Information Quality in the Cloud, What IT Managers Need to Know

BIOGRAPHY----------

Robert Grossman
Founder and Managing Partner
Open Data Group

Robert Grossman is the Managing Partner of Open Data Group, which he founded in 2002. Open Data provides management consulting, outsourced analytic services, and analytic staffing so that companies and organizations can analyze data, build analytic models, and make predictions about future events, enabling them to increase revenues, decrease costs, reduce risk, and improve business operations.

Grossman is also the Director of the National Center for Data Mining and the Laboratory for Advanced Computing at the University of Illinois at Chicago. The Lab and Center perform research, sponsor standards, manage an international data mining testbed, and engage in outreach activities in the areas of data mining and data intensive computing.

He has been involved in the development of standards, including chairing the working group that develops the Predictive Model Markup Language (PMML), which is a standard for statistical and data mining models. More recently, he has been involved in the development of standards for cloud computing through the Open Cloud Consortium (OCC), which he is the co-chair.
Robert Grossman became a faculty member at the University of Illinois at Chicago in 1988 and is a Professor of Mathematics, Statistics, and Computer Science. From 1984-1988 he was a faculty member at the University of California at Berkeley. He received a Ph.D. from Princeton in 1985 and a B.A. from Harvard in 1980. He has had a half time appointment at UIC since 1996.
Prior to founding the Open Data Group, he founded Magnify, Inc. in 1996. Magnify provides data mining solutions to the insurance industry. Grossman was Magnify's CEO until 2001 and its Chairman until it was sold to ChoicePoint in 2005. ChoicePoint was acquired by LexisNexis in 2008.

He has published over 150 papers in refereed journals and proceedings.

Information Quality in the Cloud, What IT Managers Need to Know

Robert Grossman
Open Data Group
& University of Illinois at Chicago

July 16, 2009

Part 1 – Introduction to Clouds

2

1.1 What is a Cloud?

- Clouds provide on-demand resources or services over a network, often the Internet, with the scale and reliability of a data center.
- No standard definition.
- Cloud architectures are not new.
- What is new:
  - Scale
  - Ease of use
  - Pricing model.

3

Scale is new.

Elastic, Usage Based Pricing Is New

costs the same as

1 computer in a rack
for 120 hours

120 computers in three racks for 1 hour

- Elastic, usage based pricing turns capex into opex.
- Clouds can be used to manage surges in computing needs.

Simplicity Offered By the Cloud is New

+

.. and you have a computer ready to work.

A new programmer can develop a program to process a container full of data with less than day of training using MapReduce.

Two Architectural Models

- On-demand resources & services over a network at the scale of a data center
- On-demand **computing instances (IaaS)**
  - IaaS: Amazon EC2, S3, etc.; Eucalyptus
  - supports many Web 2.0 applications/users
- On-demand **computing capacity (PaaS)**
  - Clouds services to support large data clouds
  - GFS/MapReduce/Bigtable, Hadoop, Sector, …
  - Manage 10 TB, 100 TB, 500 TB, 1PB, 5PB, …

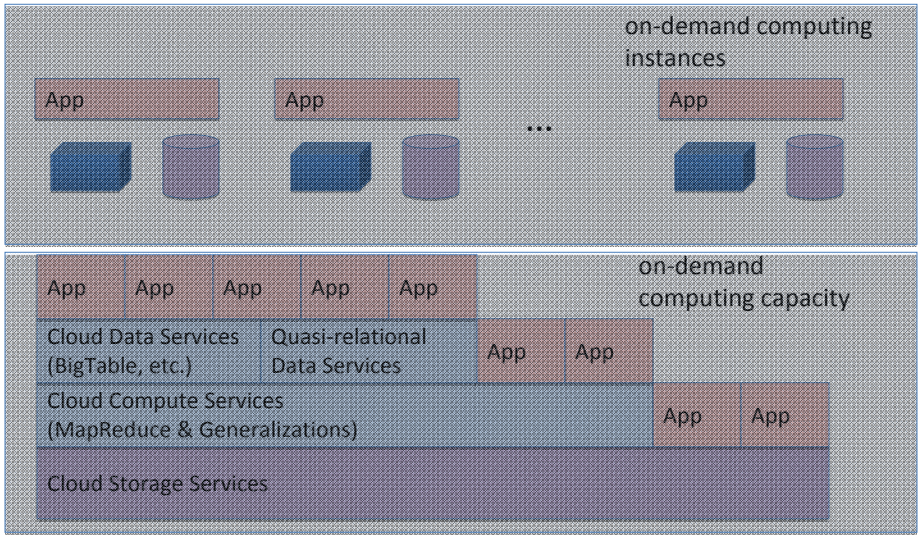
7

Ease of use – With Google's GFS & MapReduce, it is simple to compute with 10 terabytes of data over 100 nodes. With Amazon's AMIs, it is simple to respond to a surge of 100 additional web servers.

