

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



עורך: אבי ישר
תאריך: 16.9.25
גרסא: 1.1

חלק ראשון – איסוף וניתוח מידע מקבוצות טלגרם

מקורות מידע

- קבוצות טלגרם ציבוריות ופרטיות העוסקות בתחומי מניות, מסחר והשקעות, המהוות פלטפורמה מרכזית להפצת מידע, המלצות והשפעות על שוק ההון.

מטרות

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

אתגרים

- חיפוש רציף ומבוסס מילות מפתח בהיקפים רחבים ובתדירות גבוהה.
- זיהוי קשרים וקישוריות בין קבוצות שונות, כולל "גלגול הודעות" בין ערוצים ומעקב אחרי דפוסי הפצה.
- איתור אוטומטי של קבוצות חדשות שאינן מוכרות מראש והוספתן למערכת.

אפיון פונקציונלי

1. אינדקס מילות חיפוש

- יצירת רשימת מילות מפתח ייעודית בתחום ההשקעות והמניות (מניות, stocks, איתותים, המלצות, בורסה, שוק ההון וכו').
- יצירת רשימה של מילות מפתח של אנשים שיש להם דעה במניות (לדוגמה מיכה-סטוקס)
- שימוש במנוע NLP (כגון Gemini) ליצירת אינדקס חכם והרחבתו על בסיס הקשרים סמנטיים.
- שמירת האינדקס בקובץ CSV המועלה ל-BigQuery לצורך ניתוח עתידי.
- ריענון חודשי של מילות החיפוש לזיהוי ביטויים חדשים ורלוונטיים.

2. חיפוש קבוצות רלוונטיות

חיפוש קבוצות המכילות את מילות המפתח שהוגדרו באינדקס. ביצוע **אנליזה מבוססת Gemini** על מדגם הודעות מכל קבוצה כדי להעריך את רלוונטיותה לתחום ההשקעות. עדכון אוטומטי של טבלת קבוצות הכוללת:

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

רק קבוצות שסווגו כרלוונטיות נכנסות לתהליך העיבוד השוטף.

3. איסוף נתונים

- הורדת כלל ההודעות מהשנה האחרונה עבור **קבוצות רלוונטיות**. שמירתן במחסן נתונים (BigQuery DWH) תחת סכמה ייעודית.

4. עיבוד ML וסיווג פוסטים

- חלוקה לקטגוריות:
 - א. פרסומות שיווקיות.
 - ב. חדשות/מידע כללי (לדוגמה: "הבורסה ננעלה בירידות").
 - ג. המלצות/איתותי מסחר.
- עבור הודעות המסווגות כ-"המלצות":
 - i. עבור הודעה:
 1. ביצוע Sentiment Analysis לזיהוי מגמות חיוביות/שליליות.
 2. האם ההודעה נכונה או לא
 3. רשימה של המניות (הסימון שלה) המוזכרות
 4. קיטלוג לפי **אזור גיאוגרפי**
 - a. ארהב
 - b. אירופה
 - c. אסיה
 - d. ישראל
 - e. שאר העולם
 5. קיטלוג לפי **סוג נייר ערך**
 - a. מניות בודדות
 - b. אגרות חוב (אג"ח)
 - c. קרנות סל (ETFs)
 - d. קרנות נאמנות
 - e. כתבי אופציה ונגזרים
 - f. פקדונות
 - g. קריפטו
 6. האם ההודעה מוגדרת כחשודה- כן/לא
 7. למה הוא הגדיר אותה כחשודה? (משפט אחד)
 - ii. עבור קבוצה:
 1. חישוב **מדד אמינות הקבוצה** בהתבסס על אחוז ההודעות האמינות לעומת החשודות.
 2. זיהוי **קבוצות בעלות פעילות חריגה** (תדירות פרסומים, סוגי תוכן, מקור המידע).
 - iii. אפיון אמינות של מפרסם
 1. אפיון רמת אמינות וסיכון של כל מפרסם על סמך איכות ההודעות, תדירות הפרסומים ודפוסי פעילות.

5. UI ראשוני

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

- חיתוכים לפי קבוצה, מפרסם, תאריך וקטגוריה.
- סיכומי מגמות לפי סוגי פרסומים וניירות ערך.
- הצגת רמות סיכון וחשד לפעילות חריגה.

מה הלקוח מקבל בסיום הפרויקט

1. מערכת אוטומטית לניטור טלגרם

- איסוף רציף של מידע מקבוצות טלגרם ציבוריות ופרטיות בנושאי מניות, מסחר והשקעות.
- הרחבה דינמית של מקורות המידע באמצעות מנגנון אינדקס חכם המתעדכן מדי חודש.

