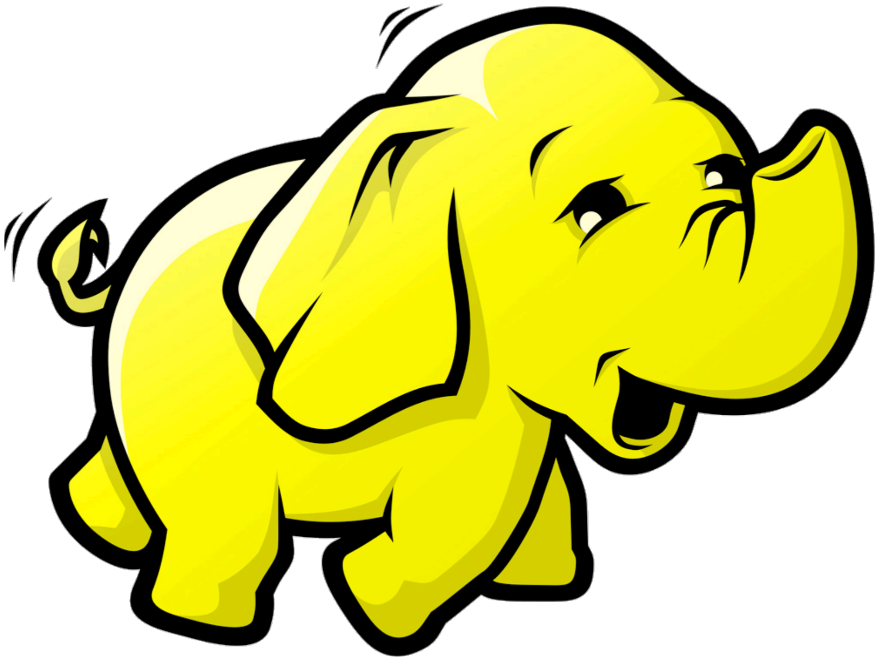


# Hadoop Architecture

# ⚠ Note

These notes provide a general introduction to Hadoop.  The content is pretty self-explanatory and is meant to help you understand Hadoop conceptually.  Pay attention to the page that discusses how to write python code so that it works with Hadoop. This is explained in more detail in the **Hadoop Introduction** activity.

# What are we learning?

Problem Solving

Parallel Programming

MapReduce

Hadoop

Pig

Hive

NoSQL

# Hadoop Framework

## Four modules:

**Hadoop Common:**
Libraries and utilities needed by other Hadoop modules

**Hadoop Distributed File System (HDFS):**
Distributed file system that stores data across (potentially) many machines

**Hadoop YARN:**
Resource management platform responsible for managing computing resources and scheduling applications.

**Hadoop MapReduce:**
Programming model for processing big data.

# Hadoop Family

MapReduce code:

- written in Java.
- can also be written in python (see **Hadoop Introduction** activity).

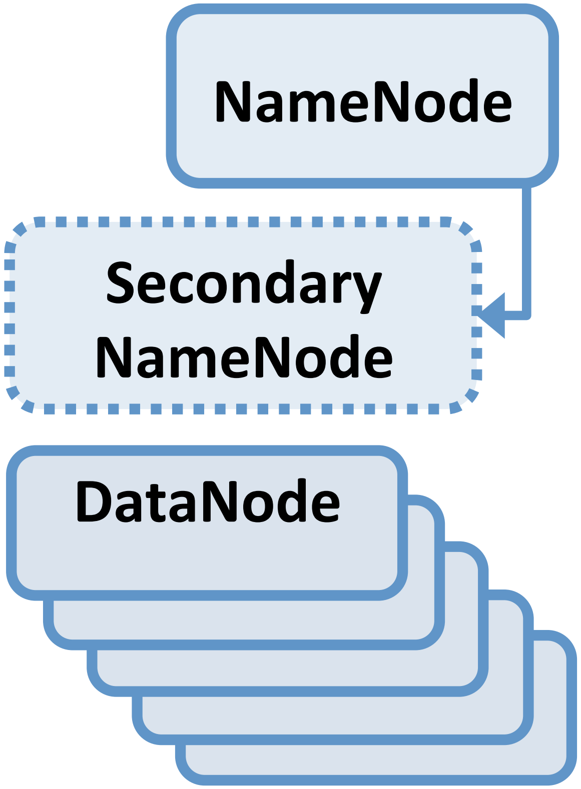
**Pig**: scripting language written in Pig Latin.

