RSR/ RED SHOES REDUX

Same journey, new roads.

From Ingestion to Consumption:

Preparing Data for Seamless AI



Matt Nunn

matt@rsr.dev
matt@foguetech.com
mattnunn@mcwtech.com

www.rsr.dev
mcwtech.com
www.foguetech.com

THE ARTWORK

I own an art gallery... But I am not an artist.



I brought the ideas.

AI helped with the artwork.

There is no AI Strategy without a Data Strategy

WHY BUSINESS DATA MATTERS

Business decisions need your business data

Much of the information you need is private to your business.

PUBLIC DATA

Understand the market

Market trends, industry research and competitor information.

PRIVATE BUSINESS DATA

Understand your business

Your orders, costs and inventory.
Your contracts and operations.

Combine them: compare market trends with your own sales.

Give AI controlled access to the business data it needs.

Making data AI-ready

From Ingestion to Consumption



SIX STAGES

1. **INGEST**
   Get the raw data
2. **TRANSFORM**
   Prepare it for use
3. **STORE**
   Store the prepared data
4. **DESCRIBE**
   Add context, remove ambiguity
5. **CONSUME**
   Make it accessible to AI
6. **TEST**
   Check for appropriate answers

Trace an answer back to the stage that needs attention.

RSR

THE STARTING POINT

Start with one business decision

The decision tells us which data to prepare.

DECISION

Who uses the answer and what they'll do

EVIDENCE

Records the business relies on

CHECK

A result a person can verify

MAKING DATA AI READY

01

Ingest

Capture raw records and track what arrived.

WHAT THIS STAGE PRODUCES

A source manifest and replayable inputs



RSR/ RED SHOES REDUX

Same journey, new roads.

01 INGEST

How data and knowledge get into your AI system

Use the right ingestion method for each source.









Real-time streaming

Live events from applications, devices and sensors

Stream events continuously.

Scheduled loads

Business records and reference tables

Load files or API extracts on a schedule.

File discovery and import

Historical and reference documents in personal work drives and archives

Find, approve and import business-relevant content.

Knowledge capture

Procedures and exceptions known by a few people

Interview, document, validate and ingest.

You may not control every source.

Agree access and ownership. Track freshness and missing inputs.

RSR / RED SHOES REDUX

01 INGEST

A missing input can look like a real drop

A successful import can still leave a gap.

3 of 4

expected files arrived

THE PITFALL

A lower total can come from a missing file.

WHAT TO DO

Compare arrivals with the delivery list. Flag the gap and replay when the file arrives.

Hand off the data with its coverage and freshness.

01 INGEST

A source manifest and replayable inputs

Handoff to Transform

PEOPLE DECIDE	Choose the sources, owners, permitted uses and freshness targets.
SYSTEM DOES	Capture originals or stable references, log versions and offsets, and reconcile arrivals.
PASS WHEN	An input can be replayed safely and a missing delivery is visible.

An owner can trace every input and say what was missing.

MAKING DATA AI READY

02 Transform

Prepare records to work together.

WHAT THIS STAGE PRODUCES

Validated records and approved tags



RSR/ RED SHOES REDUX

Same journey, new roads.

Transformation can happen at more than one point

Apply essential controls early. Refine the data for its intended use.

AT INGESTION

Validate format, capture source and access; quarantine unsafe inputs.

ETL Prepare → Load

ELT Load → Prepare

Preserve permitted originals and version the rules for reprocessing.

FOR AI

Extract structure
Create useful passages
Add metadata + indexes

AT QUESTION TIME

Apply access and scope filters, then retrieve and rank evidence.

The stages describe responsibilities; they are not a rigid execution order.

02 TRANSFORM

Join records using identity and time

Match each record to the business entity valid at the time.

Incoming record

Source ID and event time

→

Verified mapping

Match effective at that date

→

Business entity

Approved entity ID

COMMON FAILURE	A reused identifier can join an older record to the wrong entity.
PREPARATION	Standardise dates and units. Check duplicates and route uncertain matches for review.

Retain the input, the rule version and any reason for exclusion.

The meaning hidden in a red cell



Hidden in a red cell

Meaning known to the sheet owner

Explicit availability

Available from 13:00

02 TRANSFORM

Make implicit labels explicit

Capture business meaning as a field or tag that can travel with the record.

ORIGINAL RECORD

A colour, note or shorthand

The source owner knows how to interpret it.

PREPARED RECORD

A named field or tag

Confirm the meaning, record the value and link to the source.

A person reviews ambiguous tags. Keep the original available.

From document pile to usable evidence

Prepare the content and its context before you build the index.

1 Inventory

Sources, owners, access

→

2 Choose versions

Approved, current, canonical

→

3 Extract structure

Text, headings, tables

4 Classify

Topic, scope, authority

→

5 Index

Passages + source links

→

6 Test

Known questions + limits

Keep headings, table context and exceptions with each passage.

