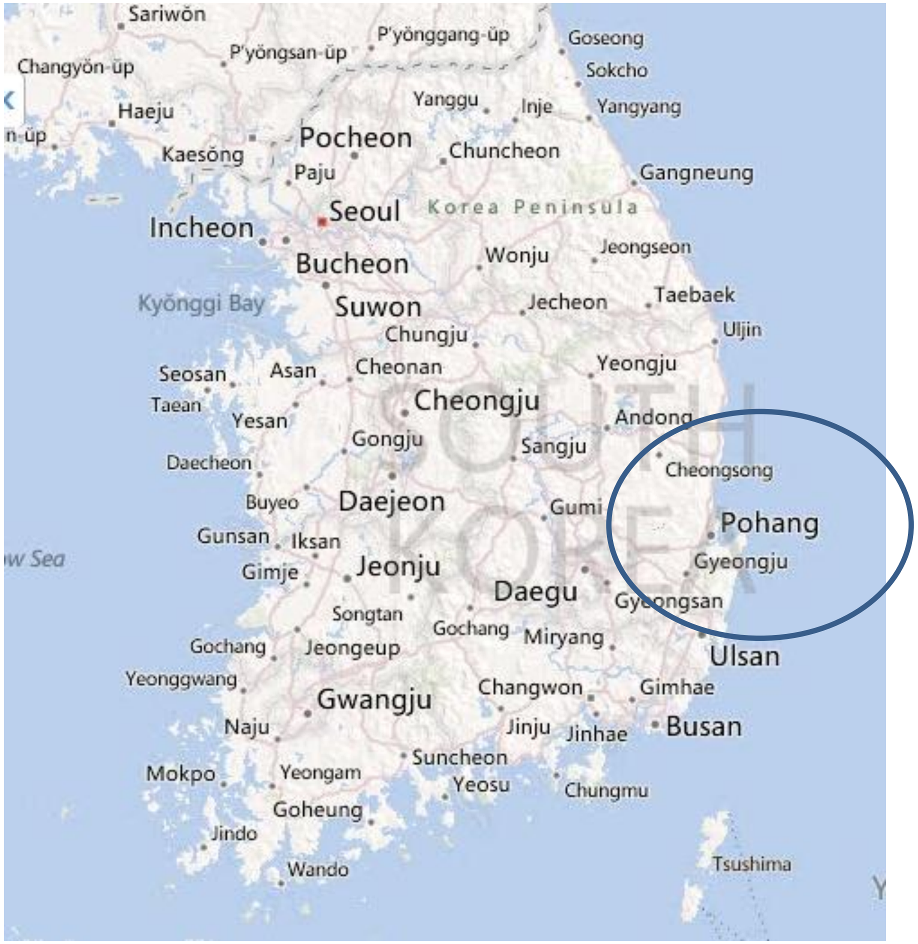
Teaching Web-scale Data Management using Microsoft Azure: POSTECH Experiences

Seung-won Hwang
Associate Professor
CSE, POSTECH, Korea

Background I: POSTECH



- **PO**hang university of **S**cience and **TECH**nology
- 25-yr old
- ~10 depts (engineering+scence)
- ~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

cameras@amazon

Model	Price	Review
D3100	$549	4.5
D5100	$699	4.5

Select *
from
where
price<600

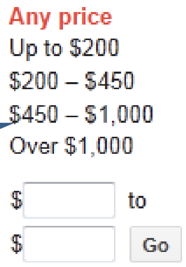
Model	Price	Review
D3100	$549	4.5

Lab: SQL/DBMS
(SQL Server,
Oracle)

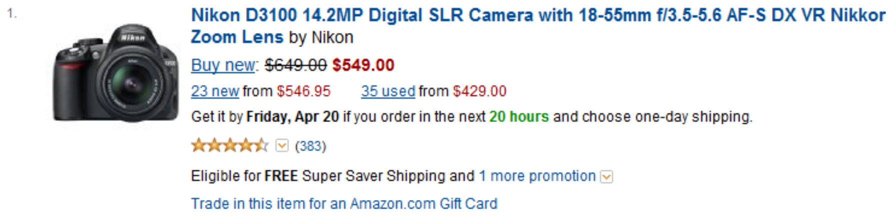
Background II:
Classic DB Lab Projects

- DB on Web: DB-powered Web app

Model	Price	Review
D3100	$549	4.5
D5100	$699	4.5



Any price
Up to $200
$200 – $450
$450 – $1,000
Over $1,000
$ to
$ Go



1.

Nikon D3100 14.2MP Digital SLR Camera with 18-55mm f/3.5-5.6 AF-S DX VR Nikkor Zoom Lens by Nikon

Buy new: ~~$649.00~~ $549.00

23 new from $546.95 35 used from $429.00

Get it by **Friday, Apr 20** if you order in the next 20 hours and choose one-day shipping.

★★★★☆ (383)

Eligible for **FREE** Super Saver Shipping and 1 more promotion

Trade in this item for an Amazon.com Gift Card

- DB under the hood
  - Minibase: DBMS for educational use (@wisc)
  - ProgresSQL: Open-source DBMS (@UCB)
- Example projects
  - Index trees
  - Buffer manager
  - Rank query processing

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?

