Platzi

BigData en Cloud

Carlos Zambrano

Cloud Computing en proyectos de BigData

Cantidad

Procesar grandes cantidades de datos

MB -> GB -> TB -> PB -> EB.

Escalabilidad

Crecimiento por demanda. Servicios escalables de acuerdo a la cantidad de información.

Automatización

Procesos automatizados para procesamiento de información.

Eficiencia

Recursos aprovisionados fácilmente y al alcance de todos.

Ahorro

Reducción de costos en proyectos.

Flexibilidad

Diferente set de servicios y cloud providers para realizar proyectos de BigData.

Datos en Cloud.

Almacenamiento

Seleccionar el mejor tipo de almacenamiento.

Extracción

Extraer la información de otras fuentes, cloud providers, herramientas de terceros u on-premise.

Ingesta

Tomar la información y alimentar otros sistemas (cloud providers, servicios o herramientas de terceros)

Validación

Proporcionar ciertas garantías bien definidas de aptitud, precisión y consistencia.

Verificación

Verifican diferentes tipos de datos para asegurar su exactitud e inconsistencias después de que se realiza la migración de datos

Test

Los test se corren sobre un porcentaje de la data para garantizar el proceso que involucra los pasos descritos.

Arquitectura Lambda.

Origen

Es atribuida a Nathan Marz, para una arquitectura escalable, tolerante a fallos y de procesamiento de datos.

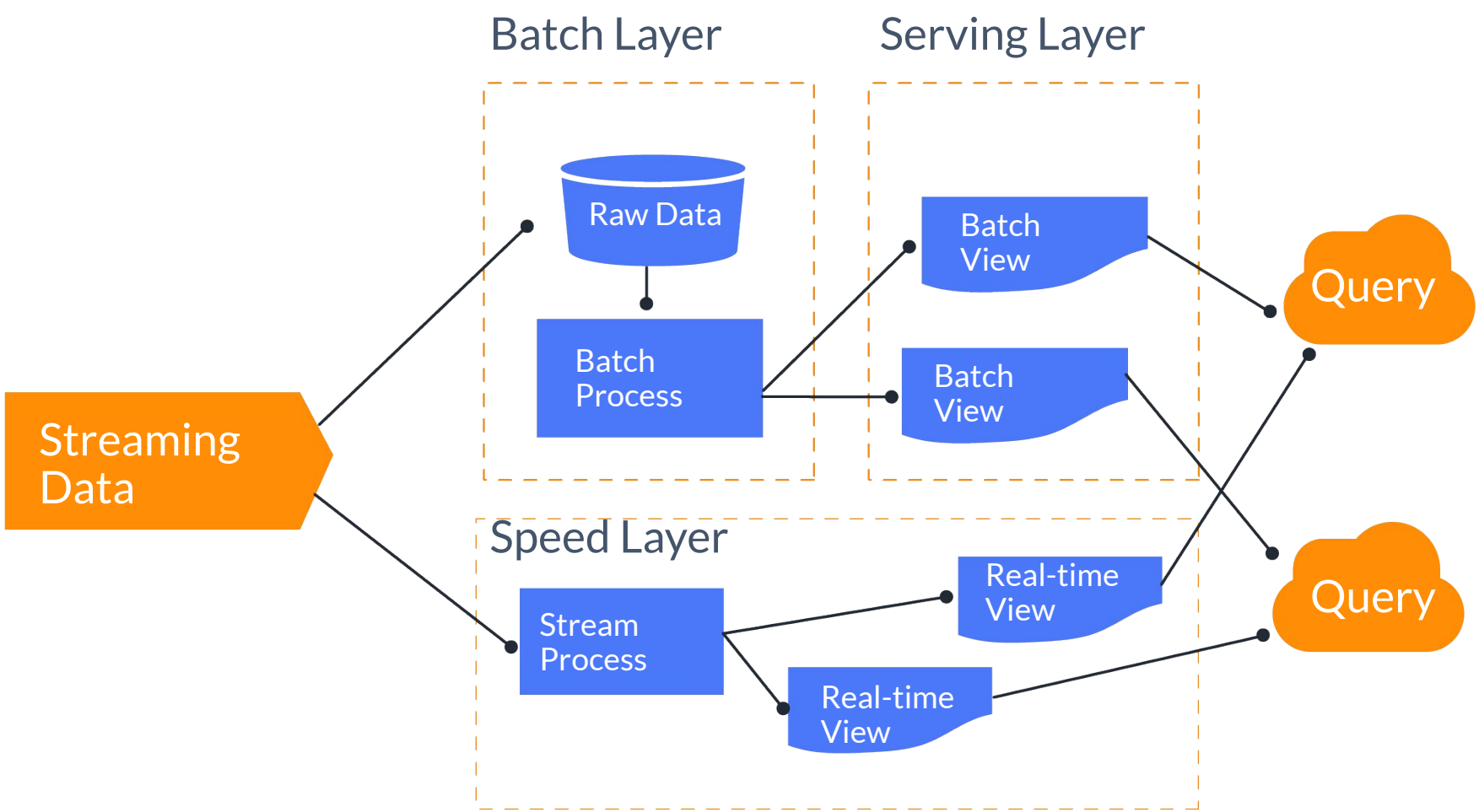
Verificación

Busca satisfacer las necesidades de un sistema robusto capaz de soportar múltiples cargas de trabajo.

Composición

Compuesto de 3 capas: Batch, Serve y Speed.

Arquitectura Lambda



Arquitectura Kappa.

Origen

Presentada por Jay Krepsen en el 2014. Es una evolución de la arquitectura lambda.

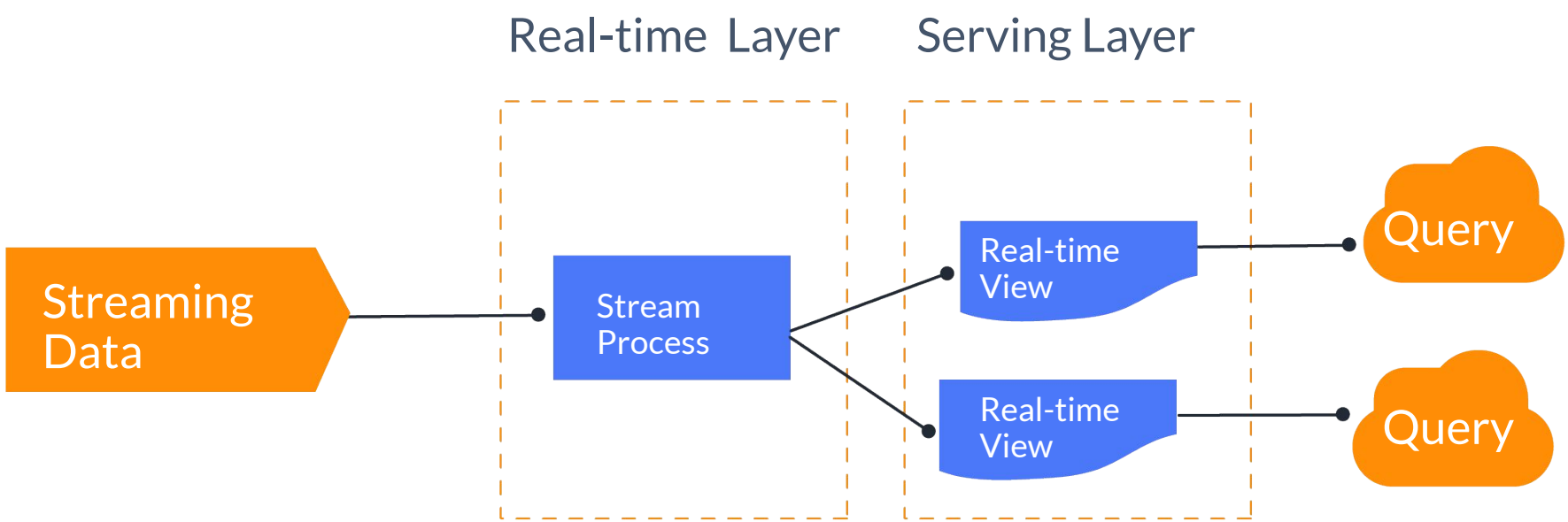
Evolución

Elimina la capa batch, dejando solo la capa de streaming.

Pilares

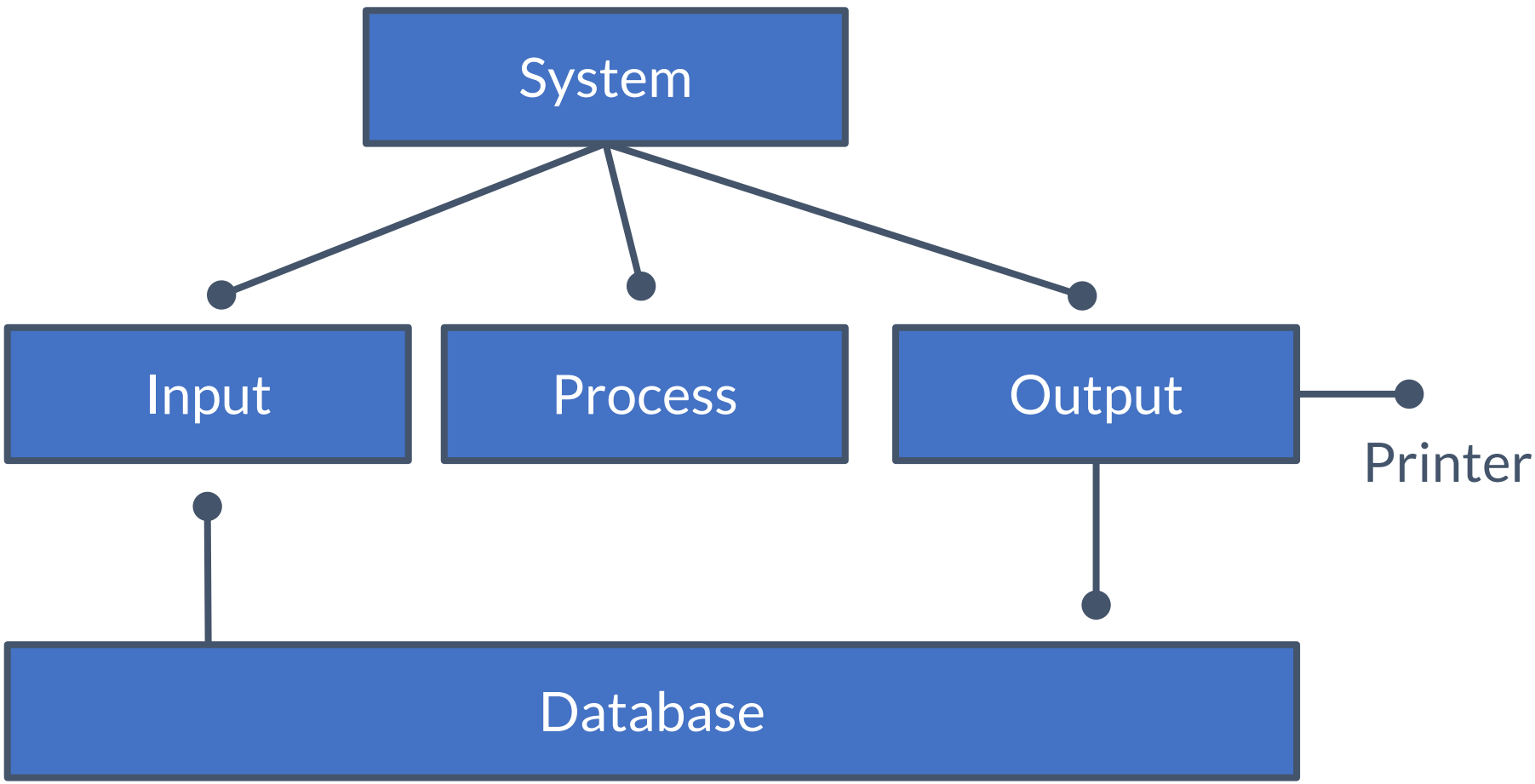
1. Todo es stream.
2. Información origen no modificada.
3. Solo un flujo de procesamiento.
4. Capaz de reprocesar.

Arquitectura Kappa



Arquitectura Batch.

Arquitectura Batch



Extracción de Información

Llevar tu información a Cloud.

SDK

Python, NodeJS, Java, .NET, Go y Ruby entre otros.

CLI

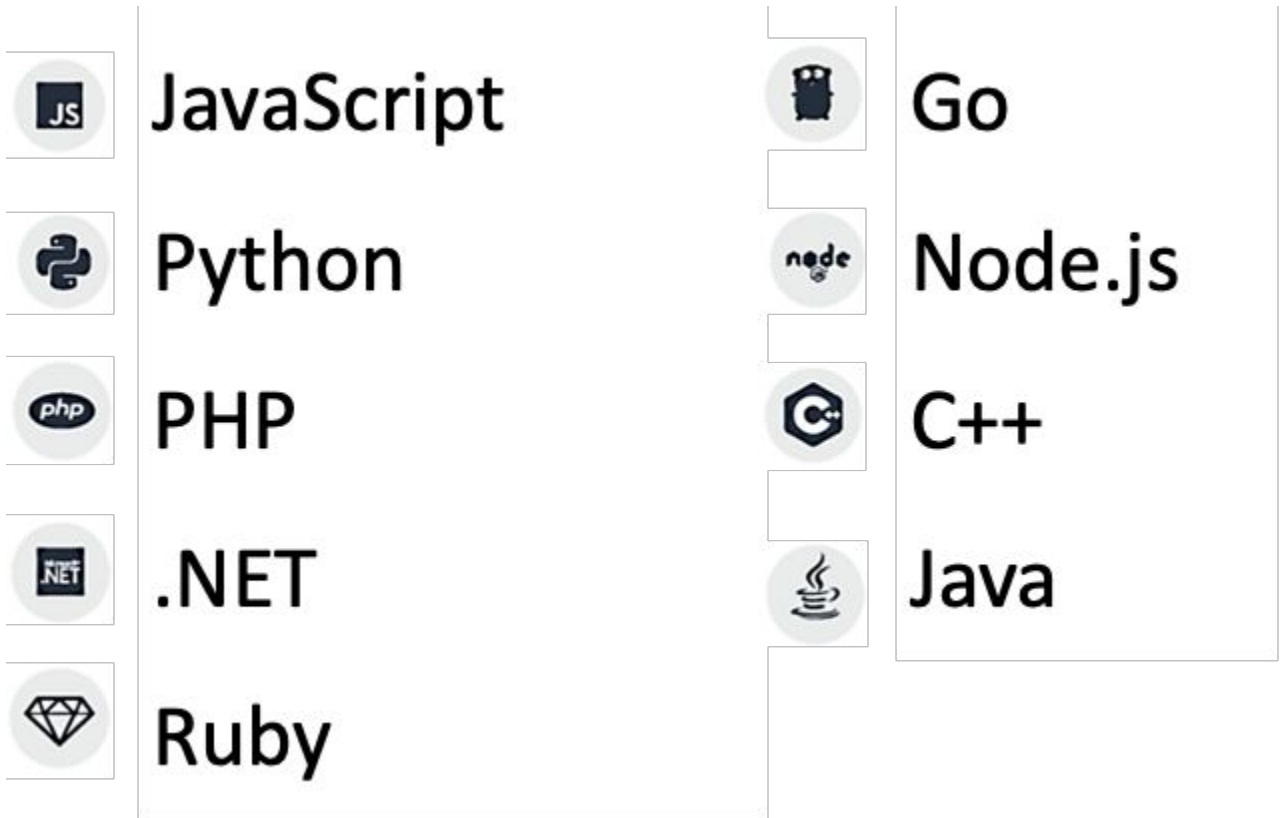
Utilización de la CLI para conectarse a la nube y enviar los eventos.

