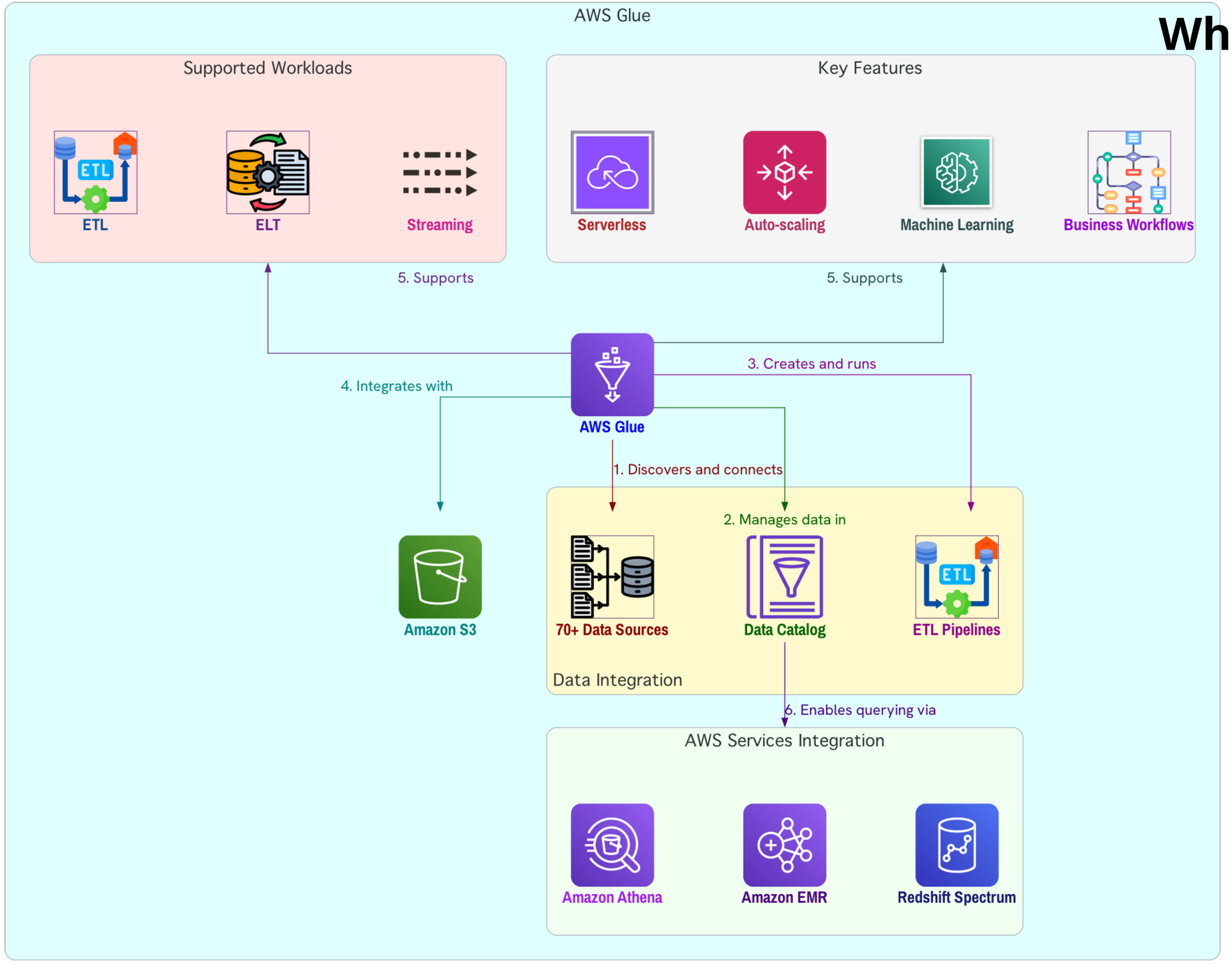
AWS Glue

Table of Contents

What is AWS Glue?



1️⃣. 🔍 Discover, prepare, move & integrate data

☁️ Serverless data integration service

🔬 Analytics, ML, app development

2️⃣. 📊 70+ data sources & centralized catalog

🔌 Connect diverse sources

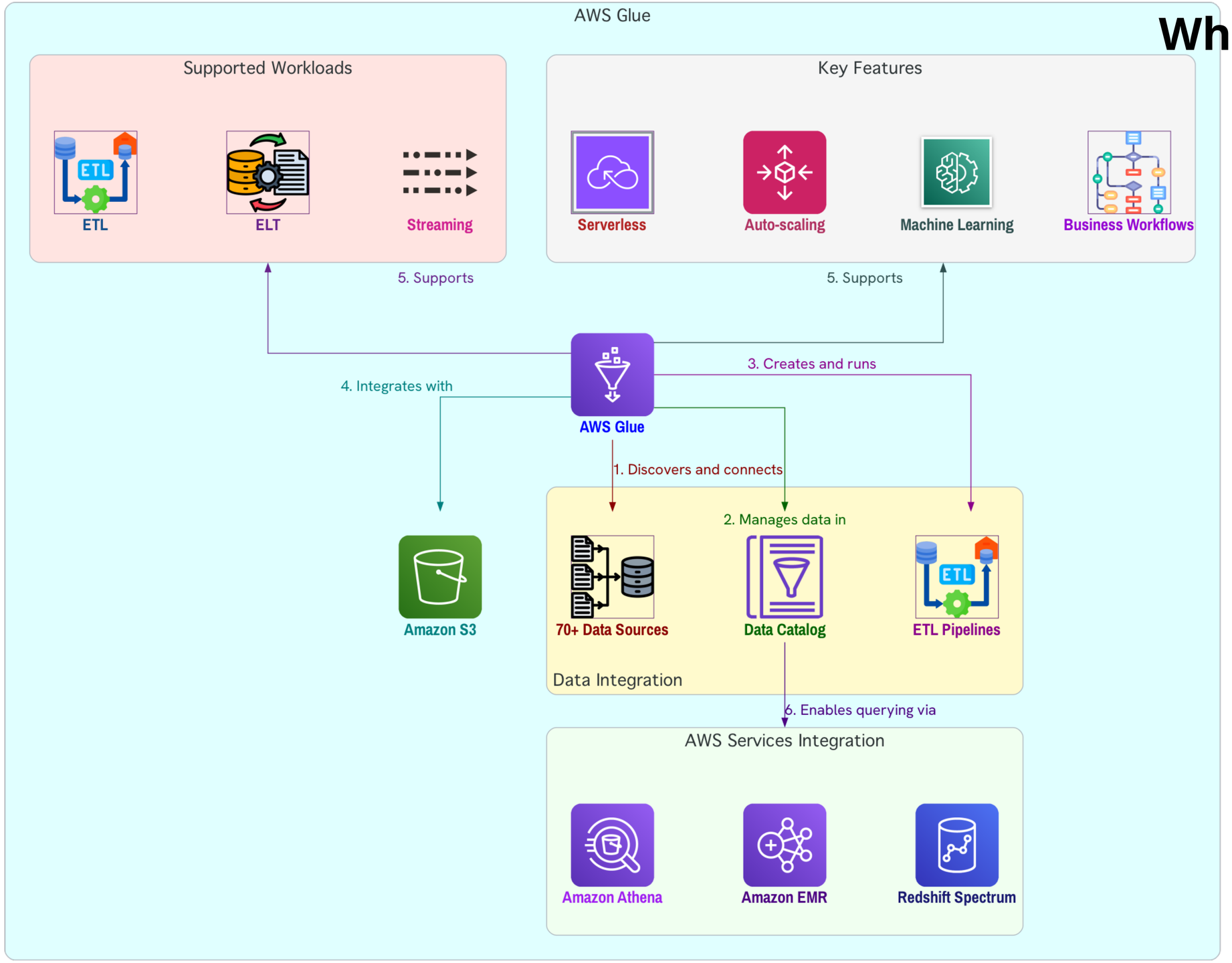
📚 Centralized data management

3️⃣. 🔄 Visual ETL pipeline creation

👁️ Visual creation and monitoring

🏞️ Efficient data lake loading

What is AWS Glue?



