TEI OF CENTRAL MACEDONIA
SERRES-GREECE

Διπλωματική Εργασία
Αθανάσιος Ντόγαρης
A.E.M 53

Διαχείριση μεγάλου όγκου δεδομένων με το οικοσύστημα Hadoop και τη NoSQL βάση δεδομένων Hbase

"Big Data management with Hadoop ecosystem and HBase"

Ιούνιος 2018

Σκοπός της παρούσης εργασίας

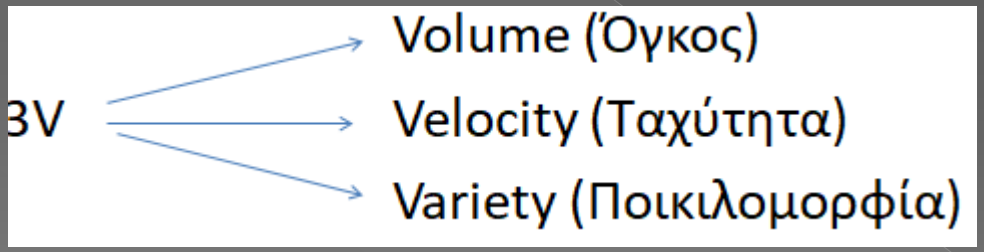
- Η μελέτη των δυνατοτήτων του οικοσυστήματος Hadoop και της HBase που είναι μία ανοιχτού λογισμικού NoSQL βάση δεδομένων ιδανική για αποθήκευση δεδομένων μεγάλου όγκου (Big Data).

Γιατί το Hadoop ήρθε στο προσκήνιο?

- Τα τελευταία χρόνια ο όγκος των δεδομένων έχει αυξηθεί εκθετικά διότι αυτά προέρχονται από διάφορες πηγές όπως τα Social Media, Κινητές συσκευές κ.α
- Αυτά τα δεδομένα τα οποία προέρχονται από πολλές και διαφορετικές πηγές ονομάζονται Δεδομένα Μεγάλου Όγκου (Big Data)

Ποιο ήταν το πρόβλημα?

- Οι παραδοσιακές βάσεις δεδομένων δεν είναι σχεδιασμένες για να χειρίζονται τα 3V των μεγάλων δεδομένων



Πως λύνει το Hadoop το πρόβλημα?

- Το Hadoop μπορεί να χειριστεί μεγάλου όγκου δεδομένα, μπορεί να επεξεργαστεί δεδομένα με μεγάλη ταχύτητα, και μπορεί να αποθηκεύσει οποιουδήποτε είδους δεδομένα.

Επομένως τι είναι το Hadoop?

- Το Hadoop ουσιαστικά είναι μία πλατφόρμα η οποία βασίζεται σε ένα κατανεμημένο σύστημα αρχείων το οποίο μας επιτρέπει να αποθηκεύσουμε δεδομένα πολύ μεγάλου όγκου σε διαφορετικούς υπολογιστές και μας παρέχει ένα API για την επεξεργασία αυτών των δεδομένων.

- Η βασική του ιδέα είναι ότι από τη στιγμή που τα δεδομένα αποθηκεύονται σε πολλούς ξεχωριστούς υπολογιστές μπορούμε να τα επεξεργαστούμε με έναν παράλληλο και κατανεμημένο τρόπο κατά τον οποίο ο κάθε υπολογιστής επεξεργάζεται τα δεδομένα τα οποία είναι αποθηκευμένα σε αυτόν.

Γιατί οι εταιρείες προτιμούν το Hadoop?

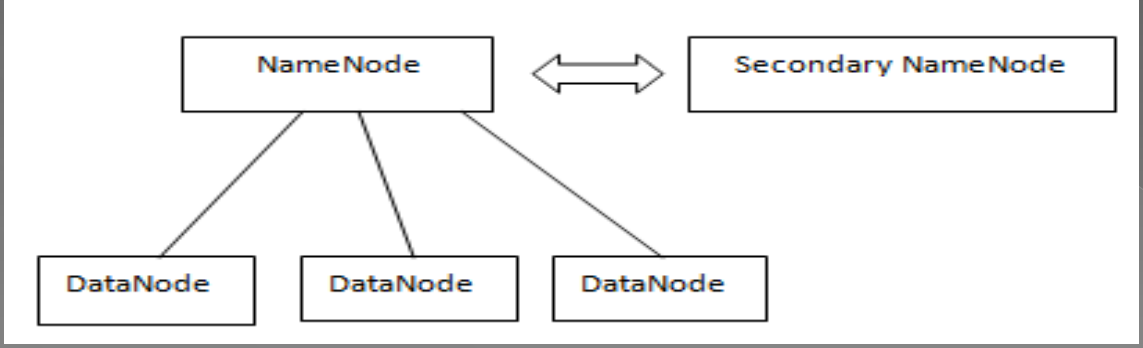
- 1) Γιατί είναι λογισμικό ανοιχτού κώδικα και δωρεάν
- 2) Γιατί μπορεί και χειρίζεται δεδομένα μεγάλου όγκου
- 3) Γιατί σχεδιάστηκε για να ‘τρέχει’ σε χαμηλού κόστους hardware

Ποια είναι τα κύρια συστατικά του Hadoop?

- Τα κύρια χαρακτηριστικά του Hadoop είναι:
- 1) Το HDFS
- 2) Το MapReduce
- 3) Το Yarn

Τι είναι το HDFS?

- Το HDFS (Hadoop Distributed File System) είναι είναι ένα κατανεμημένο σύστημα αρχείων το οποίο μας επιτρέπει να αποθηκεύσουμε πολύ μεγάλα αρχεία σε διάφορους nodes ενός Hadoop cluster.



Τι είναι το MapReduce?

- Το MapReduce είναι μία προγραμματιστική πλατφόρμα η οποία μας επιτρέπει να υλοποιήσουμε παράλληλη και κατανεμημένη επεξεργασία σε δεδομένα μεγάλου όγκου σε ένα κατανεμημένο περιβάλλον όπως αυτό του Hadoop.
- Αποτελείται από δύο διεργασίες, την Map και την Reduce. Η Reduce διεργασία γίνεται μετά από την Map, διότι στην Reduce γίνεται η επεξεργασία των δεδομένων που προέρχονται από την Map.

Τι είναι το YARN?

- Το YARN (Yet Another Resource Negotiator) είναι ουσιαστικά η κεντρική μονάδα επεξεργασίας του Hadoop και είναι υπεύθυνο για την διαχείριση των διεργασιών και των πόρων του συστήματος

Τι είναι το Cloudera Quickstarts VM?

