



Top Ten Steps to Curate Your Data for A.I. Assistants July 2024



1. Why is good data important for A.I.?

2. Top 10 Steps to Curate Your Data for AI Co-Pilots and Assistants

1. Define the Problem and the Data Requirements
2. Data Collection
3. Data Cleaning
4. Data Transformation and Enrichment
5. Data Structuring
6. Data Annotation
7. Data Splitting
8. Data Validation
9. Data Storage and Accessibility
10. Continuous Monitoring and Updates

3. A brief AI Assistant demo in eyko



Why is good data important for A.I. Assistants?

1. What is an AI Assistant or Copilot?

1. An automated chat like interface or assistant embedded into an application that provides human like answers to human generated questions.
2. Use cases: Chat bots on websites, chat bots for support, more recently A.I. Assistants to generate results from human generated natural language queries.
 1. Examples:
 - Support: How do I perform a task?
 - Analytics: Who are my top sellers last quarter? What customers at risk of late payments or not renewing (Churn)?

2. What is good data for A.I.?

- Depends on the use case and the user – to be discussed in the next slides

3. Consequences of bad data?

- **User Frustration:** Repeatedly receiving unhelpful responses can frustrate users.
- **Increased Workload:** Poor performance of the AI copilot means more queries will need to be handled manually, increasing the workload for users/humans.
- **Reputational Damage:** Persistent issues with AI-generated responses can damage the application's reputation.
- **Lack of Adoption:** AI Assistants could fail to be adopted.



1. Define the problem and the data requirements

Different use cases require different data sets.

- Massive difference in data set requirements between these two analytic questions.
 - Q1: Show me the top 10 sellers from last quarter
 - Q2: Show me what customers are at risk of churn

Data Requirements

a. Identify Objectives

- Determine the specific tasks the AI copilot will perform.
- Set clear objectives for the data to meet these needs.

b. Specify Data Needs

- Decide the types of data required (e.g., what data sources structured and unstructured will you require).
- Determine data sources (e.g., databases, web scraping, SaaS, Cloud, files, APIs).

- **For example:**

- Churn: sales, payment, customer support/success data, sentiment data, forecast data, historical adoption, information about their business such as quarterly reports



1. Define the problem and the data requirements

For example:

- **Churn:** sales, payment, customer support/success data, sentiment data, forecast data, historical adoption, information about their business such as quarterly reports

eyko | Customer 360

Summary ERP Salesforce Hubspot Jira Shopify Toggl

Customer	Total Billed	Open Amount	Campaigns	Last Activity	Open Cases	Risk	Online Orders	Online Spend	Billable Time
Action Bike Specialists	\$128,281.25	\$212,150.00	3	3/31/2022	1	Medium Risk	2	\$2,533.05	\$150.00
Engineered Bike Systems	\$668,160.00	\$176,980.00	3	12/14/2021	1	Medium Risk	3	\$3,451.88	\$78.75
Fitness Cycling	\$142,757.50	\$323,200.00			2	High Risk	1	\$2,383.67	\$130.00
Journey Sporting Goods	\$318,256.00	\$100,000.00	4	8/28/2021	4	High Risk	3	\$1,775.30	
Major Sports Suppliers	\$1,748,175.00	\$225,000.00	3	5/3/2022		Low Risk	2	\$4,527.02	\$220.00
Novelty Bikes	\$293,203.75	\$345,340.00	4	9/15/2021	2	High Risk	1	\$2,677.02	\$100.00
Professional Cycle Store	\$137,256.25	\$377,981.00				Medium Risk	1	\$2,533.05	\$1,400.00
Total	\$3,994,489.75	\$2,792,651.00	17		12		23	\$34,358.12	\$3,018.75

ERP Salesforce Hubspot Jira Shopify Toggl

All All All All All All All All All

Invoice ID	Customer	Invoice Date	Amount	Tax	Total	Due Date	Paid	Paid Date
INV-0001	Journey Sporting Goods	1/1/2021	\$13,250.00	\$1,612.50	\$14,862.50	3/2/2021	\$14,862.50	3/28/2021
INV-0002	The Bike Shop	1/5/2021	\$102,475.00	\$11,308.75	\$113,783.75	2/4/2021	\$113,783.75	1/18/2021
INV-0003	Journey Sporting Goods	1/15/2021	\$96,850.00	\$9,977.50	\$106,827.50	3/16/2021	\$106,827.50	4/5/2021
INV-0004	Engineered Bike Systems	2/1/2021	\$63,900.00	\$9,585.00	\$73,485.00	2/16/2021	\$73,485.00	2/13/2021
INV-0005	Fitness Cycling	2/1/2021	\$23,500.00	\$3,525.00	\$27,025.00	3/3/2021	\$27,025.00	2/25/2021
INV-0006	Journey Sporting Goods	2/1/2021	\$45,540.00	\$4,781.00	\$50,321.00	4/2/2021	\$50,321.00	5/27/2021
INV-0007	Major Sports Suppliers	2/1/2021	\$507,000.00	\$50,700.00	\$557,700.00	3/3/2021	\$557,700.00	2/11/2021
Total			\$3,600,665.00	\$393,824.75	\$3,994,489.75		\$3,644,331.00	



eyko – Churn example with AI Assistant

View
eyko 360



What do you want to know?

what customers are...

