BASIC SPARK CONFIGURATION & SETUP

CLASS-V

Instructor: Palash Gupta

Re-Cap

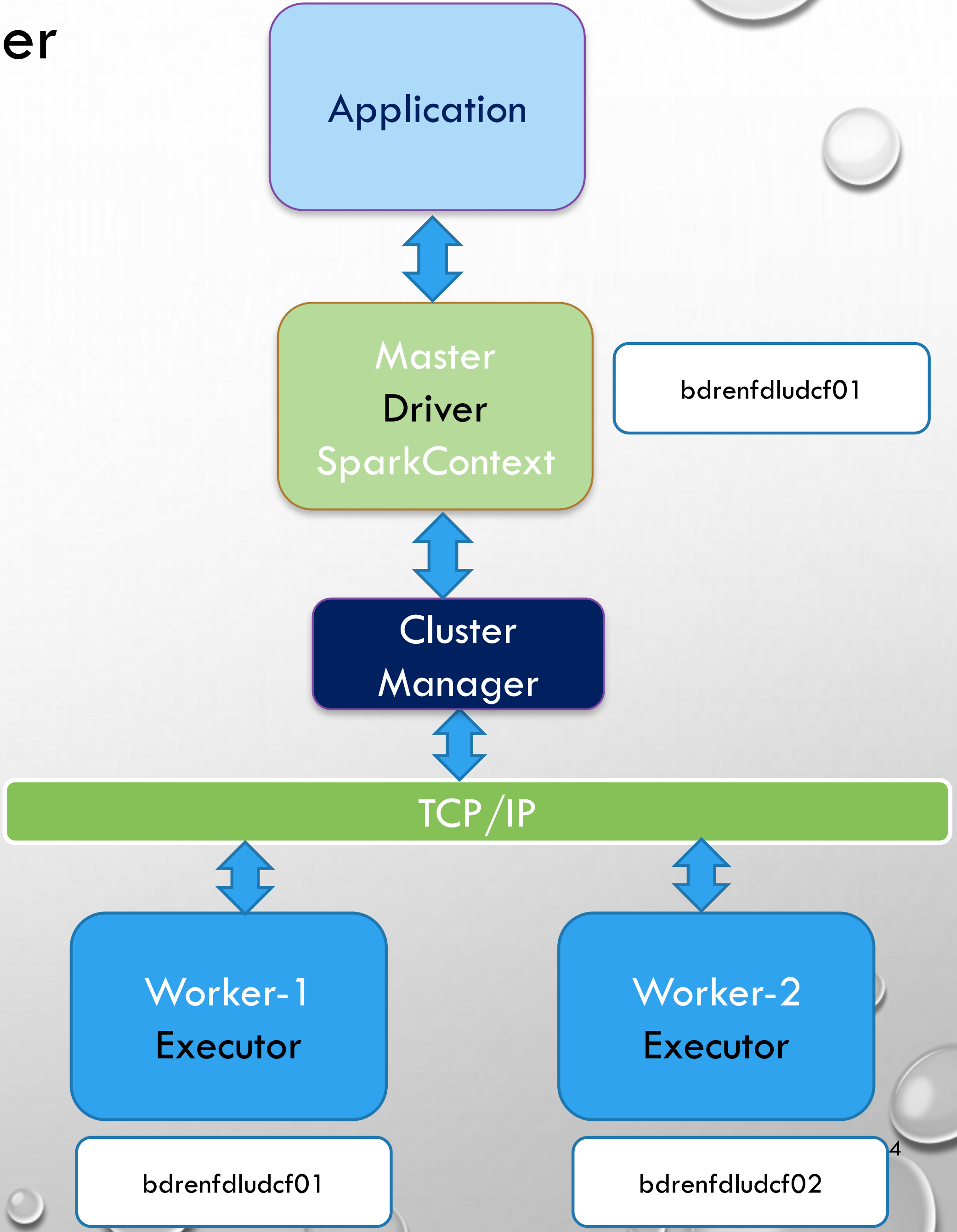
- Introduction to Apache Spark
- Why Spark where we have Hadoop?
- Spark Architecture
- Introduction to Spark Component
- Introduction to Spark RDD, Dataset, DataFrame and DAG
- Understanding Spark Execution Model

What we are going to Cover today?

