Big Data

Spark YARN

Spark YARN

❖ Spark YARN Characteristics

- Spark is compatible with Hadoop by using YARN to run Hadoop clusters
- Spark can process data on any Hadoop input format
  - HDFS, HBase, Cassandra, Hive, etc.

Spark YARN

❖ Spark YARN Client Mode Example

1. Driver (Scalar or Python shell) program for the Application runs on the Client's computer (PC or Laptop)
2. YRM (YARN Resource Manager) selects a node (with sufficient resources) and makes a request to the node's YNM (YARN Node Manager) to setup a YAM (YARN App Master)
   - YRM's Apps Master starts to monitor the YAM

Spark YARN

❖ Spark YARN Client Mode Example

3. YAM makes a request to the YRM to have multiple YCs (YARN Containers) with Executors to be setup in the cluster
4. YRM informs the YAM which nodes have sufficient resources to setup YCs and sends the required information to the YAM to enable the setup process

Spark YARN

❖ Spark YARN Client Mode Example

5. YAM makes a request to the YNMs to setup YCs on the nodes
6. Inside each YC has an Executor is setup to process Tasks on the RDD partitions

Spark YARN

❖ Spark YARN Client Mode Example

7. Client can use the Driver to be directly interactive with the Executors in the YCs on the nodes
   - Client can check the results of the Transforms/Actions in the Executors in real-time using the Driver interface

Spark YARN

❖ Spark YARN Client Mode Example

8. Two Level Scheduler Control operations

- YRM Scheduler decides which nodes will have Containers to run the Executor JVMs
- Driver Scheduler decides the actual Task that will run on each Executors
  - Driver Scheduler attempts to optimize the Task assignment based on the Cached RDD dataset (e.g., HDFS Blocks) stored on each node

Spark YARN

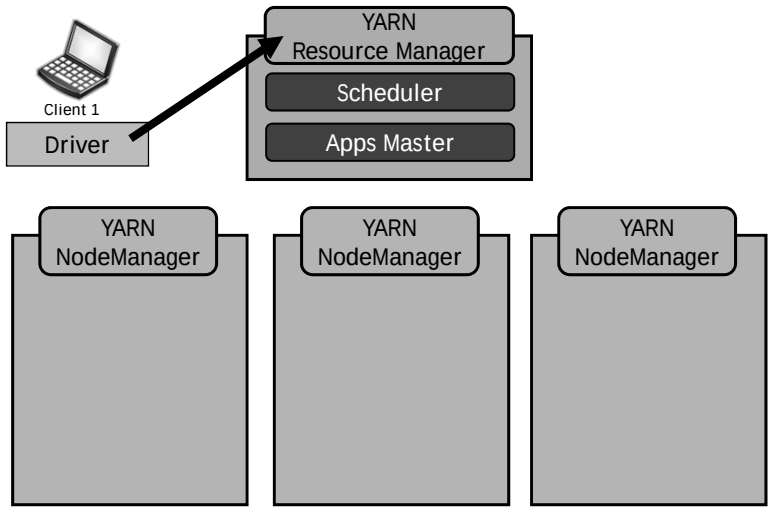
❖ Spark YARN Client Mode Example

9. Spark YARN Client mode process will End when a Driver program is closed

- Client's Driver shell is terminated
- Client's computer (PC or Laptop) is turned off or moves away

