Smarter education for smarter planet

IBM Innovation Centre for Education

Welcome to:

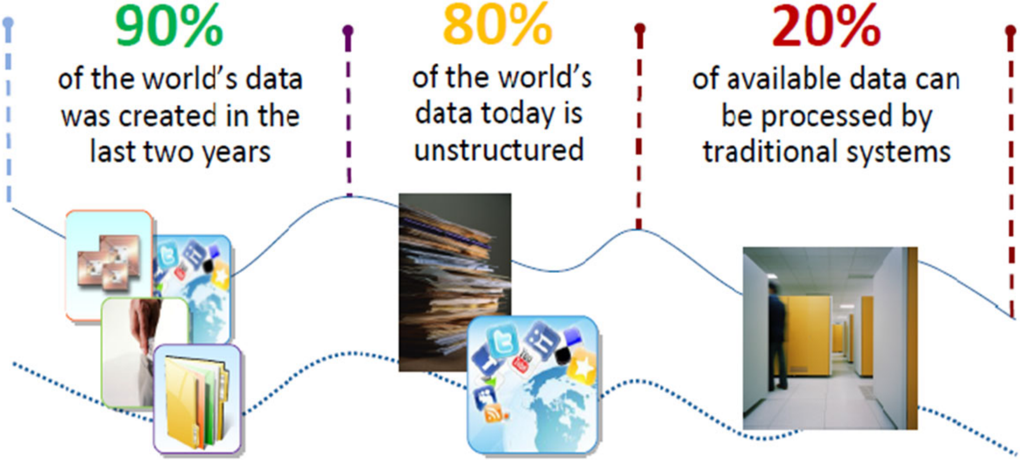
Unit 1 - Introduction to Big Data Analytics

Unit objectives

After completing this unit, you should be able to:

- Understand Need of Big Data
- Understand Big Data – Concept, Characteristics and Dimensions
- Understand Fundamentals of Big Data Architecture
- Understand Big Data Tools and Techniques
- Understand Big Data - Applications

Big Data growth story



90% of the world's data was created in the last two years

80% of the world's data today is unstructured

20% of available data can be processed by traditional systems

1 in 2 business leaders don't have access to data they need

83% of CIO's cited BI and analytics as part of their visionary plan

5.4X more likely that top performers use business analytics

Big data Sources

500+ Million users posting 55 Million tweets every day

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

1.2 Trillion searches

100s of millions of GPS enabled devices sold annually

1+ Billion active users spending 700 Million minutes per month

76 million smart meters in 2009… 200M by 2014

2+ billion people on the Web by end 2011

Big Data adoption drivers

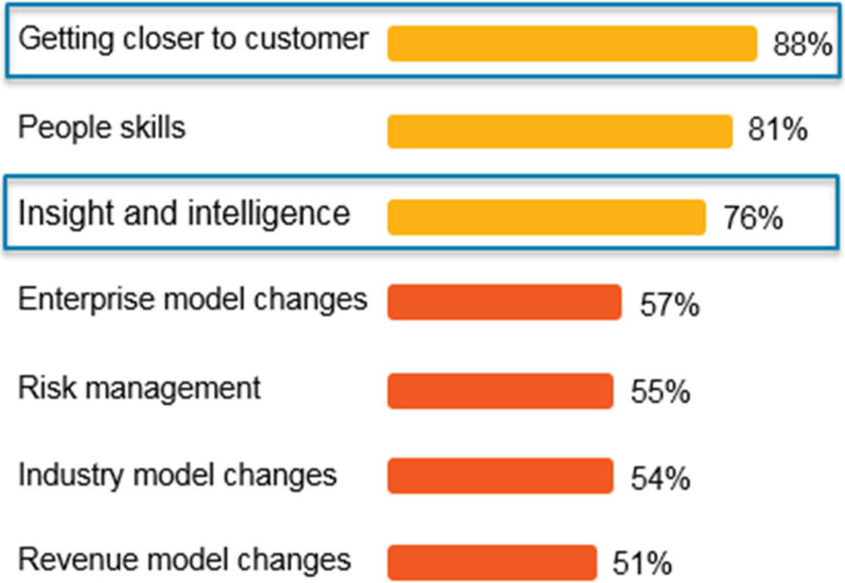
Data
is the new basis of competitive advantage
(What)

Cloud
is the path to new business models
(How)