- Είναι μία πλατφόρμα χτισμένη σε Linux η οποία δημιουργήθηκε για επιχειρήσεις αλλά και για άτομα που θέλουν μία ολοκληρωμένη λύση για επεξεργασία μεγάλων δεδομένων.
- Περιέχει τα κύρια συστατικά του Hadoop(HDFS, MapReduce, YARN) καθώς και άλλα προγράμματα ανοιχτού λογισμικού όπως είναι η HBase και το Hive με τα οποία επίσης θα ασχοληθούμε.

Πλεονεκτήματα Cloudera VM

- Προσφέρει έναν πολύ απλό τρόπο για να ξεκινήσει κάποιος με το Hadoop. Όλα τα προγράμματα είναι προεγκατεστημένα.
- Έχει ένα απλό και διαχειρίσιμο interface.

Μειονεκτήματα Cloudera VM

- Έχει υψηλές απαιτήσεις συστήματος σε περίπτωση που εγκατασταθεί σε προσωπικό υπολογιστή:

1) 64 – bit σύστημα
2) Ελάχιστη μνήμη τουλάχιστον 4GB, αλλά στην πραγματικότητα χρειάζονται 8GB τουλάχιστο για επεξεργασία δεδομένων
3) 4-πύρηνο επεξεργαστή
4) Έρχεται με εγκαταστημένη την έκδοση 1.7 της Java η οποία χρειάζεται αναβάθμιση για να τρέξει το ElasticSearch

Προγράμματα που είναι εγκατεστημένα στο Cloudera VM και με τα οποία θα ασχοληθούμε

- Hue
- HBase
- Hive

Άλλα προγράμματα που είναι εγκατεστημένα στο Hadoop και δεν θα ασχοληθούμε είναι

- **Impala** – χρησιμοποιείται για τη δημιουργία ερωτημάτων SQL κατευθείαν στο HDFS και στην HBase
- **Pig** – προσφέρει μία γλώσσα υψηλού επιπέδου την PigLatin η οποία κάνει ποιο εύκολη την συγγραφή MapReduce κώδικα
- **Spark** – ένα ανοιχτού λογισμικού περιβάλλον για γρήγορη ανάλυση μεγάλων δεδομένων
- **Oozie** – ένα σύστημα προγραμματισμού ροών εργασιών
- **Solr** – Η αντίστοιχη μηχανή αναζήτησης που προσφέρεται με το Cloudera Quickstarts VM
- **Sqoop** – Επιτρέπει την αλληλεπίδραση των RDBMS με το HDFS

Τι είναι το Hue?

- Το Hue (Hadoop User Interface) παρέχει ένα web interface για το Hadoop και για τα προγράμματα που χρησιμοποιούνται για ανάλυση δεδομένων όπως είναι το Hive και η HBase.
- Επίσης με το Hue το ανέβασμα και η διαγραφή αρχείων από το HDFS γίνεται πιο εύκολα χωρίς τη χρήση της γραμμής εντολών

Τι είναι το Hive?

- Το Hive αρχικά είχε υλοποιηθεί από το Facebook και στη συνέχεια από την Apache Software Foundation. Μας παρέχει μία γλώσσα παρόμοια με την SQL, η οποία ονομάζεται Hive Query Language (ή αλλιώς HiveQL ή HQL) για τη δημιουργία ερωτημάτων πάνω στα δεδομένα που έχουν αποθηκευτεί σε μία Hadoop συστοιχία (cluster). Αυτό που κάνει είναι να μεταφράζει αυτά τα ερωτήματα σε MapReduce εργασίες, επιτρέποντας έτσι σε μη Java προγραμματιστές να επεξεργαστούν τα δεδομένα τους.
- Το Hive χρησιμοποιείται κυρίως σε εφαρμογές αποθηκών δεδομένων (data warehouse applications) όπου τα δεδομένα είναι σχεδόν στατικά και δεν αλλάζουν με γρήγορους ρυθμούς.

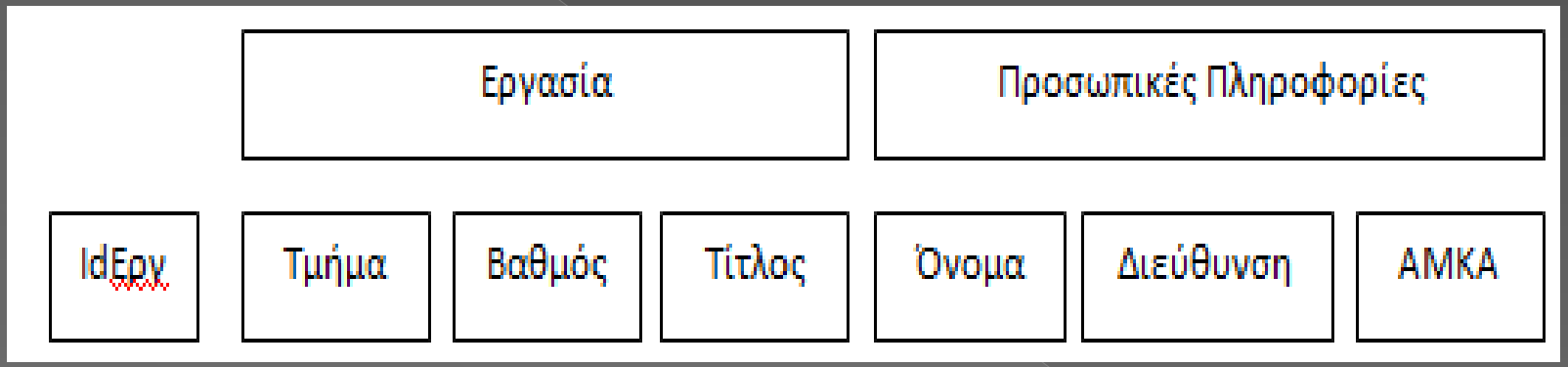
Τι είναι η HBase?

- Είναι η NoSQl πλατφόρμα του Hadoop γραμμένη σε Java
- Είναι column oriented δηλαδή οι τιμές μίας στήλης αποθηκεύονται όλες μαζί

Στην HBase οι πίνακες είναι τεσσάρων διαστάσεων. Σε αυτούς υπάρχουν:

- το κλειδί σειράς (row key)
- η οικογένεια στηλών (column family)
- η στήλη
- η χρονοσφραγίδα (timestamp)

