CS15-319 / 15-619 Cloud Computing

Recitation 7

October 7th and 9th, 2014

15-619 Project

- Students **enrolled in 15-619**
  - Since 12 units, an extra project worth 3-units
  - Project will be released this week
- **Team project** for a team up to 3 students
  - The team registration was due on last Saturday, Oct 4th

    ~~https://docs.google.com/forms/d/1SGmnh0yII81lZIHDy-lf-lj_MrJXAvj0WE7Y23c1dSg/viewform~~ (Closed)

    - If not on a team, contact the Prof.
  - Keep record of the team name, you will need it!

Upcoming deadlines

- Project 3:

Project 3			
	Files vs. Databases		
	File vs. Database	Checkpoint	Available 10/6/14 12:01 AM Due 10/12/14 11:59 PM

- Unit 3:

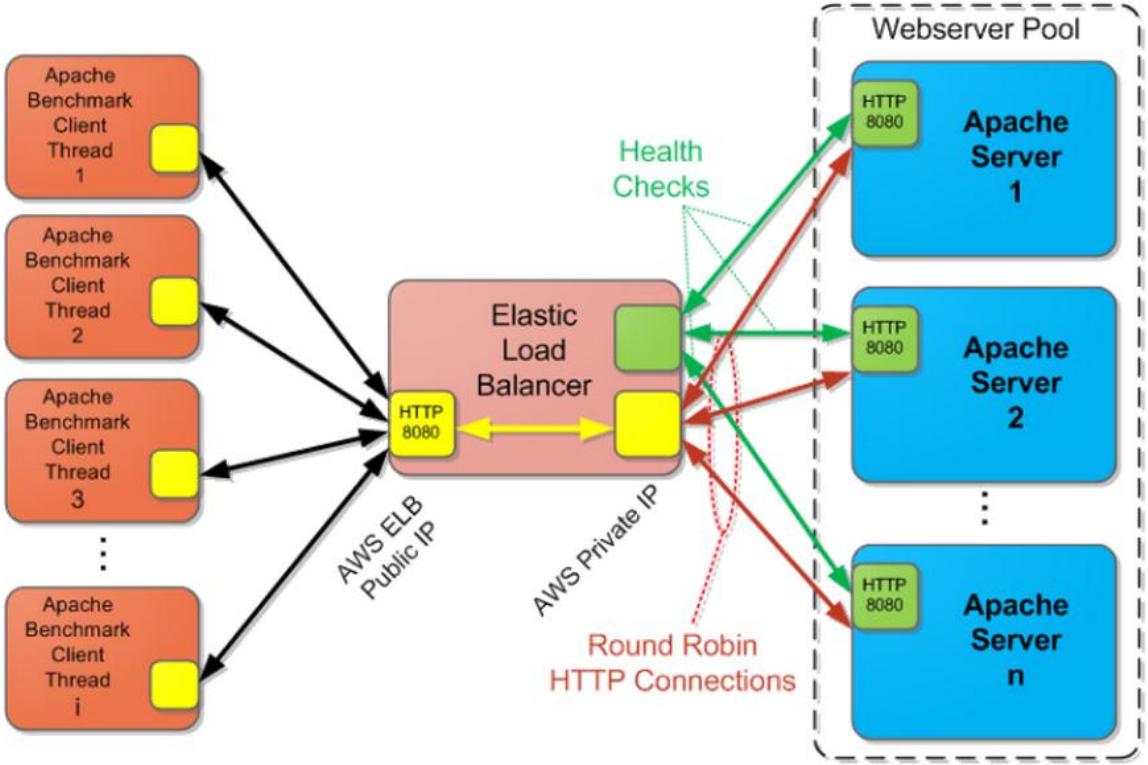
UNIT 3: Virtualizing Resources for the Cloud			
	Module 6: Introduction and Motivation		
	Module 7: Virtualization		
	Module 8: Resource Virtualization - CPU		
	Module 9: Resource Virtualization - Memory		
	Module 10: Resource Virtualization – I/O		
	Module 11: Case Study		
	Quiz 3: Virtualizing Resources for the Cloud	Checkpoint	**Available Now**. Due 10/9/14 11:59PM Pittsburgh

Due on Thursday!
Pay attention to the duration change!