Finding business knowledge in user folders

Important decks, spreadsheets and documents are scattered across personal work directories.

BUSINESS DISCOVERY CRITERIA

Reuse definitions from the business glossary / semantic layer.

Shared vocabulary + evidence of business use

Project codes, customer IDs, product names, specialist terms and synonyms.
Add document purpose, owner, scope and effective date.

1 Discover

Scan permitted folders. Read slides, cells and document content.

→ 2 Match meaning

Match IDs and terms. AI suggests topics. Keep matching evidence.

→ 3 Review value

Owner confirms value, authority and access. Resolve copies and drafts.

→ 4 Ingest

Store approved content. Keep source, version and access. Track changes.

A business-language match finds a candidate. It does not prove importance or authority.

02 TRANSFORM

Validated records and approved tags

Handoff to Store

PEOPLE DECIDE	Agree the joins, normalisation rules and tag vocabulary. Review exceptions.
SYSTEM DOES	Apply preparation rules, suggest tags, track exceptions and preserve source links.
PASS WHEN	Known duplicates, unit mismatches and ambiguous IDs behave as expected.

Clean output stays traceable to the input and the rule that changed it.

MAKING DATA AI READY

03

Store

Store prepared data for reliable use.

WHAT THIS STAGE PRODUCES

Named authoritative stores and refresh rules



RSR/ RED SHOES REDUX

Same journey, new roads.

03 STORE

Storage follows the way you use the data

Four familiar storage roles behind AI applications.



Transactional database

Current business records

Warehouse / lakehouse

Tables for analysis

Object storage

Documents, images and logs

Search / vector index

Find relevant passages

03 STORE

What does your AI need to retrieve?

Start with the answer you want the AI to give.

DOCUMENT RETRIEVAL

Find the relevant passage

The answer is in a Word file, PDF, slide deck or other document. The AI needs an excerpt it can cite.

Document source → Search index

Output: passages and source links

STRUCTURED DATA QUERIES

Get the right numbers

The answer is in business records. The AI needs a current value or a calculation across many rows.

Database or analytics store → Query

Output: records, totals or trends

One answer can combine both: show the figures and cite the policy that explains them.

Storage options that underpin AI

Choose according to the data you have and what the AI needs to return.

STORAGE OPTION	WHAT IT HOLDS	WHEN AI USES IT
Transactional database	Live structured records. Frequent, consistent updates.	Look up current status through a read API or view.
Data warehouse	Curated structured tables for analytical SQL and BI.	Query totals, trends and comparisons across many rows.
Document repository	Managed originals, versions and permissions.	Keep approved documents and link answers to the originals.
Object storage / data lake	Files, images, logs and exports. A lake organises data for reuse.	Keep raw inputs and prepared files for processing or replay.
Lakehouse	Files and queryable tables on a data lake foundation.	Process mixed data and query the resulting curated tables.
Search / vector index	Searchable text, embeddings, metadata and source references.	Retrieve relevant passages for AI to read and cite.

A single platform can support several of these jobs.

03 STORE

From scattered files to searchable evidence

Start with the locations and permissions you already have.

01 SOURCE

Find approved files

Inventory permitted folders. Identify business documents, owners and current versions.

Shared drives · SharePoint
Word · PDFs · slide decks

›

02 PREPARATION

Prepare passages

Extract text and structure. Split it into useful passages. Keep source and access data.

Document ID · version
Page or section · access

›

03 INDEX

Index for retrieval

Index text for exact terms. Add embeddings for matches based on meaning.

Keyword · vector · hybrid
Source links for citations

Keep originals in the approved source. Refresh the index when content or access changes.

RAG: Retrieval-Augmented Generation

At question time, the AI uses passages from the search index.

01	Retrieve	Find relevant passages using keyword search, vector search or both. Apply the user’s access permissions before passing text to the model.
02	Augment	Give the model the question and the selected passages. Include the document title, version and page or section.
03	Generate	Answer using those passages and cite the source. Say when the available evidence cannot support an answer.

The index helps find the evidence. The original document is what the user can verify.

03 STORE

Where should AI query structured data?

Use the system that can answer at the right level of detail and freshness.

CURRENT RECORD

Operational lookup

Use a read API or a controlled view over the transactional database.

Order status · stock availability

Output: a current record or value

MANY RECORDS

Analytical query

Query curated warehouse or lakehouse tables with approved metric definitions.

Revenue by region · inventory trend

Output: a total, comparison or trend

Give the AI controlled read access and return the source, scope and freshness of the result.

What the Store stage hands to AI

Decide what counts as authoritative, then make it available through a controlled read path.