4️⃣. 🔎 Instant data querying with AWS services

🔍 Immediate search and query

🤝 Integration with Athena, EMR, Redshift Spectrum

5️⃣. ☁️ Serverless & scalable architecture

🏗️ No infrastructure management

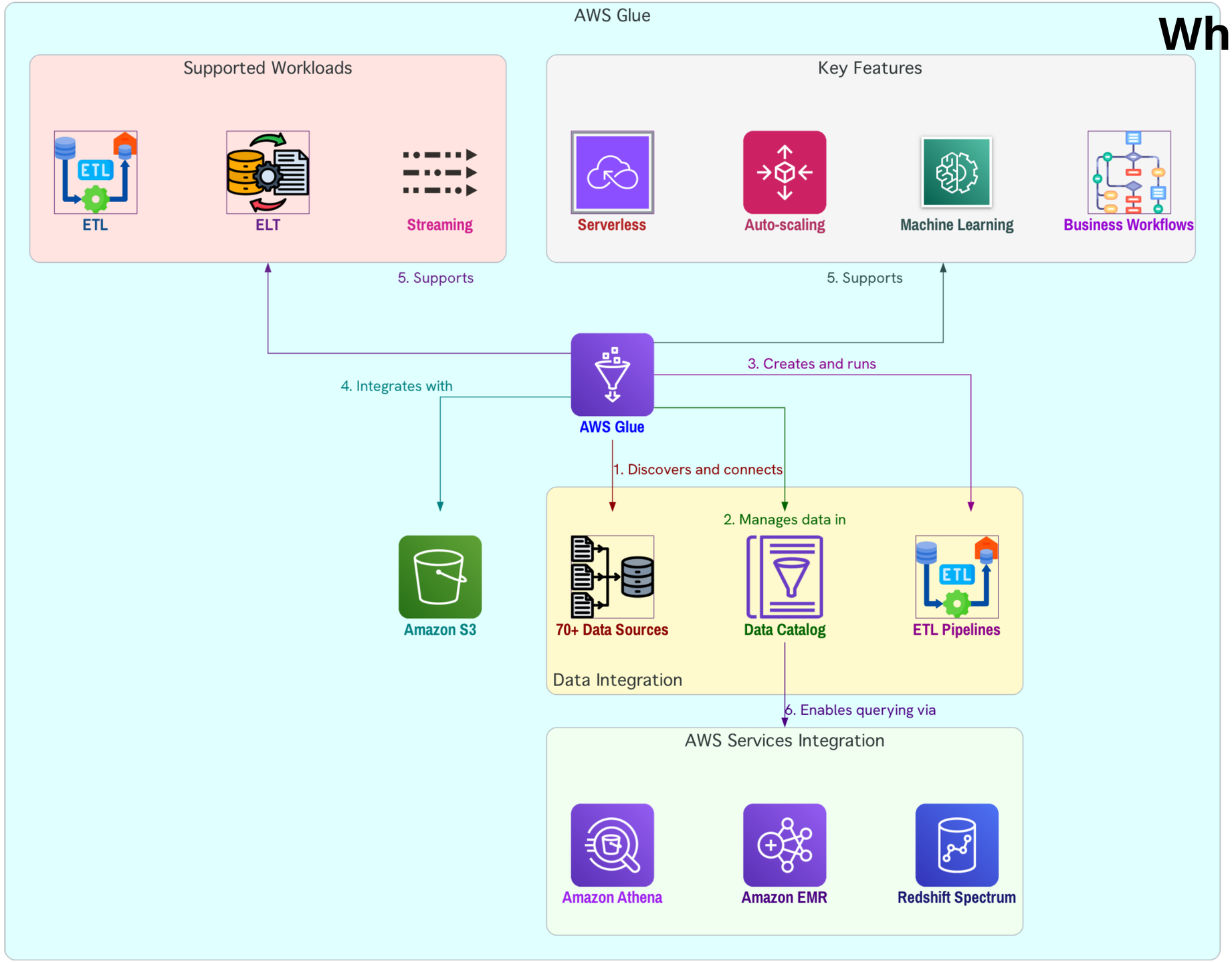
🎛️ Consolidated data integration capabilities

6️⃣. 👥 Supports various user types

🔧 Flexible workload support

📊 ETL, ELT, streaming in one service

What is AWS Glue?



7️⃣. 💼 Easy-to-use interfaces & tools

🛠️ Job-authoring tools

👷‍♂️ Tailored for varied skill sets

8️⃣. 📈 On-demand scaling for any data size

🚀 Handles all data types and schemas

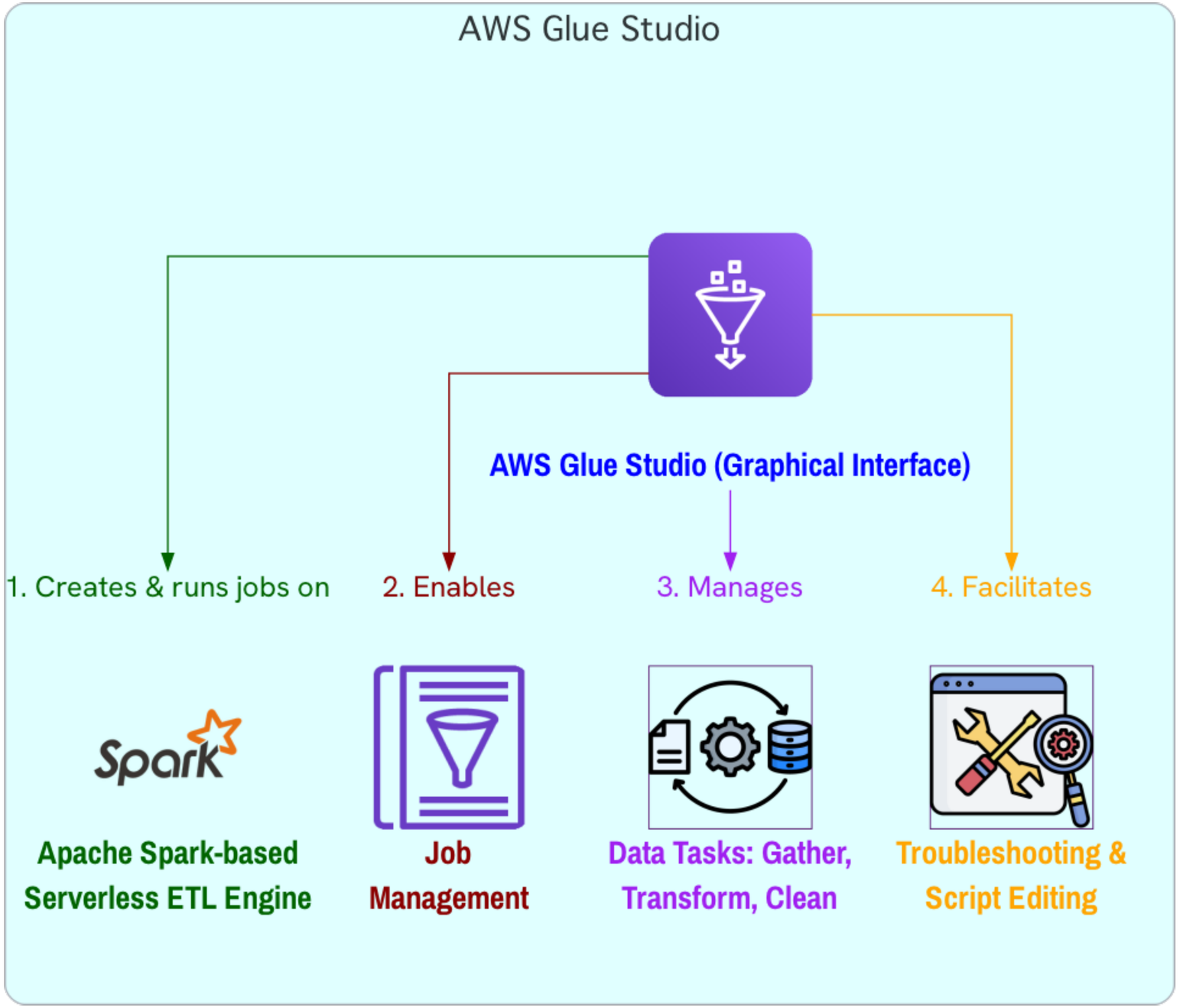
🎯 Focus on high-value activities

9️⃣. 💰 Pay-as-you-go billing

💸 Cost optimization

🔒 Built-in high availability

What is AWS Glue Studio?



1️⃣. 🖥️ Graphical interface for AWS Glue

👥 User-friendly interface

🔓 Accessible to various skill levels

2️⃣. 🔄 Create, run, and monitor data integration jobs

🏗️ Easy setup

▶️ Execute processes

📊 Track integration workflows

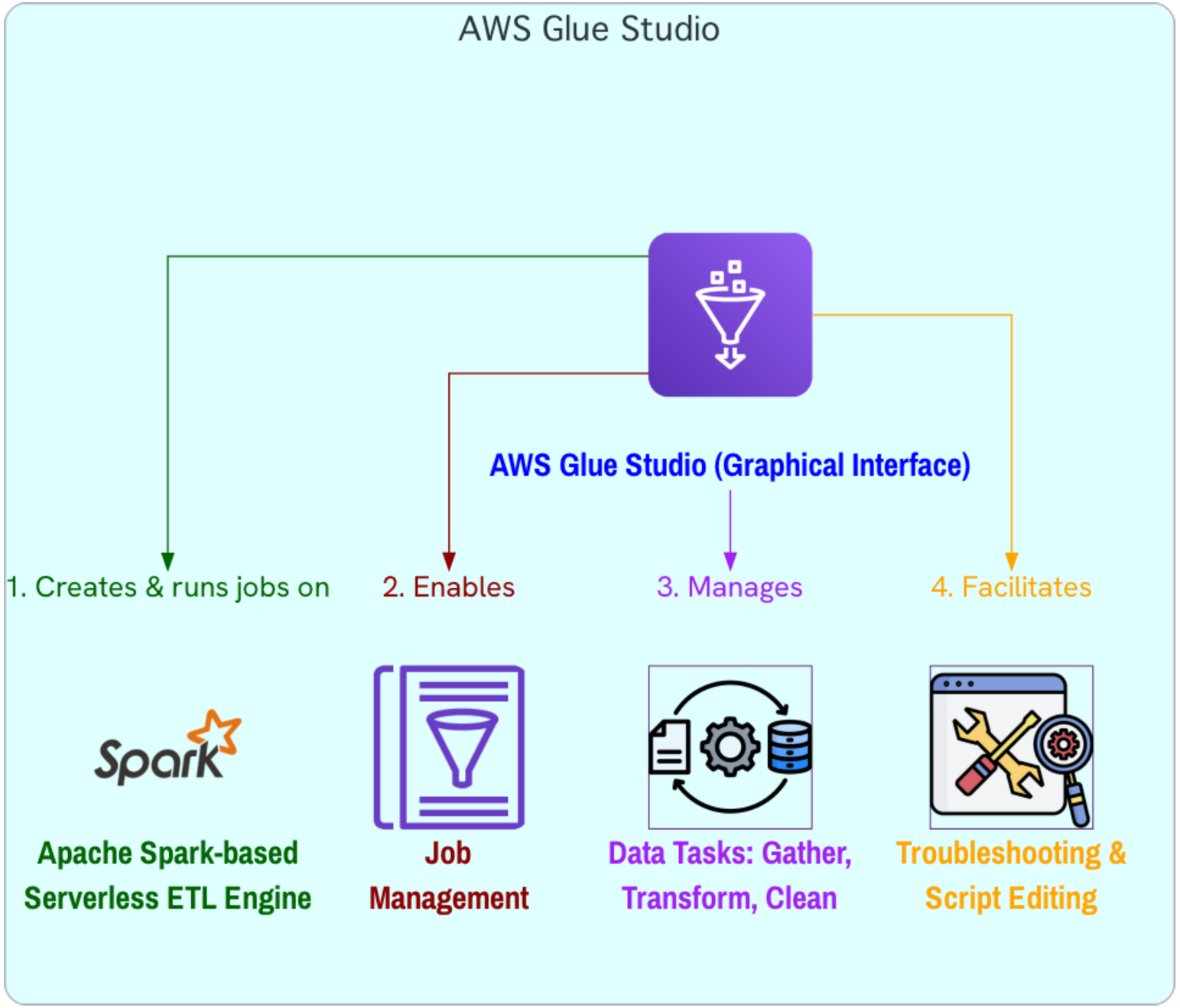
3️⃣. 🎨 Visual composition of data transformation workflows

🖼️ Design visually

🧩 Manage complex data flows

📉 Reduce code complexity

What is AWS Glue Studio?



