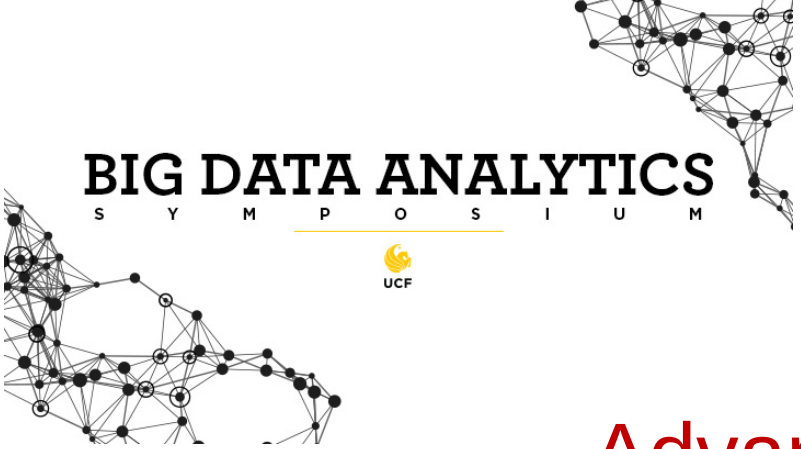
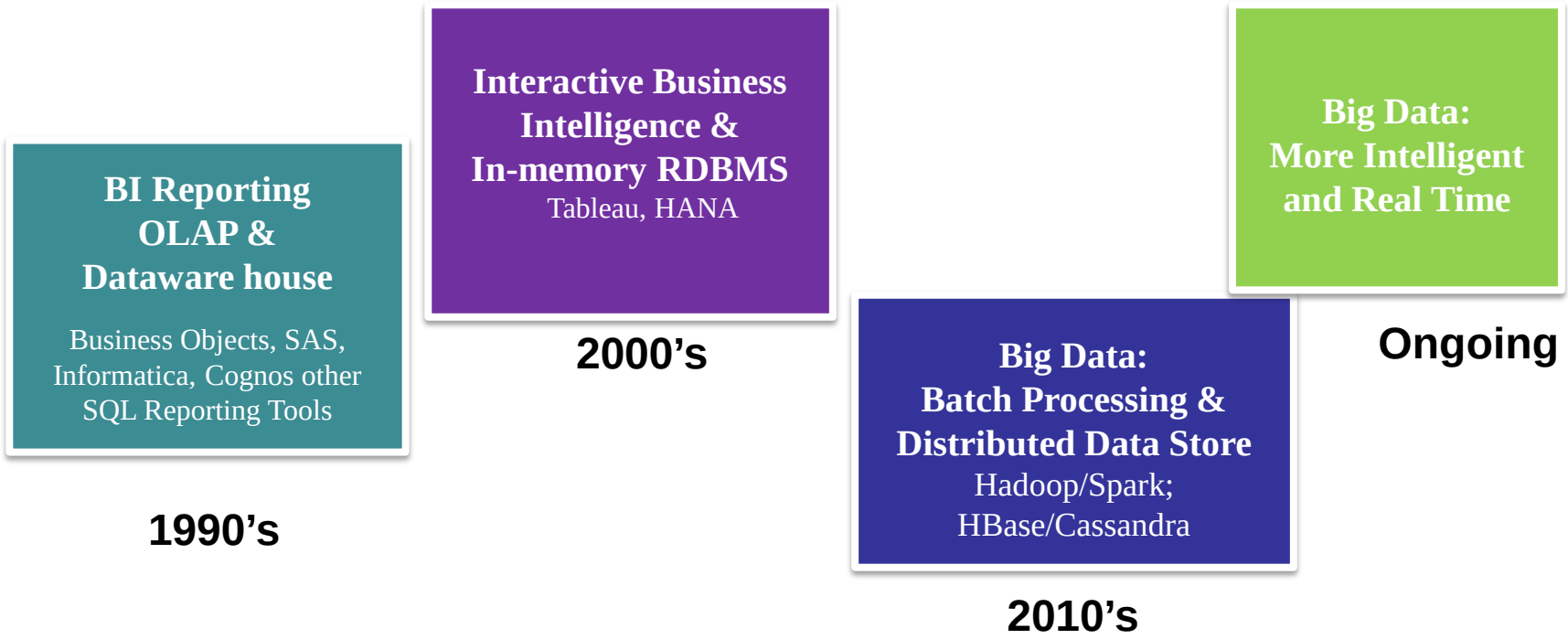


