Overview Big Data on

Which tools should I use?

Let’s solve the first question that might come to your mind: what’s the right tools for building that pipeline? And the answer I found while building mine was:

“ There is not a right tool or architecture, it will always depend on your needs! “

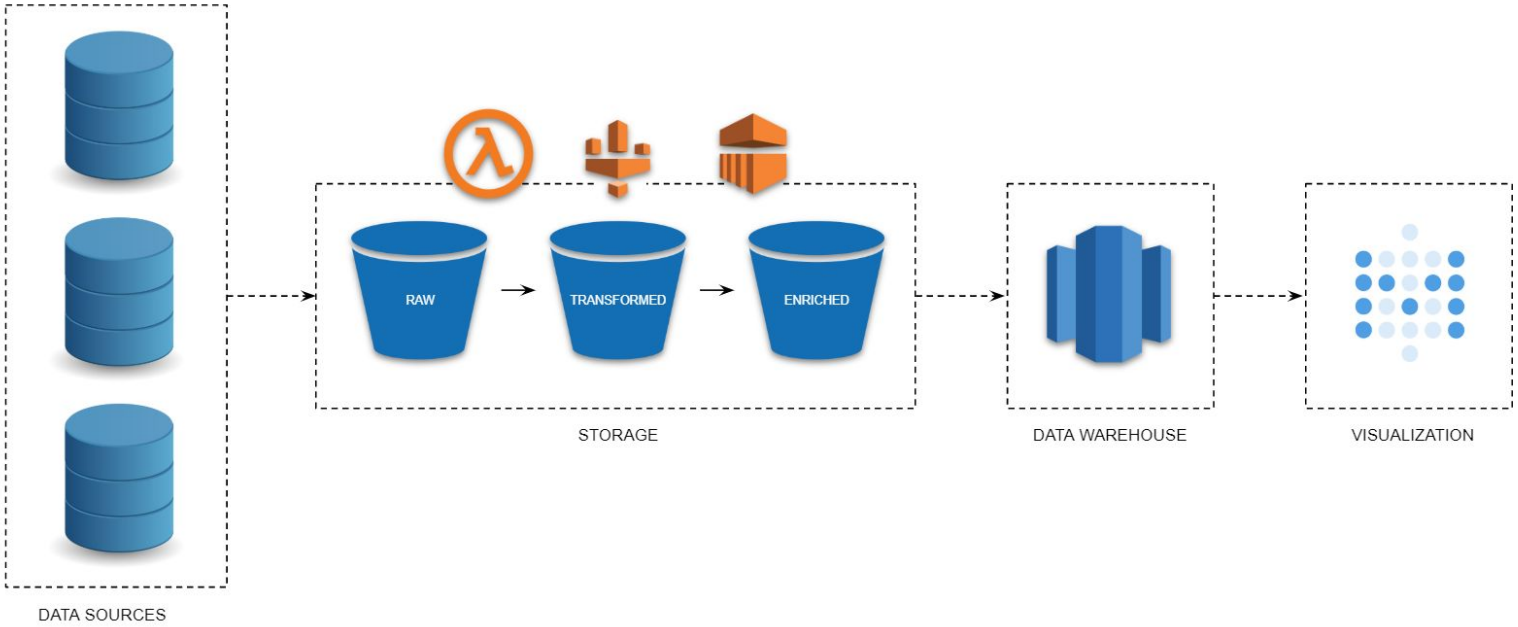
The challenge

By the time I got into the company, there was a big problem: the data was too isolated. Analyzing data was too slow and difficult that people could not find the motivation to do it.

And the challenge was: centralize that data and promote **data democratization** on the company in order to **empower** people! A big challenge, right?

The pipeline

The proposed pipeline architecture to fulfill those needs is presented on the image bellow, with a little bit of improvements that we will be discussing.

