Introduction to Hadoop

Hadoop and Ecosystem (part 1)

- Tuhin Mahmud

About Me

- Name: **Tuhin Mahmud**
- Advisory Software Engineer @ IBM
- Worked as software developer for various companies(IBM, Doubleclick, Microsoft and Telcordia)
- Linkedin profile: www.linkedin.com/in/tuhinmahmud/
- Email: tuhinm@hotmail.com
- Current interest is Big Data and Cloud.
- Cloudera Ceritified Apache Hadoop developer.
- Recent project on Hadoop –Coverting a legacy Report generation system that processes 1 terabyte daily logs and statistic into a scalable solution . Hadoop/Hive solution.

Today's Discussion Topics

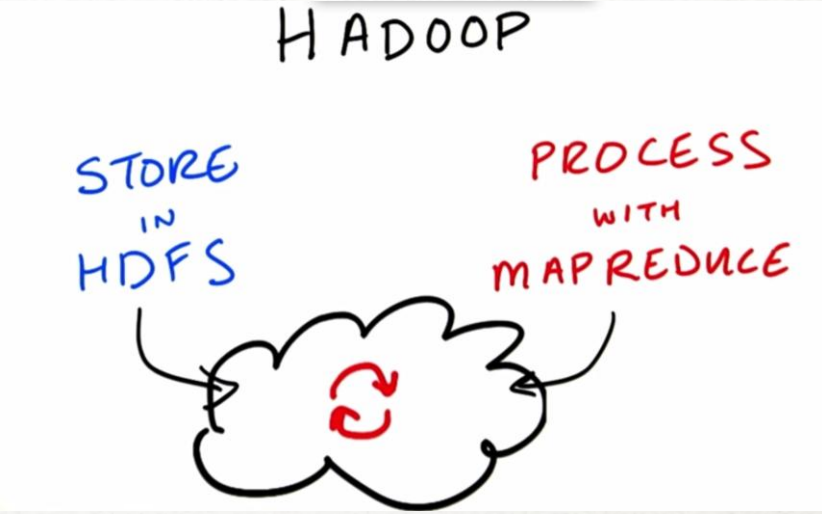
- Big Data and Hadoop
- HDFS
- MapReduce
- Code demo on Hadoop MapReduce

What is Big Data?

- According to IBM: "Every day, **2.5 billion gigabytes** of high-velocity data are created in a variety of forms, such as social media posts, information gathered in sensors and medical devices, videos and transaction records"
- Definition of Big Data is very subjective and changing .
- ***3 Vs for Big Data***
  - Volume
  - Variety
  - Velocity
  - 4th V is also sometime talked about is ***Veracity*** (correctness)

What is Hadoop?

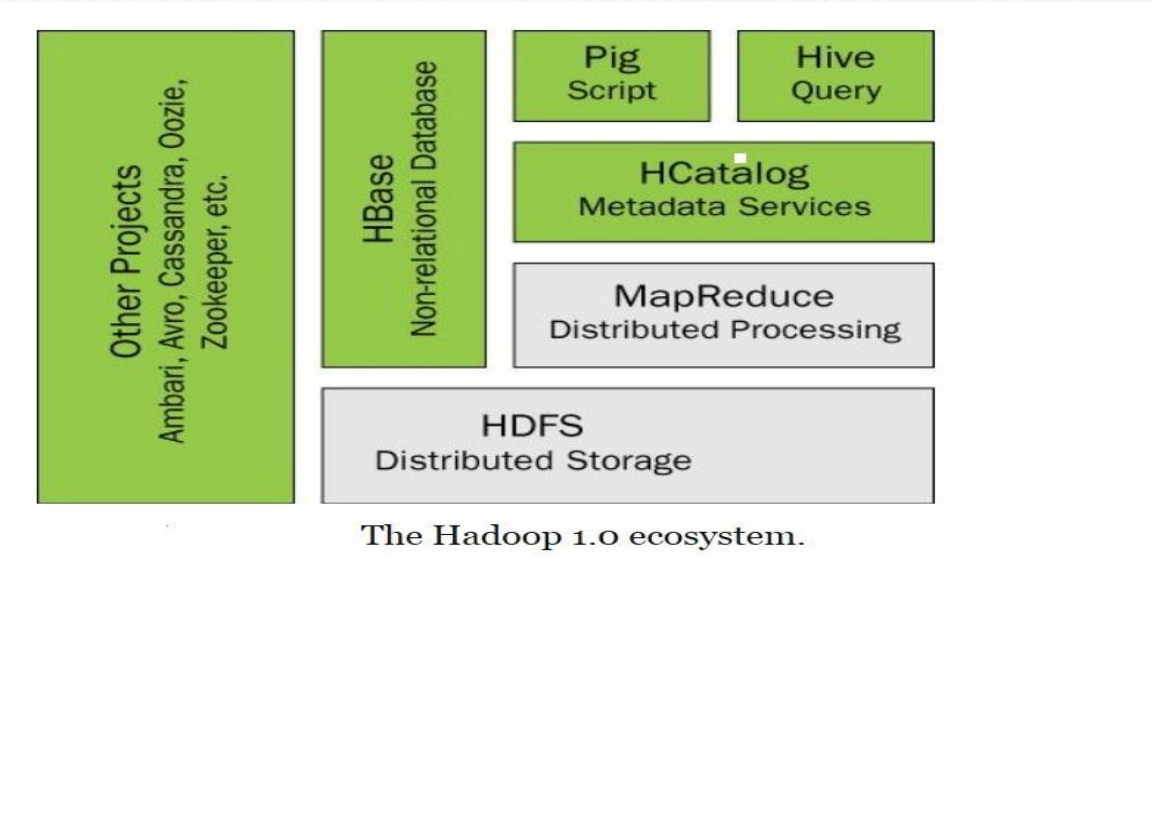
- Hadoop is an open-source project overseen by the Apache Software Foundation
- Based on papers published by Google in 2003 and 2004.
- The name “Hadoop” came from creator Doug Cutting’s son’s stuff elephant name.
- Written in Java
- Runs on commodity hardware
- Two main components
  - Hadoop Distributed File System (**HDFS**)
  - Mapreduce – to Process data