DECISION	DOCUMENT RETRIEVAL	STRUCTURED DATA QUERIES
Authoritative source	Approved document and version with a stable source link.	Live record, curated table or published reporting snapshot.
Refresh and access	Update the index as files and permissions change.	Set freshness targets and apply row and column access rules.
Verify the result	Open the cited passage in the approved source.	Reproduce the value using the same definition, filters and cutoff.

Next, Describe gives AI the business definitions it needs to interpret the evidence.

04

Describe

Give the data its business meaning.

WHAT THIS STAGE PRODUCES

Approved definitions and metadata



What a business definition needs

Publish approved terms in the semantic layer and glossary.

Part of the definition	What to capture
Meaning	The term the business uses and what it means.
Grain and scope	What each row represents, which records count and the time period.
Rules	Calculation, units, valid joins and exceptions.
Owner and version	Who approves changes and when the definition applies.

For documents, record source, effective date and scope.

04 DESCRIBE

One word, two definitions

Two teams use “active” for different groups of accounts.

USAGE

Signed in recently

Specify the activity and the period it counts.

CONTRACT

Has a current agreement

Specify the agreement status and effective date.

Name each measure and publish the rules that go with it.

AI needs your business definitions

A semantic view connects internal language to data and rules.

How many **Lighthouse** sites were **live** at **FY26 P3** close?

Illustrative company terminology



OPERATIONS SEMANTIC VIEW

WHAT WE MEAN	WHAT THE MODEL DEFINES
Lighthouse Our pilot rollout sites	Sites.cohort_code = 'LH'
Live Commissioned and signed off	Both approval dates at or before cutoff
FY26 P3 Our fiscal year and period	Cutoff from the fiscal calendar
How many sites One count per site	Count distinct site_id Join sites to approvals on site_id

AI queries using our definitions





Keep the meaning. Focus the model.

Expose the relevant business view and reuse shared definitions.
Reduce unnecessary complexity. Test accuracy and response time.

RSR / RED SHOES REDUX

04 DESCRIBE

Which documents should AI use?

Classify authority and applicability for the question being answered.

PREFER	**Approved + applicable**	Correct scope, version and effective date
SUPPORT	**Useful supporting evidence**	Keep provenance; distinguish from policy
REVIEW	**Uncertain or conflicting**	Ask the owner to resolve authority
EXCLUDE BY DEFAULT	**Duplicate, superseded or out of scope**	Retain where required; do not simply delete

TYPE · TOPIC · OWNER · APPROVAL · DATES · SCOPE · VERSION · ACCESS

AI can suggest labels. Owners approve authority. Popularity is not importance.

04 DESCRIBE

An owned semantic model and evidence dictionary

Handoff to Consume

PEOPLE DECIDE	Name metric owners and agree definitions, valid joins, reporting periods and claim rules.
SYSTEM DOES	Serve versioned measures and glossary terms to approved queries and retrieval.
PASS WHEN	Two people get the same result from the approved definition and period.

Describe the data in the language of the people making the decision.

05 Consume

Give AI a controlled path to the evidence.

WHAT THIS STAGE PRODUCES

A scoped answer with sources and limits



05 CONSUME

Two paths to usable evidence

Use the caller’s permissions on both paths.

GOVERNED QUERY

Counts, totals and trends

Query approved measures for the requested scope and period.

PERMITTED RETRIEVAL

Policies, reports and notes

Retrieve approved passages with their source and version.



Return the evidence and citations used to answer.

Source · Microsoft Foundry documentation on retrieval-augmented generation

05 CONSUME

An answer people can check

Show enough of the evidence for someone to verify the conclusion.

Conclusion	What the evidence supports
Scope	Which records and time period the answer covers
Sources	Links to the approved records and versions used
Limits	What’s missing, stale or uncertain

05 CONSUME

A scoped answer with sources and limits

Handoff to Test

PEOPLE DECIDE	Choose which roles can see which aggregates or passages, and when a user needs to clarify or seek approval.
SYSTEM DOES	Authorise before querying and retrieving. Recheck derived copies, mask unnecessary details and log evidence use.
PASS WHEN	An allowed user can verify the answer. A denied user gets no restricted evidence, including through the cache.

When evidence is thin or contradictory, say what would resolve it.

06

Test

Verify the answer and its limits before release.

WHAT THIS STAGE PRODUCES

Known cases and a release decision



Testing the route to the answer



Source	Transform	Retrieval	Outcome
Expected inputs	Meaning preserved	Evidence in scope	User task succeeds

06 TEST

Relevant does not always mean trustworthy

Illustrative question: “How should I maintain this equipment?”

CURRENT + APPROVED

Procedure v3

Applies to this model and location

Use as primary evidence

EXACT WORD MATCH

Procedure v1

Superseded by v3

Exclude from current answer

RECENT DRAFT

Proposed procedure

Not approved; different process

Do not treat as policy

Check approval, scope, date, access and citation. Abstain when evidence is unresolved.

Test known results and boundaries

Exercise the data, access rules and final response together.