4️⃣. ⚡ Apache Spark-based serverless ETL engine

🚀 Efficient, scalable processing

☁️ No infrastructure management

5️⃣. 📊 Gather, transform, and clean data

📥 Collect from various sources

🔄 Apply transformations

🧼 Clean for quality and consistency

6️⃣. 🛠️ Troubleshoot and edit job scripts

🐞 Debug tools

✏️ Modify scripts

🔧 Fine-tune processes

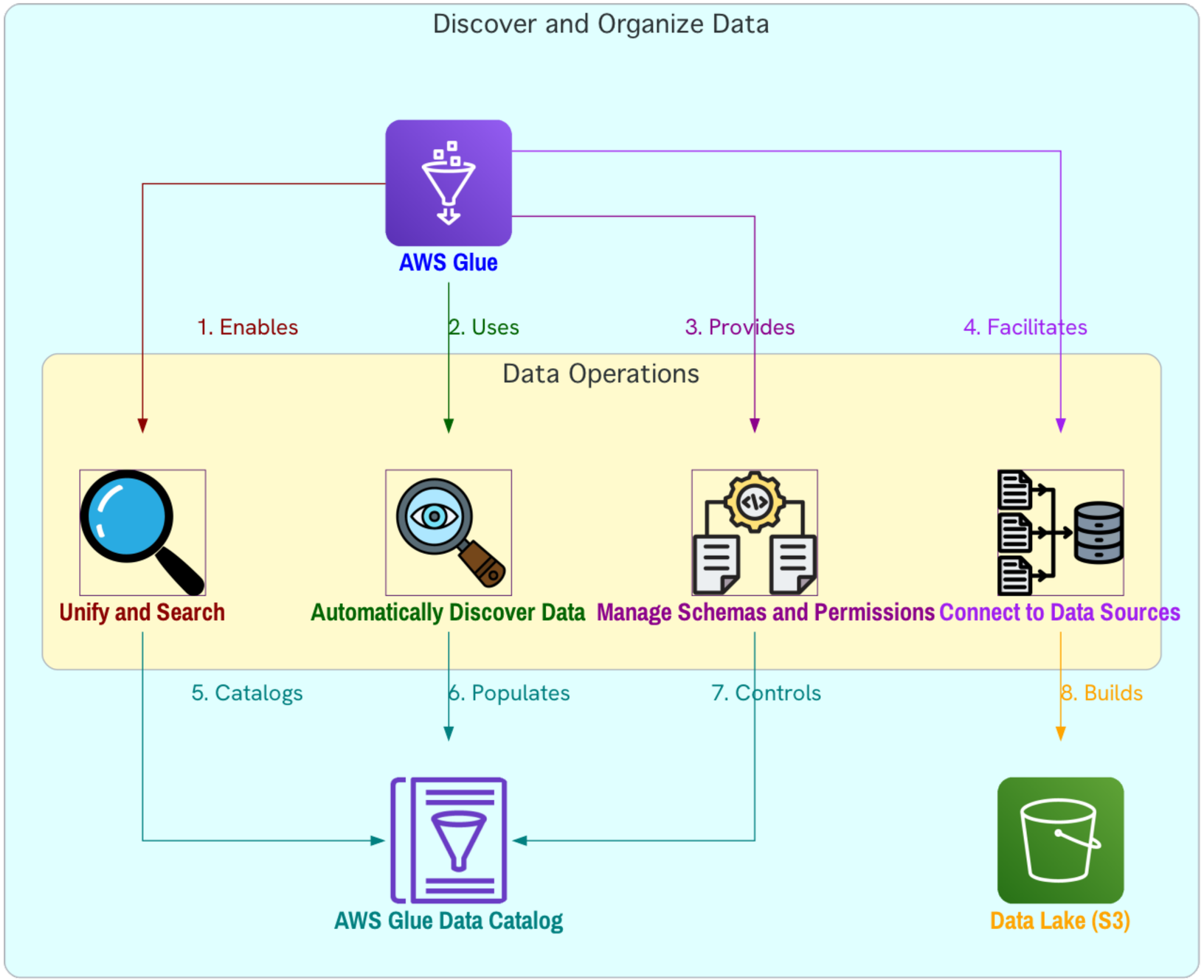
AWS Glue features - 3 categories

Discover and organize data

Transform, prepare, and clean data for analysis

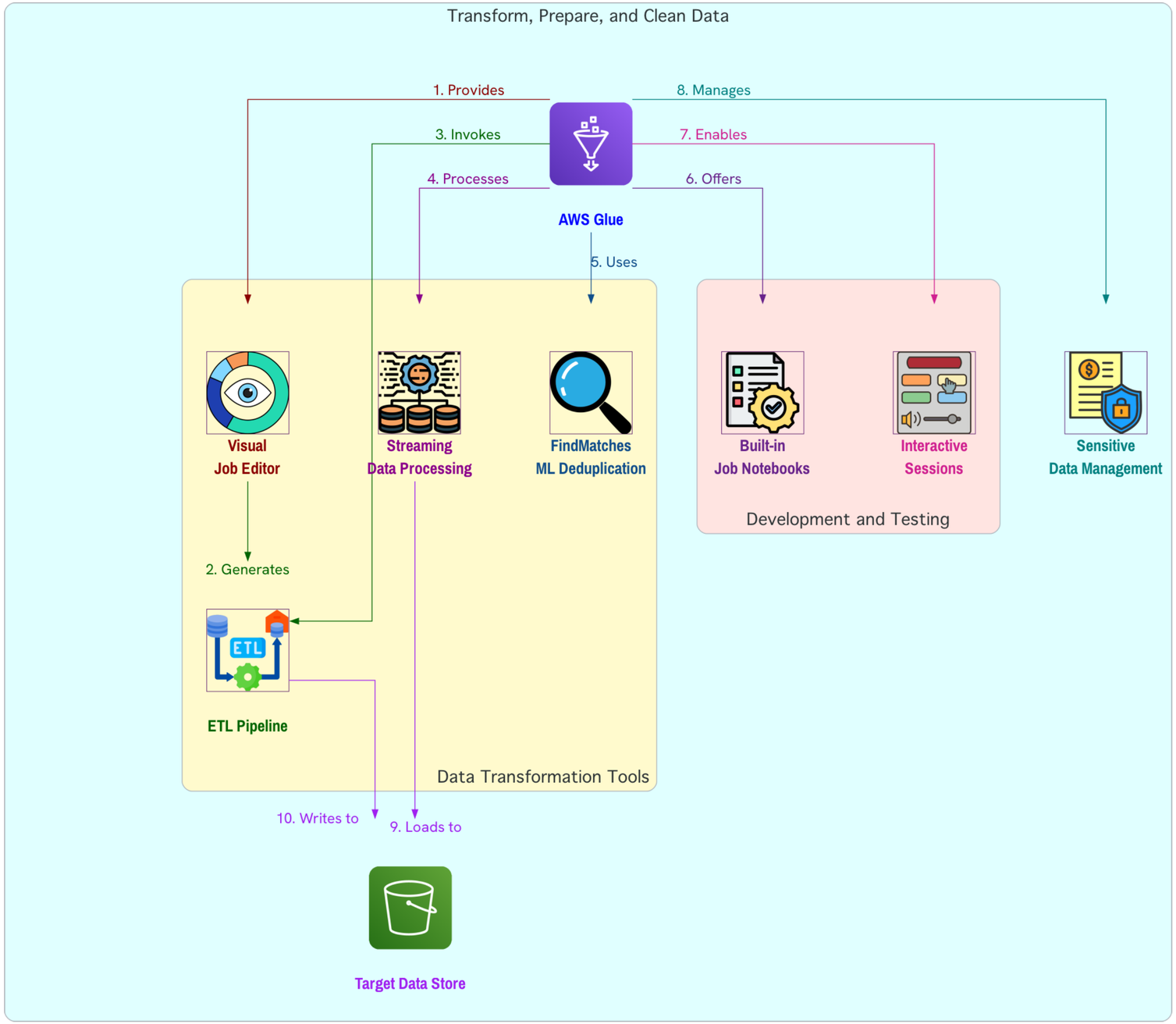
Build and monitor data pipelines

Discover and organize data



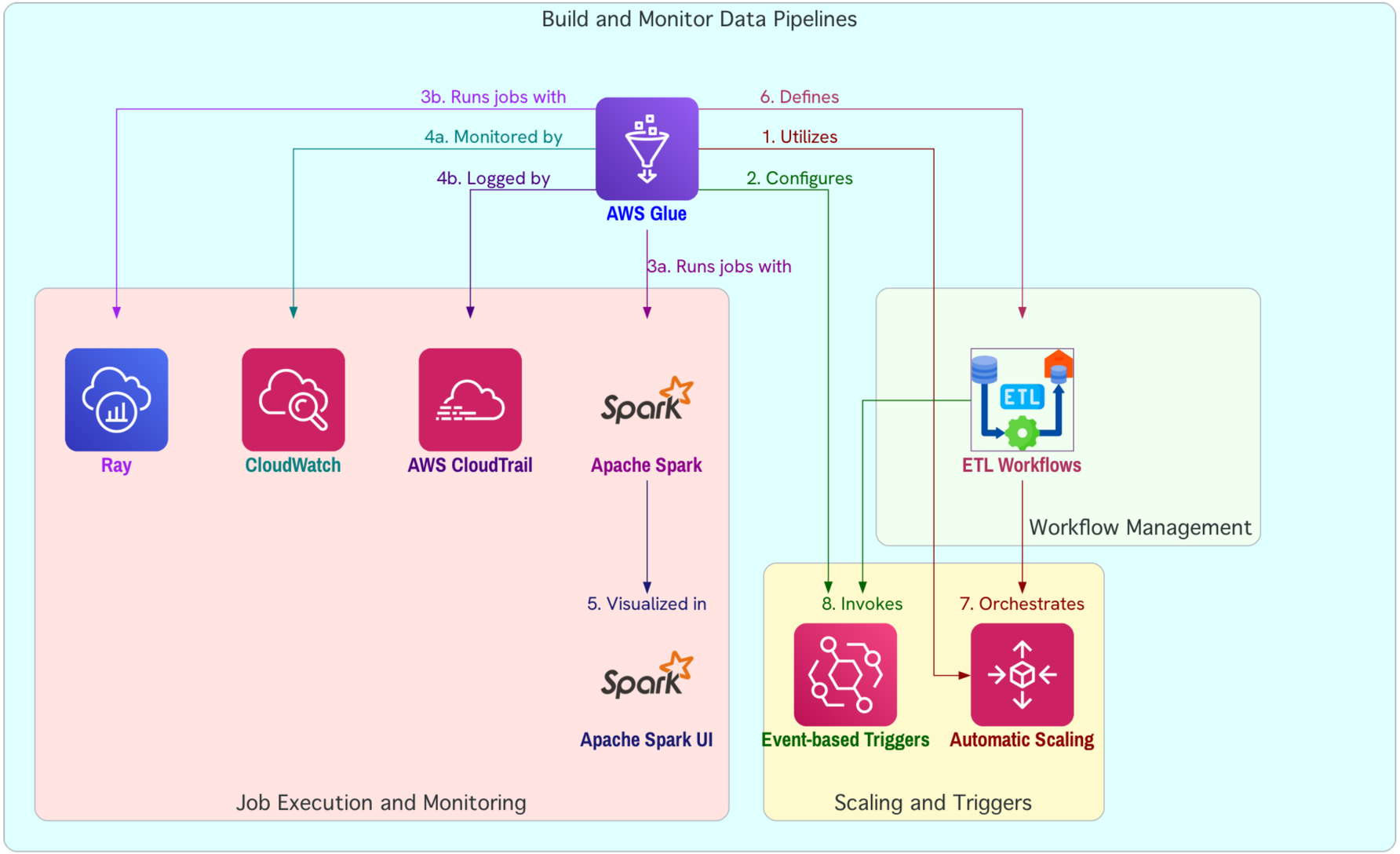
1️⃣. 🔍 Discover and organize data	
🗂️ Unify and search	📊 Store, index, search data
	📚 Catalog in AWS
🕷️ Automatically discover data	🔬 Infer schema with crawlers
	📥 Integrate into Data Catalog
🔐 Manage schemas and permissions	✅ Validate access
	🛡️ Control database/table access
🔌 Connect to data sources	🏢 On-premises sources
	☁️ AWS sources
	🌊 Build data lake

Transform, prepare, and clean data for analysis



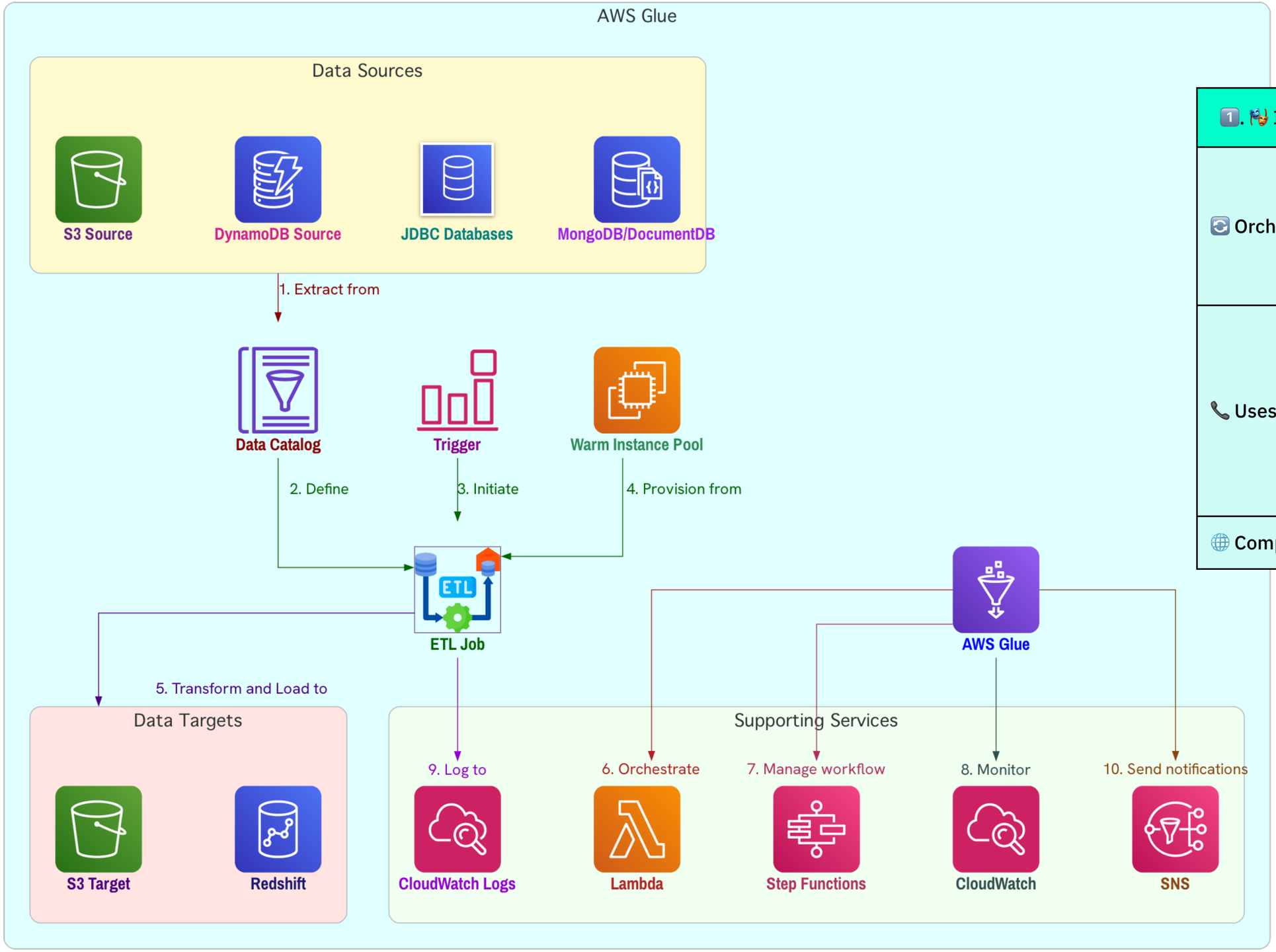
2. 🧹 Transform, prepare, and clean data	
🎨 Visual transformation	✏️ Define ETL in visual editor