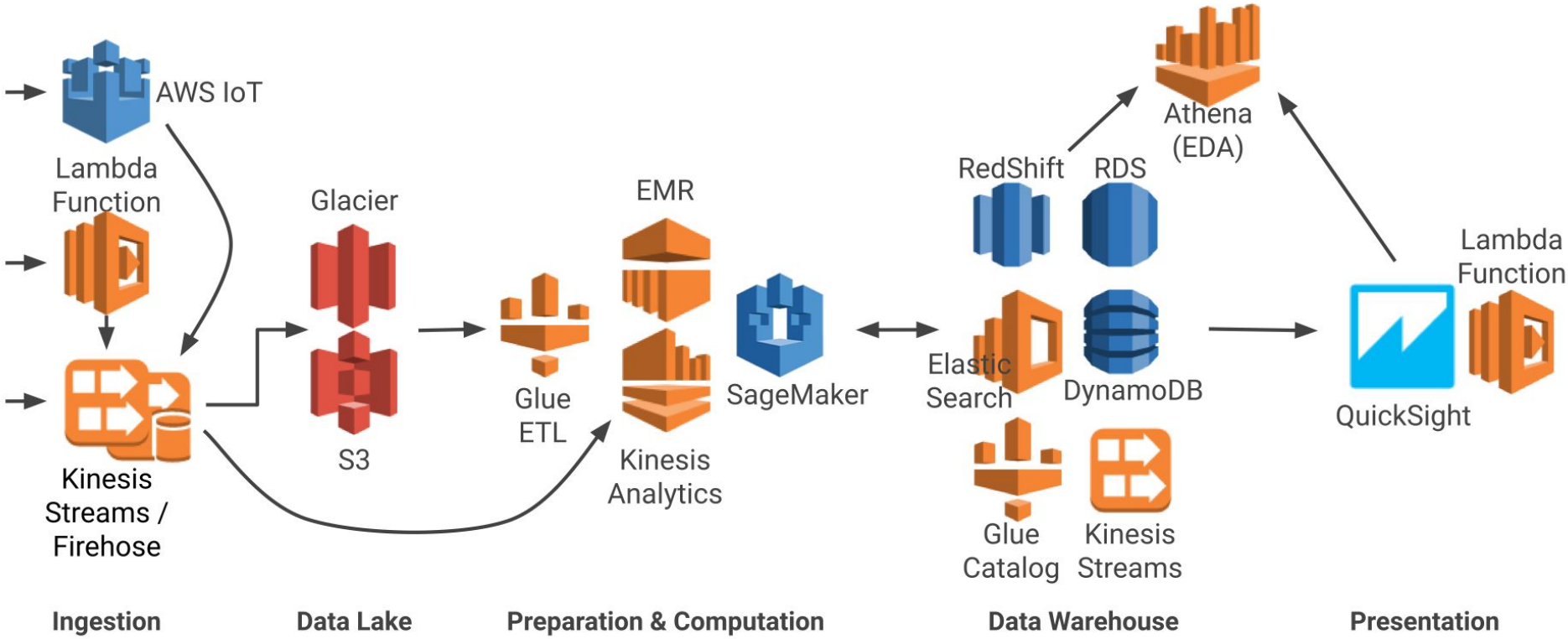


Category	Use cases	AWS service
Analytics	Interactive analytics Big data processing Data warehousing Real-time analytics Operational analytics Dashboards and visualizations Visual data preparation	Amazon Athena Amazon EMR Amazon Redshift Amazon Kinesis Amazon Elasticsearch Service Amazon Quicksight AWS Glue DataBrew

Category	Use cases	AWS service
Data movement	Real-time data movement	AWS Glue Amazon Managed Streaming for Apache Kafka (MSK) Amazon Kinesis Data Streams Amazon Kinesis Data Firehose Amazon Kinesis Video Streams

Category	Use cases	AWS service
Data lake	Object storage	Amazon S3 AWS Lake Formation
	Backup and archive	Amazon S3 Glacier AWS Backup
	Data catalog	AWS Glue AWS Lake Formation
	Third-party data	AWS Data Exchange

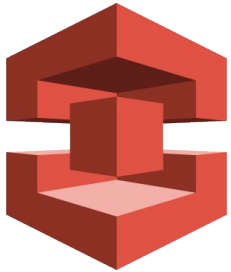
Category	Use cases	AWS service
Predictive analytics and machine learning	Frameworks and interfaces Platform services	AWS Deep Learning AMIs Amazon SageMaker


AWS IoT
Lambda Function
Kinesis Streams / Firehose
Glacier
S3
Glue ETL
EMR
Kinesis Analytics
SageMaker
RedShift
RDS
Athena (EDA)
Elastic Search
DynamoDB
Glue Catalog
Kinesis Streams
Lambda Function
QuickSight
Ingestion
Data Lake
Preparation & Computation
Data Warehouse
Presentation

Data Movement

AWS Direct Connect

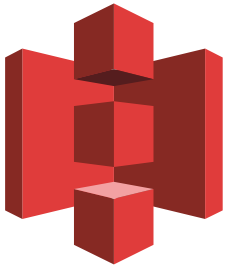
AWS Database Migration Service



AWS Import/Export & Snowball

AWS Storage Gateway

Storage and Databases

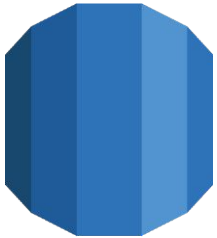


Amazon S3

- Store unlimited number of objects
- Designed for 99.999999999% durability
- As Data Lake with integration with other AWS services (Amazon Kinesis, Amazon Redshift, Amazon EMR, etc.)
- Low cost with tired-storage (Standard, IA, Amazon Glacier) via life-cycle policy
- Secure – SSL, client/server-side encryption at rest

Amazon DynamoDB

- Fully Managed NoSQL Database
- Fast consistent performance (single-digit millisecond latency at any scale)
- Highly scalable - automatic scaling of throughput capacity
- Highly available and durability
- Store unlimited number of data



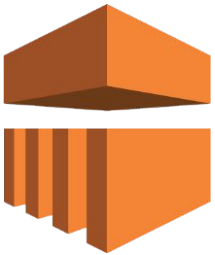
Amazon Aurora

- Fully Managed Relational Database Service
- MySQL and PostgreSQL compatible relational database with up to
  5x better performance running on the same hardware
- Security, availability, and reliability of commercial databases at 1/10th the cost
- Designed to offer greater than 99.99% availability.
- Automatically grows storage as needed, from 10GB up to 64TB
- Achieve up to 500,000 reads and 100,000 writes per second

Amazon Redshift

- Fully managed petabyte-scale relational, MPP, data warehousing
- Built-in end-to-end security, including SSL connections and cluster
  encryption
- Fault-tolerant - automatically recovers from disk and node failures
- Data automatically backed up to Amazon S3
- $1,000/TB/Year; start at $0.25/hour. Provision in minutes; scale
  from 160 GB to 2 PB of compressed data with just a few clicks

