BIG DATA INTRODUCTION

The easy approach

Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

By – Sudhanshu Saxena

Agenda

In this session, you will learn about:

- The Rise of Bytes
- Data Explosion and its Sources
- Types of Data – Structured, Semi-structured, Unstructured data
- Characteristics of Big Data
- Limitations of Traditional Large-Scale Systems
- Use Cases for Big-Data
- Challenges of Big-Data
- Hadoop Introduction - What is Hadoop? Why Hadoop?
- Supported Operating Systems
- Organizations using Hadoop
- Hadoop Job Trends
- History of Hadoop
- Hadoop Core Components – MapReduce& HDFS

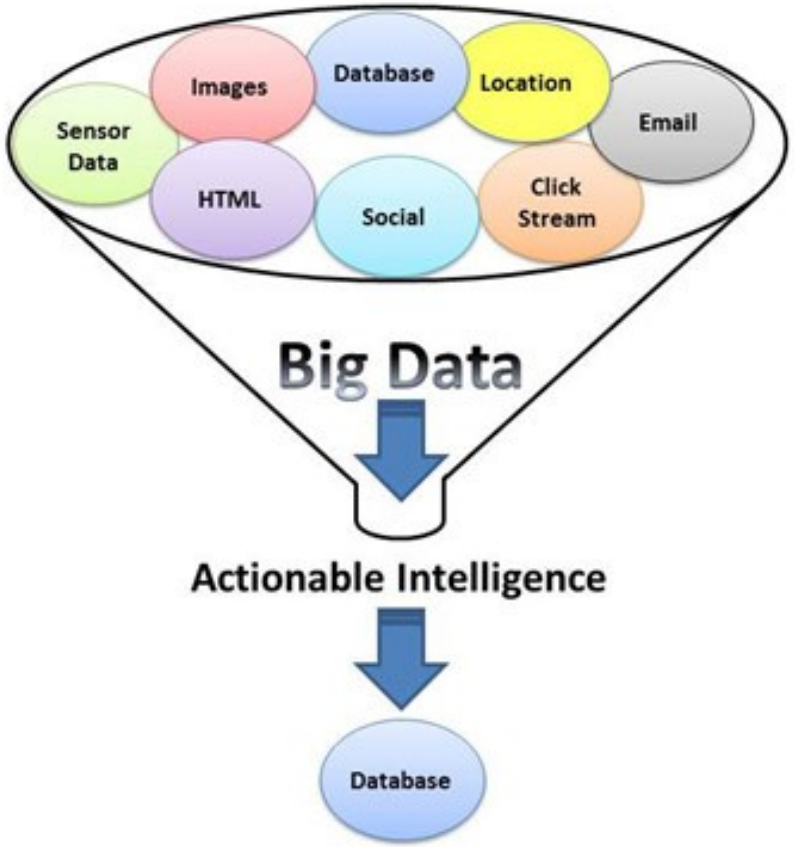
- **What is Big Data?**

Large and complex data, difficult to process them using traditional data processing applications.

Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

Definition

Big data is the term for a collection of data sets so large and complex that it becomes 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.



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable that technology and services that the world relies on every day. But what exactly is big data be used?

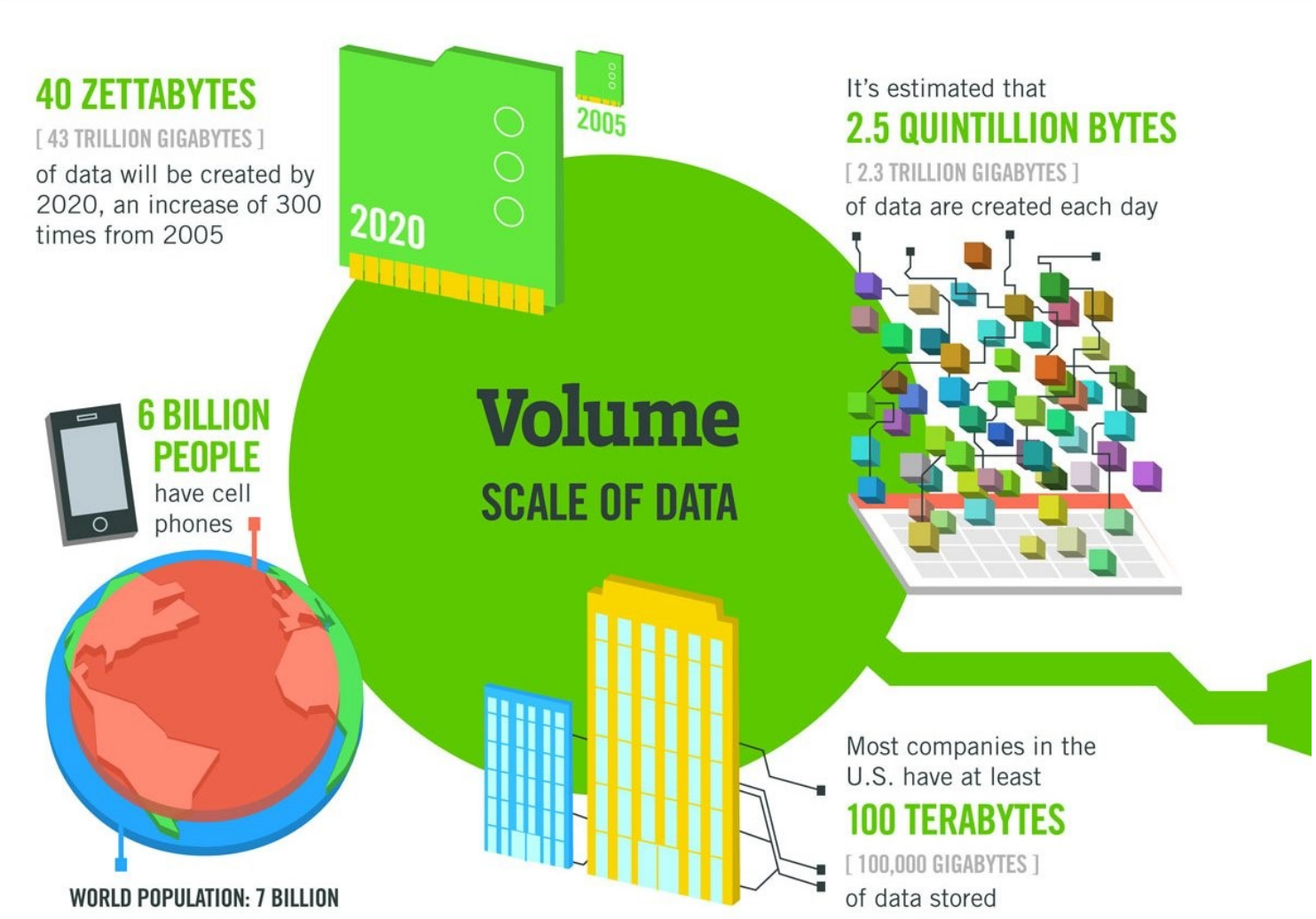
According to IBM scientists big data can be break into four dimensions:

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

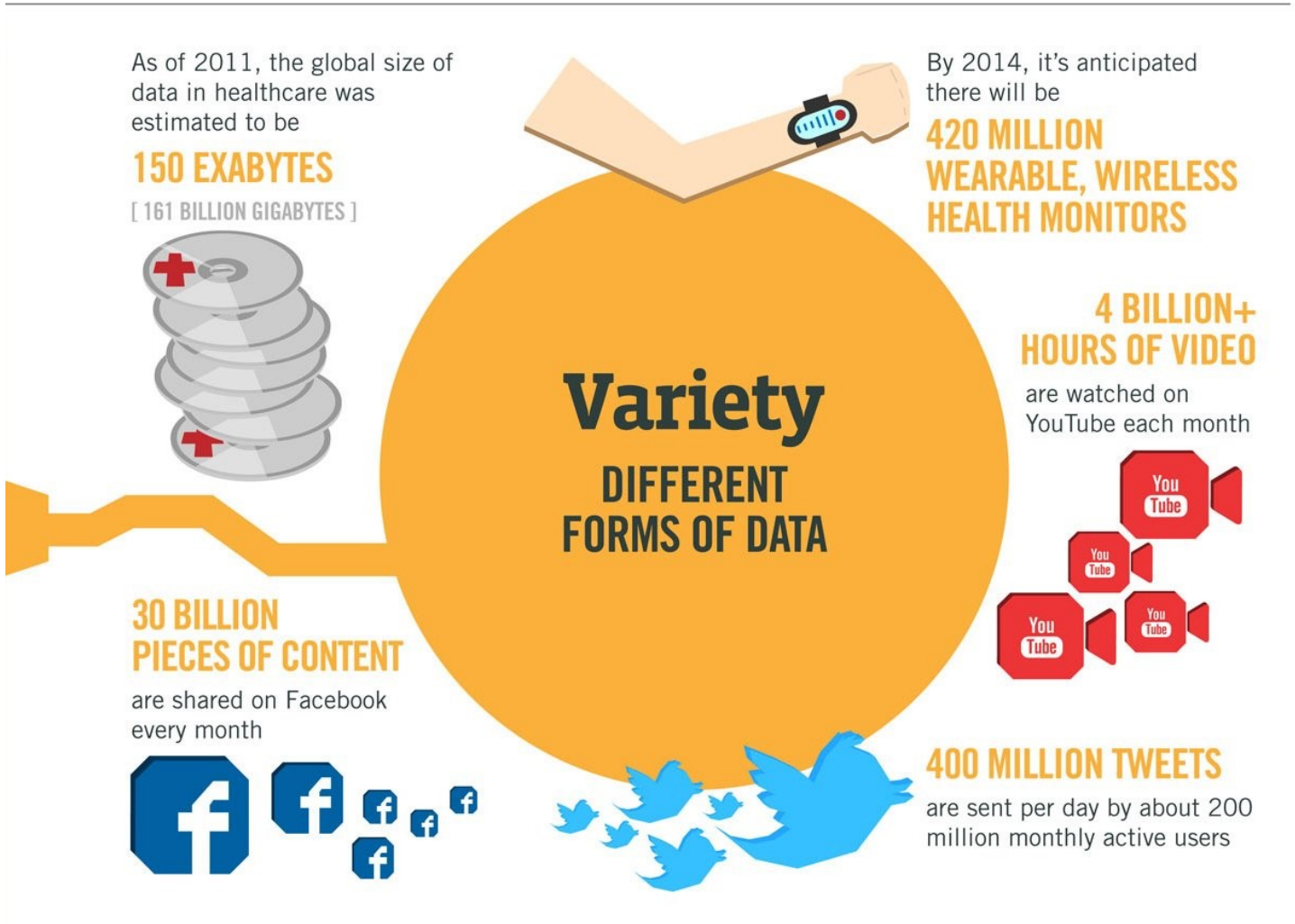
The FOUR V's of Big Data

Volume. Many factors contribute to the increase in data volume. Transaction-based **data stored through the years**. Unstructured data streaming in from social media. Increasing amounts of **sensor and machine-to-machine data** being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