Google

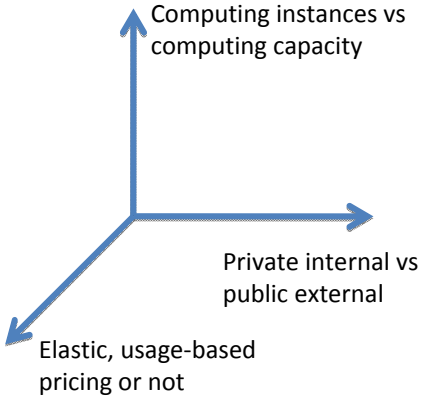
amazon web services

Cloud Architectures – How Do You Fill a Data Center?



Varieties of Clouds

- Architectural Model
  - Computing Instances vs Computing Capacity
- Payment Model
  - Elastic, usage based pricing, lease/own, …
- Management Model
  - Private vs Public; Single vs Multiple Tenant; …
- Programming Model
  - Queue Service, MPI, MapReduce, Distributed UDF



All combinations occur.

10

Part 1.2 Cloud Computing Industry

Cloud computing is approaching the top of the Gartner hype cycle.

- "Cloud computing has become the center of investment and innovation."
  Nicholas Carr, 2009 IDC Directions

11

Cloud Computing Eco-System

- No agreed upon terminology
- Vendors supporting **data centers**
- Vendors providing cloud apps & services to **end users**
- Vendors supporting the **industry** i.e. those developing cloud applications and services for themselves or to sell to end users
- Communities developing software, standards, benchmarks, etc.

12

Cloud Computing Ecosystem

Consumers of Software as a Service

Providers of Software as a Service

Consumers of Cloud Services

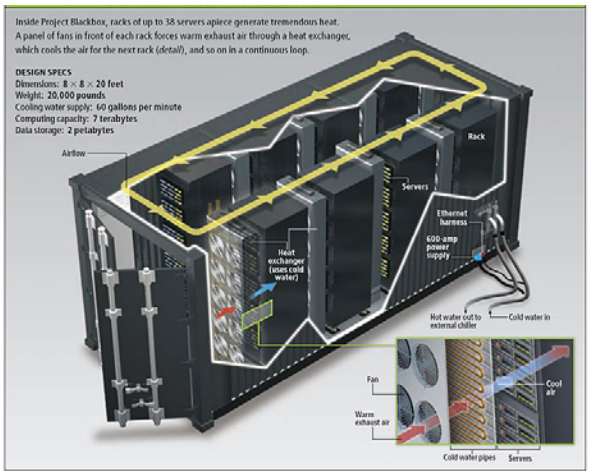
Providers of Cloud Services

Data Centers

- Berkeley RAD Report on cloud computing divides industry into these layers.

13

Building Data Centers

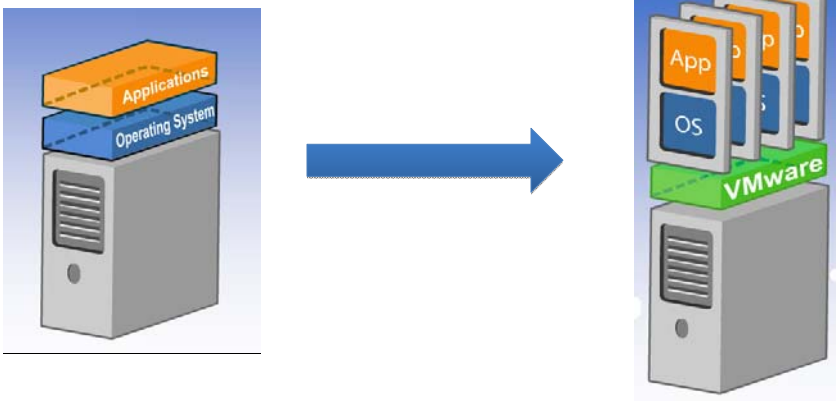


- Sun's Modular Data Center (MD)
- Formerly Project Blackbox
- Containers used by Google, Microsoft & others
- Data center consists of 10-60+ containers.