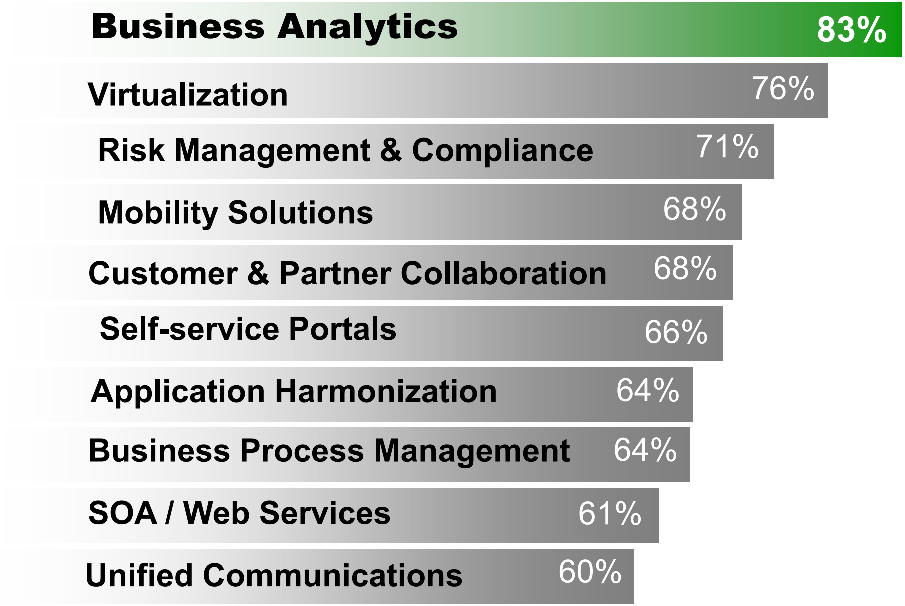
Engagement
changes our expectations
(Why)

Need of Big Data

CEO Focus Over Next 5 Years



Growth drivers for IT industry



IBM Global CIO Study.

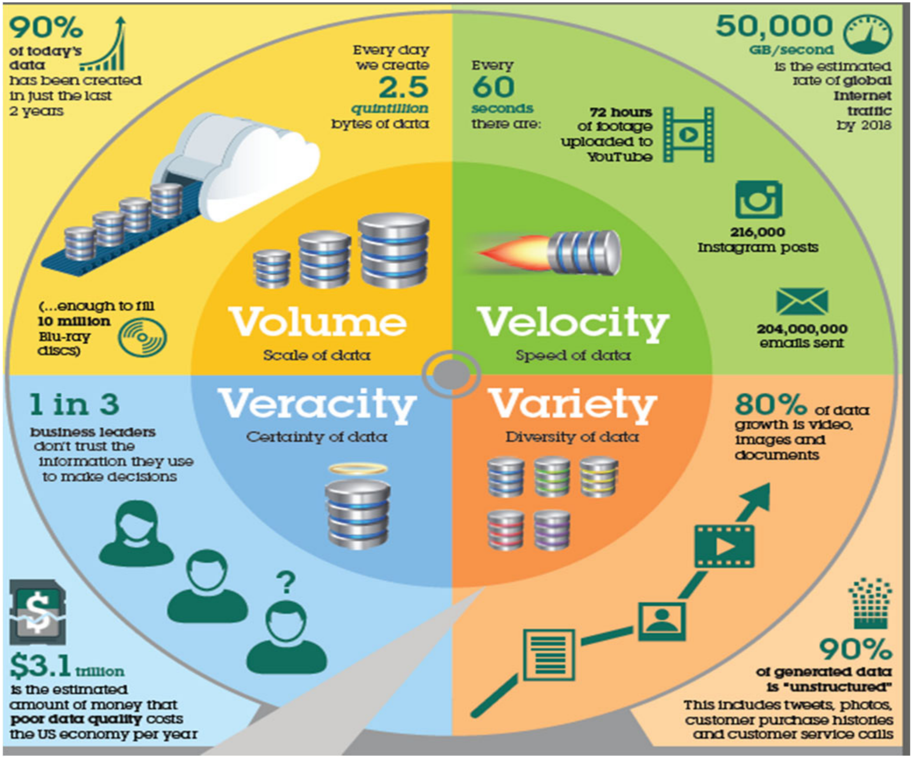
Big Data - Definition

- **Oxford English Dictionary** defines big data as
- “ **Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.”**
- Big-data is just like Small-data, but bigger in terms of
- techniques, tools & architectures and it is used to provide solution for:

- New problems
- Old problems in a better way

Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile, and to answer questions that were previously considered beyond your reach.

