CISC 432/CMPE 432/CISC 832 Advanced Database Systems

Course Info

Instructor:	Patrick Martin Goodwin Hall 630 613 533 6063 martin@cs.queensu.ca Office Hours: Wednesday 11:00 – 12:00 or by appointment
Schedule:	Tuesday 11:30 - 12:30 Botterell Hall 143 Wednesday 1:30 - 2:30 Botterell Hall 143 Friday 12:30 - 1:30 Botterell Hall 147
TAs:	Nafiseh Kahani (kahani@cs.queensu.ca) Reza Ahmadi (ahmadi@cs.queensu.ca)

Assumed Background

- CISC 332, previous course in DBMS or equivalent experience
- Relational model, SQL, relational algebra, schema design
- File structures, indexes (tree and hash)

Reference Materials

Textbook (recommended):

Database System Concepts(6th Edition) by A. Silbershatz, H. Korth and S. Sudarshan, McGraw-Hill.

(a copy is on reserve in the library)

Research papers

Links will be provided on OnQ.

Learning Outcomes

- Apply optimization algorithms to SQL queries to produce efficient query plans.
- Apply concurrency control and recovery algorithms to sample transaction workloads to ensure ACID properties are maintained.
- Assess the use of relational DBMSs and NoSQL systems for different types of data and applications.
- Apply a NoSQL system to the creation of a sample database.
- Apply the MapReduce framework to a sample big data problem.
- Evaluate the use of a big data approach to a sample application area or problem.

Marking Schemes

- **Undergraduate students (432)**
  - 3 assignments (60 %).
  - 2 term tests (40%).

Marking Schemes (Cont.)

- **Graduate students (832)**
  - 3 assignments (45%).
  - 2 term tests (30%).
  - Term paper (25%).

Marking Schemes (Cont.)

- **Grading Method**
  - In this course, some components will be graded using numerical percentage marks. Other components will receive letter grades, which for purposes of calculating your course average will be translated into numerical equivalents using the Faculty of Arts and Science approved scale (see below). Your course average will then be converted to a final letter grade according to Queen's Official Grade Conversion Scale.
- **Late Policy**
  - Assignments should be handed in by 4:00 pm on the day they are due. Late assignments are subject to a 10% per day late penalty, with weekends counted as one day. Late assignments will not be accepted beyond 5 days past the date due.

