Teaching Web-scale Data Management using Microsoft Azure: POSTECH Experiences

Seung-won Hwang
Associate Professor
CSE, POSTECH, Korea
Background I: POSTECH

- POhang university of Science and TECHNOlogy
- 25-yr old
- ~10 depts (engineering+science)
- ~30 undergrads/yr @CS
- ~20 CS faculty
Background II: Database

- Research: DB+Web
- Teaching: Undergraduate- and Graduate-level database
- Taught in Fall, 2011
  - ~40 students
  - Each week consists of:
    - 3 hrs of classroom teaching
    - 1.5 hrs of lab
Background II: Classic DB Curriculum

- Data representation: ER diagram, Relational model
- Query processing: SQL

<table>
<thead>
<tr>
<th>Model</th>
<th>Price</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3100</td>
<td>$549</td>
<td>4.5</td>
</tr>
<tr>
<td>D5100</td>
<td>$699</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Lab: SQL/DBMS (SQL Server, Oracle)
Background II: Classic DB Lab Projects

- DB on Web: DB-powered Web app
  - Minibase: DBMS for educational use (@wisc)
  - ProgresSQL: Open-source DBMS (@UCB)

- DB under the hood
  - Example projects
    - Index trees
    - Buffer manager
    - Rank query processing

<table>
<thead>
<tr>
<th>Model</th>
<th>Price</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3100</td>
<td>$549</td>
<td>4.5</td>
</tr>
<tr>
<td>D5100</td>
<td>$699</td>
<td>4.5</td>
</tr>
</tbody>
</table>
Why DB+Azure?

• The classic curriculum has remained (more or less) unchanged for many years
• Meanwhile, research and industry needs have changed drastically
  – Academia-industry gap?
Industry Buzzword: BigData

• **Wikipedia definition:**

  In information technology, **big data** consists of data sets that grow so large that they become **awkward** to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analytics, and visualizing.
Why awkward? 3Vs of Big Data

- **Volume**: Too large to store in one machine
- **Velocity**: Search/analytics is time sensitive
- **Variety**: Combines structured and unstructured (e.g., table+logs/text/video/audio)
Curriculum Design Goals

• Adding 3V challenges to projects using Azure
  – **Volume**: azure provides virtually limitless storage
  – **Velocity**: azure distributes computation over nodes
  – **Variety**: azure supports various types of storage needs
• Not losing relevance to classic materials (e.g., SQL/Web)– “backward compatible”
• Should not impose too much extra overhead
Design Specifics

• Build upon regular syllabus
  – Database Management Systems, Ramakrishnan et. al (3rd ed)

• SQL Labs (DBMS or SQL Azure)

• BigData Project using Twitter
  – Tables of user profiles
  – Social graphs of users
  – Storage/computation divided over multiple nodes
Project Specifics

• Twitter: 140M+ active users (as of 2012)

• follower relationship is asymmetric

• twitter.profiles (numeric_id int primary key, name varchar(20), screen_name varchar(16), friends_count int, followers_count int, following varchar(5), statuses_count int, favourites_count int, location varchar(40), description varchar(165), profile_image_url varchar(235), url varchar(100), created_at varchar(30), time_zone varchar(30), gender varchar(1), verified varchar(5), protected varchar(5))
Mutual Relationship Count

- Upload a graph to Azure blob
- Store the relationships to Azure table
- Read/Join tables to count mutual friends
  - Distribute/Parallelize the storage/workload!
- Join the result with the profile
- Build into an Web application
Web app code + project template provided

```csharp
// ToDo.UploadDataFileToBlobStorage
//
// input : fileContent - input stream of the file to be uploaded
//          blobContainer - blob container in which the file stored
// Return : none of Blob file created in Azure blob storage
// Purpose : Upload a local data file to Azure blob storage
// Note : 
//
// - To upload a file to blob storage:
//
// CloudBlockBlob blob = blobContainer.GetBlockBlobReference(blobAddressUri);
// blob.UploadFromStream(fileContent);
public static string UploadDataFileToBlobStorage(Stream fileContent,
    CloudBlobContainer blobContainer)
{
    string blobAddressUri = "follow_date";

    return blobAddressUri;
}
```
Upload to Azure blob storage

Web Role

Storage

Worker Role

Worker Role

Worker Role
Find mutual relationship

- Web Role
  - Find
  - Storage
  - Worker Role
    - 12 13
    - 12 14
    - 12 16
    - 12 19
    - 17 30
  - Worker Role
    - 510 18
    - 510 27
    - 510 320
    - 572 347
    - 607 419
    - 1076 573
    - 1076 589
    - 1077 101
  - Worker Role
    - Self-join
  - Worker Role
    - Self-join
  - Worker Role
    - Self-join
Count mutual relationships

Count

Web Role

Count

Worker Role
12 19 17 30 ...

Count

Worker Role
572 347 607 419 ...

Count

Worker Role
1087 2097 1090 1573 ...

Storage
Count mutual relationships
Get the profile of the user

SELECT name
FROM profiles
WHERE numeric_id = 247;

Get name

Web Role

SQL Azure

Storage

Worker Role

Worker Role

Worker Role
Emulator Screenshot
Video

- Emulator
- Azure Deployment
Outcome

• **High student satisfaction: 4.64/5.0**
  – 91% found exposure to Azure and SQL Azure useful for the course
  – 88% expected this would be useful for future careers

• **Experiences/findings disseminated:**
  – http://facultyresourcecenter.com
Summary

• DB+Azure was helpful for:
  – Motivating 3V challenges
  – Seeing DB problems in a new angle
• Developing/providing education resources were helpful significantly reducing learning curves
• Students find projects relevant and helpful
• Bigdata is relevant to all CS: mini-segment in other courses would be similarly effective
• Graduate project idea?
Thanks

http://www.postech.ac.kr/~swhwang
Elasticity (@CloudFuture’11)
Elasticity goal I – load balancing

Capacity expansion to deal with high load – Guarantee good performance
Elasticity goal II – power management

Capacity reduction to deal with low load
– Power saving
Thanks

http://www.postech.ac.kr/~swhwang