Intro to Big Data

By: Shahab Safaee

Software Engineering PhD Candidate

Agenda

- Introduction
- What is Big Data?
- Big Data Sources
- Why Big Data
- Characteristic of Big Data
- Big Data Paradigm
- Big Data History
- Big Data Use Cases
- Big Data Investment-by Industry
- Big Data Anatomy and Stack Layers
- Hadoop Ecosystem
- Big Data Frameworks
- Integrated Cloud & Big Data

Introduction (1)

- Data
  - In computing, data is Information that has been translated into a form that is efficient for movement or processing.
- Operations
  - Generating and Consuming
  - Processing
  - Store
  - Retrieval

Introduction (2)

- Traditional Paradigm
  - Few companies are generating data, all others are consuming data

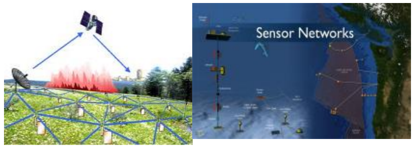
- New Paradigm
  - all of us are generating data, and all of us are consuming data

Who’s Generating Data

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

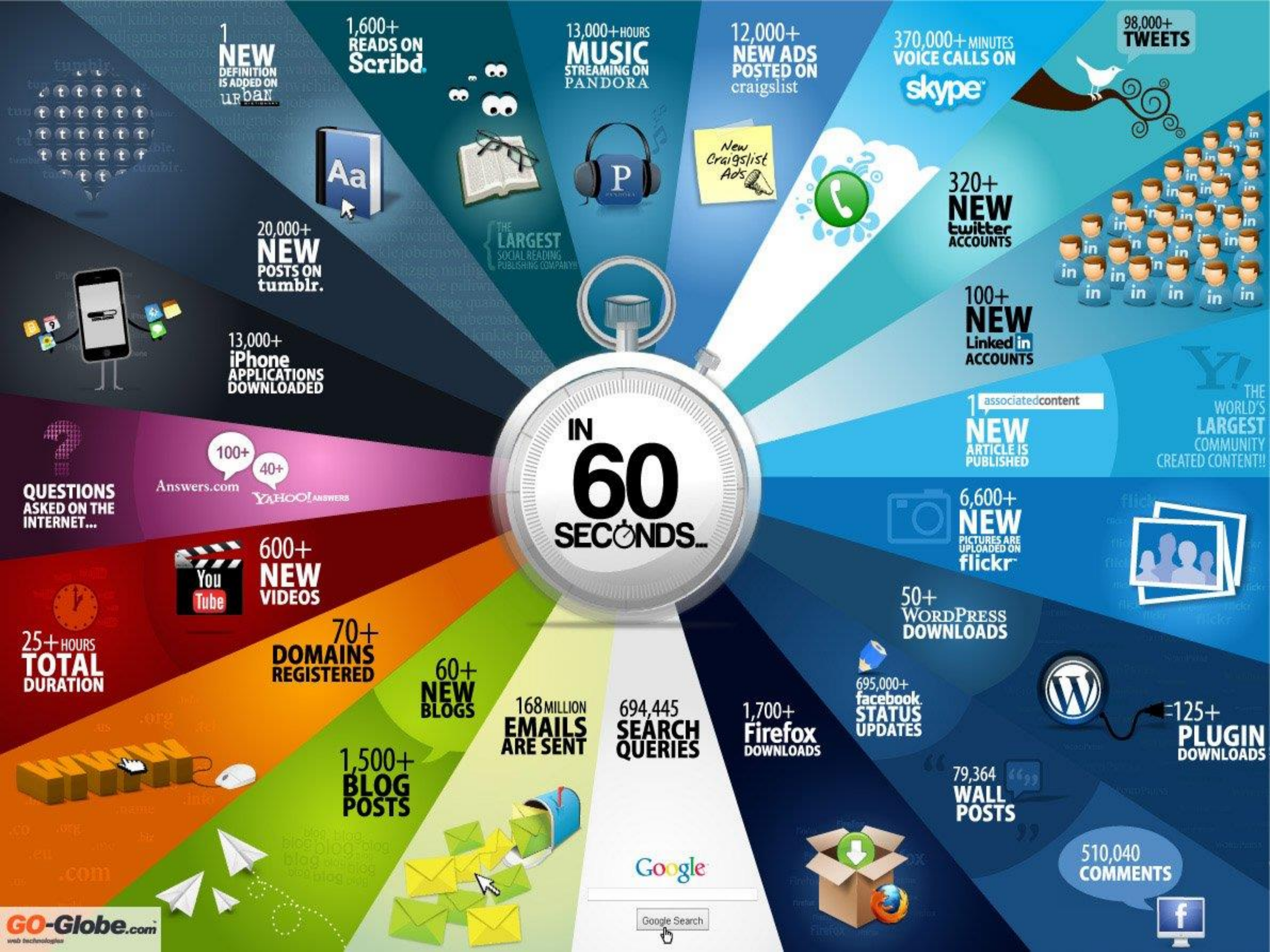
Scientific instruments
(collecting all sorts of data)

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



IN 60 SECONDS...

1 NEW DEFINITION IS ADDED ON urban DICTIONARY

1,600+ READS ON Scribd.

THE LARGEST SOCIAL READING PUBLISHING COMPANY!!

13,000+ HOURS MUSIC STREAMING ON PANDORA

12,000+ NEW ADS POSTED ON craigslist

New Craigslist Ads

370,000+ MINUTES VOICE CALLS ON skype

98,000+ TWEETS

320+ NEW twitter ACCOUNTS

100+ NEW LinkedIn ACCOUNTS

1 associatedcontent NEW ARTICLE IS PUBLISHED

THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!

6,600+ NEW PICTURES ARE UPLOADED ON flickr

50+ WORDPRESS DOWNLOADS

125+ PLUGIN DOWNLOADS

695,000+ facebook STATUS UPDATES

79,364 WALL POSTS

510,040 COMMENTS

1,700+ Firefox DOWNLOADS

694,445 SEARCH QUERIES

Google

Google Search

168 MILLION EMAILS ARE SENT

60+ NEW BLOGS

1,500+ BLOG POSTS

70+ DOMAINS REGISTERED

25+ HOURS TOTAL DURATION

600+ NEW VIDEOS

You Tube

QUESTIONS ASKED ON THE INTERNET...

100+ Answers.com

40+ YAHOO! ANSWERS

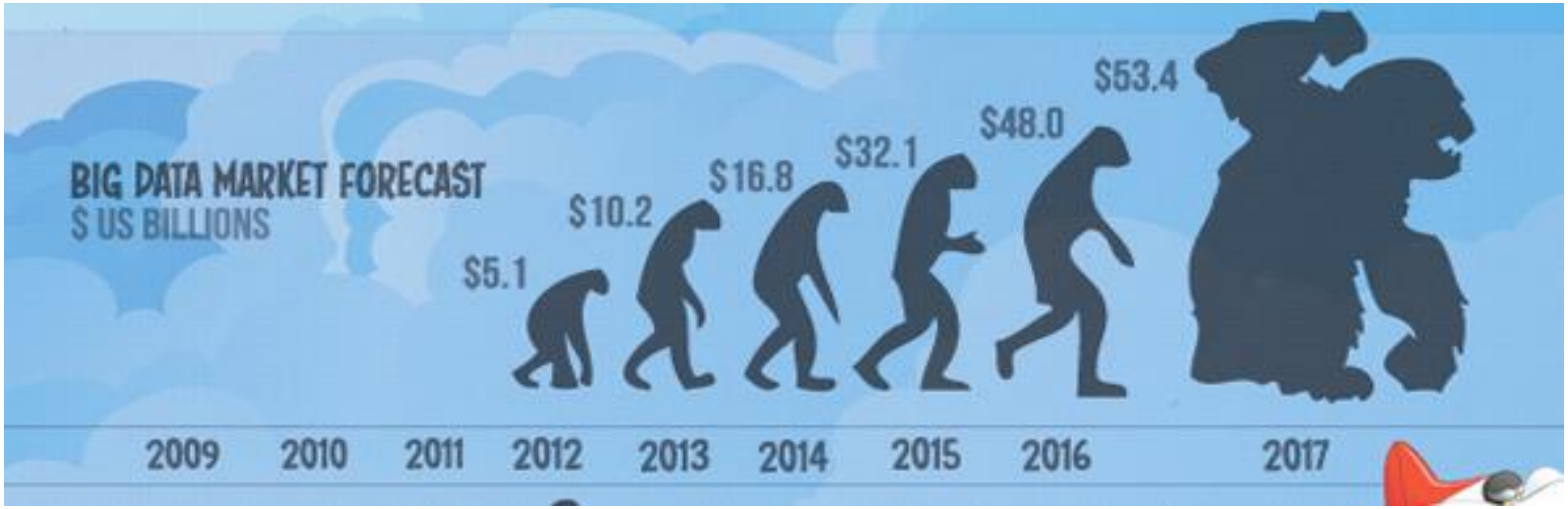
13,000+ iPhone APPLICATIONS DOWNLOADED

20,000+ NEW POSTS ON tumblr.

GO-Globe.com web technologies

What is Big Data? (1)

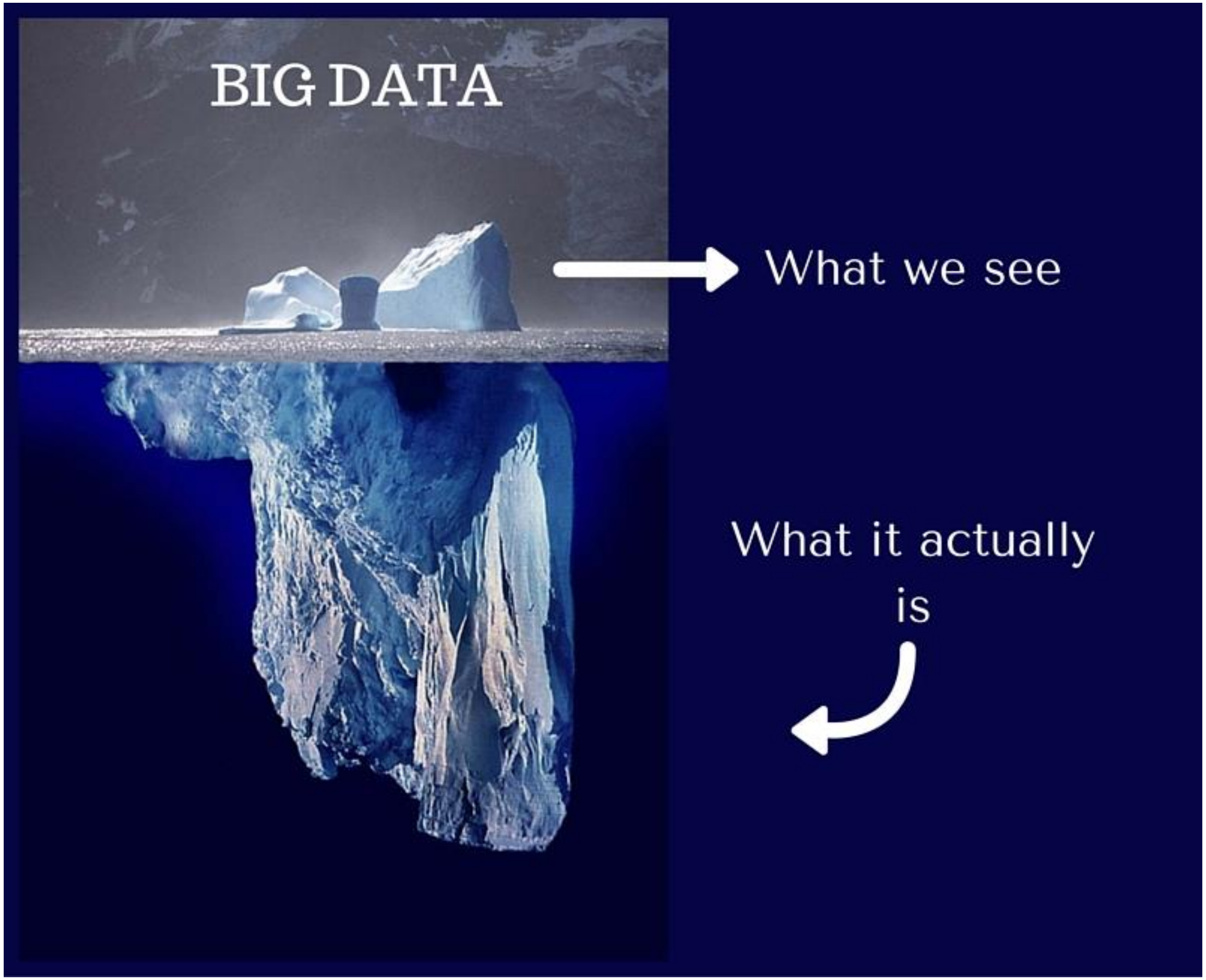
- Walmart handles more than 1 million customer transactions every hour.
- Facebook handles 40 billion photos from its user base.



