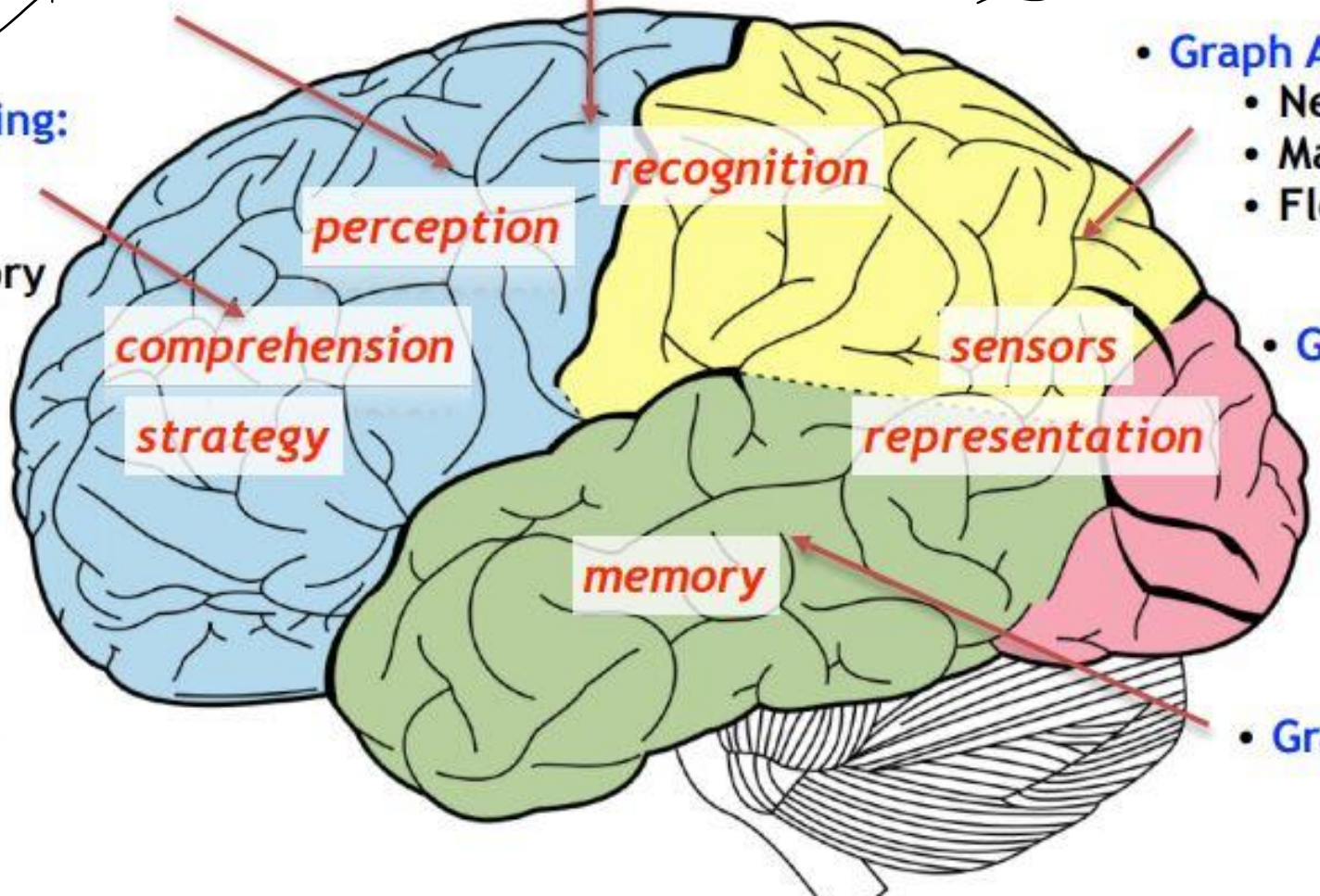
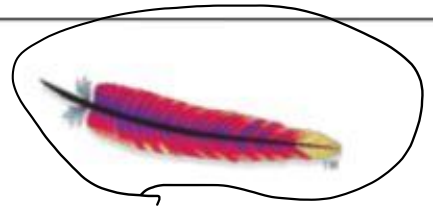
Human brain is a graph/network of 100B nodes and 700T edges.

