Introduction to Big Data and Analytics

University of California, Berkeley

School of Information

INFO 257: Database Management

Lecture Outline

- Big Data (introduction)
- Analytics (introduction)
  - https://www.ischool.berkeley.edu/courses/info/154
  - https://www.ischool.berkeley.edu/courses/info/247

Introduction

- **Big Data**
  - Data that exist in very large volumes and many different varieties (data types) and that need to be processed at a very high velocity (speed).
- **Analytics**
  - Systematic analysis and interpretation of data—typically using mathematical, statistical, and computational tools—to improve our understanding of a real-world domain.

Big Data and Databases

- "640K ought to be enough for anybody."
  - Attributed to Bill Gates, 1981

Big Data and Databases

- We have already mentioned some Big Data
  - The Walmart Data Warehouse
  - Information collected by Amazon on users and sales and used to make recommendations
- Most modern web-based companies capture EVERYTHING that their customers do
  - Does that go into a Warehouse or someplace else?

Other Examples

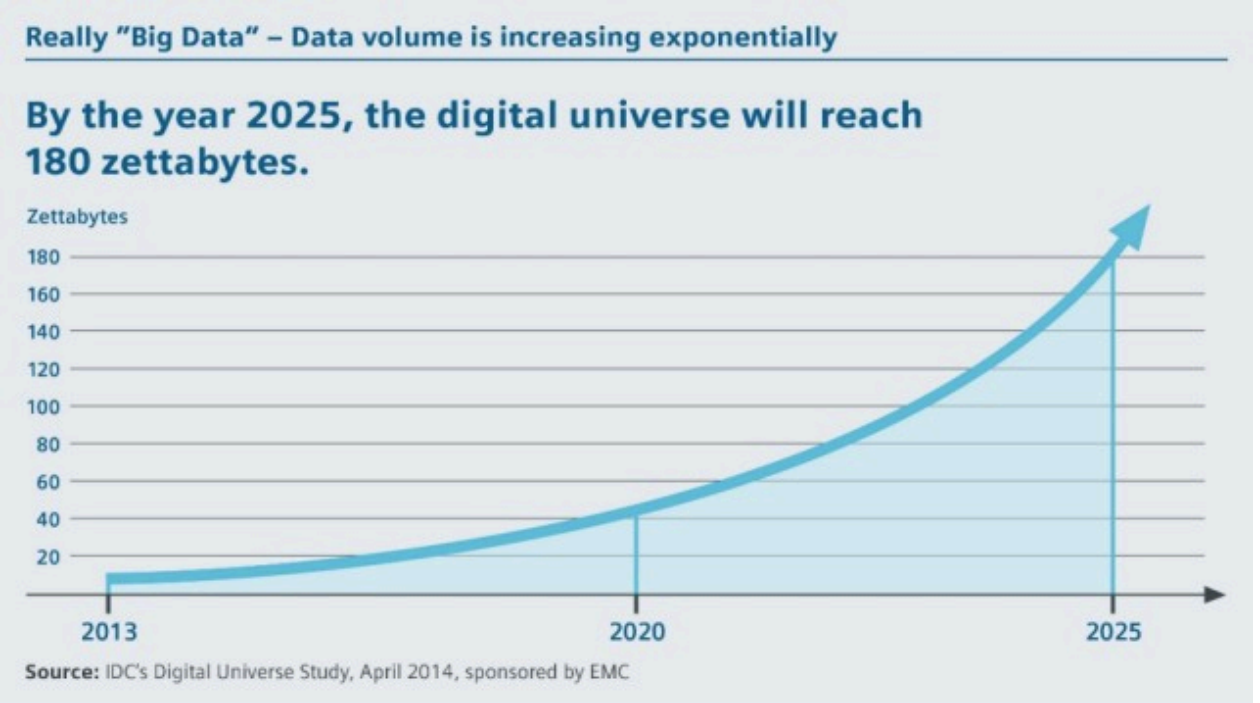
- NASA EOSDIS
  - Estimated 10^{18} Bytes (Exabyte)
- Computer-Aided design
- The Human Genome
- Department Store tracking
  - Mining non-transactional data (e.g. Scientific data, text data?)
- Insurance Company
  - Multimedia DBMS support

Table 1.1: How Big is an Exabyte?	
Kilobyte (KB)	*1,000 bytes OR 10^3 bytes* 2 Kilobytes: A Typewritten page. 100 Kilobytes: A low-resolution photograph.
Megabyte (MB)	*1,000,000 bytes OR 10^6 bytes* 1 Megabyte: A small novel OR a 3.5 inch floppy disk. 2 Megabytes: A high-resolution photograph. 5 Megabytes: The complete works of Shakespeare. 10 Megabytes: A minute of high-fidelity sound. 100 Megabytes: 1 meter of shelved books. 500 Megabytes: A CD-ROM.
Gigabyte (GB)	*1,000,000,000 bytes OR 10^9 bytes* 1 Gigabyte: a pickup truck filled with books. 20 Gigabytes: A good collection of the works of Beethoven. 100 Gigabytes: A library floor of academic journals.
Terabyte (TB)	*1,000,000,000,000 bytes OR 10^{12} bytes* 1 Terabyte: 50000 trees made into paper and printed. 2 Terabytes: An academic research library. 10 Terabytes: The print collections of the U.S. Library of Congress. 400 Terabytes: National Climactic Data Center (NOAA) database.
Petabyte (PB)	*1,000,000,000,000,000 bytes OR 10^{15} bytes* 1 Petabyte: 3 years of EOS data (2001). 2 Petabytes: All U.S. academic research libraries. 20 Petabytes: Production of hard-disk drives in 1995. 200 Petabytes: All printed material.
Exabyte (EB)	*1,000,000,000,000,000,000 bytes OR 10^{18} bytes* 2 Exabytes: Total volume of information generated in 1999. 5 Exabytes: All words ever spoken by human beings.

Source: Many of these examples were taken from Roy Williams "Data Powers of Ten" web page at Caltech.

