**פרויקט בינה עסקית**

1. **שאלות מחקר וKPIS**

המחקר שלנו עוסק בגידול זחלים ל-2 מטרות, פיטום ורבייה. המחקר מורכב מניסויים רבים כאשר התוצאות בניסויים שלנו הם דגימות יומיות.

שאלת המחקר שלנו היא, מהם תנאי הגידול (דיאטה /מצע / גודל מיכל/ טמפרטורה/ לחות) מתאימה לכל צורך (הפקת חלבון מקסימלית/ רבייה יעילה).

הגדרת KPI’S

* 1. **מציאת תנאי גידול אופטימליים לשלב הפיטום:**
     + משקל זחל מקסימלי עד שלב הגולם
     + הרכב התנאים (טמפרטורה, לחות, יחס מצע:רימה, שטח המצע – גובה ורוחב)
  2. **מציאת תנאי גידול אופטימליים לשלב הרבייה:**
     + צמצום אחוז אובדן הגלמים הבוקעים
     + צמצום אחוז אובדן הזבובים לפני זיווג
     + צמצום אחוז אובדן הביצים המוטלות (מביצה לרימה)
     + הרכב התנאים (טמפרטורה, לחות, יחס מצע:רימה, שטח המצע – גובה ורוחב)

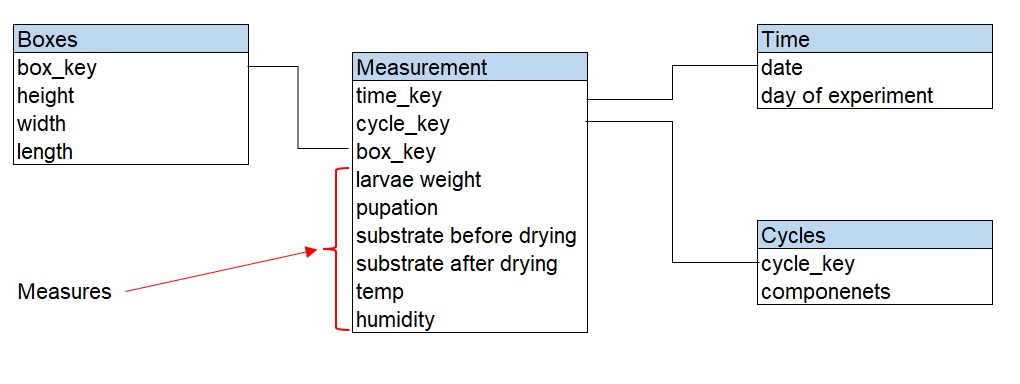
1. **הגדרת Data Warehouse**
2. מחסן נתונים הוא מאגר מרכזי המאחסן ומארגן כמויות גדולות של נתונים מובנים ולעיתים לא מובנים ממקורות שונים בתוך הארגון. הוא נועד לתמוך בפעילויות בינה עסקית, דיווח וניתוח נתונים. בהקשר של הפרויקט שלנו, בחרנו ליישם את **סכימת Star** עבור מחסן הנתונים שלנו.

סכמת Star קלה להבנה. היא מורכזת מטבלה מרכזית המכילה את מדדי המפתח וסביבה יש טבלאות ממדים המספקות הקשר ותכונות נוספות. המבנה הפשוט הזה הופך את זה לאינטואיטיבי למשתמשים לנווט ולהבין את היחסים בין הנתונים.

בנוסף, סכמת Star מפחיתה את מספר פעולות ה-JOIN הנדרשות בעת שאילתת הנתונים. במקרה שלנו, אותם נתונים משמשים לבדיקות שונות, שעשויות לכלול טבלאות נפרדות.

על ידי מינוף היתרונות של סכימת Star, נוכל לייעל את הביצועים של מחסן הנתונים שלנו ולאפשר ניתוח נתונים יעיל. הסכימה מפשטת את תהליך איסוף וניתוח הנתונים, ומאפשרת לנו לחלץ תובנות ולקבל החלטות מושכלות על סמך המידע שנאסף.

1. ERD:



III. נניח שאנו רוצים להציג את המשקל הממוצע של הזחלים ביום 5 של הניסויים, מסווג לפי מצע, ונפח קופסא.

בסכימה חלופית, אם לא היינו בוחרים בסכמת Star, היינו נדרשים לבצע לפחות שלוש פעולות JOIN. זה היה גורם להאטה ניכרת בזמן העיבוד ובביצועי השאילתות. כל פעולת JOIN מציגה מורכבות נוספת ותקורה חישובית, במיוחד כאשר עוסקים במערכי נתונים גדולים.

1. **הגדרה ומימוש ETL**

ETL (Extract, Transform, Load) מתייחס לתהליך של חילוץ נתונים ממקורות שונים, הפיכתם לפורמט רצוי וטעינתם למחסן נתונים (DW) או למערכת יעד אחרת. במקרה זה, מקור הנתונים הוא טבלת אקסל, ואנו נדון בפעולות הספציפיות שיש לבצע במהלך תהליך ה-ETL.

Extract:

השלב הראשון בתהליך ה-ETL הוא לחלץ את הנתונים מטבלת האקסל. נעשה זאת באמצעות כתיבת סקריפטים מותאמים אישית לקריאת הנתונים מקובץ האקסל.

Transform:

לאחר חילוץ הנתונים, עלינו לבצע עליו טרנספורמציות. במקרה זה,:

א. סינון שורות ריקות:

תהליך ה-ETL יכלול סינון של כל שורות ריקות מטבלת Excel. שורות ריקות בדרך כלל אינן מכילות נתונים משמעותיים וניתן להסיר אותן בבטחה. שלב זה מבטיח שרק שורות רלוונטיות ומאוכלסות ייחשבו להמשך עיבוד.

