Hadoop

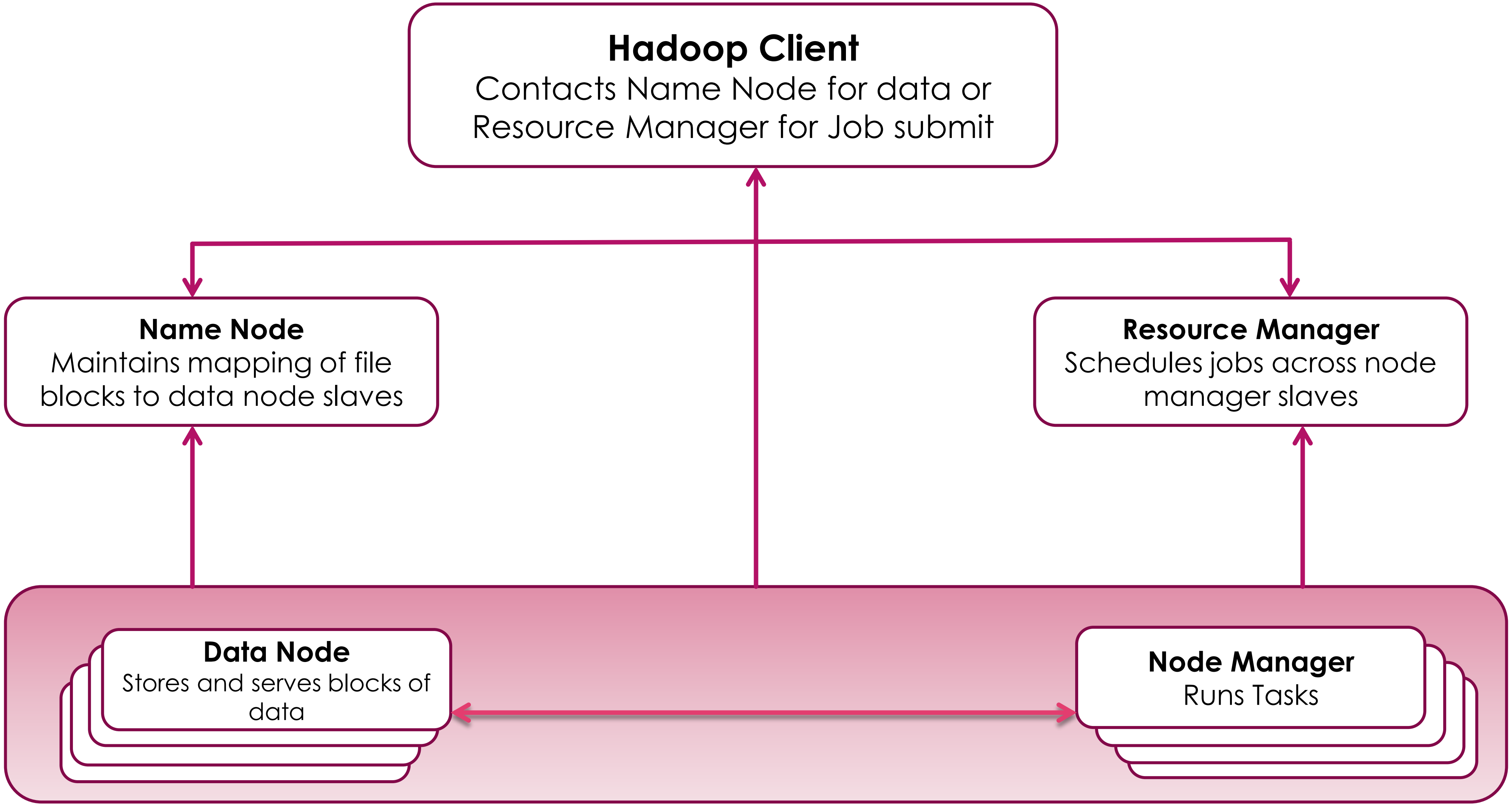
Hadoop Distributions

- Cloudera – CDH
- Hortonworks – HDP
- IBM – BigInsights
- Mapr – Mapr
- EMC – Greenplum
- Hp – Vertica, Haven
- Teradata – Aster
- Oracle – Exalytics
- Microsoft - HDInsights

- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- Hadoop runs on commodity hardware
- Completely written in java
- Robust, Self-healing and Resilient
- Hadoop core components : HDFS and Map Reduce

About Hadoop

- Released in 2008 under Apache
- A well publicised feat, the New York times used Hadoop on EC2 to crunch 4 TB of scanned archives from paper, converting the to PDF for web. This processing took less than 24 hours on 100 machines
- In April 2008, Hadoop broke a world record to become fastest system to sort a terabyte of data. Running on a 910 node cluster, Hadoop sorted 1TB in 209 seconds beating previous years winner of 297 sec.
- Later in November of the same year, Google reported that itd Map-reduce implementation sorted 1TBin 68 seconds.
- In April 2009, a team at Yahoo! Used Hadoop to sort 1TB in 62 seconds



Hadoop Client
Contacts Name Node for data or Resource Manager for Job submit

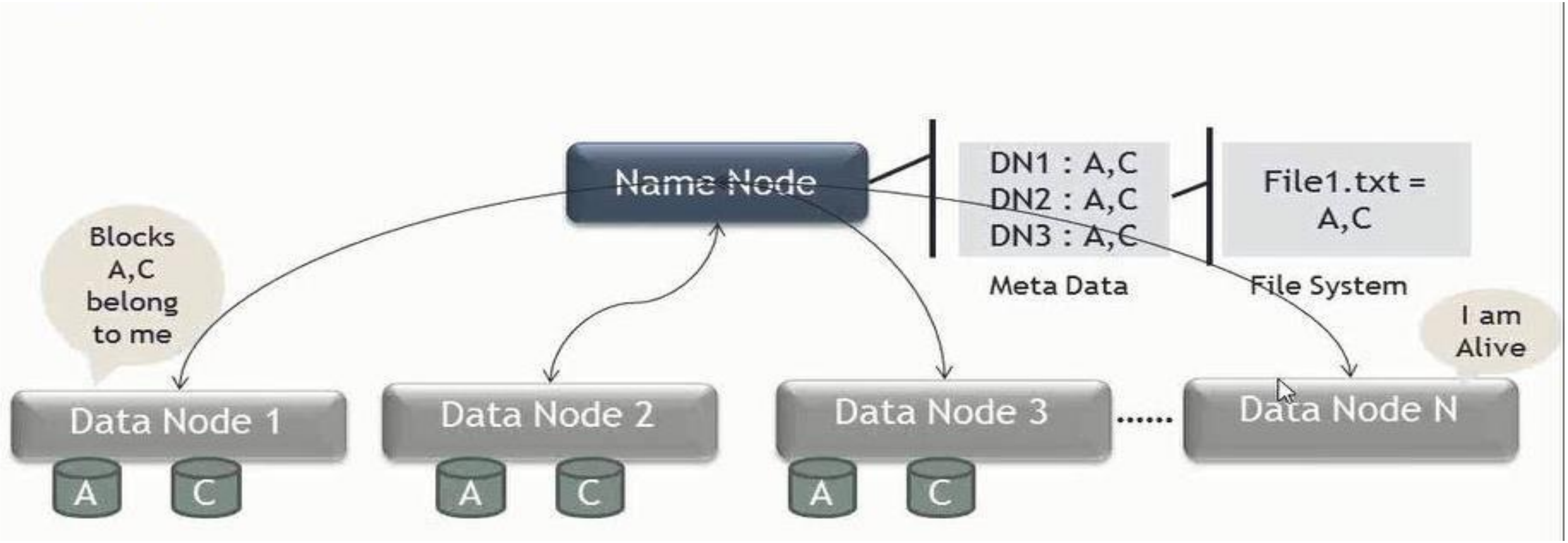
Name Node
Maintains mapping of file blocks to data node slaves

Resource Manager
Schedules jobs across node manager slaves

Data Node
Stores and serves blocks of data

Node Manager
Runs Tasks

HDFS Architecture



Thank you

Cloudlytics