Digitization of Everything: the Zettabytes are here

- **Soon most everything will be recorded and indexed**
- **Much will remain local**
- **Most bytes will never be seen by humans.**
- **Search, data summarization, trend detection, information and knowledge extraction and discovery are key technologies**
- **So will be infrastructure to manage this.**

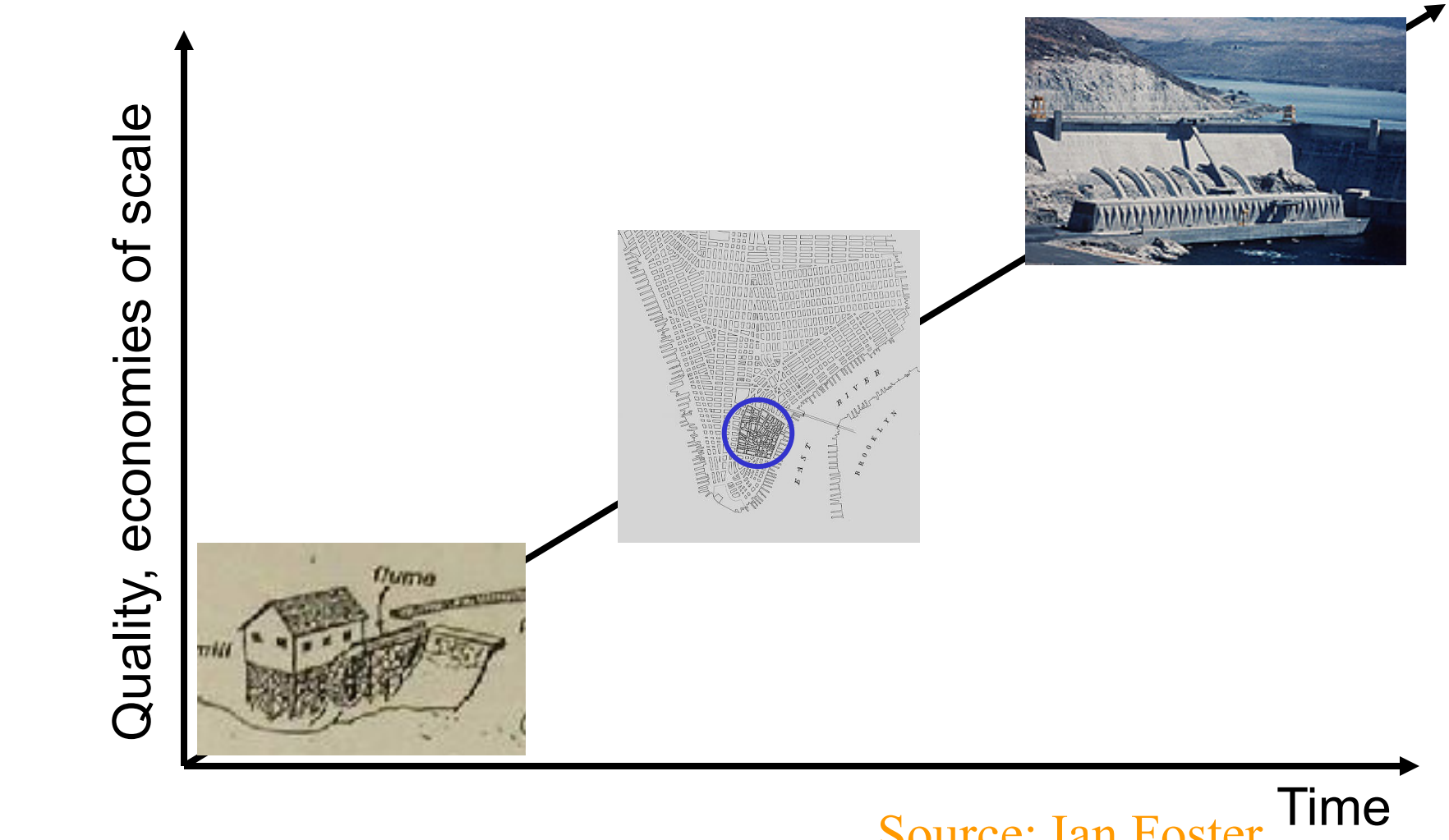


Before the Cloud there was the Grid

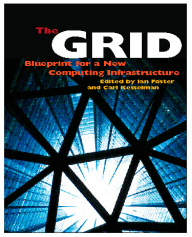
- So what's this Grid thing anyhow?
- Data Grids and Distributed Storage
- Grid vs "Cloud"

The following borrows heavily from presentations by Ian Foster (Argonne National Laboratory & University of Chicago), Reagan Moore and others from San Diego Supercomputer Center

The Grid: On-Demand Access to Electricity



By Analogy, A Computing Grid



- Decouples production and consumption
  - Enable on-demand access
  - Achieve economies of scale
  - Enhance consumer flexibility
  - Enable new devices
- On a variety of scales
  - Department
  - Campus
  - Enterprise
  - Internet

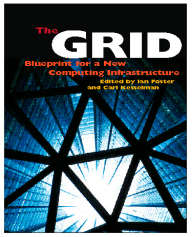
Source: Ian Foster

What is the Grid?

"The short answer is that, whereas the Web is a service for sharing information over the Internet, the Grid is a service for sharing computer power and data storage capacity over the Internet. The Grid goes well beyond simple communication between computers, and aims ultimately to turn the global network of computers into one vast computational resource."