Project 2 Reflections

You have experienced:

- Vertical and Horizontal Scaling
- Elastic Load Balancing
- Auto Scaling Group
- Cloud Watch
- SNS
- Failure



Project 2.3 Piazza Questions

- Zero RPS:
  - did not open all ports on ELB, or
  - did not detect instance failure
- Instances may automatically shut down
  - you need to be able to track this
- Low CPU utilization
  - warm up the ELB

This Week

- UNIT 3: Virtualizing Resources for the Cloud
  - Module 6: Introduction and Motivation
  - Module 7: Virtualization
  - Module 8: Resource Virtualization - CPU
  - Module 9: Resource Virtualization - Memory
  - Module 10: Resource Virtualization – I/O
  - Module 11: Case Study
  - Quiz 3: Virtualizing Resources for the Cloud

Project 3

Project 3.1

- Files vs. Databases
  - File vs. Database

Project 3.1 Hints

- Troubles with loading data using the keyword LOCAL

For those interested, another option that enables the use of the LOCAL INFILE command is detailed in documentation here: http://dev.mysql.com/doc/refman/5.5/en/load-data-local.html

- LOAD DATA INFILE failed

Try using LOAD DATA LOCAL INFILE instead, and then just provide "million_songs_metadata.csv" rather than the full path.

The LOCAL keyword will specifically tell the MySQL CLI to read the file from the client host (to search locally). Without the LOCAL keyword, this query will search for a remote server by default.

Project 3.1 Overview

- Given two files as datasets
- Run basic Unix commands like grep, awk etc to extract certain data
- Use databases to extract similar data
- Answer questions in checkpoint
  - Compare flat files to databases

Files vs. Databases

- Eg. Students' course registration
- Fields
  - First name
  - Last name
  - Course number
  - Section
  - …

Flat Files

- Computer-based flat files
- Program to access
- Store all data in a large table

Ex: Students' courses

- A comma-separated 'csv' file
- Layout:

  Lina, Li, 15619, A
  Jialiang, Lin, 15319, A
  Jiten, Mehta, 15619, A

- Problem?

Flat Files

- Accessing data is inconvenient
- Lacking knowledge of file-layout
- Data redundancy
- Concurrent-access
- …

Databases

- Database
  - Organized collection of data
  - Refers to the data themselves and supporting data structures
- Database management system (DBMS)
  - A software application
  - Interface between user and databases
  - Capture and analyze data

Databases

- Relational databases
  - Data is stored as fixed-length fields in normalized tables
  - MySQL
  - SQL (Structured Query Language)

ID	FirstName	LastName	Course	Section
1	Lina	Li	15619	A
2	Jialiang	Lin	15319	A
3	Jiten	Mehta	15619	A

Databases

- Advantages
  - Logical and physical data independence
  - Concurrent access and transaction support

- Disadvantages
  - Price
  - Additional expertise
  - Complex, difficult and time consuming to design
  - Damage affects all programs

SQL Database Introduction

- SQL History
  - Structured Query Language
  - ANSI Standard in 1986, ISO in 1987
- Relation Model
  - Structure data model

ID	Name	Email	Age
45637	Mike	mike@example.com	23
12456	Jake	jake@example.com	25

- Relational language
  - DDL: Database Definition Language
  - DML: Database Manipulation Language

MySQL Introduction

- Data Definition Language
  - Table definition:
    - column definition(name, length, data type), constraints(key)
  - Data type:
    - INT, CHAR,NUMERIC, DATE etc
- Example: create a table “students”
  - How many columns do we need?
  - What’s the name of the column?
  - What datatype should be applied?
  - What’s their relationship?