Source: udacity

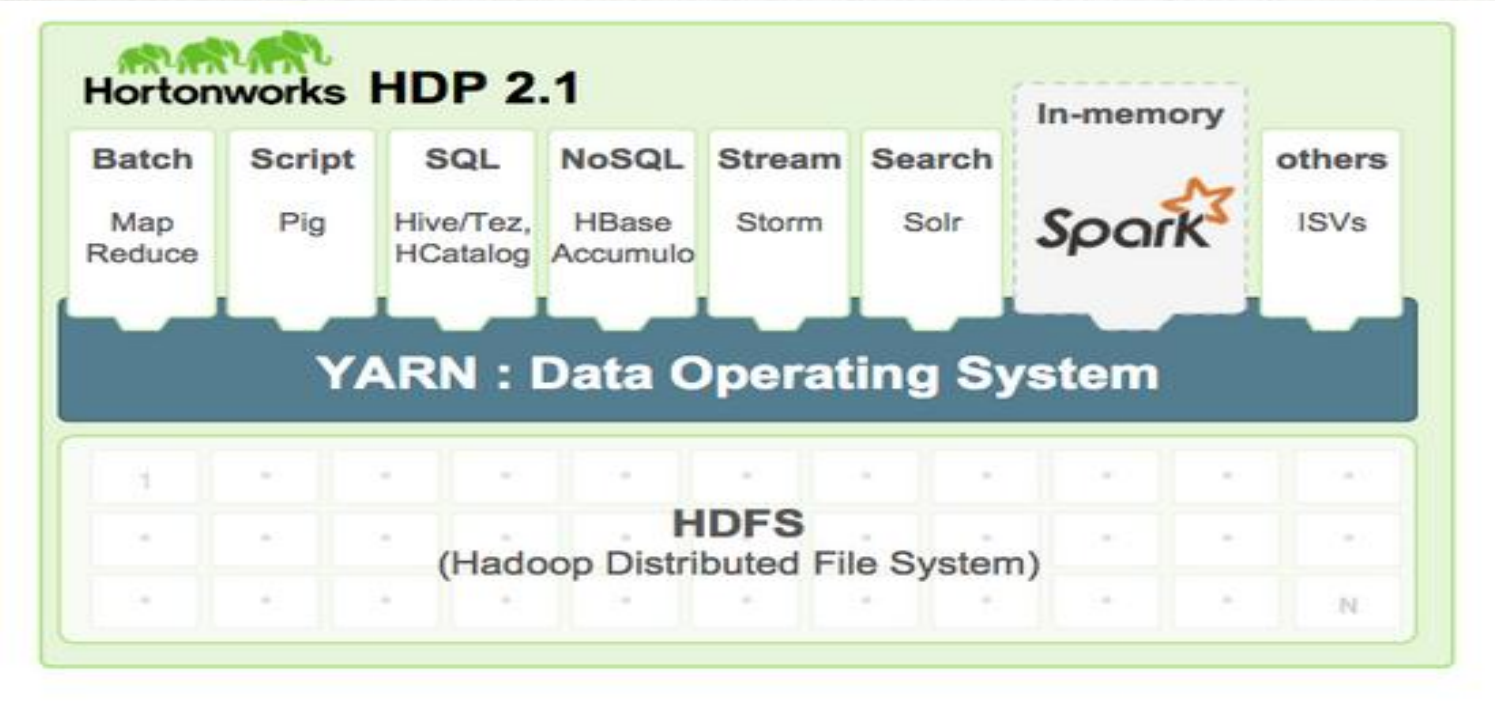
Hadoop Ecosystem

- Pig,
- Hive,
- HBase,
- Flume ,
- Oozie,
- Sqoop



The Hadoop 1.0 ecosystem.

Hadoop Ecosystem with YARN



YARN adds a more general interface to run non-MapReduce jobs within the Hadoop frameworks

Hadoop Termnilogies

- Each Machine is called a **Node.**
- A **Cluster** can have one or several thousand nodes.
- A **job** is a unit of work that the client wants to be performed.
