



Spark All Optimizations & Code



Ganesh R

Azure Data Engineer

Lets discuss all the Spark Optimizations which are required to perform in your spark project. Also these are extremely important for you interview preparation.

Let's discuss everything one by one with code.



1. Partitioning:

Partitioning refers to dividing the data into smaller, manageable chunks (partitions) across the cluster's nodes. Proper partitioning ensures parallel processing and avoids data skew, leading to balanced workloads and improved performance.

Code Example



```
# Repartitioning DataFrame to 10 partitions based on a column  
df_repartitioned = df.repartition(10, "column_name")
```



2. Caching and Persistence:

Caching and persistence are used to store intermediate results in memory, reducing the need for recomputation. This is particularly useful when the same DataFrame is accessed multiple times in a Spark job.

Code Example

```
# Caching DataFrame in memory
df.cache()
df.show()

# Persisting DataFrame with a specific storage level (Memory and Disk)
from pyspark import StorageLevel
df.persist(StorageLevel.MEMORY_AND_DISK)
df.show()
```



3. Broadcast Variables:

Broadcast variables allow the distribution of a read-only variable to all nodes in the cluster, which can be more efficient than shipping the variable with every task. This is particularly useful for small lookup tables.

Code Example

```
# Broadcasting a variable  
broadcastVar = sc.broadcast([1, 2, 3])
```



4. Avoiding Shuffles:

Shuffles are expensive operations that involve moving data across the cluster. Minimizing shuffles by using map-side combine or careful partitioning can significantly improve performance.

Code Example

```
# Using map-side combine to reduce shuffle  
rdd = rdd.map(lambda x: (x, 1)).reduceByKey(lambda x, y: x + y)
```



5. Columnar Format:

Using columnar storage formats like Parquet or ORC can improve read performance by allowing Spark to read only the necessary columns. These formats also support efficient compression and encoding schemes.

Code Example

```
# Saving DataFrame as Parquet  
df.write.parquet("path/to/parquet/file")
```



6. Predicate Pushdown:

Predicate pushdown allows Spark to filter data at the data source level before loading it into memory, reducing the amount of data transferred and improving performance.

Code Example

```
# Reading data with predicate pushdown
df = spark.read.parquet("path/to/parquet/file").filter("column_name >
100")
```



7. Vectorized UDFs (Pandas UDFs):

Vectorized UDFs, also known as Pandas UDFs, utilize Apache Arrow to process batches of rows, improving performance compared to row-by-row processing in standard UDFs.

Code Example



```
from pyspark.sql.functions import pandas_udf, PandasUDFType

@pandas_udf("double", PandasUDFType.SCALAR)
def vectorized_udf(x):
    return x + 1

df.withColumn("new_column",
vectorized_udf(df["existing_column"])).show()
```



8. Coalesce:

Coalesce reduces the number of partitions in a DataFrame, which is more efficient than repartitioning when decreasing the number of partitions.

Code Example

```
# Coalescing DataFrame to 1 partition  
df_coalesced = df.coalesce(1)
```



9. Avoid Using Explode:

Explode is an expensive operation that flattens arrays into multiple rows. Using it should be minimized, or optimized by reducing the size of the DataFrame before exploding.

Code Example

```
# Using explode function
from pyspark.sql.functions import explode
df_exploded = df.withColumn("exploded_column",
explode(df["array_column"]))
```



10. Tungsten Execution Engine:

Tungsten is Spark's in-memory computation engine that optimizes the execution plans for DataFrames and Datasets, utilizing memory and CPU more efficiently.

Code Example



Tungsten is enabled by default in Spark, so no specific code is needed. However, using DataFrames and Datasets ensures you leverage Tungsten's optimizations.



11. Using DataFrames/Datasets API:

The `DataFrames/Datasets` API provides higher-level abstractions and optimizations compared to `RDDs`, including Catalyst Optimizer for query planning and execution.

