PJ LOKESH

211720104084

HOME WORK 9

1.##WhatisApacheSparkStreaming

ApacheSparkStreamingisascalablefault-tolerantstreamingprocessingsystemthat natively supports both batch and streaming workloads

2.##DescribehowSparkStreamingprocessesdata?

SparkStreamingreceivesliveinputdatastreamsanddividesthedataintobatches,which arethenprocessedbytheSparkenginetogeneratethefinalstreamofresultsinbatches.

SparkStreamingprovidesahigh-levelabstractioncalleddiscretizedstreamorDStream,which representsacontinuousstreamofdata.

3.##WhatareDStreams?

DiscretizedStreams(DStreams):

DiscretizedStreamorDStreamisthebasicabstractionprovidedbySparkStreaming.It representsacontinuousstreamofdata,

Eithertheinputdatastreamreceivedfromsource,ortheprocesseddatastreamgeneratedby transformingtheinputstream.

4.##WhatisaStreamingContextobject?

PublicclassStreamingContextextendsObjectimplementsLogging.Mainentrypointfor SparkStreamingfunctionality.

ItprovidesmethodsusedtocreateDStreamsfromvariousinputsources.Itcanbeeither createdbyprovidingaSparkmasterURLandanappName,orfromaorg.apache.

5.##WhataresomeofthecommontransformationsonDStreams SupportedbySparkStreaming?

ItappliesoneverybatchmeaningeveryRDDinaDStream.ItincludescommonRDD transformationslikemap(),filter(),reduceByKey()etc.

Althoughthesefunctionsseemlikeapplyingtothewholestream,eachDStreamisacollection ofmanyRDDs(batches).Asaresult,eachstatelesstransformationappliestoeachRDD.

6.##Whataretheoutputoperationsthatcanbeperformedon

DStreams?

Someoftheoutputoperationsareprint(),save()etc..Thesaveoperationtakesdirectoryto savefileintoandanoptionalsuffix.

Theprint()takesInthefirst10elementsfromeachbatchoftheDStreamandprintstheresult.