14

Part 1.3

Virtualization



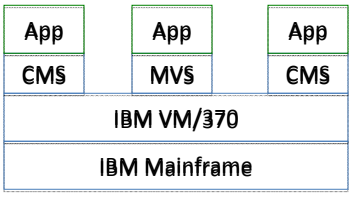
15

Virtualization

- Virtualization separates logical infrastructure from the underlying physical resources to decrease time to make changes, improve flexibility, improve utilization and reduce costs
- Example - server virtualization. Use one physical server to support multiple logical virtual machines (VMs), which are sometimes called logical partitions (LPARs)
- Technology pioneered by IBM in 1960s to better utilize mainfames

16

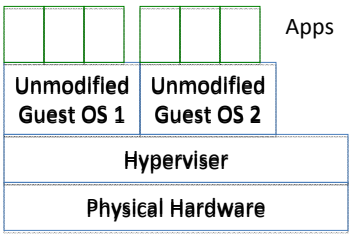
Idea Dates Back to the 1960s



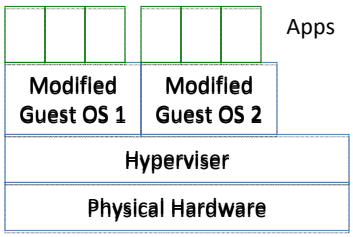
Native (Full) Virtualization
Examples: Vmware ESX

17

Two Types of Virtualization



Native (Full) Virtualization
Examples: Vmware ESX



Para Virtualization
Examples: Xen

- Using the hypervisor, each guest OS sees its own independent copy of the CPU, memory, IO, etc.

18

Part 2

Cloud Architectures for Analytics to Support Data Quality …

… and why you should care.

Part 2.1

What is Analytics Infrastructure?

What is the Size of Your Data?

- Small
  - Fits into memory
- Medium
  - Too large for memory
  - But fits into a database
  - N.B. databases are designed for safe writing of rows
- Large
  - To large for a database
  - But can use specialized file system (column-wise)
  - Or storage cloud (Google File System, Hadoop DFS)

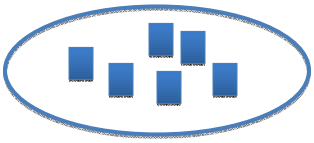
21

What is the Shape of Your Data

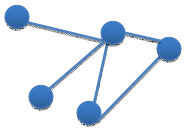
rows

events

semi-structured



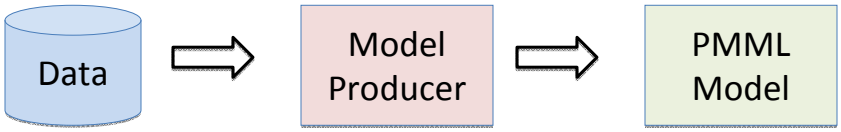
unstructured



graphs

22

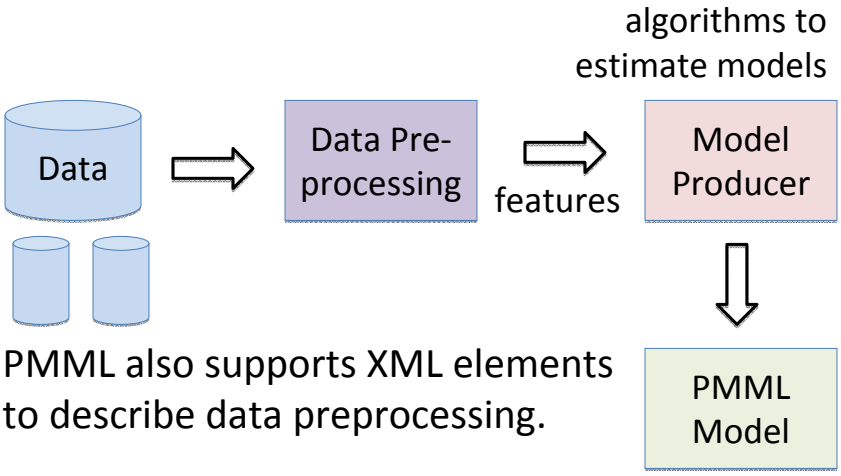
(Very Simplified) Architectural View



- The Predictive Model Markup Language (PMML) is an XML language for statistical and data mining models.
- With PMML, it is easy to move models between applications and platforms.

