Introduction to

Prosenjit Chandra Biswas
ID: 203-15-14568
Tamjid Mahmud Mahin
ID: 203-15-14498

Topics

- Scope: Big Data
- Topics:
  - Foundation of Data Analytics and Data Mining
  - Hadoop/Map-Reduce Programming and Data Processing & BigTable/Hbase/Cassandra
  - Graph Database and Graph Analytics

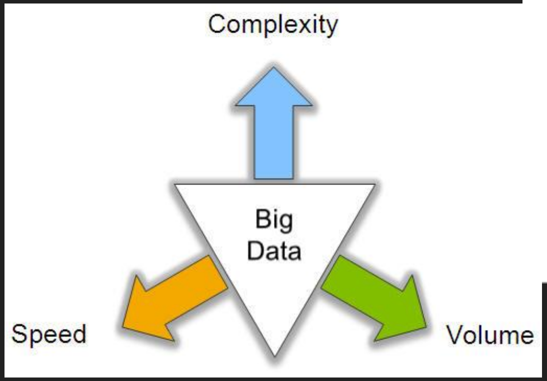
What's Big Data?

No single definition; here is from Wikipedia:

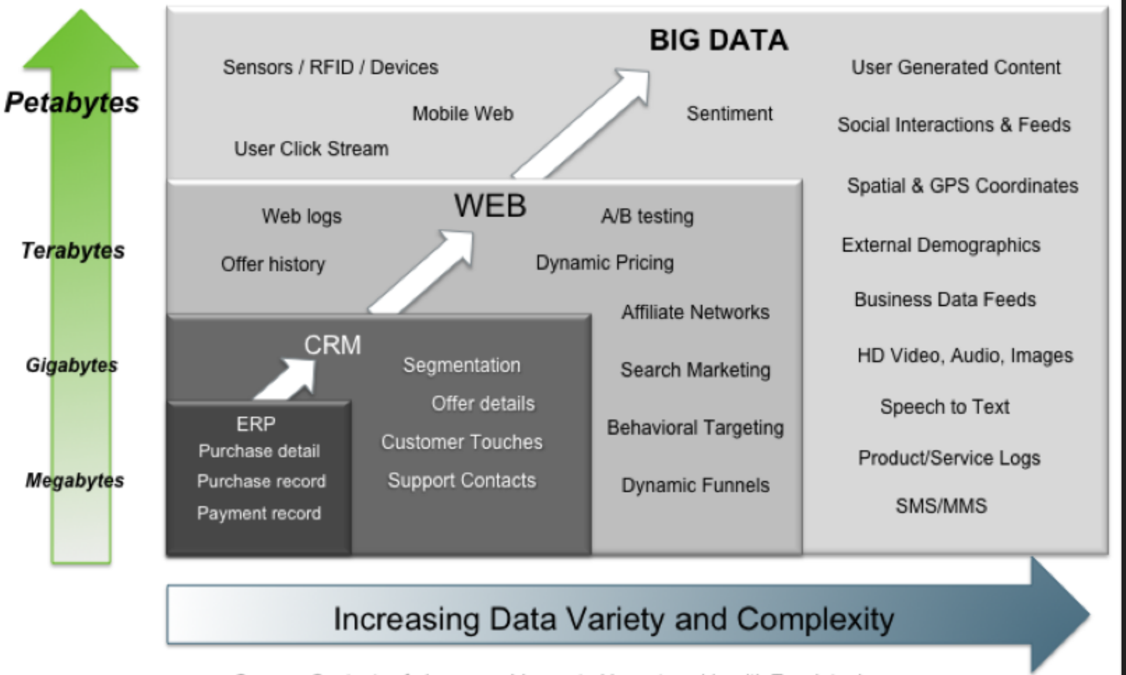
- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
- 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."

Big Data: 3V's

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



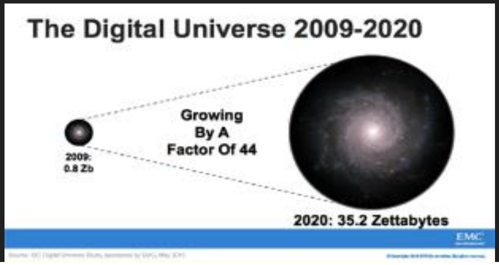
Big Data = Transactions + Interactions + Observations

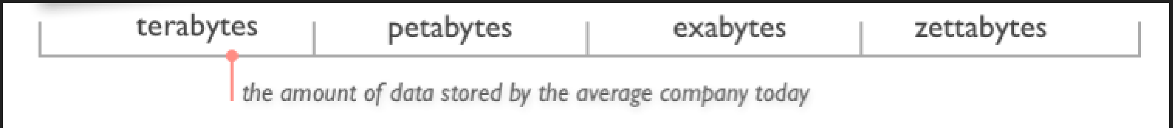


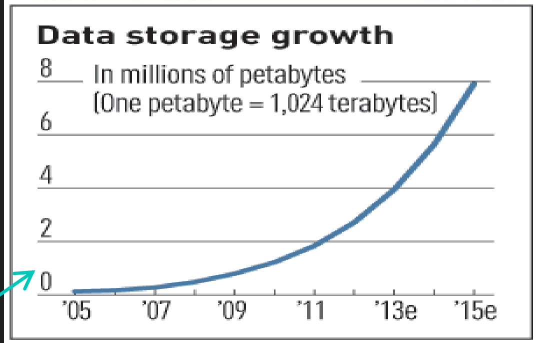
Source: *Contents of above graphic created in partnership with Teradata, Inc.*

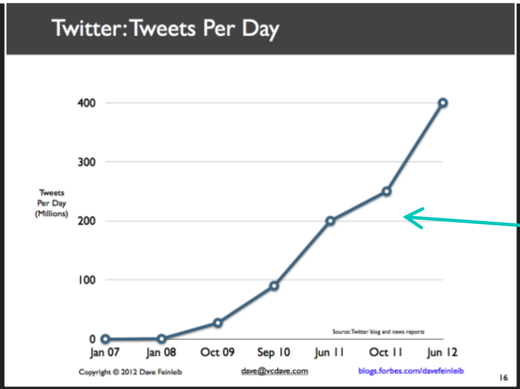
Volume (Scale)