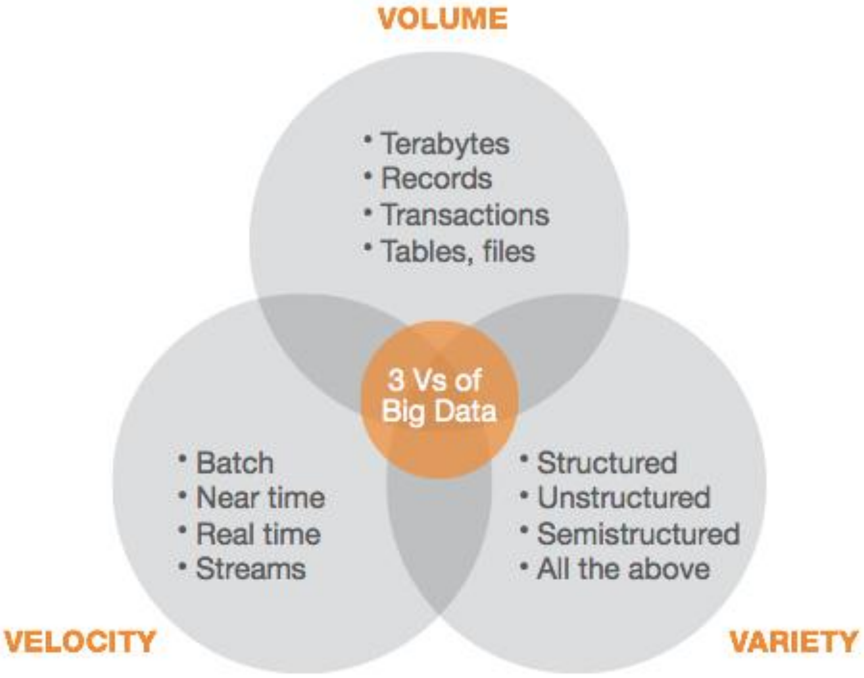
Industry Buzzword: BigData

- 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 BigData

- 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

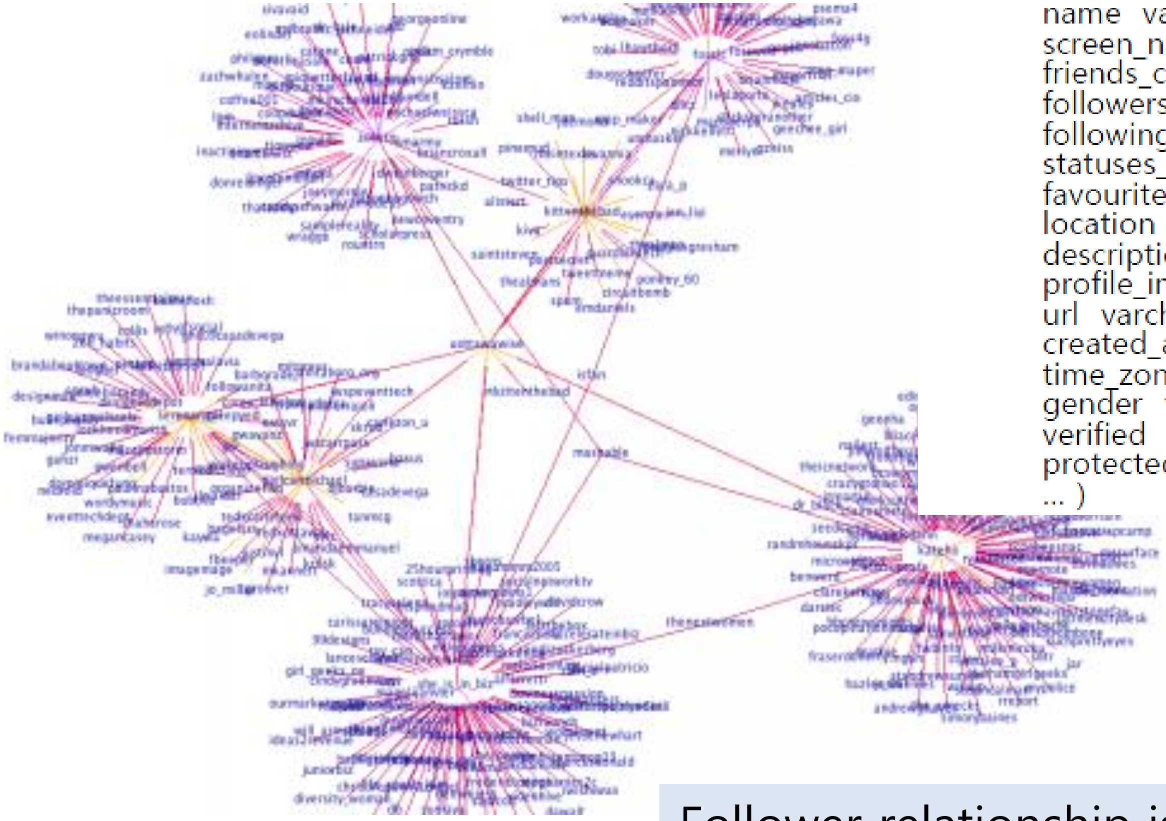
- Adding 3V challenges to projects using Azure
  - Volume: azure provides virtually limitless storage
  - Velocity: azure distributes computation over nodes
  - Variety: azure supports various types of storage needs
- Not losing relevance to classic materials (e.g., SQL/Web)– "backward compatible"
- 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)