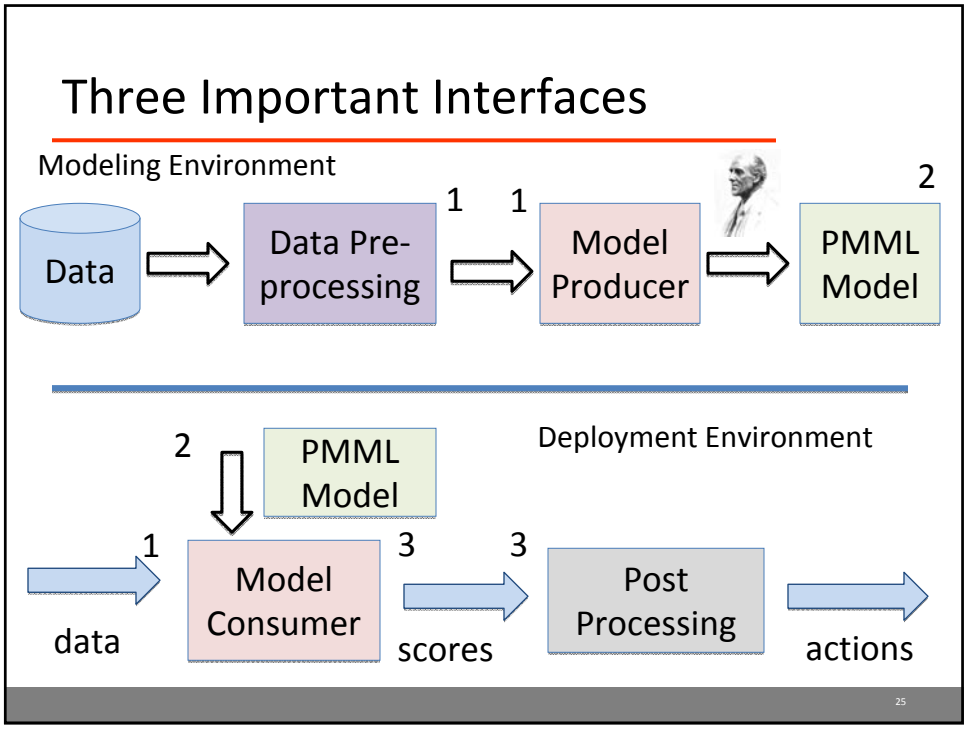
23

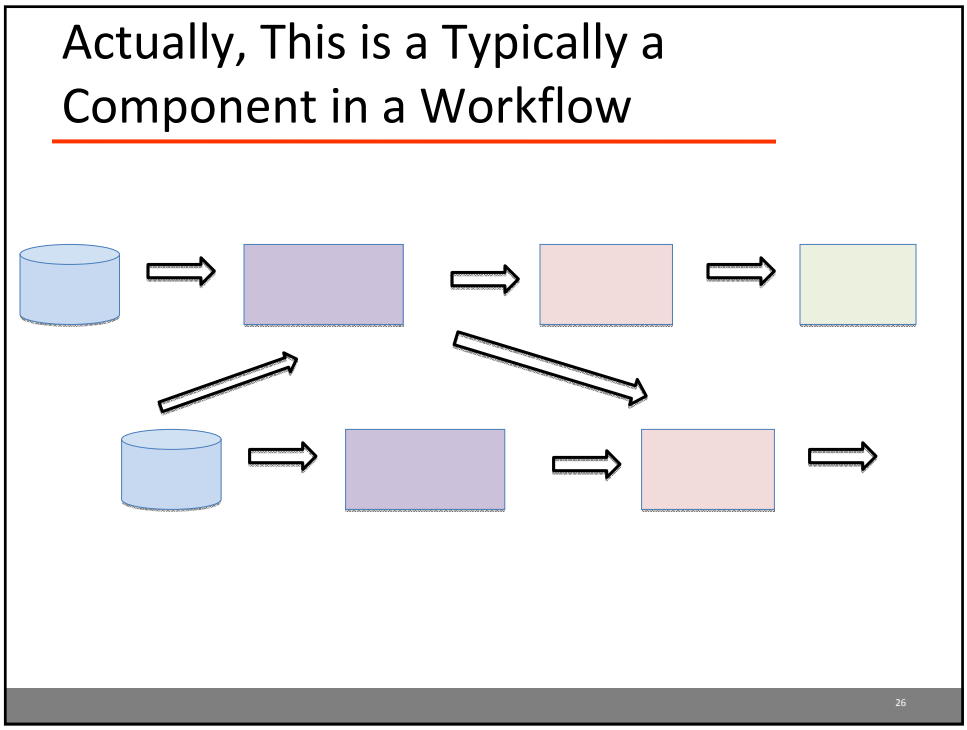
(Simplified) Architectural View



- PMML also supports XML elements to describe data preprocessing.

24

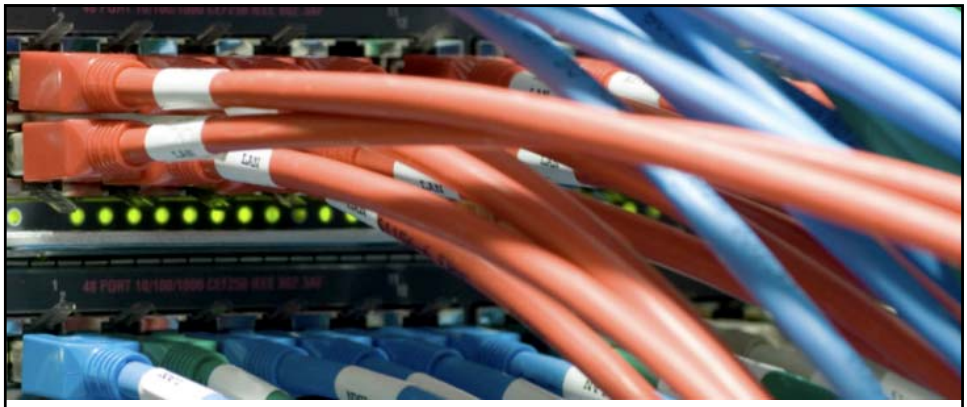

Three Important Interfaces
Modeling Environment
Data
Data Pre-processing
1
1
Model Producer
2
PMML Model
2
PMML Model
Deployment Environment
1
data
Model Consumer
3
scores
3
Post Processing
actions
25


