15-319 / 15-619
Cloud Computing

Recitation 7

February 28, 2017

Overview

- **Last week's reflection**
  - Project 3.1, OLI Unit 3 Module 10, 11, 12, Quiz 5
- **This week's schedule**
  - Quiz 6 - Friday, March 3rd (Module 13)
  - Project 3.2 - Sunday, March 5th
- **Twitter Analytics: The team project is upon us**

Last Week : A Reflection

- UNIT 3: Virtualizing Resources for the Cloud
  - Module 10: Resource virtualization (memory)
  - Module 11: Resource virtualization (I/O)
  - Module 12: Case Study
  - Quiz 5

- Project 3.1
  - Files, Databases (SQL & NoSQL)
    - MySQL
    - HBase
      - Read the NoSQL Primer

Project 3.1 Feedback

Please provide feedback on P3.1

Click here

This Week: Content

UNIT 3: Virtualizing Resources for the Cloud

- Module 10: Resource virtualization (memory)
- Module 11: Resource virtualization (I/O)
- Module 12: Case Study
- **Module 13: Storage and network virtualization**
- Software Defined Data Center (SDDC)
- Software Defined Networking (SDN)
  - Device virtualization (Router and NIC virtualization)
  - Link virtualization (Bandwidth/datapath virtualization)
- Software Defined Storage (SDS)
  - IOFlow

