



# מחשבי נתונים

עקרונות, ארכיטקטורה,  
עיצוב ויישום

רז הייפרמן

הוצאת הוד-עמי

לספרי מחשבים 09-9564716

www.hod-ami.co.il

הוצאת הוד-עמי

לספרי מחשבים



עורך ראשי: יצחק עמיהוד



עריכה: צור לוין

© כל הזכויות שמורות

**הוצאת הוד-עמי לספרי מחשבים בע"מ**

ת.ד. 6108 הרצליה 46160

טלפון: 09-9564716 פקס: 09-9571582

דואר אלקטרוני: [info@hod-ami.co.il](mailto:info@hod-ami.co.il)

אתר באינטרנט: <http://www.hod-ami.co.il>

[www.hod-ami.co.il](http://www.hod-ami.co.il)

# הקדמה

**מחסני הנתונים** (data warehouse) הפך בשנים האחרונות לאחד הנושאים החמים ביותר בתחום טכנולוגיות המידע. על פי סקרים שונים עולה שמספר הארגונים שעוסקים בצורה כלשהי בהקמת מחסני נתונים צמח במהירות מ-10% בשנת 1994 עד ל-90% בשנת 1998. תוך פרק זמן קצר יחסית הפכו מחסני הנתונים לפתרון המועדף לבניית סביבת מיחשוב ייעודית לניהול ואספקת המידע הדרוש לניתוח נתונים ולתמיכה בתהליכי קבלת החלטות עסקיות. ההתפתחות המהירה של תפיסת מחסן הנתונים באה לאחר שנים רבות בהם ארגונים פיתחו ויישמו מערכות תומכות החלטה (DSS) לצד המערכות התפעוליות, אולם ללא הצלחה משמעותית ועם הרבה בעיות וקשיים. רק בשנים האחרונות הבשילו התנאים להצלחת מערכות תומכות החלטה: הלחצים הבלתי פוסקים על הארגון לקבל החלטות מורכבות ומהירות מחד, ומאידך - ההתפתחויות בטכנולוגיות המידע בכל הקשור לחומרה, תוכנה וכלים מתקדמים לניתוח הנתונים. תנאים אלה הביאו להתפתחות תפיסה כוללת לפיתוח ויישום סביבת מיחשוב ייעודית לאספקת מידע לתהליכי קבלת החלטות – תפיסת **מיחסון הנתונים - Data Warehousing**.

הגדרת המושג "מחסן נתונים" עמומה במידת מה. זו סביבת מיחשוב ייעודית לניתוח ולעיבוד אנליטי של הנתונים והפיכתם למידע **תומך לקבלת החלטות**. מחסן הנתונים מורכב משלושה רכיבים עיקריים: רכיב העוסק בתהליכי אחסון הנתונים ממערכות המקור, רכיב העוסק בניהול הנתונים ורכיב העוסק באספקת הנתונים ובביצוע ניתוח המידע. רק לאחר שמתחילים לעכל את משמעות המושג מיחסון נתונים, להעמיק ולנתח את מלוא משמעות תפיסת מחסן הנתונים, מתחילה להתבהר התמונה האמיתית. לפנינו,

- ❖ סביבה עתירת טכנולוגיות מידע שונות,
- ❖ סביבה שצריכה לענות לדרישות משתנות שקשה להגדירן מראש באופן מפורש,
- ❖ סביבה שצריכה לנהל היקף נתונים גדול,
- ❖ סביבה שמכניסה מימד בלתי ידוע עד כה של כפילות נתונים אל הארגון,
- ❖ סביבה העוסקת במיגוון רחב של תהליכים העוסקים בגירת נתונים, טיוב הנתונים וטעינתם, ועוד.

כמו כל הליט ומגמה טכנולוגית חדשנית, גם מיחסון הנתונים מלווה בסיסמאות רבות ובציפיות שחלקן אינן ברות השגה. יש חשיבות רבה בהבנת המציאות והאתגרים האמיתיים שהארגון יצטרך להתמודד אייתם כשיצטרך ליישם מחסן נתונים וללמוד את התועלות שהוא צופה להשיג. יחד עם התלהבות הארגונים באימוץ תפיסת מחסן הנתונים, שסוף כל סוף מאפשרת להפוך את הנתונים האגורים במערכות המידע השונות של הארגון למידע תומך החלטות, באה המציאות ומוכיחה שתפיסת מחסן הנתונים היא תפיסה מורכבת ויישומה דורש משאבים רבים. לארגונים התבררו מספר עובדות:

- ❖ מחסן נתונים לא קונים אלא בונים,
- ❖ מעט מאוד ספקים מסוגלים לספק את הקשת הרחבה של המוצרים הדרושים להקמת מחסן הנתונים,
- ❖ תהליך הקמת מחסן נתונים הוא תהליך מורכב יחסית של שילוב מערכות,

❖ הנתונים במערכות התפעוליות אינם במבנה וברמת ניקיון מתאימה לתמיכה בהחלטות,

❖ הקמת מחסן נתונים הוא פרויקט ארגוני שנמשך זמן,

❖ הפרויקט צורך משאבים רבים והוא בעל מאפיינים שונים מתהליך הקמת מערכת מידע חדשה.

במקביל לסיפורי ההצלחה החלו להישמע גם לא מעט סיפורים על כישלונות ופרויקטים שננטשו לאחר שהושקעו משאבים רבים בהקמתם. בדיעבד הסתבר שחלק מהארגונים קפצו על העגלה במהירות רבה מדי והתייחסו אל פרויקט מחסן נתונים כאל עוד פרויקט פיתוח מערכת מידע. הם לא טרחו ללמוד ולנסות להבין את מלוא המשמעויות של הקמת מחסן נתונים ובסופו של דבר נכשלו ביישום נושא חשוב זה והפסידו הרבה זמן וכסף.

למרות האתגרים והקשיים בהקמת מחסן הנתונים, ארגונים צריכים ללמוד את התועלות שבתפיסת מחסן הנתונים ואת המשמעויות האמיתיות של הקמת סביבת מיחשוב מודרנית לתמיכה בקבלת החלטות. ארגונים אלה יצליחו להמשיך לתפקד ולהצליח בסביבה העסקית המודרנית, התחרותית והדינמית כל כך. יש ארגונים שעלולים להיבהל מהאתגרים ולהתמקד בעיקר במערכות התפעוליות, ולא במערכות תומכות החלטה. גישה כזו מוליכה לסיכון של איבוד אחד היתרונות התחרותיים החשובים ביותר: היכולת לנצל את הנתונים והמידע כמשאב תחרותי. כאשר מסתכלים על הנתונים כמשאב תחרותי, אפשר להבין טוב יותר ומהר יותר מה קורה לארגון, מה קורה למוצרים שהוא מוכר, מה לקוחותיו באמת רוצים וכיצד ניתן להגביר את נאמנותם למוצרי החברה, מה קורה לשוק שבו הארגון פועל ובסופו של דבר - כיצד לקבל החלטות עסקיות טובות יותר.

מטרת ספר זה להציג ולסקור את תפיסת מחסן הנתונים על כל היבטיה: מדוע המערכות התפעוליות ומערכות לתמיכה בהחלטות שונות זו מזו ויש צורך להקים סביבת מיחשוב ייעודית למיחסון הנתונים. הקמת סביבת עיבוד חדשה זו מחייבת:

❖ להכיר את הארכיטקטורות השונות לניתוח נתונים והכלים האנליטיים הפועלים בסביבה זו,

❖ להכיר את טכניקות העיצוב של בסיסי הנתונים המשרתים את מחסן הנתונים,

❖ לבחון את התמונה הכוללת של התהליכים הקיימים בסביבת מחסן הנתונים,

❖ ללמוד על החלופות השונות לפיתוח מחסני נתונים ואת מתודולוגיות היישום שלהם.

הכרת מכלול נושאים מקיף זה תאפשר לכל ארגון להבין בצורה טובה את מלוא המשמעויות של פרויקט מחסן הנתונים, ולקבל את החלטות בהתבסס על הבנה מוצקה הן של התועלות והן של הקשיים.