Παράδειγμα πίνακα HBase

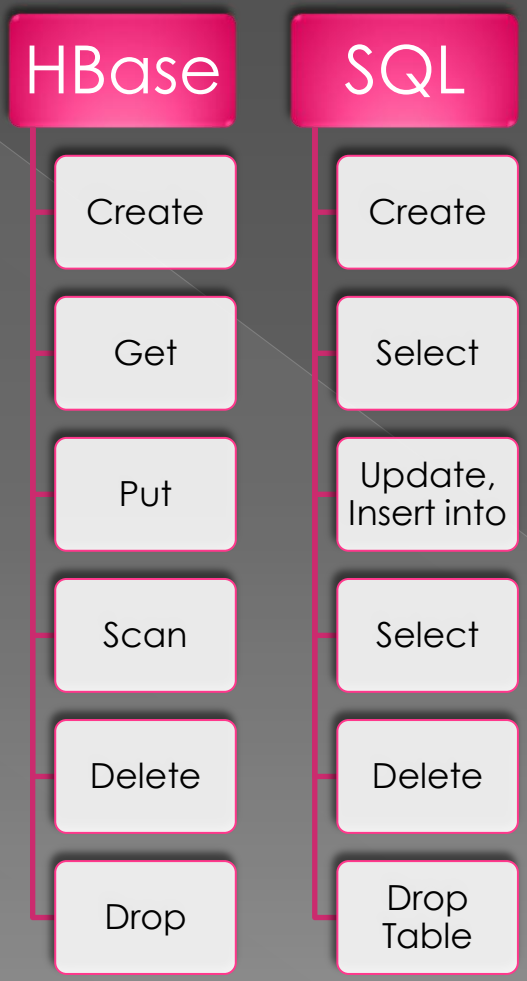


Σύγκριση πίνακα RDBMS με πίνακα της HBase

Id	Όνομα	Επίθετο	Τηλέφωνο	Πόλη
1	Γιώργος	Παπαδόπουλος	6999999999	Σέρρες
2	Γιάννης	Μακρυγιάννης	6999999998	Αθήνα
3	Φωτεινή	Ζησοπούλου	6999999997	Λάρισα
4	Νίκη	Δημητρίου	6999999996	Καβάλα
5	Αθανάσιος	Παυλίδης	6999999995	Σέρρες

Id	Στήλη	Τιμή
1	Όνομα	Γιώργος
1	Επίθετο	Παπαδόπουλος
1	Τηλέφωνο	6999999999
1	Πόλη	Σέρρες
2	Όνομα	Γιάννης
2	Επίθετο	Μακρυγιάννης
2	Τηλέφωνο	6999999998
2	Πόλη	Αθήνα
3	Όνομα	Φωτεινή
3	Επίθετο	Ζησοπούλου
3	Τηλέφωνο	6999999997
3	Πόλη	Λάρισα
4	Όνομα	Νίκη
4	Επίθετο	Δημητρίου
4	Τηλέφωνο	6999999996
4	Πόλη	Καβάλα
5	Όνομα	Αθανάσιος
5	Επίθετο	Παυλίδης
5	Τηλέφωνο	6999999995
5	Πόλη	Σέρρες

HBase εντολές vs SQL εντολές



TEI OF CENTRAL MACEDONIA
SERRES-GREECE

HBase vs RDBMS

HBase

- Μέγεθος δεδομένων: Terrabytes, Petabytes
- Column oriented
- Ευέλικτο schema (μπορούμε να προσθέτουμε δεδομένα χωρίς περιορισμό)
- Εντολές API, εκτός κι αν συνδυαστεί με το Hive
- Ευρετήρια: Μόνο κλειδί σειράς, εκτός αν συνδυαστεί με το Hive ή το ElasticSearch
- Αρχιτεκτονική: Βασισμένη στο Hadoop, εύκολα διαβαθμίσιμη

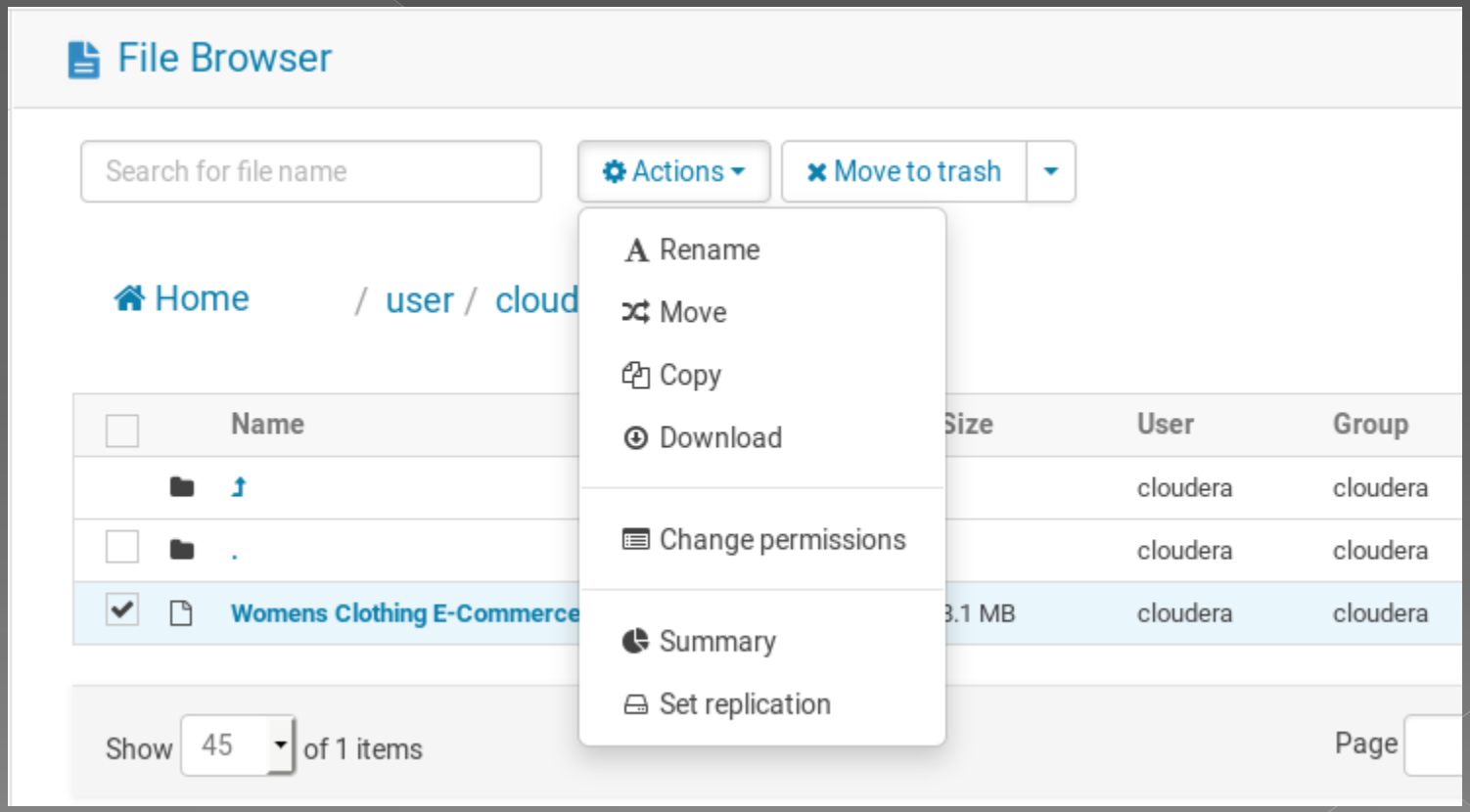
SQL

- Μέγεθος δεδομένων: Gigabytes, Terrabytes.
- Row oriented(τις περισσότερες φορές)
- Σταθερό schema: ορίζεται κατά τον σχεδιασμό της βάσης
- SQL εντολές
- Ευρετήρια: Υπάρχουν
- Αρχιτεκτονική: Για πολύ μεγάλα δεδομένα χρειάζεται ακριβά συστήματα.