	🤖 Auto-generate ETL code
🔄 Complex ETL pipelines	⏰ Schedule-based jobs
	🖱️ On-demand jobs
	🤹 Event-based jobs
🌊 Clean streaming data	📩 Continuous data consumption
	🧼 Clean/transform in transit
	⚡ Quick analysis in target store
🧬 Deduplicate with machine learning	🔍 FindMatches feature
	🧹 Clean and prepare data
	🔀 Find imperfect matches
📓 Built-in job notebooks	☁️ Serverless notebooks
	🚀 Quick setup
🛠️ Edit, debug, test ETL code	🔬 Interactive sessions
	🔧 Explore/prepare interactively
	💻 Use preferred IDE/notebook
🔒 Sensitive data management	🏷️ Define sensitive data
	🔍 Identify sensitive data
	⚙️ Process in pipeline/data lake

Build and monitor data pipelines



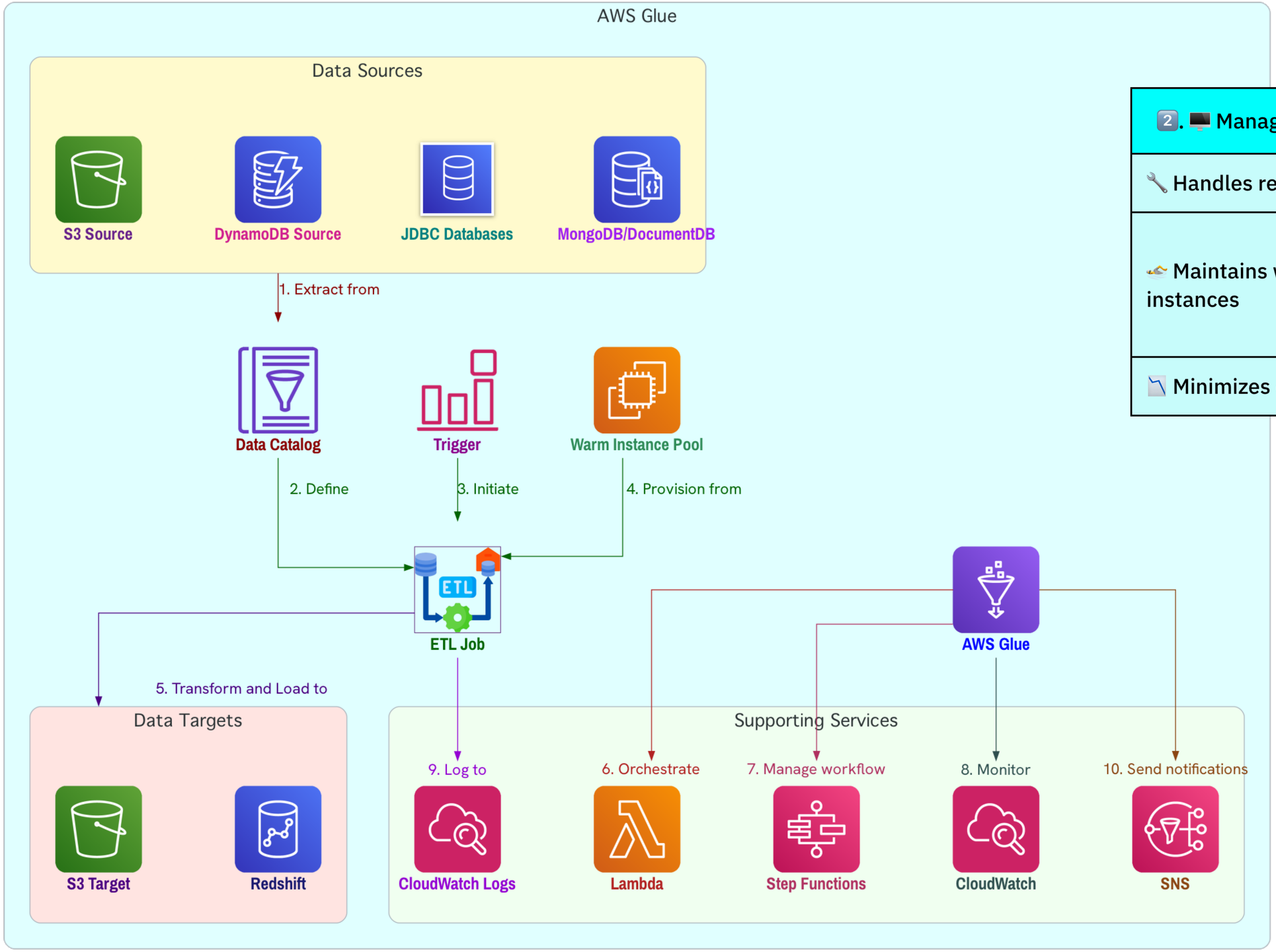
3️⃣. 🏗️ Build and monitor data pipelines	
⚖️ Automatic scaling	📈 Dynamic resource scaling
	👷 Assign workers as needed
🎬 Event-based triggers	▶️ Start crawlers/jobs
	🔗 Chain dependent jobs/crawlers
🏃 Run and monitor jobs	⚡ Spark or Ray engines
	📊 Automated monitoring tools
	🔍 Job run insights
	🧭 AWS CloudTrail
	🕸️ Apache Spark UI
🔀 Define workflows	🔄 ETL activities
	🤝 Integration activities
	🕷️ Multiple crawlers
	💼 Multiple jobs
	🚦 Multiple triggers