Test case	Expected behaviour
Known result	Match the approved result for the same scope and period.
Missing or stale data	State the gap and give only conclusions the evidence supports.
Ambiguous request	Clarify the term, scope or time window.
Restricted data	Respond without exposing restricted evidence.

Replay when a source, definition, access rule, index or model changes.

06 TEST

Where a wrong answer starts

Use the answer, the sources it cited and the data it received to trace the fault.

Missing evidence	Check capture, freshness and joins.
Wrong total	Check the snapshot and the business definition.
Unsupported answer	Check retrieval, permissions and the cited source.

Fix the failing stage and rerun the same test.

06 TEST

Data readiness after release

Keep a set of questions with known answers and replay them as things change.

Source or pipeline changes	Reconcile records and rerun coverage checks.
Definitions or access change	Recheck expected answers and denied cases.
Index or model changes	Replay questions and inspect citations.

Watch for stale evidence and failures in cases that used to pass.

06 TEST

Evidence for release

Handoff to the team using the AI

PEOPLE DECIDE	Agree expected answers, allowed access and when to hold.
SYSTEM DOES	Replay known cases and flag stale data or failures.
PASS WHEN	Known results match. Incomplete and restricted requests behave as agreed.

Keep the tests so you can run them again after a change.

REFERENCE ARCHITECTURES

Six stages across four platforms

The same data preparation work shows up in different tools.

STRUCTURED PATH

Business records and measures

Capture live and scheduled inputs. Join, version and serve the measures through an authorised read path.

DOCUMENT PATH

Approved documents and notes

Keep permitted originals. Prepare and tag excerpts, retrieve within scope and cite the source.

Look for the same checks on each platform: coverage, meaning, access and evidence.

Microsoft

SIX-STAGE PRODUCT MAP

Fabric data foundation with Microsoft Foundry for AI applications

1 Ingest

Data Factory

Pipelines and Copy job

Eventstreams

Streaming events

Mirroring

Database replication

OneLake shortcuts

Access without copying

2 Transform

Dataflow Gen2

Low-code preparation

Fabric notebooks

Spark and Python

Warehouse T-SQL

Relational transformations

3 Store

OneLake

Shared data foundation

Lakehouse / Warehouse

Files and analytical tables

Eventhouse

Real-time event data

4 Describe

Power BI semantic models

Measures and relationships

Fabric IQ

Shared business context

Ontology (preview)

Entities and operational rules

5 Consume

Fabric data agents

Questions over governed data

Microsoft Copilot

Business-user experience

Microsoft Foundry

Custom AI agents and apps

6 Test

Fabric MLV checks

Data constraints and quality

Fabric notebooks

SQL / Python assertions

Foundry evaluations

Answer quality and safety

Azure DevOps

Repeat tests in CI/CD

Document retrieval

A parallel path for RAG

OneLake files → **Azure AI Search** → **Foundry agents**

OneLake catalog and Microsoft Purview — Discovery, access policies, lineage and information protection

Test throughout the lifecycle. People define expected answers, access checks and release criteria.

Snowflake

SIX-STAGE PRODUCT MAP

Snowflake AI Data Cloud with governed semantics and Cortex Agents

1 Ingest

Snowflake Openflow
Source connectors

Snowpipe
Continuous file loading

Snowpipe Streaming
Low-latency row ingestion

COPY INTO
Bulk file loading

2 Transform

Dynamic Tables
Declarative SQL pipelines

Snowpark
Python, Java and Scala

Streams and Tasks
Change-driven processing

dbt Projects
SQL models and tests

3 Store

Snowflake tables
Managed analytical storage

Iceberg tables
Open table format

Stages
Files and documents

4 Describe

Semantic Views
Metrics and relationships

Semantic Studio
Build semantic definitions

Horizon Context (preview)
Collect and activate context

5 Consume

Cortex Agents
Coordinate queries and tools

Cortex Analyst
SQL using business semantics

Cortex AI Functions
AI inside SQL

Streamlit
Custom user experiences

6 Test

Data Metric Functions
Data quality and thresholds

dbt tests
Model and business rules

AI Observability
TruLens RAG evaluations

Cortex Agent evaluations
Agent task and tool quality

Document retrieval
A parallel path for RAG

Prepared document text → **Cortex Search** → **Cortex Agents**

Snowflake Horizon Catalog — Discovery, lineage, role-based access and data policies

Test throughout the lifecycle. People define expected answers, access checks and release criteria.