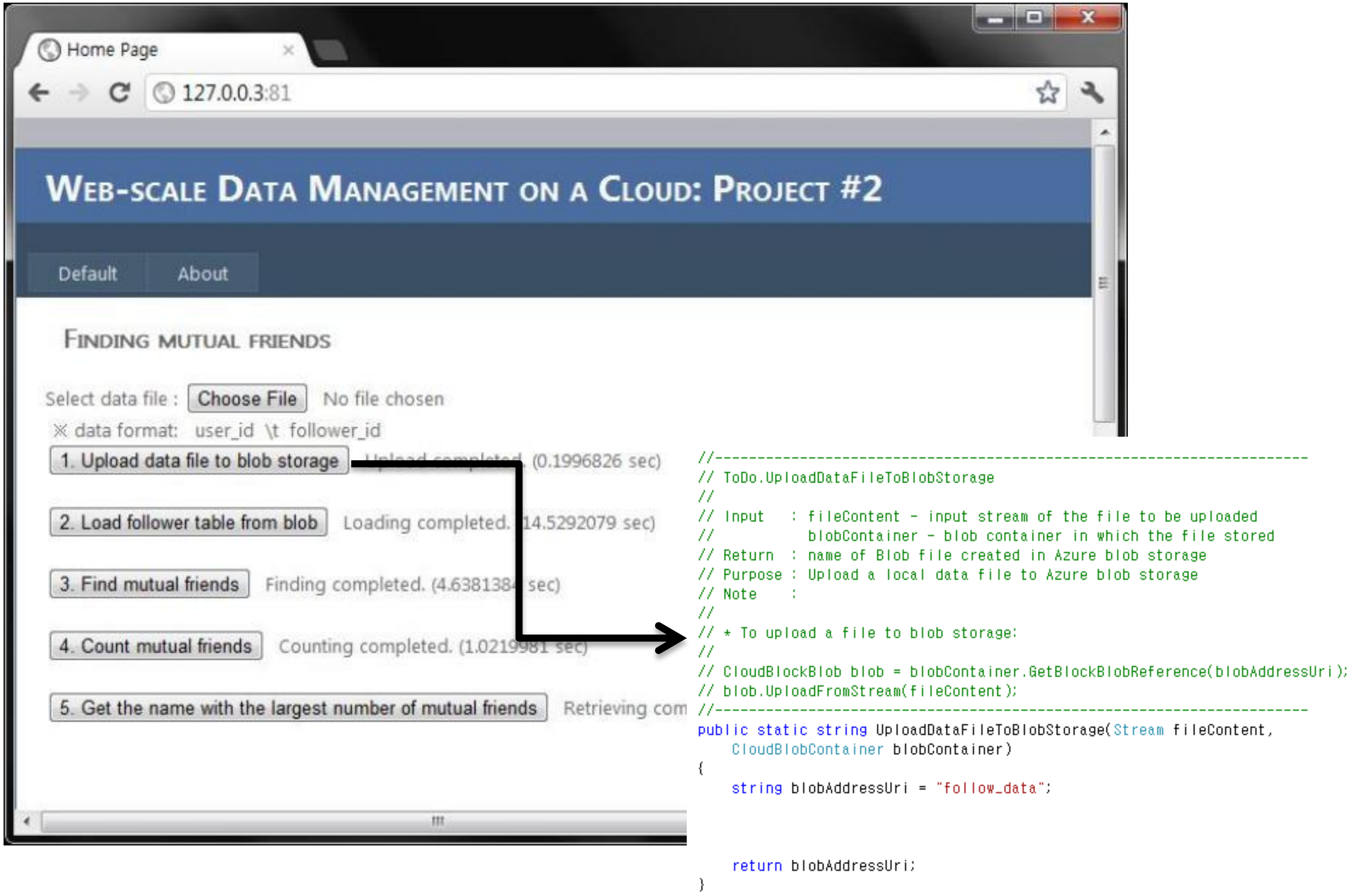
- twitter.profiles (
  numeric_id int primary key,
  name varchar(20),
  screen_name varchar(16),
  friends_count int,
  followers_count int,
  following varchar(5),
  statuses_count int,
  favourites_count int,
  location varchar(40),
  description varchar(165),
  profile_image_url varchar(235),
  url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

Follower relationship is asymmetric

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



FINDING MUTUAL FRIENDS

Select data file : Choose File No file chosen

※ data format: user_id \t follower_id

1. Upload data file to blob storage   Upload completed. (0.1996826 sec)

2. Load follower table from blob   Loading completed. (14.5292079 sec)

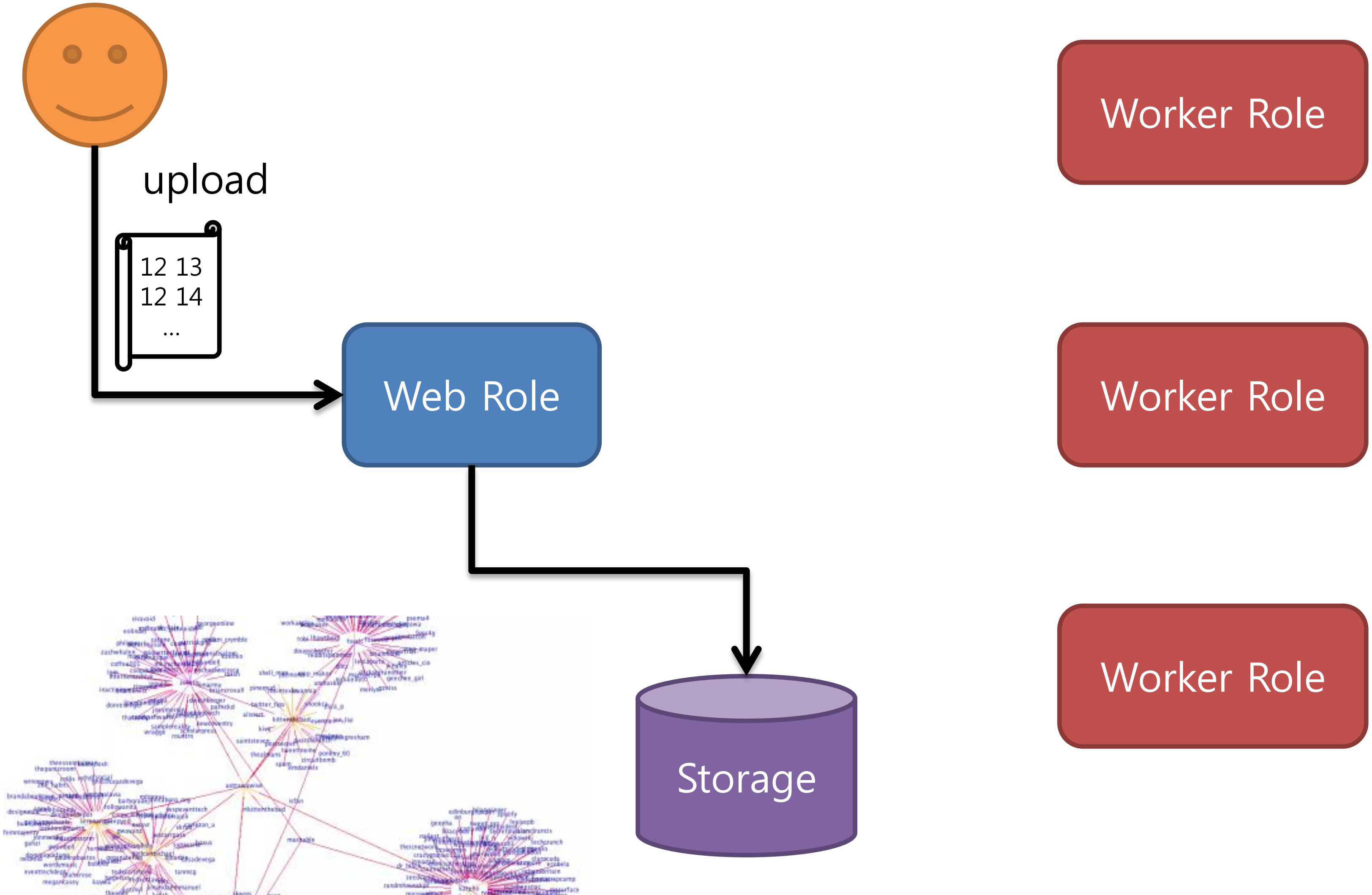
3. Find mutual friends   Finding completed. (4.6381384 sec)