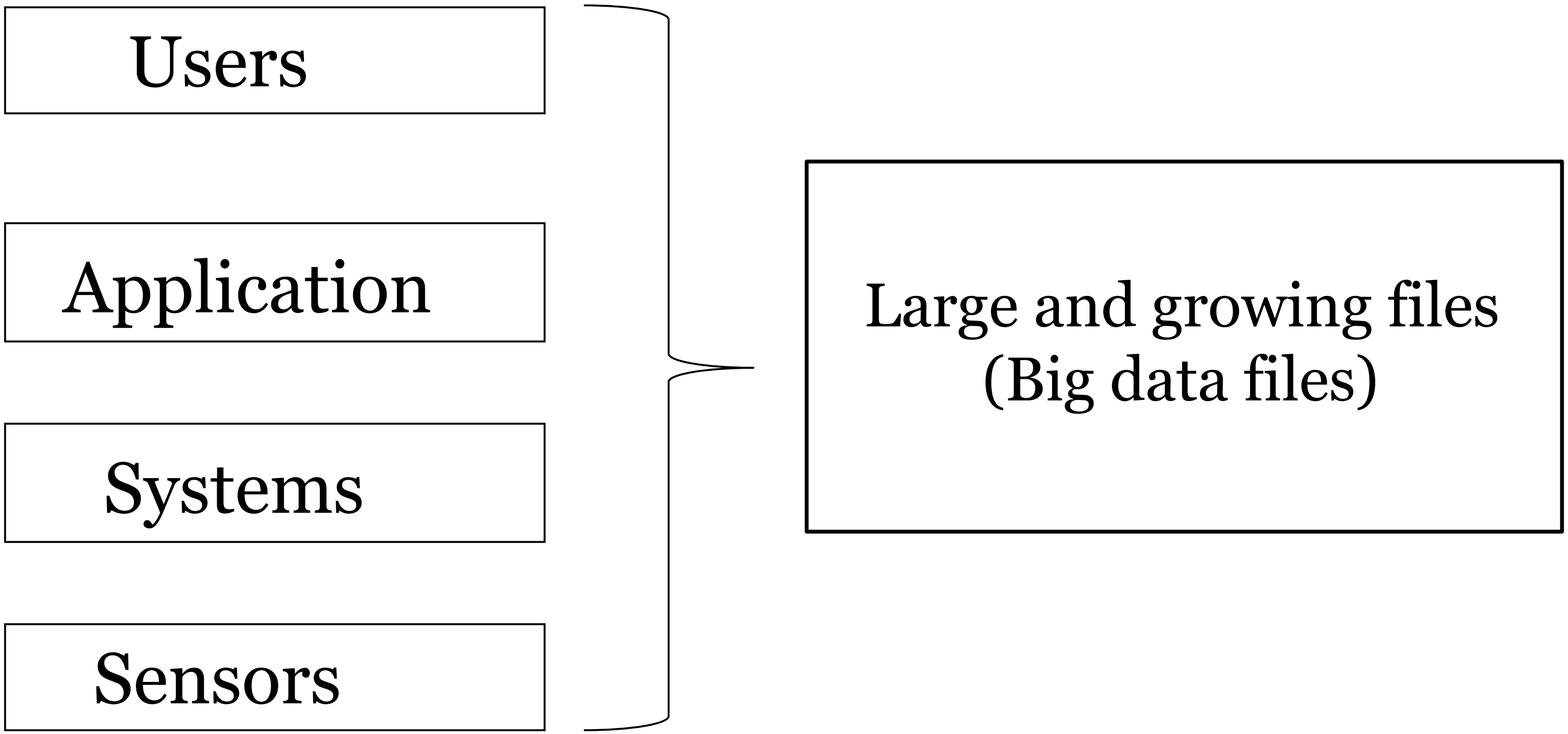
What is Big Data? (2)

- No single standard definition…
- "***Big Data***" is similar to 'small data', but bigger in size
    - but having data bigger it requires different approaches:
        - Techniques, tools and architecture
- "***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…
- an aim to solve new problems or old problems in a better way
- Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

What is Big Data? (3)



Big Data Sources



Users

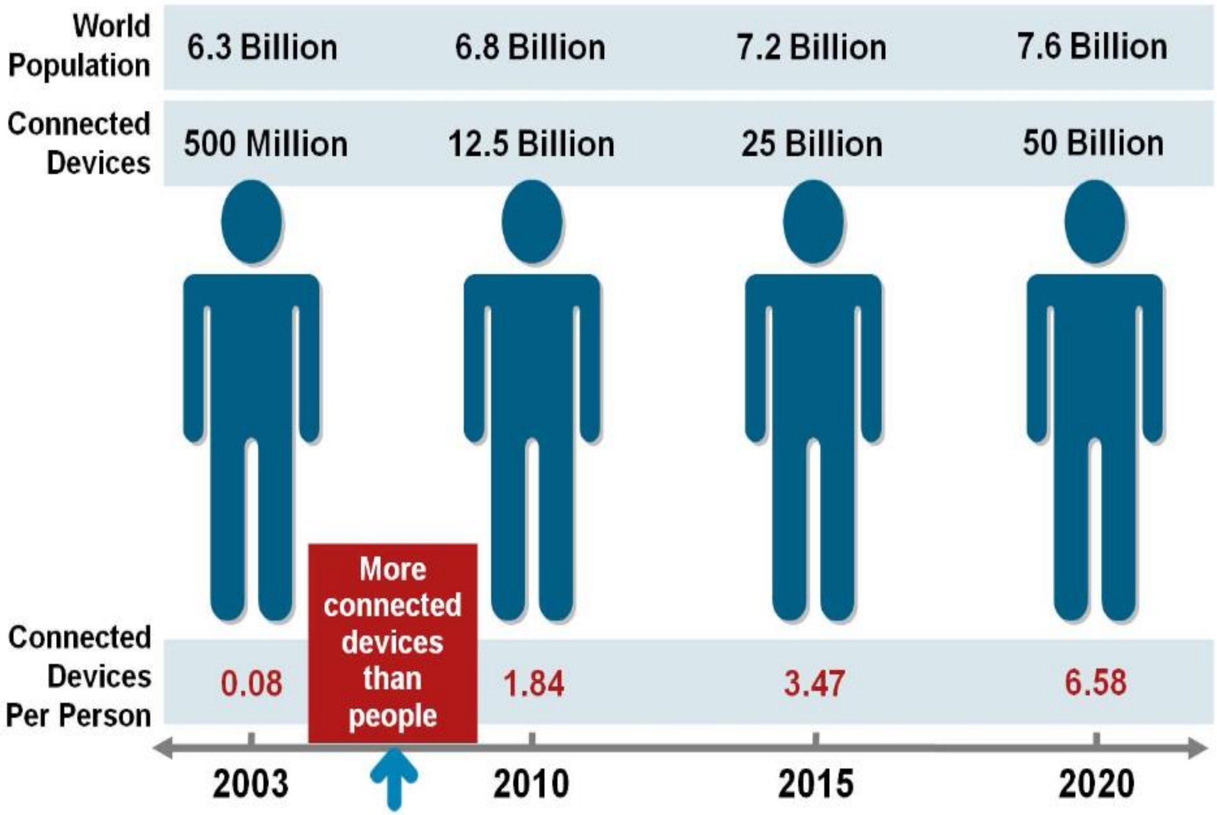
Application

Systems

Sensors

Large and growing files
(Big data files)

Why Big Data (1)



	2003	2010	2015	2020
World Population	6.3 Billion	6.8 Billion	7.2 Billion	7.6 Billion
Connected Devices	500 Million	12.5 Billion	25 Billion	50 Billion
Connected Devices Per Person	0.08	1.84	3.47	6.58

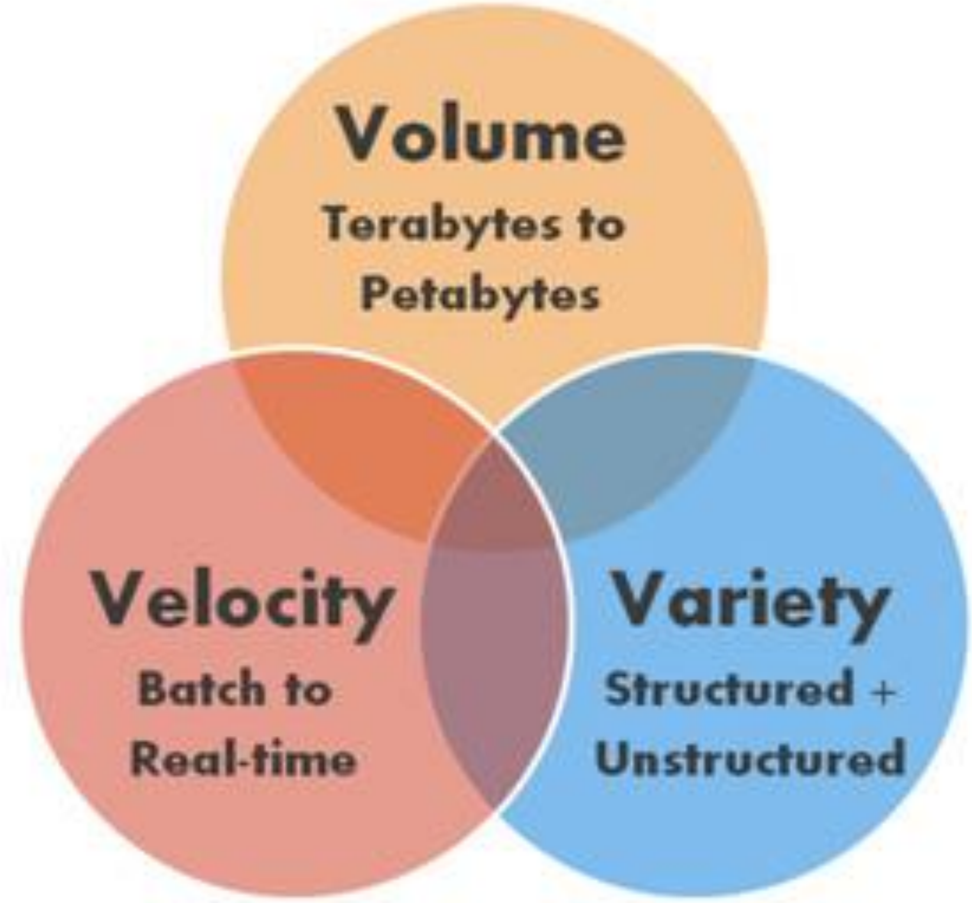
More connected devices than people

Source: Cisco IBSG, April 2011

Why Big Data (2)

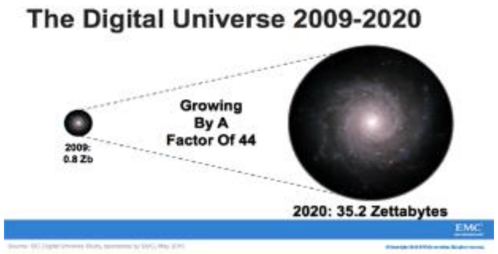
- Growth of Big Data is needed
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data(different data types)
  - Every day we create 2.5 quintillion (10^{18}) bytes of data; 90% of the data in the world today has been created in the last two years alone
- FB generates 10TB daily
- Twitter generates 7TB of data Daily
- IBM claims 90% of today’s stored data was generated in just the last two years.

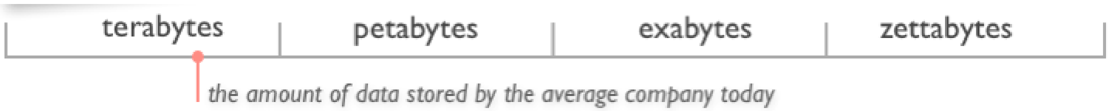
Characteristic of Big Data

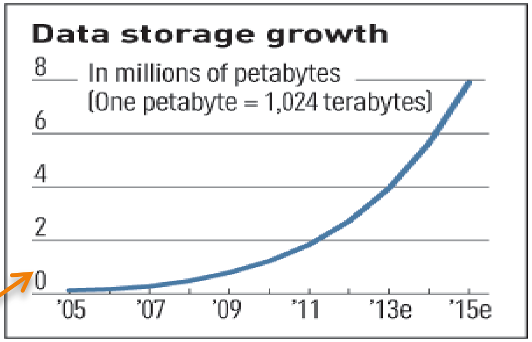


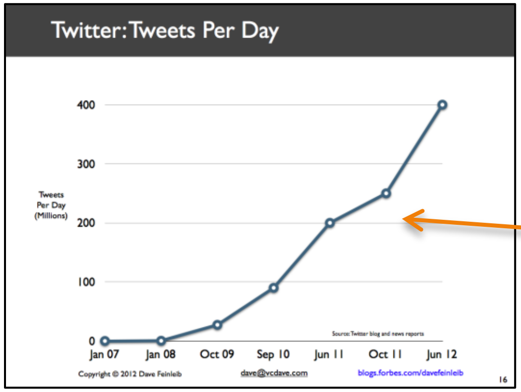
1st Character of Big Data: Volume (Scale) (1)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zetta bytes to 35zb
- Data volume is increasing exponentially









Exponential increase in collected/generated data

1st Character of Big Data: Volume (Scale) (2)

- A typical PC might have had 10 gigabytes of storage in 2000.
- Today, Facebook ingests 500 terabytes of new data every day.
- Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.
- The smart phones, the data they create and consume.
- sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video.

