Hadoop and Spark Developer - CCA 175

File Formats

Quick reference table for reading and writing into several file formats in hdfs.

File Format	Action	Procedure and points to remember
TEXT FILE	READ	sparkContext.textFile(<path to file>);
	WRITE	sparkContext.saveAsTextFile(<path to file>,classOf[**compressionCodecClass**]); //use any codec here org.apache.hadoop.io.compress.(**BZip2Codec** or GZipCodec or SnappyCodec)
SEQUENCE FILE	READ	sparkContext.sequenceFile(<path location>,classOf[<class name>],classOf[<**compressionCodecClass** >]); //read the head of sequence file to understand what two class names need to be used here

	WRITE	rdd.saveAsSequenceFile(<path location>, Some(classOf[compressionCodecClass]))
		//use any codec here (**BZip2Codec**,GZipCodec,SnappyCodec)
		//here rdd is MapPartitionRDD and not the regular pair RDD.

PARQUET FILE	READ	//use data frame to load the file. sqlContext.read.parquet(<path to location>); //this results in a data frame object.

	WRITE	sqlContext.setConf("spark.sql.parquet.compression.codec","gzip") //use gzip, snappy, lzo or uncompressed here
		dataFrame.write.parquet(<path to location>);

ORC FILE	READ	sqlContext.read.orc(<path to location>); //this results in a dataframe
	WRITE	df.write.mode(SaveMode.Overwrite).format("orc") .save(<path to location>)
AVRO FILE	READ	import com.databricks.spark.avro._; sqlContext.read.avro(<path to location>); // this results in a data frame object

	WRITE	sqlContext.setConf("spark.sql.avro.compression.codec","snappy") //use snappy, deflate, uncompressed;
		dataFrame.write.avro(<path to location>);

JSON FILE	READ	sqlContext.read.json();
	WRITE	dataFrame.toJSON().saveAsTextFile(<path to location>,classOf[Compression Codec])

G+

35 comments:

amitrock May 6, 2017 at 9:41 PM

Arun , Are you preparing contents for Configurations as per the last module of CCA175 syllabus ???

Reply

▾ Replies

Arun Kumar Pasuparthi May 10, 2017 at 5:58 PM

Yes, this is as per new syllabus. Please go through the prep plan page and the corresponding video.

http://arun-teaches-u-tech.blogspot.com/p/certification-preparation-plan.html

https://www.youtube.com/playlist?list=PLRLUm7no962j8cf-mpXjrQqusWvw-glJx

amitrock May 22, 2017 at 4:36 AM

Arun , I was expecting the Spark-Submit scenarios.

Is there any specific deadline by which you will finish the entire PlayList ?

I have already paid for the certification but due to syllabus change I have not yet given the certification exam.

Arun Kumar Pasuparthi May 22, 2017 at 5:12 AM

I am working on a scenario that combines spark streaming and submit. Will have them ready by end of this week.

Anjaneya July 28, 2017 at 12:56 PM

Hi Arun...Wonderful work. Thanks for sharing with us.
As mentioned above, could you please post streaming and submit questions?

Reply

Sandeep K May 21, 2017 at 6:50 AM

Arun, Sequence file write working only if rdd.saveAsSequenceFile(,Some(classOf[])).

Also is there a way to compress ORC file ?

Reply

▾ Replies

Arun Kumar Pasuparthi May 21, 2017 at 7:16 PM

Thank you for the correction. I updated the blog accordingly. Given that ORC files come with excellent compression ration and by default are written as snappy files i did not think about a need to use something else. I will update the blog if i find something soon.

Reply

Search This Blog

Search

CCA 175 Hadoop and Spark Developer Preparation

- Home
- CCA 175 Prep Plan
- Problem Scenario 1
- Problem Scenario 2
- Problem Scenario 3
- Problem Scenario 4
- Problem Scenario 5 [SQOOP]
- Problem Scenario 6 [Data Analysis]
- Problem Scenario 7 [FLUME]
- **File Formats**
- Youtube Playlist

A leader with a unique blend of deep technology expertise and strong management skil

Arun Kumar Pasuparthi

G+ Follow 212

View my complete profile

Report Abuse

Blog Archive

April 2017 (1)

MPS June 11, 2017 at 7:26 PM

This comment has been removed by the author.

Reply

MPS June 11, 2017 at 7:27 PM

Thank you so much for this blog. It helped me a lot to clear the exam.

Reply

Replies

Arun Kumar Pasuparthi June 25, 2017 at 8:43 AM

you are welcome. Congratulations!

Reply

Min Li June 16, 2017 at 10:43 AM

Hi Arun, just wanted to say a big thank you. Really appreciate your great work. This blog helped me a lot.

Reply

Replies

Arun Kumar Pasuparthi June 25, 2017 at 8:43 AM

you are welcome. Congratulations!

Reply

adarsh singh June 25, 2017 at 12:58 PM

```
sqlContext.setConf("spark.sql.parquet.compression.codec","lzo")
```

```
ordersDF.write.format("parquet").mode("overwrite").save("/data/output/orders_parquet_lzo")
=> failed with error Caused by: parquet.hadoop.BadConfigurationException: Class com.hadoop.compression.lzo.LzoCodec was not found
```

facing above issue . could you please help?