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





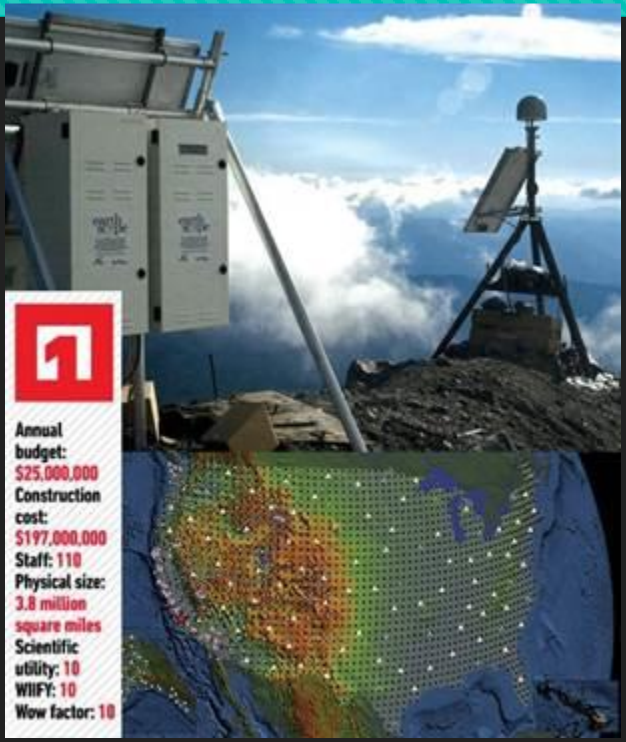




Exponential increase in collected/generated data

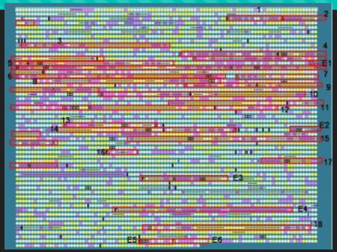
The Earthscope

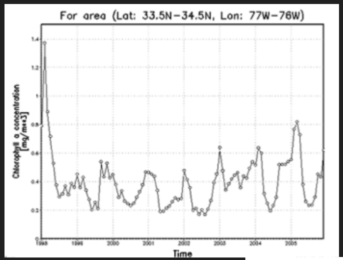
- The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more. (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI)

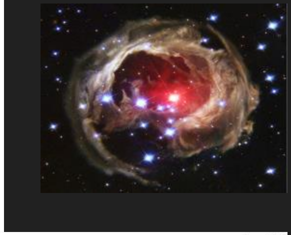

1
Annual budget: $25,000,000
Construction cost: $197,000,000
Staff: 110
Physical size: 3.8 million square miles
Scientific utility: 10
WIIFY: 10
Wow factor: 10

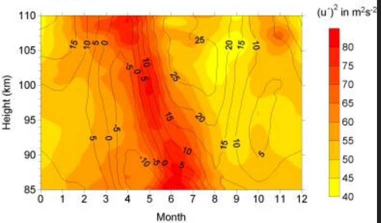
Variety (Complexity)

- 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)



