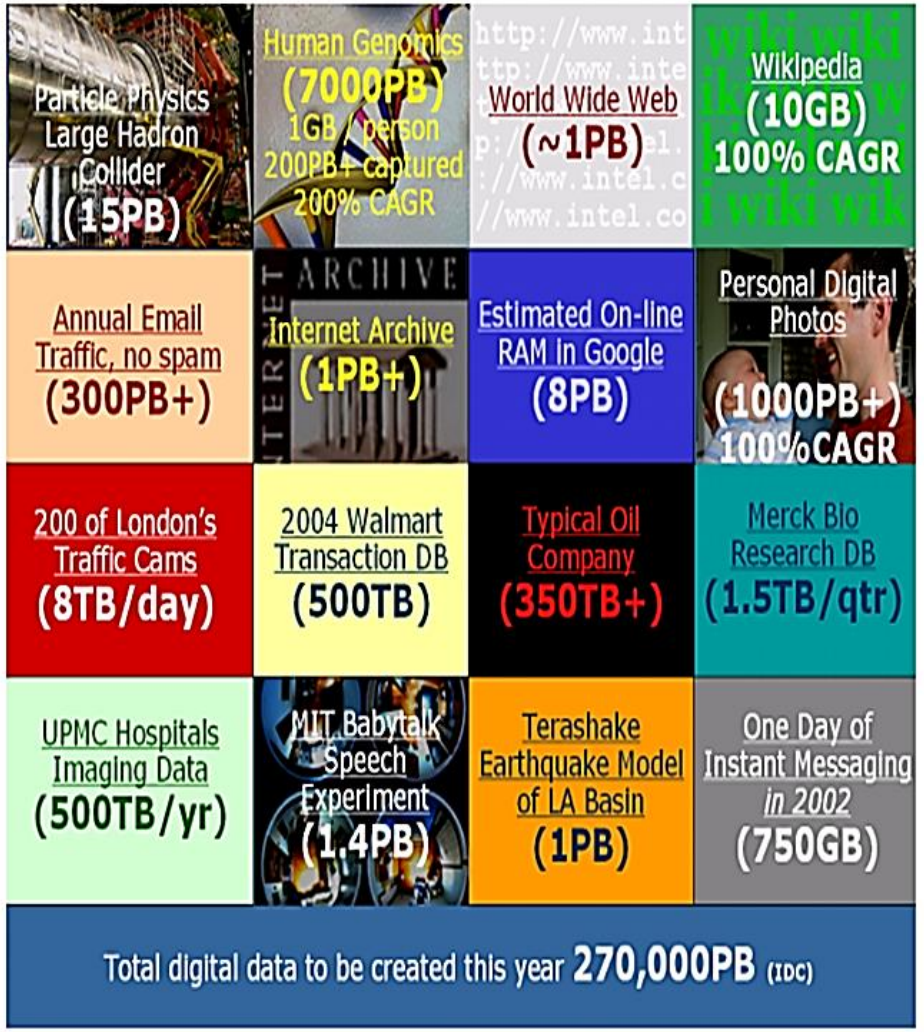


- According to IBM, the Big Data technology has helped turn the 12 terabytes of tweets created daily into improved product sentiment analysis
- Big Data technology has scrutinized 5 million trade events created daily to identify potential frauds. It has helped in analyzing 500 million daily call detail records in real time to predict the “customer churn” faster
- Big Data technology has helped monitor hundreds of live video feeds from surveillance cameras to target points of interest for security agencies. It has also been able to exploit the 80% data growth in images, videos, and documents to improve customer satisfaction

