BIG DATA ANALYSIS

PRESENTED BY:

P. SHARMILA

M.VANITHA

A.VIMALADEVI

Theme of this Course

Large-Scale Data Management

Big Data Analytics

Data Science and Analytics

- How to manage very large amounts of data and extract value and knowledge from them

Big Data Definition

- No single standard definition…

"*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…
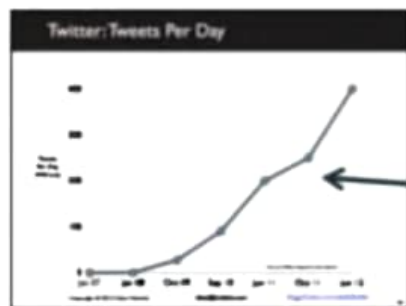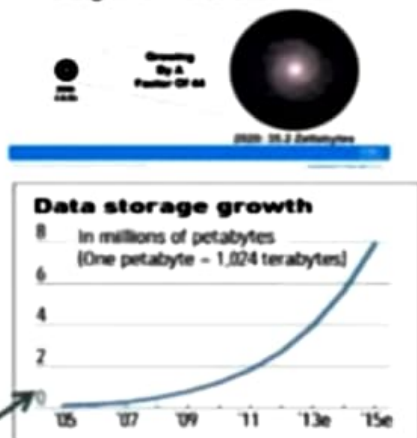
Characteristics of Big Data:
1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb

- Data volume is increasing exponentially

The Digital Universe 2009-2020

Growing
By A
Factor Of 44

2020: 35.2 Zettabytes

Data storage growth

In millions of petabytes
(One petabyte – 1,024 terabytes)

| terabytes | petabytes | exabytes | zettabytes |

*the amount of data stored by the average company today*

Twitter: Tweets Per Day

Exponential increase in collected/generated data

Characteristics of Big Data:
2-Complexity (Varity)

- 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

To extract knowledge➔ all these types of data need to linked together
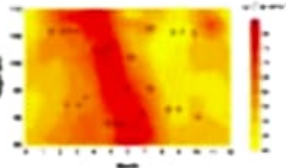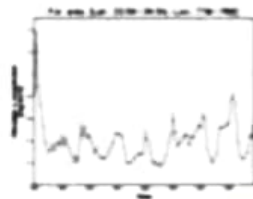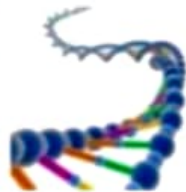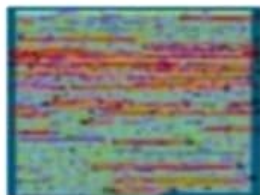
Characteristics of Big Data:
3-Speed (Velocity)

- 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

Big Data: 3V's

Big Data = Transactions + Interactions + Observations
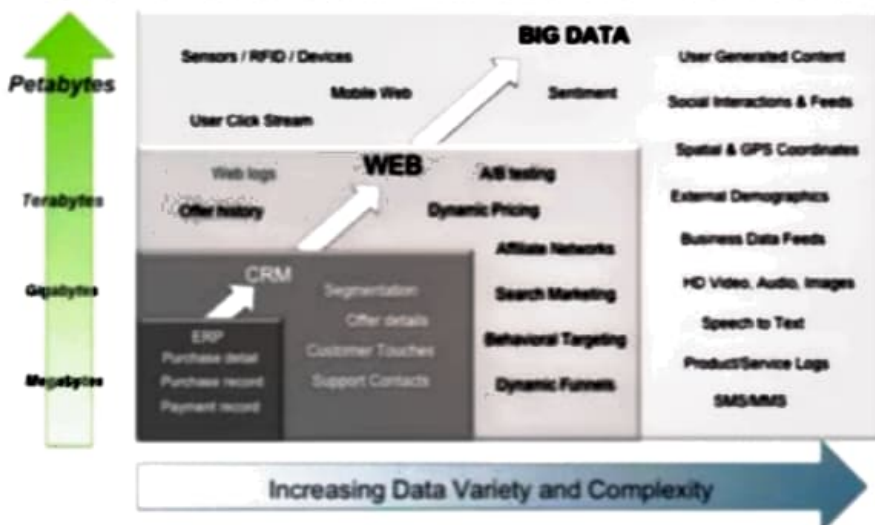
Some Make it 4V's

