

מסמך אפיון – ניטור פרסומים והמלצות בשוקים דיגיטליים- רשות ניירות ערך



עורך: אבי ישר
תאריך: 16.9.25
גרסת: 1.1

חלק ראשון – איסוף וניתוח מידע מקבוצות טלגרם

מקורות מידע

- קבוצות טלגרם ציבריות ופרטיות העוסקות בתחום מנויות, מסחר והשקעות, המהוות פלטפורמה מרכזית להפצת מידע, המלצות והשפעות על שוק ההון.

מטרות

- איתור פרטומים והמלצות בנוגע לניריות ערך אשר טרם פורסמו במדיות רשמיות או העולמים להיות לא חוקיים.
- פיתוח יכולת אוטומטית לניטור והצלבת מידע לטובת רשות ניריות ערך, לשם זיהוי מוקדם של מידע פנים, מניפולציות שוק או פעילות חשודה אחרת.

אתגרים

- חישוב רציף ומbose מילוט מפתח בהיקפים רחבים ובדירוג גבואה.
- זיהוי קשרים וקישוריות בין קבוצות שונות, כולל "אלגגוריתם" בין ערוצים ומעקב אחרי דפוס הפצתם.
- איתור אוטומטי של קבוצות חדשות שאינן מוכרות מראש והוספתן למערכת.

אפקין פונקציונלי

1. אינדקס מילוט חיפוש

- יצירת רשימה מילוט מפתח ייעודית בתחום ההשקעות והמנויות (מנויות, stocks, איטותים, המלצות, בורסה, שוק ההון וכו').
- יצירת רשימה של מילוט מפתח של אנשים שיש להם דעה במניות (לדוגמה מכיה-סטוקס)
- שימוש במנווע NLP (כגון **Gemini**) ליצירת אינדקס חכם והרחבתו על בסיס הקשרים סמנטיים.
- שימוש האינדקס בקובץ CSV המועלה ל-**BigQuery** לצורך ניתוח עתידי.
- ריענון חדש של מילוט החיפוש לזרחי ביטויים חדשים ורלוונטיים.

2. חישוב קבוצות רלוונטיות

חישוב קבוצות המכילות את מילוט המפתח שהוגדרו באינדקס. ביצוע **אנליה מבוססת** **Gemini** על מוגם הודיעות מכל קבוצה כדי להעיר את רלוונטיותה בתחום ההשקעות. עדכון אוטומטי של טבלת קבוצות הכלולות:

- מילוט מפתח
- שם הקבוצה
- מזהה קבוצה
- רמת רלוונטיות
- תאריך עדכון אחרון

רק קבוצות שסוגו כרלוונטיות נכנסות לתהיליך העיבוד השוטף.

3. איסוף נתונים

- וורדת כל ההודעות מהשנה الأخيرة עבור **קבוצות רלוונטיות**.
שמירtan במחסן נתונים (BigQuery DWH) תחת סכמה ייודית.

4. עיבוד ML ויזוג פוטיטים

- חולקה לקטגוריות:
 - א. פרסום שיווקיות.
 - ב. חדשות/מיעד כללי (לדוגמה: "הברסה ננעלה בירידות").
 - ג. המלצות/איתות מסחר.
- עבור הודעות המסוגות כ-"**המלצות**":
 - ו. עבור הودעה:
 1. ביצוע Sentiment Analysis ליהיו מגמות חיוביות/שליליות.
 2. האם ההודעה נcona או לא
 3. רשימה של המניות (הסימן שלה) המוזכרות
 4. **קייטלוג לפי אזור גיאוגרפי**
 - a. ארהב
 - b. אירופה
 - c. אסיה
 - d. ישראל
 - e. שאר העולם
 5. **קייטלוג לפי סוג נייר ערך**
 - a. מניות בודדות
 - b. אגרות חוב (אג"ח)
 - c. קרנות סל (ETFs)
 - d. קרנות נאמנות
 - e. כתבי אופציה ונגזרים
 - f. פקדיונות
 - g. קרייטפו
 6. האם ההודעה מוגדרת כחשודה- כן/לא
 7. למה הוא הגדר אותה כחשודה? (משפט אחד)
 - ii. עבור קבוצה:
 1. חישוב **מדד אמינות הקבוצה** בהסתמך על אחוז ההודעות האמינות לעומת החשודות.
 2. **זיהוי קבוצות בעלות פעילות חריגה** (תדרות פרסום, סוג תוכן, מיקור המידיע).
 - iii. אפיון אמינות של מפרסם
 1. אפיון רמת אמינות וסיכון של כל מפרסם על סמך איות ההודעות, תדרות הפרסונים ודפוסי פעילות.