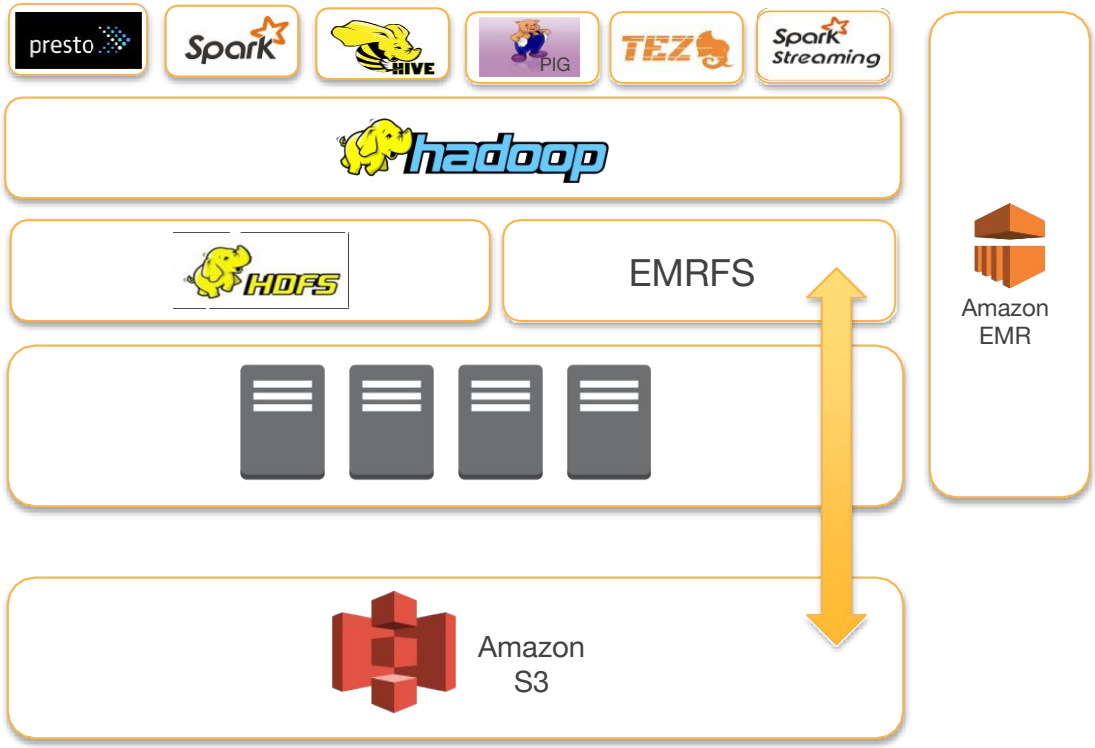
Analytic Frameworks



Amazon EMR

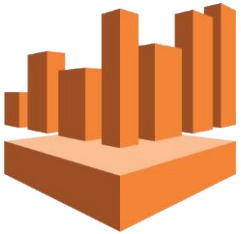
- Managed Hadoop framework
- Apache Hadoop, Hive, Spark, Zeppelin, Presto, HBase, Phoenix, Tez, Flink, etc.
- Auto Scaling clusters with support for on-demand and spot pricing
- Support for end-to-end encryption, IAM/VPC, S3 client-side encryption with customer managed keys and AWS KMS
- Integrates with Amazon S3, Amazon DynamoDB, Amazon Kinesis and Amazon Redshift

Amazon EMR


presto
Spark
HIVE
PIG
TEZ
Spark Streaming
hadoop
HDFS
EMRFS
Amazon EMR
Amazon S3

Amazon Elasticsearch

- Fully managed, reliable, and scalable Elasticsearch service
- Support for ELK
- Integration options with other AWS services (CloudWatch Logs, Amazon DynamoDB, Amazon S3, Amazon Kinesis)
- Use Case: log analytics, full text search, application monitoring, and more.



Amazon Athena

- Serverless query service for querying data in S3 using standard SQL with no infrastructure to manage
- Support for multiple data formats include text, CSV, TSV, JSON, Avro, ORC, Parquet
- Pay per query only when you’re running queries based on data scanned. If you compress your data, you pay less and your queries run faster

Familiar Technologies Under the Covers

Used for SQL Queries

In-memory distributed query engine

ANSI-SQL compatible with extensions

Used for DDL functionality

Complex data types

Multitude of formats

Supports data partitioning

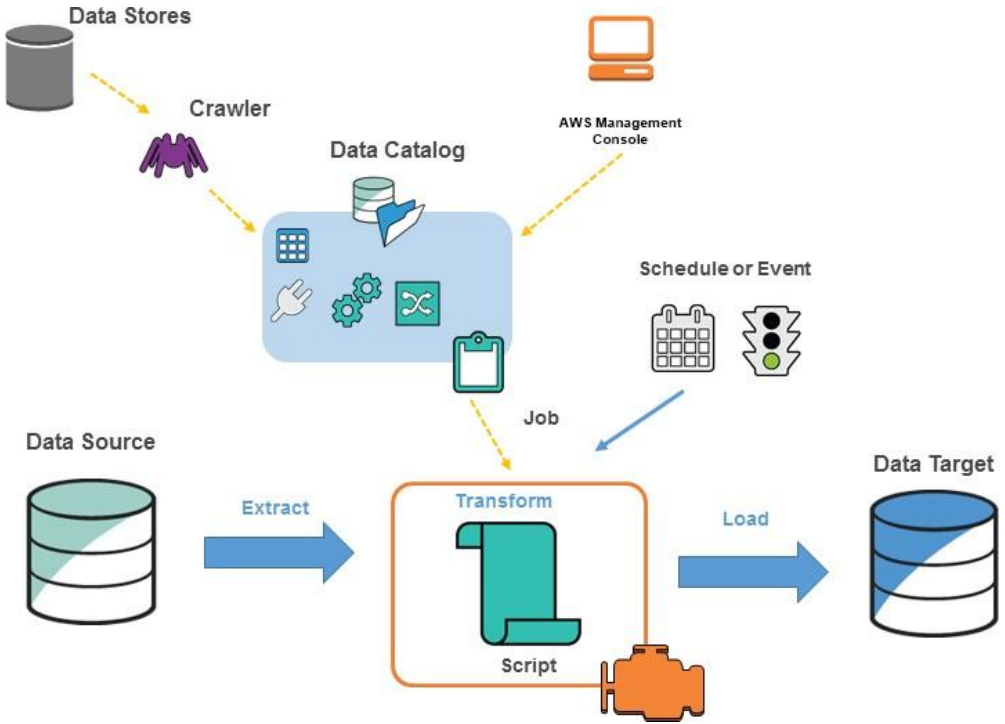
Amazon Quicksight

- Fast and cloud-powered Business Analytics
- Easy to use, no infrastructure to manage
- Quick calculations with SPICE
- 1/10th the cost of legacy BI software
- Accessed from any browser or mobile device

AWS Glue

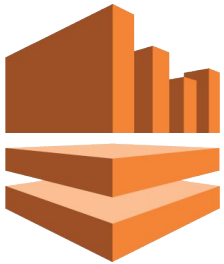
- Fully managed ETL (extract, transform, load) service
- Integrated data catalog, automatic schema discovery, ETL code generation, flexible job scheduler
- Integrated across a wide range of AWS services (Amazon RDS, Database running on Amazon EC2, Amazon Athena, etc.)