הגישה של הספר היא מעשית ולא תיאורטית. הנושאים מוצגים בלשון פשוטה ולא טכנית או מתמטית, הטקסט מלווה במספר רב של דוגמאות ושל תרשימים. נקודת המוצא היא שהעקרונות והארכיטקטורה של מחסן הנתונים חשובים בהרבה ממוצרים ספציפיים שונים. על כן, חלק מהמוצרים מזכרים באופן כללי וכמעט ללא פירוט של מוצר מסוים כלשהו. טכנולוגיית מחסני הנתונים חדשה יחסית ומורכבת מאוסף גדול של מוצרים שונים המשתנים כל הזמן. לכן יש חשיבות רבה יותר להבנת העקרונות, התמונה הכוללת והמקום של כל מוצר במסגרת הארכיטקטורה, מאשר לאופן הפעולה הספציפי של מוצר מסוים. בנוסף, ככל שהזמן שעובר והספר יהיה מונח על שולחן עבודתך הפרטים לגבי מוצר זה או אחר כבר לא יהיו תקפים, בשעה שהעקרונות ישארו איתנו זמן רב. הנקודה העיקרית המודגשת בספר היא החשיבות בבניית מחסן נתונים גמיש המסתגל בקלות יחסית לשינויים שהארגון עובר, לדרישות המידע החדשות, לטיפול בנפחי

נתונים הולכים וגדלים ולמוצרים החדשים. חוסר הבנה של התהליכים העיקריים של סביבת מחסן הנתונים והתבססות על ארכיטקטורה שלא הושקעה בה המחשבה הנדרשת, עלולים לגרום לכישלון התפיסה ולעוגמת נפש, גם אם בשלב ההקמה הארגון השתמש במוצרים מהשורה הראשונה.

הספר התגבש כתוצאה מפעילות בתחום טכנולוגיות המידע, ובעיקר בנושאי בסיסי נתונים, טכנולוגיות שרת/לקוח ומערכות מוכוונות אובייקטים. בשנים האחרונות נחשפתי לנושא מחסני נתונים, נתתי מספר רב של הרצאות במסגרות שונות של מוסדות הוראה וימי עיון מקצועיים שונים ועסקתי בייעוץ להקמת מערכות כאלו. זו אחת מטכנולוגיות המידע החשובות ביותר ויחד עם זאת אחת המורכבות למימוש ויישום, ולכן אני מאמין בחשיבות ספר בשפה העברית שיפרוס בפני הקורא את כל רוחב היריעה של הנושא.

## מבנה הספר

נסקור בקצרה את תוכן הפרקים השונים של הספר, כדי לאפשר לקורא לקבל התרשמות ראשונית ומהירה של תכולת הספר.

### פרק 1 - רקע ומושגי יסוד

פרק זה מציג את הרקע להתפתחות תפיסת מחסן הנתונים תוך הדגשת שתי המגמות העיקריות. מצד אחד, השינויים בסביבה העסקית שהביאה ללחצים מתמידים לקבלת החלטות מהירות ומבוססות; ומצד שני, ההתפתחויות המהירות בטכנולוגיית המידע עם הופעת המחשבים האישיים, השרתים המקביליים רבי העצמה והזולים יחסית, בסיסי הנתונים היחסיים המסוגלים לטפל בהיקפי נתונים גדולים והופעת רשתות התקשורת המהירות. הפרק מפרט את ההבדלים המהותיים בין היישומים התפעוליים של הארגון התומכים בתהליכי העבודה השוטפים לבין היישומים האנליטיים העוסקים בניית מידע ותמיכה בתהליכי קבלת ההחלטות. לסיום, מוצגת ההתפתחות רבת השנים של המערכות לקבלת החלטות, שהחלה בשנים הראשונות של המיחשוב והשתכללה עם הזמן עד להגדרת התפיסה המודרנית של מחסן הנתונים.

### פרק 2 - ארכיטקטורת מחסן הנתונים (Data Warehouse Architecture)

פרק זה סוקר את הארכיטקטורות השונות למימוש מחסן הנתונים: מחסן הנתונים הארגוני; מרכז הנתונים; מחסן הנתונים הרב-שכבתי המשלב את היתרונות של המחסן הארגוני עם היתרונות של מרכזי הנתונים; ומאגר הנתונים התפעולי שהוא סוג מסוים של מחסן נתונים המתאים לאספקת מידע לתהליכים תפעוליים.

### פרק 3 - סקירת מרחב הטכנולוגיות לניתוח מידע

#### (Analytical Processing Technology Space)

פרק זה מציג את כל המרחב של נושא ניתוח המידע והעיבודים האנליטיים לתמיכה בקבלת החלטות. כדי להבהיר את תהליך ניתוח המידע, מוצגת שרשרת הפעילויות ההופכת נתונים למידע, את המידע לידיע ואת הידיע להחלטות עסקיות. המרחב המלא של שיטות ניתוח מידע שונות מוצג גם הוא: מחוללי שאילתות ודוחות המשמשים לניתוחים פשוטים יחסית, כלי OLAP המורכבים יותר והמאפשרים ניתוח רב מימדי והכלים לכריית נתונים המאפשרים לבצע ניתוחי מידע מורכבים מאוד ולגלות חוקיות ותבניות בלתי ידועות בנתונים. הפרק מסתיים בהצגת מפה של מרחב הטכנולוגיות השונות לניתוח מידע על פי דרישות המידע וסוגי משימות הניתוח הנדרשים.

### פרק 4 - ניתוח מידע רב מימדי (Multi Dimensional Analysis)

פרק זה מעמיק ומסביר את הדרכים לניתוח רב-מימדי של מידע, אחת מפעילויות הניתוח והעיבוד האנליטי הנפוצות ביותר. הפרק עוסק בסקירת מושגי היסוד בנייתוח רב מימדי: מה זה מימד, מה זה עובדה, מהן ההיררכיות בתוך המימדים, מהן הפעולות העיקריות בנייתוח רב-מימדי, מהי הקובייה הרב-מימדית ועוד.

**פרק 5 - ארכיטקטורות של מוצרים לניתוח רב-מימדי (OLAP Architectures)**

הפופולריות הרבה של שיטת הניתוח הרב-מימדי תרמה לפיתוח מספר רב של מוצרים התומכים בה. פרק זה סוקר את הארכיטקטורות השונות למימוש הניתוח הרב-מימדי ובאופן מיוחד את שתי הארכיטקטורות העיקריות: ארכיטקטורת MOLAP המבוססת על בסיס נתונים רב-מימדי ייעודי, וארכיטקטורת ROLAP המבוססת על מנוע רב-מימדי המבצע את מירב העיבודים הדרושים אולם מנהל את הנתונים בבסיס נתונים טבלאי רגיל. הפרק סוקר את ההבדלים בצורת המימוש של ארכיטקטורות אלו ומדגיש שלמעשה הן אינן נוגדות, כמו שלעיתים יצרנים היו רוצים לגרום לנו לחשוב, אלא משלימות.

**פרק 6 - עיצוב רב מימדי של בסיסי נתונים יחסיים (Multi Dimensional Design of Relational Databases)**

פרק זה סוקר את שיטת העיצוב המיוחדת של בסיסי הנתונים היחסיים הפועלים בסביבת מחסן הנתונים. חלק גדול מתפישות העיצוב הרגילות של בסיסי נתונים טבלאיים המבוססות על ניתוח קל ופשוט של המידע על ידי המשתמשים אינו מתאים לתפישת הייעוד של מחסן הנתונים. פרק זה מציג את תפיסת העיצוב המיוחדת של בסיסי נתונים המשרתים מחסן נתונים תוך הצגת מושגי היסוד: טבלת עובדות, טבלת מימד, טבלאות לא מנורמלות, סכמות כוכב, סכמות פתיתי שלג, טבלאות סיכומים, שיטות לניהול שינויים במימדים ועוד.

**פרק 7 - כריית נתונים (Data Mining)**

פרק זה מציג את אחד הנושאים ה'חמים' ביותר כיום במסגרת תפיסת מחסן הנתונים. לרבים נושא זה נשמע כמו מדע בדיוני שרלוונטי בעיקר לחוקרים ומדענים. המציאות כמובן שונה מאוד. כריית הנתונים מורכבת ממספר טכנולוגיות מיחשוב ומוצרי תוכנה שונים המאפשרים לארגון לנתח את הנתונים, לגלות חוקיות פנימית ובלתי ידועה מראש, לגלות תבניות ולבצע תחזיות מבוססות. לכריית נתונים משתמשים במספר טכניקות סטטיסטיות וטכניקות הלקוחות מתחומי מחקר מתקדמים כמו אינטליגנציה מלאכותית, מערכות לומדות וכד'. הפרק סוקר את מיגוון היישומים האפשריים, את הטכניקות הנפוצות לכריית נתונים ואת מקומו של תהליך כריית הנתונים במסגרת תפיסת מחסן הנתונים.

