METROPOLITAN COLLEGE

CENTRE OF EXCELLENCE

# Streaming Analytics:
# Πώς διαμορφώνουν το παρόν και το μέλλον.

Κουγιανός Νικόλαος

**Μάθημα: Data Ecology**

MSc Data Science
Σχολή Πληροφορικής
Μητροπολιτικό Κολλέγιο
Αθήνα, Ελλάδα

Ημερομηνία:
05/01/2020

# Περιεχόμενα

# Περίληψη

Ζούμε σε έναν κόσμο ο οποίος καθορίζεται από τα δεδομένα. Ο όγκος των δεδομένων που δημιουργήθηκε το 2018 ήταν 33 Zettabytes[1] (Armstrong, 2019) ή 33 δισεκατομμύρια Terrabytes σε ένα πιο κατανοητό μέγεθος, ενώ έως το 2025 το μέγεθος αυτό θα αυξηθεί στα 175 Zettabytes (IDC, 2018). Οι ίδιες μελέτες δείχνουν ότι το ποσοστό των δεδομένων που θα έχει νόημα να περάσουν από την διαδικασία της ανάλυσης θα είναι "μόλις" το 3% (περίπου 5.25 ZB), καθώς το υπόλοιπο 97% θα αποτελεί άχρηστη πληροφορία δίχως δομή που δεν είναι δυνατό να αναλυθεί με κάποιο τρόπο. Το γεγονός αυτό δεν δημιουργεί μόνο προβλήματα στην αποθήκευση, διαχείριση και ανάλυση αυτού του 3% (που είναι ένας τεράστιος όγκος δεδομένων), αλλά βάζει και μεγάλες προκλήσεις στον τρόπο με τον οποίο το υπόλοιπο 97% θα πρέπει να φιλτραριστεί.

Ο σκοπός αυτής της εργασίας είναι να αναλύσει την επιστήμη των Streaming Analytics, τα εργαλεία τα οποία χρησιμοποιούνται για την περισυλλογή, αποθήκευση και ανάλυση δεδομένων σε πραγματικό χρόνο, τον τρόπο με τον οποίο επηρεάζουν την λήψη αποφάσεων σε εταιρίες που κινούν την αγορά εργασίας σήμερα, καθώς και το αντίκτυπο που όλα αυτά έχουν στην κοινωνία και την καθημερινότητα μας. Περιλαμβάνεται επίσης και μια κριτική ανάλυση από μια πιο ανθρωποκεντρική οπτική γωνία, στην οποία περιγράφεται ο τρόπος με τον οποίο πλέον οι άνθρωποι δεν είμαστε τίποτα παραπάνω από ένα σύνολο δεδομένων και αριθμών στην σύγχρονη βιομηχανία.

# Μεθοδολογία

Η μεθοδολογία που χρησιμοποιήθηκε για την συγγραφή της παρούσας εργασίας ήταν ενδελεχής έρευνα βασισμένη σε έμπιστες πηγές του διαδικτύου, άρθρα και δημοσιεύσεις σχετικές με το θέμα των Streaming Analytics, η οποία σε συνδυασμό με τις ήδη υπάρχουσες γνώσεις που έχω από τον χώρο της πληροφορικής διαμόρφωσαν το τελικό αποτέλεσμα. Υποστηρίζω ακράδαντα την άποψη που λέει ότι στην σύγχρονη εποχή μας, οποιοσδήποτε έχει πρόσβαση στο διαδίκτυο και γνωρίζει πώς να κάνει σωστή έρευνα μπορεί να βρει πληροφορίες για οποιοδήποτε θέμα, επιστημονικό ή μη, αλλά και να διασταυρώσει αυτές τις πληροφορίες βασιζόμενος σε επιστημονικές μελέτες και ακαδημαϊκές δημοσιεύσεις. Προσωπικά εκμεταλλευόμενος το πιο χρήσιμο εργαλείο που έχει να επιδείξει η τεχνολογία για τον συγκεκριμένο σκοπό, τις μηχανές αναζήτησης[2], κατάφερα να βρω πληθώρα επιστημονικών άρθρων τα οποία εμπλούτισαν κατά πολύ τις γνώσεις μου πάνω στο συγκεκριμένο θέμα.

[1] 1 ZB = 1.000.000.000 TB = $10^{21}$ bytes

[2] Μια μηχανή αναζήτησης είναι μια εφαρμογή που επιτρέπει την αναζήτηση κειμένων και αρχείων στο Διαδίκτυο (Wikipedia, 2019).

# Εισαγωγή

Streaming Analytics ορίζεται ως η δυνατότητα του συνεχούς υπολογισμού στατιστικών δεδομένων ταυτόχρονα με την ροή των δεδομένων (Freeman, 2016). Με πιο απλά λόγια, είναι η συνεχής επεξεργασία και ανάλυση δεδομένων σε πραγματικό χρόνο καθώς αυτά περισυλλέγονται προερχόμενα από μια ευρεία γκάμα συσκευών, γνωστή και ως Internet of Things (IoT). Το Internet of Things, ή αλλιώς Διαδίκτυο των πραγμάτων, είναι ένα δίκτυο που περιλαμβάνει μια πληθώρα συσκευών, μηχανικών και ψηφιακών, αυτοκινήτων καθώς και οποιουδήποτε αντικειμένου που ενσωματώνει ηλεκτρονικά μέσα και συνδεσιμότητα σε δίκτυο έτσι ώστε να επιτρέπεται η σύνδεση και η ανταλλαγή δεδομένων (Wikipedia, 2019). Απλούστερα, η φιλοσοφία του IoT είναι η σύνδεση όλων, ηλεκτρονικών ή μη, των συσκευών μεταξύ τους σε ένα κοινό δίκτυο.

Πως ακριβώς προέκυψε η ανάγκη ανάπτυξης αυτής της νέας τεχνολογίας; Μέχρι σήμερα με ποιον τρόπο αναλύονται τα δεδομένα; Και εν τέλει, ποια γεγονότα οδήγησαν στην αλλαγή των τεχνικών και εργαλείων με τα οποία λαμβάνονται κρίσιμες αποφάσεις που καθορίζουν το μέλλον των μεγάλων αλλά και των μικρότερων εταιριών;

Σε μια μικρή ιστορική αναδρομή, αξίζει να αναφερθεί και ο τρόπος με τον οποίο γινόταν στο - όχι και τόσο μακρινό - παρελθόν (και γίνεται μέχρι και σήμερα) ανάλυση και επεξεργασία των big data. Αυτός ο τρόπος ονομάζεται Batch Processing, και ορίζεται ως η συλλογή και επεξεργασία παρτίδων (batches) δεδομένων ανά συγκεκριμένα χρονικά διαστήματα, χωρίς να απαιτείται αλληλεπίδραση με τον χρήστη (Barone, 2019). Το Batch Processing χρησιμοποιείται από τις μεγάλες επιχειρήσεις για οργάνωση δεδομένων και δημιουργία αναφορών (report generation) από τα μέσα του 20ου αιώνα, με την εμφάνιση των mainframe computers[3]. Με το πέρασμα των χρόνων όμως, και καθώς πλέον μπαίνουμε σε μια πλήρως ψηφιακή εποχή όπου τα πάντα κινούνται με ραγδαίους ρυθμούς, η ανάγκη για επεξεργασία και εξαγωγή συμπερασμάτων και στατιστικών σε πραγματικό χρόνο, έχει προκαλέσει την δημιουργία της νέας επιστήμης των Streaming Analytics και τον παραγκωνισμό του Batch Processing το οποίο δεν εξυπηρετεί πλέον και τόσο καλά τις ανάγκες της αγοράς εργασίας.