Actually, This is a Typically a Component in a Workflow
26

Analytic Infrastructure

- We'll use the term *analytic infrastructure* to refer to the components, services, applications and platforms for managing data, preprocessing data, producing models, consuming models, post-processing scores, managing workflow, and related processes.

27



With the proper analytic infrastructure, cloud computing can be used for data preprocessing, for scoring, for producing models, and as a platform for other services in the analytic infrastructure.

28

Part 2.2

Cloud Programming Models for Analyzing Information Quality in Large Data

29

The Google Data Stack

- The Google File System (2003)
- MapReduce: Simplified Data Processing… (2004)
- BigTable: A Distributed Storage System… (2006)

30

Map-Reduce Example

- Input is file with one document per record
- User specifies map function
  - key = document URL
  - Value = terms that document contains

("doc cdickens",
"it was the best of times")

map

"it", 1
"was", 1
"the", 1
"best", 1

Example (cont'd)

- MapReduce library gathers together all pairs with the same key value (shuffle/sort phase)
- The user-defined reduce function combines all the values associated with the same key

key = "it"
values = 1, 1
key = "was"
values = 1, 1
key = "best"
values = 1
key = "worst"
values = 1

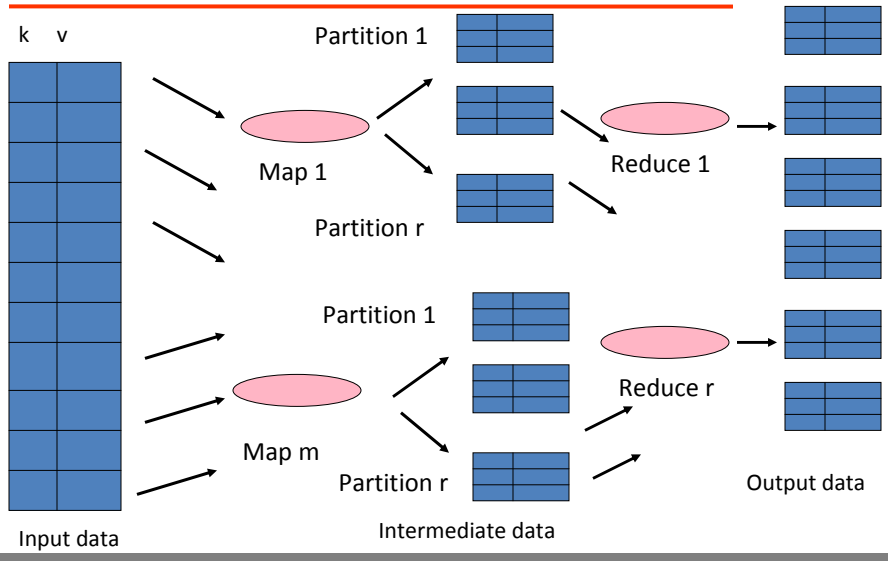
reduce

"it", 2
"was", 2
"best", 1
"worst", 1

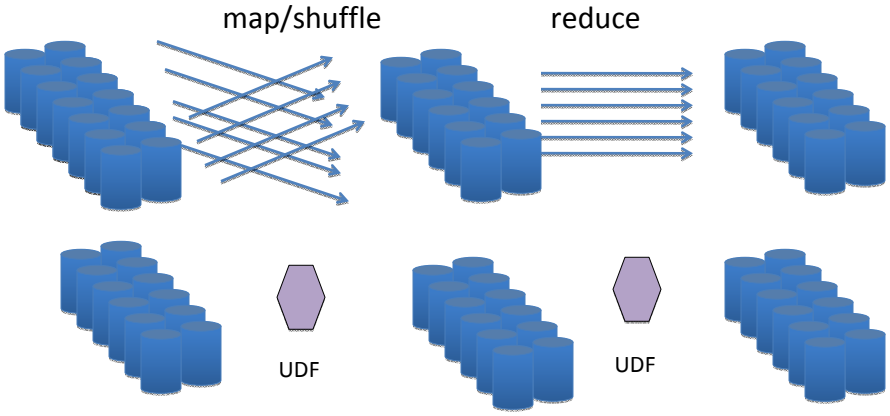
Map Reduce Summary

- All data is sequence of <key, value> pairs.
- Programmer specifies
  - Data is split and provided to Map.
  - Map takes input pair and produces a set of intermediate key-value pairs
  - MapReduce library takes all intermediate key value pairs with same key K and passes them to reduce function
  - Reduce function takes key K and collection of values and merges these values together. Input to Reduce need not fit in memory.
  - The Reduce functions produce the output.