4. Count mutual friends   Counting completed. (1.0219981 sec)

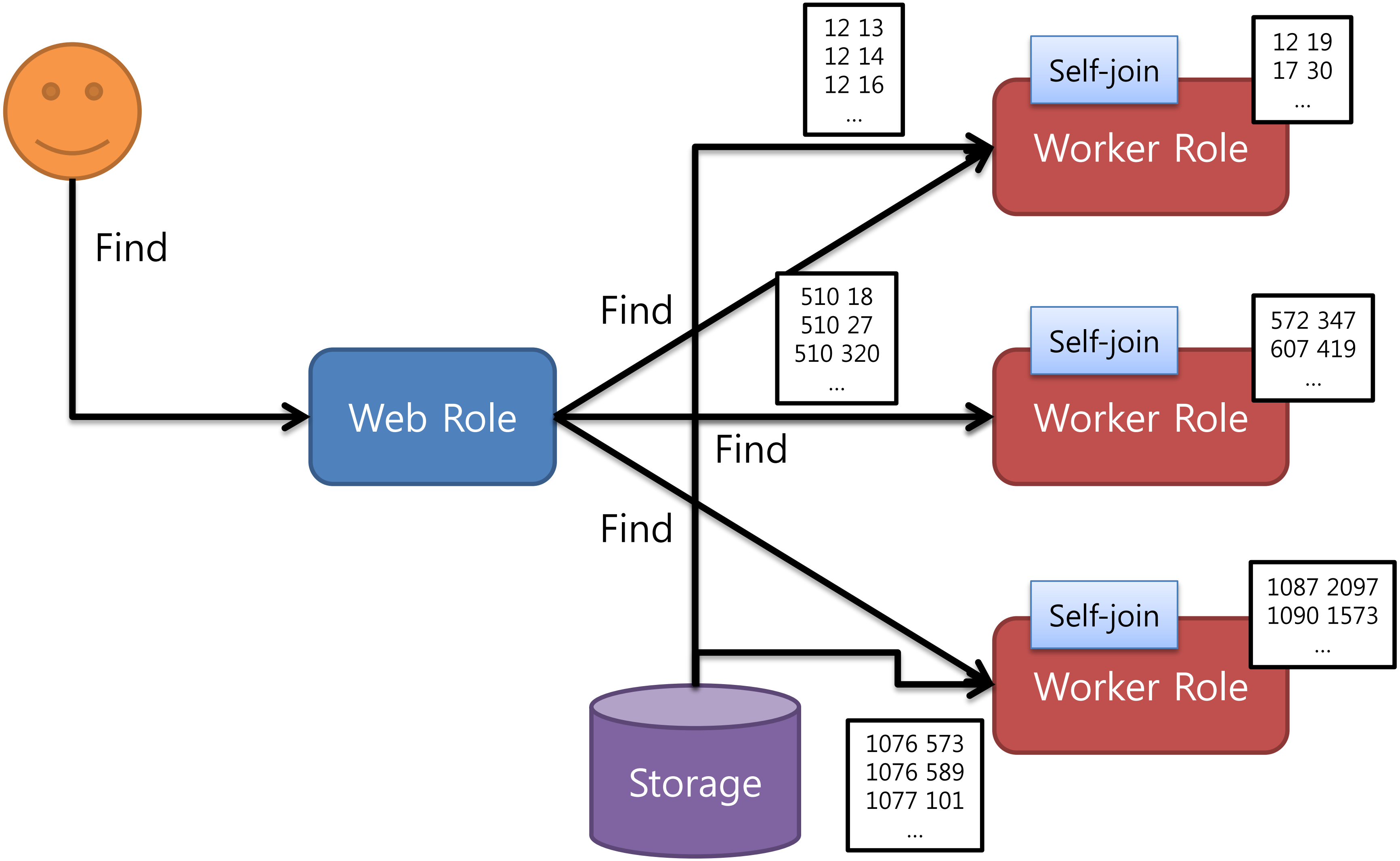
5. Get the name with the largest number of mutual friends   Retrieving com

```
//------------------------------------------------------------------------
// ToDo.UploadDataFileToBlobStorage
//
// Input   : fileContent - input stream of the file to be uploaded
//           blobContainer - blob container in which the file stored
// Return  : name 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_data";



    return blobAddressUri;
}
```

