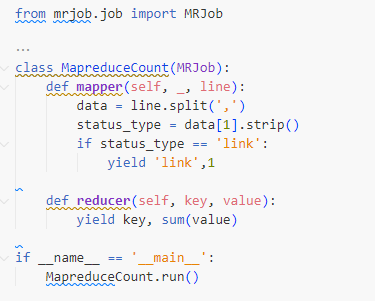
MapReduce Count

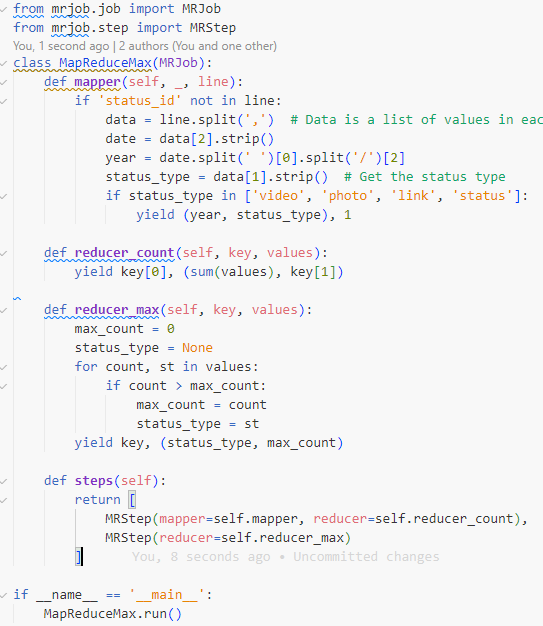


MapReduce Average

A computer screen shot of a code

Description automatically generated

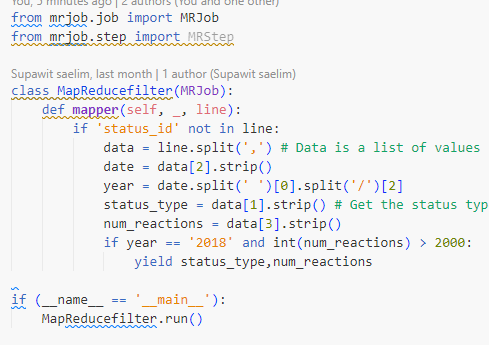
MapReduce MAX



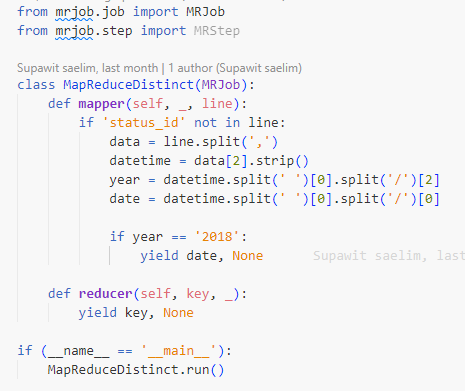
MapReduce Top-N



MapReduce Filter



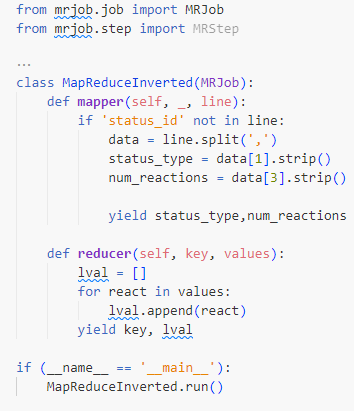
MapReduce Distinct



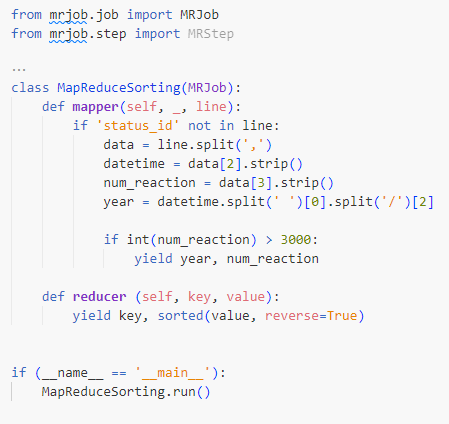
MapReduce Binning



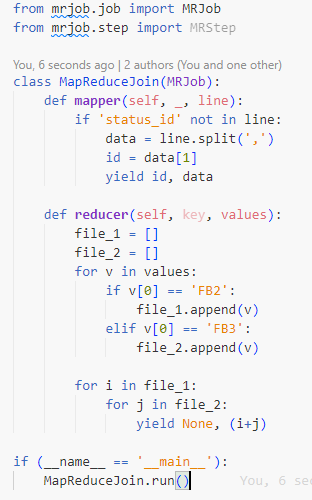
MapReduce Inverted Index



MapReduce Sorting



Inner Join



MapReduce Left Outer Join

A screenshot of a computer code

Description automatically generated

MapReduce Right Outer Join

A screenshot of a computer code

Description automatically generated

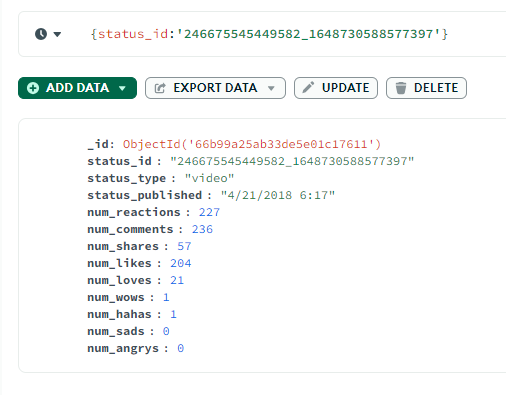
Full Outer Join

A screenshot of a computer program

