Flarion Low-Level Engineer Interview Task

The following task includes two parts: Greatest implementation and Spark GroupBy Analysis. Together the tasks should take approximately 4 hours to complete. Please read both task descriptions before starting and plan your time accordingly.

Regex Implementation

Introduction

In this task you will be required to build and run a low level software library, make changes to it, and test the changes you made to guarantee robust software. Please share your thinking in documentation whenever you encounter design or implementation dilemmas.

For engineers that are proficient in Rust, the task should take 3 hours to complete.

Background

1. Datafusion is a leading query engine implemented in Rust, allowing users to maximize the efficiency of SQL operations.
2. Spark is a unified analytics engine for large-scale data processing, allowing users to run complex SQL operations at scale.
3. Spark offers a wide range of functions for data processing, as detailed here.

Task

Implement the Spark function "regexp_extract" in Datafusion. Please do not use Datafusion's SQL API but feel free to use any other tool or interface that Datafusion provides.

Consider edge cases and write tests to validate your solution.

Spark GroupBy Analysis

GroupBy is a key data processing function. While naive GroupBy implementations are inefficient, it's possible to use more sophisticated implementations to massively improve performance.
Please explain what Spark does to optimize GroupBy and point to the relevant locations in the Spark codebase where this is done.

Good luck!