Data Science
(CS3206)

Lecture #1

Introduction

Quote of the day..

“We are what our thoughts have made us; so take care of what you think. Words are secondary. Thoughts live; they travel far.”

Swami Vivekananda

In today's discussion…

- Introduction to data
- Current trend
- Data and Big data
- Big data vs. small data
- Tools and techniques

Ref:https://cse.iitkgp.ac.in/~dsamanta/courses/da/index.html

Introduction to data

- Example:

  10, 25, …, Kharagpur, 10CS3002, namo@gov.in

  Anything else?

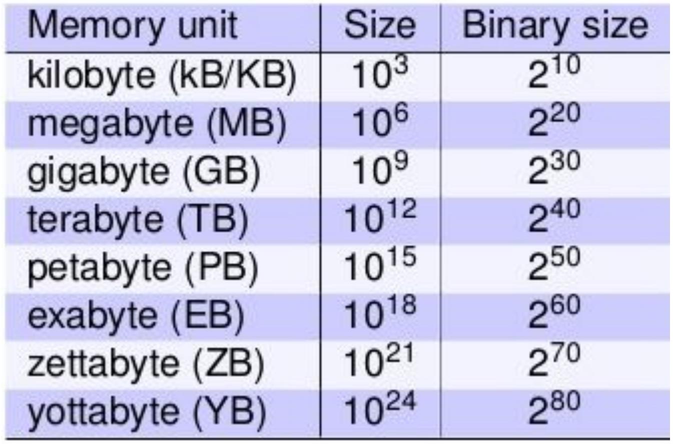
- Data vs. Information

  100.0, 0.0, 250.0, 150.0, 220.0, 300.0, 110.0

  Is there any information?

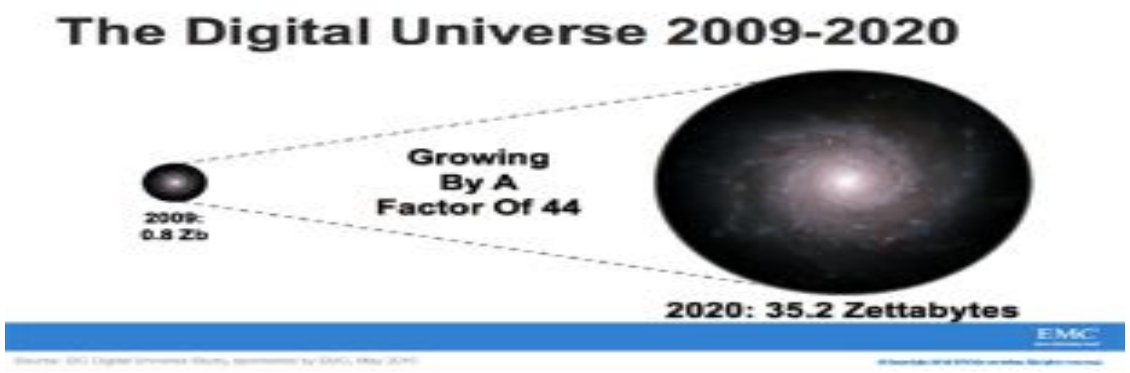
How large your data is?

- What is the maximum file size you have dealt so far?
  - Movies/files/streaming video that you have used?
- What is the maximum download speed you get?
  - To retrieve data stored in distant locations?
- How fast your computation is?
  - How much time to just transfer from you, process and get result?

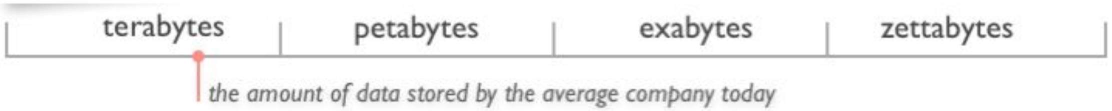


Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Growth of data



Source: IDC Digital Universe Study, sponsored by EMC, May 2010



Sources of data

- “Every day, we create 2.5 quintillion bytes of data
  - So much that 90% of the data in the world today has been created in the last two years alone.
  - The data come from several sources
    - sensors used to gather climate information
    - posts to social media sites,
    - digital pictures and videos
    - purchase transaction records
    - cell phone GPS signals

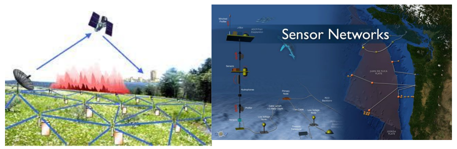
etc. …… to name a few!

