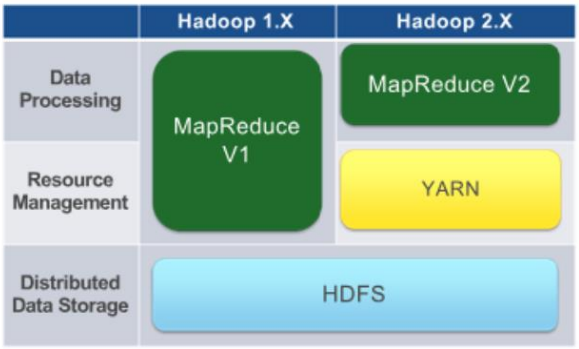
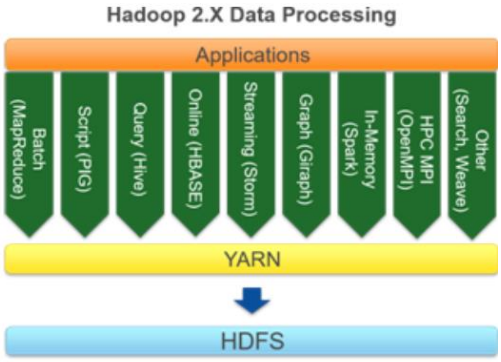
Module 3: Hadoop and MapReduce

1. What's Hadoop?

- Framework that processes large datasets via distributed computing.

2. Hadoop Architecture Overview





- The key advantages of Hadoop 2.x over 1.x.
  I. Faster recovery from Name Node failure due to Standby Name Node.
  II. The inclusion of YARN allows for greater scaling of jobs.

3. MapReduce

- **What's MapReduce**
  - It's parallel processing software framework for distributed data processing.
- **Phases**

  Input > Input Splits > Mapping > Shuffling > Reducer > Final Output
- **Join**
  - Map-Side Join is faster than Reduce-Side Join because join operation is done.
  - Compared with Map-Side Join, Reduce-Side Join has fewer constraints.
- **Note**
  - Sometimes MapReduce could be substituted with Apache Spark. Because all MapReduce operational process needs to read and write files, its operational efficiency would be slower.

Reference:

http://pcse.pw/9C637
https://ithelp.ithome.com.tw/articles/10190756