2st Character of Big Data: Velocity (Speed) (1)

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

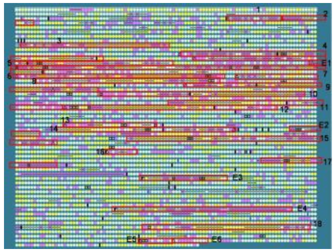
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

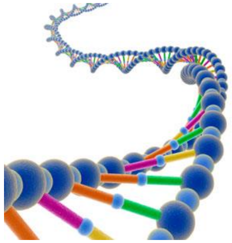
2st Character of Big Data: Velocity (Speed) (2)

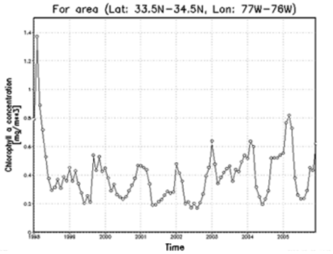
- Clickstreams and ad impressions capture user behavior at millions of events per second
- high-frequency stock trading algorithms reflect market changes within microseconds
- machine to machine processes exchange data between billions of devices
- infrastructure and sensors generate massive log data in real-time
- on-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

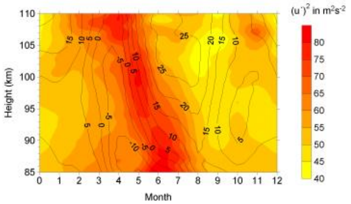
3st Character of Big Data: Variety (Complexity) (1)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data






For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m3]
Time

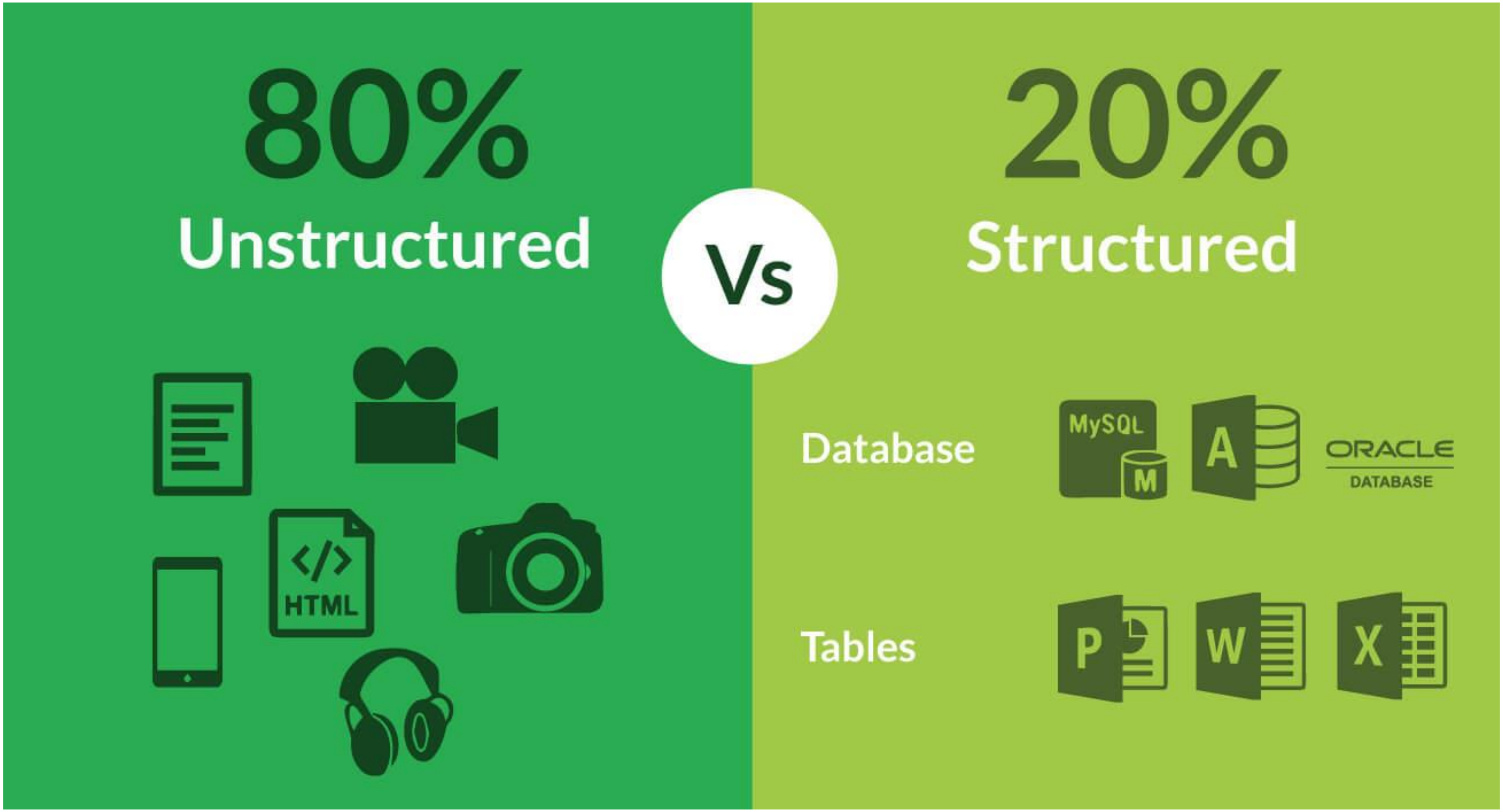


To extract knowledge➔ all these types of data need to linked together

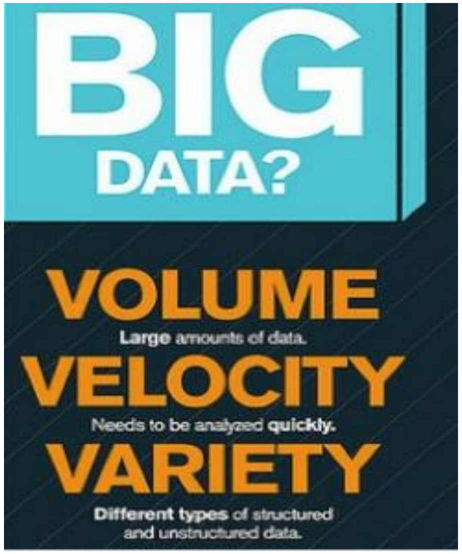
3st Character of Big Data: Variety (Complexity) (2)

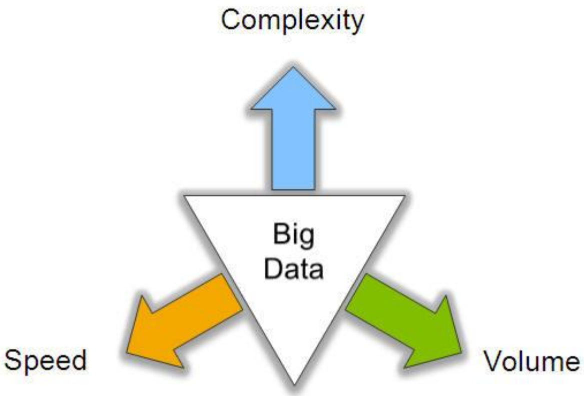
- Big Data isn't just numbers, dates, and strings. Big Data is also geospatial data, 3D data, audio and video, and unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.
- Big Data analysis includes different types of data

The Structure of Big Data



Big Data: 3V's (1)


BIG
DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.

