Hadoop

Presented by: Ashly Horner, Jillian Hart, Douglas Williams

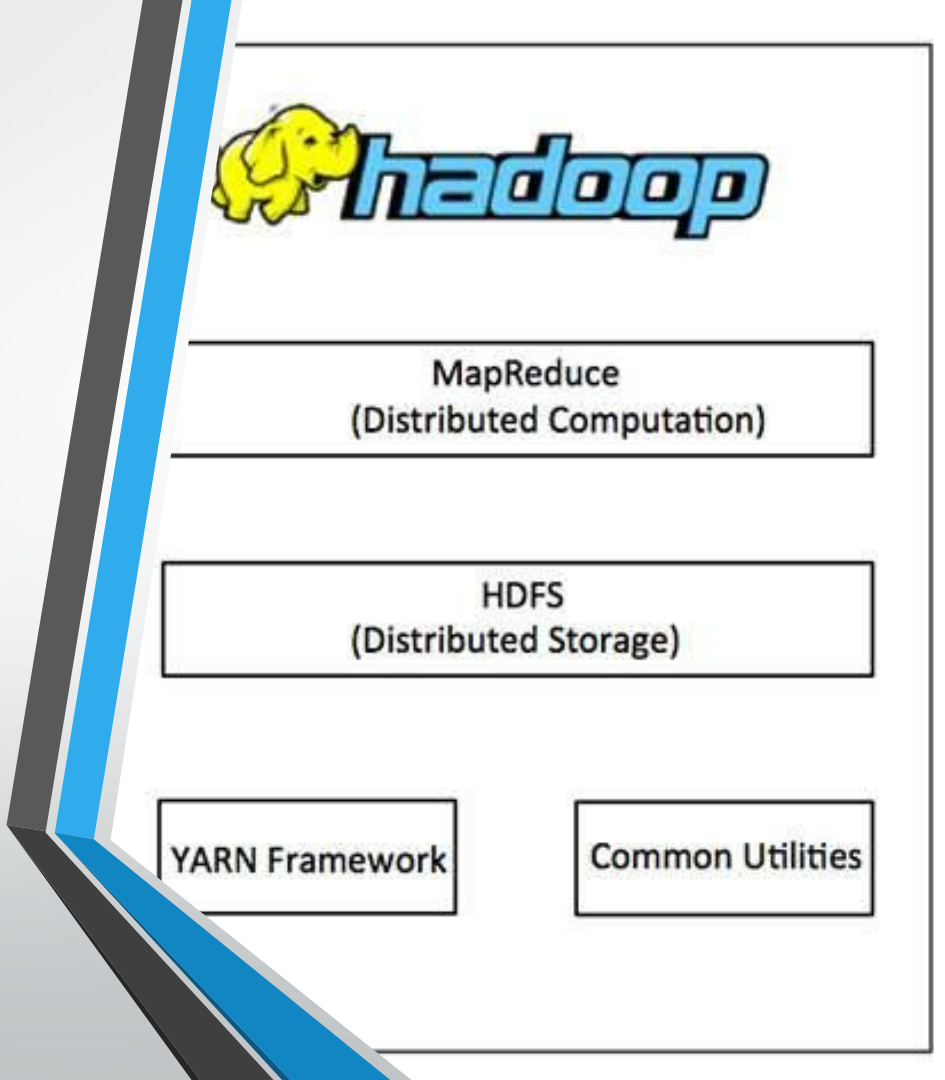
What is Hadoop?

- Open source big data management system
- Used to process large data sets
- Fast and reliable data storage/processing
- Parallel processing

*(Doug Cutter's son had a toy elephant named Hadoop. Doug Cutter and Mike Carafella named the product after the toy.)