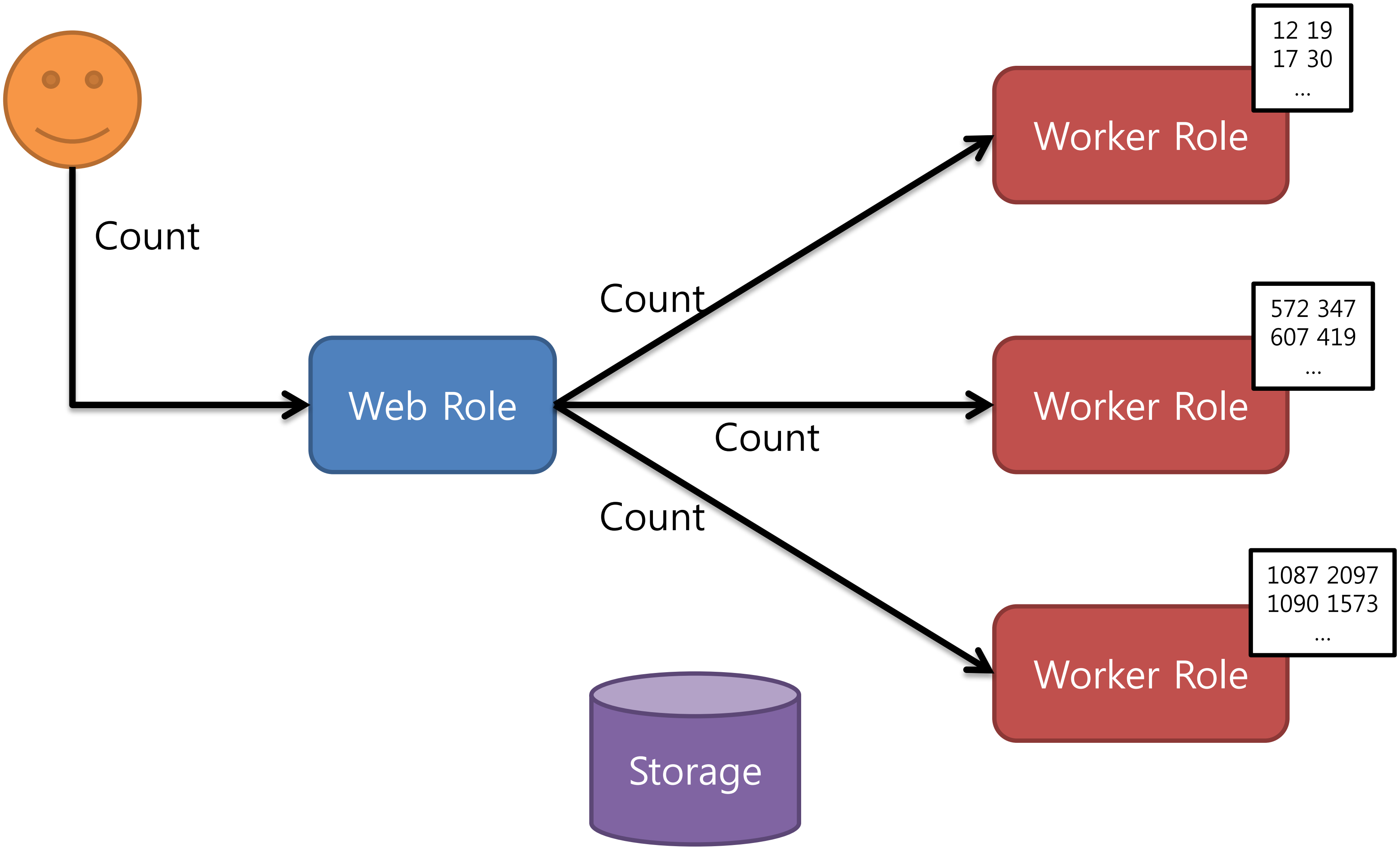
Upload to Azure blob storage



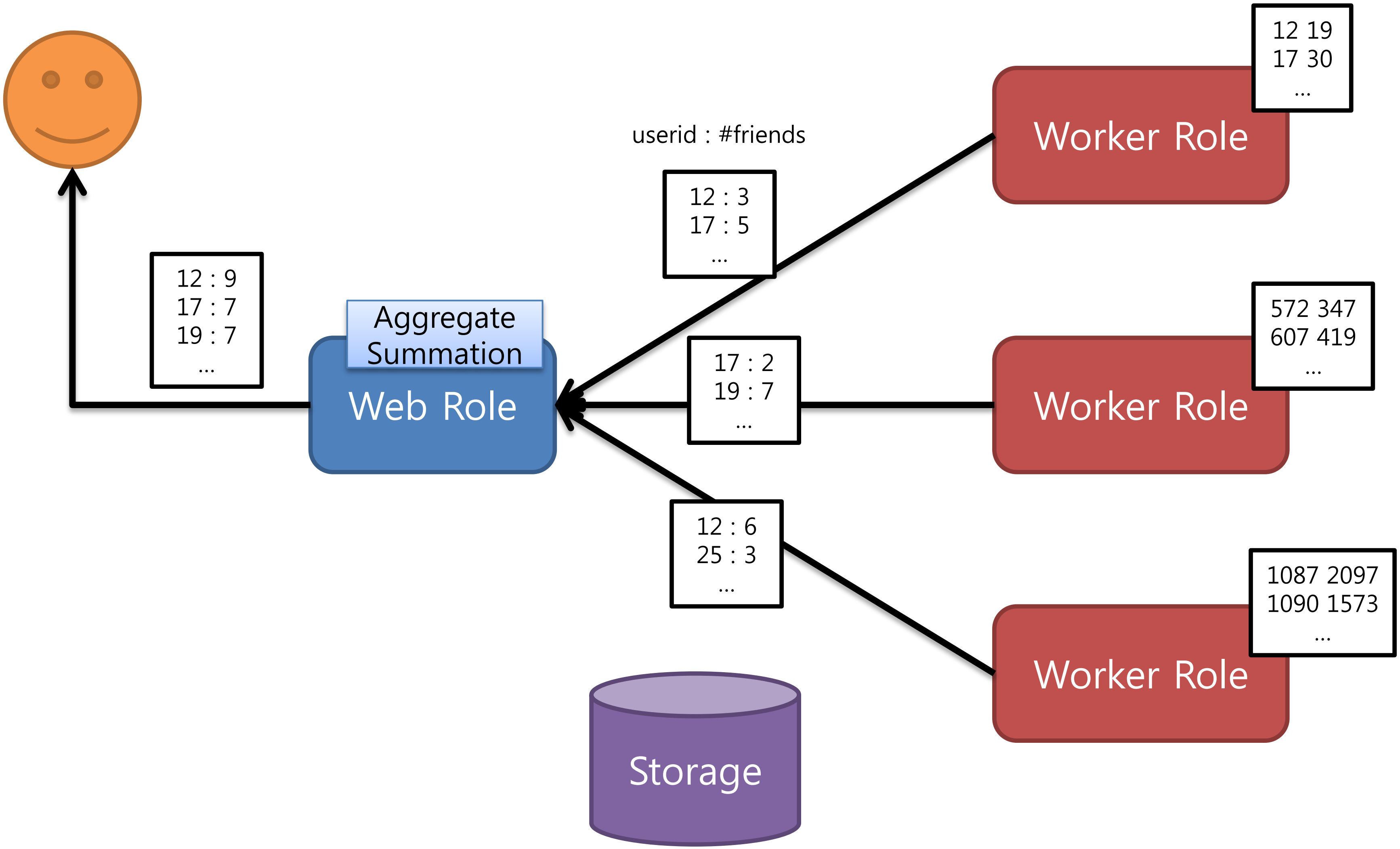
Find mutual relationship



Count mutual relationships



Count mutual relationships



12 : 9
17 : 7
19 : 7
…

Aggregate Summation

Web Role

userid : #friends

12 : 3
17 : 5
…

17 : 2
19 : 7
…

12 : 6
25 : 3
…

Storage