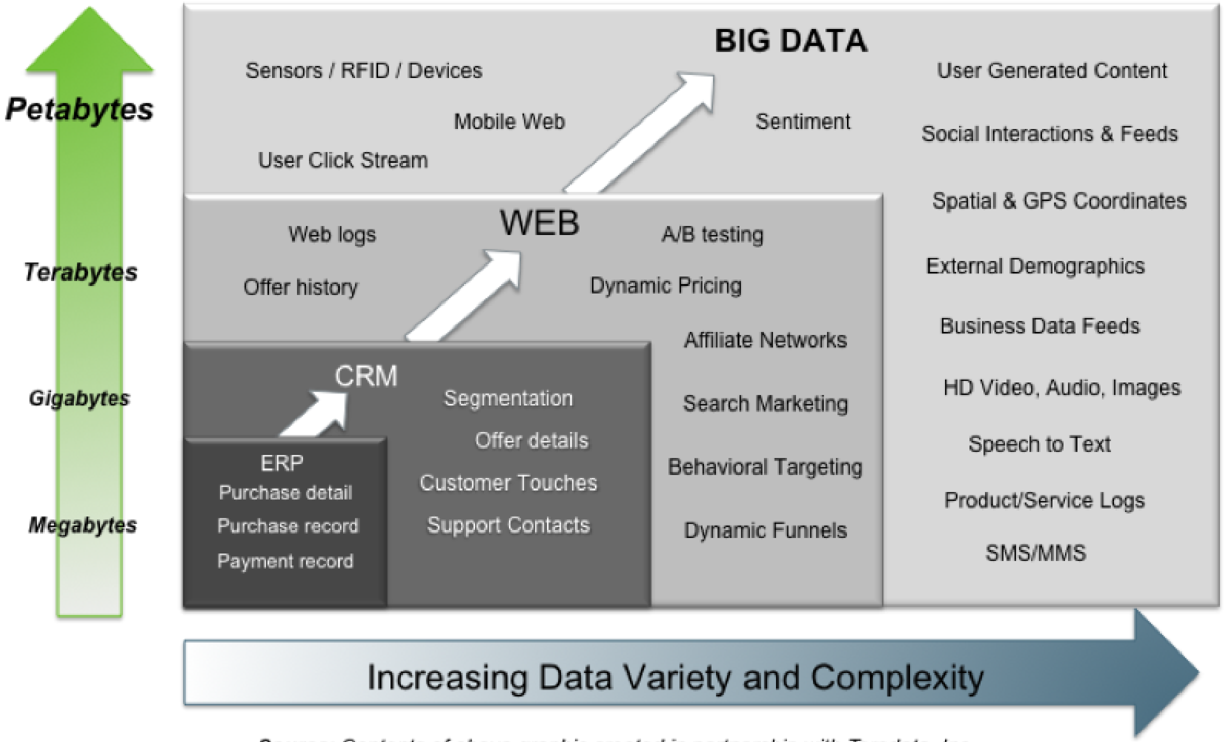

Complexity

Big Data

Speed

Volume

Big Data = Transactions + Interactions + Observations


Petabytes

Terabytes

Gigabytes

Megabytes

BIG DATA

Sensors / RFID / Devices

Mobile Web

User Click Stream

Sentiment

User Generated Content

Social Interactions & Feeds

Spatial & GPS Coordinates

External Demographics

Business Data Feeds

HD Video, Audio, Images

Speech to Text

Product/Service Logs

SMS/MMS

WEB

Web logs

Offer history

A/B testing

Dynamic Pricing

Affiliate Networks

Search Marketing

Behavioral Targeting

Dynamic Funnels

CRM

Segmentation

Offer details

Customer Touches

Support Contacts

ERP

Purchase detail

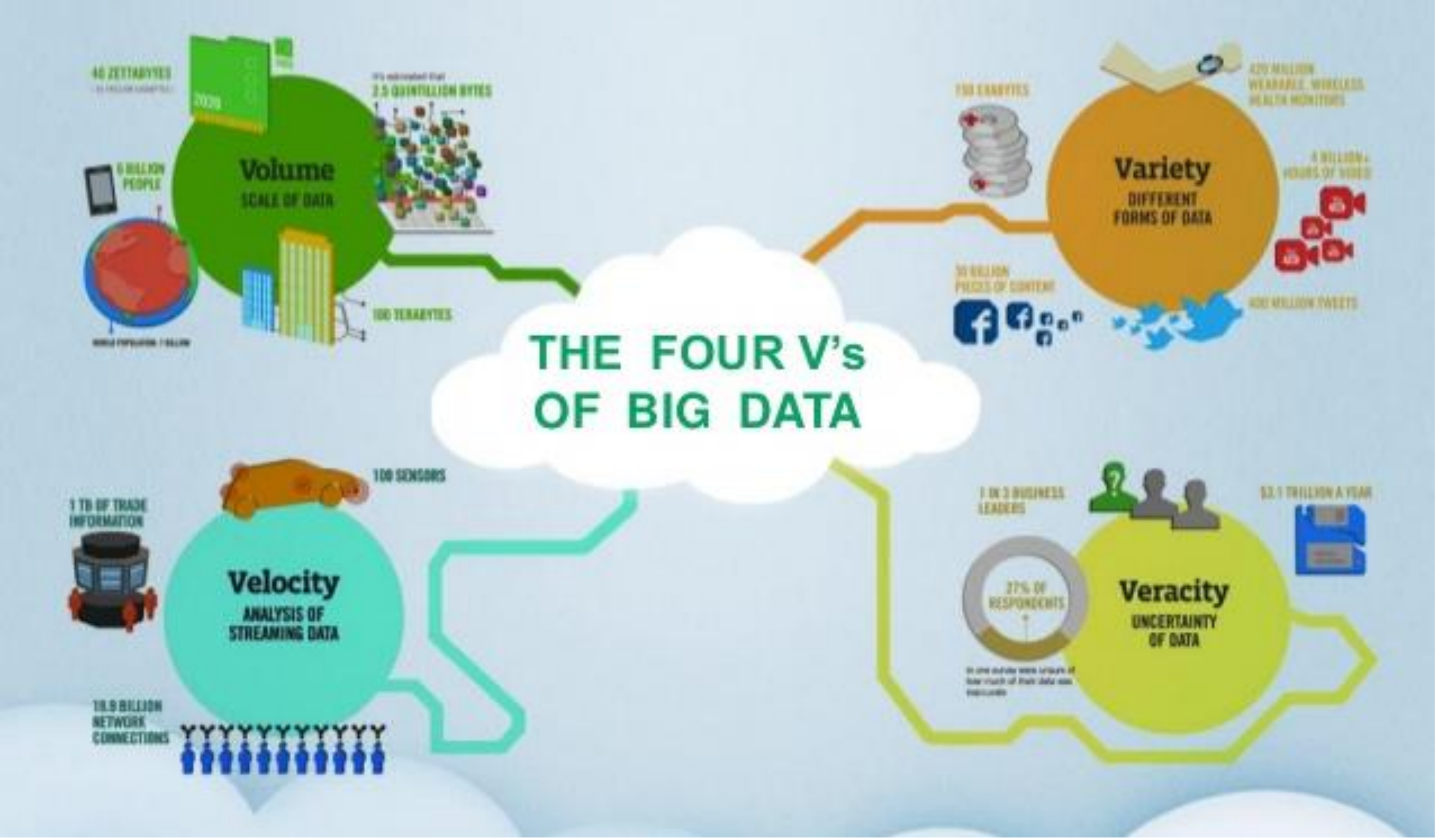
Purchase record

Payment record

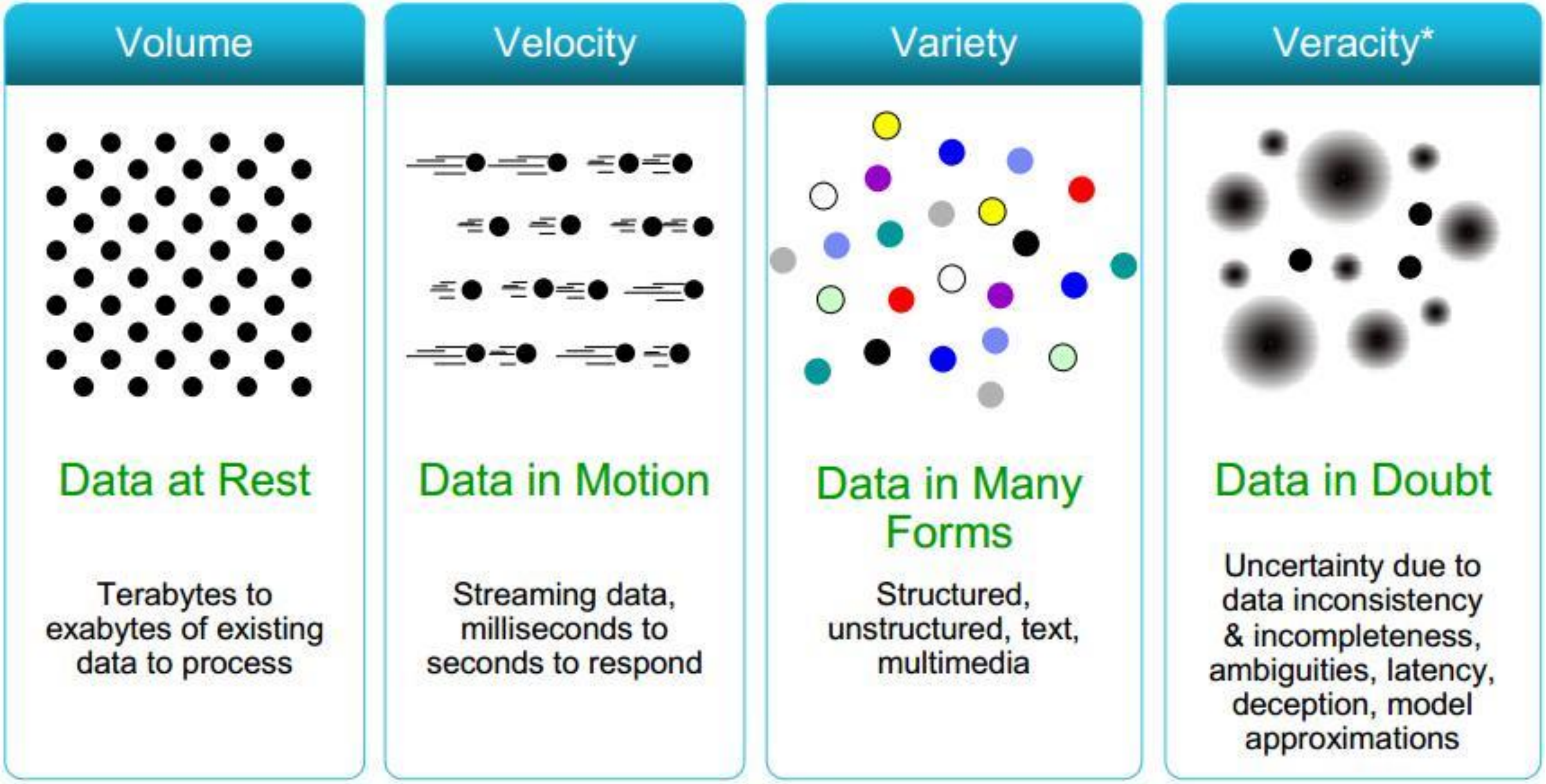
Increasing Data Variety and Complexity

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

Some Make it 4V's (1)

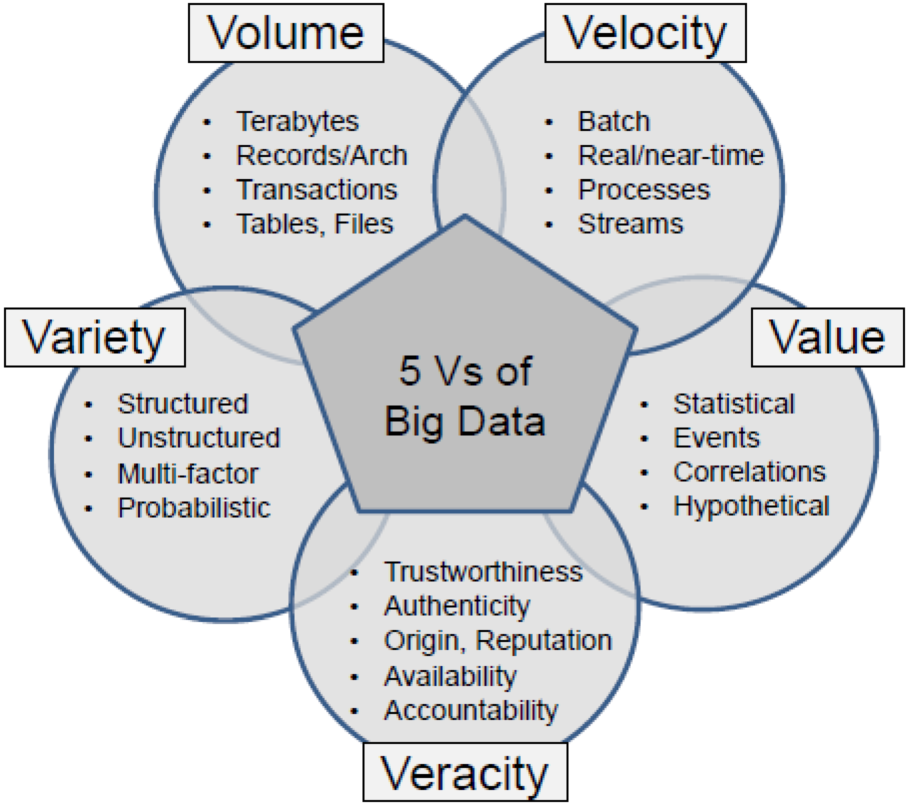


Some Make it 4V's (2)

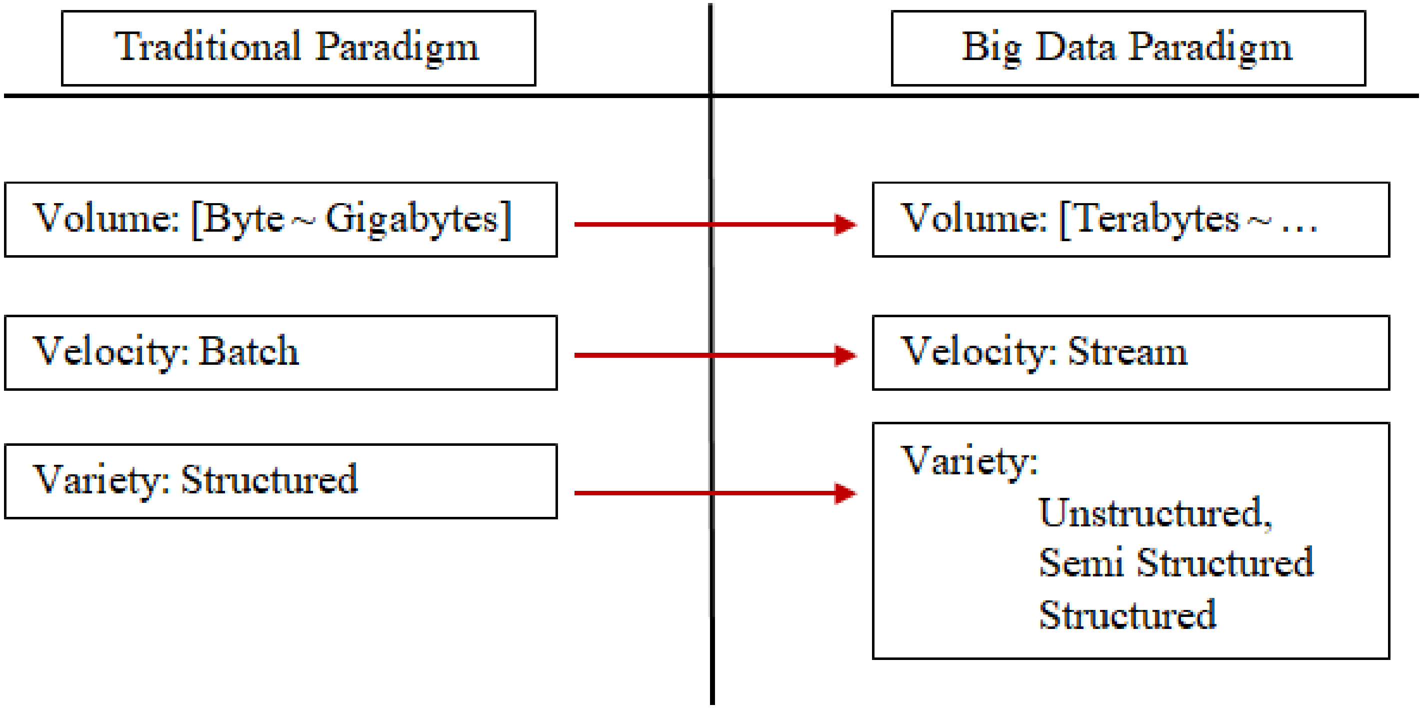


Volume	Velocity	Variety	Veracity*
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Some Make it 5V’s (1)



Big Data Paradigm



Traditional Paradigm	Big Data Paradigm
Volume: [Byte ~ Gigabytes]	Volume: [Terabytes ~ …
Velocity: Batch	Velocity: Stream
Variety: Structured	Variety: Unstructured, Semi Structured Structured

Big Data History

- Two major milestones in the development of Hadoop also added confidence into the Power of open source and Big Data Technologies.
- Only two years after its first release, in 2008, Hadoop won the terabyte sort benchmark in big data history. This is the first time that either a Java or an open source program has won.
- In 2010 Facebook claimed that they had the largest Hadoop cluster in the world with 21 PB of storage for their social messaging platform.