2. מחסן נתונים מרכזי (BigQuery DWH)

- שמירת כלל ההודעות הרלוונטיות בשכבה מאורגנת וניתנת לניתוח.
- תיעוד מלא של פרסומים משנה אחרונה לטובת חקירה ובקרה.

3. יכולות ניתוח מבוססות ML

- סיווג אוטומטי של פוסטים לקטגוריות: פרסומות, חדשות כלליות, והמלצות מסחר.
- ביצוע Sentiment Analysis ייעודי על המלצות מסחר, כולל זיהוי מגמות חיוביות/שליליות.
- אפיון רמת אמינות של קבוצות ושל מפרסמים בודדים.

4. זיהוי קשריות ודפוסי פעילות

- איתור קשרים בין קבוצות שונות ("גלגול הודעות" והפצת מידע).
- ניתוח עומק של פעילות מפרסמים, כולל הערכת השפעה ומידת סיכון.

5. ממשק משתמש (UI)

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

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

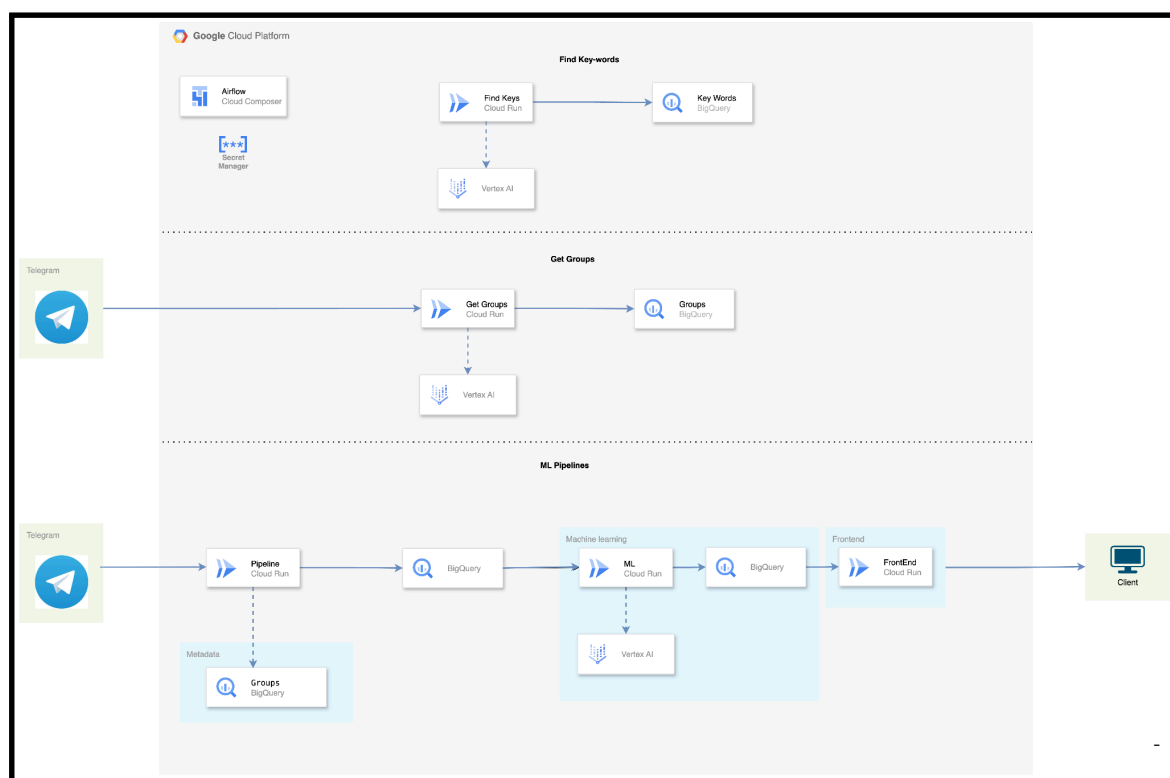
תקציר מנהלים

המערכת מאפשרת לקלוט בקשות מטלגרם, לאתר קבוצות רלוונטיות ומילות-מפתח, לשמור מטא-דאטה ב-BigQuery, להריץ צנרת ML (כולל אינטראקציה עם Vertex AI), ולהגיש תוצרים דרך FrontEnd ללקוח קצה. הארכיטקטורה מבוססת-ענן (GCP), סקייבלילית, מודולרית, ומותאמת להפרדה ברורה בין שכבות איסוף, עיבוד, ML ופרזנטציה.