Characteristics of Big Data



90% of today's data has been created in just the last 2 years

Every day we create 2.5 quintillion bytes of data

(...enough to fill 10 million Blu-ray discs)

Volume

Scale of data

Every 60 seconds there are:

72 hours of footage uploaded to YouTube

216,000 Instagram posts

204,000,000 emails sent

50,000 GB/second is the estimated rate of global Internet traffic by 2018

Velocity

Speed of data

Veracity

Certainty of data

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

$3.1 trillion is the estimated amount of money that poor data quality costs the US economy per year

Variety

Diversity of data

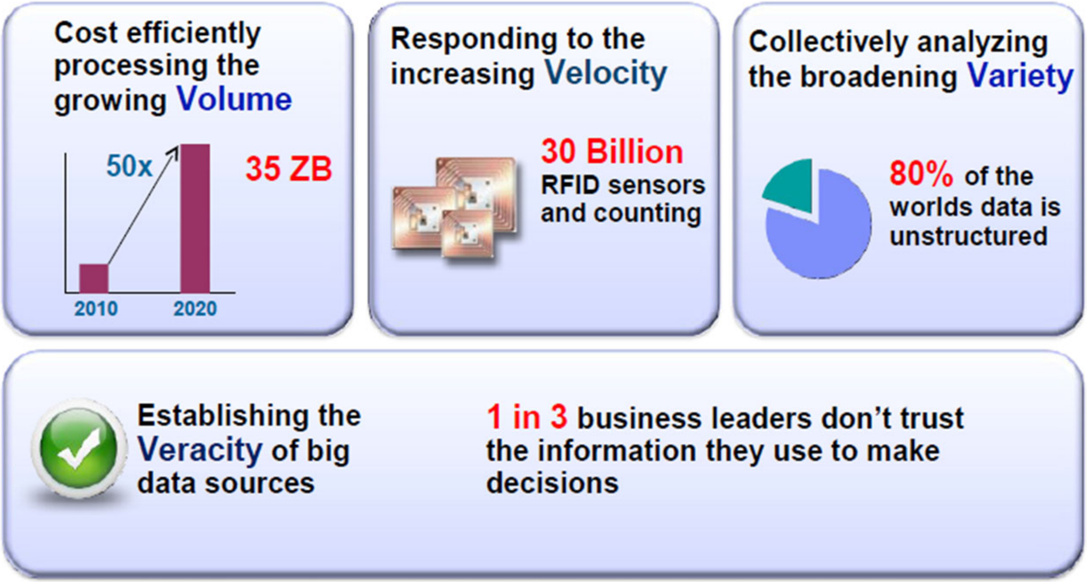
80% of data growth is video, images and documents

90% of generated data is "unstructured"

This includes tweets, photos, customer purchase histories and customer service calls

Characteristics of Big Data

- **V4 =** Volume Velocity Variety Veracity



Cost efficiently processing the growing Volume

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

Units to measure Big data

- Terabyte
- Petabyte
- Exabyte
- Zettabyte
- Yottabyte
- Brontobyte

Big data types

- The Structure of Big Data
- Today’s big data is noisy, unstructured, and dynamic rather than static. It may also be corrupted or incomplete.

- Structured data maintains hegemony over other data types. The majority of data handled via analytic platforms today falls under the rubric structured data. This is primarily about the tables and other data structures of relational databases. But other sources yield predictable structures, such as the record formats of most applications and the character-delimited rows of many flat files.

- Semi structured and complex data are coming on strong. The hegemony of structured data types will eventually be challenged by a wider range of data types.
  - Semi structured data (XML and similar standards)
  - Complex data (hierarchical or legacy sources).

Benefits & Barrier of Big Data Analytics

- Benefits of Big Data Analytics
  - Anything involving customers could benefit from big data analytics
  - Business intelligence in general can benefit from big data analytics
  - Specific analytic applications are likely beneficiaries of big data analytics

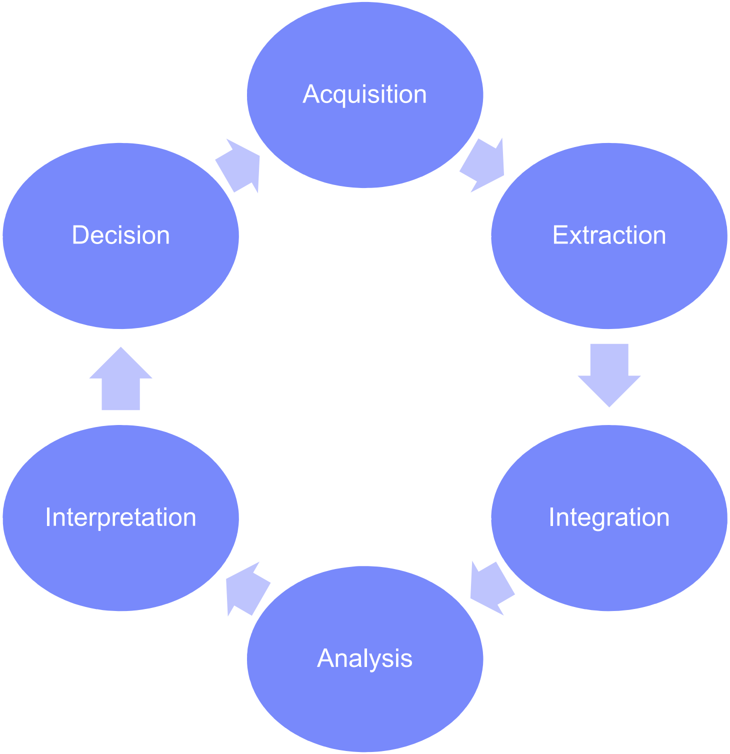
- Barriers to Big Data Analytics
  - Inadequate staffing and skills are the leading barriers to big data analytics
  - A lack of business support can hinder a big data analytics program
  - Problems with database software can be barriers to big data analytics