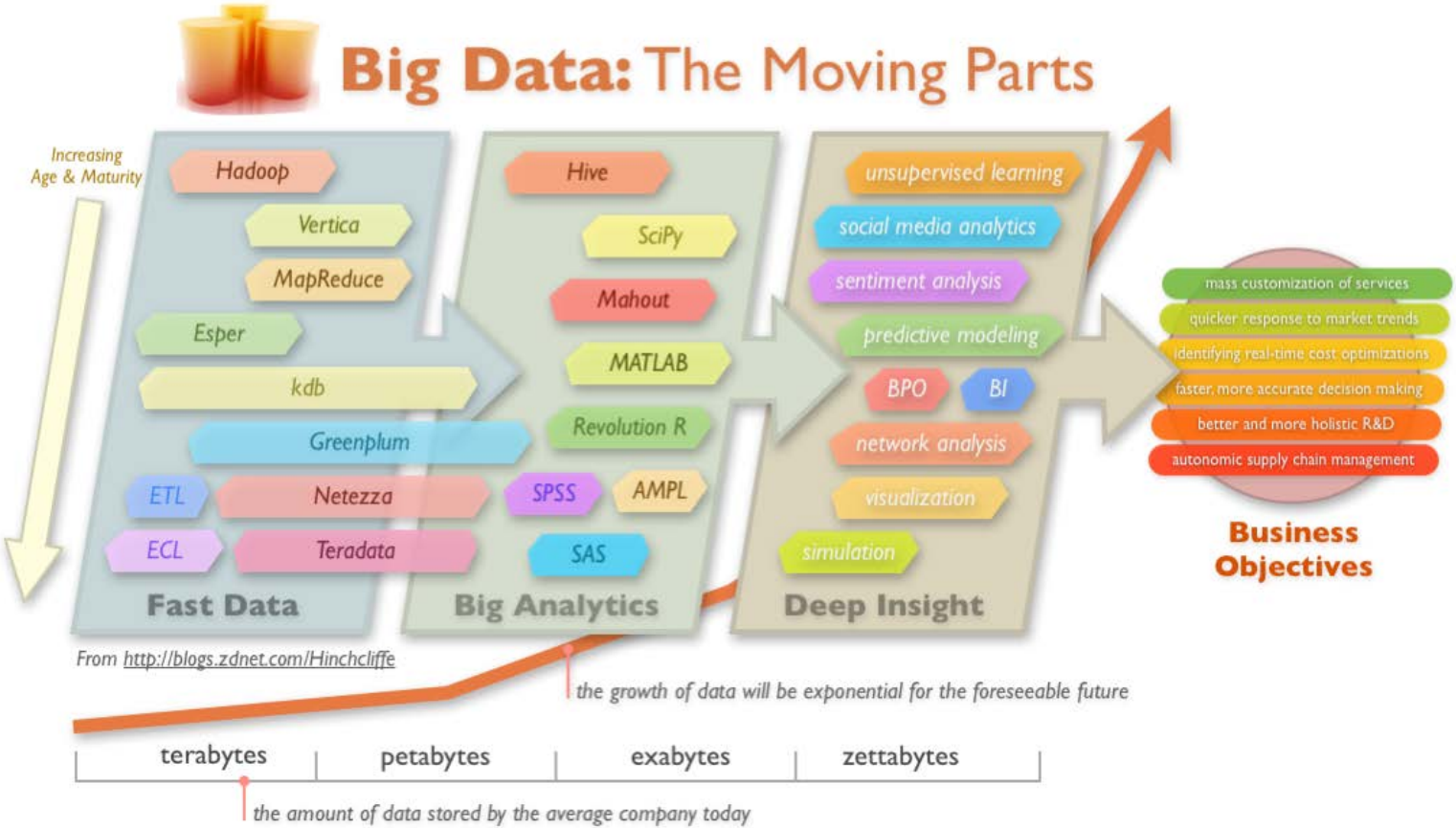
Advances and Challenges of Big Data Computing Platforms