Examples

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)

Now data is Big data!

- No single standard definition!

- ‘Big-data’ is similar to ‘Small-data’, but bigger

  …but having data bigger consequently requires different approaches

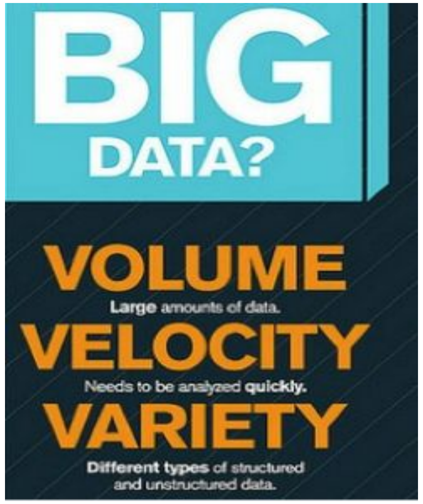
  - techniques, tools and architectures

  …to solve: new problems

  …and, of course, in a better way

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: **V3**



BIG DATA?

VOLUME

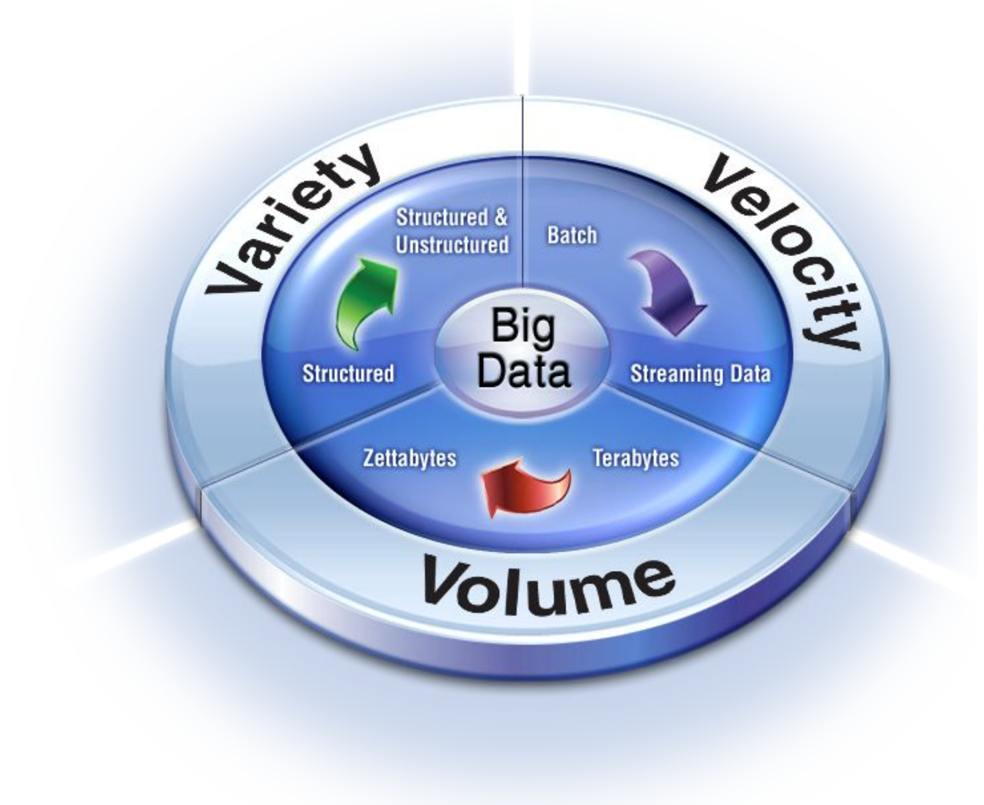
Large amounts of data.

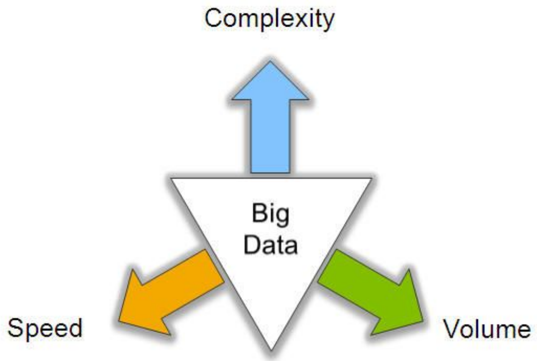
VELOCITY

Needs to be analyzed quickly.