Reply

Replies

Arun Kumar Pasuparthi June 25, 2017 at 1:47 PM

there is nothing wrong in the code your wrote. the problem is lzo libraries are not available in the CDH you are using. They will likely not be available in the environment you use during the exam as well OR the exam will only ask you to perform lzo compression only if the lzo libraries are configured and available. Hence, you will not face this issue during the exam.

If you are trying to solve this for your project or for your company and your focus is not about the exam then please follow the solution available here.

https://stackoverflow.com/questions/23441142/class-com-hadoop-compression-lzo-lzocodec-not-found-for-spark-on-cdh-5

adarsh singh June 26, 2017 at 11:22 AM

Thanks Arun ... your expertise is helping me a lot to boost up my confidence...

Reply

kiran kumar Mudradi August 6, 2017 at 4:56 AM

Hi Arun,

I want to appreciate and want to say big thank you for all videos and tutorials. please continue doing the same in future also and keep us motivated.

Reply

RAJU C August 14, 2017 at 12:33 AM

Hi Arun,

Great work!! Much needed for guys like me to prepare for the Exam.

I have query?

Can We refer local documents(existing documents) during the exam? to cross check the syntax of the PIG/Hive commands.

Please let know

Reply

sathya September 1, 2017 at 5:51 AM

I'm not sure where you're getting your information, but good topic. I need to spend some time learning more or understanding more. Thanks for fantastic info I was looking for this information for my mission.

Informatica Training in Chennai

Dataware Housing Training in Chennai

Reply

rushya nathan September 7, 2017 at 12:02 PM

Thanks Arun for consolidating all the file formats. Just figured that parquet writing method works for orc and json as well. Just thought of sharing with others.

```
var dataFile = sqlContext.read.avro("");
```

.write.format works for parquet,orc and json

```
dataFile.write.format("parquet") .save("")
dataFile.write.format("orc").save("")
dataFile.write.format("json").save("")
```

Reply

Unknown September 15, 2017 at 11:43 AM

This comment has been removed by the author.

Reply

Shubham Aggarwal September 21, 2017 at 1:39 AM

Hi Arun,

Thanks for providing such a wonderful compilation of problems and solutions. It enabled be to clear the CCA 175 certification yesterday. Many thanks.

Regards
Shubham

Reply

sathya September 29, 2017 at 5:01 AM

This is a great post. I like this topic.This site has lots of advantage.I found many interesting things from this site. It helps me in many ways.Thanks for posting this again.

Hadoop Training in Chennai

Base SAS Training in Chennai

MSBI Training in Chennai

Reply

MS PRASAD November 1, 2017 at 3:55 AM

Click here

Download Here

Visit here

Reply

Trep Helix November 3, 2017 at 4:25 AM

I definitely appreciate your blog. Excellent work!
Startups

Reply

Freedom Apk December 1, 2017 at 2:15 AM

Tech tricks

Click here

Download Here

Useful tricks

Best tricks

Reply

venkatesh gurram December 12, 2017 at 9:17 AM

Hi Arun, Your blog is definitely the stepping stone towards successful CCA175 certification. Why you are not actively updating the posts. I see that the last updated was may-2017. Can you add few more for us? It was really helpful.

Reply

▾ Replies

Arun Kumar Pasuparthi December 12, 2017 at 10:21 AM

it was reported by many who follow my blog that the content was sufficent to clear the current version of CCA 175 exam. I will update the content if i receive any feedback on the coverage of the formulated problems in addressing CCA 175 exam needs.

Reply

Pavan Tammina December 16, 2017 at 12:36 PM

Hi Arun,
Thanks for the good work, All problems are very well designed to cover all the important scenarios.

I need a clarification, Can the data in Data frame be saved as text file or sequence file. It works for json/parquet and Avro.

Reply

Rajesh K December 22, 2017 at 4:14 PM

Hi Arun,

Thanks for the content. It's really helpful.
I am confused on dataframes. The videos from itversity state that we shouldn't use data frames. Is it because the videos are older and at the time of recording there was Spark V1.2?

Now, I see Spark 1.6 is being provided on CCA175 page from Cloudera. So can we use DataFrames?

Regards
Rajesh K

Reply

Jeyassri Balachandran December 29, 2017 at 11:54 PM

what about csv files? how to read and write them with compression? there is no API for csv in cloudera quickstart VM.

Reply

SWAMI AYYAPPA January 1, 2018 at 11:13 PM

how to delete apple id permanently
delete apple id permanently
how to delete an apple id account

Reply

Rajesh K January 8, 2018 at 5:44 AM

How to save a DataFrame (or) an RDD as a text file with the delimiter as "|" (pipe) or "\t" (tab). Is there any API to do it? or it needs to be done manually in a map transformation?

Reply

Jatin January 9, 2018 at 2:54 AM

Hi I need to read json data which is on s3 in tar.gz format , can you help how to read it in spark using scala.

sqlContext.read.json is not working/reading .

Reply

Jatin January 9, 2018 at 2:55 AM

This comment has been removed by the author.

Reply

Jatin January 9, 2018 at 2:55 AM

This comment has been removed by the author.

Reply

Enter your comment...

Comment as: Select profile... ▾

Publish Preview

Home

Subscribe to: Posts (Atom)

If you have landed on this page then you are most likely aspiring to learn Hadoop ecosystem of technologies and tools. Why not make you...

(no title)
If you have landed on this page then you are most likely aspiring to learn Hadoop ecosystem of technologies and tools. Why not make you...

Simple theme. Powered by Blogger.