of 19



Apache Spark =============

Apache spark is a ==================

general purpose in memory compute engine

compute engine

hadoop provides 3 things:

1. hdfs – Storage

2. mapreduce – Computation

3. YARN - Resource Manager

Spark is a replacement/alternative of mapreduce.

Spark is a plug and play compute engine which needs 2things to work with.

1. Storage - local storage, hdfs, Amazon S3

2. Resource Manager - YARN, Mesos, Kubernetes in-memory

mr1 mr2 mr3 mr4 mr5

HDFS for each mapreduce job we require 2 disk access one is for reading and other is for writing.

Spark

V1 V2 V3 v4 v5

HDFS only 2 disk IO's are required.

spark is said to be 10 to 100 times faster than mapreduce

General Purpose

================

pig for cleaning

hive for querying

mahout sqoop

only bound to use map and reduce

learn just one style of writing the code and all the things like cleaning, querying, machine learning, data ingestion all these can happen with that. filter map reduce

Spark Session - 2

The basic unit which holds the data in spark is called as RDD

resilient distributed dataset

List

RDD is nothing but in-memory distributed collection.

rdd1 = load file1 from hdfs

rdd1 = rdd1.map

rdd1 = rdd2.filter

rdd1.collect()

DAG - Directed Acyclic graph There are 2 kind of operations in spark

1. Transformation

2. Action Transformations are lazy Action are not. whenever you call transformations an entry to the execution plan is added.

we need to find average number of connections for each age 33, 100 33 , 200 33, 300 output 33,200 42, 200 42, 400 42, 500 42 ,700 output 42, 450 input 0,Will,33,385output (33,385) //input //(33,100)