Introduction to Big Data

The 5 V's and Why It Matters

Professor Anis Koubaa

SE 446
Alfaisal University
https://github.com/aniskoubaa/big_data_course

Spring 2026

جامعة الفيصل

Outline

Motivating Question

How much data is generated every minute?

Every 60 seconds:

- 500 hours of YouTube video uploaded
- 6 million Google searches
- 500,000 tweets posted
- 200 million emails sent

The Challenge:

- Too large for one machine
- Too fast for batch processing
- Too complex for simple queries
- Traditional DBs can't cope

Welcome to the Big Data Era

We need new tools and techniques to handle this scale!

What is Big Data?

Definition

Big Data refers to datasets that are too large, fast, or complex for traditional data processing tools.

- Cannot fit on a single machine
- Cannot be processed in reasonable time
- Requires **distributed computing**

Key Insight

It's not just about *size* — it's about the *challenges* of handling the data.

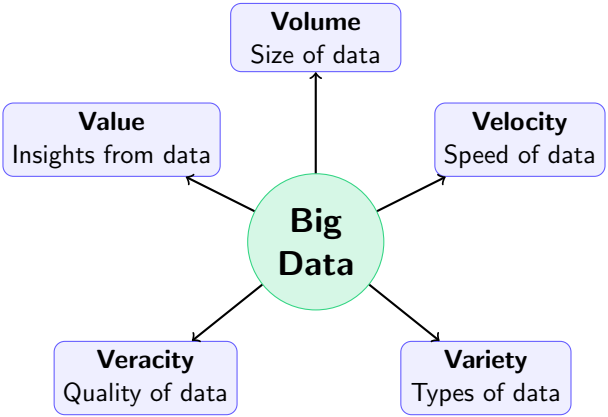
The Scale of Big Data

Company	Data Generated	Scale
Facebook	4 PB / day	250 billion photos
YouTube	500 hours video / minute	1 billion hours watched/day
Twitter	500 million tweets / day	6,000 tweets / second
Google	20 PB processed / day	3.5 billion searches / day

Perspective

1 Petabyte = 1,000 Terabytes = 1,000,000 Gigabytes

The 5 V's of Big Data



Volume: Size of Data

Challenge

Datasets too large to store or process on a single machine.

Examples:

- Genomic data: 100 GB per genome
- CERN: 1 PB / month
- Autonomous cars: 4 TB / day

Solution

Distributed Storage

- HDFS (Hadoop)
- Amazon S3
- Google Cloud Storage

Velocity: Speed of Data

Challenge

Data arrives too fast for batch processing.

Examples:

- Stock market: millions/second
- IoT sensors: continuous stream
- Social media: real-time feeds

Solution

Stream Processing

- Apache Kafka
- Spark Streaming
- Apache Flink

Variety: Types of Data

Type	Description	Examples
Structured	Rows & columns, fixed schema	SQL tables, Excel
Semi-structured	Flexible schema, self-describing	JSON, XML, logs
Unstructured	No predefined format	Images, videos, emails

Key Insight

80% of enterprise data is **unstructured**!

Veracity & Value

Veracity: Data Quality

- Missing values
- Inconsistent formats
- Noise and outliers
- Fake data (bots, spam)

"Garbage in, garbage out"

Value: Extracting Insights

- Predictive analytics
- Customer segmentation
- Fraud detection
- Recommendation engines

"The goal of Big Data"

Limitations of RDBMS

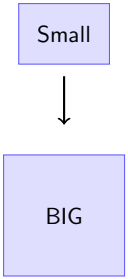
Challenge	RDBMS	Big Data
Scaling	Vertical (bigger server)	Horizontal (add nodes)
Schema	Fixed, predefined	Flexible, schema-on-read
Data Types	Structured only	All types
Cost	Expensive hardware	Commodity hardware
Speed	Slow for massive writes	Parallel distributed writes

The Solution

Distributed Systems: Hadoop, Spark, NoSQL databases

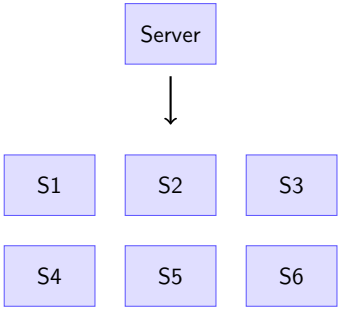
Vertical vs. Horizontal Scaling

Vertical Scaling



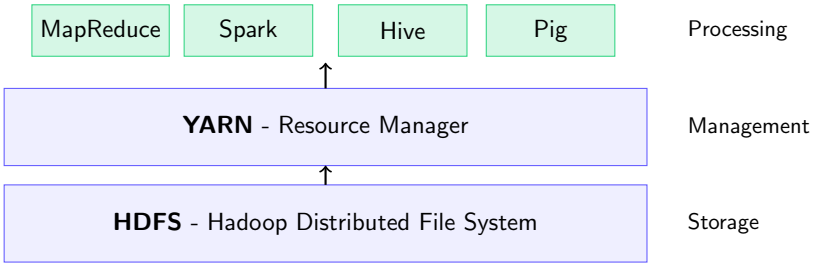
Expensive, limited

Horizontal Scaling



Scalable, cost-effective

The Hadoop Ecosystem



This Course Covers

HDFS, MapReduce, Hive, Spark, Kafka (Streaming)

Summary: Key Takeaways

1. **Big Data** = Volume + Velocity + Variety + Veracity + Value
2. **Three data types**: Structured, Semi-structured, Unstructured
3. **RDBMS limitations** solved by distributed systems
4. **Hadoop Ecosystem**: HDFS (storage), YARN (resources), Spark/Hive (processing)

Next Session

HDFS Architecture: NameNode, DataNode, Replication

Homework

1. Watch the pre-class video for Session 2B:
   - "HDFS Tutorial" - Edureka (20 min)
2. Setup your accounts (if you haven't):
   - Google Colab: `colab.google.com`
   - GitHub: `github.com`
3. Review the notebook from today's session

Questions?

Prof. Anis Koubaa
akoubaa@alfaisal.edu