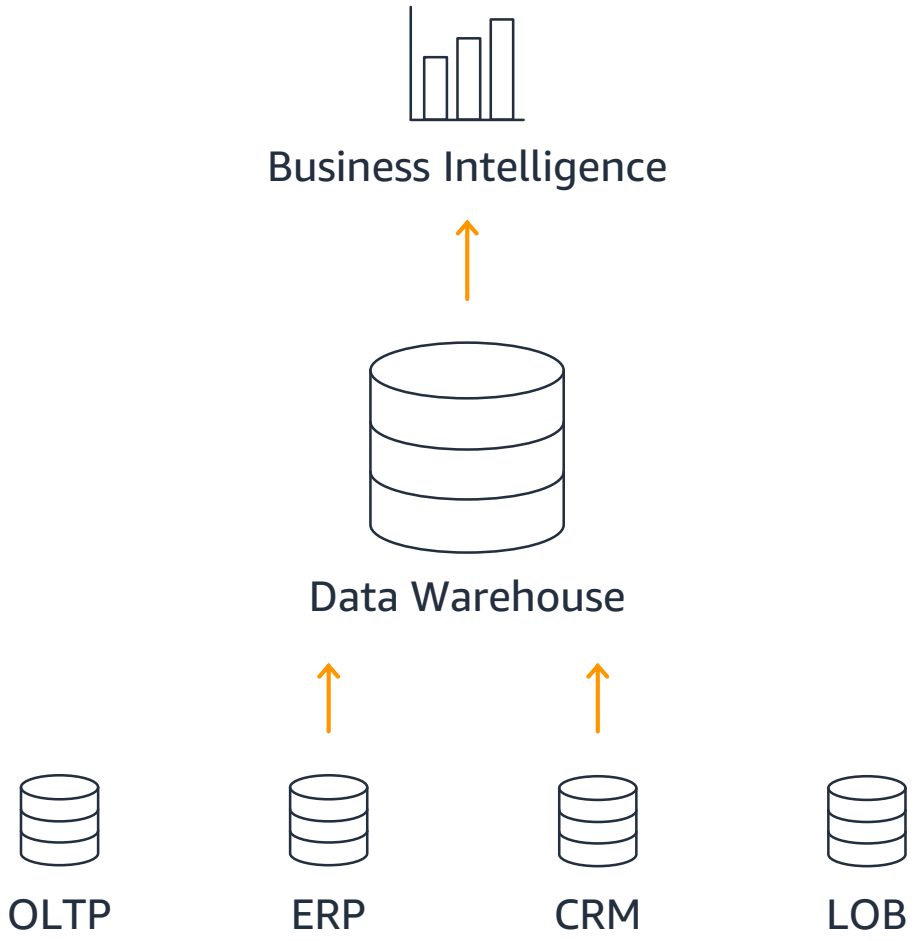
Designing Data Lakes: Best Practices (Level 200)

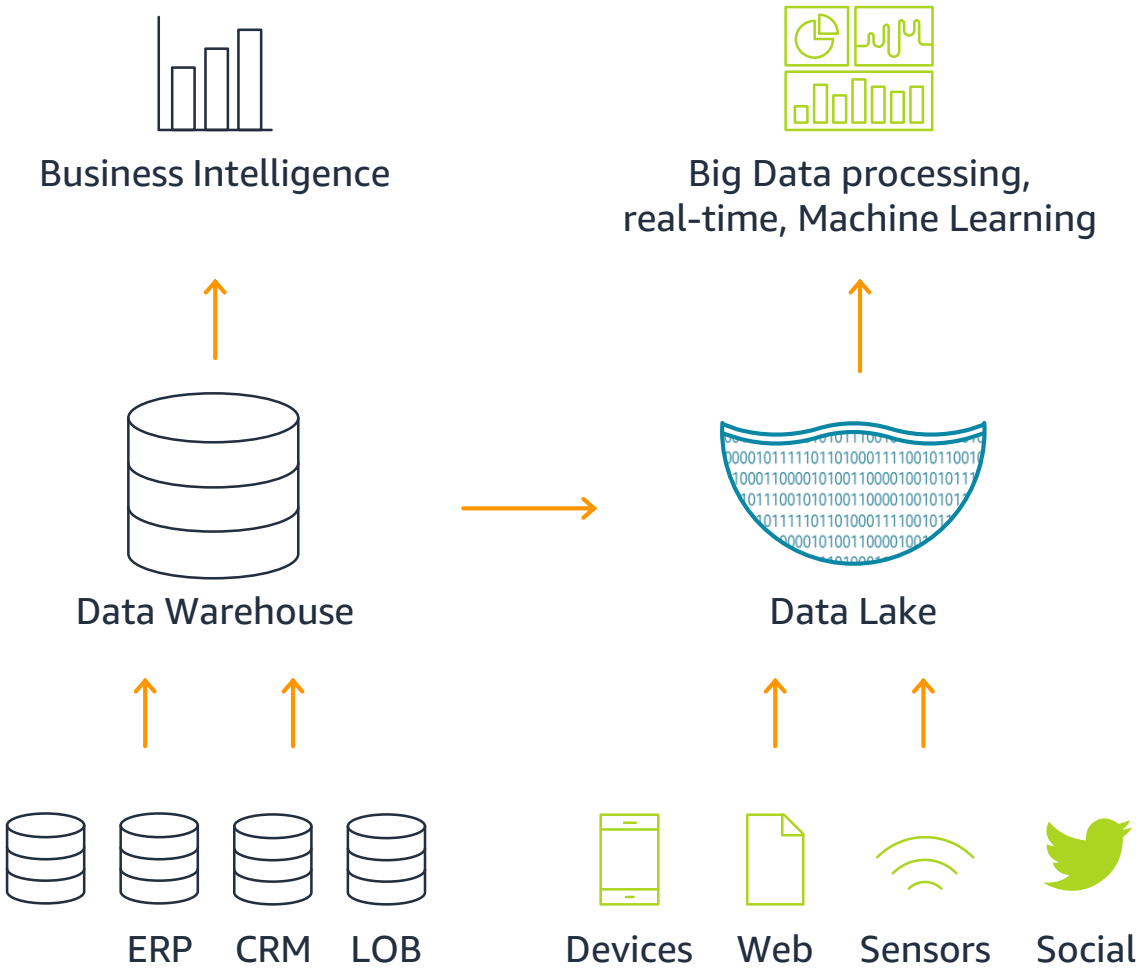
Ganesh Raja, Solution Architect

Traditionally, Analytics Used to Look Like This



- Relational data
- TBs–PBs scale
- Schema defined prior to data load
- Operational reporting and ad hoc
- Large initial CAPEX + $10K–$50K/TB/Year

Data Lakes Extend the Traditional Approach



- Relational and non-relational data
- TBs–EBs scale
- Diverse analytical engines
- Low-cost storage & analytics

