edureka!

Big Data Processing With Spark and Scala

http://www.edureka.co/apache-spark-scala-training

Objectives of this Session

edureka!

→What is Big Data?

→What is Spark?

→Why Spark?

→Spark Ecosystem

→A note about Scala

→Why Scala?

→MapReduce vs Spark

→Hello Spark!

Big Data

edureka!

- → Lots of Data (Terabytes or Petabytes)
- → Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
- → The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization

What is Spark?

edureka!

→ Apache Spark is a general-purpose cluster in-memory computing system

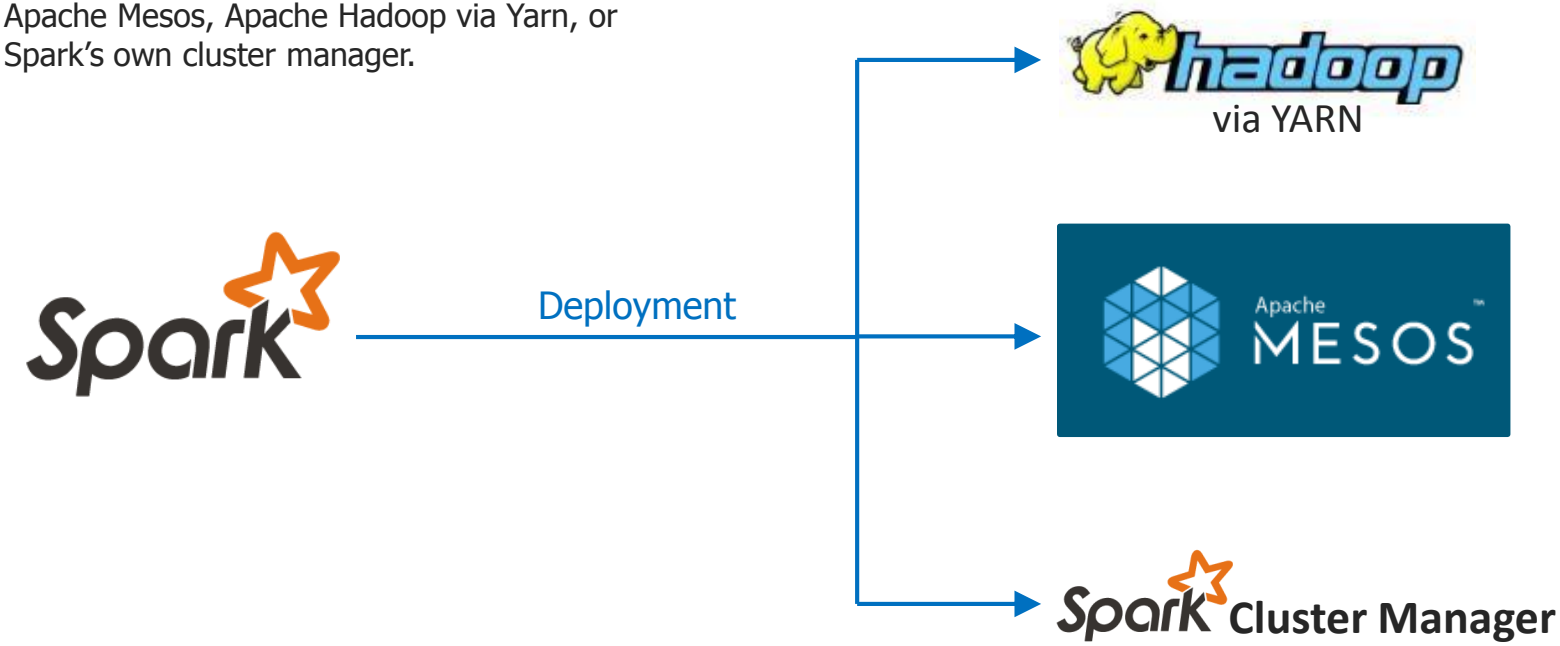
→ Provides high-level APIs in Java, Scala and Python, and an optimized engine that supports general execution graphs

→ Provides various high level tools like Spark SQL for structured data processing, Mlib for Machine Learning and more..