The FOUR V's of Big Data



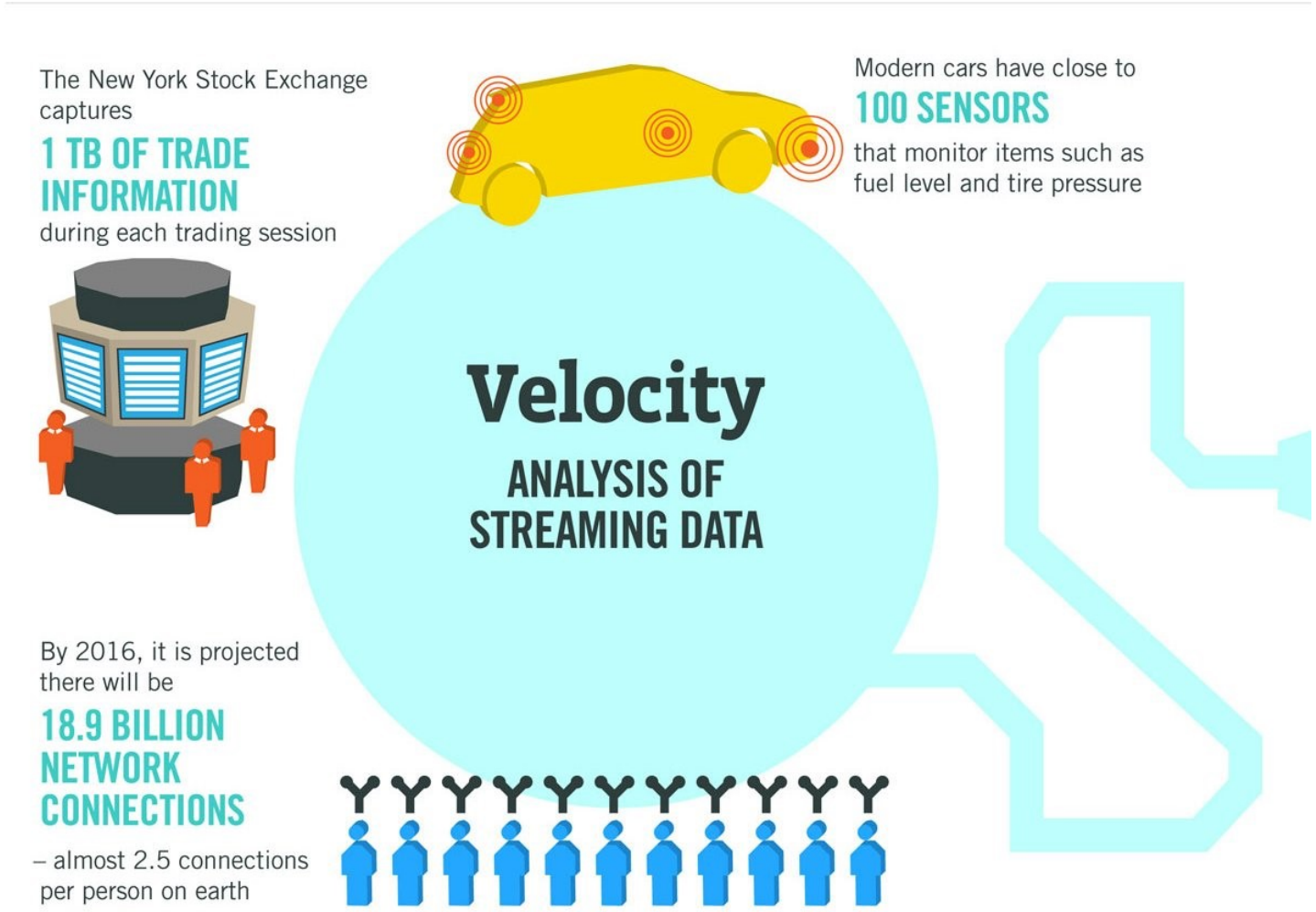
The FOUR V's of Big Data



The FOUR V's of Big Data

Variety. Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

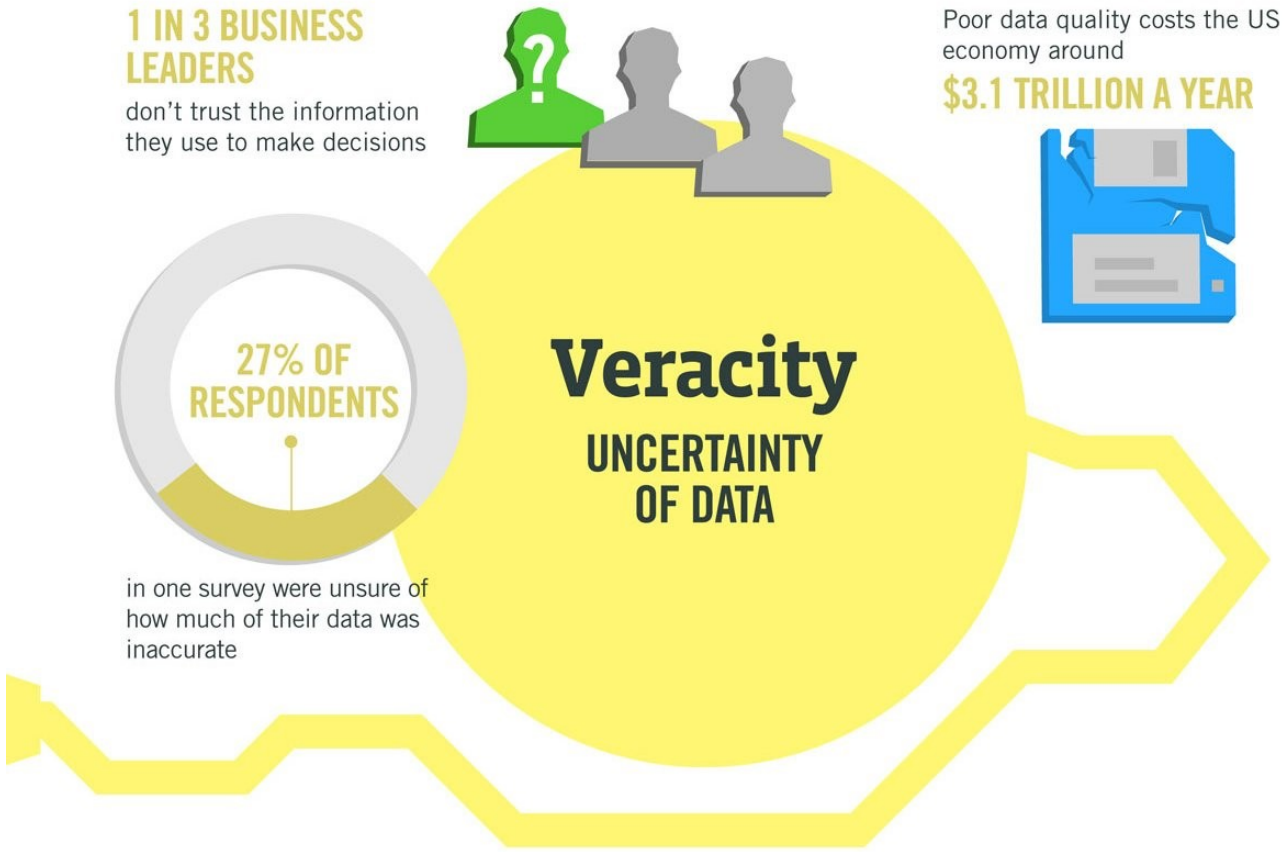
The FOUR V's of Big Data



The FOUR V's of Big Data

Velocity. Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. **Reacting quickly enough** to deal with data velocity is a challenge for most organizations.

The FOUR V's of Big Data



The FOUR V's of Big Data

Veracity - Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. **Veracity in data analysis is the biggest challenge** when compares to things like volume and velocity. In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems.

Who's Generating Big 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)

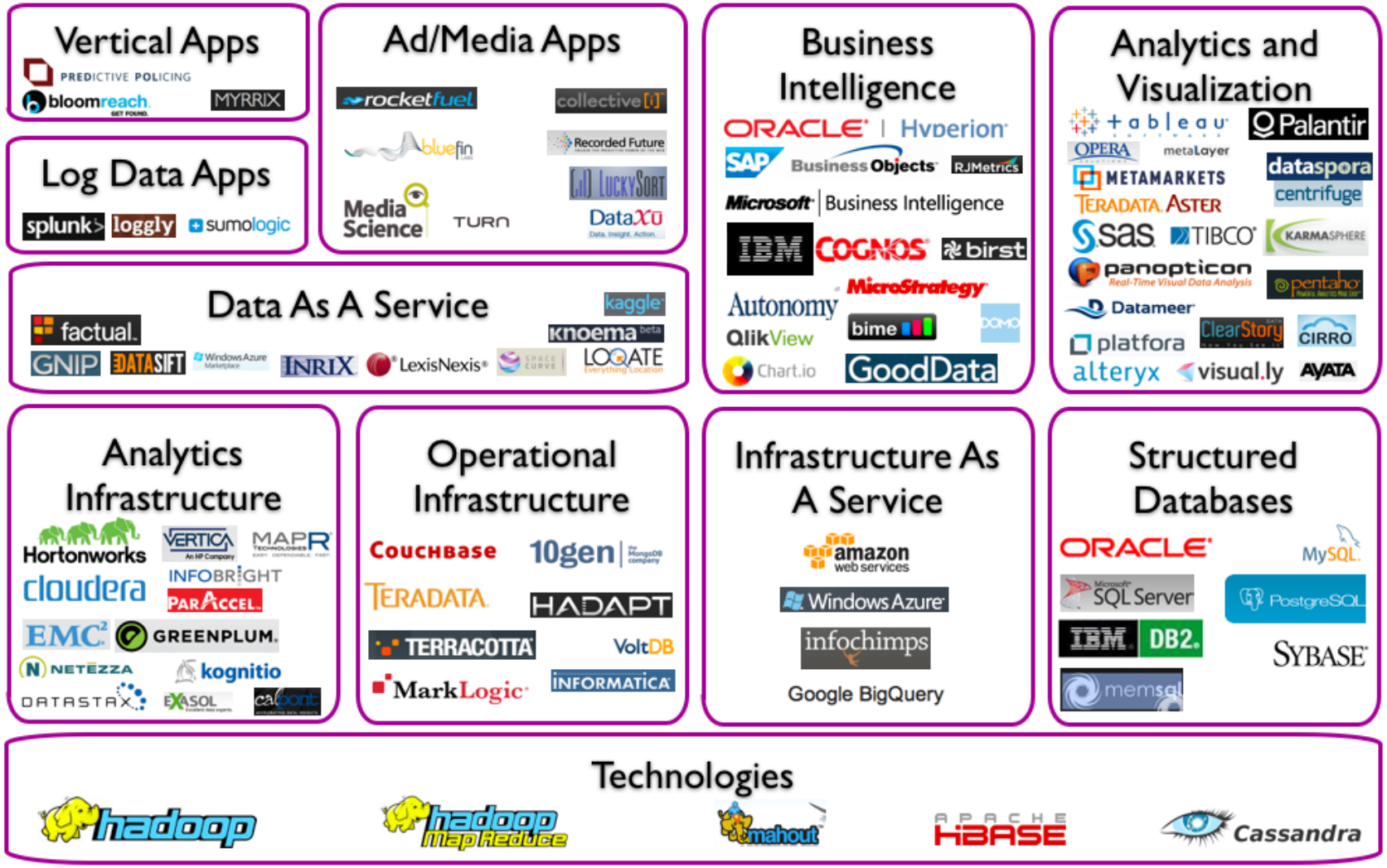
- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

The importance of Big Data

The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable:

- Cost reductions
- Time reductions
- New product development and optimized offerings
- Smarter business decision making

Big Data Landscape



dave@vc.dave.com

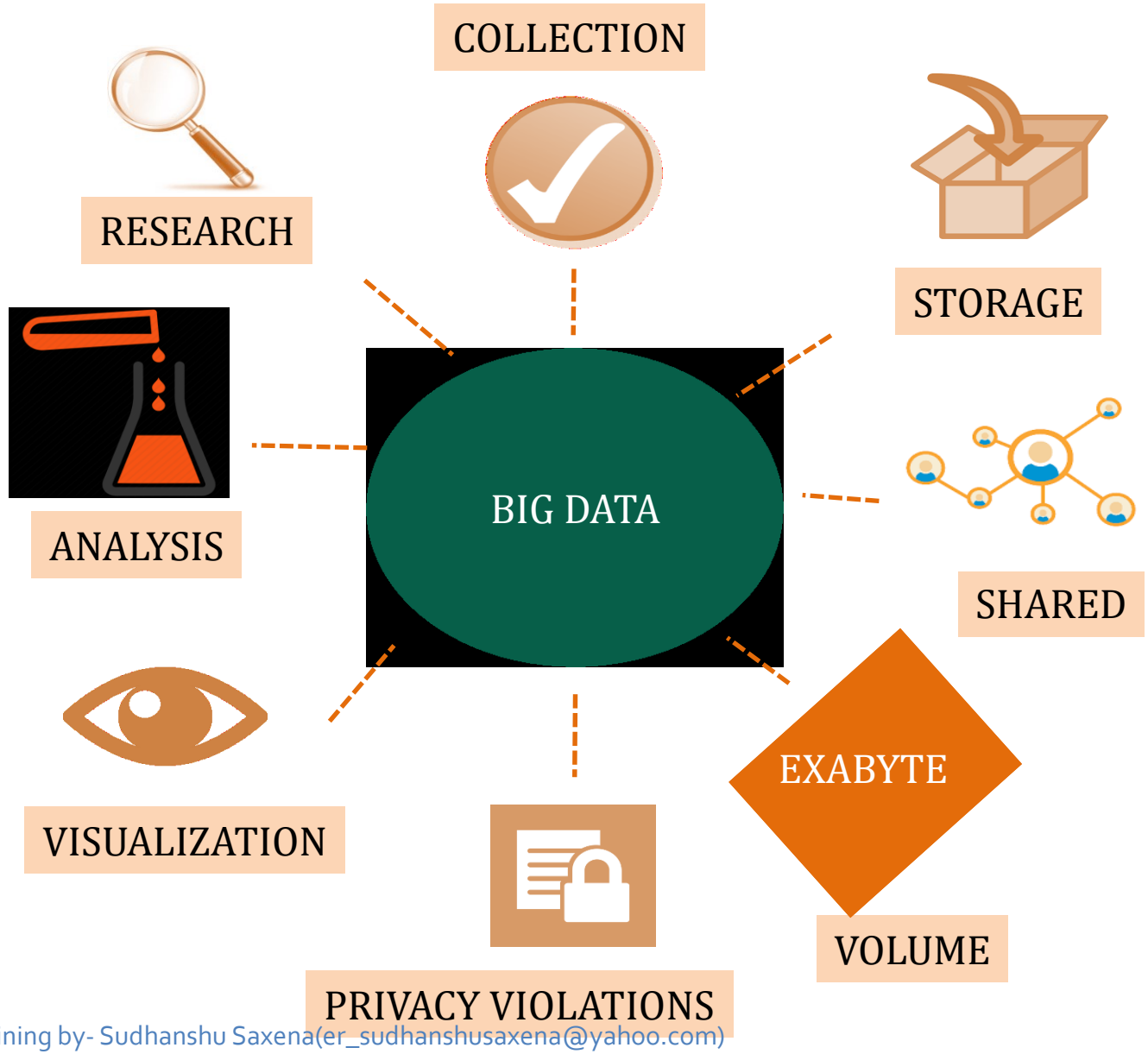
Training by- Sudhanshu Saxena (sudhanshusaxena@yahoo.com) blogs.forbes.com/davefeinleib

The importance of Big Data

For instance, by combining big data and high-powered analytics, it is possible to:

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Analyze millions of SKUs to determine prices that maximize profit and clear inventory.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use clickstream analysis and data mining to detect fraudulent behavior