ב. טיפול בערכים ריקים:

נגדיר ערך ברירת מחדל של 0 עבור תאים ריקים בשורות שבהן יש נתונים. זה מבטיח עקביות ומאפשר ניתוח וחישובים נכונים בתהליכים במורד הזרם.

Load:

לאחר שהנתונים עברו טרנספורמציה, הם מוכנים לטעינה למחסן הנתונים. טעינה כרוכה באחסון הנתונים שעברו טרנספורמציה באופן מאורגן המתאים למבנה ה-DW. תהליך הטעינה התבצע באמצעות סקריפטים מותאמים אישית שמתחברים ל-DW ומכניסים את הנתונים שעברו טרנספורמציה.

1. **ניתוח ה- Data Warehouse ומסקנות.**
2. **Average weight on day 15 for each diet, in descending order:**

SELECT Diet, AVG("10 larvae weight (mg)") AS AvgWeight

FROM (SELECT \*, ROW\_NUMBER() OVER (PARTITION BY Diet ORDER BY Cycle DESC) AS rn

FROM bi\_project\_data.json

WHERE Cycle = 15) AS subquery

WHERE rn = 1

GROUP BY Diet

ORDER BY AvgWeight DESC;

1. **Average larvae weight per day 12-15 for each humidity level:**

SELECT Humidity, AVG("10 larvae weight (mg)") AS AvgWeight

FROM (SELECT \*, ROW\_NUMBER() OVER (PARTITION BY Humidity ORDER BY Cycle) AS rn

FROM bi\_project\_data.json

WHERE Cycle BETWEEN 12 AND 15) AS subquery

WHERE rn = 4

GROUP BY Humidity;

1. **Average larvae weight per day 12-15 for each temperature:**

SELECT Temperature, AVG("10 larvae weight (mg)") AS AvgWeight

FROM (SELECT \*, ROW\_NUMBER() OVER (PARTITION BY Temperature ORDER BY Cycle) AS rn

FROM bi\_project\_data.json

WHERE Cycle BETWEEN 12 AND 15) AS subquery

WHERE rn = 4

GROUP BY Temperature;

1. **Average larvae weight on day 13-15, sorted by average temperature:**

SELECT AVG("10 larvae weight (mg)") AS AvgWeight, AVG(Temperature) AS AvgTemperature

FROM (SELECT \*, ROW\_NUMBER() OVER (ORDER BY Cycle) AS rn

FROM bi\_project\_data.json

WHERE Cycle BETWEEN 13 AND 15) AS subquery

GROUP BY rn

ORDER BY AvgTemperature DESC;

1. **Average larvae weight on day 13-15, sorted by average humidity:**

SELECT AVG("10 larvae weight (mg)") AS AvgWeight, AVG(Humidity) AS AvgHumidity

FROM (SELECT \*, ROW\_NUMBER() OVER (ORDER BY Cycle) AS rn

FROM bi\_project\_data.json

WHERE Cycle BETWEEN 13 AND 15) AS subquery

GROUP BY rn

ORDER BY AvgHumidity DESC;

1. **Average weight on day 15 for each diet, along with substrate weight and their relation:**

SELECT Diet, AVG("10 larvae weight (mg)") AS AvgWeight, "Substrate after drying, gr", "10 larvae weight (mg)" / "Substrate after drying, gr" AS Relation

FROM bi\_project\_data.json

WHERE Cycle = 15

GROUP BY Diet, "Substrate after drying, gr";

1. **Average weight on day 15 for each diet, in descending order, along with larvae-substrate relation:**

SELECT Diet, AVG("10 larvae weight (mg)") AS AvgWeight, "10 larvae weight (mg)" / "Substrate after drying, gr" AS Relation

FROM bi\_project\_data.json

WHERE Cycle = 15

GROUP BY Diet

ORDER BY AvgWeight DESC;

1. **Show first-day weight, calculate estimated number of eggs, show pupation percentage, calculate estimated pupation, and their relation per diet:**

SELECT Diet, "10 larvae weight (mg)" AS FirstDayWeight,

("10 larvae weight (mg)" / EstimatedEggWeight) AS EstimatedEggs,

"pupation, %" AS PupationPercentage,

("pupation, %" / 100) \* ("pupation, X/50" \* 50) AS EstimatedPupation,

"10 larvae weight (mg)" / ("pupation, %" / 100) AS Relation

FROM bi\_project\_data.json

WHERE Cycle = 1;

1. **Show first-day weight, calculate estimated number of eggs, show larvae weight, calculate estimated number of larvae, and their relation per diet:**

SELECT Diet, "10 larvae weight (mg)" AS FirstDayWeight,

("10 larvae weight (mg)" / EstimatedEggWeight) AS EstimatedEggs,

"10 larvae weight (mg)" AS LarvaeWeight,

("10 larvae weight (mg)" / EstimatedLarvaeWeight) AS EstimatedLarvae,

"10 larvae weight (mg)" / ("10 larvae weight (mg)" / EstimatedLarvaeWeight) AS Relation

FROM bi\_project\_data.json

WHERE Cycle = 1;

1. **Show average temperature, personal temperature, larvae weight on day 15, average humidity, personal humidity, and their larvae-substrate relation:**

SELECT AVG(Temperature) AS AvgTemperature, Temperature AS PersonalTemperature,

"10 larvae weight (mg)" AS LarvaeWeightOnDay15,

AVG(Humidity) AS AvgHumidity, Humidity AS PersonalHumidity,

"10 larvae weight (mg)" / ("Substrate after drying, gr" - "Substrate before drying, gr") AS Relation

FROM bi\_project\_data.json

WHERE Cycle = 15;