# Class Questions - SSW 2016 Spark Essentials - TE2

**Question from Zoltan Toth**: What is this question parking lot?

**Likes this question:** Jacob Parr, Doug Bateman

**Answer from Jacob Parr:** With a very large class, we’ll have students post questions here. The TAs will then help answer the questions. And the questions with the most likes will also be answered by the instructor. So put your questions here using the format below.

**Question from Zoltan Toth:** Here’s a sample question.

**Likes this question:** Taggart McCurdy, Andy Konwisky, +1 from Anonymous

**Answer from Jacob Parr:** Here’s a sample answer

Question from Kalyan: Will there be topic covering DataSet?

**Likes this question:** Mainul, +1, Tyler

**Answer:** Unfortunately datasets are not part of the Essentials class syllabus

Question from Stu: Will Spark 2.0 still support RDDs within classes/objects/libraries as emphasis seems to be shifting to Dataframes/SQLContexts?

**Likes this question:** Mainul, Monica, Paul

**Answer:** Yes, Spark 2.0 will not drop RDD support

Question from Mainul:

1.When to use RDD, Dataframe? What are the differences between them?

2. Dataframes are documented to be faster than RDD. But under the hood dataframes are actually saved as RDD. Then how are dataframes faster? +1

**Likes this question**: Monica, Paul, Stu, Park, Tyler

**Answer:** The Catalyst optimizer does a great job in optimizing your data pipeline. Furthermore all the data will run as Java bytecode. Use a Dataframe wherewhere you can!

Question from Mainul:

If I am reading data from disk already, why would I want to persist data into disk?

What is the difference between writing data into disk and persist data into disk?

**Answer:** When you persist data to disk, Spark can use a serialisation format that are quick to write and read. Writing i.e. JSON and reading it back manually takes much more time.

Question from Monica: does Databricks notebook not support java?

**Answer:** No, not at the moment. It’s more of a Java limitation. Java doesn’t have REPL (an interpreter).

Question from Monica: Does spark use data blocks created on HDFS for a file to partition the data for execution or does it read the entire file in memory and then parallelize it?

Answer from Tulasi P (participant): Yes, Spark uses underlying HDFS blocks to partition the data and perform reads in parallel. My understanding is that, number of partitions in rdd read from HDFS file is = # of blocks in HDFS for that file. <- this is the right answer (Zoltan)

Question from IP: Is it possible to execute a shell command in the notebook and see the output from the command? Is it possible to install additional python modules in the shell?

**Answer:** Yes, Databricks notebook supports executing shell commands. You would use **%sh** <shell command>. Yes, you can upload python library. For more information on Databricks notebook feature, please visit [here](https://docs.cloud.databricks.com/docs/latest/databricks_guide/02%20Product%20Overview/04%20Libraries.html).

Q: Will the class slides be shared? And where?

**Answer:** Yes, they will be available on Google drive after the class. You will receive instructions later.

Q: Where can I access the document(slides) on the screen?

**Answer:** It will be uploaded in Google drive later. You will receive instructions on this.

Question from Ashwini: Which NoSql database works best for the internal data store with some external data coming from Oracle?

Question from Ashwini: What are the best practices in Spark and where can we find them?

Answers: Some best practices are available [here](https://github.com/databricks/spark-knowledgebase).

Qn: How do we know how many partitions are there after a ‘filter’? Or do we call “coalesc” regardless?

**Answer:** You can use someRDD.getNumPartitions(). For more information on RDD partition, please visit [here](https://github.com/databricks/spark-knowledgebase/blob/master/performance_optimization/how_many_partitions_does_an_rdd_have.md). In general, coalescing is recommended when you have sparse data across partitions (e.g. after a filter transformation is applied on RDD).

Qn: Can ‘cached’ data be shared across sessions? I.e. other users can access that RDD?

**Answer:** Cached data is only available within the SparkContext it was created. It cannot be shared across SparkContexts.

Question from Kalyan: When the driver program does sc.textFile(“someFile”) then will all the data in the someFile loads into driver as RDD’s and then partition happens to worker nodes?

**Answer:** Input data never gets loaded into the driver machine.

Qn form Vinay: Is the caching happening on the executors or on the Driver

**Answer:** Caching is done on the worker nodes(executors)

Question from Stu: Does the cached data get expunged when you close the associated SparkContext?

**Answer:** Yes, the cached data is no longer available when you kill the sparkContext

Answer from Yuxing: I think the answer is yes. But a better practice is to call unpersist(boolean) when you no longer need it to save some memory.

Qn: what if cache cannot fit into memory? +1

**Answer:** It depends on the storage option that has been set as part of the configuration.

MEMORY\_ONLY: Store RDD as deserialized Java objects in the JVM. If the RDD does not fit in memory, some partitions will not be cached and will be recomputed on the fly each time they're needed. This is the default level.