Conversation (Text-Based Q&A)

1. what customers are at ri...

Prompt

Enter...



what customers are at risk of churn

Based on the provided data, customers at risk of churn can be identified by looking at factors such as high "Open Amount", "Open Cases", and "Risk" levels. Here are the customers at high or medium risk:

High Risk Customers:

1. Fitness Cycling

- Open Amount: \$323,200.00
- Open Cases: 2

2. Journey Sporting Goods

- Open Amount: \$100,000.00
- Open Cases: 4

3. Novelty Bikes

- Open Amount: \$345,340.00
- Open Cases: 2

Medium Risk Customers:

1. Action Bike Specialists

- Open Amount: \$212,150.00
- Open Cases: 1

2. Engineered Bike Systems

- Open Amount: \$176,980.00
- Open Cases: 1

3. Professional Cycle Store

- Open Amount: \$377,981.00
- Open Cases: 0

Summary:

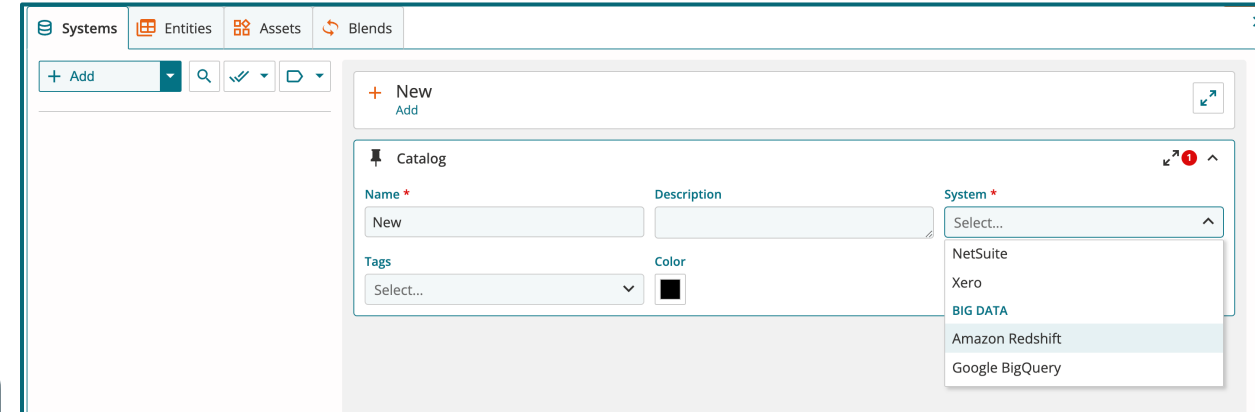
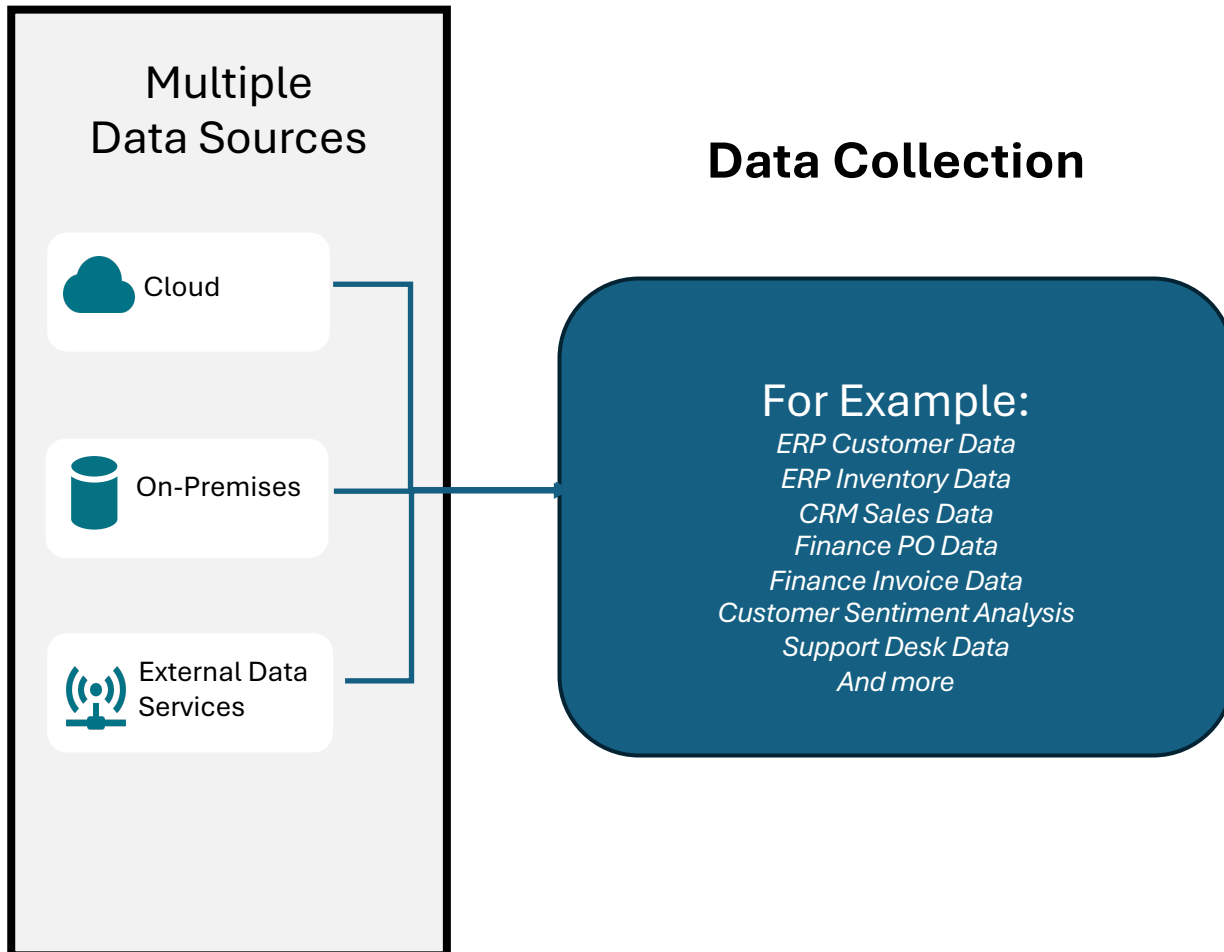
- High-Risk Customers** have significant open amounts and multiple open cases, indicating dissatisfaction or unresolved issues.
- Medium-Risk Customers** also have noticeable open amounts but fewer open cases, indicating potential but less severe risks.