Code Example



```
# Using DataFrames API
df = spark.read.csv("path/to/csv/file")
df = df.groupBy("column_name").agg({"value_column": "sum"})
```



12. Join Optimization:

Broadcast joins are more efficient than shuffle joins when one of the DataFrames is small, as the small DataFrame is broadcasted to all nodes, avoiding shuffles.

Code Example



```
# Broadcast join
from pyspark.sql.functions import broadcast
df = df1.join(broadcast(df2), df1["key"] == df2["key"])
```



13. Resource Allocation:

Properly allocating resources such as executor memory and cores ensures optimal performance by matching the resource requirements of your Spark jobs.

Code Example



Resource allocation is typically done through Spark configurations when submitting jobs:

```
spark-submit --executor-memory 4g --executor-cores 2  
your_script.py
```



14. Skew Optimization:

Handling skewed data can improve performance. Techniques like salting (adding a random key) can help distribute skewed data more evenly across partitions.

Code Example



```
# Handling skewed data by salting
from pyspark.sql.functions import rand
df_salts = df.withColumn("salt", (rand() * 10).cast("int"))
df_salts_repartitioned = df_salts.repartition("salt")
```



15. Speculative Execution:

Speculative execution re-runs slow tasks in parallel and uses the result of the first completed task, helping to mitigate the impact of straggler tasks.

Code Example

```
# Enabling speculative execution  
spark.conf.set("spark.speculation", "true")
```



16. Adaptive Query Execution (AQE):

AQE optimizes query execution plans dynamically based on runtime statistics, such as the actual size of data processed, leading to more efficient query execution.

Code Example

```
# Enabling AQE
spark.conf.set("spark.sql.adaptive.enabled", "true")
```



17. Dynamic Partition Pruning:

Dynamic Partition Pruning improves the performance of join queries by dynamically pruning partitions at runtime, reducing the amount of data read.

Code Example

```
# Enabling dynamic partition pruning  
spark.conf.set("spark.sql.dynamicPartitionPruning.enabled", "true")
```



18. Reduce Task Serialization Overhead:

Using Kryo serialization can reduce the overhead associated with task serialization, improving performance compared to the default Java serialization.

Code Example

```
# Enabling Kryo serialization
spark.conf.set("spark.serializer",
"org.apache.spark.serializer.KryoSerializer")
```



19. Reduce Shuffle Partitions:

By default, Spark has a high number of shuffle partitions (200). Reducing this number can improve performance, especially for small datasets.

Code Example

```
# Reducing shuffle partitions  
spark.conf.set("spark.sql.shuffle.partitions", "50")
```



20. Using Data Locality:

Ensuring that data processing happens as close to the data as possible reduces network I/O, leading to faster data processing.

Code Example



Data locality is handled by Spark's execution engine, but users can influence it by configuring their cluster properly and using locality preferences in their code.



