Apache Spark for Beginners

Introduction, Architecture, RDD, DataFrame, DataSet, etc.

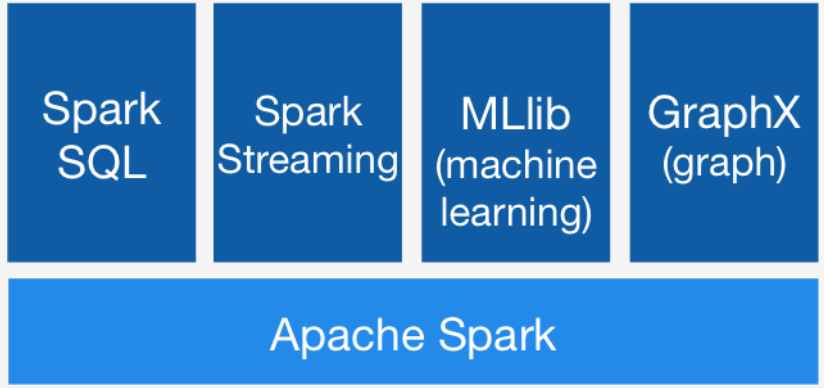
By **Cloud TV India**

Apache Spark - Introduction

Apache Spark™ is a unified analytics engine for large-scale data processing.

Apache Spark ecosystem components/libraries are,

- Spark Core API(RDD)
- Spark SQL(SQL, DataFrame)
- Spark Streaming
- MLlib/Spark ML(Machine Learning
- GraphX



Apache Hadoop vs Apache Spark

- Hadoop has two core components HDFS(Hadoop Distributed File System), MapReduce
- HDFS - Reliable and Scalable storage solution for storing big datasets
- MapReduce - Distributed programming model which helps big data computation

Let us compare Apache Hadoop and Apache Spark based on below four key points:

Key Points	Apache Hadoop	Apache Spark
Storage	HDFS	Uses/Leverages existing solutions like HDFS, S3, NoSQL, etc.
Computation	MapReduce - Designed in the form of map, reduce, shuffle phase	Its own implementation of computation engine which is much more faster and efficient
Computational Speed	Fast	Run workloads 10x - 100x faster
Resource Management	YARN	Standalone

Advantages of Apache Spark

Hadoop lacks in two below areas,

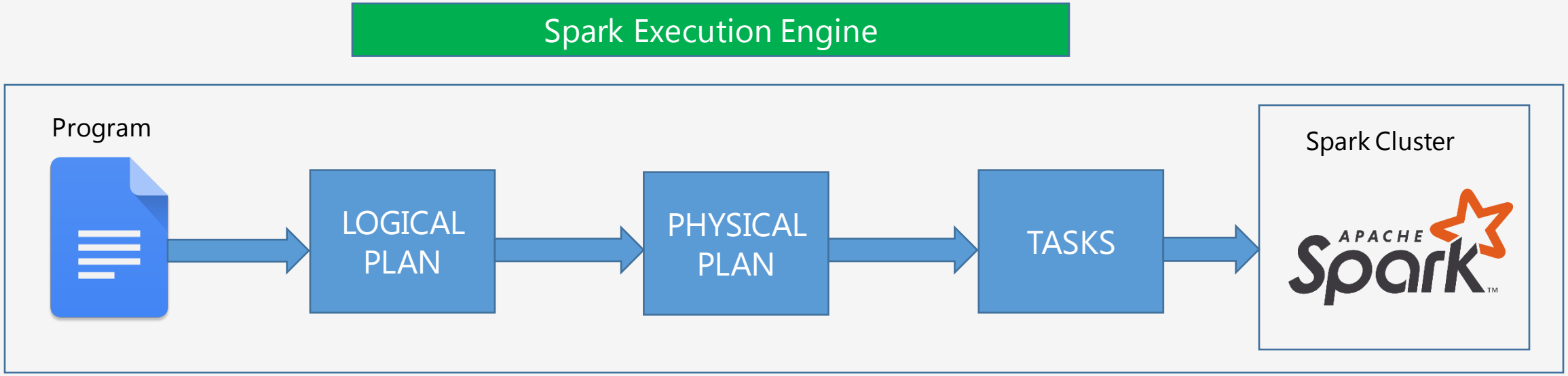
- Iterative Machine Learning
- Interactive Data Analysis

What Apache Spark provides,

- Iterative Machine Learning - Intermediate data is kept in memory to reduce no. of read and write to disk which helps to perform the computation faster and efficiently
- Interactive Data Analysis - Rich set of functions to do data analysis, speed up data analysis by caching your data in memory