**פרק 8 - הטכנולוגיה המקבילית בשירות מחסן הנתונים (Parallel Technology in Data Warehouses)**

ההתפתחויות המהירות בשנים האחרונות בכל הקשור לארכיטקטורות חומרה מרובות המעבדים השפיעו רבות על תפיסת מחסן הנתונים. התברר שניתן לנהל ולעבד באופן יעיל נפחי נתונים עצומים וגם לעשות זאת בעלויות סבירות. לטכנולוגיה המקבילית הייתה השפעה על מספר רב מאוד של תחומים הקשורים למחסן הנתונים: ניהול בסיס נתונים, אופטימיזציה של גישה לבסיס נתונים, טעינה מהירה של נתונים ועוד. הפרק סוקר את הארכיטקטורות העיקריות של ריבוי מעבדים, את השיטות לעיבוד מקבילי של שאילתות ואת שיטת חלוקת בסיס הנתונים למחיצות כשיטה לניהול יעיל של נפחי נתונים גדולים.

**פרק 9 - סקירת התהליכים העיקריים במחסן הנתונים (Main Processes in Data Warehouses)**

מחסן הנתונים הינו עתיר תהליכי עבודה: תהליכי גזירת הנתונים מהמערכות התפעוליות, תהליכי סיכום נתונים לייעול ניתוח הנתונים, תהליכי אספקת הנתונים לכלי הניתוח השונים ותהליכי גיבוי והעברת הנתונים למדיה זולה יותר. המציאות היא, שארגון משקיע אנרגיה רבה בהקמת ותחזוקה שוטפת של תהליכים אלה. פרק זה סוקר את מכלול התהליכים תוך מתן דגש מיוחד לתהליכים המורכבים לגזירה, לטיוב ולהמרה של הנתונים, שנעשים לרוב באמצעות תוכנה ייעודית.

**פרק 10 - מידע על נתונים (Meta Data)**

פרק זה מציג את תחום המידע על הנתונים. נושא זה הינו בעל חשיבות רבה להצלחת מחסן הנתונים, אולם בגלל מורכבות והעדר כלים מספיק טובים ומקיפים, הוא אינו זוכה לתשומת הלב המתאימה מצד מפתחי מחסן הנתונים. הפרק סוקר את הסוגים הרבים והמגוונים של מידע על הנתונים שיש לנהל בסביבת מחסן נתונים: משמעות הנתונים, מבנה הנתונים, היכן הם מאוחסנים, למתי הם נכונים, מה מקורם ועוד.

### פרק 11 - מרכולי נתונים (Data Mining)

מרכול הנתונים הפך בשנים האחרונות לאחד הנושאים החשובים והבולטים ביותר במסגרת תפיסת מחסן הנתונים. למעשה, זהו מחסן נתונים ממוקד במחלקה מסוימת, בנושא מסוים או במטרה מסוימת. תפיסה זו התפתחה על רקע הקשיים בהקמת ומימוש תפיסת מחסן הנתונים הארגוני, שהינה מקיפה ורחבה. מרכול הנתונים אינו מתיימר לתת פתרון מקיף לארגון בתחום התמיכה בקבלת החלטות. זהו גם מקור ההתלהבות והעניין בו: קל יחסית להקימו, עלות ההקמה זולה יותר, הבעלות עליו ברורה, קל יותר להצדיקו מבחינת עלות/תועלת. אך בצד התועלות הברורות של תפיסת מרכול הנתונים, יש לא מעט סיכונים בתפיסה זו. פרק זה סוקר את תפיסת מרכול הנתונים, מה הרקע לפופולריות שלו, מהם הסיכונים במרכולי נתונים, מה הארכיטקטורה שלהם ומהם המודלים השונים להקמת מרכולי נתונים.

### פרק 12 - מתודולוגיה ליישום מחסן הנתונים

#### (Data Warehouse Implementation Methodology)

הפרק האחרון של הספר מוקדש למתודולוגיות היישום של מחסני נתונים. הפרק בוחן מדוע יש צורך במתודולוגיה ייחודית ומותאמת, מהם עקרונות המתודולוגיה, הוא מפרט את השלבים השונים בתהליך הקמת מחסן הנתונים, מהם בעלי התפקידים השונים הפועלים בסביבת מחסן הנתונים. לסיכום ניתנות מספר עצות מעשיות לבניית מחסן הנתונים.

# קהל היעד של הספר

הספר מיועד לקהל קוראים מגוון, וביניהם:

- ❖ מנהלים ומשתמשים עתידיים של מחסן הנתונים. חשוב שציפיותיהם ממחסן הנתונים תהיינה מעשיות וניתנות למימוש, ורצוי שיהיו מסוגלים להגדיר את דרישותיהם מתוך הבנת התהליך הכולל לבניית מחסן הנתונים.
- ❖ מנהלי מיחשוב המבקשים להקים מחסן נתונים בארגון. אלה מבקשים להבין טוב יותר את מלוא המשמעויות של פרויקט ההקמה והיישום ובכלל זה: מהו מחזור החיים של הפרויקט ובמה הוא שונה ממחזור חיים של מערכת מידע רגילה, מה המתודולוגיה להקמה ויישום, איזה סוגי עובדים צריך להקצות, מה האתגרים הטכנולוגיים שהם צפויים להתמודד איתם, ועוד.
- ❖ מנתחי מערכות המבקשים להבין את הטכנולוגיה החדשה ולהצטרף לפרויקט מחסן נתונים. עליהם ללמוד על תהליכי ההקמה, המושגים העיקריים, שיטת העבודה לניתוח המערכת והערכתה, הכלים הדרושים ועוד.
- ❖ מעצבי בסיס נתונים הרגילים לעצב בסיסי נתונים למערכות עיבוד תנועות רגילות ומבקשים להבין מה ההבדל בין טכניקות העיצוב הרגילות לבין טכניקות העיצוב המיוחדות לסביבת מחסן הנתונים.
- ❖ ארכיטקטים של מערכות מידע המבקשים לשלב את תפיסת מחסן הנתונים במסגרת הארכיטקטורה הכוללת של מערכות המידע בארגון.

הטבלה הבאה מציגה את הפרקים השונים המומלצים ביותר לסוגי הקוראים השונים.

מעצבי מערכות מידע	מעצבי בסיסי נתונים	מנתחי מערכות	מנהלי מיחשוב	מנהלים/ משתמשים	שם פרק
✓	✓	✓	✓	✓	רקע ומושגי יסוד
✓	✓	✓	✓	✓	ארכיטקטורה של מחסן נתונים
✓		✓	✓	✓	סקירת מרחב הטכנולוגיות לניתוח מידע
✓		✓	✓	✓	ניתוח מידע רב-מימדי
✓	✓	✓	✓		ארכיטקטורות מוצרים לניתוח רב-מימדי
	✓				עיצוב רב-מימדי של בסיסי נתונים טבלאיים
		✓	✓	✓	כריית נתונים
	✓		✓		הטכנולוגיה המקבילית
		✓	✓	✓	סקירת התהליכים העיקריים
		✓	✓	✓	מידע על נתונים
✓		✓	✓	✓	מרכולי נתונים
		✓	✓		מתודולוגיה ליישום מחסן נתונים

www.hod-ami.co.il



## פרק 2:

# ארכיטקטורת מחסן הנתונים

## כללי

בפרק זה נגדיר את תפיסת מחסן הנתונים, נציג את הארכיטקטורה הכוללת שלו ונפרט את מכלול הטכנולוגיות והכלים המרכיבים אותו.

בשנים האחרונות החלה להתגבש תפיסה חדשה לפיתוח וניהול יישומים לאספקת מידע לתמיכה בקבלת החלטות. תפיסה זו מבוססת על אוסף טכנולוגיות וכלי תוכנה חדישים שנקראת **מחסן נתונים** (Data Warehouse). תפיסת מחסן הנתונים מספקת מסגרת כוללת לניהול נתונים המיועדים לתמיכה בקבלת החלטות ומתן דגש על היעילות, העקביות והנוחות. מחסן הנתונים מבוסס על בסיס נתונים ייעודי ונפרד, המכיל נתונים אודות פעילות הארגון וסביבתו. הוא משמש באופן בלעדי את תהליכי ניתוח הנתונים ותמיכה בקבלת החלטות. בסיס נתונים ייעודי זה מכיל נתונים הנשלפים ממקורות שונים, תוך ביצוע תהליכי גירה, שיפור ושילוב מיוחדים. מטרתו אחת ויחידה - לספק את צרכי המידע של קהיליית משתמשים חשובה - **מקבלי ההחלטות בארגון**. בסיס נתונים זה חייב להיות זמין, עדכני, אמין ומאורגן באופן התואם את הישיות העסקיות העיקריות ואת תהליכי קבלת החלטות. חשוב להבין שבסיס נתונים, שמטרתו לתמוך בתהליכי קבלת החלטות מורכבות ודינמיות, חייב להיות מעוצב ומאורגן באופן מיוחד, התואם את מטרותיו. ארגון זה שונה באופן מהותי למדי מהדרך בה מאורגנים הנתונים המשמשים ומשרתים את המערכות התפעוליות.