These customers should be prioritized for follow-up to address their issues and reduce the risk of churn.



2. Data Collection

Once you've defined your data requirements you need to collect your data.



a. Gather Data

- Collect data from specified sources.
- Ensure data variety to cover different scenarios the AI might encounter.

b. Ensure Data Quality

- Check data for completeness.
- Ensure data accuracy by cross-referencing.



3. Data Cleansing










Prepare and clean data for blending, transformation, and analysis

a. Remove Inconsistencies

- Identify and correct errors (e.g., typos, incorrect values).
- Remove duplicate records.

b. Handle Missing Values

- Decide on a strategy (e.g., imputation, removal) for dealing with missing data.
- Implement the chosen strategy to fill or remove gaps.

 Spelling Fix spelling mistakes in data instantly.	 Text Case Standardize any text into Sentence or Title cases.	 Phone Numbers Format phone numbers as national or international .
 States and Provinces Normalize States & Provinces as abbreviations or full names .	 Postal Codes Fix postal codes. For the US, choose 5 or 9-digit format.	 Countries Standardize countries into 2 or 3-letter codes or full names .
 E-Mail Addresses Fix and validate e-mail addresses.	 Web Addresses Standardize web addresses into a short or full format.	 Custom Use Masks to standardize codes, IDs, and more...



4. Data Transformation and Enrichment

Transform and enrich your sources into business ready data

a. Normalize Data

- Convert data to a consistent format (e.g., date formats, measurement units).
- Scale numerical values if necessary for uniformity.

b. Data Enrichment

- Enrich data with new calculations, classifications, hierarchies, time intelligence, and more to enhance model performance.
- Select relevant enrichment to improve efficiency.

Available Smarts:	Name	Applies To	Packs	Description
<input type="checkbox"/>	Data Smart			
<input type="checkbox"/>	Blends	Entities		Automatically generate relationships between the selected entities and determine the relationship type.
<input type="checkbox"/>	Names	Fields and Entities		Use Natural Language Processing and eyko dictionaries to rename entities and fields from technical terms to user-friendly names.
<input type="checkbox"/>	Flows	Fields and Entities		Performs a sequence of operations to clean and standardize your data to make it more suitable for use.

