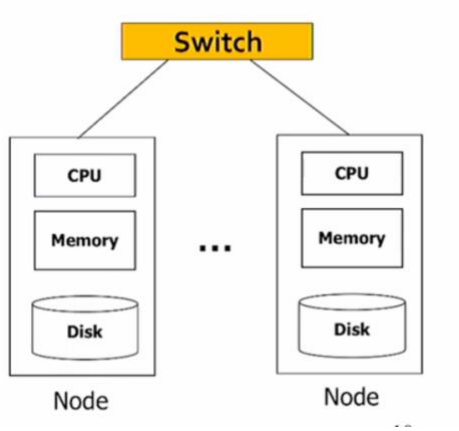
- Week 1:
- Data engineer: develops, constructs, tests, and maintains architectures.
- Data Scientist: cleans, organizes, perform descriptive statistics.
- Rack contains 16-64 nodes.
- 

- Node= cpu memory and disk
- Chuck servers: files split into contiguous chunks
- Master node: stores meta data about where files are stored

- Map Reduce: Dive and conquer
- Mapper:
    - Divide: divide file to words. Output one or multiple thing for each record
- Combiner: after mapper and before reducer
    - If a reduce function is commutative and associative then can be used as combiner
    - Commutative: a+b=b+a
    - Associative (a+b)+c=a+(b+c)
- Group by key: sort and shuffle
- Reducer: aggregate, summarize, filter or transform (one reduce function call per unique k)

- Data flow: input and output are stored on a distributed file system
    - Schedular tries to schedule map tasks close to physical storage location of input data
    - Intermediate results are stored on local file system of map workers
- Master:
    - Coordination
        - Task status,
        - idle tasks
        - master pushes this info to reducer
- Map worker failure: even if worker failed after job, it needs to be redone.
- Reducer worker failure: only in process are needed to redo/reset to idel.
- How many map and reducer needed: more mapper needed than reducer
- Map-reduce:
    - Pro: best for key-value pairs, summarize data.
    - Con: no graph data, random access, cannot do intermediate steps, gradient based learning
- Spark:
    - Batch processing
    - Interactive data query
    - Real time analysis
    - Streaming data
- Resilient distributed dataset: immutable, in-memory collection, parallel data structure
    - Distributed
    - Resilient: to data failure
    - Built in data structure
- Partitions: split by hashing function to 64mb
- Shuffling:
- Two types of operations on RDD:
    - Transformation: lazy not immediate
    - Action: immediate
- Transformation:
    - Map: map[T](f:A=>B):RDD[t]
    - Flatmap
    - Filter
    - Distinct
    -

- Action:
    - Collect
    - Count
    - Take
    - Reduce
    - foreach