Need of Big Data

- Big Data has generated tremendous value for the organizations by equipping them with rare insights and information about the consumer behavior.
- It allows us to do the segmentation of the customers to be more accurate thus resulting in tailor made product and services for the individual.
- Transactional data can be created, stored and accessed in the digital format.
- We can improve decision making thereby minimizing the risk by performing state of the art analytics to provide correct insights to the management.
- New era of manufacturing and services can ushered with introduction of smart devices.

BM Global CEO Study.

Big data process



Big data Framework

for Business Users and Analysts

IBM Social Media Analytics

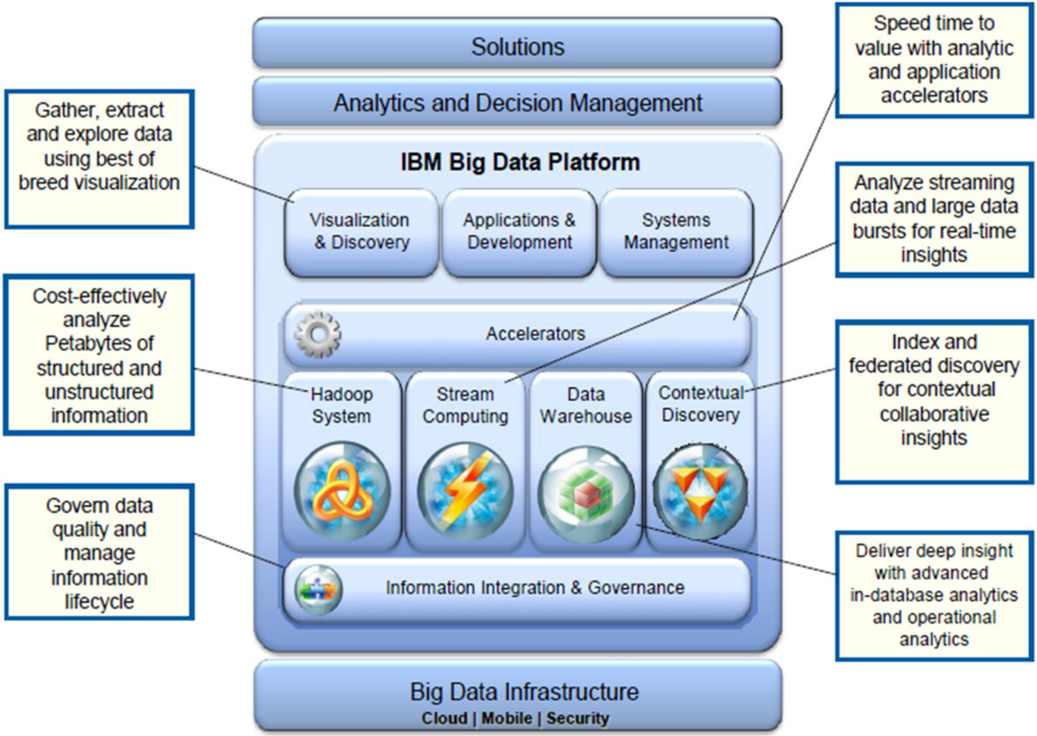
- Watson Analytics
- Social Media Analytics

for Developers, Data Professionals and Scientists

Big Insights

- IBM DataWorks
- IBM BigInsights for Cloud
- IBM Bluemix Cloud
- IBM Watson Developer Cloud