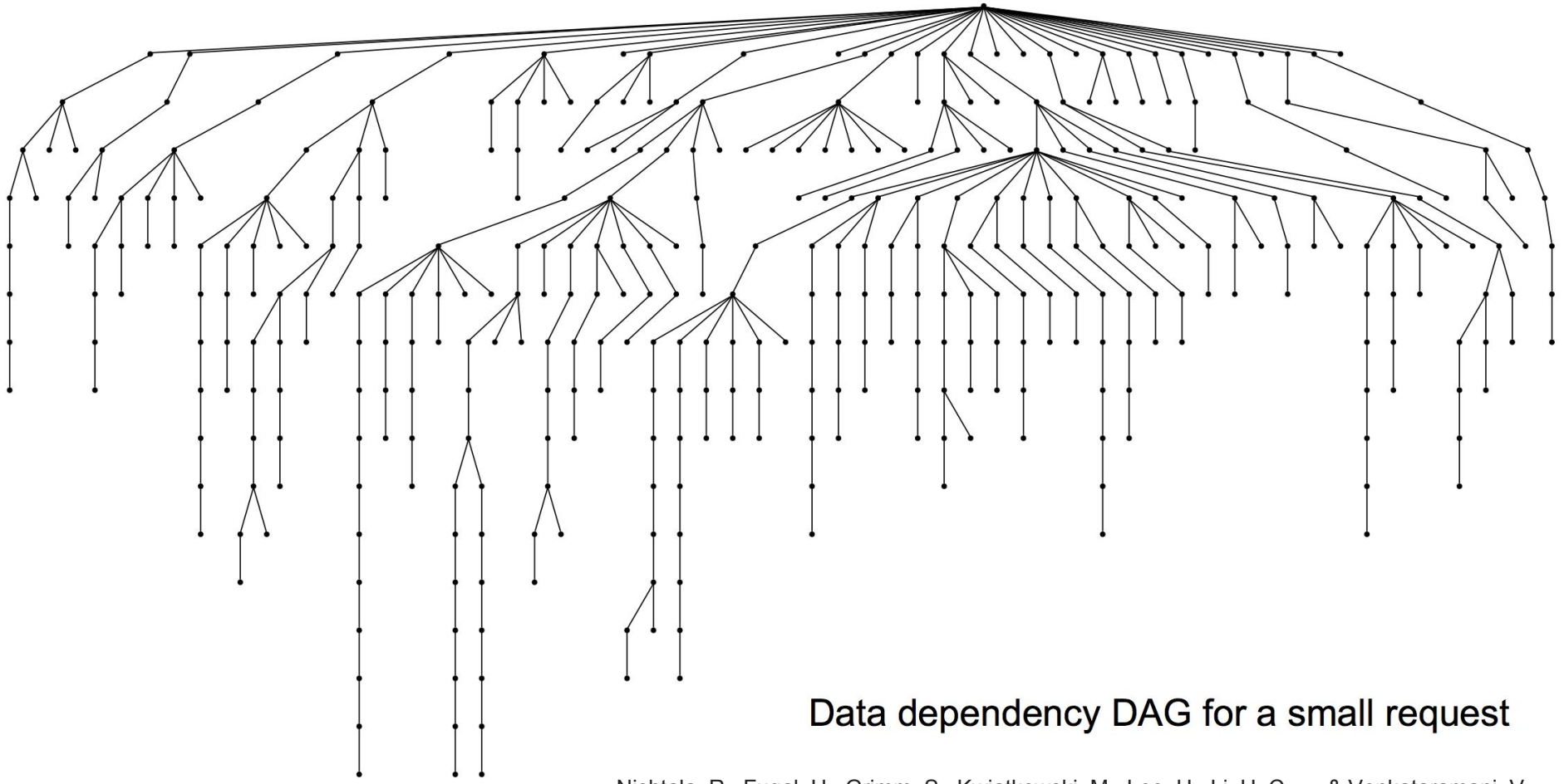
- **Quiz 6, Friday, March 3rd**

Project 3 Weekly Modules

- P3.1: Files, SQL and NoSQL
  - Primer: Storage Benchmarking
- **P3.2: Social network with heterogeneous backend storage**
- P3.3: Replication and Consistency models
  - Primer: Intro. to Java Multithreading
  - Primer: Thread-safe programming
  - Primer: Intro. to Consistency Models

High Fanout and Multiple Rounds of Data Fetching

A single Facebook page, requires many data fetch operations



Data dependency DAG for a small request

Nishtala, R., Fugal, H., Grimm, S., Kwiatkowski, M., Lee, H., Li, H. C., ... & Venkataramani, V. (2013, April). Scaling Memcache at Facebook. In *nsdi* (Vol. 13, pp. 385-398).

Distributed Databases

- In 2006, Google published details about their implementation of BigTable
- Designed as a “sparse, distributed multi-dimensional sorted map”
- HBase stores members of “column families” adjacent to each other on the file system - columnar data store

MongoDB

- Document Database
  - Schema-less model
- Scalable
  - Automatically shards data among multiple servers
  - Does load-balancing
- Complex Queries
  - MapReduce style filter and aggregations
  - Geo-spatial queries

Project 3.2

Review

Project 3.2 : Introduction

Build a social network about Reddit comments:

