Instituto de Ciências Exatas
Departamento de Ciência da Computação
Pós-Graduação em Computação Aplicada

Universidade de Brasília

How to mitigate data skew and optimize Spark jobs: The 3 biggest challenges when creating and maintaining Spark pipelines on Big Data

(A palestra será em Português, com slides em Inglês)

Otto von Sperling

Senior Data Scientist
ML & Data Engineering Platform Team
Wise (formerly TransferWise), Estonia

Local
Sala Multiuso do CIC no Prédio CIC/EST
(entre o Pavilhão João Calmon e o Centro Comunitário Athos Bulcão)
Universidade de Brasília, Campus Darcy Ribeiro

Dia 22/02/2024, às 16:45 horas.

Abstract

Otto von Sperling is a Data Scientist with over 5 years of experience in Computer Science and Financial Crime prevention. He is currently at TransferWise (Wise) as a Senior Data Scientist, where he is one of the go-to people when it comes to batch data pipelines and Apache Spark. In this talk he will talk about his professional experience in mitigating data distortions and optimizing Spark jobs in a large Fintech. The golden rule is to always try to find opportunities instead of complaining about problems and, when looking for solutions, always try to keep in mind the analysis of impact versus effort/cost to avoid over-optimization and never shy away from taking responsibility. Otto moved from financial crime risk modeling throughout 2022 to starting a new ML and data engineering team in Q1 2023 to focus on:
1. Increase the productivity of data professionals on risk teams
2. Enable product teams to create and consume data products quickly and easily
3. increase governance and observability of data and ML systems

The talk about the 3 biggest challenges in creating and maintaining Spark pipelines in Big Data is structured as follows:

1. Introduction
2. Parallelism and Partitioning
3. Data Distortion
4. Data Structure for Storage
5. Bonus: AWS EMR vs GCP Dataproc

.

Otto von Sperling, former CIC/UnB graduate student, is a Senior Data Scientist at Wise. His team's most recent major development is Wise's first internal Feature Platform to scale the impact and speed of Data Science and Analytics across the company

otto@vonsperling.com.br, linkedin.com/in/otto-von-sperling