Reasons for building a data lake

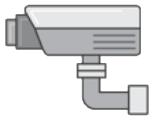
Exponential growth in data

Transactions

Billing

ERP

Web logs



Sensor Data

Infrastructure logs

Social

Reasons for building a data lake

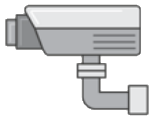
Exponential growth in data

Transactions

Billing

ERP

Web logs



Sensor Data

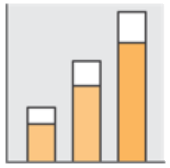
Infrastructure logs

Social

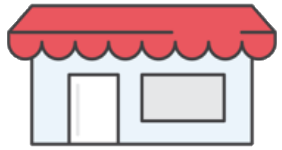
Diversified consumers

Data Scientists

Applications



Business Analyst



External Consumers

Reasons for building a data lake

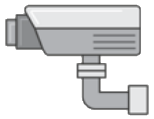
Exponential growth in data


Transactions


Billing


ERP


Web logs


Sensor Data

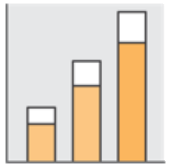

Infrastructure logs

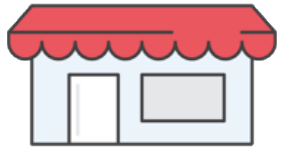

Social

Diversified consumers


Data Scientists

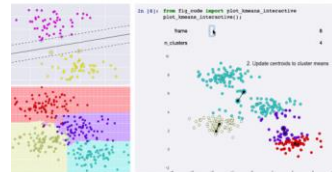

Applications


Business Analyst


External Consumers

Multiple access mechanisms


API Access


Notebooks


BI Tools

Characteristics of a data lake

Collect
Anything

Characteristics of a data lake

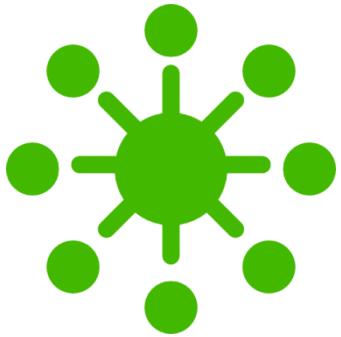
Collect Anything

Dive in Anywhere

Characteristics of a data lake

Collect Anything

Dive in Anywhere

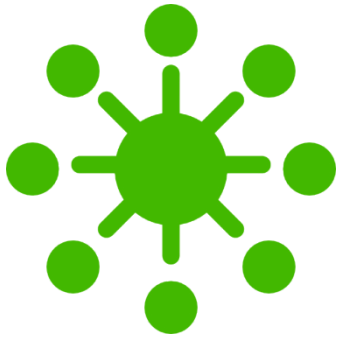


Flexible Access

Characteristics of a data lake

Collect Anything

Dive in Anywhere



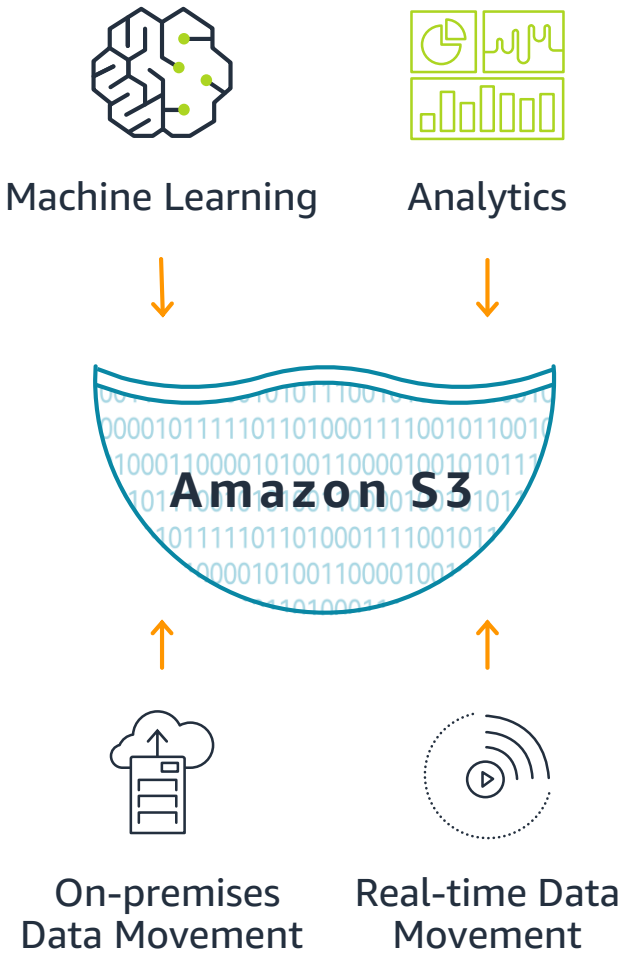
Flexible Access

Future Proof

Data Lakes and Analytics from AWS

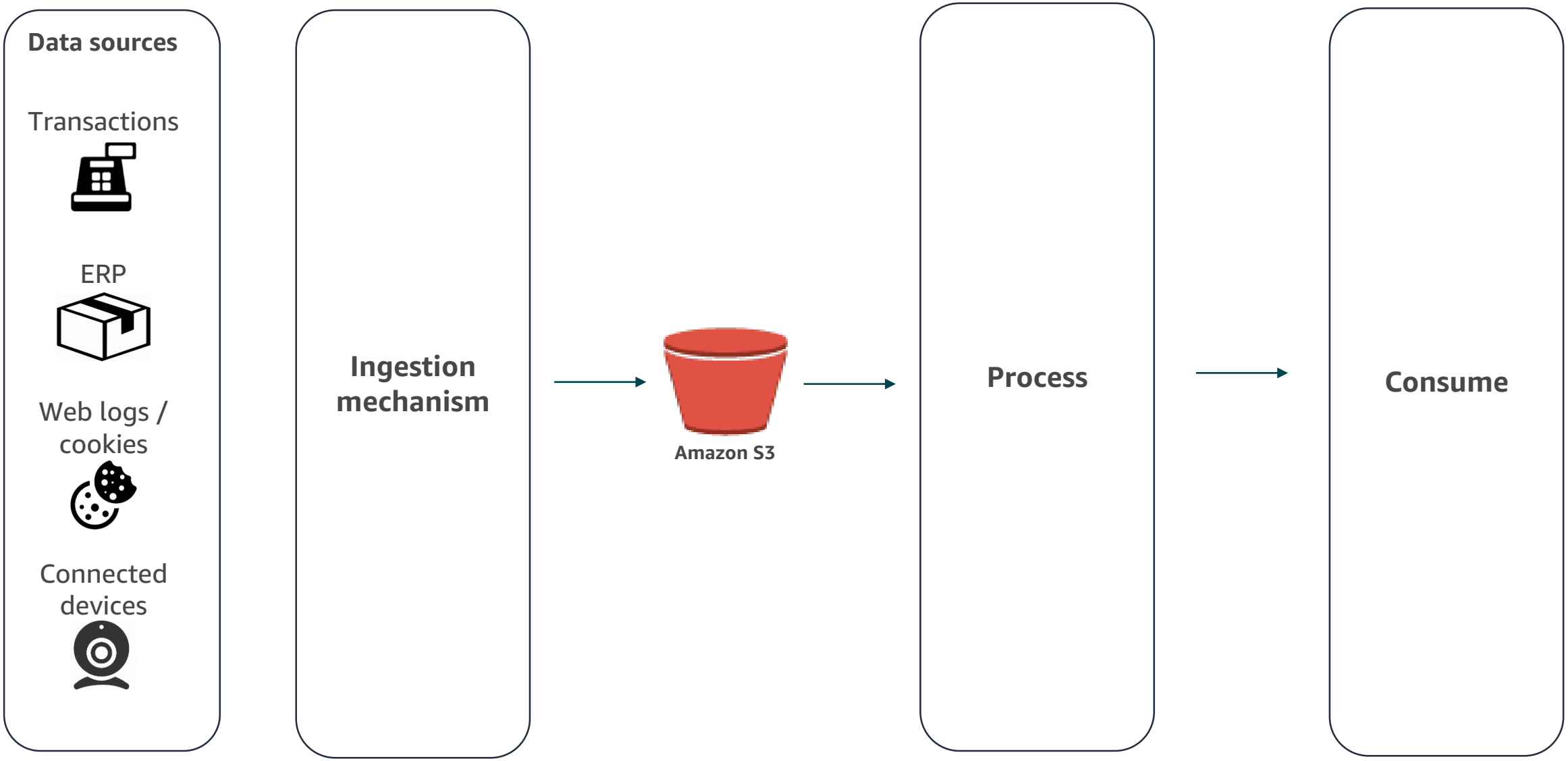
Amazon S3 as the data lake

Data Lakes on AWS

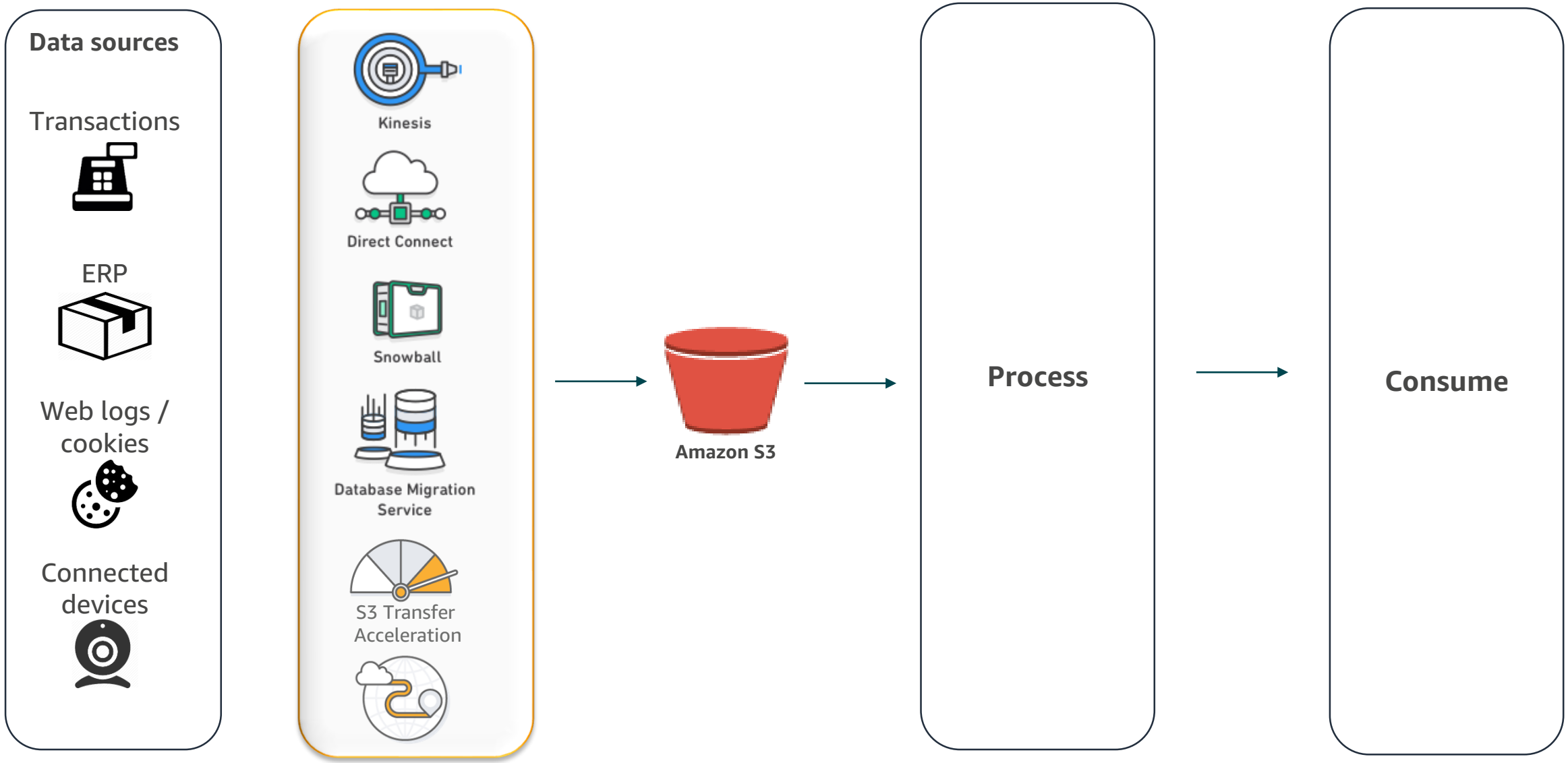


- Unmatched durability, and availability at EB scale
- Best security, compliance, and audit capabilities
- Object-level controls for fine-grain access
- Fastest performance by retrieving subsets of data
- The most ways to bring data in
- 2x as many integrations with partners
- Analyze with broadest set of analytics & ML services