Source: The Global Grid Forum

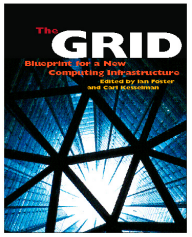
Not Exactly a New Idea …



- "The time-sharing computer system can unite a group of investigators …. one can conceive of such a facility as an … intellectual public utility."
  - Fernando Corbato and Robert Fano , 1966
- "We will perhaps see the spread of 'computer utilities', which, like present electric and telephone utilities, will service individual homes and offices across the country." Len Kleinrock, 1967

Source: Ian Foster

But, Things are Different Now



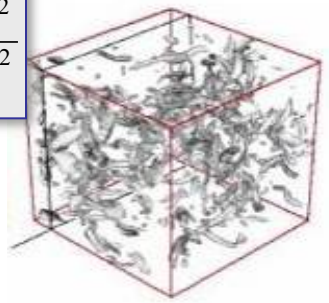
- Networks are far faster (and cheaper)
  - Faster than computer backplanes
- "Computing" is very different than pre-Net
  - Our "computers" have already disintegrated
  - E-commerce increases size of demand peaks
  - Entirely new applications & social structures
- Weve learned a few things about software
- But, the needs are changing too…

Source: Ian Foster

Progress of Science

- Thousand years ago:
  science was **empirical**
    describing natural phenomena
- Last few hundred years:
  theoretical branch
    using models, generalizations
- Last few decades:
  a **computational** branch
    simulating complex phenomena
- Today: **(big data/information)**
  data and information exploration (eScience)

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

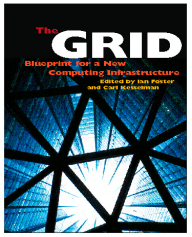


unify theory, experiment, and simulation - information driven

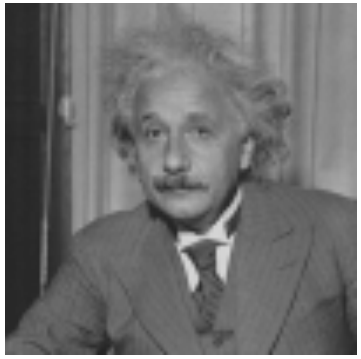
- Data captured by sensors, instruments or generated by simulator
- Processed/searched by software
- Information/Knowledge stored in computer
- Scientist analyzes database / files using data management and statistics
- Network Science
- Cyberinfrastructure

Source: Jim Gray

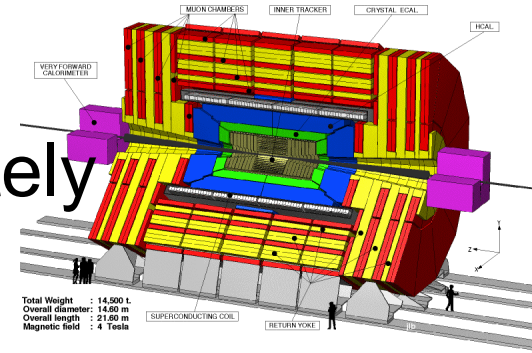
Why the Grid?
(1) Revolution in Science



- Pre-Internet
  - Theorize &/or experiment, alone or in small teams; publish paper



- Post-Internet
  - Construct and mine large databases of observational or simulation data
  - Develop simulations & analyses
  - Access specialized devices remotely
  - Exchange information within distributed multidisciplinary teams

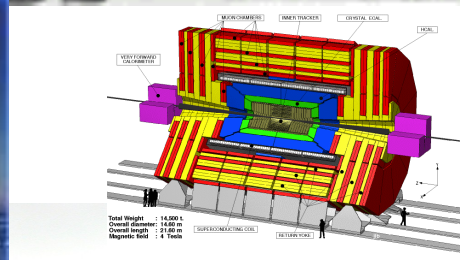


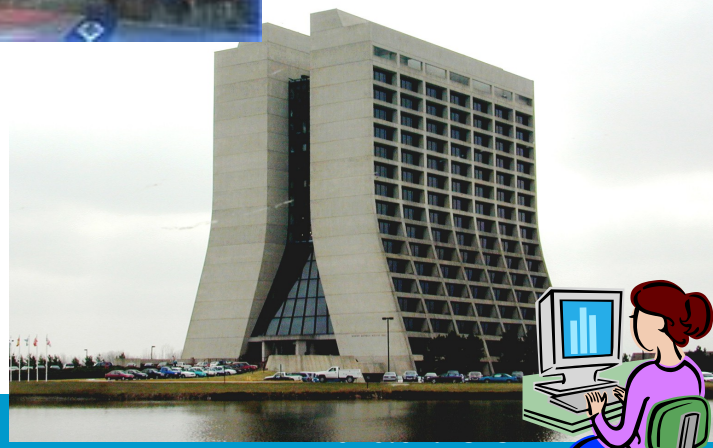
Source: Ian Foster

Computational Science

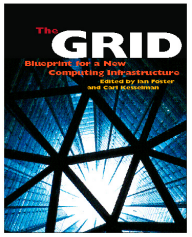
- **Traditional Empirical Science**
  - Scientist gathers data by direct observation
  - Scientist analyzes data

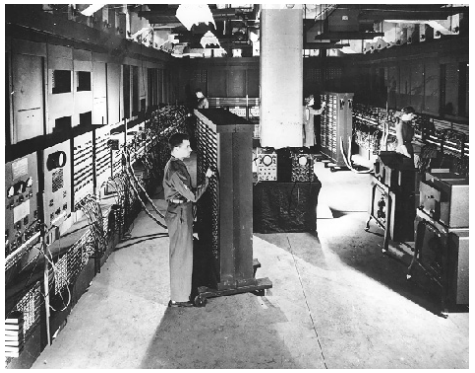
- **Computational Science**
  - Data captured by instruments
    Or data generated by simulator
  - Processed by software
  - Placed in a database
  - Scientist analyzes database
  - `tcl scripts`
    - `or C programs`
      - `on ASCII files`





Why the Grid?
(2) Revolution in Business





- Pre-Internet
  - Central data processing facility
- Post-Internet
  - Enterprise computing is highly distributed, heterogeneous, inter-enterprise (B2B)
  - Business processes increasingly computing- & data-rich
  - Outsourcing becomes feasible => service providers of various sorts

Source: Ian Foster

The Information Grid

Imagine a web of data

- Machine Readable
  - Search, Aggregate, Transform, Report On, Mine Data – using more computers, and less humans
- Scalable
  - Machines are cheap – can buy 50 machines with 100Gb of memory and 100 TB disk for under $100K, and dropping
  - Network is now *faster* than disk
- Flexible
  - Move data around without breaking the apps

Source: S. Banerjee, O. Alonso, M. Drake - ORACLE

Current Environment

- "Big Data" is becoming ubiquitous in many fields
  - enterprise applications
  - Web tasks
  - E-Science
  - Digital entertainment
  - Natural Language Processing (esp. for Humanities applications)
  - Social Network analysis
  - Etc.
- Berkeley Institute for Data Science (BIDS)

Current Environment

- Data Analysis as a profit center
  - No longer just a cost – **may be the entire business** as in Business Intelligence

Current Environment

- Ubiquity of Structured and Unstructured data
  - Text
  - XML
  - Web Data
  - Crawling the Deep Web
- How to extract useful information from "noisy" text and structured corpora?

