

DAY-11 (Spark Assignment - 1)

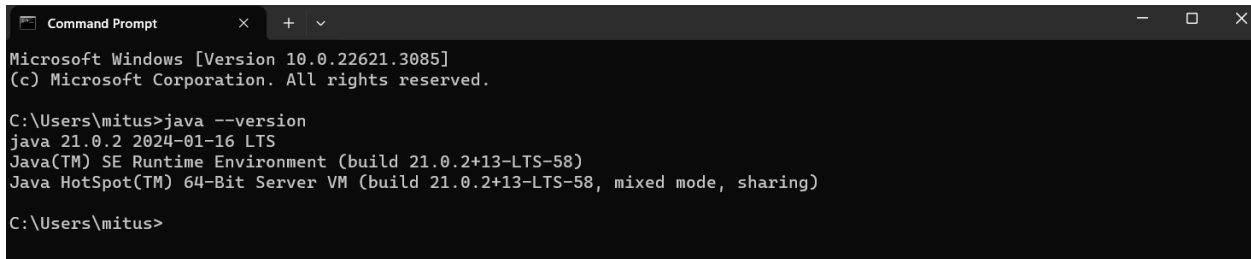
Mitushi Vishwakarma

INSTALLATION of Apache Spark

Apache Spark is an open-source distributed computing system that provides a fast and general-purpose cluster-computing framework for big data processing. PySpark is the Python API for Apache Spark, allowing you to write Spark applications using Python.

Step 1 : Install JDK

To check the installation, type `java -version` in cmd prompt. It will display the version installed on your system. Install it in `C:\java\jdk\` folder.



```
Command Prompt
Microsoft Windows [Version 10.0.22621.3085]
(c) Microsoft Corporation. All rights reserved.

C:\Users\mitus>java --version
java 21.0.2 2024-01-16 LTS
Java(TM) SE Runtime Environment (build 21.0.2+13-LTS-58)
Java HotSpot(TM) 64-Bit Server VM (build 21.0.2+13-LTS-58, mixed mode, sharing)

C:\Users\mitus>
```

Step 2 : Install Python

To check the installation, type `python --version` in cmd prompt. It will display the version installed on your system.

```
C:\Users\mitus>python --version
Python 3.10.13
```

Step 3 : Download Apache spark

Go to <https://spark.apache.org/downloads.html> and download the latest version for spark. The .tar file be downloaded and extract that in your system inside spark folder in C: drive. After this download winutils file for your Hadoop version from here <https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/winutils.exe> in the Hadoop folder in C: drive.



Download Apache Spark™

1. Choose a Spark release: 3.5.0 (Sep 13 2023) ▾
2. Choose a package type: Pre-built for Apache Hadoop 3.3 and later ▾
3. Download Spark: [spark-3.5.0-bin-hadoop3.tgz](#)
4. Verify this release using the 3.5.0 [signatures](#), [checksums](#) and [project release KEYS](#) by following these [procedures](#).

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

Link with Spark

Spark artifacts are [hosted in Maven Central](#). You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.12
version: 3.5.0
```

Step 4 : Setting up environment variables

Environment Variables

User variables for Mitushi

Variable	Value
HADOOP_HOME	C:\Hadoop
JAVA_HOME	C:\java\jdk
OneDrive	C:\Users\mitus\OneDrive
OneDriveConsumer	C:\Users\mitus\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin;C:\Users\mitus\...
PyCharm Community Editi...	C:\Program Files\JetBrains\PyCharm Community Edition 2022...
SPARK_HOME	C:\spark\spark-3.5.0-bin-hadoop3
TEMP	C:\Users\mitus\AppData\Local\Temp

New...

Edit...

Delete

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	16
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIN...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTU...	AMD64
PROCESSOR_IDENTIFIER	AMD64

New...

Edit...

Delete

OK

Cancel

Successfully installed spark. A spark session is created.

```
Microsoft Windows [Version 10.0.22621.3085]
(c) Microsoft Corporation. All rights reserved.

C:\Users\mitus>spark-submit --version
Welcome to

  /---/  ---/  /---/  /---/
 /  \ /  \  /  \ /  \ /  \
/---/  /---/  /---/  /---/  version 3.5.0
/_/_/  /---/  /---/  /---/

Using Scala version 2.12.18, Java HotSpot(TM) 64-Bit Server VM, 21.0.2
Branch HEAD
Compiled by user ubuntu on 2023-09-09T01:53:20Z
Revision ce5ddad990373636e94071e7cef2f31021add07b
Url https://github.com/apache/spark
Type --help for more information.