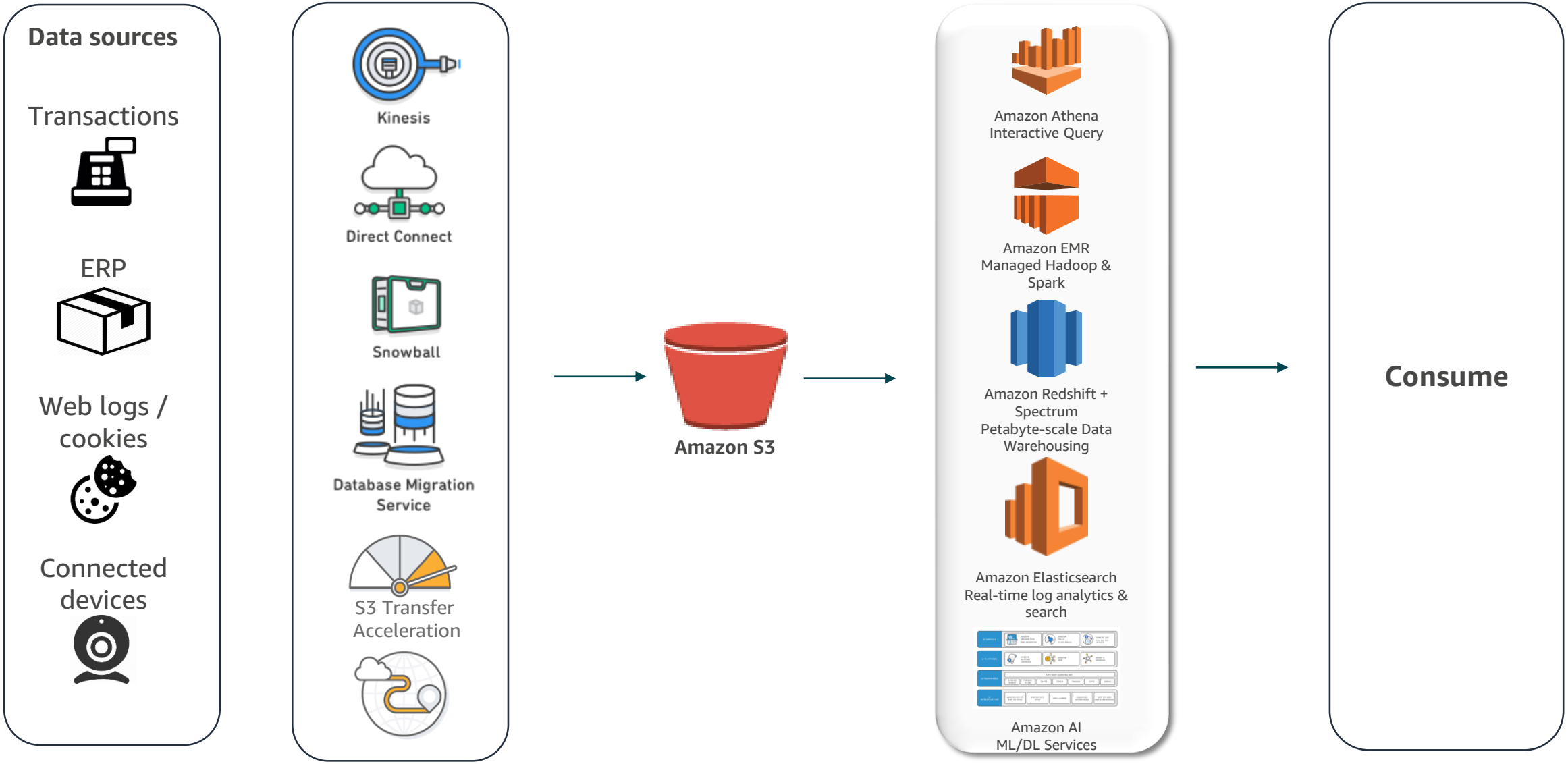
Simplified architectural view



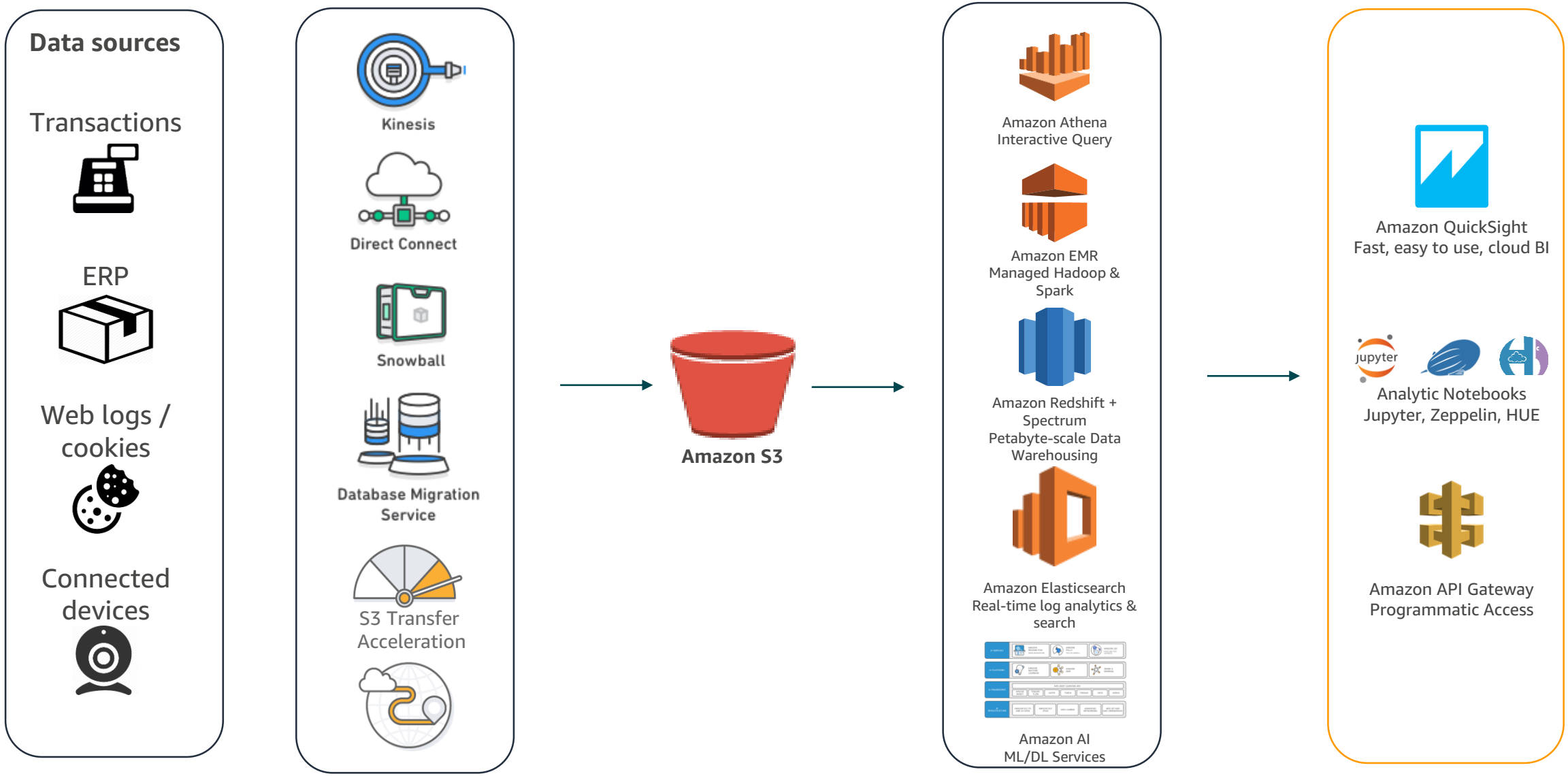
There are lots of ingestion tools



Variety of data processing tools



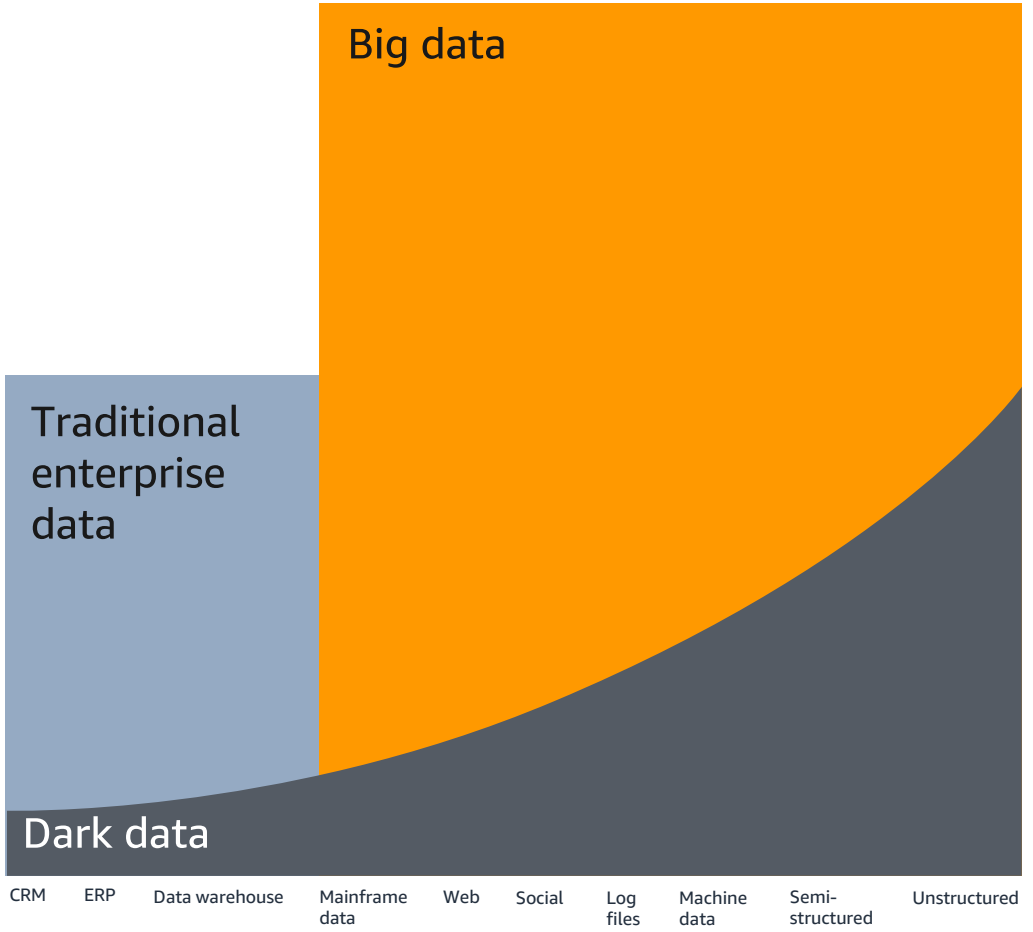
And multiple ways to consume the data



Storing is Not Enough, Data Needs to Be Discoverable

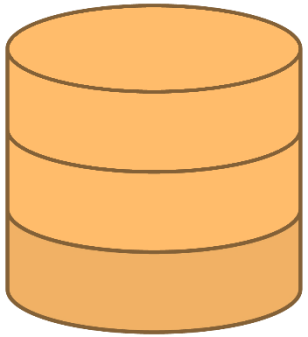
"Dark data are the information assets organizations collect, process, and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing)."

Gartner IT Glossary, 2018
https://www.gartner.com/it-glossary/dark-data



AWS Glue Data Catalog

Make data discoverable



AWS Glue Data Catalog
Central Metadata Catalog for the data lake

One per account

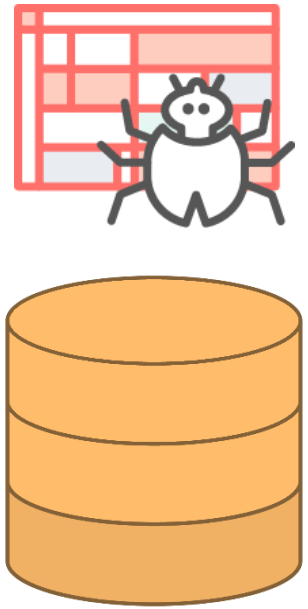
Allows you to share metadata between Amazon Athena, Amazon Redshift Spectrum, EMR & JDBC sources

We added a few extensions:

- **Search** over metadata for data discovery
- **Connection info** – JDBC URLs, credentials
- **Classification** for identifying and parsing files
- **Versioning** of table metadata as schemas evolve and other metadata are updated

AWS Glue Data Catalog

Crawlers



AWS Glue Data Catalog - Crawlers
Helping Catalog your data

Crawlers automatically build your Data Catalog and keep it in sync

Automatically discover new data, extracts schema definitions

- Detect schema changes and version tables
- Detect Hive style partitions on Amazon S3

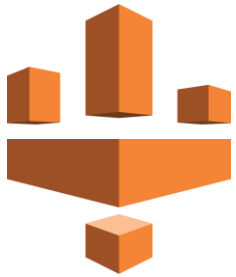
Built-in classifiers for popular types; custom classifiers using Grok expression