How AWS Glue Works



1. Build your data catalog
2. Generate and Edit Transformations
3. Schedule and Run Your Jobs

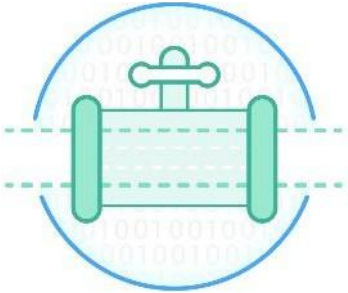
Real-time Analytics



Amazon Kinesis

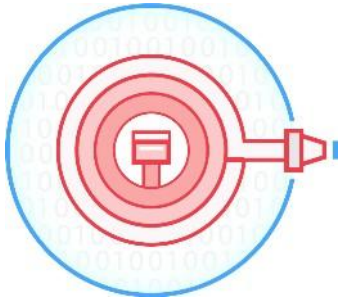
- Fully managed streaming application
- Scalable – handle any amount of streaming data
- Ingest, buffer and process data in real-time
- React quickly – derive insight in seconds

Amazon Kinesis



Amazon Kinesis Streams

Build your own custom applications that process or analyze streaming data



Amazon Kinesis Firehose

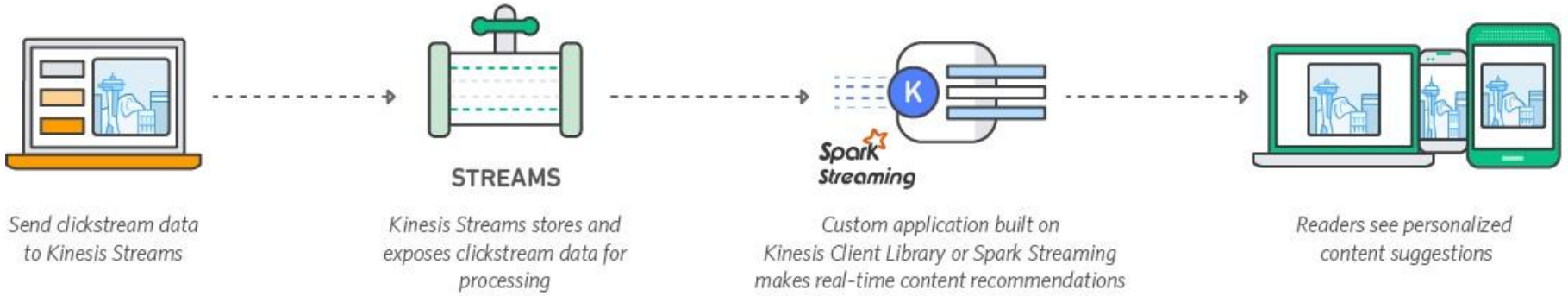
Easily load massive volumes of streaming data into Amazon S3, Amazon Redshift, and Amazon Elasticsearch

Amazon Kinesis Analytics

Easily analyze data streams using standard SQL queries

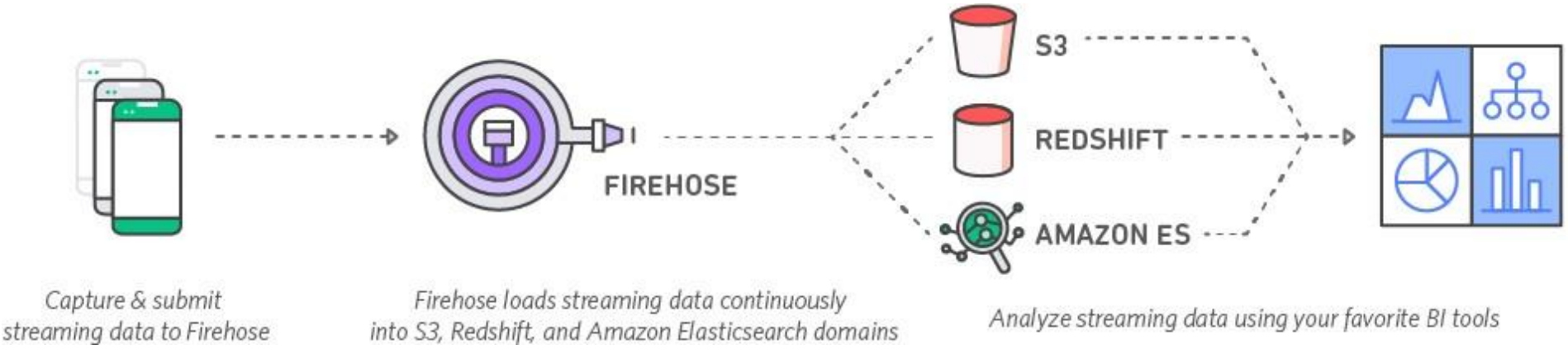
Amazon Kinesis Streams

- Reliably **ingest and durably store** streaming data at low cost
- Build **custom real-time applications** to process streaming data



Amazon Kinesis Firehose

Reliably **ingest** and **deliver batched, compressed, and encrypted data** to **S3, Amazon Redshift, and Amazon Elasticsearch Service**



Amazon Kinesis Analytics

Interact with streaming data in real time using SQL

Capture streaming data with Kinesis Streams or Kinesis Firehose

Run standard SQL queries against data streams

Kinesis Analytics can send processed data to analytics tools so you can create alerts and respond in real-time

AWS Market Place for Big Data Solution

Hundreds of big data products are immediately available through the AWS marketplace