MapReduce



Generalization: Apply User Defined Functions (UDF) to Files in Storage Cloud

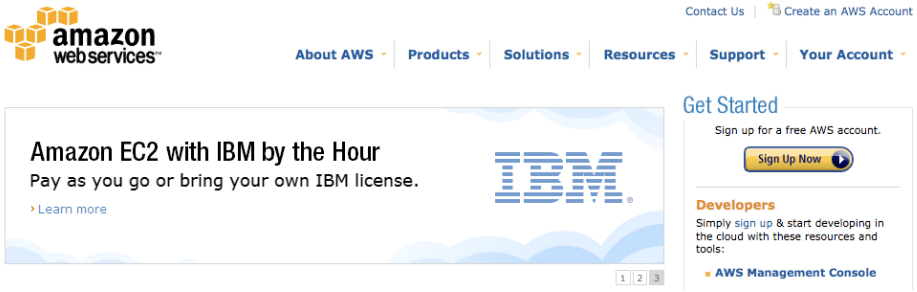


Sector (sector.sf.net) is an open source cloud with security that supports UDFs over the data in a storage cloud.

35

Part 2.3

Using Clouds for Information Quality Scoring (Model Consumers)



36

What is a Statistical/Data Mining Model?

- Infrastructure
  - Inputs: data attributes, mining attributes
  - Outputs, targets
  - Transformations
  - Segmented models, ensembles of models
- Models that are part of a standard
  - Trees, SVMs, neural networks, cluster models, etc.
  - In this case, only need to specify parameters
- Arbitrary models
  - e.g. arbitrary code that takes inputs to outputs

37

From an Architectural Viewpoint

- In an operational environment in which models are being deployed, it may be useful to "Just so no to viewing models as arbitrary code"
- The deployment can be much shorter if a scoring engine reads a PMML file instead of integrating a new piece of code containing a model.

38

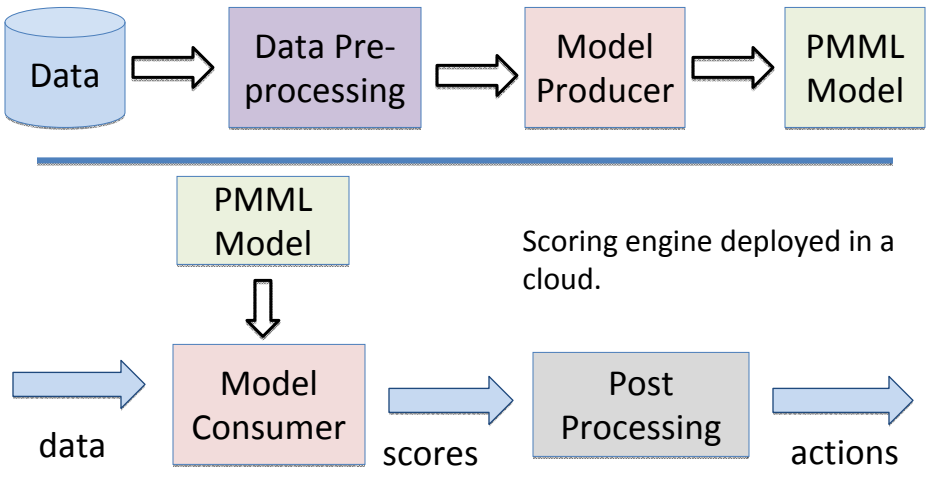
Model Producers/Consumers in Clouds

- Model Consumers take analytic models and use them to score data
  - Very easy to deploy in a cloud
  - Deploy a scoring engine in a cloud and then simply read PMML files
  - Very easy to scale up with cloud surges
- Model Producers take data and produce models
  - The decision to use a cloud requires weighing several factors
  - More difficult to parallelize

39

Sometimes it makes sense to the pre-processing in the cloud, especially if the data is there.

Modeling can be done in-house.

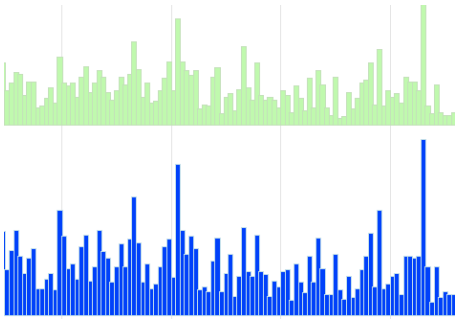


Scoring Engines

- IBM and SPSS have scoring engines
- Augustus (www.sourceforge.net/projects/augustus) is an open source PMML-compliant scoring engine.
- Zemantis has a scoring application that is already deployed in the Amazon cloud.