Servicios

Existen diferentes servicios para recibir/extraer información de diferentes fuentes.

Python - Boto3

SDK AWS



SDK

AWS SDK para Python.

Recursos

Manejo de todos los recursos de AWS y aprovechar las librerías que brinda Python para procesamiento de datos.

Integración

Integrar servicios de AWS con librerias para proyectos de BigData

Demo - Creando nuestro IDE en la Nube.

Demo - Cómo usar Boto3.

AWS - API Gateway

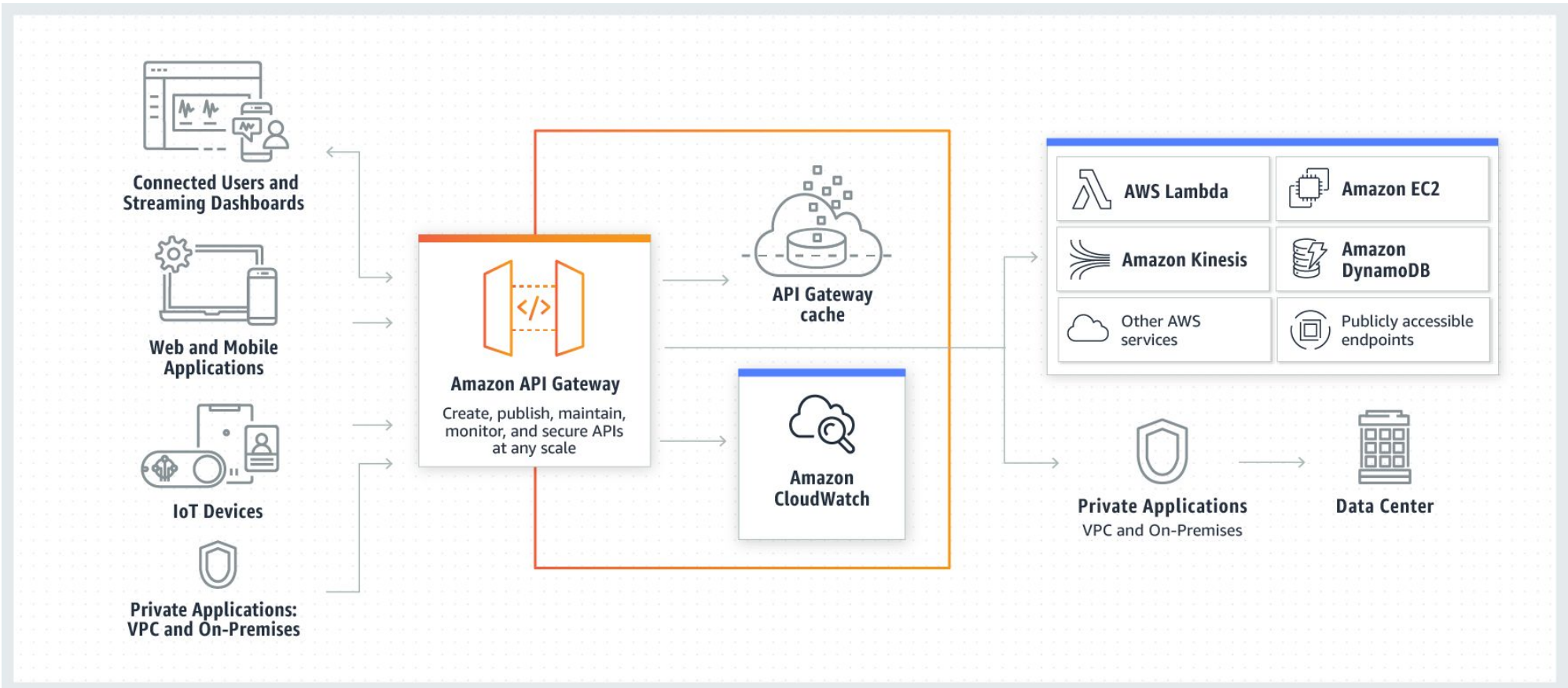
1. Crea un "front door" de nuestras aplicaciones.

2. Puede manejar cientos de miles de llamadas concurrentes a la API.

3. Previene ataques de DDoS y exponer nuestras aplicaciones.

Arquitectura con API Gateway

Arquitectura con API Gateway



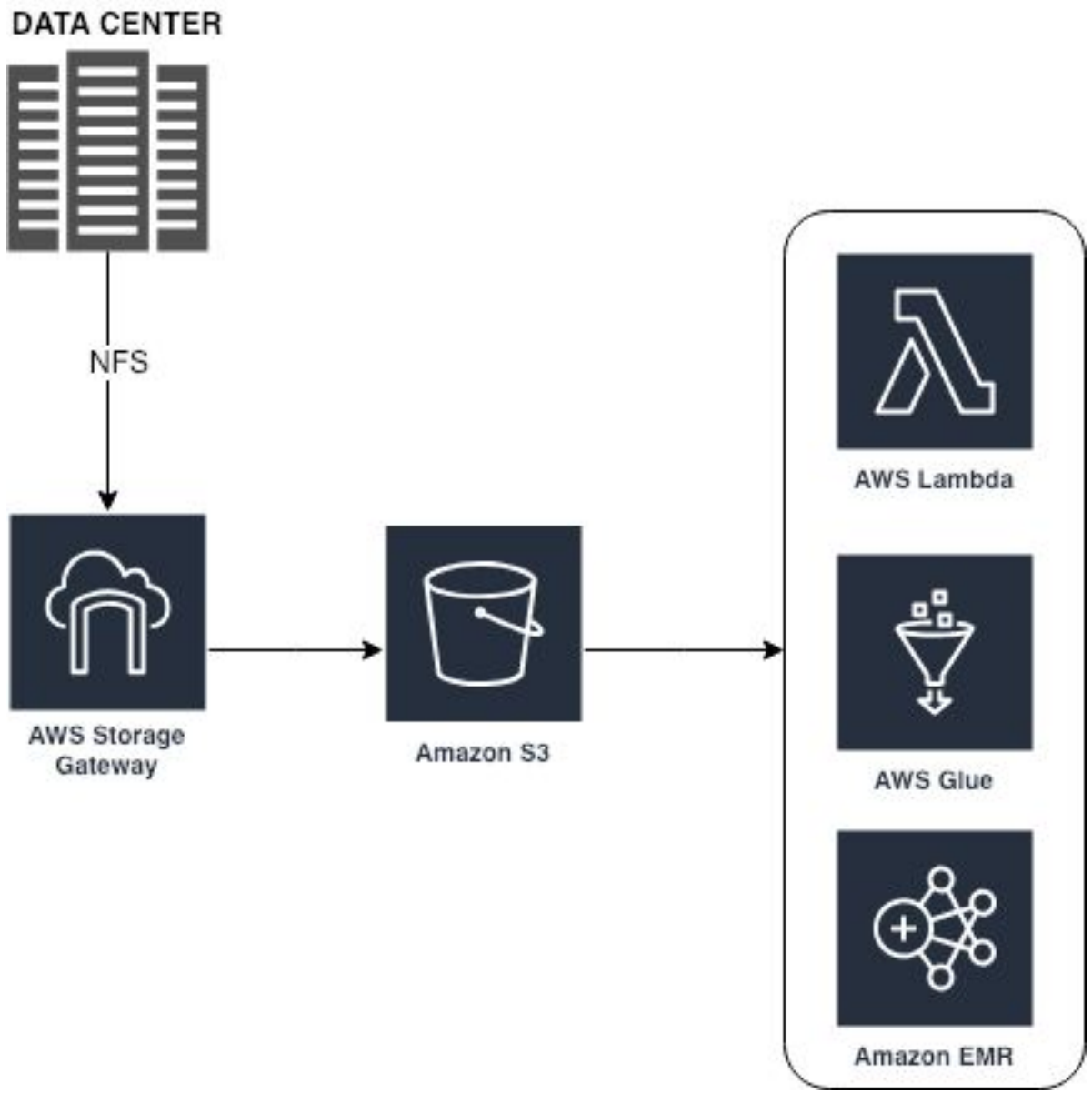
AWS - Storage Gateway.

1. Permite enviar información desde on-premise a AWS.

2. Se podrían enviar los logs de una aplicación que corra on-premise a S3 para ser procesados.

3. Funciona en una VM instalada en nuestro datacenter.

Storage Gateway



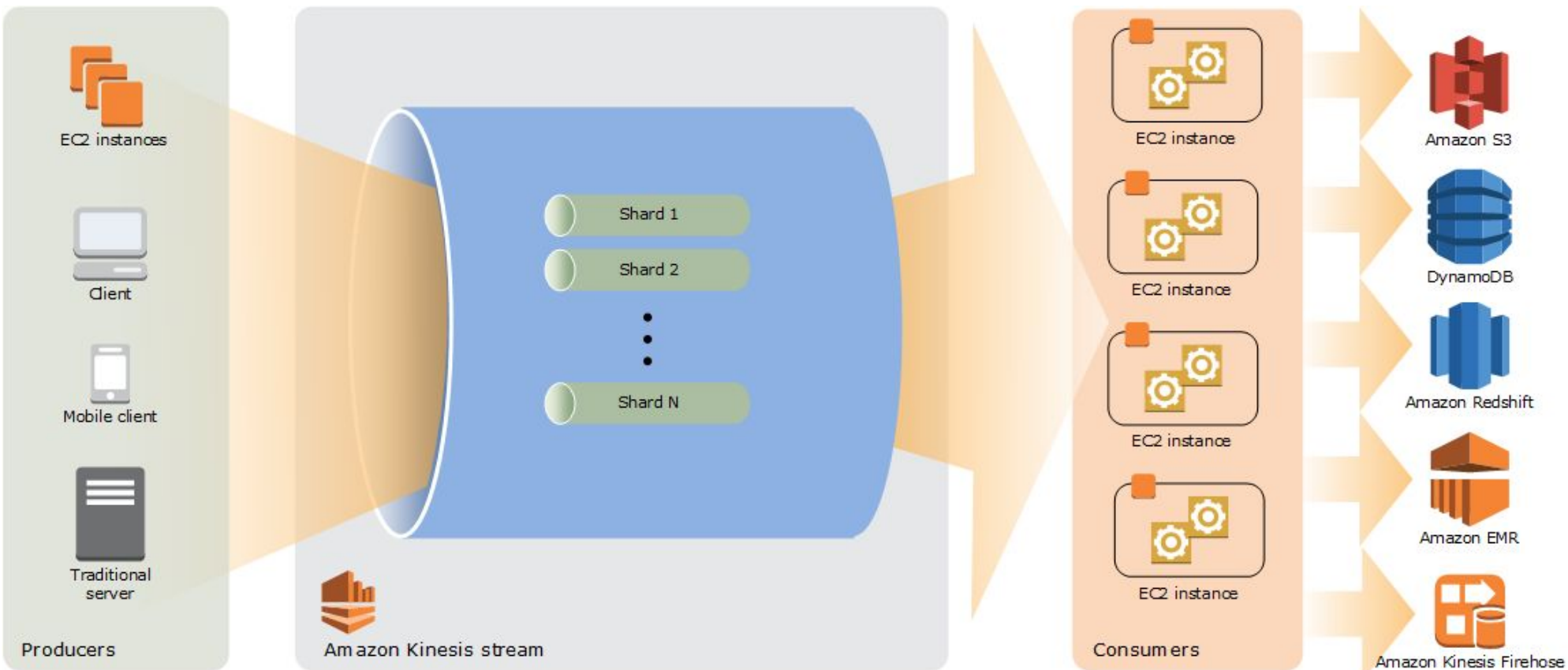
AWS - Kinesis Data Streams

1. Recopilar y procesar grandes cantidades de stream de datos en tiempo real.

2. Logs, social media, market data feeds y web clickstream.

3. Se utiliza también para hacer agregación de datos.

Kinesis Data Streams



Data Record

Es la unidad de dato almacenada en Kinesis Data Streams.

Retention Period

El tiempo que la data es accesible después que se agrega al stream, por defecto es 24 Horas.

Producer

Es el encargado de poner el Data Record en Kinesis Stream.

Consumer

Toma los Data records de kinesis streams para procesarlos.

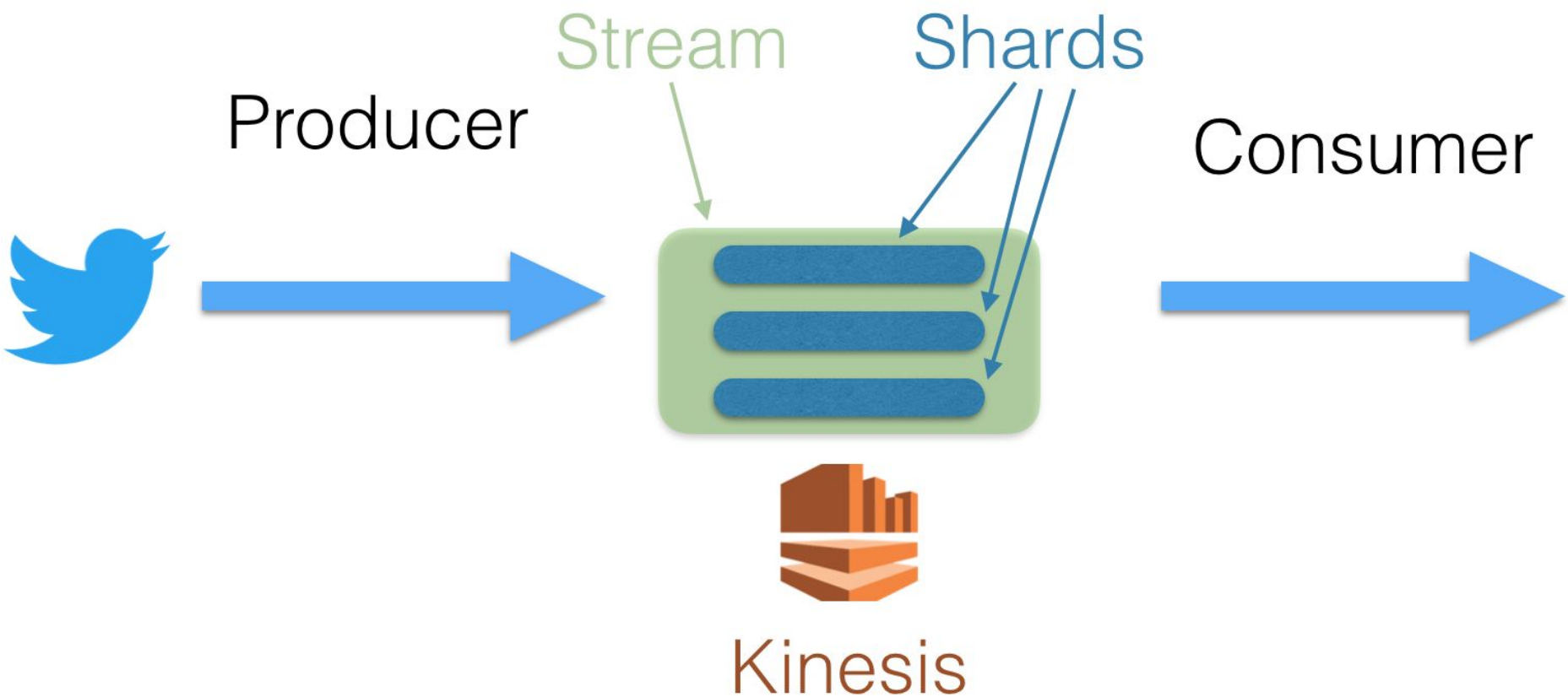
Shard

Es una secuencia de Data Records dentro de un stream.