Big Data Use Cases (1)

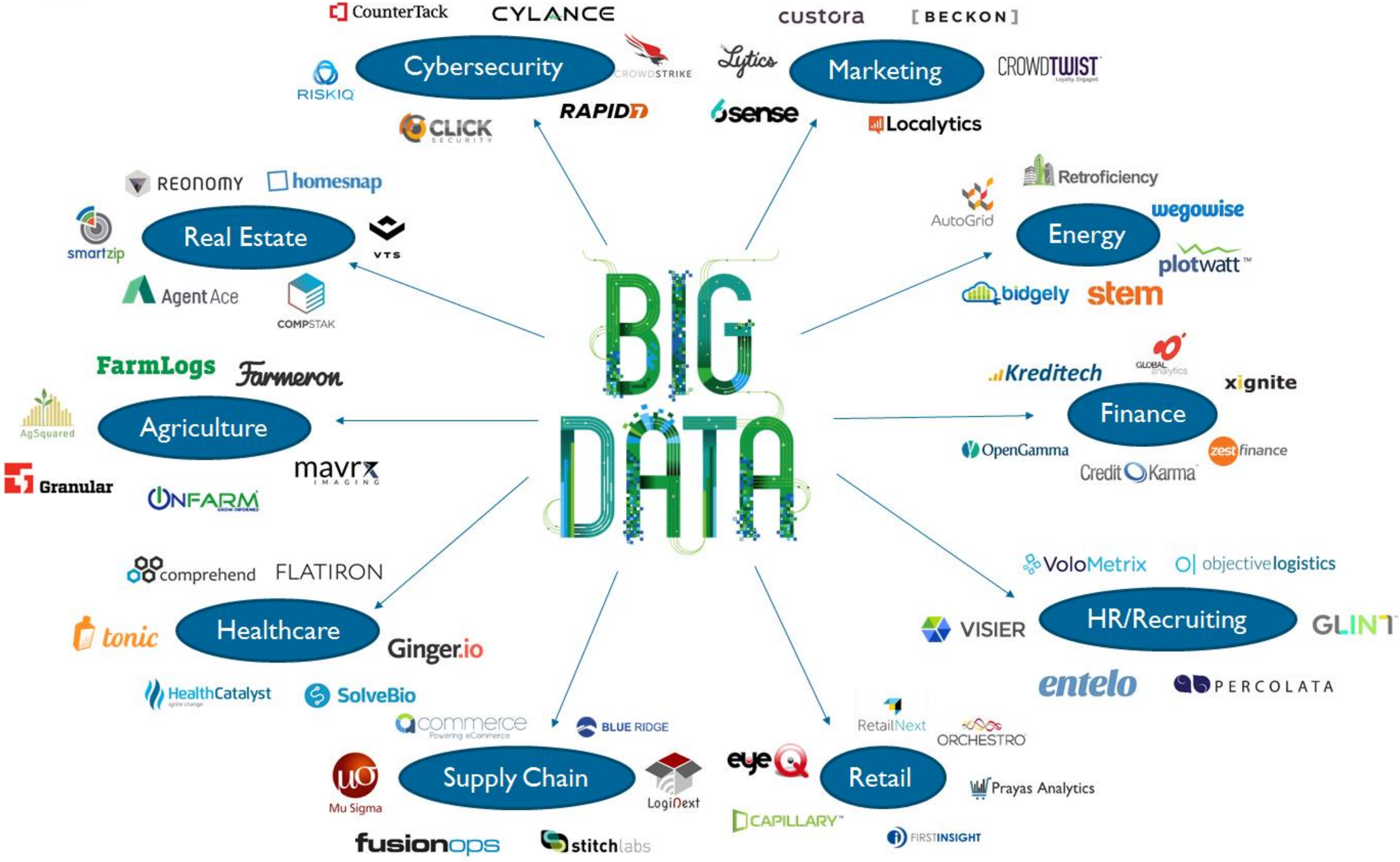
- Financial services
  - 360-degree complete view of customer
  - Risk and fraud monitoring and management
  - Real-time transaction tracking and analytics
- Healthcare/Life sciences
  - Disease diagnosis analysis
  - Medical record text analysis
  - Genomic analytics
- Telecommunications
  - Real-time Call detail record CDR processing and analysis
  - Customer profile monetization and analysis
  - Real-time network element monitoring
  - Real-time Network fault analysis

Big Data Use Cases (2)

- Digital media
  - Real-time ad matching, analysis, and targeting
  - Website analytics and conversion tracking
- Retail
  - Cross-channel marketing
  - Customer Clustering and Segmentation
  - Click-stream analysis
  - Market Basket Analytics
  - Real-time Recommendation
  - Sentiment Analysis

CB INSIGHTS

Startups Using Big Data



BIG DATA

Cybersecurity: CounterTack, CYLANCE, RISKIQ, CROWDSTRIKE, CLICK SECURITY, RAPID7

Marketing: custora, [BECKON], Lytics, CROWDTWIST Loyalty. Engaged., 6sense, Localytics

Real Estate: REONOMY, homesnap, smartzip, VTS, Agent Ace, COMPSTAK

Energy: Retroficiency, AutoGrid, wegowise, plotwatt™, bidgely, stem

Agriculture: FarmLogs, Farmeron, AgSquared, mavrx IMAGING, Granular, ONFARM GROW INFORMED

Finance: Kreditech, GLOBAL analytics, xignite, OpenGamma, zest finance, Credit Karma

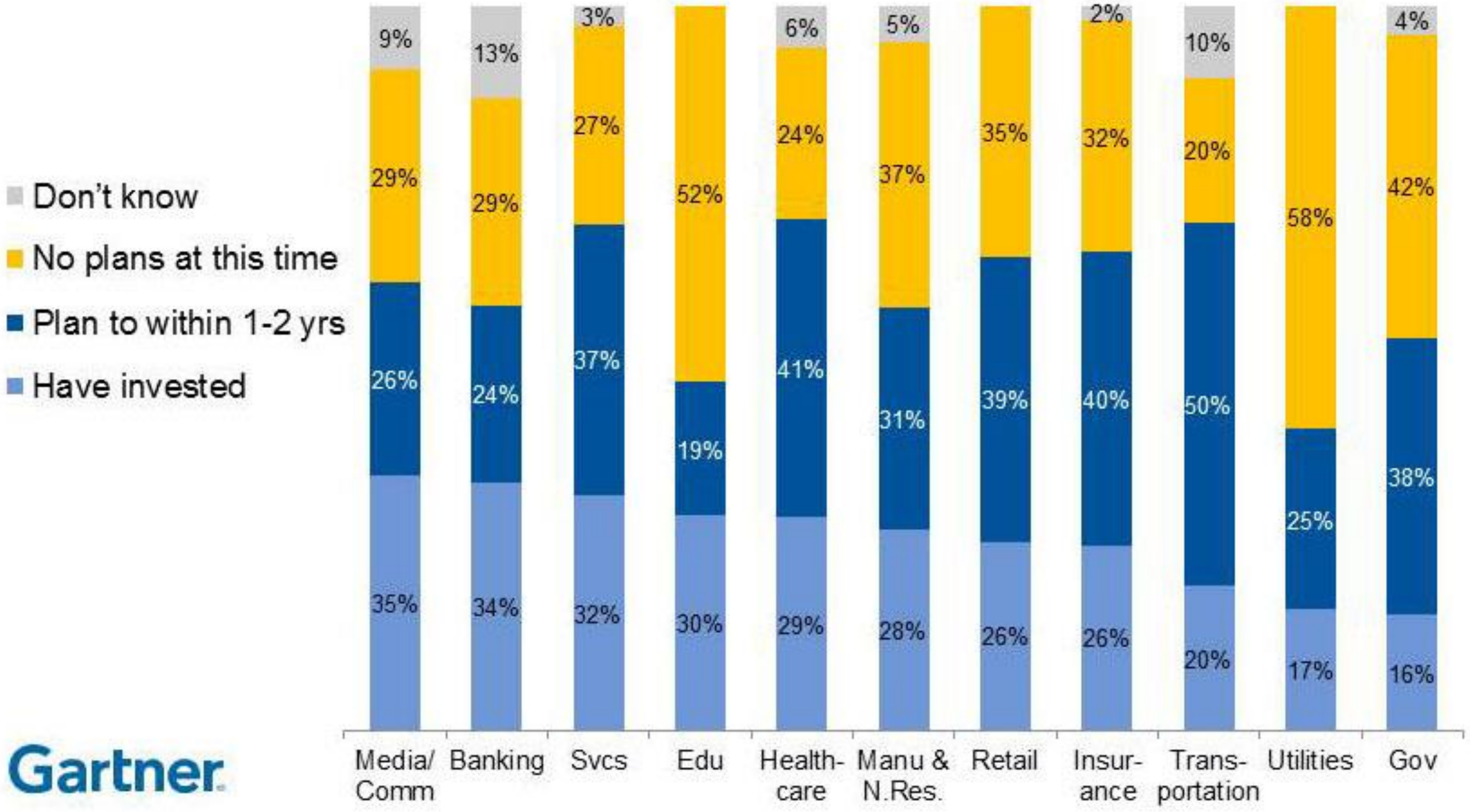
Healthcare: comprehend, FLATIRON, tonic, Ginger.io, HealthCatalyst, SolveBio

HR/Recruiting: VoloMetrix, objective logistics, VISIER, GLINT, entelo, PERCOLATA

Supply Chain: commerce Powering eCommerce, BLUE RIDGE, Mu Sigma, LogiNext, fusionops, stitchlabs

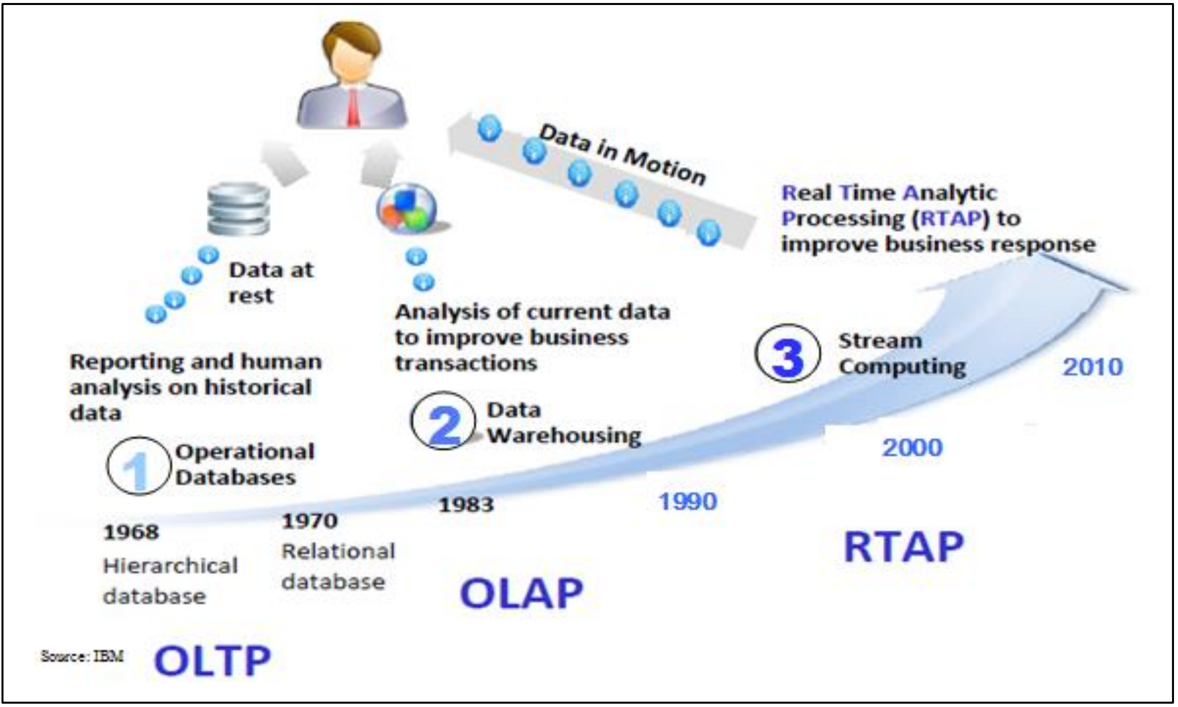
Retail: RetailNext, ORCHESTRO, eyeQ, Prayas Analytics, CAPILLARY™, FIRSTINSIGHT

