9th International Conference "Distributed Computing and Grid Technologies in Science and Education" (GRID'2021)

GRID
2021
July 5 - 9 Dubna
Distributed Computing and Grid-technologies in Science and Education
JINR
MLIT
GRID 2021

Contribution ID: **115** Type: **Sectional reports**

SQL query execution optimization on Spark SQL

Tuesday, 6 July 2021 16:20 (15 minutes)

The Spark –Hadoop ecosystem includes a wide variety of different components and can be integrated with any tool required for Big Data nowadays. From release-to-release developers of these frameworks optimize the inner work of components and make their usage more flexible and elaborate.
Anyway, since inventing MapReduce as a programming model and the first Hadoop releases data skew was and remains the main problem of distributed data processing. Data skew leads to performance degradation i.e., common slowdown of application execution and idle of the resources. The newest Spark framework versions allow handling this situation easily from the box. However, there is no opportunity to upgrade versions of tools and appropriate logic in the case of huge projects in which development was started years ago.
In this article, we consider approaches to execution optimization of SQL query in case of data skew on concrete example with HDFS and Spark SQL 2.3.2 version usage.

Summary

Primary author: MOZHAISKII, Gleb

Co-authors: KORKHOV, Vladimir (St. Petersburg State University); GANKEVICH, Ivan (Saint Petersburg State University)

Presenter: MOZHAISKII, Gleb

Session Classification: Big data Analytics and Machine learning.

Track Classification: 9. Big data Analytics and Machine learning