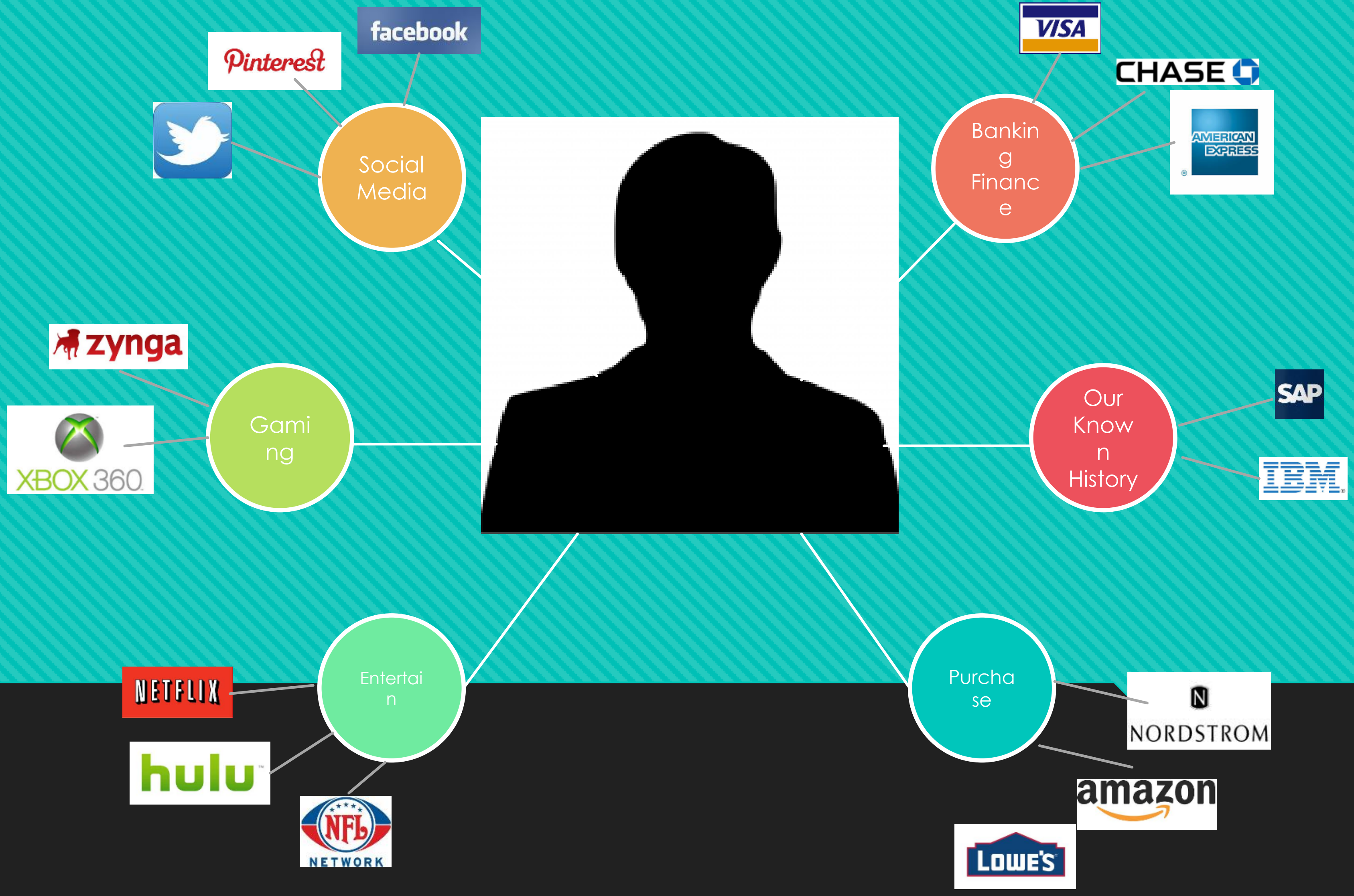

For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m3]
Time





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

A Single View to the Customer



Social Media

Bankin
g
Financ
e

Gami
ng

Our
Know
n
History

Entertai
n

Purcha
se

Velocity (Speed)

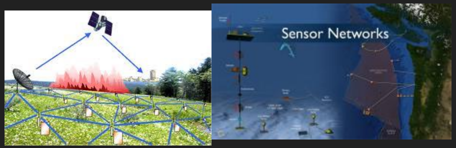
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - E-Promotions: Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - Healthcare monitoring: sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

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 tim

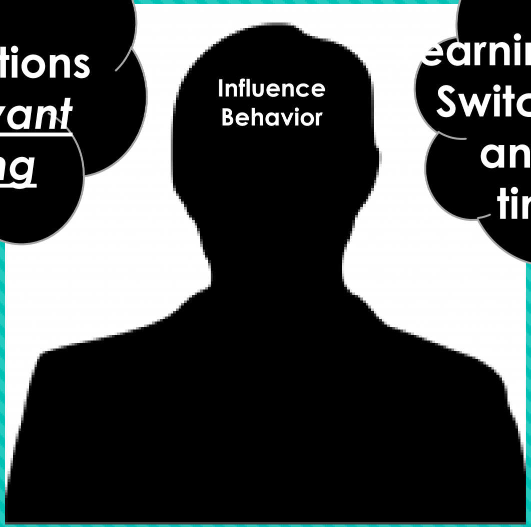


Sensor technology and networks
(measuring all kinds of data)

- 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

Real-Time Analytics/Decision Requirement

Product Recommendations that are *Relevant* & *Compelling*



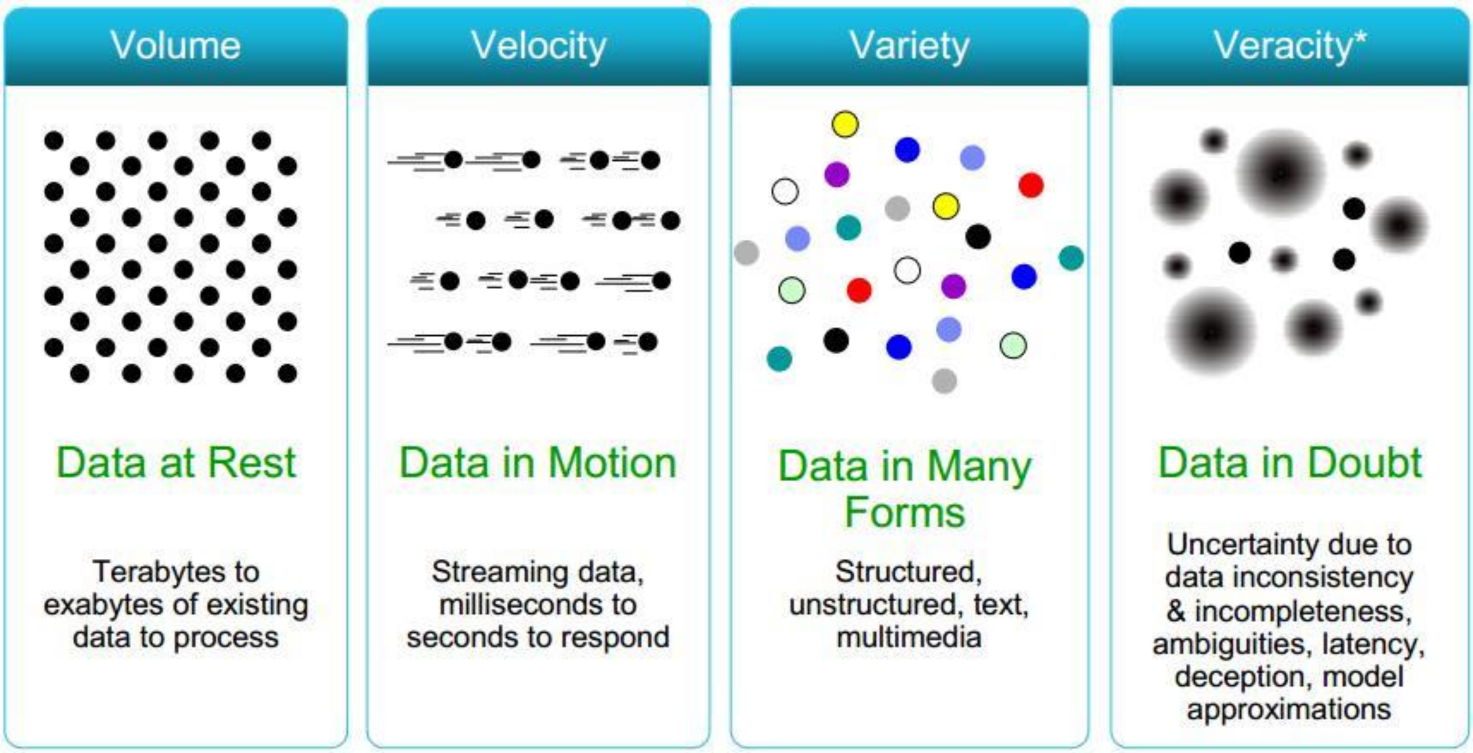
earning why Customers Switch to competitors and their offers; in time to Counter

