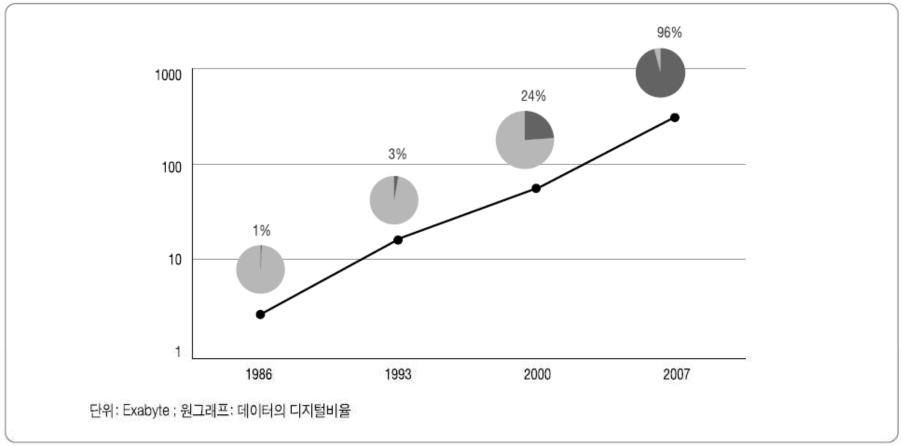
The Evolution of Analytic Scalability

Outline

- **Introduction**
- The Convergence of the Analytic and Data Environment
- Massively Parallel Processing System (MPP)
- Cloud Computing
- Grid Computing
- MapReduce
- Conclusion

Introduction

- The amount of data organizations process continues to increase



*출처: Hilbert & Lopez(2011) 재구성

[그림 1] 전 세계 정보량의 변화(로그 스케일)

The old methods for handling data won’t work anymore

- Important technologies to tame the big data tidal wave possible

MPP

The cloud

Grid computing

MapReduce

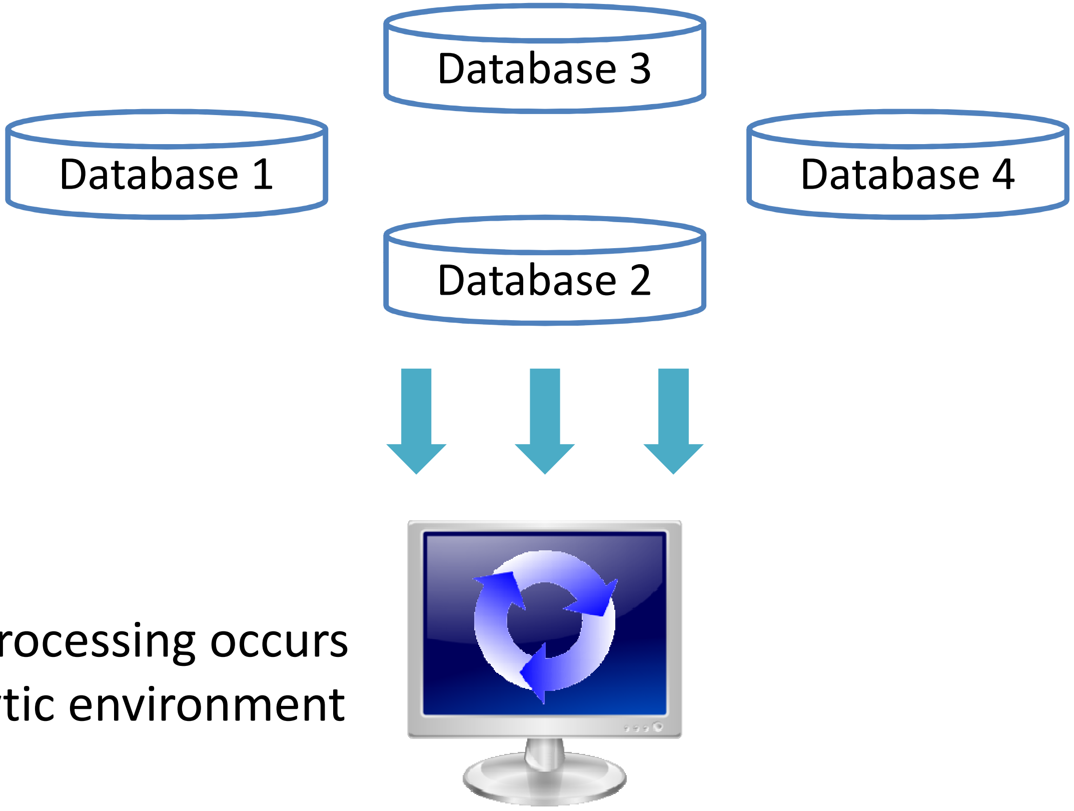
Outline

- Introduction
- **The Convergence of the Analytic and Data Environment**
- Massively Parallel Processing System (MPP)
- Cloud Computing
- Grid Computing
- MapReduce
- Conclusion

The Convergence of the Analytic and Data Environment (1/2)

Traditional Analytic Architecture

- We had to pull all data together into a separate analytics environment to do analysis



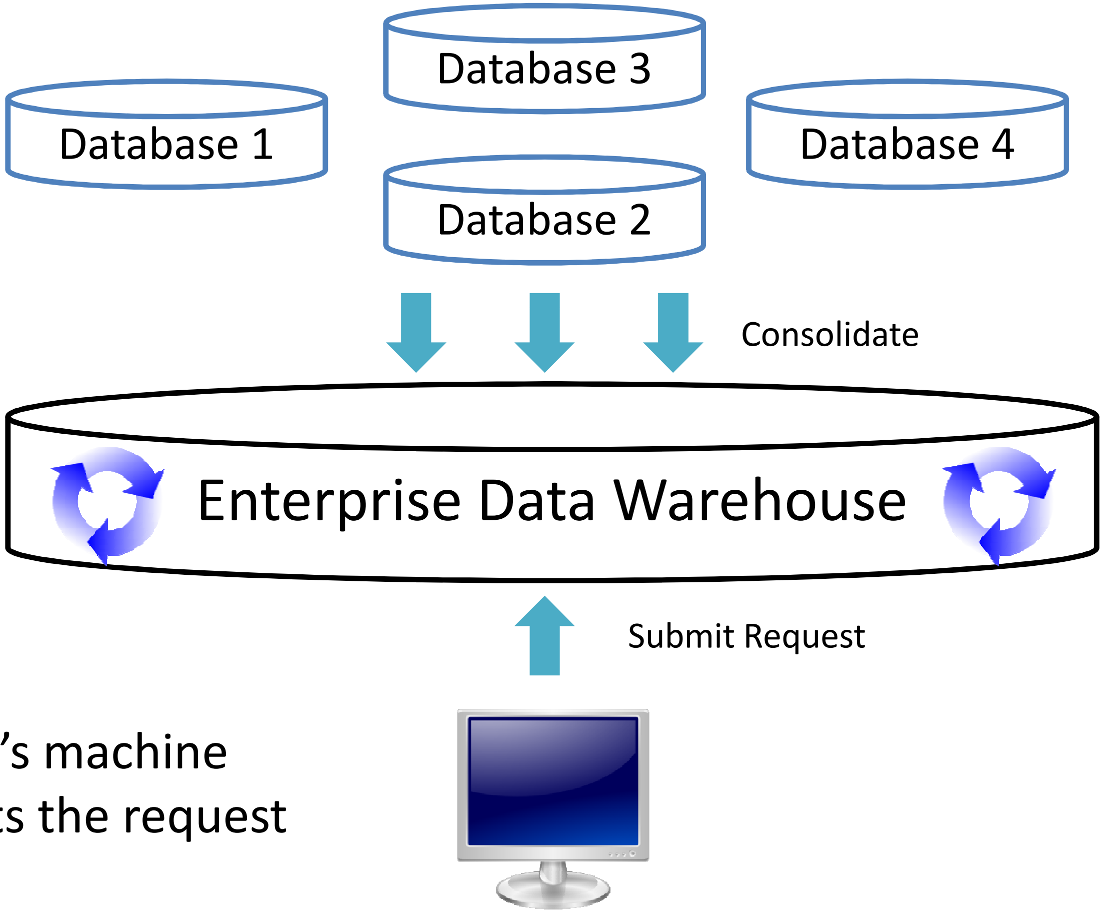
The heavy processing occurs in the analytic environment

Analytic Server
Or PC

The Convergence of the Analytic and Data Environment (2/2)

Modern In-Database Architecture

- The processing stays in the database where the data has been consolidated



The user's machine just submits the request

Analytic Server Or PC

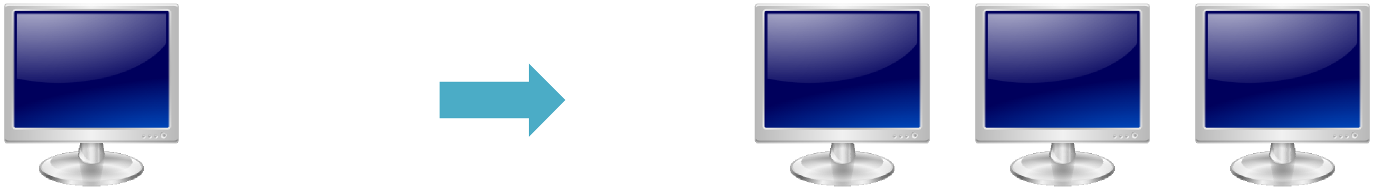
Outline

- Introduction
- The Convergence of the Analytic and Data Environment
- **Massively Parallel Processing System (MPP)**
- Cloud Computing
- Grid Computing
- MapReduce
- Conclusion

Massively Parallel Processing (1/3)

What is an MPP Database?