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
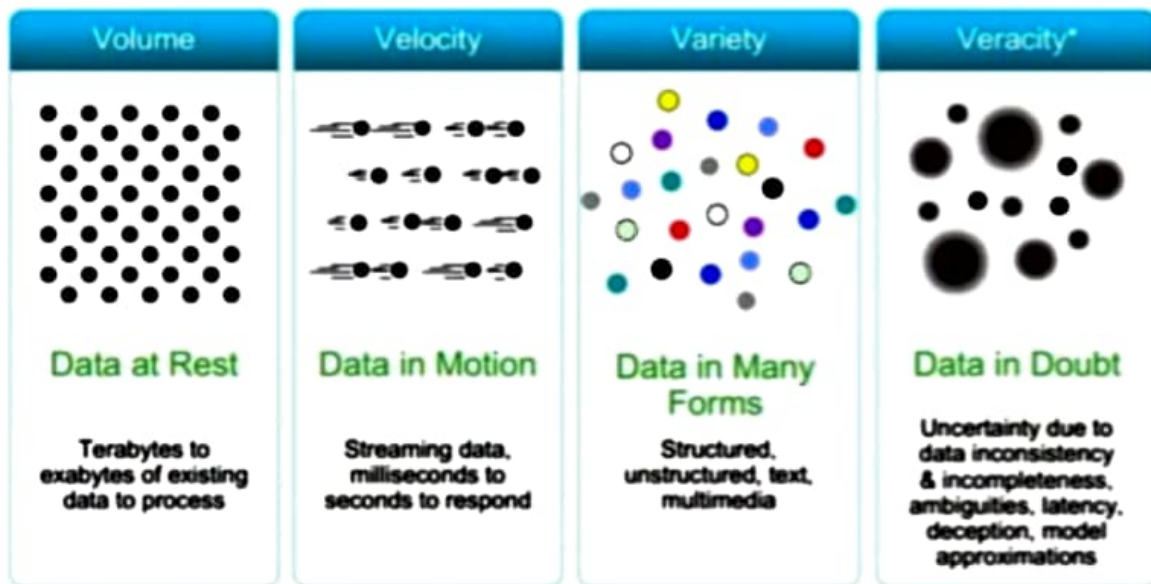
Harnessing Big Data
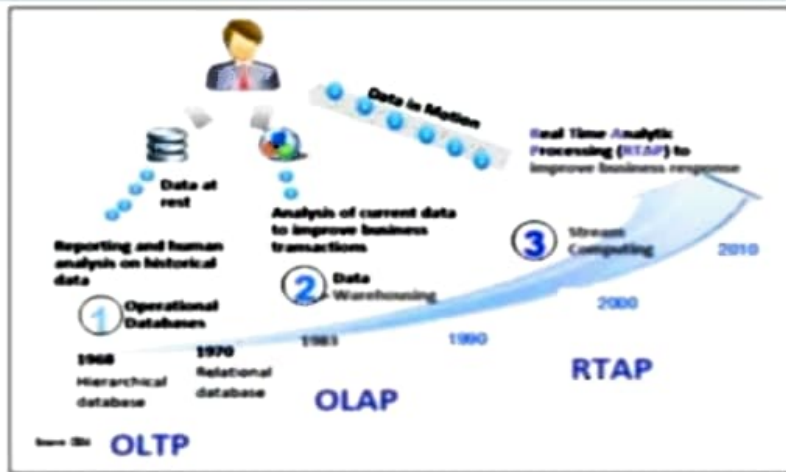
- **OLTP:** Online Transaction Processing   (DBMSs)

- **OLAP:** Online Analytical Processing   (Data Warehousing)

- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

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 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'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