מן הראוי לציין שתחום מחסני נתונים הוא תחום צעיר יחסית, ולכן הוא עדיין בתהליך התגבשות. המושגים עדיין אינם מוגדרים עד הסוף, הטכנולוגיה ממשיכה להשתנות, הארגונים עדיין נמצאים בתהליך יישום והפקת לקחים ולכן גם מתודולוגיות היישום הן עדיין בחיתוליהן. מכיון שתחום זה מתפתח בקצב מהיר, יש לצפות לכך שדברים ישתנו, הגדרות מדויקות יותר תתגבשנה וטכנולוגיות חדשות תופענה ותחלפנה טכנולוגיות קיימות. אם זו חשיפתו הראשונה של הקורא לתחום מחסן הנתונים, אני מקווה שדינמיות וחוסר ההתגבשות הסופית של התפיסה לא תרפה את ידיו. זהו אך תהליך טבעי בהתפתחותה של תפיסה הצומחת ותופסת פופולריות עצומה.

## מהו מחסן נתונים?

ברמה הבסיסית ביותר, **מחסן נתונים** היא תפיסה כוללת העוסקת באיסוף, אחסון, הפצה ושיתוף של נתונים למטרת תמיכה בקבלת החלטות. ניתן לומר שזו תפיסת קצה לקצה המטפלת בכל מיגוון התהליכים והפעילויות הדרושות כדי לספק את הנתונים הדרושים לתמיכה בקבלת החלטות עסקיות. מחסן הנתונים שואב את הפונקציונליות שלו מתוך העובדה שהנתונים נגזרים ממקורות שונים, עוברים תהליכי שיפור ומיטוב ומשולבים יחדיו. מחסן הנתונים שואב את עוצמתו מהקלות בה מוצגים נתונים אלה ומנותחים על ידי המשתמשים בו.

אחת ההגדרות המצוטטות ביותר שייכת למי שנחשב לאבי התפיסה, Bill Inmon, שבמאמר משנת 1992 הציג את ההגדרה הבאה:

**מחסן הנתונים** מוגדר כאוסף נתונים ייעודיים ומשולבים, מאורגנים לפי נושא, בעלי עומק היסטורי ושאינם מתעדכנים - אשר מיועדים לתמוך בתהליכי קבלת החלטות.

נתעכב בקצרה על כל אחד ממרכיבי ההגדרה הזו:

❖ **אוסף נתונים ייעודי (Dedicated Data Base):** בסיס הנתונים המשמש את מחסן הנתונים נפרד מבסיסי הנתונים התפעוליים של הארגון. החשיבות בהפרדה זו כבר נדון בפרק הראשון בספר זה, בו הוסברה האבחנה בין יישומים תפעוליים לבין יישומים לאספקת מידע. העובדה שזהו בסיס נתונים נפרד וייעודי מאפשרת לארגן ולעצב אותו במבנה מיוחד ותואם למטרות נגישות ישירה של המשתמשים. כלומר, שימת דגש מיוחדת על מבנה קל ונוח להבנה מצד אחד, ויעיל - לצורך קבלת זמני תגובה טובים, מצד שני. בסיס נתונים נפרד זה יפעל בשרת ייעודי כדי שלא להעמיס את המערכות התפעוליות, וכדי לאפשר זמני תגובה טובים למשתמשי מחסן הנתונים.

❖ **אוסף נתונים משולב (Integrated Data):** הנתונים של מחסן הנתונים נמצאים בבסיס נתונים משולב שבו כל הנתונים הדרושים ליישומים השונים המספקים מידע, ללא תלות במקורם. סביר להניח שבארגון יהיו מספר בסיסי נתונים שונים הפועלים על פלטפורמות חומרה ותוכנה שונות. על כן, היכולת לשלב את כולם לסביבה משולבת אחת היא בעלת חשיבות רבה. שילוב זה יבטיח אחידות מנקודת מבט המשתמש, ויאפשר לעבוד בנוחות וקלות יחסיים מול כמה בסיסי נתונים שונים. לעיתים, בסיסי הנתונים פועלים על פלטפורמות חומרה שונות ומנהלים נתונים במבני נתונים שונים ולא תואמים.

❖ **ארגון לפי נושא (Subject Oriented):** הנתונים במחסן הנתונים מאורגנים לפי נושא, ולא לפי היישום המייצר אותם. בדרך כלל, הנתונים במערכות התפעוליות מאורגנים לפי צרכי היישום, לפי דרישות הביצועים ולפי אופי התנועות של היישום. סביר להניח שנתונים הקשורים ליישום עסקית אחת, למשל לקוח, יהיו מפוזרים במספר יישומים תפעוליים שונים, שכל אחד מהם פותח בזמן שונה ולעיתים גם על פלטפורמה שונה. המשתמש הטיפוסי של מחסן הנתונים זקוק לנתונים אודות היישויות העסקיות לרוחב הארגון, כמו למשל כל הנתונים על לקוח מסוים. הדוגמה הנפוצה ביותר היא נתוני לקוח במערכת בנקאית. אלה מפוזרים ביישומים שונים (עובר ושב, חסכונות, השקעות, משכנתאות וכדומה). דרישה בסיסית של משתמש במחסן הנתונים היא להיות מסוגל לנתח את נתוני הלקוח השונים ואת פעילותו בבנק. כפי שנראה בהמשך, ארגון הנתונים לפי נושא מציב אתגר מיוחד, מכיון שהוא דורש עיצוב מיוחד וגירת הנתונים ושילובם לקראת הכנסתם למחסן הנתונים.

❖ **בעל עומק היסטורי (Time Variant):** מחסן הנתונים מאופיין בעומק היסטורי נרחב יותר מאשר המערכות התפעוליות. חלק ניכר מהניתוחים המתבצעים דורשים ניתוח מגמות, ולכן יש חשיבות מרובה לעומק ההיסטורי. ככל שנעמיק בהיסטוריה, בסיס הנתונים של מחסן הנתונים ילך ויגדל וידרוש משאבי אחסון וניהול גדולים יותר. חשוב להבין שגם במחסן הנתונים רצוי להעריך מהו העומק ההיסטורי הנדרש, כדי שלא לסבך יתר על המידה את תהליכי הניהול ולהאריך את זמני העיבוד ומשאבי המחשב הנדרשים.

❖ **לא מתעדכן (Non volatile):** להבדיל מבסיס הנתונים התפעולי (שצריך לשקף בכל זמן את התמונה העדכנית ביותר, ולכן הוא מתעדכן על ידי תנועות באופן שוטף), מחסן הנתונים אינו מתעדכן על ידי תנועות. הדרך המקובלת לעדכון מחסן הנתונים היא על ידי עדכון במנות תקופתיות (יומיות, שבועיות,

חודשיות וכדומה). עדכון מנות זה מתבצע בדרך של **טעינה מסיבית** (Bulk Load), תוך שימוש בתוכניות שירות של בסיס הנתונים, או על ידי הפעלת תוכניות עדכון מיוחדות. רצוי לא לאפשר עדכון ישיר של הנתונים במחסן הנתונים באמצעות תנועות, למעט במקרים קיצוניים מיוחדים.

❖ **תמיכה בקבלת החלטות (Decision Support):** תמיכה בקבלת החלטות היא מטרת העל של מחסן הנתונים. המאמץ הכרוך בבניית מחסן הנתונים, שיקולי העיצוב, כלי התוכנה המשמשים אותו והארגון התומך והמזין אותו - כולם נועדו למימוש הגישה הקלה והנוחה של משתמשים לסיוע בתהליכי קבלת ההחלטות. מחסן הנתונים מבוסס על אוסף כלי תוכנה מתוחכמים המיועדים להצגה ולניתוח נתונים. כלים אלה מאפשרים גישה ישירה לבסיס הנתונים או בנייה של יישומים מותאמים לתמיכה בתהליכי קבלת ההחלטות.

למרות שהגדרה זו ממצה למדי, חשוב שנדגיש גם מה חסר בה: **המימד הדינמי**. ההגדרה ממוקדת בעיקר בהיבט הסטטי של מחסן הנתונים ומתעלמת מההיבט הדינמי, שהוא אוסף התהליכים והפעילויות הקשורים בפיתוח, תפעול, ניהול, תמיכה ותחזוקת מחסן הנתונים. על רקע זה מצאתי לנכון להביא הגדרה נוספת למחסן הנתונים, המתייחסת אל מחסן הנתונים כאל תהליך, ולא כאל מקום:

מחסן הנתונים הוא תהליך **קצה לקצה** (End-to-End) המארגן נתונים שמקורם במספר יישומים תפעוליים שונים, בהתאם למימד הזמן ובהתאם לנושאים בעלי משמעות למקבלי ההחלטות.