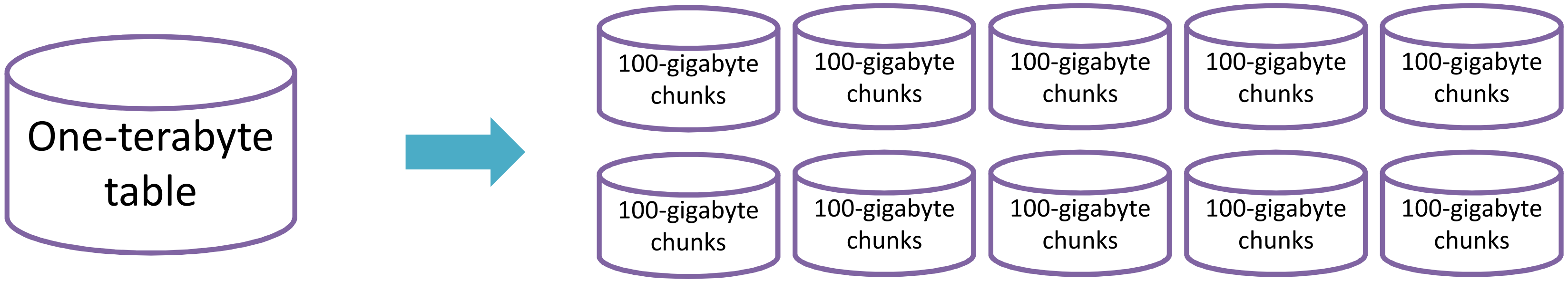
- An MPP database breaks the data into independent chunks with independent disk and CPU



Single overloaded server

Multiple lightly loaded servers

Shared Nothing!



One-terabyte table

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

100-gigabyte chunks

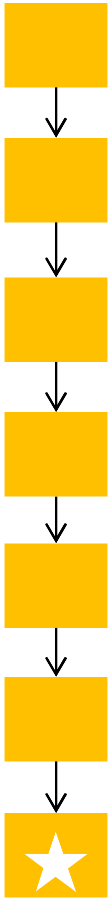
A Traditional database will query
a one-terabyte table one row at time

10 simultaneous 100-gigabyte queries

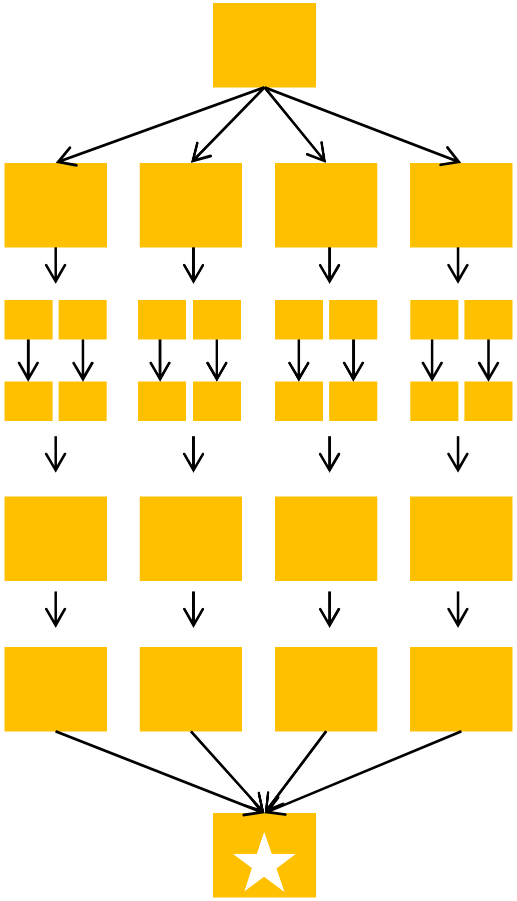
Massively Parallel Processing (2/3)

Concurrent Processing

- An MPP system allows the different sets of CPU and disk to run the process concurrently



Single Threaded Process



An MPP system breaks the job into pieces

Parallel Process

Massively Parallel Processing (3/3)

Others

- MPP systems build in redundancy to make recovery easy

- MPP systems have resource management tools
  - Manage the CPU and disk space
  - Query optimizer

Outline

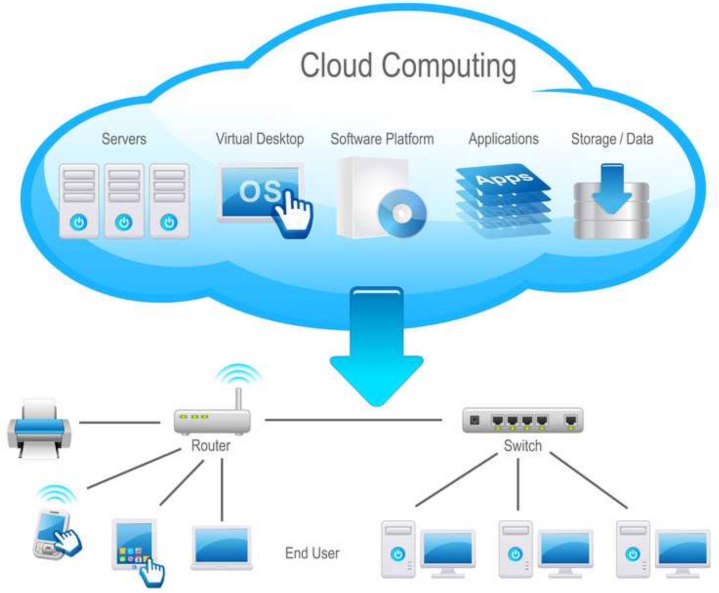
- Introduction
- The Convergence of the Analytic and Data Environment
- Massively Parallel Processing System (MPP)
- **Cloud Computing**
- Grid Computing
- MapReduce
- Conclusion

Cloud Computing (1/2)

What is Cloud Computing?

- McKinsey and Company paper from 2009[1]
  - Mask the underlying infrastructure from the user
  - Be elastic to scale on demand
  - On a pay-per-use basis

- National Institute of Standards and Technology (NIST)
  - On-demand self-service
  - Broad network access
  - Resource pooling
  - Rapid elasticity
  - Measured service



[1] McKinsey and Company, 'Clearing the Air on Cloud Computing," March 2009.

Cloud Computing (2/2)

Two Types of Cloud Environment

1. **Public Cloud**
   - The services and infrastructure are provided off-site over the internet
   - Greatest level of efficiency in shared resources
   - Less secured and more vulnerable than private clouds

