Data Science Lifecycle

Business Understanding, Data Engineering (Data Mining, Data Cleaning, Data Exploration and Feature Engineering), Predictive Modelling, Data Visualization

Data Size

in memory, memory io, disk io,

Data Distribution

cluster
horizontal scaling, vertical scaling, data volume,
distributed processing, parallel processing

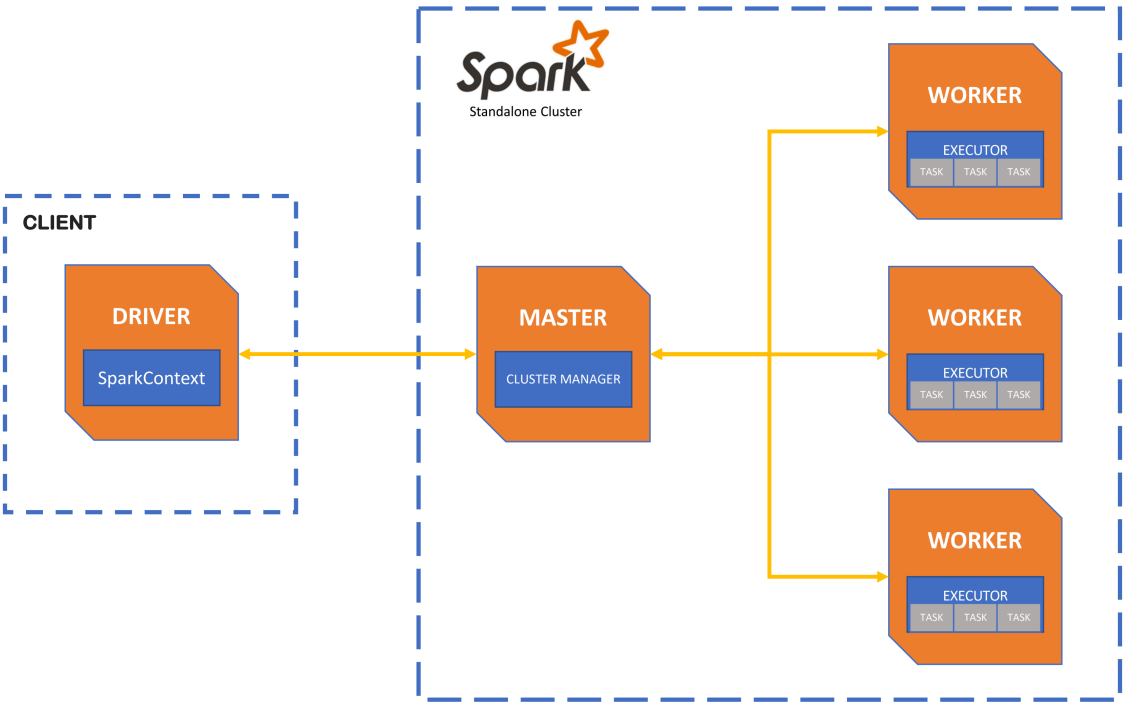
framework vs library

Hadoop
Spark

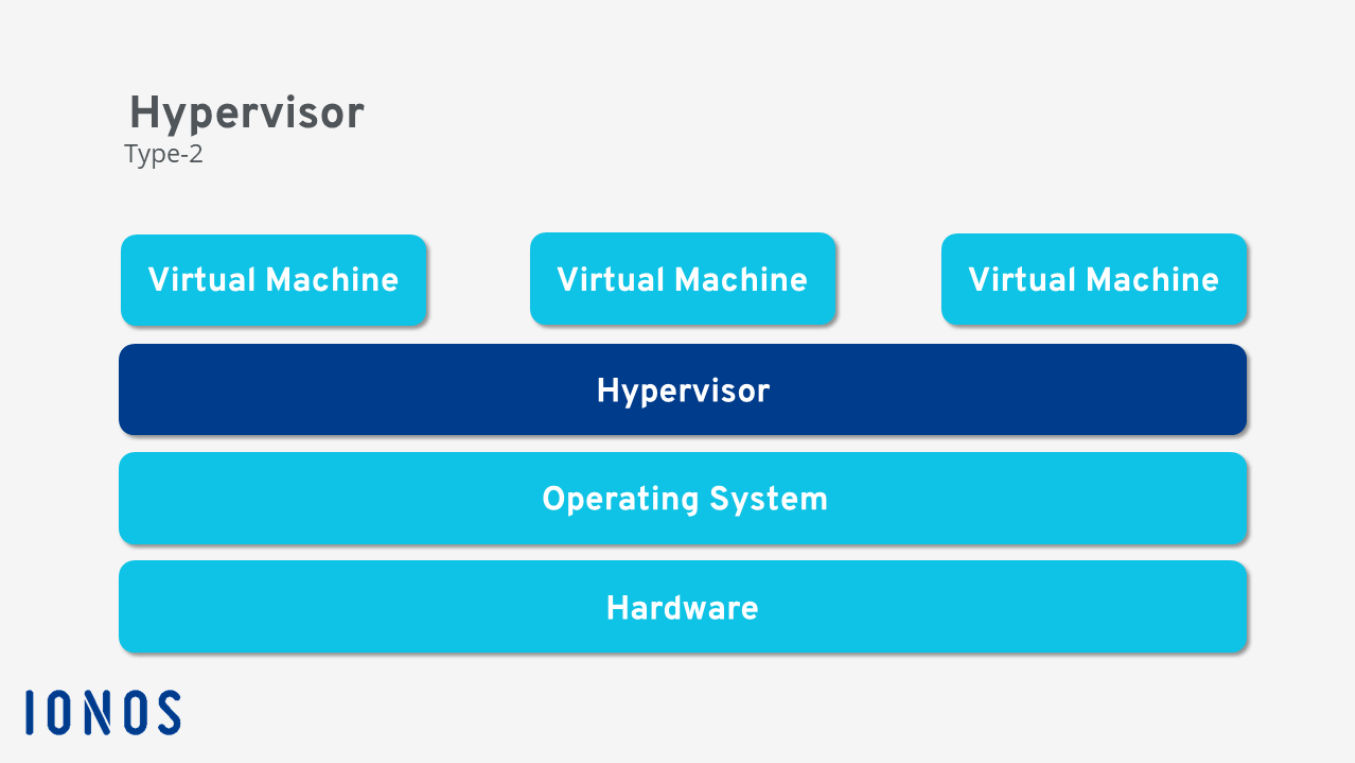
Spark

distributed compute
driver <-> cluster manager <-> node
process should work at data block itself
network bandwidth
cluster based environment

driver <-> cluster manager <-> node



Host OS
Guest OS
Mount Point



Streaming
Spark REPL

Spark -V 2.4.6
Python -V 3.6

File Format

csv
xml
json
avro
parquet file - columnar file format

java > scala > spark > pyspark
jps -> java processing stack

Big Data

3Vs: Volume, Variety, and Velocity

clustered computing, parallel computing, distributed computing, batch processing, real-time processing

