

The Internals of Spark SQL

Introduction

Spark SQL — Structured Data Processi...

Datasets vs DataFrames vs RDDs

Dataset API vs SQL

HIVE INTEGRATION / HIVE DATA SOURCE

Hive Data Source

Demo: Connecting Spark SQL to Hiv...

Demo: Hive Partitioned Parquet Tabl...

Configuration Properties

Hive Metastore

DataSinks Strategy

HiveFileFormat

HiveClient

HiveClientImpl

HiveUtils

IsolatedClientLoader

HiveTableRelation

CreateHiveTableAsSelectCommand

SaveAsHiveFile

InsertIntoHiveDirCommand

InsertIntoHiveTable

Case Study: Number of Partitions for groupBy Aggregation

Important	As it fairly often happens in my life, right after I had described the discovery I found out I was wrong and the "Aha moment" was gone.
	Until I thought about the issue again and took the shortest path possible. See Case 4 for the definitive solution.
	I'm leaving the page with no changes in-between so you can read it and learn from my mistakes.

The goal of the case study is to fine tune the number of partitions used for `groupBy` aggregation.

Given the following 2-partition dataset the task is to write a structured query so there are no empty partitions (or as little as possible).