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

Friend Invitations to join a Game or Activity that expands business

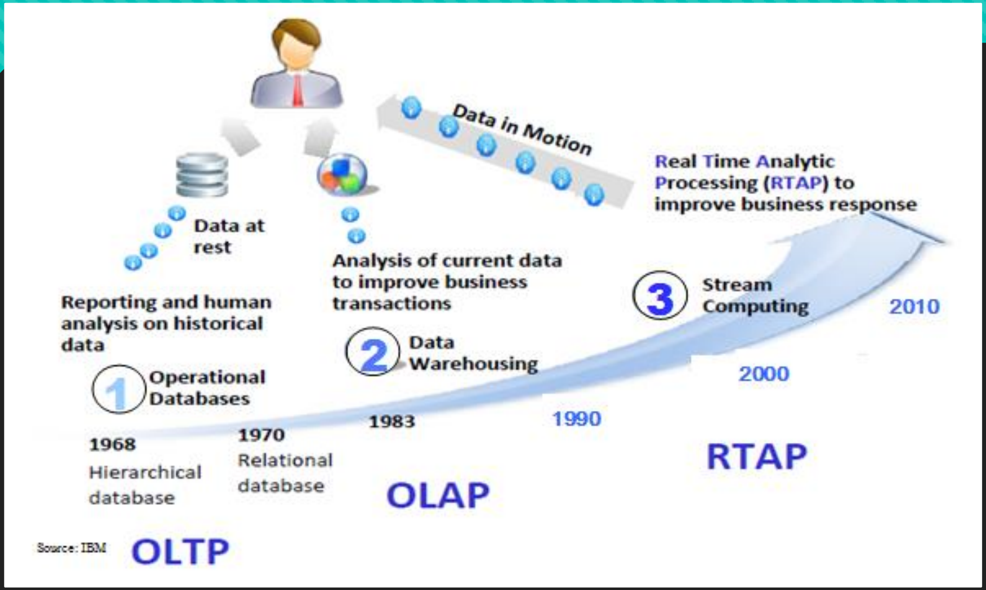
Preventing Fraud as it is *Occurring* & preventing more proactively

Some Make it 4V's



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

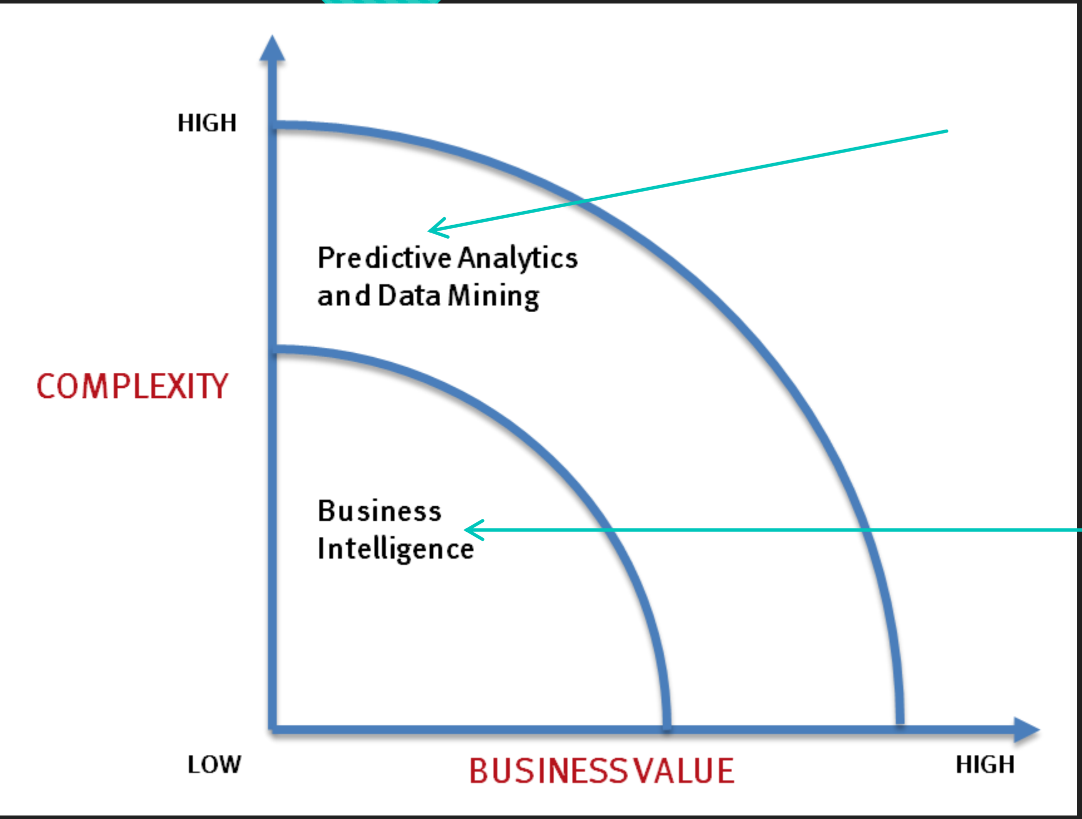
Harnessing Big Data


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
1983
1990
1968 Hierarchical database
1970 Relational database
RTAP
OLAP
OLTP

