Apache Spark - written in Scala & runs in JVM
general purpose cluster computing system. high-level API in Java Scala a though
general purpose cluster computing system. high-level API in Java, Scala, Python, R. It also has abundant high-level tools for structured data processing, machine learning, graph processing distressing.
6 components of Apache Spark ecosystem
- Spark core
- Spark SQL
- Spark streaming
- MLib
- Cropphx - SparkR
Spark Spark M Llib Graph X
Spark Core
Spark Madoop YARN Mesos
Standalone
Spark Components
Apache Spark Core
It delivers speed by providing in-memory computation
parallel & distributed computing

and the second s	
In charge of essential. I/o functionalities.	6
	6
Fault recovery Task dispatching It overcomes the snag of mop-reduce by using in-memory computation.	6
last dispatching	6
it overcomes the snag of map-ruduce by using in-memory	
computation.	6
Coople con is a 111 11 11 11 11	
Spark core is embedded with special collection called RDD (resilient distributed dataset).	-
desilien asimbuled dataset).	-
Spark ROD handly and basing data amount the	t
Spark RDD handles partioning data across all the nodes in a cluster.	•
	ê
Two operations performed on RDDs;	e
· · · ·	è
Transformation > function that creater new RDD from the existing	ans
	¢
Action > When we want to work with actual dataset, then we we have Action.	È
we use Action.	E
	C
	C
Spank SQL	
71: 1:4741 £ 1 € 1 4 114	
It is a distributed framework for structured data process	ang &
	6

Key features -

Features of Spark SQL
Cost band Optimiser Mid-query fault-folerany
Mid-query fault-folerany
Full compatibility with existing hire data. Datat
Datafrances le SQL provide ways to access a variety of data sources.
Cooph Str.
Spark Streaming
Add-on to cose spook API which allows scalable - high-throughput,
- high - throughput, fault - tolerant stream processing of line data streams
3 phases of Spark
(a) Crathering > Basic sources - Sources which are
Advanced sources sources like Katha, Flume
(b) Procerning
3(c) Data Storage de processed datais pushed out to file systems, databases & hix dashboards.
DStocam in Spark signifier continuous stream of data.
Dits cam is internally a require of RDD's.

Melib (Machine learning library) Machine learning library RDD graph Chriter Threads DAG Schiduler Block Makager Spank Context Management Task Schedules Scheduler backend 7 Llib (Madine tearning How Spark Works Task Scheduler Worker DAG Scheduler RDD Objects Towset | Tork DAG Execute tasks launch tash via Split graph into cluster management stages of task build operator DAh Submit each stage rutoy failed or ntraggling tasks as rucky