Challenges of Big Data



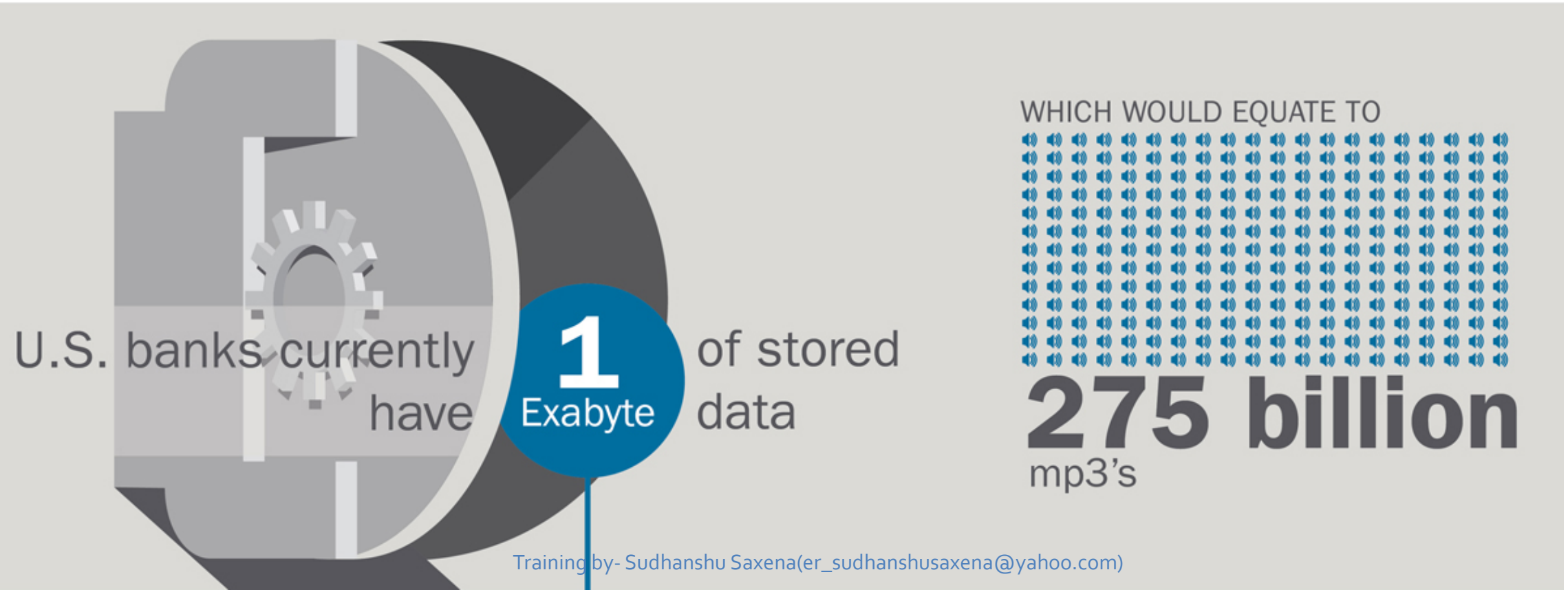
HOW BIG IS BIG DATA?

Not the usual way

Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

BIG DATA in banking

THE BANKING BUSINESS FINDS A NEW ASSET



Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

In 2012

50%

Of coustomers changed banks or were planning to change banks

SOLUTION

A customer churn prediction model based on big data analysis of social media customer sentiment and purchasing power helps identify customers at risk of leaving

Understanding Customer Needs

Staffing costs account for

66%

of a branch bank's costs

SOLUTION

Staffing models based on transaction times and account holder traffic patterns data make annual resource planning easier and more effective

Managing Rising Security Costs

Insights for Product Development

The costs of developing new products and services can be staggering

SOLUTION

Customer transactional data such as timing of visits and duration of teller transactions can be analyzed to find gaps in product offerings

Scoring Credit Risks

Banks need to cut lending risks while improving customer marketing

SOLUTION

Consumer payment patterns and law enforcement databases supply the data

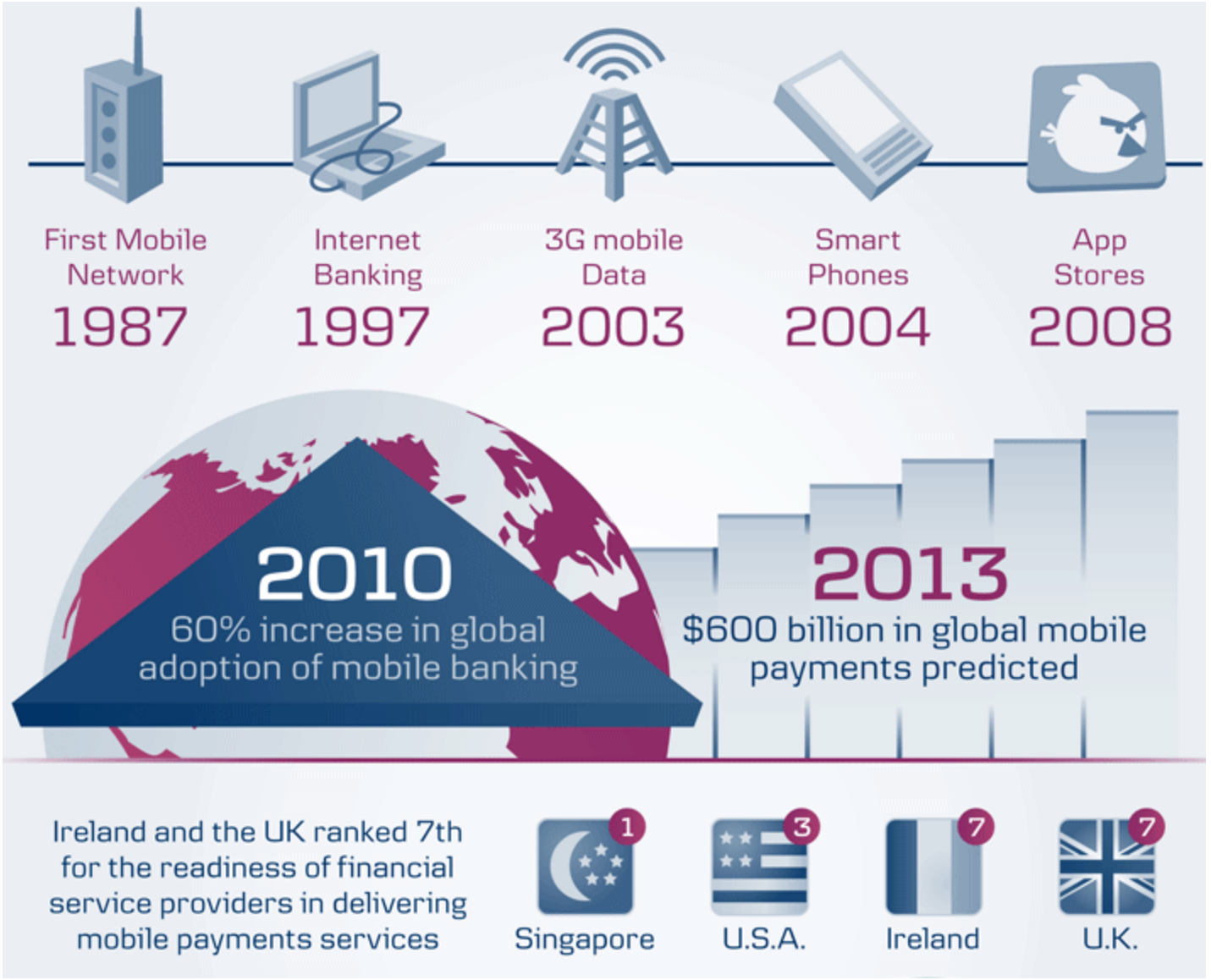
credit bureaus store over

800 billion records

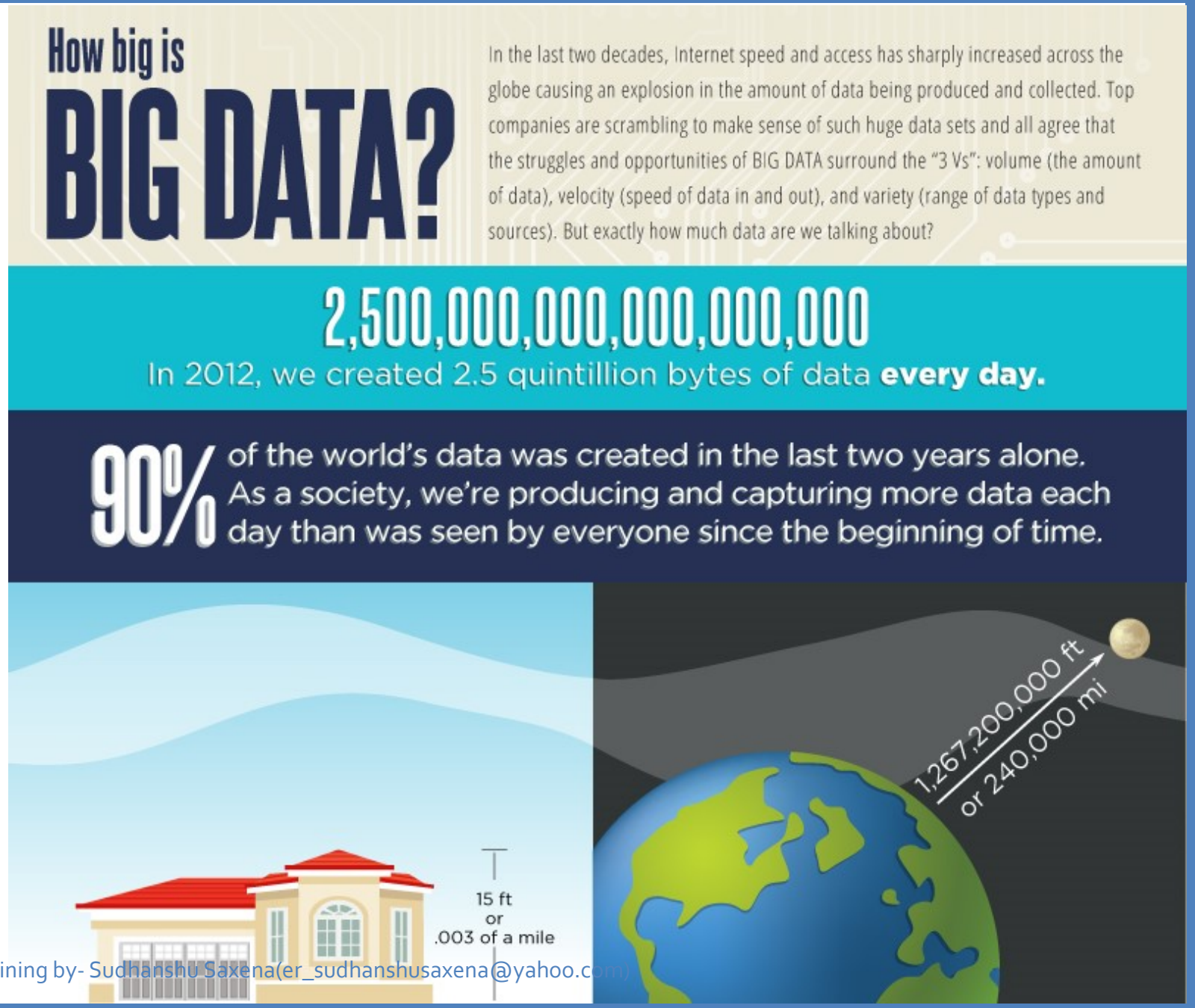
to be sliced, diced and analyzed for more accurate credit risk scores.

For comparison, the FBI's Investigative Data Warehouse has only 1.5 billion documents.