Run ad hoc or on a schedule; serverless – only pay when crawler runs

Because data is ~~not~~ never perfect

Because data is ~~not~~ never perfect

AWS Lambda
Trigger-based Code Execution

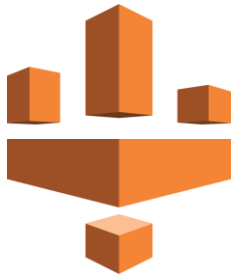


AWS Glue
Event based Server-less ETL engine

Amazon EMR
Spark and Hive running on EMR

Because data is ~~not~~ never perfect

AWS Lambda
Trigger-based Code Execution



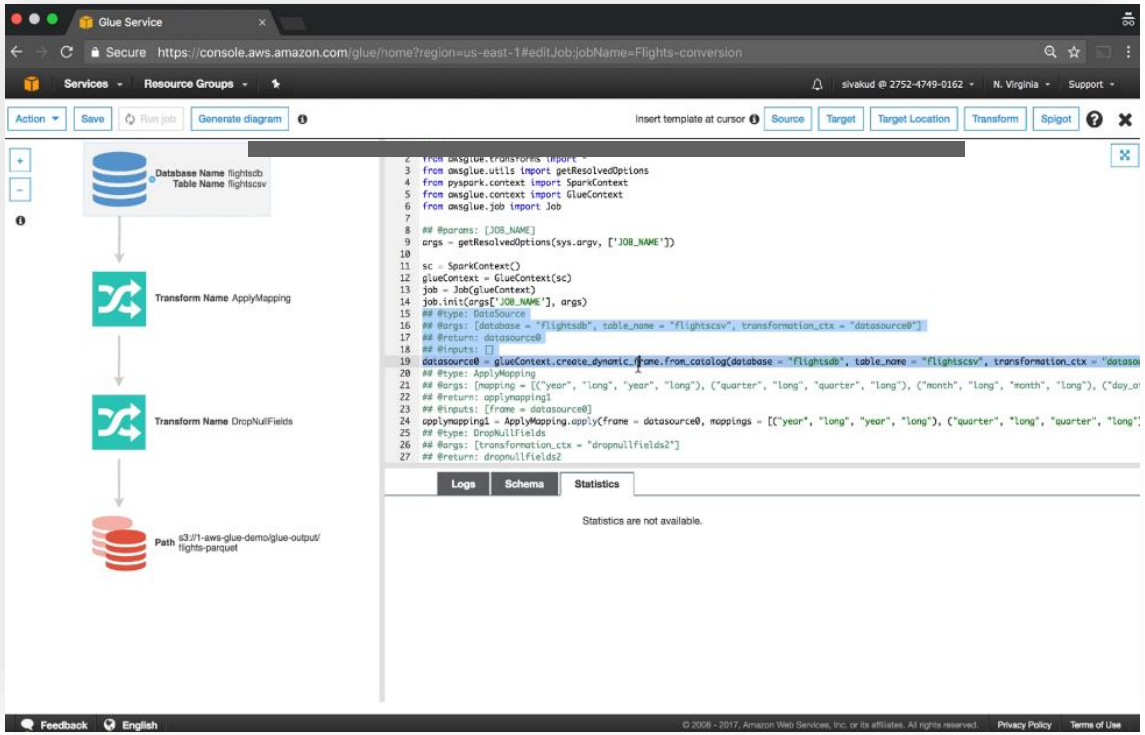
AWS Glue
Event based Server-less ETL engine

Amazon EMR
Spark and Hive running on EMR

Clean
Transform
Concatenate
Convert to better formats
Schedule transformations
Event-driven transformations
Transformations expressed as code