- **Machine Cognition:**
  - Robot Cognition Tools
  - Feeling
- **Machine Reasoning:**
  - Bayesian Networks
  - Game Theory Tools
- **Machine Learning:**
  - Machine Learning Tools
  - Deep Learning Tools
- **Graph Analytics:**
  - Network Analysis
  - Matching and Search
  - Flow Prediction
- **Graph Visualization:**
  - Dynamic Graph
  - Big Graph
- **Graph Database:**
  - Large-Scale Native Store



Course Main Thrust 1: Apache Hadoop and Big Data



The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

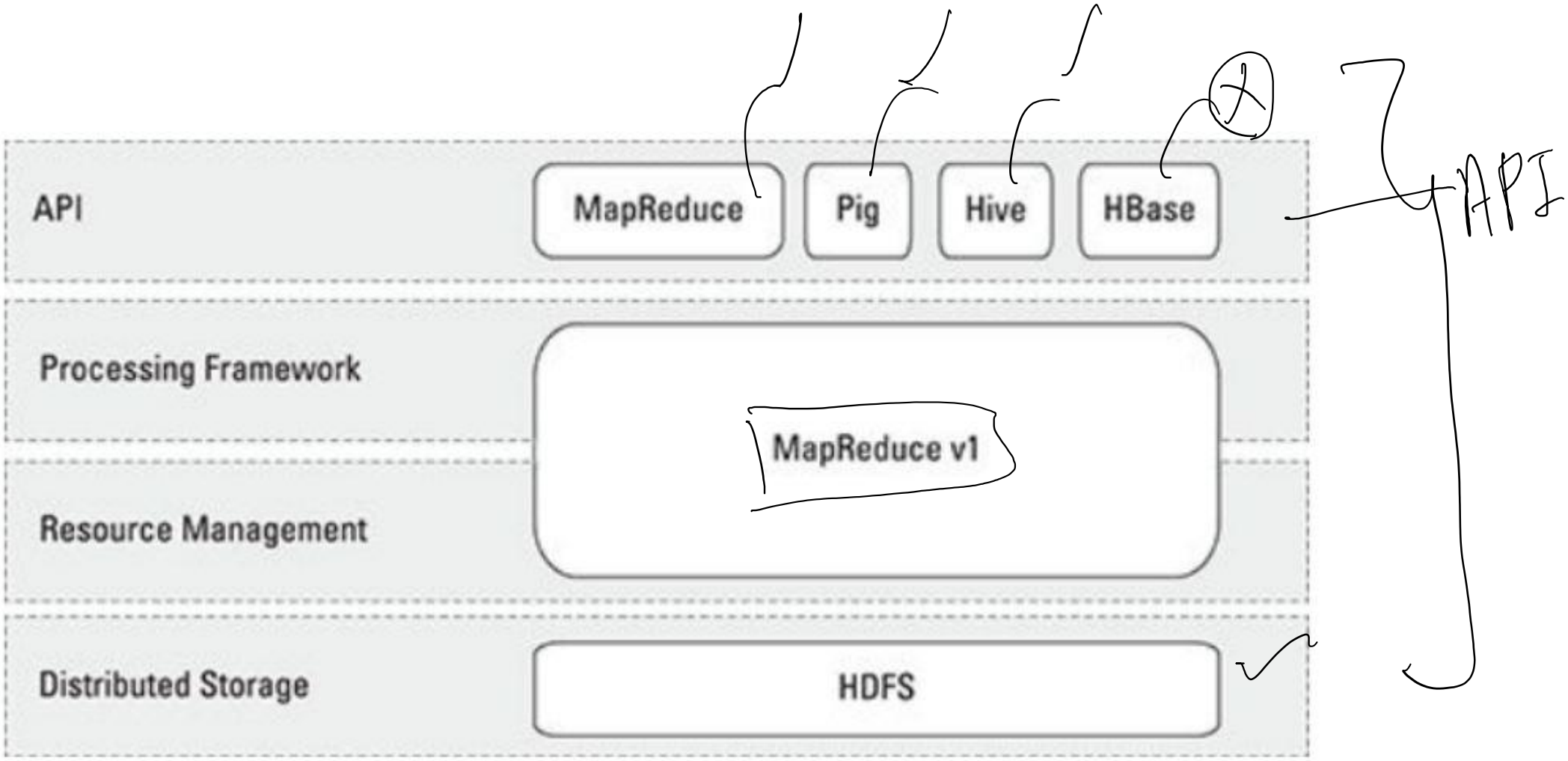
The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