How big the Big data is in Telecom



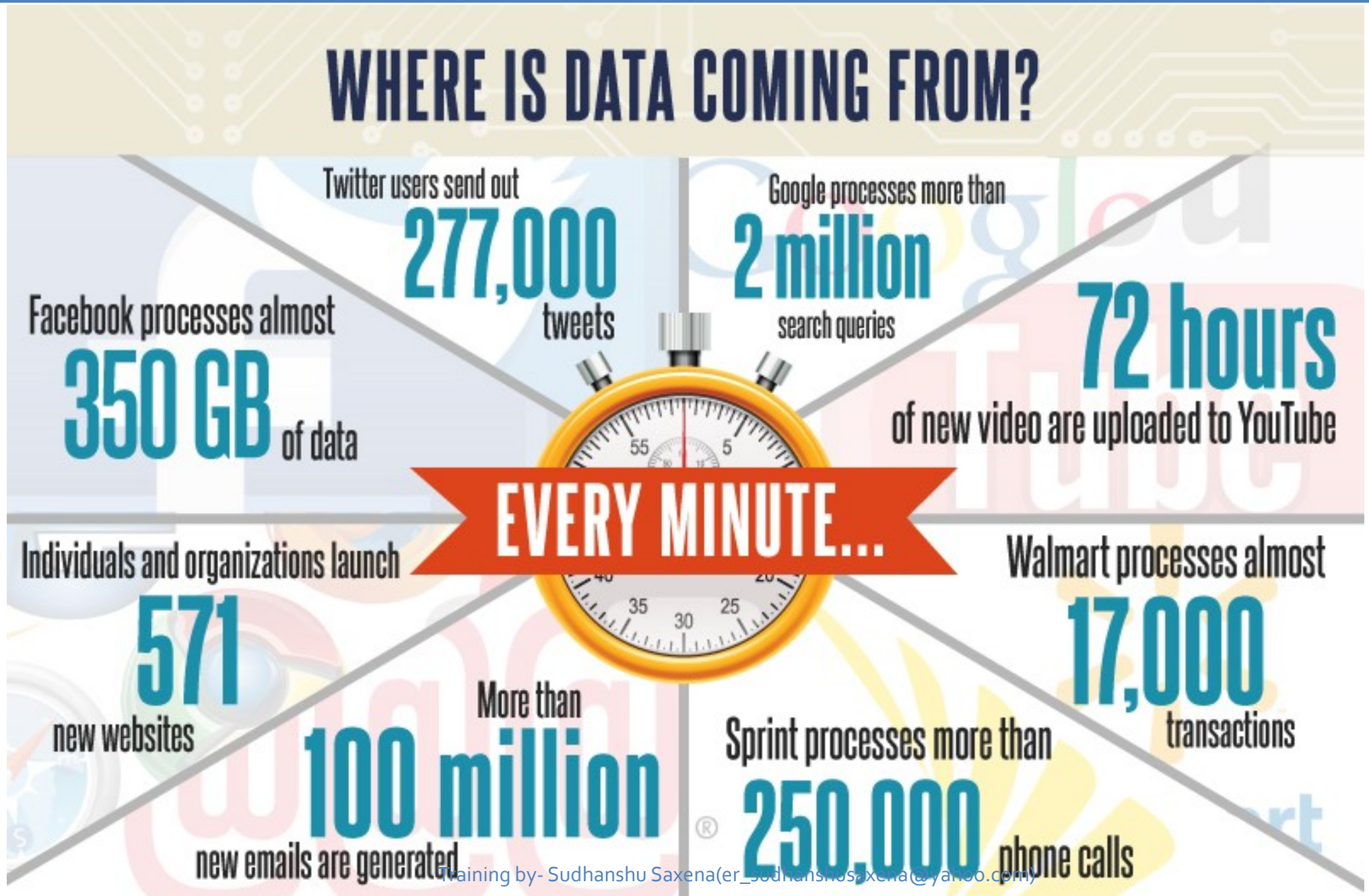
How big the Big data is actually

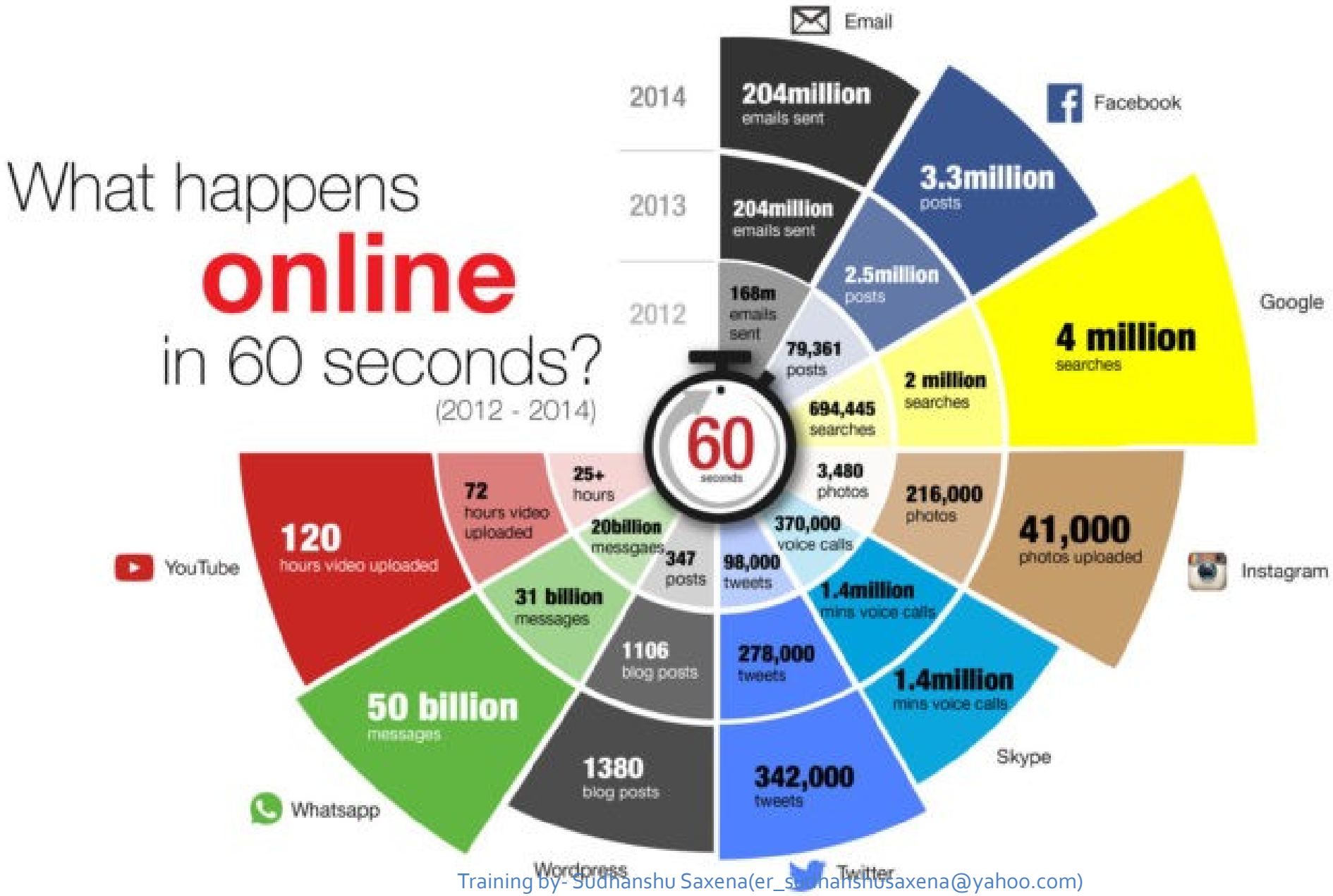


WHERE IS BIG DATA COMING FROM ?

Not the usual way

General trends for Big Data


WHERE IS DATA COMING FROM?
Twitter users send out
277,000
tweets
Google processes more than
2 million
search queries
Facebook processes almost
350 GB
of data
72 hours
of new video are uploaded to YouTube
EVERY MINUTE...
Individuals and organizations launch
571
new websites
Walmart processes almost
17,000
transactions
More than
100 million
new emails are generated
Sprint processes more than
250,000
phone calls


What happens
online
in 60 seconds?
(2012 - 2014)
2014
2013
2012
Email
204million
emails sent
204million
emails sent
168m
emails
sent
Facebook
3.3million
posts
2.5million
posts
79,361
posts
Google
4 million
searches
2 million
searches
694,445
searches
60
seconds
YouTube
120
hours video uploaded
72
hours video
uploaded
25+
hours
3,480
photos
216,000
photos
41,000
photos uploaded
Instagram
20billion
messages
370,000
voice calls
347
posts
98,000
tweets
1.4million
mins voice calls
31 billion
messages
1106
blog posts
278,000
tweets
1.4million
mins voice calls
Skype
50 billion
messages
Whatsapp
1380
blog posts
342,000
tweets
Wordpress
Twitter

The Model Has Changed…

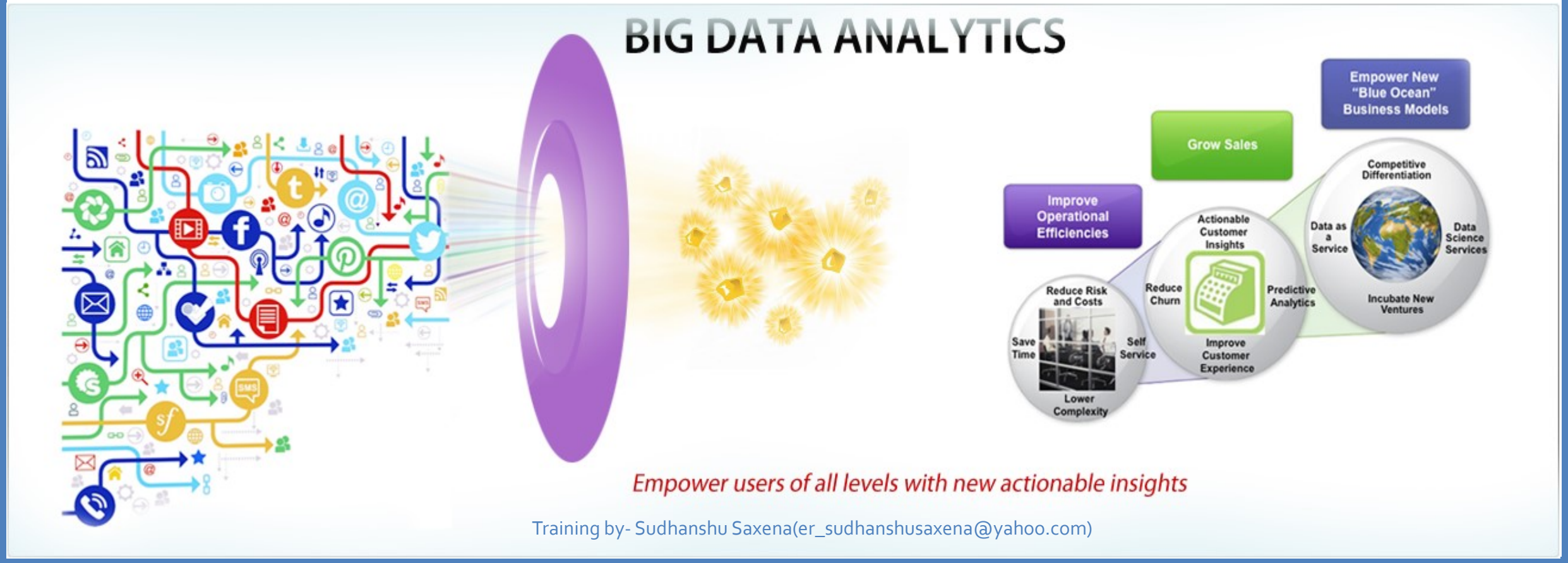
- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

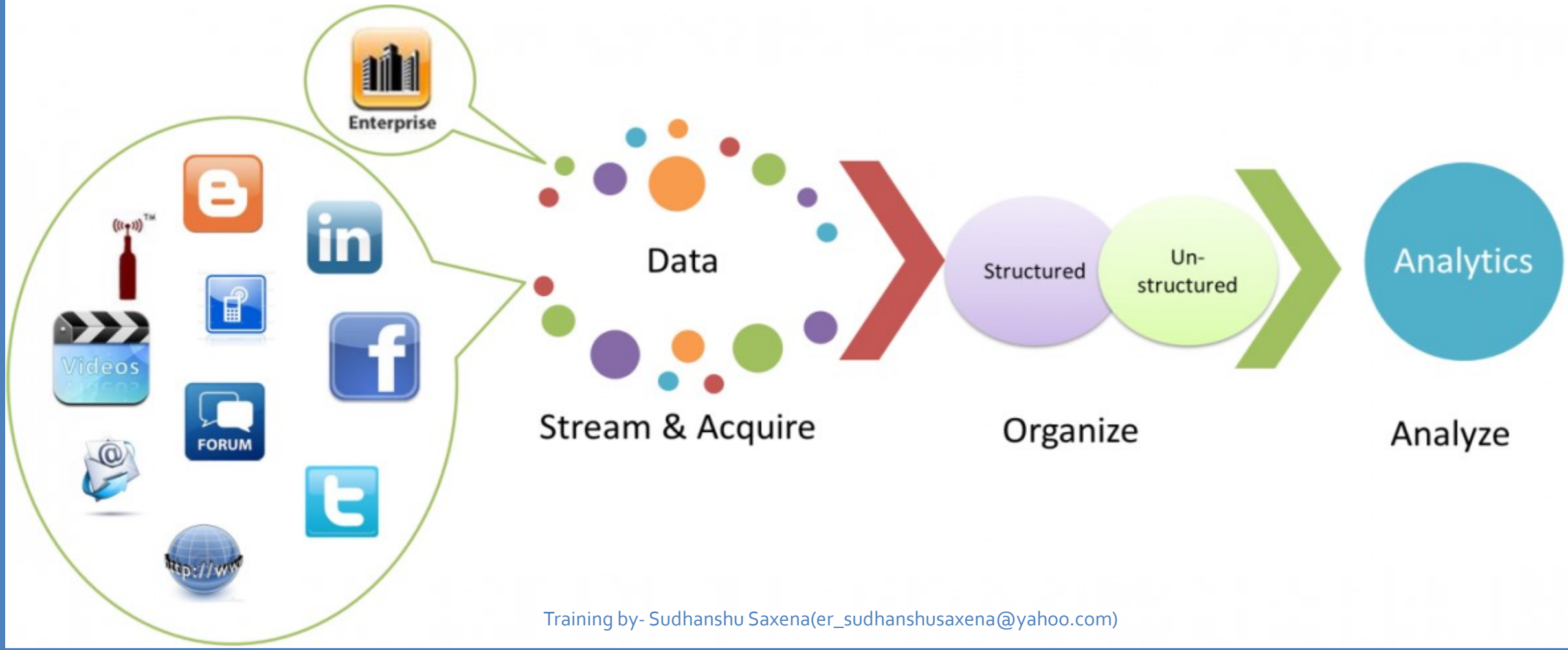
New Model: all of us are generating data, and all of us are consuming data

What is Big Data Analytics?

Big data analytics is the process of examining large and varied **data** sets -- i.e., **big data** -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.

