Tomas Polacek (tompo940), Paul Thiele (pauth144)

Lab BDA1 – Spark – Exercises

Task 1

spark with station numbers:
tempMax:
(u'1975', (36.1, u'86200'))
(u'1992', (35.4, u'63600'))
(u'1994', (34.7, u'117160'))
…
(u'1965', (28.5, u'116500'))
(u'1951', (28.5, u'75040'))
(u'1962', (27.4, u'86200'))

tempMin:
(u'1990', (-35.0, u'147270'))
(u'1952', (-35.5, u'192830'))
(u'1974', (-35.6, u'166870'))
…
(u'1978', (-47.7, u'155940'))
(u'1999', (-49.0, u'192830'))
(u'1966', (-49.4, u'179950'))

Description	Submitted	Duration
saveAsTextFile at NativeMethodAccessorImpl.java:-2	2018/04/20 10:51:19	0.5 s
saveAsTextFile at NativeMethodAccessorImpl.java:-2	2018/04/20 10:51:18	0.6 s
sortBy at temp-min-max.py:13	2018/04/20 10:51:18	94 ms
sortBy at temp-min-max.py:13	2018/04/20 10:50:15	1.0 min
sortBy at temp-min-max.py:12	2018/04/20 10:50:15	0.1 s
sortBy at temp-min-max.py:12	2018/04/20 10:49:12	1.1 min

Time needed: 2.3 min

sequential program:
Maximum Temperatures:
(1975, (36.1, '86200'))
(1992, (35.4, '63600'))
(1994, (34.7, '117160'))
...
(1951, (28.5, '75040'))
(1965, (28.5, '116500'))
(1962, (27.4, '76380'))
Minimum Temperatures:
(1990, (-35.0, '147270'))
(1952, (-35.5, '192830'))
(1974, (-35.6, '166870'))

...
(1978, (-47.7, '155940'))
(1999, (-49.0, '192830'))
(1966, (-49.4, '179950'))
Time needed: 269.84s

How does the runtime compare to the Spark version?
When executing the task locally we have a runtime of 270 seconds. As we can observe this is about twice as long as the execution on hdfs spark. As we parallelize the computation we expect a faster execution and shorter runtimes. However, we also add overhead, therefore, the computation is only twice as fast.

Task 2

(u'1996-10', 22811)
(u'1974-07', 66277)
(u'2003-05', 48264)
(u'1986-11', 1198)
(u'1978-03', 306)
(u'1981-10', 9882)
…

with only one count per station:
('1982-04', 246)
('1963-04', 283)
('1994-05', 299)
('1990-09', 312)
('2000-08', 325)
('1953-04', 104)
…

Task 3

(u'99090;1977-04', 1.6933333333333331)
(u'71360;2005-01', 3.7822580645161294)
(u'72160;1992-11', 4.2283333333333335)
(u'64520;1979-11', 2.9616666666666664)
(u'172770;1998-01', -9.948387096774194)
(u'132030;2000-10', 3.124193548387097)
…

Task 4

The result is empty after the join.

Task 5

('2009-05', 54.166666666666686)
('2016-04', 26.90000000000001)
('2009-08', 61.566666666666684)
('2014-09', 48.45000000000001)
('1998-05', 38.36666666666668)
('1999-05', 27.383333333333344)
('1999-01', 61.933333333333394)
('2008-11', 46.75000000000003)
…

Task 6

(u'2003-12', 1.8367157746866383)
(u'1974-07', -1.1745224038304158)
(u'2003-05', 0.2756803862110839)
(u'1986-11', 2.776498170716614)
(u'2004-04', 1.2260719468703334)
(u'1981-10', -0.6171810863434173)
(u'1981-03', -0.42682936903540447)
(u'1999-06', -0.4124652877693684)
...