Η αγορά εργασίας λοιπόν, η οποία διαμορφώνεται κατά κύριο λόγο από τις εταιρίες κολοσσούς που έχουν παγκόσμιο πεδίο δράσης, έχει δημιουργήσει άλλες ανάγκες σε αυτές τις εταιρίες. Το Streaming Analytics προσφέρει τεράστιο όφελος, σε πολλούς τομείς οι οποίοι θα αναλυθούν εκτενέστερα παρακάτω. Ενδεικτικά, μια επιχείρηση που επενδύει στην προαναφερθείσα τεχνολογία, μπορεί να δει τεράστιο όφελος σε δύο κρίσιμους τομείς: Ο σημαντικότερος εξ' αυτών είναι η προστασία από κακόβουλες επιθέσεις και η παρακολούθηση διαδικτυακών απειλών, πράγματα τα οποία μπορούν να καταστρέψουν ολοκληρωτικά οποιαδήποτε επιχείρηση αν δε δοθεί η απαιτούμενη σημασία. Ο δεύτερος τομέας έχει να κάνει περισσότερο με την αύξηση των κερδών, αφού επιτρέπει στην επιχείρηση να αποκτήσει μια

[3] Οι κεντρικοί υπολογιστές (mainframes) είναι κατηγορία υπολογιστών που χρησιμοποιούνται κυρίως από κυβερνητικές υπηρεσίες και μεγάλες εταιρίες για κρίσιμες εφαρμογές, όπως μαζική επεξεργασία συναλλαγών και δεδομένων σε απογραφή πληθυσμού, στατιστικές έρευνες βιομηχανιών/καταναλωτών, σχεδιασμός και διαχείριση πόρων κλπ. (Wikipedia, 2019)

σφαιρική και άμεση εικόνα των πελατών και των αναγκών τους, με σκοπό την μετατροπή του marketing σε μια άκρως εξατομικευμένη και στοχευμένη δράση (Hans, 2017).

Όλα τα παραπάνω, αν αναλυθούν από μια κοινωνιολογική οπτική γωνία, οδηγούν στο συμπέρασμα πως η κοινωνία, οι άνθρωποι και οι μεταξύ τους σχέσεις διαμορφώνονται σε τεράστιο βαθμό από το ψηφιακό τους αποτύπωμα, από τα δεδομένα δηλαδή που συλλέγονται και αναλύονται για τον κάθε ένα από εμάς.

# Εργαλεία Streaming Analytics

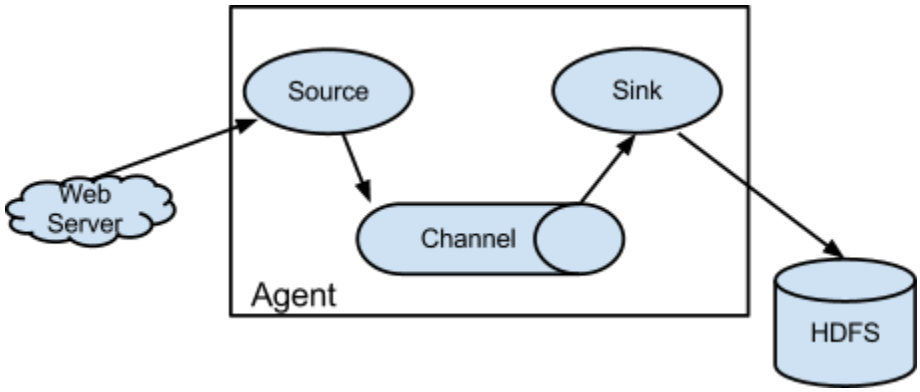
Η ανάγκη για Streaming Analytics έχει δημιουργήσει πρόσφορο έδαφος για την ανάπτυξη διαφόρων τεχνολογιών και εργαλείων, προκειμένου να επιτευχθεί η περισυλλογή, επεξεργασία και ανάλυση δεδομένων σε πραγματικό χρόνο. Παρακάτω περιγράφονται τα κυριότερα εξ' αυτών, και στο τέλος λαμβάνει χώρα και μια σύγκριση με το Batch Processing και τις κύριες διαφορές που έχει με το Streaming Analytics.

**Apache Flume**

Το Apache Flume είναι ένα κατανεμημένο (distributed[4]), αξιόπιστο και συνεχώς διαθέσιμο (high availability[5]) service για την αποδοτική συλλογή, συγκέντρωση και μεταφορά μεγάλων ποσών δεδομένων προερχόμενα από διαφορετικές πηγές σε ένα Hadoop Distributed File System (HDFS[6]) (Apache, 2019). Το Hadoop είναι και αυτό εργαλείο του Apache το οποίο όμως χρησιμοποιείται για Batch Processing. Συνεπώς το Apache Flume συνδυάζει τα θετικά στοιχεία και των 2 κόσμων με αποτέλεσμα μια σχετικά γρήγορη αλλά ταυτόχρονα και δομημένη ανάλυση των δεδομένων που συλλέγονται. Τα κύρια χαρακτηριστικά του είναι:

1. Εγγυημένη παράδοση όλων των δεδομένων.
2. Οριζόντια κλιμάκωση (horizontal scaling) που σημαίνει ότι σε περίπτωση ανάγκης μπορούν να προστεθούν και άλλες ροές για να ικανοποιηθεί ο αυξημένος ρυθμός συλλογής των δεδομένων.
3. Buffering. Σε περίπτωση που ο ρυθμός των δεδομένων υπερβεί τον ρυθμό με τον οποίο το σύστημα μπορεί να τα επεξεργαστεί, τότε αυτά μπαίνουν σε μια προσωρινή μνήμη (buffer[7]) και παραδίδονται αργότερα.

[4] Distributed computing ορίζεται ως ένα μοντέλο στο οποίο συμμετέχουν πολλοί υπολογιστές με αποτέλεσμα την αύξηση της αποδοτικότητας (Rouse, 2015).
[5] High availability είναι η ικανότητα ενός συστήματος να είναι συνεχώς λειτουργικό για μεγάλο χρονικό διάστημα (Rouse, 2019).
[6] Σύστημα με το οποίο ένα ενοποιημένο σύνολο υπολογιστών σπάει σε πολλούς, μικρότερους ευέλικτους κόμβους.
[7] Buffer είναι μια προσωρινή μνήμη στην οποία αποθηκεύονται προσωρινά δεδομένα καθώς μεταφέρονται από ένα μέρος σε ένα άλλο.




*Εικόνα 1. Αρχιτεκτονική Apache Flume*

**Apache Storm**

Είναι ένα open source[8] λογισμικό που κάνει αναλύσεις σε ροές (streams) δεδομένων καθώς αυτά συλλέγονται. Είναι επεκτάσιμο, με ανοχή σφαλμάτων (fault tolerant[9]) και σύμφωνα με μετρήσεις είναι ικανό να διαχειριστεί και να αναλύσει εκατομμύρια bytes δεδομένων ανά δευτερόλεπτο, ανά κόμβο (Shoro & Soomro, 2015). Είναι επίσης σχετικά απλό και εύκολο στην χρήση του, μπορεί να χρησιμοποιηθεί με οποιαδήποτε γλώσσα προγραμματισμού και εγγυάται σίγουρη επεξεργασία των δεδομένων (Apache, 2019). Το κύριο χαρακτηριστικό του είναι η πολύ μεγάλη ταχύτητα του, η οποία το καθιστά κατάλληλο σε κλάδους όπου η αμεσότητα φαντάζει αδήριτη ανάγκη. Χαρακτηριστικό παράδειγμα είναι το cyber security analytics and threat detection, τομέας στον οποίο ακόμα και ελάχιστα δευτερόλεπτα έχουν τεράστια σημασία και μπορούν να καθορίσουν την ασφάλεια μιας επιχείρησης από κακόβουλες επιθέσεις.