Sources of Data Flood

ACADGILD

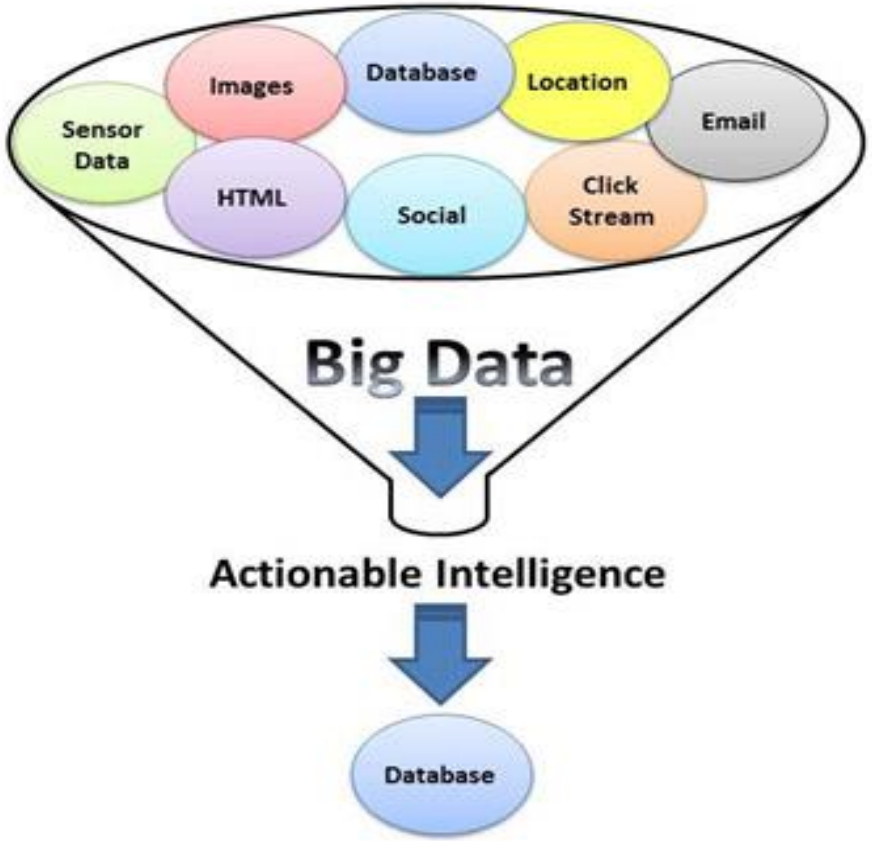
- This flood of data is coming from many sources
- The New York stock exchange generates about 4-5 terabytes of data everyday
- Facebook hosts more than 240 billion photos, growing at 7 petabytes of data everyday
- Ancestory.com, the genealogy site stores around 10 petabytes of the data
- The Internet Archive stores around 18.5 petabytes of data
- The Large Hadron Collider near Geneva produces about 30 Petabytes of data every year