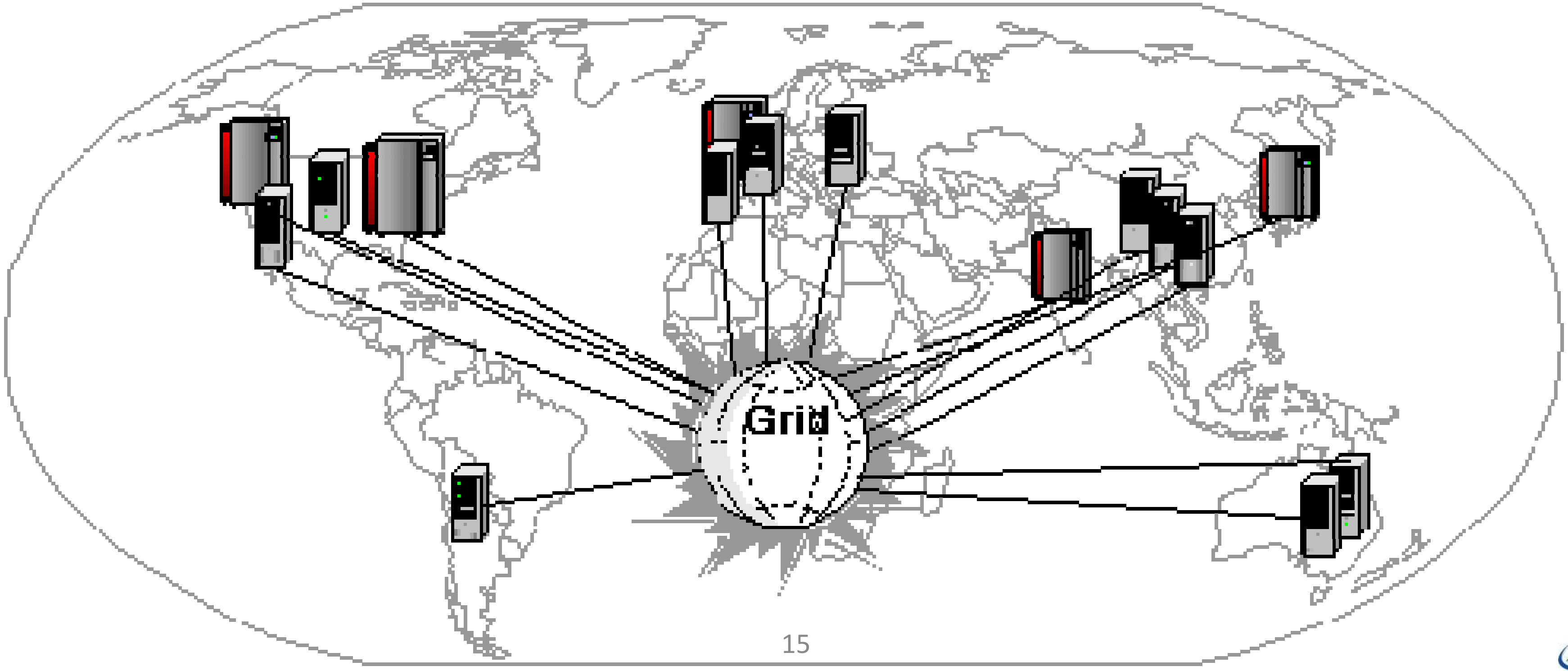
2. **Private Cloud**
   - Infrastructure operated solely for a single organization
   - The same features of a public cloud
   - Offer the greatest level of security and control
   - Necessary to purchase and own the entire cloud infrastructure

Outline

- Introduction
- The Convergence of the Analytic and Data Environment
- Massively Parallel Processing System (MPP)
- Cloud Computing
- **Grid Computing**
- MapReduce
- Conclusion

Grid Computing

- The federation of computer resources to reach a common goal
  - E.g., SETI@Home (Search for Extraterrestrial Intelligence)
    - An Internet-based public volunteer computing project
  - Grid computing harnesses the idle processing power of various computing units, and uses that processing power to compute one job. The job itself is controlled by one main computer, and is broken down into multiple tasks which can be executed simutaneously on different machines and results are again integrated.



Cloud Computing vs. Grid Computing

- The difference between grid computing and cloud computing is hard to grasp because they are not always mutually exclusive. In fact, they are both used to economize computing by maximising existing resources. A dditionally, both architectures use abstraction extensively, and both have distinct elements which interact with each other.
- However, the difference between the two lies in the way the tasks are computed in each respective environment. In a computational grid, one large job is divided into many small portions and executed on multiple machines. This characteristic is fundamental to a grid; not so in a cloud.
- The computing cloud is intended to allow the user to avail of various services without investing in the underlying architecture. While grid computing also offers a similar facility for computing power, cloud comp uting isn't restricted to just that. A cloud can offer many different services, from web hosting, right down to word processing. In fact, a computing cloud can combine services to present a user with a homogenous optimized result.

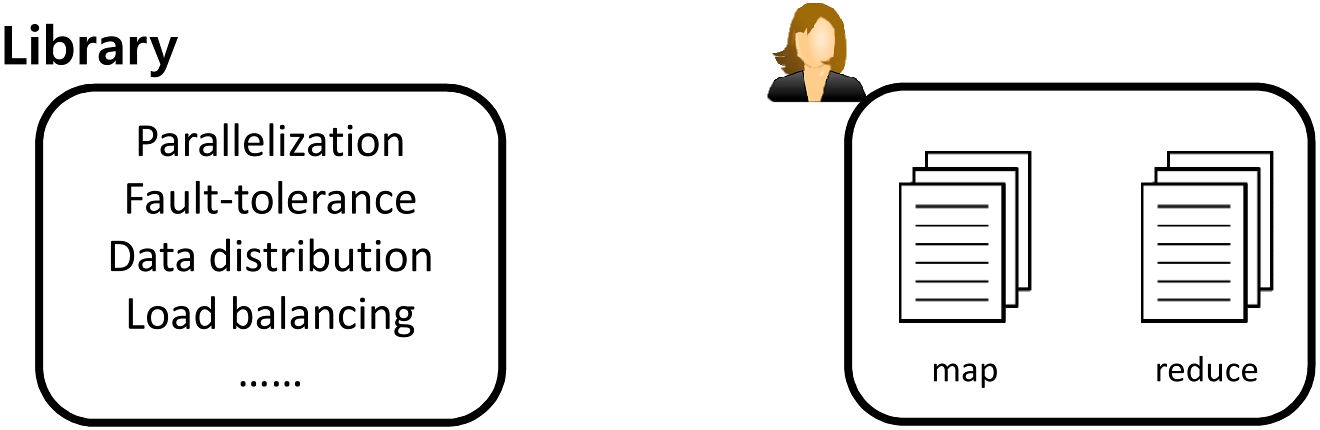
Outline

- Introduction
- The Convergence of the Analytic and Data Environment
- Massively Parallel Processing System (MPP)
- Cloud Computing
- Grid Computing
- **MapReduce**
- Conclusion

MapReduce (1/3)

What is MapReduce?

- A Parallel programming framework[1]



Library

Parallelization
Fault-tolerance
Data distribution
Load balancing
……

  - *Map function*
    - Processing a key/value pairs to generate a set of intermediate key/value pairs
  - *Reduce function*
    - Merging all intermediate values associated with the same intermediate key

[1] MapReduce: Simplified Data Processing on Large Clusters – OSDI 2004

MapReduce (2/3)

How MapReduce Works

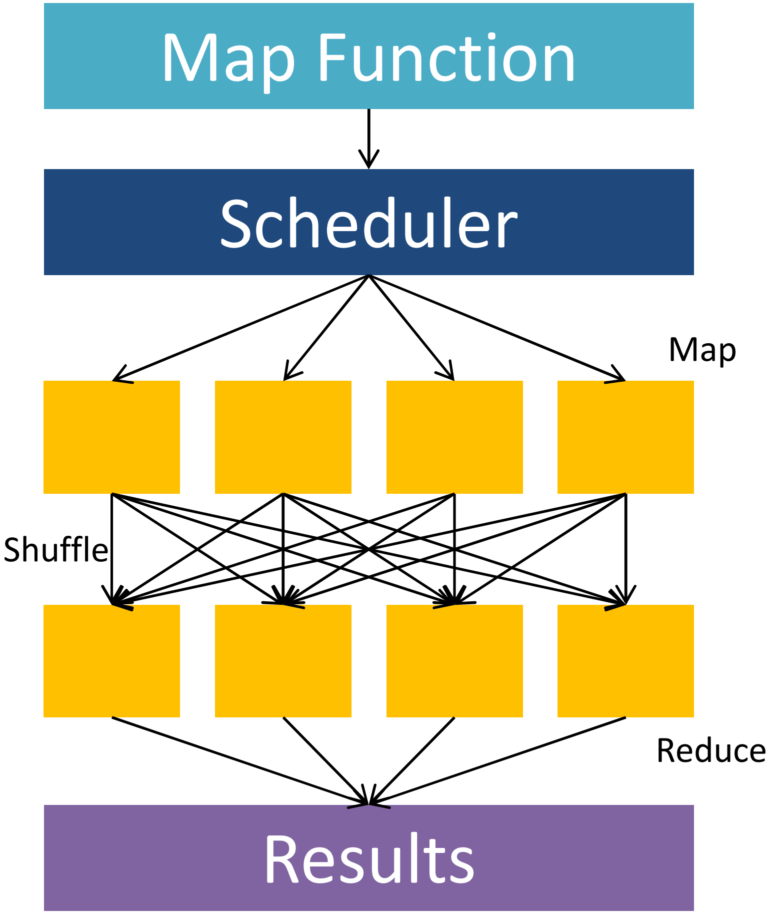
- Let's assume there are 20 terabytes of data and 20 MapReduce server nodes for a project

1. Distribute a terabyte to each of the 20 nodes using a simple file copy process

2. Submit two programs(Map, Reduce) to the scheduler

3. The map program *finds the data* on disk and *executes* the logic it contains

4. The results of the map step are then passed to the reduce process to *summarize* and *aggregate* the final answers



MapReduce (3/3)