כפי שנראה בהמשך, חלק ניכר מהאנרגיה שמשקיע הארגון בהקמת ובתפעול מחסן הנתונים מושקעת בעבודה היומיומית להכנסת הנתונים למחסן הנתונים ובתפעול השוטף. לכן, מן הראוי שהגדרת מחסן הנתונים תדגיש גם את ההיבט התהליכי, ולא רק את ההיבט הסטטי.

בפרק הקודם סקרנו את ההבדלים העיקריים בין המערכות התפעוליות לבין מחסן הנתונים. להבדיל ממערכות תפעוליות, שהן בדרך כלל בעלות רמת יציבות גבוהה יחסית, מחסן הנתונים מתפתח כל הזמן ומתאים את עצמו באופן שוטף לדרישות המידע המשתנות של הארגון. משמעות הדבר היא שעיצוב מחסן הנתונים חייב לקחת בחשבון את האופי הדינמי - הדרישות של היום בוודאי שלא תהיינה הדרישות של מחר. הבעיה היא שלא ניתן אפילו לחזות מה תהיינה הדרישות העתידיות. ובכל זאת, כפי שנראה בהמשך, ניתן לעצב את מחסן הנתונים כך שתהיה בו מידה רבה של יכולת הסתגלות לשינויים ולדרישות בלתי צפויות. עובדה זו באה לעיתים כהלם למעצבי מערכות, שכל חינוכם בא מתחום המערכות התפעוליות: הם למדו שיש להקדיש זמן ניכר לניתוח הבעיה לפני בניית המערכת, מכיון ששינוי מערכת תפעולית הוא תהליך מורכב ויקר. כאשר הם עוסקים בהקמת מחסני נתונים, הם מנסים להשקיע זמן רב מדי בניתוח ובניסיון לחיזוי הצרכים העתידיים ולעיתים - עד כדי שיתוק (Paralysis by Analysis). לבסוף, כשהם לומדים ומבינים את הדרישות של היום ומנסים ליישם אותן, הדרישות משתנות ולכן מתפתחת אצלם תחושה של מרדף בלתי פוסק אחר דבר שאין להשיגו. כפי שנראה בהמשך, חלק מהתשובה לתכונה זו של מחסני הנתונים טמונה במתודולוגיית הבנייה, העיצוב והיישום.

## ארכיטקטורת מחסן הנתונים (Data Warehouse Architecture)

כל מי שנחשף לנושא מחסני נתונים יוצא מבולבל מהאוסף הגדול של השמות השונים: מחסן נתונים ארגוני, מחסן נתונים מחלקתי, מרכז נתונים, מאגר

נתונים תפעולי וכדומה. ננסה לעשות סדר באוסף מושגים זה, ונראה שההבדלים ביניהם נובעים מהעושר הרב של ארכיטקטורות ושל היעדים של מחסני הנתונים.

אם נתבונן במחסני הנתונים שהוקמו על ידי הארגונים השונים, סביר להניח שלא נמצא שניים זהים לחלוטין. הסיבה לכך היא שכל מחסן נתונים מותאם לסביבת המיחשוב המיוחדת של הארגון, לסביבת החומרה ולכלי התוכנה בהם בחר הארגון להשתמש, ולמיגוון הייחודי של היישומים התפעוליים הפועלים בו. עם זאת, ניתן לומר שקיימים מספר קווי דמיון בולטים בין רוב מחסני הנתונים ורובם כוללים, בצורה זו או אחרת, מספר מרכיבים בסיסיים.

מקובל להתייחס למספר ארכיטקטורות עקרוניות של מחסני נתונים, שההבדלים ביניהם (בין מחסני הנתונים) נובעים מיעדיהם ומקהל המשתמשים שהם צריכים לשרת:

❖ מחסן נתונים כלל ארגוני (Enterprise Data Warehouse)

❖ מרכז נתונים (Data Mart)

❖ מחסן נתונים רב-שכבתי (Multi Tiered Data Warehouse)

❖ מאגר נתונים תפעולי (Operational Data Store)

❖ מחסן נתונים מדומה (Virtual Data Warehouse)

נסקור כל אחת מארכיטקטורות אלו בפירוט.

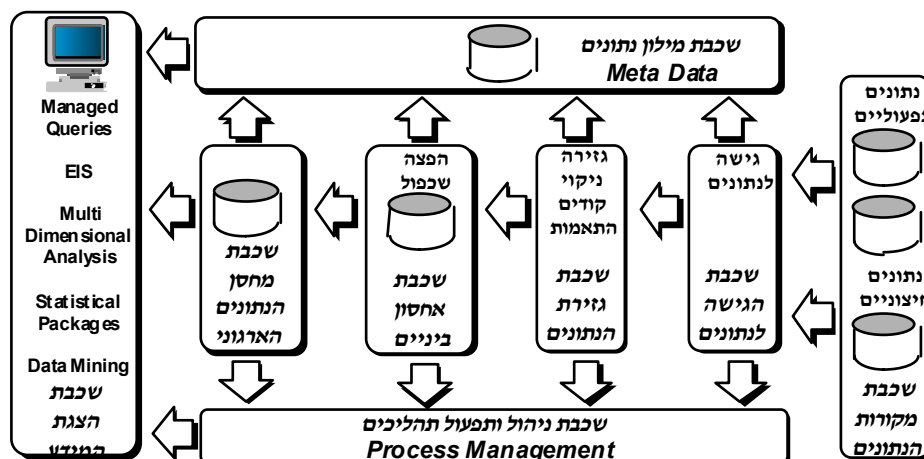
## ארכיטקטורת מחסן נתונים ארגוני (Enterprise Data Warehouse Architecture)

ארכיטקטורת מחסן הנתונים ארגוני היא המקיפה ביותר מבין ארכיטקטורות מחסני הנתונים. נתחיל בהגדרתה.

**הגדרה:** מחסן הנתונים הארגוני הוא מחסן נתונים מרכזי המיועד לשרת את כל המשתמשים בארגון, ולכן הוא משלב בבסיס נתונים אחד את כל הנתונים הנדרשים עבור כל יישומי התמיכה בקבלת החלטות.

באופן טבעי, מחסן נתונים כזה הוא גדול, מקיף ומשרת מספר רב של משתמשים, ולכן תהליך הקמתו מורכב וארוך יחסית. בדרך כלל, יישום מחסן נתונים ארגוני מתבצע בשלבים ולא בבת אחת, ובסופו של תהליך ההקמה והיישום יעמוד לרשות הארגון מחסן נתונים מרכזי אחד.

תרשים 1 מציג את הארכיטקטורה של מחסן הנתונים הארגוני.



תרשים 1: ארכיטקטורה של מחסן נתונים ארגוני

כפי שניתן לראות מתרשים XX.XX, הארכיטקטורה הכללית של מחסן הנתונים הארגוני מורכבת ממספר שכבות. נציג אותן בשלב זה ובהמשך הפרק נקדיש לחלק מהן הסבר מעמיק יותר.

❖ **שכבת מקורות הנתונים (Data Source Layer):** את שכבת מקורות הנתונים ניתן לחלק לשתי קטגוריות עיקריות - מקורות נתונים פנימיים ומקורות נתונים חיצוניים. המקורות הפנימיים הם אוסף כל הנתונים שמנהל הארגון במסגרת היישומים התפעוליים; המקורות החיצוניים הם אוסף הנתונים הנדרשים לתהליכי קבלת החלטות ונאספים מחוץ לארגון. הנתונים החיצוניים עשויים לכלול סטטיסטיקות המתפרסמות על ידי גופים רשמיים (כגון הלשכה המרכזית לסטטיסטיקה או בנק ישראל), מאגרי מידע של חברות המנתחות שווקים שונים (כגון Dun&Breadstreet), נתונים דמוגרפיים ועוד. מכיון שאלה הם נתונים חיצוניים לארגון, הם אינם מנוהלים במסגרת המערכות התפעוליות שלו. מקורות הנתונים יכולים להיות מגוונים מאוד. למרות העושר של המקורות החיצוניים, מקור הנתונים העיקרי למחסן הנתונים, אם כי לא הבלעדי, הוא הקבצים ובסיסי הנתונים המשמשים את מערכות המידע התפעוליות. הבעיה העיקרית היא שמערכות אלו עשויות להיות ישנות מאוד, ולהימצא על פלטפורמות מיחשוב מסוגים שונים, שאינם בהכרח תואמים.