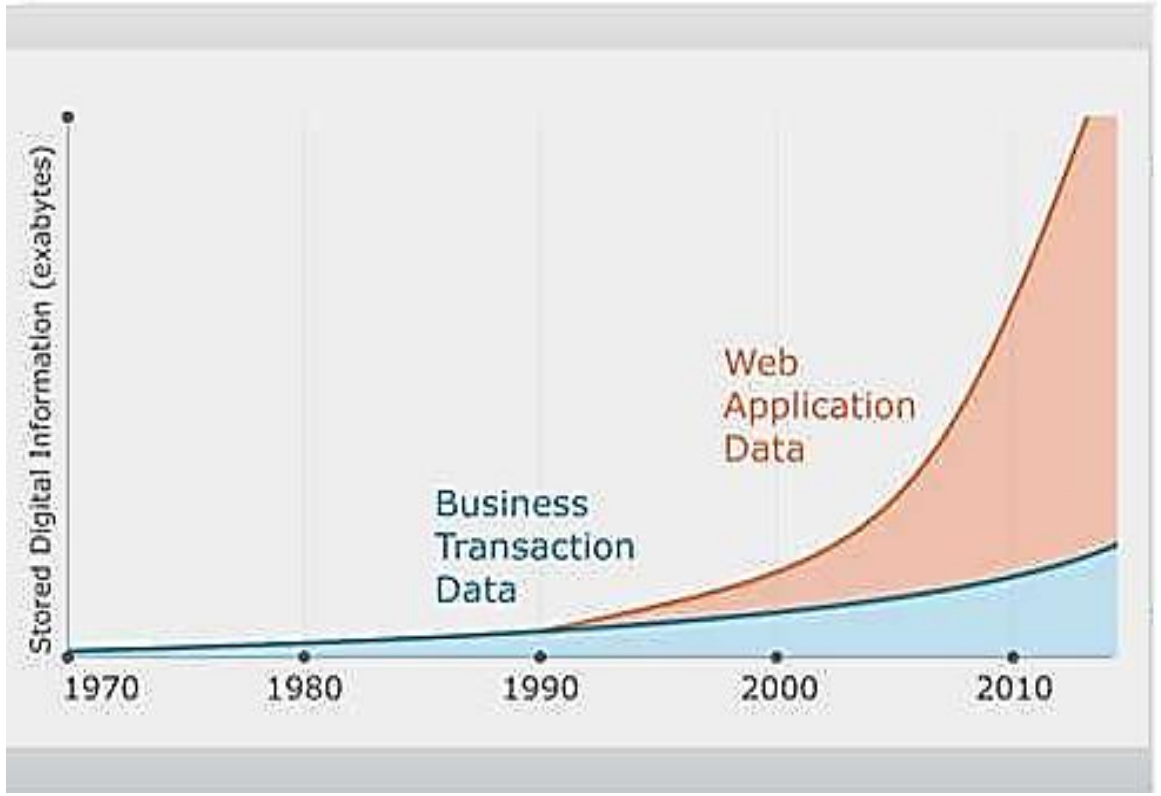


Exploding Data Problem

ACADGILD

- Big Data constitutes large data sets in petabytes & zettabyte which cannot be processed by a single machine within an expected timeframe