Liqiang Wang
Associate Professor
Department of Computer Science
UCF

UCF

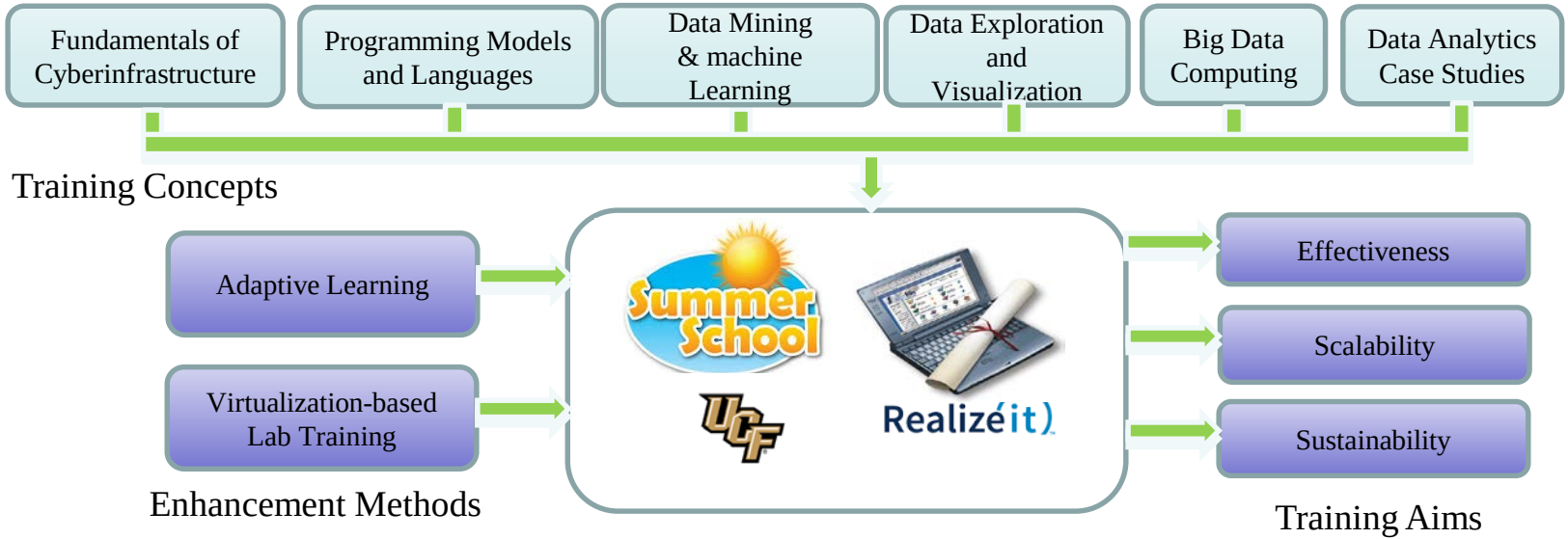
THE EVOLUTION OF BUSINESS INTELLIGENCE





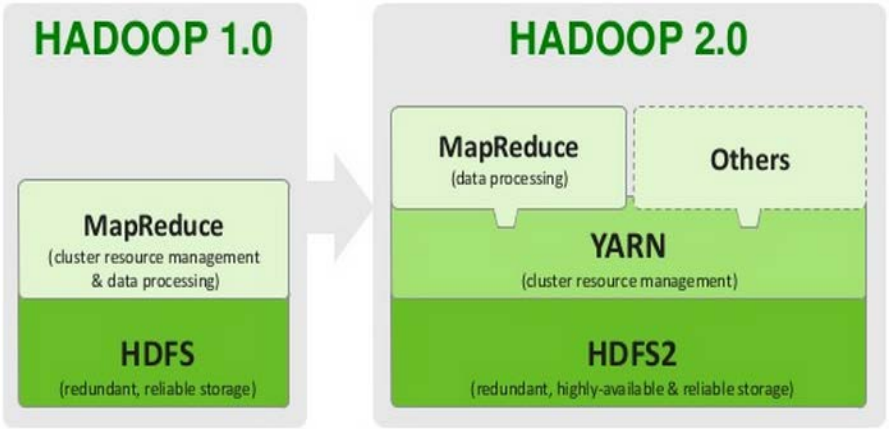
Source: Dion Hinchcliffe, "The enterprise opportunity of Big Data: Closing the 'clue gap,'"

