Σχολη Επιστημης Υπολογιστων-Πανεπιστημιο Κρητης

CS543-Software Systems and Technologies for Big Data Applications

Εαρινο Εξαμηνο 2020-2021

Programming Assignment 1

Φοιτητής
Καλογερακης Στεφανος
AM:1205
Email:skalogerakis97@gmail.com

Διδάσκων
Χ. Κοζανιτης

Εισαγωγή

Στα πλαίσια της πρώτης εργασίας του μαθήματος CS543 σκοπός ήταν η εξοικείωση με την γλώσσα προγραμματισμού Scala και το προγραμματιματιστικό μοντέλου του functional programming. Παράλληλα ήρθαμε σε πρώτη επαφή με το framework του Apache Spark. Το πρώτο μέρος της εργασίας, αφορούσε κάποια απλά σενάρια με την γλώσσα Scala, ενώ στο δεύτερο ασχοληθήκαμε με την επεξεργασία ενός dataset και πάλι στην γλώσσα Scala. Η συγκεκριμένη αναφορά ασχολείται κυρίως με το δεύτερο κομμάτι εργασίας, παρουσιάζοντας τα αποτελέσματα όπως ζητείται από την εκφώνηση.

Άσκηση 1

Το πρώτο σκέλος ήταν αφιερωμένο αποκλειστικά στην γλώσσα Scala. Στην συγκεκριμένη αναφορά δεν αναγράφονται παραπάνω λεπτομέρειες για την υλοποίηση της κάθε συνάρτησης αφού κάτι τέτοιο δεν απαιτείται απο την εκφώνηση. Ο κώδικας περιέχει επεξηγηματικά σχόλια για τον τρόπο σκέψης και υλοποίησης του κάθε σεναρίου, καθώς και μια σειρά απο test cases που αποδεικνύουν την ορθότητα λειτουργίας.

Οδηγίες εκτέλεσης

Η εκτέλεση κατά το development πραγματοποιήθηκε κατά κύριο λόγο στο περιβάλλον IntelliJ με έκδοση scala 2.13. Η εκτέλεση απο terminal πραγματοποιείται με τις εντολές

- **scalac ScalaWamUp.scala** και
- **scala ScalaWarmUp**

Στην main συνάρτηση εχουν δημιουργηθεί μια σειρά από demo test που επιβεβαιώνουν την ορθή λειτουργία βάσει της εκφώνησης

Άσκηση 2

Το δεύτερο κομμάτι της εργασίας περιλαμβάνει την χρήση του εργαλείου Apache Spark. Λόγω του μεγέθους του dataset και χαμηλών απαιτήσεων hardware, αρκούσε για την εξαγωγή των αποτελεσμάτων τοπική εκτέλεση με standalone cluster του Spark.

Οδηγίες εκτέλεσης

Η εκτέλεση κατά το development πραγματοποιήθηκε κατά κύριο λόγο στο περιβάλλον IntelliJ IDEA με την χρήση pom.xml αρχείο για την φόρτωση των απαραίτητων dependencies. Για την αποφυγή περαιτέρω προβλημάτων με τα dependencies παρακάτω παρατίθενται οδηγίες για την εκτέλεση του κώδικα μέσω spark shell σε κατάλληλα τροποποιημένο αρχείο.

- Μετάβαση στον φάκελο bin του directory του spark
- Εκτέλεση **./spark-shell -i <File_dir>/FunWithApacheSpark.scala**
- Στο περιβάλλον του spark shell εκτέλεση της εντολής **SimpleApp.main(Array(""))**

Η main συνάρτηση περιέχει όλα τα υποερωτήματα και τα αποτελέσματα θα εμφανίζονται σειριακά καθώς εκτελούνται.

Στα τελικά παραδοτέα το αρχείο **FunWithApacheSpark.scala** είναι το τροποποιημένο αρχείο που λειτουργεί με **spark shell**, ενώ το **FunWithApacheSpark(INTELLIJ).scala** είναι το αρχείο που εκτελείται στο **IntelliJ**.