How Apache Spark achieves faster computation compare to Hadoop

In-memory computing at distributed scale - Caching data in memory

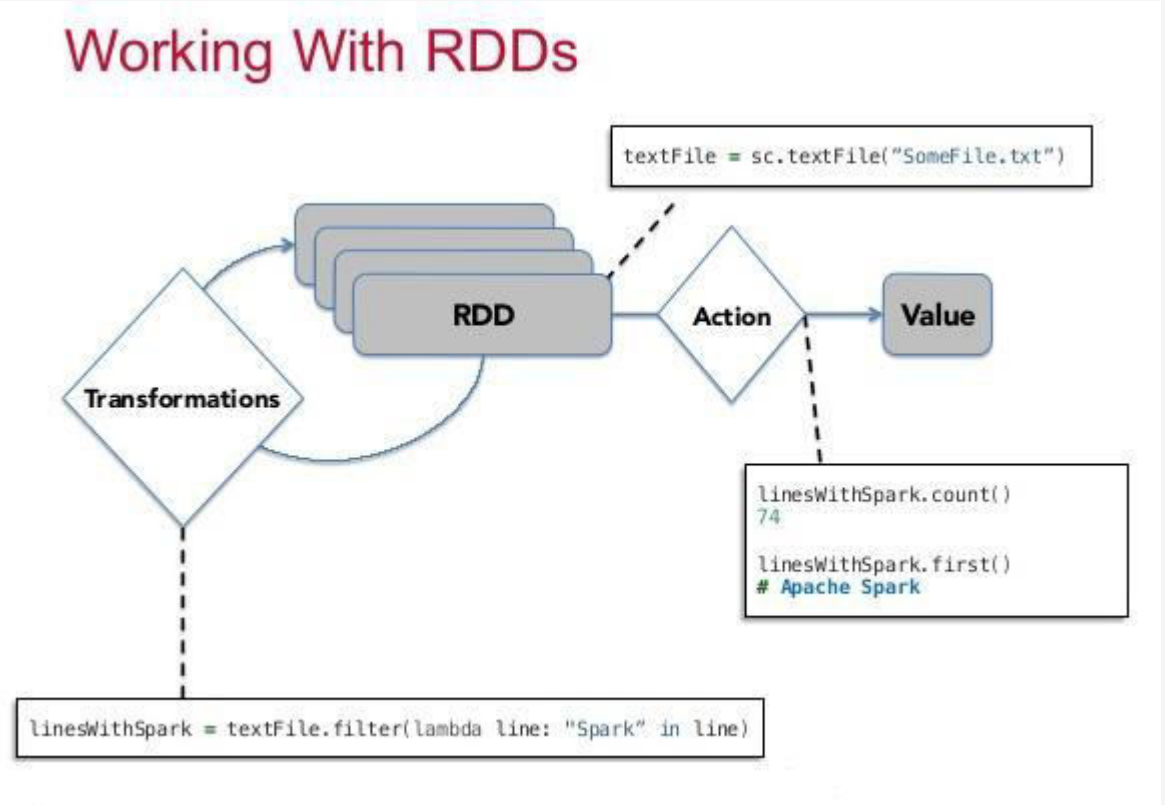


- Spark execution engine translates user code into series of tasks, that task or operation is DAG(Directed Acyclic Graph) in nature, meaning execution flow goes from one operation to another/one task to another task, but never come back and re-execute same task again(no cyclic flow)
- Spark achieves this tracking of each operation on dataset using concept called RDD(Resilient Distributed Datasets)
- Spark Architecture is built from the ground up for speed and efficiency