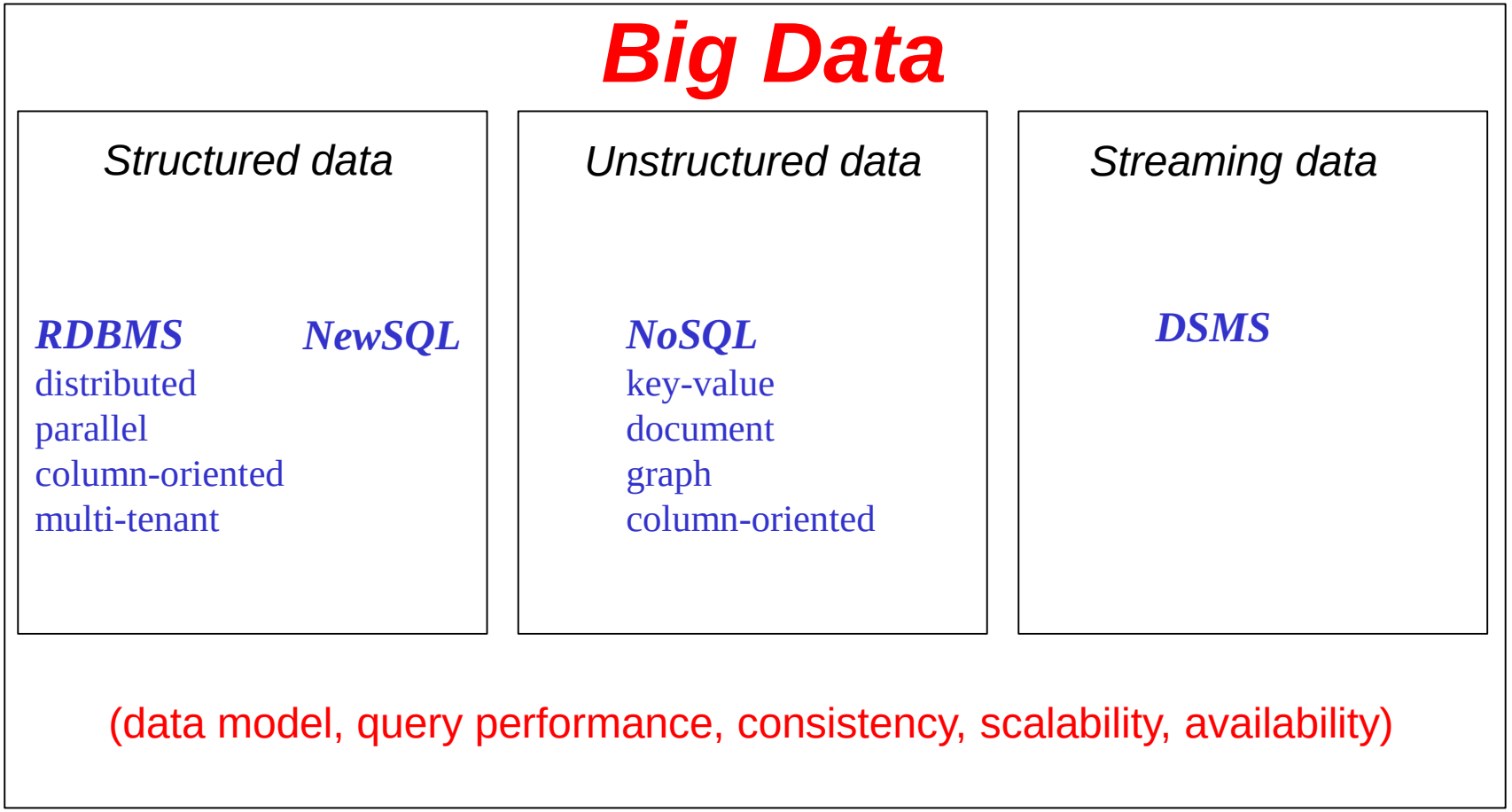
Requirements Schedule

Requirement	**Due Date**
Assignment 1	October 4
Grad Paper Proposal (832 students only)	October 14
Term test 1	Oct 18
Assignment 2	Nov 1
Term test 2	Nov 25
Assignment 3	Dec 2
Grad Research Paper (832 students only)	Dec 9

Academic Integrity

- Academic integrity is constituted by the five core fundamental values of honesty, trust, fairness, respect and responsibility (see www.academicintegrity.org). These values are central to the building, nurturing and sustaining of an academic community in which all members of the community will thrive. Adherence to the values expressed through academic integrity forms a foundation for the "freedom of inquiry and exchange of ideas" essential to the intellectual life of the University (see the Senate Report on Principles and Priorities)
- Students are responsible for familiarizing themselves with the regulations concerning academic integrity and for ensuring that their assignments conform to the principles of academic integrity. Information on academic integrity is available in the Arts and Science Calendar (see Academic Regulation 1), on the Arts and Science website (see http://www.queensu.ca/artsci/sites/default/files/Academic%20Regulations.pdf), and from the instructor of this course.
- Departures from academic integrity include plagiarism, use of unauthorized materials, facilitation, forgery and falsification, and are antithetical to the development of an academic community at Queen's. Given the seriousness of these matters, actions which contravene the regulation on academic integrity carry sanctions that can range from a warning or the loss of grades on an assignment to the failure of a course to a requirement to withdraw from the university.

