



בינה עסקית

פרויקט

שובל כהן, יהונתן כהן, אורן הייקאני

I. שאלה עסקית + KPIs

מטרת פרויקט BI זה היא לנתח את מערך הנתונים של IMDB ולקבל תובנות עסקיות חשובות הקשורות לתעשיית הקולנוע. על ידי בחינת גורמים שונים כגון תקציבי סרטים, הכנסות גולמיות, דירוגים ומאפיינים אחרים, אנו שואפים לזהות דפוסים ומגמות שיכולים להשפיע על תהליכי קבלת החלטות עבור חברות הפקת סרטים, מפיצים ובעלי עניין אחרים בתעשייה. פרויקט זה יעזור לנו להבין את גורמי המפתח התורמים להצלחת סרטים, להעריך את הביצועים של ז'אנרים וסוגים יצירתיים שונים, ולספק המלצות לאופטימיזציה של אסטרטגיות הפקת סרטים.

שני מדדי KPI פוטנציאליים

1. החזר על השקעה (ROI): מדידת הרווחיות של סרטים על ידי השוואת ההכנסות שנוצרו (כגון ברוטו ברחבי העולם או ברוטו בארה"ב) מול תקציב ההפקה. זה יעזור להעריך את ההצלחה הפיננסית של סרטים ולזהות אילו סרטים מספקים את ההחזר הגבוה ביותר על ההשקעה.

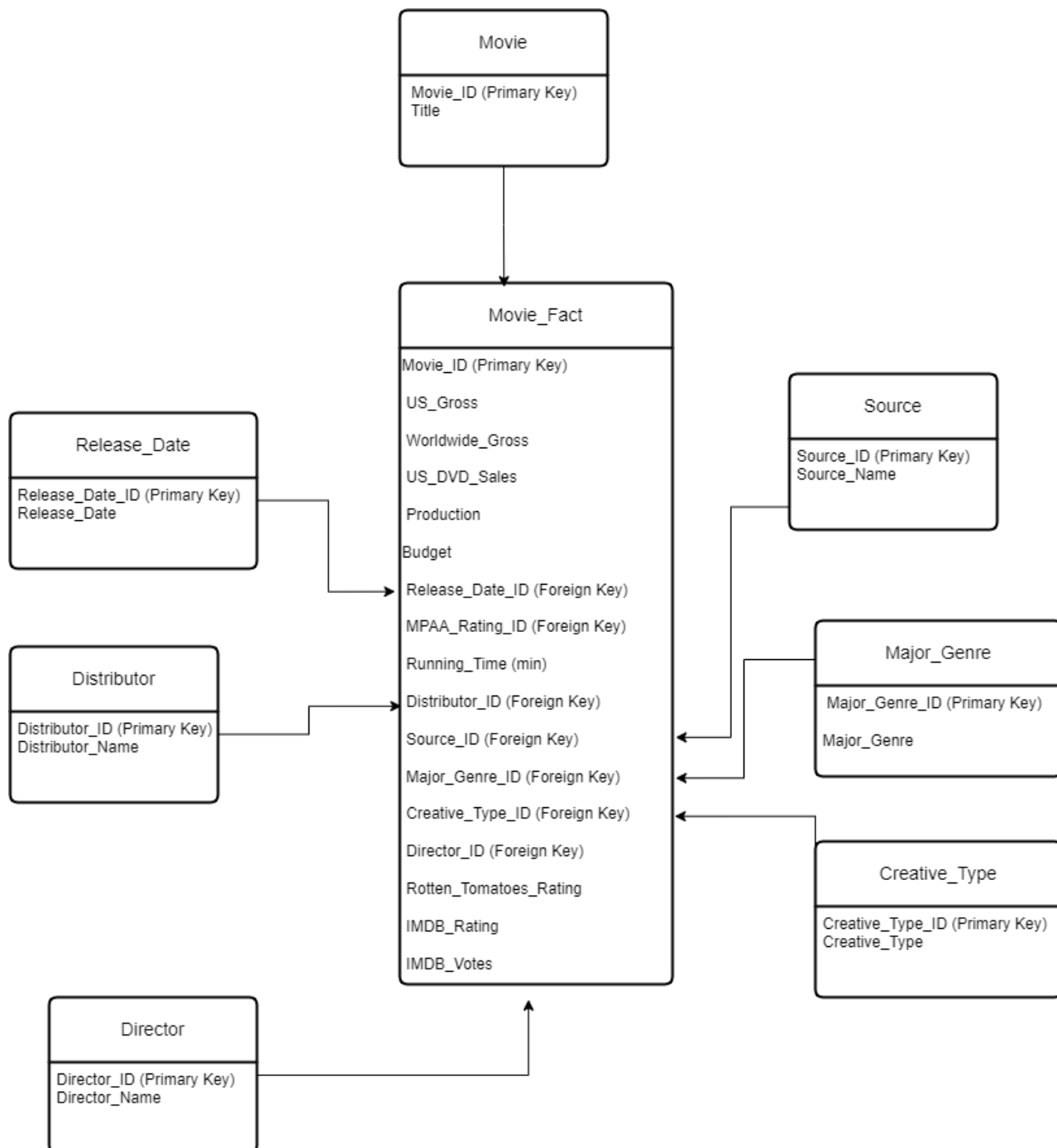
2. דירוג IMDB ממוצע: ייצוג דירוג ה-IMDB הממוצע שקיבלו סרטים. זה מציין את הקליטה הקריטית הכוללת ואת האיכות של הסרטים במערך הנתונים. על ידי ניטור KPI זה, אנו יכולים לזהות את הסרטים שזוכים בעקביות לביקורות חיוביות ולהבין את המתאם בין דירוגים וביצועים פיננסיים.