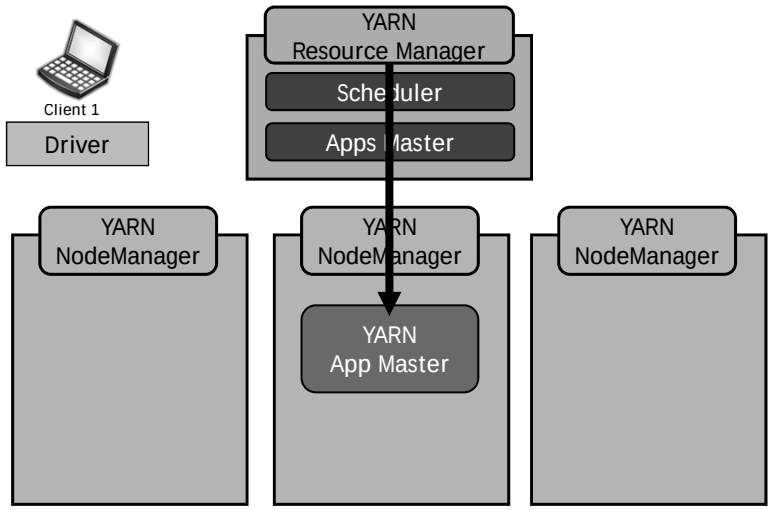
Spark YARN

❖ Spark YARN Client Mode Example



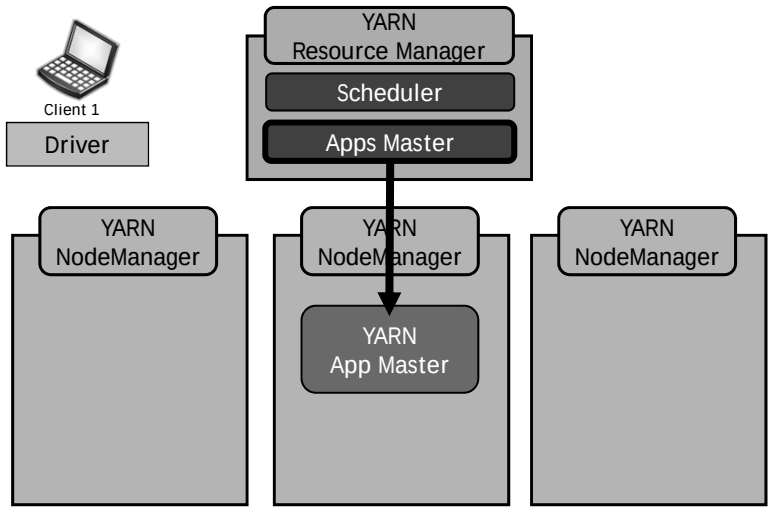
Spark YARN

❖ Spark YARN Client Mode Example



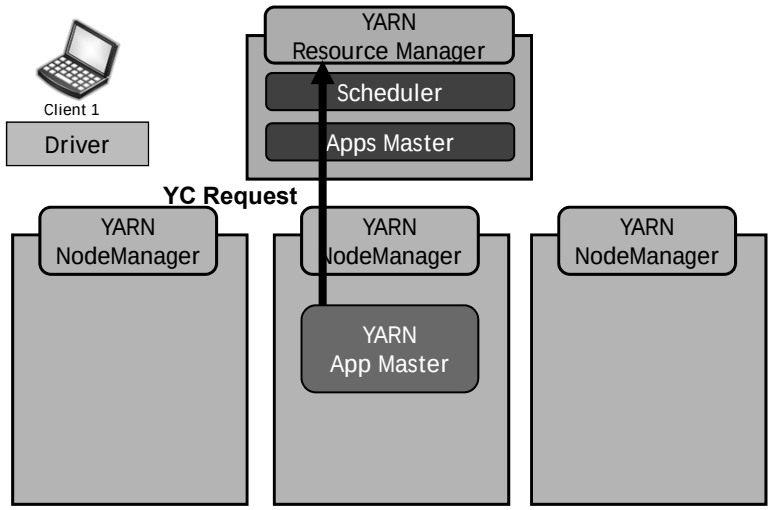
Spark YARN

❖ Spark YARN Client Mode Example



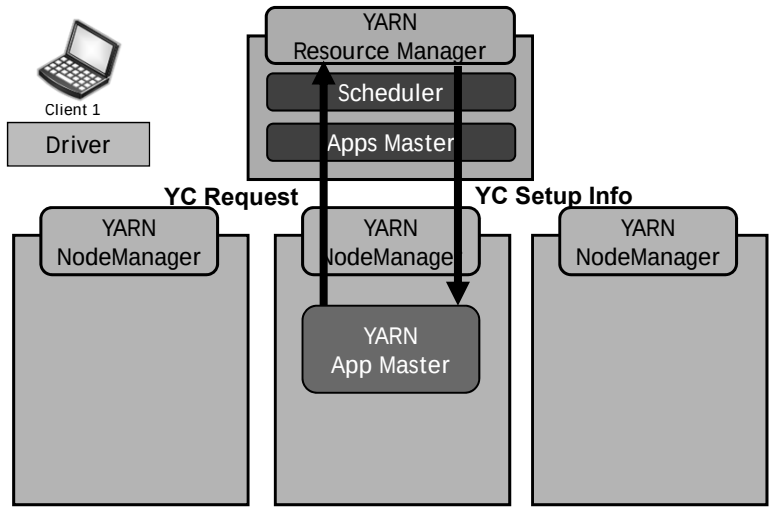
Spark YARN

❖ Spark YARN Client Mode Example



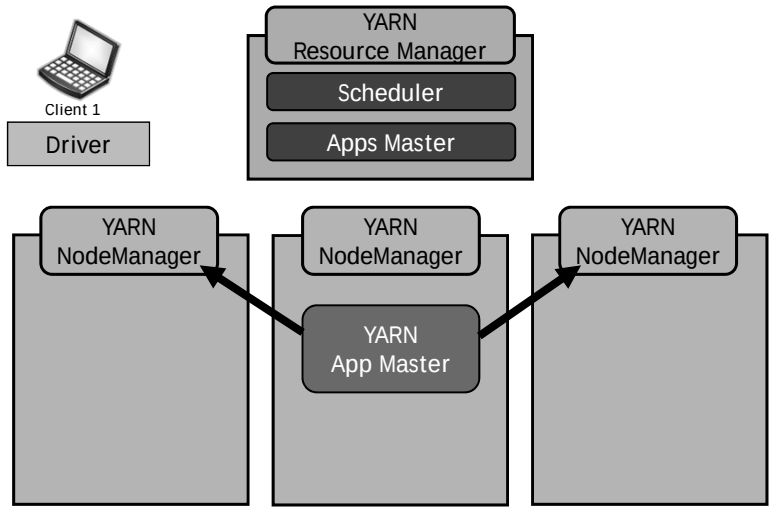
Spark YARN

❖ Spark YARN Client Mode Example



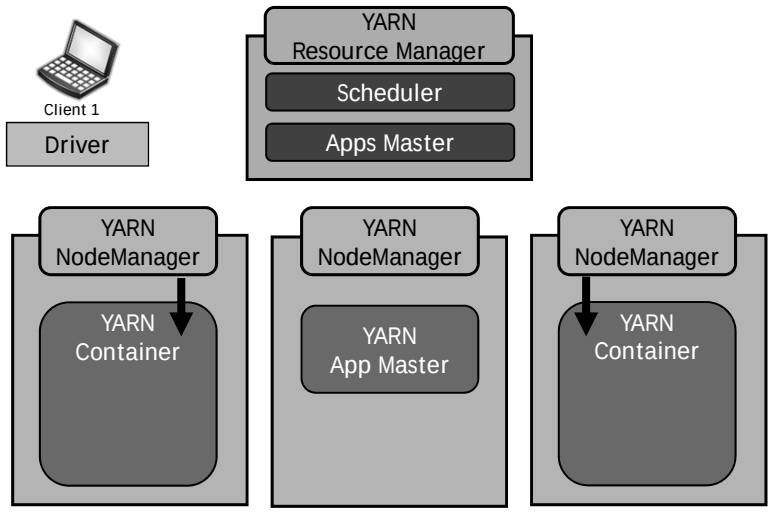
Spark YARN

❖ Spark YARN Client Mode Example



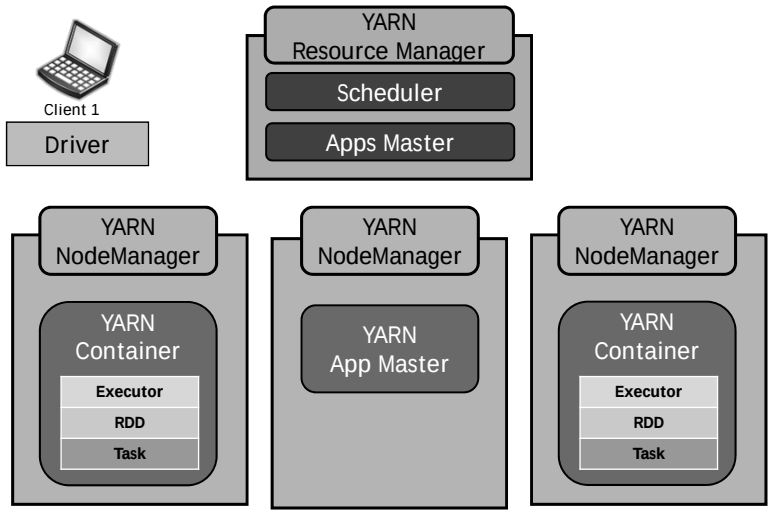
Spark YARN

❖ Spark YARN Client Mode Example



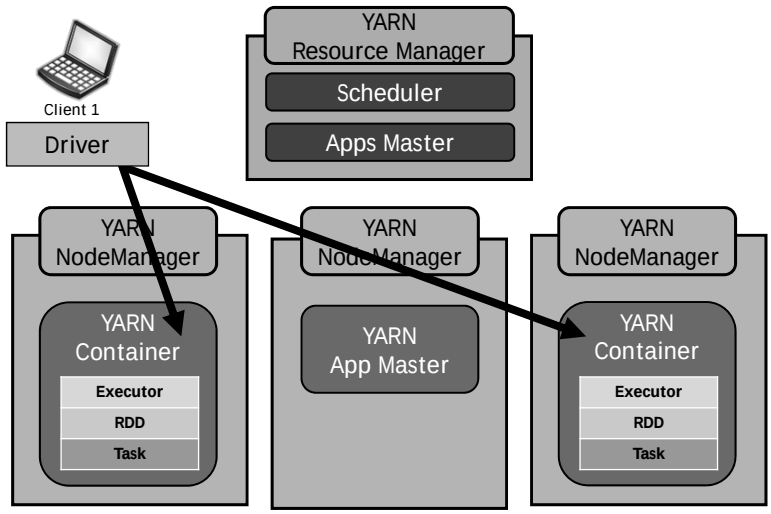