VARIETY

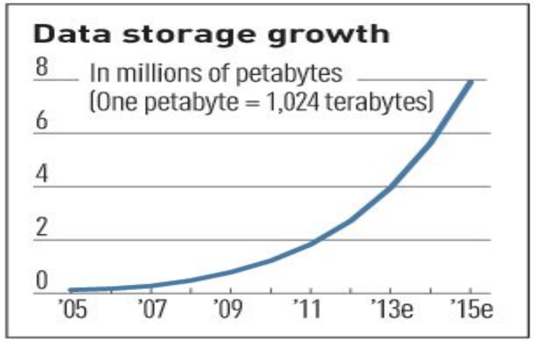
Different types of structured and unstructured data.

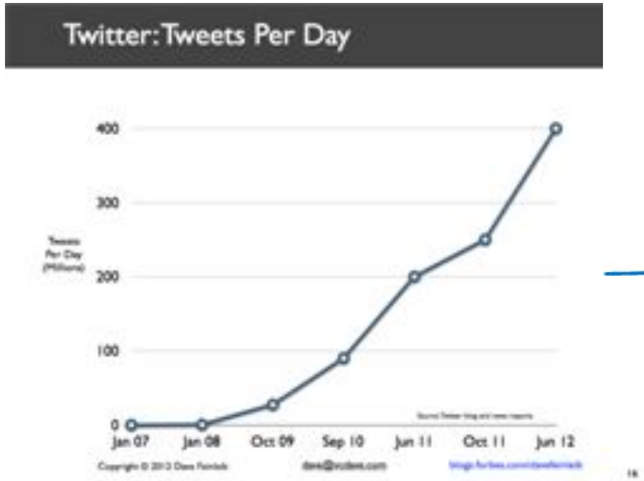




V3 : V for Volume

- Volume of data, which needs to be processed is increasing rapidly
  - More storage capacity
  - More computation
  - More tools and techniques



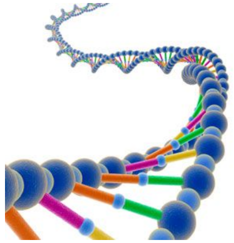


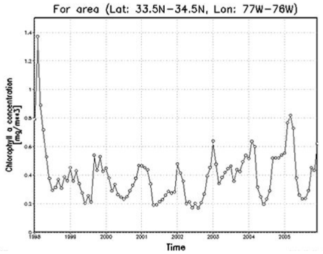
Exponential increase in collected/generated data

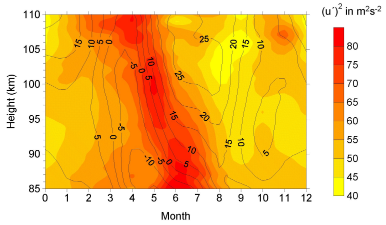
V3: V for Variety

- Various formats, types, and structures
  - Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional 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 be linked together



