Using Big Data to Explore New Opportunities

Fandhy Haristha Siregar,
M.Kom, CIA, CRMA, CISA, CISM, CISSP, CEH, CEP-PM, QIA, COBIT5

Introduction to Big Data

The Myth About Big Data

- Big Data Is *New*
- Big Data Is Only About *Massive Data Volume*
- Big Data Means *Hadoop*
- Big Data Means *Unstructured Data*
- Big Data Is for *Social Media & Sentiment Analysis*

Source: Big Data: New Era of Analytic, Omer Sever, IBM SWG TR, Enterprise Content Manager

Big Data Is..

It is all about **better Analytic** on a **broader** spectrum of **data**, and therefore represents an **opportunity** to **create** even more **differentiation** among **industry peers.**

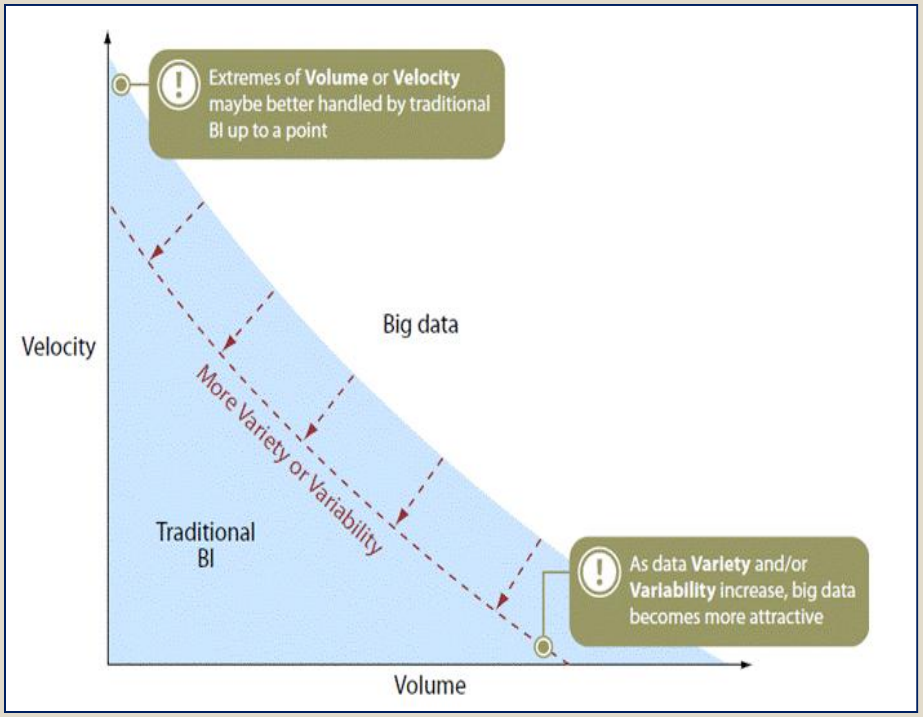
Source: Big Data: New Era of Analytic, Omer Sever, IBM SWG TR, Enterprise Content Manager

What's Big Data?

- **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 trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Source: Big Data: New Era of Analytic, Omer Sever, IBM SWG TR, Enterprise Content Manager

What is Big Data?



Volume

- Exceeds physical limits of vertical scalability

Velocity

- Decision window small compared to data change rate

Variety

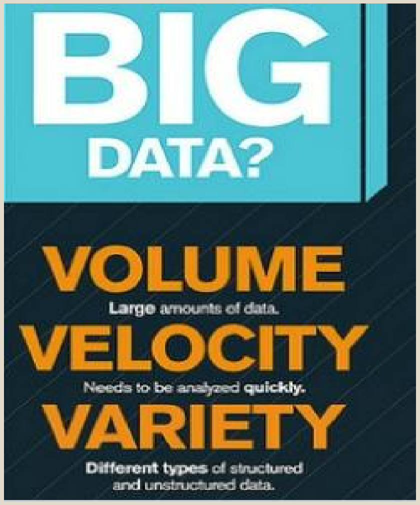
- Many different formats makes integration expensive

Variability

- Many options or variable interpretations confound analysis

Source: Ensuring Compliance of Patient Data with Big Data and BI, Ayad Shammout & Denny Lee, PASS Business Analytics Conference

Big Data: From 3V's to 4V's

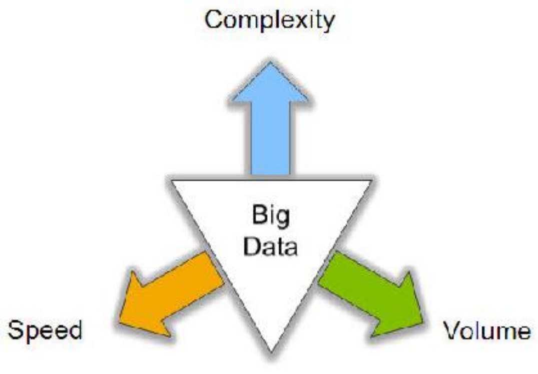


BIG DATA?

VOLUME
Large amounts of data.

VELOCITY
Needs to be analyzed quickly.

VARIETY
Different types of structured and unstructured data.

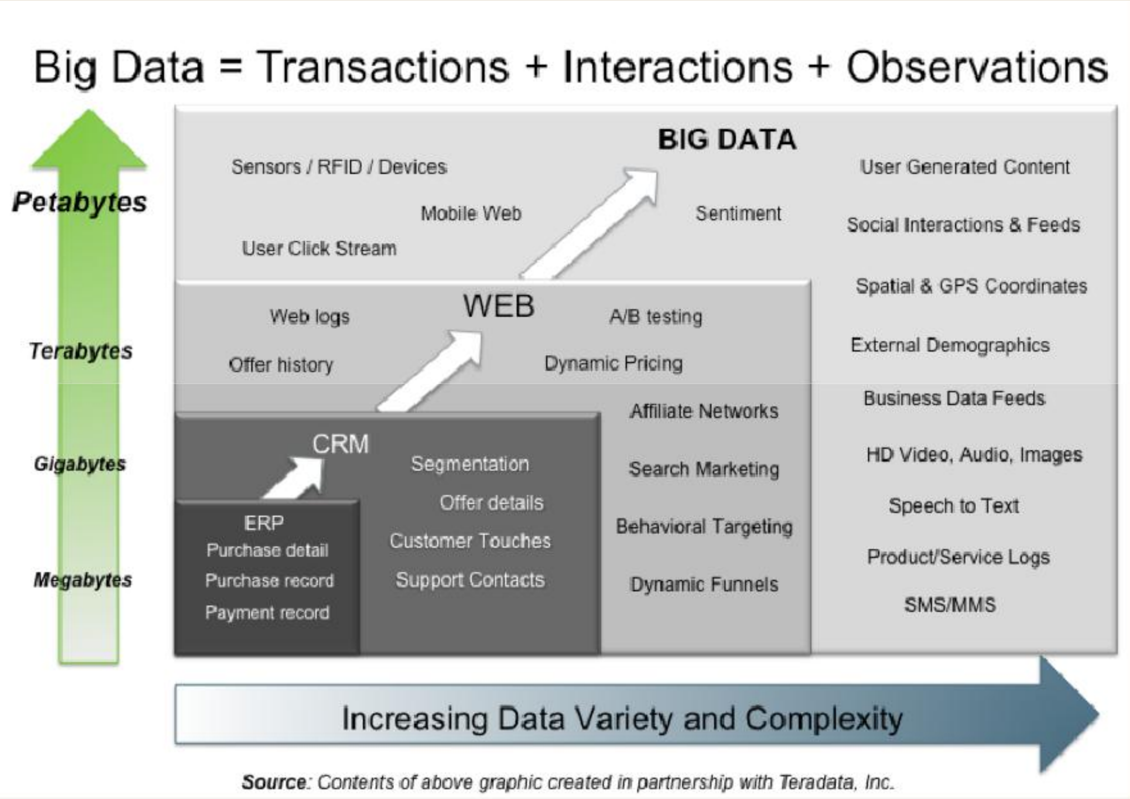


Complexity

Big Data

Speed

Volume



Big Data = Transactions + Interactions + Observations

Petabytes

Terabytes

Gigabytes

Megabytes

BIG DATA

Sensors / RFID / Devices

Mobile Web

Sentiment

User Click Stream

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

A/B testing

Offer history

Dynamic Pricing

Affiliate Networks

Search Marketing

Behavioral Targeting

Dynamic Funnels

CRM

Segmentation

Offer details

Customer Touches

Support Contacts

ERP

Purchase detail

Purchase record

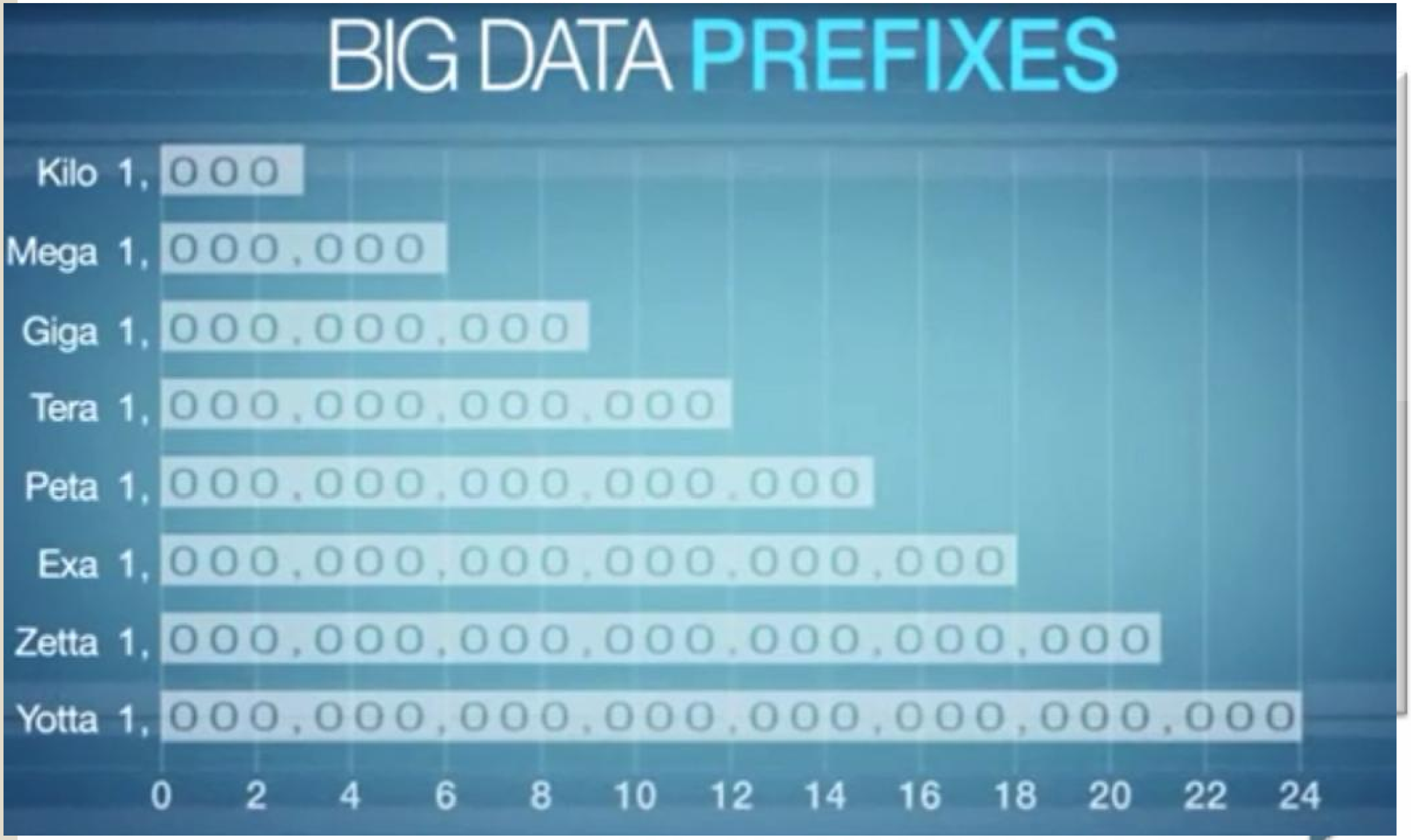
Payment record

Increasing Data Variety and Complexity

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

Source: Introduction to Big Data and Basic Data Analysis, Ruoming Jin, Kent University