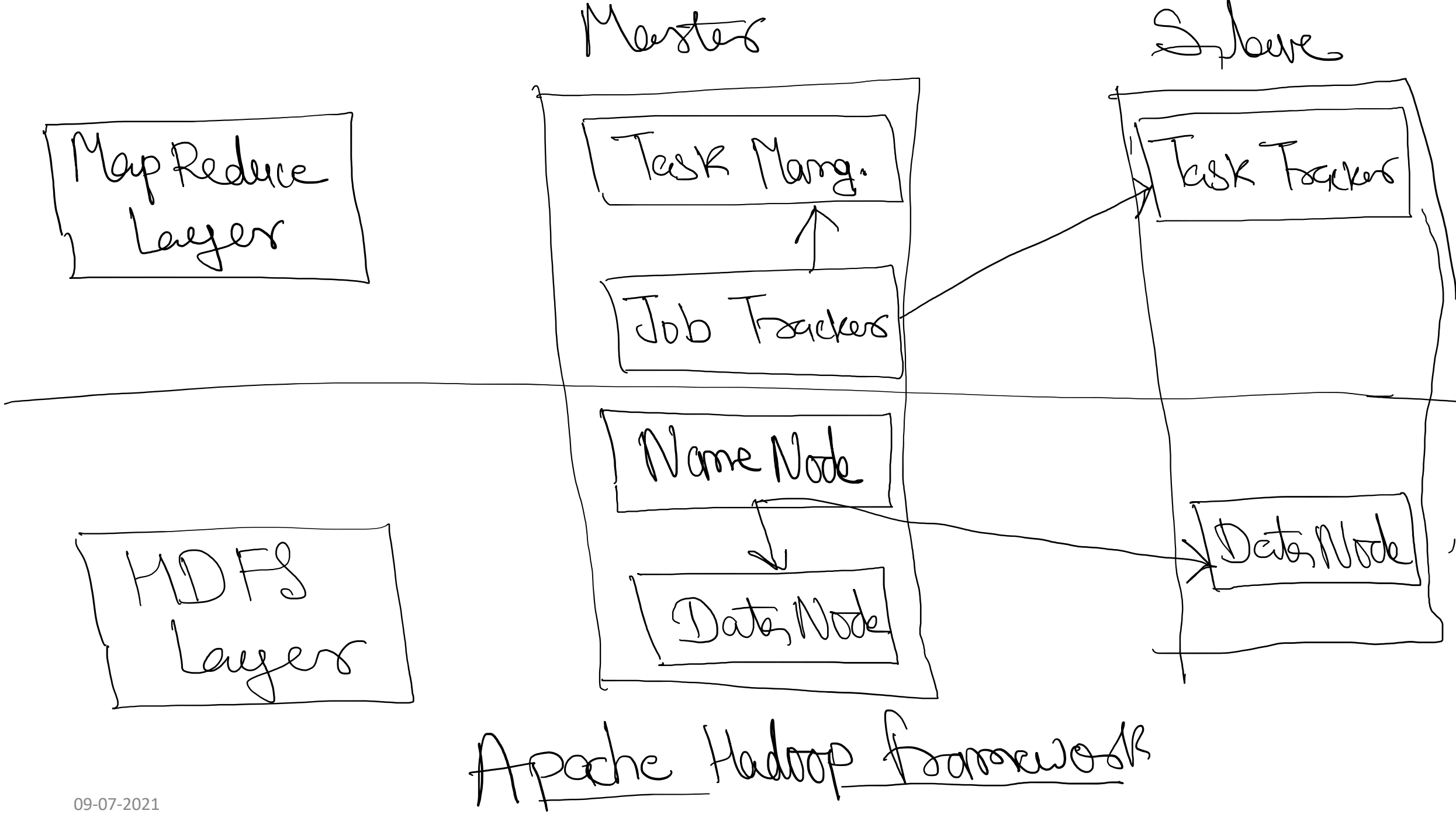
The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

