Notes: Introduction to Big Data for Machine Learning and Deep Learning

1. What is Big Data: https://en.wikipedia.org/wiki/Big_data
2. Why do we need new tools now?
    a. Tons of data generated.
    b. Traditional Databases do not scale well.
    c. Need for Cheap, scalable and efficient ways to process data.
3. No single tool can achieve everything. Each task might need its own tools.
4. Pre 2004: Parallel processing using MPI and PVM for compute intensive tasks.
    a. https://en.wikipedia.org/wiki/Parallel_Virtual_Machine
    b. https://en.wikipedia.org/wiki/Message_Passing_Interface
5. Traditional Databases from Oracle, MySQL, PostgreSQL, IBM DB2, Microsoft SQL Server
    a. Core idea: Hard-disk costs were not low.
    b. Mostly single box, multi-core.
6. 2004: Google publishes MapReduce paper: https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
7. Hadoop starts in ~ 2006/2007. Key features:
    a. Cheap hardware connected with fast local networks.
    b. Cost of storage is plummeting: http://www.mkomo.com/cost-per-gigabyte https://www.backblaze.com/blog/farming-hard-drives-2-years-and-1m-later/
    c. Data-intensive vs compute intensive tasks:
    d. HDFS:http://akceptor.blogspot.com/2014/10/hadoop-distributed-file-system-hdfs.html
    e. PIG (Yahoo! Research, ~2007-2008): http://guyharrison.squarespace.com/blog/2012/1/6/getting-started-with-apache-pig.html https://www.tutorialspoint.com/apache_pig/
    f. HIVE(Facebook, ~2010) : Hive-QL (Very similar to SQL) https://www.tutorialspoint.com/hive/hiveql_select_where.htm
    g. RAM available over time: https://www.reddit.com/r/hardware/comments/5cr2j2/how_much_ram_did_computers_have_over_time_timeline/
8. Spark:
    a. RAM sizes have grown. Why not use them more effectively?
    b. 2012-2014 (V1.0)
    c. Very good for ML and iterative algorithms: SparkMLib https://spark.apache.org/docs/latest/ml-guide.html
    d. SparkSQL, Hive, PIG
9. Distributed Deep-Learning:
    a. NVIDIA: 1999 Graphics processors, 2007: CUDA
    b. Operate on a vector/array simultaneously, Distributed VRAM.

    c. 2009 Rajat Raina, Andrew Ng http://robotics.stanford.edu/~ang/papers/icml09-LargeScaleUnsupervisedDeepLearningGPU.pdf
    d. TensorFlow: 2015
    e. Multiple GPUs on the same box
    f. Multiple GPU cluster.
10. NoSQL Databases:
    a. Traditional Databases: Minimize the disk space to store data using normalization.
    b. Example: Text documents , MongoDB ; Search using ElasticSearch
    c. Example: Key-Value store [Simple Login] Redis, MemCached,
    d. Columnar stores: HBase
    e. Many more models.
11. Cloud computing:
    a. 2002 : AWS (Infrastructure as a service), EC2, S3, …...
    b. Compute infrastructure on demand.
    c. Platform as a service [Databases]
    d. Google Cloud Platform: 2008
    e. Microsoft Azure: 2010
    f. IBM, Oracle,.....
12. Constantly changing landscape as technologies (CPU, GPU, RAM, Storage) and applications evolve over time.