**Hive**: SQL variant.

# Hadoop Distributed File System

Written in Java and is:

- Distributed
- Scalable
- Portable



Single, controls all DataNodes

Builds snapshot of NameNode in case it fails

- Contains actual data in very large chunks.
- Default replication of data is 3

# HDFS Advantages

Data awareness between the data and the job tracker.

- If node A contained data of a,b,c and node Z contained data of x,y,z, then the job tracker would assign map processes accordingly.

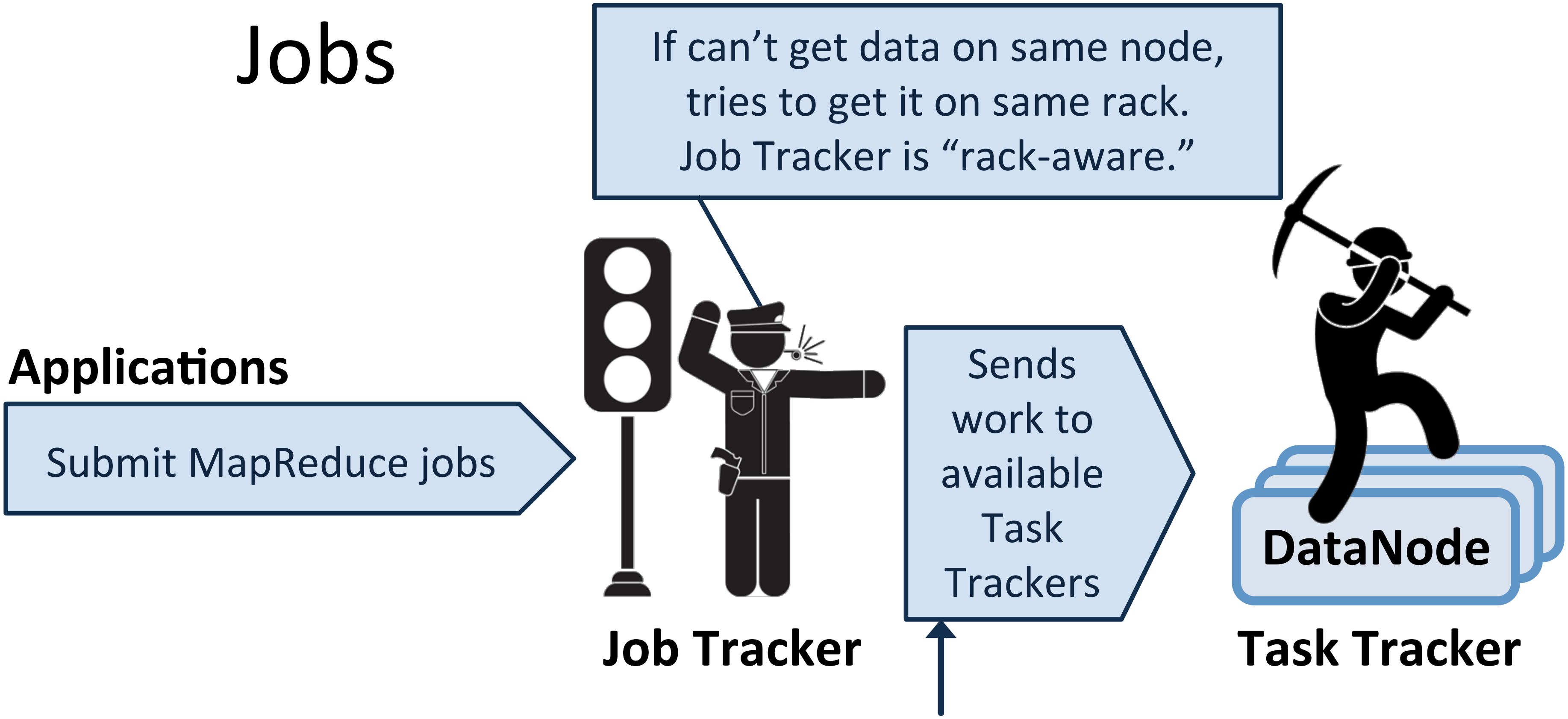
Reduces network traffic and data transfer.