❖ **שכבת הגישה לנתונים (Data Access Layer):** שכבה זו עוסקת בגישה למקורות הנתונים, הן במערכות התפעוליות והן לנתונים החיצוניים. מכיון שמקורות הנתונים יכולים להיות מגוונים מאוד, קיים לעיתים קושי לגשת אליהם לצורכי גיר הנתונים. בשנים האחרונות התפתחו כלי גישה שונים לנתונים, כגון EDA/SQL של חברת IBI, מוצרים מבוססי תקן ODBC וכדומה. מוצרים אלה מאפשרים גישה שקופה למיגוון המקורות המבוססים על מבני נתונים שונים (קבצי VSAM; בסיסי נתונים המבוססים על IMS או IDMS; בסיסי נתונים טבלאיים המבוססים על Oracle, Informix, Sybase, SQL Server ואחרים). מעבר לתמיכתם במבני נתונים שונים, תומכים כלים אלה בפרוטוקולי תקשורת רבים, במערכות הפעלה וחומרות שונות.

❖ **שכבת גזירה, מיטוב והמרה של הנתונים (Data Transformation and Enhancement Layer):** שכבה זו עוסקת בתהליך הגזירה וההכנה של הנתונים שצריך לאחסן במחסן הנתונים. מכיון שהנתונים המאוחסנים בבסיסי הנתונים התפעוליים אינם מתאימים למטרות תמיכה במידע לקבלת החלטות, יש לבצע בהם מספר פעולות מקדימות. בדרך כלל, בסיום שלב זה נמצאים הנתונים במבנה מתאים לטעינה למחסן הנתונים.

❖ **שכבת אחסון ביניים והעברה (Staging Data Layer):** שכבה זו מאחסנת לתקופת ביניים את הנתונים המוכנים לטעינה. לעיתים נוח לטפל בנתונים מראש ולצבור אותם לפני טעינתם. סיבה עיקרית לכך היא נפחי הנתונים ו/או ניצול חלון זמן מתאים. הנתונים נשמרים בשכבת ביניים זו עד להכנסתם למחסן הנתונים והיא גם יכולה להפיץ אותם לאחר מכן. משכבה זו מועברים הנתונים אל מחסן הנתונים, לשם טעינתם לבסיס הנתונים. שכבת פעילות זו אינה דרושה בכל מימוש של מחסן נתונים.

❖ **שכבת בסיס הנתונים (Data Base Layer):** שכבה זו מנהלת את בסיס הנתונים המרכזי של מחסן הנתונים. ברוב יישומי מחסן נתונים מפעילים בסיס נתונים טבלאי. הנתונים נטענים אל בסיס הנתונים באמצעות תוכניות שרות או יישומים מיוחדים המעדכנים אותו. בדרך כלל, עם תום שלב הטעינה נבנים האינדקסים, מחושבים סיכומים שונים ועוד פעולות ניהול, כנדרש.

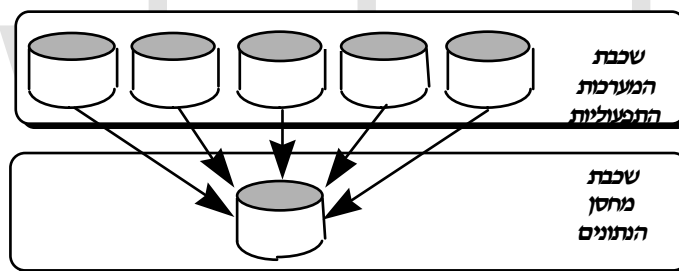
❖ **שכבת הצגת המידע (Information Presentation Layer):** שכבה זו עוסקת בהצגת המידע המנוהל במחסן הנתונים, וכוללת אוסף של כלי תוכנה ויישומים המיועדים למשתמשי מחסן הנתונים. בדרך כלל, אין בנמצא כלי אחד ויחיד המתאים לצורכי כל המשתמשים. על כן, מקובל לראות בשכבה זו מספר כלים שונים המיועדים לצרכים שונים: מחוללי שאילתות ודוחות, כלים

לניתוח רב-מימדי, כלים לניתוחים סטטיסטיים, כלים לבניית מודלים פיננסיים, גיליונות אלקטרוניים, כלי כריית נתונים (Data Mining) וכדומה. בחלק מהמקרים המשתמשים עובדים באופן ישיר עם כלים אלה ובמקרים אחרים נבנים עבורם יישומים ייחודיים. הכל בהתאם ליכולות המשתמש וצרכי המידע שלו. קיימים מצבים בהם המשתמשים זקוקים לאותו מבנה מידע באופן תקופתי. לכן, מקובל להריץ אוסף דוחות מול מחסן הנתונים ולהעמיד את התוצאות לרשות אותם משתמשים. כך, המשתמשים אינם צריכים לנסח את המבנה מחדש בכל פעם ואינם צריכים לזכור כיצד להפעילו. טכנולוגיית האינטרנט הפכה לטכנולוגיה נפוצה מאוד לצורך הפצת המידע המופק ממחסן הנתונים למשתמשים השונים.

❖ **שכבת קטלוג הנתונים (Meta Data Layer):** שכבה זו משמשת כקטלוג מרכזי המאפשר למשתמש להתמצא בנבכי מחסן הנתונים, במשמעות הנתונים ובעדכניותם. קטלוג הנתונים מכיל את ההסבר לכל הטבלאות, את ההסבר לעמודות שבכל טבלה, כיצד חושבו הנתונים, מה המקור ממנו נגזרו או חושבו הנתונים, למתי הם מעודכנים וכדומה. ללא שכבה זו יש למשתמשים קושי ניכר בהתמצאות במחסן הנתונים ובהבנת משמעות הנתונים האגורים בו.

❖ **שכבת ניהול ותפעול תהליכים (Process Management Layer):** שכבה זו עוסקת בהפעלת התהליכים השונים הפועלים בסביבת מחסן הנתונים, בתזמון שלהם, בבדיקה שהסתיימו בהצלחה וכדומה.

מכיון שמבחינת ניהול הנתונים בארגון קיימת אבחנה ברורה בין שתי שכבות נתונים - נתונים תפעוליים ונתונים לתמיכה בקבלת החלטות - מקובל גם לקרוא לארכיטקטורה זו, מבחינת הנתונים, **ארכיטקטורה דו-שכבתית (Two Tier Architecture)**.



תרשים 2: שתי שכבות של ניהול נתונים

היתרון הגדול בגישת מחסן הנתונים הארגוני היא, כמובן, מעצם היותו מאגר מרכזי המנהל נתונים אודות כלל היישויות העסקיות המעניינות את הארגון והמשתמש את כלל המשתמשים. מכיון שזהו בסיס נתונים מרכזי, ניתן לקבל ממנו חתכי מידע רוחביים חשובים, מעניינים ומגוונים.

הבעיה העיקרית בגישת מחסן הנתונים הארגוני הוא הזמן הארוך הדרוש להקמתו, תהליך שיכול להימשך לפעמים מספר שנים. הניסיון להקים מחסן נתונים ארגוני דורש להגיע להסכמה בין מספר רב של משתמשים ומחלקות בארגון באשר להגדרת ומשמעות הנתונים, ובאשר לעדיפויות הנושאים השונים שיש לנהל בו. הארגון צריך להשקיע מאמץ זמן רב בשלב העיצוב והבניה של מודל הנתונים. לעיתים מגיעים לסתירות בין הדרכים השונות בהן משתמשים שונים רוצים לנהל ולהסתכל בנתונים. לרוב הארגונים והמשתמשים אין את הזמן ואת הסבלנות הנדרשים להקמת מחסן נתונים ארגוני, כי הלחצים לקבלת מידע גבוהים ואין אפשרות להשקיע את המשאבים. בנוסף, התנאים והדרישות משתנים כל הזמן, נפח הנתונים של המחסן הארגוני עצומים וכל זה - מרפה את ידי המתחילים לעסוק בנושא. לפעמים מגיעים למסקנה שאין זה מעשי לבנות מחסן

נתונים ארגוני עבור ארגונים גדולים שבהם יש שונות גדולה של מוצרים ושירותים.

בגלל קשיים אלה נטשו, לפחות באופן זמני, חלק מהארגונים את הגישה הזו והגדירו מטרה צנועה יותר: מחסן נתונים מחלקתי, או מחסן נתונים נושאי ובפשטות - מרכז נתונים.