Database and Data Enablement

ATTUNITY

syncsort

Matillion
Business Intelligence

informatica

TERADATA

splunk>

Advanced Analytics

ZEMENTIS
EMPOWERING BIG DATA INSIGHTS

sas

XPATTERNS
BY ATIGEO

Digital Reasoning

bigml

PredictionIO

Business Inteligence

TIBCO Spotfire

+ableau
SOFTWARE

MicroStrategy

SAP

Logi
ANALYTICS

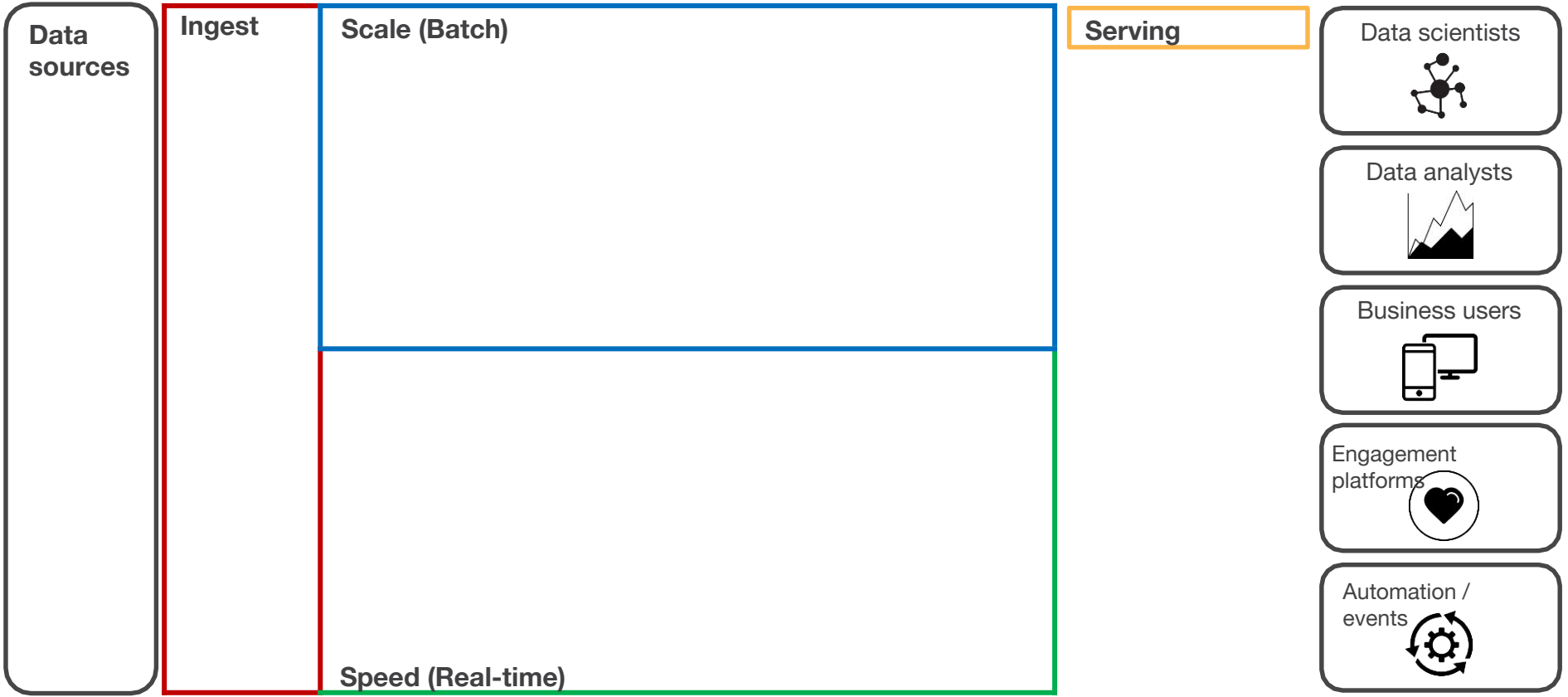
looker

Fully Integrated | 1-click deployment |
Pay-as-you-go pricing

Modern Data Analytics Architecture on AWS

Modern data architecture

Insights to enhance business applications, new digital services



Modern data architecture

Insights to enhance business applications, new digital services

Data sources

Transactions

ERP

Web logs / cookies

Connected devices

Social media

Ingest

Scale (Batch)

Speed (Real-time)

Serving

Data scientists

Data analysts

Business users

Engagement platforms

Automation / events

Modern data architecture

Insights to enhance business applications, new digital services

Modern data architecture

Insights to enhance business applications, new digital services

Data sources

Transactions

ERP

Web logs / cookies

Connected devices

Social media

Ingest

Internet Interfaces

AWS Direct Connect

AWS Database Migration

Amazon Kinesis

Scale (Batch)

Raw Data

Amazon S3

EMR

ETL

Amazon

AWS Glue

Staged Data (Data Lake)

Amazon S3

Speed (Real-time)

Serving

Data scientists

Data analysts

Business users

Engagement platforms

Automation / events

Modern data architecture

Insights to enhance business applications, new digital services

Data sources

Transactions

ERP

Web logs / cookies

Connected devices

Social media

Ingest

Internet Interfaces

AWS Direct Connect

AWS Database Migration

Amazon Kinesis

Scale (Batch)

Raw Data
Amazon S3
EMR

ETL
Amazon
HIVE

Staged Data (Data Lake)
Amazon S3

Advanced Analytics
Spark MLlib

Amazon ML

Deep Learning
mxnet

Speed (Real-time)

Serving

Data scientists

Data analysts

Business users

Engagement platforms

Automation / events

Modern data architecture

Insights to enhance business applications, new digital services

Data sources
Transactions
ERP
Web logs / cookies
Connected devices
Social media

Ingest
Internet Interfaces
AWS Direct Connect
AWS Database Migration
Amazon Kinesis

Scale (Batch)
Raw Data
Amazon S3
EMR
ETL
Amazon
Amazon ML
Advanced Analytics
Spark MLlib
Deep Learning
mxnet
Staged Data (Data Lake)
Amazon S3

Speed (Real-time)