C:\Users\mitus>
```


Notes :

Apache Spark

Components :-

Spark SQL
Spark Streaming

ML Librai
GraphX, Spark

Sparks run on :-

standalone
Hadoop YARN
Mesos.

Features -

Spark → written in Scala
runs in JVM.

API : Scala, Java, Py, R.

Data Sources :- SQL, Local file sy, S3

Interactive shell → Scala, Python.

Components :-

Spark SQL, Spark Streaming, ML lib, GraphX
↑ Spark Core ↓

Apache Spark :-

open source data-processing engine for large datasets.

In real-time, databases are extremely large.

Data processing speed depends on :-

- processor
- RAM
- Storage devices.
- softwares.

MOBILE



PHONES



NAME, ADDRESS & E-MAIL



(1)
Data servers are the machines stored physically at data centres.

Apache Spark

→ general purpose cluster computing system.

~~xxxxxx~~

No GUI in Spark.

Apache Spark Core

→ runs query parallelly

→ parallel and distributed processing of huge dataset.

Spark SQL

→ data sources are Hive, Avro, Parquet, ORC, JSON, JDBC

→ distributed framework for structured data processing.

In

→ works to access structured and semi-structured information

→ Using Spark SQL, Spark gets more info. of data

MOBILE



PHONES



NAME, ADDRESS & E-MAIL

Spark Streaming :-

- add-on to core Spark.
- 3 phases of Spark Streaming

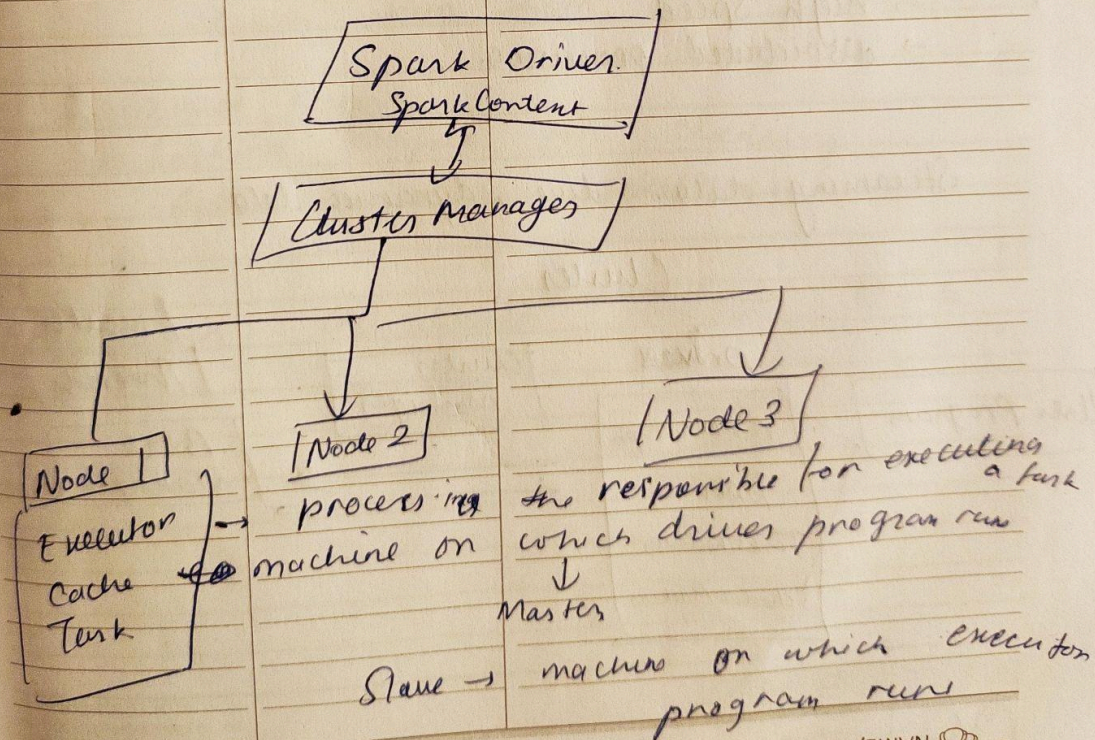
1. Gathering.
2. Processing.
3. Data Storage.

processed data pushed out to file systems, databases,

1. Gathering.

Ⓑ Collecting data
Basic sources
Advanced

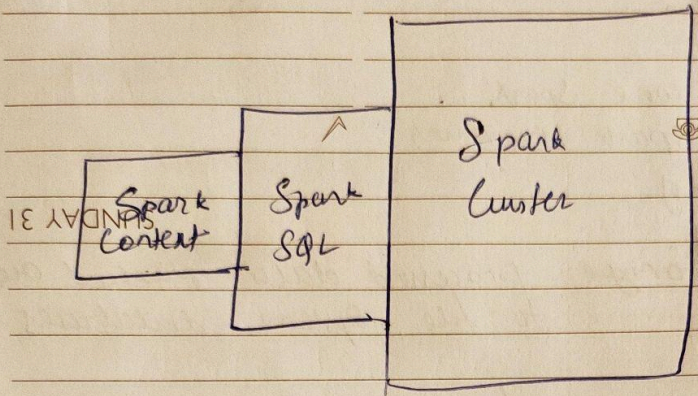