II. הגדרת Data Warehouse

1. עבור הנתונים שנבחרו, סכימת הכוכבים תהיה בחירה מתאימה. סכימה זו מורכבת מטבלת עובדות מרכזית (למשל, מידע על סרט) מוקפת בטבלאות מימדים (למשל, תכונות כמו ז'אנר, במאי, דירוגים). סכימה זו נבחרה מכיוון שהיא מפשטת שאלות וניתוח על ידי מזעור מספר הצירופים הנדרשים ומאפשרת צבירה יעילה של נתונים סביב טבלת העובדות המרכזית.

2. תיאור ה-Datawarehouse לאחר שילוב הסכימה:



3. Use Case 1: ניתוח פשוט

סכימת הכוכבים מאפשרת ניתוח פשוט על ידי הפחתת מורכבות השאילתות. מכיוון שטבלת העובדות מכילה את המידע המרכזי של הסרט, משתמשים יכולים לאחזר נתונים ולבצע ניתוח עם פחות פעולות הצטרפות. לדוגמה, אם משתמש רוצה לנתח את הקשר בין דירוגי סרטים (Rotten Tomatoes) לבין הכנסות ברוטו ברחבי העולם, הם יכולים לבצע שאילתות ישירות בטבלת העובדות ולצבור את הנתונים ללא צורך במספר הצטרפות, מה שייביל לאחזור מהיר יותר וקבלת החלטות קלה יותר.

2. Use Case 2: צבירה יעילה

סכימת Star מאפשרת צבירה יעילה של נתונים סביב טבלת העובדות המרכזית. משתמשים יכולים לחשב במהירות מדדי מפתח כגון הכנסה גולמית ממוצעת ברחבי העולם לפי ז'אנר, במאי או סוג קריאייטיב על ידי קיבוץ של הממדים הרלוונטיים וסיכום המדדים המתאימים מטבלת ה-Facts. יכולת צבירה זו מאפשרת למשתמשים לזהות מגמות, להשוות ביצועים בין תכונות שונות ולקבל החלטות מושכלות ללא צורך ב-Joins מורכבים וגוזלי זמן.