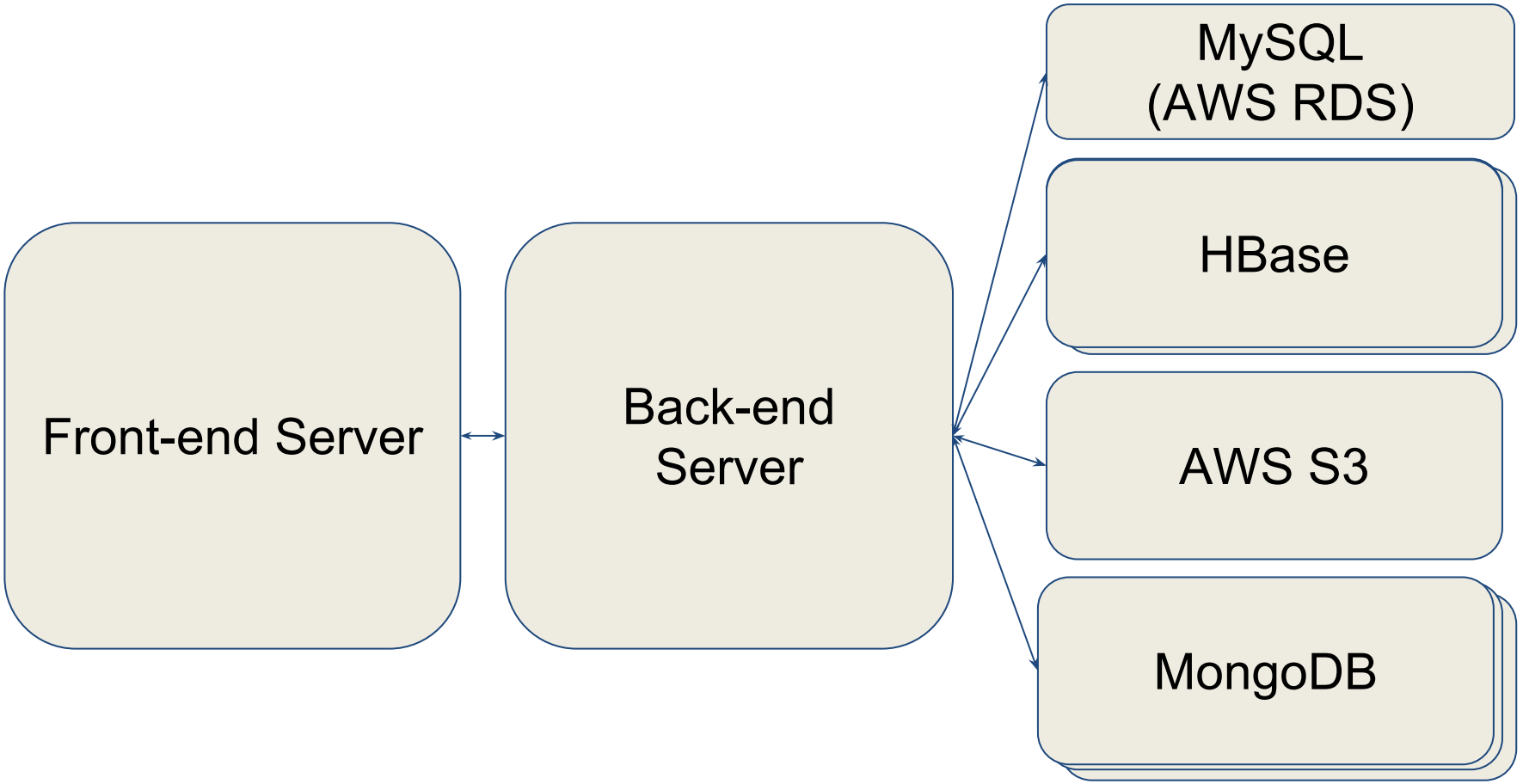
- Dataset generated from Reddit.com
  - users.csv, links.csv, posts.json
- Five Tasks
  - Task1: Basic login
  - Task2: Social graph
  - Task3: Rank user comments
  - Task4: Timeline
  - Task5: Recommendation

P3.2, Reddit Dataset

- User Profiles
  - User Authentication System (e.g., Single-Sign-On or SSO)
    - AWS RDS MySQL (users.csv)
  - User Info / Profile
    - AWS RDS MySQL
- Social Graph of the Users
  - Follower, followee, etc.
    - HBase (links.csv)
- User Activity System
  - All user generated comments