How it works



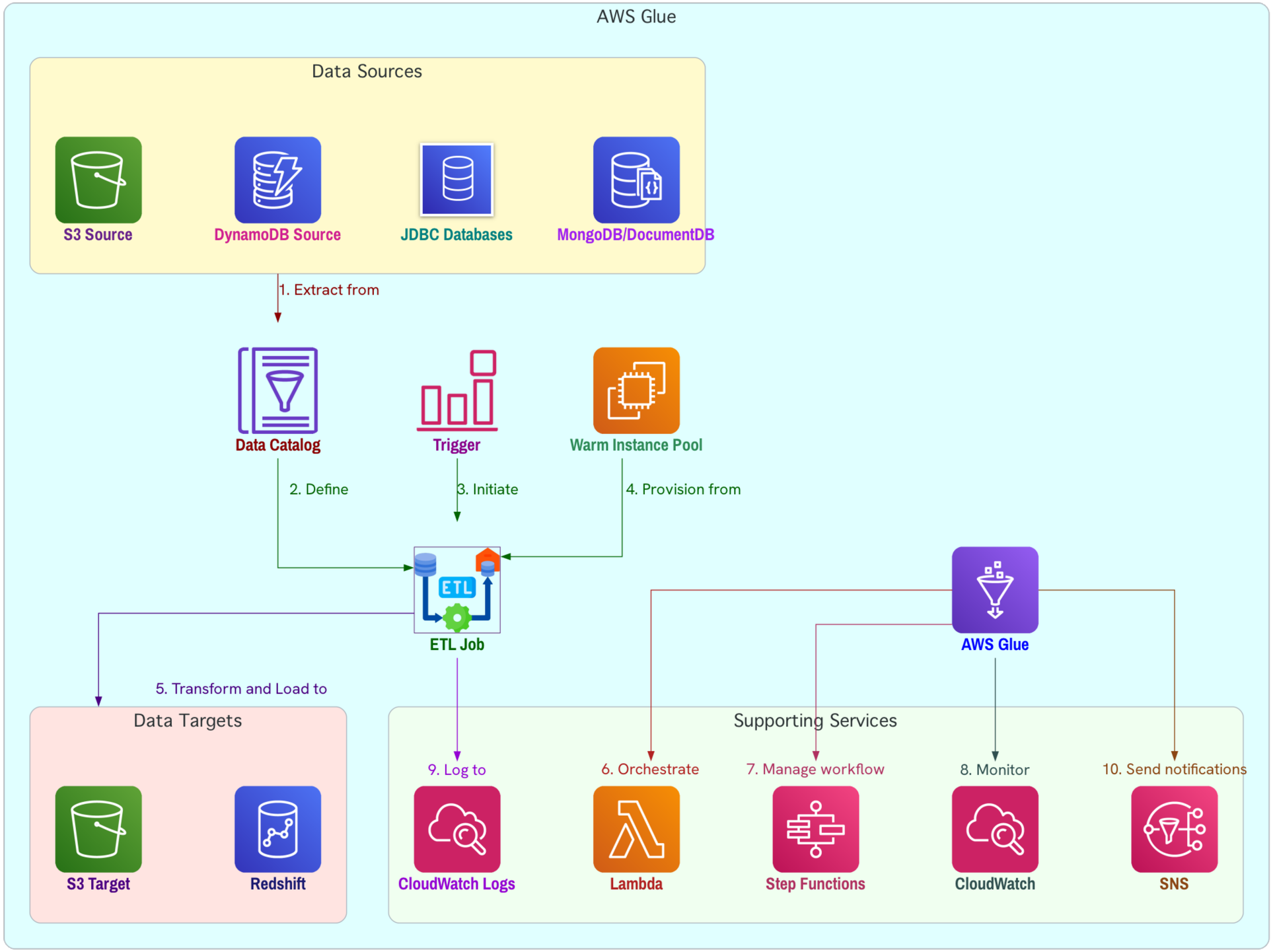
1. 🤝 Integrates AWS services for ETL operations	
🔄 Orchestrates ETL jobs	📥 Extract
	🔁 Transform
	📤 Load
📞 Uses API calls	🔄 Transform data
	📝 Create logs
	💾 Store job logic
	🔔 Generate notifications
🌐 Comprehensive ETL solution in AWS ecosystem	

How it works



2️⃣. 🖥️ Manages infrastructure automatically
🔧 Handles resource provisioning and management
🏊 Maintains warm pool of instances
⏱️ Reduces startup time
📊 Ensures efficient resource allocation
📉 Minimizes operational overhead

How it works



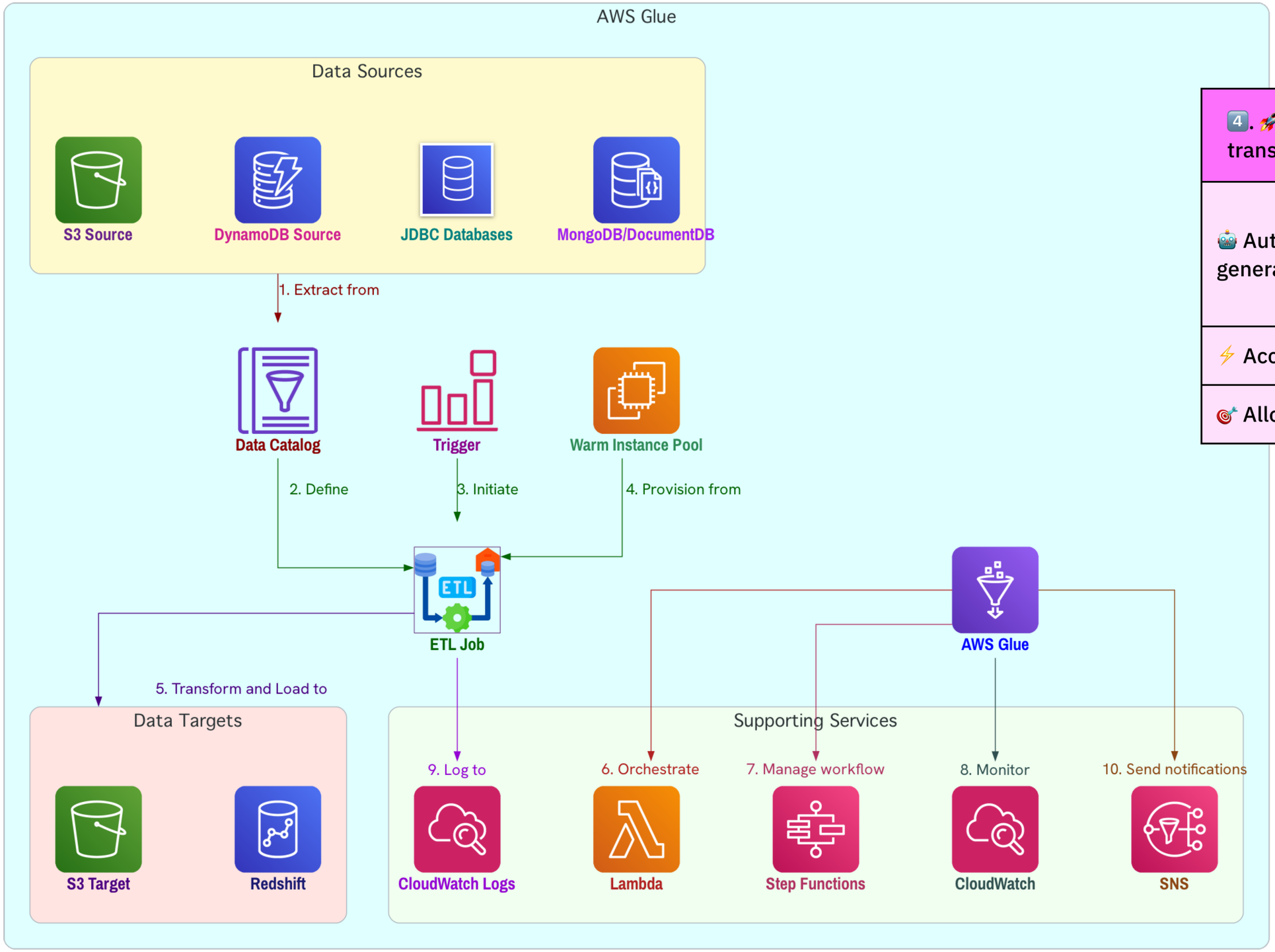
3️⃣. 📊 Utilizes Data Catalog for job structuring