Essential Training at UCF (Pending)



Hadoop Architecture

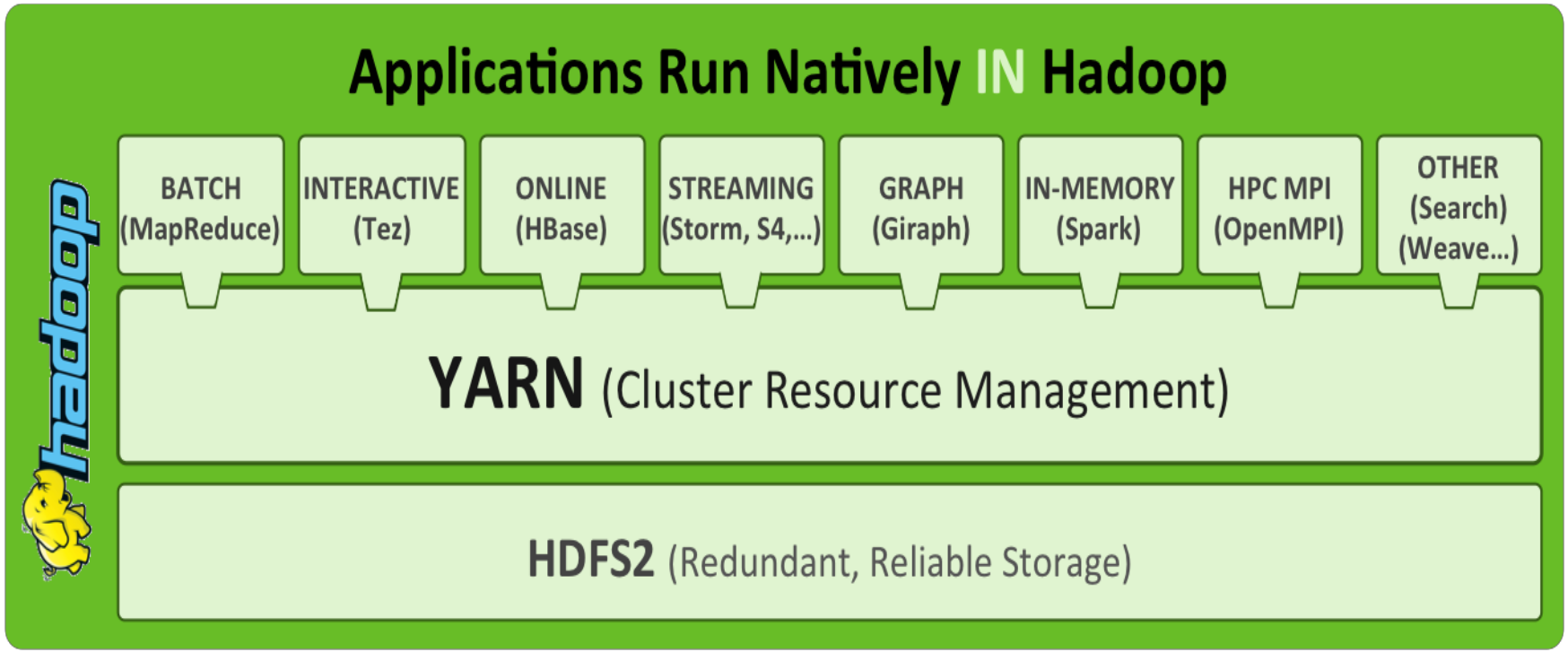
- Hadoop consists of
  - Hadoop 1.0: HDFS and MapReduce
  - Hadoop 2.0: HDFS, Yarn, and MapReduce