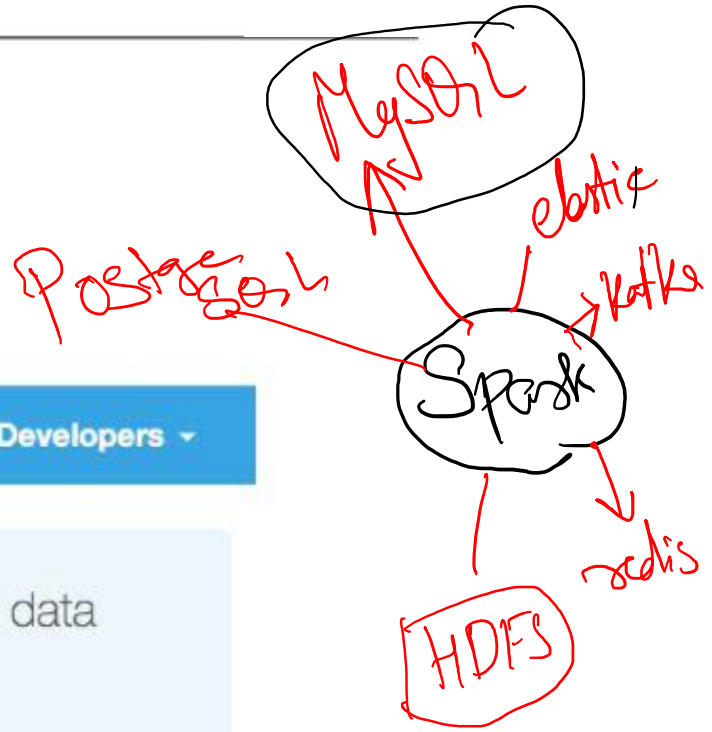
http://hadoop.apache.org

Four distinctive layers of Hadoop




Master
Slave
Map Reduce Layer
Task Mang.
Task Tracker
Job Tracker
Name Node
Data Node
Data Node
HDFS Layer
Apache Hadoop framework

Course Main Thrust 2: Apache Spark and ML



Lightning-fast unified analytics engine

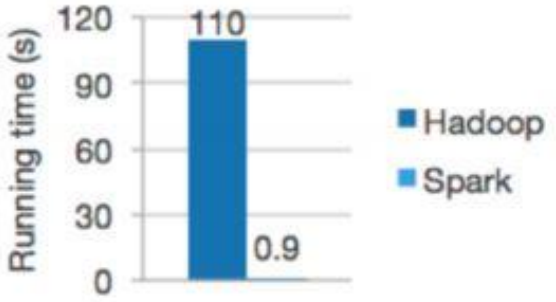
Download | Libraries | Documentation | Examples | Community | Developers