Source: IBM

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

What’s driving Big Data

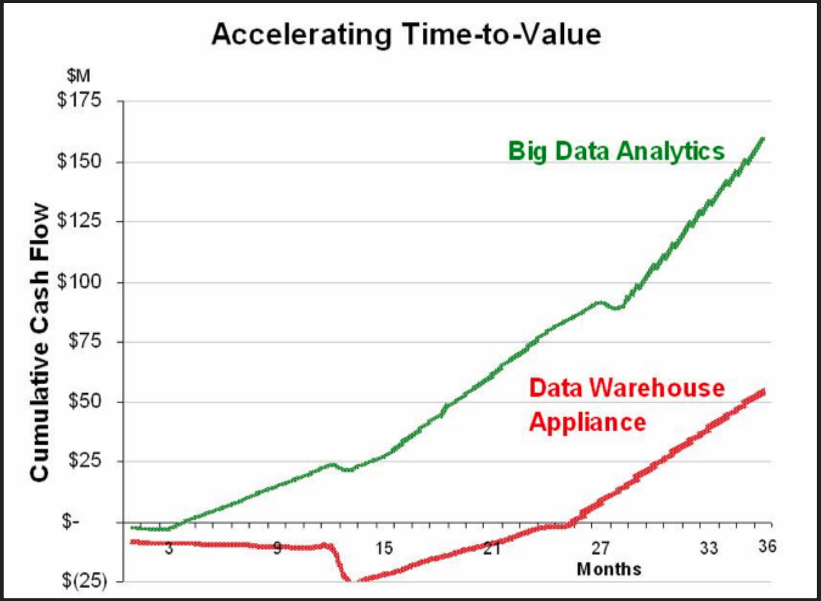


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

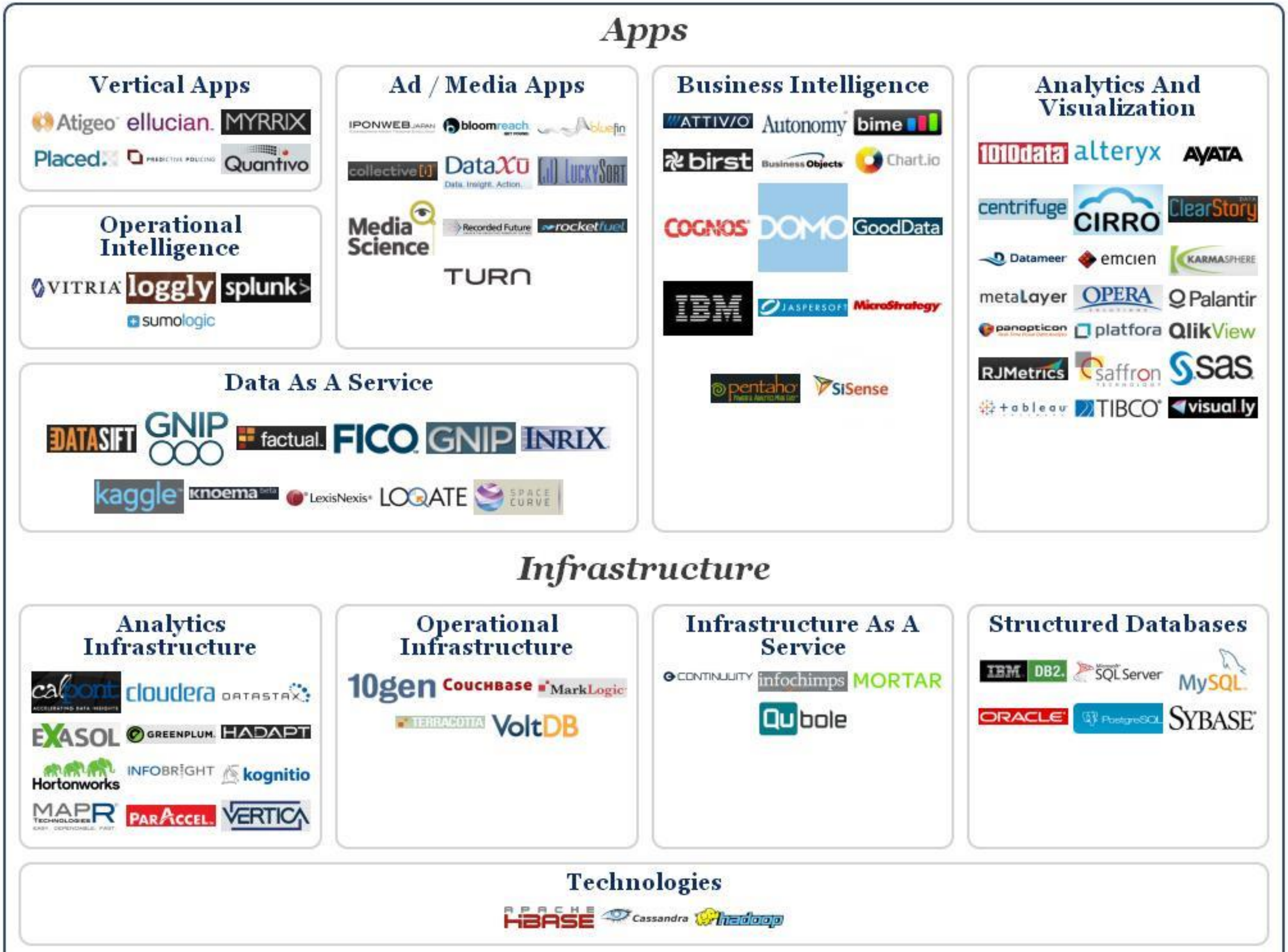
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

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



The Big Data Landscape



Apps

Vertical Apps

Atigeo, ellucian, MYRRIX, Placed, Predictive Policing, Quantivo

Operational Intelligence

VITRIA, loggly, splunk>, sumologic

Ad / Media Apps

IPONWEB JAPAN, bloomreach, bluefin, collective[i], DataXu Data. Insight. Action., LuckySort, Media Science, Recorded Future, rocketfuel, TURN

Business Intelligence

ATTIVIO, Autonomy, bime, birst, Business Objects, Chart.io, COGNOS, DOMO, GoodData, IBM, JASPERSOFT, MicroStrategy, pentaho, SiSense

Analytics And Visualization

1010data, alteryx, AYATA, centrifuge, CIRRO, ClearStory DATA, Datameer, emcien, KARMASPHERE, metaLayer, OPERA SOLUTIONS, Palantir, panopticon, platfora, QlikView, RJMetrics, saffron TECHNOLOGY, sas, tableau, TIBCO, visual.ly

Data As A Service

DATASIFT, GNIP, factual., FICO, GNIP, INRIX, kaggle, knoema beta, LexisNexis, LOQATE, SPACE CURVE