- Design a Spark Cluster
- Key Procedure Setting up a cluster
- Configuring a Cluster
- Basic Administration
- Under practically how spark works

Design a Hadoop Spark Cluster

SN	Hostname	SSH Port	IP	Operating System	vCPU	vRAM	vHDD
1	bdrenfdludcf01	22	192.168.0.104	Cent OS 7 64 bit	**2**	1 GB	25 GB
2	bdrenfdludcf02	22	192.168.0.105	Cent OS 7 64 bit	**2**	1 GB	25 GB



Key Procedure Configuring a Cluster Contd.

SN	Step	Impacted Nodes	Type	Remarks
1	Know your machine resources & their IP and credential	All	Mandatory	We will use two hosts.
2	Configure Hostname	All	Mandatory	Naming convention is preferable.
3	Create a Repository	All	Optional	
4	Check & fix your date time	All	Mandatory	
5	Setup a NTP	All	Optional	
6	Stop Firewall	All	Mandatory	This is only for our Lab, not for a production system.
7	Step SELINUX	All	Mandatory	This is only for our Lab, not for a production system.
8	Create hadoop User	All	Mandatory	
9	Create Password less login within cluster nodes	All	Mandatory	

Key Procedure Configuring a Cluster Contd.

SN	Step	Impacted Nodes	Type	Remarks
10	Install targeted Java version	All	Mandatory	
11	Install Python & update hadoop user environment	All	Mandatory	
Start Hadoop if we need to interact with HDFS				
Start Main Spark Installation				
1	Upload Spark Binaries & Unzip	All	Mandatory	
2	Configure spark-env.sh	All	Mandatory	

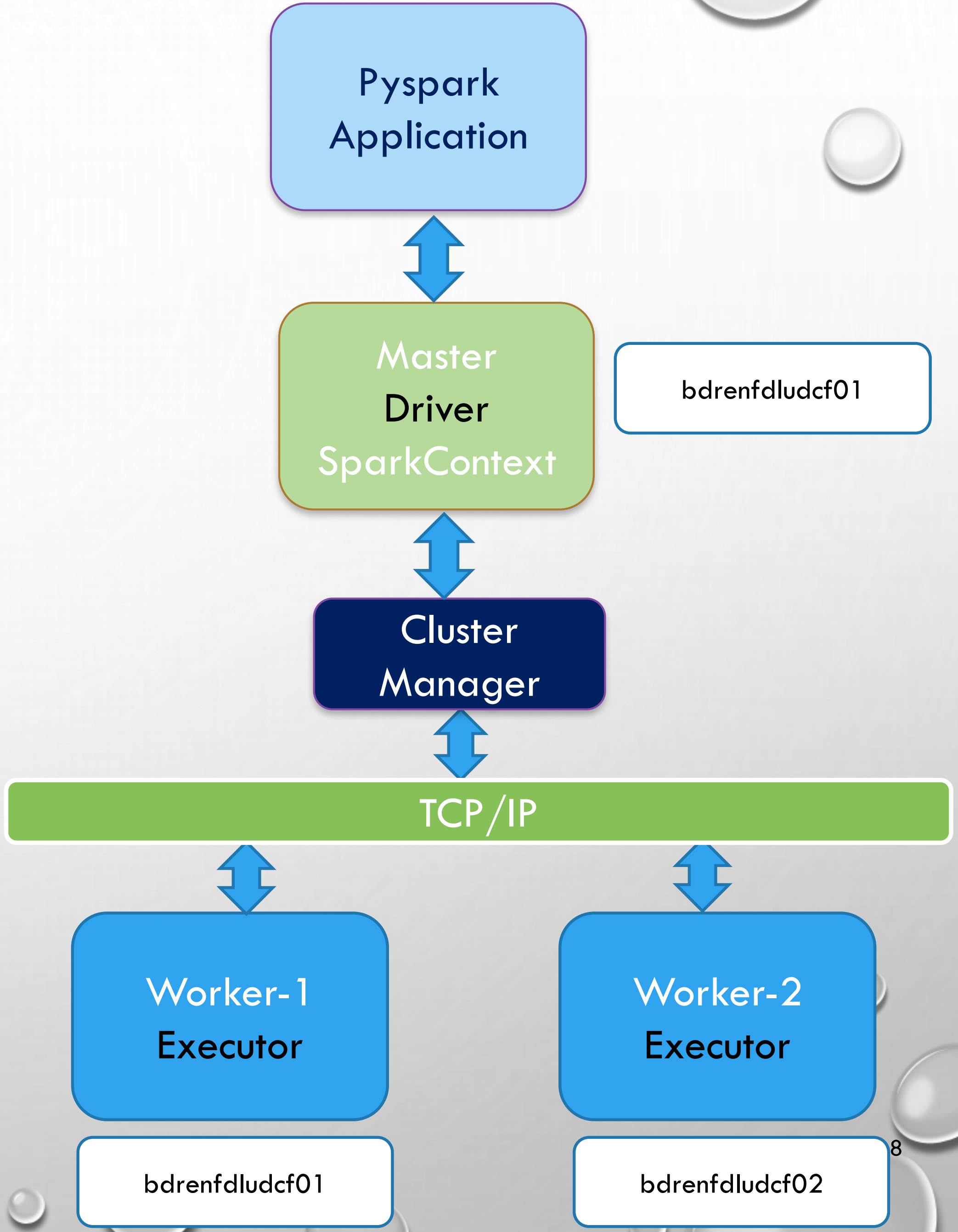
Key Procedure Configuring a Cluster Contd.