Partition Key

Se usa para agrupara la data por shard dentro de un stream.

Kinesis Data Streams



Demo - Configuración de Kinesis Data Stream.

Demo - Desplegando Kinesis con Cloudformation.

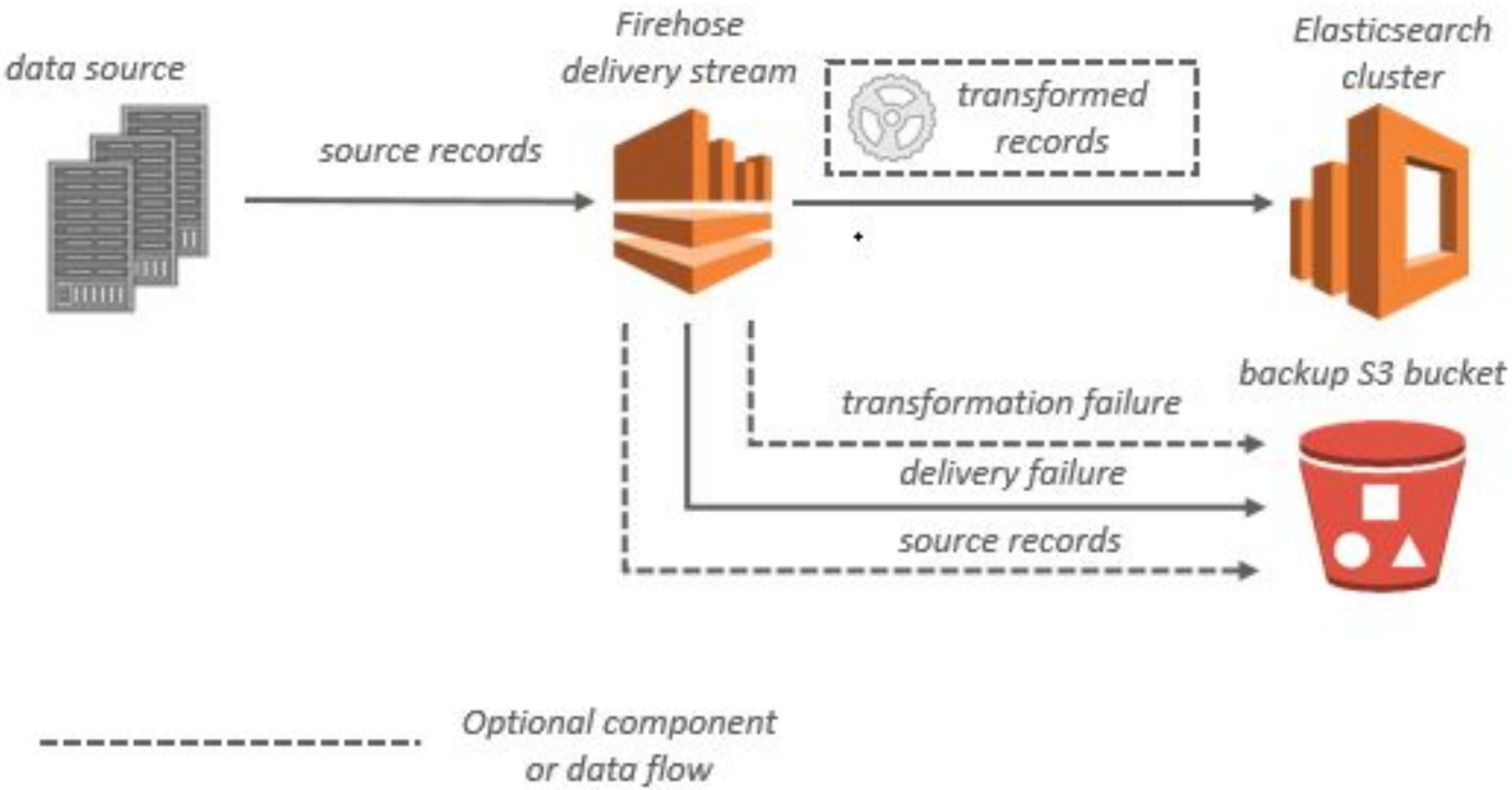
AWS - Kinesis Firehose

1. Es un servicio completamente administrado para la entrega de datos de streaming en tiempo real.

2. Puede usar una lambda para hacer transformaciones a la data.

3. Puede ingestar a diferentes servicios: S3, redshift, ElasticSearch y Splunk.

Kinesis Firehose



Kinesis Firehose

Demo - Configuración de Kinesis Firehose.

AWS - MSK

1. Es un servicios que permite tener Apache Kafka administrado en AWS.

2. Se despliega en clúster y tiene autoreparación de los nodos.

3. Viene con la versión de Apache Kafka 1.1.1

Broker Nodes

Al crear el clúster debemos especificar la cantidad de nodos por AZ.

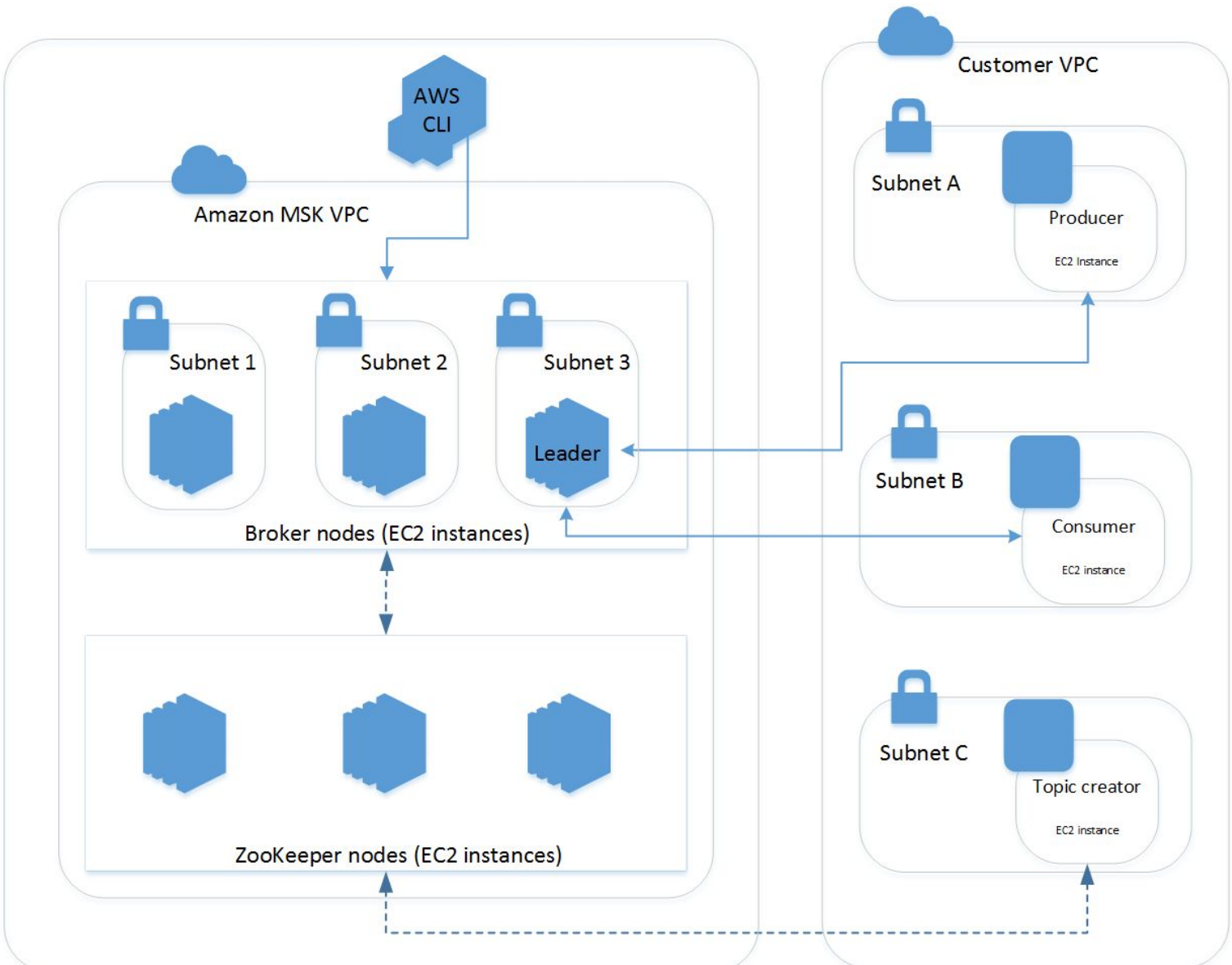
Zookeeper

Mantener datos de nombres y configuración. y para proporcionar una sincronización flexible y robusta.

Zookeeper Nodes

Por defecto al crear el clúster de MSK se crea un nodo de Zookeeper.

MSK


AWS
CLI
Amazon MSK VPC
Subnet 1
Subnet 2
Subnet 3
Leader
Broker nodes (EC2 instances)
ZooKeeper nodes (EC2 instances)
Customer VPC
Subnet A
Producer
EC2 Instance
Subnet B
Consumer
EC2 instance
Subnet C
Topic creator
EC2 instance

Demo - Despliegue de un clúster con MSK.

Transformación de Información

AWS - Glue

1. Servicio administrado para implementación de ETL (Extract, Transform, Load)

2. Provee un contexto de spark para ejecutar trabajos en Python o Scala.

3. Se encarga de la creación del glue catalog para poder ser consultado por otros servicios como Athena.

DPU

Es la unidad de procesamiento de Glue equivalente a 4 vCPU y 16GB RAM.

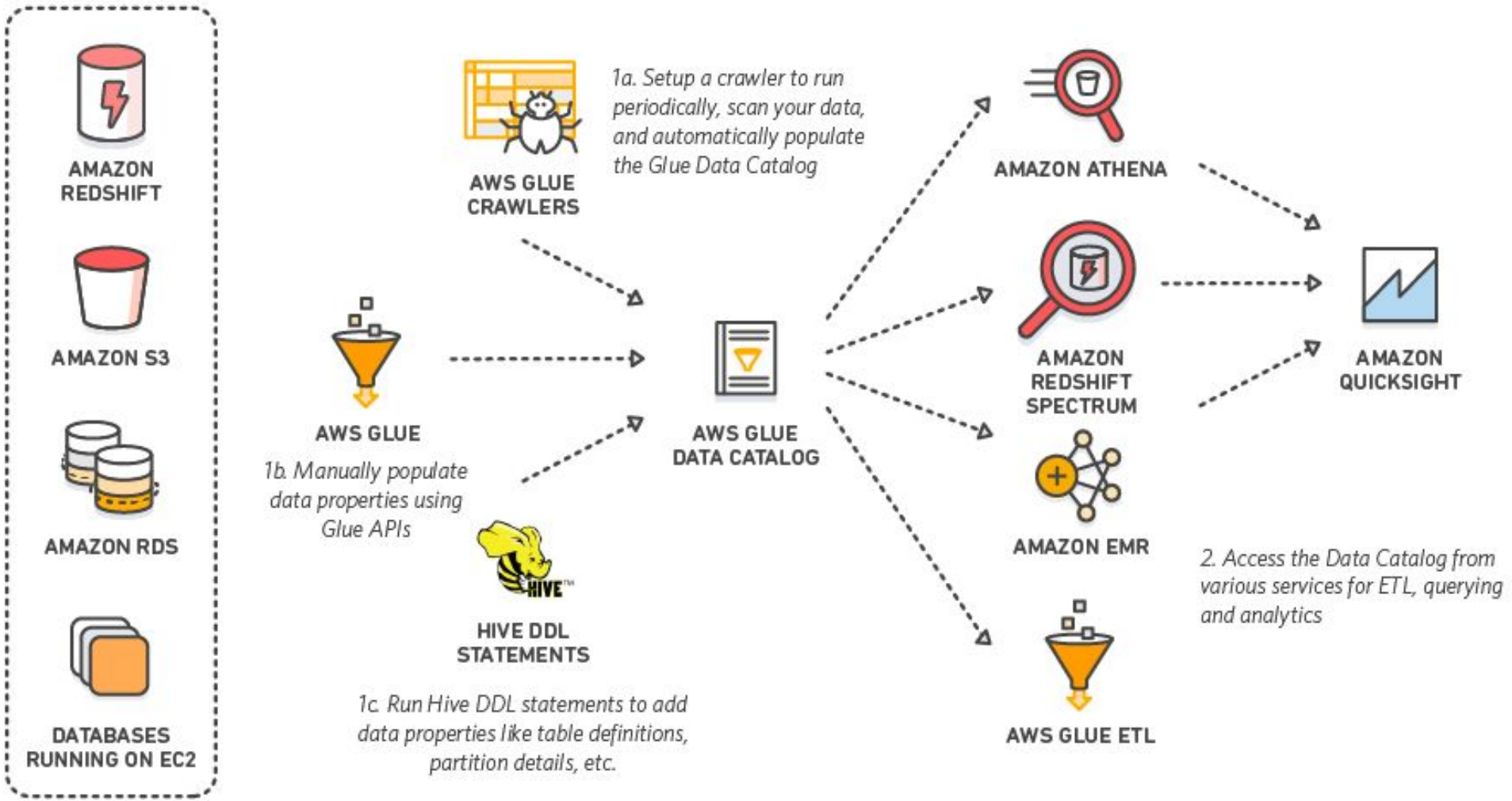
Glue Catalog

Es un almacén de metadatos persistentes. Cada cuenta tiene un Glue Catalog.

Crawler y Classifier

Escanea e identifica la información de origen y crea el glue catalog.