Databricks

SIX-STAGE PRODUCT MAP

Lakeflow data engineering with Unity Catalog semantics and Agent Bricks

1 Ingest

Lakeflow Connect
Managed source connectors

Auto Loader
Incremental file ingestion

Structured Streaming
Streaming sources

2 Transform

Lakeflow pipelines
Spark Declarative Pipelines

Lakeflow Designer
Visual data preparation

Notebooks / SQL
Custom transformations

Lakeflow Jobs
Orchestrate the workflow

3 Store

Delta Lake tables
Bronze, silver and gold data

Unity Catalog volumes
Governed files

Cloud object storage
Underlying storage

4 Describe

Unity Catalog semantics
Business context

Metric Views
Reusable measures and joins

Semantic metadata
Descriptions and synonyms

5 Consume

Genie
Natural-language analytics

Agent Bricks
Build and deploy agents

Model Serving
Model inference endpoints

Databricks Apps
Custom user experiences

6 Test

Pipeline expectations
Data rules and violations

MLflow evaluation
AI scorers and test sets

MLflow traces
Inspect retrieved evidence

Lakeflow Jobs
Schedule regression runs

Document retrieval
A parallel path for RAG

Prepared source tables → **AI Search** → **Agent Bricks agents**

Unity Catalog — Data and AI discovery, permissions, lineage and governance

Test throughout the lifecycle. People define expected answers, access checks and release criteria.

AWS

SIX-STAGE PRODUCT MAP

AWS data services with Amazon Bedrock and AWS Quick for AI consumption

1 Ingest	2 Transform	3 Store	4 Describe	5 Consume	6 Test
AWS DMS Database migration and CDC	**AWS Glue** Serverless ETL jobs	**Amazon S3** Files and lake storage	**AWS Quick Sight Topics** Semantic metadata and joins	**Amazon Bedrock** Models and custom AI apps	**AWS Glue Data Quality** DQDL rules and checks
Amazon Data Firehose Streaming delivery	**Amazon EMR** Spark processing	**Amazon S3 Tables** Managed Iceberg tables	**Amazon Redshift views** Approved SQL definitions	**Bedrock Agents** Tools and orchestration	**Bedrock evaluations** Model and RAG quality
AWS Glue connectors Batch source ingestion	**Amazon Redshift SQL** SQL transformations	**Amazon Redshift** Analytical warehouse	**AWS Glue Data Catalog** Technical metadata	**AWS Quick** Business-user AI workspace	**Amazon CloudWatch** Operational signals
Amazon AppFlow SaaS data integration	**Bedrock Data Automation** Extract document content	**Amazon OpenSearch** Text and vector indexes	**Amazon DataZone** Business glossary	**AWS Quick Sight** BI and natural-language Q&A	**AWS CodeBuild** Automate regression tests

Document retrieval

A parallel path for RAG

S3 + prepared documents → **Bedrock Knowledge Bases** → **Bedrock agents / apps**

AWS Lake Formation, IAM and DataZone — Data permissions, access control and business discovery

Test throughout the lifecycle. People define expected answers, access checks and release criteria.

TAKE IT BACK TO WORK

Six checks before launch

Use these as milestones when you explain the investment to management.

Stage	Evidence to show	Risk it addresses
01 Ingest	Source register, coverage and replay	Silent gaps bias answers
02 Transform	Verified joins and approved tags	Wrong records are linked
03 Store	Named source, version and refresh	Two copies disagree
04 Describe	Approved measures and examples	Teams count differently
05 Consume	Scoped evidence and citations	Leaks or unsupported claims
06 Test	Result and access checks	Bad answers drive decisions

Each proof has an owner, a check and a cost of skipping it.

TAKE IT BACK TO WORK

A first project to sponsor

Fund the data work for one decision the business already makes.

SCOPE	**One business decision** Name the question, the people who use the answer and the sources it needs.
INVESTMENT	**Time for the data work** Give source, definition and access owners time to address the first gap.
RETURN	**Evidence that it works** Show a verified answer and the checks that keep it dependable.

Use the six stages to explain what the pilot needs and why.

TAKE IT BACK TO WORK

A first data readiness project

From raw records to an answer you can check

One business decision gives the work a clear scope.

The team follows the evidence through six stages, gives the first gap an owner and tests the fix.

Release when a user can verify the answer from evidence they're allowed to see.

RSR/ RED SHOES REDUX

Same journey, new roads.

Thank you.

Matt Nunn

matt@rsr.dev

www.rsr.dev

www.foguetech.com

mcwtech.com



APPENDIX

Five-stage architecture maps

Alternative product views for reference

AI-ready data with Fabric and Azure

Microsoft

Structured data and document evidence across the same six stages

1 Ingest	2 Transform	3 Store	4 Describe	5 Consume
Fabric workspace and Azure AI services				
Data Factory Eventstreams	Fabric notebooks Prepare records	Lakehouse OneLake tables	MEASURES Power BI semantic model	AI APP Fabric data agent Foundry app
OneLake shortcuts Approved files	Fabric notebooks Extract + chunk	OneLake files Prepared passages	METADATA Source metadata Scope + access	RETRIEVAL Azure AI Search Foundry app