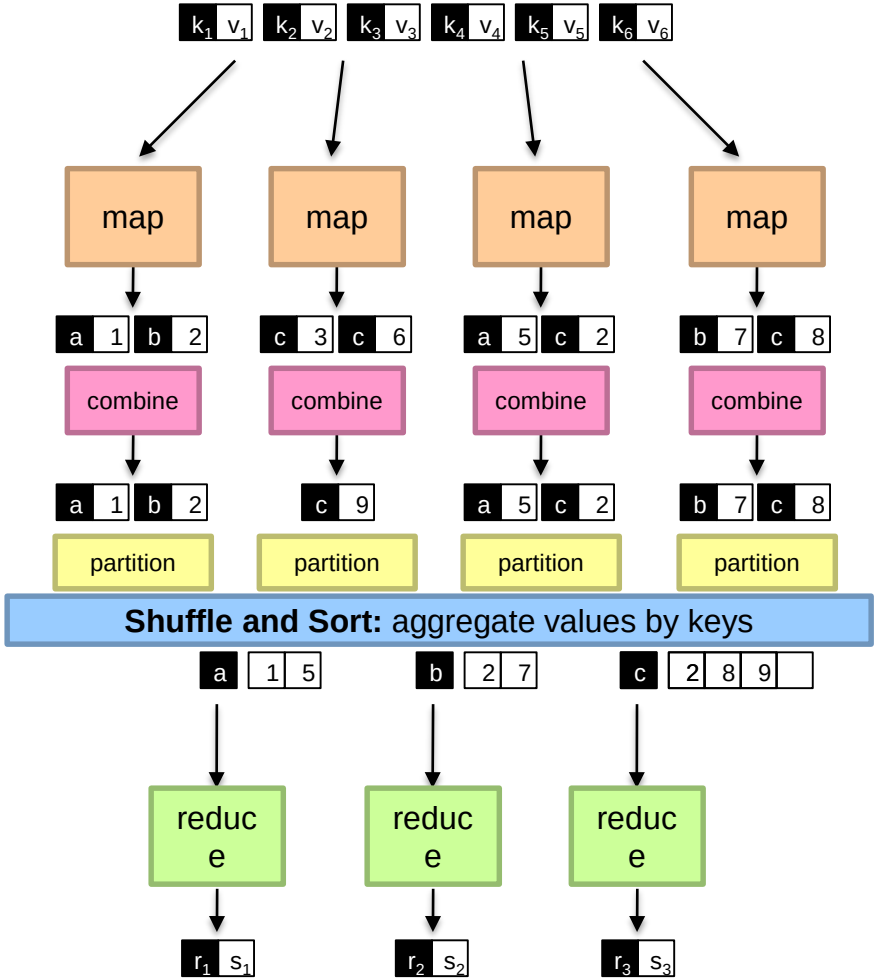


Hadoop 1 vs 2

	Hadoop1	Hadoop 2
Components	HDFS, MapReduce	HDFS, Yarn, MapReduce, other module
Scalability	Less	More
Name Node	Single	Multiple
Resource Management	Slot	Container
Job Type	MapReduce	MapReduce, MPI, Spark
Reliability	Worse	Better
JVM re-use	Yes	No

Yarn & HDFS


Applications Run Natively IN Hadoop
BATCH (MapReduce)
INTERACTIVE (Tez)
ONLINE (HBase)
STREAMING (Storm, S4,...)
GRAPH (Giraph)
IN-MEMORY (Spark)
HPC MPI (OpenMPI)
OTHER (Search) (Weave...)
YARN (Cluster Resource Management)
HDFS2 (Redundant, Reliable Storage)
hadoop