- Hadoop runs the job by dividing it into **tasks.**

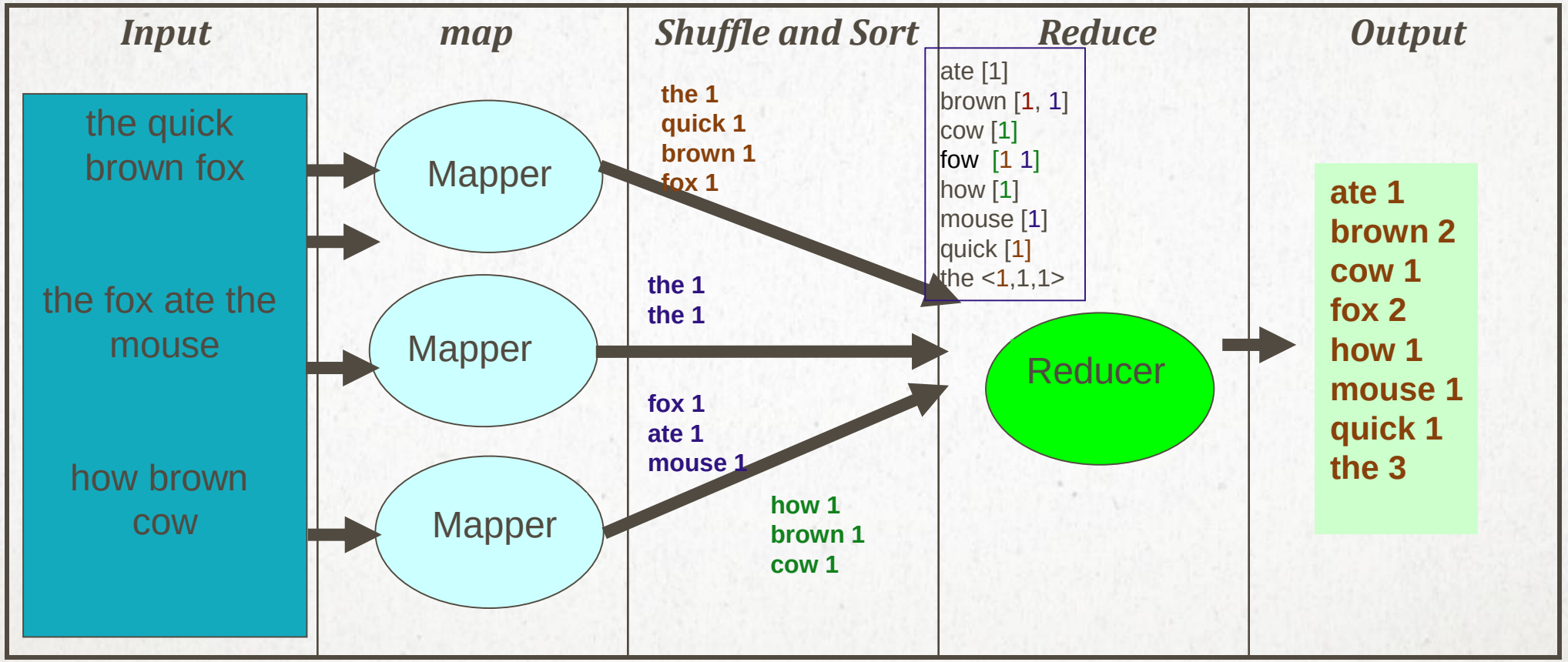
Mapper

- ***The Map***
- A map transform is provided to transform an input data row of key and value to an output **key/value**:
- map(key1,value) -> list<key2,value2>
- That is, for an input it returns a list containing zero or more (key,value) pairs:
- The output can be a different key from the input
- The output can have multiple entries with the same key