    - MongoDB (posts.json)
- Social Data Analytics System
  - Search System
  - User Behavior Analysis
  - Recommender System

Project 3.2: Architecture

- **Build a social network akin to Reddit.com:**



Project 3.2 : Tasks, Datasets & Storage

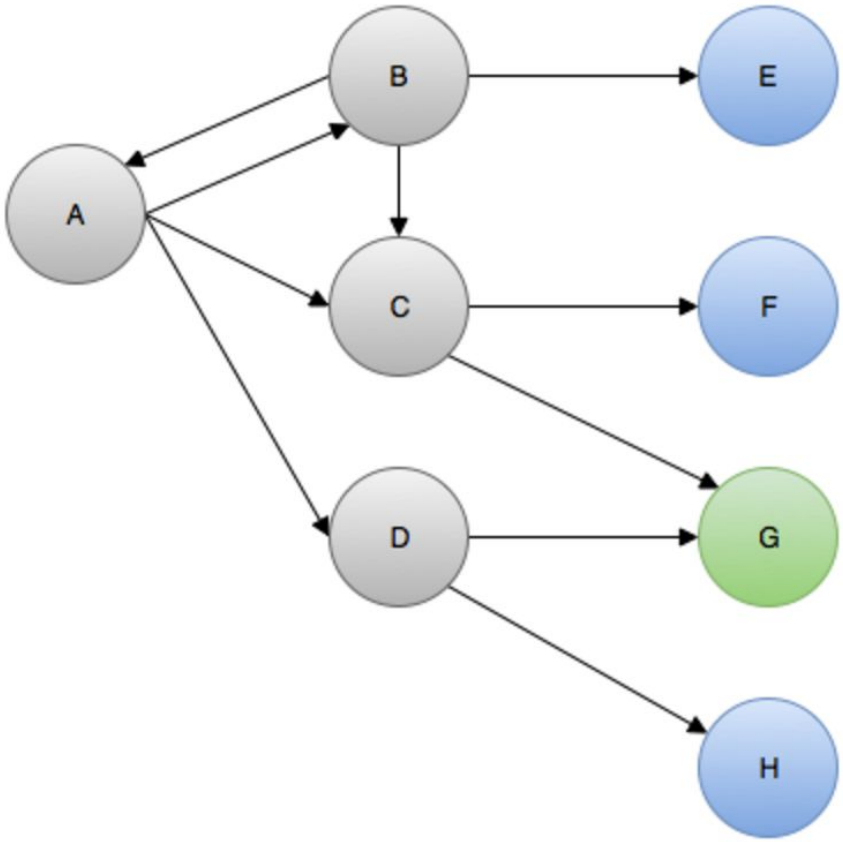
- Introduction
- The Scenario: Build Your Own Social Website
- Task1: Implementing Basic Login with MySQL on RDS
- Task2: Storing Social Graph using HBase
- Task3: Build Homepage using MongoDB
- Task4: Put Everything Together
- Task 5: Basic Recommendation
- Summary

Dataset Name	Data Store Used
Login Information	MySQL (RDS)
User Profile	MySQL (RDS)
Relation	HBase
Posts	MongoDB
Profile and Post Images	S3

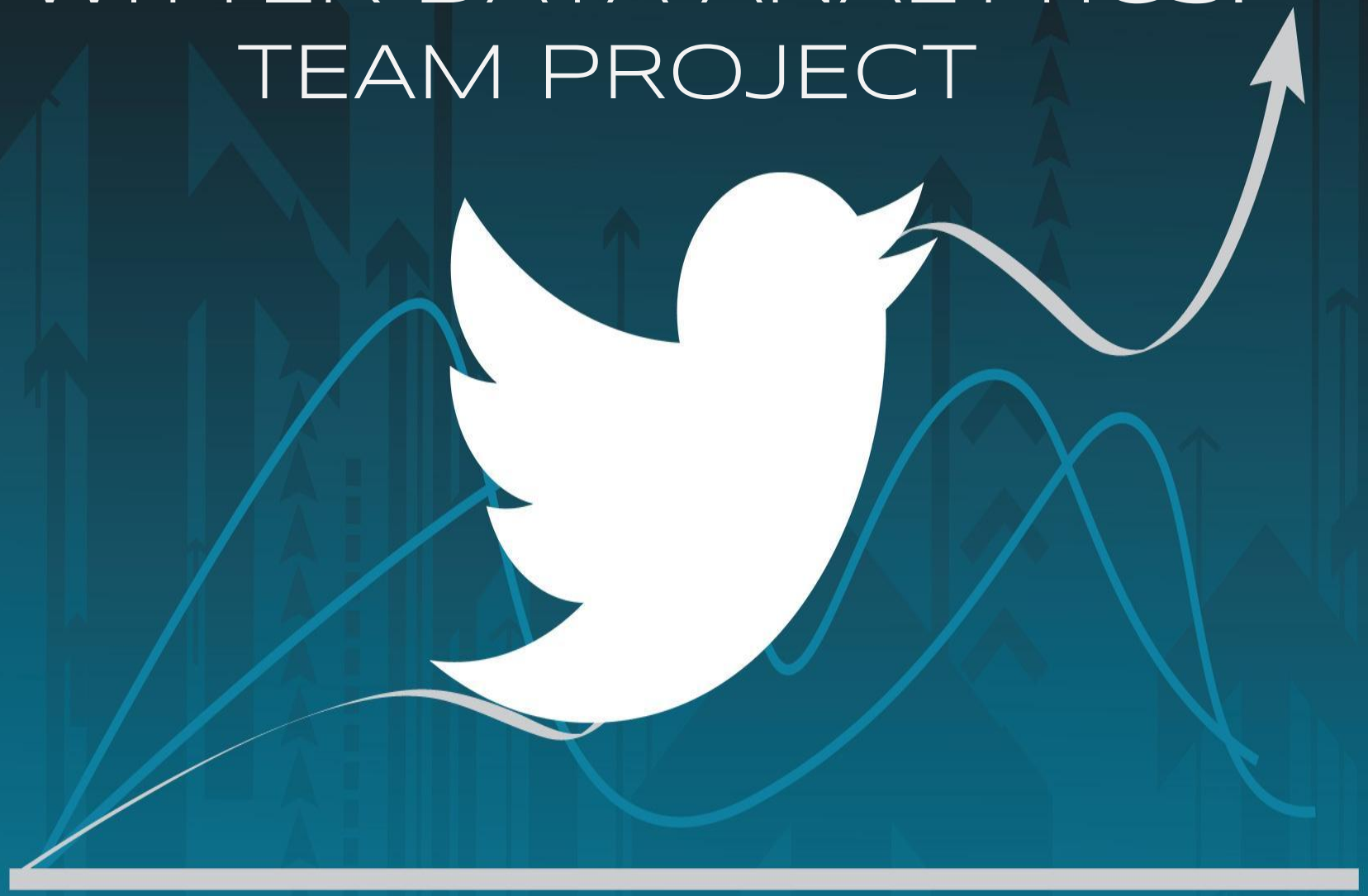
Project 3.2 : Task 5

Friend recommendation

- For a given user, recommend 10 users with distance = 2 to follow.