📚 Central metadata repository

🏗️ Simplifies job creation and management

🗂️ Provides clear structure for data assets

How it works



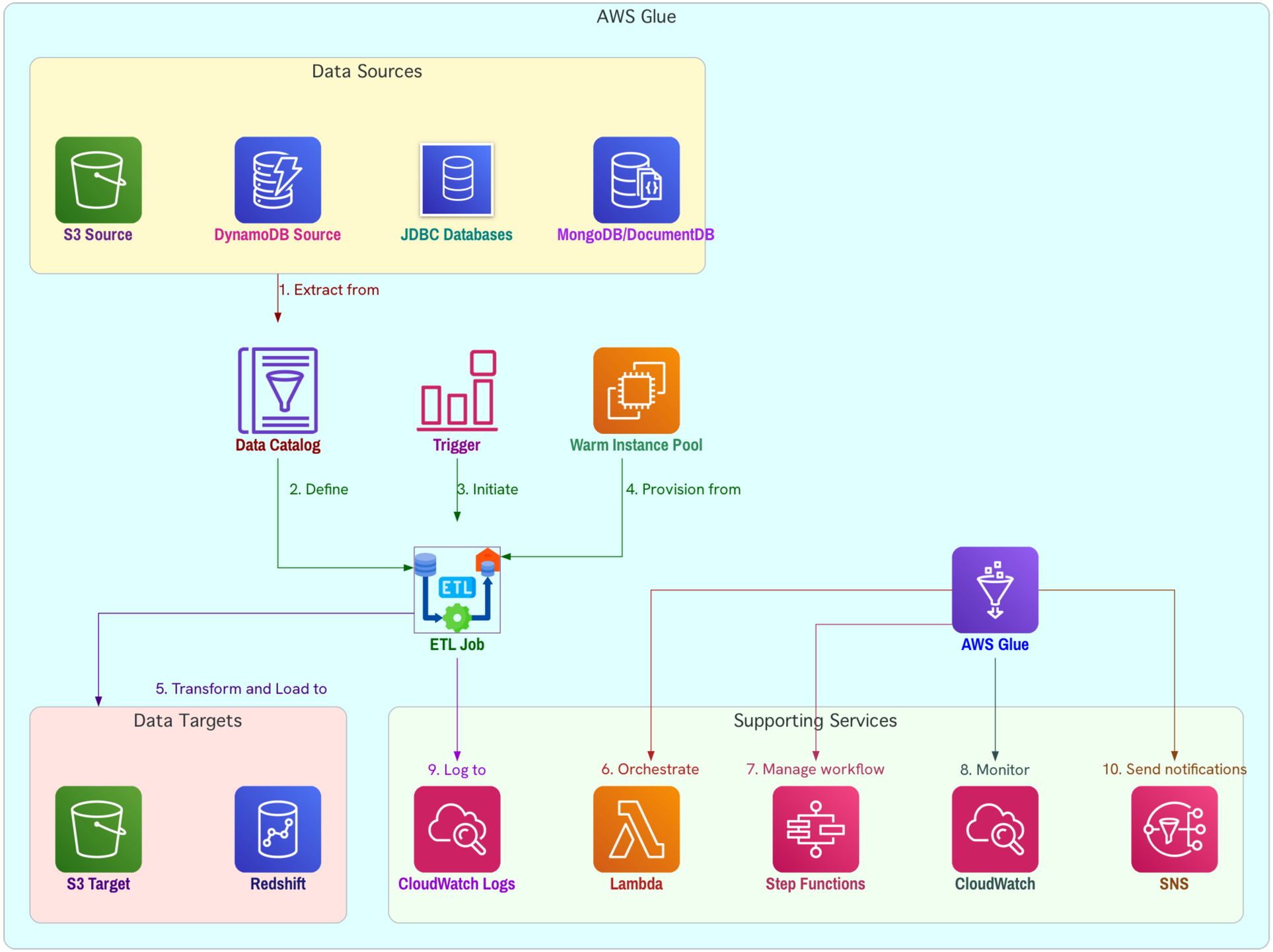
4️⃣. 🚀 Generates and executes transformation code

🤖 Automatic code generation	📩 Based on user input
	📩 For source and target data

⚡ Accelerates development process

🎯 Allows focus on data flow definition

How it works



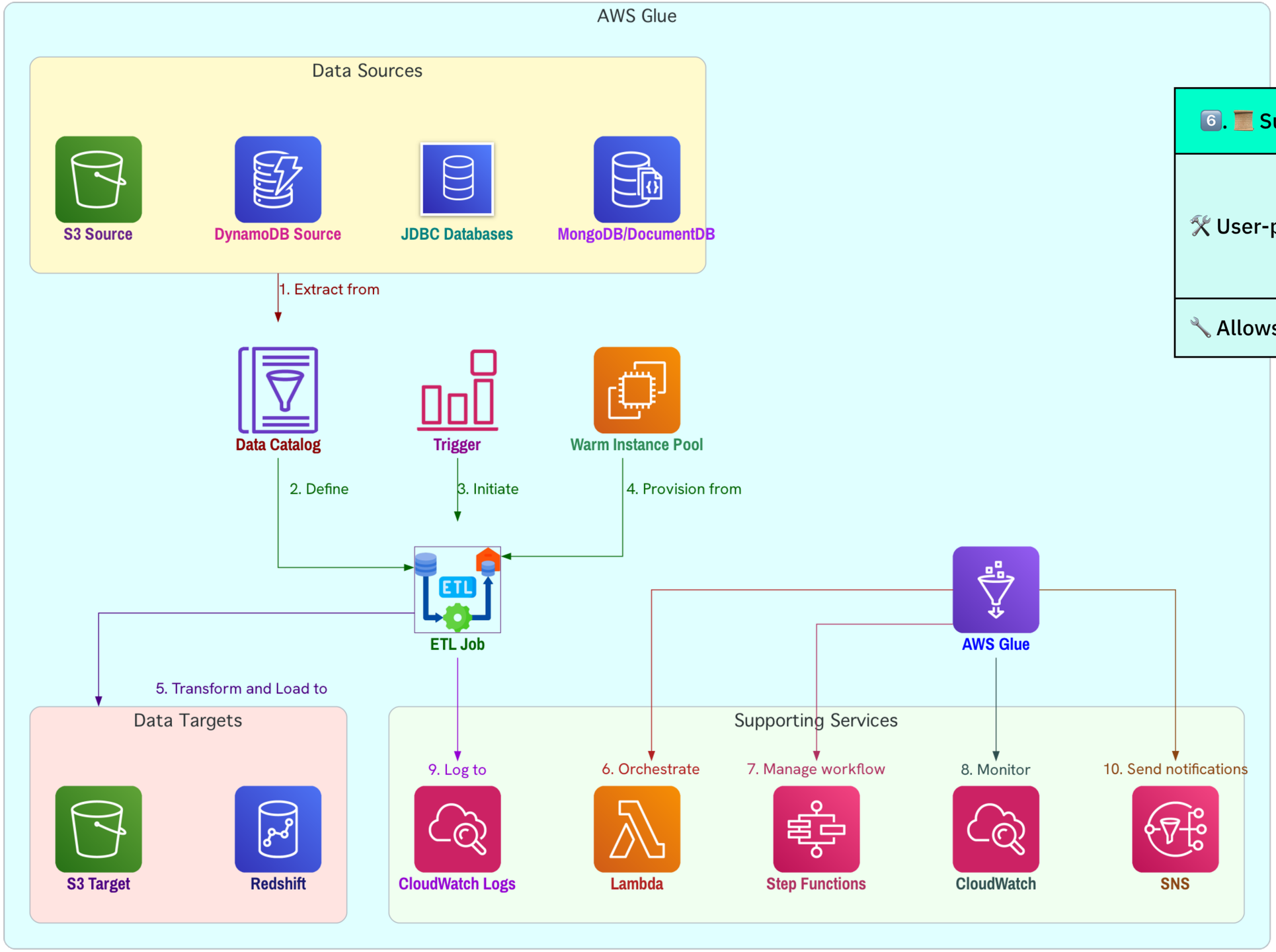
5️⃣. ⏱️ Schedules and triggers jobs flexibly