Glue


AMAZON REDSHIFT
AMAZON S3
AMAZON RDS
DATABASES RUNNING ON EC2
AWS GLUE CRAWLERS
1a. Setup a crawler to run periodically, scan your data, and automatically populate the Glue Data Catalog
AWS GLUE
1b. Manually populate data properties using Glue APIs
HIVE DDL STATEMENTS
1c. Run Hive DDL statements to add data properties like table definitions, partition details, etc.
AWS GLUE DATA CATALOG
AMAZON ATHENA
AMAZON REDSHIFT SPECTRUM
AMAZON EMR
AWS GLUE ETL
AMAZON QUICKSIGHT
2. Access the Data Catalog from various services for ETL, querying and analytics

Apache Zeppelin

1. Notebook web que permite Análisis de datos interactivos con SQL, Scala y más.

2. Permite ejecutar consultas usando SQL, Python y Spark.

3. Proyecto de Apache que puede ser ejecutado en los Developer Endpoint y en EMR de AWS.

Demo - Instalando Apache Zeppelin.

Demo - Developer Endpoint.

Demo - Creando nuestro primer ETL Parte 1 - Glue.

Demo - Creando nuestro primer ETL Parte 2 - Glue.

AWS - EMR

1. Crear y escalar clúster administrados con Hadoop en EC2.

2. Correo aplicaciones basadas en: MapReduce, Spark, Pig, Presto, Hive, Impala, Flink, TensorFlow.

3. Provee interacción con otros servicios de AWS como S3, Redshift, DynamoDB y Kinesis.

Bootstrap

Ejecución de algún script que se hace en el clúster antes de iniciar.

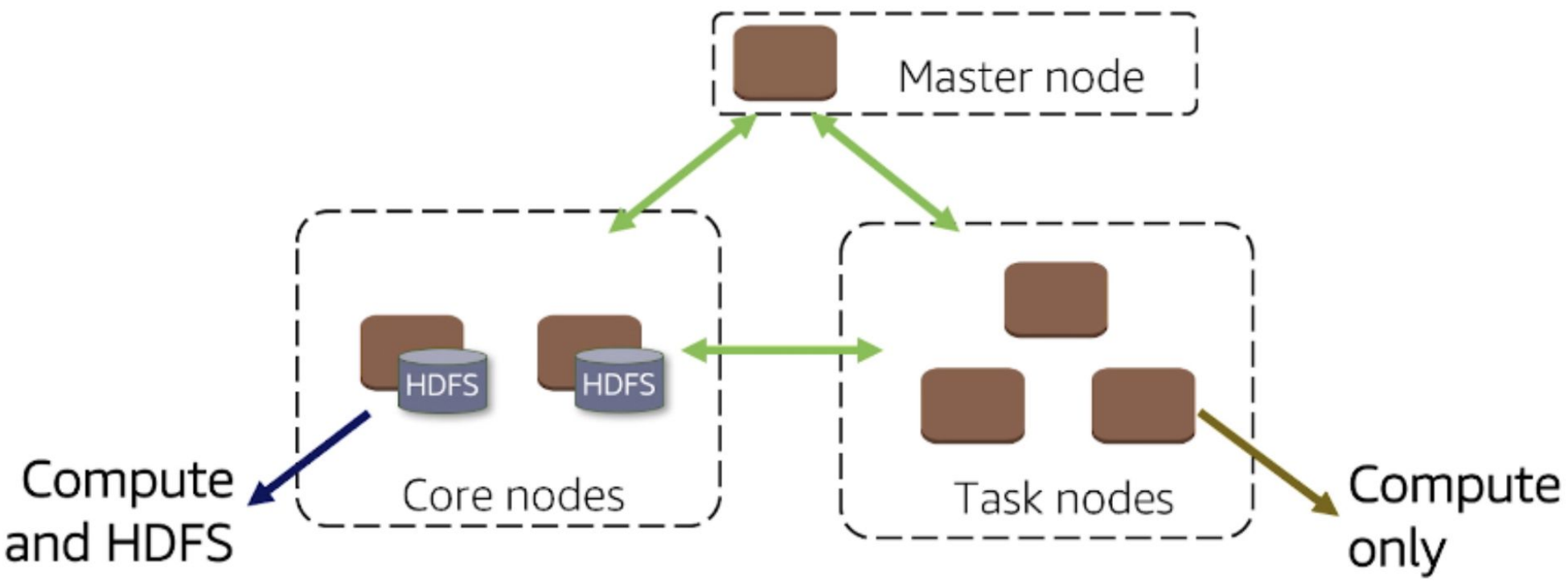
Step

Pasos que se ejecutan para procesar la información en el clúster.

Clúster

Compuesto de un Master node, Core nodes para la información en HDFS y Task nodes de procesamiento.

EMR


Master node
HDFS
HDFS
Core nodes
Task nodes
Compute and HDFS
Compute only

Demo - Desplegando nuestro primer clúster EMR.

Demo - Conectandonos a Apache Zeppelin en EMR.

Demo - EMR y Cloudformation.

AWS - Lambda

Límites

Por cuenta se puede llegar a 20.000 de concurrencia en lambdas.

Firehose

Se puede integrar con Kinesis firehose para realizar transformaciones de datos.

SQS

Suele utilizarse para entornos de alto procesamiento para evitar throtles en las lambdas.

Deployment

Actualización de código de lambdas usando Codepipeline y Boto3.

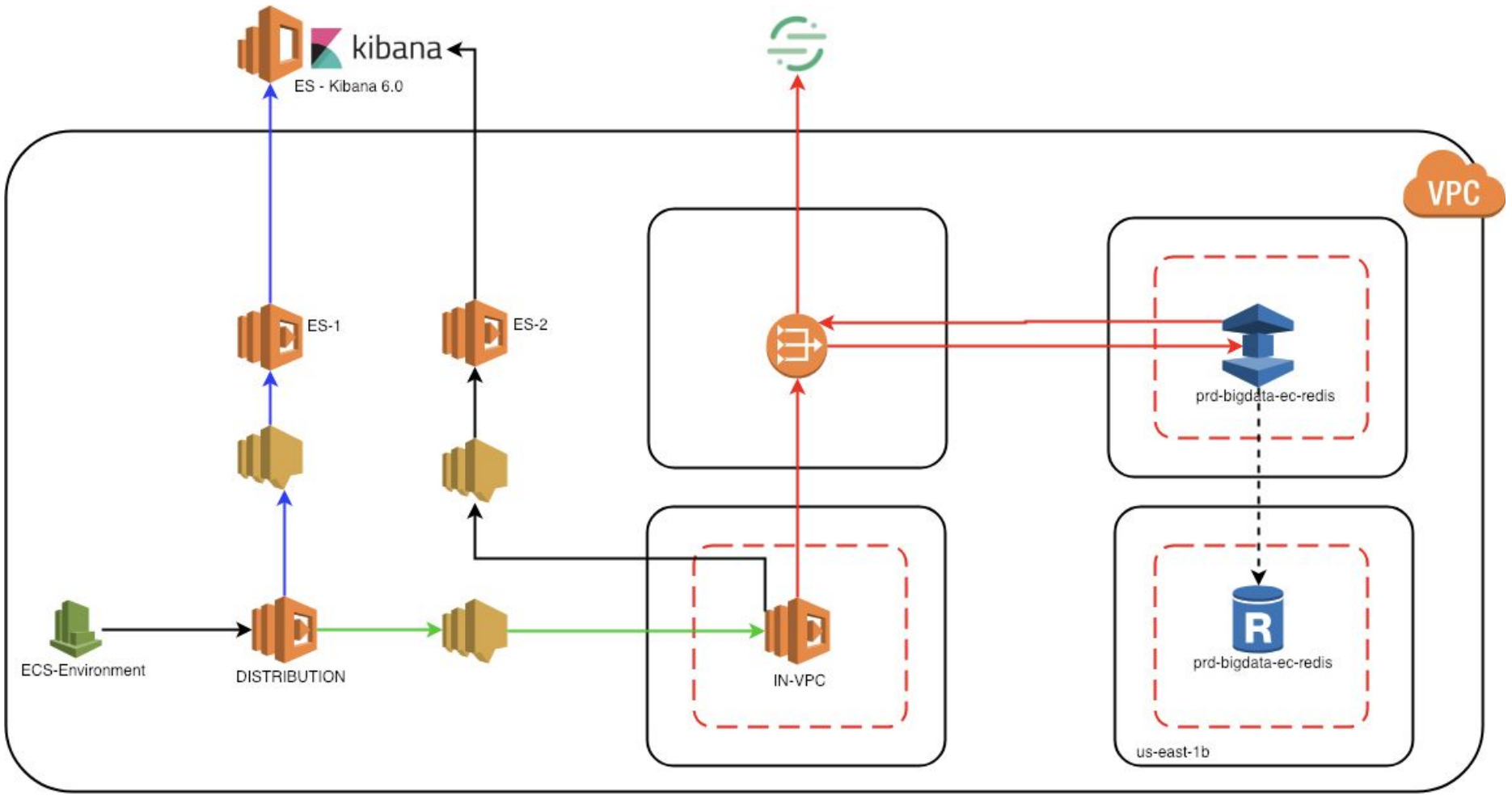
Monitoreo

Librerías de Python para monitoreo de ejecución de código como Rollbar.

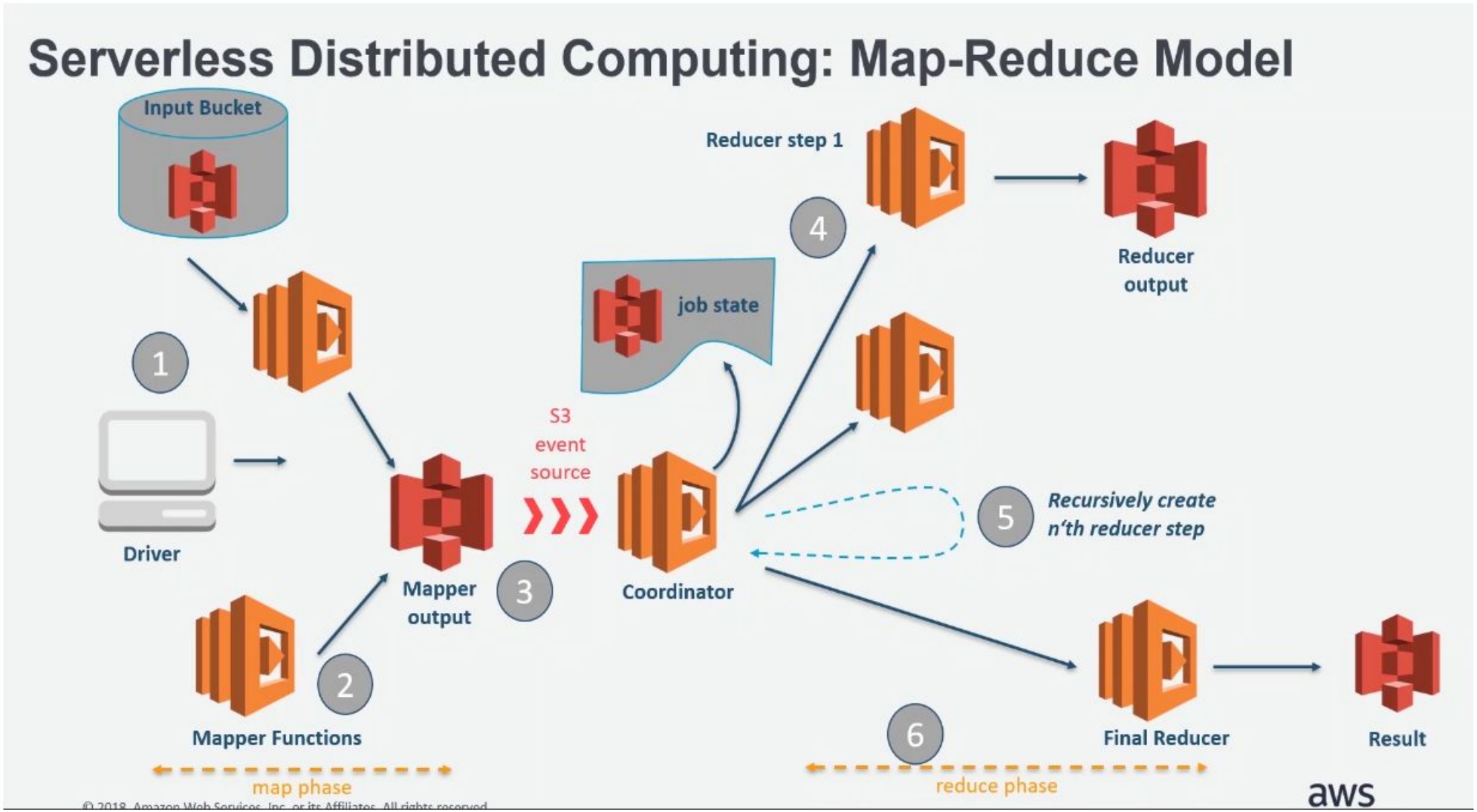
Errores

Al superar los re-intentos se pueden enviar los eventos a una cola SQS o a un topic SNS.

Lambda


kibana
ES - Kibana 6.0
VPC
ES-1
ES-2
prd-bigdata-ec-redis
ECS-Environment
DISTRIBUTION
IN-VPC
prd-bigdata-ec-redis
us-east-1b

Lambda - EMR


Serverless Distributed Computing: Map-Reduce Model
Input Bucket
Reducer step 1
Reducer output
job state
1
4
S3 event source
Driver
Mapper output
3
Coordinator
5
Recursively create n'th reducer step
2
Mapper Functions
6
Final Reducer
Result
map phase
reduce phase
aws

Demo - Configuración de Lambda para BigData.