Course Content

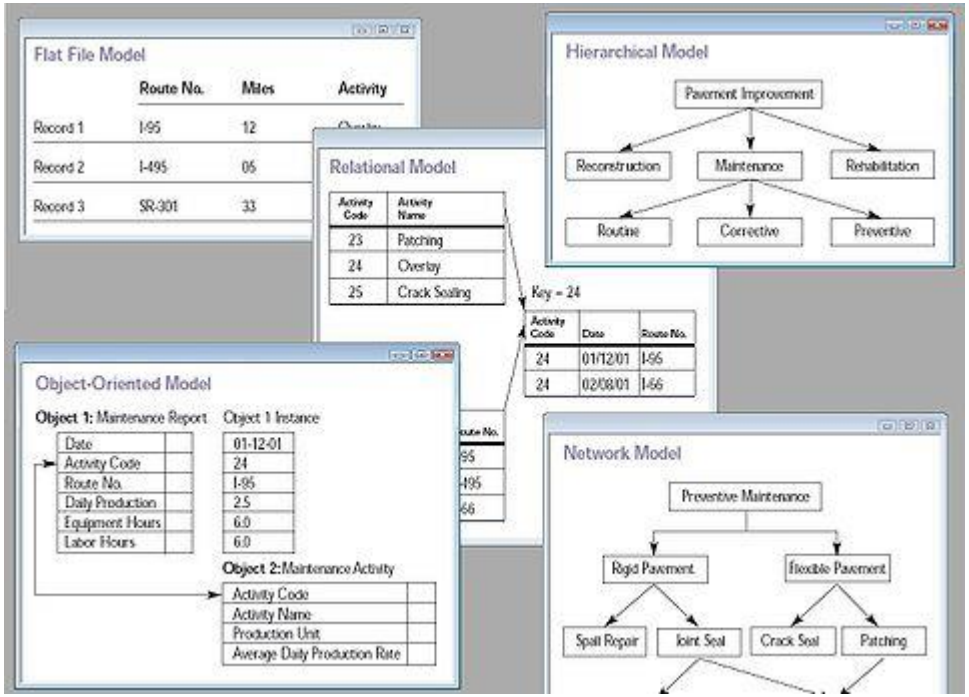


Course Schedule

- **Week 1:** Course information; introductions to big data, cloud computing
- **Week 2 - 3:** Bluemix tutorial; RDBMS review
- **Week 4 - 5:** RDBMS architectures
- **Week 6:** Distributed storage systems
- **Week 7:** NoSQL systems
- **Week 8:** Hadoop ecosystem & Map-Reduce
- **Week 9:** BigSQL / BigInsights
- **Week 10 - 12:** Big data topics

Focus Issues

- **Data model:**
  - language and logical constructs that determine the logical structure of a database and consequently how data is stored, organized, and manipulated.



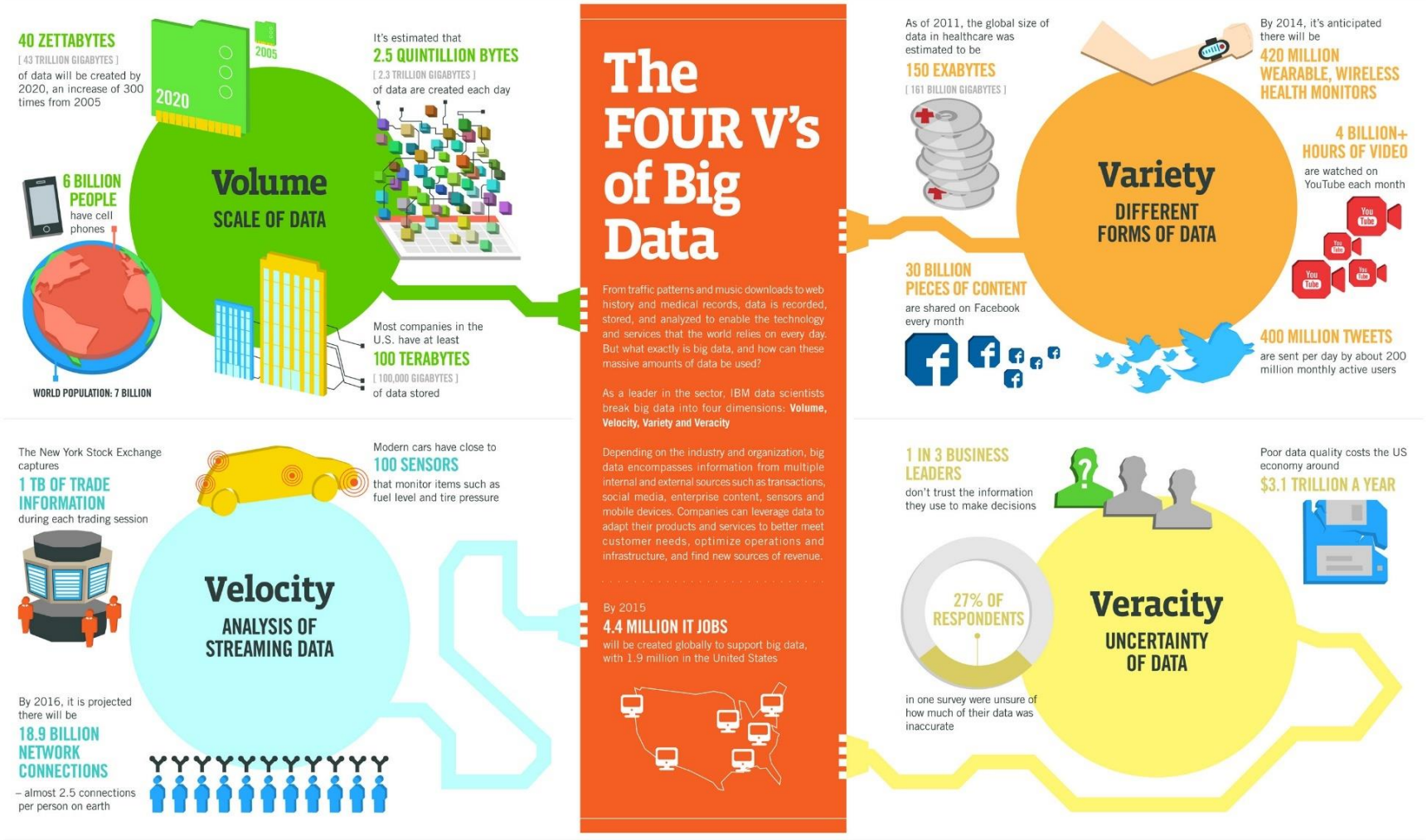
Focus Issues (cont)