Description automatically generated

Mongo DB

Where:



Where … And …

A computer code with red and blue text

Description automatically generated

Or

A computer code with text

Description automatically generated with medium confidence

And with Or

{num\_likes: {$eq:500}, $or:[{num\_reactions:{$gt:3000}},{num\_comments:{$gt:10}}]}

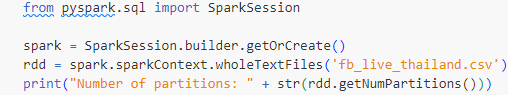
RDD

Simple RDD with alphabet and fb\_live\_thailand ถ้าไม่อยากกำหนด patitions ก็ไม่ต้องใส่

A screen shot of a computer program

Description automatically generated

**wholeTextFiles(path)**: ใช้เพื่ออ่านไฟล์หลายไฟล์ (หรือไฟล์เดียว) โดยจะโหลดทั้งไฟล์เป็นแต่ละ record ใน RDD และแต่ละ record จะเป็นคู่ (key-value) โดยที่ key คือ path ของไฟล์และ value คือเนื้อหาทั้งหมดของไฟล์



Distinct



Filter กรองบรรทัดที่มีค่าในคอลัมน์ที่สอง (คอลัมน์ที่แยกจากกันด้วยคอมมา) เป็น 'link' และแสดงผลลัพธ์ที่กรองได้