Big Data Platform and Application Frameworks



Gather, extract and explore data using best of breed visualization

Cost-effectively analyze Petabytes of structured and unstructured information

Govern data quality and manage information lifecycle

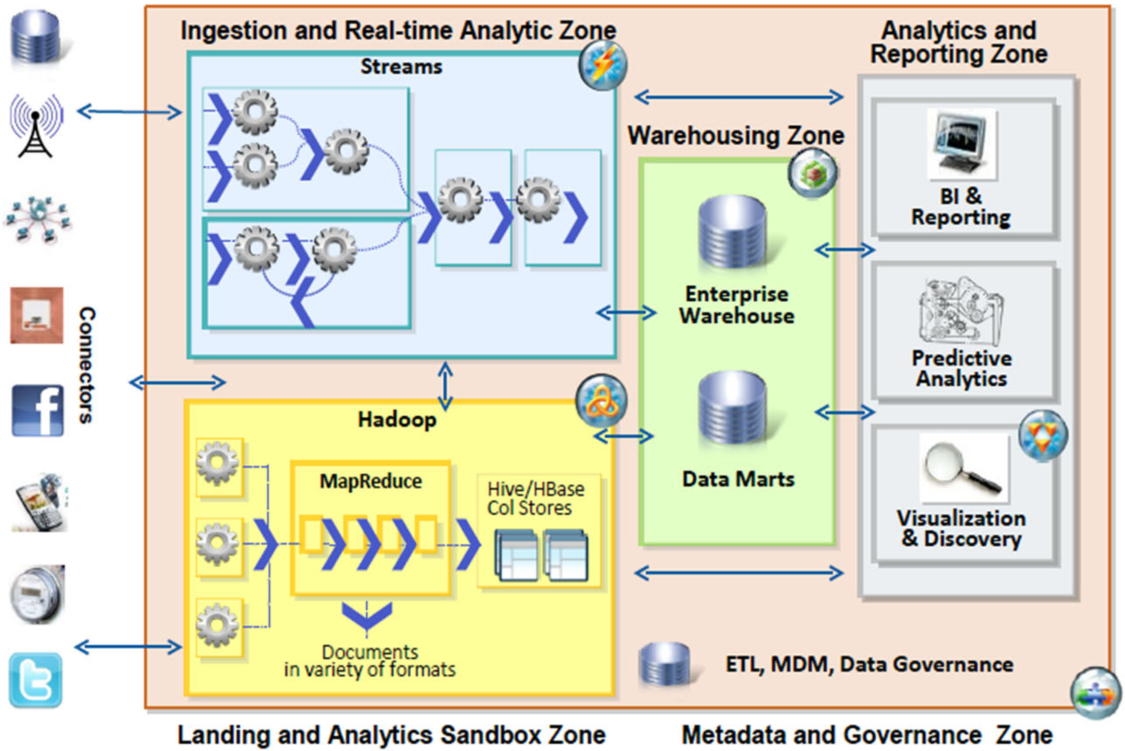
Speed time to value with analytic and application accelerators

Analyze streaming data and large data bursts for real-time insights

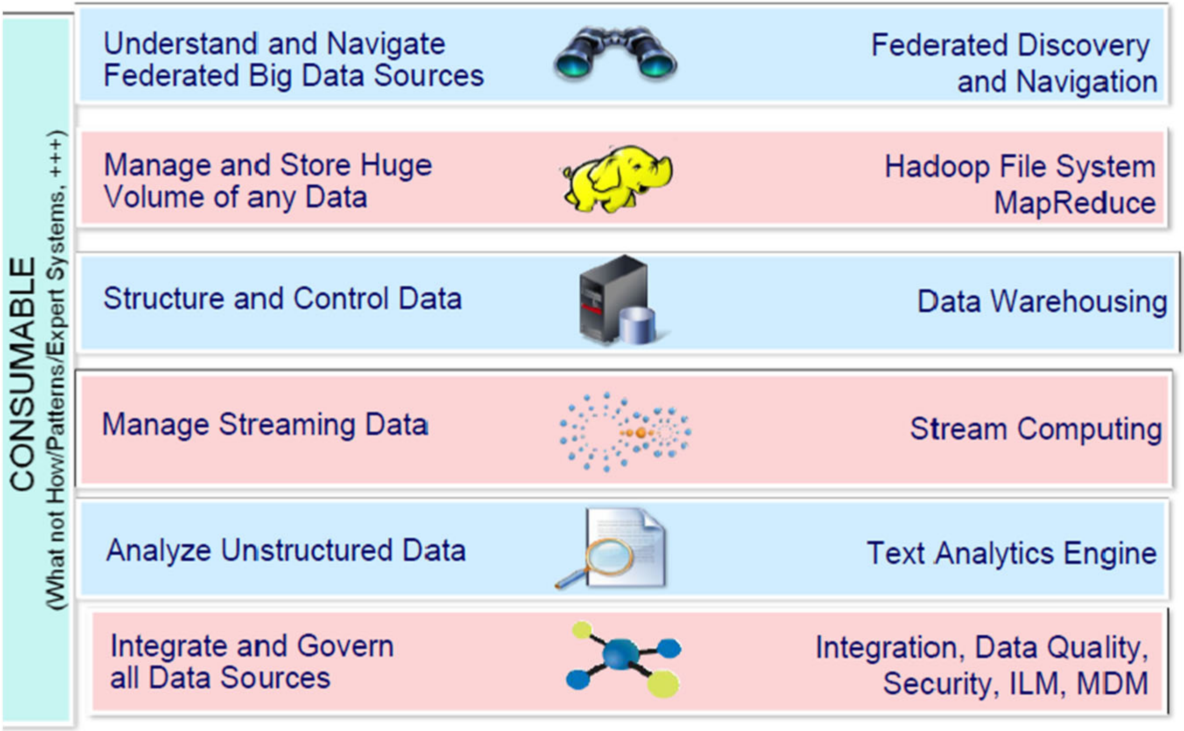
Index and federated discovery for contextual collaborative insights