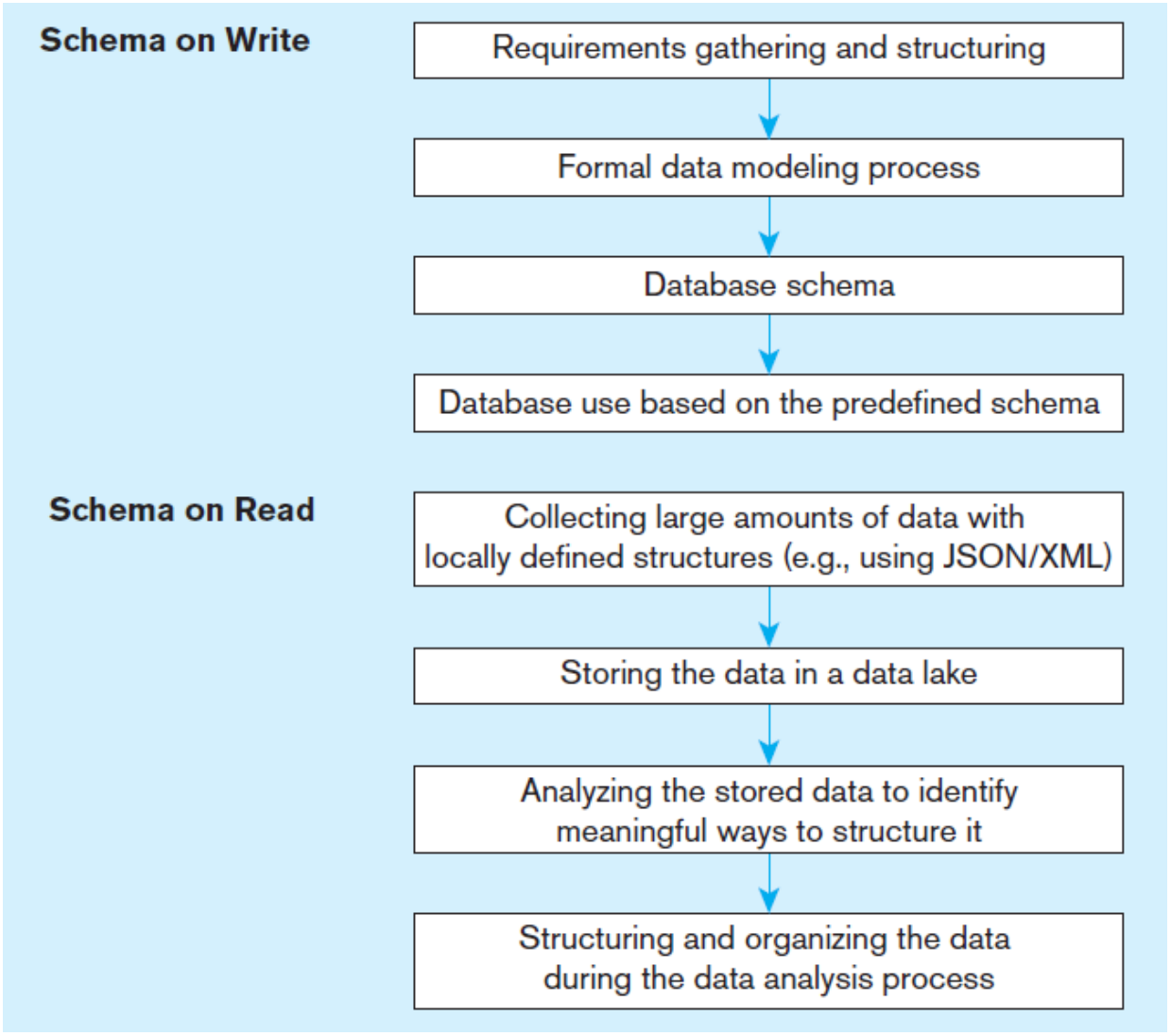
Current Environment

- Expanded developer demands
  - Wider use means broader requirements, and less interest from developers in the details of traditional DBMS interactions
- Architectural Shifts in Computing
  - The move to parallel architectures both internally (on individual chips)
  - And externally – Cloud Computing
  - https://aws.amazon.com/financial-services/grid-computing/

Characteristics of Big Data

- **Schema on Read, rather than Schema on Write**
  - ✖ Schema on Write– pre-existing data model, how traditional databases are designed (relational databases)
  - ✖ Schema on Read – data model determined later, depends on how you want to use it
  - ✖ Capture and store the data, and worry about how you want to use it later
- **Data Lake**
  - A large integrated repository for internal and external data that does NOT follow a predefined schema
  - Capture everything, dive in anywhere, flexible access

Figure 11-2 Schema on write vs. schema on read

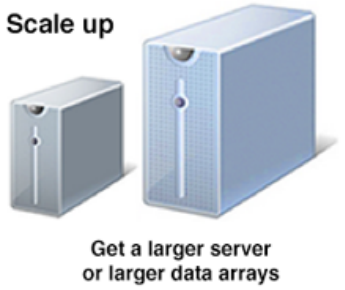


Traditional database design

The big data approach

NoSQL Database

- NoSQL = Not Only SQL (most also support SQL)
- A category of recently introduced data storage and retrieval technologies not based on the relational model
- Scaling out rather than scaling up
- For a cloud environment
- Largely open source



- Supports schema on read
- BASE (basically available, soft state, eventually consistent) instead of ACID properties

NoSQL Classifications

- Key-value stores
  - A simple pair of a key and an associated collection of values. Key is usually a string. Database has no knowledge of the structure or meaning of the values.
- Document stores
  - Like a key-value store, but "document" goes further than "value". Document is structured so specific elements can be manipulated separately.
- Wide-column stores
  - Rows and columns. Distribution of data based on both key values (records) and columns, using "column groups/families"
- Graph-oriented database
  - Maintain information regarding the relationships between data items. Nodes with properties, Connections between nodes (relationships) can also have properties.