Infrastructure

Analytics Infrastructure

calpont ACCELERATING DATA INSIGHTS, cloudera, DATASTAX, EXASOL, GREENPLUM, HADAPT, Hortonworks, INFOBRIGHT, kognitio, MAPR TECHNOLOGIES EASY. DEPENDABLE. FAST., ParAccel, VERTICA

Operational Infrastructure

10gen, Couchbase, MarkLogic, TERRACOTTA, VoltDB

Infrastructure As A Service

CONTINUUITY, infochimps, MORTAR, Qubole

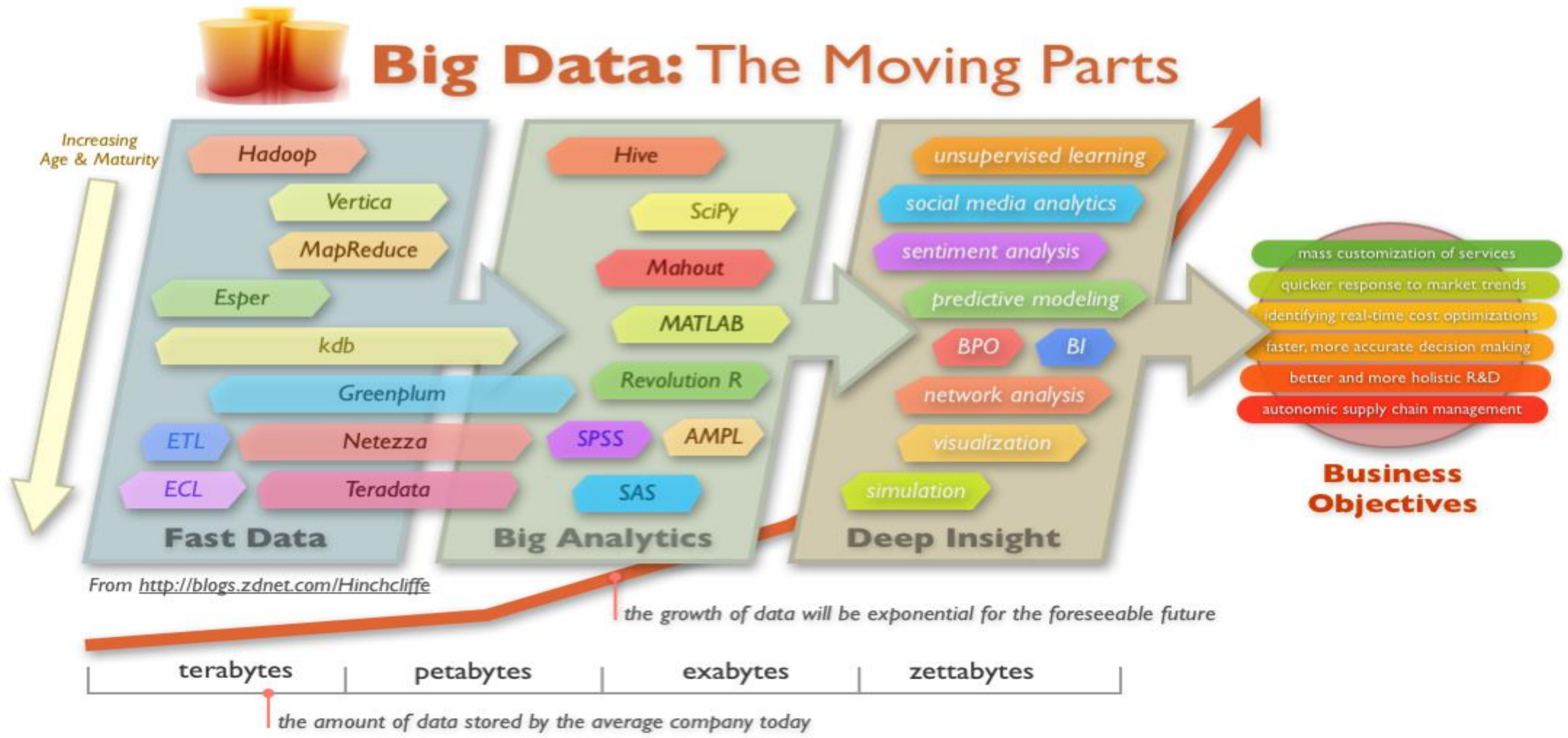
Structured Databases

IBM DB2, Microsoft SQL Server, MySQL, ORACLE, PostgreSQL, SYBASE

Technologies

APACHE HBASE, Cassandra, hadoop

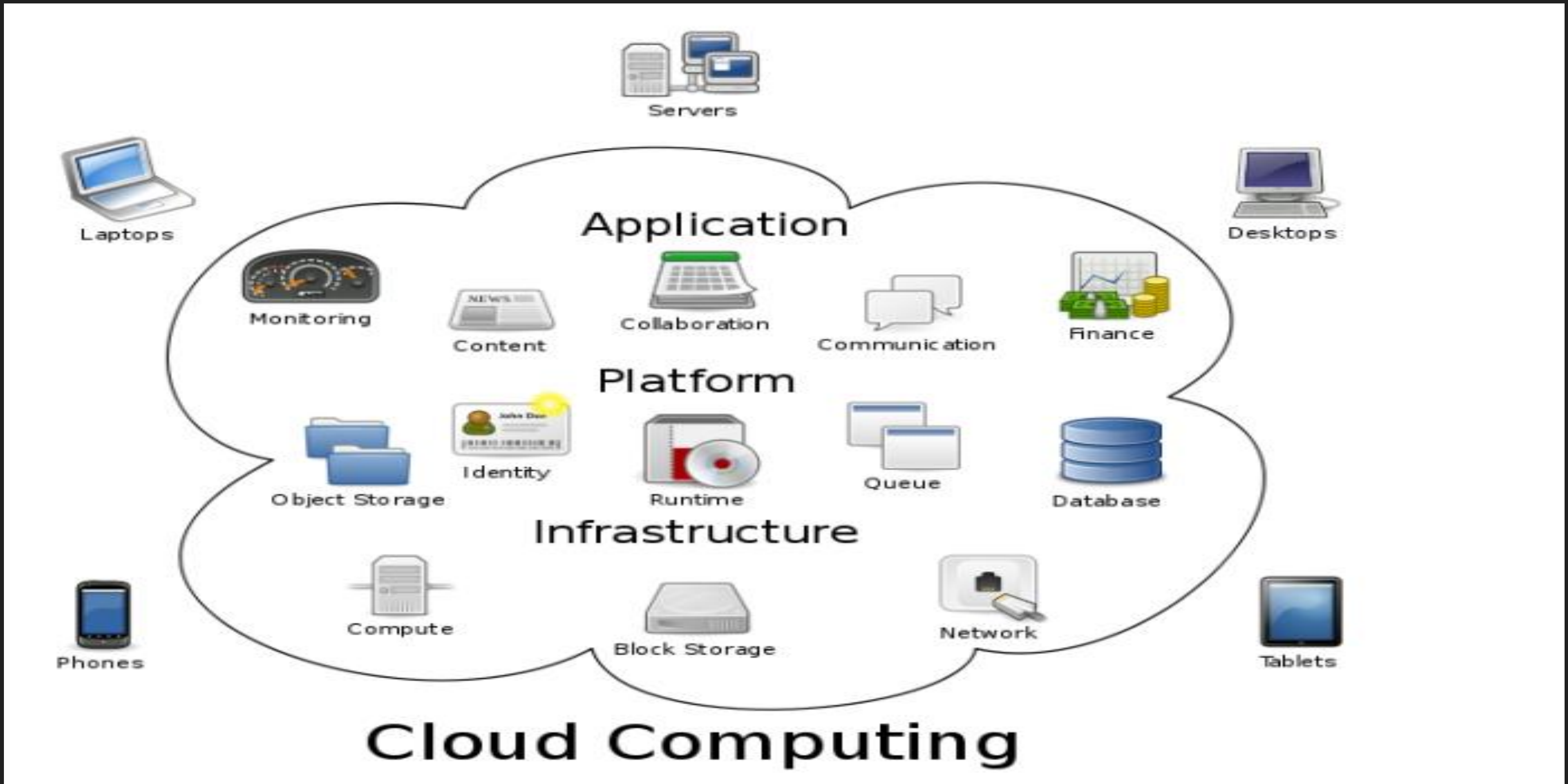
Big Data Technology


Big Data: The Moving Parts
Increasing Age & Maturity
Hadoop
Vertica
MapReduce
Esper
kdb
Greenplum
ETL
Netezza
ECL
Teradata
Fast Data
Hive
SciPy
Mahout
MATLAB
Revolution R
SPSS
AMPL
SAS
Big Analytics
unsupervised learning
social media analytics
sentiment analysis
predictive modeling
BPO
BI
network analysis
visualization
simulation
Deep Insight
mass customization of services
quicker response to market trends