41

Part 3
Case Study: Baseline Models for Data Quality



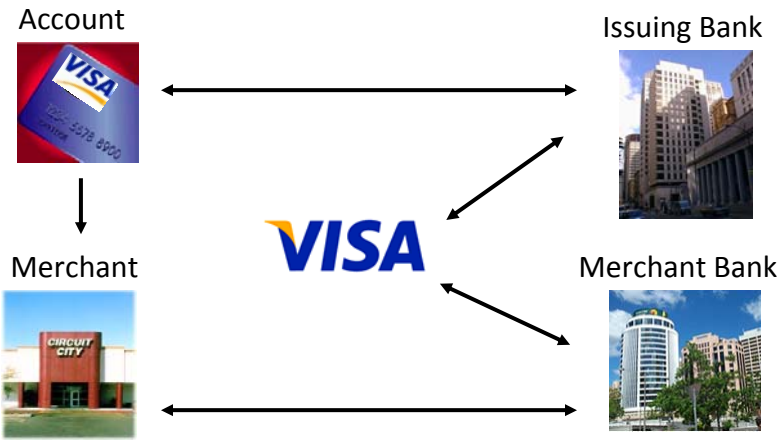
Joint work with Joe Bugajski, Burton Group

3.1 Problem

- How can we monitor, baseline, alert, and ameliorate data quality for high volume transaction systems?
- How can we tell when there are statistically significant changes with significant business or operational value?
- Examples:
  - Payment systems
  - Distributed sensor systems
  - Cyber defense systems

43

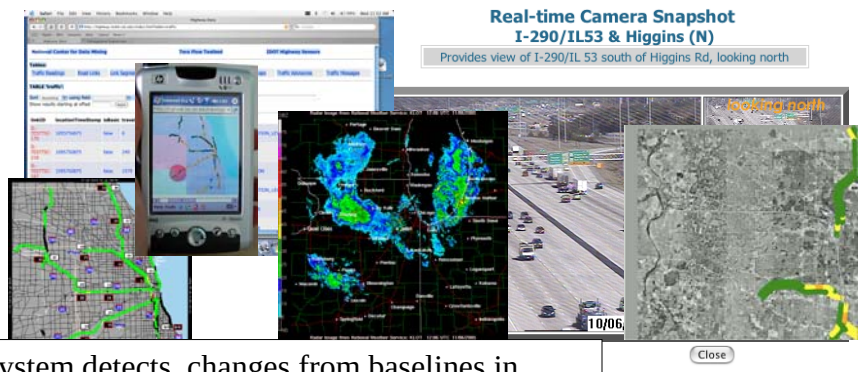
Example 1: Payment Systems



- 8000+ peak transactions per second.

44

Example 2: Highway Traffic Data



System detects changes from baselines in real-time and distributes them as alerts.

- 833 traffic sensors, 170,000 new sensor readings per day
- also image, text & semi-structured data (about 1 TB)

45

Challenges

- Large, high volume, complex, heterogeneous, distributed streaming data
- Multiple parties involved, each of which can modify the data in subtle ways
- System is sufficiently complex that establishing accuracy and other data quality dimensions is a challenge

46

Motivating Questions

- Are the payment field values and payment fields exceptions from this merchant (Starbucks, Cambridge, November, weekend etc.) ***different*** than the baseline?
- Is this traffic speed and volume today leaving this meeting (Thursday, July 16, 9 am, no special events, no rain) ***different*** than the baseline?
- Approach:
  - Establish baselines for each data quality dimension (completeness, declines, consistency, etc.)
  - Detect deviations from the baselines

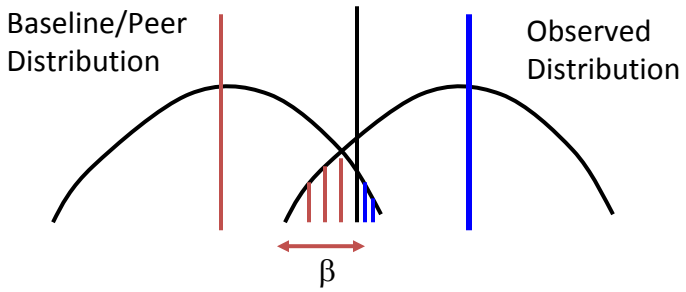
47

Change in Perspective

	Small, well understood systems	**Large, less well understood systems**
Begin	Specify what an accurate, complete, current & consistent (etc.) record	Divide and conquer using a data cube; establish baselines for each cell in data cube; compare baselines across cells
Monitor	Data quality dimensions (accuracy, completeness, etc.)	Monitor changes to baselines
Root cause analysis & amelioration	Identify root causes of identified data quality problems and ameliorate	Identify root causes of identified differences in baselines and ameliorate