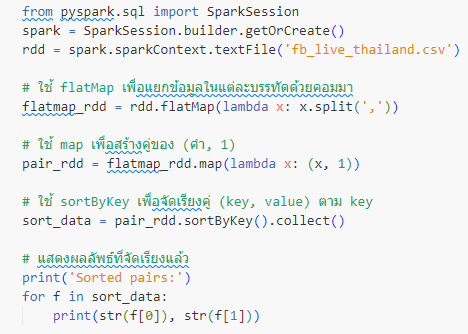
Map



Flapmap



SortByKey



SortBy



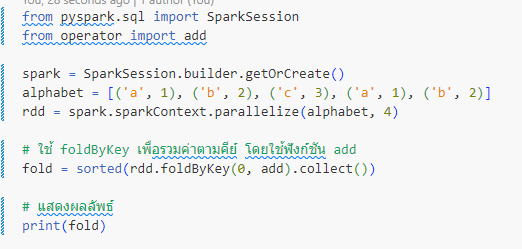
ReduceByKey



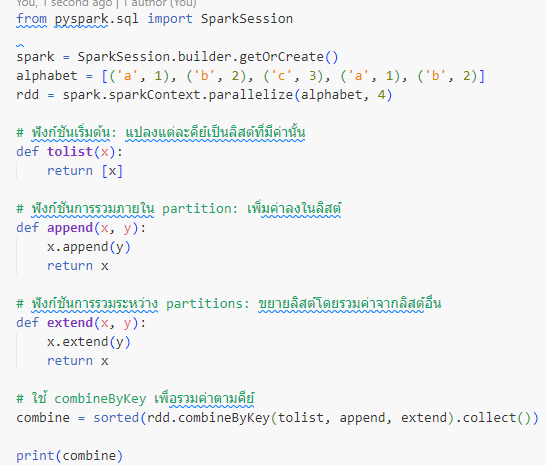
aggregateByKey



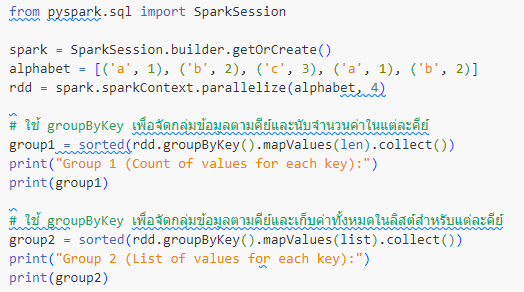
foldByKey



combineByKey



groupByKey



Join leftOuterJoin rightOuterJoin

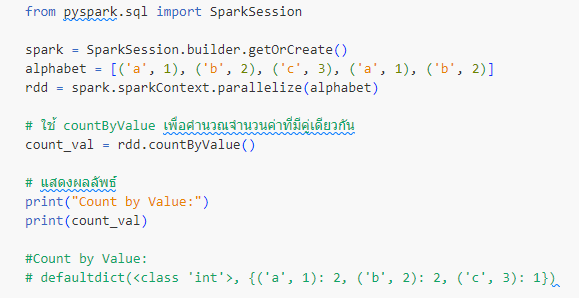


Action

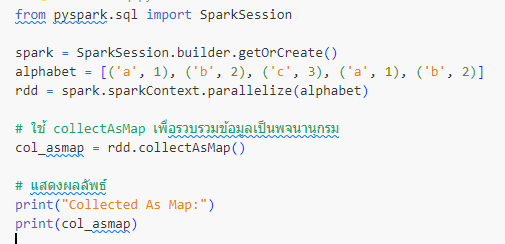
countByKey



countByValue



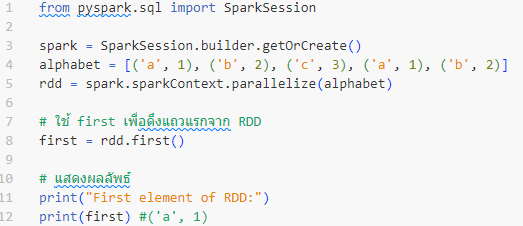
collectAsMap



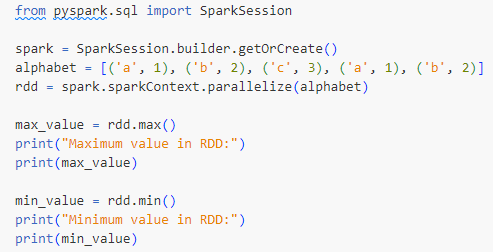
lookup



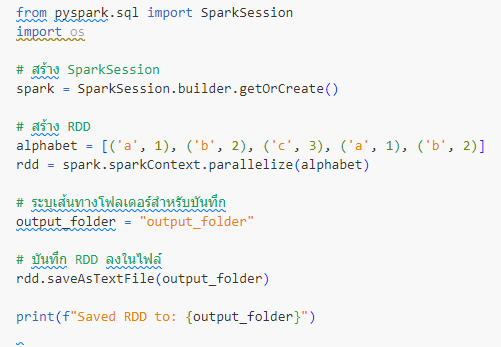
First



Mix-Max

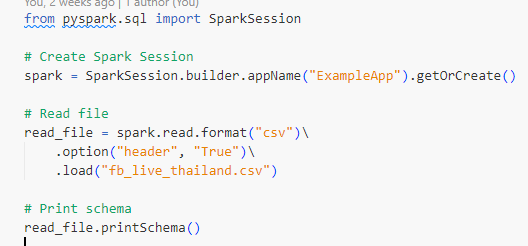


saveAsTextFile



High Level

Dataframes



Spark SQL query

A screen shot of a computer code

Description automatically generated

Where SQL



Temporary Views

A screenshot of a computer program

Description automatically generated

Save Modes (errorifexists, append, overwrite, ignore)