מטרות עסקיות

- איסוף וריכוז נתונים מטלגרם לקבוצות רלוונטיות להשקעות ומסחר.
- יצירת אינדקס מילות מפתח והרחבתו אוטומטית.
- סיווג תכנים (פרסומות/חדשות/המלצות), ניתוח Sentiment, והפקת תובנות.
- בניית ממשק לקוח תפעולי להצגת הממצאים, פילוחים ותרשימי מגמה.

תרשים על (High Level)



- **(Orchestration: Cloud Composer (Airflow**
- **Secrets & Keys: Secret Manager**
- **Compute סרר-מצב:** Cloud Run (סרוויסים: ML, Pipeline, Get Groups, Find Keys, FrontEnd)
- **ML Services: Vertex AI** (מודלי NLP/טקסט/קלאסיפיקציה)
- **Data Platform: BigQuery** (טבלאות: Keywords, Groups, Messages/Posts, Classifications, Aggregates)
- **מקור נתונים:** Telegram (API/Bot/Client)
- **צרכן קצה:** FrontEnd ממשיך (לקוח/אנליסט)

תיאור סרוויסים ותפקידם

- **(Airflow (Cloud Composer**
 - **תפקיד:** תזמון ותזמור Job-ים מחזוריים/אד-הוק (DAGs).
 - **טריגרים עיקריים:**
 - חודשי: הרצת **Find Keys** (בניית והרחבת אינדקס מילות מפתח).
 - שבועי: הרצת **Get Groups** (איתור/עדכון קבוצות רלוונטיות).
 - יומי: הרצת **Pipeline** (איסוף הודעות חדשות) ו-**ML** (סיווג/ניתוח).
 - **אחריות נוספת:** ניהול תלויות, רטריי, SLA, התראות.
- **Secret Manager**
 - **תפקיד:** אחסון מאובטח של סודות (Token/Key של OAuth/SA, Telegram ל-Vertex/BigQuery, מפתחות API חיצוניים).
 - **דפוס שימוש:** הסרוויסים ב-Cloud Run מושכים סודות בזמן ריצה, עם הרשאות IAM עקרון המינימום.
- **(Find Keys (Cloud Run**
 - **קלט:**
 - אינדקס קיים (BigQuery: **analytics.keywords**) – אופציונלי
 - פרומפטים/Seed Terms (מוגדר ב-Airflow/קובץ קונפיג)
 - **תהליך:**
 - פונה ל-Vertex AI ליצירת/הרחבת רשימות מונחים וסינוניהם (כולל מילים גזורות, מילים נרדפות, ורמזים מרובי-שפה).
 - מנרמל (Normalization) ו-Deduplication.

- פלט:

- עדכון טבלת `Keywords` ב-BigQuery (שדות לדוגמה: `term`, `lang`, `category`, `source`, `score`, `updated_at`).

- מדיניות ריצה: חודשי/אד-הוק.

- (Get Groups (Cloud Run

- קלט:

- טבלת `Keywords`
- Telegram API (חיפוש קבוצות לפי מונחים)

- תהליך:

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

- פלט:

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

- מדיניות ריצה: יומי/שעתי; תומך Incremental Update.

- (Pipeline (Cloud Run

- קלט:

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

- תהליך:

- משיכת הודעות משנה אחרונה/חלון זז לפי `last_ingested_ts`.
- ניקוי, העשרה (Parsing Tickers/Links/Hashtags), נרמול שפה וקידוד.

- פלט:

- 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:

- `Messages` חדשים
- מודלים מאומנים/Prompt-Templates ב-Vertex AI

- תהליך:

- **Content Classification**: פרסומת / חדשות / המלצה.
- **Sentiment Analysis**: בעיקר לקטגוריית המלצות (חיובי/שלילי/נייטרלי + confidence).
- **Suspiciousness Scoring**: זיהוי דפוסים חשודים והסבר קצר (rationale).
- **Enrichment**: זיהוי שוק יעד (US/EU/ASIA/IL/ROW), סוג נייר ערך (מניה/ETF/אג"ח/נגזרים/קריפטו/וכו').
- **Group/Publisher Scoring**: חישוב מדדי אמיונות/סיכון ברמת קבוצה ומפרסם.