Deliver deep insight with advanced in-database analytics and operational analytics

An Example of Big Data Platform in Practice


Ingestion and Real-time Analytic Zone
Streams
Analytics and Reporting Zone
Warehousing Zone
Connectors
Enterprise Warehouse
Data Marts
BI & Reporting
Predictive Analytics
Visualization & Discovery
Hadoop
MapReduce
Hive/HBase Col Stores
Documents in variety of formats
ETL, MDM, Data Governance
Landing and Analytics Sandbox Zone
Metadata and Governance Zone

A Big Data Platform Manifesto



CONSUMABLE

(What not How/Patterns/Expert Systems, +++)

Understand and Navigate Federated Big Data Sources	Federated Discovery and Navigation
Manage and Store Huge Volume of any Data	Hadoop File System MapReduce
Structure and Control Data	Data Warehousing
Manage Streaming Data	Stream Computing
Analyze Unstructured Data	Text Analytics Engine
Integrate and Govern all Data Sources	Integration, Data Quality, Security, ILM, MDM

Big Data Technologies

Hadoop

- Hadoop framework is based upon the principles given by Map Reduce and Big table. It follows the principle of distributed computing where data is distributed, managed and stored on different systems known as nodes. It was first used by Yahoo to support its business.
- Hadoop is designed to parallelize data processing across computing nodes to speed computations and hide latency.

Map Reduce

- Map reduce is designed to process a large quantity of data in a batch mode. The process follows the principle of distributed computing in which each and every task is ‘Mapped’ to a large number of systems for processing in a way that manages the recovery from failure and balances the load. This system was developed by Google.
- Reduce operates to provides the aggregator function. It aggregates back all the result to provide a result. Suppose you come across text in which recording of multilingual people is present. Map Reduce job is to determine the exactly how many recordings are captured in each and every language in text form.

Big Table

- Data storage was solved with the help of Big table. It is a distributed storage system used to manage vast quantity of highly scalable structured data.
- Big table is like multidimensional sorted map. Data captured is stored in different nodes across the systems. It is unlike the traditional databases where data is organized in rows and Columns.

Big Data Tools

Databases

- MongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper

- **MapReduce**

- Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum

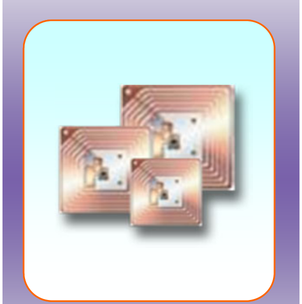
- **Storage**
  - S3, Hadoop Distributed File System
- **Servers**
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- **Processing**
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Big Data & Analytics

Transactional & Application Data

- Volume
- Structured
- Throughput

Machine Data



- Velocity
- Semi-structured
- Ingestion

Social Data

- Variety
- Highly unstructured
- Veracity

Enterprise Content

- Variety
- Highly unstructured
- Volume

Merging the Traditional and Big Data Approaches

Traditional Approach

Structured & Repeatable Analysis

Business Users

Determine what question to ask

IT

Structures the data to answer that question

Monthly sales reports
Profitability analysis
Customer surveys

Big Data Approach

Iterative & Exploratory Analysis

IT

Delivers a platform to enable creative discovery

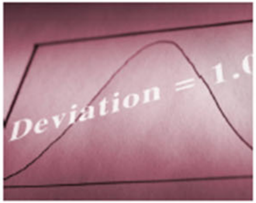
Business

Explores what questions could be asked

Brand sentiment
Product strategy
Maximum asset utilization

More Ways – Wide Ranging Analytics and Techniques

Spatial Analysis



Statistics

Text Analysis

Machine Learning

Temporal Analysis

Image Analysis

Audio Analysis

Video Analysis

The 5 Key Big Data Use Cases

Big Data Exploration

Find, visualize, understand all big data to improve decision making

Enhanced 360º View of the Customer

Extend existing customer views by incorporating additional internal and external data sources