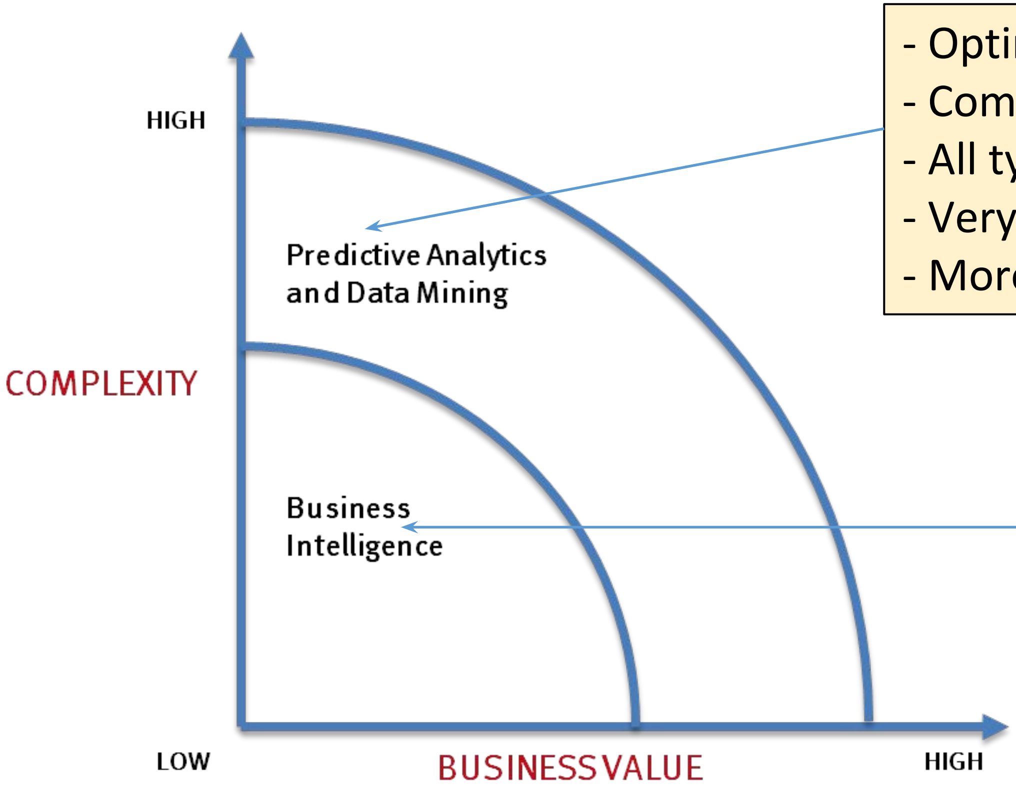




V3: V for Velocity

- Data is being generated fast and need to be processed fast
    - For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value
    - Scrutinize 5 million trade events created each day to identify potential fraud
    - Analyze 500 million daily call detail records in real-time to predict customer churn faster
- Sometimes, 2 minutes is too late!
    - The latest we have heard is 10 ns (nano seconds) delay is too much

Big data vs. small data



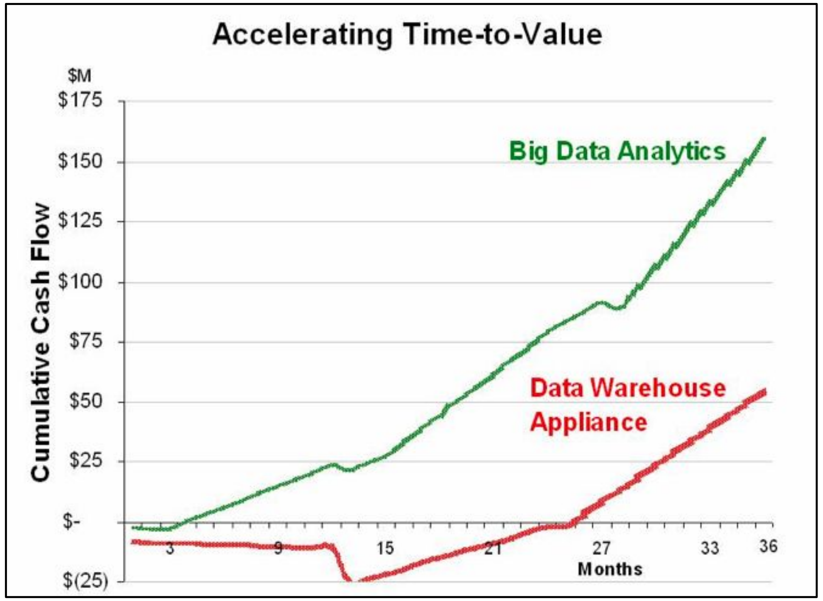
- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

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

Big data vs. small data

- Big data is more real-time in nature than traditional applications

- Big data architecture
  - Traditional architectures are not well-suited for big data applications (e.g. Exa-data, Tera-data)
  - Massively parallel processing, scale out architectures are well-suited for big data applications