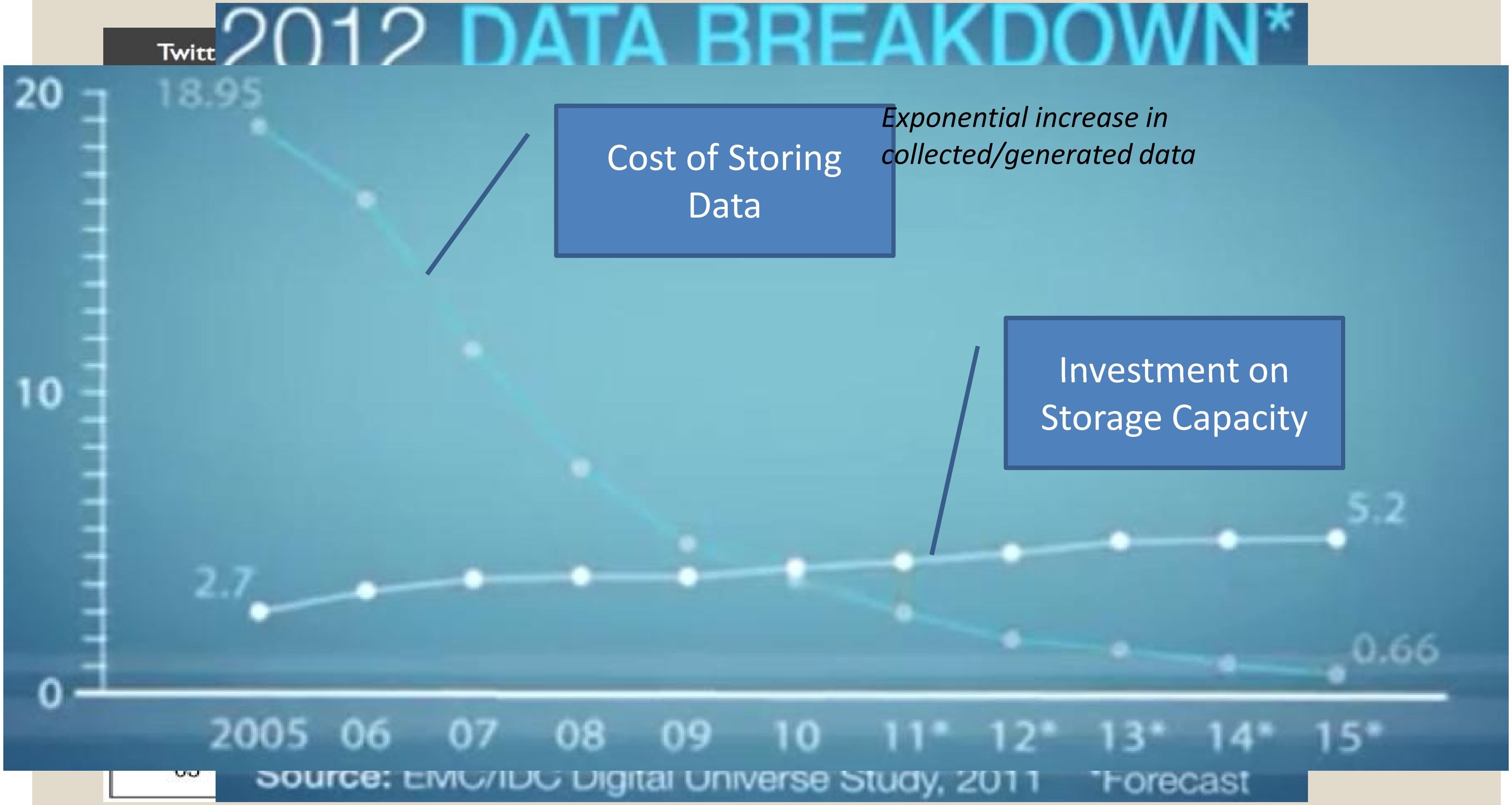
Volume (Scale)



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

Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

Volume (Scale)



Cost of Storing Data

Exponential increase in collected/generated data

Investment on Storage Capacity

Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

Where Is This “Big Data” Coming From ?

Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

? TBs of
data every day

12+ TBs
of tweet data
every day

25+ TBs of
log data every
day

30 billion RFID tags today
(1.3B in 2005)

4.6 billion
camera
phones
world wide

100s of
millions of
GPS enabled
devices sold
annually

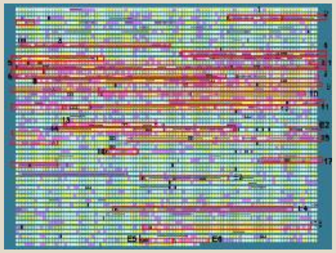
2+ billion
people on
the Web by
end 2011

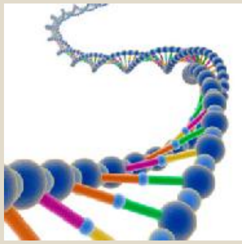
76 million smart meters
in 2009…
200M by 2014

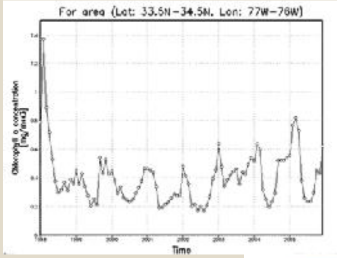
Variety (Complexity)

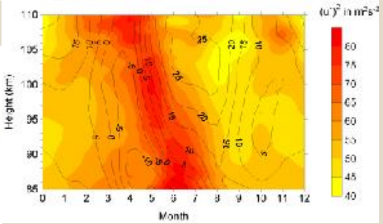
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online,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

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



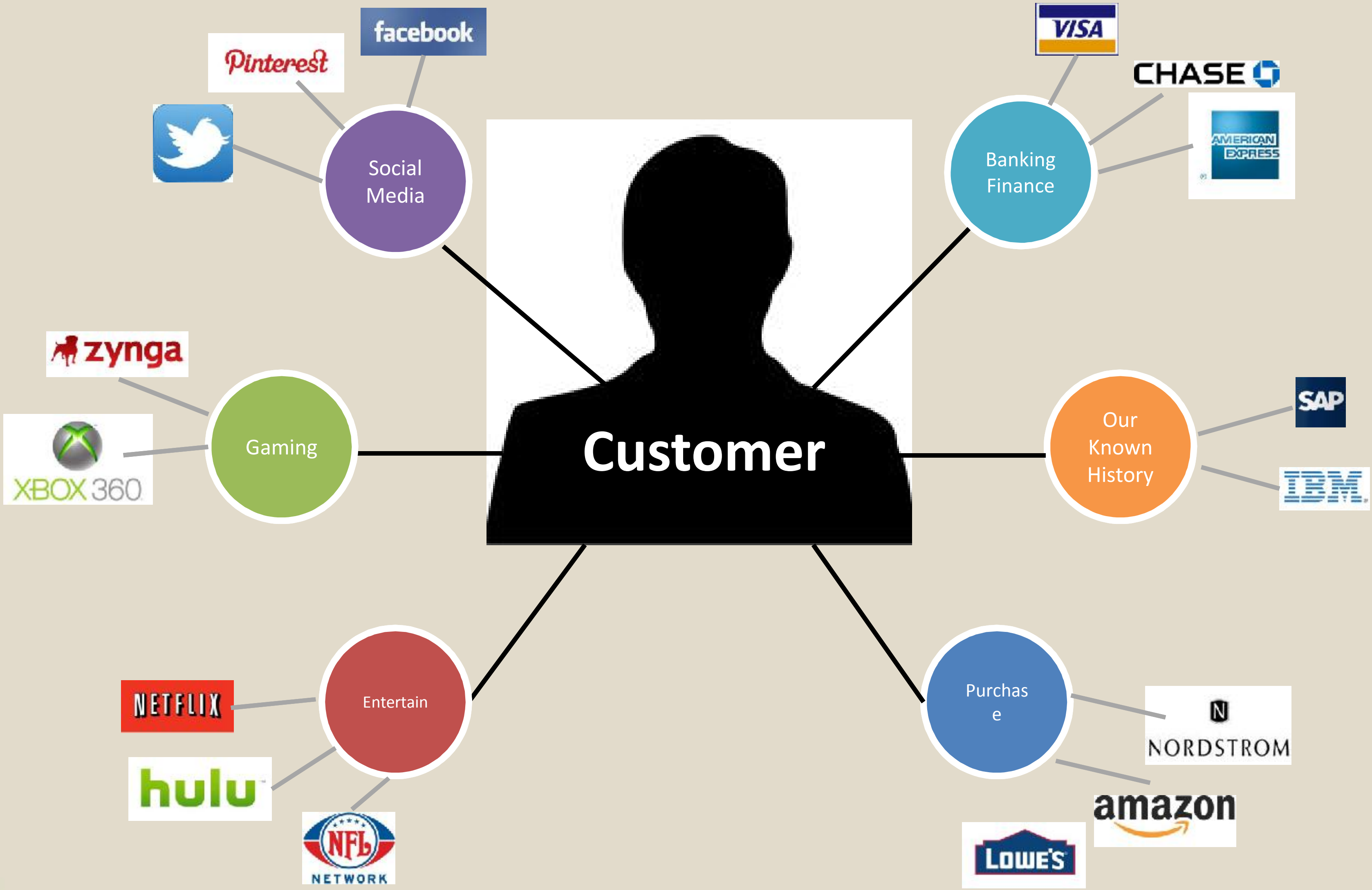



For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration
Time



Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

A Single View to the Customer



Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

Velocity (Speed)

- 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

Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

Real-time/Fast Data

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

Scientific instruments
(collecting all sorts of data)