ΠΑΡΑΤΗΡΗΣΗ: Πραγματοποιήθηκαν κάποιες δοκιμές στο spark-shell και το αρχείο ScalaWarmUp.scala παράγει σωστά αποτελέσματα. Σε περίπτωση οποιασδήποτε ασυνέπειας με τα αποτελέσματα στις επόμενες ενότητες ή πρόβλημα εκτέλεσης παρακαλώ επικοινωνήστε μαζί μου για περισσότερες πληροφορίες σχετικά με το build στο περιβάλλον Intellij

2.2 - Data Exploration

Ακολουθώντας τις οδηγίες της εκφώνησης στην baseRDD φορτώθηκε το dataset προς επεξεργασία (όνομα αρχείου: NASA_access_log_Jul95) σε μορφή RDD[String]. Σαν επόμενο βήμα ήταν ο μετασχηματισμός της baseRDD σε RDD[Log] μορφή, με την συνάρτηση **getLogFiles**. Την στιγμή που έγινε trigger με την count function προέκυψαν αρκετά σφάλματα της μορφής **java.util.NoSuchElementException**, το οποίο ήταν αναμενόμενο όπως ανέφερε η εκφώνηση.

2.2.1

Αναλύοντας τόσο την συμπεριφορά των patterns και του dataset, και με την βοήθεια της συνάρτησης **findFirstIn()** παρατηρήθηκαν τα προβλήματα που απέτρεπαν από την επιτυχή εκτέλεση του προγράμματος. Χωρίς να διαφοροποιηθεί καθόλου το dataset όλα τα πεδία παρουσιάζουν προβλήματα για διαφορετικούς λόγους που αναλύονται στην επόμενη ενότητα. Το μεγαλύτερο πρόβλημα παρατηρείται στο πεδίο Bytes, όμως για λόγους πληρότητας παρουσιάζονται όλα τα διαφορετικά προβλήματα που συναντήθηκαν.

2.2.1

Στον παρακάτω πίνακα παρουσιάζονται τα διαφορετικά πεδία και ο λόγος που δημιουργείται πρόβλημα στο συγκεκριμένο πεδίο

Πεδίο	**Πρόβλημα**
Bytes	Στο συγκεκριμένο πεδίο το regular expression αναμένει κάποιο αριθμό(συμβολίζει τα bytes) και να ακολουθεί το EOL(End Of Line). Σε αρκετές περιπτώσεις συναντάται ο χαρακτήρας "-" αντί για τον ζητούμενο αριθμό
requestURI	Κάποιες λίγες περιπτώσεις συναντάται πρόβλημα στο encoding αντί για το κανονικό κείμενο που αναμένεται και αναπόφευκτα προκαλείται exception
Host, Date, Status	Πρόβλημα σε αυτά τα πεδία δημιουργεί η τελευταία γραμμή του εγγράφου η οποία περιέχει την εγγραφή "alyssa.p". Το πεδίο Host δεν αναμένει το τέλος της γραμμής σε αυτό το σημείο και αναμένει επόμενο πεδίο, ενώ αντίστοιχα τα πεδία Date και Status είναι κενά ενώ αναμένεται κάποιας μορφής κείμενο

Χαρακτηριστικό παράδειγμα από τα διαφορετικά προβλήματα που επισημάνθηκαν στον πίνακα

- **Bytes**
  dynip42.efn.org - - [01/Jul/1995:00:02:14 -0400] ”GET /software HTTP/1.0” 302 -
- **requestURI**
  128.159.122.20 - - [20/Jul/1995:15:28:50 -0400] ”k??tx??tG??t??” 400 -

2.2.3

Η συνάρτηση **getImprovedLogFields** χρησιμοποιήθηκε για την αντιμετώπιση όλων των παραπάνω προβλημάτων. Η συγκεκριμένη συνάρτηση είναι τροποποιημένη έκδοση της **getLogFiles**, με την βελτίωση ότι όποιο πεδίο δεν συμβαδίζει με το pattern του εκάστοτε regular expression αντικαθίσταται στην περίπτωση των αριθμών με 0 και στην περίπτωση των Strings με κενό “”

2.3 - Walk through on the Web Server Log File

2.3.1 - Explore content size

Function	Result
MAX	6823936
MIN	0
AVG	20455.49857721697

2.3.2 - HTTP status analysis

Status	Frequency
STATUS: 200	FREQ: 1701534
STATUS: 304	FREQ: 132627
STATUS: 302	FREQ: 46573
STATUS: 404	FREQ: 10845
STATUS: 500	FREQ: 62
STATUS: 403	FREQ: 54
STATUS: 501	FREQ: 14
STATUS: 400	FREQ: 5
STATUS: 0	FREQ: 1

