[Section 2 Lecture 8]

What is spark?

A fast and general engine for large scale data processing

Working Architecture

Driver Program:

The **spark driver** is the **program** that declares the transformations and actions on RDDs of data and submits such requests to the master

Cluster Manager:

Spark supports pluggable cluster management. The cluster manager is responsible for starting executor processes. ... A central master service decides which applications get to run executor processes, as well as where and when they get to run

Executors:

Executors are worker nodes' processes in charge of running individual tasks in a given Spark job

Features Of Spark

Spark works much faster than hadoop . As it is a process of in memory computation.

Uses DAG(Directed Acyclic Graph) that optimizes result.

Components Of Spark

Spark Streaming:

This component allows Spark to process real-time streaming data. It provides an API to manipulate data streams that matches with RDD API. It allows the programmers to understand the project and switch through the applications that manipulate the data and giving outcome in real-time. Similar to Spark Core, Spark Streaming strives to make the system fault-tolerant and scalable.

Spark SQL

Spark SQL is a component on top of Spark Core that introduces a new set of data abstraction called Schema RDD, which provides support for both the structured and semi-structured data.

Spark MLLIB

Apache Spark is equipped with a rich library known as MLlib. This library contains a wide array of machine learning algorithms, classification, clustering and collaboration filters, etc

GraphX

Spark also comes with a library to manipulate the graphs and performing computations, called as GraphX. Just like Spark Streaming and Spark SQL, GraphX also extends Spark RDD API which creates a directed graph. It also contains numerous operators in order to manipulate the graphs along with graph algorithms.

[Section 2 Lecture 9]

Resilient Distributed Dataset

Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

Basic Operations On RDD

map -> One to one arrangement of elements
flatMap -> One to many arrangement of elements
filter -> Filtering data by specific feature
distinct -> Removing redundant data
join -> Joining dataset

Some Functions Of RDD

collect
count
countByValue
top
reduce

NB: Nothing actually happens in your driver program until an action or a function is called from RDD