- פלט:

- **Classifications** (למשל `analytics.classifications`):
 - `message_id, category, sentiment, suspicious_flag, suspicious_reason, tickers[], market, asset_class, model_version, scored_at`
- **Group_Aggregates / Publisher_Aggregates** (מדדים מצטברים לעדכונים מהירים ל-UI).
- **מדיניות רצה**: לפי טריגר מ-Airflow או Pub/Sub (אופציונלי), מיד לאחר הטענה.

- **(FrontEnd (Cloud Run**

- **תפקיד**: חשיפת UI ללקוח/אנליסט.
- **תצוגות עיקריות**:
 - **דאשבורד קבוצות**: חוזק קבוצה, מגמות, סוגי פרסומים, sentiment, תאריכי פעילות.
 - **דאשבורד מפרסמים**: פרופיל סיכון, קצב פעילות, חשיפה, sentiment על כל ההיסטוריה.
 - **Explorer**: חיתוכים לפי קבוצה/מפרסם/תאריך/קטגוריה/נייר ערך; חיפוש טקסטואלי.
- **מקורות נתונים**: קריאות SQL/Views/Materialized Views ב-BigQuery (דרך API שרת/שכבת BFF).
- **אבטחה וגישות**: OAuth/IAP לפי פרופיל משתמש; RBAC (אנליסט/מנהל/צופה).

זרימות נתונים (Data Flow)

1. **Find Keys** → כותב ל-**KeyWords**.
2. **Get Groups** קורא **KeyWords**, מושך מטלגרם, מעריך ב-Vertex AI, כותב ל-**Groups**.
3. **Pipeline** קורא **Groups** (רק רלוונטיים), מושך הודעות מטלגרם, כותב ל-**Messages**.
4. **ML** קורא **Messages**, עושה קלאסיפיקציה/סנטימנט/סיכון ב-Vertex AI, כותב ל-**Classifications** ו-**Aggregates**.
5. **FrontEnd** קורא **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-Service Accounts ייעודיים.
- **IAM (Principle of Least Privilege)**:
 - לכל סרוויס ב-Cloud Run SA משלו עם הרשאות מצומצמות ל-BQ/Vertex/Secrets
 - הדרושים בלבד.
- **תעבורה**: HTTPS בלבד; שקילת VPC-SC/Private Service Connect עבור BQ/Vertex (לפי סיווג המידע).
- **Audit & Logging**: Cloud Audit Logs לכל גישה למשאבים; שמירת לוגים לפי מדיניות ארגונית.

תפעול, ניטור ו-SLA

SLA מוצע (פנים-ארגוני):

- אימות הרצות DAGs לפי לוח זמנים.
- זמינות FrontEnd $\geq 99.5\%$ (חודשי).
- עיכוב נתונים (Ingestion \rightarrow Classification) יעד > 30 דק' בחלון רגיל.

סקיילינג, עלויות ו-מיטוב

- **Cloud Run: Auto-Scaling** לפי בקשות; הגבלת Concurrency ושיוך משאבים (CPU/Memory) בהתאם ל-Throughput.
- **Vertex AI**: שימוש במודלים מנוהלים; שקילת Batch Prediction להפחתת עלויות; Caching ל-Prompt-Templates.
- **BigQuery**: שימוש ב-Partitioning/Clustering לפי Views `ts/group_id`; מצטברות; Materialized Views לדאשבורדים.

תאימות, פרטיות ורגולציה

- אנונימיזציה/פסאודונימיזציה של מזהי משתמשים ככל שנדרש.
- **Cookie/PII**: זיהוי/סינון/השחרה של שדות אישיים לפי מדיניות רגולטורית.
- **Data Governance**: קטלוג סכמות, תיוג רגישויות, ורשומות נתיבי עיבוד (Data Lineage).

ממשקי API פנימיים (סקיצה)

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

תרשים רצף (מילולי)

1. Airflow מזניק Vertex AI `Find Keys` \rightarrow כתיבה ל-`Keywords`.
2. Airflow מזניק Telegram + Vertex AI `Get Groups` \rightarrow כתיבה ל-`Groups`.
3. Airflow/טריגר זמן מזניק Telegram `Pipeline` \rightarrow כתיבה ל-`Messages`.
4. Airflow מזניק Vertex AI `ML` \rightarrow כתיבה ל-`Classifications/Aggregates`.
5. FrontEnd שואל Views (BigQuery) ומציג ללקוח.

הנחות ותלויות

- גישה חוקית ל-Telegram API בהתאם לתנאי שימוש.
- זמינות Vertex AI באזורים הנתמכים ובעלות צפויה בהתאם לנפחים.
- סכמה ב-BigQuery עשויה להתעדכן בשלבי ה-POC על סמך ממצאים.