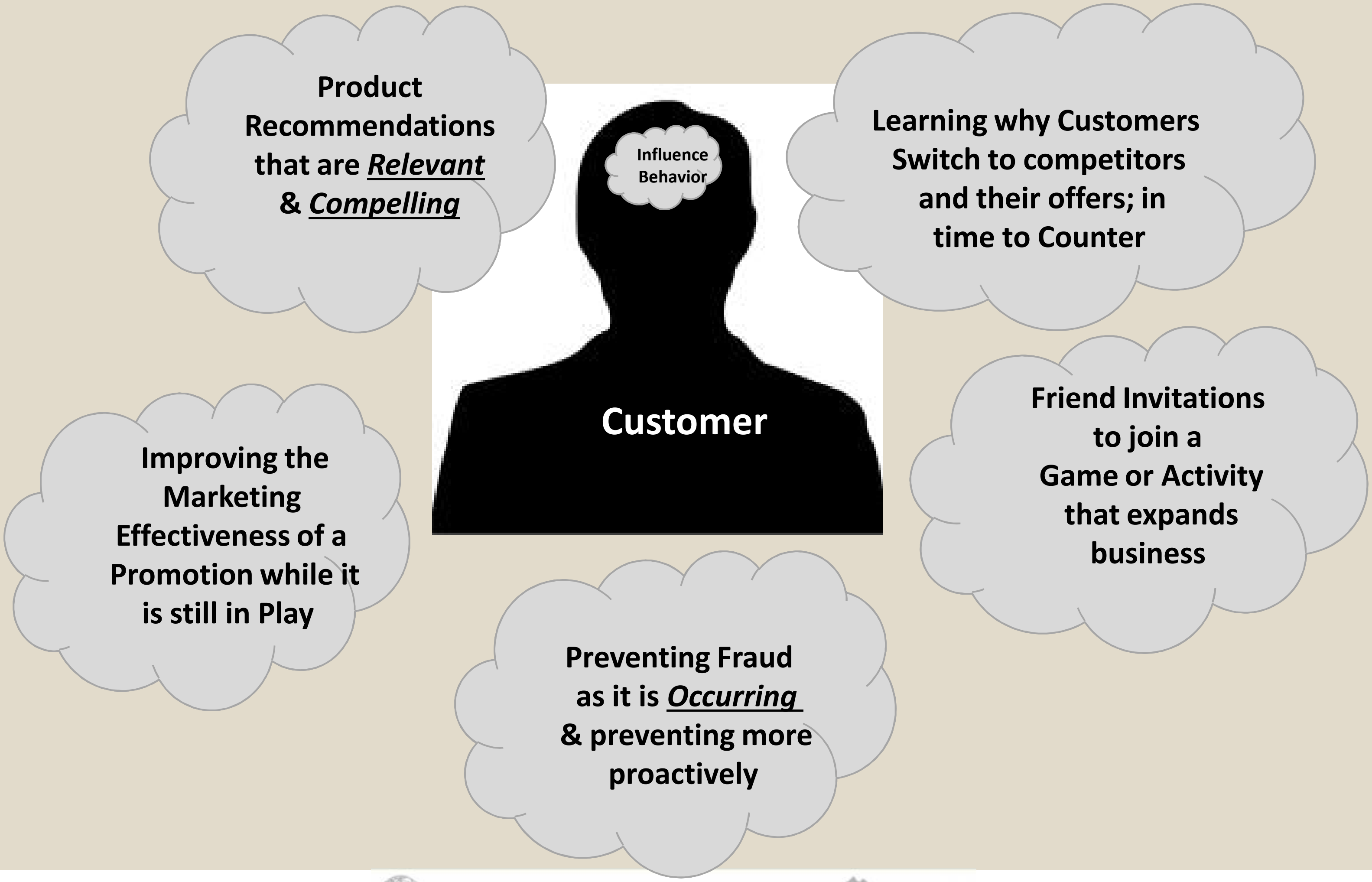
Mobile devices
(tracking all objects all the time)

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

Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

Real-Time Analytics/Decision Requirement



Customer

Influence Behavior

Product Recommendations that are ***Relevant*** & ***Compelling***

Learning why Customers Switch to competitors and their offers; in time to Counter

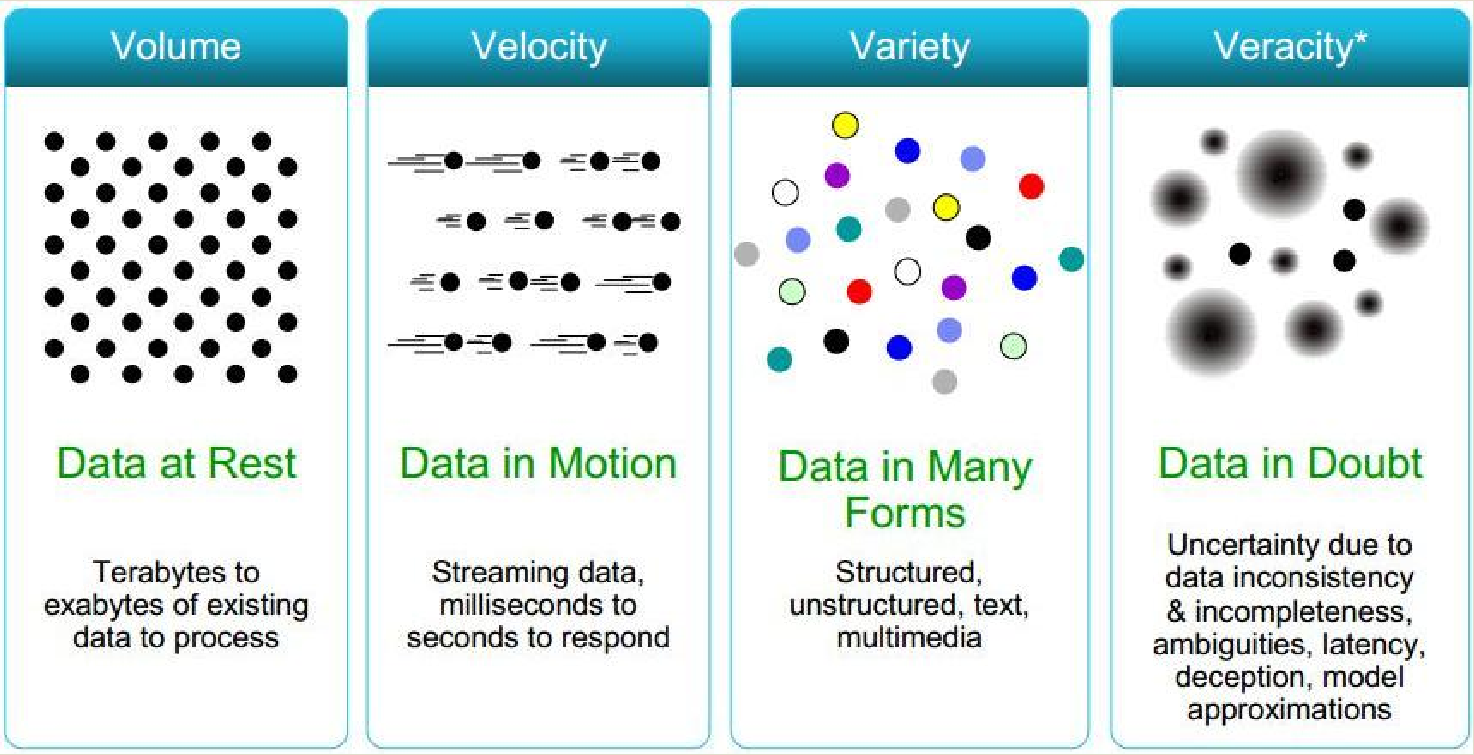
Improving the Marketing Effectiveness of a Promotion while it is still in Play

Friend Invitations to join a Game or Activity that expands business

Preventing Fraud as it is ***Occurring*** & preventing more proactively

Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

Some Make it 4V's



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

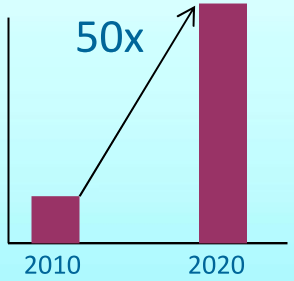
Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Source: Big Data in Engineering Applications, Jasti Aswini

Summary: Four Characteristics of Big Data

Cost efficiently processing the growing **Volume**

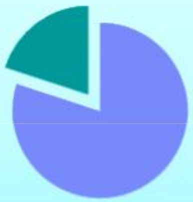


35 ZB

Responding to the increasing **Velocity**

30 Billion RFID sensors and counting

Collectively Analyzing the broadening **Variety**



80% of the worlds data is unstructured

Establishing the **Veracity** of big data sources

1 in 3 business leaders don’t trust the information they use to make decisions

Source: Big Data: New Era of Analytic, Omer Sever, IBM SWG TR, Enterprise Content Manager

Big Data:

Big Opportunity, Big Challenge

Data explosion

10x increase every five years

85% from new data types

Volume
Velocity
Variety

Hadoop

Cheap, Distributed Storage & Processing

Cloud

Easy Accessibility of External Data

> By 2015, organizations that build a modern information management system will outperform their peers financially by 20 percent.
>
> – Gartner, Mark Beyer
> “Information Management in the 21st Century”

Source: Ensuring Compliance of Patient Data with Big Data and BI, Ayad Shammout & Denny Lee, PASS Business Analytics Conference

The Big Potential Opportunities of Big Data

- **Bollen, Mao, and Zeng (2011) -** use Twitter data to predict daily fluctuations of the Dow Jones Industrial Average (DJIA). Google's Profile of Mood States and OpinionFinder (Wilson et al. 2005) measure public mood based on 10 million public tweets.
- **Chan (2003) and Mittermayer (2004) -** use news articles to predict stock price movements. Social and Mass Media data can be utilized to help measure financial health of a firm, and better evaluate the audit engagement
- **Mofitt and Vasarhelyi (2013) -** propose to use news, audio and video streams, cell phone recordings, social media comments to obtain new forms of audit evidence, confirm existence of events, and validate reporting elements.

Source:Big Data Analytics in Financial Statement Audits, Min Cao, Roman Chychyla,Trevor Stewart, 2015

Big Data Business Value

15 out of 17

sectors in the US have more data stored per company than the US Library of Congress

140,000-190,000

more deep analytical talent positions

1.5 million

more data savvy managers in the US alone

€250 billion

Potential annual value to Europe's public sector

50-60%

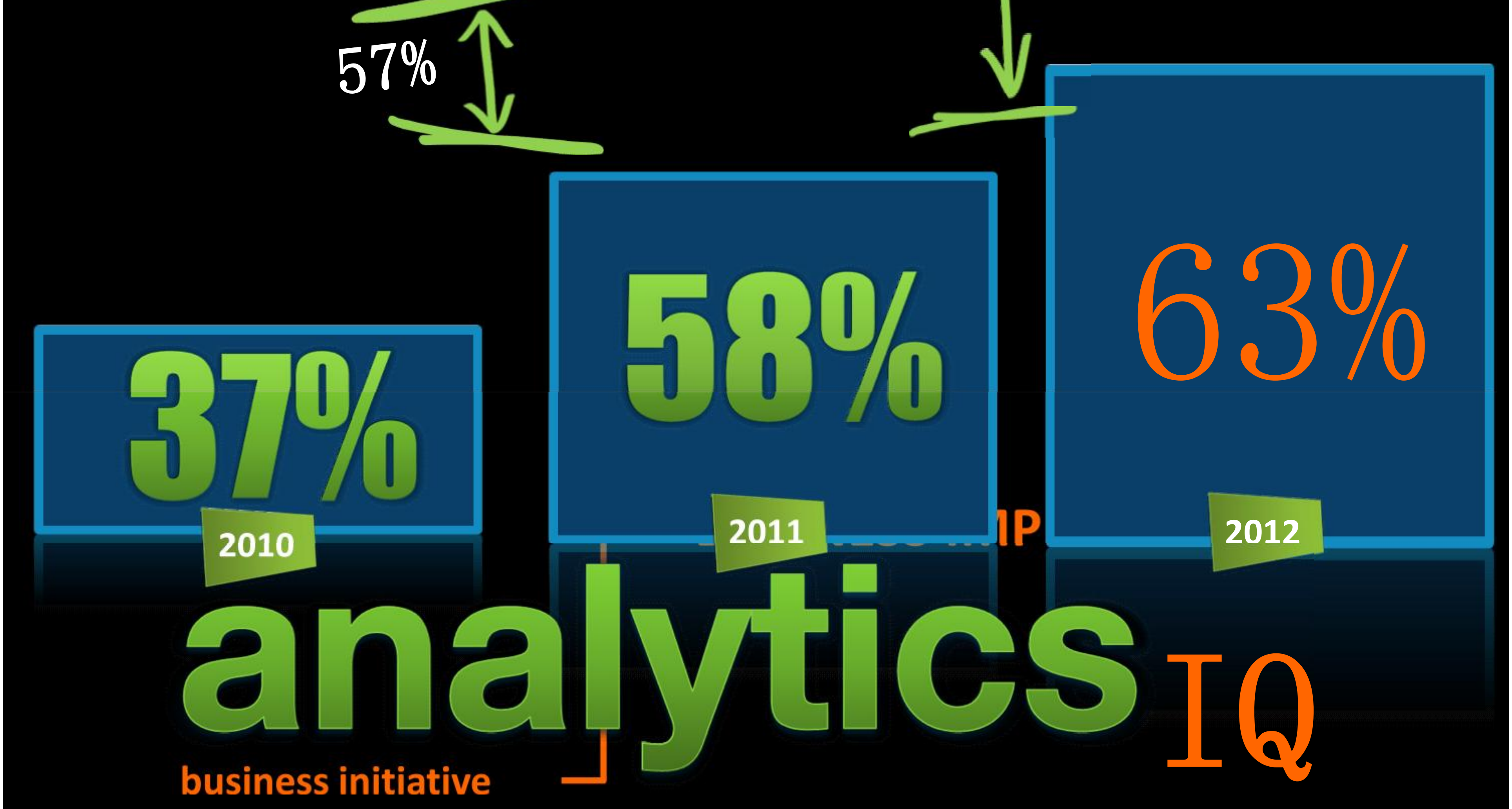
increase in the number of Hadoop developers within organizations already using Hadoop within a year