Spark YARN

❖ Spark YARN Client Mode Example



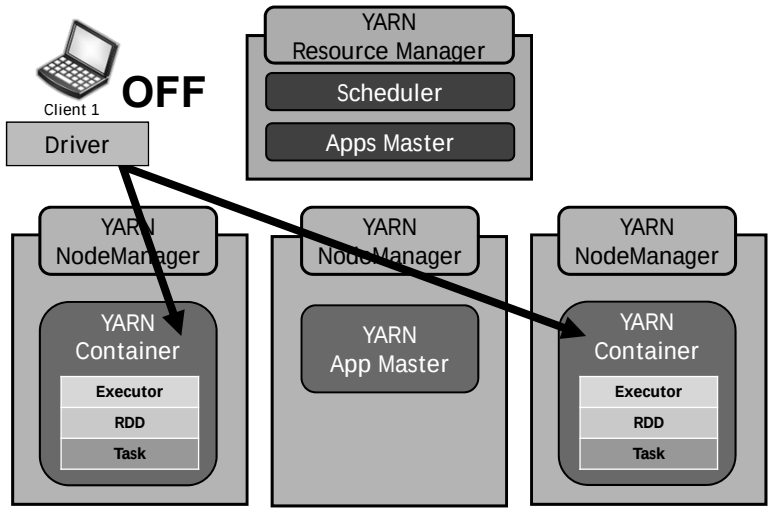
Spark YARN

❖ Spark YARN Client Mode Example



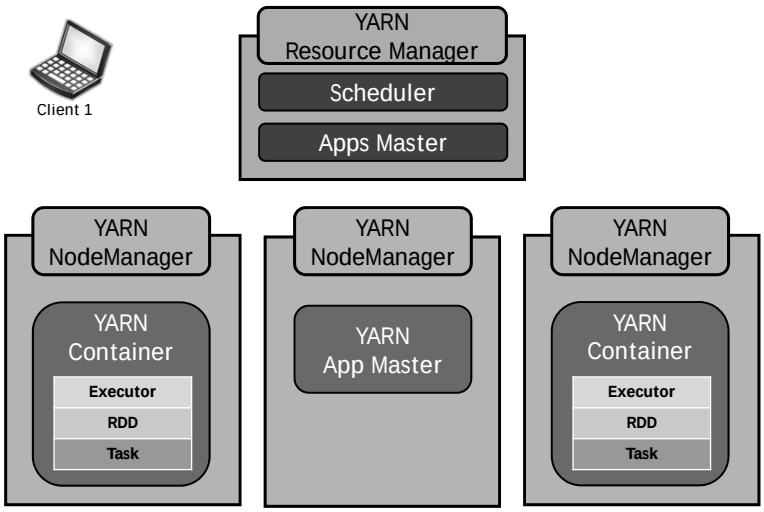
Spark YARN

❖ Spark YARN Client Mode Example



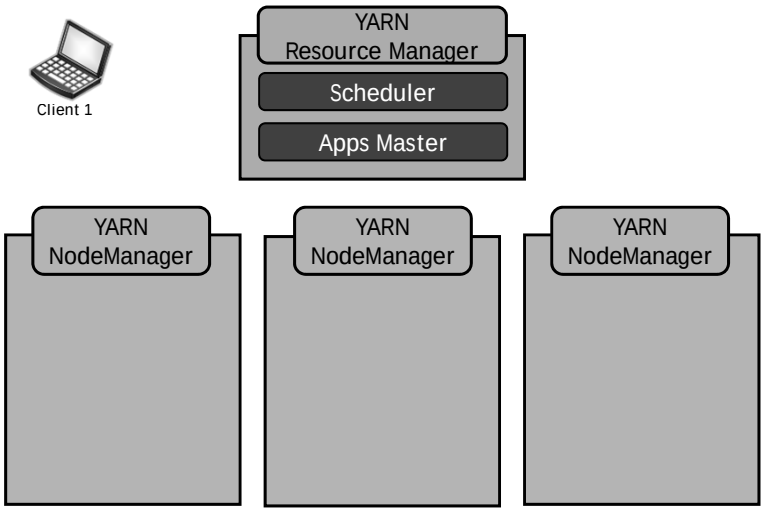
Spark YARN

❖ Spark YARN Client Mode Example



Spark YARN

❖ Spark YARN Client Mode Example



Spark YARN

❖ Spark YARN Cluster Mode Example

1. Client submits the App and the driver (e.g., Python script, JAR file) program to the cluster's RM
2. RM decides the location of the YAM (YARN App Master)
   - YRM's Apps Master starts to monitor the YAM
3. YAM starts to execute the Driver program internally