Why Spark?

edureka!

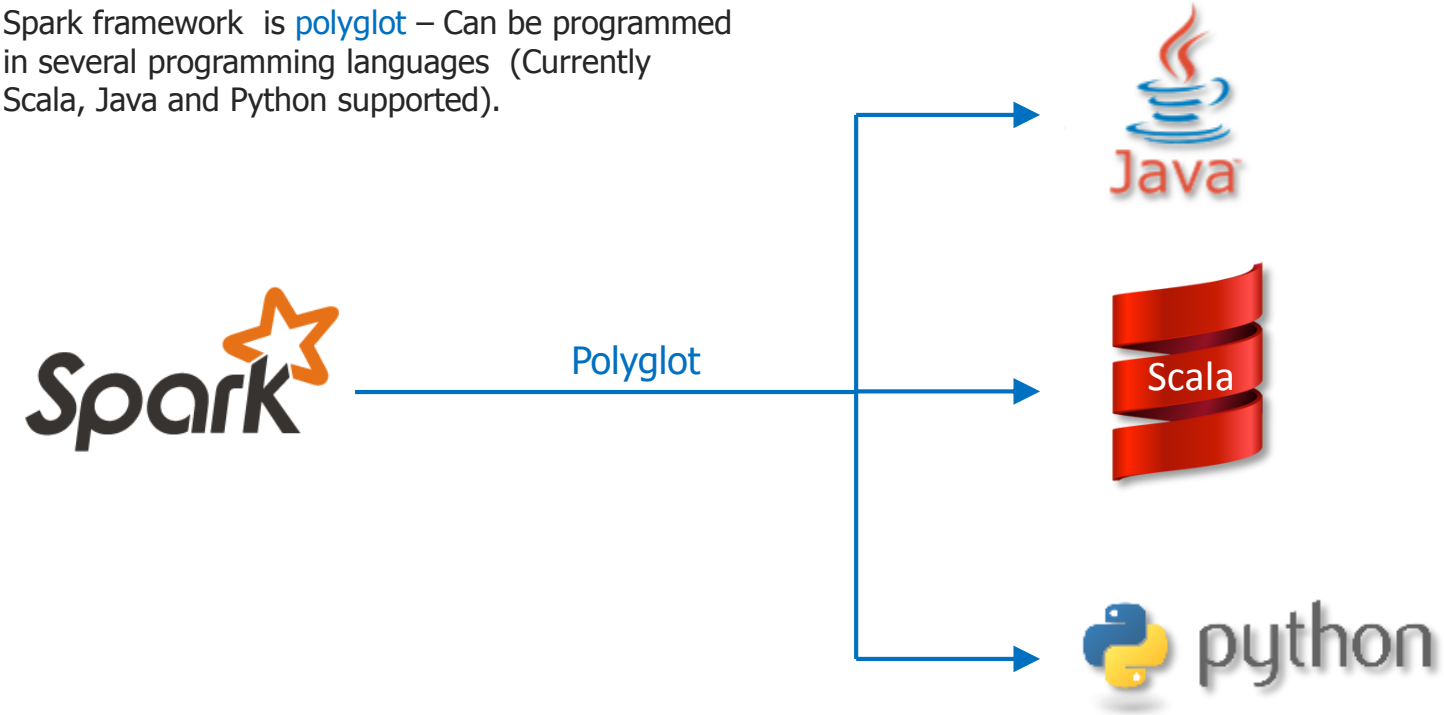
→ The Spark framework can be deployed through Apache Mesos, Apache Hadoop via Yarn, or Spark's own cluster manager.



Why Spark?

edureka!

→ Spark framework is polyglot – Can be programmed in several programming languages (Currently Scala, Java and Python supported).



Why Spark?

A fully **Apache Hive** compatible data warehousing system that can run **100x faster** than Hive.

100x faster than

for certain applications.