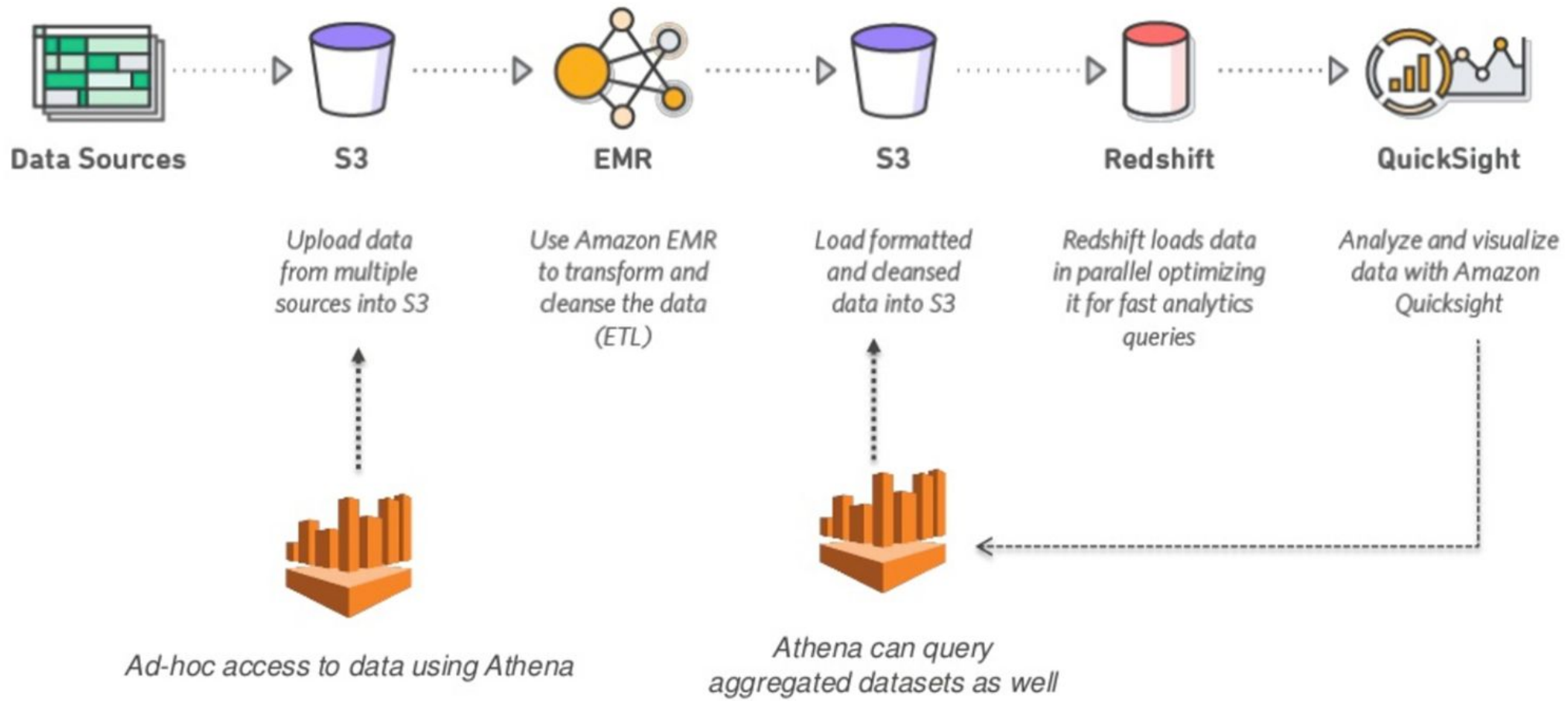
AWS - Athena

1. Servicio de consultas interactivo para data en S3 utilizando SQL.

2. Es serverless, creado en presto y soporta diferentes formatos csv, json, parquet, ORC, tsv…)

3. Provee interacción con otros servicios de AWS como S3, Redshift, DynamoDB y Kinesis.

Athena


Data Sources
S3
EMR
S3
Redshift
QuickSight
Upload data from multiple sources into S3
Use Amazon EMR to transform and cleanse the data (ETL)
Load formatted and cleansed data into S3
Redshift loads data in parallel optimizing it for fast analytics queries
Analyze and visualize data with Amazon Quicksight
Ad-hoc access to data using Athena
Athena can query aggregated datasets as well

JDBC / ODBC

Conexión con herramientas usando JDBC ejemplo: SQL Workbench.

Queries

Consultas pueden ser guardadas para utilizar más adelante.

Seguridad

Permisos granulares por base de datos y por tabla.

Demo - Consultando data de S3 con Athena.

AWS - RedShift

Datawarehouse

Repositorio de datos centralizado que contiene data de múltiples fuentes dentro de una organización.

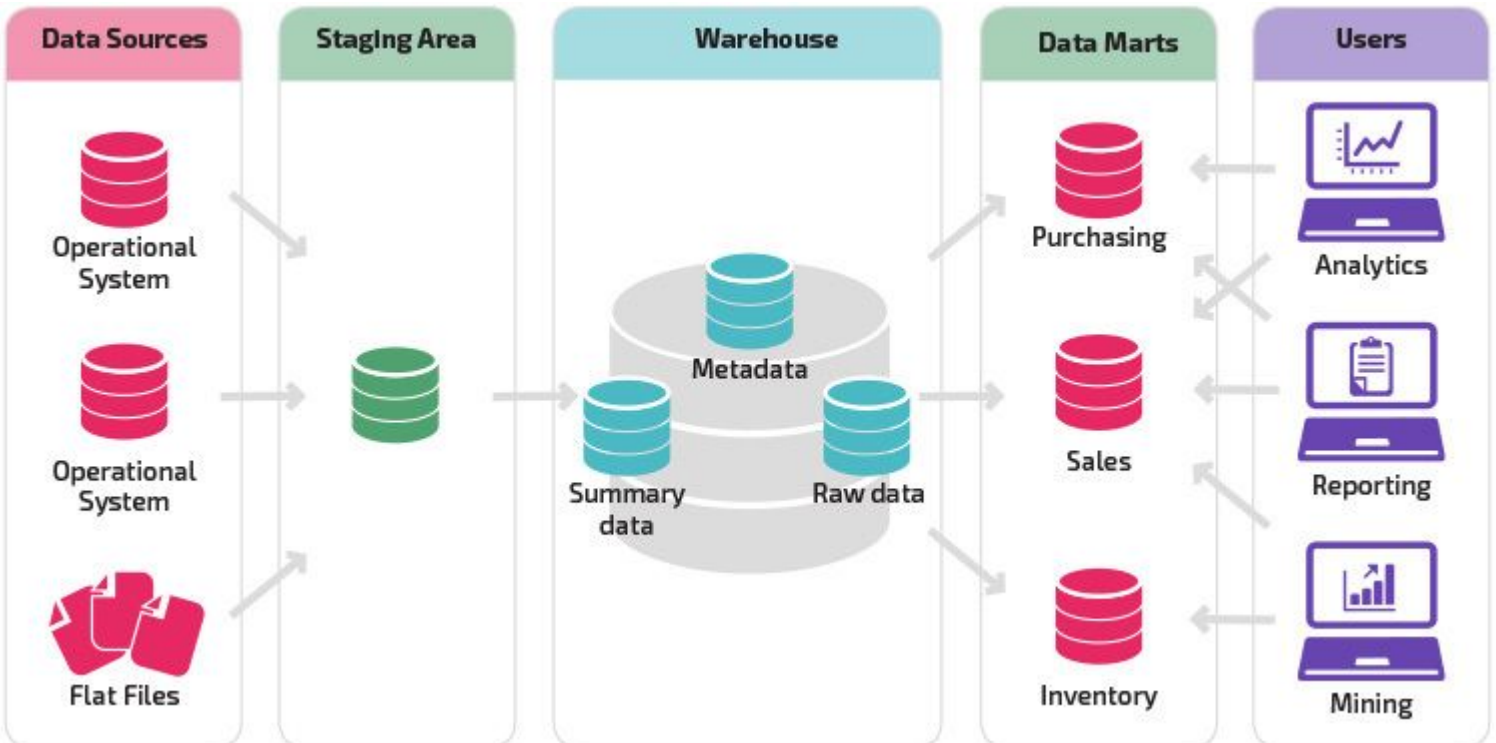
Data Lake

Repositorio de almacenamiento que guarda una cantidad muy grande de raw-data en formato nativo.

Data Mart

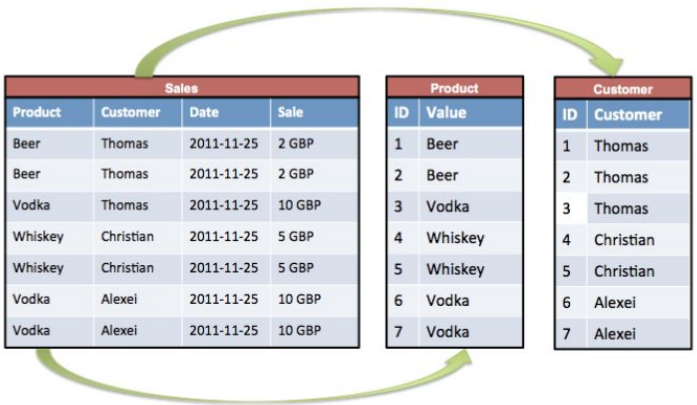
Es un subset del Datawarehouse orientada a una específica línea de negocio.

Datawarehouse



- Sirve para analizar y tomar mejores decisiones.
- Diferentes fuentes de datos.
- Diferentes stakeholders.

Columnar Database



Sales			
Product	Customer	Date	Sale
Beer	Thomas	2011-11-25	2 GBP
Beer	Thomas	2011-11-25	2 GBP
Vodka	Thomas	2011-11-25	10 GBP
Whiskey	Christian	2011-11-25	5 GBP
Whiskey	Christian	2011-11-25	5 GBP
Vodka	Alexei	2011-11-25	10 GBP
Vodka	Alexei	2011-11-25	10 GBP

Product	
ID	Value
1	Beer
2	Beer
3	Vodka
4	Whiskey
5	Whiskey
6	Vodka
7	Vodka

Customer	
ID	Customer
1	Thomas
2	Thomas
3	Thomas
4	Christian
5	Christian
6	Alexei
7	Alexei

Product	
ID	Value
1	Beer
2	Beer
3	Vodka
4	Whiskey
5	Whiskey
6	Vodka
7	Vodka

Customer	
ID	Customer
1	Thomas
2	Thomas
3	Thomas
4	Christian
5	Christian
6	Alexei
7	Alexei

Date	
ID	Date
1	2011-11-25
2	2011-11-25
3	2011-11-25
4	2011-11-25
5	2011-11-25
6	2011-11-25
7	2011-11-25

Sale	
ID	Sale
1	2 GBP
2	2 GBP
3	10 GBP
4	5 GBP
5	5 GBP
6	10 GBP
7	10 GBP

- La estructura columnar optimiza la analítica.
- Reduce los requerimientos de I/O.

1. Servicio de almacenamiento de datos en nube administrado a escala de PB.

2. El servicio se lanza en un clúster de instancias.

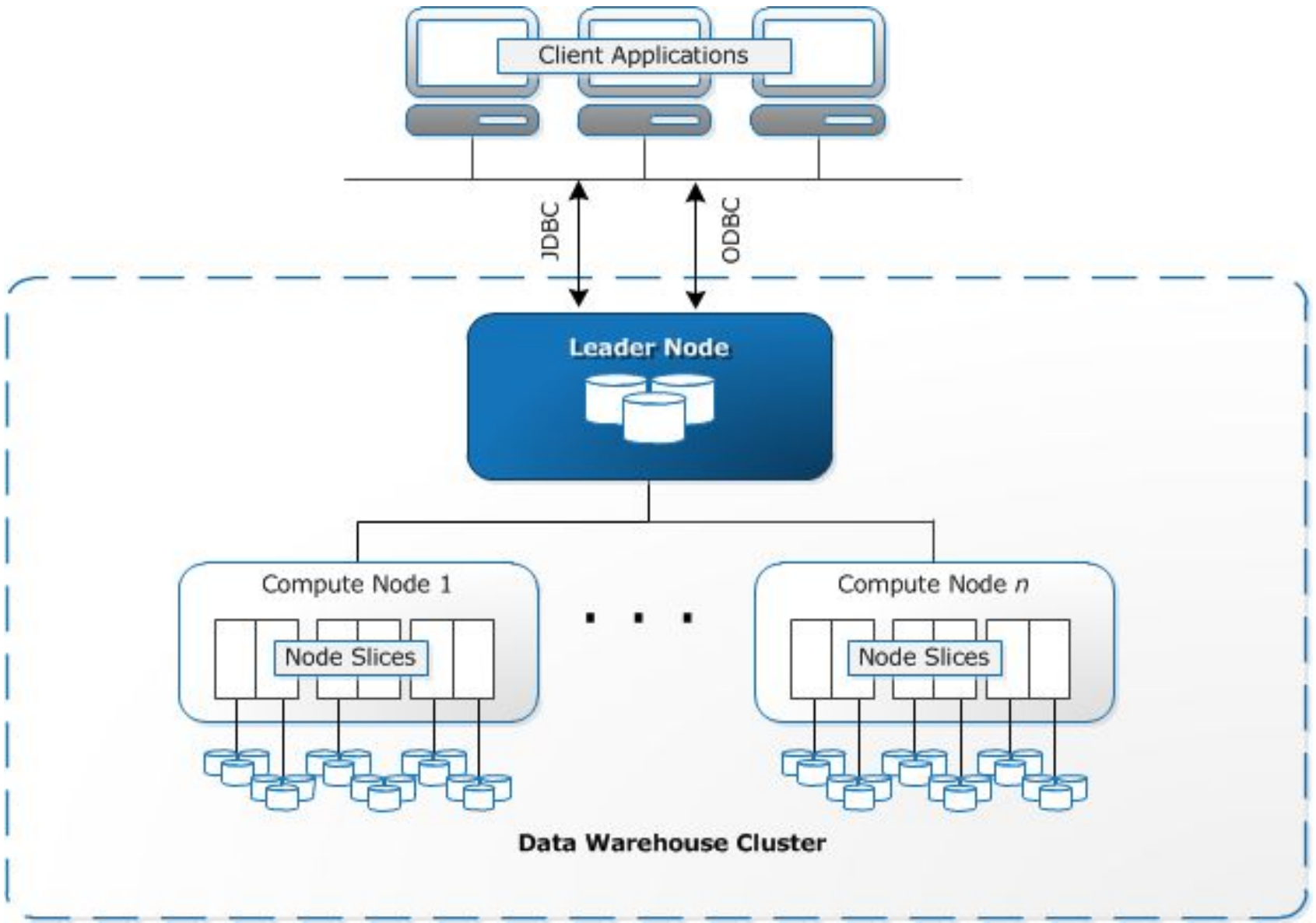
3. Sirve para consultas complejas SQL sobre cantidades grandes de datos a nivel columnar.

4 Está basado en PosgreSQL y esta disenado para OLAP y aplicaciones de BI.

5 Hace compresión de la data para optimizar el I/O.

6 Utiliza caché para ciertos tipos de consulta y no tener que volver a procesar información.

RedShift


Client Applications
JDBC
ODBC
Leader Node
Compute Node 1
Node Slices
Compute Node n
Node Slices
Data Warehouse Cluster

Seguridad

Servicio de almacenamiento de datos en nube administrado a escala de PB.

Costos

El servicio se lanza en un clúster de instancias.

Queries

Sirve para consultas complejas SQL sobre cantidades grandes de datos.

Demo - Lanzando nuestro primer clúster de RedShift.

AWS - Lake Formation

1. Facilita y permite la creación de un data lake en días con muy buena seguridad.

2. Tiene integración con diferentes fuentes, hasta On-premise usando JDBC.

3. Identifica los orígenes y crea las tablas basado en su estructura (Crawlers).

ETL

Después de hacer el Crawl, se encarga de orquestar el ETL en Glue para transformar la data.

Clean

Limpia y elimina data duplicada utilizando Machine Learning llamado FindMatch.

Optimización

Optimiza las particiones de S3 para consultar más eficientemente la data.