Security/Intelligence Extension

Lower risk, detect fraud and monitor cyber security in real-time

Operations Analysis

Analyze a variety of machine data for improved business results

Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

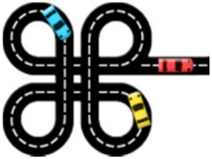
Big data Usage

Low-latency network analysis

Fraud & risk detection

Understand and act on customer sentiment

Real-time Traffic Flow Optimization



Accurate and timely threat detection

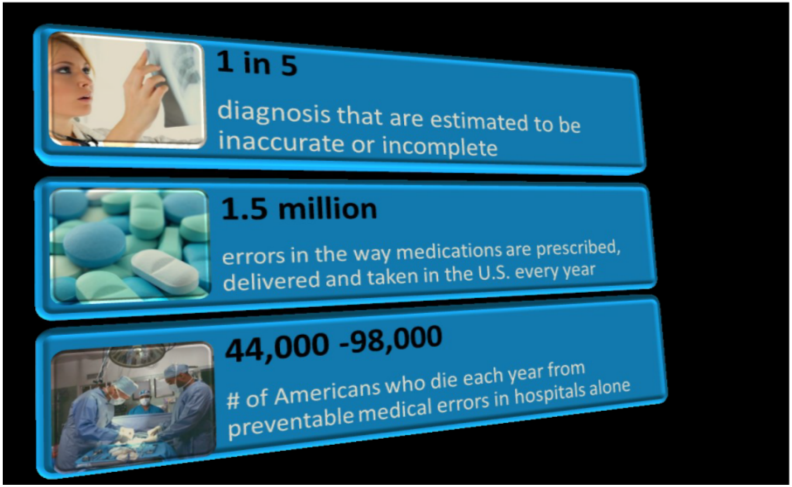
Predict and act on intent to purchase

Use Cases

Today's Challenge	New Data	What's Possible
Healthcare **Expensive office visits**	Remote patient monitoring	Preventive care, reduced hospitalization
Manufacturing **In-person support**	Product sensors	Automated diagnosis, support
Location-Based Services **Based on position**	Real time location data	Geo-advertising, traffic, local search
Public Sector **Standardized services**	Citizen surveys	Tailored services, cost reductions
Retail **One size fits all marketing**	Social media	Sentiment analysis segmentation

Big Data and Complexity in Health Care

- Medical information is doubling every 5 years, much of which is unstructured
- 81% of physicians report spending 5 hours or less per month reading medical journals



1 in 5

diagnosis that are estimated to be inaccurate or incomplete

1.5 million

errors in the way medications are prescribed, delivered and taken in the U.S. every year

44,000 -98,000

# of Americans who die each year from preventable medical errors in hospitals alone

“Medicine has become too complex (and only) about 20 percent of the knowledge clinicians use today is evidence-based”

– Steven Shapiro, Chief Medical and Scientific Officer, UPMC

…to keep up with the state of the art, a doctor would have to devote 160 hours a week to perusing papers…”

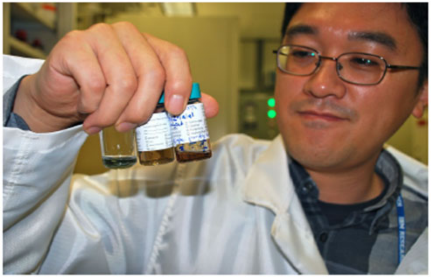
– The Economist Feb 14th 2013

Source - http://www.ibm.com/software/data/infosphere/use-cases

Use Cases for a Big Data Platform

- Healthcare and Life Sciences

  - Problem:
    - Vast quantities of real-time information are starting to come from wireless monitoring devices that postoperative patients and those with chronic diseases are wearing at home and in their daily lives.
  - How big data analytics can help:
    - Epidemic early warning
    - Intensive Care Unit and remote monitoring



Source - http://www.ibm.com/software/data/infosphere/use-cases

Big data success story

Children's hospital strives to accelerate health warning alerts for premature infants

- **USD150,000 reduction** in the cost of unnecessary neonatal surgery
- **USD24,000 reduction** in hospital costs per neonatal bed per year
- **24 hours advance** in real-time alerts to begin antibiotics and other treatment for premature infants

Source - http://www.ibm.com/software/data/infosphere/use-cases

Use Cases for a Big Data Platform

- Transportation services
  - Problem:
    - Traffic congestion has been increasing worldwide as a result of increased urbanization and population growth reducing the efficiency of transportation infrastructure and increasing travel time and fuel consumption.
  - How big data analytics can help:
    - Real time analysis to weather and traffic congestion data streams to identify traffic patterns reducing transportation costs.