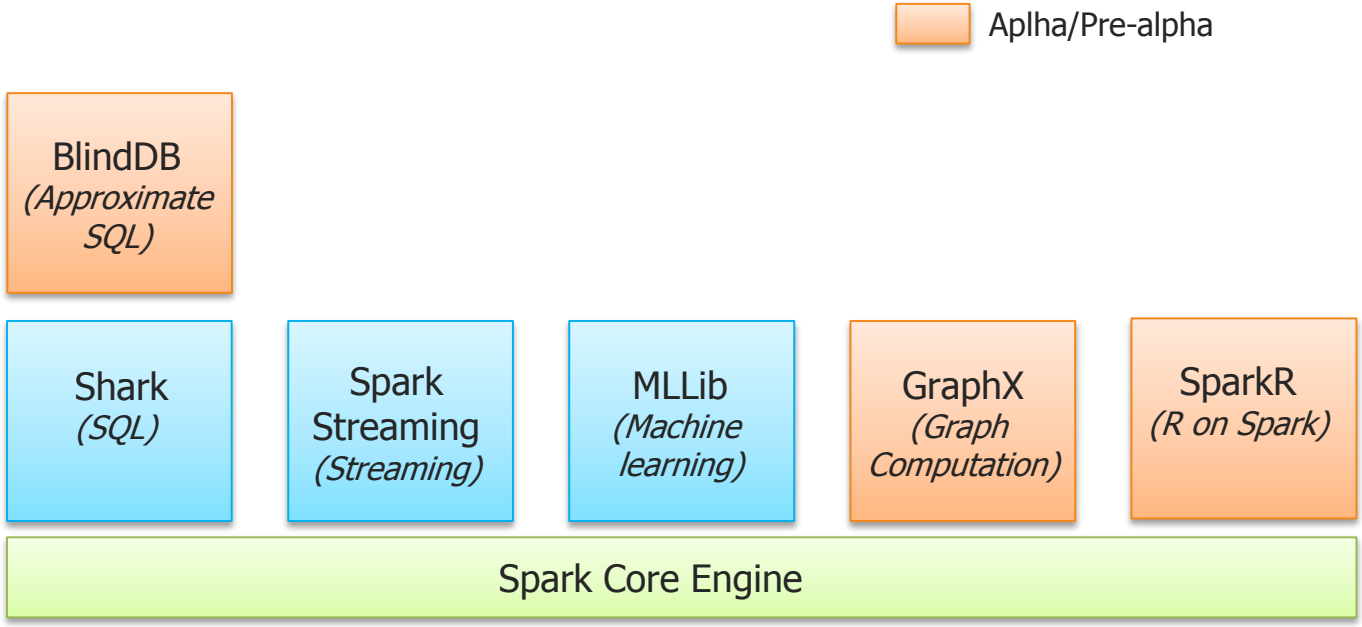
Why Spark?

edureka!

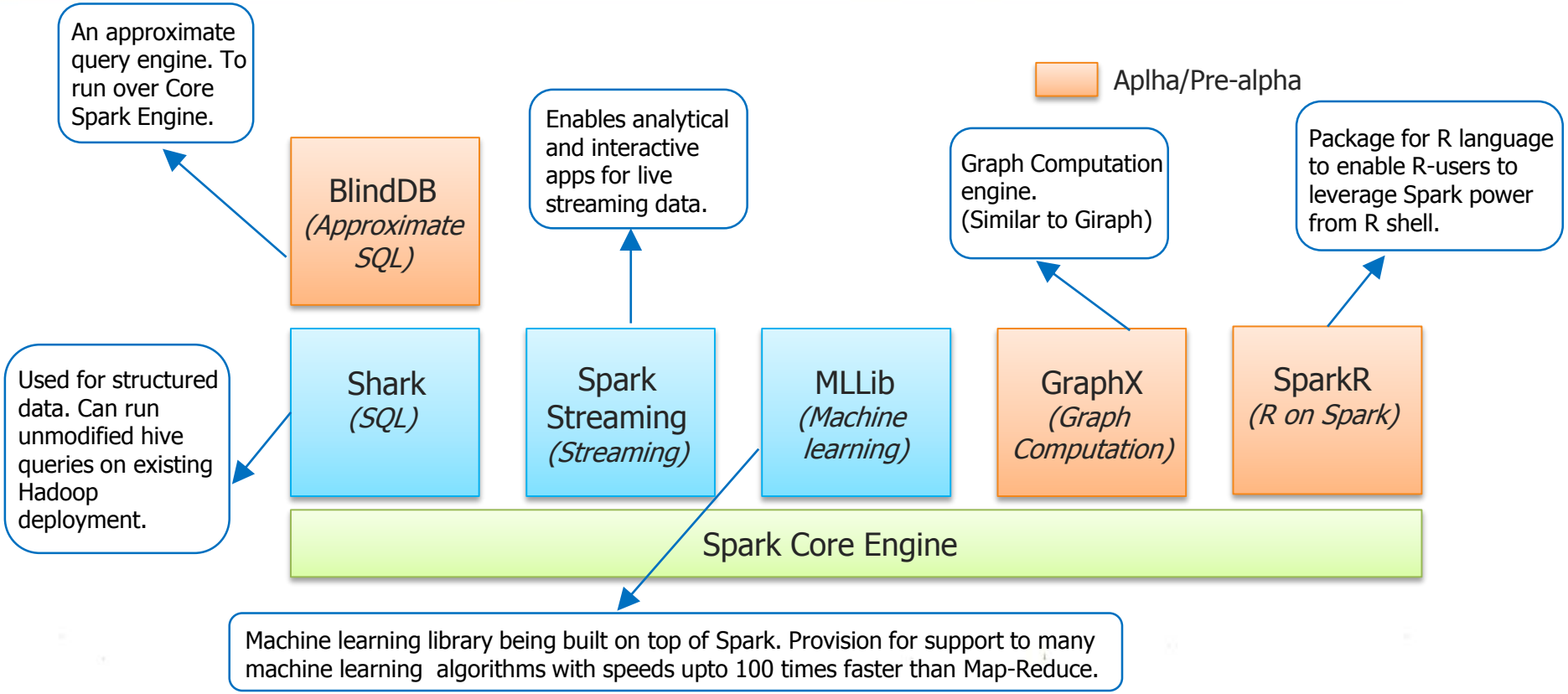
- → Provides powerful caching and disk persistence capabilities
- → Interactive Data Analysis
- → Faster Batch
- → Iterative Algorithms
- → Real-Time Stream Processing
- → Faster Decision-Making

Spark Community is Super Active!

Spark Ecosystem



Spark Ecosystem (Contd.)



A Note on Scala

edureka!

- → Scala is a general-purpose programming language designed to express common programming patterns in a concise, elegant, and type-safe way
- → Scala supports both Object Oriented Programming and Functional Programming
- → Scala is very much in fabric of present and Future Big Data frameworks like Scalding, Spark, Akka

- » All examples of Spark in class will be covered in Scala
- » Scala would be covered before Spark coverage as part of course!

Why Scala?

edureka!

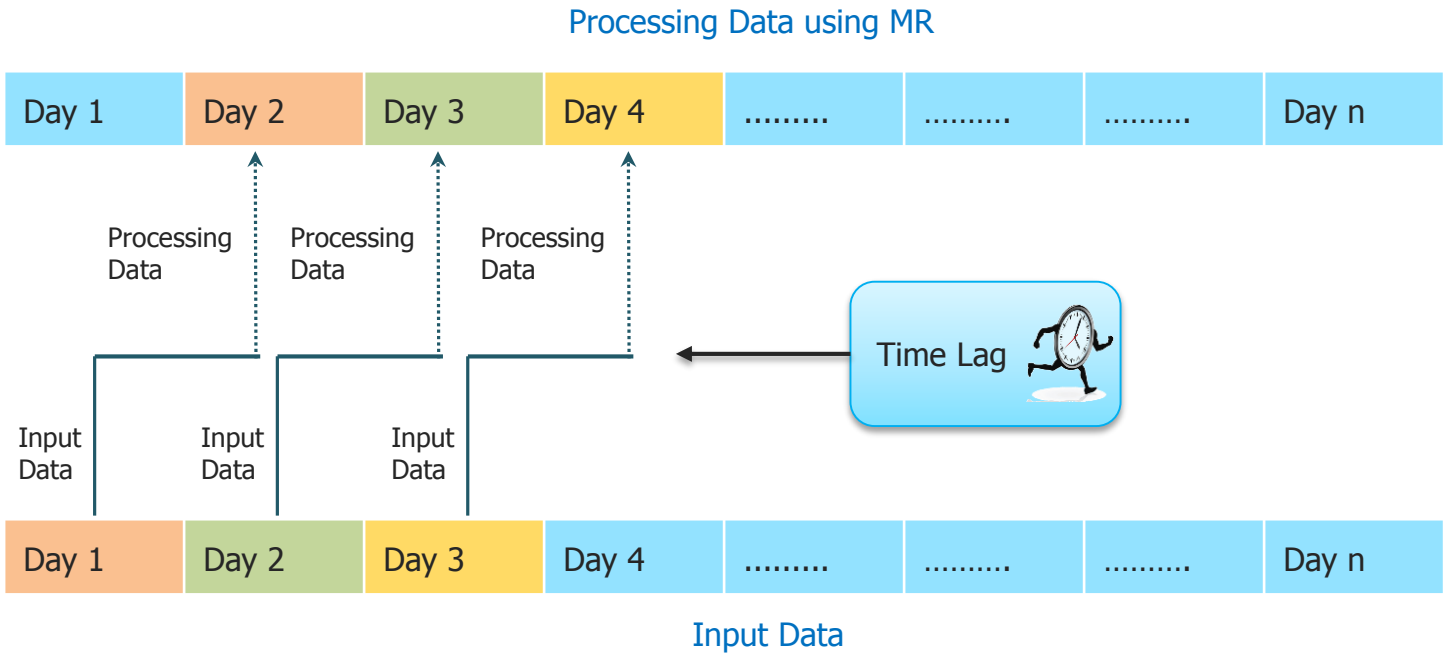
- → Scala is a pure object-oriented language. Conceptually, every value is an object and every operation is a method-call. The language supports advanced component architectures through classes and traits
- → Scala is also a functional language. Supports functions, immutable data structures and preference for immutability over mutation
- → Seamlessly integrated with Java
- → Being used heavily for future Big data and developments frameworks like Spark, Akka, Scalding, Play etc

Real Time Analytics

edureka!

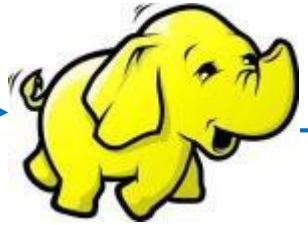
→ If you want to do some Real Time Analytics, where you are expecting result quickly, Hadoop should not be used directly

→ Hadoop works on Batch processing, hence response time is high



Real Time Analytics – Accepted Way

Storing



Streaming Data

MapReduce vs Spark

Application Overview

User:	edureka
Name:	mywc
Application Type:	MAPREDUCE
State:	FINISHED
FinalStatus:	SUCCEEDED
Started:	8-Dec-2014 17:02:22
Elapsed:	14sec
Tracking URL:	History
Diagnostics:	

14 sec

0.6 sec

```
14/12/08 04:10:06 INFO spark.SparkHadoopWriter: attempt_201412080410_0005_m_000000_5: Committed
14/12/08 04:10:06 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 5.0 (TID 5) in 296 ms on localhost (1/1)
14/12/08 04:10:06 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
14/12/08 04:10:06 INFO scheduler.DAGScheduler: Stage 5 (saveAsTextFile at <console>:14) finished in 0.279 s
14/12/08 04:10:06 INFO spark.SparkContext: Job finished: saveAsTextFile at <console>:14, took 0.626043309 s
14/12/08 04:10:06 INFO executor.Executor: Finished task 0.0 in stage 5.0 (TID 5). 826 bytes result sent to driver
wordCounts: Unit = ()
```

Spark Demo!

edureka!

Spark Demo!

Questions?

edureka!