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

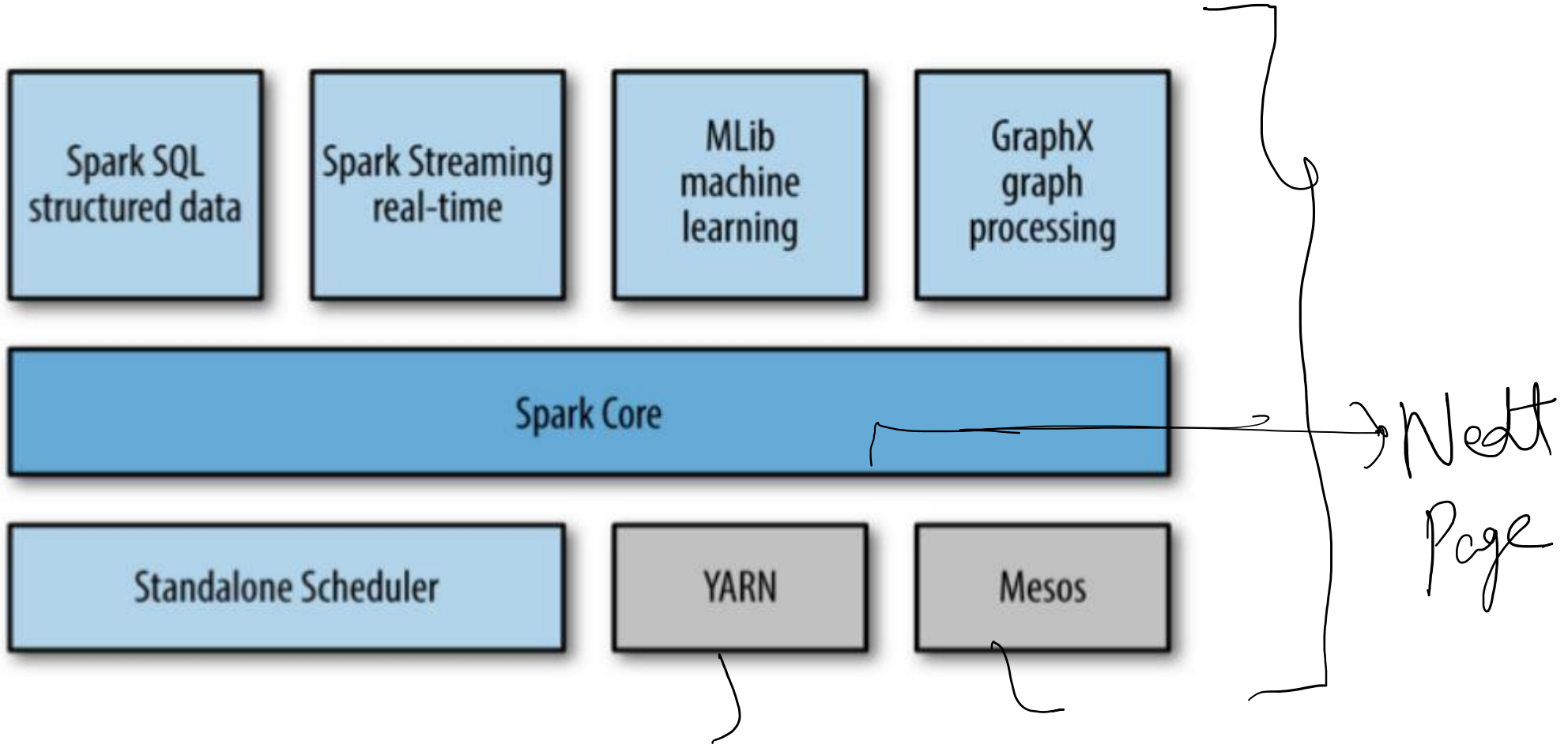
Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Logistic regression in Hadoop and Spark

Main Spark Stack



Spark Core API (Speed, Ease of Use, Engine)