$300 billion

Potential annual value to US healthcare

Source: Ensuring Compliance of Patient Data with Big Data and BI, Ayad Shammout & Denny Lee, PASS Business Analytics Conference

The number of organizations who see analytics as a competitive advantage is growing.



Source: Big Data: New Era of Analytic, Omer Sever, IBM SWG TR, Enterprise Content Manager

Studies show that organizations competing on analytics outperform their peers

substantially **outperform**

220%

1.6x

Revenue Growth

2.5x

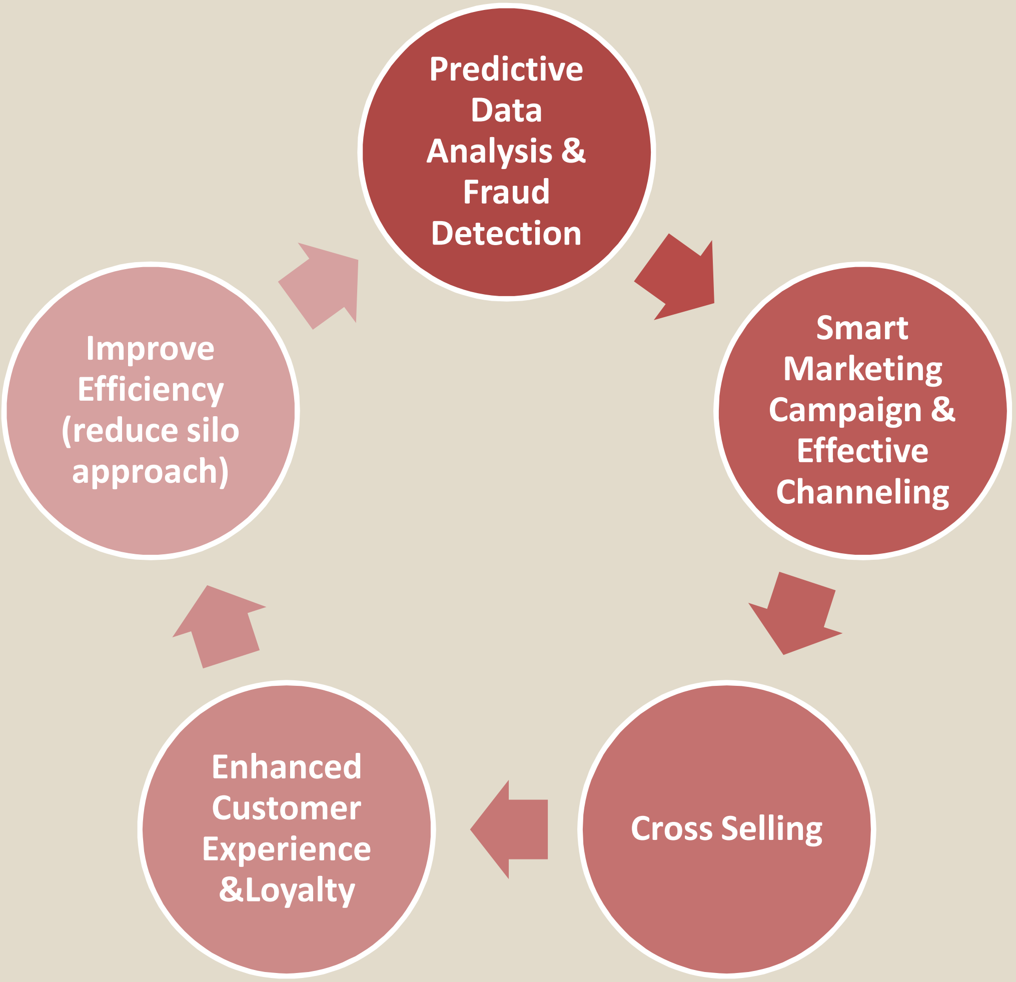
Stock Price Appreciation

2.0x

EBITDA Growth

Source: Big Data: New Era of Analytic, Omer Sever, IBM SWG TR, Enterprise Content Manager

The Big Opportunities for Financial Industry



Predictive Data Analysis & Fraud Detection

Smart Marketing Campaign & Effective Channeling

Cross Selling

Enhanced Customer Experience &Loyalty

Improve Efficiency (reduce silo approach)

The Big Challenge of Big Data

McKinsey & Co. recently reported that two-thirds of C-suite executives surveyed consider big data to be a top strategic priority.

McKinsey survey, however, noted a big gap between what organizations want to do with big data and their capabilities to do so, given their existing IT infrastructure and expertise.

Source: Big Data, Collect It, Respect It, IIA Tone at the Top, Aug 2013

The Big Challenge for Internal Auditor

- The Los Angeles Police Department analyzes data from crime scenes, including time, location, nature, and actors in order to predict the most likely timing and location of crimes on that d

- Similar analyt and direct audit effort aimed a

- Population vs Integration with continuous au

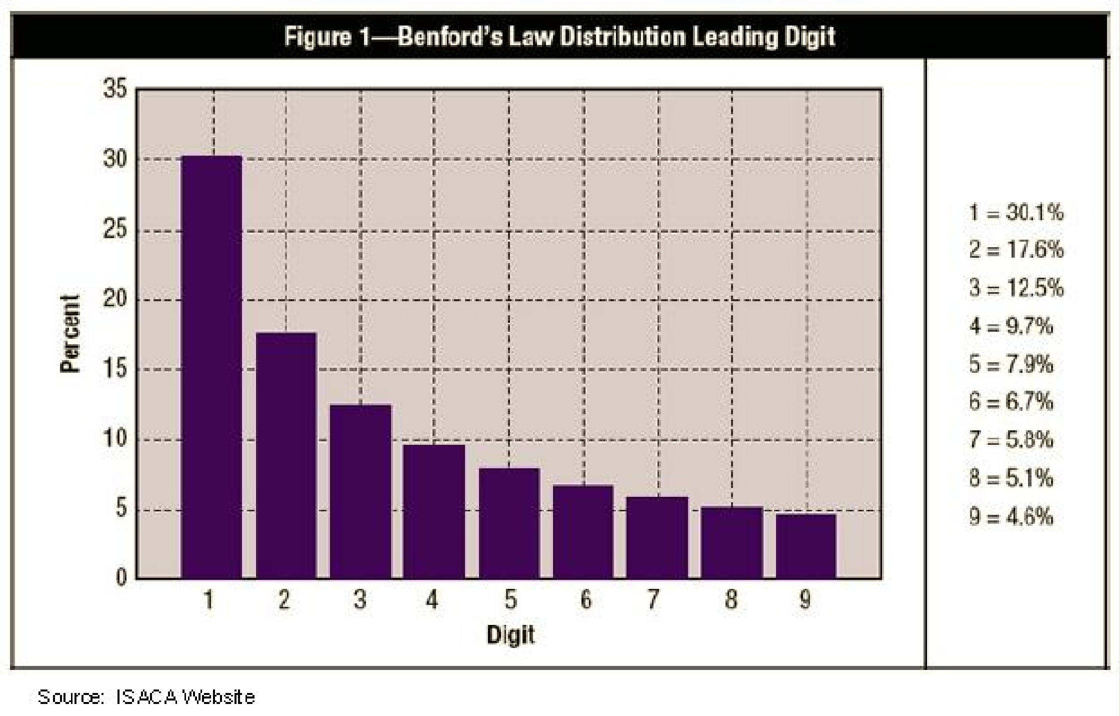
- The IIA's *Inter* nplications of big data in its Feb ticle, author Russell Jacks uld address when their org ance **with applicable priv ions, and data destruction policies**.

> **"The audit tool kit looks the same as it did 50 or 60 years ago," he says. "If we were doctors, that would be pretty frightening. This has tremendous potential, but it's still early. We're still experimenting."**
>
> **Dorsey Baskin, Grant Thornton**

Source:Big Data Analytics in Financial Statement Audits, Min Cao, Roman Chychyla,Trevor Stewart, 2015

Benford Law

- Mathematical theory of leading digits. Leading digits are distributed in aspecific, non-uniform way.
- Simon Newcomb, 1881: Described theoretical frequency that is Benford’s Law
- Frank Benford, 1938: Numbers starting with 1, 2, or 3 are more common in nature than those with initial digits 4 – 9.

