ΑΝΩΤΑΤΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΠΕΙΡΑΙΑ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΤΟΜΕΑΣ ΑΡΧΙΤΕΚΤΟΝΙΚΗΣ Η/Υ, ΠΛΗΡΟΦΟΡΙΚΗΣ & ΔΙΚΤΥΩΝ

Εργ. Τεχνολογίας Λογισμικού & Υπηρεσιών

S^2E Lab

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

Σπουδάστρια: Κοντζοπούλου Παναγιώτα

Θέμα:

Μελέτη-σχεδίαση εφαρμογής σε υπολογιστικό νέφος (cloud computing) με έμφαση στην κατασκευή δέντρων.

Εισηγητής:

Δ. Ν. Καλλέργης, MSc.

Εργ. Συνεργάτης

Πειραιάς S^2E Lab – Μάρτιος 2013

Πνευματικά δικαιώματα

Τα πνευματικά δικαιώματα χρησιμοποίησης του μη πρωτότυπου υλικού της εργασίας ανήκουν στο/στη φοιτητή/-τρια και τον επιβλέποντα εις ολόκληρον, δηλαδή εκάτερος μπορεί να κάνει χρήση αυτών χωρίς τη συναίνεση του άλλου. Τα πνευματικά δικαιώματα χρησιμοποίησης του πρωτότυπου μέρους διπλωματικής εργασίας ανήκουν στο/στη φοιτητή/-τρια και τον επιβλέποντα από κοινού, δηλαδή δεν μπορεί ο ένας από τους δύο να κάνει χρήση αυτού χωρίς τη συναίνεση του άλλου. Κατ' εξαίρεση, επιτρέπεται η δημοσίευση του πρωτότυπου μέρους της εργασίας σε επιστημονικό περιοδικό ή πρακτικά συνεδρίου από τον ένα εκ των δύο, με την προϋπόθεση να αναφέρονται τα ονόματα και των δύο ως συν-συγγραφέων. Στην περίπτωση αυτή, προηγείται γραπτή ενημέρωση του μη συμμετέχοντα στη συγγραφή του επιστημονικού άρθρου. Δεν επιτρέπεται η κατά οποιοδήποτε τρόπο δημοσιοποίηση υλικού το οποίο έχει δηλωθεί εγγράφως ως απόρρητο.

Οι υπογράφοντες

Περίληψη

Η παρούσα εργασία αφορά την μελέτη της απόδοσης ενός κατανεμημένου συστήματος βασισμένου στο μοντέλο του υπολογιστικού νέφους (Cloud Computing), υποδομή ως υπηρεσία (Infrastructure As a Service). Γίνεται χρήση του προγραμματιστικού μοντέλου Hadoop MapReduce υλοποιώντας τη παράλληλη κατασκευή δέντρων επιθεμάτων σύμφωνα με τον αλγόριθμο Ukkonen. Τέλος πραγματοποιούνται μετρήσεις οι οποίες δείχνουν τη συμπεριφορά του αλγόριθμου σε παράλληλη εκτέλεση.

Abstract

The aim of this Thesis was to study the performance of a distributed system based on the model of Cloud Computing, Infrastructure as a Service. It uses the Hadoop MapReduce programming model by implementing the parallel construction of suffix trees according to Ukkonen's algorithm. Finally, experiments were conducted to evaluate the behavior of the algorithm with parallel execution.

Περιεχόμενα

Κατάλογος εικόνων

Κατάλογος πινάκων

ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της Πτυχιακής Εργασίας

Η ανάγκη για αποθήκευση και επεξεργασία δεδομένων έχει αυξηθεί κατακόρυφα τα τελευταία χρόνια. Καθημερινά δημιουργείται τεράστιος όγκος πληροφορίας ενώ μέσω του διαδικτύου διακινούνται μεγάλες ποσότητες δεδομένων. Αυτή την ανάγκη έρχονται να καλύψουν τα κατανεμημένα συστήματα παράλληλης επεξεργασίας και το υπολογιστικό νέφος(cloud computing).

Το υπολογιστικό νέφος είναι η παροχή υπολογιστικού χώρου υπό τη μορφή υπηρεσίας και όχι προϊόντος, όπου κοινόχρηστοι πόροι, λογισμικό και πληροφορίες παρέχονται σε υπολογιστές και άλλες συσκευές διαμέσου ενός δικτύου. Στη πτυχιακή εργασία, αρχικά, γίνεται εκτενής ανάλυση σε θεωρητικό επίπεδο του υπολογιστικού νέφους έτσι ώστε να γίνει κατανοητή η έννοια, τα διαθέσιμα μοντέλα, τα χαρακτηριστικά κάθε μοντέλου καθώς και τα πλεονεκτήματα και τα μειονεκτήματα του καθενός. Τα μοντέλα χωρίζονται σε δύο κατηγορίες ως προς το είδος της υπηρεσίας και ως προς τον τρόπο ανάπτυξης. Βασιζόμενοι στο είδος της υπηρεσίας, τα διαθέσιμα μοντέλα είναι τα λογισμικό ως υπηρεσία (Software as a service), πλατφόρμα ως υπηρεσία (Platform as a service) και υποδομή ως υπηρεσία (Infrastructure as a service). Τα μοντέλα ανάπτυξης είναι τα δημόσιο υπολογιστικό σύννεφο (public cloud), ιδιωτικό υπολογιστικό σύννεφο (private cloud), κοινοτικό υπολογιστικό σύννεφο (community cloud) και υβριδικό υπολογιστικό σύννεφο (hybrid cloud).

Ακολουθεί μελέτη για το ανοιχτό λογισμικό Apache Hadoop που χρησιμοποιήθηκε για την υλοποίηση του προγράμματος και του μοντέλου υποδομή ως υπηρεσία του υπολογιστικού νέφους. Το Hadoop είναι ένα λογισμικό γραμμένο σε Java που υποστηρίζει την επεξεργασία μεγάλων συνόλων δεδομένων σε ένα κατανεμημένο υπολογιστικό σύστημα. Προσφέρει ένα επίπεδο απλοποίησης στη διαδικασία ανάπτυξης παράλληλων προγραμμάτων αναλαμβάνοντας να διαχειριστεί το διαμοιρασμό των δεδομένων, την συγκέντρωση των αποτελεσμάτων, τις πιθανές αποτυχίες κόμβων και

άλλα θέματα. Το προγραμματιστικό μοντέλο Hadoop MapReduce είναι υπεύθυνο για την κατασκευή και την εκτέλεση παράλληλων προγραμμάτων. Με τη βοήθειά του υλοποιείται η παράλληλη κατασκευή δέντρων επιθεμάτων (suffix trees) σύμφωνα με τον αλγόριθμο του Ukkonen.

Το δέντρο επιθεμάτων είναι δομή δεδομένων που αναπαριστά ένα επίθεμα που μπορεί να ανήκει σε μία ή παραπάνω συμβολοσειρές επιτρέποντας τη γρήγορη υλοποίηση πολλών διεργασιών συμβολοσειράς. Συγκεκριμένα για την εφαρμογή χρησιμοποιήθηκε ο αλγόριθμος Ukkonen ο οποίος είναι ένας γραμμικός, online αλγόριθμος που προτάθηκε από τον Esko Ukkonen το 1995.

1.2 Στόχος

Η παρούσα πτυχιακή, αρχικά, δίνει στον αναγνώστη μια γενική εικόνα για την έννοια του λεγόμενου υπολογιστικού νέφους. Οι βασικοί στόχοι της όμως είναι δύο:

- Η υλοποίηση ενός παράλληλου προγράμματος κατασκευής δέντρων επιθεμάτων βασισμένο στον παραπάνω αλγόριθμο και στο προγραμματιστικό μοντέλο του MapReduce για την επίλυση του προβλήματος εύρεσης της μεγαλύτερης κοινής υπό-συμβολοσειράς (Longest common substring problem).
- Η ανάλυση της απόδοσης του παραπάνω προγράμματος σε πραγματικό δίκτυο υπολογιστών που περιλαμβάνει ένα αυξανόμενο πλήθος κόμβων.

1.3 Δομή

Το κείμενο της πτυχιακής αποτελείται από πέντε κεφάλαια. Το παρόν κεφάλαιο είναι το πρώτο και σε αυτό γίνεται αναφορά στο αντικείμενο της πτυχιακής και στον τρόπο οργάνωσης του υπόλοιπου κειμένου.

Στο *κεφάλαιο 2* επιχειρείται η παροχή του απαραίτητου υπόβαθρου που χρειάζεται ο αναγνώστης για την κατανόηση του συνόλου της εργασίας. Συγκεκριμένα, γίνεται θεωρητική ανάλυση του υπολογιστικού νέφους, των μοντέλων του και των χαρακτηριστικών του. Στη συνέχεια παρουσιάζονται οι πιο σύγχρονες τάσεις παροχής

δικτυακών υπηρεσιών που βασίζονται σε υπολογιστικά νέφη. Τέλος εισάγεται στον αναγνώστη η έννοια της τεχνολογία του framework Hadoop MapReduce καθώς περιγράφεται ο τρόπος λειτουργίας του Hadoop MapReduce και του κατανεμημένου συστήματος αρχείων HDFS.

Στο *κεφάλαιο 3* παρουσιάζεται η συνεισφορά της εργασίας σε επίπεδο υλοποιήσεων. Αναλυτικότερα, παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν, γίνεται μια θεωρητική προσέγγιση στον αλγόριθμου του Ukkonen και έπειτα αναλύονται σημαντικά κομμάτια κώδικα του προγράμματος που υλοποιήθηκε.

Το *κεφάλαιο 4* περιλαμβάνει τον τρόπο εγκατάστασης και πραγματοποίησης ενός Hadoop MapReduce cluster και τις μετρήσεις που έγιναν για τη παραπάνω υλοποίηση με σκοπό την κατανόηση της συμπεριφοράς του και της απόδοσης του σε πραγματικές συνθήκες.

Ακολουθεί το *κεφάλαιο 5* που αποτελεί την πρόταση για μελλοντικές εργασίες. Συγκεκριμένα αναφέρονται διάφορες εκδοχές υλοποίησης που δίνουν τη δυνατότητα περαιτέρω έρευνας.