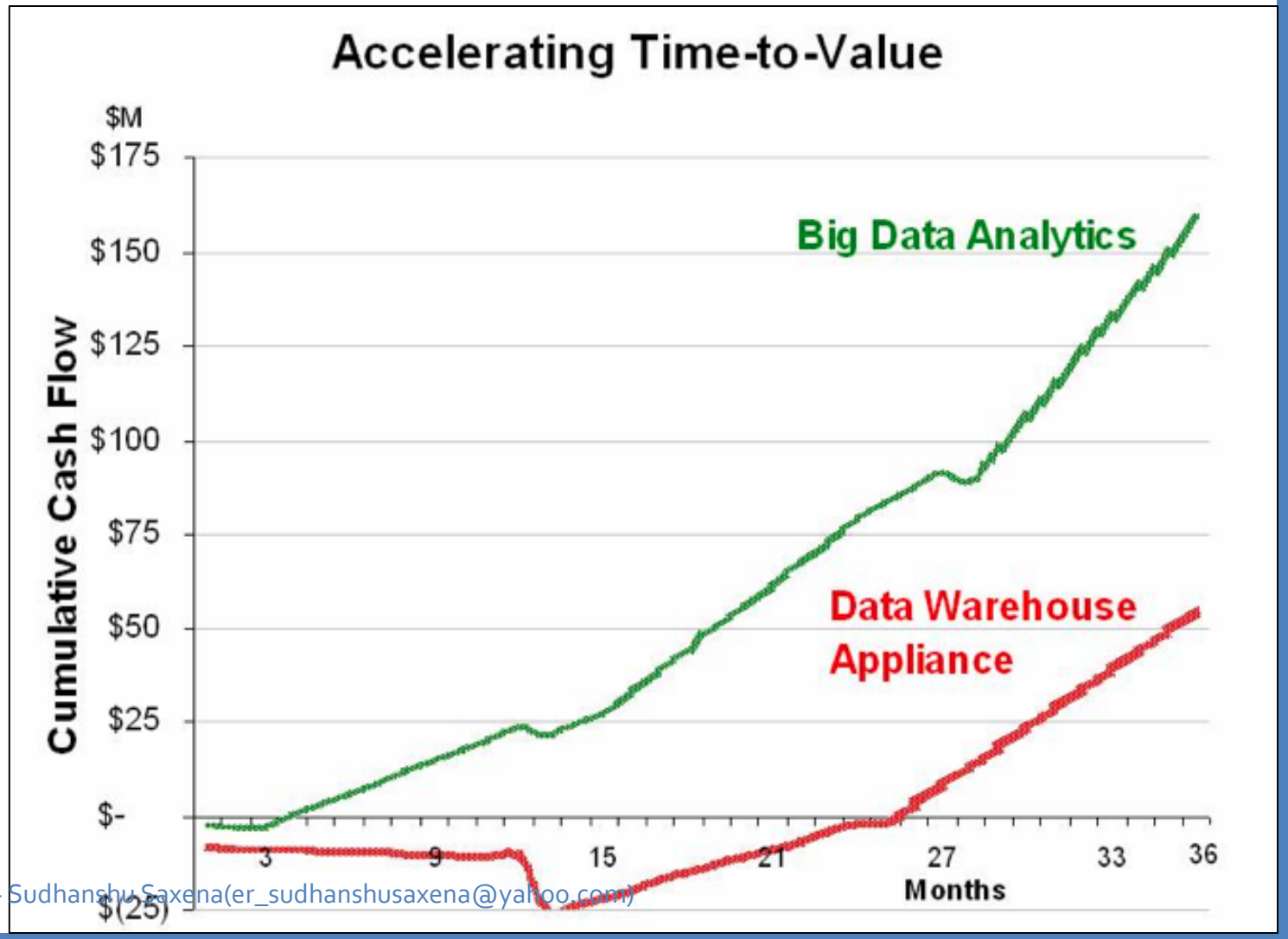


Lets Look little closely..

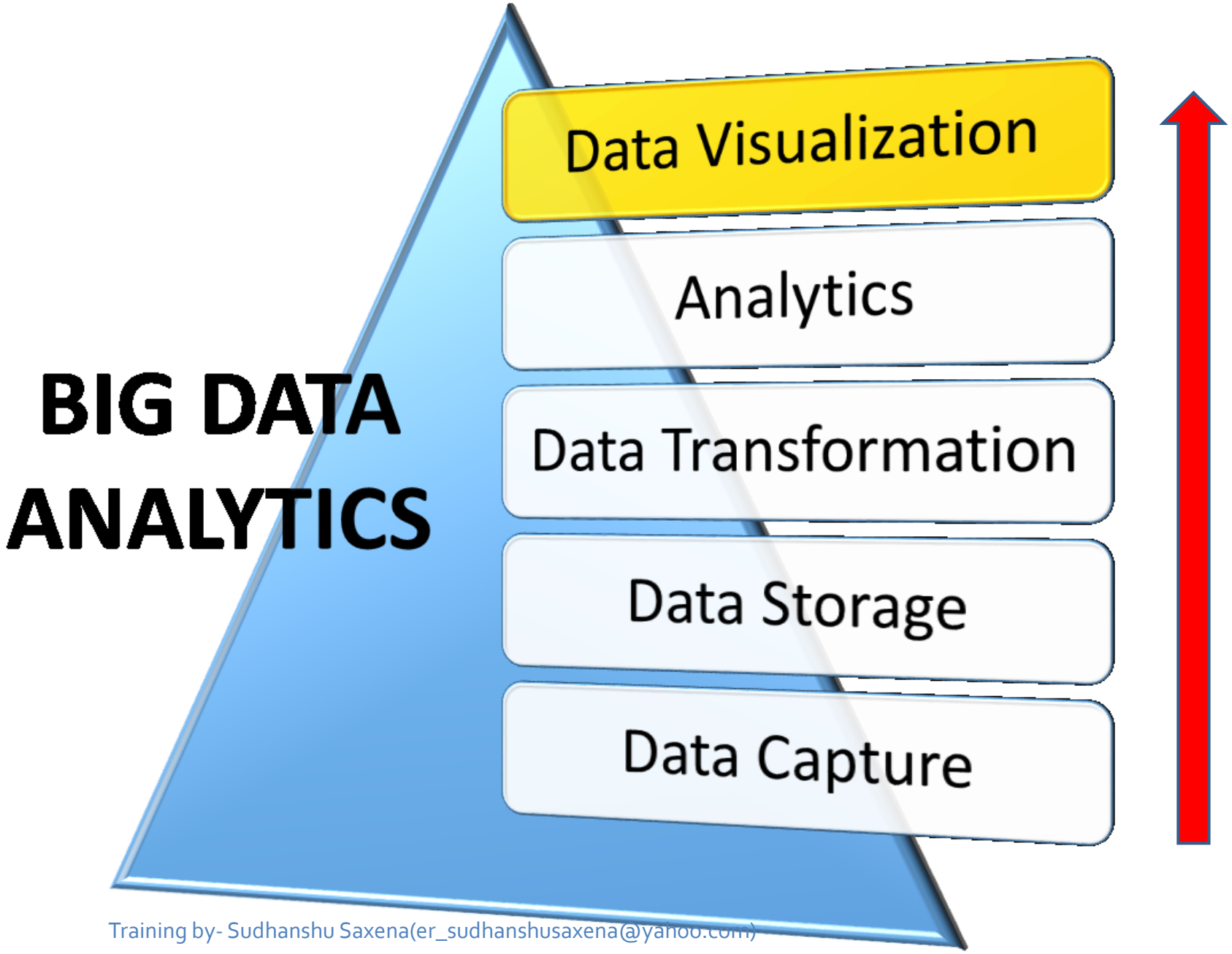


Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Getting for the Analyti



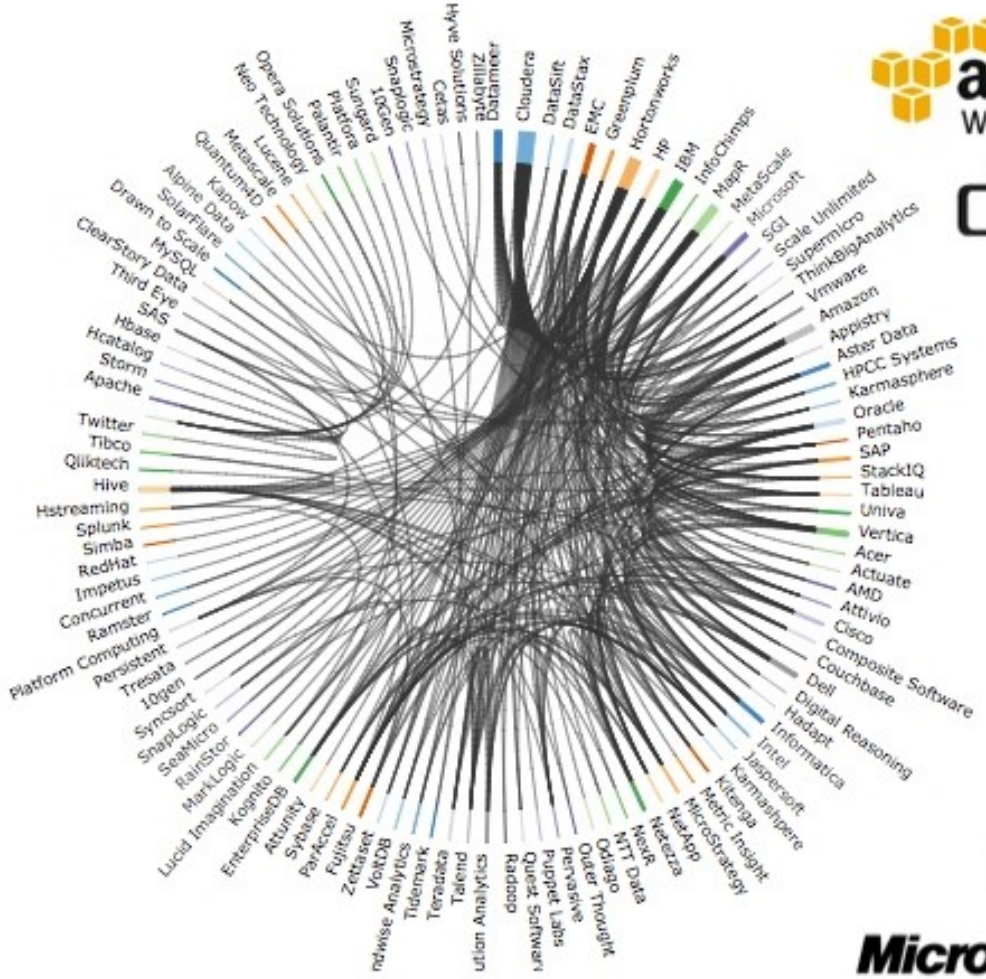
Big Data Analytics

Tools & Trends

Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)


Data Analysis & Platforms
HPCC Systems
Storm
Distributed and fault-tolerant realtime computation
Hadoop
Dremel
hadoop MapReduce
Spark
HD
APACHE DRILL
SAMOA
IKANOW | BRILLIANT DECISIONS
Hortonworks
Databases / Data warehousing
bigdata
INFOBRIGHT
Cassandra
4store
H2
GlobalsDB
InfiniDB
riak
Infinispan
HYPERTABLE
SQLite
RethinkDB
MariaDB
Drizzle
terrastore
HIVE
Firebird
ORACLE
BERKELEY DB
HyperSQL
monetdb
In-Memory Computing
GridGain
IN-MEMORY COMPUTING
hazelcast
TERRACOTTA
GORA
NMemory
ERP BI Solutions
Open Source Business Intelligence Solutions for ERP
Business Intelligence
openI.org
Open Intelligence
talend
open data solutions
Jaspersoft
Palo
Open Source Business Intelligence
spagobi
pentaho
jedox.
BIRT
Data Mining
orange
rapidminer
KNIME
Data Analytics Made Easy
mahout
WEKA
The University of Waikato
KEEL
togaware
SPMF
Big Data search
Lucene
Apache Solr
elasticsearch.
Multivalue database
Rocket
U2
REVELATION SOFTWARE
Open QM
Ladybridge Systems
northgate Reality
jBASE INTERNATIONAL
FREEDOM. OPENNESS. POWER.
ScarletDME
Programming
R
julia
Data aggregation
cloudera
FLUME
chukwa
KeyValue
AEROSPIKE
leveldb
redis
Chordless
Beta
Tokyo Cabinet
8192PIB
MEMCACHED
SCALIEN
Project Voldemort
A distributed database.
hamsterdb
RAPTORDB
FairCom
STSDB
DATABASE & VIRTUAL FILE SYSTEM
HyperDex
OpenLDAP
IQLECT
ioremap.net
STORAGE AND BEYOND
Scalaris
Document Store
mongoDB
CouchBase
Raven DB
CLUSTERPOINT
Tokutek
RaptorDB
EJDB
djondb
JasDB
SchemafreeDB
sisodb
Apache CouchDB
relax
Graph databases
Gephi
makes graphs handy
Gremlin
GraphBuilder
FRANZ INC.
*Sparksee
INFOGRID
InfiniteGraph
Powered by Objectivity
HYPERGRAPHDB
Neo4j
the graph database
FlockDB
GraphBase
BrightstarDB
Operational
VOLTDB
SMART DATA FAST.
Social
Apache Kafka
A high-throughput distributed messaging system.
ThinkUp
Corona
Multidimensional
FIS
SciDB
rasdaman
raster data manager
db4objects
Object databases
ZOPE
mobject
precision data management
Magma
Picolisp
NEOPPOD
Distributed Transactional NoSQL for the Cloud
siaqodb
RAMER D
PERSEVERE
EyeDB
STARCOUNTER
Sterling
NDatabase
Multimodel
ArangoDB
alchemydatabase
A Hybrid Relational-Database/NOSQL-Datastore
XML Databses
existdb
BASEX
Qizx
sedna
LIQUIBASE
Grid Solutions
GIGASPACES
Galaxy
Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