MySQL Example

- Create a table

```
CREATE TABLE students
(
ID int,
Name varchar(255),
email varchar(255)
);
```

id	name	email

MySQL Example

- Insert into a table

```
INSERT INTO students
VALUES ('1','Mike','mike@example.com');
```

id	name	email
1	Mike	mike@example.com

MySQL Example

- Insert into a table

```
INSERT INTO Persons
VALUES ('1','Mike','mike@example.com');

INSERT INTO Persons
VALUES ('2','Jake','jake@example.com');
```

id	name	email
1	Mike	mike@example.com
2	Jake	jake@example.com

MySQL Example

- Insert into a table

```
INSERT INTO Persons
VALUES ('1','Mike','mike@example.com');

INSERT INTO Persons
VALUES ('2','jake','jake@example.com    ');

INSERT INTO Persons
VALUES ('3','Bob','bob@example.com');
```

id	name	email
1	Mike	mike@example.com
2	Jake	jake@example.com
3	Bob	bob@example.com

MySQL Manipulation

- Data Manipulation Language
  - select, from, where, renaming
  - set operation
  - ordering
  - aggregate function
  - nested subqueries
  - join: left, right, inner, full
  - group by

Files vs. Databases

- Compare flat files vs. MySQL
- Answer:
  - What are the advantages and disadvantages using flat files / databases?
  - In what situation would you use a flat file / database?
  - How to build your own databases? How to manipulate it?

NoSQL Databases

- NoSQL means non-relational
  - Organized as Key-Value pairs: DynamoDB, Apache Cassandra

Key

Key (ID)
45637
12456

Name	Email	Age
Mike	mike@example.com	23
Jake	jake@example.com	25

value

  - Columnar DB: HBase, Vertica

redis

mongoDB

APACHE HBASE

cassandra

Statistical Data:

- Number of posts in 2013: 40 million
- Number of comments in 2013: 404 million
- Number of visitors: 100 million monthly unique
- Number of employees: 28

Technic:

- Python, 2005
- AWS, 2009
- Cloud Watch, 2012
- Database?

- Database Choices
  - Flat Files
    - Think of a spreadsheet
  - Relational Databases
    - MySQL
    - PostgreSQL
  - Non-relational Databases (NoSQL)
    - HBase
    - DynamoDB
    - Apache Cassandra

ID	UPS	DOWNS	TITLE	URL
12345	120	34	Buffins Create Zombie Dog!	www.someaussiesite.co.au/dog.html
12346	3	24	Check out my new blog!	noobspamer.blogspot.com
12347	509	167	Pee in a sink if you've ever voted up.	self

*Steve Huffman - Lessons learned while at reddit.com

reddit

ID	UPS	DOWNS	TYPE
12345	120	34	Link
12346	3	24	Link

Thing

Data

THING_ID	KEY	VALUE
12345	Title	Boffins Create Zombie Dog!
12345	URL	www.someaussiesite.com.au/zombiedog.html
12346	Title	Pee in a sink if you've ever voted up.
12346	URL	self

*Steve Huffman - Lessons learned while at reddit.com

Reminders

- Tag your instance: Key: **Project**, Value: **3.1**
- Be sure not to terminate the instance before finishing the checkpoint. Make sure to close the instance after finishing the checkpoint.
- You can also take snapshots of the instance if you want to keep the data for later work.

Important announcements (P3.1)

- Due to an OLI bug, we request that you work carefully and try to constrain your attempts at the checkpoint to 2 as much as possible
  - Incase of case you need a third attempt, email suhailr@andrew.cmu.edu by following the template given below
    - Subject: P3.1 Attempt Cancellation Request
    - Include your Andrew ID in the email
    - Include the attempt number you wish to cancel (either 1 or 2)

Important announcements (P3.1)

- Check the P3.1 Bug Tracker on Piazza:
  https://piazza.com/class/hxs4b3o2lox5f9?cid=917

- There are additional updates and bugs including a new AMI for P3.1.

- Question 4 also has issues and will have to be cancelled. You will receive full credit for this question after manual grading is complete.
  - We encourage you to do the work and learn.

Upcoming Deadlines

- Quiz 3 Due THIS WEEK!
  - Thursday 10/9/2014 11:59PM Pittsburgh
  - Late submissions are NOT accepted
- Timed
  - **180mins** once started
  - Remember to click SUBMIT before deadline