Big Data (DS-GA 1004) – Ultimate Finals Preparation Notes

Fully based on Week 1 Slides + Lecture Transcripts

Spring 2025

1 Introduction to Big Data

Big Data is about scaling, not just trendy tools. This course addresses scaling in data storage, computation, and communication across distributed systems.

1.1 Warm-Up Metaphor

Imagine an ant trying to carry a giant **Colocasia gigantea** (Elephant Ear) leaf: too large for one. Solution? **Team effort.** This parallels Big Data: complex tasks distributed among many units.

2 What is Big Data?

- **Scaling** challenges in storage, computation, and processing.
- **Operational Definition**: Data that does not fit comfortably on a laptop.
- **Deeper Definition**: Big Data often requires **coordinated processing across multiple computers**.

2.1 Importance of Scale

- **J.B.S. Haldane analogy**: Small mass → minor effects; Large mass → catastrophic consequences.
- More data (rows/columns) improves:
  - Statistical power
  - Parameter estimation
  - Personalized recommendations
  - Subgroup analysis

  - Leveraging high dimensionality

- But: **too much data unhandled** becomes a showstopper.

3 CS vs DS View of Data

- **DS View**: Data are immutable givens to derive insights.
- **CS View**: Data are bits/bytes needing efficient movement, storage, transformation.

4 The Five V's of Big Data (Laney, 2001; Hurwitz et al., 2013)

- **Volume**: Sheer amount of data.
- **Velocity**: Speed of incoming data.
- **Variety**: Structured vs. unstructured formats.
- **Veracity**: Uncertainty, noise, errors.
- **Value**: Potential actionable insights.

5 Why Study Big Data?

- Machine Learning and statistics perform better with more data.
- However, scaling creates new issues.
- The course covers **underlying principles** to adapt to evolving tools.

6 Class Context

- **DS-GA 1001**: Small datasets, core concepts.
- **DS-GA 1004**: Large datasets, practical scaling.
- Evolution from ideas to massive implementation.

7 Expected Outcomes

- Familiarity with distributed computing and storage.
- Understanding technical scaling challenges.
- Judging the right tool for the right scale.
- Tools: Git, SQL, Hadoop, MapReduce, HDFS, Spark, Dremel, Parquet, Dask, CUDA, etc.

8 How the Class Works

- Platform: Brightspace.
- Homework via GitHub Classroom.
- Weekly assigned readings.
- Grading: 25% HW, 25% Capstone, 25% Final (T/F), 9% Quizzes, 16% Low-stakes work.

9 Key Concepts in Resource Management

9.1 Storage

- Storage costs have dropped exponentially (e.g., 540TB costs $11,292 today).
- Storage rarely the main bottleneck anymore.

9.2 Communication

- **Latency critical**:
  - L1 Cache: 1 ns
  - L2 Cache: 4 ns
  - RAM: 100 ns
  - SSD: 16,000 ns
  - HDD: 2,000,000 ns
- **Main memory access often bottleneck**. Minimize data movement.

9.3 Computation

- **Moore's Law slowing down**. CPU clock speeds plateau.
- **Parallelism** (multi-core, GPUs) is key to future computation.

10 Communication is the Hidden Enemy

- Communication cost grows **super-linearly**.
- **Brooks' Law**: Adding manpower to a late project delays it further.
- Coordination becomes a primary cost in scaling complex systems.

11 Principles of Scaling

11.1 Tasks Easy to Parallelize

- Example: Moving stones for Pyramids (nearly linear scaling).

11.2 Tasks Hard to Parallelize

- Example: Building Cathedrals, software.
- **Communication dominates**.
  "Data Cathedral" metaphor: Big Data requires disciplined coordination.

11.3 Lecture Emphasis

- **Carrying the Leaf**: Distributed effort needed for massive tasks.
- **Brutality of Parallelization**: Expect high effort from students and instructors.
- **Pyramids vs. Operating Systems**: Simple vs. complex parallelization.

12 Big Data Strategy

- Distribute storage and processing.
- Minimize communication overhead.
- Introduce hierarchies where necessary.

13 Critical Quotes

- *"We shouldn't be trying for bigger computers, but for more systems of computers."* – Rear Admiral Grace Hopper
- *"Adding manpower to a late project makes it later."* – Brooks' Law

14 Next Steps

- **Read**: Garcia-Molina, Ullman, & Widom, 2009, Chapter 2.
- Topics: Centralized Systems, File Systems, Relational Databases.