Transforming Data

Disclaimer: This function uses the OpenAI service. By pressing Save, you agree that the selected data will be sent to this service for processing in accordance with the [Terms of Service](#).

Enriching Data



5. Data Structuring

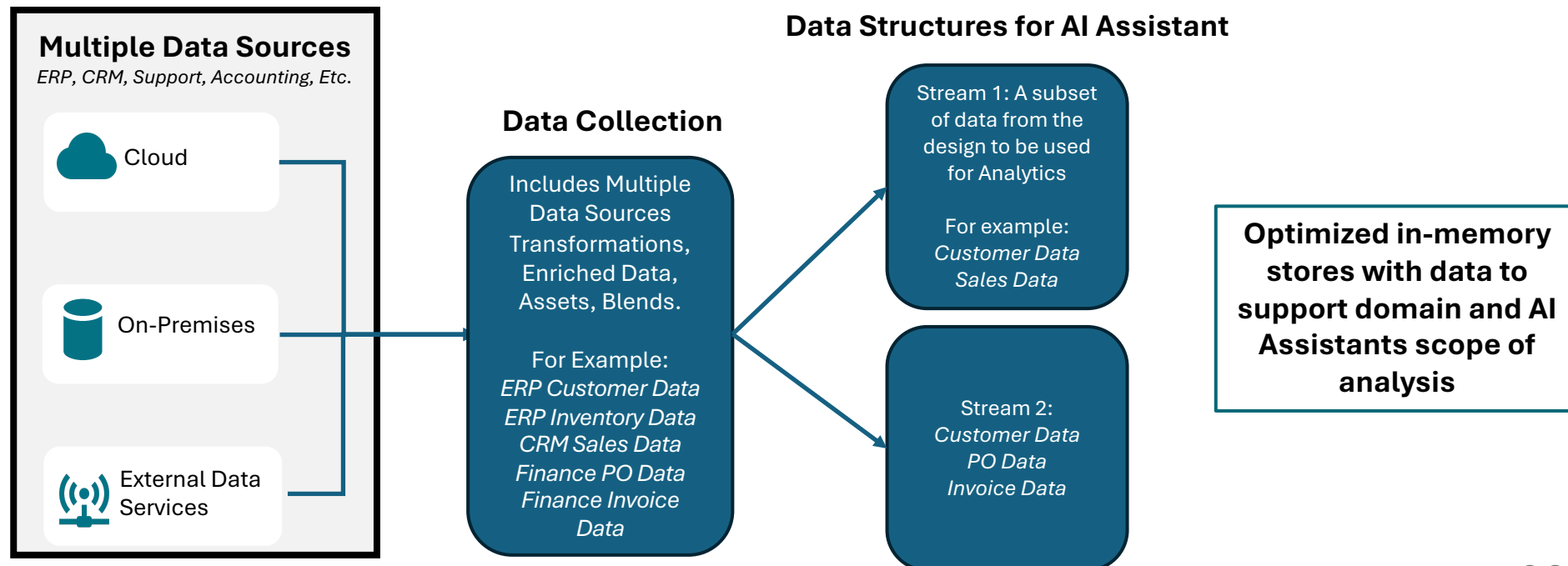
Data structuring for AI assistants

a. Choose a Suitable Format

- Select an appropriate data format (e.g., in-memory, LLM, XML) based on the AI copilot's requirements.

b. Organize Data

- Structure the data logically with clear labels and organization.
- Ensure relational integrity if using databases.





6. Data Annotation


Tagging, labelling, and categorizing

a. Label Data

- Manually or automatically annotate data (e.g., tagging images, categorizing text).
- Ensure high-quality annotations by using reliable annotators or tools.

b. Validate Annotations

- Cross-check annotations for consistency and correctness.
- Use a sample review to ensure annotation quality.

 Asset



Name *

Customer Risk Category

Description

Sentiment Classification



Tags

Customer Categories x  

Sentiment x

Risk Indicator x

Type *

 Classifications 



7. Data Splitting and 8. Data Validation

7. Data splitting for training and validation

a. Divide Data

- Split data into training, validation, and test sets.
- Ensure random and representative sampling for each subset.

b. Ensure Non-Leakage

- Check that no data leakage occurs between training and test sets.

8. Data validation

a. Verify Data Integrity

- Check data consistency across the dataset.
- Validate data against known benchmarks or standards. (for example, pose a question using the AI Assistant and compare the answer to a query you define using a traditional query building process).

b. Conduct Quality Assurance

- Perform statistical analysis to ensure data quality.
- Visualize data to detect any hidden issues.