Characteristics of Big Data

- **The Five Vs of Big Data**
  - **Volume** – much larger quantity of data than typical for relational databases
  - **Velocity** – data comes at very fast rate (e.g. mobile sensors, web click stream)
  - **Variety** – lots of different data types and formats
  - **Veracity** – traditional data quality methods don't apply; how to judge the data's accuracy and relevance?
  - **Value** – big data is valuable to the bottom line, and for fostering good organizational actions and decisions

High Volume Data

- "Big Data" in the sense of large volume is becoming ubiquitous in many fields
  - enterprise applications
  - Web tasks
  - E-Science
  - Digital entertainment
  - Natural Language Processing (esp. for Humanities applications – e.g. Hathi Trust)
  - Social Network analysis
  - Etc.

High Volume Data Examples

- The Walmart Data Warehouse
  - Often cited as one of, if not the largest data warehouse
- The Google Web database
  - Current web
- The Internet Archive
  - Historic web
- Flickr and YouTube
- Social Networks (E.g.: Facebook)
- NASA EOSDIS
  - Estimated 10^{16} Bytes (Exabyte)
- Other E-Science databases
  - E.g. Large Hadron Collider, Sloan Digital Sky Survey, Large Synoptic Survey Telescope (2016)

Difficulties with High Volume Data

- Browsability
- Very long running analyses
- Steering Long processes
- Federated/Distributed Databases
- IR and item search capabilities
- Updating and normalizing data
- Changing requirements and structure

High Velocity Data

- Examples:
  - Harvesting hot topics from the Twitter "firehose"
  - Collecting "clickstream" data from websites
  - System logs and Web logs
  - High frequency stock trading (HFT)
  - Real-time credit card fraud detection
  - Text-in voting for TV competitions
  - Sensor data
  - Adwords auctions for ad pricing
    - http://www.youtube.com/watch?v=a8qQXLby4PY

High Velocity Requirements

- Ingest at very high speeds and rates
  - E.g. Millions of read/write operations per second
- Scale easily to meet growth and demand peaks
- Support integrated fault tolerance
- Support a wide range of real-time (or "near-time") analytics
- Integrate easily with high volume analytic datastores (Data Warehouses)

Put Differently

You need to ingest a firehose in real time

You need to process, validate, enrich and respond in real-time (i.e. update)

You often need real-time analytics (i.e. query)

High velocity and you

How Big is Big Data

- How big is big?

1 Kilobyte	1,000 bits/byte
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000,000
1 petabyte	1,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

What is Big Data?

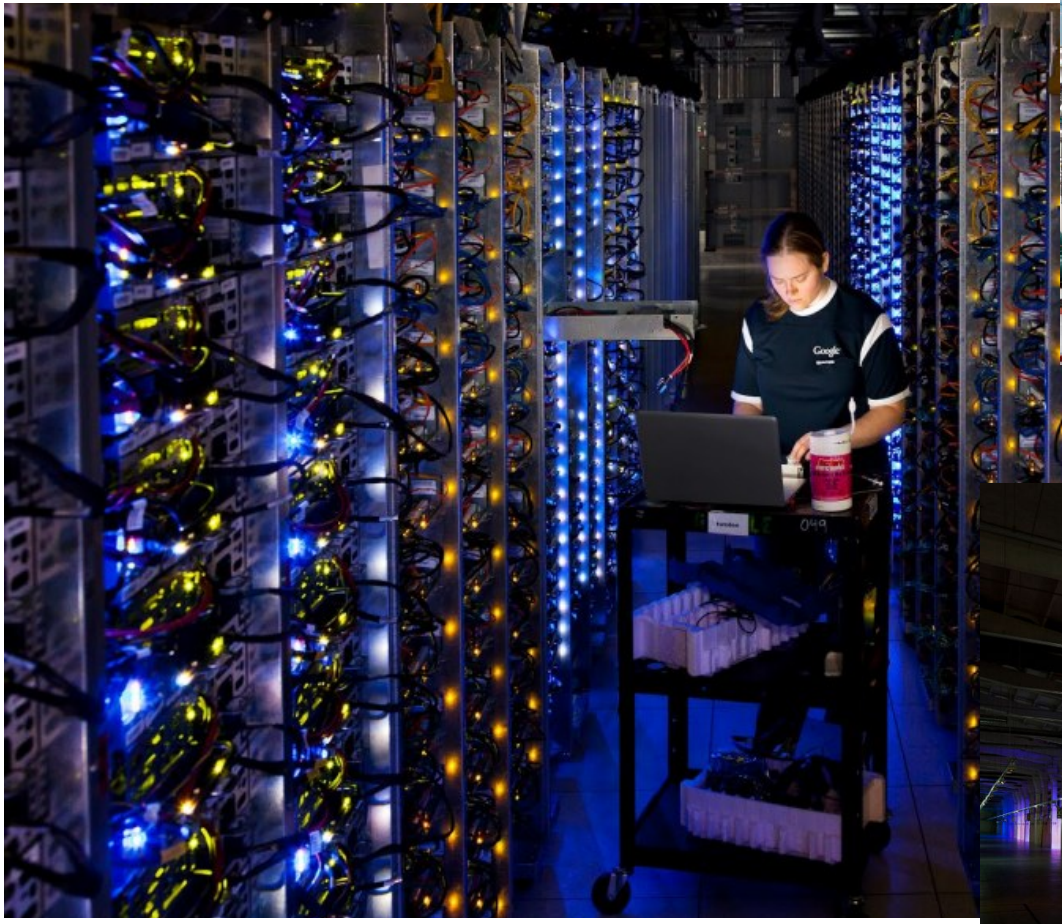
- Ran across some interesting slides from a decade ago that already frame the problem and did a fair job of predicting where we are today
  - Slides by Jim Gray and Tony Hey : “In Search of Petabyte Databases” ca. 2001