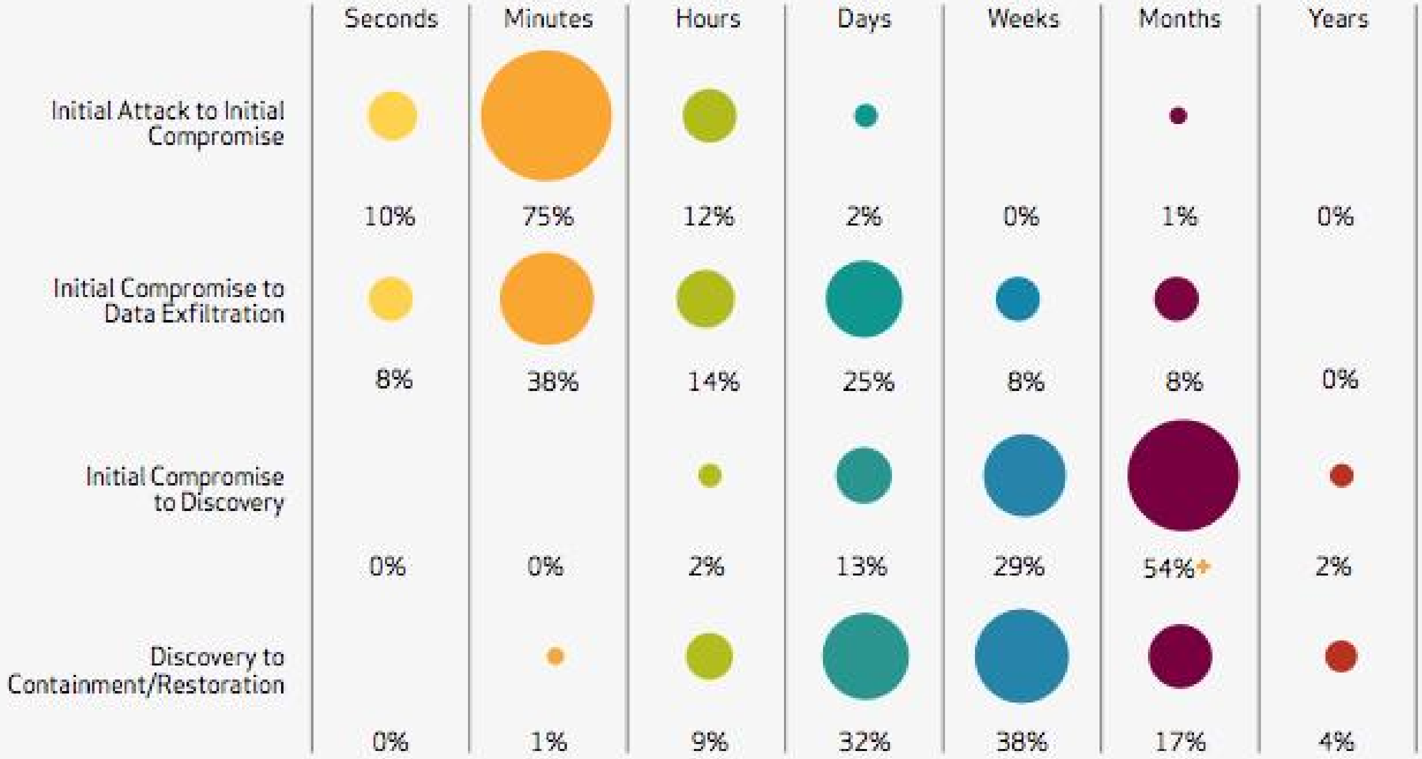


Source: ISACA Website

Source: An Auditors Guide to Data Analysis, Natasha DeKroon, Duke University

Velocity of Data Generation vs Fraud/Breach Detection

Figure 40. Timespan of events by percent of breaches

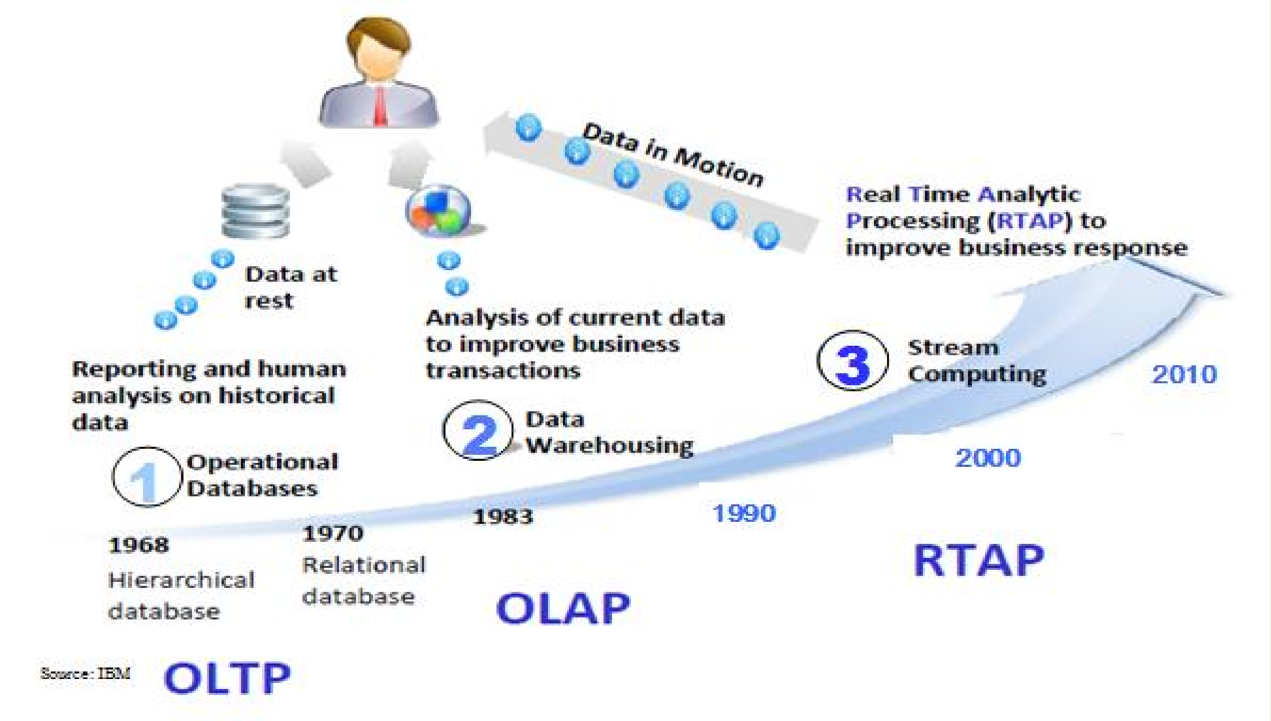


	Seconds	Minutes	Hours	Days	Weeks	Months	Years
Initial Attack to Initial Compromise	10%	75%	12%	2%	0%	1%	0%
Initial Compromise to Data Exfiltration	8%	38%	14%	25%	8%	8%	0%
Initial Compromise to Discovery	0%	0%	2%	13%	29%	54%+	2%
Discovery to Containment/Restoration	0%	1%	9%	32%	38%	17%	4%

Source: Verizon Data Breach Investigation Report 2013

Development of Big Data

Harnessing Big Data



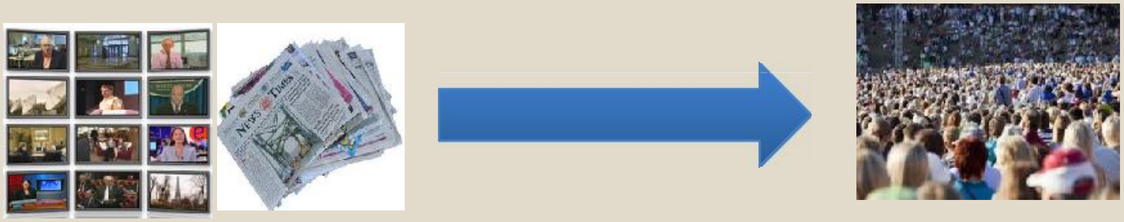
Source: Big Data in Engineering Applications, Jasti Aswini

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

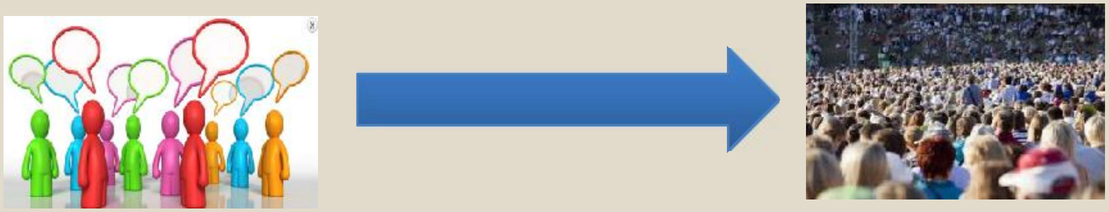
The Model Has Changed…

The Model of Generating/Consuming Data has Changed

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

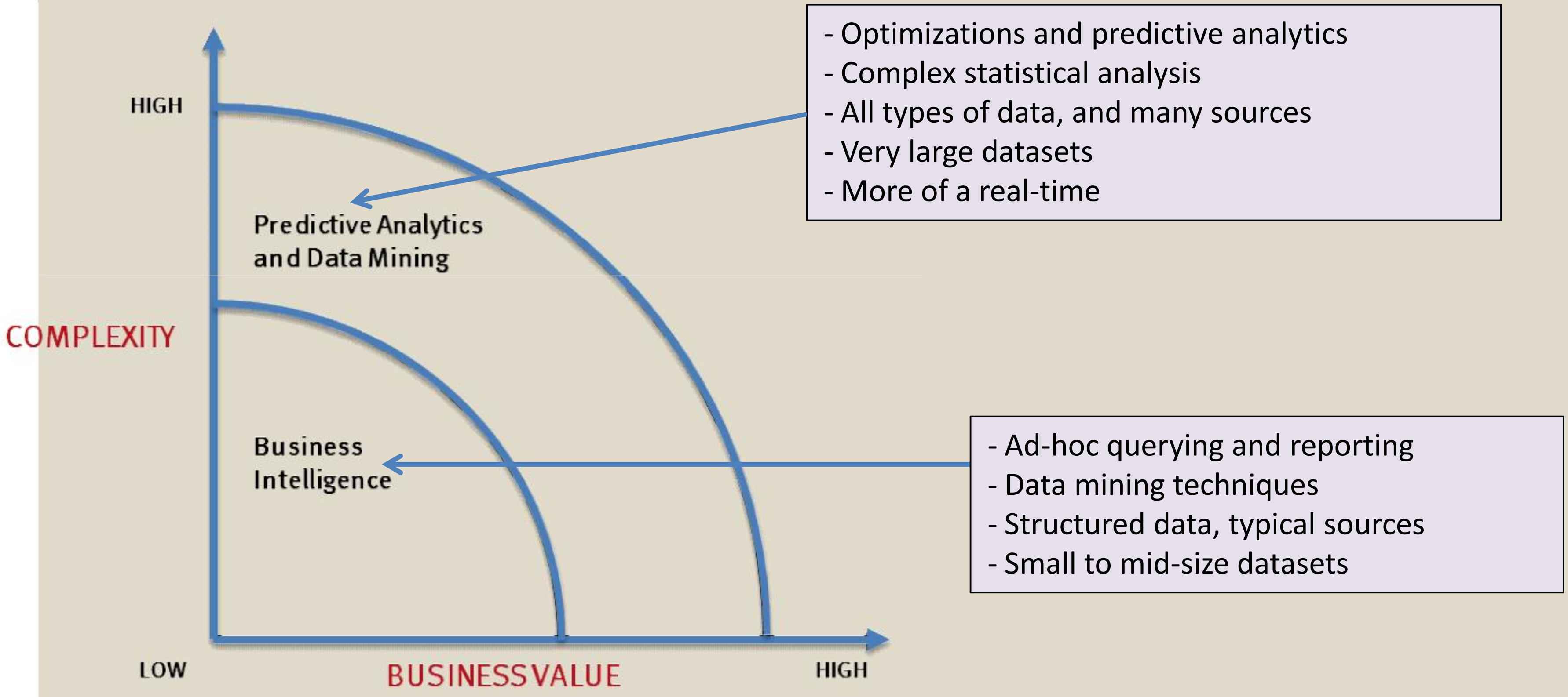


New Model: all of us are generating data, and all of us are consuming data



Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

What’s driving Big Data



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

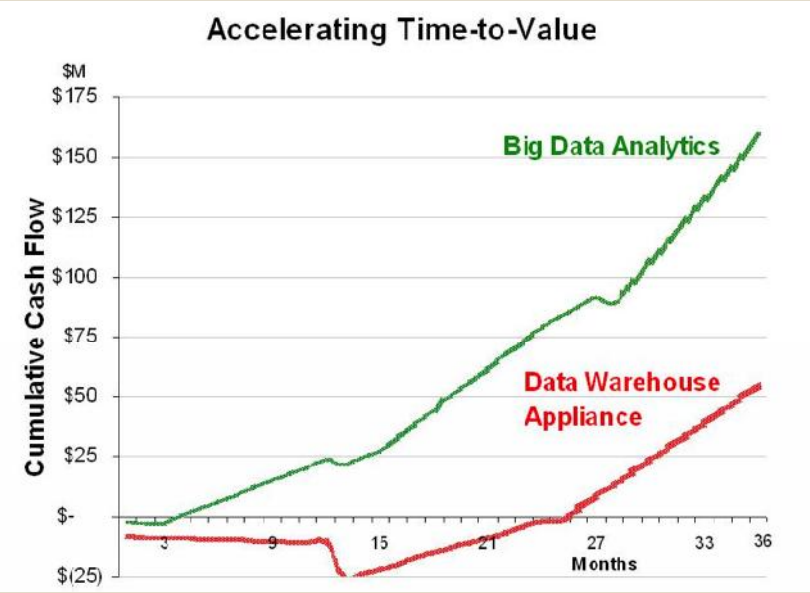
Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