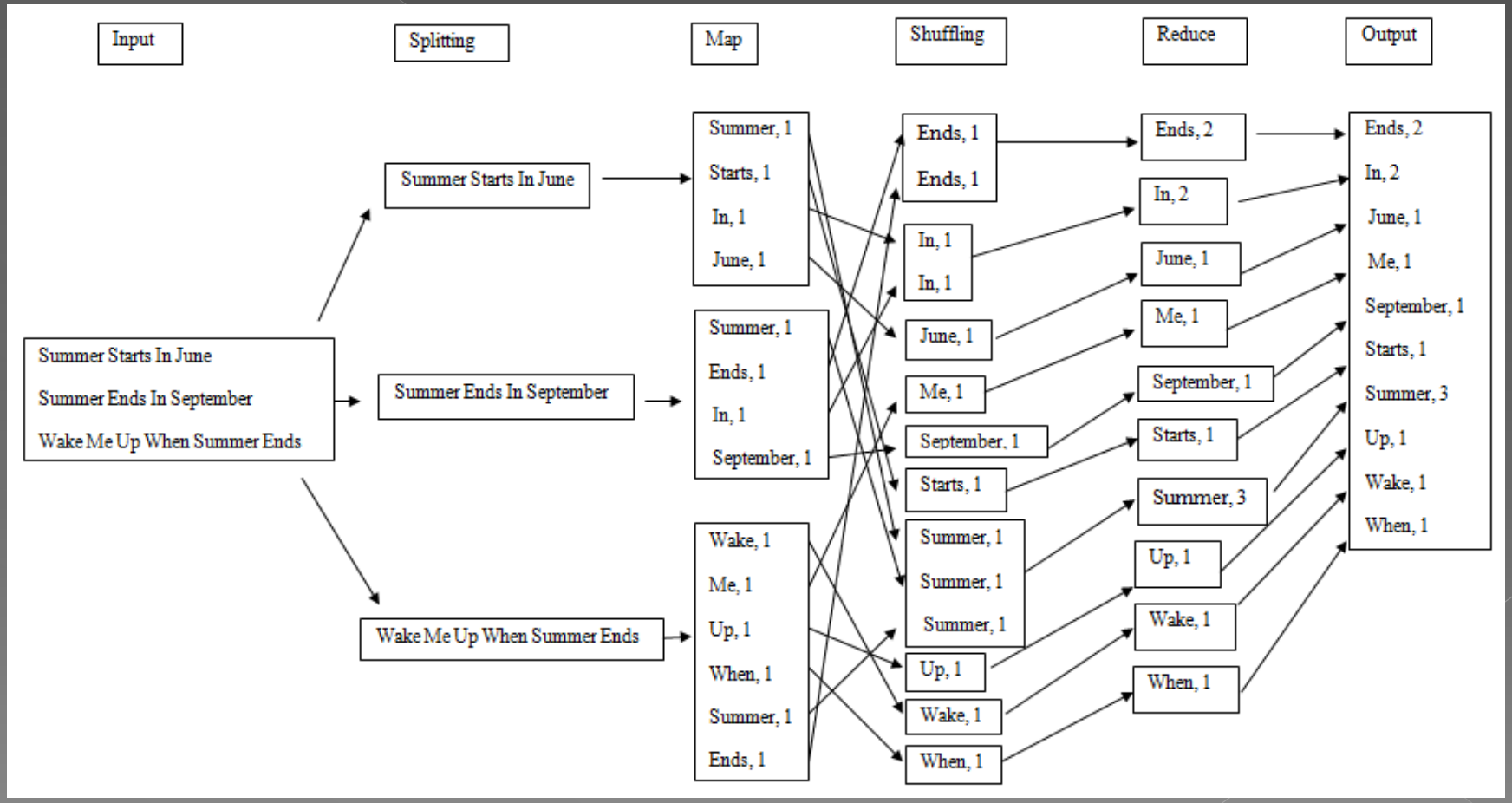
Τι είναι το ElasticSearch?

- Το ElasticSearch είναι μία ανοιχτού λογισμικού μηχανή αναζήτησης κειμένου καθώς και ανάλυσης δεδομένων η οποία μας επιτρέπει να αποθηκεύσουμε, να αναλύσουμε και να αναζητήσουμε πληροφορίες μέσα σε δεδομένα πολύ μεγάλου όγκου σε πολύ μικρό χρονικό διάστημα.
- Είναι γραμμένο στην γλώσσα Java και βασίζεται στο Apache Lucene. Είναι όμως πιο εύχρηστο, γιατί προσφέρει ένα απλό HTTP RESTful API το οποίο χρησιμοποιεί την JSON για την διαχείριση των δεδομένων. Επίσης με το ElasticSearch μπορεί να χρησιμοποιηθεί και το Kibana το οποίο είναι ένα εργαλείο ανάλυσης των δεδομένων με πολύ απλό τρόπο.