from sources
file system
Kafka flume



NAME, ADDRESS & E-MAIL

MOBILE

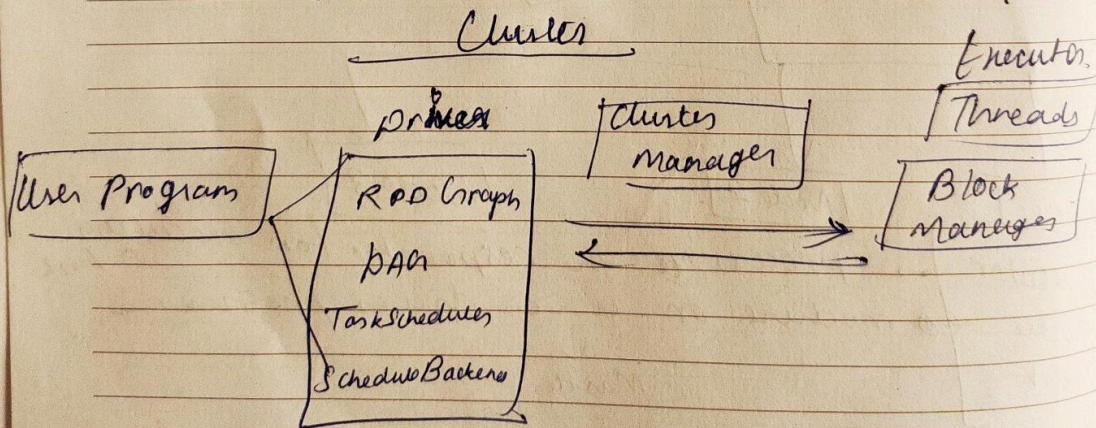
PHONES



RDD (Resilient Distributed Dataset)

Spark. →
 → high speed
 → distributed processing.

Streaming data: live, dynamic data.



30
 2023 SATURDAY
 WK 52 • DAY 364-001
 DECEMBER 30 - 12 - 2023

JANUARY 2024
 S M T W T F S
 1 2 3 4 5 6 7
 8 9 10 11 12 13 14
 15 16 17 18 19 20 21
 22 23 24 25 26 27 28
 29 30 31

023
 S
 3
 10
 17
 24
 31

Spark Content :-

- connection to a Spark cluster
- can create RDDs.

DAG Scheduler

- computes a DAG of stages for each job and submits them to Task scheduler for finding min. scheduler to run the job.

Task Scheduler

- responsible for sending tasks to cluster, run them, retry if failure

Scheduler Backend

- backend interface for scheduling system that allow plugging

RDD Obj's	DAG Scheduler	Task Sch.	Worker
build operators	Split graph into tasks	Launch tasks via clients	manage
DAG			

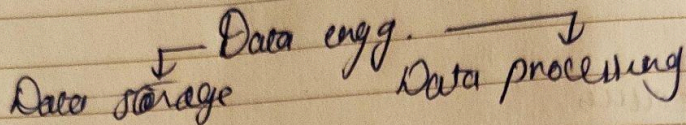
52	51	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	S	F	T	W	T	M	WK	2023
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	------

2023
FRIDAY
WK 52 • DAY 363-002
DECEMBER

29-12-2023
29

Key features

A cluster in spark is the group of machines



Apache Spark Core

1. deliver speed by ~~pro~~ parallel-distributed processing of huge dataset
2. Every functionalities are built on top of core

Features :-

- Change of I/O functionalities.
- task dispatching
- fault tolerance