Η αρχιτεκτονική του Apache Storm είναι η παρακάτω:

- Spouts. Αποτελούν σημεία που λειτουργούν σαν “πύλες εισόδου” για τα δεδομένα. Ένα spout θα συλλέξει δεδομένα από εξωτερικές πηγές (δημόσια APIs[10], εξωτερικές βάσεις δεδομένων) και θα τα δρομολογήσει προς τα bolts.
- Bolts. Τμήματα στα οποία γίνεται το φιλτράρισμα, ο καθαρισμός (data sanitation), η ανάλυση και τελικά η αποστολή των δεδομένων στο UI[11] για να τα δει μορφοποιημένα ο τελικός χρήστης.
- Topologies. Είναι τα δίκτυα τα οποία περιέχουν σύνολα από spouts και bolts.

Ένα πολύ σημαντικό χαρακτηριστικό της προαναφερθείσας αρχιτεκτονικής είναι πως τα δεδομένα κινούνται προς τα μια κατεύθυνση για συμβατικούς κυρίως λόγους (Ashraf, 2018).

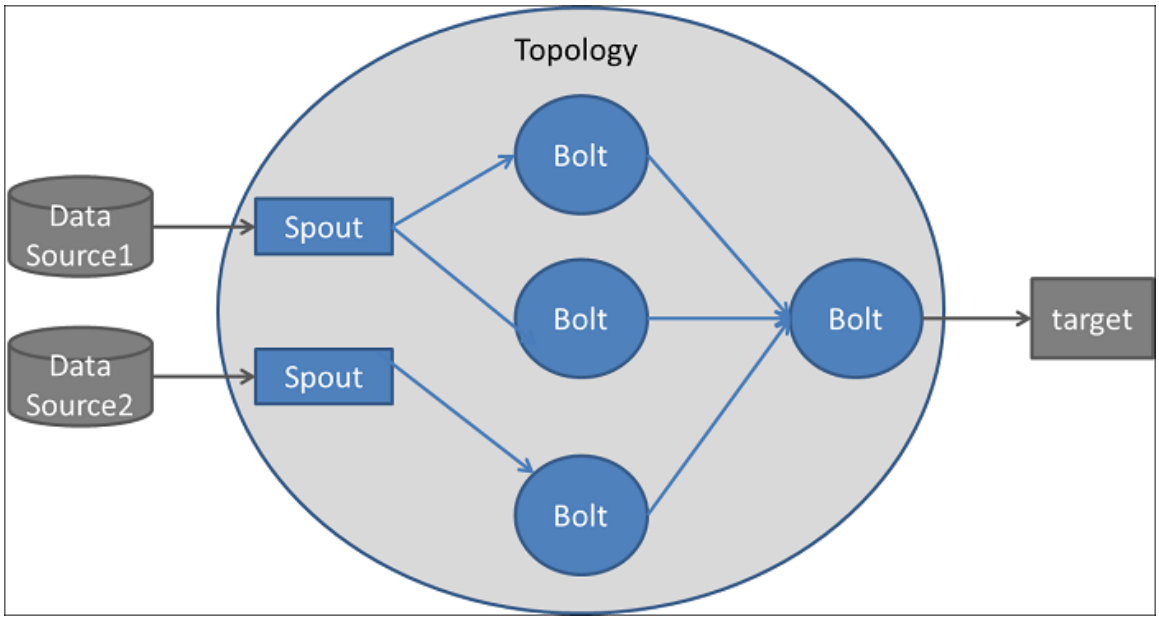
---

[8] Open source ονομάζεται ένα λογισμικό που μας επιτρέπει να δούμε τον πηγαίο του κώδικα, να τον τροποποιήσουμε και να τον μοιραστούμε ελεύθερα.

[9] Χαρακτηριστικό που επιτρέπει σε ένα σύστημα να συνεχίζει να λειτουργεί απροβλημάτιστα ακόμα και σε περίπτωση λάθους.

[10] Application Programming Interface. Είναι ένα ενδιάμεσο λογισμικό που επιτρέπει την επικοινωνία μεταξύ δύο εφαρμογών.

[11] User Interface. Το σύνολο εικόνων, γραφημάτων και πληροφοριών που εμφανίζονται στην οθόνη και προκαλούν αλληλεπίδραση ανάμεσα στον χρήστη και σε μια εφαρμογή.




*Εικόνα 2. Αρχιτεκτονική Apache Storm*

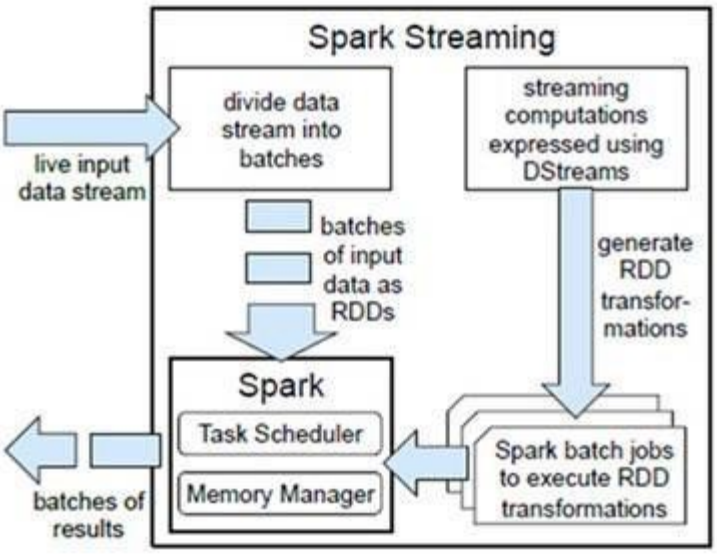
**Spark Streaming**

Το Spark Streaming είναι μια επέκταση του Apache Spark[12] API η οποία επιτρέπει ζωντανή επεξεργασία δεδομένων που προέρχονται από πολλές διαφορετικές πηγές[13] (Apache, 2019). Το συγκεκριμένο framework έχει ένα πιο περιεκτικό API σε σχέση με τα υπόλοιπα και ως αποτέλεσμα υπάρχει μικρότερη λογική (και μικρότερο learning curve[14]) στις εφαρμογές του και παρατηρούνται αυξημένες επιδόσεις χάρη στο in-memory caching σύστημα που διαθέτει. Αυτό συμβαίνει γιατί τα δεδομένα δε χρειάζεται να γράφονται και να διαβάζονται από τον δίσκο σε κάθε υπολογιστικό βήμα.

Είναι γραμμένο σε Scala, αλλά υποστηρίζει επίσης και τις γλώσσες Java, Python, SQL και R. Έγινε open source το 2010, και ανήκει στο Apache software foundation από το 2014. Η ικανότητα του Spark να συνδυάζεται με την Java μέσω Java libraries σε συνδυασμό με τις εκπληκτικές επιδόσεις του το έχουν καταστήσει μια δημοφιλή επιλογή για εταιρίες κολοσσούς όπως το eBay, το TripAdvisor και το Netflix (Dezyre, 2016).

Η κύρια αρχιτεκτονική του Spark Streaming είναι κατανεμημένες και αμετάβλητες συλλογές (distributed & immutable collections) που ονομάζονται resilient distributed datasets (RDDs), οι οποίες μπορούν να υποστούν επεξεργασία μέσω ντετερμινιστικών διεργασιών.

[12] Apache Spark: Ενοποιημένο σύστημα ανάλυσης και επεξεργασίας δεδομένων με Batch Processing και Streaming Analytics, που βασίζεται σε προσωρινή μνήμη (cache).
[13] Ενδεικτικές πηγές: Kafka, Flume, Kinesis, TCP sockets.
[14] Learning curve ή καμπύλη εκμάθησης είναι ένα διάγραμμα δύο διαστάσεων εκμάθησης και εμπειρίας που δείχνει κατά πόσο μαθαίνεις πως λειτουργεί κάτι όσο ασχολείσαι με αυτό. Με πιο απλά λόγια είναι η ευκολία εκμάθησης.