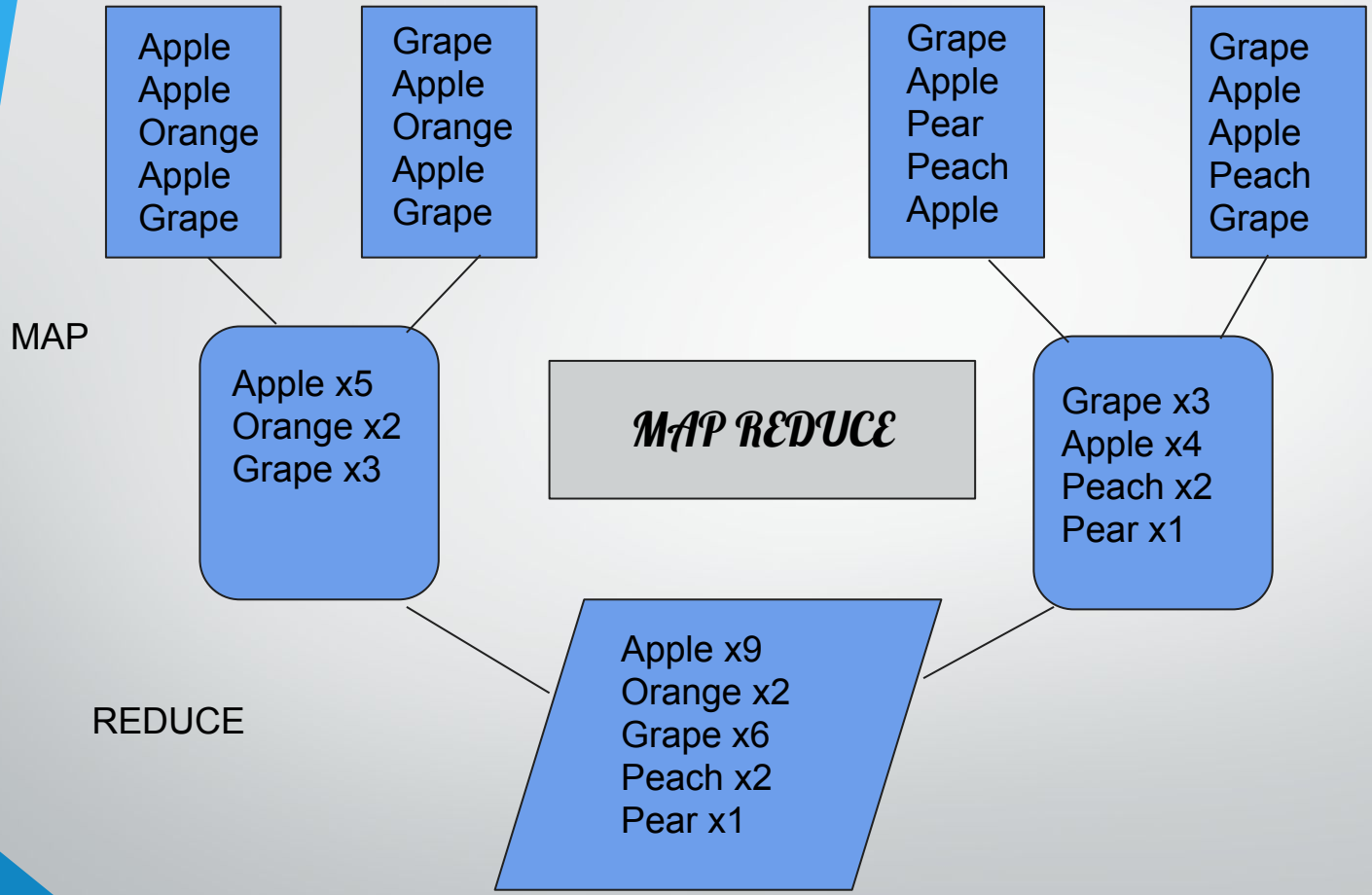
Architecture

- 4 main components:
- HDFS
- MapReduce
- YARN
- Hadoop Common



Map Reduce

- It is a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster.
- The term "MapReduce" refers to two separate jobs
- first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples
- The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples


Apple
Apple
Orange
Apple
Grape
Grape
Apple
Orange
Apple
Grape
Grape
Apple
Pear
Peach
Apple
Grape
Apple
Apple
Peach
Grape
MAP
Apple x5
Orange x2
Grape x3
MAP REDUCE
Grape x3
Apple x4
Peach x2
Pear x1
REDUCE
Apple x9
Orange x2
Grape x6
Peach x2
Pear x1