🗓️ Predefined schedules

🤹 Event-driven triggers

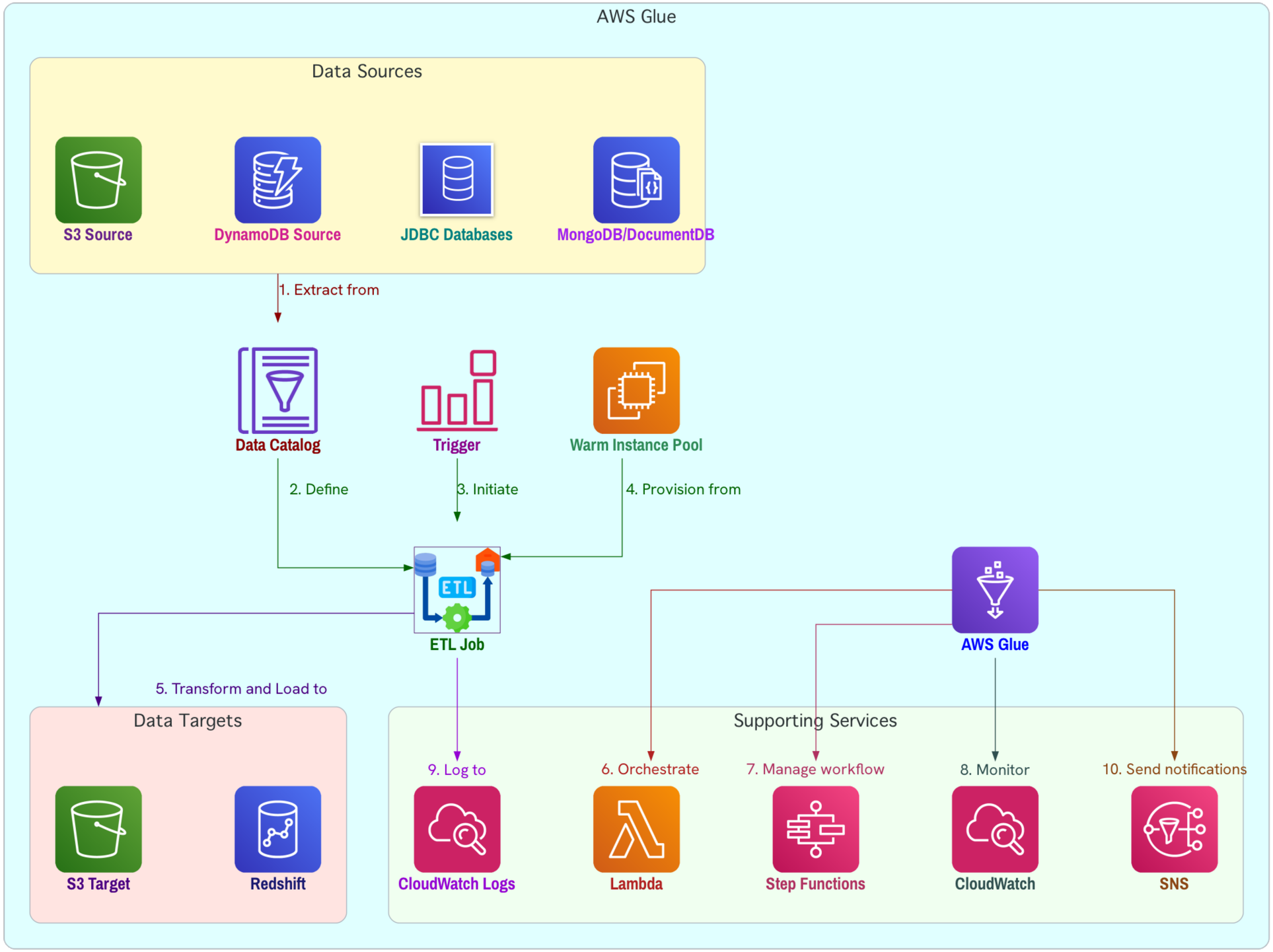
🔄 Supports batch and event-driven architectures

How it works



6. 📜 Supports custom scripts and logic	
🛠️ User-provided scripts	💻 Through AWS Glue console
	🔌 Via API
🔧 Allows complex, customized ETL processes	

How it works



7️⃣. 🔄 Monitors and logs job performance

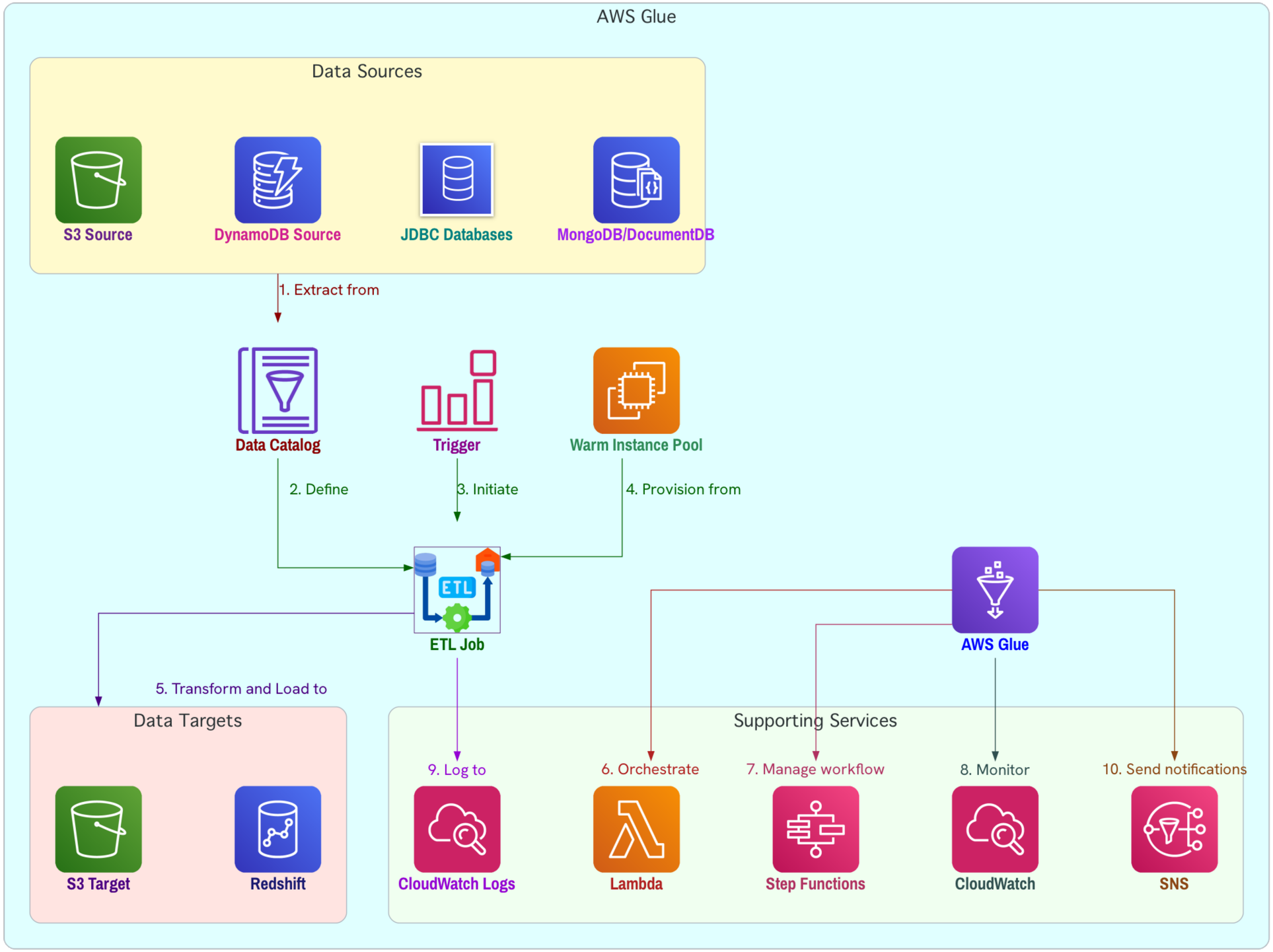
📊 Creates runtime logs

🔔 Generates notifications

👁️ Provides visibility into ETL processes

🔍 Enables quick issue identification

🚀 Facilitates performance optimization



How it works

8️⃣. 💼 Streamlines data warehouse and lake creation
🏗️ Simplifies building process
🌊 Supports output stream generation
⚡ Enables real-time data processing
📊 Feeds into AWS analytics services

AWS Integrations

Data Sources and Destinations

Amazon S3

Amazon DynamoDB

Amazon Redshift

Amazon RDS

JDBC Databases

MongoDB/DocumentDB

Marketplace Connectors

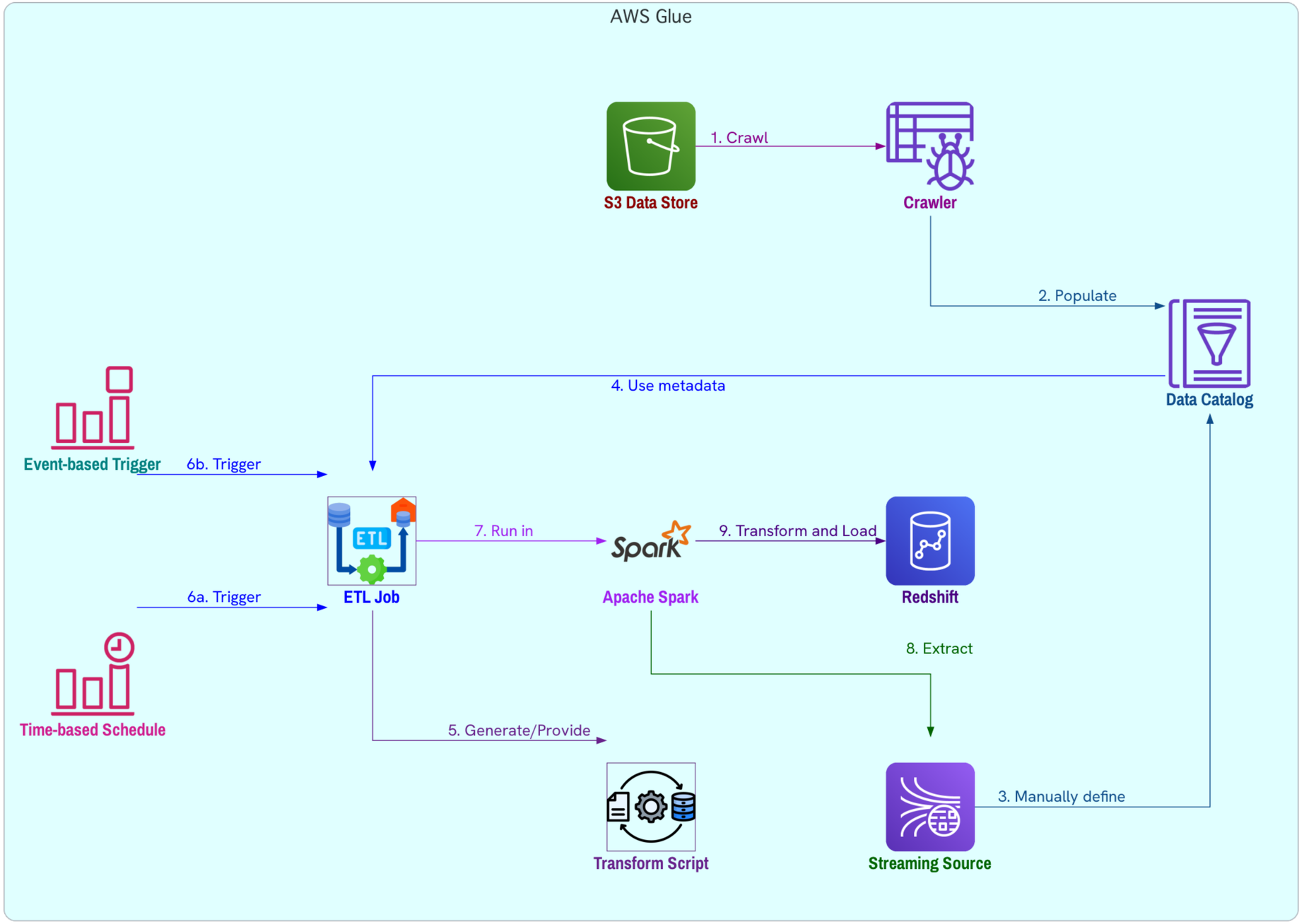
Apache Spark Plugins

Data Streams

Kinesis Data Streams

Apache Kafka

Architecture of an AWS Glue environment



1️⃣. 📋 Define ETL jobs for data processing

- 🔄 Perform Extract, Transform, Load
- 🚚 Move data efficiently
- 🎯 Apply necessary transformations