*Εικόνα 3. Αρχιτεκτονική Spark Streaming*

Όπως φαίνεται και στην εικόνα 3, το Spark Streaming δέχεται σε πραγματικό χρόνο εισερχόμενες ροές δεδομένων, τις οποίες χωρίζει σε batch[15] και τις σώζει στην μνήμη σαν RDDs. Στην συνέχεια το σύστημα επεξεργάζεται αυτά τα batch και παράγει τις τελικές ροές αποτελεσμάτων σε παρτίδες.

**Apache Flink**

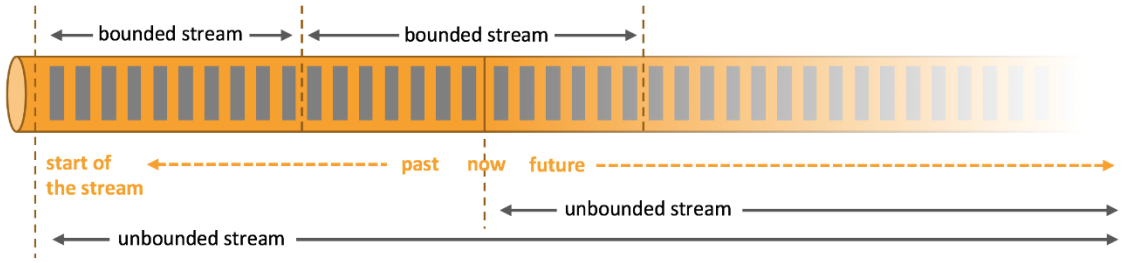
Άλλο ένα open source framework που έχει αναπτυχθεί και υποστηρίζεται από την Apache. Το συγκεκριμένο χαρακτηρίζεται ως μια επεξεργαστική μηχανή η οποία δέχεται οριοθετημένες και χωρίς όρια (bounded and unbounded) ροές δεδομένων (Flink, 2019). Το Flink έχει σχεδιαστεί με τέτοιο τρόπο ώστε να τρέχει σε όλα τα κοινά περιβάλλοντα, και να εκτελεί υπολογισμούς στην προσωρινή μνήμη οποιουδήποτε μεγέθους.

Όπως προαναφέρθηκε, το συγκεκριμένο εργαλείο δέχεται δύο είδη δεδομένων:

- Unbounded streams. Έχουν οριοθετημένη αρχή αλλά δεν έχουν τέλος. Δεν τερματίζουν ποτέ και εξάγουν αποτελέσματα συνεχώς και σε πραγματικό χρόνο. Ο συγκεκριμένος τύπος δεδομένων πρέπει να περνάει μέσα από αδιάκοπη επεξεργασία χωρίς να περιμένει μια ροή να τελειώσει, αφού το τέλος είναι αόριστο. Σημαντικό είναι να τηρείται η σειρά με την οποία έρχονται τα δεδομένα, έτσι ώστε να είναι δυνατόν να εξαχθούν λογικά αποτελέσματα.
- Bounded streams. Έχουν όρια, συγκεκριμένη αρχή και συγκεκριμένο τέλος. Οι οριοθετημένες ροές, εφόσον το τέλος τους είναι γνωστό, μπορούν πρώτα να προσληφθούν στο 100% πριν περάσουν οποιαδήποτε επεξεργασία. Εδώ δεν είναι σημαντικό να διατηρείται η αρχική σειρά με την οποία παραλαμβάνονται τα δεδομένα,

[15] Στην πληροφορική batch ονομάζουμε μια παρτίδα δεδομένων, εντολών ή αρχείων τα οποία είναι ομαδοποιημένα και μπορούν να τρέξουν ή να υποστούν επεξεργασία χωρίς την παρέμβαση του χρήστη.

καθώς μόλις τελειώσει το data ingestion[16] τα δεδομένα μπορούν να ταξινομηθούν εκ των υστέρων.




*Εικόνα 4. Bounded and unbounded streams.*

Το Flink μπορεί να τρέξει σε εφαρμογές οποιουδήποτε μεγέθους. Η βελτιστοποίηση απόδοσης που έχει γίνει του επιτρέπει να τρέχει παράλληλα σε χιλιάδες πυρήνες οι οποίοι ταυτόχρονα επεξεργάζονται και αναλύουν δεδομένα μέσα σε ένα cluster[17] (όπως Hadoop YARN, Apache Mesos, Kubernetes). Είναι αρκετά εύκολο στη χρήση και επιτρέπει στον χρήστη να το χειρίζεται και να αλλάζει τις ρυθμίσεις του ανάλογα με τις ανάγκες του, με την χρήση απλών REST API κλήσεων. Το γεγονός ότι μπορεί να εκμεταλλευτεί θεωρητικά απεριόριστους πόρους (μνήμη, επεξεργαστή, σκληρούς δίσκους, δίκτυο) και να αποδίδει το ίδιο καλά, το έχει κάνει ευρέως διαδεδομένο (Flink, 2019).

Εν κατακλείδι, το Apache Flink περιλαμβάνει δύο APIs για την ανάλυση και επεξεργασία δεδομένων:

- Dataset API: Χρησιμοποιείται σε συνδυασμό με τα bounded streams ως batch processing. Τα datasets δημιουργούνται από διάφορες πηγές και τα αποτελέσματα επιστρέφονται μέσω sinks τα οποία μπορούν τα γράψουν είτε σε ένα αρχείο, είτε στο standard output (για παράδειγμα το command line terminal).
- DataStream API: Χρησιμοποιείται σε συνδυασμό με τα unbounded streams ως Streaming Analytics. Οι ροές δεδομένων προέρχονται και αυτές από διαφορετικές πηγές οι οποίες είναι διαφορετικές από αυτές του Dataset API (ουρές μηνυμάτων, sockets). Τα αποτελέσματα επιστρέφονται με τον ίδιο τρόπο.

**Apache Kafka**

Η συγκεκριμένη τεχνολογία είναι σχετικά καινούργια σε σχέση με τις υπόλοιπες, και χρησιμοποιείται κυρίως για logging και για ανταλλαγή σημαντικών μηνυμάτων μεταξύ εφαρμογών. Είναι γρήγορη, fault tolerant και προσφέρει τεράστιο ρυθμό ανταλλαγής μηνυμάτων. Αναπτύχθηκε από την LinkedIn το 2011 και στη συνέχεια έγινε δωρεά στην Apache (Wikipedia, 2019). Επίσης έχει γραφτεί σε Scala και Java και προσφέρει εύκολη σύνδεση με

[16] Data ingestion είναι η διαδικασία παραλαβής και εισαγωγής δεδομένων σε κάποιο υπολογιστικό σύστημα, προκειμένου να χρησιμοποιηθούν άμεσα για κάποιο σκοπό ή να αποθηκευτούν σε μια βάση δεδομένων.

[17] Ένα σύνολο υπολογιστικών συστημάτων τα οποία δουλεύουν μαζί για κοινό σκοπό και μπορούν να παρουσιαστούν ως ένα ενιαίο σύστημα.