Source - http://www.ibm.com/software/data/infosphere/use-cases

Big data success story

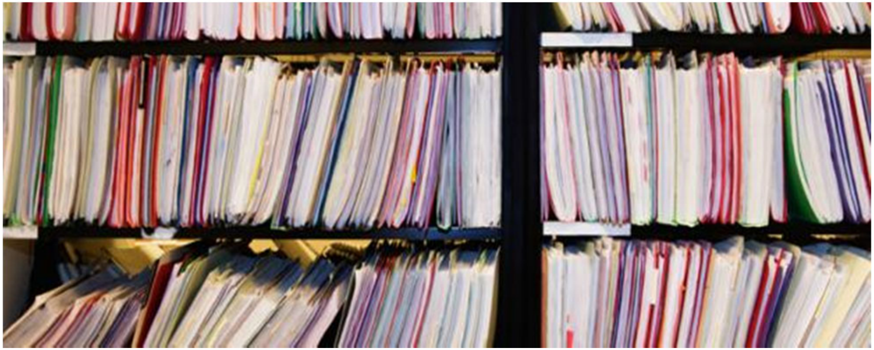
FleetRisk Advisors helps trucking operators prevent accidents by building stronger and faster risk prediction models

- **80% reduction** in serious accidents
- **20% reduction** in minor accidents
- **30% increase** in driver retention rates

Source - http://www.ibm.com/software/data/infosphere/use-cases

Use Cases for a Big Data Platform

Life insurer in Japan pays customers with greater speed and accuracy using content analytics to standardize medical terms



- **22% fewer** mistakenly unpaid claims—from 435 cases to 339 in the first year
- **90% accuracy** in coding medical terms and treatments during claim assessment
- **20% reduction** in assessment workforce, saving several hundred million yen each year

Source - http://www.ibm.com/software/data/infosphere/use-cases

Use Cases for a Big Data Platform

Vestas turns climate into capital, optimizing turbine placement with big data

- **97% faster** response times for wind forecasting information
- **40% reduction** in energy consumption
- **Cuts cost per kilowatt hour** increasing ROI

Source - http://www.ibm.com/software/data/infosphere/use-cases

Use Cases for a Big Data Platform

- Financial services
  - Problem:
    - Manage the several Petabytes of data which is growing at 40-100% per year under increasing pressure to prevent frauds and complain to regulations.
  - How big data analytics can help:
    - Fraud detection
    - Risk management
    - 360°View of the Customer

Source - http://www.ibm.com/software/data/infosphere/use-cases

Use Cases for a Big Data Platform

- Telecommunication services
  - Problem:
    - Legacy systems are used to gain insights from internally generated data facing issues of high storage costs, long data loading time, and long administration process.

  - How big data analytics can help:
    - CDR processing
    - Churn prediction
    - Geomapping / marketing
    - Network monitoring

Source - http://www.ibm.com/software/data/infosphere/use-cases

IBM's Big Data Success story

Bloomberg

IBM CEO Says 'Big Data' Is Company's Top Priority

- $16 Billion in Big Data acquisitions – 35 new acquisitions in the last 5 years
- More than 1000 developers focused on Big Data technology development
- **2014 IBM joins Apple & Twitter in strategic partnerships**

- Largest patent portfolio in the industry
- IBM has the largest commercial research organization on Earth
  - 200+ mathematicians developing breakthrough analytics
- IBM's Big Data Business grew over 150% in 2014

Checkpoint (1 of 2)

1. Big data generally refers to

   a. Voluminous structured data

   b. Voluminous unstructured data

   c. Voluminous semi structured data

   d. Both b & c

2. The three V’s in Big data refers to

   a. Velocity

   b. Volume

   c. Variety

   d. All the above

Checkpoint solution (1 of 2)

1. Big data generally refers to

   a. Voluminous structured data

   b. Voluminous unstructured data

   c. Voluminous semi structured data

   d. Both b & c

2. The three V's in Big data refers to

   a. Velocity

   b. Volume

   c. Variety

   d. All the above

Checkpoint (2 of 2)

3. Big Data Analytics Adoption structure is

    a. Educate ->Explore -> Engage -> Execute

    b. Explore -> Educate -> Engage -> Execute

    c. Explore -> Engage -> Execute

    d. Explore -> Educate -> Execute -> Engage

4. Benefits of Big data includes the following

    a. Analytics

    b. Business Intelligence

    c. Handling Volumes and data

    d. All the above

5. The barriers to Big Data includes

    a. Lack of skilled staff

    b. Lack of Sufficient Knowledge

    c. Lack of software to handle the volume of data

    d. Both A & B

Checkpoint solution (2 of 2)

3. Big Data Analytics Adoption structure is

   a. Educate ->Explore -> Engage -> Execute

   b. Explore -> Educate -> Engage -> Execute

   c. Explore -> Engage -> Execute

   d. Explore -> Educate -> Execute -> Engage

4. Benefits of Big data includes the following

   a. Analytics

   b. Business Inteligence

   c. Handling Volumes and data

   d. All the above

5. The barriers to Big Data includes

   a. Lack of skilled staff

   b. Lack of Sufficient Knowledge

   c. Lack of software to handle the volume of data

   d. Both A & B

Unit Summary

Having completed this unit, you should be able to:

- Understand Need of Big Data
- Understand Big Data – Concept, Characteristics and Dimensions
- Understand Fundamentals of Big Data Architecture
- Understand Big Data Tools and Techniques
- Understand Big Data - Applications