

Introduction to Data Management CSE 344

Lecture 27: Misc Topics

Magda Balazinska - CSE 344, Fall 2012

1

Plan For Today

- Overview of various topics in data mgmt
 - A bit more about data integration
 - Data management and cloud computing

Magda Balazinska - CSE 344, Fall 2012

2

Data Integration and Data Cleaning

Magda Balazinska - CSE 344, Fall 2012

3

Motivation

- See first few slides of:
<http://www.cs.washington.edu/education/courses/cse544/09au/lecture-notes/lecture18/lecture18.pdf>
- Data integration raises the problem of duplicate records
- Goal: "resolve" the entities
 - Find matching entities

Magda Balazinska - CSE 344, Fall 2012

4

Step 1: Find *Similar* Items

- Define similarity function between two entities
 - Similarity estimates " $x \sim y$ "
 - For each attribute, use a string similarity function
 - Edit Distance
 - Jaccard Similarity of k-grams
 - Other
 - Combine similarity results for different attributes
- Return pairs with similarity above threshold

Magda Balazinska - CSE 344, Fall 2012

5

More on Similarity Functions

- No universally good similarity function
 - Cosine similarity
 - Hamming distance
 - Q-gram
 - Smith-Waterman distance
 - Soundex distance
 - TF/IDF
 - many more
- An interesting approach: compute all these distances! Each distance becomes a feature. Train a classifier to decide on similarity based on all these features.

Magda Balazinska - CSE 344, Fall 2012

6

Step 2: Merge Similar Items

While similar records exist

Identify two similar records r and s

Replace them with the output of $\text{merge}(r,s)$

Desirable properties of $\text{merge}(r,s)$

- Merge record with itself, get the record back
- Merging r with s = merging s with r
- Merge ($\text{merge}(r,s), t$) = merge($r, \text{merge}(s,t)$)
- If x is similar to y then $\text{merge}(x,y)$ is defined

Magda Balazinska - CSE 344, Fall 2012

7

Data Management as a Cloud Service

Magda Balazinska - CSE 344, Fall 2012

8

Examples Services Today

- Amazon SimpleDB, RDS, Elastic MapReduce Websites
 - Part of Amazon Web services
- Google App Engine Datastore Website
 - Part of the Google App Engine
- Microsoft SQL Azure
 - Part of Windows Azure
- Very dynamic space! Need to check docs regularly!

Magda Balazinska - CSE 344, Fall 2012

9

Cloud Computing

- A definition
 - "Style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet"
- Basic idea
 - Developer focuses on application logic
 - Infrastructure, software, and data hosted by someone else in their "cloud"
 - Hence all operations tasks handled by cloud service provider

Magda Balazinska - CSE 344, Fall 2012

10

Cloud Computing History

- "Computation may someday be organized as a public utility" (John McCarthy – 1960)
- Late 1990's: Infrastructure as a Service (i.e., rent machines)
- Late 1990's: Software as a service (e.g., Hotmail, Salesforce)
- Early 2000s: Web services
- 2006: Amazon Web Services
- And now it's a craze!

Magda Balazinska - CSE 344, Fall 2012

11

Levels of Service

- Infrastructure as a Service (IaaS)
 - Example Amazon EC2
- Platform as a Service (PaaS)
 - Example Microsoft Azure, Google App Engine
- Software as a Service (SaaS)
 - Example Google Docs

Magda Balazinska - CSE 344, Fall 2012

12

How About Data Management as a Service?

- **Running a DBMS is challenging**
 - Need to hire a skilled database administrator (DBA)
 - Need to provision machines (hardware, software, configuration)
 - If business picks up, may need to scale quickly
 - Workload varies over time
- **Solution: Use a DBMS service**
 - All machines are hosted in service provider's data centers
 - Data resides in those data centers
 - **Pay-per-use policy**
 - **Elastic scalability**
 - **No administration!**

Magda Balazinska - CSE 344, Fall 2012

13

Basic Features for Data Management as a Service

- Data storage and query capabilities
- Operations and administration tasks handled by provider
 - Include high availability, upgrades, etc.
 - **Elastic scalability**: Clients pay exactly for the resources they consume; consumption can grow/shrink dynamically
 - No capital expenditures and fast provisioning

Magda Balazinska - CSE 344, Fall 2012

14

Types of Data Management as a Service

Three different types exist at the moment

- Relational data management systems (e.g., SQL Azure)
- Simplified data mgmt systems (e.g., Amazon SimpleDB)
 - Also called "NoSQL" systems.
- Analysis services such as Amazon Elastic MapReduce

Magda Balazinska - CSE 344, Fall 2012

15



Amazon Web Services

- Since 2006
- "Infrastructure web services platform in the cloud"
- **Amazon Elastic Compute Cloud (Amazon EC2™)**
- **Amazon Simple Storage Service (Amazon S3™)**
- **Amazon SimpleDB™**
- **Amazon Elastic MapReduce™**
- And more...
- And growing...

Magda Balazinska - CSE 344, Fall 2012

16



Amazon EC2

- **Amazon Elastic Compute Cloud (Amazon EC2™)**
- Rent compute power on demand ("server instances")
 - Select required capacity: small, large, or extra large instance
 - Share resources with other users (multitenant): **Virtual machines**
 - Variety of operating systems
- Includes: Amazon Elastic Block Store
 - Off-instance storage that persists independent from life of instance
 - Highly available and highly reliable

Magda Balazinska - CSE 344, Fall 2012

17



Amazon S3

- **Amazon Simple Storage Service (Amazon S3™)**
 - "Storage for the Internet"
 - "Web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web."
- Some key features
 - Write, read, and delete uniquely identified objects containing from 1 byte to 5 TB of data each
 - Objects are stored in buckets. User chooses geographic area
 - A bucket can be accessed from anywhere
 - Authentication
 - Reliability

Magda Balazinska - CSE 344, Fall 2012

18



Amazon RDS

- Amazon Relational DB Service (Amazon RDS™)
 - Web service that facilitates set up, operations, and scaling of a relational database in the cloud
 - Full capabilities of a familiar MySQL or Oracle DBMS
- Some key features
 - Automated patches of DBMS
 - Automated backups for user-defined retention period
 - Elastic scalability but can only **scale-up**
 - Make your instance more powerful (CPU and memory)
 - Attach more storage to your instance
 - Can scale-out only by adding **read replicas**

Magda Balazinska - CSE 344, Fall 2012

19



Amazon SimpleDB

- An example of a NoSQL data management system
- See NoSQL Lecture

Magda Balazinska - CSE 344, Fall 2012

20

Some Current Research Topics

Magda Balazinska - CSE 344, Fall 2012

21

Some Research Topics in Database Community

- Big Data management and analytics
- Scaling OLTP (consistency issues, etc.)
- Multi-tenancy for OLTP and OLAP
- Privacy and security
- Exploiting new hardware
- Social data, crowd-sourcing + data management
- Pricing + data management
- Data stream processing
- etc.... see CIDR, SIGMOD, and other db confs.

Magda Balazinska - CSE 344, Fall 2012

22