Worker Role

12 19
17 30
…

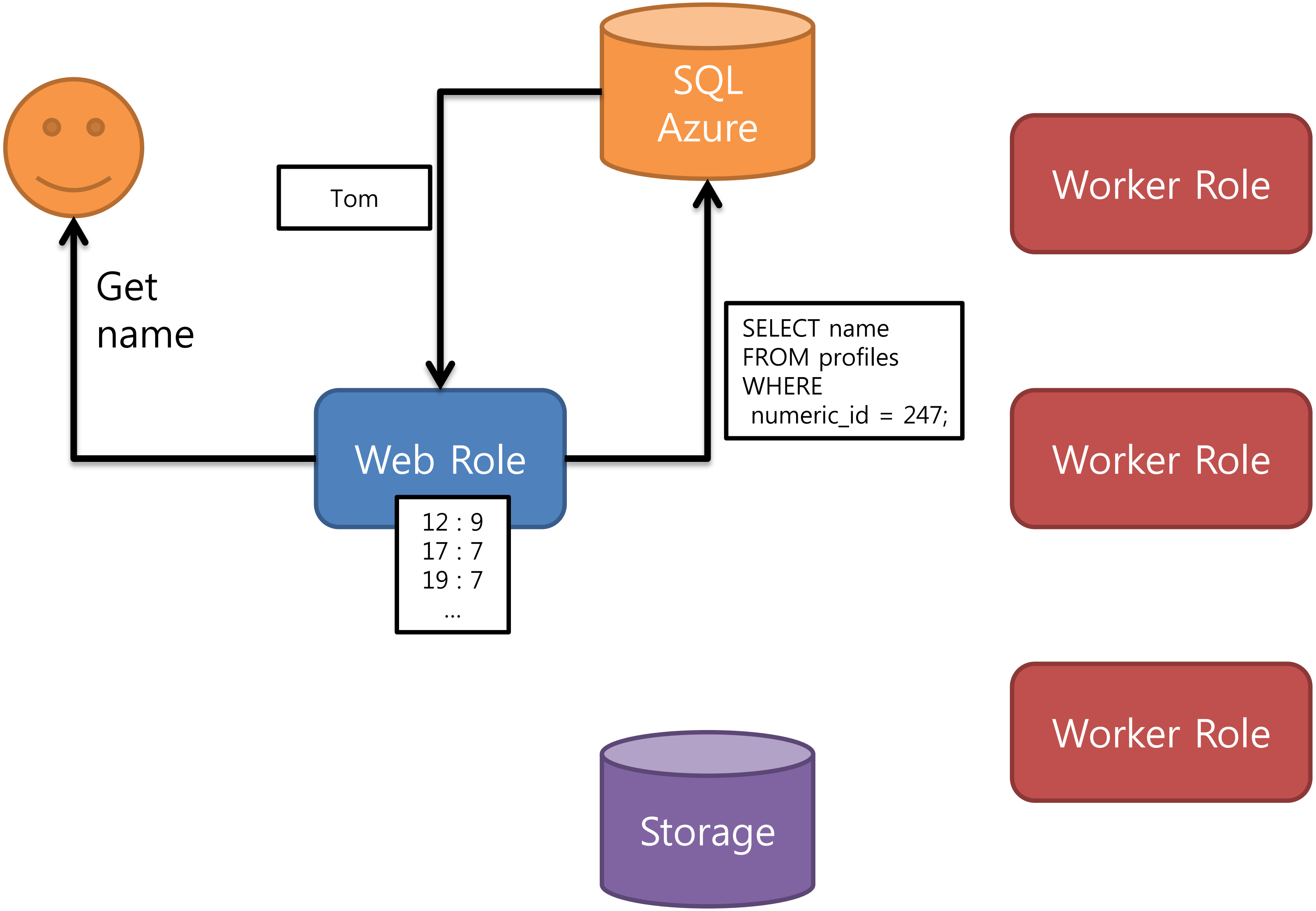
Worker Role

572 347
607 419
…

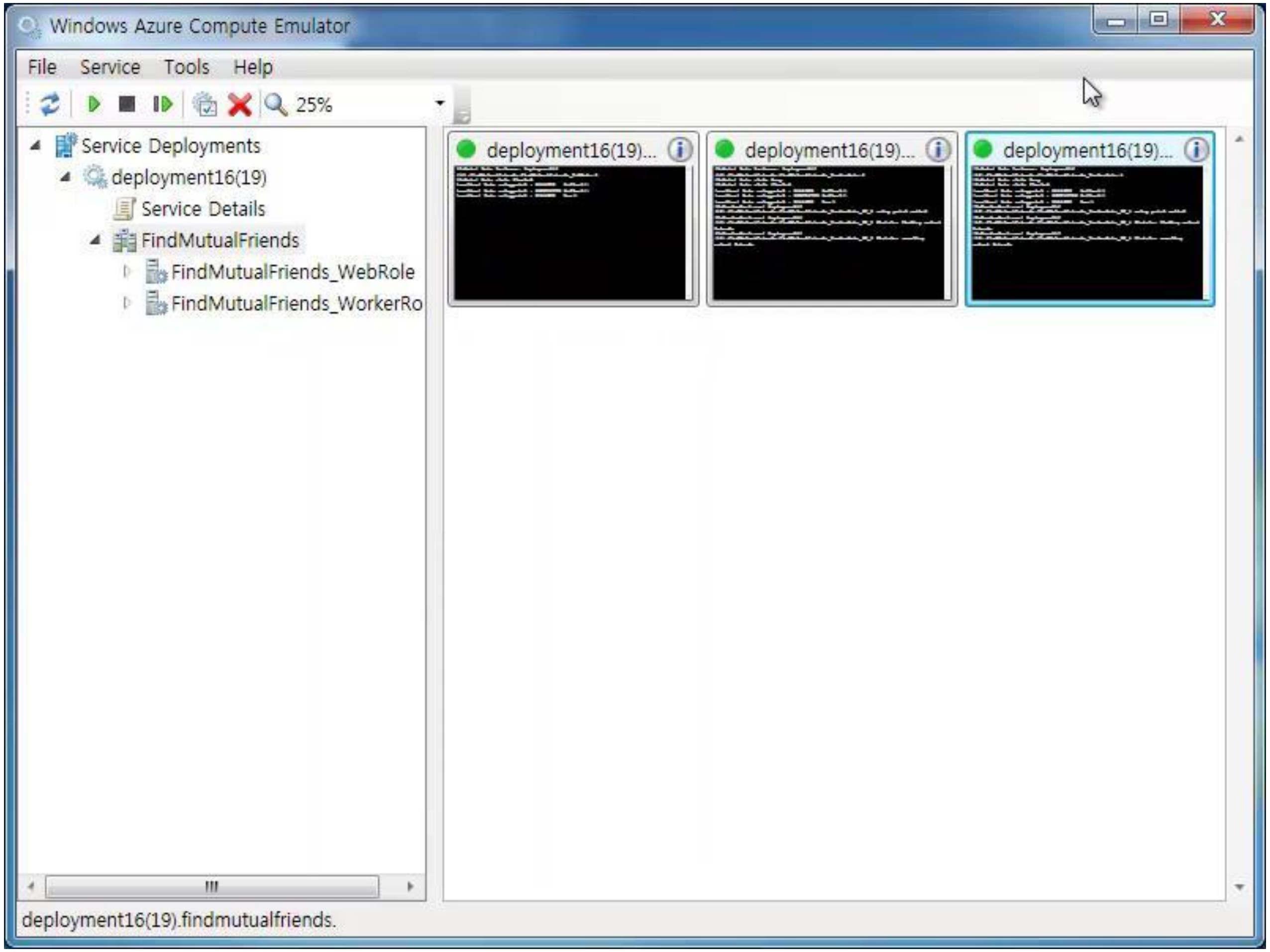
Worker Role

1087 2097
1090 1573
…

Get the profile of the user



Emulator Screenshot

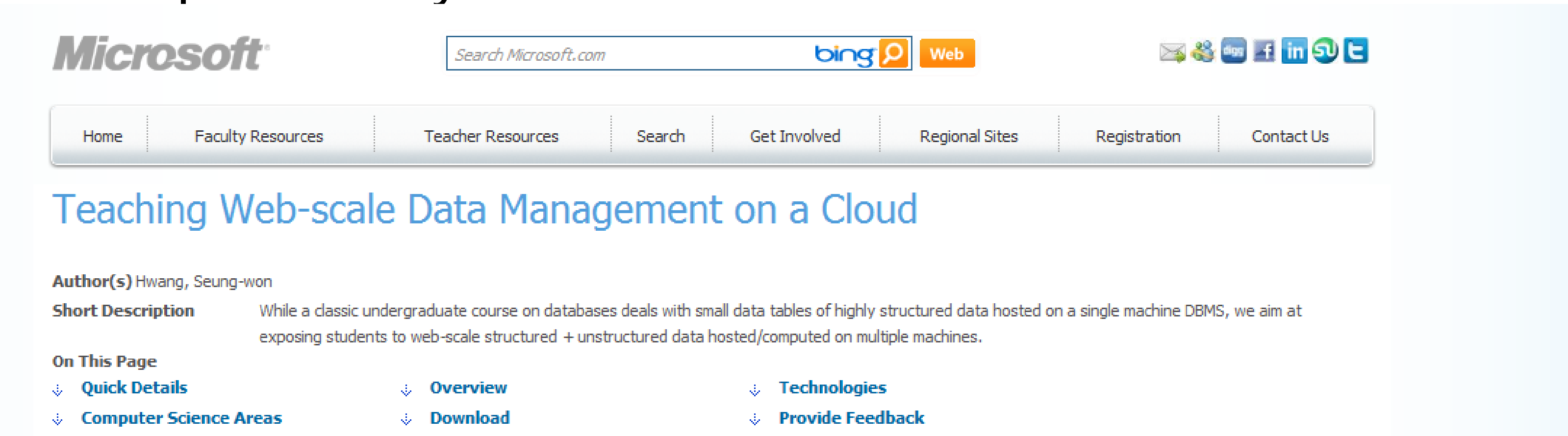


Video

- 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



Teaching Web-scale Data Management on a Cloud

Author(s) Hwang, Seung-won

Short Description While a classic undergraduate course on databases deals with small data tables of highly structured data hosted on a single machine DBMS, we aim at exposing students to web-scale structured + unstructured data hosted/computed on multiple machines.