Big Data Tools Market share



amazon web services™

cloudera

DataStax

EMC² where information lives

Hortonworks

IBM

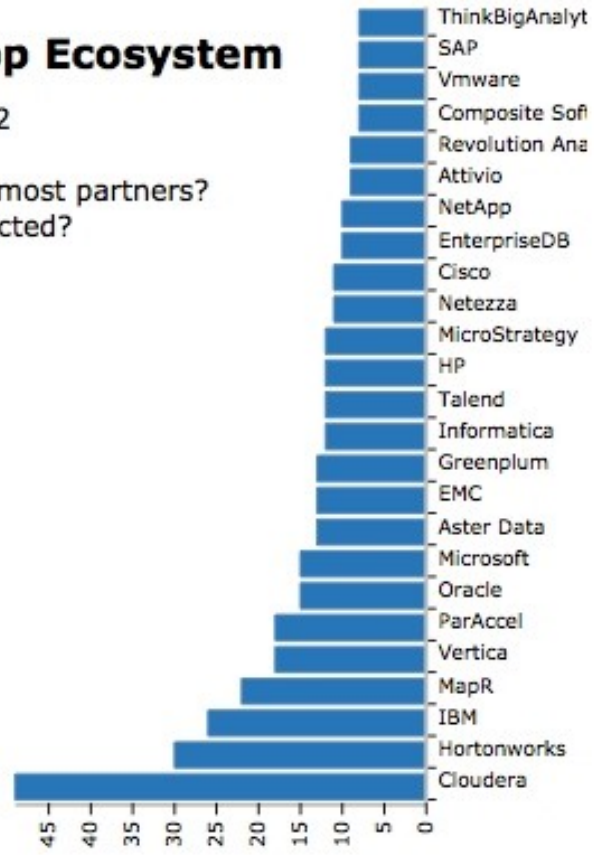
MAPR TECHNOLOGIES

Microsoft

The Hadoop Ecosystem

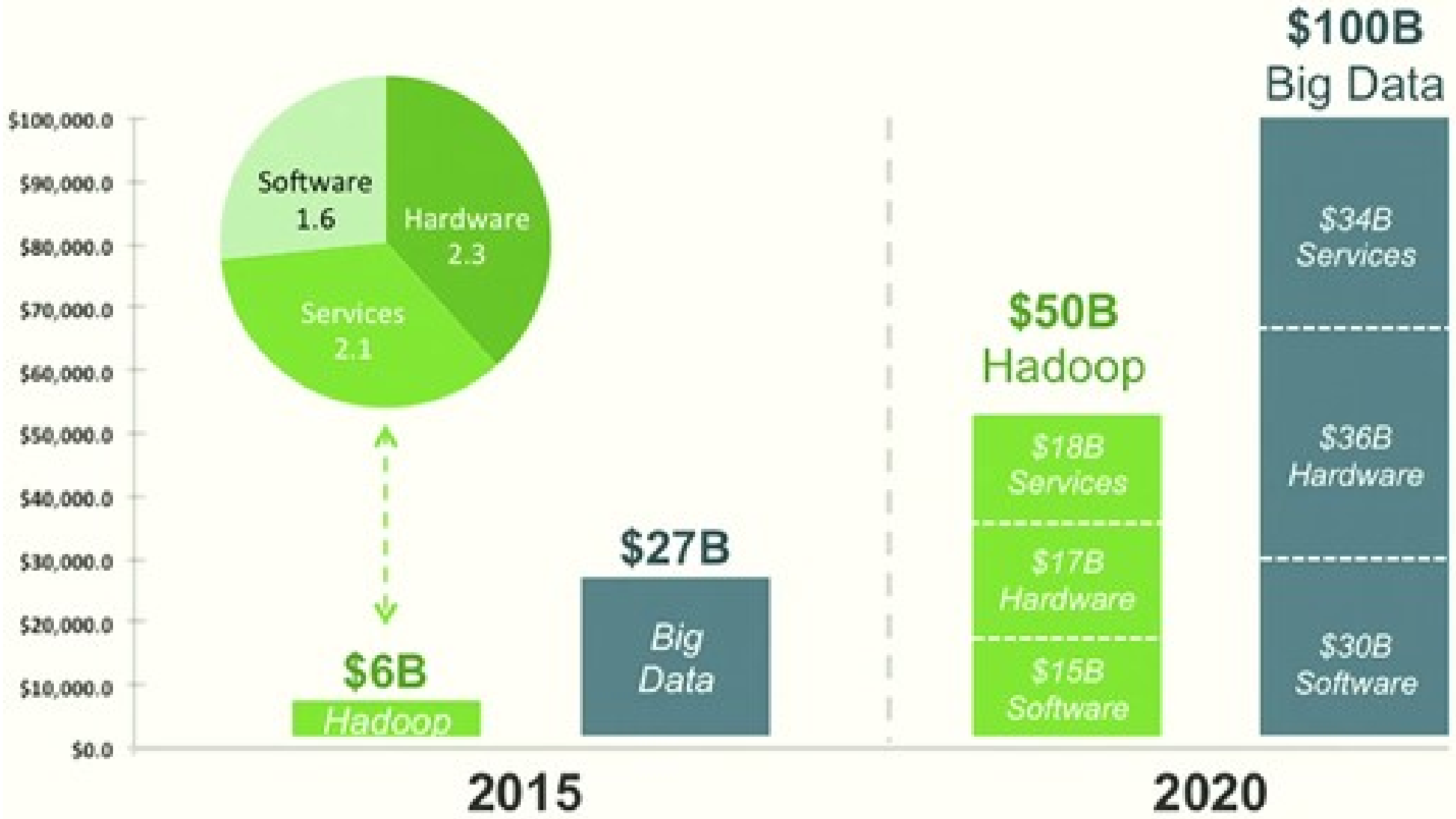
June 21, 2012

Who has the most partners?
Who is connected?



brought to you by Datameer Powerfully Simple™

Big Data and Hadoop Markets Growing Sharply

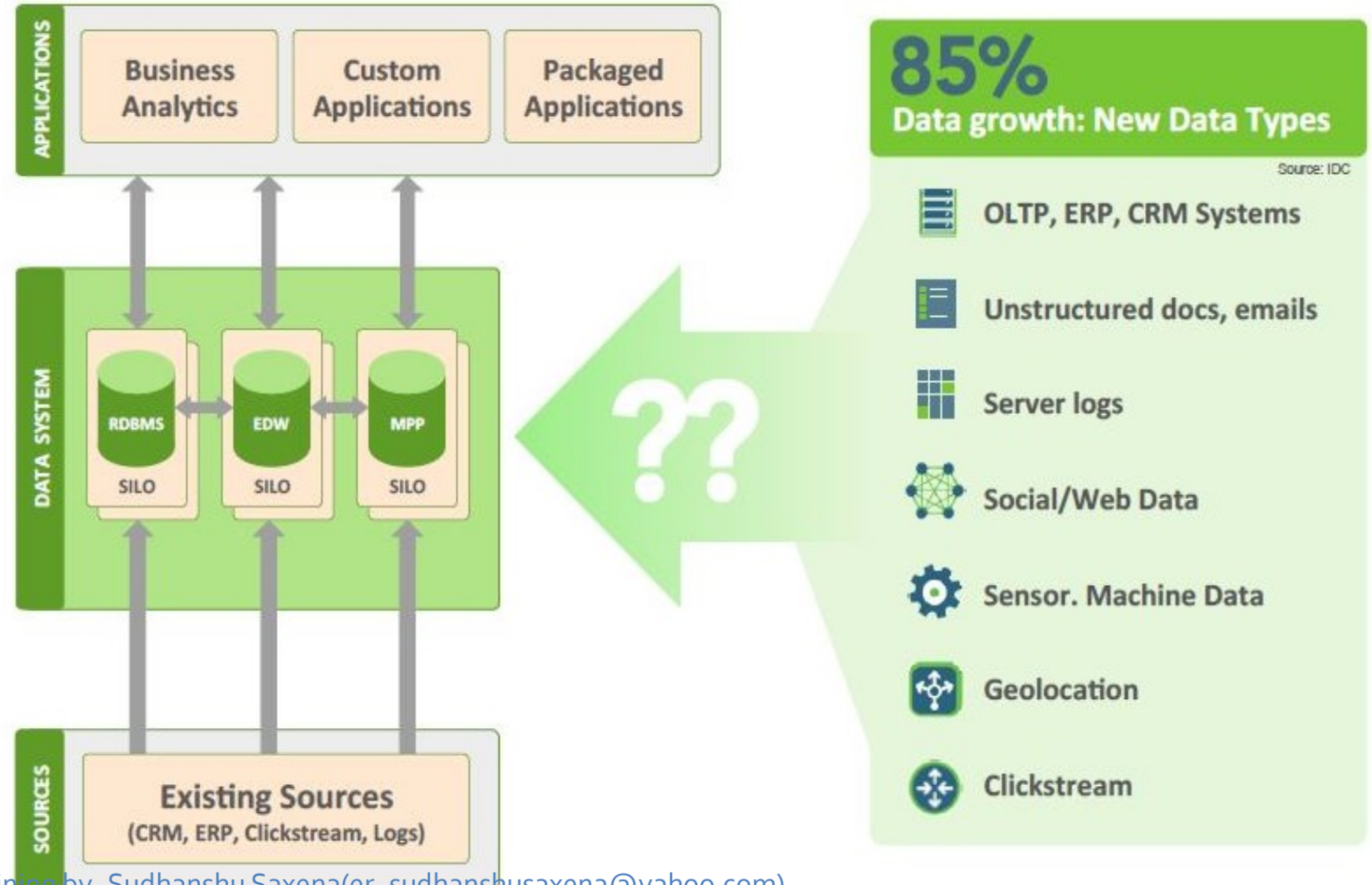


Hortonworks

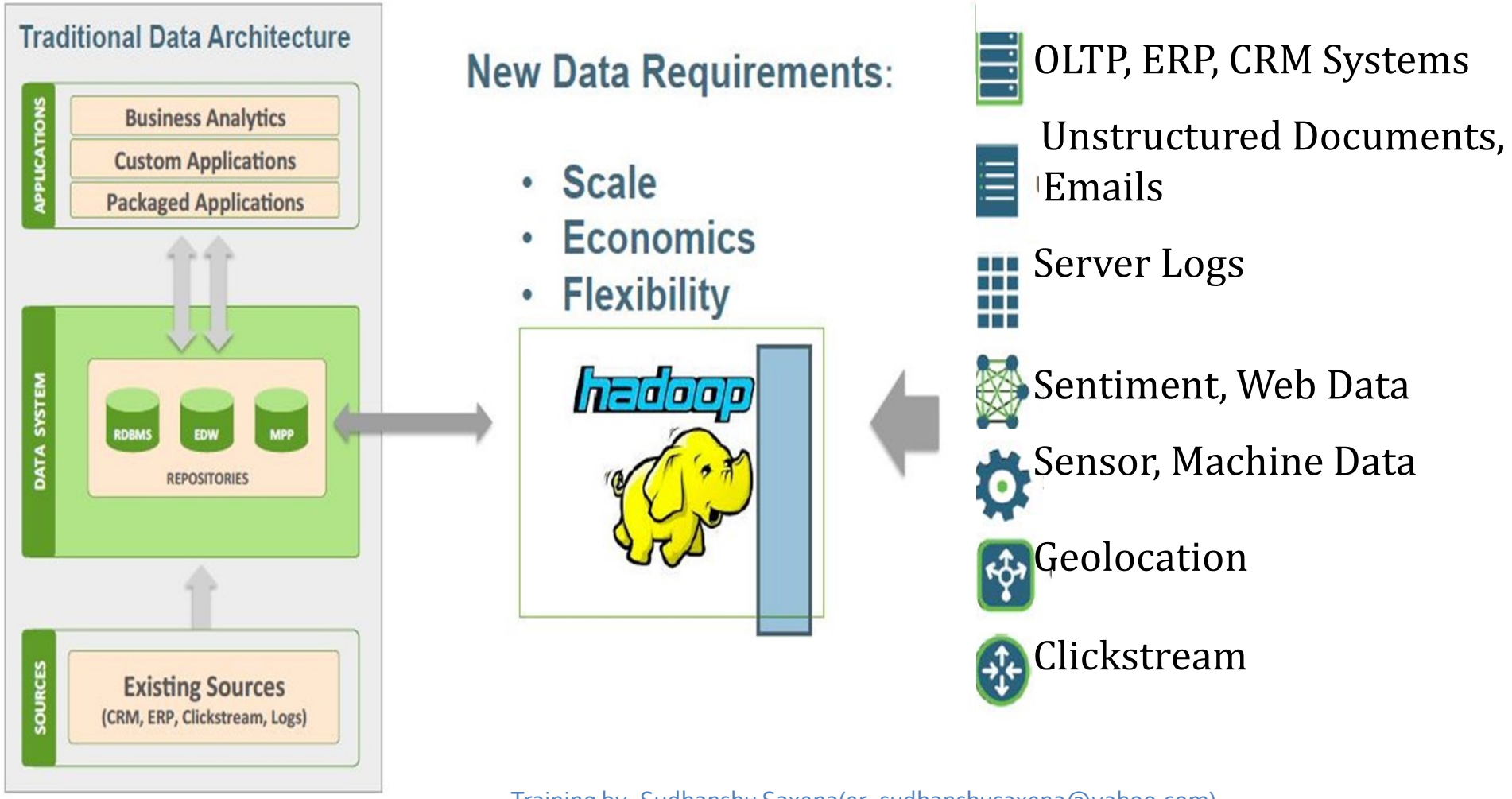
Traditional Data Architecture Under Pressure

Disadvantages of traditional data architecture under pressure:

- Can’t manage new data paradigm
- Constrains data to specific schema
- Siloed data
- Limited scalability
- Economically unfeasible
- Limited analytics