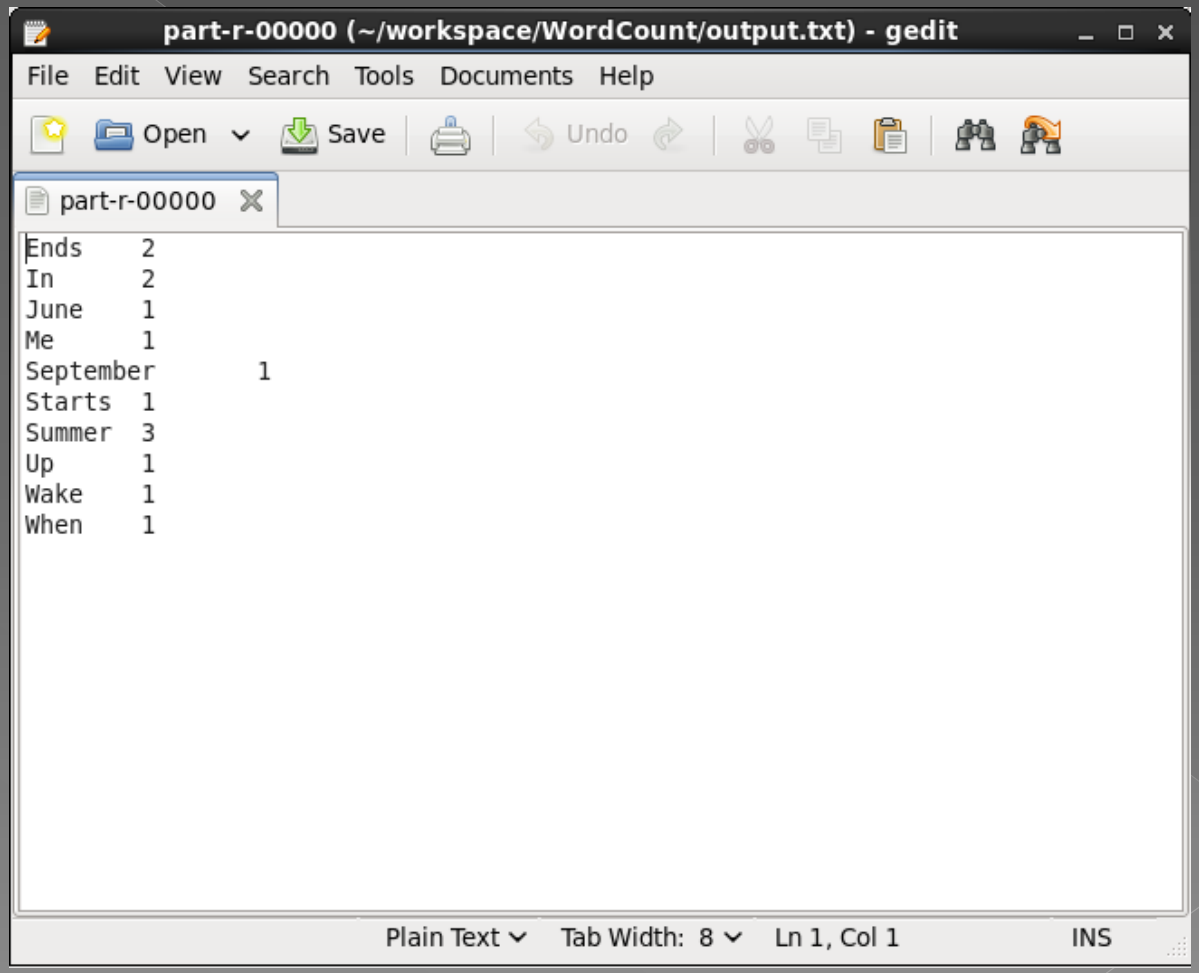
Παράδειγμα HDFS



Παράδειγμα MapReduce



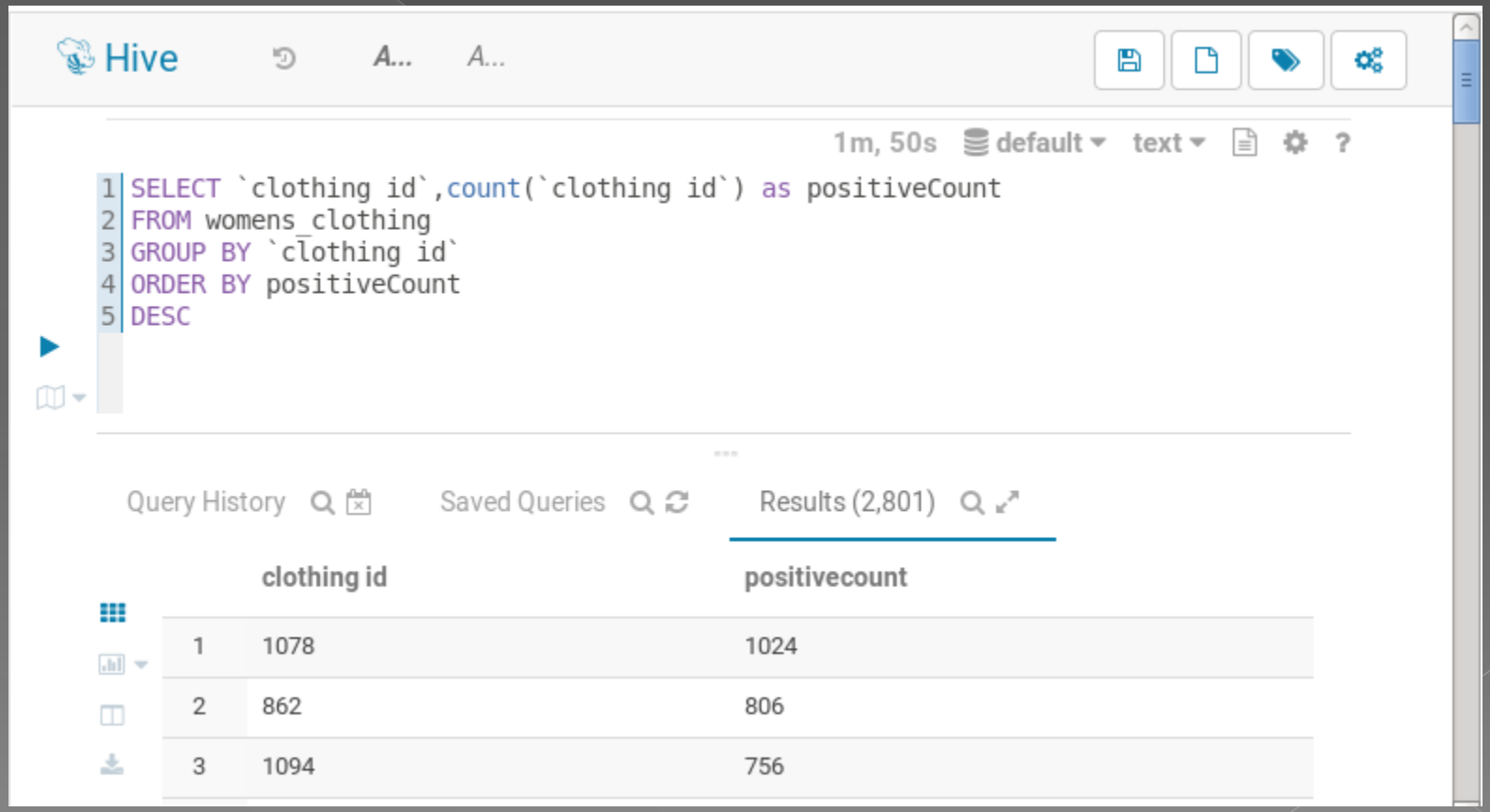
Παράδειγμα MapReduce



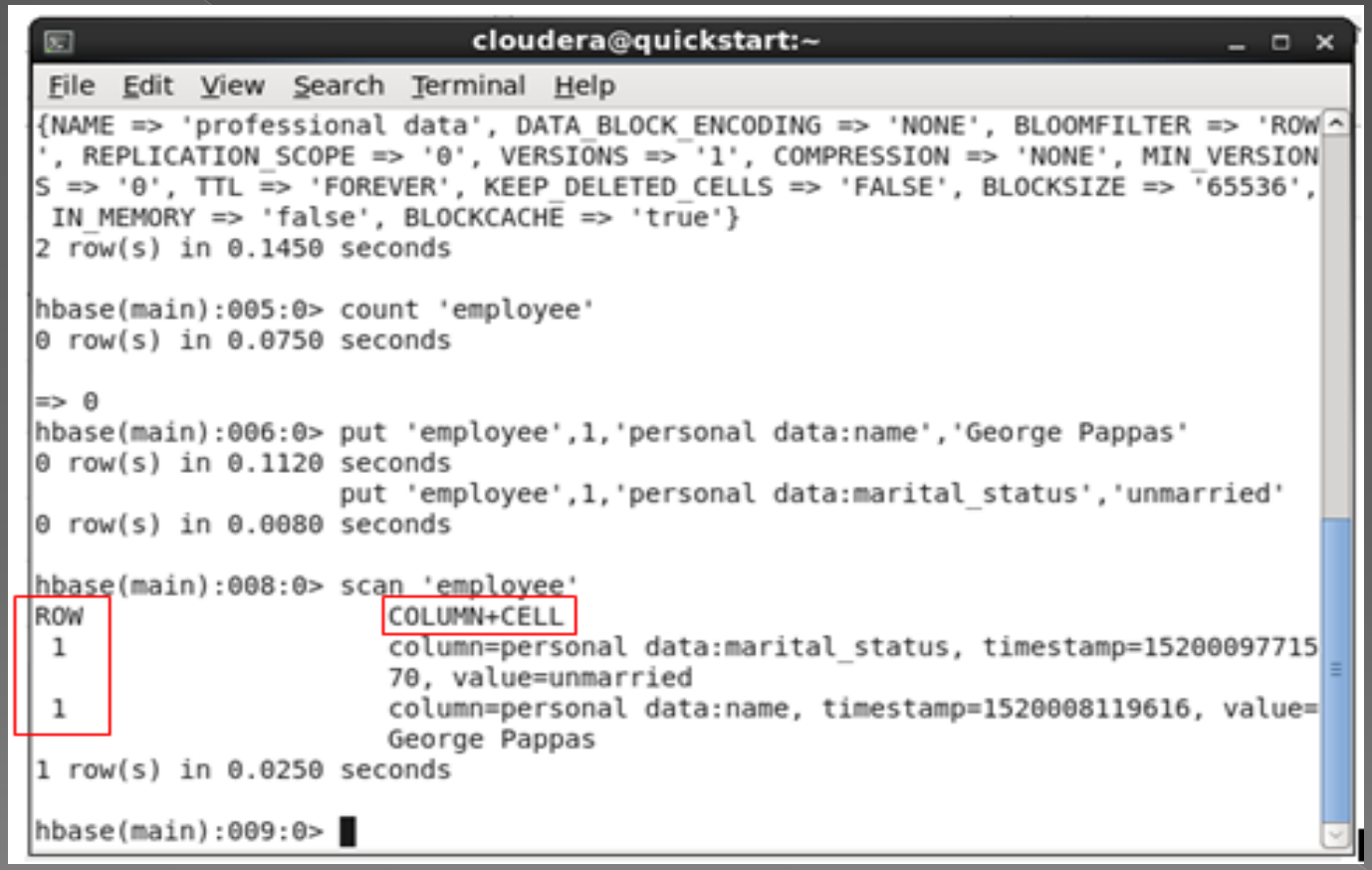
Παράδειγμα Hive

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
,Clothing ID,Age,Title,Review Text,Rating,Recommended IND,Positive Feedback Count,Division Name,Department Name,Class Name																				
0,767,33,,Absolutely wonderful - silky and sexy and comfortable,4,1,0,Initmates,Intimate,Intimates																				
1,1080,34,,"Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i never would have ordered it online bc it's petite. i bought a petite and am 5'8"". i love the length on me- hits just a li																				
2,1077,60,Some major design flaws,"I had such high hopes for this dress and really wanted it to work for me. i initially ordered the petite small (my usual size) but i found this to be outrageously small. so small in fact																				
3,1049,50,My favorite buy!,"I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothing but great compliments!",5,1,0,General Petite,Bottoms,Pants																				
4,847,47,Flattering shirt,This shirt is very flattering to all due to the adjustable front tie. it is the perfect length to wear with leggings and it is sleeveless so it pairs well with any cardigan. love this shirt!!!,5,1,6,Genera																				