OneLake catalog, Entra ID and Purview information protection

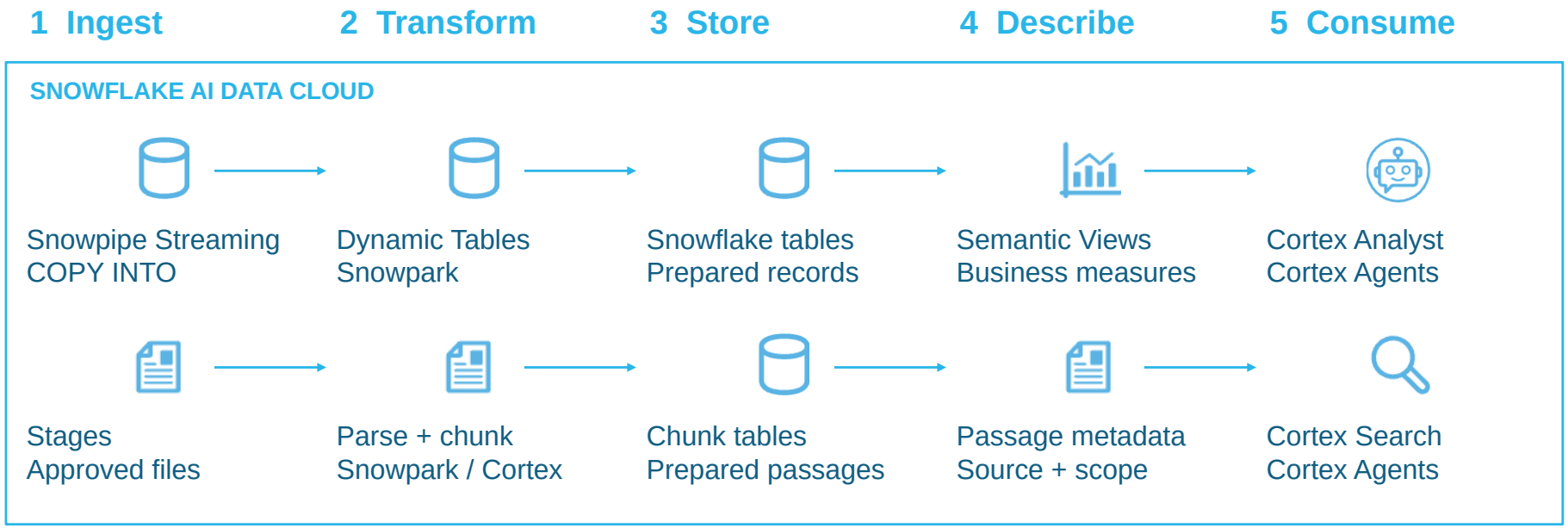
6 Test

Fabric MLV checks / notebook assertions + Foundry evaluations

People approve expected answers, citations and access tests. Checks run across all stages.