21. Leveraging Built-in Functions

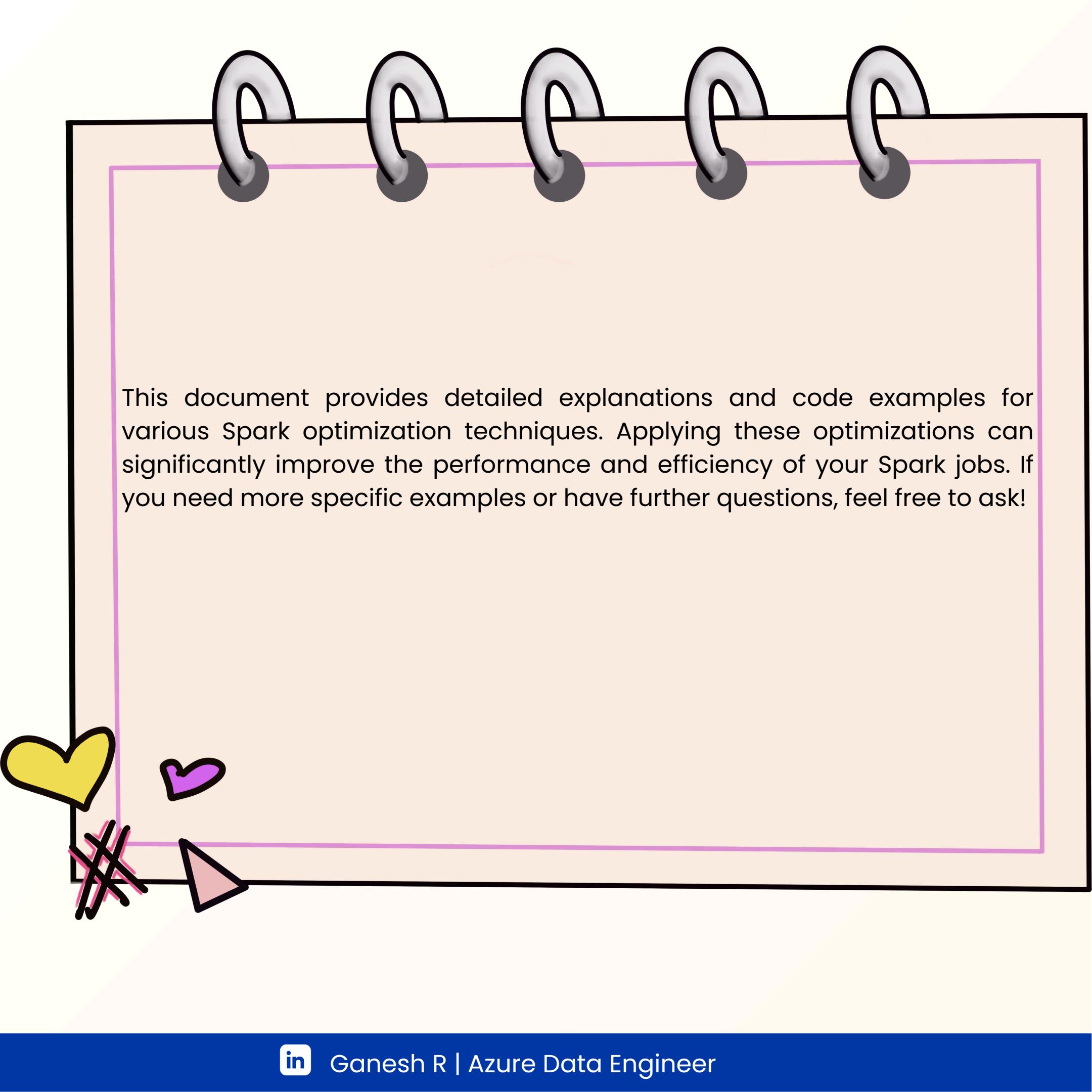
Explanation:

Built-in functions are optimized for performance and should be preferred over custom UDFs, which can introduce significant overhead.

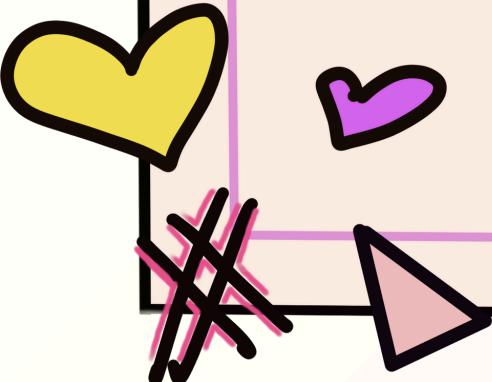
Code Example

```
# Using built-in functions
from pyspark.sql.functions import col, expr
df.select(col("column_name").alias("new_column_name")).show()
```





This document provides detailed explanations and code examples for various Spark optimization techniques. Applying these optimizations can significantly improve the performance and efficiency of your Spark jobs. If you need more specific examples or have further questions, feel free to ask!



**Follow for more content
like this**



Ganesh R
Senior Azure Data Engineer

