



- 1. Apache Spark Professional Training with Hands On Lab Sessions
- 2. Oreilly Databricks Apache Spark Developer Certification Simulator

SPARK ADVANCED : DATA PARTITIONING

By www.HadoopExam.com

Note: These instructions should be used with the HadoopExam Apache Spark: Professional Trainings.

Where it is executed and you can do hands on with trainer.

1. Hadoop Training

- 2. Spark Training
- 3. HBase Training
- 4. MapR Developer
- 5. MapR HBase
- 6. CCA500 Certification
- 7. Spark Certification Send of the Offer Prices
- 8. EMC Data Science

<u>Hadoop Specialization offer == 50% + 35% off</u>

doop 52000INR == 16900INR Only pert \$1150 == \$373 Only

Hadoop Specialization offer

Limited Time Offee (Less Than 5 Days Remain)



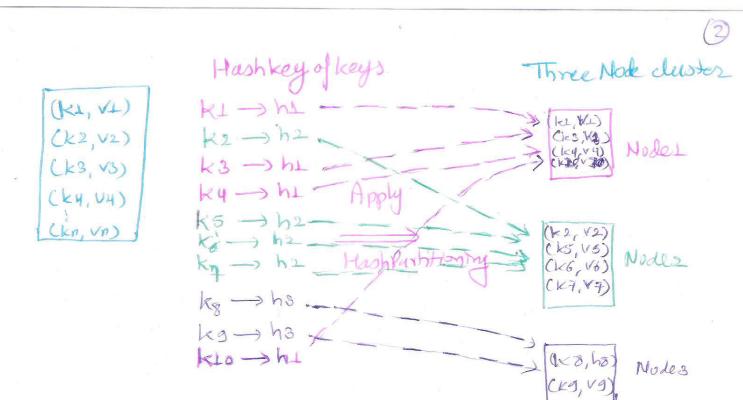
Cloudera CCA175 (Hadoop and Spark Developer Hands-on Certification available with total 75 solved problem scenarios. Click for More Detail)

- **1.** What is Partitioning and why?
- 2. Data Partitioning example using Join (Hash Partitioning)
- 3. Understand Partitioning using Example for get Recommendations for Customer
- 4. Understand Partitioning code using Spark-Scala
- 5. Operations which create Partitioned RDD
- 6. Operation which get benefit of Partitioning
- 7. Operation that affect the partitioning

Spank: Data Pastitioning

- > Distributing Jula Across nodes in cluster, to improve efficiency.
- =) To biggest performance construint, while working on distributed platform.
 - -> Disk Read & Write
 - -> Network Traffic
- > Data pastitioning will reduce now traffic.
- => Always useful: No
- => If given RDD is 8 canned only once, then
 no use of pustitioning
- Hhen=) In key-oriented operations (maybe Pain RD9), and distaset is re-used multiple times e.g. Lookup Data, used in Join operation.
 - => Hashkey bused pastitioning: -

WWW. Hadoop Exam-com



=) Now calling persistes method on RDD to Mondatory

Example: - Let's tuke on example from an Amazon kindle online, web reading application.

Leasyez

Customer_Info (RDD Structure)

Constrainted, list [books, books, books, books --- books]

I, [books, books, books, books, books, books]

2, [books, books, books, books, books]

=) Customer Info RDD is huge in size.

-> Another RDD, which we receive every to minutes using Streaming Applications. With following structure

Constances Id, book Visited

1; book 1

3, book 44

3, book 1001

20, book 1001

1001, book 42

Www.Hadropexom.com

Problem Statement:=>

Customer-Interest

Find out the book name, which is not substricted by a Customer, but he had visited.

For Example: - Ceestoner Id -> 1001 Book VIsited -> book 72 (Not subscribed)

Join both the RDD, so Amazon can find, which book is not subscribed by customer, but he has interest 80 it can send recommendation in his kindle box.

=> Why pastitioning will help here.

Partitioning of Customer-Info RDD, is required. Because this RDD, we will be using in every tominutes to calculate customer Interest. Tuo RDDS Customer Info Customer Interest 1319 Table Re-generated every (No need to re-calculate to minutes from Stocaroning application again and again) Pastition this ROD & persist () is state Customer-Interestinas Customer: Info (RDD) (1,60LL) (2,6007) Part-2 (3) 4 3.600ksd 1,000k7 1, took 10 Parts 8,2 (1, buskey) (7, bude 7) (8,600LB)

Application Code looklike

Val user

Val Customer Deeta = Se. tentfile [Custad, 45t [600k]]

Dessist ()

partition By (new Hash Partitioner (100))

persist ()

3

1 -> partition by is a method to prestition data

2 -> Algorithm (technique) used to do pastitioning.

3 -> Save the pustition created (Nort miss it)

det generate Data (Streamed File: String) {

val now = Sc. tertfile [cust 2d, book] (stromfile)

val result = customer Desta. join (2000).

// (CustId, [List(Book), Book)] RDS structure Joined

val new Visits = result. filter &

Case (custed, (list, book)) =>

! list. books. Contains (book.name)

3

new Visits. Save Astentile ("path"),

=> Operations creete furtitioned RSD:-

There are many operations which will create partitioned RDD.

20st Bykey() -> Range Partition.
group Bykey() -> hash-partitioned RDD.

mape) -> torget furent RDD partitions

Why? => This mape operation can

modify the key

=> Operations that Benefit from Partitioning: -

-> Operations which requires shuffling data by key.

-> Cognoup() Sright Outer Join() Wolkap()

group with() group Bykay()

Join() reduce Bykey()

left Outer Join() Combine Bykey()

-> reduceByliey(): - Will work only on single RDD.

k7 Part-L (do operation loadly in k8
k8
Cerch partition)

K3
Result to master node

Operation That affect Partitioning: -Calling reduceByley() often join() Significantly faster. Why? Join will creek hash-pastitioned RDD and reduce Bylieges will take benefit of this. =) mup() can change key. => may Values & flootmap Values (): will not change the Dinary Operation: Partitioned by Treault ROO [it will be perhitioned by Hach] Hash (2) & Hash Partitioner Ceiston-Partitioner result ROD [It will be with centon gatitoner] [No pudition] [No-pathron] RODL 2002 Treal ARDO [Sefout Partition: Hish-partition]

WWW. HADOOPEXAM. COM

cloudera CCA175 Is Now Available, with Hands On Sessions

CCA Spark and Hadoop Developer Certification

HadoopExam Learning Resource provides the following material for the Advanced Technologies. Please visit www.HadoopExam.com for more detail this is just a few products from portfolio.

Price start for training with Just \$79/3500INR



Apache Spark Professional Training with HandsOn Session

+ Certification Material



Hadoop Professional Training with HandsOn Sessioon

> + Certification Material



HBase Professional Training with HandsOn Session

> + Certification Material



Certification Material





Certification Material



Certification Material



Certification Material



Certification Material