- **Query performance**
  - Efficiency of a DBMS typically expressed with metrics like query response time and/or query throughput
- **Consistency**
  - Guarantee concerning state of a data item when accessed
  - Eg ACID, BASE

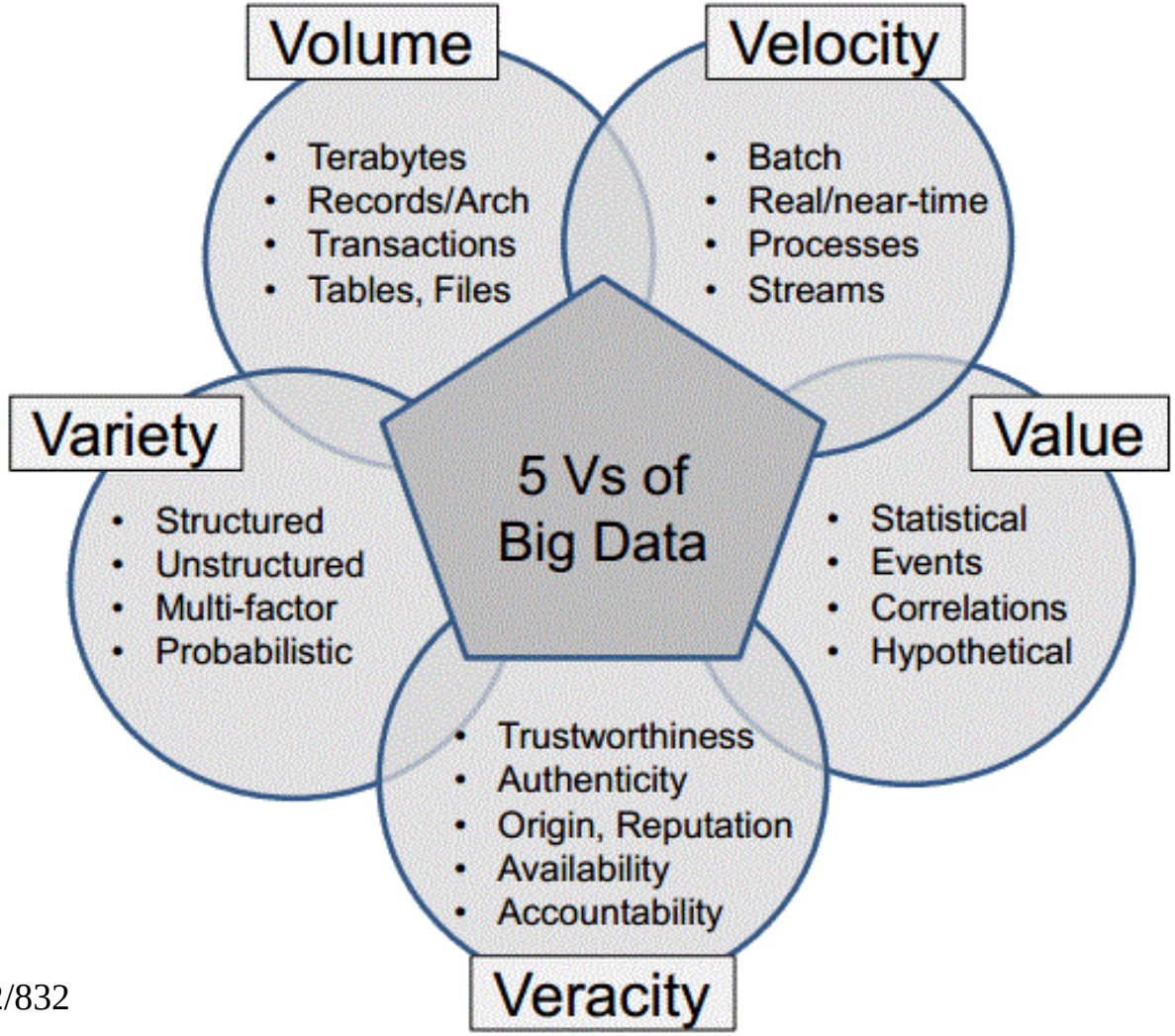
Focus Issues (cont)