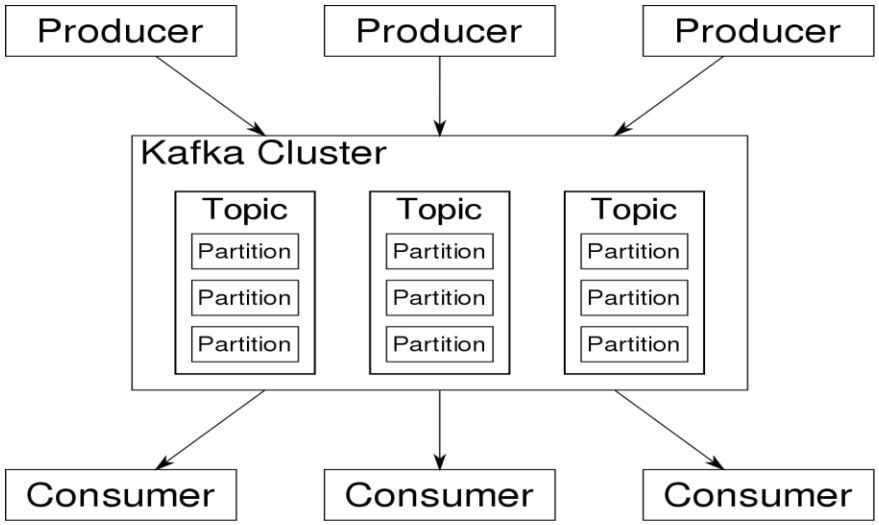
εξωτερικά συστήματα μέσω του Connector API (περισσότερα παρακάτω) και παρέχει και το Kafka Streams, που είναι μια βιβλιοθήκη της Java η οποία μας επιτρέπει να εκμεταλλευτούμε τις δυνατότητες του Kafka.

Η αρχιτεκτονική του συγκεκριμένου εργαλείου διαφέρει σε σχέση με τα υπόλοιπα. Μπορούμε να φανταστούμε το μέρος όπου αποθηκεύονται τα μηνύματα σαν ένα ενιαίο χώρο (Kafka cluster) μέσα στον οποίο τοποθετούν μηνύματα οι παραγωγοί (producers) τα οποία διαβάζονται από τους καταναλωτές (consumers) (Kafka, 2019). Το εκπληκτικό χαρακτηριστικό του Kafka Streams είναι το μοντέλο publish-subscribe, το οποίο μας εξασφαλίζει ότι ένα μήνυμα μπορούν να το διαβάσουν όσοι καταναλωτές έχουν κάνει εγγραφή (subscribe) πάνω στο συγκεκριμένο θέμα (topic).

Ένα σχετικό παράδειγμα για να γίνει πιο κατανοητή η συγκεκριμένη τεχνολογία είναι το εξής: Έστω μια εφαρμογή η οποία παράγει δεδομένα κάθε μια ώρα, τα οποία αντλεί από μια εξωτερική πηγή, σχετικά με τον καιρό. Γράφει αυτά τα δεδομένα σε ένα Kafka topic, στο οποίο έχουν κάνει subscribe δύο ανεξάρτητες και διαφορετικές εφαρμογές. Αυτές οι εφαρμογές λοιπόν μπορούν ταυτόχρονα να καταναλώνουν τα ίδια δεδομένα αλλά για διαφορετικό σκοπό. Έστω λοιπόν ότι η μια θέλει να κάνει ημερήσια πρόβλεψη της θερμοκρασίας της επόμενης μέρας, βασισμένη σε δεδομένα προηγούμενων ημερών, και η άλλη να βγάζει συγκεντρωτικές αναλύσεις της υγρασίας κάθε μήνα. Με βάση αυτό το μοντέλο βλέπουμε ότι εξοικονομείται μεγάλος αριθμός πόρων, καθώς με έναν producer εξυπηρετούνται περισσότεροι από έναν consumers, και η δουλειά γίνεται παράλληλα.

Τα τέσσερα APIs που περιλαμβάνει το Apache Kafka είναι τα παρακάτω:

- Producer API. Επιτρέπει σε μια εφαρμογή να δημοσιεύει σε topics ροές δεδομένων ή μηνύματα.
- Consumer API. Επιτρέπει σε μια εφαρμογή να κάνει subscribe σε topics και να καταναλώνει δεδομένα από εκεί.
- Connector API. Χρησιμοποιεί τα δύο προαναφερθέντα APIs και περιέχει όλη την λογική του μοντέλου publish-subscribe, συνδέοντας Kafka topics σε ήδη υπάρχουσες εφαρμογές.
- Streams API. Είναι μια fault tolerant βιβλιοθήκη για stream processing η οποία παρέχει πολλές δυνατότητες για επεξεργασία δεδομένων, όπως είναι τα φίλτρα, οι ομαδοποιήσεις, τα tables κ.α. Για να επιτευχθεί το fault tolerance, κάθε ενημέρωση σε δεδομένα που γίνεται τοπικά από μια εφαρμογή, γράφεται επίσης και σε ένα Kafka topic.




*Εικόνα 5. Αρχιτεκτονική Apache Kafka.*

**Σύνοψη**

Με την ραγδαία αύξηση της τεχνολογίας το μόνο σίγουρο είναι ότι θα συνεχίσουν να αναπτύσσονται συνεχώς νέα εργαλεία που εξυπηρετούν το Streaming Analytics. Παρατηρούμε πως η Apache διατηρεί τη μερίδα του λέοντος προς το παρόν, αλλά η ιστορία μας έχει διδάξει πως κάτι τέτοιο μπορεί να αλλάξει με την πάροδο των χρόνων. Όλες οι τεχνολογίες σε γενικές γραμμές εξυπηρετούν τον ίδιο σκοπό, ο οποίος είναι η γρήγορη και σωστή συλλογή, επεξεργασία και ανάλυση δεδομένων μέσα από την οποία μπορούν να εξάγονται συμπεράσματα σε πραγματικό χρόνο τα οποία θα καθορίζουν στρατηγικές κινήσεις εταιριών. Οι διαφορές τους βρίσκονται στις λεπτομέρειες, και απαιτείται ενδελεχής έρευνα από εξειδικευμένο προσωπικό προκειμένου να επιλεχθεί η σωστή τεχνολογία η οποία καλύπτει απόλυτα τις ανάγκες μιας επιχείρησης.

# Σύγκριση Streaming Analytics με Batch Processing

Όπως έχει προαναφερθεί, το Batch Processing είναι η επεξεργασία μεγάλου όγκου δεδομένων σε παρτίδες. Τα δεδομένα περιέχουν εκατομμύρια εγγραφές μέσα σε μια μέρα, οι οποίες αποθηκεύονται σε βάσεις δεδομένων ή αρχεία (Shiff, 2018). Η δουλειά της επεξεργασίας συνήθως ολοκληρώνεται με συνεχή και διαδοχικό τρόπο, δηλαδή η διαδικασία δεν σταματάει μέχρι να τελειώσει και τα δεδομένα αναλύονται με την σειρά που αποθηκεύτηκαν. Ένα κλασσικό παράδειγμα Batch Processing μπορεί να είναι η μηνιαία μισθοδοσία όλων των υπαλλήλων μιας υπηρεσίας, η οποία μπορεί να τρέχει τις πρώτες μέρες κάθε μήνα. Ο συγκεκριμένος τρόπος επεξεργασίας δεδομένων είναι εξαιρετικά αποδοτικός όταν έχουμε να κάνουμε με τεράστιους όγκους δεδομένων τα οποία συλλέγονται με την πάροδο του χρόνου. Βοηθάει στην μείωση του λειτουργικού κόστους των επιχειρήσεων καθώς αυτοματοποιείται μια διαδικασία η οποία θα απαιτούσε πολλές εργατοώρες για να ολοκληρωθεί από υπαλλήλους. Στα πολύ θετικά βρίσκεται επίσης και το γεγονός ότι δίνει

πλήρη έλεγχο στους υπεύθυνους σχετικά με το πότε ακριβώς θα τρέξουν το Batch Processing, και με ποιον ακριβώς τρόπο.