Big Data Implementation

From CAATTs, Open Source Project to Full Package Commercial

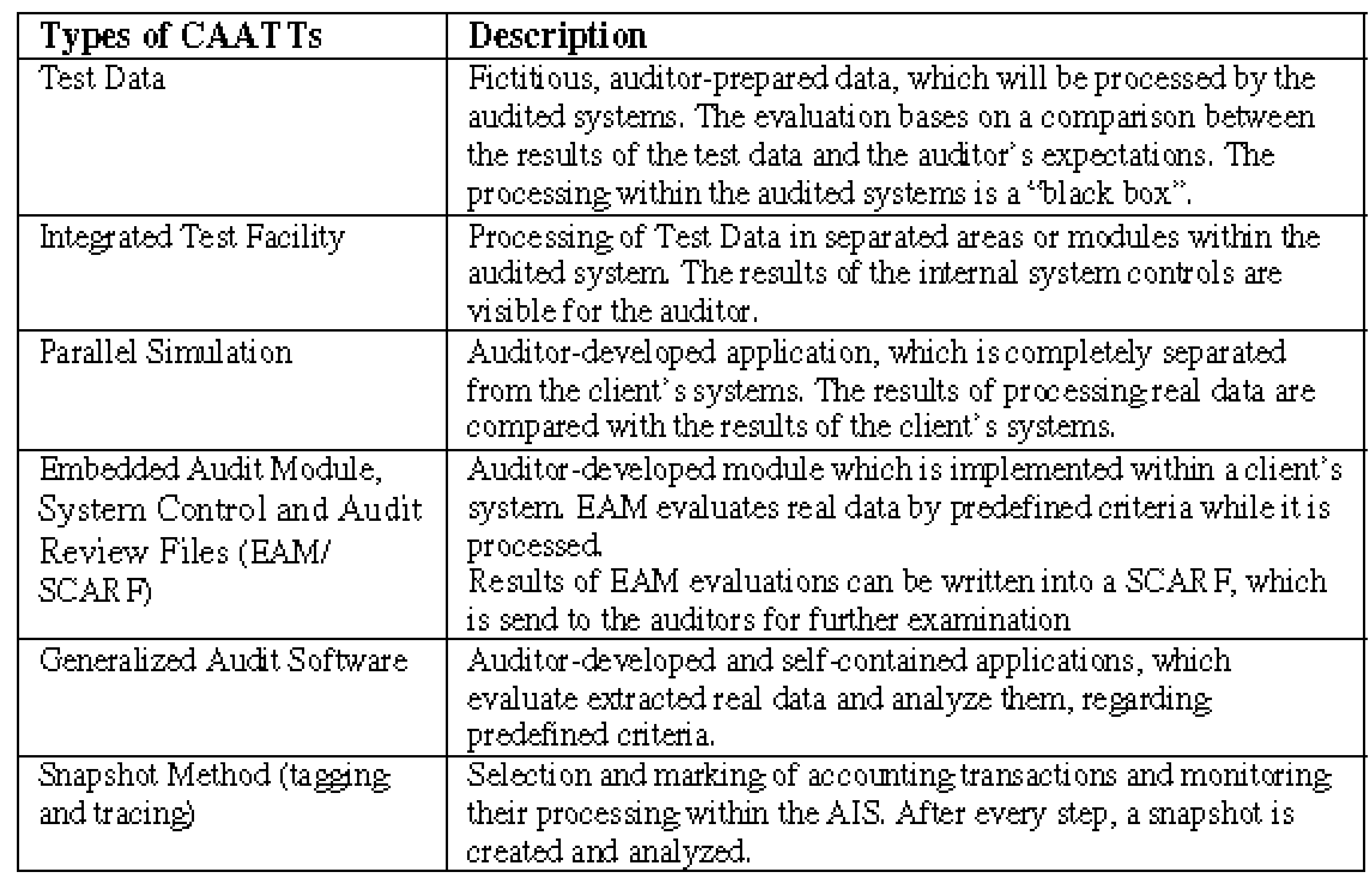
Big Data Analytics

- 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



Source: Introduction to Big Data and Basic Data, Analysis, Ruoming Jin, Kent University

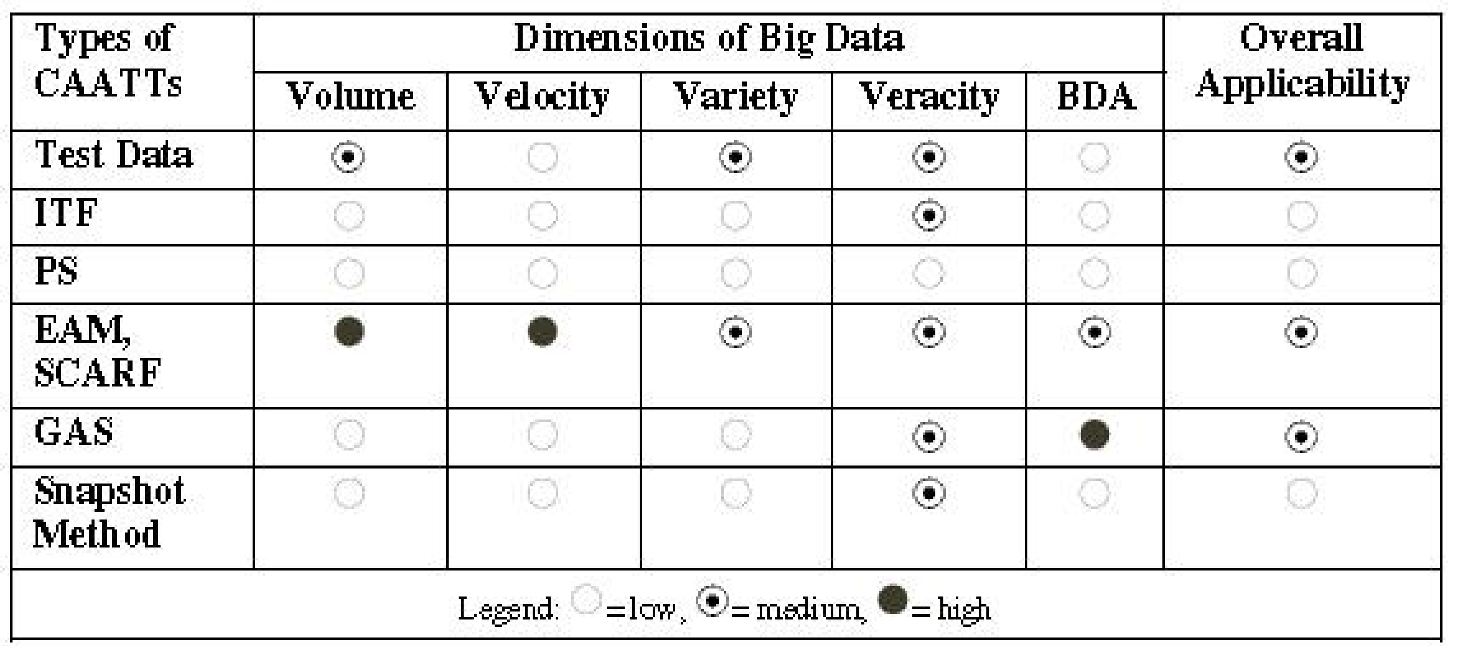
Big Data & CAATTs



Types of CAATTs	Description
Test Data	Fictitious, auditor-prepared data, which will be processed by the audited systems. The evaluation bases on a comparison between the results of the test data and the auditor's expectations. The processing within the audited systems is a "black box".
Integrated Test Facility	Processing of Test Data in separated areas or modules within the audited system. The results of the internal system controls are visible for the auditor.
Parallel Simulation	Auditor-developed application, which is completely separated from the client's systems. The results of processing real data are compared with the results of the client's systems.
Embedded Audit Module, System Control and Audit Review Files (EAM/ SCARF)	Auditor-developed module which is implemented within a client's system. EAM evaluates real data by predefined criteria while it is processed. Results of EAM evaluations can be written into a SCARF, which is send to the auditors for further examination
Generalized Audit Software	Auditor-developed and self-contained applications, which evaluate extracted real data and analyze them, regarding predefined criteria.
Snapshot Method (tagging and tracing)	Selection and marking of accounting transactions and monitoring their processing within the AIS. After every step, a snapshot is created and analyzed.

Source: Continuous Auditing in Big Data Computing Environments, Andreas Kiezow, University of Osnabruck