9. Data Storage and Accessibility

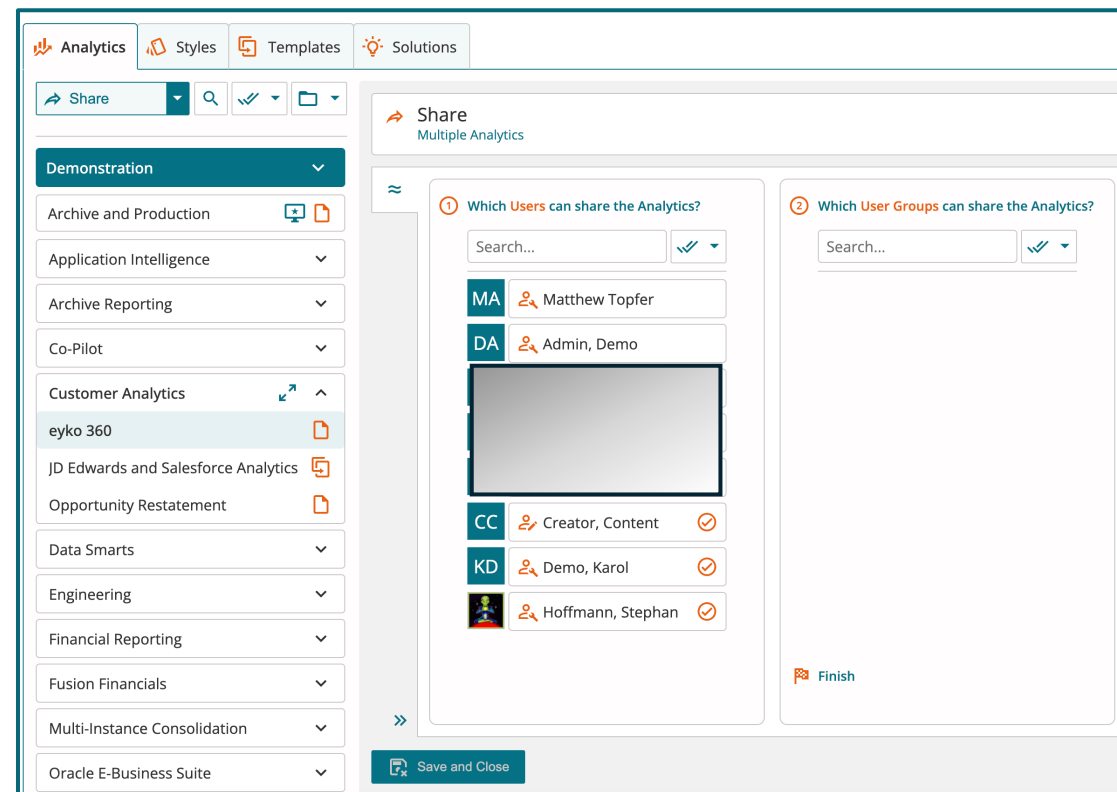
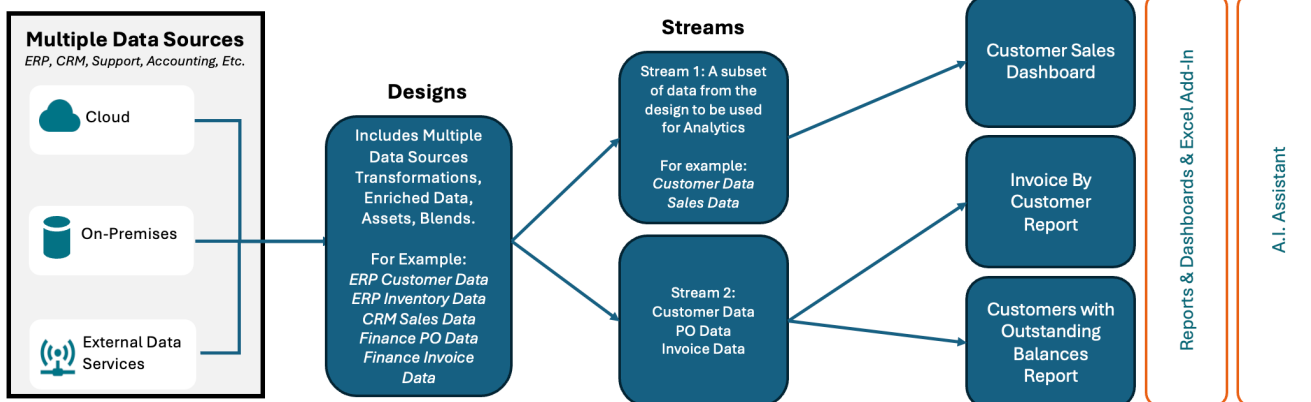
Ensure data storage is reliable, scalable, responsive, and secure

a. Store Data Securely

- Choose a reliable storage solution (e.g., cloud storage, databases).
- Ensure data security with appropriate access controls.

b. Maintain Accessibility

- Organize data for easy retrieval.
- Ensure compatibility with AI copilot systems.