identifying real-time cost optimizations
faster, more accurate decision making
better and more holistic R&D
autonomic supply chain management
Business Objectives
the growth of data will be exponential for the foreseeable future
terabytes
petabytes
exabytes
zettabytes
the amount of data stored by the average company today

From http://blogs.zdnet.com/Hinchcliffe

Cloud Computing

- IT resources provided as a service
  - Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
  - Cheap storage, high bandwidth networks & multicore processors
  - Geographically distributed data centers
- Offerings from Microsoft, Amazon, Google, …

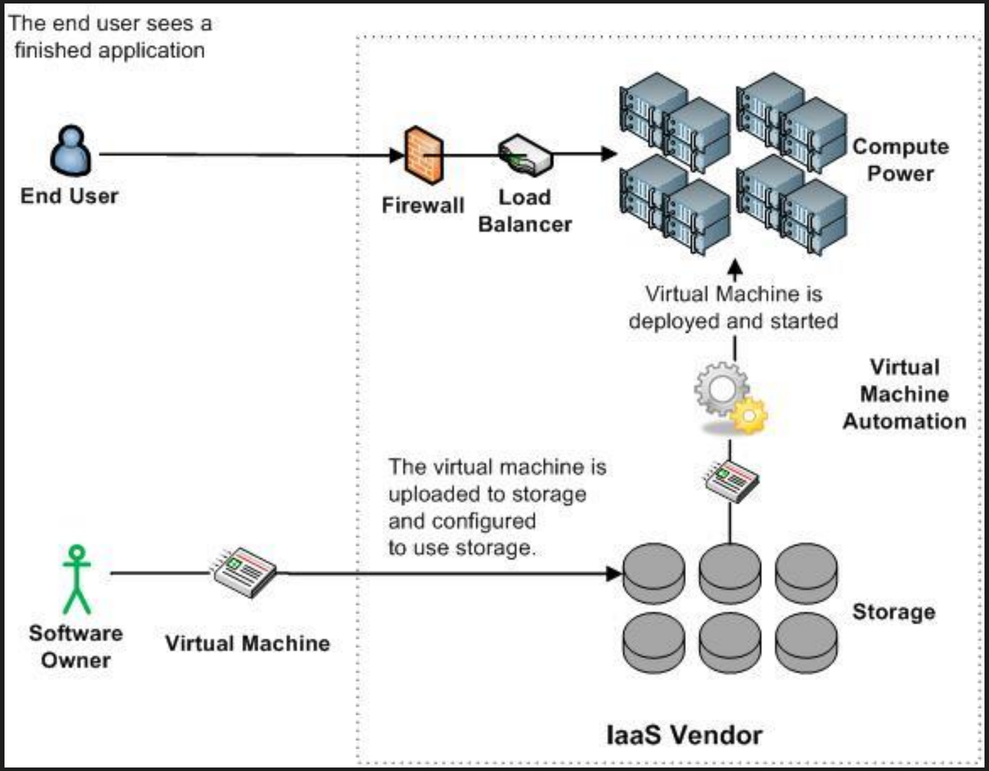


wikipedia: Cloud Computing

Benefits

  - Cost & management
  - Economies of scale, “out-sourced” resource management
- Reduced Time to deployment
  - Ease of assembly, works “out of the box”
- Scaling
  - On demand provisioning, co-locate data and compute
- Reliability
  - Massive, redundant, shared resources
- Sustainability
  - Hardware not owned

