1. RDD Exploration

Assignment 1 focuses on exploring Resilient Distributed Datasets (RDDs) within Spark, which serve as the foundational data structure in Spark's programming model. The goal of this assignment is to gain a deeper understanding of RDDs and their operations through practical exploration.

The assignment tasks may include:

1. **Creating RDDs**: Students may be tasked with creating RDDs from various data sources such as text files, CSV files, or through parallelizing collections.
2. **Transformations**: Exploring RDD transformations involves applying operations like `map`, `filter`, `flatMap`, `reduceByKey`, etc., to manipulate the data within RDDs. Students might be asked to perform transformations to preprocess data, filter out irrelevant information, or aggregate data based on certain keys.
3. **Actions**: Students will also explore RDD actions, which trigger the execution of transformations and return results to the driver program. Examples of actions include `collect`, `count`, `take`, `reduce`, etc. They might be required to perform actions to retrieve results or save RDD data to external storage.
4. **Performance Optimization**: As part of the exploration, students may be encouraged to analyze the performance of RDD operations and identify opportunities for optimization. This could involve techniques such as partitioning, caching, and leveraging built-in optimizations provided by Spark.
5. **Error Handling and Debugging**: Dealing with errors and debugging Spark applications is an essential skill. Students may encounter scenarios where they need to troubleshoot errors, understand Spark's execution model, and optimize code for better performance.
6. **Documentation and Reporting**: Finally, students might be required to document their exploration process, including the RDD operations performed, insights gained, and any challenges faced during the assignment. Clear and concise reporting is vital for conveying the learnings and observations effectively.

Overall, Assignment 1 on RDD exploration provides students with hands-on experience in working with distributed datasets in Spark, honing their skills in data manipulation, performance optimization, and troubleshooting within the Spark framework.

Link to DataBricks Notebook