pyspark -> batch processing

MapReduce

Batch Processing

data is divided into multiple blocks and process is also divided into blocks

Batch Processing System

Hadoop/MapReduce

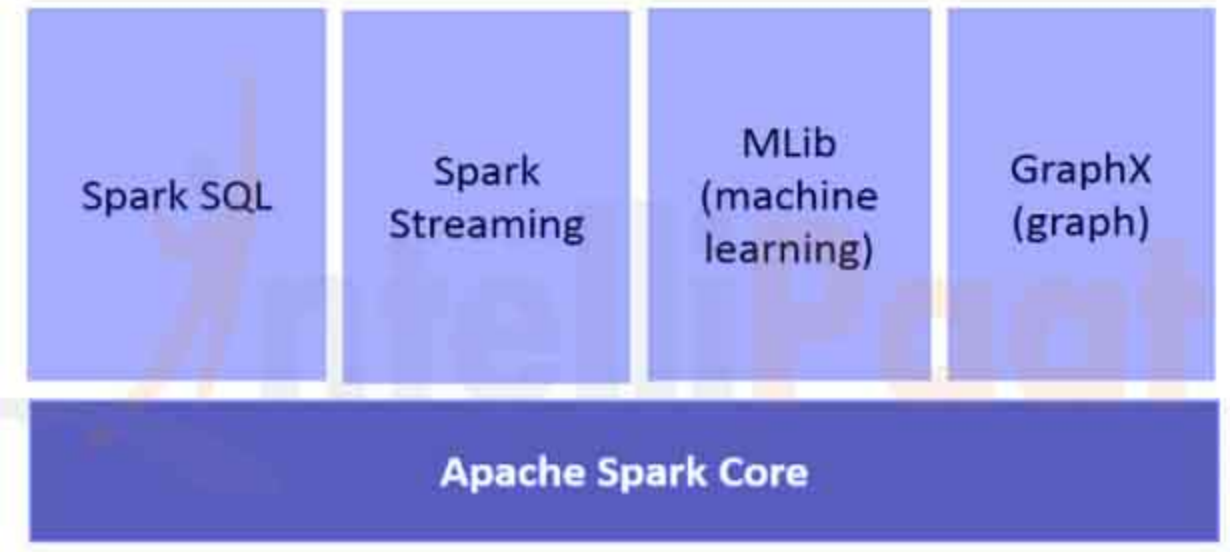
Apache Spark

general purpose cluster based

Features of Apache Spark Framework

distributed cluster computing, in memory, fast data processing, language -> java, scala, python, R, SQL

Components



Deployment Modes

local modes, cluster mode,
spark-shell, pyspark, sparkR

Shell

REPL - Read Evaluate Print Loop

spark-shell - Scala

```
spark-shell
```

```
# constant
val a = 2

# variable
var b = 3

a + b
```

pyspark - Python

```
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
```

```
unset PYSPARK_DRIVER_PYTHON
unset PYSPARK_DRIVER_PYTHON_OPTS
```

```
file unset_jupyter.sh
# unset_jupyter.sh: ASCII text

## After adding shebang statement
#!/bin/bash
file unset_jupyter.sh
#unset_jupyter.sh: Bourne-Again shell script, ASCII text executable

chmod 744 unset_jupyter.sh

source ./unset_jupyter.sh

echo $PYSPARK_DRIVER_PYTHON
echo $PYSPARK_DRIVER_PYTHON_OPTS
```

```
# unset an pwd
#!/bin/bash

unset PYSPARK_DRIVER_PYTHON
unset PYSPARK_DRIVER_PYTHON_OPTS

source ./unset_jupyter.sh
```

```
pyspark
```

```
a = 1
b = 2
a + b
```

Spark Context

entry point to spark cluster
default spark context -> sc

```
print(type(sc))
# <class 'pyspark.context.SparkContext'>
id(sc) # Memory Location
# 140269501754728
```

```
sc.version
# '2.4.5'
sc.pythonVer
# '3.6'
sc.master
# 'local[*]'

rdd = sc.parallelize([1,2,3,4,5])
print(type(rdd))
# <class 'pyspark.rdd.RDD'>

rdd2 = sc.textFile("test.txt") # residing in hdfs
```

Loading data in PySpark

file protocol

file://

s3://

```
rdd = sc.parallelize([1,2,3,4,5]) # load data from local python collection
print(type(rdd))
# <class 'pyspark.rdd.RDD'>

# file protocol
rdd2 = sc.textFile("test.txt") # residing in hdfs
rdd2 = sc.textFile("file:///home/talentum/test.txt") # residing in local
filesystem
rdd2 = sc.textFile("s3://test.txt") # residing in aws s3

rdd2 = sc.textFile("file:///home/talentum/spark/README.md")
rdd2.count()
# 104

rdd2.take(3)
# ['# Apache Spark', '', 'Spark is a fast and general cluster computing system
for Big Data. It provides']

rdd2.collect() # Print all data
```

```
talentum@talentum-virtual-machine:~/spark$ wc -l README.md
104 README.md
```

```
talentum@talentum-virtual-machine:~/spark$ head -n 3 README.md
# Apache Spark

Spark is a fast and general cluster computing system for Big Data. It provides
```

Documentation

https://spark.apache.org/docs/latest/api/python/index.html

https://archive.apache.org/dist/spark/docs/2.4.5/api/python/index.html

Linux Basics

file or directory

d means directorry

`-` means file