48

3.2 Baselines Distributions



- Question: is the observed distribution different than the baseline distribution?
- Used CUSUM, Generalized Likelihood Ratio (GLR), threshold, and contingency table tests

49

Example: Single-Variant Distribution Changes

Baseline Model	
Value	Percentage
90	76.94
01	21.60
05	0.99
00	0.27
02	0.20
Total	100.00

Observed Model	
Value	Percentage
90	76.94
01	20.67
05	0.90
00	0.25
02	1.24
Total	100.00

50

Example: Bivariate Distribution Changes

Baseline Model	
Value	Percentage
90, -	0.13
90, blank	0.21
05, -	0.01
05, blank	0.01
etc.	etc.
Total	100.00

Observed Model	
Value	Percentage
90, -	0.13
90, blank	0.63
05, -	0.01
00, blank	0.11
etc.	etc.
Total	100.00

51

Example - Conditioned Distribution Changes

- We usually condition upon business events of interest
  - For example, declines
- We also generally use a peer population to understand whether a change has significance
  - Issuers in the same region
  - Merchants of the same Merchant Category Class (MCC)
  - etc.

Build baseline models condition on declines

52

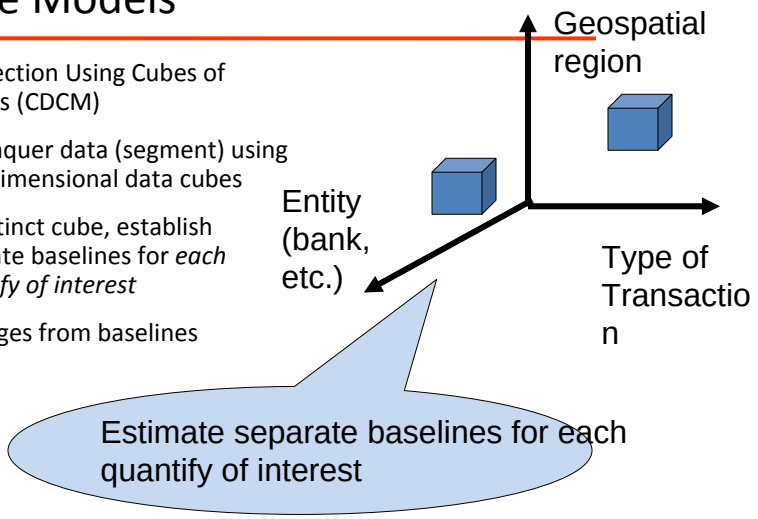
Idea 2: Scale with Data Cubes of Separate Baseline Models

Change Detection Using Cubes of Models (CDCM)

Divide & conquer data (segment) using multidimensional data cubes

For each distinct cube, establish separate baselines for *each quantify of interest*

Detect changes from baselines



PMML 4.0 has standardized multiple models to support data cubes of models applications.

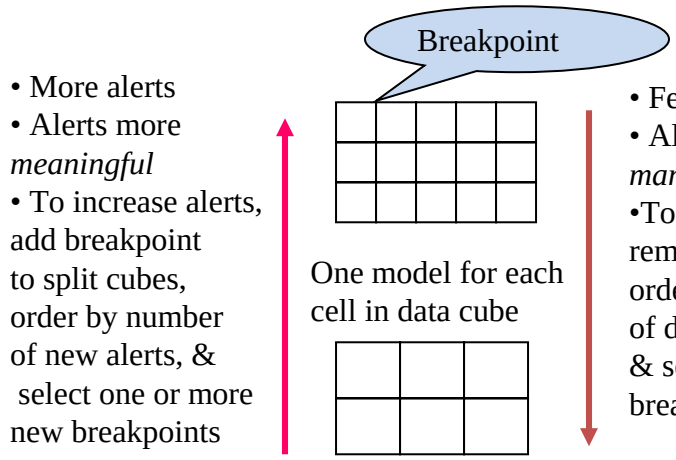
53

Examples - Change Detection Using Cubes of Models

- Visa Payment Systems
  - each field (21+) x each acquirer (thousands) x each merchants (thousands+)
  - 1,000,000+ baseline models used
- Highway Traffic Data
  - each day (7) x each hour (24) x each sensor (hundreds) x each weather condition (5) x each special event (dozens)
  - 50,000 baselines models used in current testbed

54

Greedy Meaningful/Manageable Balancing (GMMB) Algorithm



55

For More Information

Contact information:
Robert Grossman

blog.rgrossman.com

www.rgrossman.com

open data
www.opendatagroup.com

UIC
www.ncdm.uic.edu

56