R | SQL | Python | Scala | Java — Add-ON

Spark SQL + Data Frames

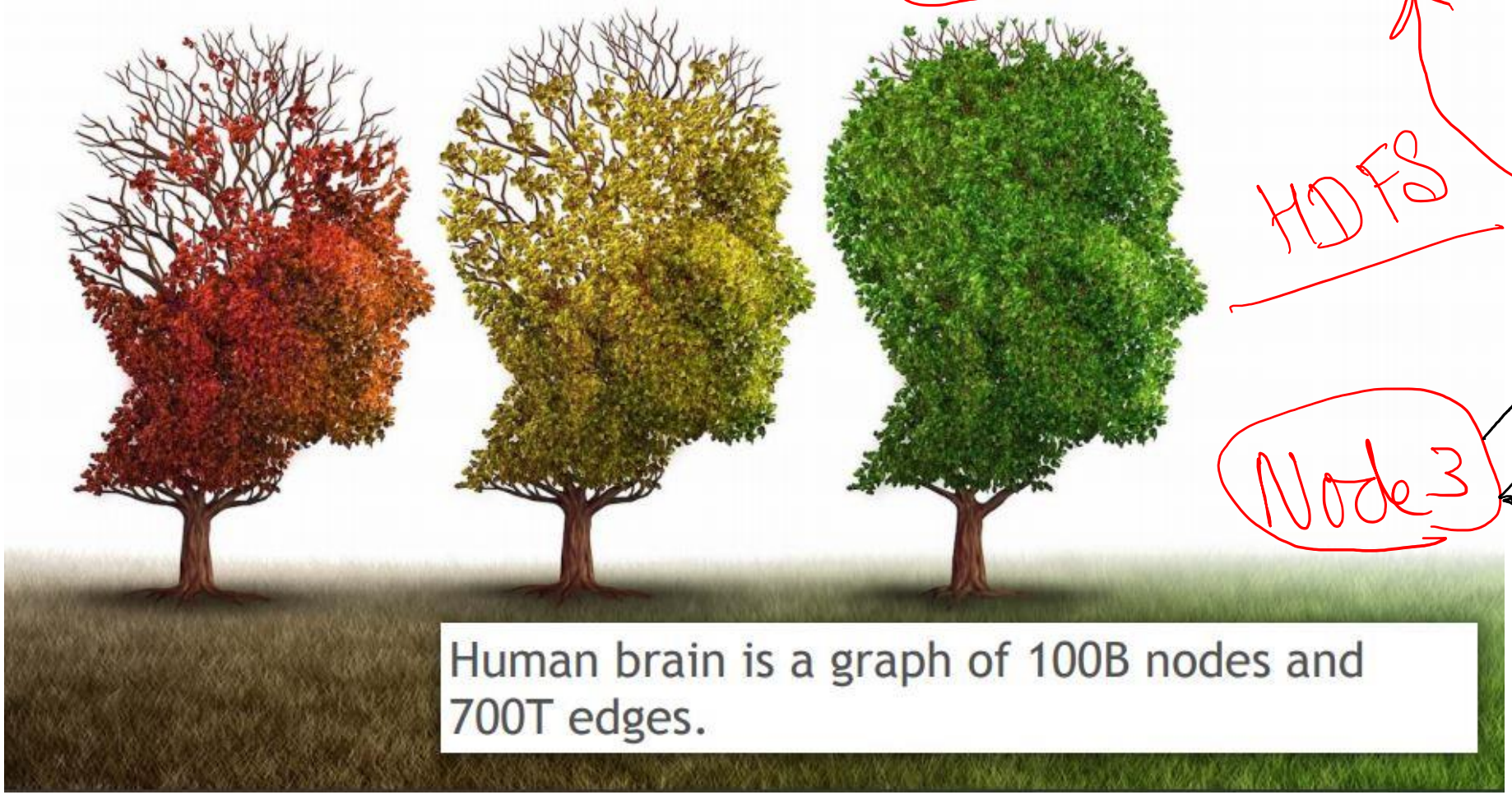
Streaming

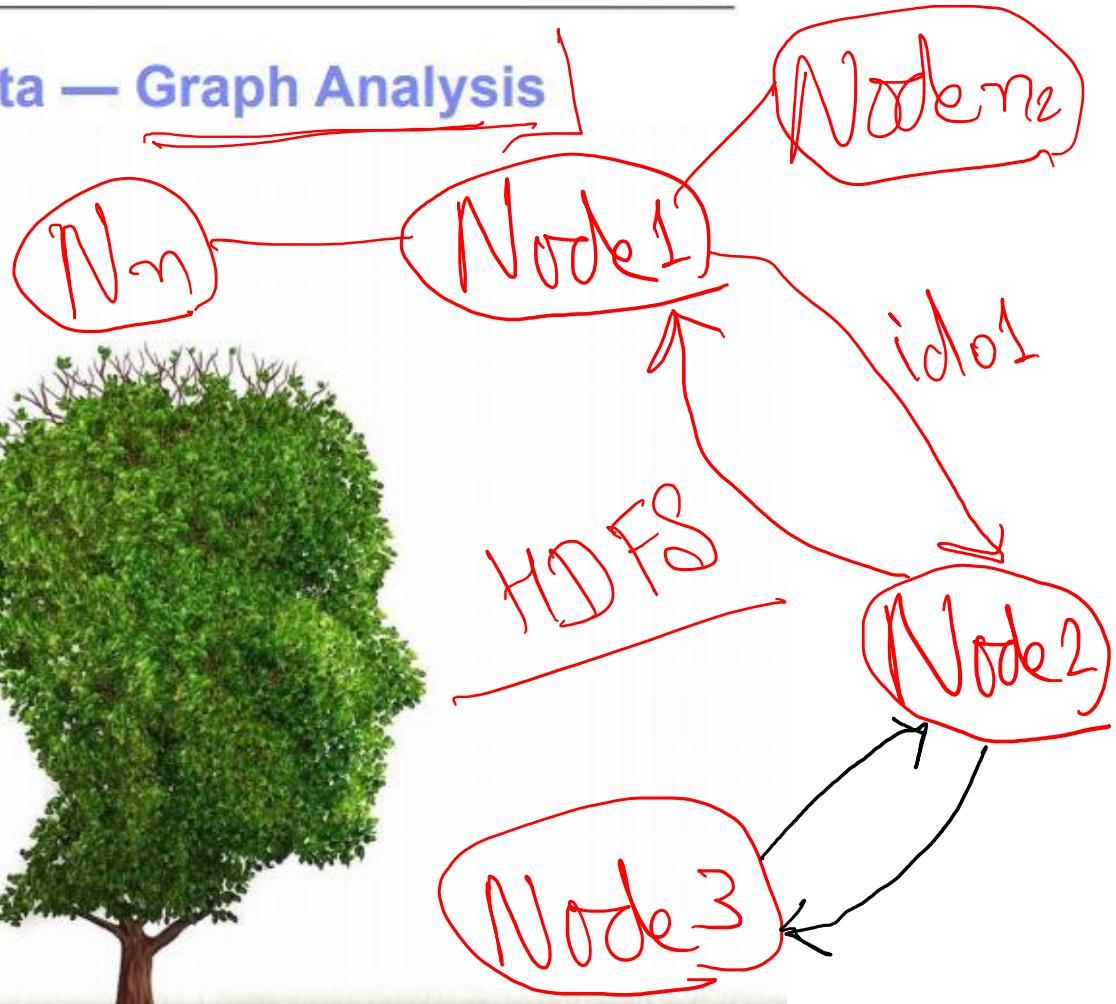
ML Lib.