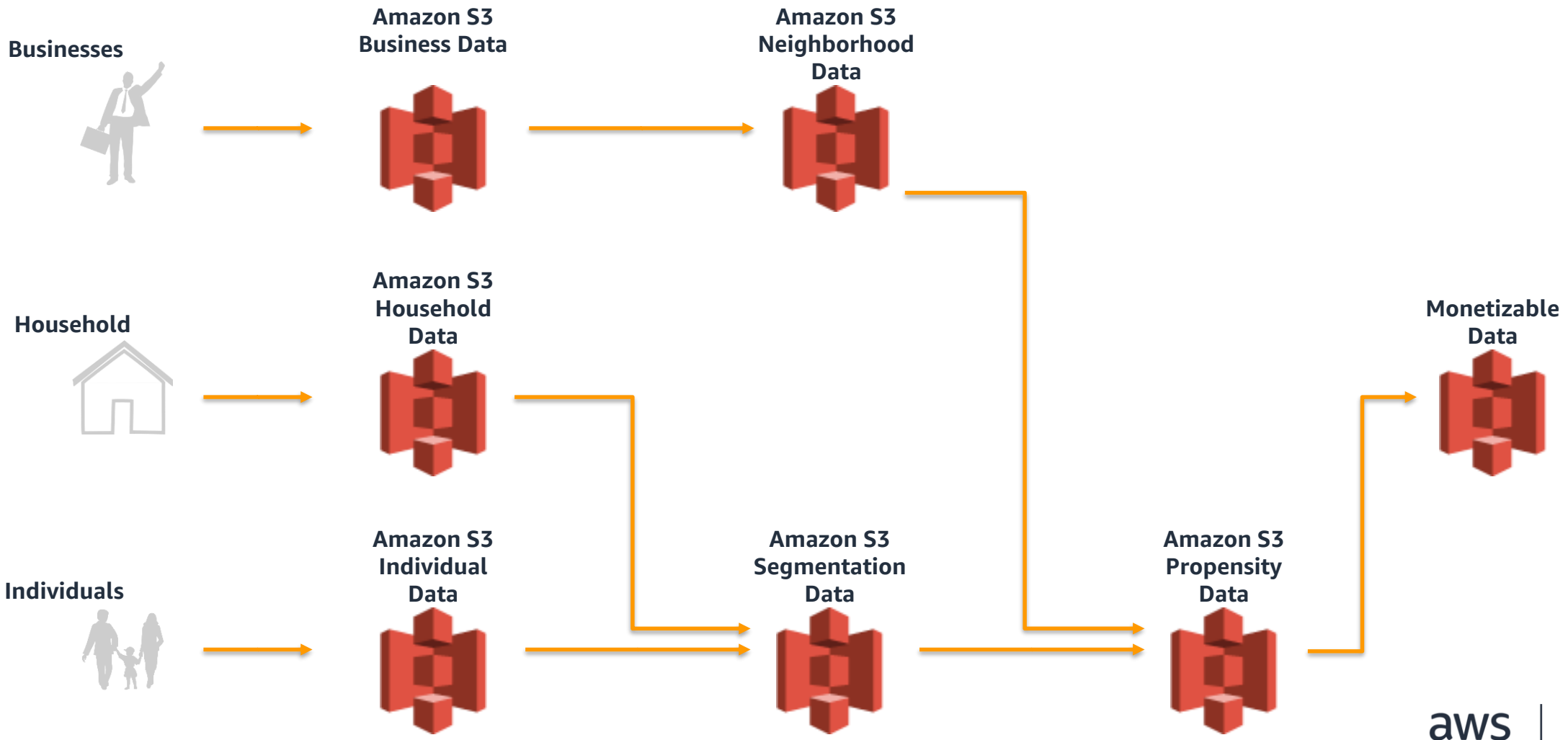
AWS Glue—ETL Service

Make ETL scripting and deployment easy

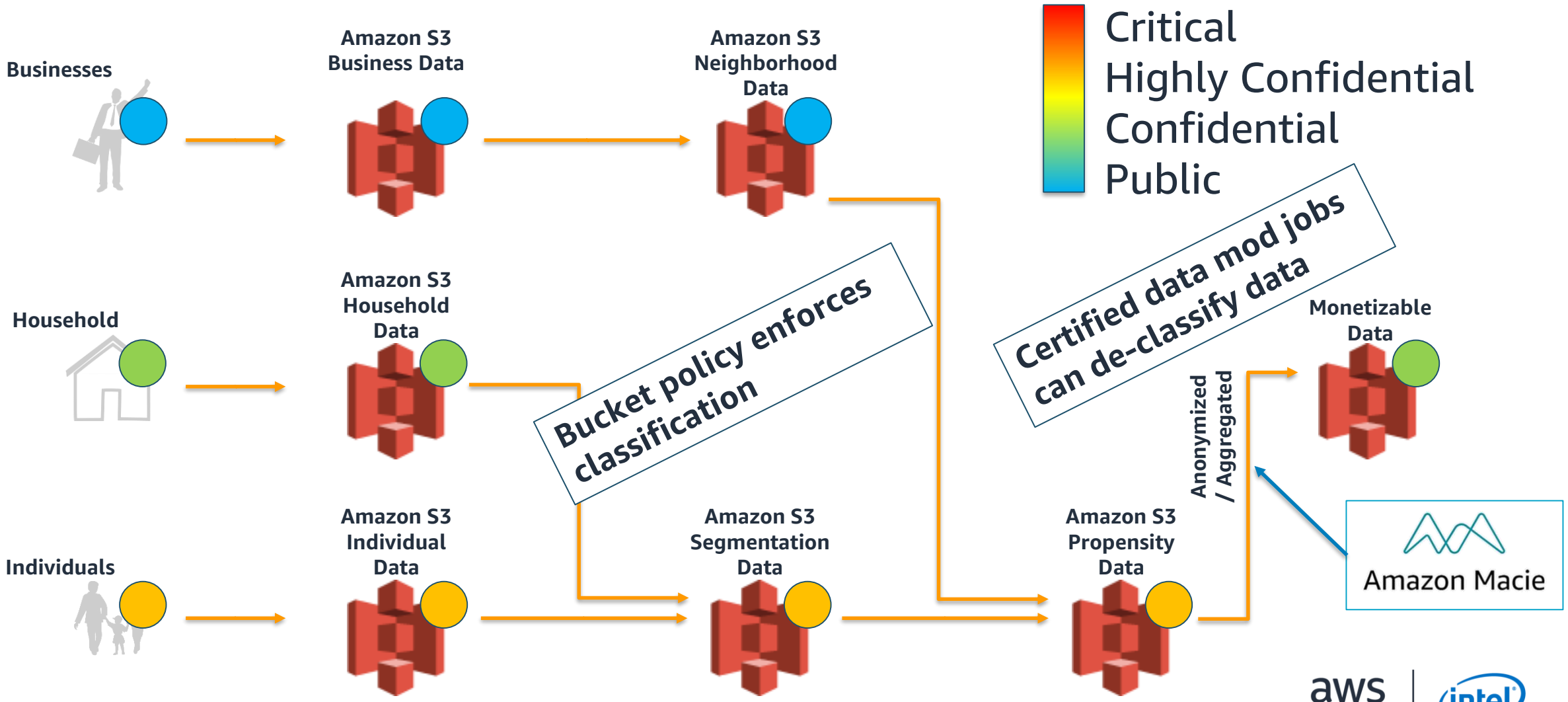


- Automatically generates ETL code
- Code is customizable with Python and Spark
- Endpoints provided to edit, debug, test code
- Jobs are scheduled or event-based
- Serverless

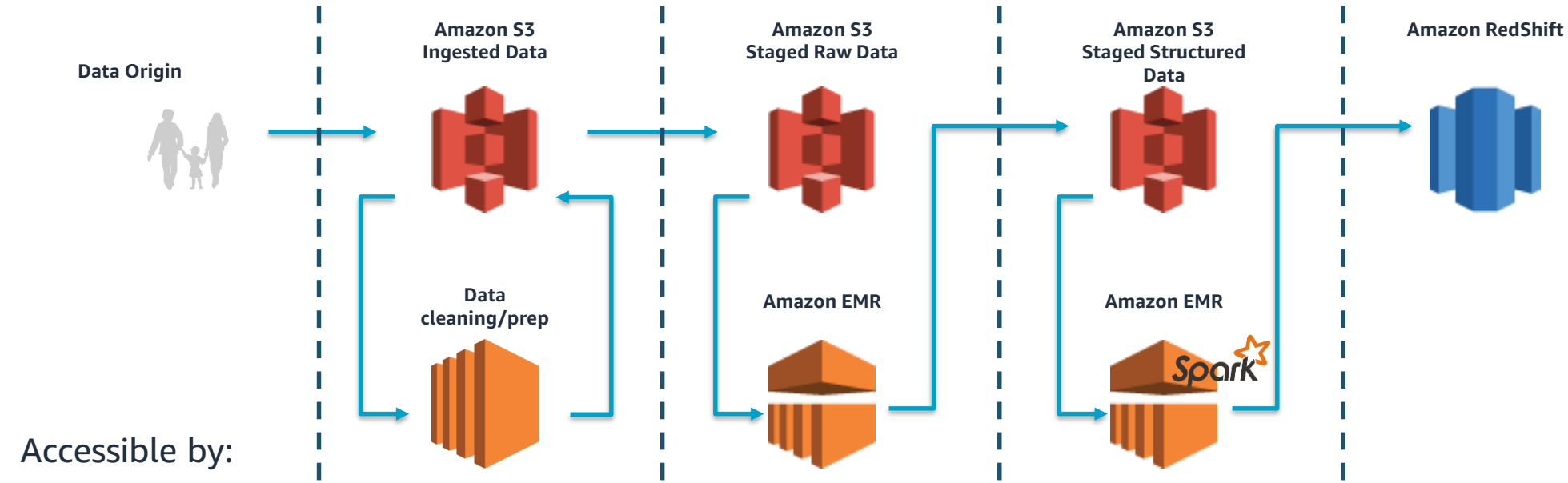
Data enrichment involves a pipelining strategy



Enriched data takes on highest classification



Access controls applied at pipeline stages



Accessible by:

Service role	Data mgmt service	Data mgmt service	Data mgmt service	Reporting service
Human role	N/A	Data scientist	Business analyst	Report builders

Data Lakes, Analytics, and ML Portfolio from AWS

Broadest, deepest set of analytic services

Machine Learning

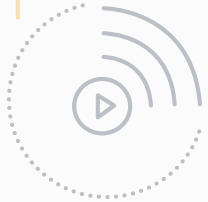
Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly

Analytics

Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch service
Amazon Kinesis
Amazon QuickSight

On-premises Data Movement

AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service
AWS Storage Gateway



Real-time Data Movement

AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

Amazon Redshift—Data Warehousing

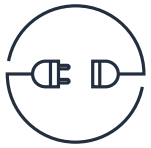
Fast, powerful, simple, and fully managed data warehouse at 1/10 the cost

Massively parallel, scale from gigabytes to petabytes

Fast at scale

Columnar storage technology to improve I/O efficiency and scale query performance

Open file formats



Analyze optimized data formats on the latest SSD, and all open data formats in Amazon S3

Secure

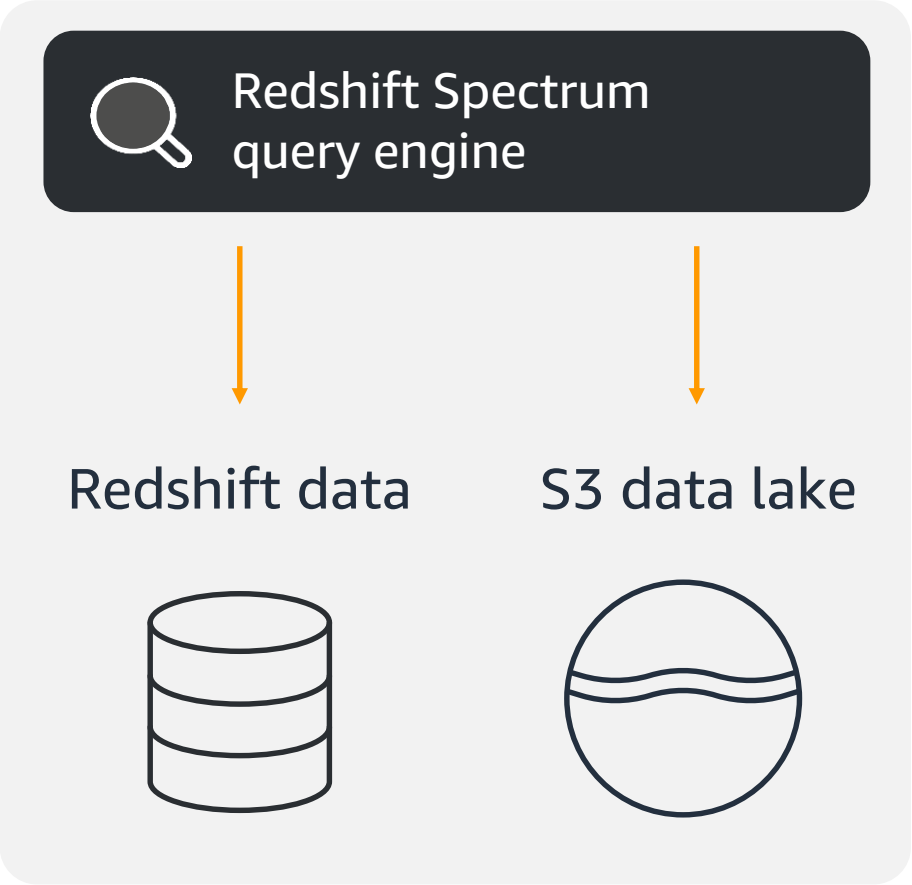
Audit everything; encrypt data end-to-end; extensive certification and compliance