Στα αρνητικά τώρα, όπως με όλα τα συστήματα έτσι και με το Batch Processing η διαδικασία του debugging[18] μπορεί να αποδειχθεί αρκετά πολύπλοκη και χρονοβόρα, και αν η εταιρία δεν διαθέτει εξουσιοδοτημένο IT τμήμα για την επίλυση, τότε η ανάθεση σε εξωτερικούς συνεργάτες θα είναι αρκετά κοστοβόρα και επιβλαβής στα οικονομικά της. Το δεύτερο αρνητικό έγκειται στο γεγονός πως όπως όλες οι τεχνολογίες, έτσι και η συγκεκριμένη απαιτεί αρκετό χρόνο εκπαίδευσης και εκμάθησης (που σημαίνει και αρκετό κόστος) προκειμένου να επιτευχθεί ένα αποδοτικό Batch Processing το οποίο παράγει σωστά αποτελέσματα.

Το Streaming Analytics από την άλλη είναι η διαδικασία της ανάλυσης δεδομένων σε πραγματικό χρόνο, για data τα οποία μεταφέρονται ως ροές από μια συσκευή σε μια άλλη (Shiff, 2018). Αυτή η μέθοδος της συνεχούς επεξεργασίας δε γίνεται σε προγραμματισμένα χρονικά διαστήματα, και προσδίδει μεγάλο όφελος σε περιπτώσεις όπου τα δεδομένα συλλέγονται συχνά και ανά τακτά χρονικά διαστήματα. Χρησιμοποιείται σε περιπτώσεις όπου τα συμβάντα πρέπει να αναγνωρίζονται αμέσως και να εκτελείται μια συγκεκριμένη δράση ανάλογα με αυτά. Κλασικό παράδειγμα είναι η προστασία από κακόβουλες ενέργειες διαδικτύου, περίπτωση στην οποία αναλύονται συνεχώς οι εισερχόμενες συνδέσεις στο δίκτυο μιας επιχείρησης, και εφόσον παρατηρηθεί ότι σχηματίζεται κάποιο σχήμα (pattern) τότε αυτόματα το σύστημα μπορεί να “κόψει” την συγκεκριμένη ip[19].

Ένα από τα αρνητικά αυτής της τεχνολογίας είναι η δυσκολία της διατήρησης του output rate στα ίδια (αν όχι καλύτερα) επίπεδα με το input rate. Το output rate είναι ο ρυθμός με τον οποίο αναλύονται τα δεδομένα και εξάγονται αποτελέσματα, και το input rate είναι ο ρυθμός με τον οποίο εισέρχονται τα δεδομένα για ανάλυση. Σε περίπτωση που το input είναι πολύ μεγαλύτερο από το output, τότε δημιουργείται πρόβλημα στην χωρητικότητα και στην μνήμη του υπολογιστικού συστήματος. Επιπροσθέτως, ένα ακόμη πρόβλημα που δημιουργείται συχνά είναι το κατά πόσο είναι εκμεταλλεύσιμα τα δεδομένα τα οποία συλλέγονται, γεγονός το οποίο πολλές φορές οδηγεί σε αυξημένη κατανάλωση πόρων χωρίς να υπάρχει κάποιο ουσιαστικό αποτέλεσμα.

[18] Η διαδικασία αναγνώρισης και εξάλειψης λαθών σε ένα υπολογιστικό σύστημα, σε επίπεδο hardware και software.

[19] Διεύθυνση διαδικτυακού πρωτοκόλλου. Μια μοναδική διεύθυνση που χρησιμοποιείται από συσκευές σε ένα δίκτυο υπολογιστών για τη μεταξύ τους αναγνώριση και συνεννόηση.

Συνοψίζοντας, ακολουθεί μια οργανωμένη παρουσίαση των θετικών και των αρνητικών στοιχείων της κάθε τεχνολογίας (Rehman, 2019) :

| ***Batch Processing*** | ***Streaming Analytics*** |
|---|---|
| ***Οι τυπικές και επαναλαμβανόμενες δουλειές γίνονται γρήγορα χωρίς την παρέμβαση ανθρώπου.*** | ***Ανάλυση δεδομένων και εξαγωγή συμπερασμάτων σε πραγματικό χρόνο.*** |
| ***Δε χρειάζεται εξειδικευμένος εξοπλισμός και συστήματα.*** | ***Απόκτηση πλήρους εικόνας για τον πελάτη.*** |
| ***Ανάλυση τεράστιου όγκου δεδομένων offline.*** | ***Διερεύνηση και επίλυση θεμάτων σε σχετικά μικρό χρόνο.*** |
| ***Πλήρης έλεγχος στο χρονοδιάγραμμα επεξεργασίας και ανάλυσης δεδομένων.*** | ***Δημιουργία προτάσεων (recommendation engines).*** |
| ***Πολύπλοκο debugging.*** | ***Πρόληψη ανεπιθύμητων συμβάντων.*** |
| ***Χρονοβόρα & κοστοβόρα εκπαίδευση και εκμάθηση της τεχνολογίας.*** | ***Έλλειψη εξειδίκευσης πάνω σε Streaming Analytics.*** |
| ***Δεν προσφέρει αποτελέσματα σε πραγματικό χρόνο, κάτι το οποίο απαιτεί η αγορά σήμερα.*** | ***Δυσκολία διατήρησης input rate & output rate.*** |
| ***Αν κάποια συγκεκριμένη δουλειά σταματήσει λόγω λάθους, εμποδίζεται ολόκληρη η διαδικασία.*** | ***Καθαρότητα δεδομένων.*** |
| | ***Μεγάλη κατανάλωση πόρων.*** |

## Πως επηρεάζονται οι επιχειρήσεις

Ζούμε σε μια εποχή όπου η πληροφορία και τα δεδομένα αποτελούν την πιο ισχυρή και πολύτιμη μορφή χρήματος (Kh, 2016). Όσο πιο πολλές πληροφορίες συλλέγει μια εταιρία, τόσο πιο αποδοτικά μπορεί να προβλέψει ορισμένες ενέργειες και να λάβει ασφαλείς αποφάσεις για το μέλλον της. Υπάρχουν 4 κύριοι πυλώνες πάνω στους οποίους οι επιχειρήσεις επηρεάζονται από τα Big Data και κατ' επέκταση το Streaming Analytics.

**Ασφάλεια**

Αδιαμφισβήτητα ο πιο σημαντικός πυλώνας (και αυτός με το μεγαλύτερο impact στο παρόν και μέλλον μιας επιχείρησης) είναι ο τομέας της ασφάλειας. Σύμφωνα με έρευνες, πολλές εταιρίες από διαφορετικούς κλάδους έχουν αναφέρει πως έχουν υπάρξει θύματα κακόβουλων ενεργειών (Hans, 2017). Κατά μέσο όρο χρειάστηκαν 245 μέρες για να συνειδητοποιήσουν πως υπήρξε εισβολή, ενώ το 41% ανέφερε πως οι επιθέσεις προήλθαν από εσωτερικές πηγές. Οι εισβολείς σε τέτοιες επιθέσεις (cyber attacks[20]) συνήθως στοχεύουν σε

[20] Κυβερνοεπίθεση είναι μια διαδικτυακή, κακόβουλη και συνειδητή ενέργεια ενός ατόμου ή οργανισμού, ώστε να παραβιάσουν τα δεδομένα, ευαίσθητα και μη, ενός άλλου οργανισμού προκειμένου να τα χρησιμοποιήσουν προς όφελος τους.

προσωπικά δεδομένα και πιστωτικές κάρτες, τα οποία μπορούν εύκολα να πουλήσουν στην μαύρη αγορά.

Με τον αριθμό αυτών των επιθέσεων να αυξάνεται συνεχώς τα τελευταία έτη, την πρόβλεψη και αντιμετώπιση τους έρχεται να λύσει το Streaming Analytics. Τα Security Operation Centers (SOCs) των μεγάλων επιχειρήσεων, δηλαδή τα κέντρα που είναι υπεύθυνα για την ασφάλεια του οργανισμού, συνήθιζαν να συλλέγουν ιστορικά δεδομένα και να βασίζονται στο Batch Processing, δηλαδή στην εκ των υστέρων ανάλυση τους, προκειμένου να πετύχουν την απαιτούμενη προστασία. Όμως η ιστορία έχει δείξει πως σε επείγουσες ανάγκες, όπως όταν διακυβεύεται το μέλλον μιας εταιρίας εξαιτίας μιας μεγάλης επίθεσης, αυτή η λύση δεν είναι αποδεκτή (Hans, 2017). Αντιθέτως, η ανάλυση δεδομένων σε πραγματικό χρόνο, σε συνδυασμό με την εκμάθηση μηχανών (machine learning) μπορούν να προσφέρουν τεράστια ασφάλεια σε αυτόν τον τομέα.

Ένα χαρακτηριστικό παράδειγμα όπου το Streaming Analytics έχει βοηθήσει στην ασφάλεια σε πραγματικό χρόνο, είναι το IP Blacklisting[21]. Οι hackers πολλές φορές προσπαθούν με μεθόδους brute force να αποκτήσουν πρόσβαση στους servers μιας εταιρίας. Η αντιμετώπιση αυτής της επίθεσης μπορεί να επιτευχθεί εάν σε πραγματικό χρόνο αναγνωριστεί ότι γίνεται αυτή η προσπάθεια από μια συγκεκριμένη IP, έτσι ώστε να αποκλειστεί αυτή η IP και τα SOCs να βρίσκονται σε εγρήγορση για επιπλέον επιθέσεις.

Το δεύτερο πιο δημοφιλές use case των Streaming Analytics στον τομέα της ασφάλειας είναι το Geolocation, και συνήθως έχει εφαρμογή στην αποτροπή χρήσης κλεμμένων πιστωτικών καρτών. Κάθε φορά που ένας χρήστης εισάγει τα στοιχεία της κάρτας του για διαδικτυακή αγορά ενός προϊόντος ή μιας υπηρεσίας, αναλύεται και αποθηκεύεται η διεύθυνση IP από το μηχάνημα που χρησιμοποίησε. Κάθε διεύθυνση IP μπορεί να δείξει και την περιοχή στην οποία βρίσκεται το σύστημα που την έχει. Επίσης, στα περισσότερα sites που σχετίζονται με διαδικτυακές αγορές, ο χρήστης έχει δηλώσει και την περιοχή στην οποία βρίσκεται. Σε κάθε αγορά που λαμβάνει χώρα λοιπόν, το Streaming Analytics αναλύει σε πραγματικό χρόνο και παράλληλα δύο συνδυασμούς δεδομένων. Αρχικά αναλύεται, εφόσον υπάρχει, το ιστορικό προηγούμενων παραγγελιών και οι IPs από τις οποίες προήλθαν. Στην συνέχεια συγκρίνεται η IP από την οποία έγινε η αγορά με την IP που έχει δηλώσει ο χρήστης, και με τις IPs του ιστορικού. Εάν υπάρχει διαφορά, τότε το σύστημα (ανάλογα με την αυστηρότητα του) είτε αποκλείει αμέσως την αγορά και ενημερώνει τον χρήστη πως πιθανότατα υπάρχει παραβίαση της κάρτας του, είτε χρησιμοποιεί μεθόδους ταυτοποίησης (όπως πχ Two Factor Authentication) για να αποκλείσει την πιθανότητα να πρόκειται για κακόβουλη ενέργεια.

Σε γενικότερα πλαίσια, το Streaming Analytics μπορεί να βοηθήσει σημαντικά στον τομέα της ασφάλειας αναλύοντας και αναγνωρίζοντας σε πραγματικό χρόνο οποιαδήποτε ανωμαλία προκύπτει, βρίσκοντας patterns εκεί που μια συμβατική ανάλυση δεδομένων είναι αδύνατον να βρει.

[21] Η απαγόρευση σύνδεσης μιας διεύθυνσης IP στους σέρβερ μιας εταιρίας.

**Παραγωγικότητα**

Η ενίσχυση των λειτουργικών διαδικασιών μιας εταιρίας είναι ένας γρήγορος τρόπος αύξησης του κέρδους, που είναι και ο απώτερος σκοπός γενικότερα. Η συλλογή και επεξεργασία δεδομένων σε μια μεγάλη γραμμή παραγωγής μπορεί να δώσει εξαιρετικά πολύτιμες πληροφορίες σε έναν business owner σχετικά με το που πονάει η επιχείρηση, ποιο τμήμα χρειάζεται περισσότερο ανθρώπινο δυναμικό, πού υπάρχουν καθυστερήσεις κλπ. Όλα τα παραπάνω μπορούν να επιτευχθούν μέσω αισθητήρων και στατιστικών τα οποία θα τροφοδοτούν συνεχώς ένα σύστημα με δεδομένα τα οποία θα αναλύονται σε πραγματικό χρόνο.

**Φήμη**

Ετήσιες έρευνες που γίνονται με θέμα την φήμη των εταιριών (Hahn-Griffiths & Friedman, 2018) αποκαλύπτουν ότι ένα μείζον θέμα είναι ο υπολογισμός και η βελτίωση της φήμης της εταιρίας. Πλέον έχει επικρατήσει η άποψη ότι κάτι τέτοιο μπορεί να επιτευχθεί επαρκώς μέσω του Streaming Analytics. Το feedback που λαμβάνεται από πελάτες μέσω social media είτε μέσω της ιστοσελίδας της ίδιας της εταιρίας αναλύεται συνεχώς, διαμορφώνοντας σιγά σιγά ένα λεπτομερές προφίλ της εταιρίας από την πολύ σημαντική οπτική γωνία του πελάτη. Με αυτόν τον τρόπο οι οργανισμοί μπορούν να δουν από τα μάτια των πελατών τους τα θετικά και τα αρνητικά χαρακτηριστικά τους, και να εξασφαλίσουν μελλοντικά την διατήρηση των πρώτων και την εξάλειψη των δεύτερων.

**Revenue Channels**[22]

Για μια επιχείρηση το να βρει το επόμενο revenue channel της είναι πολύ πιο εύκολο όταν έχει στην κατοχή της πληθώρα δεδομένων για να αναλύσει. Μπορεί επίσης να καθορίσει και πιο αποδοτικούς τρόπους marketing φτάνοντας έτσι στους καταναλωτές της με μεγαλύτερη ευκολία. Μέσω του Streaming Analytics, οι εταιρίες μπορούν να εξάγουν σημαντικά συμπεράσματα από παράπονα πελατών, αξιολογήσεις προϊόντων/υπηρεσιών και συναλλαγών, και να κινηθούν πιο αποτελεσματικά στο μέλλον.

# Ανθρωποκεντρική ανάλυση - Κριτική

Πώς επηρεάζεται όμως η κοινωνία από όλο αυτό; Είναι λίγο τρομακτικό το γεγονός πως οποιαδήποτε ενέργεια κάνουμε πλέον με το κινητό μας ή με τον υπολογιστή μας στο διαδίκτυο παρακολουθείται. Το Streaming Analytics έχει τεράστιο αντίκτυπο στην κοινωνία μας, τον τρόπο με τον οποίο διαμορφώνονται και δρουν οι εταιρίες-κολοσσοί, οι κυβερνήσεις και κατ' επέκταση η παγκόσμια οικονομία (Rijmenam, 2016).

Πολλά πειράματα που έχουν κάνει χρήστες στο διαδίκτυο έχουν δείξει ότι τα δεδομένα μας συλλέγονται και αναλύονται σε αστραπιαίο χρόνο. Για παράδειγμα εάν κάποιος χρήστης αναζητήσει οπουδήποτε (μηχανές αναζήτησης, price aggregators[23], social media) πληροφορίες για αγορά παπουτσιών, τότε μέσα σε λίγα μόλις λεπτά η πλειοψηφία των διαφημίσεων που θα

[22] Οι τρόποι με τους οποίους μπορούμε να φέρουμε ένα προϊόν ή μια υπηρεσία στην αγορά και στους καταναλωτές.

[23] Οργανισμοί σαν το skroutz που παρέχουν σύγκριση τιμών για προϊόντα από διαφορετικούς παρόχους.

βλέπει αυτός ο χρήστης θα είναι σχετικές με παπούτσια. Ούτε το ίδιο το άτομο δε γνωρίζει ποια και πόσα δεδομένα έχουν συγκεντρωθεί για τον εαυτό του, ή πού έχουν διαμοιραστεί ευαίσθητες πληροφορίες εν αγνοία του.

Συμπερασματικά, σύμφωνα με τα παραπάνω, δυστυχώς η απάντηση στην κύρια ερώτηση του θέματος της εργασίας (*Are humans being subordinated to the power of algorithms and social media?*) είναι ναι. Έχει χαθεί πλέον το μέτρο και οι περισσότεροι δεν το γνωρίζουν καν, καθώς καθημερινά συλλέγονται αδιανόητα ποσά δεδομένων για τον καθένα από εμάς, από την στιγμή που το κινητό μας θα αποκτήσει πρόσβαση στο διαδίκτυο. Από το πόση ώρα θα κοιτάξεις μια φωτογραφία στο facebook/instagram, μέχρι ένα τυχαίο προϊόν που θα αναζητήσεις στο google, τα πάντα αναλύονται και διαμορφώνουν ένα προφίλ για εσένα στο οποίο δε μπορείς να έχεις πρόσβαση.

# Σύνοψη

Συνοψίζοντας, σε αυτήν την εργασία αναλύθηκε εκτενώς η τεχνολογία του Streaming Analytics. Παρουσιάστηκαν κάποια από τα εργαλεία που χρησιμοποιούνται για την επεξεργασία και ανάλυση δεδομένων σε πραγματικό χρόνο, παρατηρώντας ότι η Apache έχει αναλάβει την ανάπτυξη και υποστήριξη πολλών εξ' αυτών. Στη συνέχεια, έγινε μια σύγκριση ανάμεσα σε Streaming Analytics και Batch Processing με έμφαση στα θετικά και αρνητικά χαρακτηριστικά της κάθε τεχνολογίας, για να δοθεί μια πιο καθαρή εικόνα σχετικά με το πότε πρέπει να χρησιμοποιείται η κάθε μία. Επιπροσθέτως, παρουσιάστηκαν οι τέσσερις κύριοι πυλώνες με τους οποίους επηρεάζονται οι μικρές και μεγάλες επιχειρήσεις εκμεταλλευόμενες τα δεδομένα που συλλέγουν. Τέλος, έλαβε μέρος η προσωπική μου κριτική σχετικά με την κατάσταση που επικρατεί στην κοινωνία, η οποία θα πρέπει να ανησυχήσει όλους μας έτσι ώστε να ενημερωθούμε καλύτερα σχετικά με τα προσωπικά δεδομένα που συλλέγονται για εμάς.

# Βιβλιογραφία


Apache, 2019. *Apache Spark.* [Online]
Available at: https://spark.apache.org/docs/latest/streaming-programming-guide.html
[Accessed October 2019].

Apache, 2019. *Flume.* [Online]
Available at: https://flume.apache.org/
[Accessed October 2019].

Apache, 2019. *Storm Apache.* [Online]
Available at: https://storm.apache.org/
[Accessed October 2019].

Armstrong, M., 2019. *Statista.* [Online]
Available at: https://www.statista.com/chart/17723/the-data-created-last-year-is-equal-to/
[Accessed October 2019].

Ashraf, U., 2018. *FreeCodeCamp.* [Online]
Available at: https://www.freecodecamp.org/news/apache-storm-is-awesome-this-is-why-you-should-be-using-it-d7c37519a427/
[Accessed October 2019].

Barone, A., 2019. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/b/batch-processing.asp
[Accessed October 2019].

Dezyre, 2016. *Dezyre.* [Online]
Available at: https://www.dezyre.com/article/top-5-apache-spark-use-cases/271
[Accessed October 2019].

Flink, A., 2019. *Flink Apache.* [Online]
Available at: https://flink.apache.org/flink-architecture.html
[Accessed October 2019].

Freeman, H., 2016. *Dataversity.* [Online]
Available at: https://www.dataversity.net/streaming-analytics-101/
[Accessed October 2019].

Hahn-Griffiths, S. & Friedman, D., 2018. *reputation institute.* [Online]
Available at: https://insights.reputationinstitute.com/blog-ri/annual-reputation-leaders-study-what-you-need-to-know
[Accessed October 2019].

Hans, J., 2017. *RTInsights.* [Online]
Available at: https://www.rtinsights.com/7-ways-your-business-can-benefit-from-streaming-analytics/
[Accessed October 2019].

IDC, 2018. *Seagate.* [Online]
Available at: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf
[Accessed October 2019].

Kafka, A., 2019. *Apache Kafka.* [Online]
Available at: https://kafka.apache.org/documentation/
[Accessed October 2019].

Kh, R., 2016. *datafloq.* [Online]
Available at: https://datafloq.com/read/how-big-data-is-affecting-business-decisions/2191
[Accessed October 2019].

Rehman, J., 2019. *itrelease.* [Online]
Available at: http://www.itrelease.com/2012/12/what-are-advantages-and-disadvantages-of-batch-processing-systems/
[Accessed October 2019].

Rijmenam, M. v., 2016. *datafloq.* [Online]
Available at: https://datafloq.com/read/big-data-will-have-a-big-impact-on-society/212
[Accessed October 2019].

Rouse, M., 2015. *Techtarget.* [Online]
Available at: https://whatis.techtarget.com/definition/distributed-computing
[Accessed October 2019].

Rouse, M., 2019. *Techtarget.* [Online]
Available at: https://searchdatacenter.techtarget.com/definition/high-availability
[Accessed October 2019].

Shiff, L., 2018. *bmc.com.* [Online]
Available at: https://www.bmc.com/blogs/batch-processing-stream-processing-real-time/
[Accessed October 2019].

Shoro, A. G. & Soomro, T. R., 2015. Global Journal of Computer Science and Technology. In: *Global Journal of Computer Science and Technology.* s.l.:s.n.

Wikipedia, 2019. *Wikipedia.* [Online]
Available at: https://en.wikipedia.org/wiki/Internet_of_things
[Accessed October 2019].

Wikipedia, 2019. *Wikipedia.* [Online]
Available at:
https://el.wikipedia.org/wiki/%CE%9A%CE%B5%CE%BD%CF%84%CF%81%CE%B9%CE%BA%CF%8C%CF%82_%CF%85%CF%80%CE%BF%CE%BB%CE%BF%CE%B3%CE%B9%CF%83%CF%84%CE%AE%CF%82
[Accessed October 2019].

Wikipedia, 2019. *Wikipedia.* [Online]
Available at:
https://el.wikipedia.org/wiki/%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%AE_%CE%B1%CE%BD%CE%B1%CE%B6%CE%AE%CF%84%CE%B7%CF%83%CE%B7%CF%82
[Accessed October 2019].

Wikipedia, 2019. *Wikipedia, Apache Kafka.* [Online]
Available at: https://en.wikipedia.org/wiki/Apache_Kafka
[Accessed October 2019].