Big Data Investment-by Industry



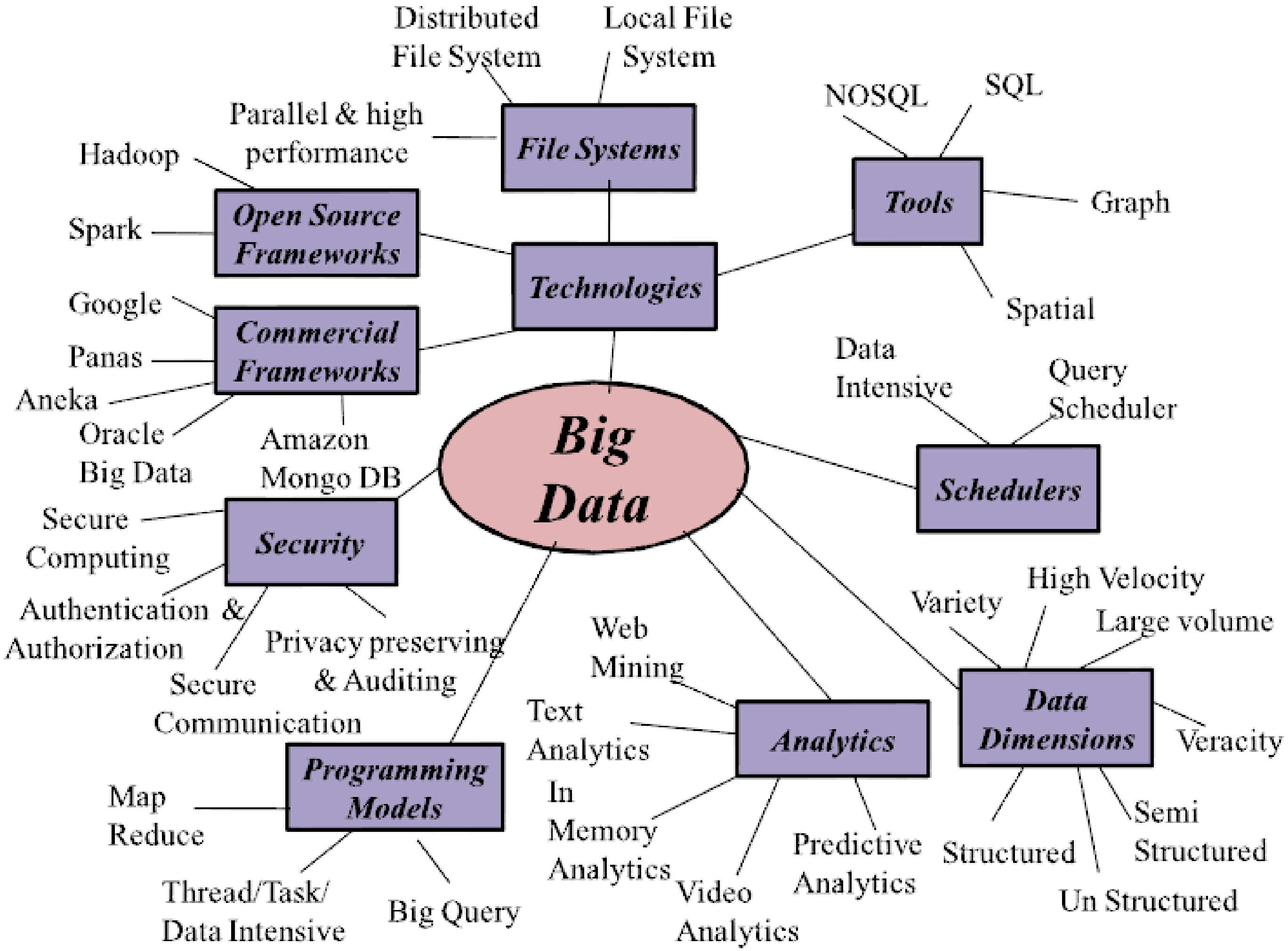
Harnessing Big Data

- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

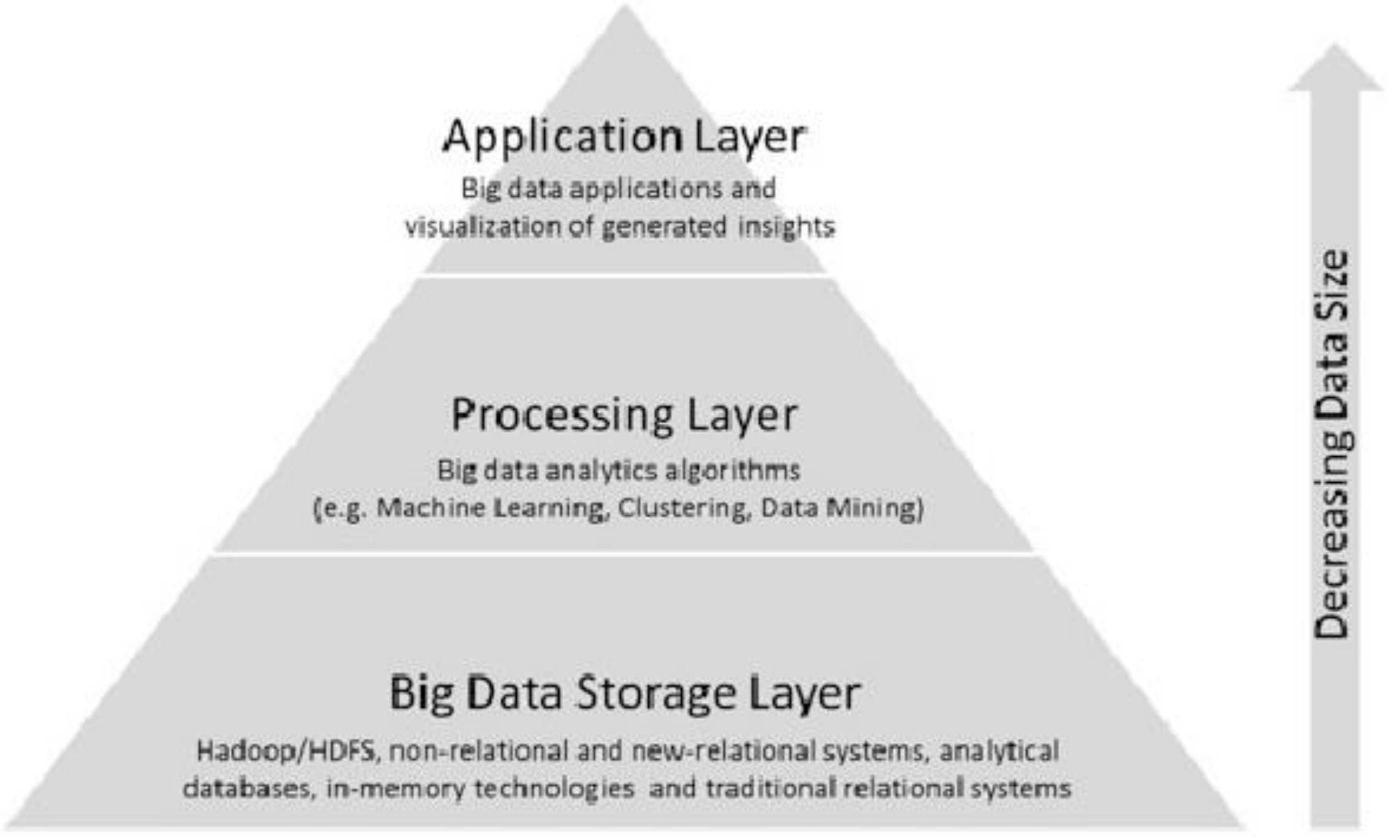

Data in Motion
Real Time Analytic Processing (RTAP) to improve business response
Data at rest
Analysis of current data to improve business transactions
Reporting and human analysis on historical data
3 Stream Computing
2010
2 Data Warehousing
2000
1 Operational Databases
1990
1983
1970 Relational database
1968 Hierarchical database
RTAP
OLAP
OLTP

Source: IBM

Big Data Anatomy



Big Data Stack Layers



Application Layer

Big data applications and visualization of generated insights

Processing Layer

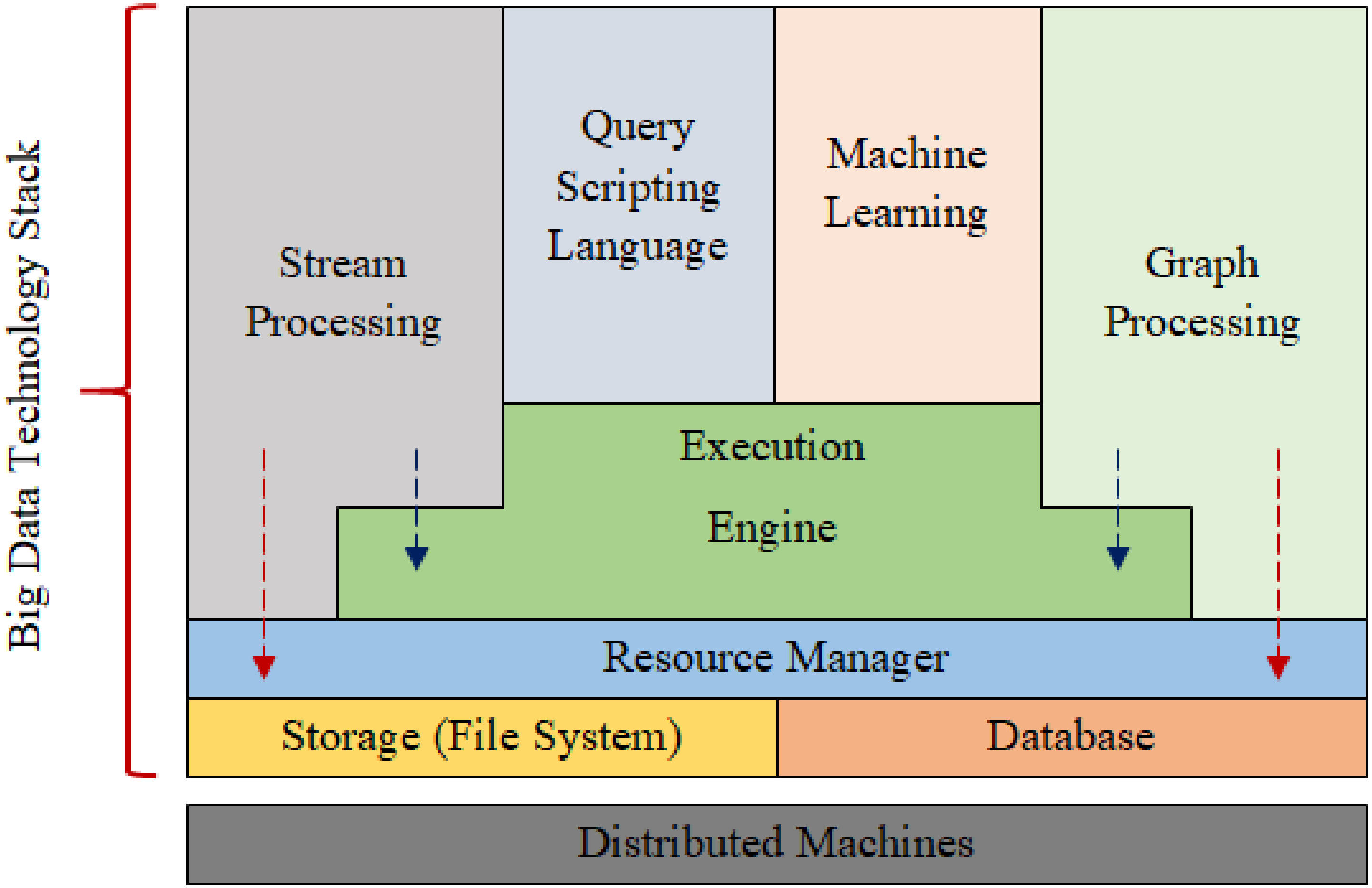
Big data analytics algorithms (e.g. Machine Learning, Clustering, Data Mining)

Big Data Storage Layer

Hadoop/HDFS, non-relational and new-relational systems, analytical databases, in-memory technologies and traditional relational systems

Decreasing Data Size

Big Data Technology Stack



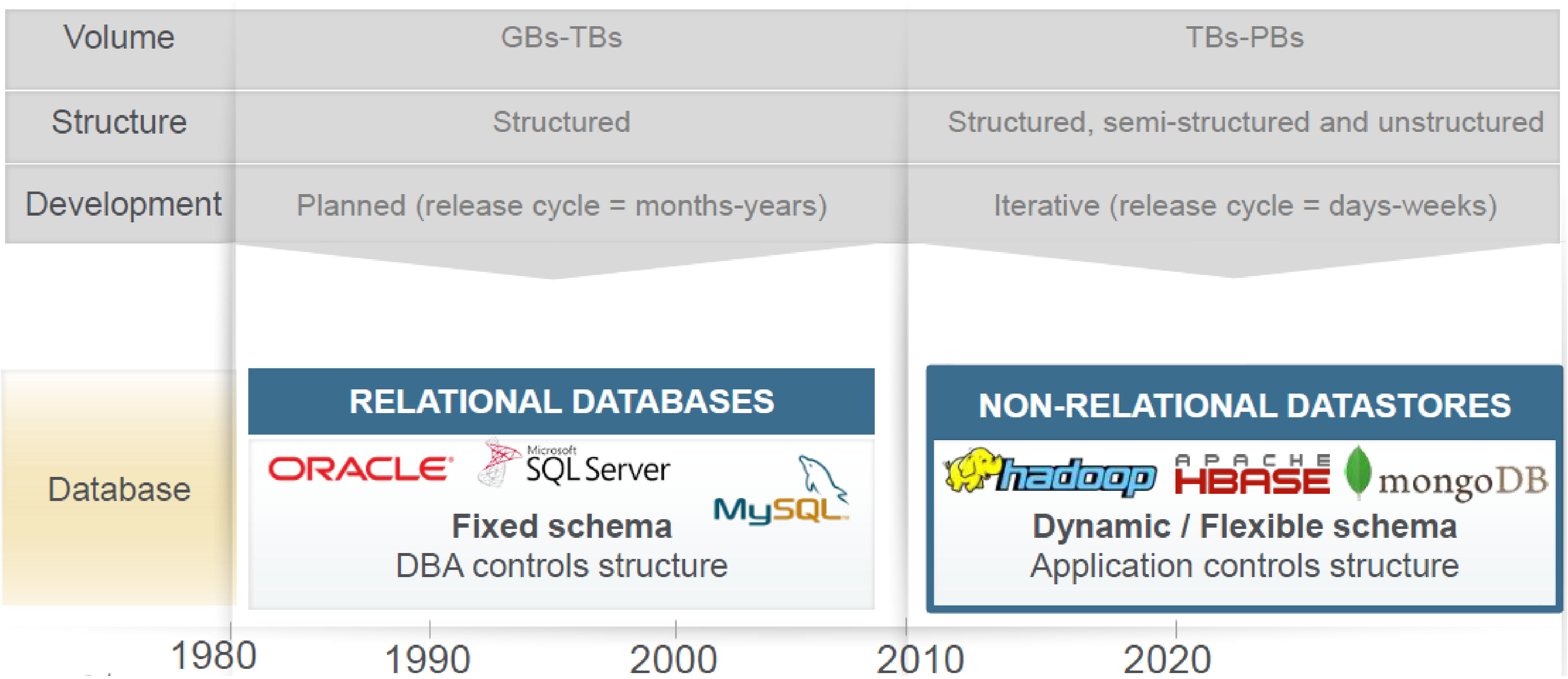
Big Data File Systems

- Features
  - Efficient Massive Data Support
  - Distributed Storing and Retrieving in Multiple Machines
- Tools
  - HDFS
  - S3
  - GPFS