Infrastructure as a Service (IaaS)


The end user sees a finished application
End User
Firewall
Load Balancer
Compute Power
Virtual Machine is deployed and started
Virtual Machine Automation
The virtual machine is uploaded to storage and configured to use storage.
Software Owner
Virtual Machine
Storage
IaaS Vendor

More Refined Categorization

- Storage-as-a-service
- Database-as-a-service
- Information-as-a-service
- Process-as-a-service
- Application-as-a-service
- Platform-as-a-service
- Integration-as-a-service
- Security-as-a-service
- Management/ Governance-as-a-service
- Testing-as-a-service
- Infrastructure-as-a-service

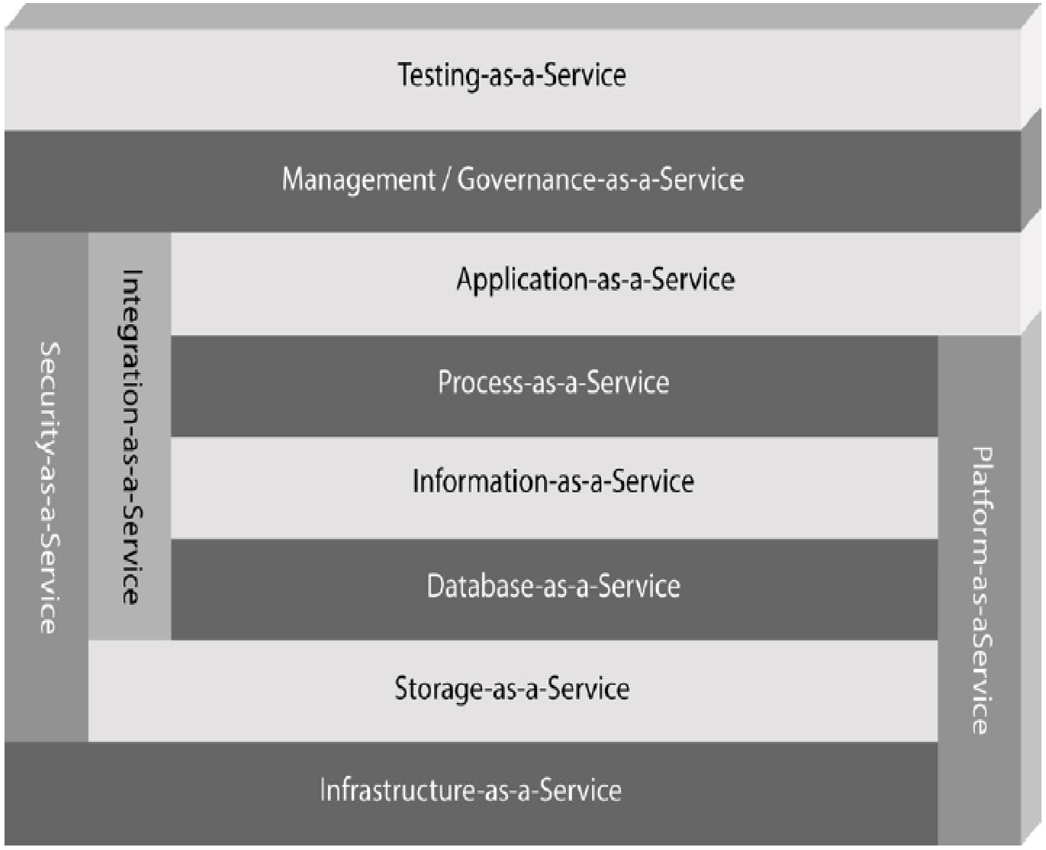


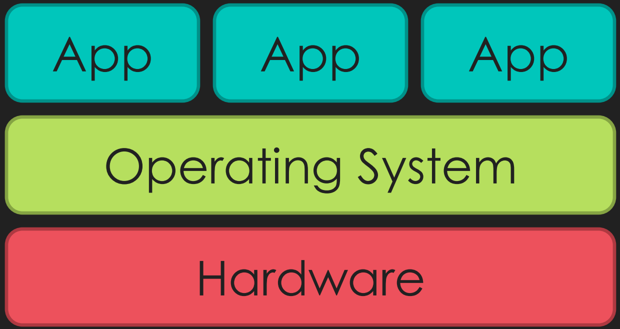
Figure 1: The patterns or categories of cloud computing providers allow you to use a discrete set of services within your architecture.

InfoWorld Cloud Computing Deep Div

Key Ingredients in Cloud Computing

- Service-Oriented Architecture (SOA)
- Utility Computing (on demand)
- Virtualization (P2P Network)
- SAAS (Software As A Service)
- PAAS (Platform AS A Service)
- IAAS (Infrastructure AS A Servie)
- Web Services in Cloud

Enabling Technology: Virtualization

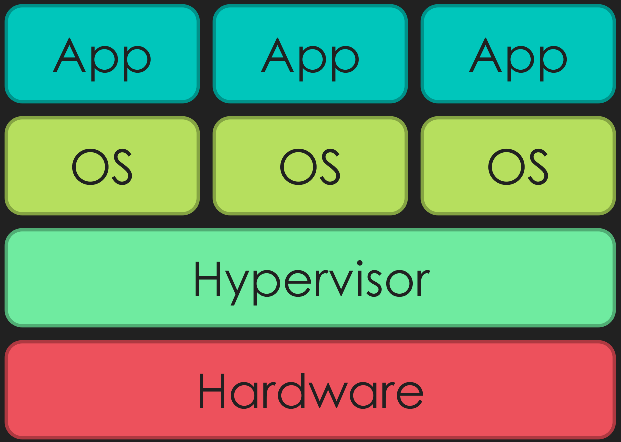


App App App

Operating System

Hardware

Traditional Stack



App App App

OS OS OS

Hypervisor

Hardware

Virtualized Stack

THANK YOU