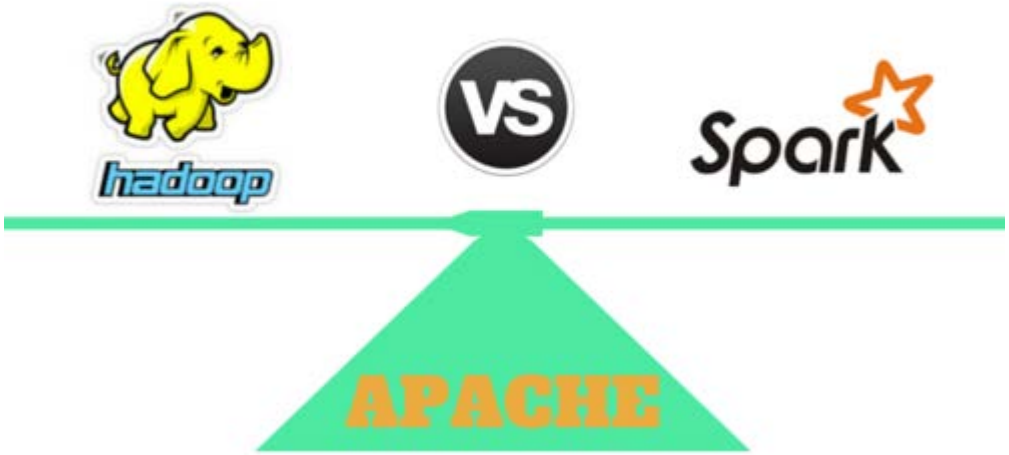

k1 v1
k2 v2
k3 v3
k4 v4
k5 v5
k6 v6
map
map
map
map
a 1 b 2
c 3 c 6
a 5 c 2
b 7 c 8
combine
combine
combine
combine
a 1 b 2
c 9
a 5 c 2
b 7 c 8
partition
partition
partition
partition
Shuffle and Sort: aggregate values by keys
a 1 5
b 2 7
c 2 8 9
reduce
reduce
reduce
r1 s1
r2 s2
r3 s3

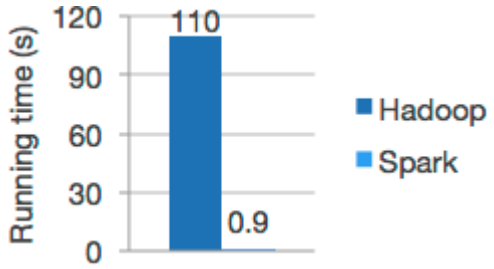
Why Use MapReduce Instead of Classical Supercomputing?

Comparison

	MPI	Hadoop/Spark
Node Communication	Supports more frequent node communication (tightly coupled)	Usually nodes do not communicate directly (loosely coupled)
Disk I/O	Usually load data once	Every nodes read/write its own data
Fault tolerance	No	Yes
Auto-Scaling	No	Yes
Applications	CPU-Intensive Scientific Computing	Data-Intensive Analytics
Challenging Research Issues	➢ Scalability ➢ Resilience (including checkpointing) ➢ Energy-efficiency	➢ Performance Tuning ➢ Integration with Edge Computing & IoT

Hadoop is Slow in Machine Learning!