On This Page

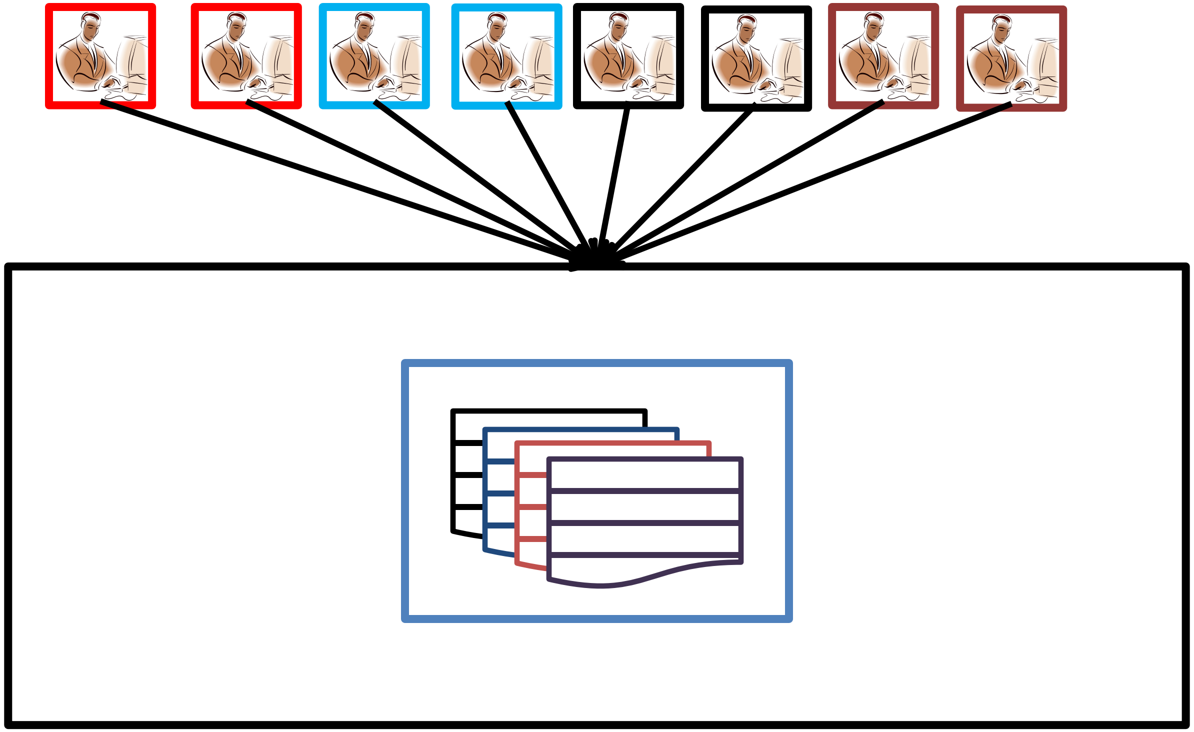
Summary

- 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

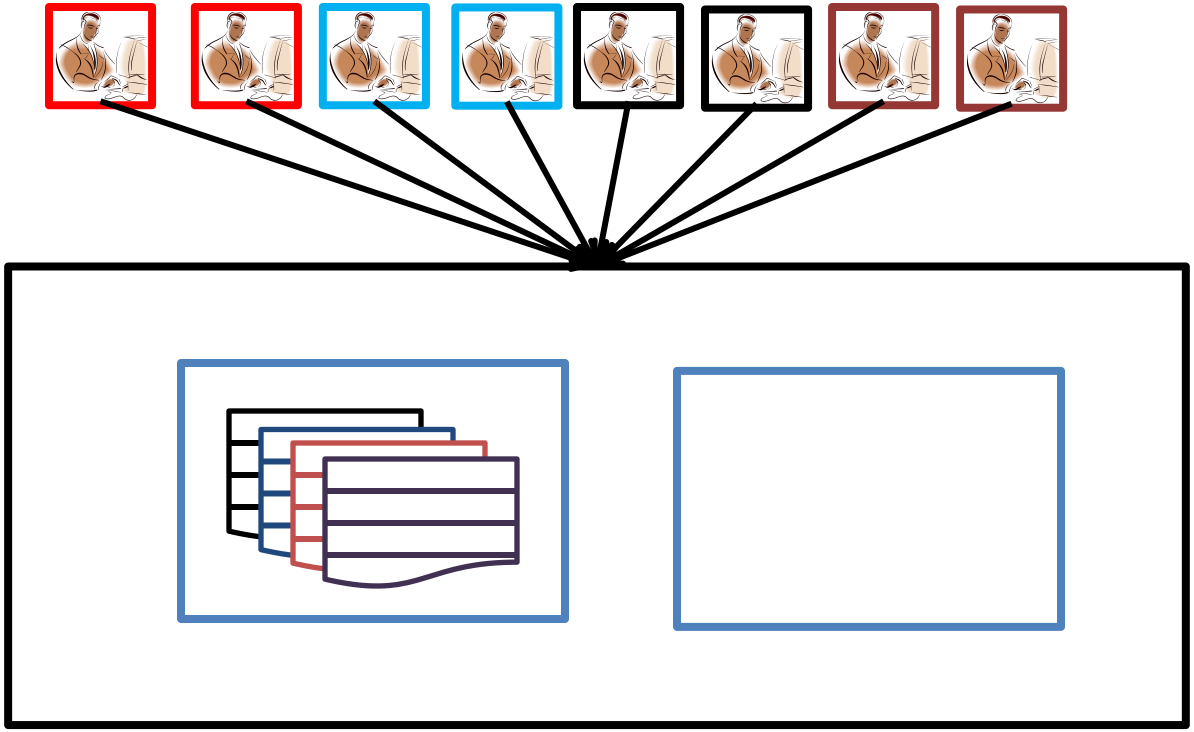
http://www.postech.ac.kr/~swhwang

Elasticity(@CloudFuture’11)



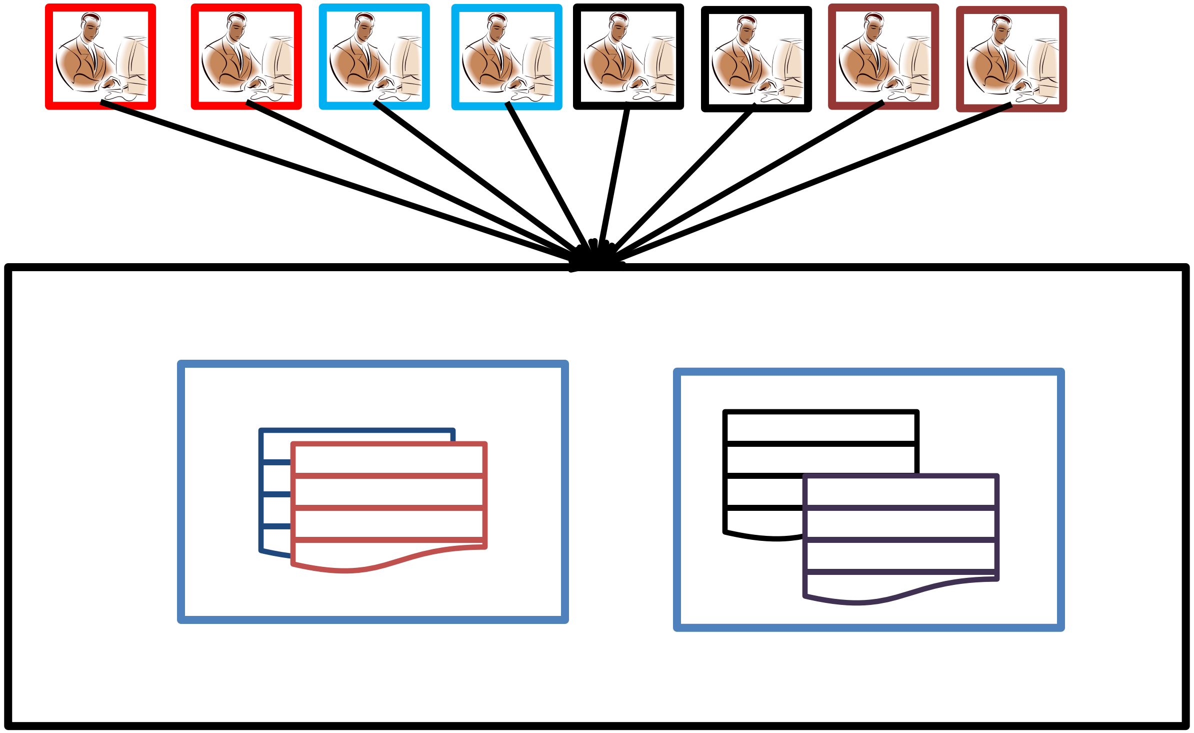
Elasticity goal I – load balancing

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