Quote

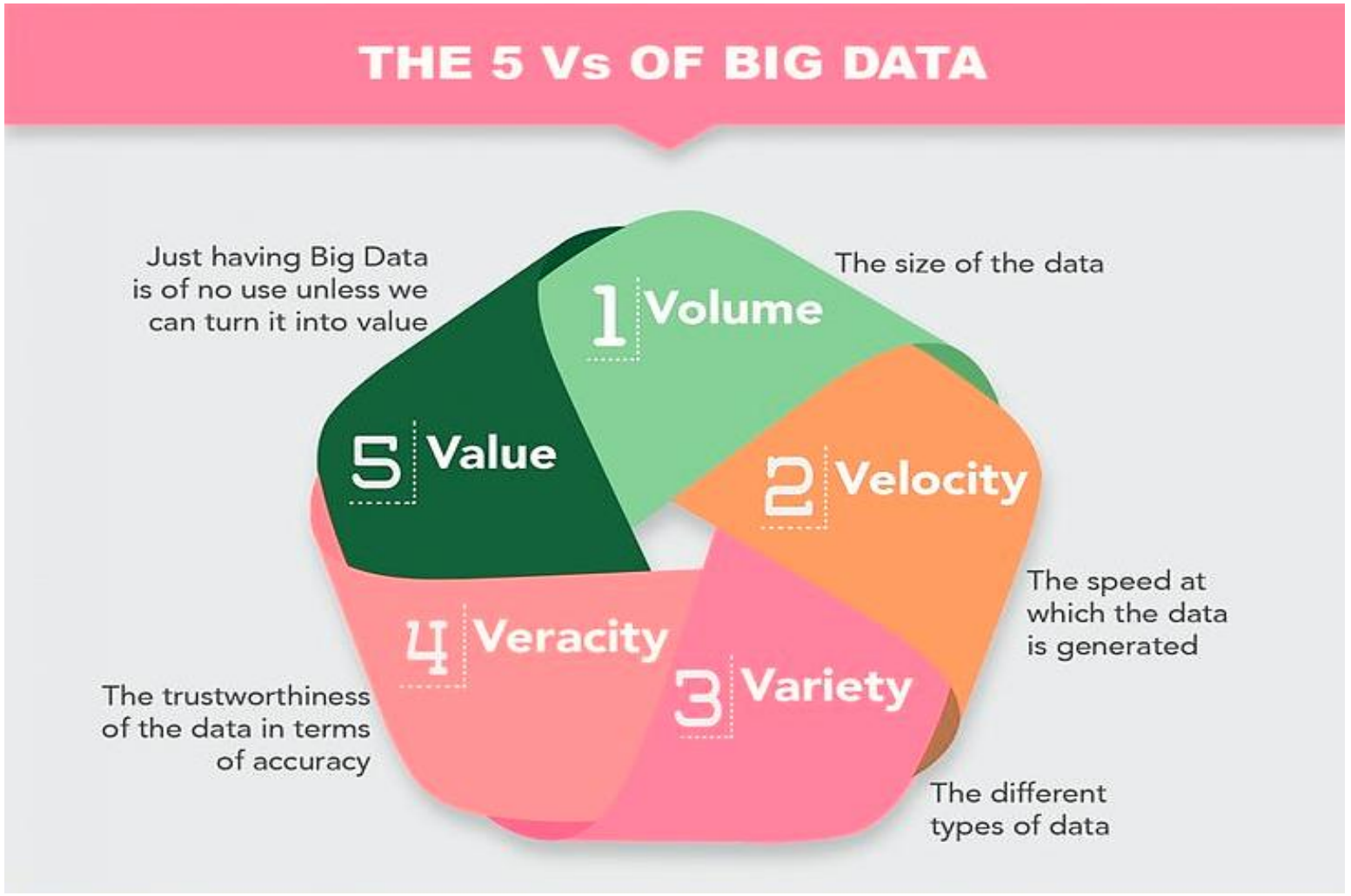
ACADGILD

- From Eric Schmidt, the ex-CEO of Google

Every two days now, we create as much information as we did from the dawn of civilization up until 2003, according to Schmidt. That's something like five exabytes of data now

Redefining the Challenges of Big Data

ACADGILD



Quiz

ACADGILD

Why is big data technology gaining so much attention?

1. To manage high volume of data in cost effective manner
2. To unify different varieties of data spread across heterogeneous systems
3. To capture data from fast-occurring events
4. To analyze high volume and wide variety of data to generate valuable insight

Ans: All the above

Quiz (Contd.)

Which of the following is not a challenge associated with Big Data?

1. High Volume
2. Large Velocity
3. Wide Variety
4. Viscosity of data

Ans: 4

Possible Solutions

ACADGILD

1. Scale Up

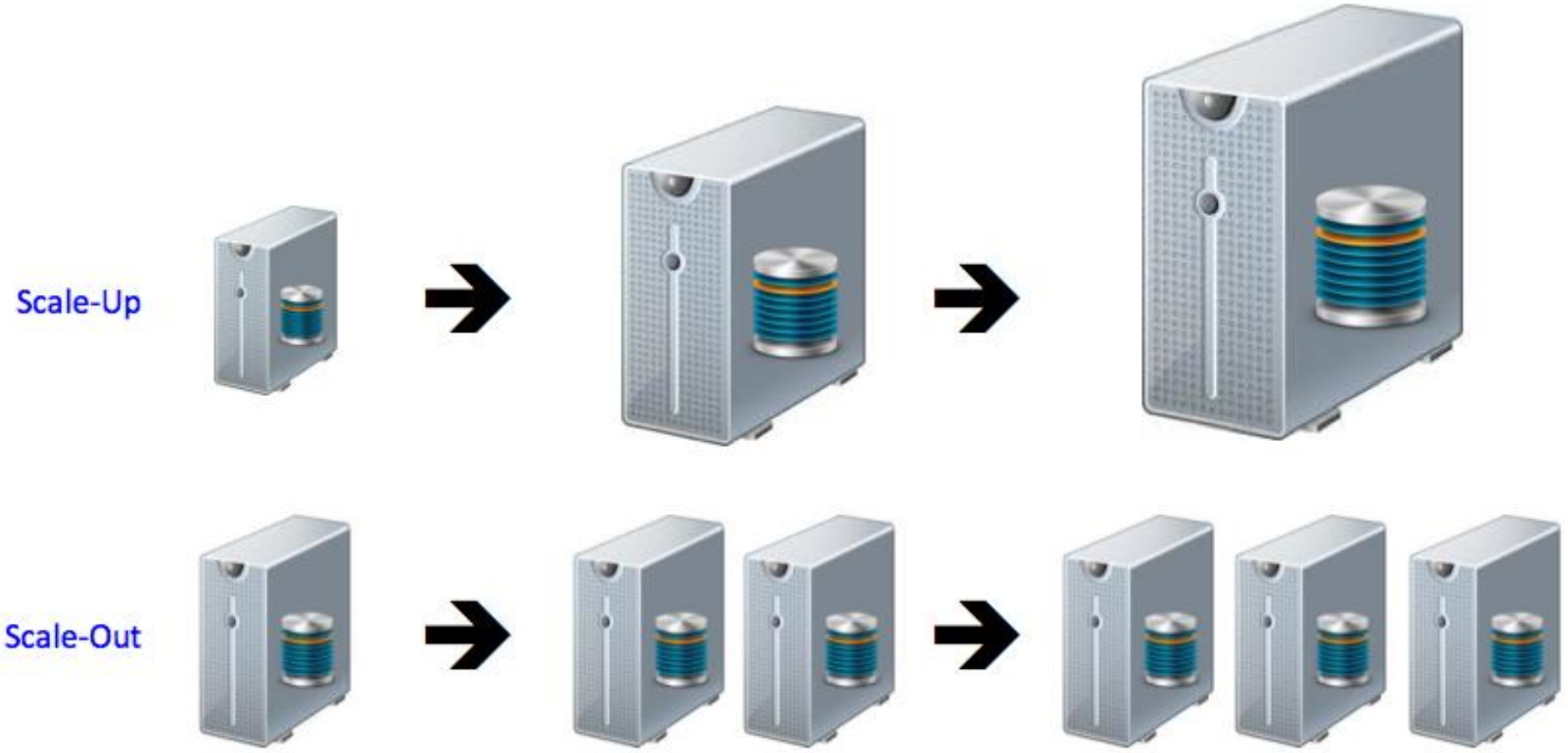
- Increase the configuration of a single system, like disk capacity, RAM, data transfer speed, etc.
- Complex, costly, and a time consuming process