randomSplit



Aggregation Functions

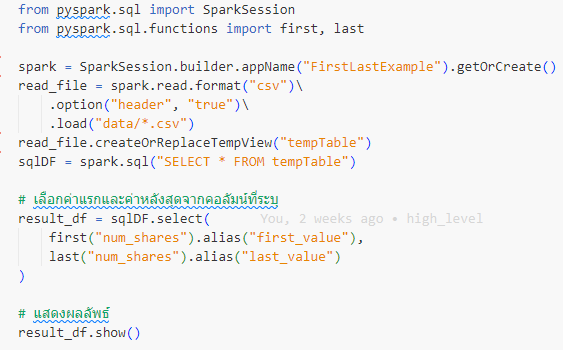
Count



countDistinct



first and last



Mix\_max



Sum and sumdistinct



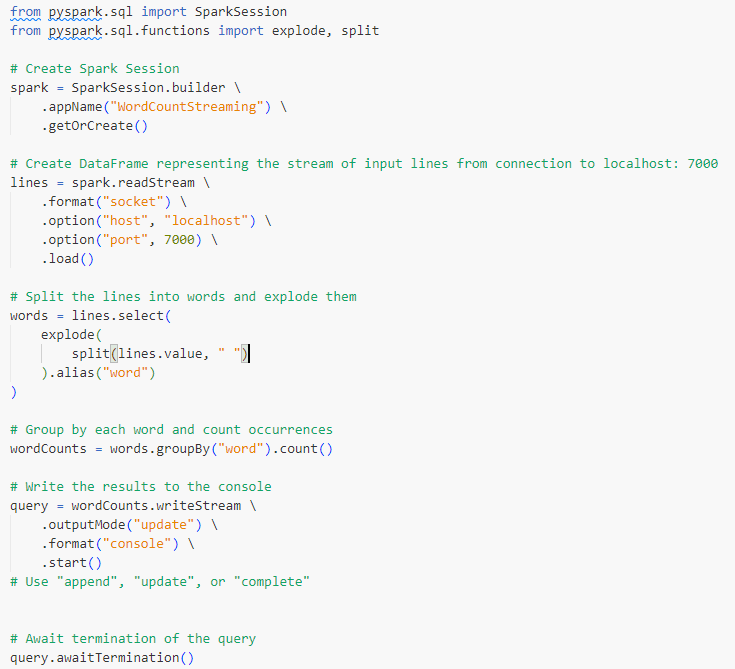
avg



Join (“left\_outer”, “right\_outer”,”inner”,”outer”



Socket Source



ncat ใช้สำหรับส่งข้อมูลเข้าไปยัง Spark Streaming ผ่าน socket

Spark Streaming รับข้อมูล, ประมวลผลข้อมูล, และแสดงผลลัพธ์

ความซิงค์กันเกิดจากการที่ Spark Streaming รับข้อมูลจาก ncat ผ่านการเชื่อมต่อ socket ที่กำหนดไว้

**File Source**



1. โค้ดนี้ใช้ Apache Spark เพื่อประมวลผลข้อมูลแบบสตรีมมิ่งจากไฟล์ CSV
2. กำหนด schema สำหรับไฟล์ CSV เพื่อให้การอ่านข้อมูลมีความถูกต้อง
3. ข้อมูลจะถูกประมวลผลโดยการแยกวันที่จากคอลัมน์ status\_published และจัดกลุ่มตามวันที่และประเภทสถานะ
4. ผลลัพธ์จะถูกเขียนไปยังคอนโซลในโหมด "complete" ซึ่งหมายถึงการแสดงผลลัพธ์ทั้งหมดในแต่ละครั้งของการประมวลผล

Source and Sink อย่าลืมเปลี่ยน path



A close-up of a computer screen

Description automatically generated

Event-time