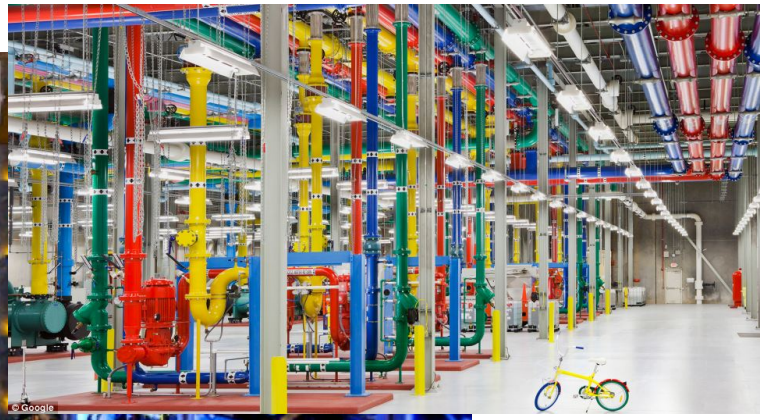
Summary from Gray & Hey

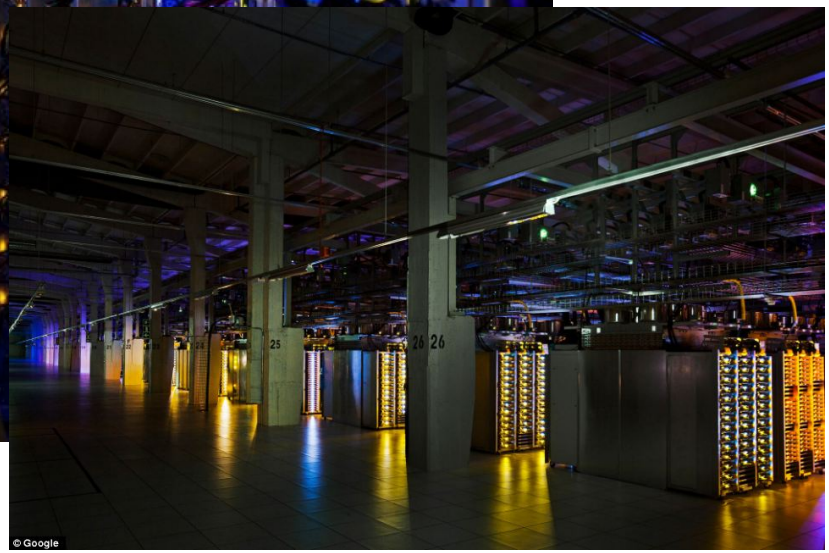
- DBs own the sweet-spot:
  - 1GB to 100TB
- Big data is ***not*** in databases
- HPTS (high performance transaction systems) crowd is not really high performance storage (BIG DATA)
- Cost of storage is people:
  - Performance goal:
    1 Admin per PB

From Jim Gray and Tony Hey : "In Search of Petabyte Databases" ca. 2001

Why People?







One row of one of Google's data centers

Also – the plumbing need for cooling, and the many rows of the data center

High Variety

- Big data can come from a variety of sources, for example:
  - Equipment sensors: Medical, manufacturing, transportation, and other machine sensor transmissions
  - Machine generated: Call detail records, web logs, smart meter readings, Global Positioning System (GPS) transmissions, and trading systems records
  - Social media: Data streams from social media sites like Facebook and miniblog sites like Twitter

High Variety

- The problem of high variety comes when these different sources must be combined and integrated to provide the information of interest
- Problems of:
  - Different structures
  - Different identifiers
  - Different scales for variables
- Often need to combine unstructured or semi-structured text (XML/JSON) with structured data

Various data sources

What Does Machine Data Look Like?

Sources

Order Processing

ORDER,2012-05-21T14:04:12.484,10098213,569281734,67.17.10.12,43CD1A7B8322,SA-2100

Middleware Error

May 21 14:04:12.996 wl-01.acme.com Order 569281734 failed for customer 10098213. Exception follows: weblogic.jdbc.extensions.ConnectionDeadSQLException: weblogic.common.resourcepool.ResourceDeadException: Could not create pool connection. The DBMS driver exception was: [BEA][Oracle JDBC Driver]Error establishing socket to host and port: ACMEDB-01:1521. Reason: Connection refused

Care IVR

05/21 16:33:11.238 [CONNEVENT] Ext 1207130 (0192033): Event 20111, CTI Num:ServID:Type 0:19:9, App 0, ANI T7998#1, DNIS 5555685981, SerID 40489a07-7f6e-4251-801a-13ae51a6d092, Trunk T451.16
05/21 16:33:11.242 [SCREENPOPEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092 CUSTID 10098213
05/21 16:37:49.732 [DISCEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092

Twitter

{actor:{displayName:"Go Boys!!",followersCount:1366,friendsCount:789,link: "http://dallascowboys.com/",location:{displayName:"Dallas, TX",objectType:"place"}, objectType:"person",preferredUsername:"B0ysF@n80",statusesCount:6072},body:"Just bought this POS device from @ACME. Doesn't work! Called, gave up on waiting for them to answer! RT if you hate @ACME!!",objectType:"activity",postedTime:"2012-05-21T16:39:40.647-0600"}