Spark YARN

❖ Spark YARN Cluster Mode Example

4. YAM makes a request to the RM to have multiple Containers (with Executors) to be setup in the cluster
5. RM informs the YAM which nodes have sufficient resources to setup Containers and sends the required information to the YAM to enable the setup process

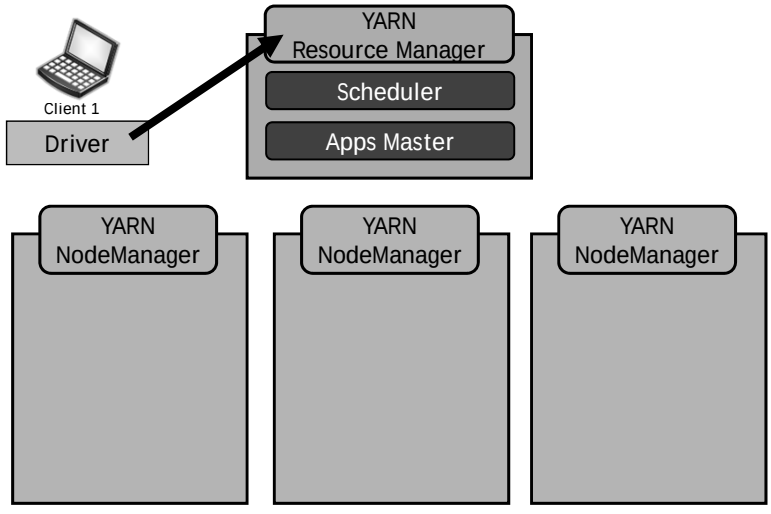
Spark YARN

❖ Spark YARN Cluster Mode Example

6. YAM makes a request to the NMs to setup Containers on their nodes
7. Inside each Container an Executor is setup to execute Tasks on the RDD partitions
8. The result (DAG's final RDD dataset) is saved to HDFS
9. Client can later check the results saved in the HDFS