Strengths and Weaknesses

- **Good for**
  - Lots of input, intermediate, and output data
  - Batch oriented datasets (ETL: Extract, Load, Transform)
  - Cheap to get up and running because of running on commodity hardware

- **Bad for**
  - Fast response time
  - Large amounts of shared data
  - CPU intensive operations (as opposed to data intensive)
  - NOT a database!
    - No built-in security
    - No indexing, No query or process optimizer
    - No knowledge of other data that exists

Outline

- Introduction
- The Convergence of the Analytic and Data Environment
- Massively Parallel Processing System (MPP)
- Cloud Computing
- Grid Computing
- MapReduce
- **Conclusion**

Conclusion

- These technologies can integrate and work together
  - Databases running in the cloud
  - Databases including MapReduce functionality
  - MapReduce can be run against data sourced from a database
  - MapReduce can also run against data in the cloud

SimpleDB

Microsoft SQL Azure

[Cloud Database]

TERADATA ASTER

[SQL-MapReduce]

ORACLE

[In-Database MapReduce][1]

CouchDB relax

mongoDB

[Running MapReduce in Database]

cloudmapreduce

[Running MapReduce in Cloud][2]

[1] https://blogs.oracle.com/datawarehousing/entry/in-database_map-reduce

[2] http://code.google.com/p/cloudmapreduce/
Cloud mapreduce: a mapreduce implementation on top of a cloud operating system – CCGRID 2011, IEEE Computer Society

The Evolution of Analytic Processes

Outline

- **Introduction**
- The Analytic Sandbox
- Analytic Data Set (ADS)
- Enterprise Analytic Data Set (EADS)
- Scoring Routines

Introduction

- Upgrading technologies won't provide a lot of value, if the same old analytical processes remain in place
  1. Change the process of configuring and maintaining workspace

     The Analytic SandBox

  2. Consistently leverage a database platform through a sandbox

     Enterprise Analytic Data Set (EADS)

  3. Necessary to keep scores up to date on a daily

     Embedded Scoring

Outline

- Introduction
- **The Analytic Sandbox**
- Analytic Data Set (ADS)
- Enterprise Analytic Data Set (EADS)
- Scoring Routines

The Analytical Sandbox (1/5)

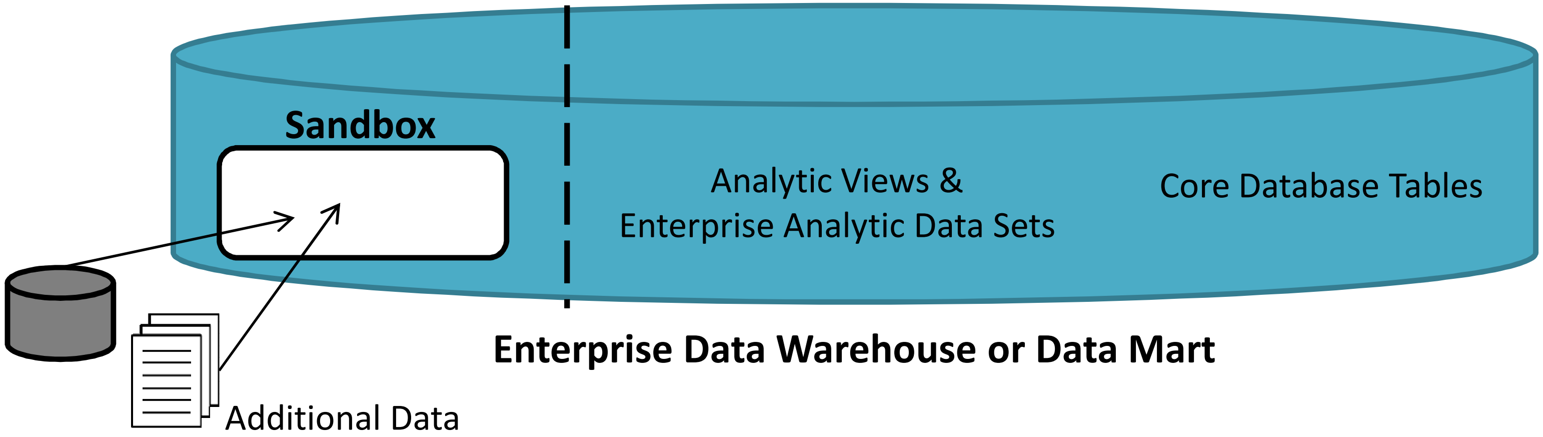
Definition

- A set of resources that enable analytic professionals to experiment and reshape data in whatever fashion they need to
  - Data exploration
  - Development of analytical processes
  - Proof of concepts
  - prototyping

The Analytical Sandbox (2/5)

An Internal Sandbox

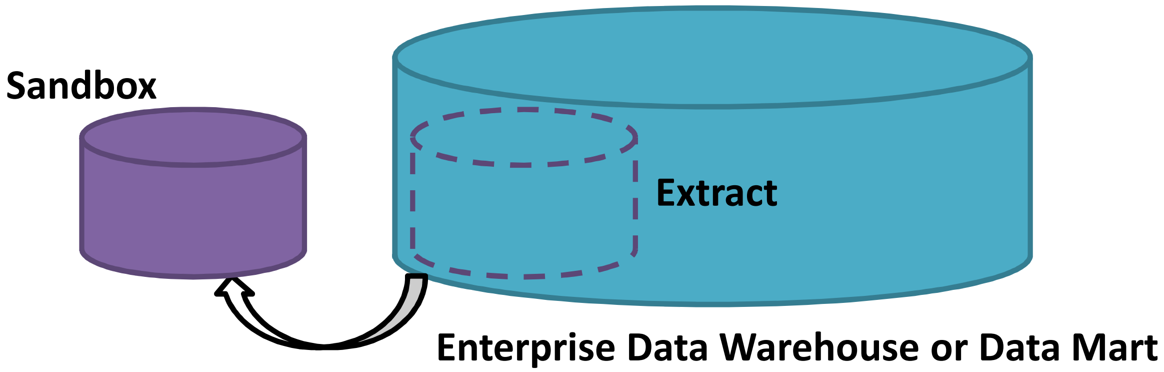
- A portion of an enterprise data warehouse or data mart is carved out to serve as the analytic sandbox
  - Strength
    - Leverage existing hardware resources and infrastructure already in place
    - Ability to directly join production data with sandbox data
    - Cost-effective since no new hardware is needed
  - Weaknesses
    - An additional load on the existing enterprise data warehouse or data mart
    - Can be constrained by production policies and procedures


Sandbox
Analytic Views &
Enterprise Analytic Data Sets
Core Database Tables
Enterprise Data Warehouse or Data Mart
Additional Data

The Analytical Sandbox (3/5)

An External Sandbox

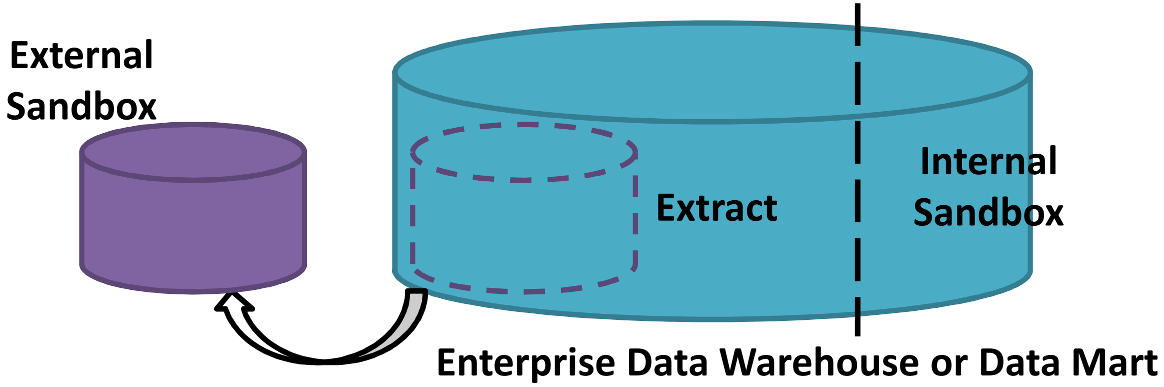
- A physically separate analytic sandbox is created for testing and development of analytic processes
  - Strength
    - A stand-alone environment, no impact on other processes
    - Reduce workload management
  - Weaknesses
    - The additional cost of the stand-alone system
    - Some data movement



The Analytical Sandbox (4/5)

A Hybrid Sandbox

- The combination of an internal sandbox and an external sandbox
  - Strength
    - Flexibility in the approach taken for an analysis
    - Can be run in a 'pseudo-production' mode temporarily
  - Weaknesses
    - Maintain both an internal and external sandbox environment
    - Two-way data feeds may be required, which adds complexity



