Notes 16: Conclusion
Topics Covered

Foundations
Distributed database design
Query processing and optimization
Concurrency control
Reliability
Replication
Network partitions
Time and clocks
Topics Covered

Systems
Peer-to-peer systems
Publish/subscribe systems
Distributed data processing
  MapReduce
  Graph processing
  Data flow (Spark)
Distribute data stores
  NoSQL (Bigtable, HBase, and Cassandra)
  NewSQL (Spanner)
Cloud Computing

What is it?
On-demand access to shared computing resources
Infrastructure/platform/software as a service
Hardware virtualization
Business model (focus on business, not infra, differentiators)
Cloud Computing

Platforms
Amazon Web Services
Google Cloud
Microsoft Azure
Cloud Computing

Platforms

- Application layer
- Big data layer
- Computation layer
- Data store
- Virtual machine
- Distributed file system
- Data centers
- Global network
Cloud Basics

**Virtual machines**
1-32 CPUs, 1-200 GB RAM
0-4 TB local and 0-64 TB network storage
Linux or Windows

**Networking**
Virtual networks
Load balancing
Cloud Basics

Application layer
Web server frameworks
Mobile backend frameworks
Caching and search
Identity management
Monitoring and analytics

Tools
Management
Development
Security
Distributed File System

*E.g., HDFS*

Concurrency control

<table>
<thead>
<tr>
<th>Locks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timestamps</td>
</tr>
</tbody>
</table>

Reliability

<table>
<thead>
<tr>
<th>Logs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed commit</td>
</tr>
</tbody>
</table>

Replication

| Primary copy |
| Distributed consensus |

Peer-to-peer system?
Computation Layer

Custom data processing

E.g., Hadoop, Spark, Pregel
Data Store

Relational database
*E.g., MySQL instances (1-32 cores, 1-200 GB RAM, ~TB HDD/SSD)*
Distributed database design
Query processing

Non-relational store
*E.g., Bigtable (~PB data)*
Big Data Layer

Managed procedural and semi-declarative data processing
*E.g.*, Hadoop, Pig, Spark

Managed declarative processing
*E.g.*, Hive (*SQL on top of Hadoop*)

Interactive processing
*E.g.*, Jupyter (*using Python, SQL, Javascript*)

Messaging
*E.g.*, publish/subscribe system
Big Data Layer

Data toolkit
Machine learning libraries
   *E.g.*, *TensorFlow*
Recommendations
Vision
Speech recognition
Translation
Logistics

Final homework
Submit hard copy together with final exam

Extra office hours
6-8pm on Thursday, June 2 (Huang Eng Center Basement)

Final exam
12:15pm on Friday, June 3 (Huang Eng Center, Nvidia Auditorium)
Same format as midterm (open notes)

Alternate final exam
7pm on Friday, June 3 (Gates Building, Room B21/B26)
Must RSVP in private Piazza post
Feedback

Official course eval

Axess > Student > Course and Section Evaluations
Submit by 8am on Monday, June 13

Piazza

Upcoming poll(s)
Comments/suggestions?
Thank you!