graph API

Ecosystem - Apache Spark

Course Main Thrust 3: Linked Big Data — Graph Analysis



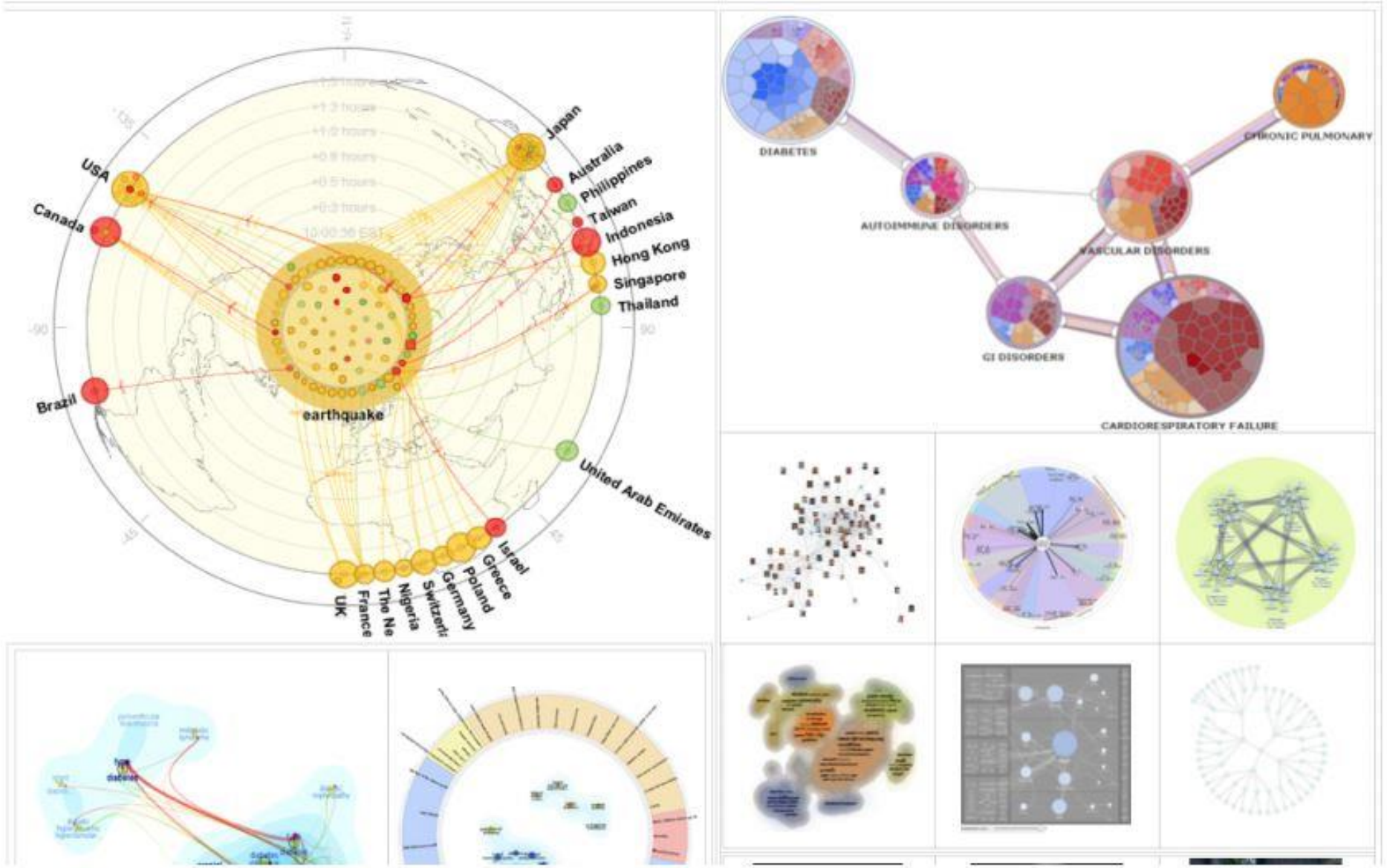


Course Main Thrust 4: Streaming Big Data Analytics

Data → DBMS → Hadoop

Process Data → Insights → DBMS → Reports

Course Main Thrust 5: Big Data Visualization



Course Main Thrust 6: Big Data System and AI Solutions

- Big Data Pipeline
- Big Data and AI for Finance
- Big Data and AI for Healthcare

Why you want to take this class — MooC / YouTube / edx / MiT

- **Key Differentiator of this class:** Focusing on building a full-spectrum understanding of the latest Big Data Analytics technologies and using them to build real industry real-world solutions.

- **Sapphire Big Data Analytics Open Source Applications:** Create a Big Data open source toolsets for various industries (and disciplines)

- **Dataset and Use Cases:** Welcome!! — DataSets