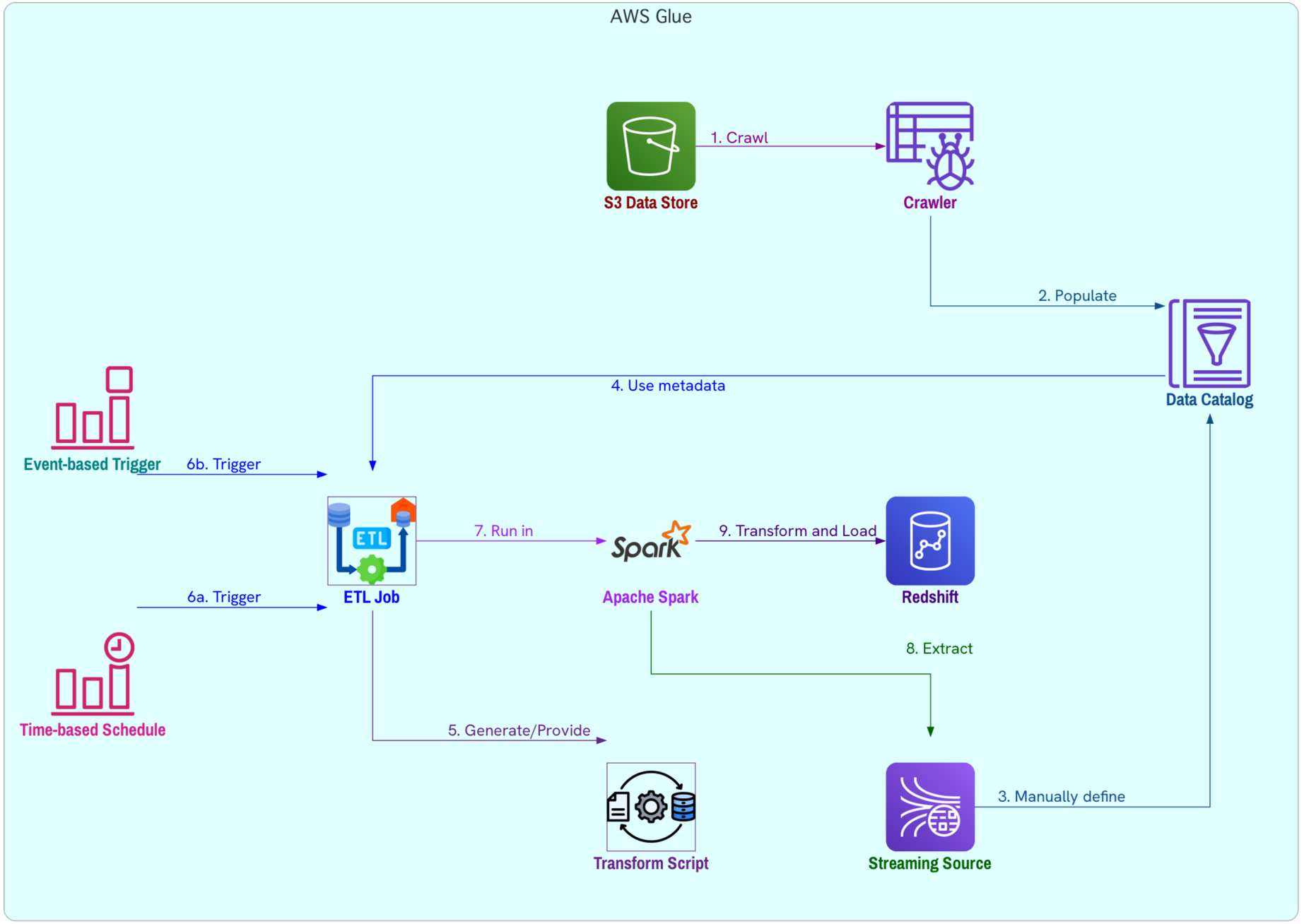
2️⃣. 🕷️ Set up crawlers for data store sources

- 🤖 Automatically populate Data Catalog
- 🔍 Scan data store
- 📊 Create table definitions with metadata

3️⃣. 📊 Manually define tables for streaming sources

- 🌊 Work with streaming data
- 📝 Specify data stream properties
- 🧩 Ensure correct data interpretation

Architecture of an AWS Glue environment



4️⃣. 📚 Leverage AWS Glue Data Catalog metadata

- 🏛️ Central repository for definitions
- 🔑 Essential for defining ETL jobs
- 🏗️ Provide data structure information

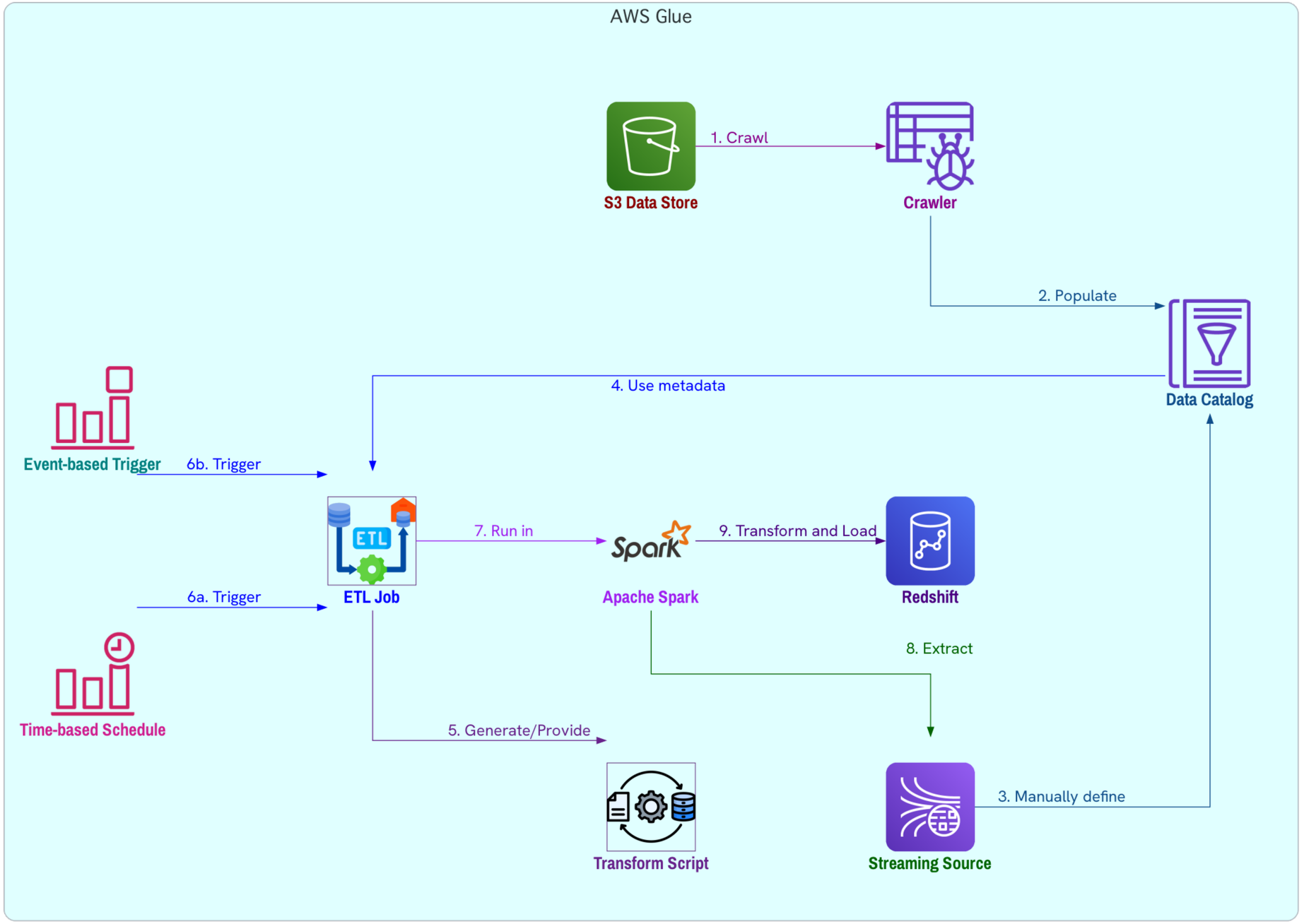
5️⃣. 📜 Generate or provide transformation script

- 🤖 Automatic script generation
- ✍️ Custom script provision
- 🛠️ Complex data transformations

6️⃣. ⏰ Configure job triggers

- 🕒 Time-based triggers
- 🎭 Event-based triggers
- 🔧 Flexible scheduling options

Architecture of an AWS Glue environment



7️⃣. 🔄 Execute job: Extract, Transform, Load

- 📩 Extract data from source
- 🔄 Apply transformations
- 📩 Load to target destination

8️⃣. ⚡ Process data in Apache Spark environment

- 🌟 Leverage Spark's capabilities
- 📊 Efficient, scalable processing
- 🚀 Optimize ETL workloads

Thanks for Watching