AI-ready data on Snowflake

Structured data and document evidence across the same six stages

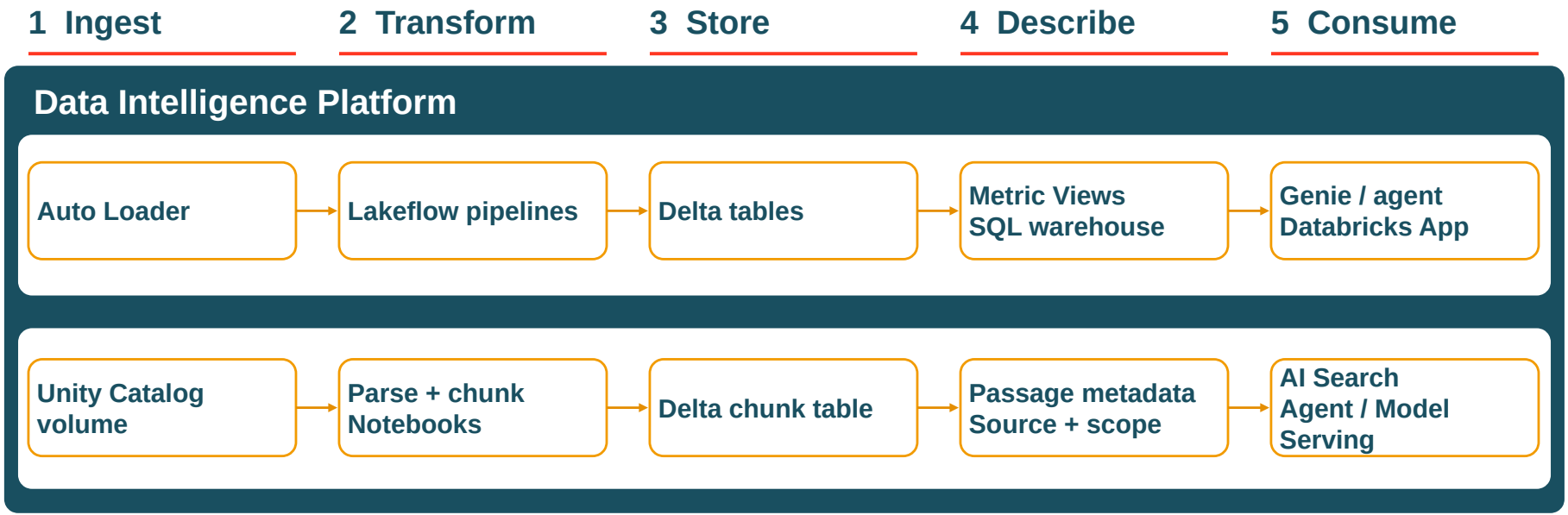


6 Test Data Metric Functions / dbt tests + AI Observability / Agent evaluations

People approve expected answers, citations and access tests. Checks run across all stages.

AI-ready data on Databricks

Structured data and document evidence across the same six stages

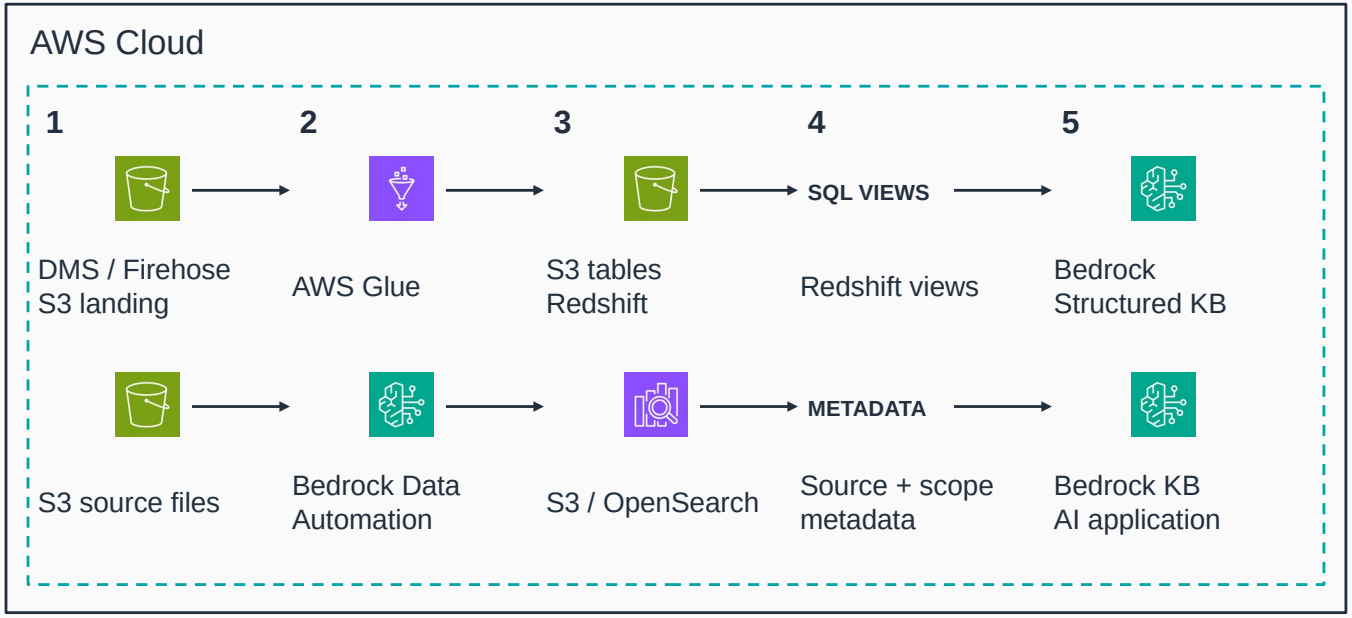


6 Test Pipeline expectations + MLflow evaluation and traces + Lakeflow Jobs

People approve expected answers, citations and access tests. Checks run across all stages.

AI-ready data on AWS

Structured and document paths through six stages



Glue Data Catalog, Lake Formation and IAM

6 Test: Glue Data Quality + Bedrock evaluations

People approve definitions, expected answers and access tests.

1 Ingest

DMS or Firehose brings records in. S3 holds source files.

2 Transform

Glue prepares records. Bedrock extracts document content.

3 Store

S3, Redshift and OpenSearch hold usable copies.

4 Describe

SQL definitions and passage metadata add meaning.

5 Consume

Bedrock queries structured data and retrieves passages.

6 Test

Glue Data Quality and Bedrock evaluations check results.