Seguridad

Cifrado automático de la data en S3 utilizando SSE-KMS.

Acceso

Control de permisos por usuarios por bases de datos, tablas y columnas.

Auditoría

Logging a nivel de auditoría registrados en Cloudtrail.

Owners

Se pueden designar data owners para controlar permisos por usuarios.

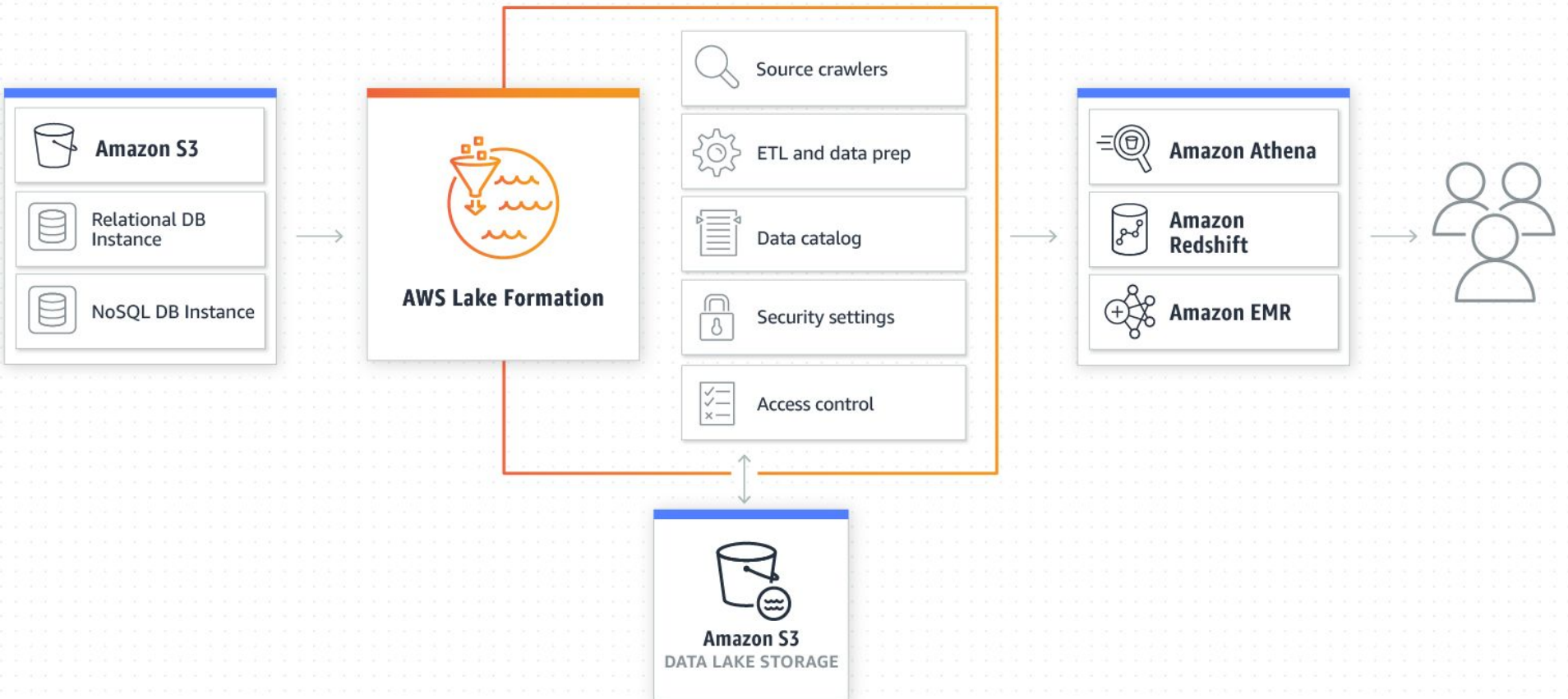
Discover

Descubre data relevante para implementar análisis.

Insights

Analytics desde otros servicios como EMR y RedShift.

Lake Formation



AWS - ElasticSearch

1. Es un motor de búsqueda basado en Lucene. Busca texto completo y JSON sin esquema.

2. Se despliega en un clúster en AWS compuesto de varios nodos.

3. La solución viene integrada con Kibana y Logstash.

Autenticación	Se puede integrar con AWS Cognito para manejar pool de usuarios o usar identity federation.
Cifrado	Se pueden cifrar los datos en reposo y en tránsito con KMS.
Integración	Puede recibir información de Kinesis Firehose y Lambda.

índice

Es como una base de datos que guarda información relacional. Es un nombre lógico que apunta a uno o más shards.

Estructura

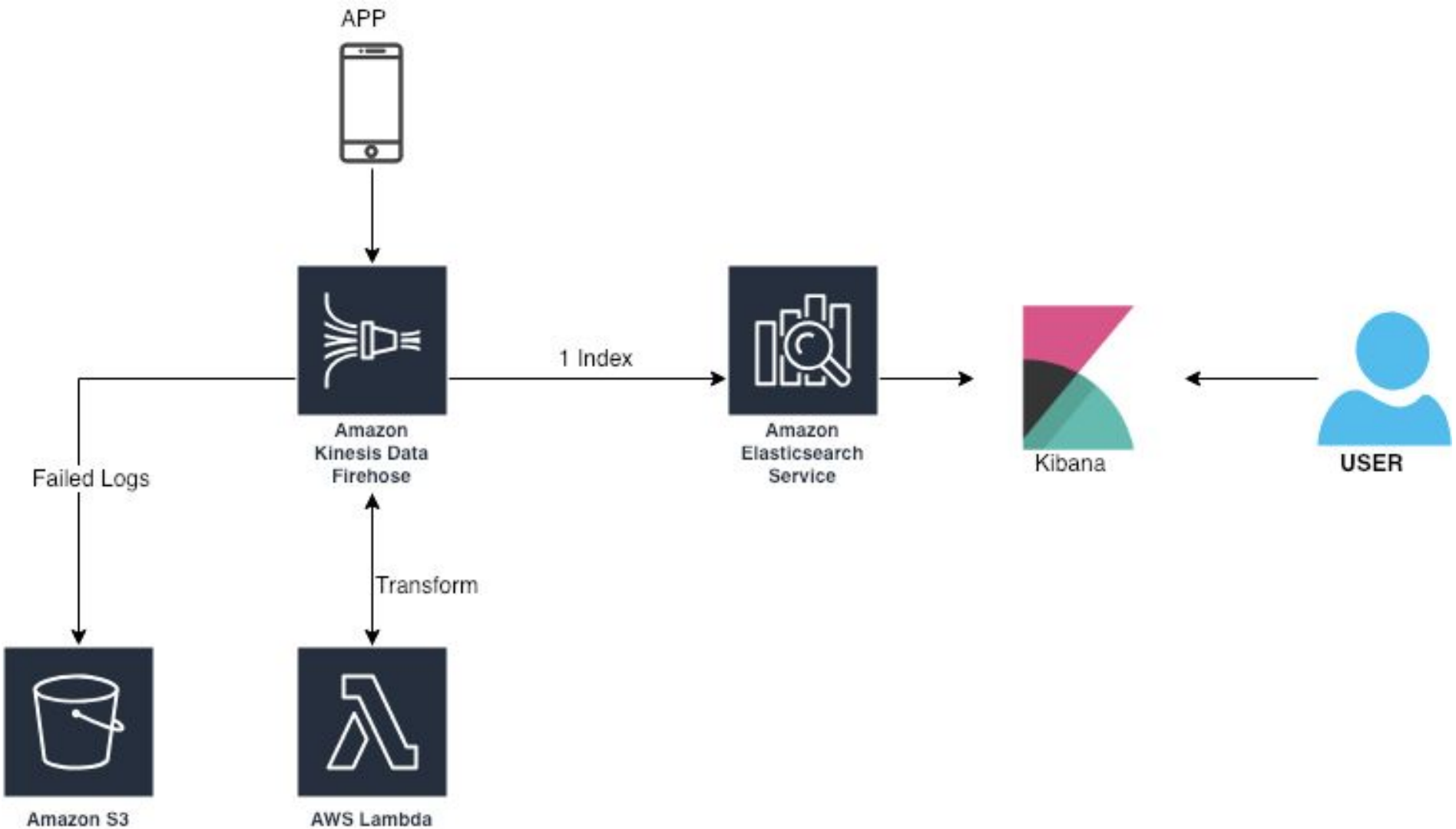
MySQL => Databases => Tables => Columns/Rows

ES => Indices => Types => Documents with Properties

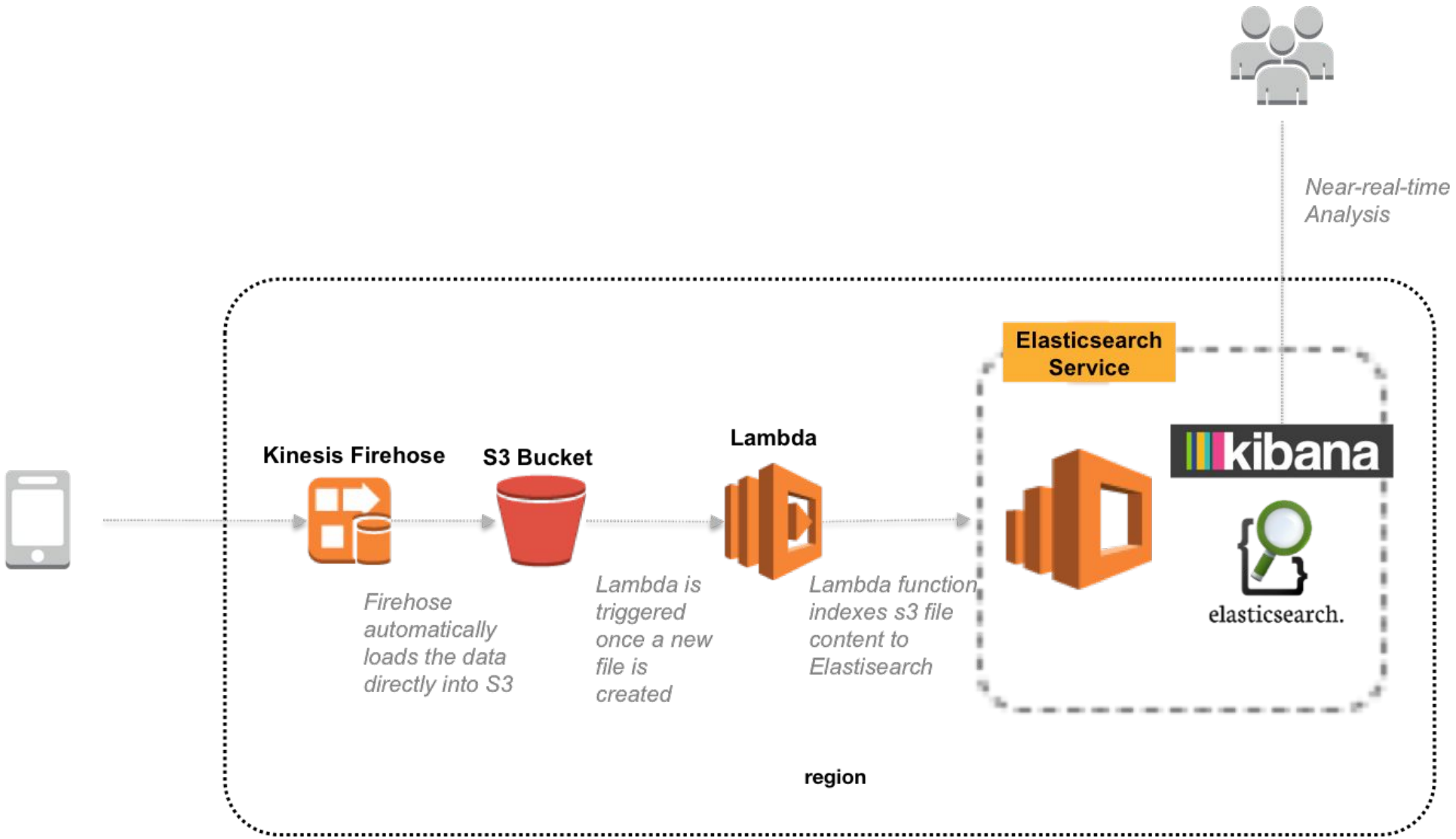
Shard

Un índice se puede dividir en múltiples shards y estos se almacenan en diferentes nodos.

ElasticSearch



ElasticSearch



Demo - Dimensionando nuestro clúster de ES .

Demo - Creando nuestro primer clúster de ES .

AWS - Kibana

Función

Permite visualizar de forma gráfica la data que tenemos en ElasticSearch.

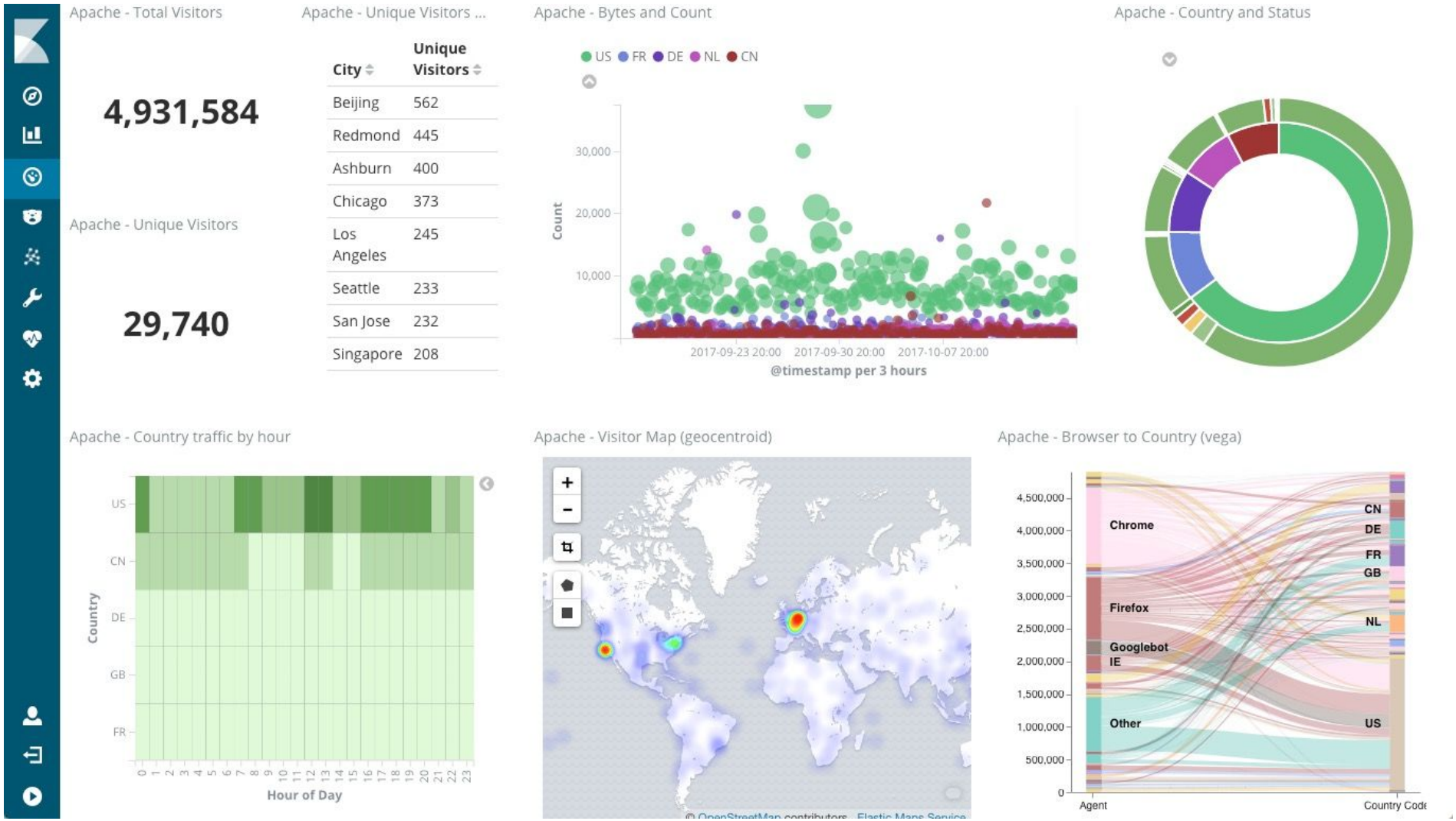
Visualización

Provee diferentes opciones de visualización (mapas de calor, barras, tortas, tendencias…)

Integración

Permite el uso de plugins de terceros para visualización y analítica.

Kibana


Apache - Total Visitors
4,931,584
Apache - Unique Visitors
29,740
Apache - Unique Visitors ...
City
Unique Visitors
Beijing 562
Redmond 445
Ashburn 400
Chicago 373
Los Angeles 245
Seattle 233
San Jose 232
Singapore 208
Apache - Bytes and Count
US FR DE NL CN
Count
30,000
20,000
10,000
2017-09-23 20:00
2017-09-30 20:00
2017-10-07 20:00
@timestamp per 3 hours
Apache - Country and Status
Apache - Country traffic by hour
US
CN
DE
GB
FR
Country
Hour of Day
Apache - Visitor Map (geocentroid)
Apache - Browser to Country (vega)
4,500,000
4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Chrome
Firefox
Googlebot
IE
Other
CN
DE
FR
GB
NL
US
Agent
Country Code

Demo - Explorando visualizaciones con Kibana .

AWS - QuickSight

1. Es un servicio de Business intelligence en Cloud para análisis y visualización.

2. Cuenta con un cliente para dispositivos móviles.

3. Puede escalar hasta 10.000 usuarios y su cobro es por demanda.

ML

Incluye funcionalidades de ML como detección de anomalías, prevención y alertas. Utiliza SPICE como motor.

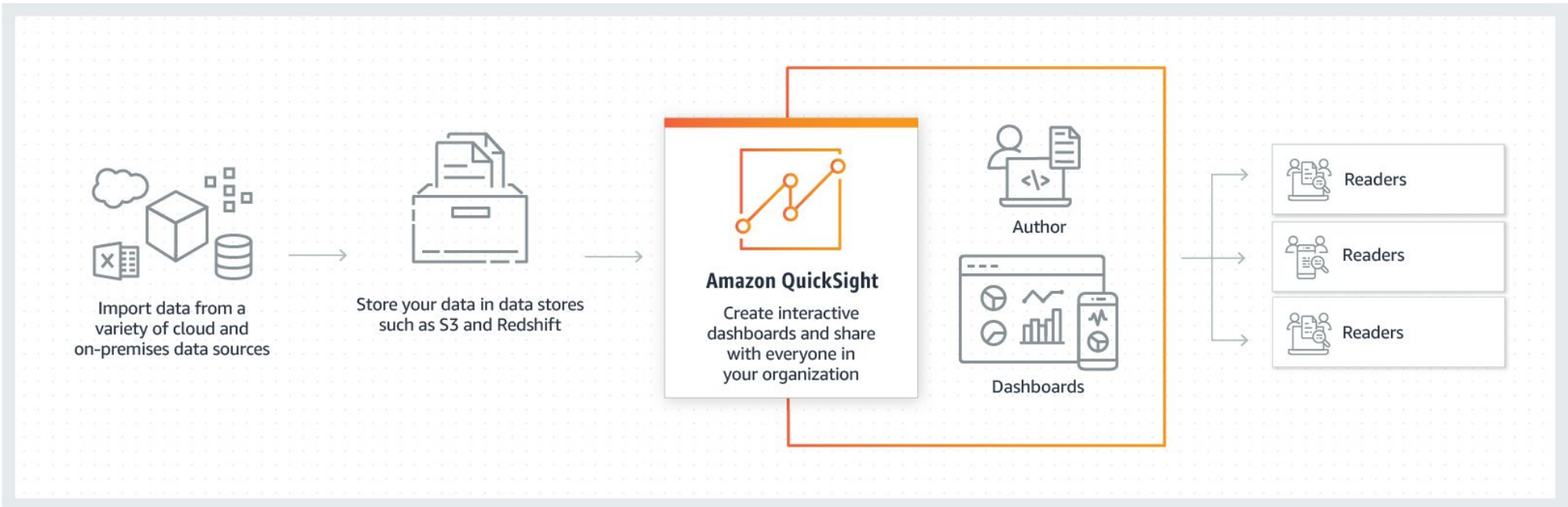
Visualización

Usando API permite realizar el embebido de Dashboards en diferentes sistemas.

Integración

Permite integración con gran variedad de servicios dentro de AWS y de terceros.

Quicksight


Import data from a variety of cloud and on-premises data sources
Store your data in data stores such as S3 and Redshift
Amazon QuickSight
Create interactive dashboards and share with everyone in your organization
Author
Dashboards
Readers
Readers
Readers

Demo - Visualización con QuickSight .

Seguridad, Orquestación y Automatización

Seguridad en los Datos.

Cifrado

Utilizar en todos los servicios cifrado en reposo y en tránsito con KMS.

Permisos

Gestionar todos los permisos basados en usuarios y en recursos de forma granular.

Servicios

Utilizar servicios administrados y servicios de seguridad.

Monitoreo

Monitoreo de los datos, dónde están, quién, cuándo, por qué….

Contingencia

Replicación de datos, pruebas de DRP, almacenar data histórica.

Test

Utilice datos para hacer las pruebas, no datos de producción.

AWS - Macie

1

Aprendizaje automático para descubrir, clasificar y proteger datos confidenciales automáticamente.

2

El servicio administrado monitoriza la actividad de acceso a los datos en busca de anomalías y genera alertas.

3

Se encuentra disponible para proteger datos almacenados en Amazon S3.

Predictivas

Lectura/Escritura en un bucket.

Compliance

Personally Identifiable Information (PII) o credenciales de acceso.

Disruption

Cambios de configuración que puedan afectar un servicio.

Ransomware

Detecta software potencialmente malintencionado.

Suspicious

Accesos a sus recursos desde IP o sistemas sospechosos.

Privileges

Identifica intentos de un usuario/role para obtener privilegios elevados.

Anonymous

Acceso a los recursos tratando de ocultarse tras una identidad verdadera.

Permissions

Identifica recursos sensibles de acuerdo a sus políticas permisivas.

Data Loss

Riesgos o anomalías de acceso a su data más importante

Credential

Credenciales de acceso comprometidas.

Location

Intentos de acceso a la información desde una ubicación desconocida.

Hosting

Almacenamiento de software riesgoso o malintencionado.

Macie

Inscriba su cuenta de AWS con Amazon Macie

Seleccione los buckets para la clasificación y la detección de contenido

Revise sus alertas en el panel de Amazon Macie

Demo - Configuración de Macie .

Demo - Generando alertas con Macie .

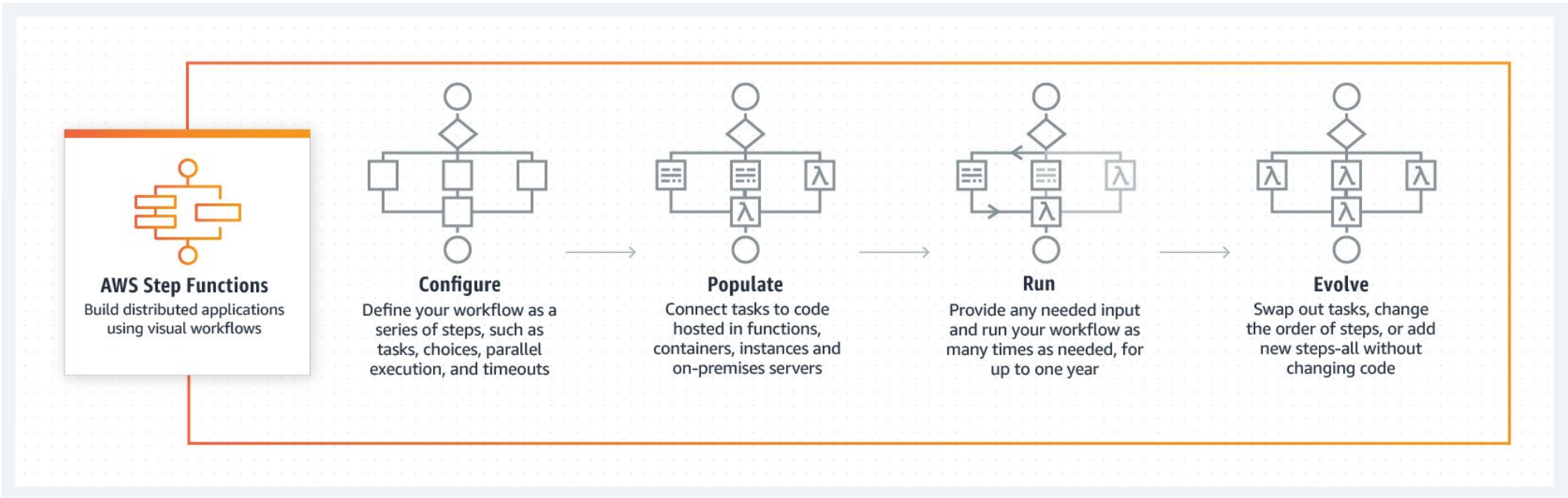
AWS - Step Functions

1. Permite coordinar múltiples servicios de AWS en flujos de trabajo sin servidor.

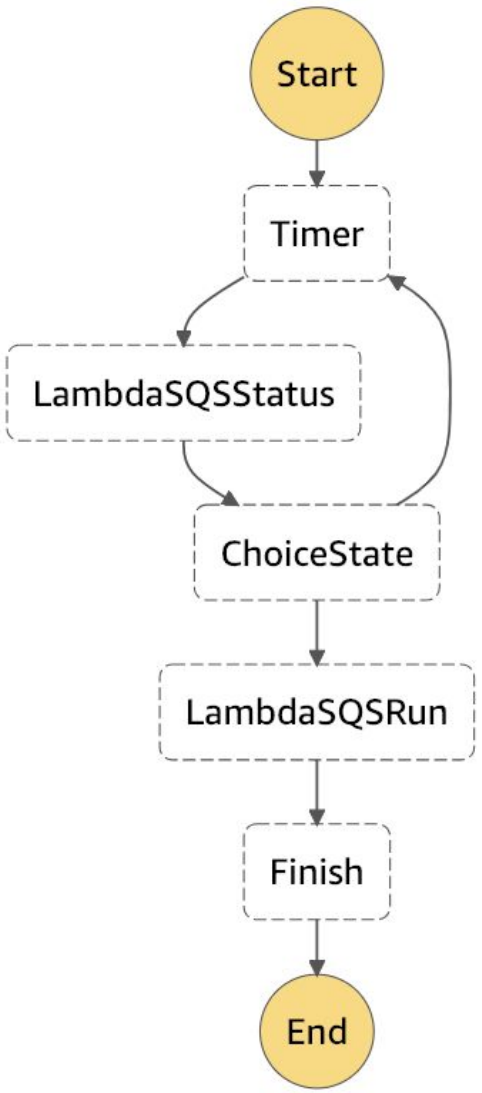
2. Se componen de una serie de pasos, con la salida de un paso que actúa como entrada en el siguiente.

3. Puede monitorear cada paso de la ejecución.

Step Functions



Step Functions



Demo - Configuración de Step Functions .

Apache Airflow

1. Permite crear, monitorear y orquestar flujos de trabajo.

2. Los pipeline son son configurados usando Python.

3. Es muy flexible, permite modificación de excutors, operators y demás entidades dentro de Airflow.

DAG

Directed Acyclic Grap, es una colección de todas las tareas que quiere correr con sus dependencias y relaciones.

Operator

Describe una tarea que corre independiente de los otros.

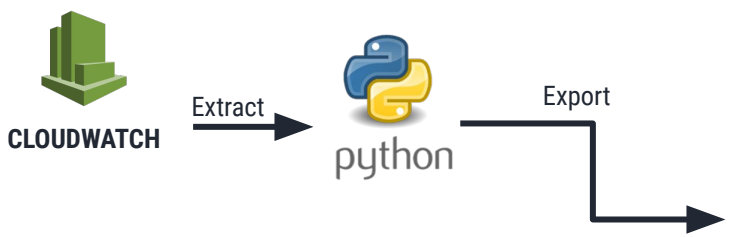
GCP

Cloud Composer es el nombre del servicio que provee Google Cloud Platform para Airflow.

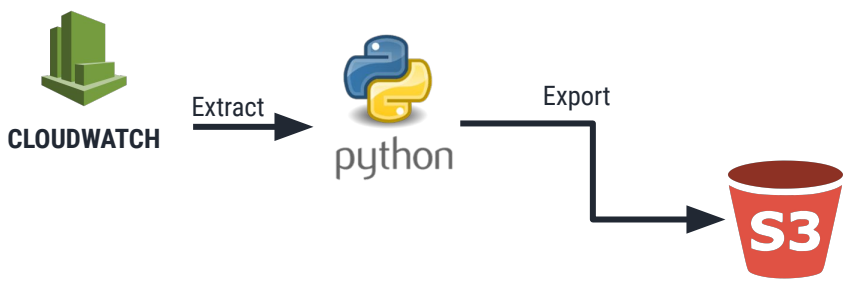
Demo - Desplegando GCP - Cloud Composer.

Arquitecturas de Referencia

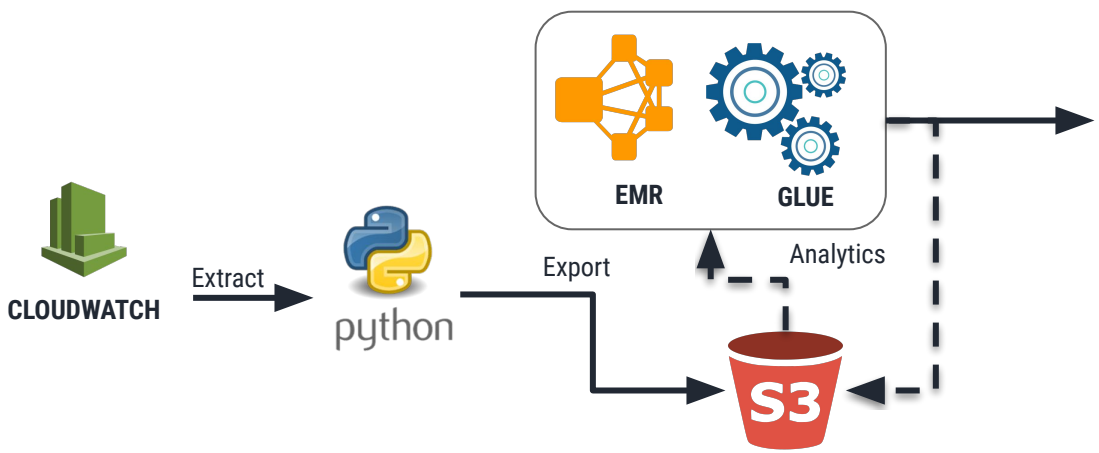
Procesamiento Batch



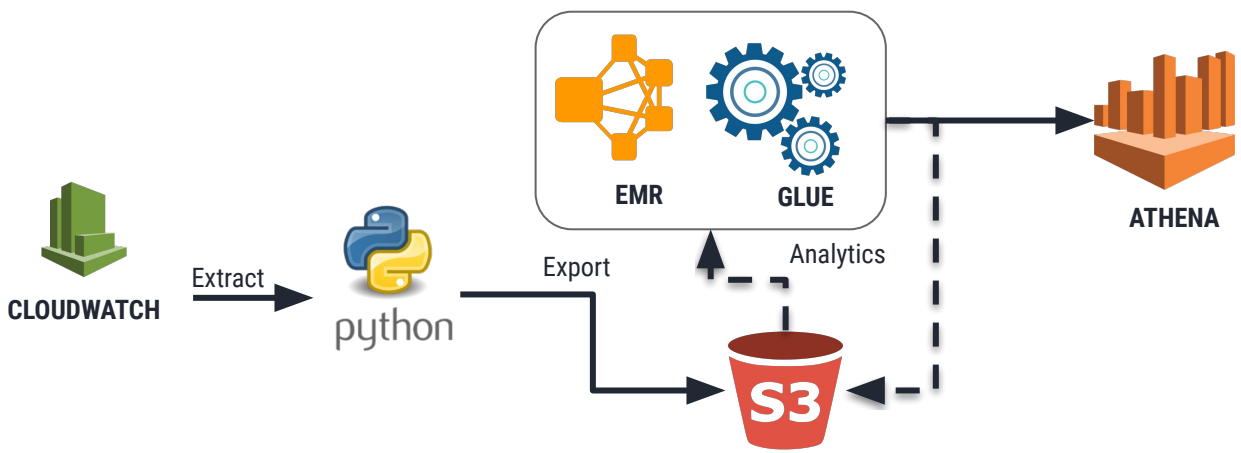
Procesamiento Batch



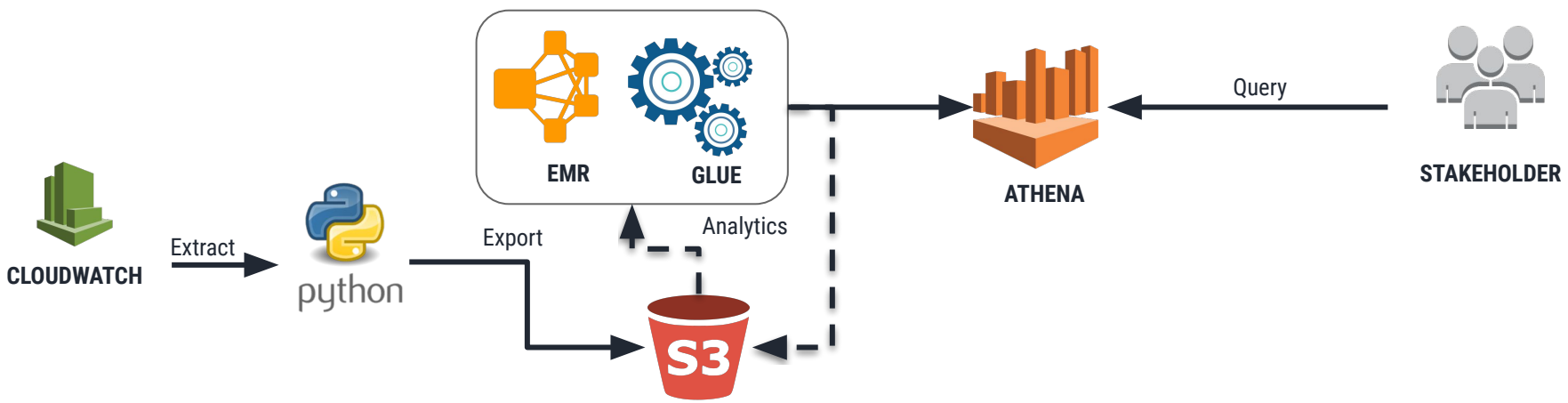
Procesamiento Batch



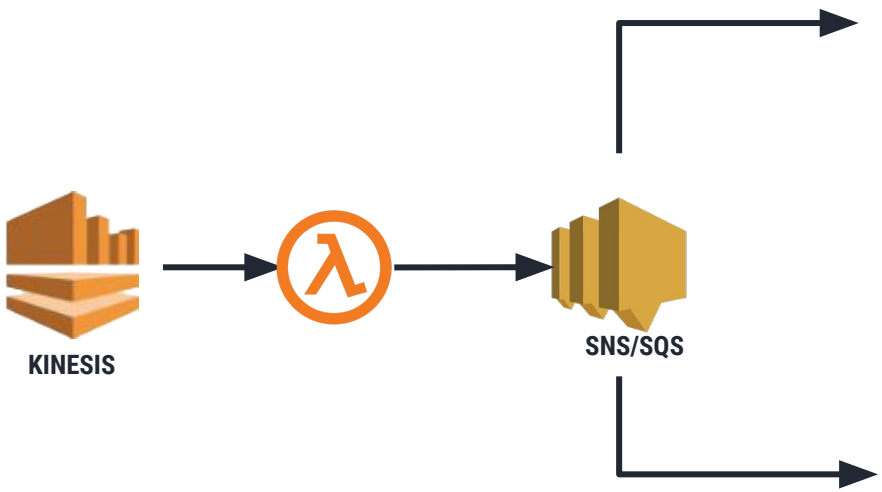
Procesamiento Batch



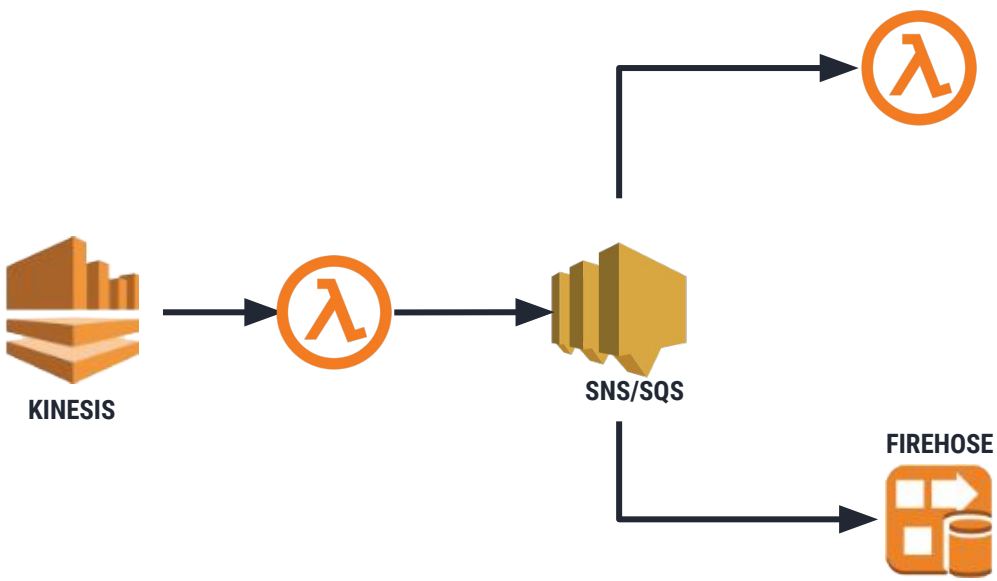
Procesamiento Batch



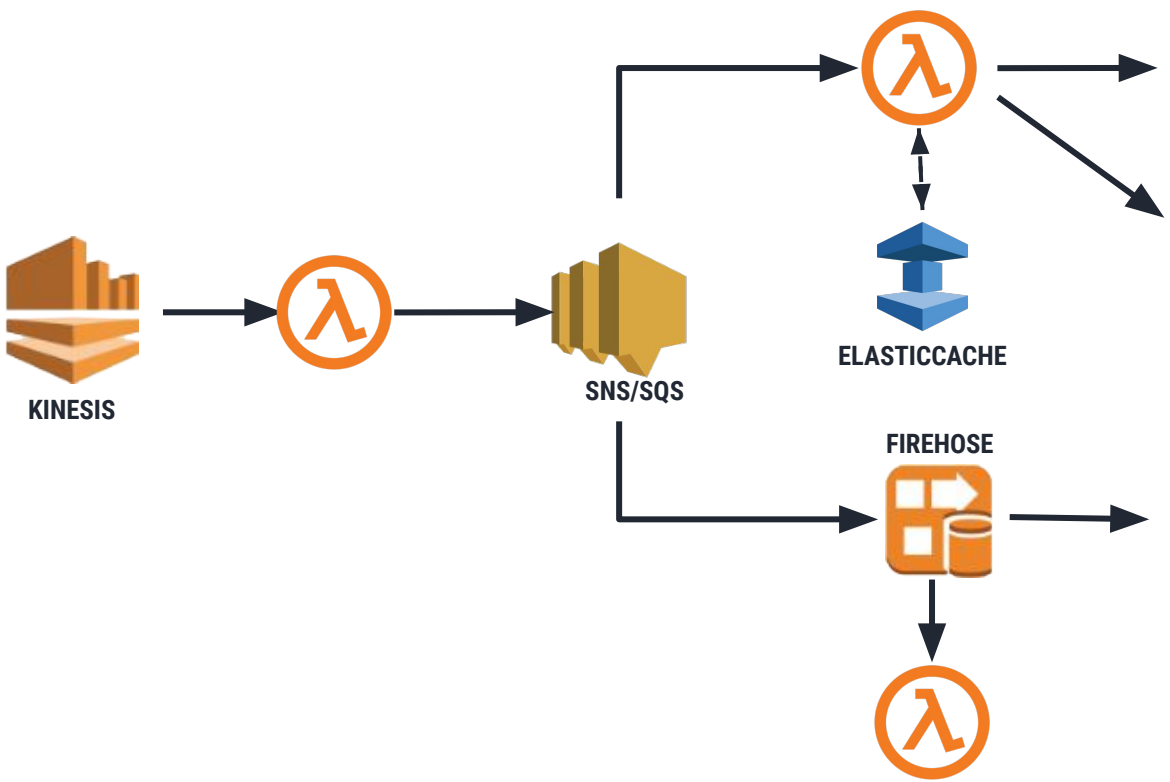
Procesamiento Tiempo Real



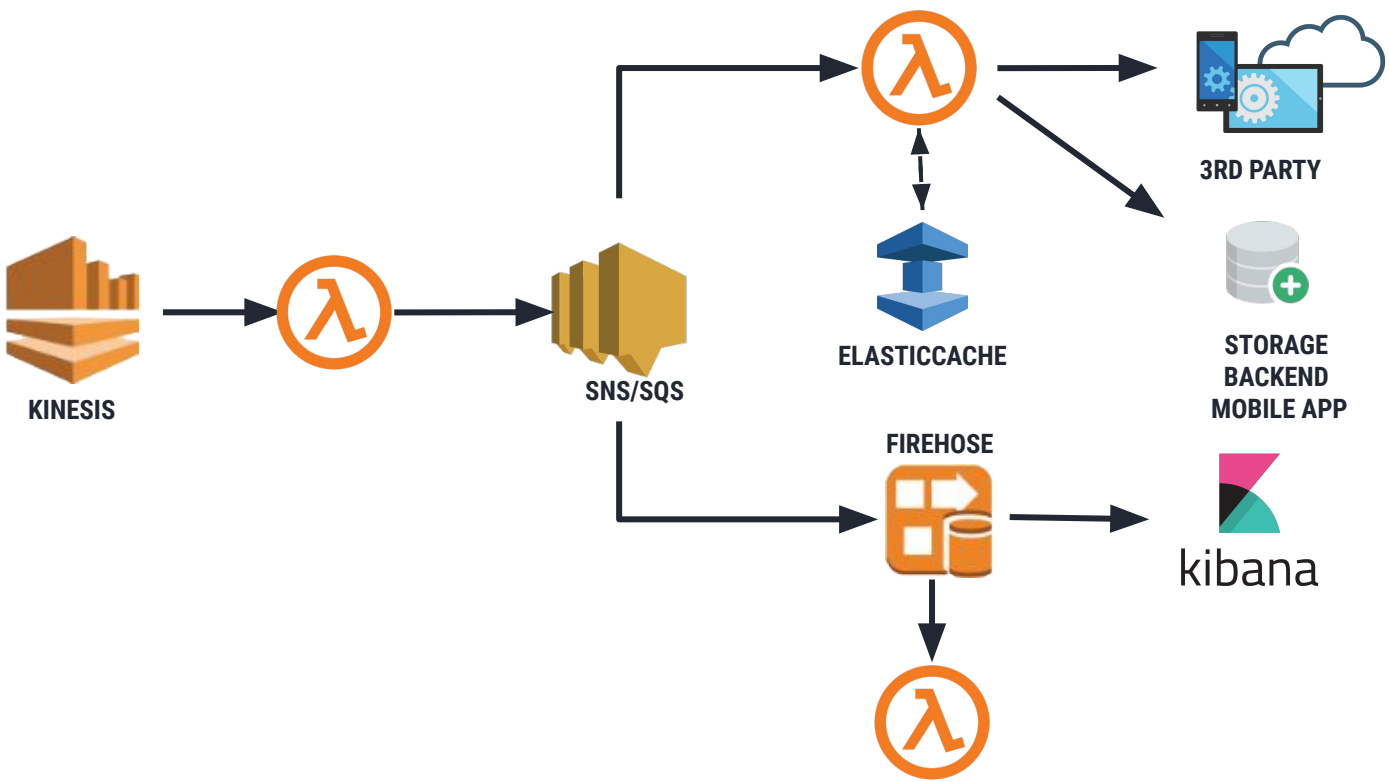
Procesamiento Tiempo Real



Procesamiento Tiempo Real



Procesamiento Tiempo Real



Procesamiento Tiempo Real