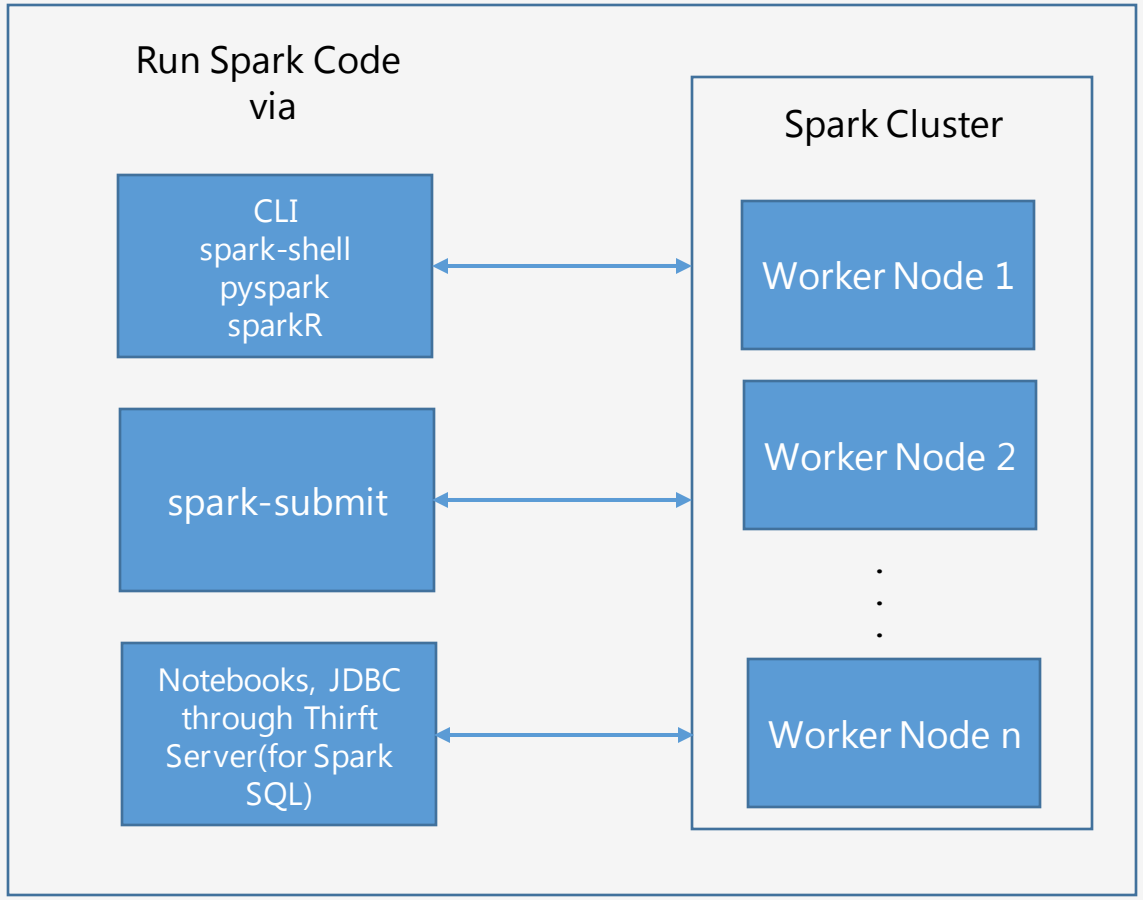
Apache Spark - RDD(Resilient Distributed Datasets)

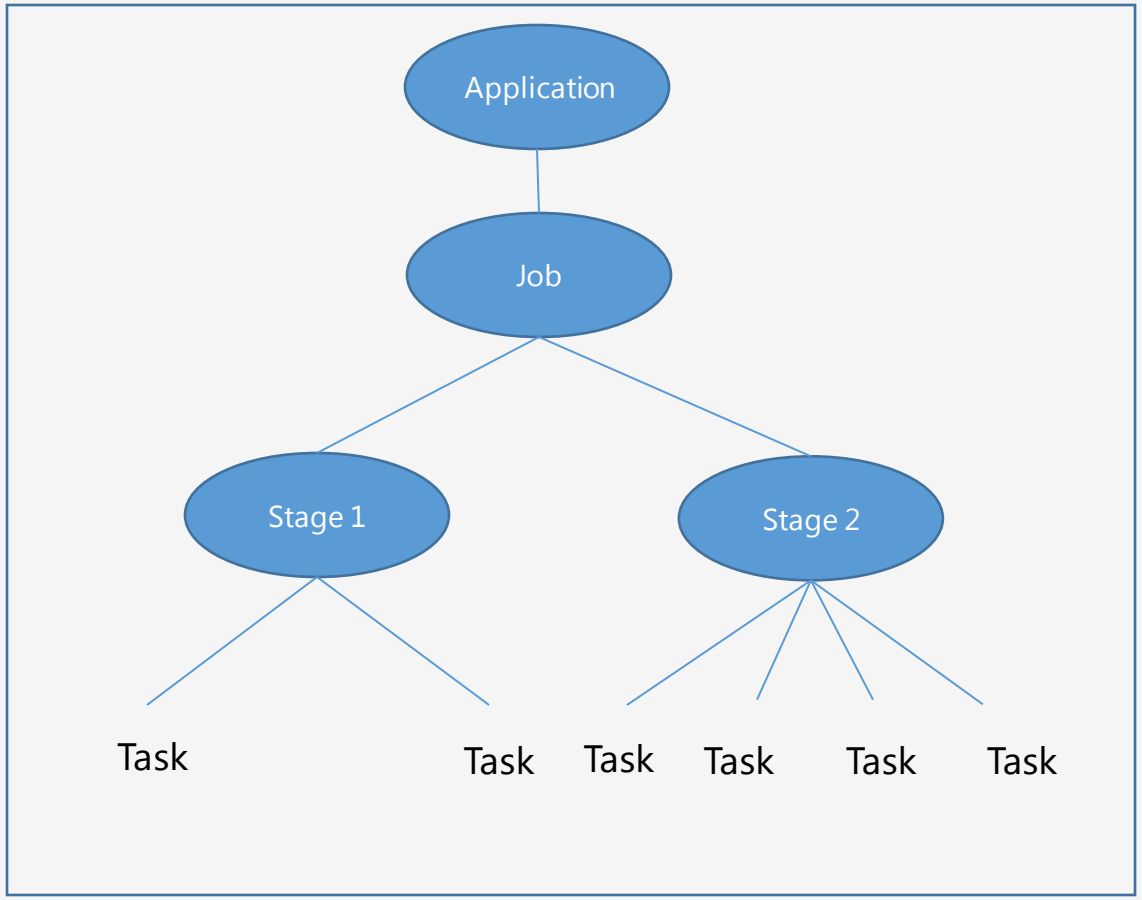
A Resilient Distributed Dataset (RDD), the basic abstraction in Spark. Represents an immutable, partitioned collection of elements that can be operated on in parallel.

