Lecture 39:
• Make-Up Quiz
• Exam 3 overview
• Wrap up (parting words)

What is left:
• HW-6 due
• Projects, Exam 3, Extra Credit

Exam 3 Overview

Basics: ... Wednesday 3:30–5:30
• Closed notes, etc.
• Worth 100 points, 7 multipart questions

General Structure: Rough question layout
1. Basic ideas: warehouses/lakehouses, performance metrics, scalability, ...
2. Foundations: job scheduling (Borg), distributed file management (GFS)
3. Platforms: map reduce, spark, dataproc, data (shared) architectures
4. Query Processing: plans, operations (e.g., joins), costs
5. Data Formats: storage schemes, parquet, benefits
6. Query Engines: basic ideas, components, hive, presto, dremel
7. Dataflow and ML: big query ML, dataflow, tensorflow

Mix of question types: concepts, system properties, technical approaches, etc.
Course Overview (From Lecture 1)

Goal 1: Cover underlying ideas of “modern” data and analytics systems
- systems span entire data and processing “lifecycle”
- focus on scalability as data, processing, users, and uses increase

Goal 2: Hands-on experience using these systems
- built for and/or leverage distributed/cloud computing (for scalability)
- rapid change and growth in last several years
- which will likely continue for several more! ... one reason for Goal 1

Common Themes

Patterns
- Scalability: many machines, data partitioning (sharding), parallel exec.
- Replication & Recovery: for fault tolerance, high availability
- Coordination: centralized components to “manage” parallel “workers”
- Consistency: concurrency control to deal with replica differences
- Placement: manage communication cost and data movement

Abstraction
- Machines to Clusters to Data Centers to ...
- Resources (Borg), Storage (GFS), Processing (MapReduce, Spark, BQ)

Trends
- Declarative D.S. (functional, SQL and extensions) ... system optimizes
- Decoupled and componentized (services) ... reuse and optimize parts
Roadmap of GCP Analytics Services (From Lecture 3)

Orchestration Overlay
Cloud Composer

Real-Time Streaming Data (Messaging)
Cloud PubSub

Cloud PubSub
Cloud Dataflow

Ingestion
Cloud Storage
BigQuery

Streaming Pipelines and Analytics
Cloud Dataflow, Cloud Data Fusion

Data Warehouse
BigQuery

Data Lake Storage
Cloud Storage

Batch Pipelines and Analytics
Dataproc, Cloud Dataflow, Cloud Data Fusion

Metadata and Discovery Overlay
Cloud Catalog, Cloud Dataplex

Resource and Job Management (Borg)

Data Consumers
Looker
Looker Studio
Jupyter
VertexAI

Physical Resources
(Datacenter Campuses, Buildings, Clusters, Cells)