Comparing Hadoop MapReduce with Apache Spark

MapReduce

- Data processing:Read/Write to a Disk
- Able to work with much larger data sets

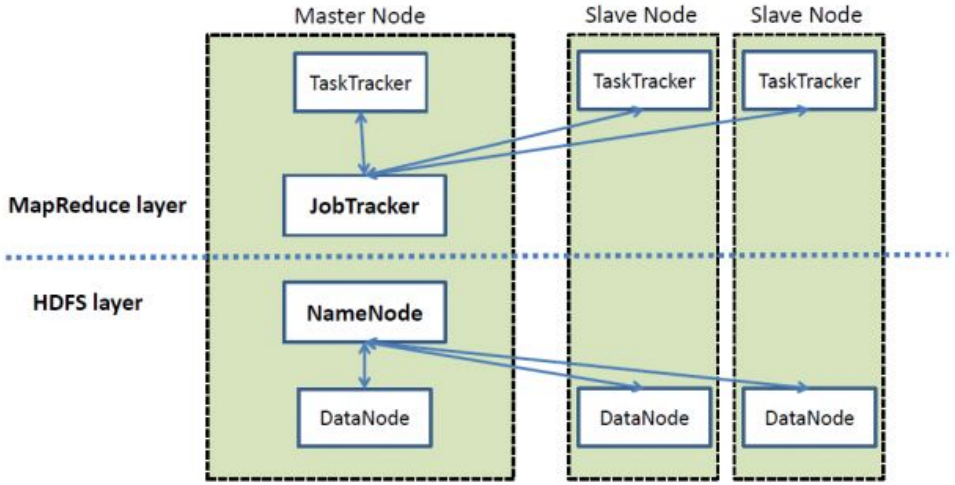
Spark

- Data processing: in-memory
- Spark processes 100 times faster

HDFS – Hadoop Distributed File System

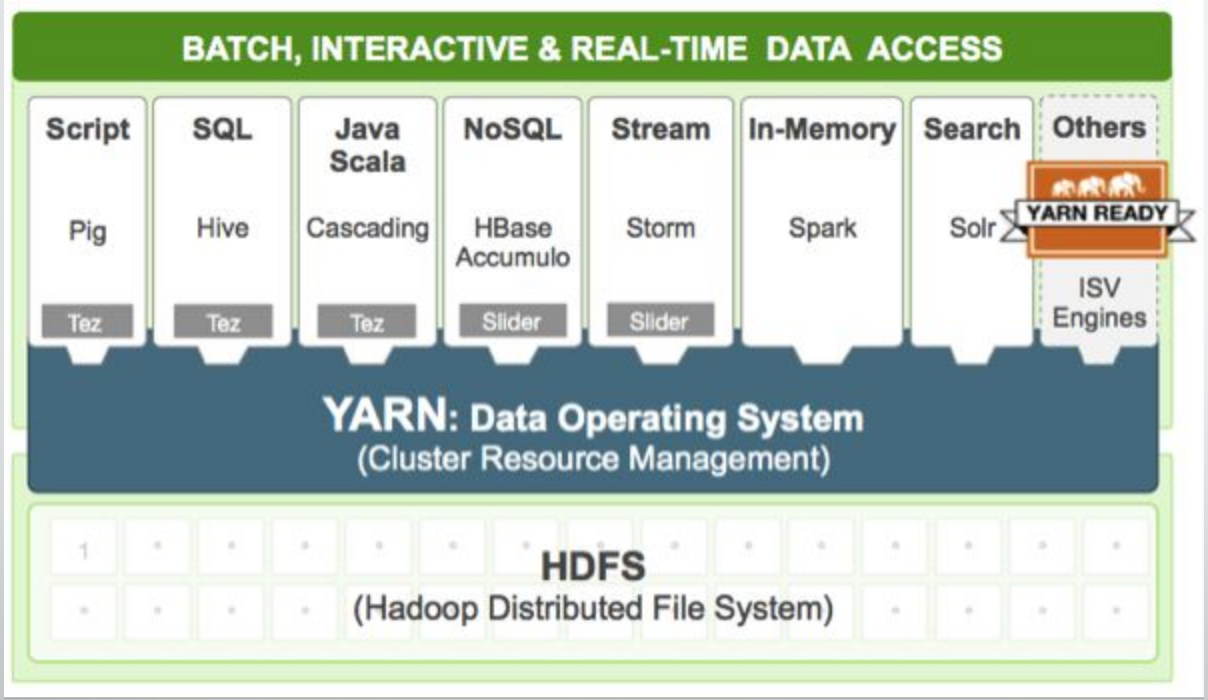
- File system of hadoop framework.
- Due to the way HDFS operates it has redundancy
- Designed to store and manage huge volumes of data efficiently.
- User space file system – runs as the user process.

High Level Architecture of Hadoop



YARN-Yet Another Resource Negotiator

- Cluster resource manager.Used to reduce bottleneck from mapReduce version 1. It splits Job Tracker in two.
  - 1)Manages resources for applications
  - 2)Managers resources for job scheduling/queue.
- Use to increase data analysis and scale resources according to client requirements.
- Supports multiple ways of processing data, like interactive query on Apache Spark, and other processing engines.


BATCH, INTERACTIVE & REAL-TIME DATA ACCESS
Script
Pig
Tez
SQL
Hive
Tez
Java
Scala
Cascading
Tez
NoSQL
HBase
Accumulo
Slider
Stream
Storm
Slider
In-Memory
Spark
Search
Solr
Others
YARN READY
ISV
Engines
YARN: Data Operating System
(Cluster Resource Management)
HDFS
(Hadoop Distributed File System)

- Installed Hadoop in Stand-Alone Mode on Ubuntu 18.04
- Ran Hadoop with map reduce example

Activities Terminal Fri 14:03

labpc1@bdlab-pc1: ~

File Edit View Search Terminal Help

```
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=34
2018-10-25 09:30:39,352 INFO mapred.LocalJobRunner: Finishing task: attempt_local752158577_0002_r_000000_0
2018-10-25 09:30:39,352 INFO mapred.LocalJobRunner: reduce task executor complete.
2018-10-25 09:30:40,294 INFO mapreduce.Job: Job job_local752158577_0002 running in uber mode : false
2018-10-25 09:30:40,294 INFO mapreduce.Job:  map 100% reduce 100%
2018-10-25 09:30:40,295 INFO mapreduce.Job: Job job_local752158577_0002 completed successfully
2018-10-25 09:30:40,299 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=1330800
                FILE: Number of bytes written=3133603
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=2
                Map output records=2
                Map output bytes=33
                Map output materialized bytes=43
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=2
                Reduce shuffle bytes=43
                Reduce input records=2
                Reduce output records=2
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=6
                Total committed heap usage (bytes)=729808896
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=147
        File Output Format Counters
                Bytes Written=34
labpc1@bdlab-pc1:~$ cat ~/grep_example/*
19      allowed.
1       allowed
labpc1@bdlab-pc1:~$
```

Questions?