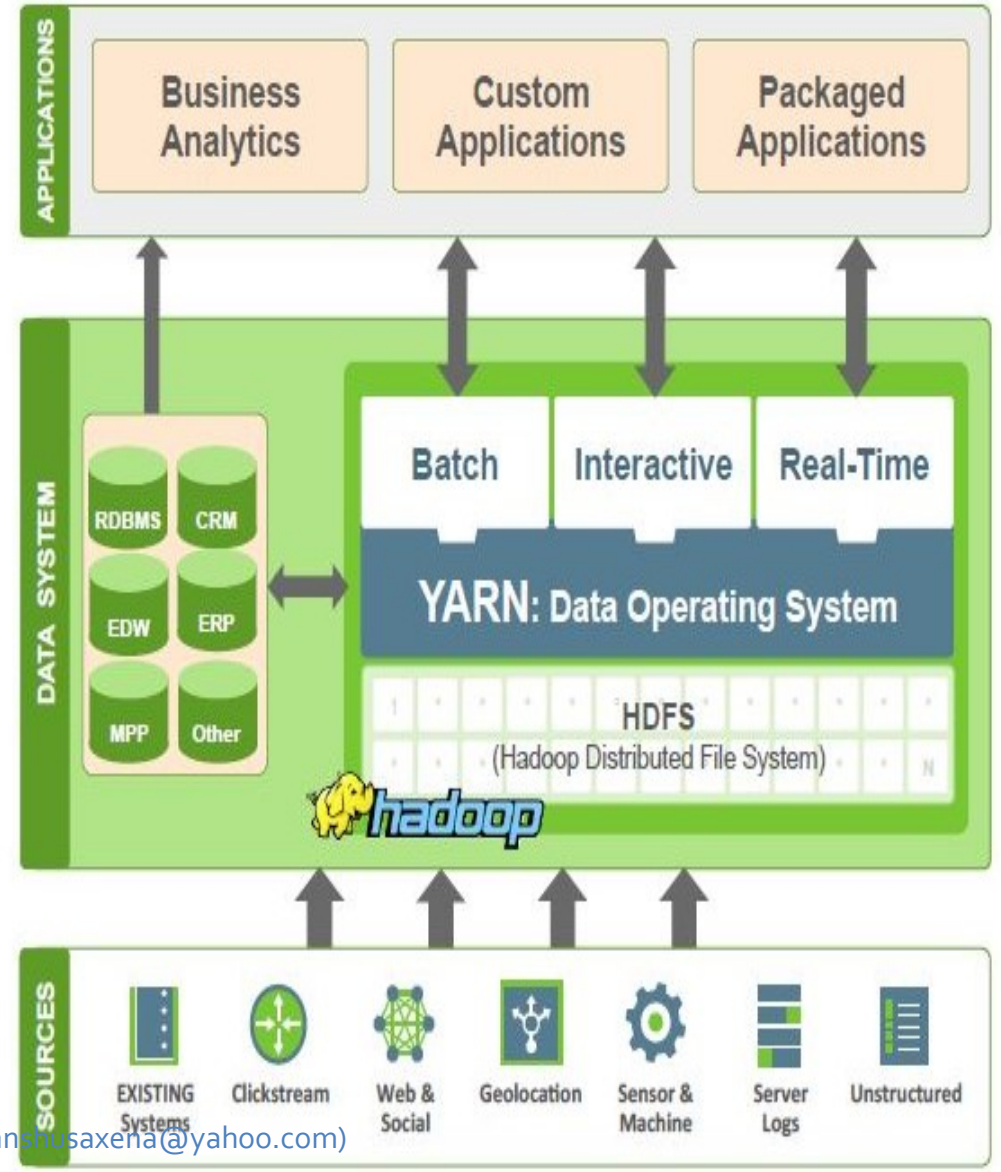


Modern Data Architecture for New Data



Enterprise Goals for Modern Data Architecture

- Centrally manage new and existing data.
- Provide single view of the customer, product, supply chain.
- Run batch, interactive & real time analytic applications on shared datasets.
- Assure enterprise grade security, operations and governance.
- Leverage new and existing data center infrastructure investments.
- Scalable and affordable; low cost per TB
- Deployment flexibility



Who Uses Hadoop?

WHAT TRIGGERED BIG DATA TECHNOLOGIES

Knowing Hadoop when it wasn't hadoop…

Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

Hadoop Creation History

Doug Cutting adds DFS & MapReduce support to Nutch

Doug Cutting & Mike Cafarella started working on Nutch

Google publishes GFS & MapReduce papers

NY Times converts 4TB of image archives over 100 EC2s

Yahoo! hires Cutting, Hadoop spins out of Nutch

Facebooks launches Hive: SQL Support for Hadoop

Fastest sort of a TB, 3.5mins over 910 nodes

Fastest sort of a TB, 62secs over 1,460 nodes
Sorted a PB in 16.25hours over 3,658 nodes

cloudera Founded

Doug Cutting joins Cloudera

Hadoop Summit 2009, 750 attendees

2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009

Source: Cloudera, Inc.

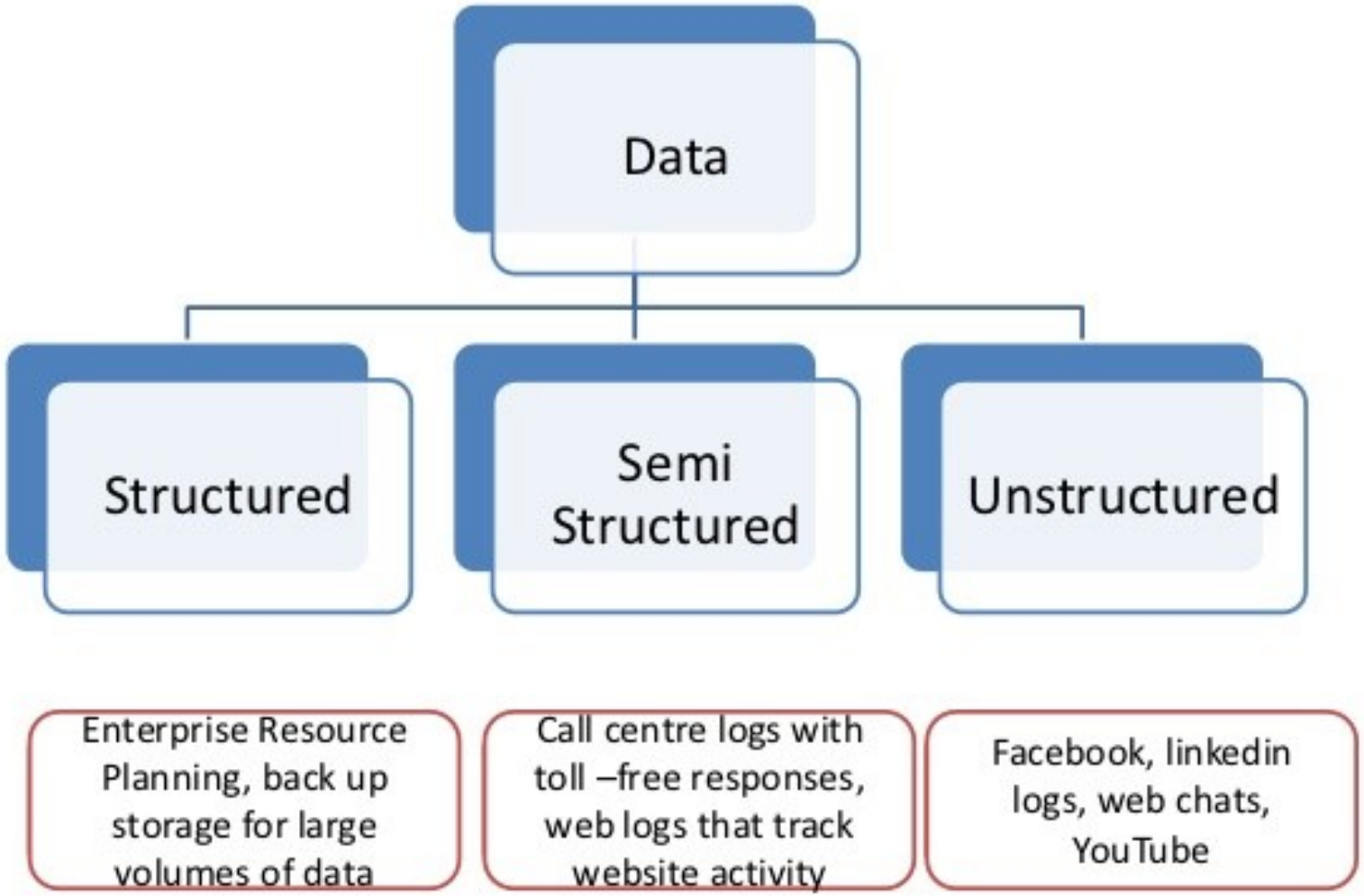
Structured semi-structured and unstructured data

Structured Data
The data which can be co-related with the relationship keys, in a geeky word, RDBMS data! Maximum processing is happening on this type of data even today but then it constitutes around **5%** of the total digital data!

Semi Structured Data
The structured data which does not conform with formal structure of data models in context of relationships is semi-structured data. Examples could be XMl, JSON, some NoSQL databases like MongoDB which store the data natively in JSON.
This again represents another 5% of the total available data.

Unstructured Data
All the remaining data having no structure at all, falls into this category and according to IDC estimate, it represents whopping 90% in share.
Examples could be sensor data (huge in percentage), social media streams, images, videos, mobile data, etc.

Types Of Data

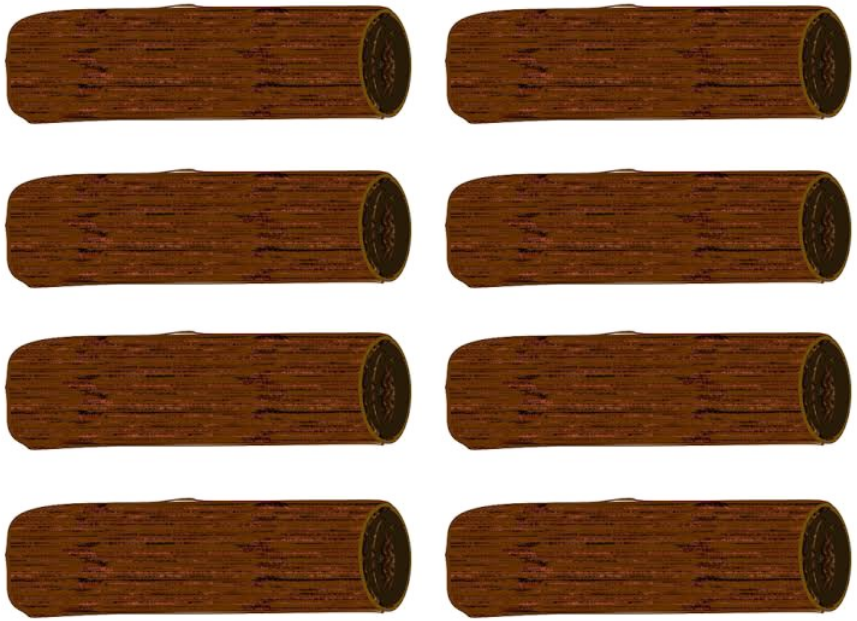


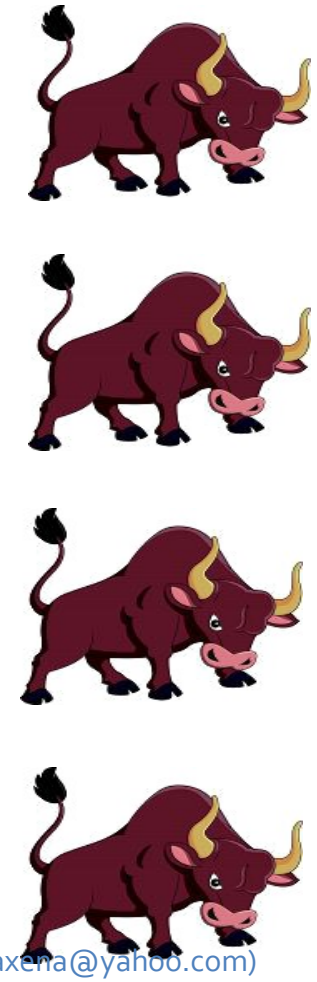
INTRODUCTION TO HADOOP MAGIC

What is the new thing that Hadoop brings to computing…

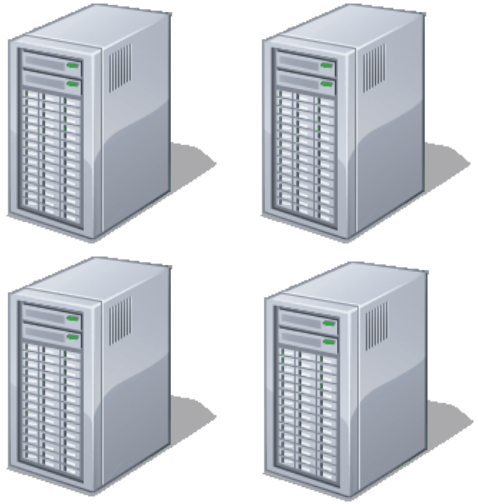
Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

