Gaurav Singh | UFID: 36398850

COP 6726: Database Systems Implementation
Spring 2018
Weekly Assignment 9

27-03-2018:

- MPPs are not parallelized
- MPPS are stupid today, green plum -> entity over Postgres
- Oracle is different, its parallelized
- You can have run entire Facebook data analytics on average high-end server racks.
- The thing is it would be very slow.
- nVidia and other GPU vendors are really jumping into analytics
- Now they build GPUs called GPGPUs .. general processing gpus
- They were actually build for parallelizing workloads for gaming but obviously can be used to parallelize all sorts of simple tasks
- have found a great market in self driving cars and all sorts of other applications
- When fast analytics over massive data is needed, use GPGPUs
- MapP is the coolest new DB which supports parallelization.
- mapReduce has two steps
    - map
    - Reduce
- It's a simple key value pair idea
- Don't use string for both keys and values, its very inefficient and stupid.
- Mapp uses a similar technique, it's called sharding.
- The idea is the same to segment the data and then work on solving them
- Hadoop uses mapReduce, it was one of the famous big things in "Big Data"

29-03-2018

- Dbs which use Iterator Model are so much slower compared to Hadoop
- No comparison as tasks are not bottlenecked
- Use map reduce on activity where you need multiple rows and operations
- Each of those should run faster on in map reduce.
- They will also follow the normal logic of map reduce.
- Biggest problem would be to convert other problem types into map reduce.
- Hadoop map reduce is good for group by. It is a bit more intuitive here
    - Just map all the different groups you need
    - Count all the values for each group
- Group by is basically inherent in the structure of map reduce
- Google also uses map reduce but adds some other things to make it even faster.