SN	Step	Impacted Nodes	Type	Remarks
4	Configure spark-defaults.conf properties	All	Mandatory	
5	Configure slaves to define worker node	All	Mandatory	
6	Creating Spark Events Folders	All	Mandatory	
7	Change ownership of hadoop Directory	All	Mandatory	
7	Start Spark Cluster (Master and Worker)	Master	Mandatory	

Configuring a Spark Cluster

❑ **Assumptions**

- ○ One Master
- ○ Two Workers
- ○ Both Master and Workers will have minimum two vCPU(s)
- ○ Will configure first host as Master
- ○ Will configure both hosts as Workers
- ○ Spark Master Port is 7077
- ○ Spark Web UI Port is 9999
- ○ Spark Application Port will be 4040,4041 and so on.
- ○ Will Follow a Method of Procedure
- ○ Will do RDD, Data frame Exercise in interactive mode
- ○ Will see how sparks works from Web GUI



Basic Administration

SN	Topic	Command Syntax
1	Start and Stop Spark Cluster	$/usr/local/spark/sbin/start-all.sh $/usr/local/spark/sbin/stop-all.sh
2	Start and Stop Master	$/usr/local/spark/sbin/start-mater.sh $/usr/local/spark/sbin/stop-master.sh
3	Start and Stop Worker	$/usr/local/spark/sbin/start-slave.sh $/usr/local/spark/sbin/stop-slave.sh

HDFS Reference Command

SN	Topic	Command Syntax
1	Start and Stop HDFS	$HADOOP_HOME/sbin/start-dfs.sh HADOOP_HOME/sbin/stop-dfs.sh
2	Browsing HDFS	$hdfs dfs –ls / $hadoop fs –ls /
3	Putting a file into HDFS	$hdfs dfs –copyFromLocal my_file.txt / $hadoop fs –put my_file.txt /
4	Getting a file from HDFS	$hdfs dfs –copyToLocal /my_file.txt /home/hadoop $hadoop fs –get /my_file.txt /home/Hadoop
5	Removing a file from HDFS	$hdfs dfs -rm /user/my_file $Hadoop fs –rm /user/my_file
6	Overwriting a file in HDFS	$hadoop fs –put –f my_file.txt /
7	Append to a file	$echo "Line-to-add" | hdfs dfs -appendToFile - /my_file.txt
8	View a file	$hdfs dfs -cat /mydir/mysecfile.log

QUESTION & ANSWER

THANKS FOR ATTENDING THE CLASS & YOUR CO-OPERATION

References

- https://spark.apache.org/docs/2.3.0/sql-programming-guide.html
- https://spark.apache.org/docs/latest/configuration.html
- https://spark.apache.org/
- https://spark.apache.org/docs/2.1.0/api/python/pyspark.html
- https://stackoverflow.com/questions/31610971/spark-repartition-vs-coalesce
- https://spark.apache.org/docs/latest/spark-standalone.html