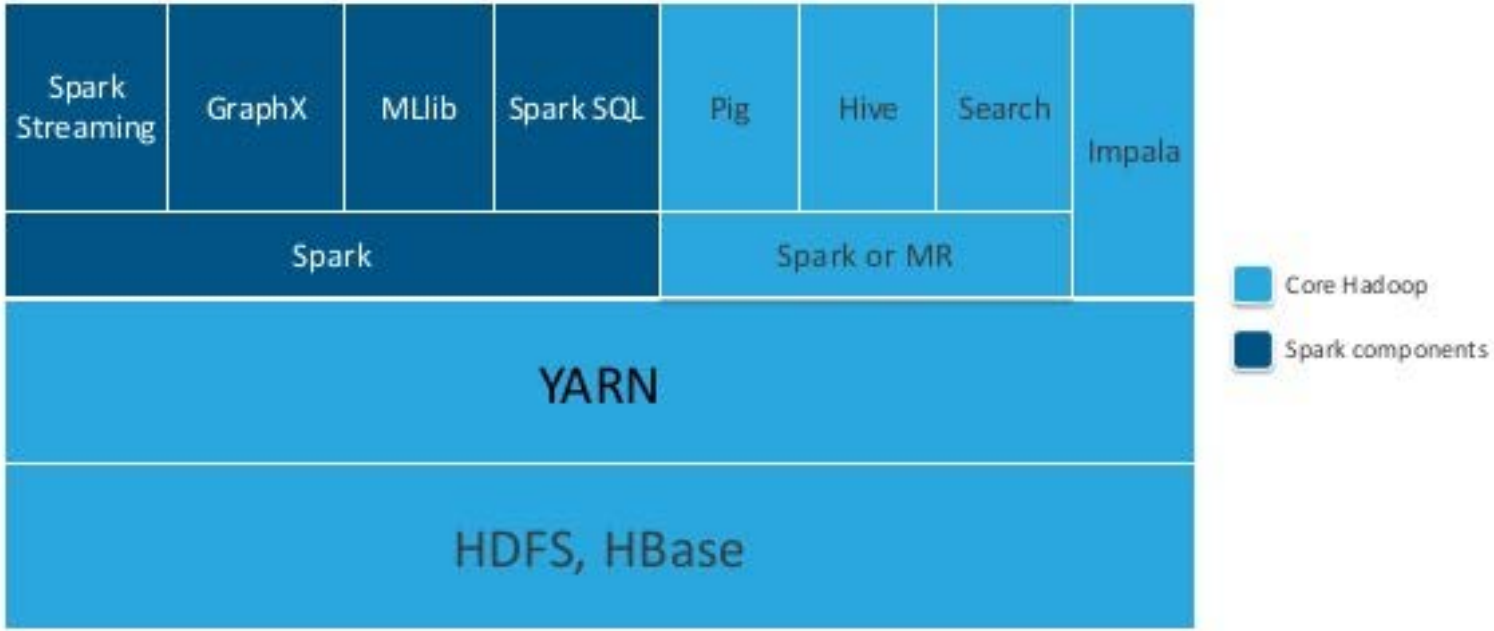


Logistic regression in Hadoop and Spark

Spark vs Hadoop

Spark key features	Apache Spark	Hadoop MapReduce
Speed	Ten to hundred times faster than MapReduce	Slower
Analytics	Supports streaming, machine learning, complex analytics, etc	Simple Map and Reduce tasks
Suitable for	Real-time streaming	Batch processing
Coding	Lesser lines of code	More lines of code
Processing location	In-memory	Local disk

Spark is Based on Hadoop

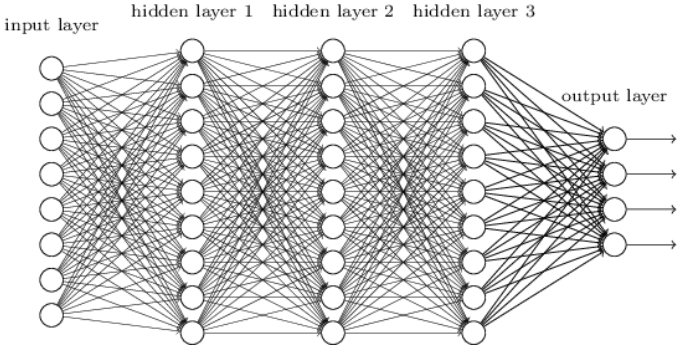


Why is Machine Learning Booming Now?

Big Data

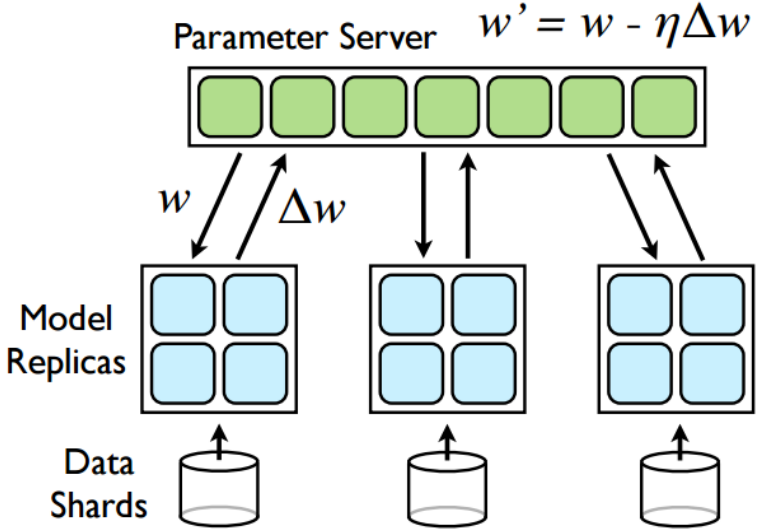
Big Computing Power

Evolution of Machine Learning



Distributed Machine Learning

- Examples: Tensorflow
- Simple structure
- Based on MPI




UNIVERSITY OF CENTRAL FLORIDA
1963
Thank you !
UCF