Big Data & CAATTs

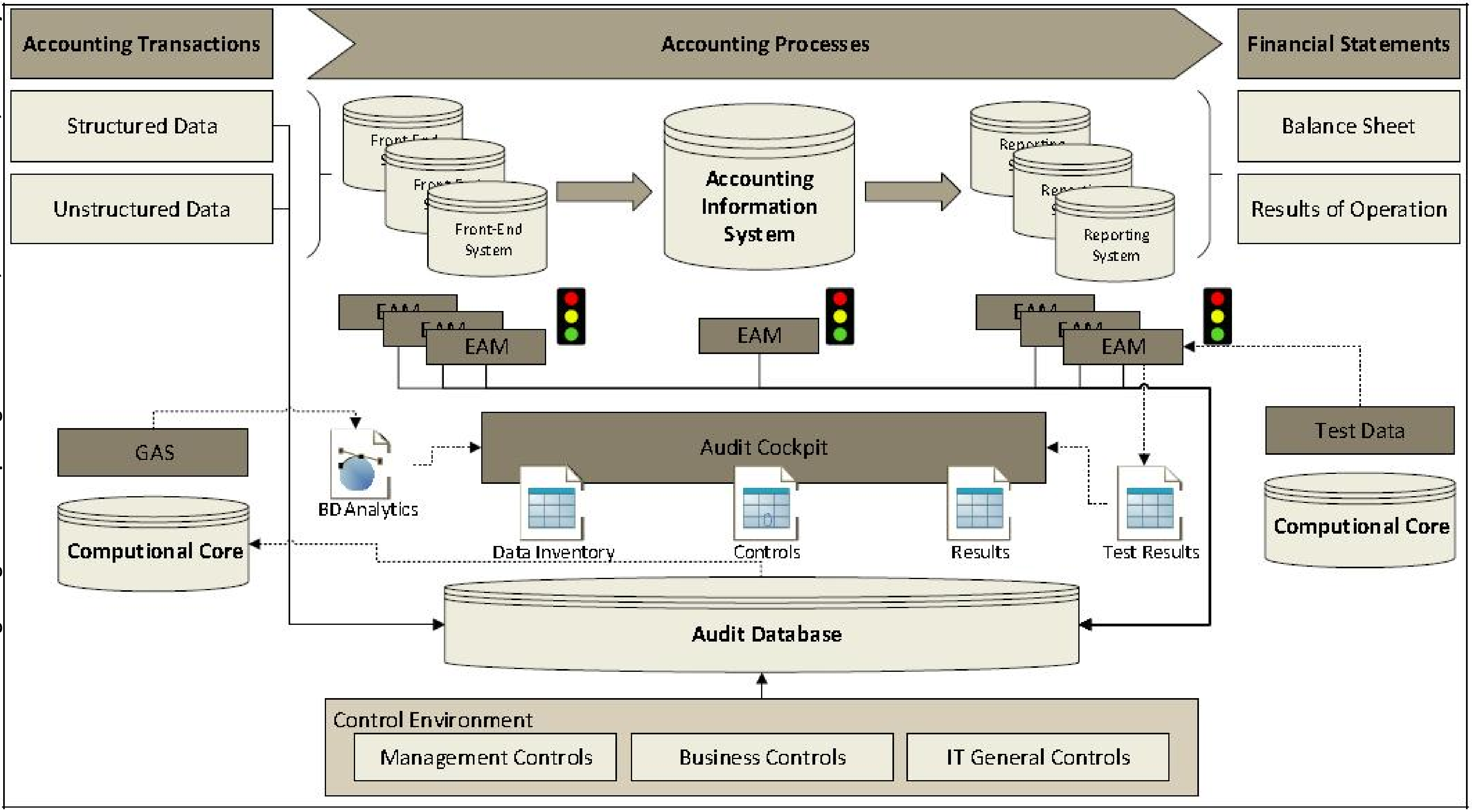


Types of CAATTs	Dimensions of Big Data					Overall Applicability
	Volume	Velocity	Variety	Veracity	BDA	
Test Data	⊙	○	⊙	⊙	○	⊙
ITF	○	○	○	⊙	○	○
PS	○	○	○	○	○	○
EAM, SCARF	●	●	⊙	⊙	⊙	⊙
GAS	○	○	○	⊙	●	⊙
Snapshot Method	○	○	○	⊙	○	○

Legend: ○ =low, ⊙= medium, ●= high

Source: Continuous Auditing in Big Data Computing Environments, Andreas Kiezow, University of Osnabruck

Embedded Audit Module - On-going Monitoring

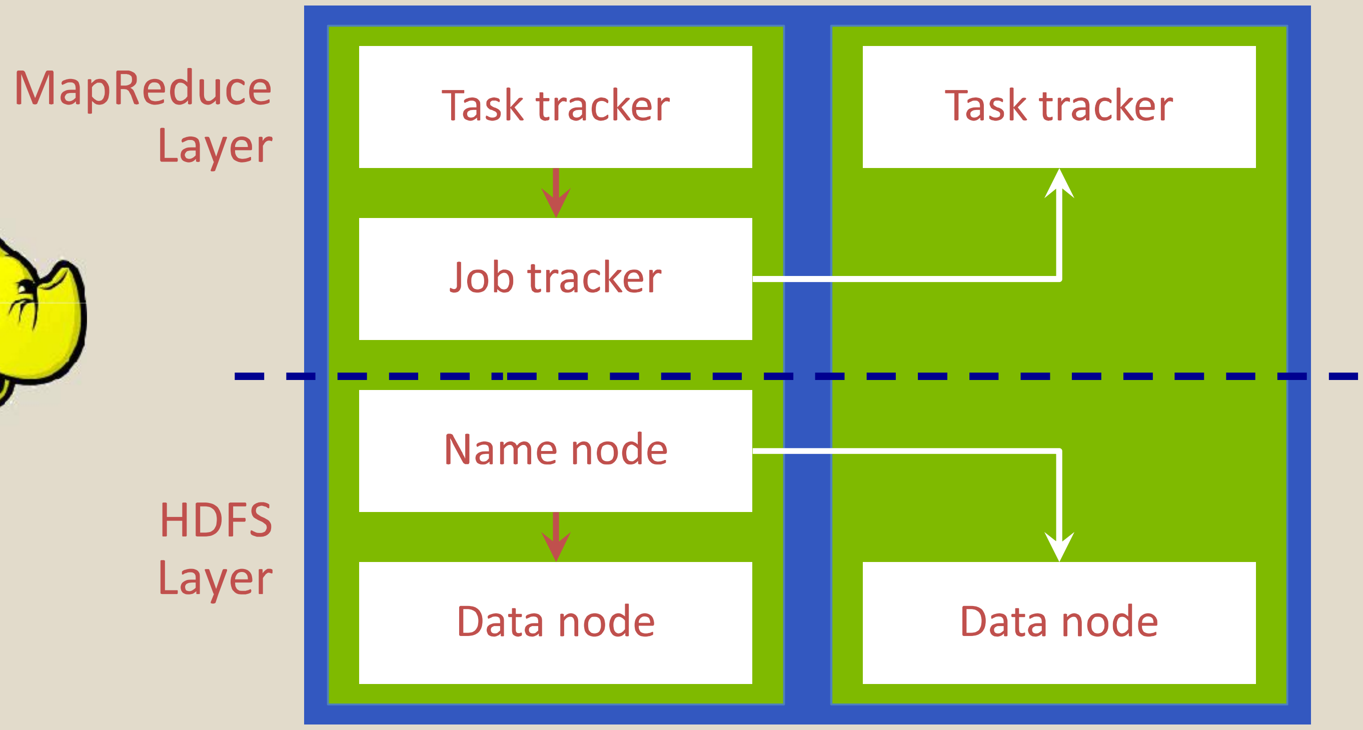


Source: Continuous Auditing in Big Data Computing Environments, Andreas Kiezow, University of Osnabruck

Cloud based implementation

- **Infrastructure as a Service (IaaS)**
  - Elastic Compute Cloud – EC2 (IaaS)
  - Simple Storage Service – S3 (IaaS)
  - Elastic Block Storage – EBS (IaaS)
- **Platform as a Service**
  - SimpleDB (SDB) (PaaS)
  - Simple Queue Service – SQS (PaaS)
  - CloudFront (S3 based Content Delivery Network – PaaS)

Hadoop: The most visible face of Big Data



MapReduce Layer

Task tracker

Job tracker

Task tracker

HDFS Layer

Name node

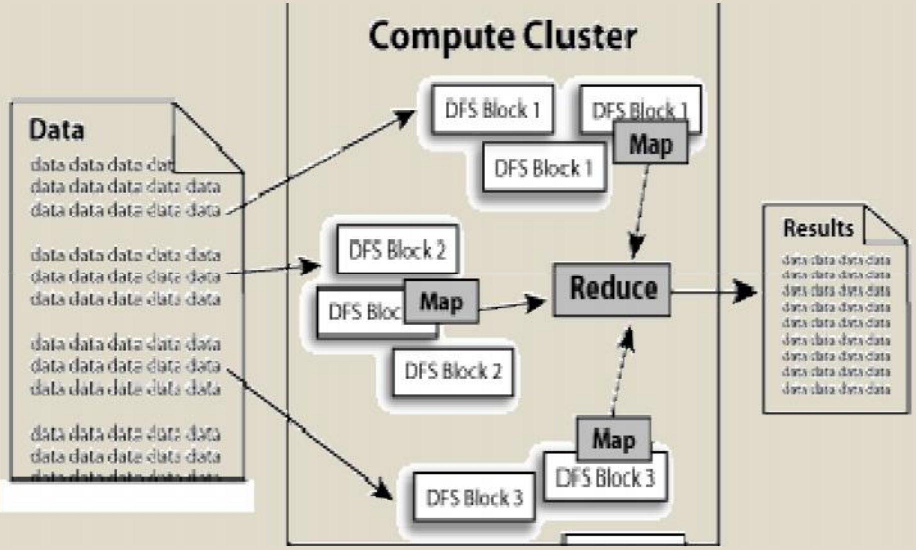
Data node

Data node

Source: Ensuring Compliance of Patient Data with Big Data and BI, Ayad Shammout & Denny Lee, PASS Business Analytics Conference

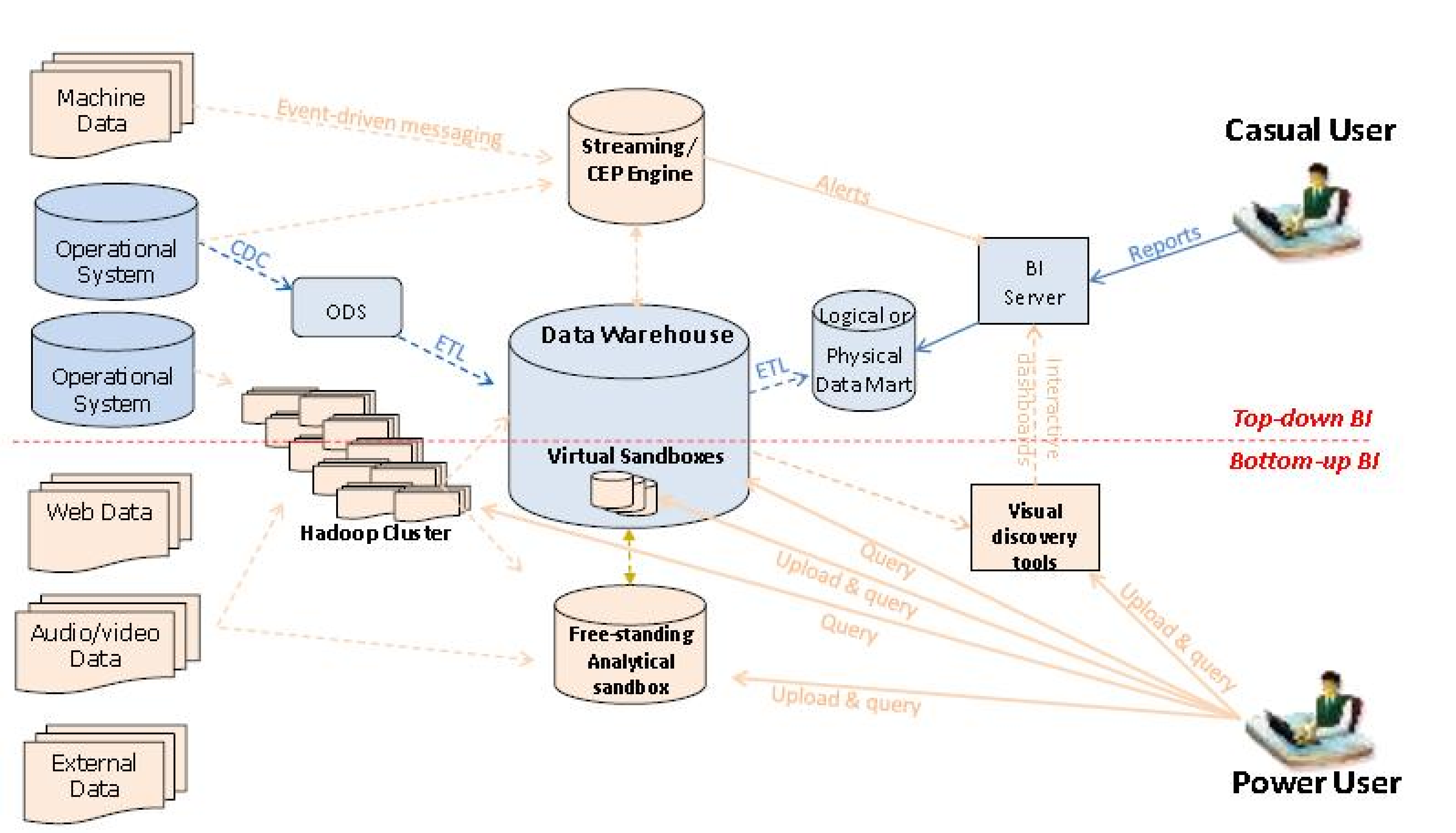
Hadoop: How does it do?