HIGH

Predictive Analytics and Data Mining

COMPLEXITY

Business Intelligence

LOW

BUSINESS VALUE

HIGH

Value of 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
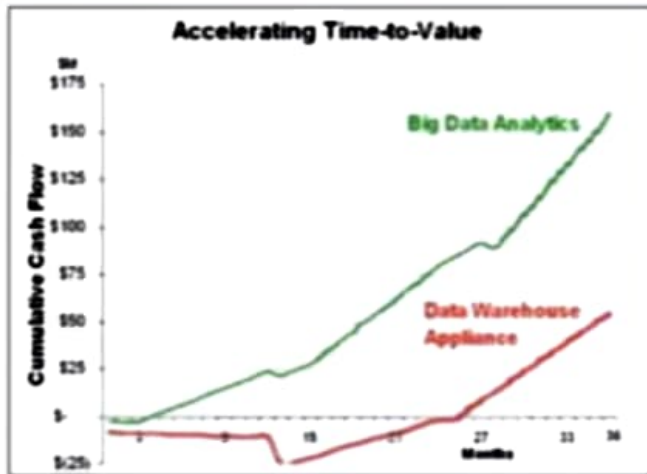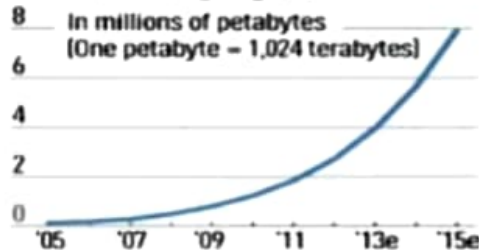

Accelerating Time-to-Value

Cumulative Cash Flow

Big Data Analytics

Data Warehouse Appliance

Challenges in Handling Big Data

Big Data Boom

Data storage growth

8 — In millions of petabytes
(One petabyte = 1,024 terabytes)
6
4
2
0
'05 '07 '09 '11 '13e '15e

Sources: IDC, DataXu

Big data challenge

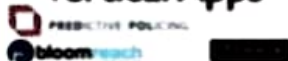|	
Lack of software/technology	30%
Lack of analytic skills	28%
Insufficient budget	25%
Already using	11%
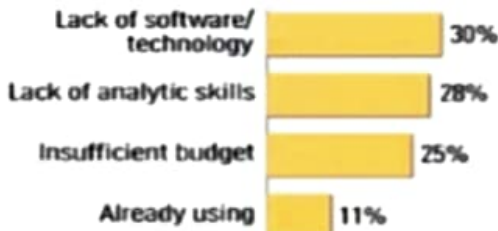
- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed

- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

What Technology Do We Have
For Big Data ??

Big Data Landscape

Vertical Apps
PREDICTIVE POLICING
bloomreach

Log Data Apps
splunk   loggly   sumologic

Ad/Media Apps
rocketfuel
Media Science   TURN   Data.xu

Data As A Service
factual
GNIP   DATASF1   INRIX   LexisNexis   knoema   LOCATE

Business Intelligence
ORACLE   Hyperion
SAP   Business Objects
Microsoft   Business Intelligence
IBM   COGNOS   birst
Autonomy   MicroStrategy
QlikView   birne
Chart.io   GoodData

Analytics and Visualization
tableau   Palantir
metaLayer   dataspora
METAMARKETS   centrifuge
TERADATA ASTER
SAS   TIBCO   KARMASPHERE
panopticon
Datameer
platfora
alteryx   visual.ly

Analytics Infrastructure
Hortonworks   VERTICA   MAPR
cloudera   INFOBRIGHT
ParAccel
EMC   GREENPLUM
NETEZZA   kognitio
DATASTAX   EXASOL

Operational Infrastructure
COUCHBASE   10gen
TERADATA   HADAPT
TERRACOTTA   VoltDB
MarkLogic   INFORMATICA

Infrastructure As A Service
amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases
ORACLE   MySQL
SQL Server   PostgreSQL
IBM   DB2
SYBASE

Technologies
hadoop   hadoop mapreduce   mahout   HBASE   Cassandra

Big Data Technology

Big Data: The Moving Parts

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

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

THANK YOU