- **Scalability**
  - Ability of a system to handle a growing amount of work in a capable manner or its ability to be enlarged to accommodate that growth
- **Availability**
  - Proportion of time that requests received by a system receive a response.

Big Data

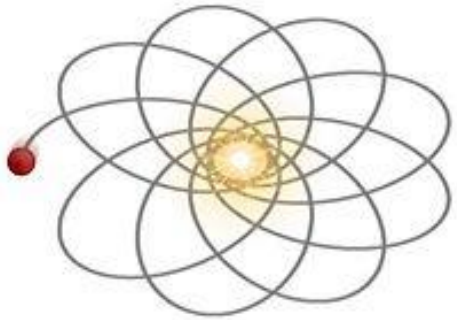

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Big Data – One More V



What is Analytics?

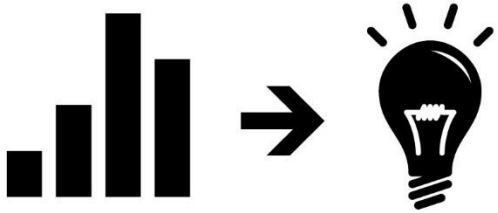
Mathematical or Scientific methods that highlight data for insight



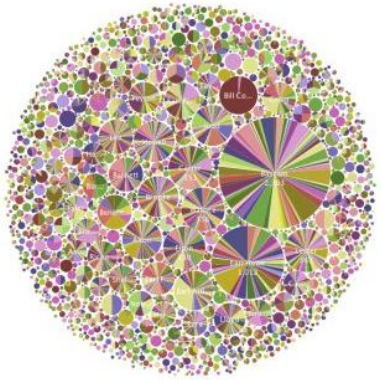
$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = r.$$

Insight = Competitive Advantage

Used to inform actions and decisions



Data is becoming the world's new natural resource



With analytics, insights are created to augment the gut feelings and intuition for decisions

Related Buzzwords

- Cloud computing
  - "A networking solution in which everything — from computing power to computing infrastructure, applications, business processes to personal collaboration — is delivered as a service wherever and whenever you need."

Related Buzzwords

- NoSQL (Not Only SQL)
  - NoSQL encompasses a wide range of technologies and architectures not based on the relational model that seeks to solve the scalability and big data performance issues of relational databases.
  - Eg Cassandra, BigTable, SimpleDB, CouchDB MongoDB, Voldemort, Neo4j

Related Buzzwords

- **NewSQL**
  - Encompasses solutions aimed at bringing to the relational model the benefits of horizontal scalability and fault tolerance provided by NoSQL solutions
  - Eg. Google Spanner, VoltDB