Ειδικά Θέματα Τεχνολογίας Λογισμικού

Περιγραφή Έργου: Oryx

Ομάδα NP-complete
Ξυνογαλά Ελένη
Χαντζάρα Γιάννης

24/3/2014

Πληροφορίες του έργου

Όνομα έργου: cloudera / oryx
Γλώσσα υλοποίησης: java
Περιβάλλον εκτέλεσης έργου : Windows, Linux
Άδεια : Apache License, Version 2.0
Μέγεθος κώδικα : πάνω από 40.000 γραμμές
Μέγεθος ομάδας ανάπτυξης: 5
Τρόπος δόμησης έργου: maven, eclipse

Περιγραφή του Oryx

Το Oryx είναι ένα έργο ανοικτού πηγαίου κώδικα, το οποίο παρέχει, σε πραγματικό χρόνο, μεγάλης κλίμακας μηχανική μάθησης και υποδομή **analytics** .

Υλοποιεί **αλγορίθμους** που χρησιμοποιούνται συνήθως στις επιχειρηματικές εφαρμογές :

- το συνεργατικό φιλτράρισμα / σύσταση ,
- την ταξινόμηση / παλινδρόμηση , και
- Ομαδοποίηση-Συσταδοποίηση .

Μπορεί να χτίσει **μοντέλα** από μια ροή δεδομένων σε μεγάλη κλίμακα , χρησιμοποιώντας Apache Hadoop .

Σερβίρει επίσης ερωτήματα των μοντέλων αυτών σε πραγματικό χρόνο μέσω HTTP API REST , και μπορεί να ενημερώσει τα μοντέλα όταν υπάρχει ροή νέων δεδομένων .

Αντιπροσωπεύει μια ενιαία συνέχιση των έργων της Myrrix και Cloudera .

Αλγόριθμοι

Το συνεργατικό φιλτράρισμα / σύσταση

Το Oryx βασίζεται στη παραγοντοποίηση και σε μια παραλλαγή του ALS (εναλλασσόμενο ελαχίστων τετραγώνων). Οι μηχανές Σύστασης χρησιμοποιούνται πλέον ευρέως για να προτείνουν αντικείμενα όπως τα βιβλία και τις ταινίες στους ανθρώπους , αλλά μπορεί σε γενικές γραμμές να χρησιμοποιηθούν για να μαντέψουν ενώσεις μεταξύ των οντοτήτων δοθέντων πολλών παρατηρούμενων ενώσεων .

Ταξινόμηση και Παλινδρόμηση

Oryx υποστηρίζει τυχαία δέντρα απόφασης για την ταξινόμηση και παλινδρόμηση των δραστηριοτήτων. Η τιμή για τις νέες εισόδους προβλέπεται βασιζόμενη σε γνωστές τιμές για τις προηγούμενες εισόδους . Αυτό περιλαμβάνει τις Δραστηριότητες ταξινόμησης- προβλέπουν μια κατηγορία όπως το «spam» (μαζική αποστολή μηνυμάτων)- και τα καθήκοντα παλινδρόμησης - προβλέπουν μια αριθμητική τιμή όπως το μισθό .

Ομαδοποίηση

Oryx εφαρμόζει κλιμακωτή k -means + + για την ομαδοποίηση . Αυτό είναι ένα είδος μάθησης χωρίς επίβλεψη , η οποία επιδιώκει να βρει τη δομή στην είσοδο του με τη μορφή των φυσικών ομάδων. .

Αρχιτεκτονική Λάμδα

Η αρχιτεκτονική αυτή χρησιμοποιείται για συστήματα Big Data, τρέχοντας ad - hoc ερωτήματα. Τα συστατικά στοιχεία της αρχιτεκτονικής του Oryx είναι:

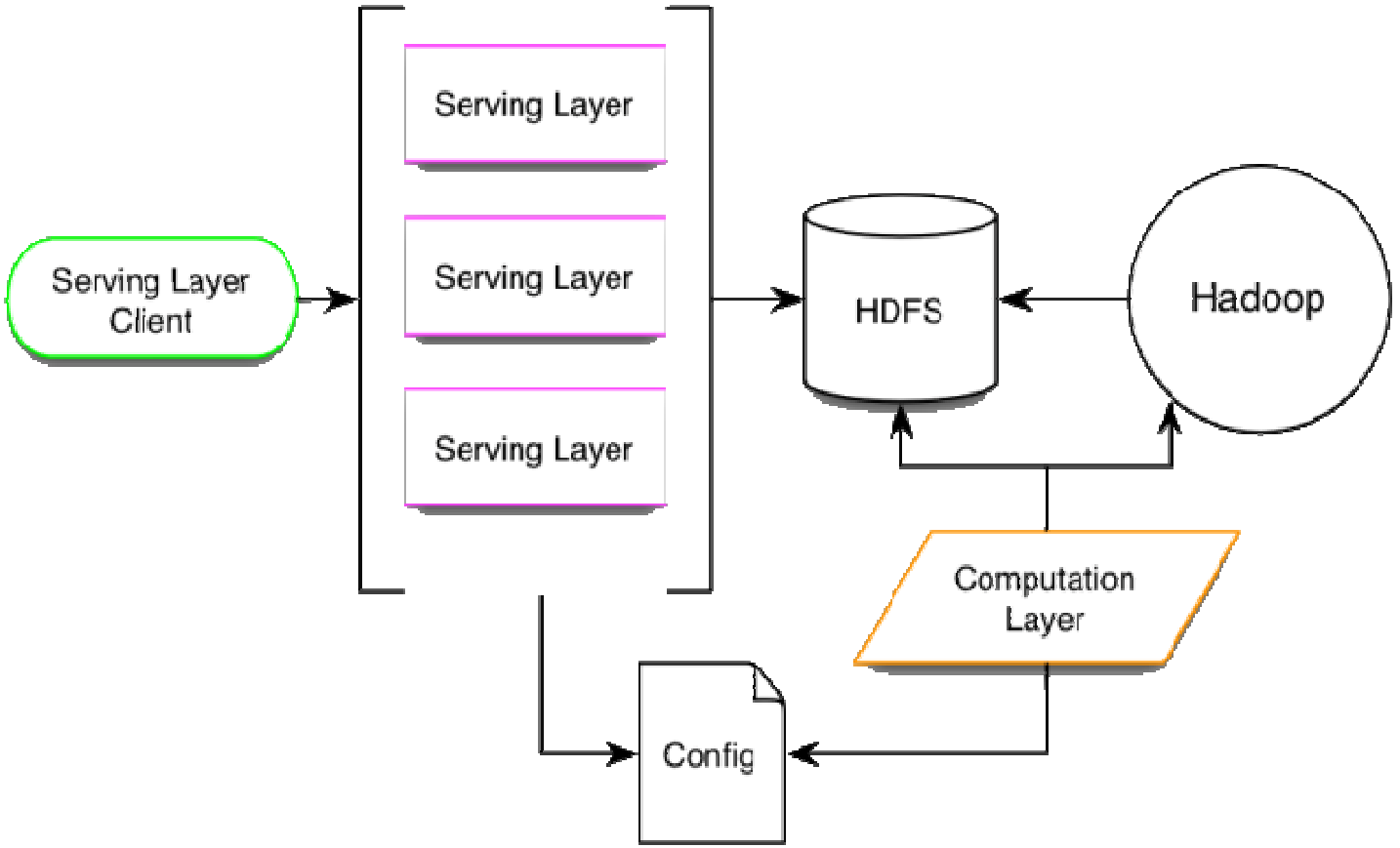
- Το **Επίπεδο Υπολογισμού:** χτίζει μοντέλα και βασίζεται σε «γενιές», δηλαδή εξάγει μια διαδοχή αποτελεσμάτων των μοντέλων στην πάροδο του χρόνου. Η είσοδος φτάνει στο HDFS και τα μοντέλα είναι γραμμένα ως αρχεία PMML.

- Hadoop Distributed File System παρέχει τη δυνατότητα κλιμάκωσης που είναι απαραίτητη για την επεξεργασία των μεγάλων δεδομένων. Predictive Model Markup Language είναι μια μορφή αρχείου σε XML που παρέχει έναν τρόπο για να περιγράψει τις εφαρμογές και την ανταλλαγή προτύπων που παράγονται από την εξόρυξη δεδομένων και τους αλγόριθμους μηχανικής μάθησης.

- Το **Επίπεδο Σερβιρίσματος**, το οποίο σερβίρει τα μοντέλα από το HDFS . Η διαδικασία βασίζεται σε Java server. Μπορεί να προσεγγιστεί από ένα πρόγραμμα περιήγησης , ή οποιαδήποτε γλώσσα ή εργαλείο που μπορεί να κάνει αιτήσεις HTTP . Πολλά περιστατικά μπορεί να τρέξουν ταυτόχρονα . Δρουν και εξυπηρετούν ανεξάρτητα το ένα από το άλλο.

- Διαμόρφωση**:** η χρήση ενός αρχείου ρυθμίσεων για να εκτελεστούν τα επίπεδα Υπολογισμού και Σερβιρίσματος. Ένα αρχείο ρυθμίσεων είναι απλά ένα αρχείο κειμένου χρησιμοποιώντας τη σύνταξη HOCON (ένας συνδυασμός JSON και απλό αρχείο ιδιοτήτων σύνταξης) .

Διάγραμμα Αρχιτεκτονικής



Τεχνολογία Κατασκευής

Κατανεμημένο

Το Επίπεδο Υπολογισμού προορίζεται κυρίως για χρήση περιβάλλοντος υπολογισμού Hadoop για τον υπολογισμό ,όπως το μοντέλο MapReduce για αλγορίθμους παράλληλους και κατανεμημένους σε συστάδες (cluster).

Τοπικό

Το Επίπεδο Υπολογισμού μπορεί επίσης να ρυθμιστεί να τρέχει τους υπολογισμούς σε τοπικό επίπεδο αντί για Hadoop , και να διαβάσει και να γράψει τα δεδομένα στο τοπικό σύστημα αρχείων. Αυτό είναι χρήσιμο για τα μικρά ή μη κρίσιμα προβλήματα , ή για την απλή δοκιμή .

Επίπεδο διάταξης **καταλόγου**

Όλα τα δεδομένα αποθηκεύονται σε ένα , διαμορφωμένο κατάλογο. Κάθε ένας από αυτούς περιέχει τα αρχεία για μια γενιά . Το ακριβές περιεχόμενο αυτών των καταλόγων διαφέρει από τον αλγόριθμο , αλλά σε όλες τις περιπτώσεις , η είσοδος φτάνει σε ένα υποκατάλογο , και το πρότυπο αρχείο δημιουργείται στον κατάλογο γενιά.

Συνεισφορά

- Διόρθωση ορισμένων σφαλμάτων (bugs) που εμφανίζονται στο έργο.
- Βελτίωση/ανάπτυξη των κομματιών CF που χρησιμοποιούν αλγόριθμους ALS και βρίσκονται σε επίπεδο beta (Initial Commits).