

Lecture Notes 6

What can you do on a single Node?

- To use a heap
 - $O(n \log n)$
- Divide the dataset for each core (multithreaded approach). Then at the end, merge all of the results.
 - **important for system design questions**
- Identify bottlenecks
 - One bottleneck of that approach will be in the merging stage

YARN Concepts (1)

Container

- YARN uses an abstraction of resources called a container for managing resources -- a unit of computation of a slave node, ie a certain amount of CPU, Memory, Disk, etc

3 Components

- Application Master - Single instance per job
- Spawned within a container when a new job is submitted by a client

Resource Manager

- Scheduler - Global scheduler, responsible for allocating resources to jobs
- Application Manager

Node Manager

- Single instance per slave node (multiple containers created)
- Responsible for monitoring and reporting on local container status (all containers on slave node) to Resource Manager