# HDFS Limitations

Designed for immutable files.

Retrieving data and storing data must go through Hadoop.

# Jobs



If can’t get data on same node, tries to get it on same rack. Job Tracker is “rack-aware.”

**Applications**

Submit MapReduce jobs

**Job Tracker**

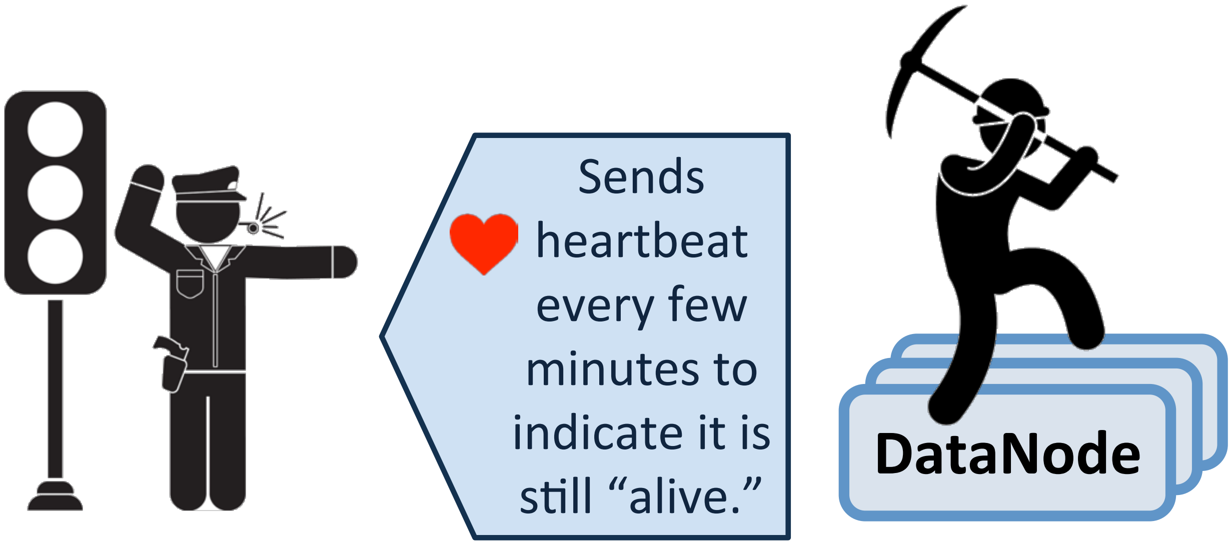
Sends work to available Task Trackers

**DataNode**

**Task Tracker**

**Goal**: Keep process as close to data as possible.

# Tasks



Sends heartbeat every few minutes to indicate it is still "alive."