2.3.3 - Frequency Hosts

Hosts	Frequency
192.188.119.226	22
192.156.154.71	12
n1122099.ksc.nasa.gov	158
line115.worldweb.net	53
194.108.167.34	12
129.108.37.35	18
ix-phx4-29.ix.netcom.com	33
ad10-013.compuserve.com	44
ac_anx_1_5.mur.csu.edu.au	12
sbd0114.deltanet.com	22

2.3.4 - Top-10 Error paths

Error Paths	Frequency
/images/NASA-logosmall.gif	21010
/images/KSC-logosmall.gif	12435
/images/MOSAIC-logosmall.gif	6628
/images/USA-logosmall.gif	6577
/images/WORLD-logosmall.gif	6413
/images/ksclogo-medium.gif	5837
/images/launch-logo.gif	4628
/shuttle/countdown/liftoff.html	3509
/shuttle/countdown/	3345
/shuttle/countdown/images/cdtclock.gif	3251

2.3.5 - Unique Hosts

Number of unique hosts: **81983**

2.3.6 - 404 Response Codes

Number of 404 Response Codes: **10845**

2.3.7 - Distinct requestURIs with 404 errors

/history/apollo/apollo15.html
/shuttle/missions/sts-63/sts-63-press-kit.tx”
/history/apollo/publications/sp-350/sp-350.txt
/de/systems.html
/shuttle/technology/ml
/history/skylab/skylab-overview.txt
/history/apollo/appollo-13/
/shuttle/technology/sts-newsref/et.html#et-safety
/history/apollo/apollo-13/apollo-13.htmlhppt:
/shuttle/missions/sts-70/tmp/images.html
/imagemap.cgi//sts-70/landing/landing.map?384,207
/history/apollo/a-001/sounds/
/elv/TITAN/elvhead2.gif
/cwis/organizations/kucia/uroulette/sig.gif
/news/sci.space.shuttle/archive/sci-space-shuttle-7-feb-1994-87.txt
/Government/Research_Labs/NASA/Kennedy_Space_Center/
/history/apollo/sa-3/news/
/finance/main.html
/history/apollo/apollo/-13/apollo-13-info.html
/history/appolo/appolo-13.html
/http//www.cs.tut.fi/ p116711/http//www.cs.tut.fi/ p116711/
/software///icons/menu.xbm
/ismap/image-maps/pccomp/webmap/overview.map?320,134
/shuttle/missions/weathico.gif

/shuttle/missions/missions.htnl
/shuttle/missions/sts-71/palace.com
/wxworld/images/IMAGE6pNgmSfc/95072712_00.GIF
//spacelink.msfc.nasa.gov:70/00/About.Spacelink/File.Transfer.Protocols
/apollo.
/ksc.htlm
/htbin.cdt_mail.pl
/shuttle/missions/sts-71/images/KSC-9
/history/apollo/a-003/movies/
/shuttle/technology/sts-newsref/sts_coord.html#sts_body
/jistory/apollo/apollo-13/apollo-13.html
/shuttle/count.gif
/history/apollo/apollo-13.htm
/history/apollo/apollo-11/apollo-13.html
/shuttle/missions/delta/
/shuttle/technology/st-newsref/stsref-toc.hyml

2.3.8 - Paths most 404 Errors

URI	Frequency
/pub/winvn/readme.txt	667
/pub/winvn/release.txt	547
/history/apollo/apollo-13.html	286
/shuttle/resources/orbiters/atlantis.gif	230
/history/apollo/a-001/a-001-patch-small.gif	230
/://spacelink.msfc.nasa.gov	215
/history/apollo/pad-abort-test-1/pad-abort-test-1-patch-small.gif	215
/images/crawlerway-logo.gif	214
/history/apollo/sa-1/sa-1-patch-small.gif	183
/shuttle/resources/orbiters/discovery.gif	180
/shuttle/missions/sts-68/ksc-upclose.gif	175
/shuttle/missions/sts-71/images/KSC-95EC-0916.txt	168
/elv/DELTA/uncons.htm	163
/history/apollo/publications/sp-350/sp-350.txt	140
/shuttle/missions/technology/sts-newsref/stsref-toc.html	107
/shuttle/resources/orbiters/challenger.gif	92
/procurement/procurement.htm	86
/history/apollo-13/apollo-13.html	73
/history/apollo/pad-abort-test-2/pad-abort-test-2-patch-small.gif	71
/shuttle/countdown/video/livevideo.jpeg	68