Week 2, May 5th

workshop/pyspark-notebooks/

S005-

Ps -e | grep python -> list all python running processor on the system

1. Coalesce: S005

Jupyterlab: S005- RepartitionCoalesce

Number of partitions = number of tasks

There’s a task even the partition is empty

“Coalesce” is used to reduce the number of partitions, check S005

1. FlatMap: , IS A TRASFORMATION METHOD. S006

can be used to extract the data into one single list or array – check S006-FlatMap

1. KeyValuePairRdd. IS AN ACTION METHOD S007?

*# Key Value pair formed from tuple, whereas first element in tuple is known as key*

*# Second element known as value*

*# (key, value)*

1. countByKey (): result is dictionary, it returns count of keys WHILE THE INPUT DATA is a tuple – check S007
2. *reduceByKey (): transformation method - reduce here refer to aggregate. If there are for example we have 2* tuples with the same key in our data, we can perform “arithmetic” operations on the values such as the sum. How? by using countByKey () method + adding to it a function, e.g.: “countByKey **(lambda** acc, value: acc **+** value)”
3. groupby() groupbykey()
4. Read data from hadoop S008-

* Include data manipulation, read records, select, header,
* convert the printed lines(list) into tuple: list is mutable, tuple immutable , can perform modification only on list, modify before you convert

S008

Graphical user interface, text, application

Description automatically generated

**S013-BroadCast:**

1. **without Boradcast**

Diagram

Description automatically generated

Diagram

Description automatically generatedDiagram

Description automatically generatedDiagram

Description automatically generated

1. **with broadcast**

**Diagram

Description automatically generated**

**S014 CACHING**

**Graphical user interface, text, application

Description automatically generated**

**When do you cache?**

**Graphical user interface

Description automatically generated**