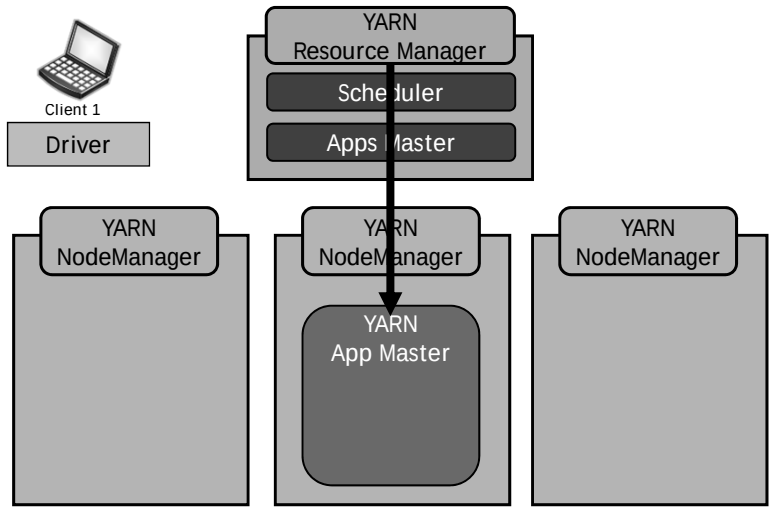
Spark YARN

❖ Spark YARN Cluster Mode Example



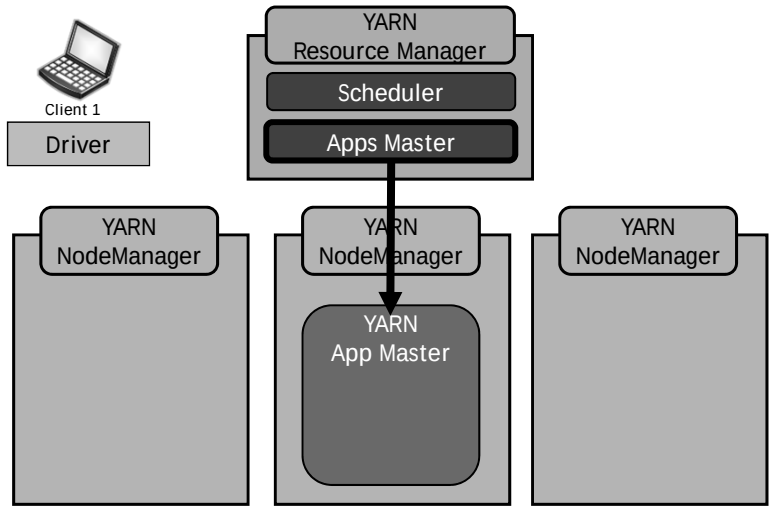
Spark YARN

❖ Spark YARN Cluster Mode Example



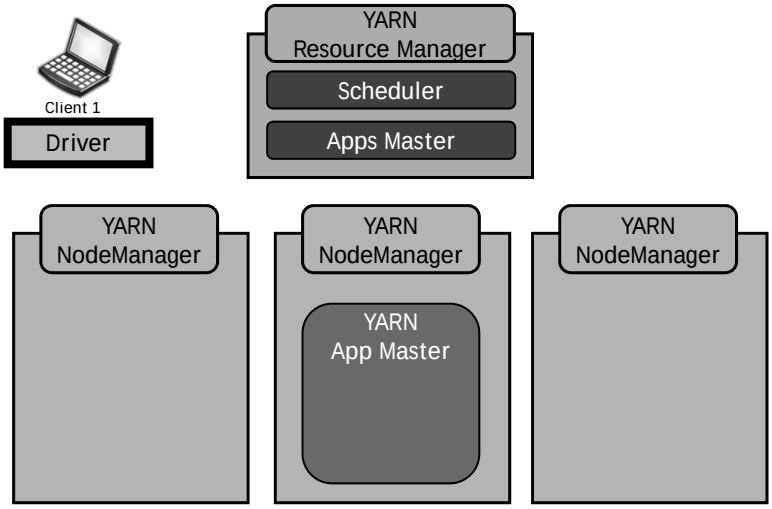
Spark YARN

❖ Spark YARN Cluster Mode Example



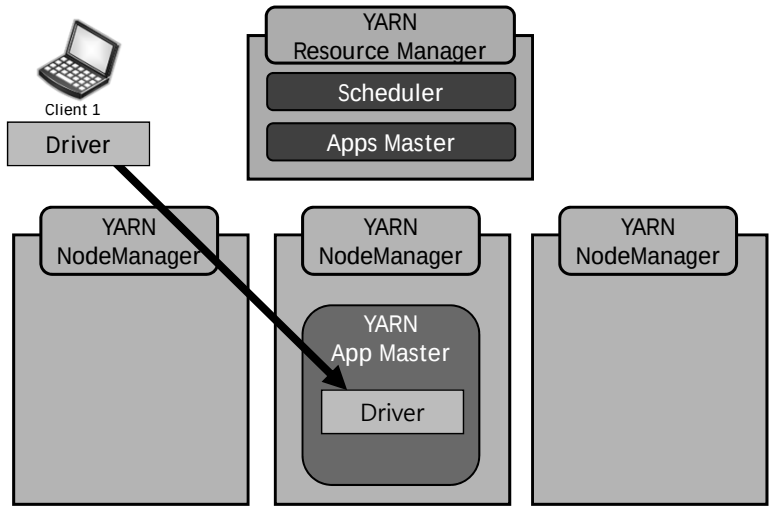
Spark YARN

❖ Spark YARN Cluster Mode Example



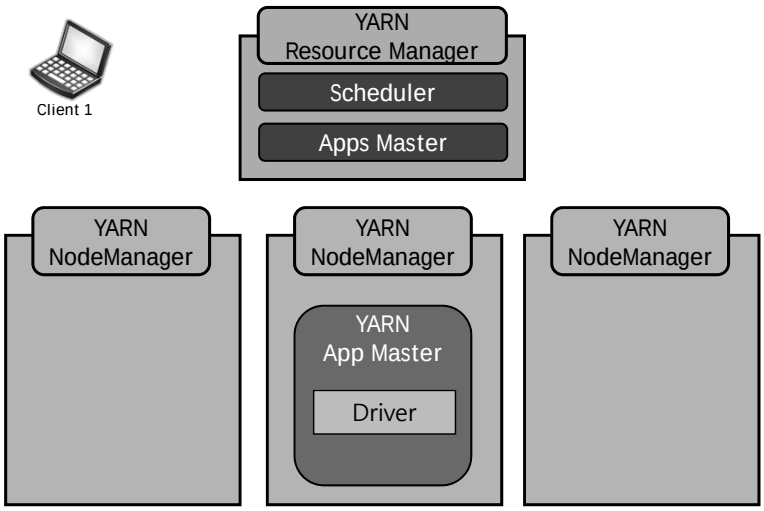
Spark YARN

❖ Spark YARN Cluster Mode Example



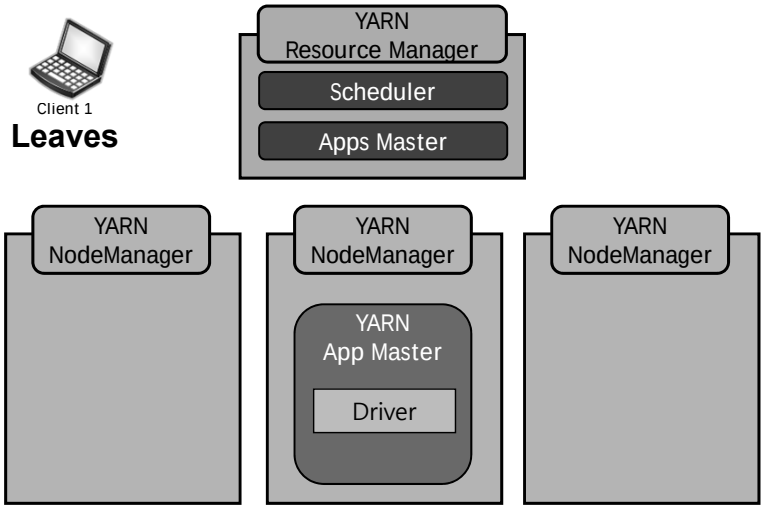
Spark YARN

❖ Spark YARN Cluster Mode Example



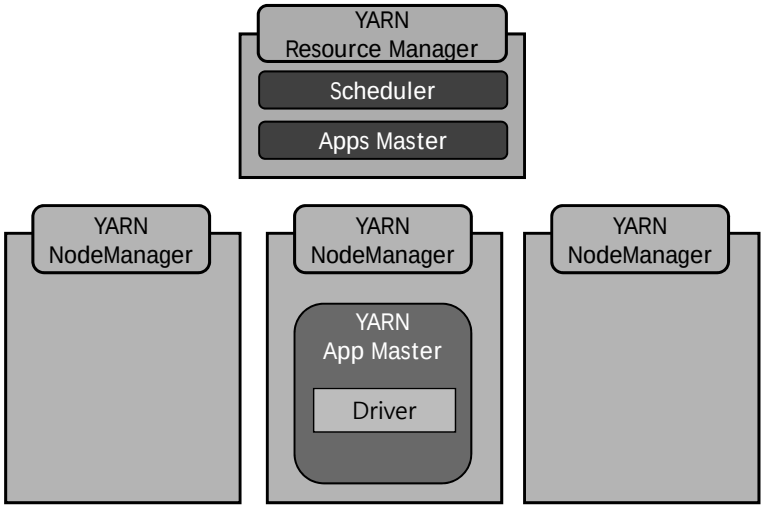
Spark YARN

❖ Spark YARN Cluster Mode Example



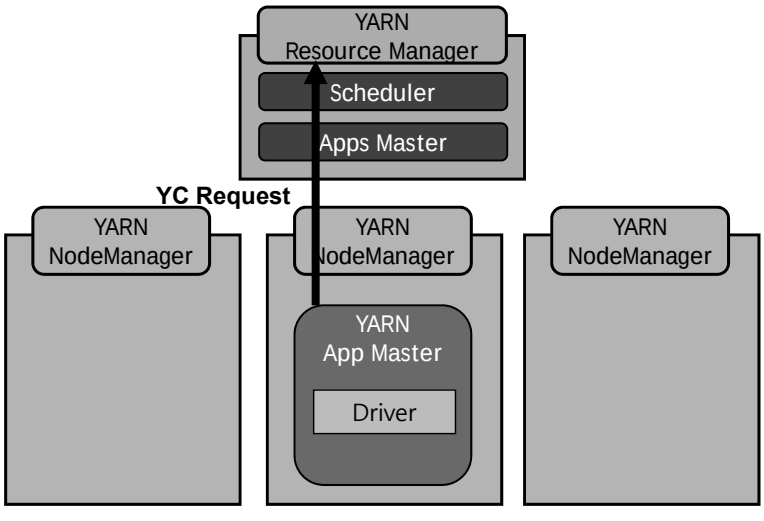
Spark YARN

❖ Spark YARN Cluster Mode Example



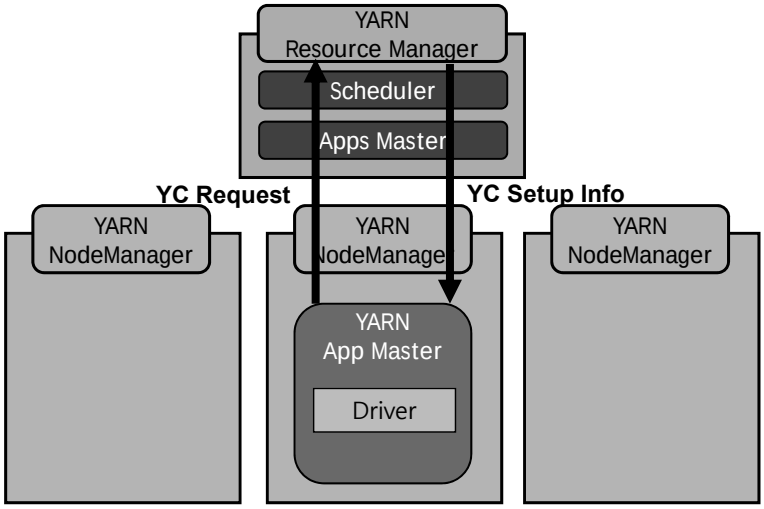