Inexpensive

As low as $1,000 per terabyte per year, 1/10th the cost of traditional data warehouse solutions; start at $0.25 per hour

Amazon Redshift Spectrum

Extend the data warehouse to exabytes of data in S3 data lake



- Exabyte Redshift SQL queries against S3
- Join data across Redshift and S3
- Scale compute and storage separately
- Stable query performance and unlimited concurrency
- CSV, ORC, Grok, Avro, & Parquet data formats
- Pay only for the amount of data scanned

Amazon Athena—Interactive Analysis

Interactive query service to analyze data in Amazon S3 using standard SQL

No infrastructure to set up or manage and no data to load

Ability to run SQL queries on data archived in Amazon Glacier (coming soon)

Query Instantly

Zero setup cost; just point to S3 and start querying

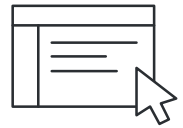
Pay per query

Pay only for queries run; save 30–90% on per-query costs through compression

Open

ANSI SQL interface, JDBC/ODBC drivers, multiple formats, compression types, and complex joins and data types

Easy



Serverless: zero infrastructure, zero administration
Integrated with QuickSight

Amazon EMR—Big Data Processing

Analytics and ML at scale

19 open-source projects: Apache Hadoop, Spark, HBase, Presto, and more

Enterprise-grade security

Latest versions

Updated with the latest open source frameworks within 30 days of release

Low cost

Flexible billing with per-second billing, EC2 spot, reserved instances and auto-scaling to reduce costs 50–80%

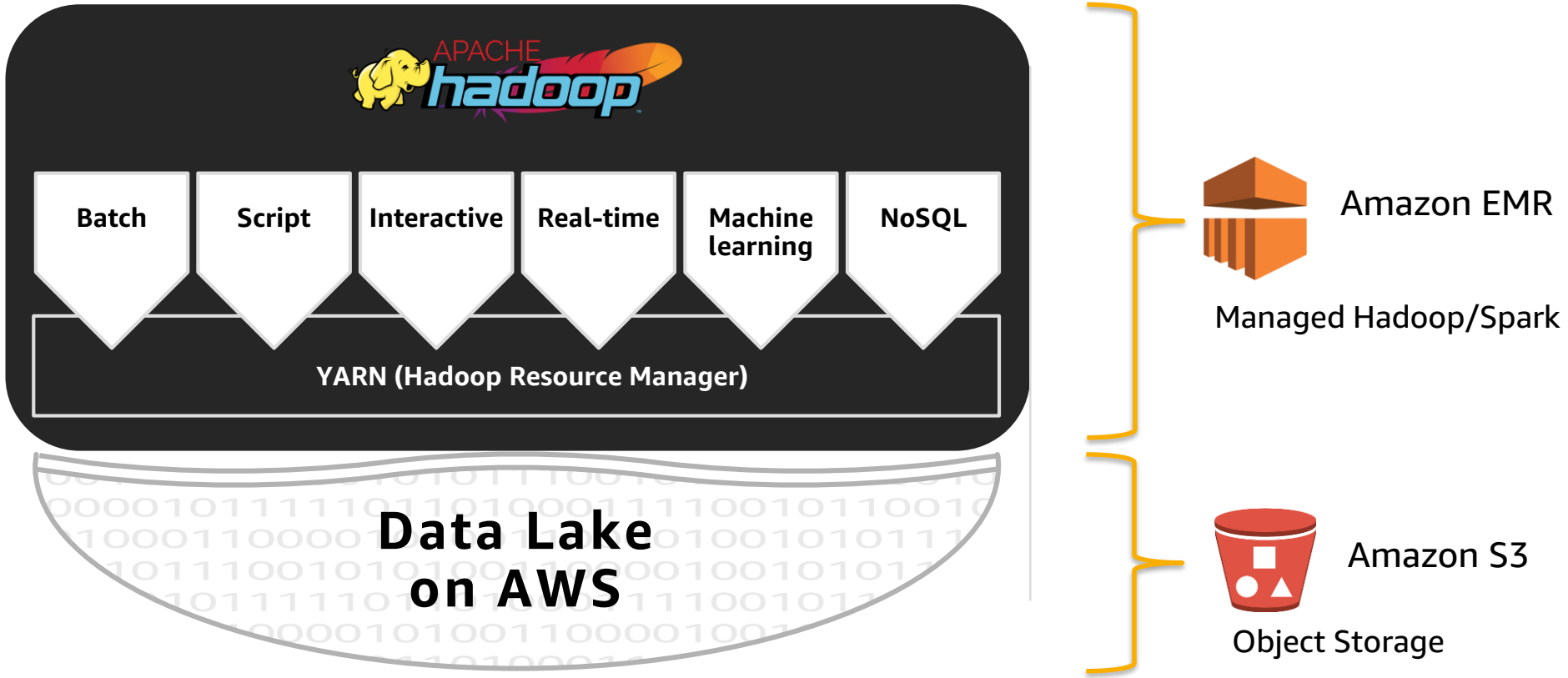
Use S3 storage

Process data directly in the S3 data lake securely with high performance using the EMRFS connector