The Analytical Sandbox (5/5)

Benefits

- From the view of an analytic professional
  - Independence
  - Flexibility
  - Efficiency
  - Freedom
  - Speed

- From the view of IT
  - Centralization
  - Streamlining
  - Simplicity
  - Control
  - Costs

Outline

- Introduction
- The Analytic Sandbox
- **Analytic Data Set (ADS)**
- Enterprise Analytic Data Set (EADS)
- Scoring Routines

Analytic Data Set (1/2)

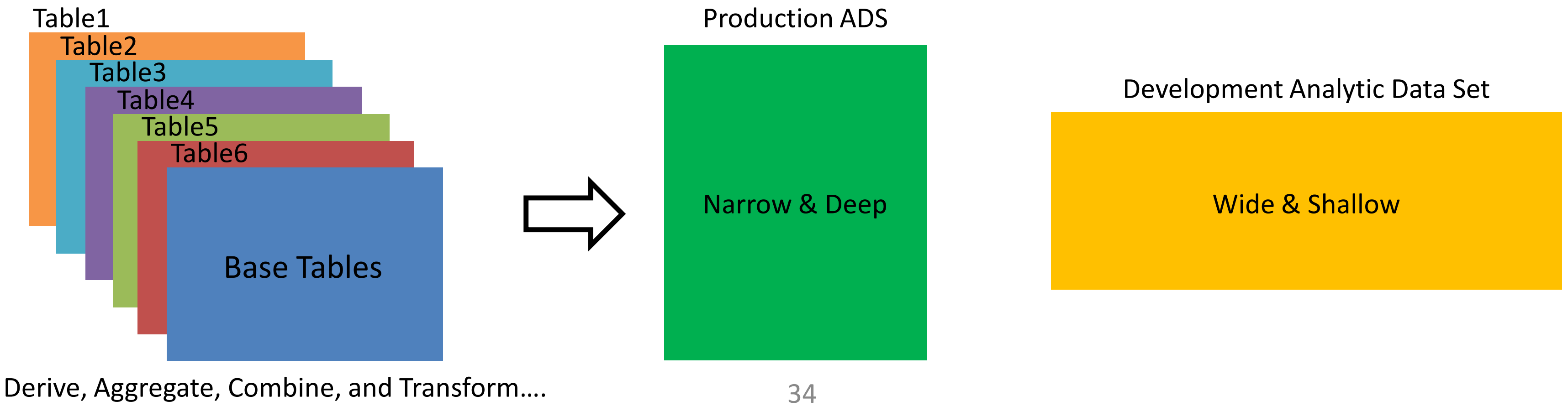
Definition

- The data that is pulled together in order to create an analysis or model
  - In the format required for the specific analysis at hand
  - Generated by transforming, aggregating, and combining data
  - Help to bridge the gap between efficient storage and ease of use

Analytic Data Set (2/2)

Two Primary kinds of Analytic Data Sets

- A development ADS
  - Used to build an analytic process
  - Have many variables or metrics within it
  - Very wide but not very deep
- Production analysis data set
  - Needed for scoring and deployment
  - Contain only the specific metrics that were actually in the final solution
  - Not very wide but very deep



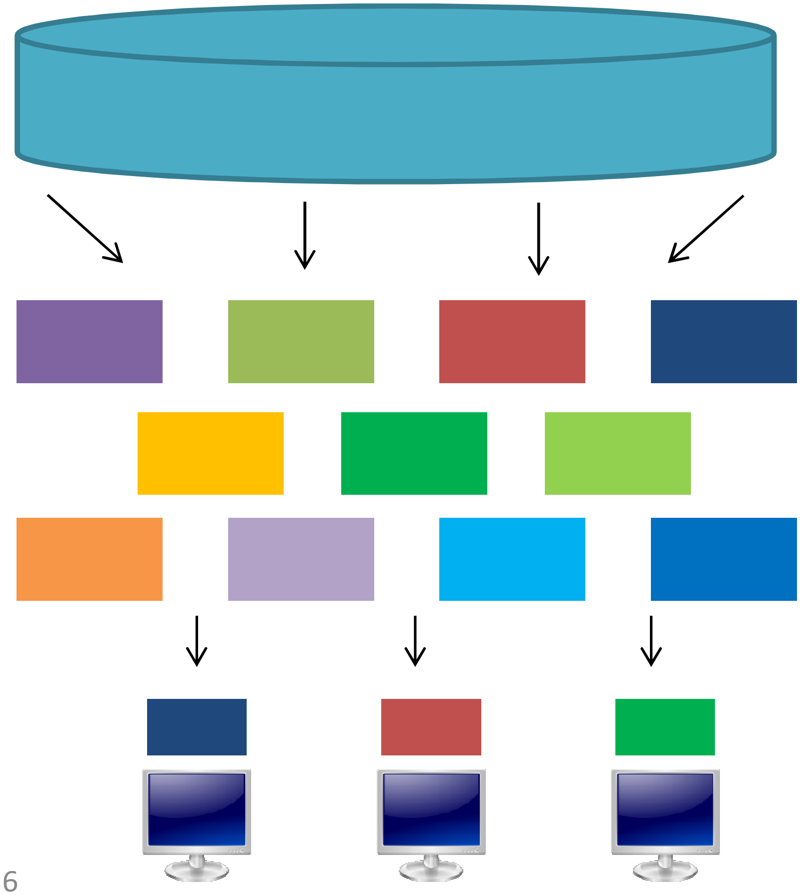
Outline

- Introduction
- The Analytic Sandbox
- Analytic Data Set (ADS)
- **Enterprise Analytic Data Set (EADS)**
- Scoring Routines

Enterprise Analytic Data Set (1/5)

Traditional Analytic Data Sets

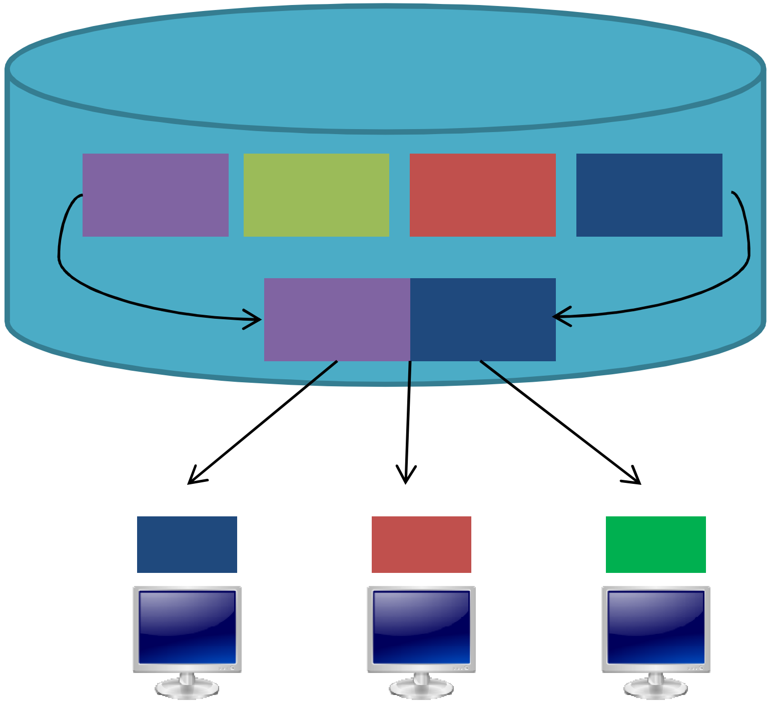
- All analytic data sets are created outside of the database
  - Each analytic professional creates their own data sets independently
  - The risk of inconsistencies
  - The repetitious work



A dedicated ADS is generated
outside the database for every project

Enterprise Analytic Data Set

- A shared and reusable set of centralized, standardized analytic data sets for use in analytics
  - A standardized view of data to support multiple analysis efforts
  - Streamline the data preparation process
  - Provide grate consistency, accuracy, and visibility to analytics processes
  - Build once, use many



Centralized ADS tables and views are utilized across many projects

Enterprise Analytic Data Set (3/5)

Structure

EADS Logical View:

Customer ADS Table

Customer	Total Sales	Total Purchases	Home-owners	Gender	Mail Responder	E-mail Opt in

EADS Potential Physical View:

Customer Sales

Customer	Total Sales	Total Purchases

Customer Demographics

Customer	Home-owner	Gender

Customer Sales

Customer	Mail Responder	E-mail Opt in

It could very well be stored differently!

For updating an EADS

Enterprise Analytic Data Set (4/5)

Summary Table or View?

- **Summary tables** that are updated via a scheduled process
  - Benefits
    - Compute once, use many
    - Most advanced analytics efforts involve a heavy use of historical data
    - Very low latency in getting data
  - Downsides
    - Not be fully up-to-date with the latest data
    - Use disk space on the system, potentially a whole lot of it

Enterprise Analytic Data Set (5/5)

Summary Table or View?