5. UI ראשוני

פיתוח **ממשק תצוגה אינטואיטיבי** המציג את תוצאות הניתנות:

- חיתוכים לפי קבוצה, מפרסם, תאሪיך וקטגוריה.
- סיכומי מגמות לפי סוג פרסומים וניירות ערך.
- הציגת רמות סיכון וחשד לפעולות חריגה.

מה הלקוח מקבל בסיום הפרויקט

1. מערכת אוטומטית לניטור טלגרם

- איסוף רציף של מידע מקבוצות טלגרם ציבוריות ופרטיות בנושא מנויות, מסחר והש侃עות.
- הרחבת דינמית של מקורות המידע באמצעות מנגנון אינדקס חכם המותעדכן מדי חדש.

2. מחסן נתונים מרכזי (BigQuery DWH)

- שבירת כלל הհודעות הרלוונטיות בשכבה מאורגנת וניתנת לניתוח.
- תיעוד מלא של פרסומים משנה אחרונה לטובות חוקירה ובקירה.

3. יכולות ניתוח מבוססות ML

- סיוג אוטומטי של פוסטים לקטגוריות: פרסומות, חדשות כלויות, והמלצות מסחר.
- ביצוע Sentiment Analysis ייעודי על המלצות מסחר, כולל דיהו מגמות חיוביות/שליליות.
- אפיון רמת אמינות של קבוצות ושל פרסומים בודדים.

4. דיהו קשריות ודפוסי פעילות

- איתור קשרים בין קבוצות שונות ("גלאול הודעות" והפצת מידע).
- ניתוח عمוק של פעילות פרסומים, כולל הערכת השפעה ומידת סיכון.

5. משק משתמש (UI)

- לוח בקרה המציג את תוצאות הניתוח בצורה ויזואלית וברורה.
- אפשרות לחיתוך מידע לפי קבוצה, מפרסם, תאሪיך וקטגוריה.
- כלי עבודה אנליטי לרשوت ניירות ערך המאפשר בקרה וניטור בזמן אמת.