2. Scale Out

- Use multiple commodity (economical) machines and distribute the load of storage/processing among them
- Economical and quick to implement as it focuses on distribution of load
- Instead of having a single system with 10 TB of storage and 80 GB of RAM, use 40 machines with 256 GB of storage and 2 GB of RAM

Scaling Up Vs. Scaling Out

ACADGILD



Quiz

ACADGILD

- What are the challenges of scaling up?

1. Complexity
2. Costly
3. Less Reliability
4. Less computational power

Ans: 1. Complex
2. Costly
3. Less Reliability

Quiz (Contd.)

ACADGILD

- What are the challenges of scaling out?

1. Low storage capacity
2. Coordination between networked machines
3. Handling failures of machines
4. Poor performance

Ans: 2. Coordination between networked machines
3. Handling failures of machines

Challenges of Scaling Out

ACADGILD

Need a new system:

- With new database management other than Relational Databases capable of handling unstructured as well as structured data
- To process huge datasets on large clusters of computers than on a single system

To manage clusters in which:

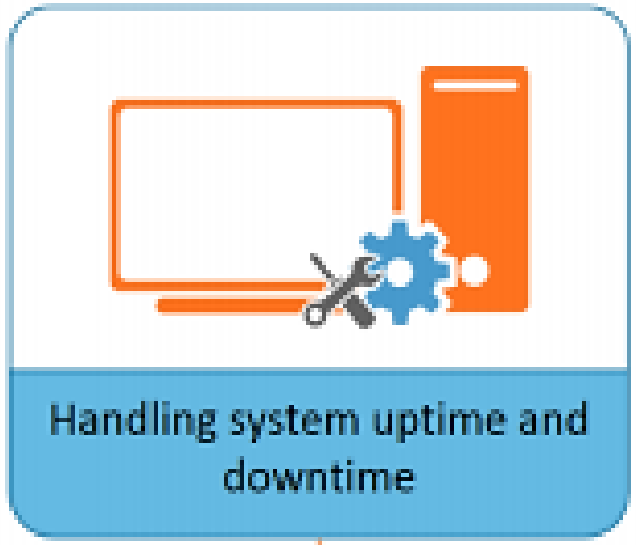
- Nodes fail frequently
- Number of nodes keep changing
- Take care of communication between the nodes
- During analysis, take results from different machines and merge/aggregate them.

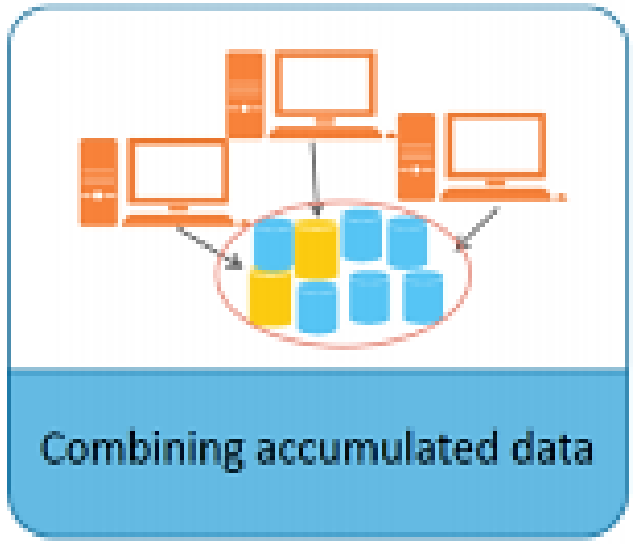
Common infrastructure which is:

- Efficient
- Easy to use
- Reliable

Challenges of Scaling Out (Contd.)