10. Continuous Monitoring and Updates

Continuously monitor and update for better results and adoption

a. Implement Monitoring

- Set up systems to monitor data quality and performance regularly.
- Detect and address issues promptly.

b. Update Data

- Regularly update the dataset to keep it relevant and accurate.
- Incorporate new data and remove obsolete information.

Create an agile and continuous process to:

- Continuously validate data sources (existing and new), collect, clean, transform, enrich, tag, augment, and secure.
- Validate adoption, run traditional reports vs AI Assistant generated reports. Compare for accuracy.
- Monitor usage and adoption – interview users, check service utilization records via billing and admin tools.



AI Assistant Demo



eyko – Insights from ALL your DATA in MINUTES

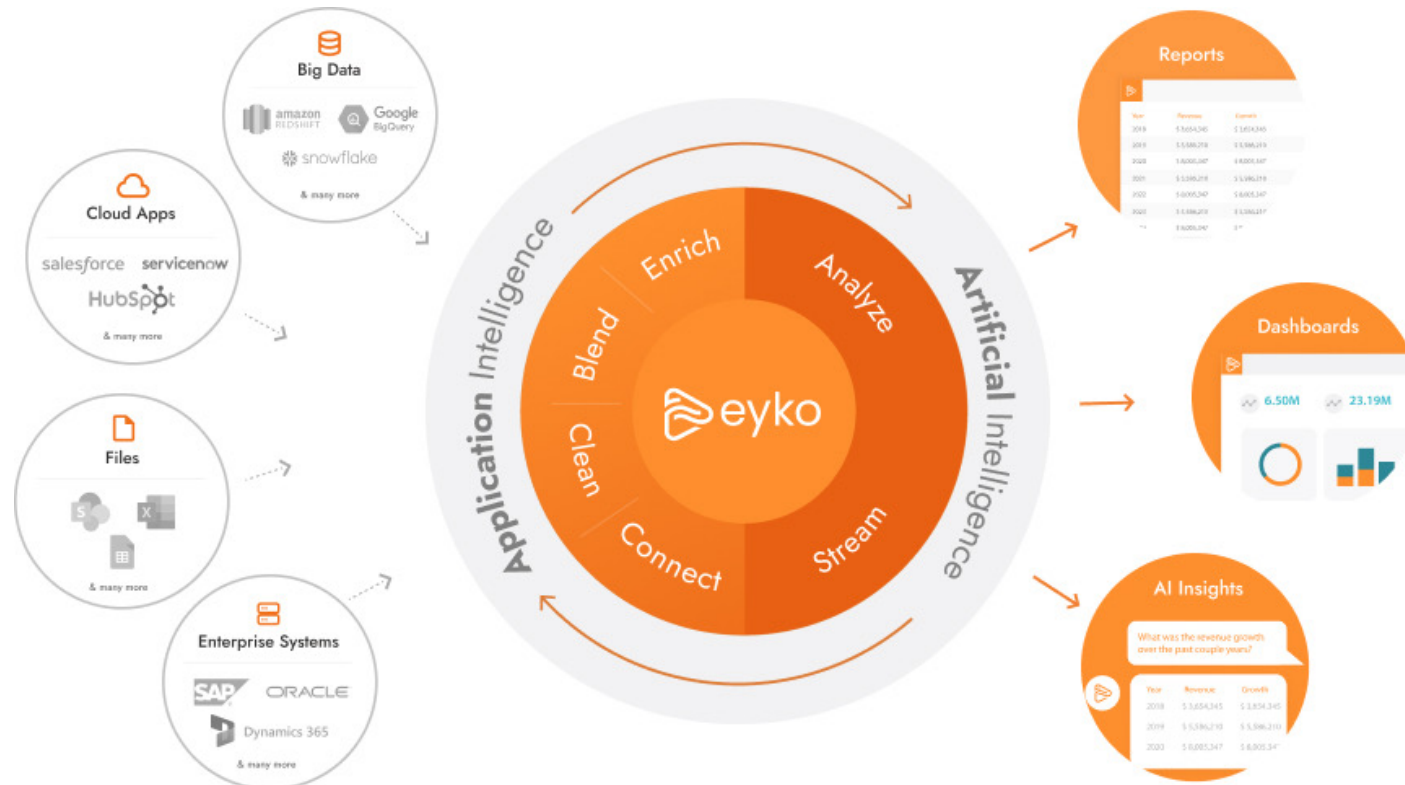
Integrated and scalable platform – all data, large volumes

AI powered – data ingestion and insights

User empowerment – intuitive end to end experience

Data quality and security – accurate and enriched data

Cost and resource efficiency – eliminate tool complexity



eyko empowers businesses to make better data-driven decisions with all their data – in minutes



eyko – What problems do we solve?