- A series of **views** that are run on demand
  - Benefits
    - be completely fresh and updated
    - Good performance in real-time analysis
    - Changes are immediately available
    - Consistency and transparency of the computations
  - Downsides
    - The system load won’t necessarily be reduced that much
    - Have to wait longer to get their data back

Outline

- Introduction
- The Analytic Sandbox
- Analytic Data Set (ADS)
- Enterprise Analytic Data Set (EADS)
- **Scoring Routines**

Scoring Routines (1/2)

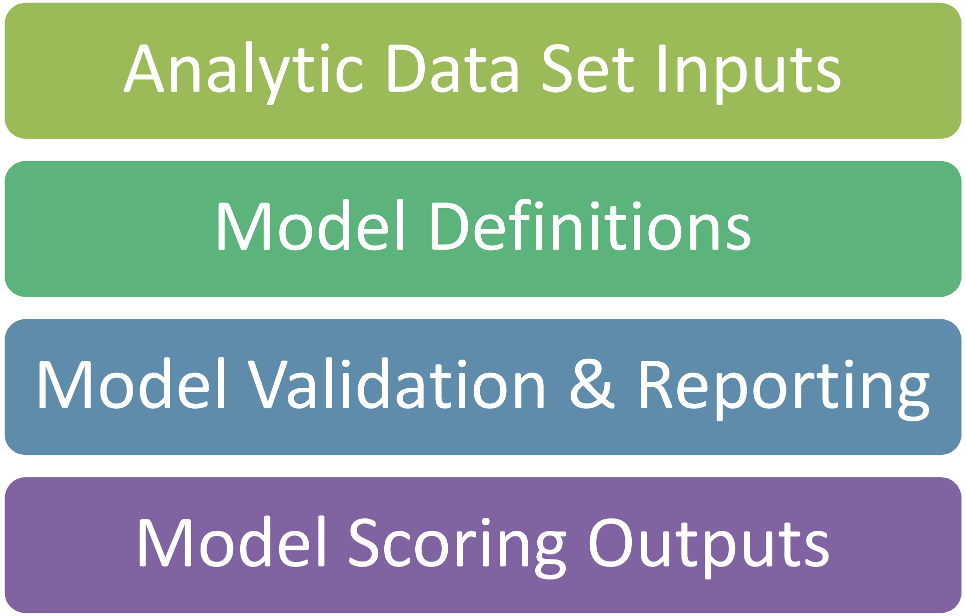
Embedded Scoring

- Score
  - Something generated from a predictive model, or any other type of output from analytic process
- Embedded Scoring
  - Deploying each individual scoring routine
  - A process to manage and track the various scoring routines
- Benefits
  - Scores run in batches will be available on demand
  - Real-time scoring
  - Abstract complexity from users
  - Have all the models contained in a centralized repository so they are all in one place

Embedded Scoring (2/2)

Model and Score Management

- Model and score management procedures will need to be in place to scale the use of models by an organization



Analytic Data Set Inputs

Model Definitions

Model Validation & Reporting

Model Scoring Outputs

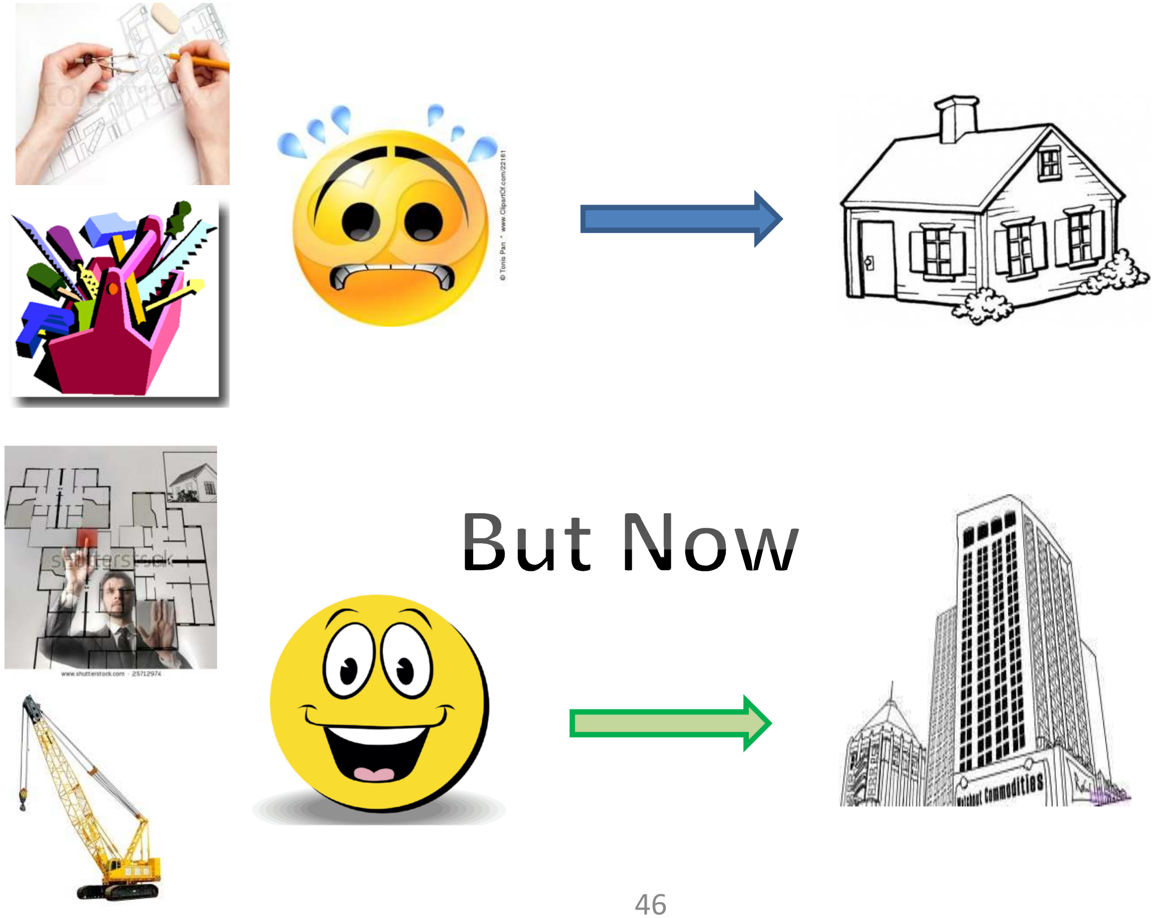
The Evolution of Analytic Tools and Methods

Outline

- **Introduction**
- The Evolution of Analytic Methods
- The Evolution of Analytic Tools

Introduction

- Analytic professionals have used a range of tools over the years
  - Execute analytic algorithms
  - Assess the results