ACADGILD





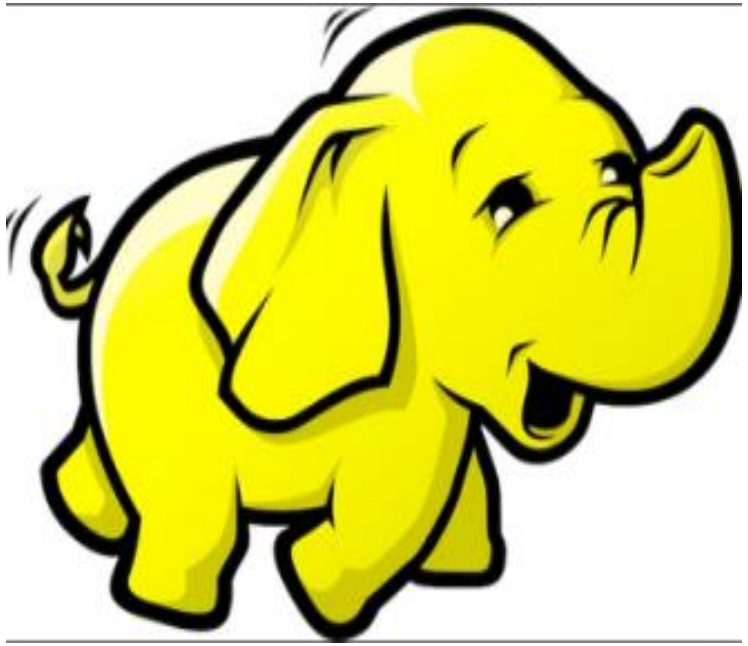
- Big Data technology has to use commodity hardware for data storage and analysis. Furthermore, it has to maintain a copy of the same data across clusters

- Big Data technology has to analyze data across different machines and then merge the data

Solution for Data Explosion - Hadoop

ACADGILD

- **Hadoop is that new framework!**



Hadoop: Introduction

ACADGILD

- Hadoop is an open source, Java-based programming framework that supports the processing of large data sets in a distributed computing environment
- Hadoop provides: A reliable, scalable platform for storage and analysis
- It is based on Google File System or GFS
- Hadoop runs a number of applications on distributed systems with thousands of nodes involving petabytes of data
- It has a distributed file system, called the Hadoop Distributed File System or HDFS, which enables fast data transfer among the nodes
- It leverages a distributed computation framework called MapReduce

Hadoop: Introduction (Contd.)

ACADGILD

Problems with distributed processing:

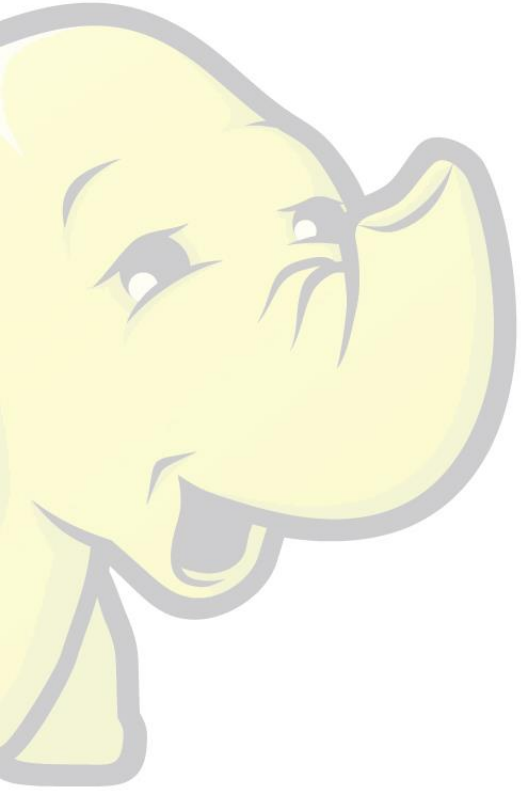
1. Hardware failure: can be solved by redundancy

2. Coordinating the tasks and combining results from all machines

- **Hadoop takes care of the above complexities and the challenges of network/distributed programming**
  - **HDFS (for storage)**
  - **Map Reduce (for processing)**

Two key concepts:

1. Storage (of data and results)
2. Processing (Analysis of data)



THANK YOU

Email us at: support@acadgild.com