מסמך ארכיטקטורה – מערכת איתור קבוצות מילוט-מפתח וצנרת ML על GCP

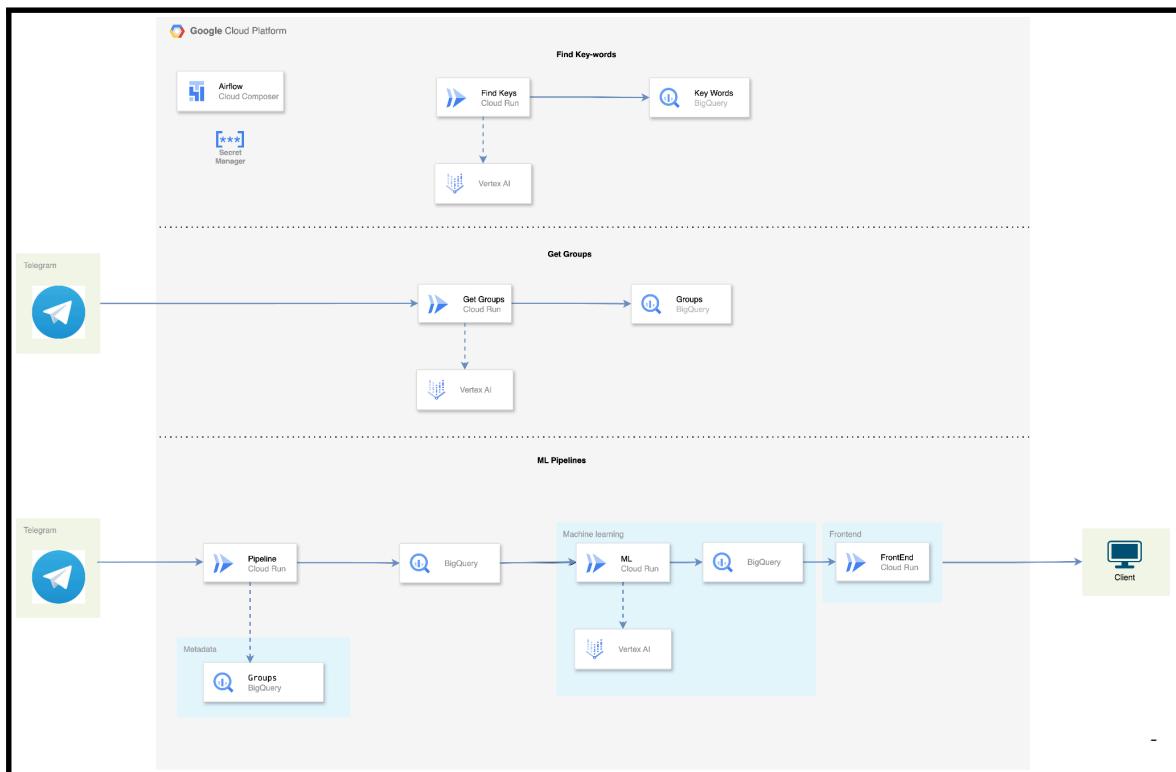
תקציר מנהלי

המערכת מאפשרת לקלוט בקשה מטלגרם, לאתר קבוצות רלוונטיות ומילוט-מפתח, לשמר מטא-דאטה ב-BigQuery, להריץ צנרת ML (כולל אינטראקציה עם Vertex AI), ולהציג תוצאות דרך FrontEnd דרך ללקוט קבוצה. הארכיטקטורה מבוססת-ענן (GCP), סקליבילית, מודולרית, ומותאמת להפרדה ברורה בין שכבות איסוף, עיבוד, ML ופרגנטציה.

מטרות עסקיות

- איסוף וריכוז נתונים מטלגרם לקבוצות רלוונטיות להשקיות ומסחר.
- יצירת אינדקס מילוט מפתח ורחיבתו אוטומטית.
- סיווג תכנים (פרסומות/חדשנות/המלצות), ניתוח Sentiment, והפקת תובנות.
- בניית משקל לכוח תפעולי להציג הממצאים, פילוחים ותרשיימי מגמה.

תרשים על (High Level)



- **(Orchestration: Cloud Composer (Airflow** •
- **Secrets & Keys: Secret Manager** •
- **Compute** •
סודות: Cloud Run (סודותים: Groups, Get Keys, Pipeline, ML, Find Keys, Compute FrontEnd (FrontEnd
- **ML Services: Vertex AI** •
מודלים/NLP/טקסט/קלאסיפיקציה (ML Services
- **Data Platform: BigQuery** •
(Classifications, Aggregates (טבלאות: KeyWords, Groups, Messages/Posts
- **מקור נתונים: Telegram (API/Bot/Client** •
- **צרכן קצה: ממשק FrontEnd (ללקוח/أنلوج)** •

תיאור שירותי ותפקידם

- **(Airflow (Cloud Composer** •
- **תפקיד: תזמון ותצמוך Job-ים מחזוריים/אד-הוק (DAGs).**
- **טריגרים עיקריים:**
 - חודשי: הרצת **Find Keys** (בנייה והרחבת אינדקס מילוט מפתח).
 - שבועי: הרצת **Get Groups** (איתור/עדכון קבוצות רלוונטיות).
 - יומי: הרצת **Pipeline** (איסוף הודעות חדשות) ו-**ML** (סיווג/ניתוח).
- **אחריות נוספת: ניהול תלויות, רטרי, SLA, התראות.**
- **Secret Manager** •
- **תפקיד: אחסון מאובטח של סודות (Token/Key/SA OAuth, BigQuery/Vertex, Telegram, Key/Token של מפתחות API חיצוניים).**
- **דף שימוש: הסרוויסים ב-Run Cloud מושכים סודות בזמן ריצה, עם הרשות IAM עקרון המינימום.**
- **(Find Keys (Cloud Run** •
- **קלט:**
 - אינדקס קיימ (BigQuery: **analytics.keywords** – אופציונלי)
 - פרומפטים/Words (Seed Terms/Airflow/Kobz Konfag) מוגדר ב-Wflow/Airflow/Cloud Run
- **תהליכי:**
 - פונה ל-AI Vertex ליצירת/הרחבת רשימות מונחים וסינוניהם (כולל מילים גזירות, מילים נרדפות, וرمזים מרובי-שפה).
 - מנרמל (Normalization) ו-**Deduplication**

- **פלט:**

`term, lang, category` (שדות לדוגמה: BigQuery KeyWords Table)
o עדכון טבלת KeyWords ב-BigQuery (שדות לדוגמה: `source, score, updated_at`)

- **מדיניות ריצה:** חדש/אד-הוק.