5,1080,49,Not for the very petite,"I love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tall and usually wear a 0p in this brand. this dress was very pretty out of the package but its a l																				
6,858,39,Cagrcoal shimmer fun,"I aded this in my basket at hte last mintue to see what it would look like in person. (store pick up). i went with teh darkler color only because i am so pale :-) hte color is really gorgeou																				
7,858,39,"Shimmer, surprisingly goes with lots","I ordered this in carbon for store pick up, and had a ton of stuff (as always) to try on and used this top to pair (skirts and pants). everything went with it. the color is rea																				
8,1077,24,Flattering,I love this dress. i usually get an xs but it runs a little snug in bust so i ordered up a size. very flattering and feminine with the usual retailer flair for style.,5,1,0,General,Dresses,Dresses																				
9,1077,34,Such a fun dress!,"I'm 5""5' and 125 lbs. i ordered the s petite to make sure the length wasn't too long. i typically wear an xs regular in retailer dresses. if you're less busty (34b cup or smaller), a s petite will																				
10,1077,53,Dress looks like it's made of cheap material,Dress runs small esp where the zipper area runs. i ordered the sp which typically fits me and it was very tight! the material on the top looks and feels very cheap																				
11,1095,39,,This dress is perfection! so pretty and flattering.,5,1,2,General Petite,Dresses,Dresses																				

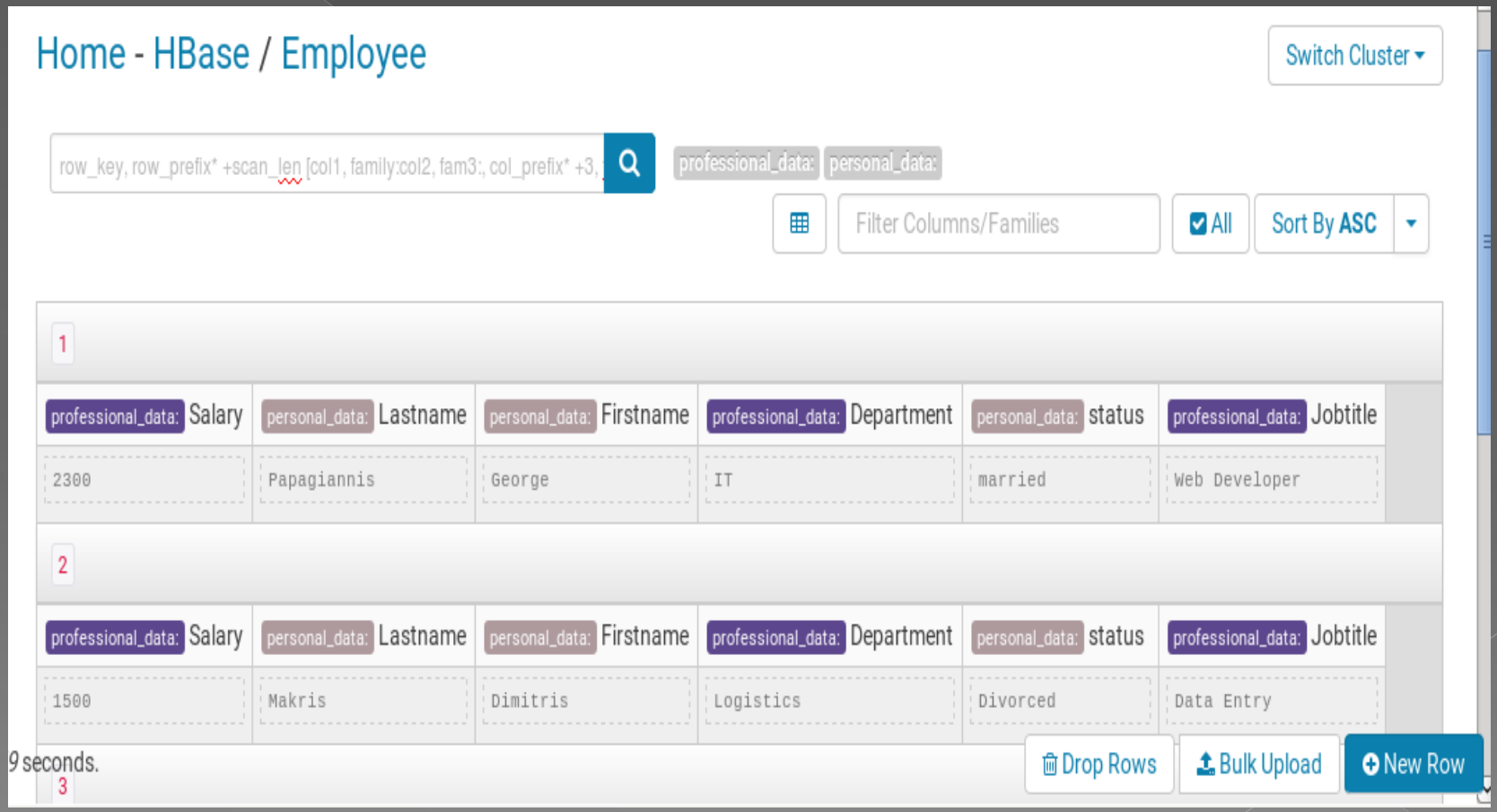
Παράδειγμα Hive



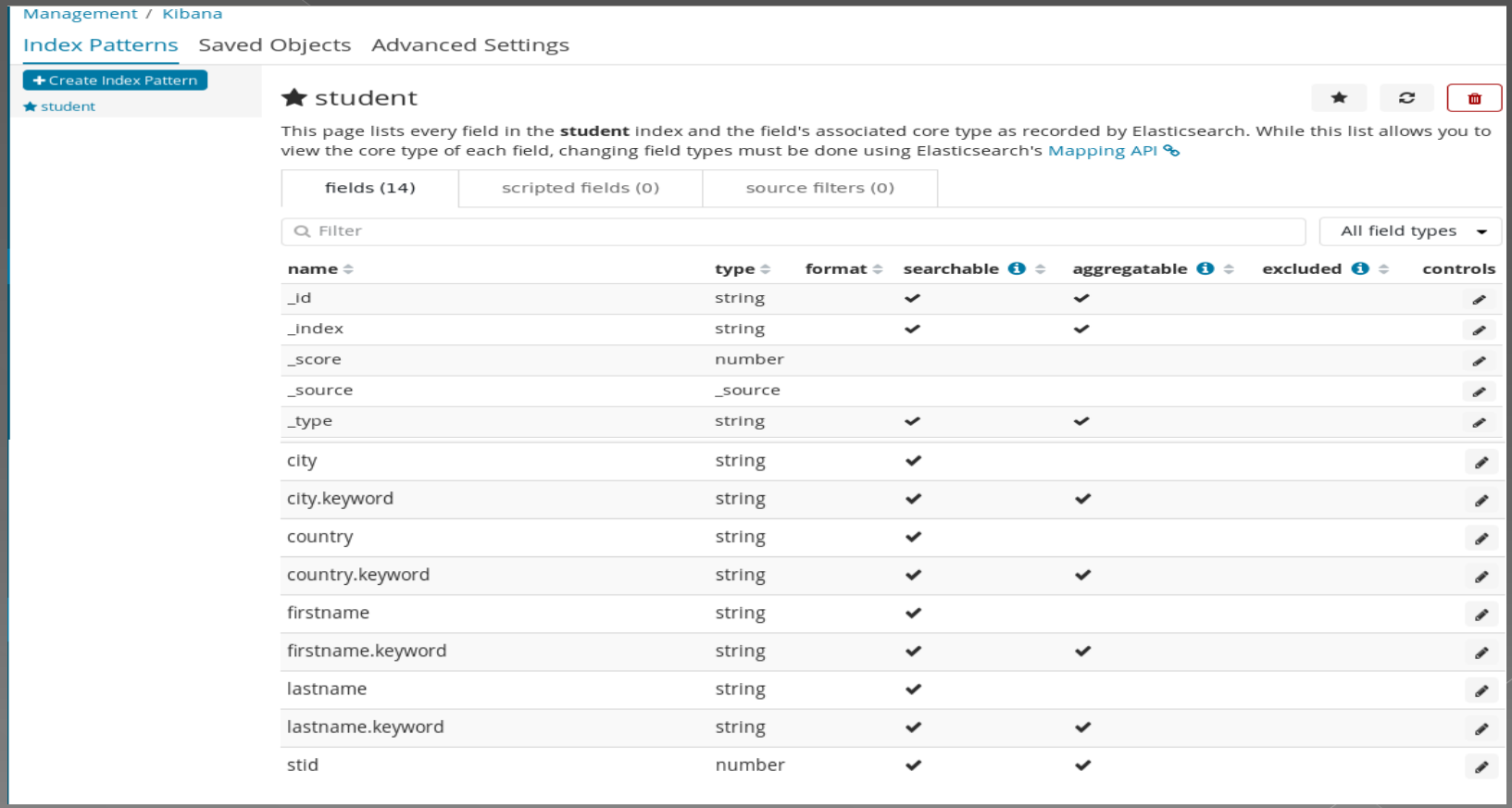
Παράδειγμα: Εντολές HBase



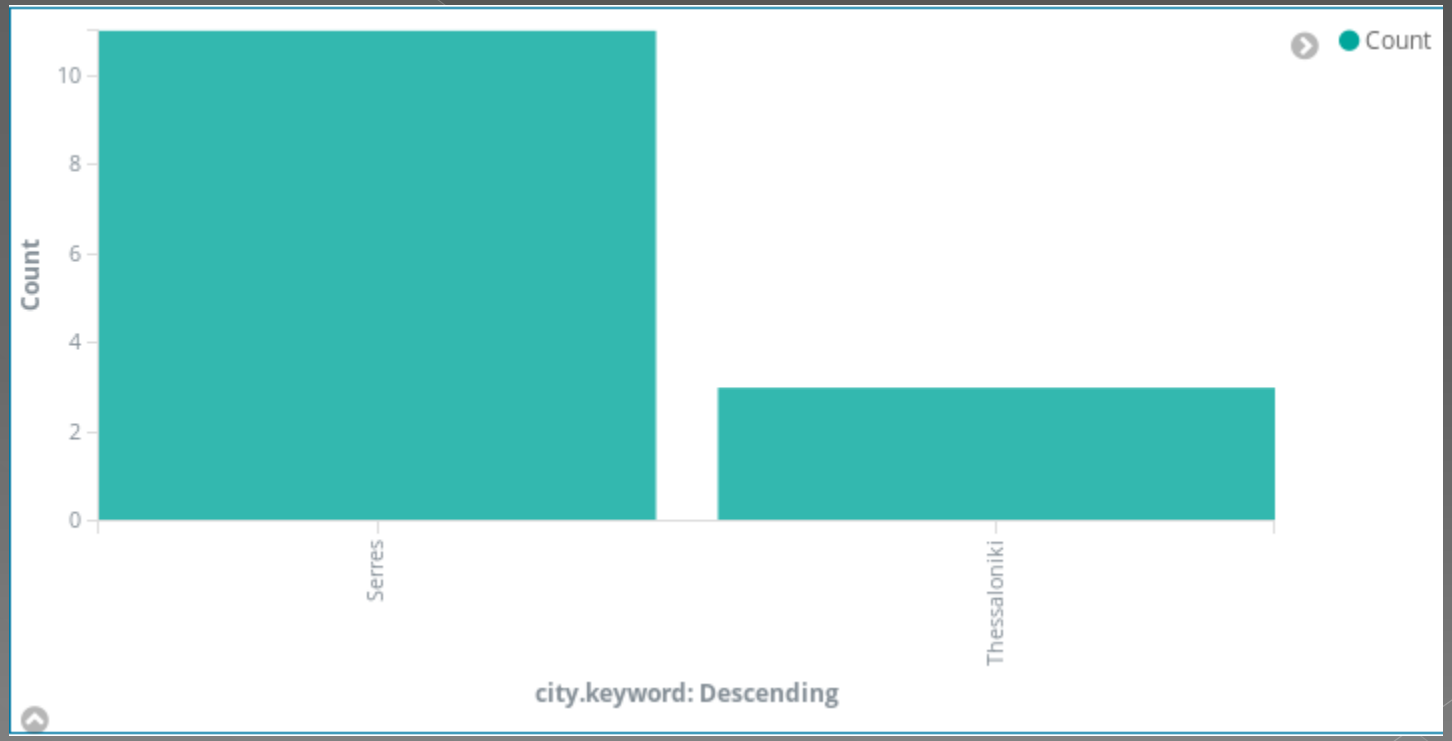
Παράδειγμα HBase μέσω Hue



Δημιουργία ευρετηρίου ElasticSearch - Kibana



Οπτικοποίηση δεδομένων με Kibana



Συμπεράσματα

- Παρουσιάστηκε το **HDFS** και η ευκολία να ανεβάσουμε κάποιο αρχείο σε αυτό μέσω του **Hue**
- Παρουσιάστηκε το **MapReduce** και πως μπορούμε να τρέξουμε ένα απλό **WordCount** πρόγραμμα
- Παρουσιάστηκε το **Hive** και πως μπορούμε να θέσουμε ερωτήματα σε μία βάση μέσω του **Hue**
- Παρουσιάστηκαν οι εντολές της **HBase Shell** καθώς και η δημιουργία πινάκων στην **HBase** μέσω του **Hue**
- Τέλος δημιουργήθηκαν ευρετήρια στα δεδομένα μας με τη βοήθεια του **ElasticSearch** και κατ'επέκταση του **Kibana**

Συμπεράσματα

- Ο συνδυασμός του ElasticSearch με την HBase (παρόλο που το ElasticSearch μπορεί να χρησιμοποιηθεί κι αυτό ως βάση δεδομένων) μπορεί να προσφέρει περισσότερη ασφάλεια, ώστε να μη χαθεί κάποιο μέρος ενός αρχείου, καθώς και μεγαλύτερη ταχύτητα στην αναζήτηση και οπτικοποίηση των δεδομένων μας.

Αναφορές

- Ιστοσελίδες
- [1] WordCount Example on CDH 5.12 διαθέσιμο online: https://www.youtube.com/watch?v=3l7qAHjVOc4 [πρόσβαση 17/06/2018]
- [2] MapReduce Tutorial διαθέσιμο online: https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v1.0 [πρόσβαση 17/06/2018]
- [3] What is Big Data https://www.edureka.co/blog/hadoop-tutorial/#WhatIsBigData [πρόσβαση 17/06/2018]
- [4] Big Data Tutorial https://www.edureka.co/blog/big-data-tutorial [πρόσβαση 17/06/2018]
- [6] HDFS Tutorial https://www.edureka.co/blog/hdfs-tutorial [πρόσβαση 17/06/2018]
- [7] HDFS Architecture https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/ [πρόσβαση 17/06/2018]
- [8] MapReduce Tutorial https://www.edureka.co/blog/mapreduce-tutorial/ [πρόσβαση 17/06/2018]
- [9] Getting started with HBase: The Hadoop database https://www.pluralsight.com/courses/hbase-hadoop-getting-started [πρόσβαση 17/06/2018]
- [10] HBase Architecture: HBase Data Model & HBase Read/Write Mechanism https://www.edureka.co/blog/hbase-architecture/ [πρόσβαση 17/06/2018]
- [11] ElasticSearch Reference: Getting Started https://www.elastic.co/guide/en/elasticsearch/reference/current/_basic_concepts.html [πρόσβαση 17/06/2018]
- Βιβλία
- [5] Aven Jeffrey (2017), Sams Teach Yourself Hadoop in 24 Hours, Εκδόσεις Sams Publishing