From Stephen Sorkin of Splunk

Integration of Variety

Machine Data Contains Critical Insights

From Stephen Sorkin of Splunk

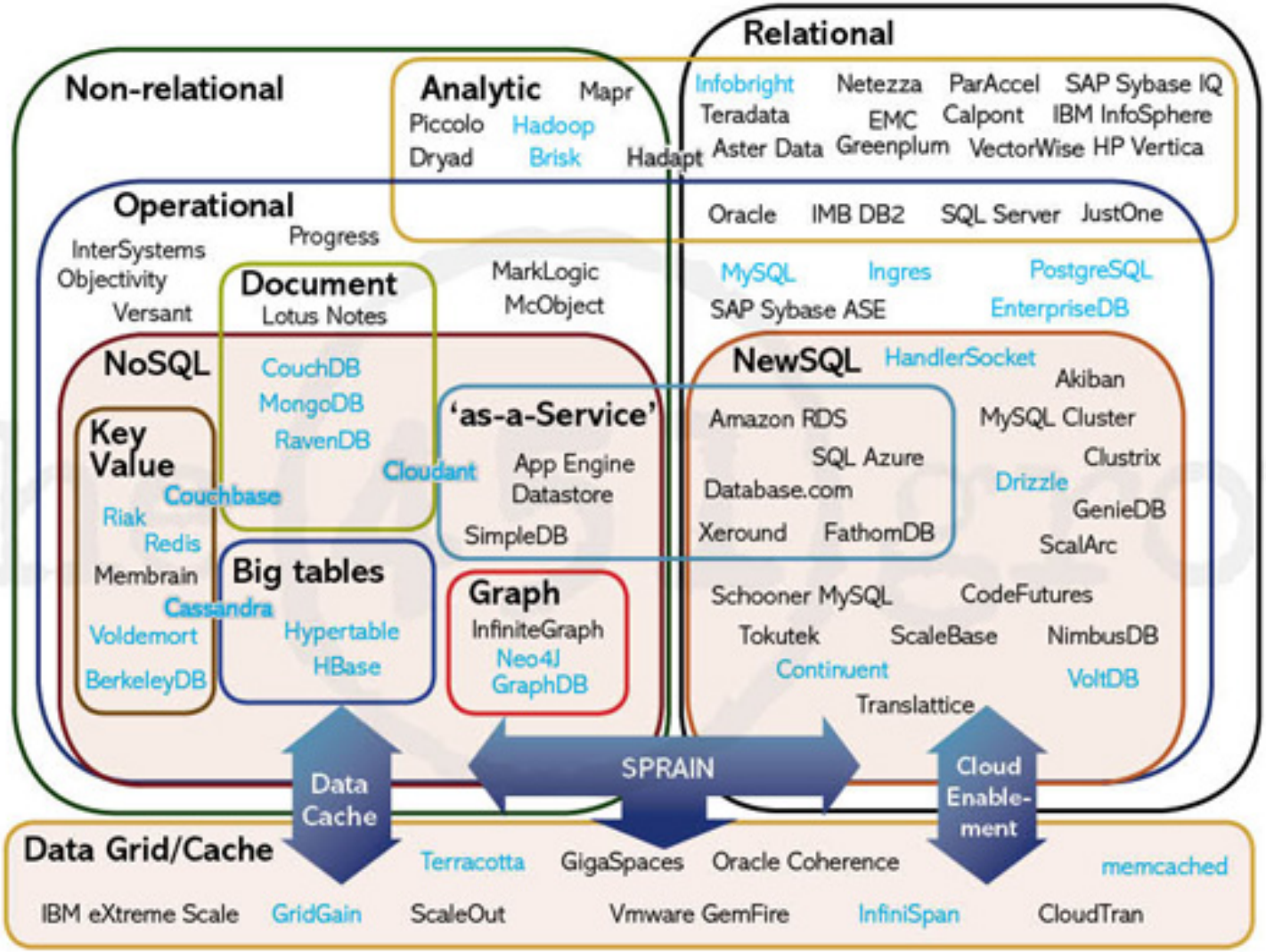
Current Environment

- Data Analysis as a profit center
  - No longer just a cost – may be the entire business as in Business Intelligence

Current Environment

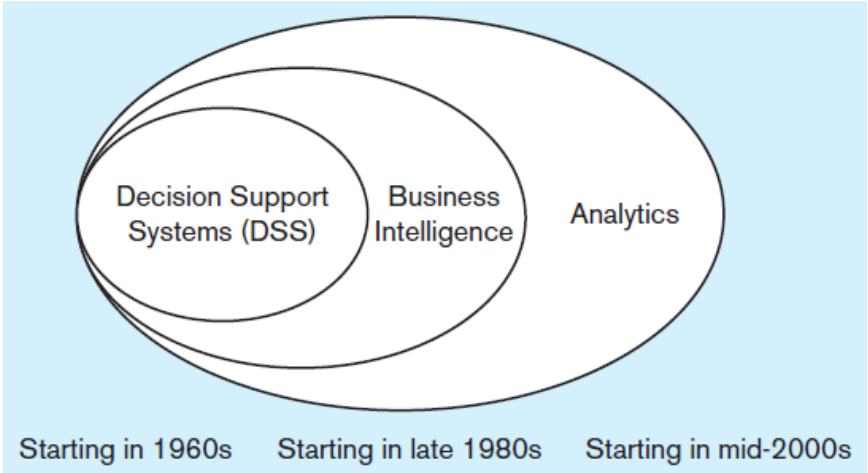
- Expanded developer demands
  - Wider use means broader requirements, and less interest from developers in the details of traditional DBMS interactions
- Architectural Shifts in Computing
  - The move to parallel architectures both internally (on individual chips)
  - And externally – Cloud Computing/Grid Computing