- Hadoop implements Google's MapReduce, using HDFS
- MapReduce divides applications into many small blocks of work/job.
- HDFS creates multiple replicas of data blocks for reliability, placing them on compute nodes around the cluster.
- MapReduce can then process the data where it is located.
- Hadoop 's target is to run on clusters of the order of 10,000-nodes.



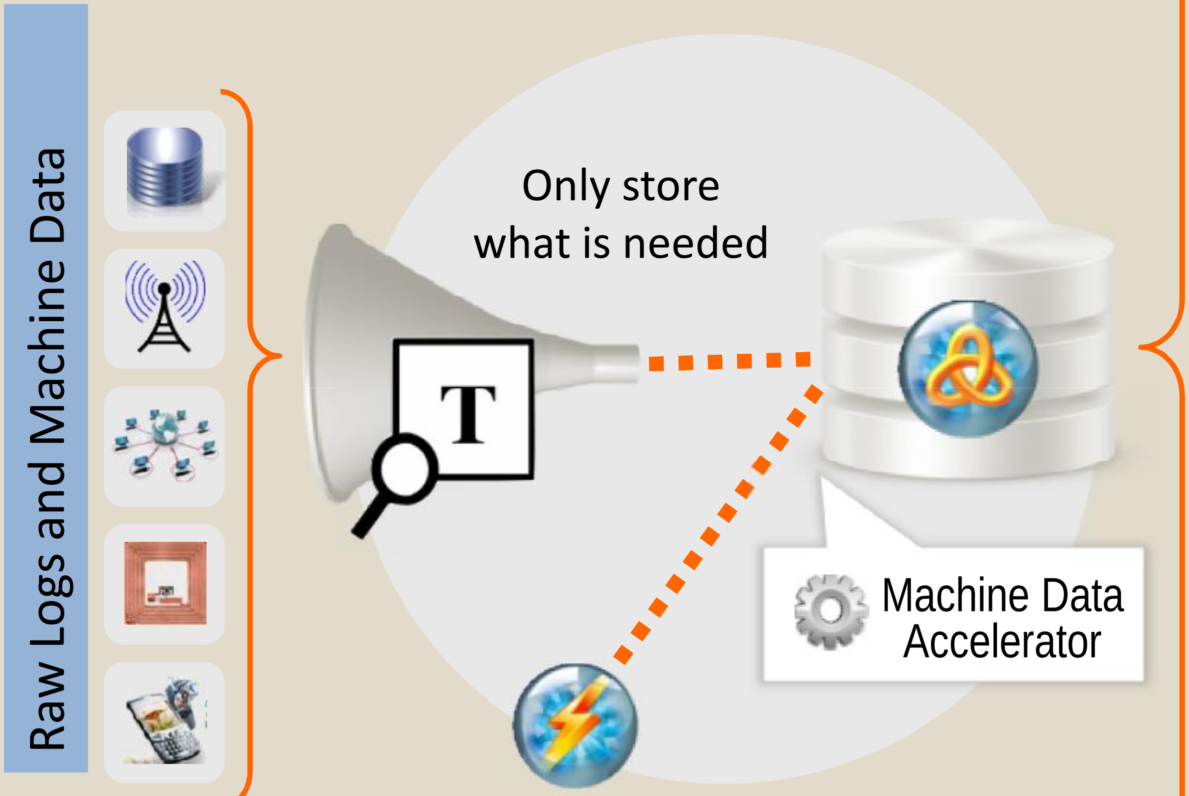
Source: Hadoop: A Software Framework for Data Intensive Computing Applications, Ravi Mukkamala, Old Dominion University

Integrated Analytics Architecture with Hadoop



Source: Watson, Tutorial Big Data Business Analytics

Sample Machine Data Refinery with IBM BigInsights®

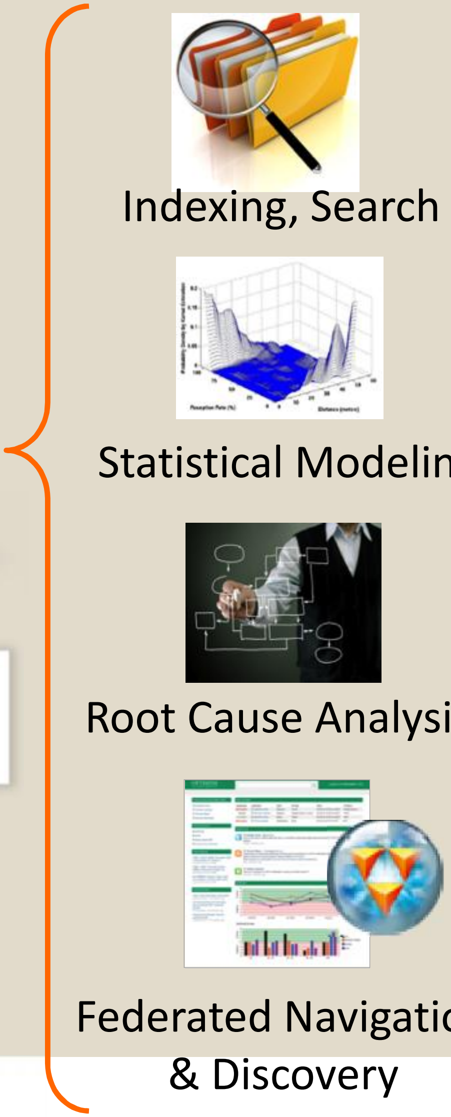


Raw Logs and Machine Data

Only store what is needed

Machine Data Accelerator

Real-time Analysis



Indexing, Search

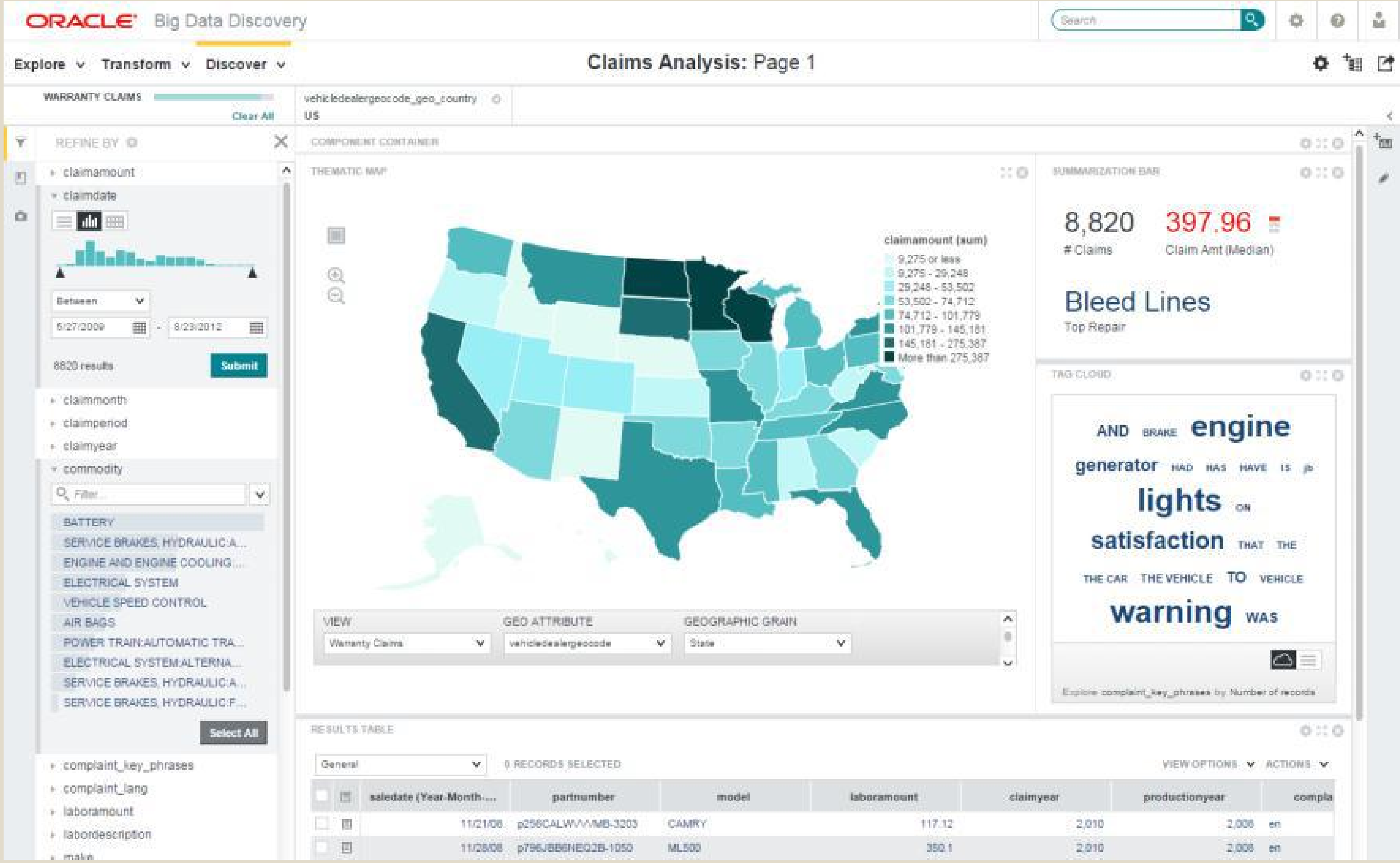
Statistical Modeling

Root Cause Analysis

Federated Navigation & Discovery

Source: Big Data, New Era of Analytic, Omer Sever, IBM

Visual Discovery with Oracle Big Data Discovery®



vehicledealergeocode_geo_country
US

- claimamount
- claimdate

Between 5/27/2009 - 8/23/2012

8820 results

- claimmonth
- claimperiod
- claimyear
- commodity

Filter...

- BATTERY
- SERVICE BRAKES, HYDRAULIC:A...
- ENGINE AND ENGINE COOLING...
- ELECTRICAL SYSTEM
- VEHICLE SPEED CONTROL
- AIR BAGS
- POWER TRAIN:AUTOMATIC TRA...
- ELECTRICAL SYSTEM:ALTERNA...
- SERVICE BRAKES, HYDRAULIC:A...
- SERVICE BRAKES, HYDRAULIC:F...

- complaint_key_phrases
- complaint_lang
- laboramount
- labordescription
- make

claimamount (sum)
9,275 or less
9,275 - 29,248
29,248 - 53,502
53,502 - 74,712
74,712 - 101,779
101,779 - 145,181
145,181 - 275,387
More than 275,387

VIEW	GEO ATTRIBUTE	GEOGRAPHIC GRAIN
Warranty Claims	vehicledealergeocode	State

8,820
# Claims

397.96
Claim Amt (Median)

Bleed Lines
Top Repair

AND BRAKE engine generator HAD HAS HAVE IS jb lights ON satisfaction THAT THE THE CAR THE VEHICLE TO VEHICLE warning WAS

Explore complaint_key_phrases by Number of records

General — 0 RECORDS SELECTED

saledate (Year-Month-...	partnumber	model	laboramount	claimyear	productionyear	compla
11/21/06	p256CALWVWMB-3203	CAMRY	117.12	2,010	2,008	en
11/28/08	p796JBB6NEQ2B-1050	ML500	350.1	2,010	2,008	en

Questions & Discussion

When do we need to be aware of?
Some of them already in the game…

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN’s Large Hydron Collider (LHC) generates 15 PB a year

640K ought to be enough for anybody.

Thank You

fandhy.haristha@iia-indonesia.org
fandhy.haristha@isaca.or.id
fandhy.haristha@commbank.co.id