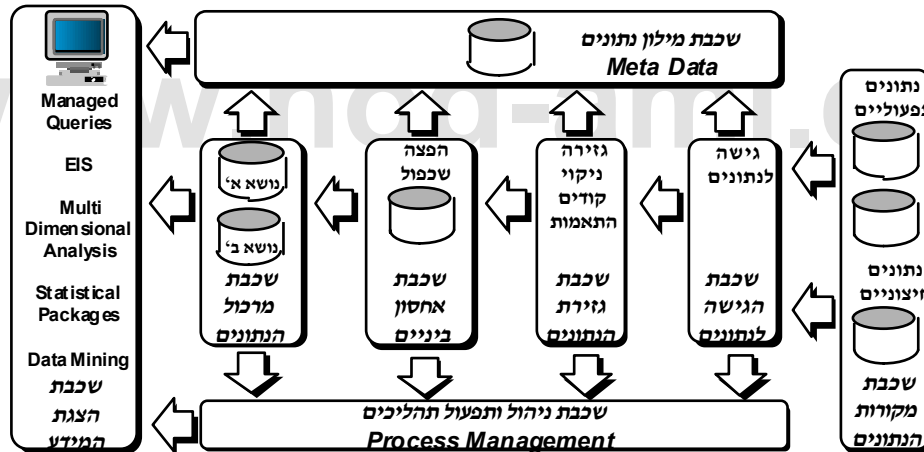
## ארכיטקטורת מרכז הנתונים (Data Mart Architecture)

מרכז הנתונים הוא הרחבה של תפיסת מחסן הנתונים הארגוני לרמה המחלקתית, ולכן הוא ממוקד בנושא או בתחום מסוים ומוגדר. באופן טבעי, הוא קטן יותר ממחסן הנתונים הארגוני.

**הגדרה:** מרכז הנתונים הוא מחסן נתונים שנועד לשרת, בדרך כלל, מחלקה אחת (או מספר מצומצם של מחלקות) וממוקד בנושא אחד בלבד.

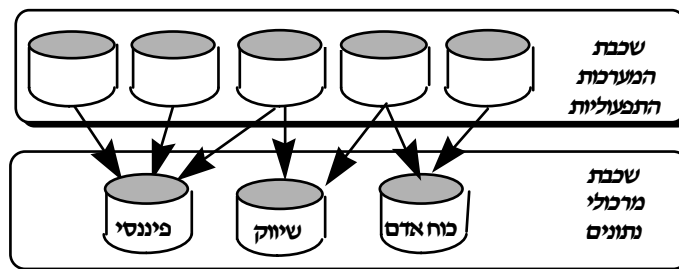
מבחינת תכולת הנתונים מרכז הנתונים הוא נגזרת חלקית כלשהי של המחסן הארגוני. הוא מיועד לשימוש מחלקה אחת או מספר מצומצם של מחלקות קשורות ולרוב הוא ממוקד בנושא אחד. אחד מפרקי הספר מוקדש לנושא זה.

ארכיטקטורת מרכז הנתונים דומה לארכיטקטורת מחסן הנתונים הארגוני, בהבדל אחד מהותי: בעוד שמחסן נתונים ארגוני יש רק אחד, מרכזי נתונים יכולים להיות רבים.



תרשים 3: ארכיטקטורה מרכז נתונים

כפי שניתן לראות מתרשים 4, כל השכבות שקיימות במחסן הארגוני קיימות גם כאן. מכיון שגם בארכיטקטורה זו קיימת אבחנה בין שכבת הנתונים התפעוליים לבין שכבת הנתונים המיועדים לתמיכה בקבלת החלטות, ארכיטקטורת מרכז הנתונים היא ארכיטקטורה דו-שכבתית מבחינת הנתונים.



תרשים 4: שכבות בארכיטקטורה של מרכזי הנתונים

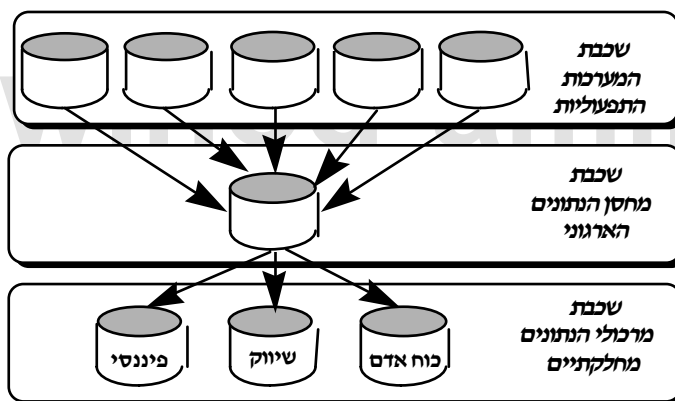
## מחסן נתונים רב-שכבתי ( Multi Tier Data Warehouse )

שתי הארכיטקטורות שהצגנו - המחסן הארגוני ומרכז הנתונים - אינן מיועדות להחליף זו את זו. הן ארכיטקטורות משלימות ומשולבות.

**הגדרה:** מחסן נתונים רב-שכבתי משלב בתוך ארכיטקטורה אחת מחסן נתונים ארגוני יחד עם מספר בלתי מוגבל של מרכזי נתונים.

להבדיל מהארכיטקטורה הדו-שכבתית של מרכזי הנתונים השואבים את הנתונים באופן ישיר מהמערכות התפעוליות, בארכיטקטורה הרב-שכבתית שואבים מרכזי הנתונים את הנתונים ממחסן הנתונים הארגוני. כל מרכז נתונים מכיל חלק מסוים מבסיס הנתונים של המחסן הארגוני. תהליכי הגירה ליצירת המרכז מתבססים, בדרך כלל, על טכנולוגיות **שכפול נתונים** (Data Replication), הנתמכות כיום על ידי רוב בסיסי הנתונים היחסיים המודרניים.

ארכיטקטורת מרכזי הנתונים היא המומלצת ביותר. מצד אחד היא מבוססת על מחסן נתונים ארגוני ומצד שני היא נותנת פתרונות למחלקות ולנושאים ומספקת כלים מיוחדים כגון **כלי ניתוח רב מימדיים** (OLAP) או **כלי כריית נתונים** (Data Mining). חלק מכלים אלה אינם פועלים באופן ישיר על בסיס הנתונים הטבלאי. עבור כלי OLAP דרוש מבנה קוביה רב-ממדית, בשעה שכלי כרייה אחרים יכולים להסתפק במבנה של קובץ שטוח.

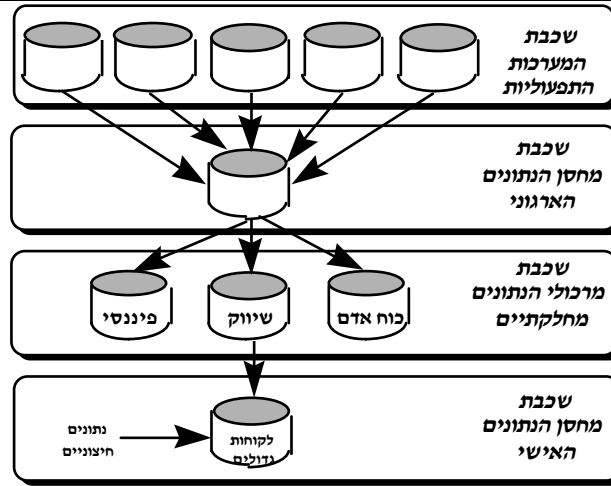


תרשים 5: ארכיטקטורה מחסן נתונים רב-שכבתי

ארכיטקטורת מחסן הנתונים הרב-שכבתי משלבת את היתרונות שבקיום מחסן נתונים כלל ארגוני ואת היתרונות שברמת עצמאות והתאמה לדרישות ייחודיות של מחלקות שונות. הבעיה היא בקושי שבבניית המחסן הארגוני ובתפעול שוטף של ארכיטקטורה מבוצרת ומורכבת יחסית.

לפעמים ניתן לבנות גם מחסן נתונים המבוסס על שכבה נוספת: שכבת הנתונים המנוהלים במחשב האישי של משתמש מסוים. בארגון קיימים מספר מועט של משתמשים המבצעים ניתוחי מידע מורכבים ומאוד ייחודיים ומסיבות שונות, כגון סודיות או שימוש בכלי ניתוח מאוד ייחודיים, עדיף לגזור את הנתונים ממרכז מחלקתי (או לפעמים ישירות מהמחסן הארגוני) ולהעבירם לתחנת העבודה האישית.





תרשים 6: מחסן נתונים ארבע שכבתי

תרשים 6 מציג דוגמה של שכבה נוספת המופעלת במחשב אישי של משתמש מסוים. שכבה זו מקבלת נתונים ממרכז הנתונים השיווקי על ידי גירת הלקוחות הגדולים בלבד וביצוע ניתוחים מיוחדים המשלבים גם נתונים חיצוניים.