Ox and the load





Distributed Computing

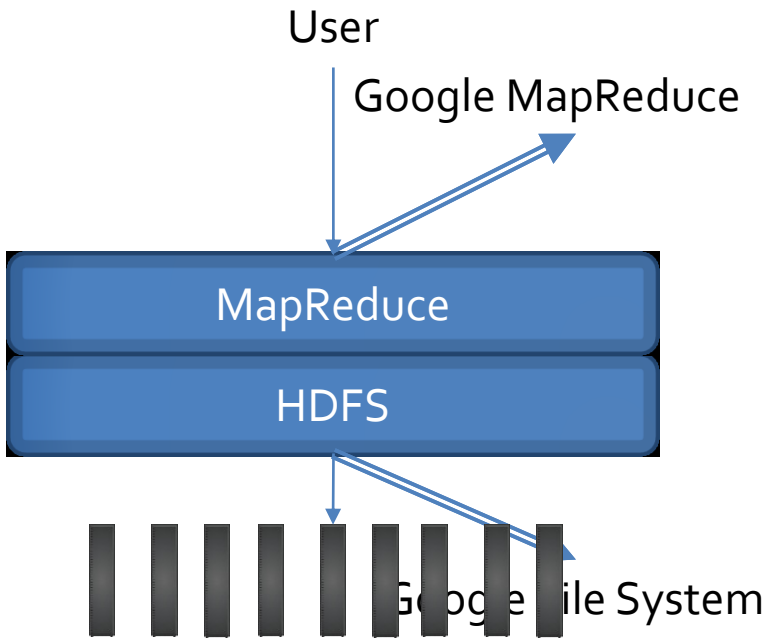


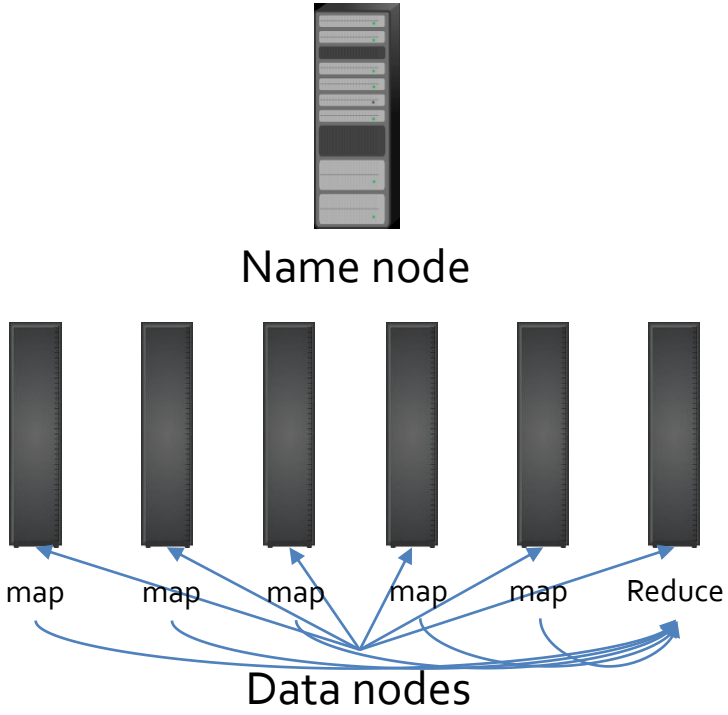
Price Advantage:

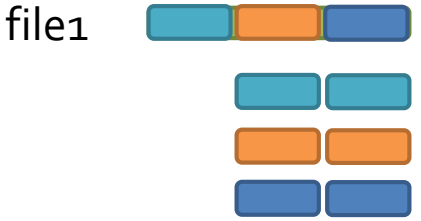
1. Clusters use commodity hardware, cheaper than one expensive server.

2. Software License is free.

Hadoop Framework – Brief overview







The new Fundamentals

- Moving the code to data

- Use of Commodity Hardware and Open Source Software against expensive proprietary software on expensive custom Hardware.

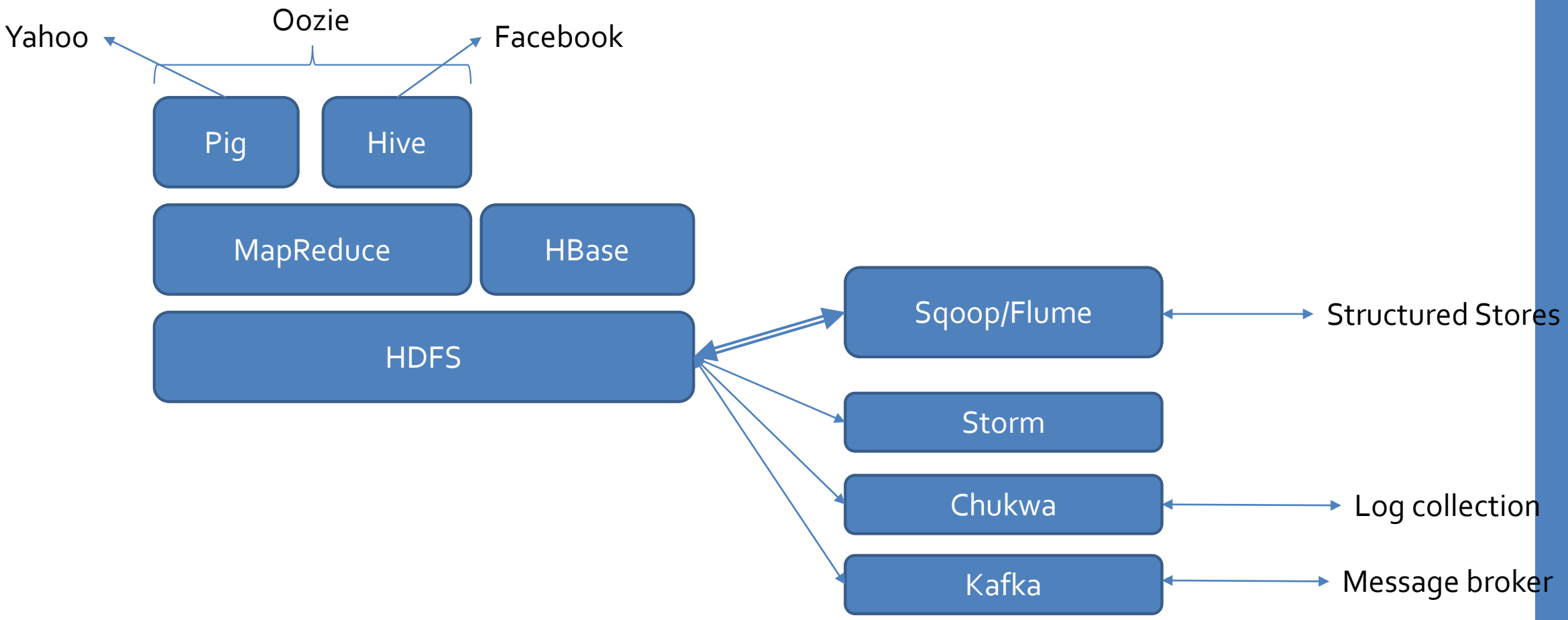
- On read schema.

HADOOP ECOSYSTEM

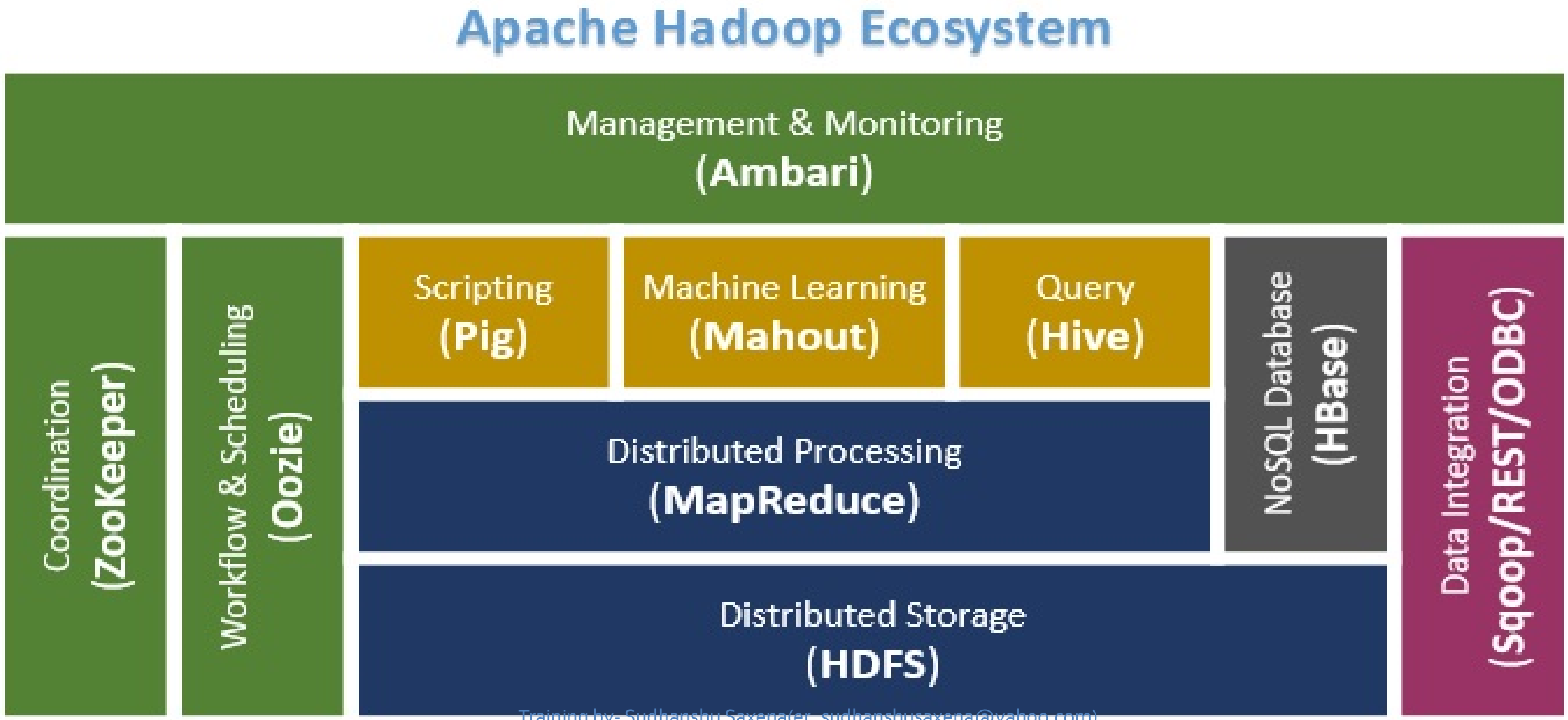
The umbrella of tools around hadoop…

Training by- Sudhanshu Saxena(er_sudhanshusaxena@yahoo.com)

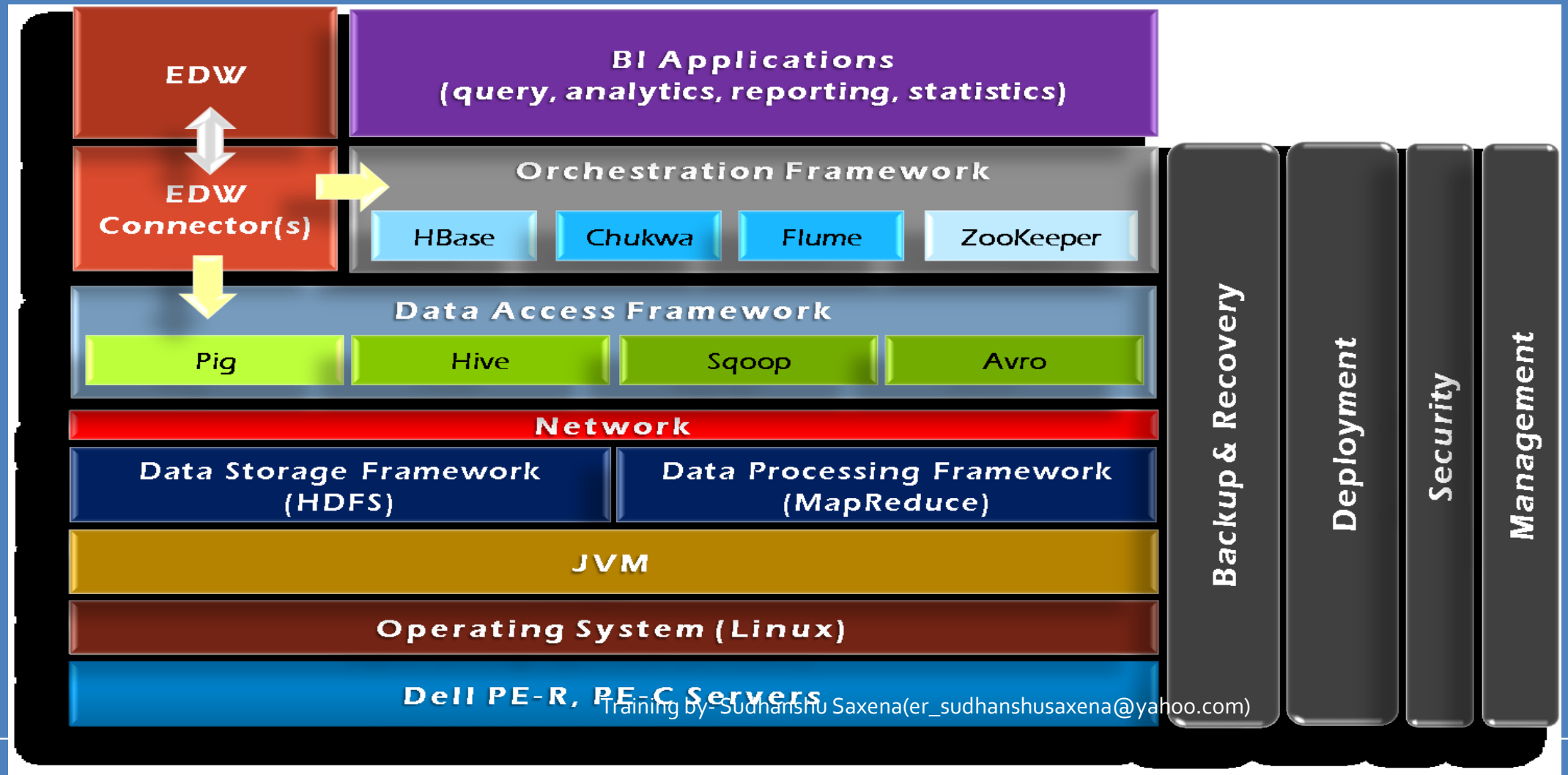
Hadoop Ecosystem



Hadoop Ecosystem



HADOOP FRAMEWORK TOOLS



Thank You!!!