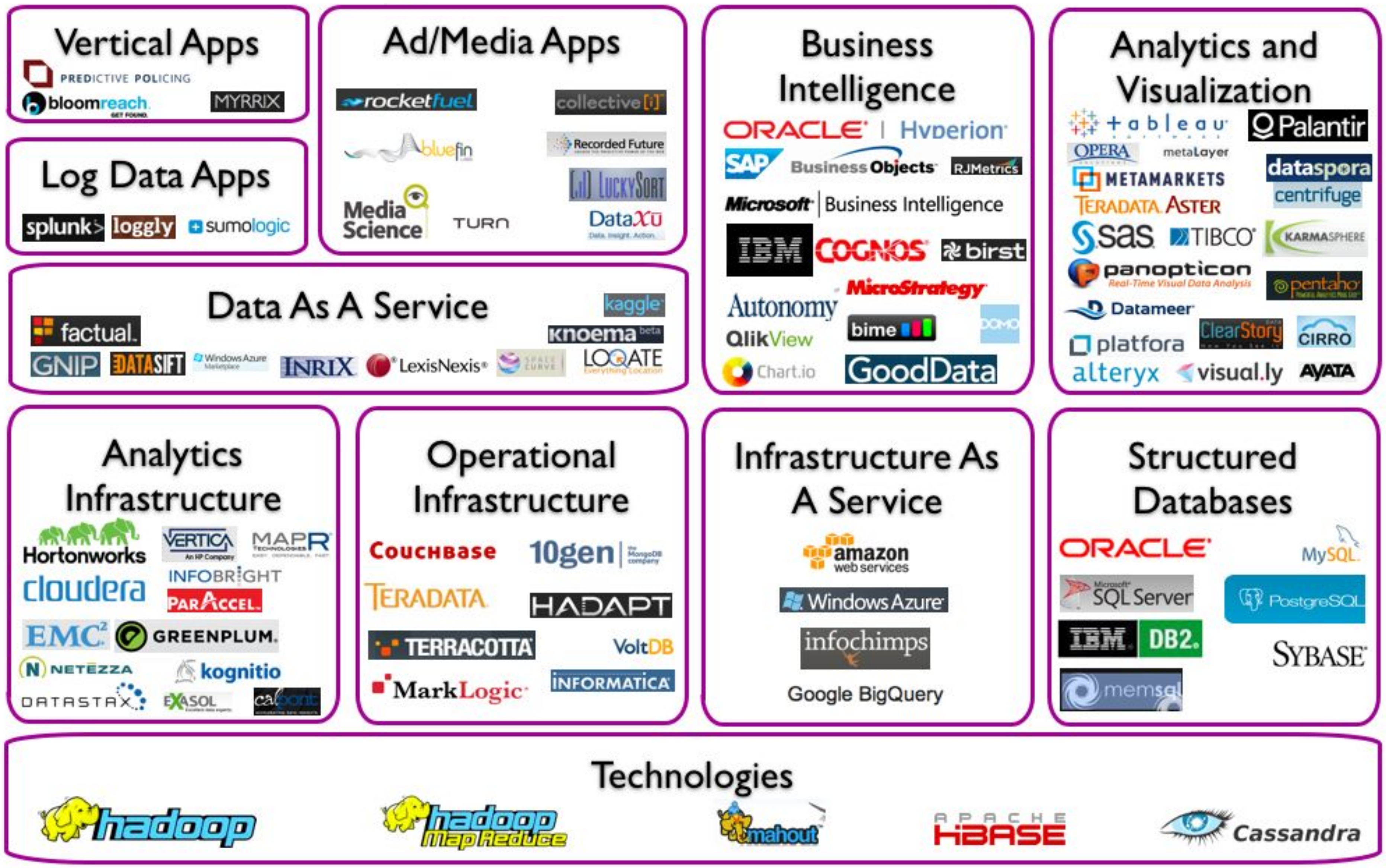


Challenges ahead…

- **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

Who are the major players in the world of Big data?

Big Data Landscape



Vertical Apps

Log Data Apps

Ad/Media Apps

Data As A Service

Business Intelligence

Analytics and Visualization

Analytics Infrastructure

Operational Infrastructure

Infrastructure As A Service

Google BigQuery

Structured Databases

Technologies

dave@vcdave.com

blogs.forbes.com/davefeinleib

Major players…

- Google
- Hadoop
- MapReduce
- Mahout
- Apache Hbase
- Cassandra

Tools available

- **NoSQL**
  - DatabasesMongoDB, 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, HDFS, GDFS
- **Servers**
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- **Processing**
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Any question?

Questions of the day…

1. What is the smallest and largest units of measuring size of data?

2. How big a Quintillion measure is?

3. Give the examples of a smallest the largest entities of data.

4. Give FIVE parameters with which data can be categorized as i) simple, ii) Moderately complex and iii) complex?

Questions of the day…

5. What type of data are involved in the following applications?

1. Weather forecasting
2. Mobile usage of all customers of a service provider
3. Anomaly (e.g. fraud) detection in a bank organization
4. Person categorization, that is, identifying a human
5. Air traffic control in an airport
6. Streaming data from all flying aircrafts of Boeing