Two types of operation that we can perform on RDD:
Transformation - which generates new RDD from existing RDD/creates new RDD Example: map, filter
Actions - which return a value to the driver program after running a computation on the RDD Example: count, collect, take

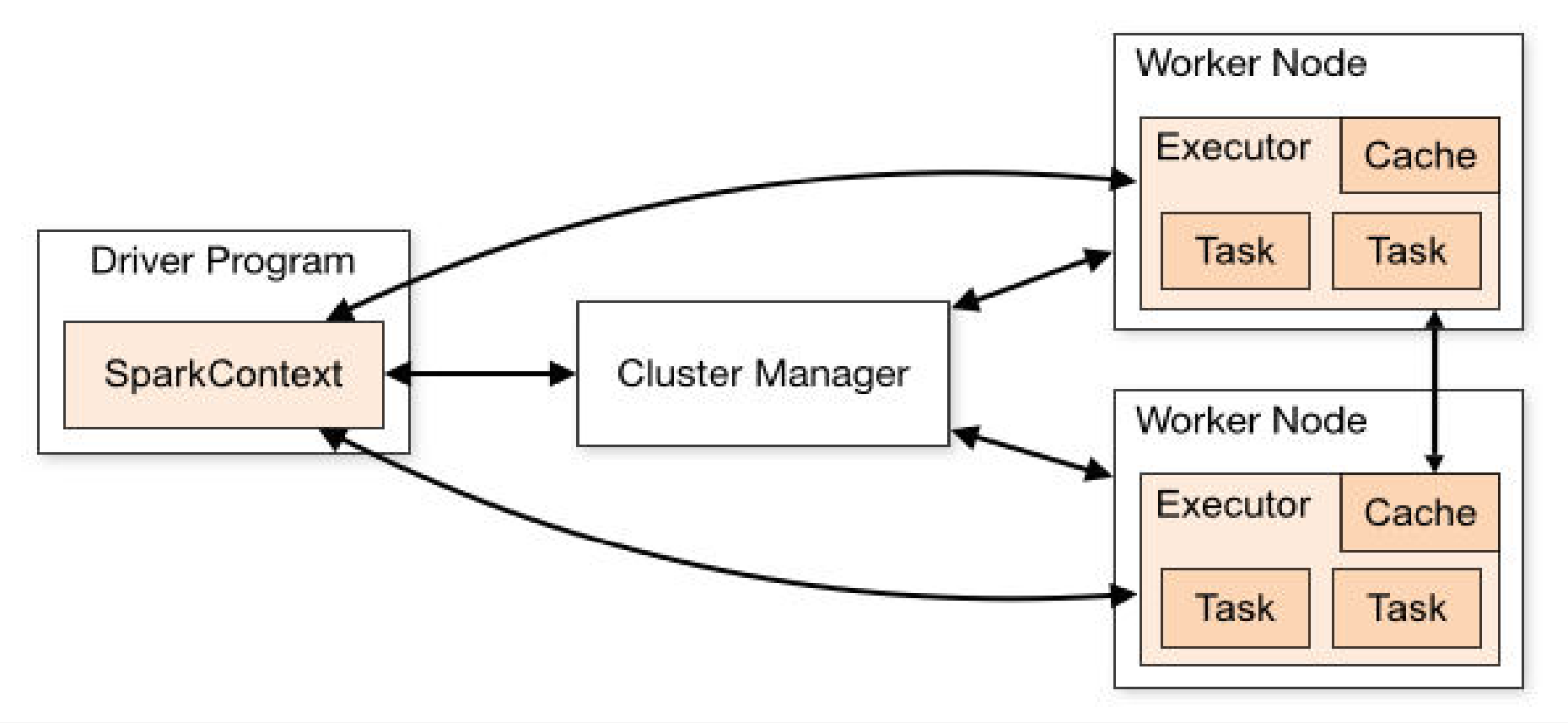


Apache Spark - Architecture

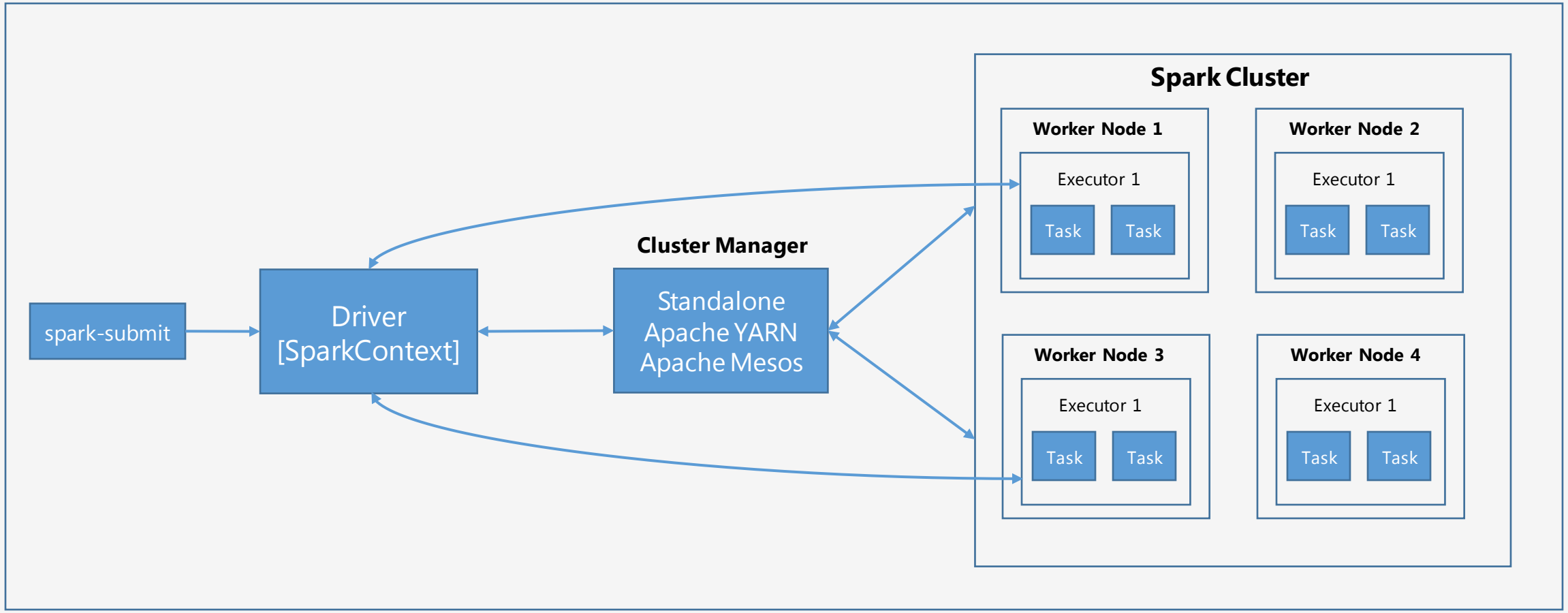




Apache Spark - Architecture



Apache Spark - Architecture



Learn more about Apache Spark …

Apache Spark official documentation: https://spark.apache.org/docs/latest/quick-start.html

Reach Us Out Here

Cloud TV India's Telegram group: https://t.me/joinchat/LH3O8022w8PIizV9NkHrag

Cloud TV India's WhatsApp group: https://chat.whatsapp.com/IGI19zxh5825a3Ewz2ciNJ

Cloud TV India's YouTube channel: https://www.youtube.com/channel/UCFQucNX7WsUwaWGNTrn6bIQ