Big Data Database

- NoSQL Database
  - NoSQL describes a fairly large number of NoSQL database technologies.
  - NoSQL databases are non-relational, distributed and schema-free.
- NoSQL Data Models
  - Key-value
  - Column-family
  - Document Oriented
  - Graph Based

RDBMS vs Non-RDBMS



Volume	GBs-TBs	TBs-PBs
Structure	Structured	Structured, semi-structured and unstructured
Development	Planned (release cycle = months-years)	Iterative (release cycle = days-weeks)
Database	**RELATIONAL DATABASES** ORACLE Microsoft SQL Server MySQL **Fixed schema** DBA controls structure	**NON-RELATIONAL DATASTORES** hadoop APACHE HBASE mongoDB **Dynamic / Flexible schema** Application controls structure

Big Data Execution Engine (1)

- Features
  - Scalable
  - Fault Tolerant
  - Parallelism
- Infrastructure Processing
  - Commodity Clustered Machines
- Programming Model
  - Divide and Conquer
- Programming Framework
  - Map/Reduce
    - It can model processing large data, split complications into different parallel tasks and make efficient use of large commodity clusters and distributed file systems.
- Tools
  - Hadoop Map/Reduce
  - Apache Spark

Big Data Query & Scripting Languages

- Features
  - High Level Language
  - Close to SQL Language
  - Translatable to Map/Reduce Functions
- Tools
  - Cassandra
  - Apache Hive
  - Apache Pig

Big Data Stream Processing

- Features
  - Fresh
  - Low Latency
- Tools
  - Storm
  - S4
  - …

Big Data Graph Processing

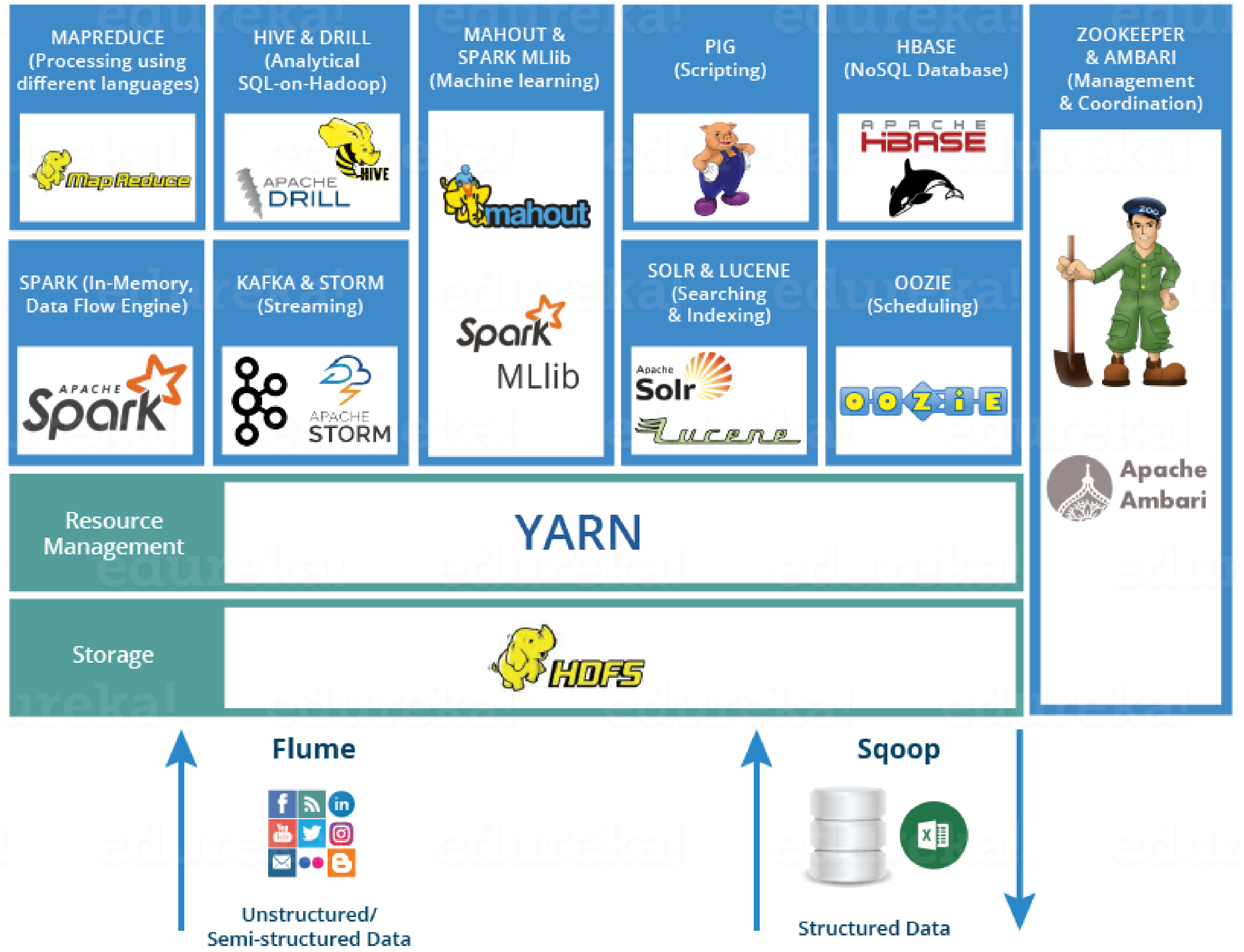
- Applications
  - Location Based Data
  - Social Networks
- Tools
  - Pregel
  - Apache Giraph
  - GraphLab
  - …

Big Data Machine Learning

- Tools
  - Mahout
  - ML Base
  - …

Big Data Other Tools

Hadoop Technology Stack



MAPREDUCE (Processing using different languages)

HIVE & DRILL (Analytical SQL-on-Hadoop)

MAHOUT & SPARK MLlib (Machine learning)

PIG (Scripting)

HBASE (NoSQL Database)

ZOOKEEPER & AMBARI (Management & Coordination)

SPARK (In-Memory, Data Flow Engine)

KAFKA & STORM (Streaming)

SOLR & LUCENE (Searching & Indexing)

OOZIE (Scheduling)

Resource Management

YARN

Storage

Flume

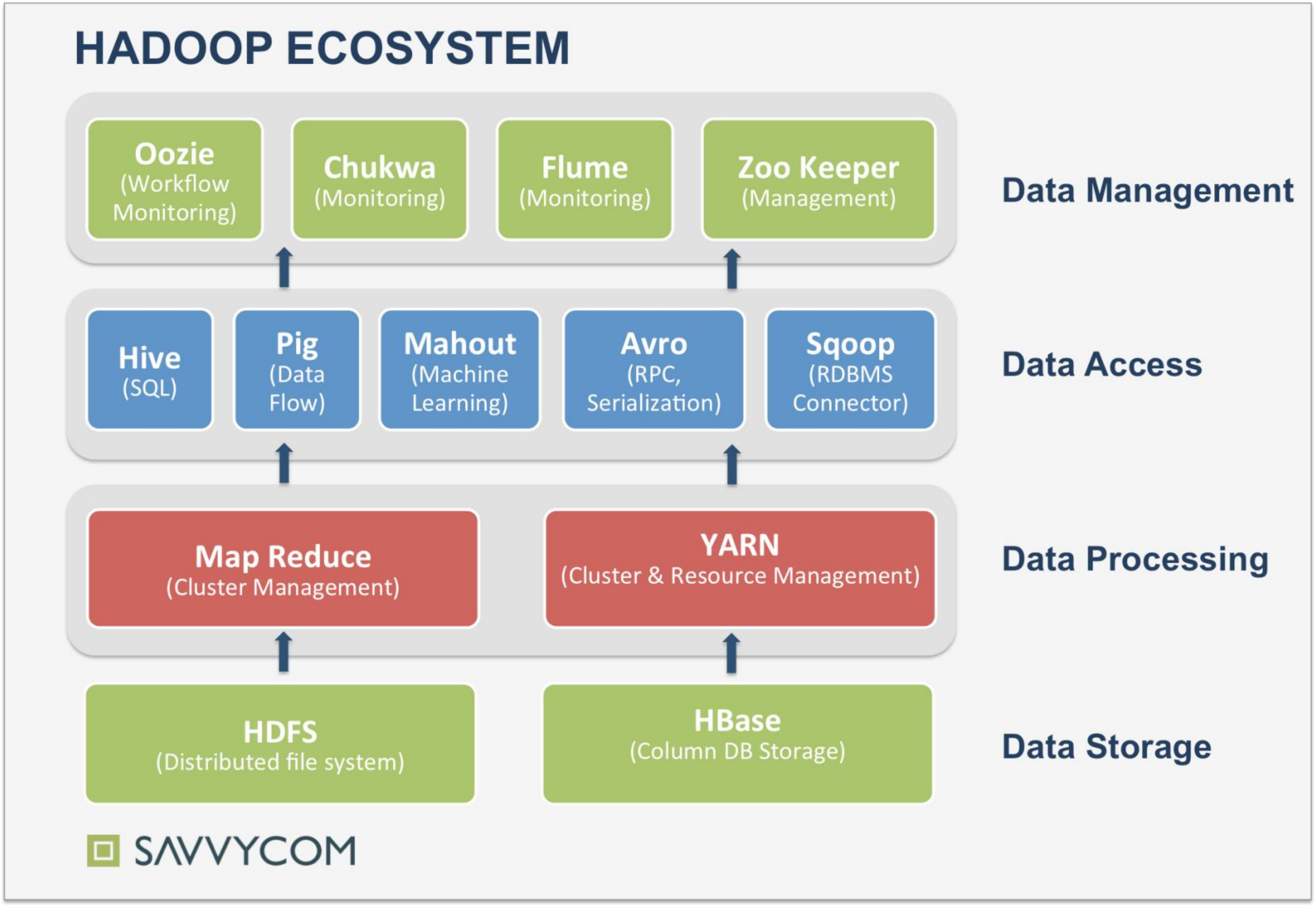
Unstructured/ Semi-structured Data

Sqoop

Structured Data

edureka!