(Get Groups (Cloud Run

- **קלט:**

o טבלת KeyWords (חיפוש קבוצות לפי מונחים)
o API Telegram (חיפוש קבוצות לפי מונחים)

- **תהליכי:**

o חיפוש קבוצות תואמות מילות מפתח.
o דגימת הودעות מהקבוצה (Rate-limit Aware).
o פניה ל-AI Vertex להערכת רלוונטיות (Classification: רלוונטי/לא).

- **פלט:**

o טבלת Groups ב-BigQuery (שדות לדוגמה: `group_id, group_name, member_count, lang, last_sample_ts, relevant:BOOL, relevance_score, topics`)

- **מדיניות ריצה:** יומי/שבתי; תומך Incremental Update.

(Pipeline (Cloud Run

- **קלט:**

o רשימת קבוצות רלוונטיות (`Groups.relevant = TRUE`)
o Telegram API

- **תהליכי:**

o משיכת הודעות משנה אחרונה/חלוןZZ ל-`ts` (Parsing Tickers/Links/Hashtags).
o ניקוי, העשרה (Parsing Tickers/Links/Hashtags), נרמול שפה וקידוד.

- **פלט:**

o טבלת Messages/Posts (למשל `:analytics.messages`)
■ `message_id, group_id, publisher_id, text, ts, lang, attachments, extracted_tickers[], region, asset_classes[], ingested_at`

(ML (Cloud Run

- **קלט:**

o טבלת Messages (חדש)
o מודלים מאומנים/Prompt-Templates ב-AI

- **תהליכי:**

- **Content Classification:** פרסום / חדשות / המלצה.
- **Sentiment Analysis:** בעיקר לקטגוריות המלצות (חיובי/שליל/нейтрלי + confidence + rationale).
- **Suspiciousness Scoring:** דיהוי דפוסים חשודים והסביר קצר (rationale).
- **Enrichment:** דיהוי שוק יעד (US/EU/ASIA/IL/ROW), סוג נייר ערך (منיה/ETF/אג"ח/נגזרים/קריפטו וכו').
- **Group/Publisher Scoring:** חישוב מודיע אמינות/סיכון בرمת קבוצה ומספרם.

- **פלט:**

- `: (analytics.classifications (لמשל Classifications`
- `message_id, category, sentiment, suspicious_flag, ■`
- `suspicious_reason, tickers[], market, asset_class,`
- `.model_version, scored_at`
- **Group_Aggregates / Publisher_Aggregates** (מדדים מוצברים לעדכנים מהירים ל-UI).
- **מדדיות רצה:** לפי טריגר m-Pub/Sub או Airflow (אופציונלי), מיד לאחר הטענה.

- **(FrontEnd) Cloud Run**

- **תפקיד:** חישפת UI ללקוח/אנליסט.
- **תצוגות עיקריות:**
 - **דאשborad קבוצה:** חזק קבוצה, מגמות, סוג פרסום, sentiment, תאריכי פעילות.
 - **דאשborad פרסום:** פרופיל סיכון, קצב פעילות, חשיפה, על כל ההיסטוריה.
 - **Explorer:** חיתוכים לפי קבוצה/ פרסום/תאריך/קטגוריה/נייר ערך; חיפוש טקסטואלי.
- **מקורות נתונים:** קריאות SQL/Views/Materialized Views ב-BigQuery (דרך API שרת/שכבה (BFF)).
- **אבטחה וגישה:** OAuth/IAP לфи פרופיל משתמש; RBAC (אנליסט/מנהל/צופה).

זרימות נתונים (Data Flow)

1. **Find Keys** → כותב ל-**KeyWords**.
2. **Get Groups** קורא **KeyWords**, מושך מטלאגרם, מעריך ב-**AI**, כותב ל-**Groups**.
3. **Pipeline** קורא **Groups** (רק רלוונטיים), מושך הודעות מטלאגרם, כותב ל-**Messages**.
4. **ML** קורא **Messages**, עושה קלאסיפיקציה/סנטימנט/סיכון ב-**AI**, כותב ל-**Classifications**.
5. **FrontEnd** קורא **FrontEnd** Views/Materialized Views ומציג ללקוח.