Serving
Direct Query
Amazon Athena
Schemaless
Amazon ElasticSearch
Semi/Unstructured
Amazon EMR
Data Warehouse
Amazon Redshift
Legacy Apps
Amazon RDS
Near-Zero Latency
Amazon DynamoDB

Data scientists
Data analysts
Business users
Engagement platforms
Automation / events

Modern data architecture

Insights to enhance business applications, new digital services

Data sources

Transactions

ERP

Web logs / cookies

Connected devices

Social media

Ingest

Internet Interfaces

AWS Direct Connect

AWS Database Migration

Amazon Kinesis

Scale (Batch)

Raw Data
Amazon S3

EMR

ETL
Amazon

Amazon ML

Advanced Analytics
Spark MLlib

Deep Learning
mxnet

Staged Data (Data Lake)
Amazon S3

Speed (Real-time)

Serving

Direct Query
Amazon Athena

Schemaless
Amazon ElasticSea

Amazon QuickSight

Semi/Unstructured
Amazon EMR

Data Warehouse
Amazon Redshift

Legacy Apps
Amazon RDS

Near-Zero Lat
Amazon Dynamo

Amazon API Gatewa
events

Data scientists

Data analysts

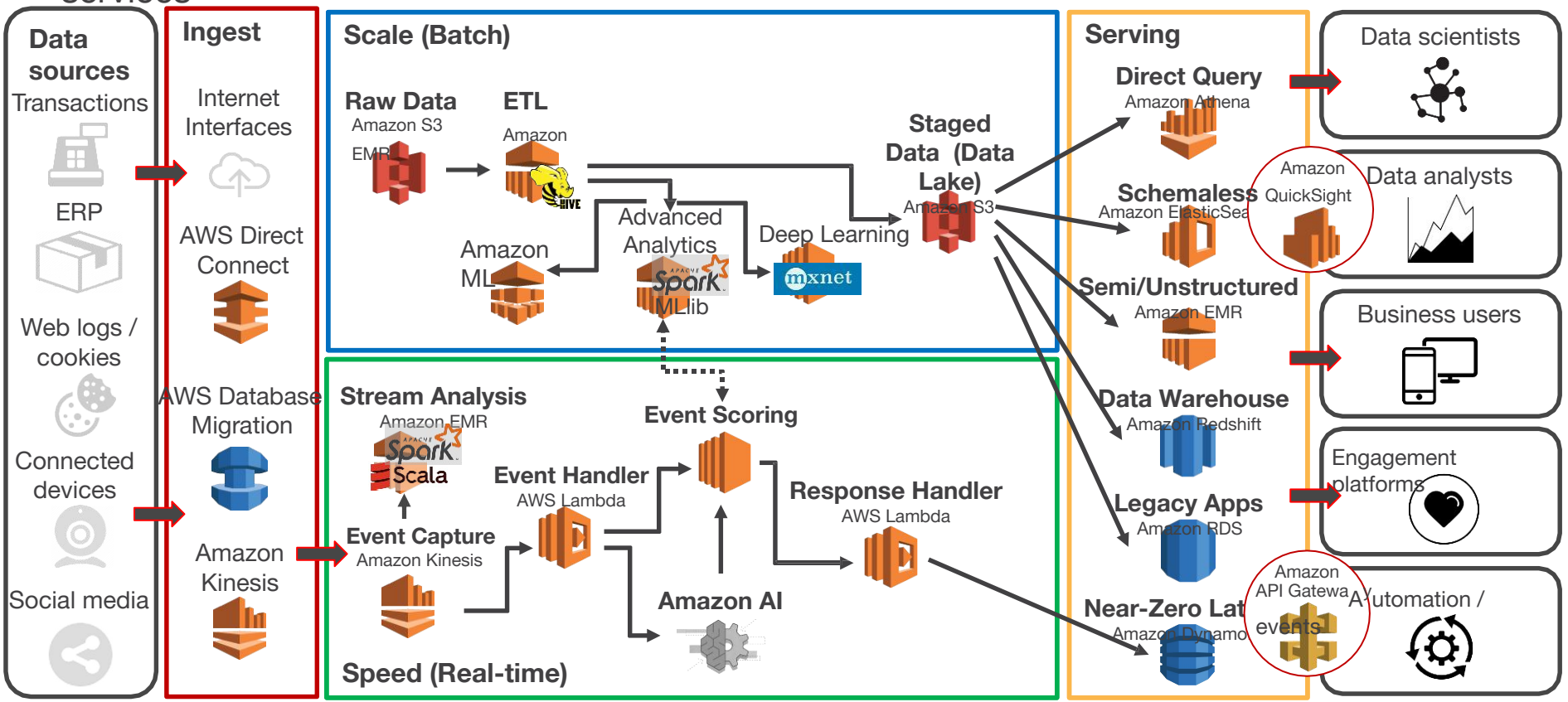
Business users

Engagement platforms

Automation /

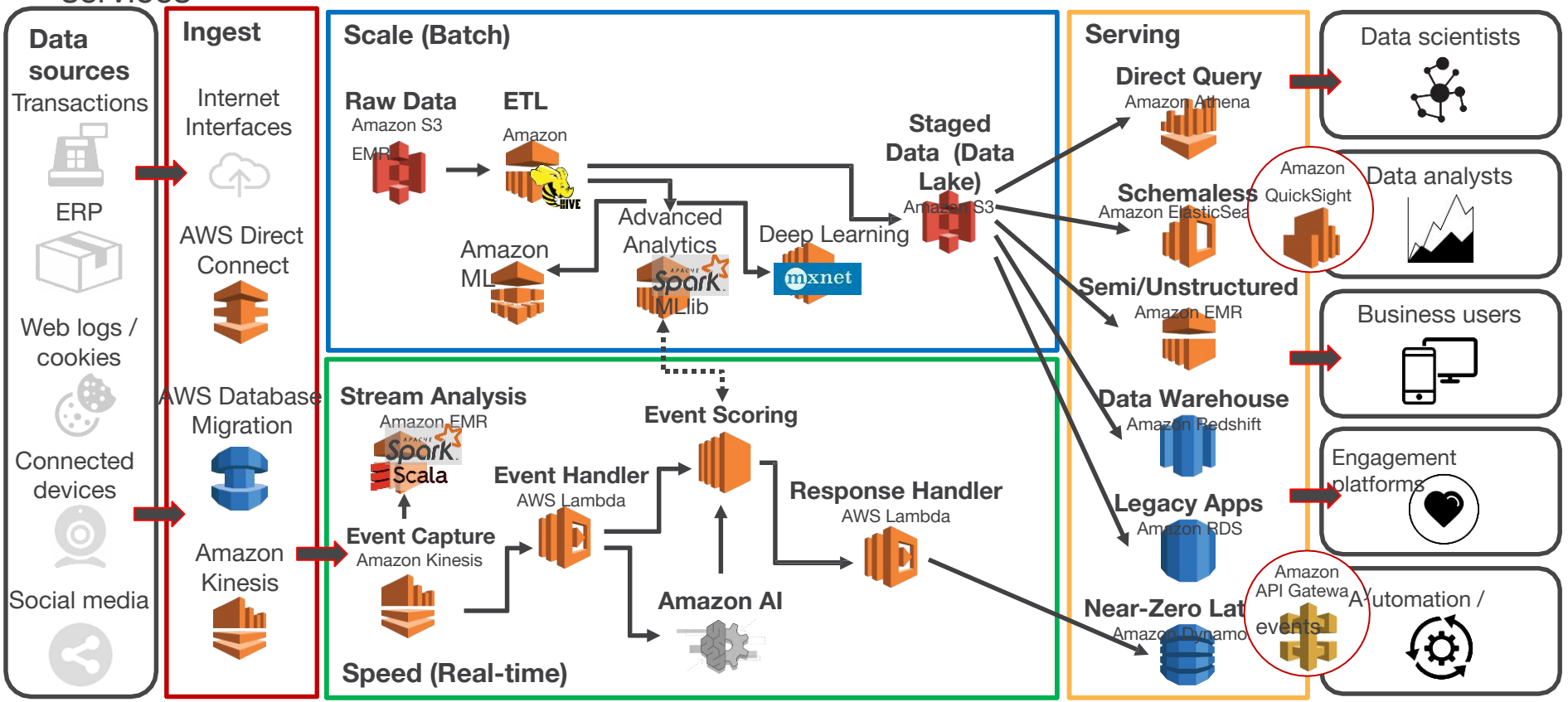