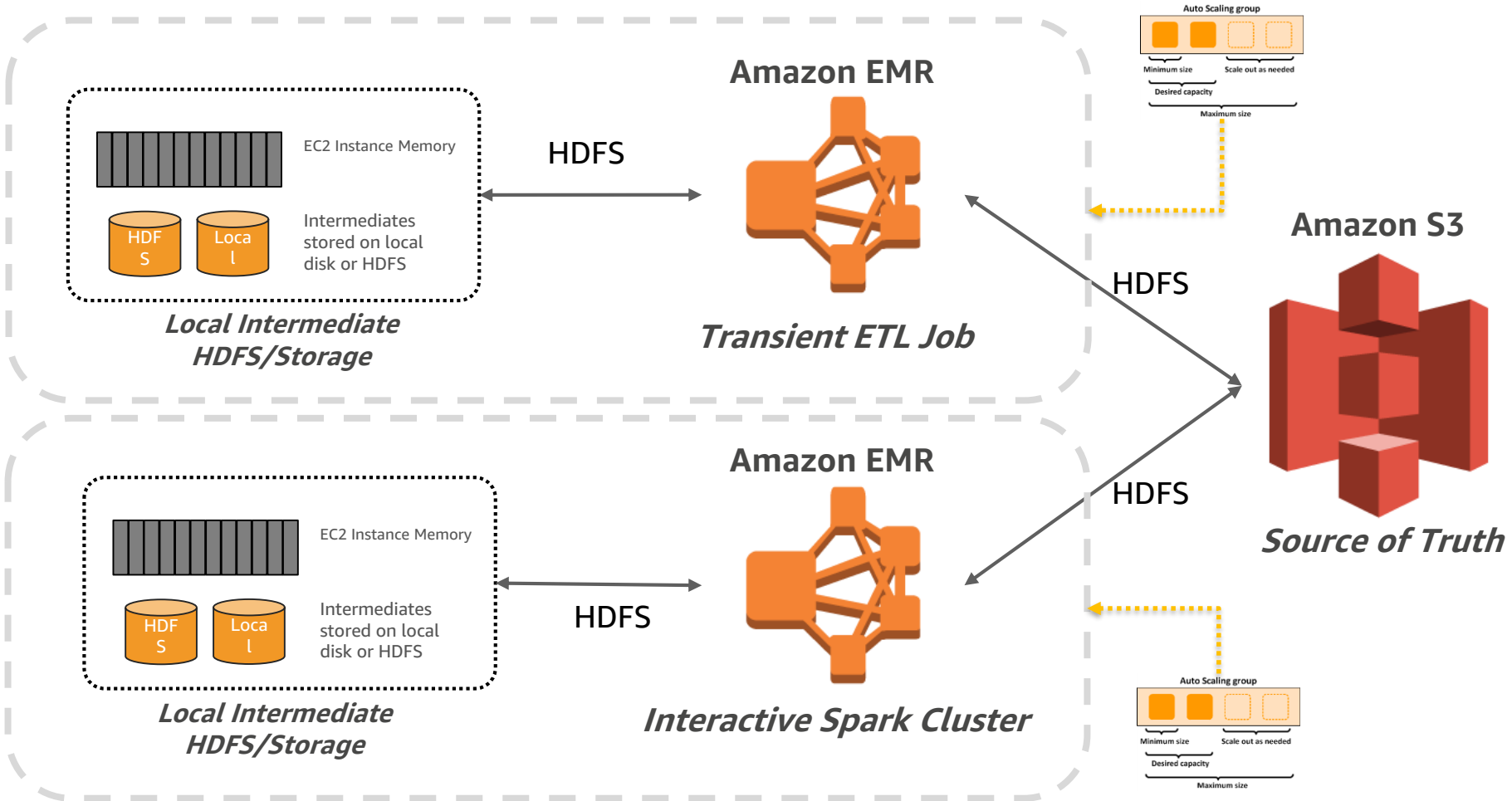
Easy

Launch fully managed Hadoop & Spark in minutes; no cluster setup, node provisioning, cluster tuning

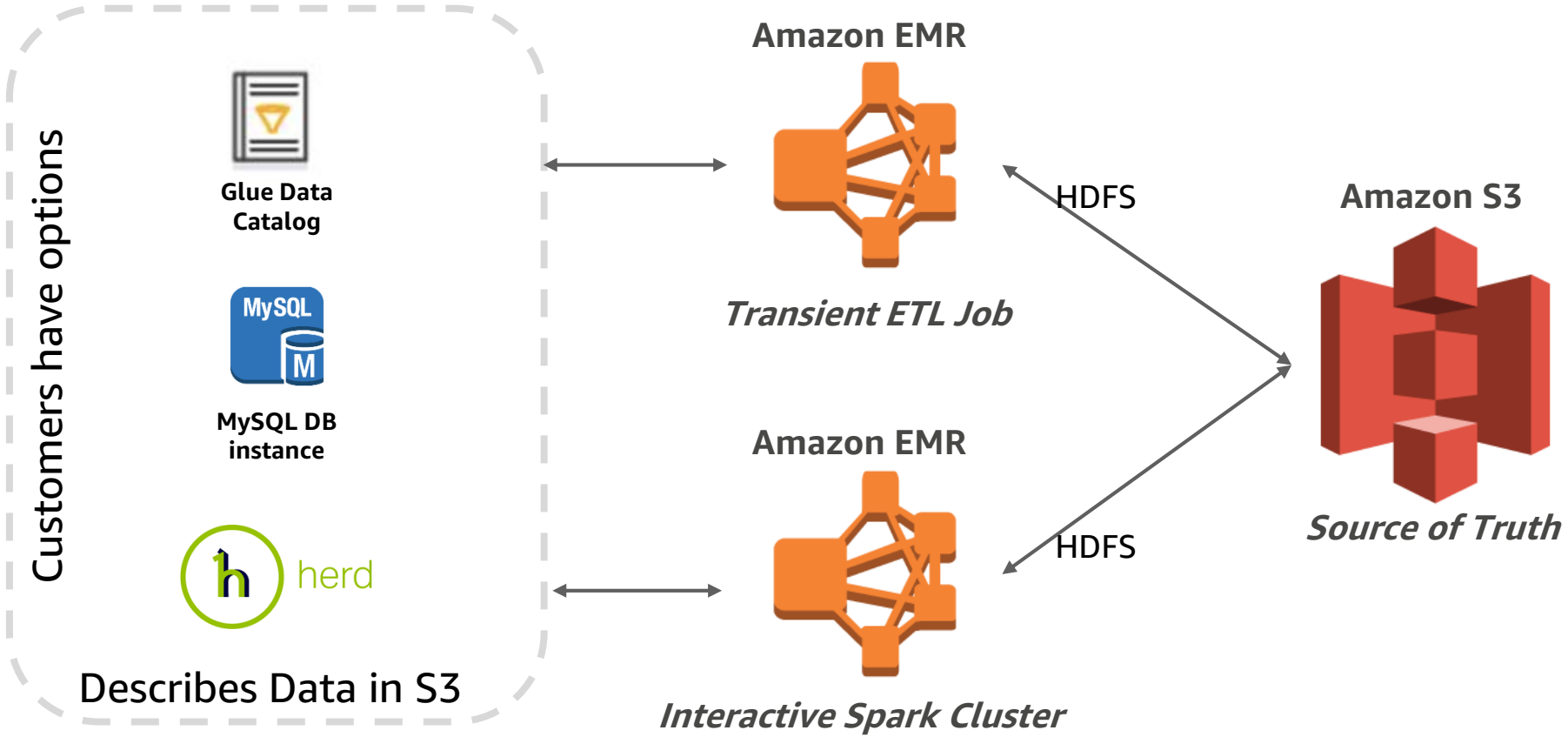
Big Data Processing



Amazon S3—Source of Truth, Multiple Clusters



External Metadata Management



Characteristics of a data lake

Collect
Anything

Characteristics of a data lake

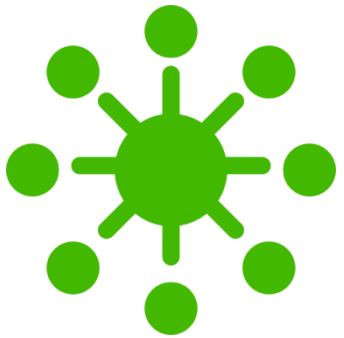
Collect Anything

Dive in Anywhere

Characteristics of a data lake

Collect Anything

Dive in Anywhere

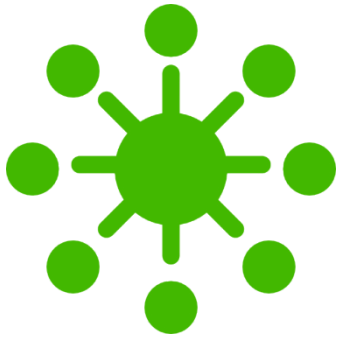


Flexible Access

Characteristics of a data lake

Collect Anything

Dive in Anywhere



Flexible Access

Future Proof

Learn from AWS experts. Advance your skills and knowledge. Build your future in the AWS Cloud.

Digital Training

Free, self-paced online courses built by AWS experts

Classroom Training

Classes taught by accredited AWS instructors

AWS Certification

Exams to validate expertise with an industry-recognized credential

Ready to begin building your cloud skills?
Get started at: https://www.aws.training/

With deep expertise on AWS, APN Partners can help your organization at any stage of your Cloud Adoption Journey.

AWS Managed Service Providers

APN Consulting Partners who are skilled at cloud infrastructure and application migration, and offer proactive management of their customer's environment.

AWS Competency Partners

APN Partners who have demonstrated technical proficiency and proven customer success in specialized solution areas.

AWS Marketplace

A digital catalog with thousands of software listings from independent software vendors that make it easy to find, test, buy, and deploy software that runs on AWS.

AWS Service Delivery Partners

APN Partners with a track record of delivering specific AWS services to customers.

Ready to get started with an APN Partner?
Find a partner: https://aws.amazon.com/partners/find/
Learn more at the AWS Partner Network Booth

Thank You for Attending AWS Innovate

We hope you found it interesting! A kind reminder to **complete the survey.**

Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apac-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws