

Fall 2017 - W205 – Storing and Retrieving Data
Week 6 Live Class Session Agenda
Kevin R. Crook

- Schedule
 - Lab 4 – due Tuesday, 10/10/2017 at 11:59 pm
 - Exercise 1
 - Due Tuesday 10/31/2017 at 11:59 pm
 - Lab 6 – due Tuesday, 11/7/2017 at 11:59 pm
 - Asynchronous for next week
 - Unit 6 – Data Processing and Aggregation
 - Project
 - Milestones
 - Start forming project teams
(no more than 3 students per project team)
 - Proposal Report
 - In class - 11/2/2017 or 11/7/2017
 - 10 minute presentation per project team
 - Progress Report
 - In class – 11/16/2017 or 11/28/2017
 - 10 minute presentation per project team
 - Final Presentations
 - In class – 12/14/2017 or 12/19/2017
 - All materials must be checked into GitHub repo prior to class time
 - Weeks in which we do not have a formal milestone – we will spend 3 to 5 minutes going round-robin through the project teams
- Take time in class to complete the mid-term survey for this course:
 - Section 4:
 - <https://www.surveymonkey.com/r/9ZDRZXP>
 - Section 5:
 - <https://www.surveymonkey.com/r/9ZJ9BP6>
 - Section 6:
 - <https://www.surveymonkey.com/r/9ZTSQQP>

- Today in class
 - Finish up last week's instructor led exercise
 - DAGs for the PostgreSQL queries from Lab 5
 - SQL – review the basics using PostgreSQL
 - “Soft Skills” Break Out Exercise
 - Content Delivery Networks
 - Big Data Architectures
 - Database Partitioning for Big Data (aka Scale Out SQL)

(next page)

Break Out Exercise
“Soft Skills” – Applying Theory from the Asynchronous to Real World Examples
Content Delivery Networks - CDN
Big Data Architectures
Database Partitioning for Big Data (aka Scale Out SQL)

(All Groups)

Content Delivery Networks - CDN

Give a real world example of a business with a world-wide presence using a CDN for a website. Focus on the database layer, which is often called the “last frontier” of CDN. We want to push as much of the database layer as we possibly can into the content delivery network, using:

- Schema-On-Read
- Immutability
- ELT instead of ETL
- Schema Behavior as a Contract (similar to Contract Programming)

For our real world business, list several pieces of the database layer that can be pushed out using a CDN, and estimate about what percentage of the database load would be reduced (what percentage of queries would hit parts of the database in the CDN).

(All Groups)

Big Data Architectures

Give a real world example of a business problem that would be appropriate for each of the following Big Data Architectures:

- Lambda Architecture
- Kappa Architecture
- “Netflix Architecture” aka Object Store Architecture

(All Groups)

Database Partitioning for Big Data (aka Scale Out SQL)

Give a real world example of a huge table that can benefit from both Horizontal Partitioning and Vertical Partitioning. What parts of a query would Horizontal Partitioning speed up? What parts of a query would Vertical Partitioning speed up?