MEMORY\_AND\_DISK: Store RDD as deserialized Java objects in the JVM. If the RDD does not fit in memory, store the partitions that don't fit on disk, and read them from there when they're needed.

Fore more information, please read RDD persistence section [here](http://spark.apache.org/docs/latest/programming-guide.html).

Qn: Can we check to see how much memory is being used?

Answer: You can view it from spark ui.

Qn: When we create an RDD, how do we tell what type of RDD it is?

**Answer:** In Databricks notebook, type the RDD name in a cell and execute the command. It should display the RDD type.

QN: If cache write to disk, is the disk area accessible only by the originating RDD?

**Answer:** The cached data is available as long as the sparkContext is running and during this active session, you can create any number of new RDDs using cached RDDs(data)

Question From Stu: Why did the number of partitions change in Exercise 4?

Qn from Vinay : Is there way to have partial count done in the executor itself (Exercise 3)

Qn: How does Caching work with DRA and Spark Map Shuffle

Question from Sammy: How do we specify the number of partitions that are used in the beginning?

Q: When Spark 2.0 would be out?

**Answer:** Late May / early June 2016. Please find information [here](https://cwiki.apache.org/confluence/display/SPARK/Wiki+Homepage).

Q: We can not find more information(eg: node, jobs,) from Spark UI when we deploy Spark in Yarn Cluster mode

Answer from Yuxing: Yarn has its own UI, for example, [http://resourceMgr:8088](http://resourcemgr:8088)

Q: Problem is that we can not see useful information with Yarn cluster instead of Spark Standalone cluster

Question from Stu: Can dataframes be cached()?

Answer from Yuxing: yes, save as RDD, use .cache for memory only or persist(level) for other storage options.

Question from Stu: Do dataframes do automatic type inference (since rows are like tuples with multiple types per column)?

Answer from Yuxing: please refer the Spark doc for details, particularly this one: <http://spark.apache.org/docs/latest/sql-programming-guide.html#interoperating-with-rdds>

Question from $$$: when we apply parallelize and specify the number of partitions, is there a threshold on number of partitions and how do we choose the number of partition?based on what metric? For example: **val** rdd **=** sc**.**textFile**(**"/some\_file"**,**3**)**

Qn: is pySpark usually slower than the Scala version?

Question: When do we use the parallelize() function?

Answer from Yuxing: when you have a sequence you want to test it around or persist onto hdfs.

Question: The temp table in data frame api, is it only available during the spark context or it is permanent.

Answer by Tulasi: temp table is available only for duration of the spark job and it is not persistent. btw, it’s sql context that is used to register a temp table.

Question: Please explain how projections are pushed to the RDBMS as well as predicates. By projections I mean column selection from the RDBMS.

Question: How spark works in distributed mode ? Suppose we have 10 nodes and one file has multiple partitions and distributed to all nodes. If we run the job then how all partitioned data will be shown after execution ?

Question: How to configure memory configs for distributed spark cluster ? Can you please mention parameters name and description.

Answer from Yuxing: I found the following 2 articles are particularly helpful:

<http://blog.cloudera.com/blog/2015/03/how-to-tune-your-apache-spark-jobs-part-1/>

<http://blog.cloudera.com/blog/2015/03/how-to-tune-your-apache-spark-jobs-part-2/>

Question:

Q from Ashwini: Data Frames should generally push the filters to the database esp. With the databases like Oracle. But I have experienced that although it does push the filters when applied to the numeric columns it doesn't do so for the date and timestamp columns. In that case it basically does a full table scan and then applies the filter which is very slow.

Q: Overuse of cache() really hits the performance. When should you not use cache()?

**Answer:** Generally, you cache an RDD when you are going to reuse it for later use (future transformations). And, it is always best to never cache an input RDD right away because you may want to perform some ETL such as clean, z

Q: Where does sc exist? On driver only? What if I want to execute some tasks on executor that needs sc, for example, running clustering on a sub-group of data? A use case: we’ve collected a large amount of user data identified by userId, then we want to do clustering per user.

Qn: How does Caching work with DRA and Spark Map Shuffle

Qn: Is their way to destroy the RDD, so that we can free up some memory to load RDD for other file?

**Answer:** There are two things here. RDD as an object and cached RDD. If you are talking about a regular RDD, then, the underlying JVM does it automatically for you if the RDD as part of the garbage collection. If you are talking about the cached RDD, then you can alway uncache it using **unpersist()**

Qn: What happens to the task performed by an Executor, if the executor get killed due to DRA.

Qn: What is a SparkContext?

**Answer:** SparkContext is the entry point to Spark for a Spark application. A SparkContext represents the connection to a Spark cluster, and can be used to create RDDs, accumulators and broadcast variables on that cluster. More information is available on Spark documentation.