## ארכיטקטורת מאגר נתונים תפעולי (ODS - Operational Data Store)

בנוסף לדרישות ניתוח מידע עבור דרג מקבלי ההחלטות, הארגון צריך לספק מידע באופן שוטף גם לדרג התפעולי. אחת האפשרויות היא לספק מידע זה ישירות מתוך המערכות התפעוליות. אפשרות אחרת היא להקים ליד המערכת התפעולית מאגר נתונים ייעודי לאספקת מידע לדרגים התפעוליים. ברמת הארכיטקטורה קיים דמיון רב בין מחסן הנתונים לבין מאגר הנתונים התפעולי, אולם כפי שנראה מיד, האופי, עדכניות הנתונים המנוהלים בו וכלי הדיווח - שונים.

**הגדרה:** מאגר נתונים תפעולי הוא מחסן נתונים המיועד לשרת את הדרג התפעולי בארגון בכל הקשור לדיווחים ולניתוחי מידע שוטפים.

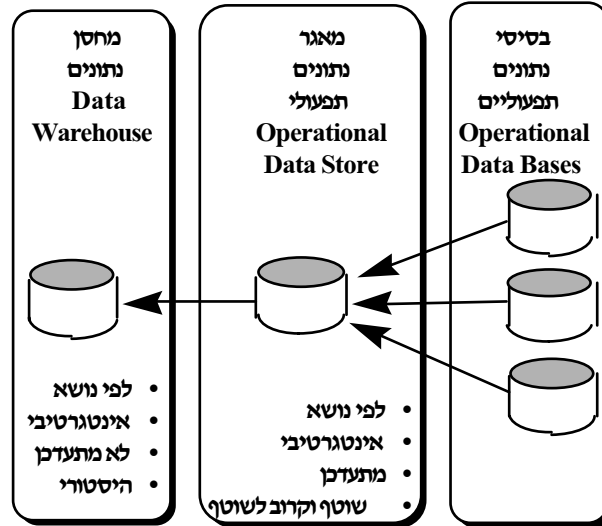
מאגר הנתונים התפעולי מיועד להפקת דיווחים וניתוחי מידע עבור הדרג התפעולי, ואולי גם טקטי, אבל עם דגש חזק על הפעילות השוטפת בארגון. נסקור את נקודות הדמיון והשוני בין מחסן הנתונים לבין מאגר הנתונים התפעולי.

❖ **ארגון לפי נושאים** - בדומה למחסן הנתונים, גם במאגר הנתונים התפעולי הנתונים מאורגנים לפי נושאים, ולא לפי תהליכים עסקיים, כמקובל במערכות תפעוליות.

❖ **מאגר אינטגרטיבי** - המאגר אוגר נתונים ממערכות תפעוליות שונות ופותר סתירות שונות.

❖ **מתעדכן** - להבדיל ממחסן נתונים המתעדכן רק דרך תהליכי טעינה, מקובל לעדכן את מאגר הנתונים התפעולי גם באמצעות תנועות המערכת התפעולית, במידת הצורך. מכיון שמאגר הנתונים התפעולי צריך להיות ברמת עדכניות גבוהה, לא תמיד תהליכי טעינה באצווה מספקים את דרישות הדרג התפעולי. יחד עם זאת, רצוי לבנות מאגר נתונים תפעולי המתעדכן באופן דומה למחסן הנתונים, במידה ורמת העדכניות מאפשרת.

❖ **שוטף וקרוב לשוטף** - להבדיל ממחסן הנתונים המנהל נתונים מפורטים לפרקי זמן ארוכים, העומק ההיסטורי של מאגר קצר יותר, כי הוא משרת בעיקר את הדרג התפעולי. בדרך כלל מנוהלים בו הנתונים השוטפים והקרובים לשוטף (Near Current).



תרשים 7: ארכיטקטורה של מאגר נתונים תפעולי ליד מחסן נתונים

כפי שניתן לראות מתרשים 7, מאגר הנתונים התפעולי יכול לשמש כמאגר ביניים לפני העברת הנתונים אל מחסן הנתונים, אם כי זה איננו תנאי הכרחי. יתרון השימוש במאגר הנתונים התפעולי כמאגר ביניים הוא בכך שתהליכי הגירה ומיטוב הנתונים מתבצעים בעת הכנסת הנתונים אליו, ולכן אין צורך לחזור על פעולה זו בעת העברתם למחסן הנתונים.

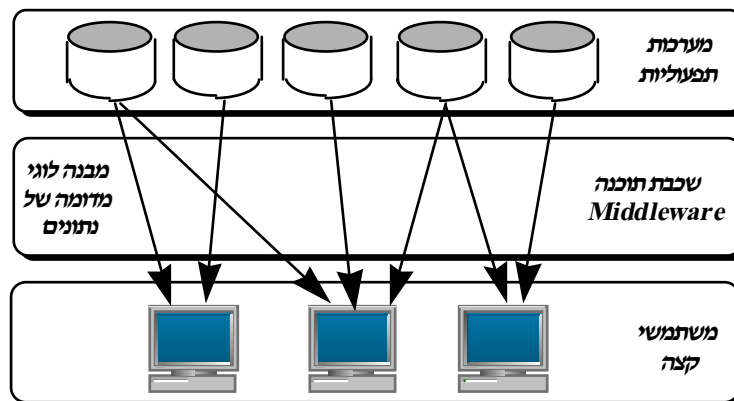
להבדיל ממחסן הנתונים, שצריך להכיל כלי ניתוח מידע מתוחכמים מאוד המיועדים לשימושם של משתמשים מתמחים, מאגר הנתונים התפעולי מכיל בדרך כלל מחוללי דוחות ושאליות, מכיון שדרישות ניתוח המידע צנועות בהרבה.

## מחסן נתונים מדומה (Virtual Data Warehouse)

למרות ההבדלים בין מערכות תפעוליות לבין מערכות לאספקת מידע, עדיין יש ארגונים המנסים ליישם תפיסה של מחסן נתונים מדומה.

**הגדרה:** מחסן נתונים מדומה הוא אוסף של כלי גישה ותוכנות קישור, המאפשרים גישה נוחה לנתונים תפעוליים מתוך כלי הצגת וניתוח מידע פשוטים יחסית.

הרעיון הבסיסי עליו מבוסס **מחסן הנתונים המדומה** הוא לאפשר גישה נוחה וקלה של המשתמשים לבסיסי הנתונים התפעוליים בארגון.



תרשים 8: ארכיטקטורה של מחסן מדומה.

בארכיטקטורה זו ניתן דגש רב על בניית **שכבת ביניים** (Middleware) המאפשרת קישור בין אוסף כלי קצה לבסיסי הנתונים התפעוליים בארגון. שכבת ביניים זו, הפועלת לפעמים על שרת ייעודי, מבוססת על תקנים כגון ODBC, DRDA, או על

תוכנות ביניים מיוחדות המאפשרות גישה לבסיסי נתונים לא טבלאיים כמו IMS, IDMS, על קבצי VSAM ואחרים.

כפי שניתן להניח, ארכיטקטורה זו מפתה מאוד, מפני שאין צורך בגירת נתונים וניהול בסיסי נתונים נפרדים עבור מערכות אספקת המידע. מתקינים אוסף של כלי קצה ותוכנות קישור מסוימות, והרי לנו מחסן נתונים. איך בכלל ניתן לעמוד בפני פיתוי שסחה? הבעיה היא שתפיסה זו פשוט "לא עובדת", ויש לכך סיבות רבות: הבדלים במבנה הנתונים, סתירות בין הנתונים, עומס על מחשבי הייצור, חוסר יכולת לבצע אינטגרציה בין נתונים המנוהלים במערכות מיחשוב שונות ורמת הסיבוכיות שהמשתמש נחשף לה. כל אלו הן רק חלק מהבעיות שארגון המנסה לאמץ תפיסה זו עלול להיתקל בהן.

ניתן לומר שארכיטקטורה המוצגת כאן מתאימה כפתרון ביניים בלבד לארגונים הנתקלים בקשיים בבניית מחסן נתונים אמיתי והמבקשים להקל, לפחות באופן זמני, על חלק מהמשתמשים על ידי מתן אפשרות קלה יותר של גישה לנתונים. "ארכיטקטורה" פשטנית זו קיימת רק בעלוני השיווק של ספקים מסוימים, המבקשים למכור את האשליה שניתן לבנות מחסן נתונים במהירות ובקלות.

זהו חלק מפרק 2 מתוך הספר

## **מחסני נתונים - עקרונות, ארכיטקטורות, עיצוב ויישום**

בהוצאת הוד-עמי.

www.hod-ami.co.il