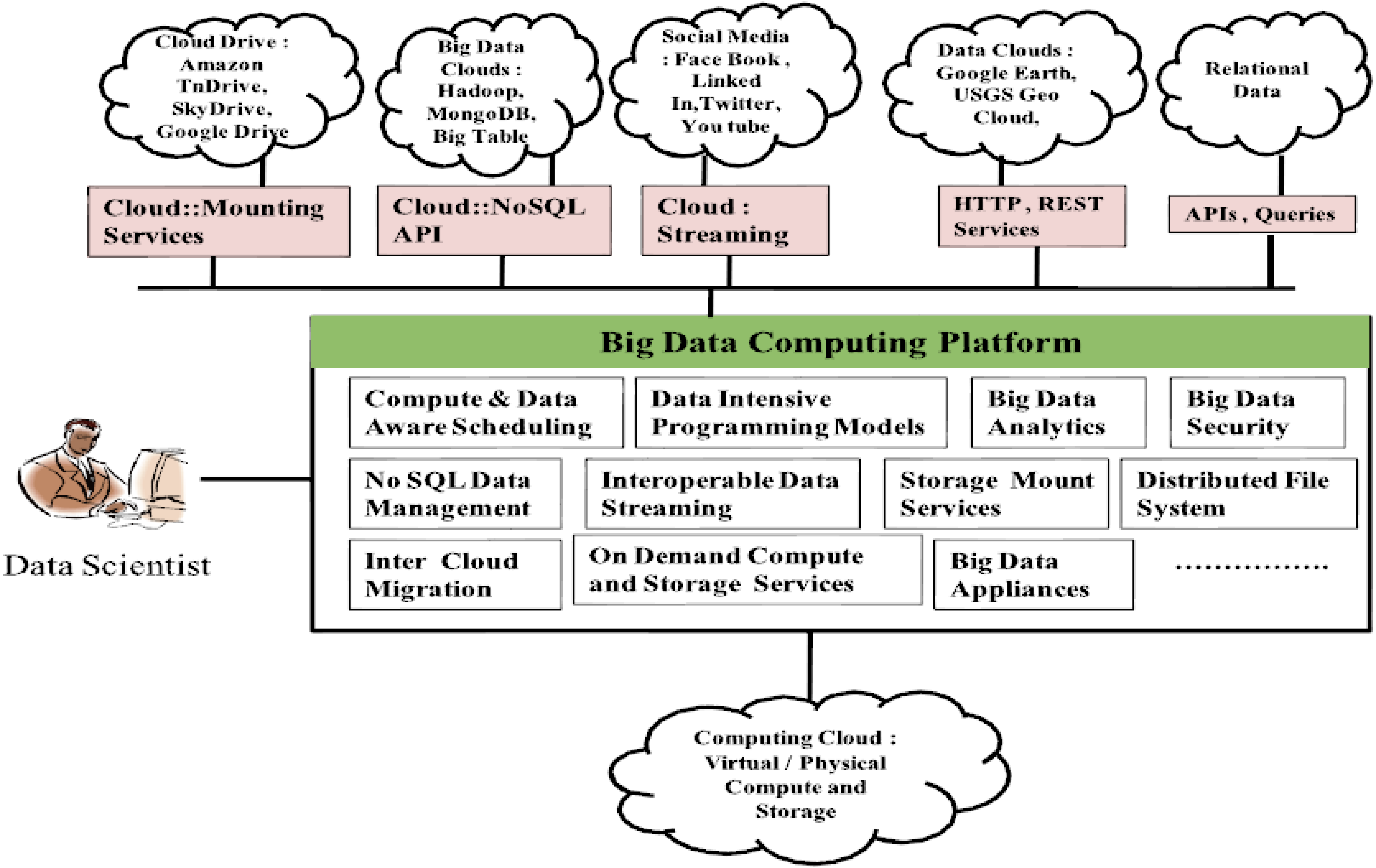
Hadoop Ecosystem


HADOOP ECOSYSTEM
Oozie (Workflow Monitoring)
Chukwa (Monitoring)
Flume (Monitoring)
Zoo Keeper (Management)
Data Management
Hive (SQL)
Pig (Data Flow)
Mahout (Machine Learning)
Avro (RPC, Serialization)
Sqoop (RDBMS Connector)
Data Access
Map Reduce (Cluster Management)
YARN (Cluster & Resource Management)
Data Processing
HDFS (Distributed file system)
HBase (Column DB Storage)
Data Storage
SAVVYCOM

Big Data Frameworks

- Open Sources
  - Apache Hadoop
  - Apache Spark
  - Apache Storm
  - Apache S4
- Commercials
  - Google Big Query
  - Amazon DynamoDB
  - Amazon Elastic MapReduce
  - Microsoft HDInsight Service
  - RackSpace Horton Hadoop on Openstack

Integrated Cloud & Big Data



Cloud Drive : Amazon TnDrive, SkyDrive, Google Drive

Big Data Clouds : Hadoop, MongoDB, Big Table

Social Media : Face Book , Linked In,Twitter, You tube

Data Clouds : Google Earth, USGS Geo Cloud,

Relational Data

Cloud::Mounting Services

Cloud::NoSQL API

Cloud : Streaming

HTTP , REST Services

APIs , Queries

Big Data Computing Platform

Compute & Data Aware Scheduling

Data Intensive Programming Models

Big Data Analytics

Big Data Security

No SQL Data Management

Interoperable Data Streaming

Storage Mount Services

Distributed File System

Inter Cloud Migration

On Demand Compute and Storage Services

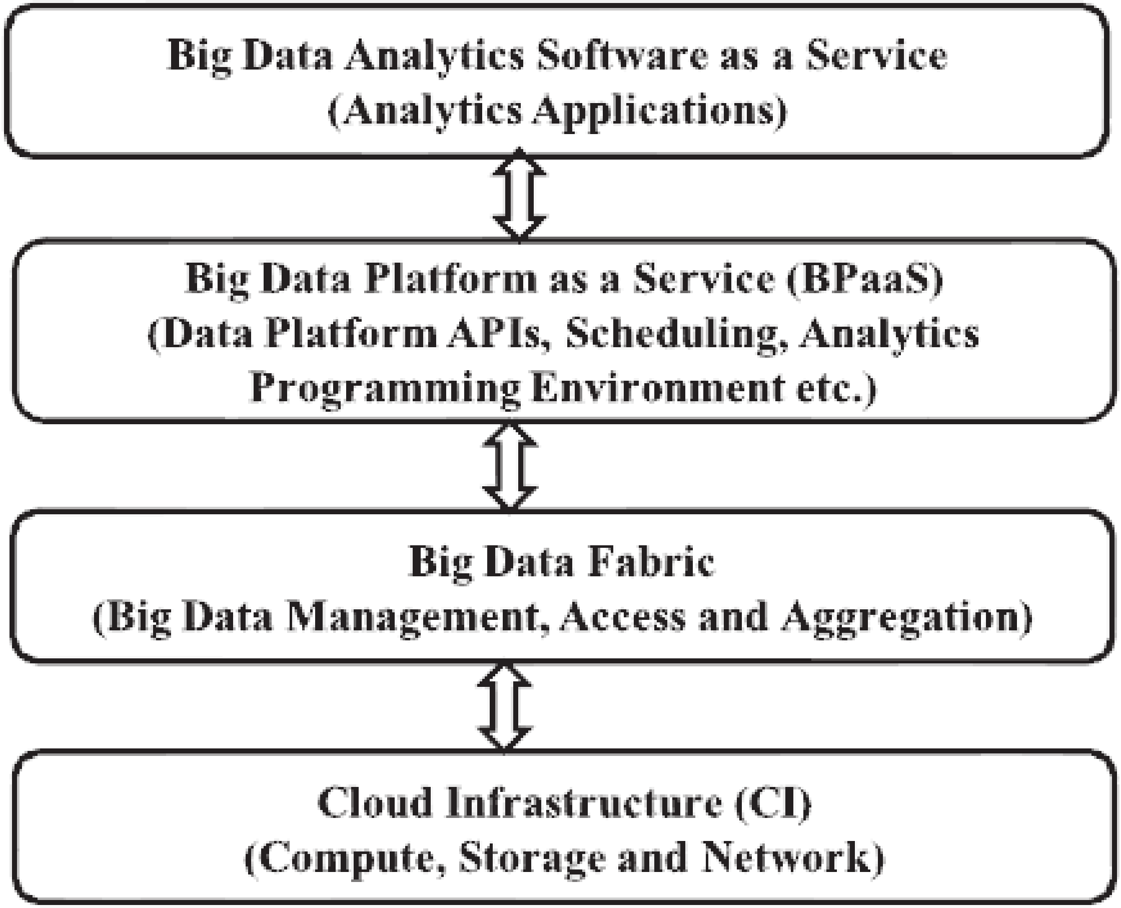
Big Data Appliances

……………

Data Scientist

Computing Cloud : Virtual / Physical Compute and Storage

Big Data Cloud Reference Architecture



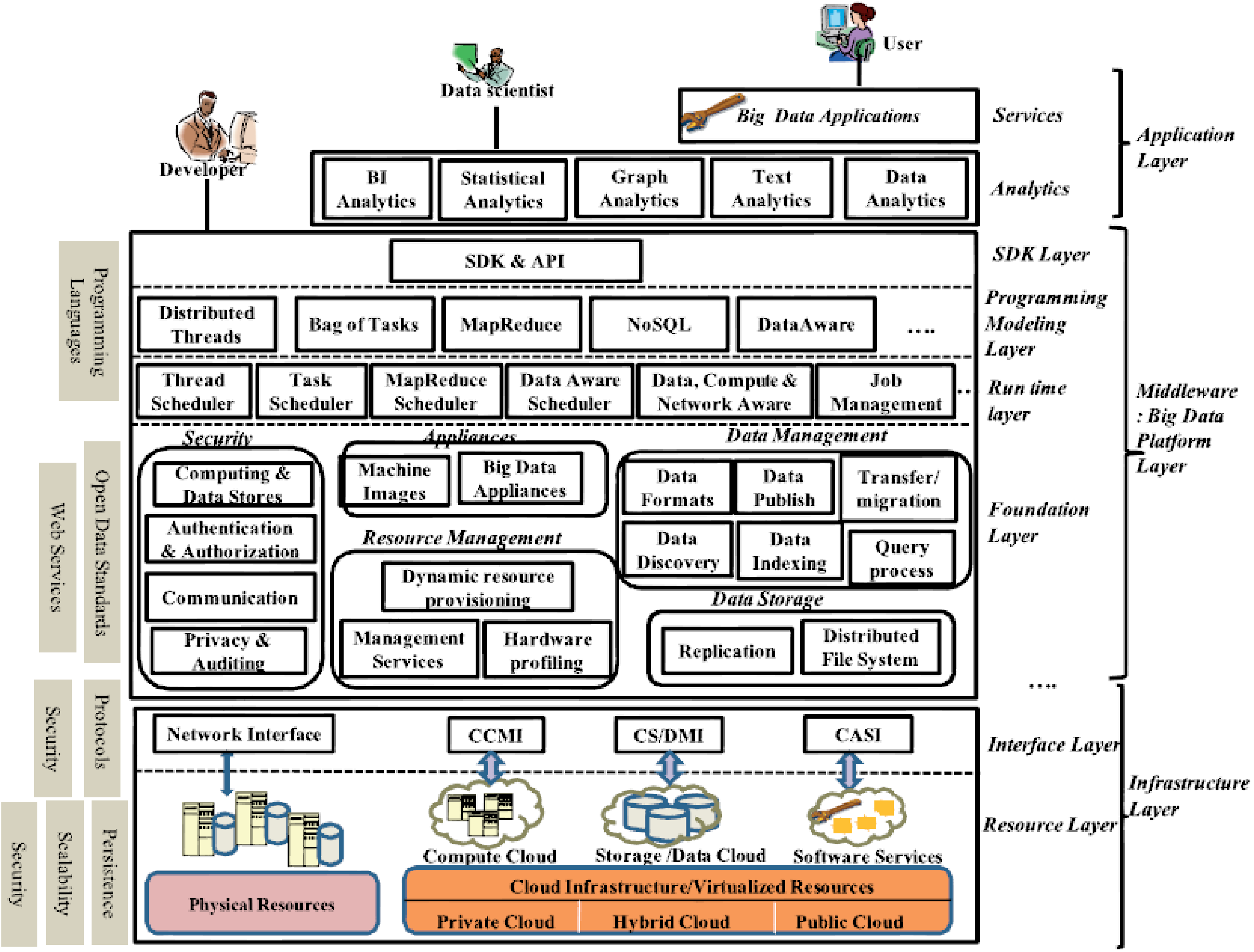
Big Data Analytics Software as a Service
(Analytics Applications)

Big Data Platform as a Service (BPaaS)
(Data Platform APIs, Scheduling, Analytics Programming Environment etc.)

Big Data Fabric
(Big Data Management, Access and Aggregation)

Cloud Infrastructure (CI)
(Compute, Storage and Network)

Big Data Cloud Layered Components



Thank
You
Merci
Gracias
TapadhLeat
Köszönöm
Murakoze
aDank
BarakAllahuFiik
Grazzi
Grazie
Shukriya
Buznyg
WaadMahadsantahay
Takk
Enkosi
Nouari
ThintKo
TapadhLeibh
FaafetaiLava
Ngiyabonga
M-Sapo
Bedankt
Zikomo
Mèsi
Chokrane
Kiitos
AsanteSana
Waita
Blagodaram
Rahmat
Dhanyavaad
Xièxie
Dziakuju
Mauruuru
GoRaibhMaithAgat
KyayTzuTinPaTe
Mochchakkeram
Trugarez
Aabhar
Matondo
Welalin
TangioTumas
Dhanyavad
TerimaKasih
Mercé
TesekkurEdirem
Mamnun
Tanan
Sukriya
TeşekkürEderim
Sobodi
AabhariAahe
Barkal
KurreSumanga
GratiasAgimus
Yekeniele
Dhanyabaad
Taiku
KeYaLeboha
Tanemirt
Vinaka
Tenki
Barka
Dankon
DankJe
KhobChaiDeu
Chnorakaloutioun
Mieskker
TänanVäga
Ačiū
KopKhunKrap
Tau
Nanni
Dankie
Motashakkeram
Dakujem
aNiKié
Toda
Paldies
DjiereDieuf
Gratzias
Trugéré
Akpé
Spacibo
Danke
Mersi
Nizzik ajr
Aguyjé
Diolch
EskerrikAsko
Dankewol
ANiKié
Misaotra
Tak
Sagolun
Dhanyavadagalu
Grandmercé
Dhanyavadalu
Sulpáy
Rahmet
GratiasAgo
Tanmirt
Grácies
Blagodaria
NajisTuke
Nandri
Hvala
Arigatô
Spas
Akiba
MaloTeOfa
Grâçie
Gracie
KopKhunKha
Merkzi
Akun
Wado
Asante
Faleminderit
Mahalo
Gracinas
KamSahHamnida
CámOn
Efharisto
DankJuWel
Marahaba
ÇoxSagOlun
Obrigado
Dìky
Saha
Multumesc
Meharbani
Nirringrazzjak
Bayarlalaa
TakkFyri
Salamat
Miigwetch