**QUESTION:** Where can I find out more how caching works? E.g. is there a caching/eviction policy? Lifetime within context only?

**Answer:** Read [here](http://spark.apache.org/docs/latest/programming-guide.html#rdd-persistence) about Spark persistence. A [blog](https://0x0fff.com/spark-misconceptions/) on LRU Cache.

Question from Stu: Re: Machine Learning -- Regarding either MLLib or ML, can we basically do a one-for-one swap of RDD syntax to DF syntax? Essentially, are DF’s a fairly seamless abstraction that is recognized by MLlib/ML?

Question- Which function, i can use for getting top 10 lines ? (as “head” function) ?

Answer: take(10)

Question: Are there good examples of sqoop on spark and its advantages?

Question -- Isn’t schema inference sometimes dangerous / imprecise in terms of semantics? For example, if I have multiple dates, what kind of dates are they?

Answer from Tulasi: Generally infer schema works well but if you need good control over the datatypes, you can define the schema within the application. If there are multiple fields that are dates, each of it will be inferred as Date datatype. Here is an example for defining a schema (reference: <https://spark.apache.org/docs/1.5.2/api/python/pyspark.sql.html>).

**>>> from** **pyspark.sql.types** **import** \*  
**>>>** schema = StructType([  
**...**  StructField("name", StringType(), True),  
**...**  StructField("age", IntegerType(), True)])  
**>>>** df3 = sqlContext.createDataFrame(rdd, schema)  
**>>>** df3.collect()  
[Row(name=u'Alice', age=1)]

Q from Ashwini: When you read a data from Oracle database, date types are automatically converted into timestamp, why is that? We can change this by updating Oracle dialect. Is there is better way to read dates as dates and not timestamps?

Q: why in the df filter uses “&” rather than “&&”?

Q; Does groupBy on DataFrame does shuffling?

Question: It was mentioned that the data needs to be static or a snapshot. What would happen if the underlying data changes? For eg. if one of the source is a database that gets updated often. Would this result in errors or exceptions being thrown by Spark?

Questions from Wei Su: Is Databrick notebook open-source? Can we use it on our own cluster?

**Answer:** Databricks notebook is only available on Databricks cloud. You cannot use outside of Databricks cloud.

Questions from Wei Su: How do we connect to remote hive cluster from spark? I have tried using JDBC connector and it does not work on spark 1.4.

This guy got the same exception I got.

<http://belablotski.blogspot.com/2016/01/access-hive-tables-from-spark-using.html>

java.sql.Exception: "Method not supported"

Method 1: Pull table into Spark using JDBC

*Conclusion*: the sqlContext.read.format("jdbc") will work with "domestic" cluster only?

Questions from Wei Su: In which scenario will you use checkpoint()?

**Answer:** You use checkpoint() with Spark Streaming so Spark can recover from failures. For more information, please read *checkpointing* section in [Spark Programming Guide](http://spark.apache.org/docs/latest/streaming-programming-guide.html)

Further questions: Thank you. I have read spark programming guide. It does not give an sample code for checkpoint(). I read somewhere that checkedpointed RDD can be reused by other driver programs, how do we reuse them? Are there any sample code for that?

Questions from Wei Su: Could you please define narrow dependency and shuffle dependency? It seems to me that some of the narrow dependencies may also involve shuffles?

Questions from Wei Su: Could you please talk about the difference between Hadoop MapReduce and Spark shuffle process? Or could you direct me to some articles I can read about that?

Q: can you talk a little bit about running Spark on EC2? I’ve been ramping up on Spark by spinning up EMR’s. Is using an EMR necessary when using AWS without DataBricks?

Question from Prince- Can you please share some docs to build spark cluster on hadoop.?

Question from Prince- What types of issues generally users face if they run on large volume of data ? Can you please share some docs.

Question from Prince- How do we know that what memory configs would be ideal for spark jobs ?

Question from prince- Is there any matrix to configure memory size with disk size ? I mean if we have 24 TB of SDD in one node and around 10 TB data is there then how much memory we should allocate for spark jobs ?

Question from Prince- WHat are tuning parameters for shuffling phase?

Question from prince- Can we get databricks kb articles info where we can see that what types of issues are being faced by other customers and what are their resolutions ?

**Answer:** You can visit <https://forums.databricks.com/>

Question from Prince- Is Pyspark slower than scala ?

**Answer:** The performance differs only at the RDD level. When you use DataFrame, there is no performance difference between Scala and Python.

Question: How do we specify/configure job chaining/dependency in Spark?

Question: is Spark UI functionality (spark job status) available through Hue interface?

Question: Is it bad to have a lot of Jobs? Should we try to minimize this number?

Question: What are the key optimizations in spark 2.0 over spark 1.6?