Consider this question: ***What customers are at risk of churning?***

Data Requirements: Invoices, support cases, forecast, payment, sentiment, and more...

Before eyko:

- Multiple data pipeline tools
 - Data source mapping, ingestion, and cleaning: ETL skills and scripts
 - MDM: Define common data definitions
 - Governance and security
 - Database: data design, warehouse
 - R/Python for modelling
 - BI: for visualization and reporting
- **Skills:** ETL, deep data source schema knowledge, modeling, scripts, R, data cleansing, BI tools, data scientists
- **Risks:** timelines, data inconsistency, performance, scale, costs

With eyko:

- One end-to-end integrated solution
 - Ingest all data
 - Clean & consolidate
 - Enrich data
 - Answer hard questions
 - Quickly build reports & dashboards (Embedded Power BI)
 - Scales with volume and new data sources
- **Skills:** business domain knowledge, source system understanding, general analytics tool knowledge
- **Benefits:** save time, better decisions, no tool complexity, user empowerment, simplicity, and scale



eyko – Churn example with AI Assistant

Churn example: sales, payment, customer support/success data, sentiment data, forecast data, historical adoption, information about their business such as quarterly reports

eyko | Customer 360

Summary ERP Salesforce Hubspot Jira Shopify Toggl

Customer	Total Billed	Open Amount	Campaigns	Last Activity	Open Cases	Risk	Online Orders	Online Spend	Billable Time
Action Bike Specialists	\$128,281.25	\$212,150.00	3	3/31/2022	1	Medium Risk	2	\$2,533.05	\$150.00
Engineered Bike Systems	\$668,160.00	\$176,980.00	3	12/14/2021	1	Medium Risk	3	\$3,451.88	\$78.75
Fitness Cycling	\$142,757.50	\$323,200.00			2	High Risk	1	\$2,383.67	\$130.00
Journey Sporting Goods	\$318,256.00	\$100,000.00	4	8/28/2021	4	High Risk	3	\$1,775.30	
Major Sports Suppliers	\$1,748,175.00	\$225,000.00	3	5/3/2022		Low Risk	2	\$4,527.02	\$220.00
Novelty Bikes	\$293,203.75	\$345,340.00	4	9/15/2021	2	High Risk	1	\$2,677.02	\$100.00
Professional Cycle Store	\$137,256.25	\$377,981.00				Medium Risk	1	\$2,533.05	\$1,400.00
Total	\$3,994,489.75	\$2,792,651.00	17		12		23	\$34,358.12	\$3,018.75

ERP Salesforce Hubspot Jira Shopify Toggl

Ticket Number	Ticket Date	Contact Name	Customer	Issue	Priority	Severity	Impact	Owner	Status
JIRA-1006	12/1/2021	Jenny Donald	Journey Sporting Goods	Bikes from invoice INV-0029 are missing reflectors. Can you please send them?	High Priority	1 - Critical	1 - Critical	Sally Whittaker	Open
JIRA-1008	12/14/2021	Jenny Donald	Journey Sporting Goods	Haven't been contacted regarding my support case about the bike reflectors.	High Priority	1 - Critical	1 - Critical	Sally Whittaker	Open
JIRA-1011	1/3/2022	Jenny Donald	Journey Sporting Goods	Third ticket about the bike reflectors! This is unacceptable.	High Priority	1 - Critical	1 - Critical	Sally Whittaker	Open
JIRA-1012	1/10/2022	Jenny Donald	Journey Sporting Goods	We are going to return the bikes purchased. The way you handle support does not meet acceptable standards. We are considering other suppliers.	High Priority	1 - Critical	1 - Critical	Sally Whittaker	Open



eyko – Churn example with AI Assistant

View
eyko 360

What do you want to know?

what customers are...

Conversation (Text-Based Q&A)

1. what customers are at ri...