III. תהליך ה-ETL

תהליך ה-ETL עבור מערך הנתונים כולל את הפעולות הבאות בכל שלב:

• Extract

- חילוץ הנתונים מקבצי המקור, כמו קבצי CSV, JSON או מסד נתונים המכיל את מערך הנתונים.

• Transform

- ניקוי והפיכת הנתונים שחולצו כדי שיתאימו לניתוח וטעינה למחסן הנתונים.
- ניקוי נתונים כולל טיפול בערכים חסרים, תיקון פורמטים לא עקביים של נתונים, הסרת כפילויות והבטחת איכות הנתונים.
- הטרנספורמציות עשויות לכלול המרת פורמטים של תאריכים, סטנדרטיזציה של ז'אנרים או דירוגים, גזירת מדדים נוספים או ביצוע נורמליזציה של נתונים.

• Load

- טעינת הנתונים שעברו טרנספורמציה למחסן הנתונים או למערכת היעד, תוך הקפדה על המבנה של סכימת Star.
- הנתונים שעברו טרנספורמציה נטענים לטבלאות המתאימות של מחסן הנתונים, כגון טבלת ה-Fact וטבלאות מימדים (למשל, סרט, תאריך יציאה, במאי וכו').
- ניתן לבצע טעינה באמצעות שאילתות SQL, כלי טעינת נתונים או תוכנת ETL התומכת במסד הנתונים היעד.

תהליך ה-ETL עבור מערך הנתונים כולל חילוץ הנתונים הגולמיים מהמקור, הפיכתם כך שיתאימו למבנה ולתקני האיכות הרצויים, ולאחר מכן טעינתם למחסן הנתונים. זה מכין את הנתונים לשאלתה, ניתוח ודיווח יעילים, ומאפשרים למשתמשים להפיק תובנות חשובות ממערך הנתונים המשולב והמשתנה.

מיפוי העמודות משכבת הODS אל DataWarehouse

Movie_Fact	
US Gross	US_Gross
Worldwide Gross	Worldwide_Gross
US DVD Sales	US_DVD_Sales
Production	Production
Budget	Budget
Running time	Running_Time
Rotten Tomatoes Rating	Rotten_Tomatoes_Rating
IMDB Rating	IMDB_Rating
IMDB Votes	IMDB_Votes

Movie	
Title	Title

Release_Date	
Release Date	Release_Date

Source	
Source	Source_Name

Major Genre	
Major Genre	Major_Genre

Creative_Type	
Creative type	Creative_Type

Distributor	
Distributor	Distributor_Name

Director	
Director	Director_Name

שאלתה 1: סך ההכנסות ברוטו ברחבי העולם עבור כל הסרטים.

```
SELECT SUM(Worldwide_Gross) AS Total_Worldwide_Gross  
FROM Movie_Fact;
```

שאלתה 2: תקציב הפקה ממוצע לסרטים לפי ז'אנר.

```
SELECT g.Major_Genre, AVG(mf.Production_Budget) AS Average_Production_Budget  
FROM Movie_Fact mf  
JOIN Major_Genre g ON mf.Major_Genre_ID = g.Major_Genre_ID  
GROUP BY g.Major_Genre;
```

שאלתה 3: 10 הסרטים המובילים עם הדירוג הגבוה ביותר ב-IMDB.

```
SELECT m.Title, mf.IMDB_Rating  
FROM Movie_Fact mf  
JOIN Movie m ON mf.Movie_ID = m.Movie_ID  
ORDER BY mf.IMDB_Rating DESC  
LIMIT 10;
```

שאלתה 4: סך ההכנסות ברוטו ברחבי העולם עבור כל במאי.

```
SELECT d.Director_Name, SUM(mf.Worldwide_Gross) AS Total_Worldwide_Gross  
FROM Movie_Fact mf  
JOIN Director d ON mf.Director_ID = d.Director_ID  
GROUP BY d.Director_Name;
```