Modern data architecture

Insights to enhance business applications, new digital services



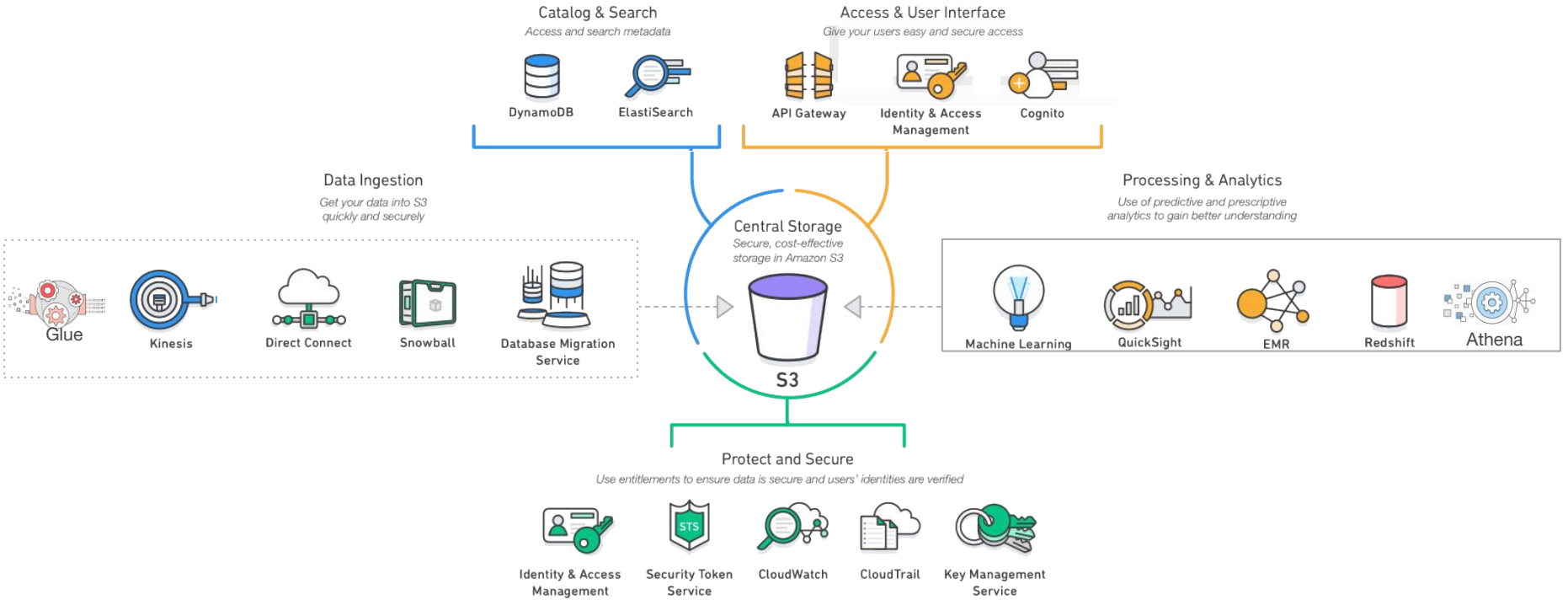
Modern data architecture

Insights to enhance business applications, new digital services

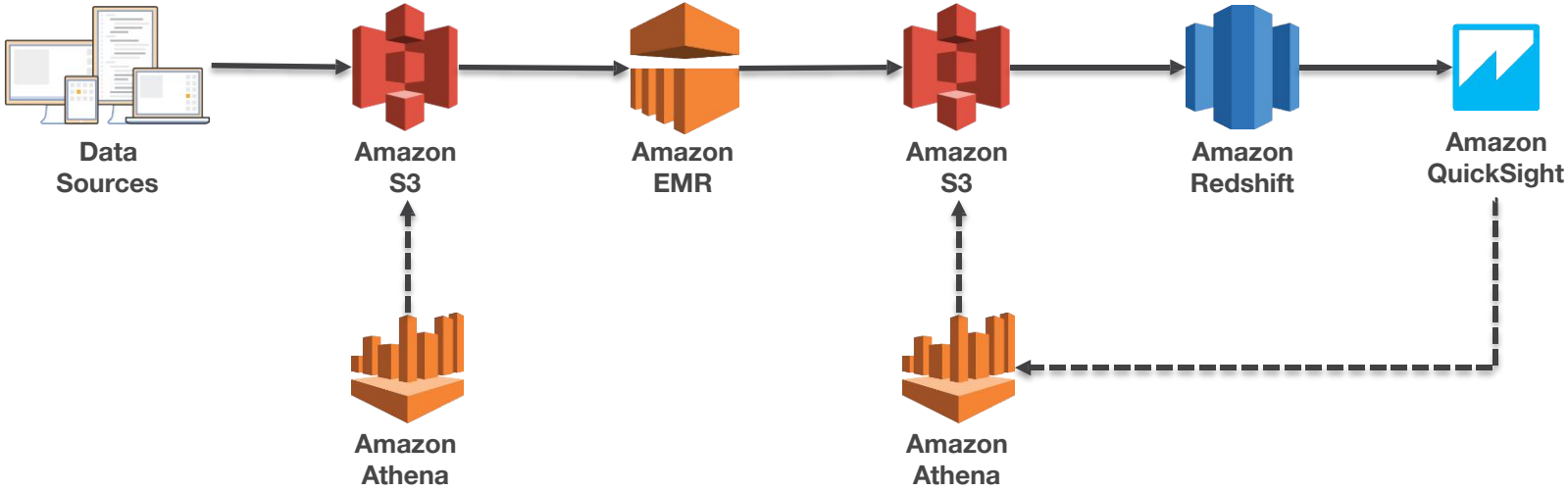


Reference Architecture

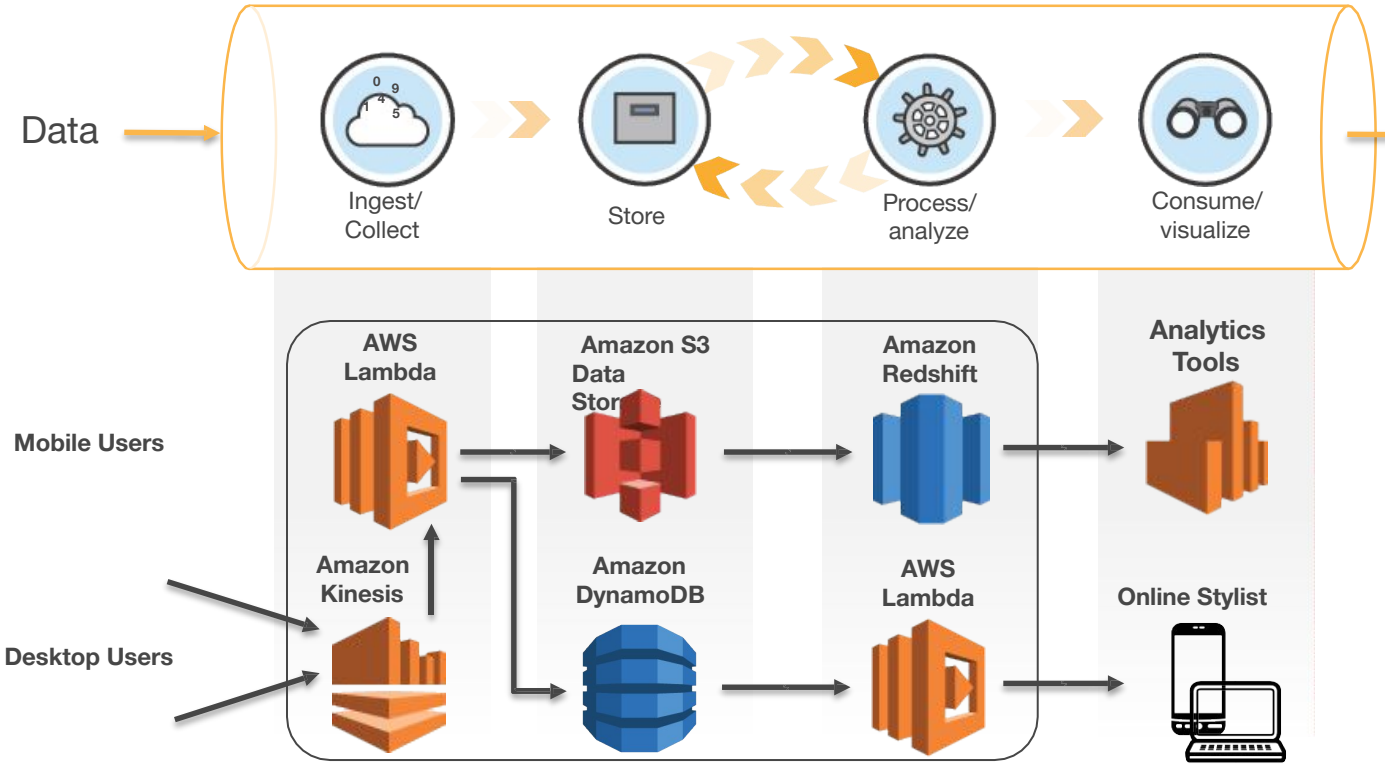
Sample Reference Architecture: Data Lake



Enterprise Data Warehouse



NORDSTROM



Personalized recommendations within seconds (from 15-20 min)
Scale the expertise of stylists to all shoppers
Reduce costs by 2X order of magnitude

...