סכמות נתונים מוצעות (BigQuery)

שמות סכמות/שדות דוגמתיים – יתוקנו מול Data Modeling.

```

analytics.keywords(term STRING, lang STRING, category STRING, •
                  (source STRING, score FLOAT64, updated_at TIMESTAMP

                  analytics.groups(group_id STRING, group_name STRING, •
                                   member_count INT64, lang STRING, relevance_score FLOAT64,
                                   relevant BOOL, topics ARRAY<STRING>, last_sample_ts TIMESTAMP,
                                   (updated_at TIMESTAMP

                  analytics.messages(message_id STRING, group_id STRING, •
                                   publisher_id STRING, text STRING, ts TIMESTAMP, lang STRING,
                                   attachments ARRAY<STRING>, extracted_tickers ARRAY<STRING>,
                                   region STRING, asset_classes ARRAY<STRING>, ingested_at
                                   (TIMESTAMP

analytics.classifications(message_id STRING, category STRING, •
                           sentiment STRING, sentiment_score FLOAT64, suspicious_flag
                           BOOL, suspicious_reason STRING, tickers ARRAY<STRING>, market
                           STRING, asset_class STRING, model_version STRING, scored_at
                           (TIMESTAMP

analytics.group_aggregates(group_id STRING, window_start •
                           TIMESTAMP, window_end TIMESTAMP, posts INT64, ads_ratio
                           FLOAT64, news_ratio FLOAT64, reco_ratio FLOAT64,
                           suspicious_ratio FLOAT64, sentiment_avg FLOAT64, rank FLOAT64,
                           (updated_at TIMESTAMP

analytics.publisher_aggregates(publisher_id STRING, posts •
                           INT64, followers_est INT64, suspicious_ratio FLOAT64,
                           sentiment_avg FLOAT64, influence_score FLOAT64, updated_at
                           (TIMESTAMP

```

אבטחת מידע ו-IAM

- **Secret Manager**: כל המפתחות וה-Tokens. גישה רק ל-Tokens.
- **IAM (Principle of Least Privilege** :

 - לכל שירות ב-Cloud Run SA משלו עם הרשות מוגבלת ל-
 - הדרושים בלבד.

- **תעבורה**: HTTPS בלבד; שיקלט VPC-SC/Private Service Connect עבור BQ/Vertex/Secrets.
- **סיווג המידע**.
- **Audit & Logging**: Cloud Audit Logs

תפועל, ניטור ו-SLA

SLA מוצע (פנימ-ארגוני):

- אימות הרצות DAGs לפי לוח זמנים.
- זמינות ≥ 99.5% (חדש).
- עיכוב נתונים (Ingestion → Classification) עד > 30 דק' בחלון רגיל.

סקילינג, עליות ו-밋וב

(CPU/Memory) Concurrency Auto-Scaling • הגבלת CPU/Memory לשירות משאבי Cloud Run: בקצבות; בהתאם ל-Throughput.

Vertex AI • Prediction: שימוש במודלים מנהליים; שיקילת Batch Prediction להפחיתת עליות; Prompt-Templates.

BigQuery • Partitioning/Clustering: שימוש ב-`ts/group_id`; Views לשירות Partitioning/Clustering; Materialized Views לדאשborad.

תאימות, פרטיות ורגולציה

- אונימיזציה/פואודונימיזציה של מידע משתמשים ככל שנדרש.
- PII: זיהוי/סיכון/השחרה של שמות אישיים לפי מדיניות רגולטורית.
- Data Governance: קטלוג סכמות, תיוג רגישויות, ורשומות נתיבי עיבוד.

ממשק API פנימיים (סקיצה)

- GET /groups?relevant=true&updated_since – לאורקסטראציה/דוחות.
- POST /ingest/messages – נקודת צ'אנקיניג.(Batch Ingest (Pipeline))
- POST /ml/classify – הרצת קלואטיפיקציה/Scoring לאד-הוק.
- GET /dashboard/summary?granularity=daily&group_id – שכבת BFF – שכבת FrontEnd.

תרשים רצף (מילולי)

- KeyWords → Find Keys → Vertex AI → Airflow .1
- Groups → Get Groups → Telegram + Vertex AI → Airflow .2
- Messages → Pipeline → Telegram → Airflow .3
- Classifications/Aggregates → ML → Vertex AI → Airflow .4
- FrontEnd → Views (BigQuery) ומציג ללקוט. .5

הנחות ותלויות

- גישה חוקית ל-API Telegram בהתאם לתנאי שימוש.
- זמינות AI Vertex באזוריים הנתמכים ובעלויות צפוייה בהתאם לנופחים.
- סכמה ב-*BigQuery* עשויה להתעדכן בשלבי ה-POC על סמך ממצאים.