Prompt

Enter...

what customers are at risk of churn

Based on the provided data, customers at risk of churn can be identified by looking at factors such as high "Open Amount", "Open Cases", and "Risk" levels. Here are the customers at high or medium risk:

High Risk Customers:

- 1. Fitness Cycling**
 - Open Amount: \$323,200.00
 - Open Cases: 2
- 2. Journey Sporting Goods**
 - Open Amount: \$100,000.00
 - Open Cases: 4
- 3. Novelty Bikes**
 - Open Amount: \$345,340.00
 - Open Cases: 2

Medium Risk Customers:

- 1. Action Bike Specialists**
 - Open Amount: \$212,150.00
 - Open Cases: 1
- 2. Engineered Bike Systems**
 - Open Amount: \$176,980.00
 - Open Cases: 1
- 3. Professional Cycle Store**
 - Open Amount: \$377,981.00
 - Open Cases: 0

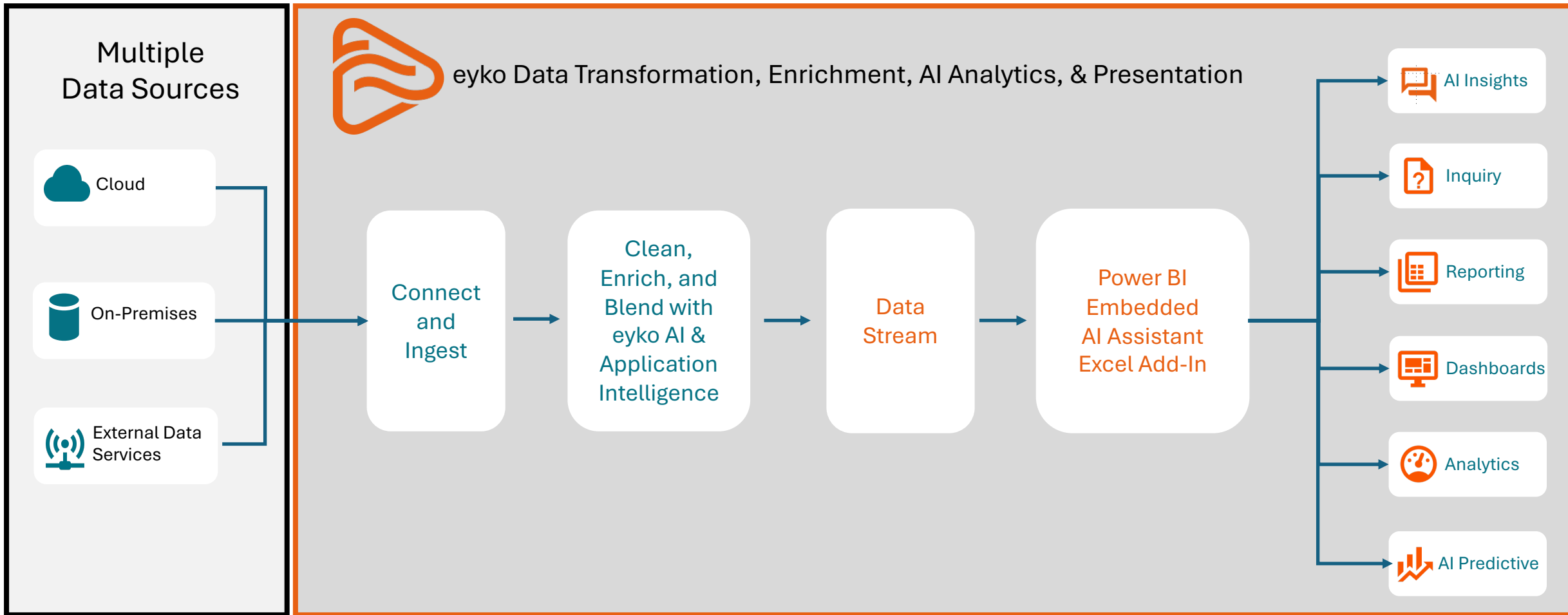
Summary:

- High-Risk Customers** have significant open amounts and multiple open cases, indicating dissatisfaction or unresolved issues.
- Medium-Risk Customers** also have noticeable open amounts but fewer open cases, indicating potential but less severe risks.

These customers should be prioritized for follow-up to address their issues and reduce the risk of churn.



eyko – A Modern and Integrated Data & Analytics Platform





THANK YOU!



eyko – Want to learn more?

1. More information on eyko and AI

1. [AI Chatbots vs AI Assistants and Why it Matters for Analytics](#)
2. [A Virtual Data Scientist Built for You](#)
3. [Get Insights and Answers – Not More Reports Demo](#)
4. [Ten Steps to Curate Your Data for AI](#)
5. [Large Language Models and Analytics](#)
6. [An Introduction to Assistive AI with eyko](#)

2. Get in touch:

- Sales: sales@eyko.io
- Website: <https://eyko.io>
- LinkedIn: <https://www.linkedin.com/company/eyko-inc/>
- YouTube: <https://www.youtube.com/@eykosoftware>
- X: https://x.com/eyko_software



THANK YOU!