But Now

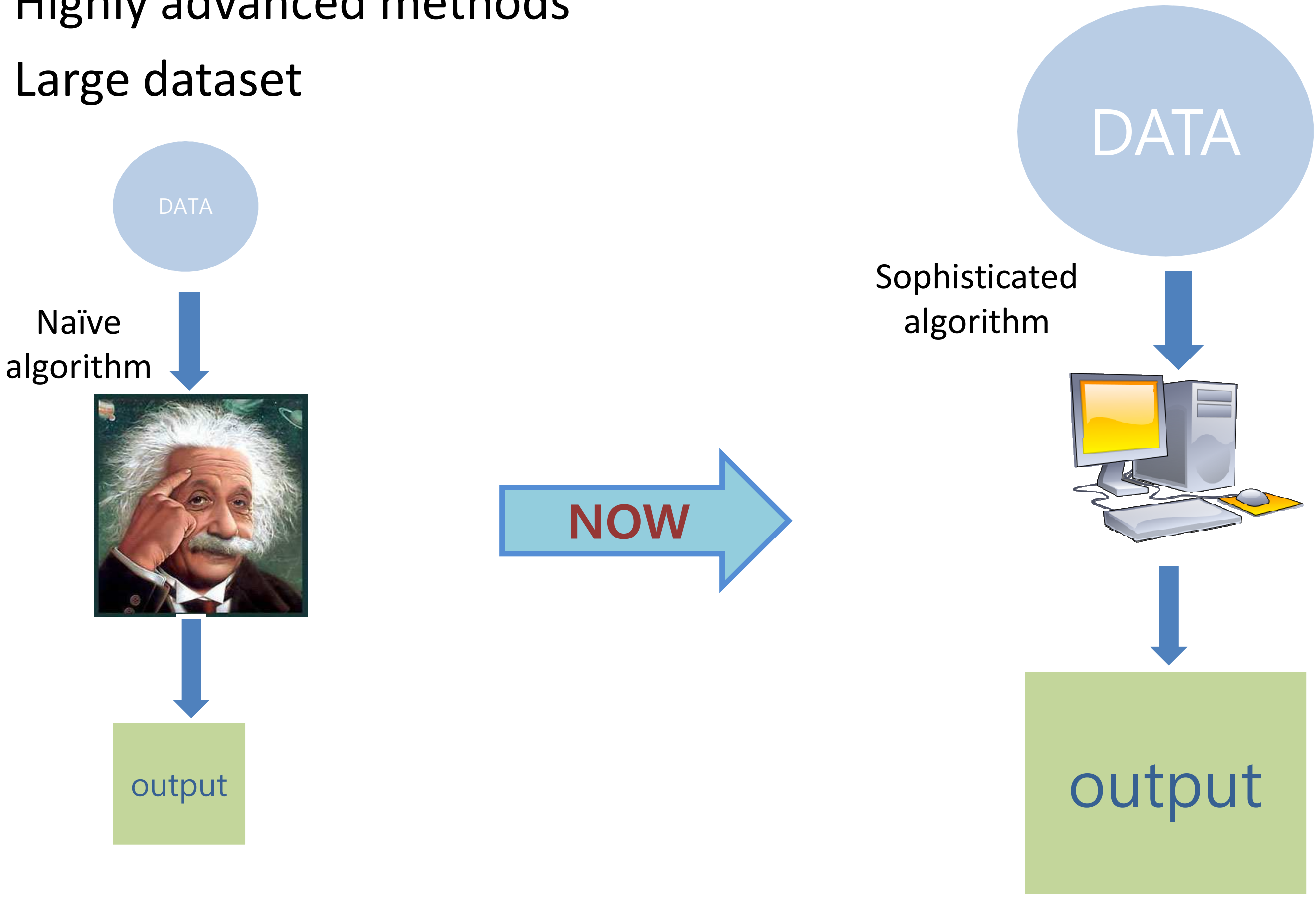
Outline

- Introduction
- **The Evolution of Analytic Methods**
- The Evolution of Analytic Tools

The Evolution of Analytic Methods(1/7)

The Evolution of Analytic Methods

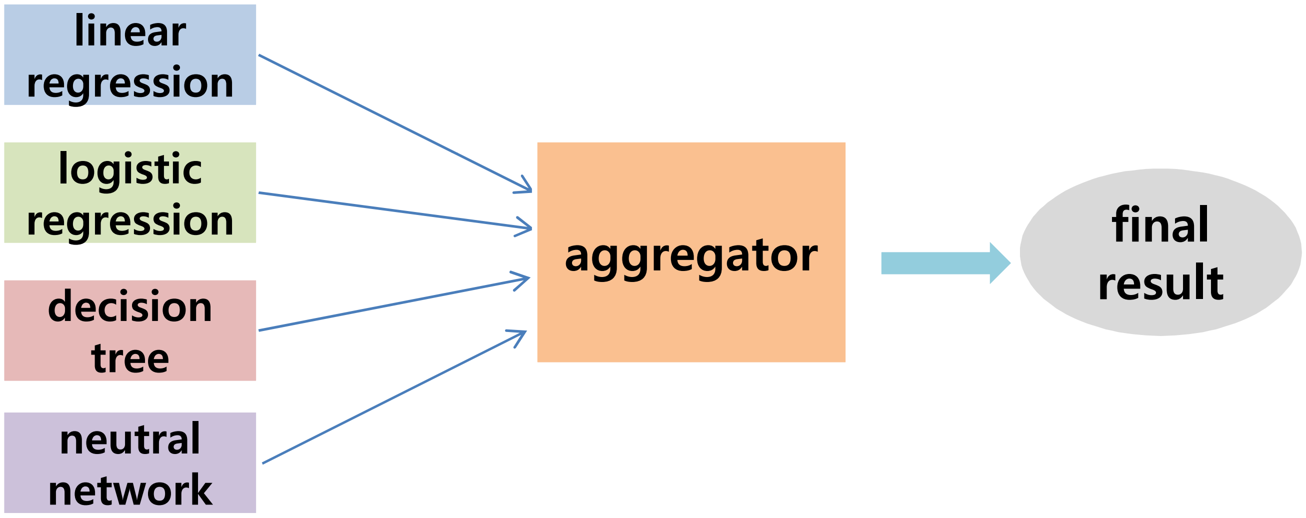
- Until the advent of computers, it wasn’t feasible to run
  - Many iterations of a model
  - Highly advanced methods
  - Large dataset



The Evolution of Analytic Methods(2/7)

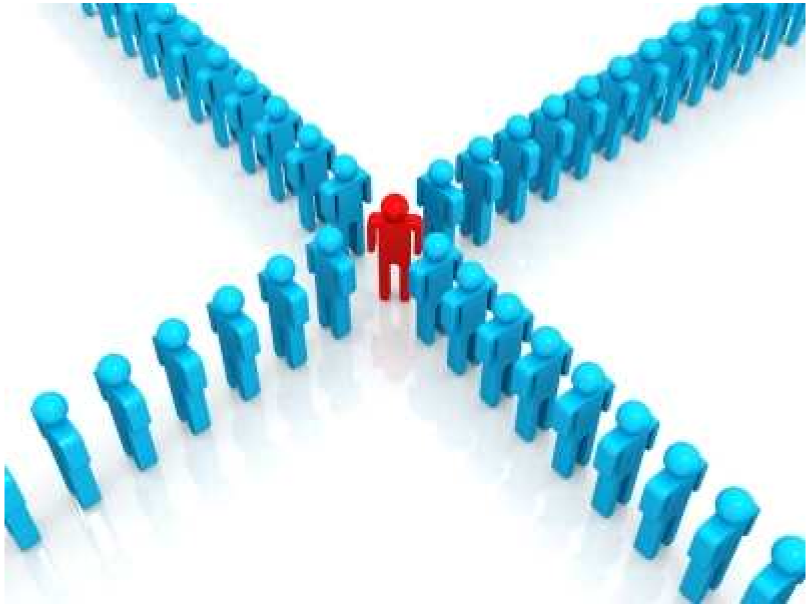
Ensemble Methods

- Ensemble methods are built using multiple techniques
  - go beyond individual performer



The Wisdom of Crowds

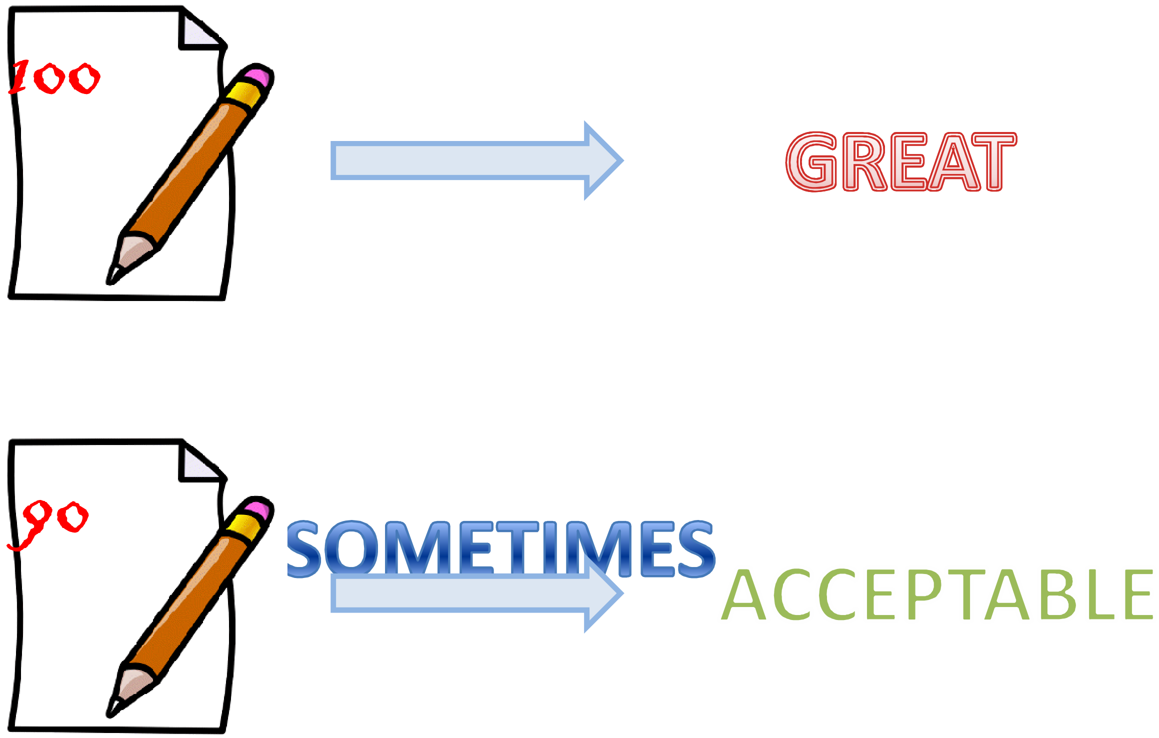
- One reason for ensemble models are gaining traction is
  - *The Wisdom of Crowds*