Reducer

- ***The Reduce***
- A reduce transform is provided to take all values for a specific key, and generate a new list of the *reduced* output.
- reduce(key2, list<value2>) -> list<value3>

Simple Word Count Example of MapReduce



Code
& Demo

Hadoop Word Count (Python using streaming)– Mapper

mapper.py

```
#!/usr/bin/python

import sys

for line in sys.stdin:

   line = line.strip()

   words = line.split()

   for word in words:

      print '%s\t%s' % (word, "1")
```

Hadoop Word Count (Python using streaming) Reducer

reducer.py

```
-----------------------------------
#!/usr/bin/python
import sys
word2count = {}

for line in sys.stdin:
    line = line.strip()

    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        continue
    try:
        word2count[word] = word2count[word]+count
    except:
        word2count[word] = count

for word in word2count.keys():
    print '%s\t%s'% ( word, word2count[word] )
```

References

1. **Tom White. *Hadoop – The Definitive Guide* (3rd Edition)**
2. Henry H. Liu . *Hadoop 2 Essentials*.
3. Arun C Murthy , Vinod Kumar Vavilapalli. *Apache Hadoop YARN – Moving beyond mapreduce and Batch Processing with Apache Hadoop 2*.
4. Donald Miner & Adam Shook . *Map Reduce Deign Patterns*
5. https://developer.ibm.com/hadoop/
6. http://bigdatauniversity.com/
7. https://www.udacity.com/course/ud617
8. http://hortonworks.com/products/hortonworks-sandbox/
9. **Michael G. Noll.** http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

Online Resources

- ***Google original papers that started it all***
  - *http://static.googleusercontent.com/media/research.google.com/en/us/archive/gfs-sosp2003.pdf*
  - *http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf*
- ***Certifications***
  - **Cloudera Hadoop Certifications**
    - http://www.cloudera.com/content/cloudera/en/training/certification.html
  - **Hortonworks Hadoop Certification**
    - http://hortonworks.com/training/certification/

THANK YOU!

- Please let me know if you want to volunteer for any presentation on this topic in future meetings.

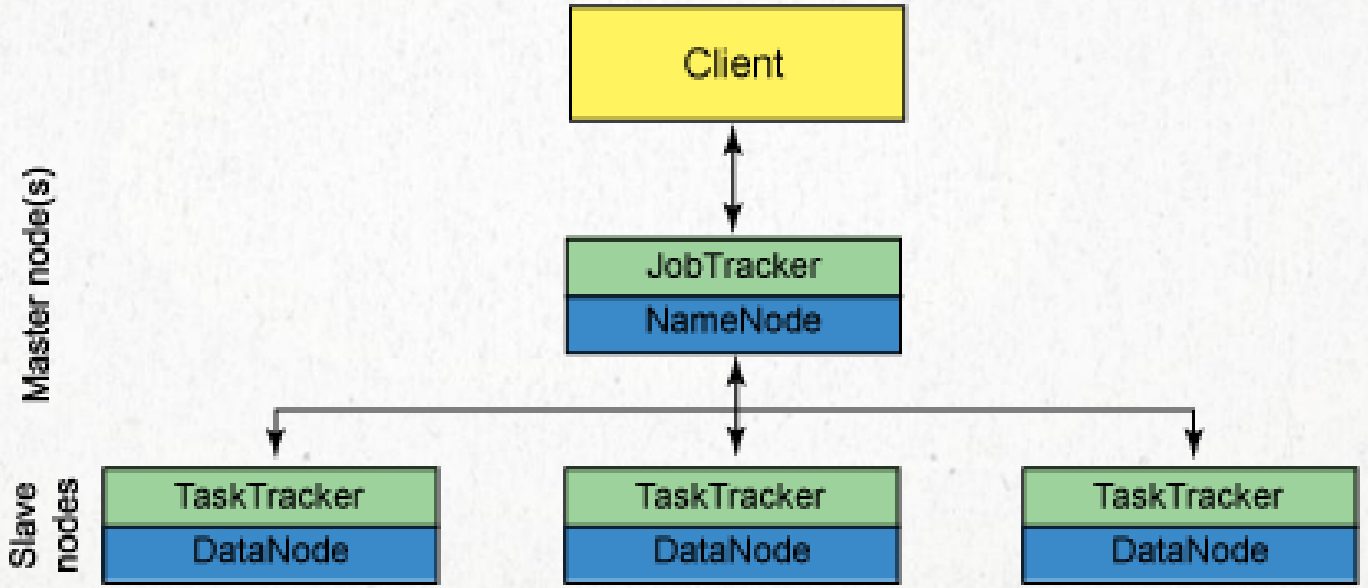
- "Anyone who stops learning is old, whether at twenty or eighty. Anyone who keeps learning stays young."
  — Henry Ford

Back Up 1
Relational database vs Hadoop

Relational Databases **Use when**	**Hadoop** **Use when**
Interactive OLAP Analytics (<1sec)	Structured or Not (Flexible)
Multistep ACID transaction	Scalabilty of storage /Compute
100% SQL compliance	Complex data processing

Simple illustration of the Hadoop cluster architecture

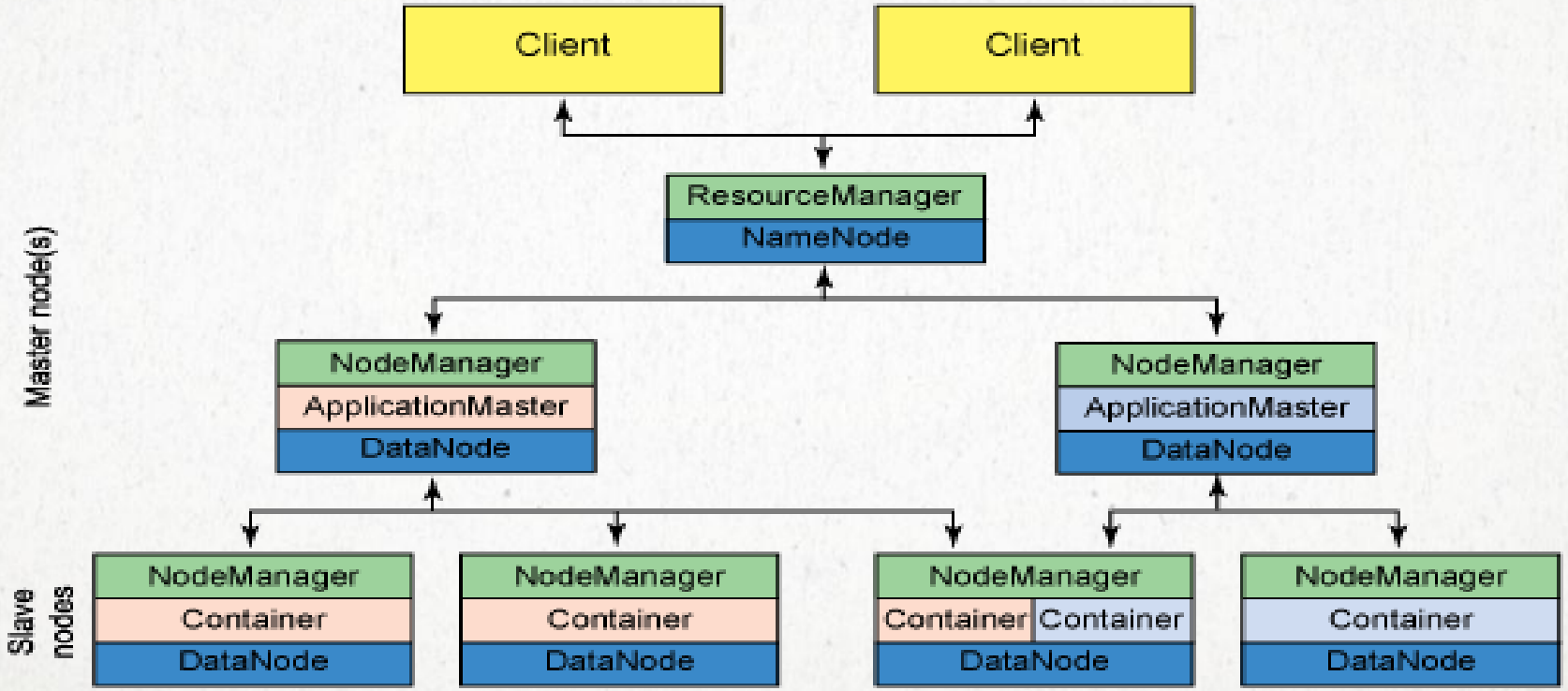
- **Hadoop and MRv1**



Source: http://www.ibm.com/developerworks/library/bd-hadoopyarn/

Simple illustration of the Hadoop cluster architecture

- **The new architecture for YARN**



Source: http://www.ibm.com/developerworks/library/bd-hadoopyarn/

Backup