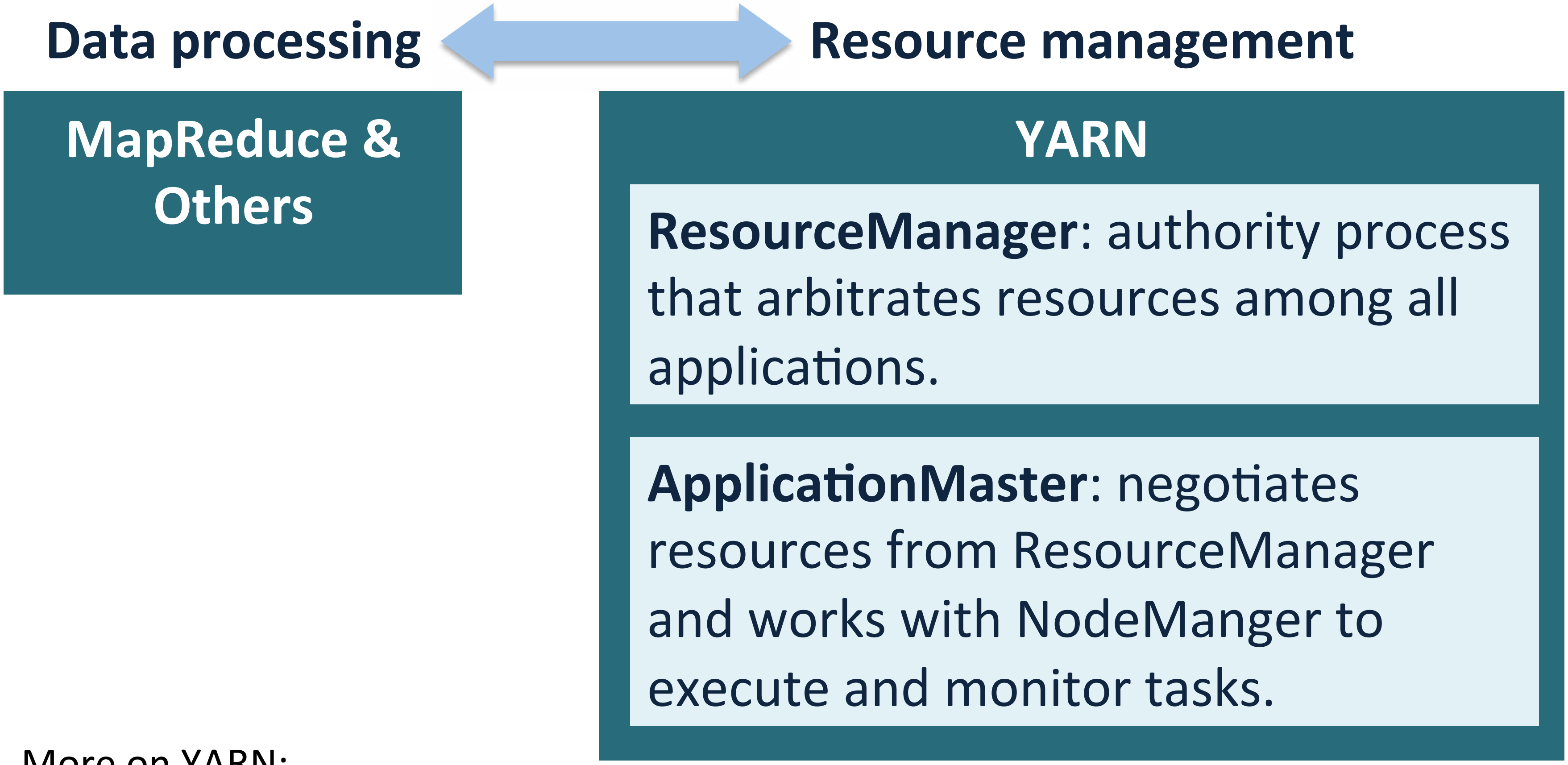
DataNode

**Job Tracker**

**Task Tracker** runs:

- on each node
- tasks given by Job Tracker

Separated in Hadoop 2.0



**Data processing**

**Resource management**

**MapReduce & Others**

**YARN**

**ResourceManager**: authority process that arbitrates resources among all applications.

**ApplicationMaster**: negotiates resources from ResourceManager and works with NodeManger to execute and monitor tasks.

More on YARN:
https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

# Writing Hadoop Programs

In python

```
#!/usr/bin/env python

#any imports will go here
#no global variables

def main(argv):
    #our map or reduce code will go here
if __name__ == "__main__":
    main(sys.argv)
```

# Cloud Computing

In order to see the true power of Hadoop, we will utilize Amazon Web Services.

You should have already created an account and have applied for educational credits.

The activity will walk through how to run Hadoop in the cloud environment.