Spark YARN

❖ Spark YARN Cluster Mode Example



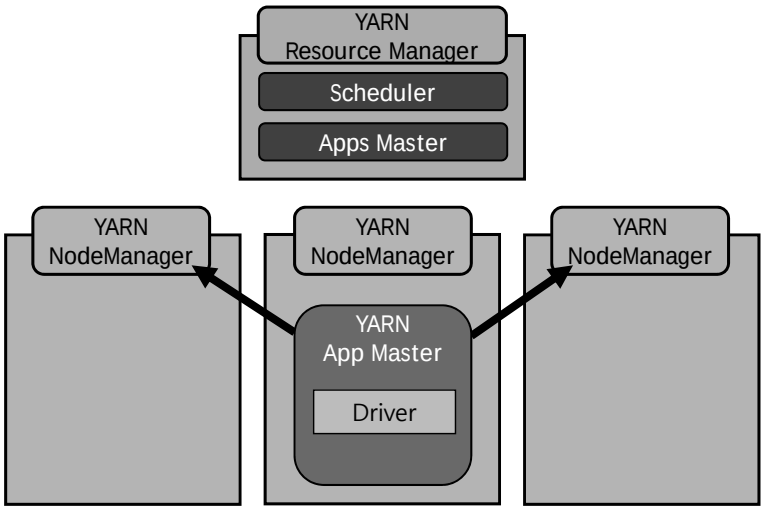
Spark YARN

❖ Spark YARN Cluster Mode Example



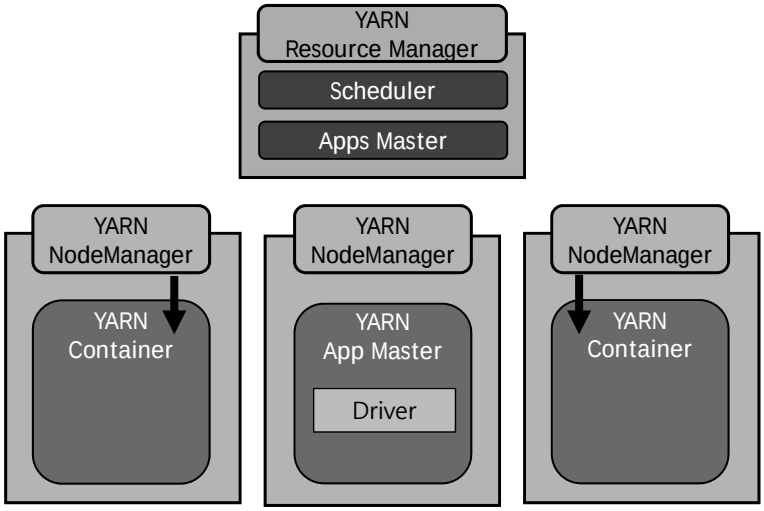
Spark YARN

❖ Spark YARN Cluster Mode Example



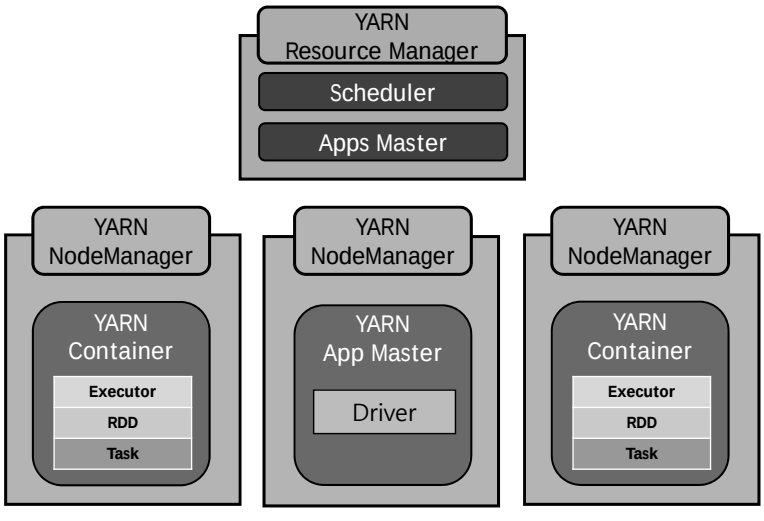
Spark YARN

❖ Spark YARN Cluster Mode Example



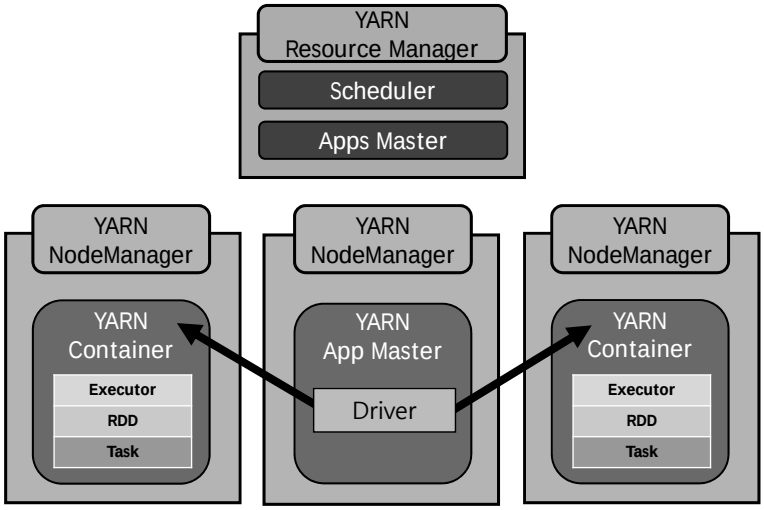
Spark YARN

❖ Spark YARN Cluster Mode Example



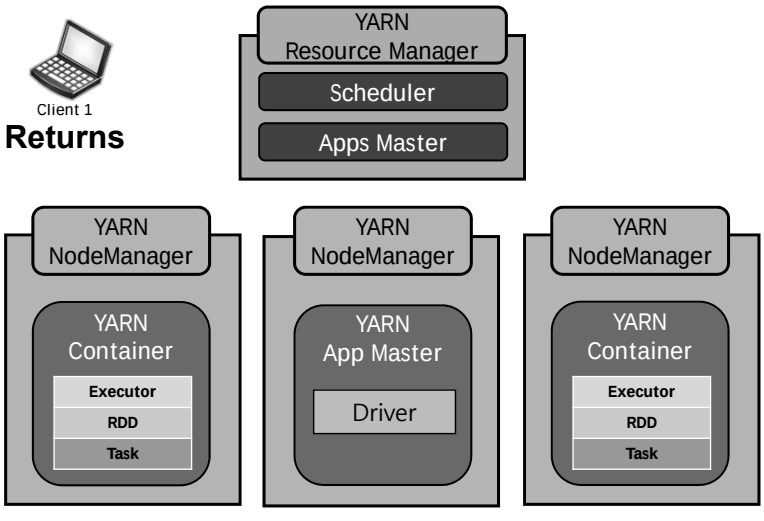
Spark YARN

❖ Spark YARN Cluster Mode Example



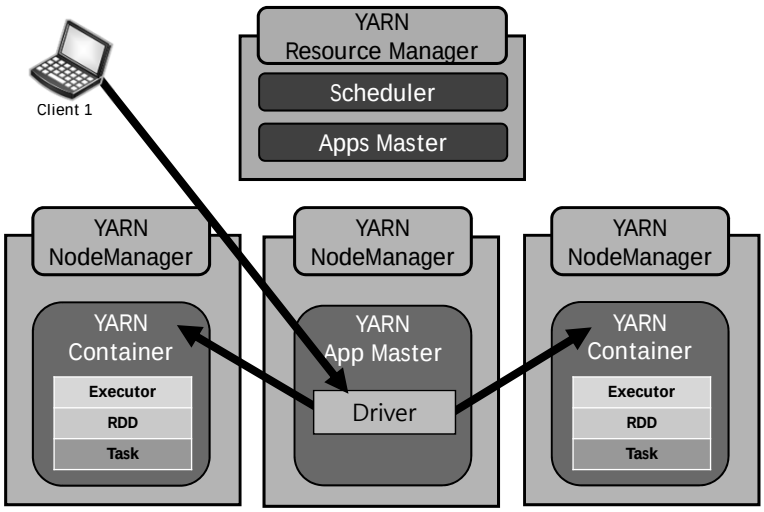
Spark YARN

❖ Spark YARN Cluster Mode Example



Spark YARN

❖ **Spark YARN Cluster Mode Example**



Big Data

References

References

- Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia, *Learning Spark: Lightning-Fast Big Data Analysis*. 1st Edition. O'Reilly, 2015.
- Sameer Farooqui, Databricks, **Advanced Apache Spark Training**, Devops Advanced Class, Spark Summit East 2015, http://slideshare.net/databricks, www.linkedin.com/in/blueplastic, March 2015.
- Apache Spark documents (all documents and tutorials were used)
  - http://spark.apache.org/docs/latest/rdd-programming-guide.html
  - http://spark.apache.org/docs/latest/rdd-programming-guide.html#working-with-key-value-pairs
  - https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html#rdd-persistence
- Wikipedia, www.wikipedia.org
- Stackoverflow, https://stackoverflow.com/questions
- Bernard Marr, "Spark Or Hadoop -- Which Is The Best Big Data Framework?," Forbes, Tech, June 22, 2015.
- Quick introduction to Apache Spark, https://www.youtube.com/watch?v=TgiBvKcGL24
- Wide vs Narrow Dependencies, https://github.com/rohgar/scala-spark-4/wiki/Wide-vs-Narrow-Dependencies

References

- Partitions and Partitioning, https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-rdd-partitions.html
- Neo4j, "From Relational to Neo4j," https://neo4j.com/developer/graph-db-vs-rdbms/ (last accessed Jan. 1, 2018).

Image Sources

- By Robivy64 at English Wikipedia [Public domain], via Wikimedia Commons
- Teravolt at English Wikipedia [CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons
- By Konradr (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/)], via Wikimedia Commons