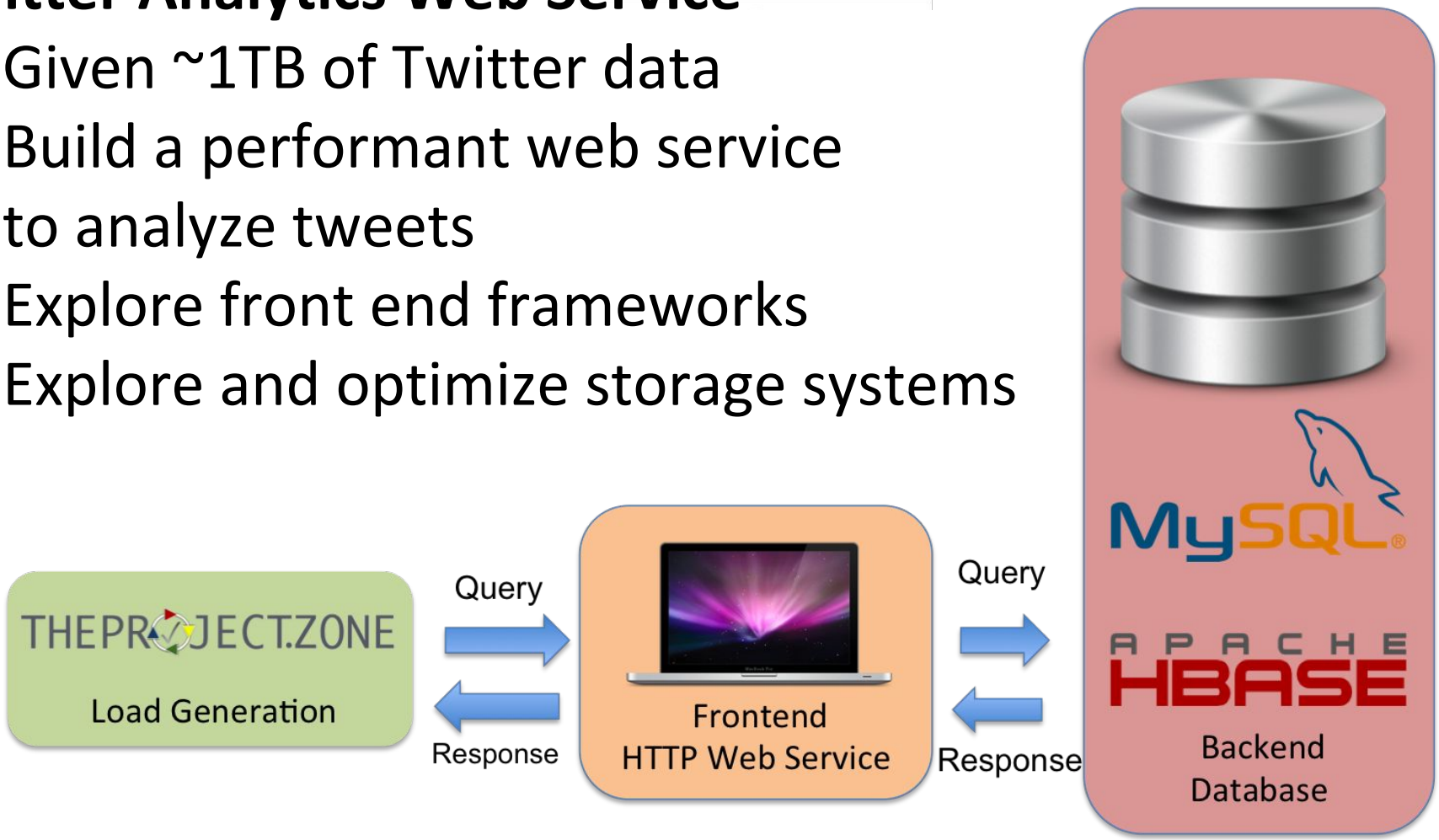
TWITTER DATA ANALYTICS: TEAM PROJECT



Team Project

Twitter Analytics Web Service

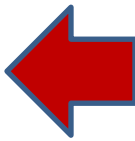
- Given ~1TB of Twitter data
- Build a performant web service to analyze tweets
- Explore front end frameworks
- Explore and optimize storage systems



Team Project

- Phase 1:
  - Q1
  - Q2 & Q3 (MySQL **AND** HBase)