The Evolution of Analytic Methods(4/7)

Commodity Model

- Commodity model has been produced rapidly
- A commodity modeling process stops when something good enough is found



GREAT

SOMETIMES ACCEPTABLE

The Evolution of Analytic Methods(5/7)

Uses for Commodity Models

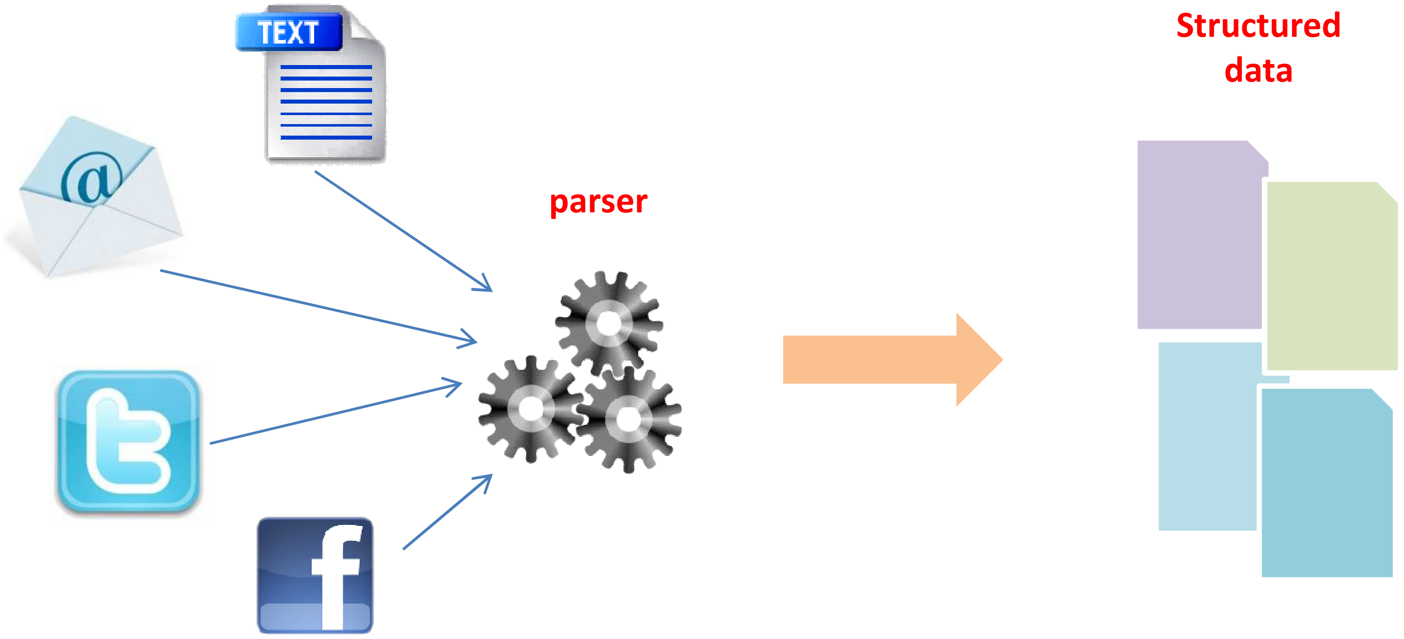
- Traditionally, building models was a time-intensive and expensive
  - Modeling for low-value problems doesn’t make sense
- Commodity model provides an option for low-value problems

VS

The Evolution of Analytic Methods(6/7)

Text Analysis

- Analysis of text and other unstructured data sources is growing rapidly
- Unstructured data is applied to some structure after being processed
  - Structured results are what is analyzed



The Evolution of Analytic Methods(7/7)

Ambiguity

- Applying context to the text is no easy task
  - read a *book* vs *book* a ticket

- Emphasis can change the meaning

Varying the emphasis	Changes the meaning
I didn't say Bill's book stinks	But my buddy Bob did!
I ***didn't*** say Bill's book stinks	How dare you accuse me of such a thing
I didn't ***say*** Bill's book stinks	But I admit that I did write it in an e-mail
I didn't say ***Bill's*** book stinks	It's that other guy's book that stinks
I didn't say Bill's ***book*** stinks	I said his blog stinks
I didn't say Bill's book ***stinks***	I simply said it wasn't my favorite

Outline

- Introduction
- The Evolution of Analytic Methods
- **The Evolution of Analytic Tools**

The Evolution of Analytic Tools(1/7)

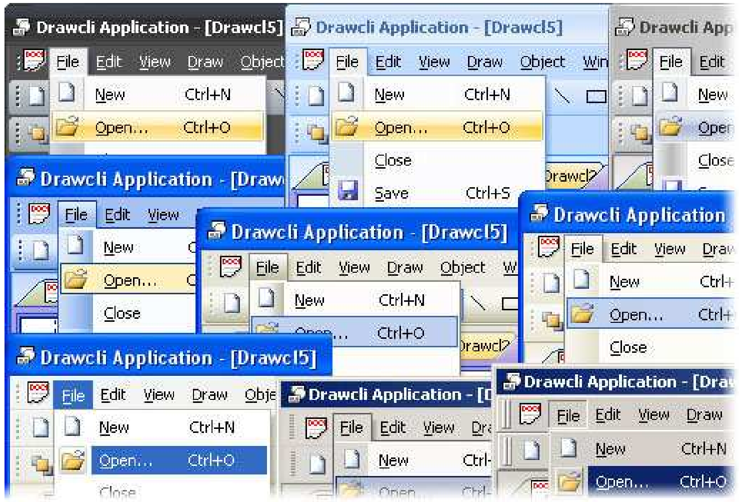
Previous Tools

- Analytics work was done against a mainframe in 1980s
  - Not user-friendly
  - Directly program code to do analytics

The Evolution of Analytic Tools(2/7)

Graphical User Interface

- Graphical user interfaces can accelerate the generation of code while ensuring it is bug-free and optimized
  - Point-and-click environment
  - Generate the code automatically
  - Users still should understand the code to validate the intention


Drawcli Application - [Drawcl5]
File Edit View Draw Object Win
New Ctrl+N
Open... Ctrl+O
Close
Save Ctrl+S

The Evolution of Analytic Tools(3/7)

The Explosion of Point Solutions

- Analytic point solutions are software package that address a set of specific problems
  - Price optimization applications
  - Fraud applications
  - Demand forecasting applications
- One downside of point solutions is the high price
  - Can be $10 million
  - Implementing point solutions in a serial way is preferred

The Evolution of Analytic Tools(4/7)

Open Source

- Open-source software have been around for some time
  - In many cases, open-source products are outside the mainstream
- Many individuals are contributing to improving the functionality
  - Bugs can be patched soon

The Evolution of Analytic Tools(5/7)

The R Project for Statistical Computing

- R Project is open source for statistical computing
- Features of R Project

More object-oriented ☺

Integrate new features faster ☺

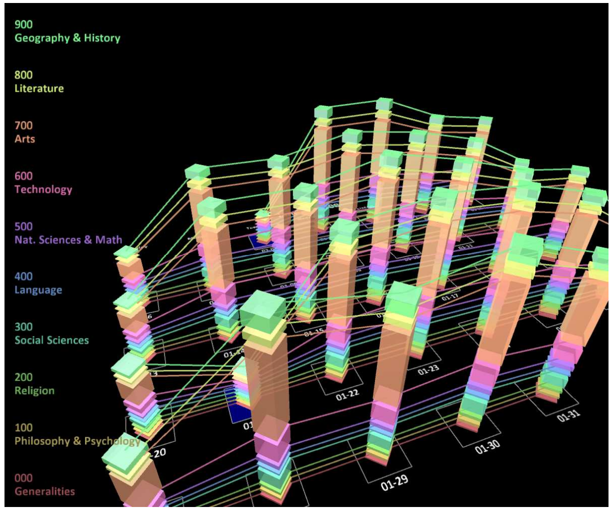
Free for charge ☺

Programming is intensive ☹

The Evolution of Analytic Tools(6/7)

Data Visualization

- An effective visualization can make a pattern jump right off the page at you
- Today’s visualization tools allows
  - Multiple tabs
  - Link the graphs and charts with underlying data
- New idea for data visualization
  - 3-D



The Evolution of Analytic Tools(7/7)

Importance of Data Visualization

- Appropriate visualization will increase an audience's comprehension
- Understanding how to visualize data will help analytic professionals become better