שאלתה 5: ספירת הסרטים לפי דירוג MPAA.

```
SELECT m.MPAA_Rating_ID, COUNT(*) AS Movie_Count
FROM Movie_Fact mf
JOIN Movie m ON mf.Movie_ID = m.Movie_ID
GROUP BY m.MPAA_Rating_ID;
```

שאלתה 6: דירוג ממוצע של Rotten Tomatoes לפי ז'אנר.

```
SELECT g.Major_Genre, AVG(mf.Rotten_Tomatoes_Rating) AS Average_Rotten_Tomatoes_Rating
FROM Movie_Fact mf
JOIN Major_Genre g ON mf.Major_Genre_ID = g.Major_Genre_ID
GROUP BY g.Major_Genre;
```

שאלתה 7: סך מכירות ה-DVD בארה"ב עבור כל מפיץ.

```
SELECT d.Distributor_Name, SUM(mf.US_DVD_Sales) AS Total_US_DVD_Sales
FROM Movie_Fact mf
JOIN Distributor d ON mf.Distributor_ID = d.Distributor_ID
GROUP BY d.Distributor_Name;
```

שאלתה 8: 5 הסרטים המובילים עם ההכנסה הגולמית הגבוהה ביותר בארה"ב.

```
SELECT m.Title, mf.US_Gross
FROM Movie_Fact mf
JOIN Movie m ON mf.Movie_ID = m.Movie_ID
ORDER BY mf.US_Gross DESC
LIMIT 5;
```



שאלתה 9: דירוג IMDB ממוצע לפי דירוג MPAA.

```
SELECT m.MPAA_Rating_ID, AVG(mf.IMDB_Rating) AS Average_IMDB_Rating
FROM Movie_Fact mf
JOIN Movie m ON mf.Movie_ID = m.Movie_ID
GROUP BY m.MPAA_Rating_ID;
```

שאלתה 10: סה"כ הכנסות ברוטו ברחבי העולם עבור כל מקור.

```
SELECT s.Source_Name, SUM(mf.Worldwide_Gross) AS Total_Worldwide_Gross
FROM Movie_Fact mf
JOIN Source s ON mf.Source_ID = s.Source_ID
GROUP BY s.Source_Name;
```

V. מסקנות

1. תקציב הפקה ממוצע לפי ז'אנר:

השאלתה המנתחת את תקציב ההפקה הממוצע לפי ז'אנר מרכזי חושפת שינויים משמעותיים בהקצאת התקציב בין הז'אנרים. תובנה זו יכולה לעזור לעסק לקבל החלטות מושכלות לגבי אסטרטגיות השקעה עבור ז'אנרים שונים. לדוגמה, הם עשויים לבחור להקצות תקציבים גדולים יותר לז'אנרים שנוטים לייצר הכנסות גבוהות יותר או שיש להם ביקוש גבוה יותר לקהל.

2. סרטים מובילים לפי דירוג IMDB:

השאלתה המזהה את הסרטים המובילים על סמך דירוגי IMDB מדגישה את הסרטים שקיבלו את הדירוג הגבוה ביותר מהצופים. מידע זה יכול להיות בעל ערך עבור העסק במונחים של הבנת העדפות הקהל וזיהוי גורמים פוטנציאליים התורמים להצלחתם של סרטים בעלי דירוג גבוה. זה יכול לעזור להנחות החלטות הקשורות ליצירת תוכן, שיווק ומיקוד לפלחי קהל ספציפיים.

3. סך ההכנסה הגולמית העולמית לפי מקור:

ניתוח סך ההכנסה הגולמית העולמית לפי מקור מספק תובנות לגבי הביצועים והיעילות של מקורות שונים ביצירת הכנסות. מידע זה יכול לסייע בהערכת ההחזר על ההשקעה (ROI) עבור כל מקור ולסייע בקבלת החלטות לגבי שותפויות, אסטרטגיות הפצה והקצאת משאבים כדי לייעל את ייצור ההכנסות.