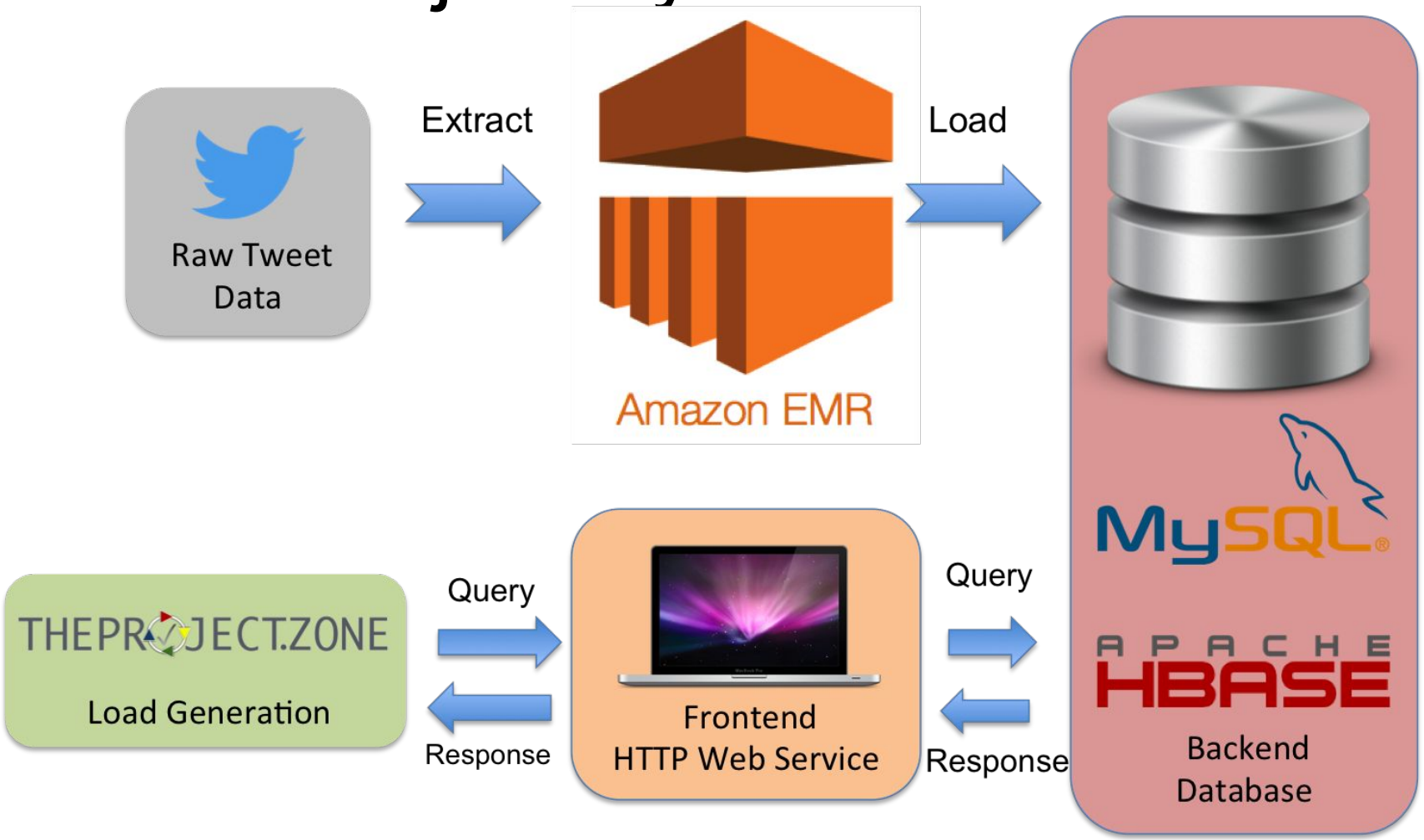
CONFIRM YOUR AWS ACCOUNT AND TEAM INFO



- Phase 2
  - Q1
  - Q2 & Q3 & Q4(MySQL **AND** HBase)

- Phase 3
  - Q1
  - Q2 & Q3 & Q4 & Q5(MySQL **OR** HBase)

Team Project System Architecture



- Web server architectures
- Dealing with large scale real world tweet data
- HBase and MySQL optimization

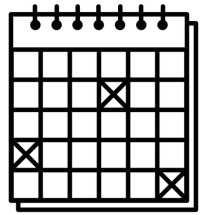
Query 1, Heartbeat

- Query 1 is a simple heartbeat & authentication query
  - Implement a heartbeat with encryption
  - You must explore different web frameworks
    - Get at least 2 different web frameworks working
    - Select the performant framework
  - In this query, there is no backend database involved

Query 2, Match Hashtag

- Query 2 is a read query
  - Find keyword matches in tweets containing a given hashtag
  - You have to perform Extract, Transform and Load (ETL)
    - Can be done on AWS, Azure or GCP
    - Dataset is available on AWS & Azure & GCP
  - Pay close attention to the ETL rules and related definitions
  - Make good use of the reference data provided
  - In this query, you will need both front end + back end

Team Project Time Table



Phase (and query due)	Start	Deadline	Code and Report Due
Phase 1 • Q1	Monday 02/27/2017 00:00:01 EST	**Sunday 03/12/2017 23:59:59 ET**	
• Q2 & Q3	Monday 02/27/2017 00:00:01 EST	Sunday 04/02/2016 23:59:59 ET	Tuesday 04/04/2016 23:59:59 ET
Phase 2 • Q1, Q2, Q3, Q4	Monday 04/03/2016 00:00:01 ET	Sunday 04/16/2016 23:59:59 ET	
Phase 2 Live Test (Hbase/MySQL) • Q1, Q2, Q3, Q4	Sunday 04/16/2016 Time TBD	Sunday 04/16/2016 Time TBD	Tuesday 04/18/2016 23:59:59 ET
Phase 3 • Q1, Q2, Q3, Q4, Q5	Monday 04/17/2016 00:00:01 ET	Sunday 04/30/2016 23:59:59 ET	
Phase 3 Live Test • Q1, Q2, Q3, Q4, Q5	Sunday 04/30/2016 Time TBD	Sunday 04/30/2016 Time TBD	Tuesday 05/02/2016 23:59:59 ET

Note:

- There will be a report due at the end of each phase, where you are expected to discuss optimizations
- **WARNING: Check your AWS instance limits on the new account (should be > 10 instances)**
- **Query 1 is due on Mar. 12, not Apr. 2nd! We want you to start early!**

Start early!

Team Project Q1 Due
Sunday 3/12

Upcoming Deadlines

- Quiz 6 - OLI (Module 13)
  Due: **Friday,** 03/03/2017 11:59PM Pittsburgh

- Project 3.2: Social Networking Timeline with Heterogeneous Backends
  Due: 03/05/2017 11:59PM Pittsburgh

- Team Project: Phase 1 - Query 1, (Next Sunday, Mar 12!)

  Due: 03/12/2017 11:59PM Pittsburgh

Q&A