COMPSCI 590S Systems for Data Science **Fall 2016**

Lecture 25: Dynamo and Basics of ML [Tensor Flow]

Lecturer: Emery Berger *Scribe(s): Namrita Pandita, Mark Saad*

25.1 Virtual Nodes [Used in Dynamo]

Why use Virtual Nodes?

- For handling heterogeneity of Nodes.
  - A Node with 1TB Memory can handle more load than the one having just 256MB. Placing more Virtual Nodes for the machine with higher capacity can proportionally handle the load.
- For load balancing.
  - Virtual Nodes ensure that if a node goes down, its load gets proportionally divided among other active nodes.

25.2 Gossip Protocol [Used in Dynamo]

Broadcast are easy but it scales horribly, gossip protocol is peer based (peer to peer) and scales infinitely.

Reason for Scalabilty:

- Gossip Protocol uses P2P. i.e. No matter how big the cluster is, one node communicates with just one other node at a time.
- How does it work?
  - It randomly chooses a peer instead of a fixed peer and they trade messages and update state
  - Repeats the process every so often
  - No deterministic guarantees, entirely probabilistic
  - Rapid Dissemination of messages
  - Eventually converges. i.e. All nodes will eventually receive all the messages.

25.3 Does Dynamo implement ACID ?

- It doesnt. The only guarantee is that the key-value operations are atomic

25.4 Basics of Machine Learning

- Two types of ML:
  - Supervised (has a label)
  - Unsupervised (no label)
- ML is used mainly for classification (sometimes binary classification)
- Cross validation:
  - you train ML algorithm on a subset of the data and then see how it classifies the rest of the data.
  - You then calculate the accuracy of that classification (precision and recall: which is a fancy way of saying false positives and false negatives)
- The goal of ML is to come up with a loss function that converges. The loss function is able to find the wrong predictions (both false positives and false negatives)
- Feature Engineering:
  - ML started with the idea of feature engineering (where you train the machine to classify based on certain features you choose, much like an if this then that tree)
  - Feature engineering proved to be hard and ineffective and thats why we transitioned to neural networks
- Neural Network
  - For neural nets to work you need a lot of data, you may also want to call it BIG DATA
  - Big training sets are slow to converge and compute intensive.
  - TPUs are specific chips[ASIC] that Google designed and built to use to run the tensor flow classifications in production