Microsoft

FIVE-STAGE PRODUCT MAP

Fabric data foundation with Microsoft Foundry for AI applications

1 Ingest

Data Factory

Pipelines and Copy job

Eventstreams

Streaming events

Mirroring

Database replication

OneLake shortcuts

Access without copying

2 Transform

Dataflow Gen2

Low-code preparation

Fabric notebooks

Spark and Python

Warehouse T-SQL

Relational transformations

3 Store

OneLake

Shared data foundation

Lakehouse / Warehouse

Files and analytical tables

Eventhouse

Real-time event data

4 Semantics

Power BI semantic models

Measures and relationships

Fabric IQ

Shared business context

Ontology (preview)

Entities and operational rules

5 AI consumption

Fabric data agents

Questions over governed data

Microsoft Copilot

Business-user experience

Microsoft Foundry

Custom AI agents and apps

Document retrieval

A parallel path for RAG

OneLake files → **Azure AI Search** → **Foundry agents**

OneLake catalog and Microsoft Purview — Discovery, access policies, lineage and information protection

Core product map. Storage can precede transformation. Sources checked 16 November 2026.

Snowflake

FIVE-STAGE PRODUCT MAP

Snowflake AI Data Cloud with governed semantics and Cortex Agents

1 Ingest

Snowflake Openflow
Source connectors

Snowpipe
Continuous file loading

Snowpipe Streaming
Low-latency row ingestion

COPY INTO
Bulk file loading

2 Transform

Dynamic Tables
Declarative SQL pipelines

Snowpark
Python, Java and Scala

Streams and Tasks
Change-driven processing

dbt Projects
SQL models and tests

3 Store

Snowflake tables
Managed analytical storage

Iceberg tables
Open table format

Stages
Files and documents

4 Semantics

Semantic Views
Metrics and relationships

Semantic Studio
Build semantic definitions

Horizon Context (preview)
Collect and activate context

5 AI consumption

Cortex Agents
Coordinate queries and tools

Cortex Analyst
SQL using business semantics

Cortex AI Functions
AI inside SQL

Streamlit
Custom user experiences

Document retrieval
A parallel path for RAG

Prepared document text → **Cortex Search** → **Cortex Agents**

Snowflake Horizon Catalog — Discovery, lineage, role-based access and data policies

Core product map. Storage can precede transformation. Sources checked 16 November 2026.

Databricks

FIVE-STAGE PRODUCT MAP

Lakeflow data engineering with Unity Catalog semantics and Agent Bricks

1 Ingest → 2 Transform → 3 Store → 4 Semantics → 5 AI consumption

Lakeflow Connect
Managed source connectors

Auto Loader
Incremental file ingestion

Structured Streaming
Streaming sources

Lakeflow pipelines
Spark Declarative Pipelines

Lakeflow Designer
Visual data preparation

Notebooks / SQL
Custom transformations

Lakeflow Jobs
Orchestrate the workflow

Delta Lake tables
Bronze, silver and gold data

Unity Catalog volumes
Governed files

Cloud object storage
Underlying storage

Unity Catalog semantics
Business context

Metric Views
Reusable measures and joins

Semantic metadata
Descriptions and synonyms

Genie
Natural-language analytics

Agent Bricks
Build and deploy agents

Model Serving
Model inference endpoints

Databricks Apps
Custom user experiences

Document retrieval
A parallel path for RAG

Prepared source tables → **AI Search** → **Agent Bricks agents**

Unity Catalog — Data and AI discovery, permissions, lineage and governance

Core product map. Storage can precede transformation. Sources checked 16 November 2026.

AWS

FIVE-STAGE PRODUCT MAP

AWS data services with Amazon Bedrock and AWS Quick for AI consumption

1 Ingest	2 Transform	3 Store	4 Semantics	5 AI consumption
AWS DMS Database migration and CDC	**AWS Glue** Serverless ETL jobs	**Amazon S3** Files and lake storage	**AWS Quick Sight Topics** Semantic metadata and joins	**Amazon Bedrock** Models and custom AI apps
Amazon Data Firehose Streaming delivery	**Amazon EMR** Spark processing	**Amazon S3 Tables** Managed Iceberg tables	**Amazon Redshift views** Approved SQL definitions	**Bedrock Agents** Tools and orchestration
AWS Glue connectors Batch source ingestion	**Amazon Redshift SQL** SQL transformations	**Amazon Redshift** Analytical warehouse	**AWS Glue Data Catalog** Technical metadata	**AWS Quick** Business-user AI workspace
Amazon AppFlow SaaS data integration	**Bedrock Data Automation** Extract document content	**Amazon OpenSearch** Text and vector indexes	**Amazon DataZone** Business glossary	**AWS Quick Sight** BI and natural-language Q&A

Document retrieval
A parallel path for RAG

S3 + prepared documents → **Bedrock Knowledge Bases** → **Bedrock agents / apps**

AWS Lake Formation, IAM and DataZone — Data permissions, access control and business discovery

Core product map. Storage can precede transformation. Sources checked 16 November 2026.