The Database Universe 201x



ANALYTICS

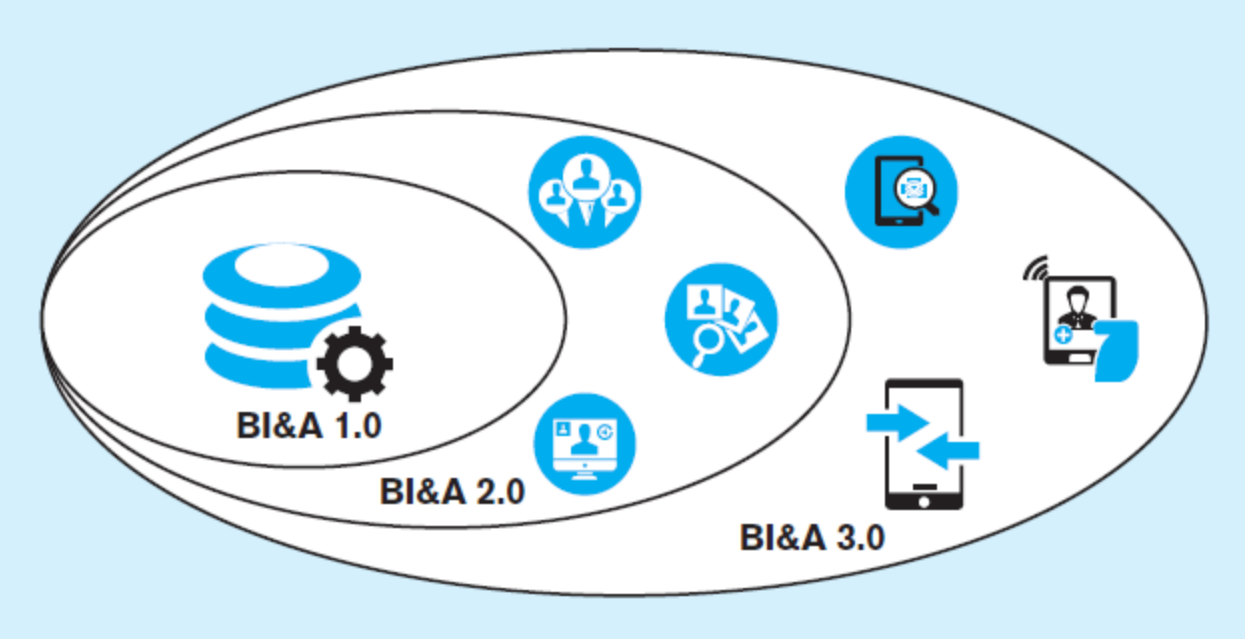
- Historical precedents to analytics:
  - Management information systems (MIS) ➔ Decision Support Systems (DSS) ➔ Executive Information Systems (EIS)
  - DSS idea evolved into Business Intelligence (BI)
- Business Intelligence – a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information.
- Analytics encompasses more than BI
  - Broader term that includes BI
  - Transform data to useful form
  - Infrastructure for analysis
  - Data cleanup processes
  - User interfaces



TYPES OF ANALYTICS

- **Descriptive analytics** – describes the past status of the domain of interest using a variety of tools through techniques such as reporting, data visualization, dashboards, and scorecards
- **Predictive analytics** – applies statistical and computational methods and models to data regarding past and current events to predict what might happen in the future
- **Prescriptive analytics** – uses results of predictive analytics along with optimization and simulation tools to recommend actions that will lead to a desired outcome

Generations of Business Intelligence and Analytics



Adapted from Chen et al., 2012

BI&A 1.0

Focus on structured quantitative data largely from relational databases

BI&A 2.0

Include data from the Web (web interaction logs, customer reviews, social media)

BI&A 2.0

Include data from mobile devices, (location, sensors, etc.) as well as Internet of Things

USE OF DESCRIPTIVE ANALYTICS

- Descriptive analytics was the original emphasis of BI
- Reporting of aggregate quantitative query results
- Tabular or data visualization displays
- Dashboard – a few key indicators
- Scorecard – like a dashboard, but broader range
- OLAP – online analytical processing

Predictive Analytics

- Statistical and computational methods that use data regarding past and current events to form models regarding what might happen in the future
- Examples: classification trees, linear and logistic regression analysis, machine learning, neural networks, time series analysis, Bayesian modeling

Online Analytical Processing (OLAP) Tools

- **Online Analytical Processing (OLAP)** -- the use of a set of graphical tools that provides users with **multidimensional** views of their data and allows them to analyze the data using simple windowing techniques
- **Relational OLAP (ROLAP)** – OLAP tools that view the database as a traditional relational database in either a star schema or other normalized or denormalized set of tables
- **Multidimensional OLAP (MOLAP)** –OLAP tools that load data into an intermediate structure, usually a three- or higher-dimensional array.

Cube Slicing, Drill down (Roll up), Pivoting, etc.

Data Mining Tools

- Knowledge discovery using a sophisticated blend of techniques from traditional statistics, artificial intelligence, and computer graphics
- Goals:
  - Explanatory – explain observed events or conditions
  - Confirmatory – confirm hypotheses
  - Exploratory –analyze data for new or unexpected relationships
- Text mining – Discovering meaningful information algorithmically based on computational analysis of unstructured textual information

TABLE 11-5 Typical Data-Mining Applications

Data-Mining Application	Example
Profiling populations	Developing profiles of high-value customers, credit risks, and credit-card fraud.
Analysis of business trends	Identifying markets with above-average (or below-average) growth.
Target marketing	Identifying rustomers (or customer segments) for promotional activity.
Usage analysis	Identifying usage patterns for products and services.
Campaign effectiveness	Comparing campaign strategies for effectiveness.
Product affinity	Identifying products that are purchased concurrently or identifying the characteristics of shoppers for certain product groups.
Customer retention and churn	Examining the behavior of customers who have left for competitors to prevent remaining customers from leaving.
Profitability analysis	Determining which customers are profitable, given the total set of activities the customer has with the organization.
Customer value analysis	Determining where valuable customers are at different stages in their life.
Upselling	Identifying new products or services to sell to a customer based upon critical events and life-style changes.

Source: Based on Dyché (2000).

Preview: Massively Parallel Processing

- MPP used to mean that you had to write a lot of code to partition tasks and data, run them on different machines, and combine the results back together
- That has now largely been replaced due to the MapReduce paradigm

Motivation

- Large-Scale Data Processing
  - Want to use 1000s of CPUs
    - But don't want hassle of ***managing*** things
- MapReduce provides
  - Automatic parallelization & distribution
  - Fault tolerance
  - I/O scheduling
  - Monitoring & status updates

From "MapReduce…" by Dan Weld