



Immersation Day

Data Analytics

Jeong, Chanhui

2022.12.01.

Agenda

Real-time analytics with Kinesis

Kinesis Data Streams

Kinesis Data Firehose

Kinesis Data Analytics

Hands on Lab

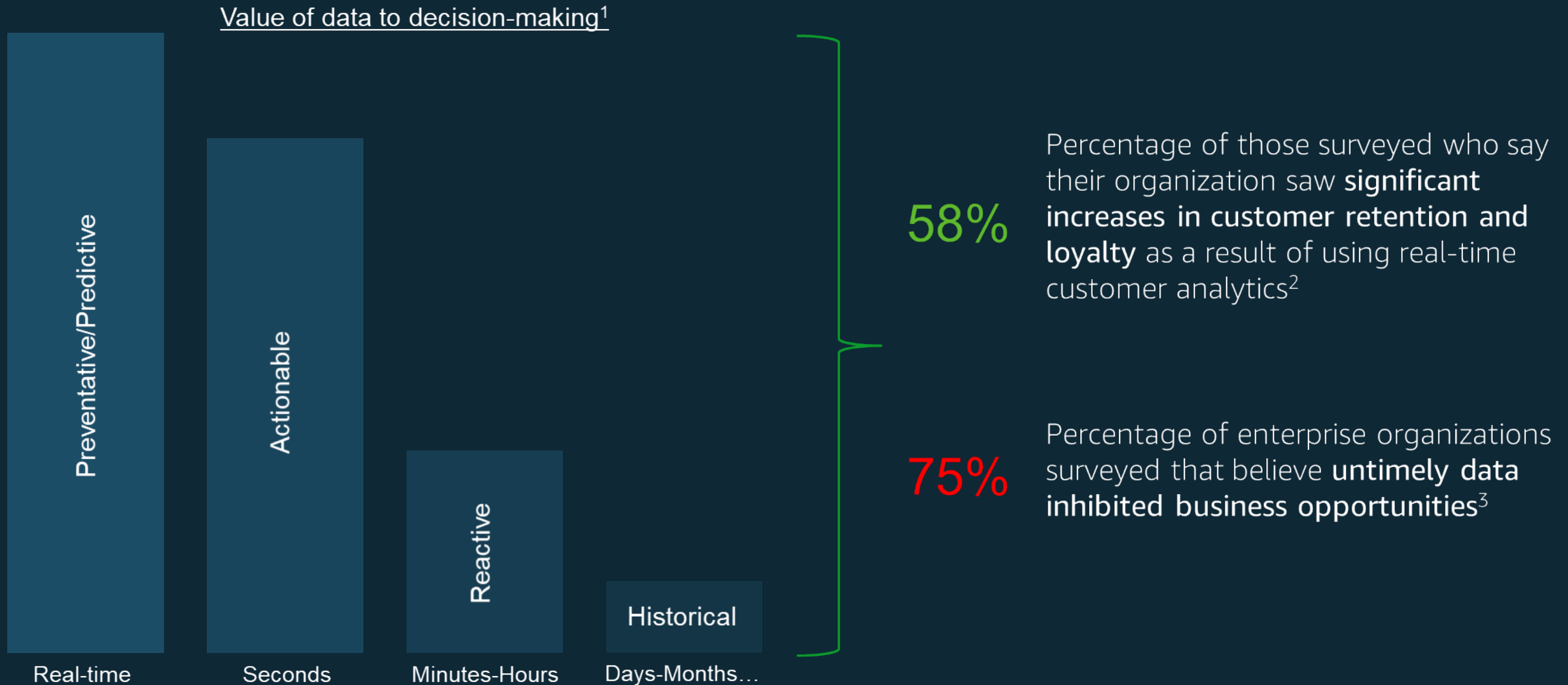




Real-time analytics with Kinesis

Failing to **act in real-time** can translate to real losses

Data have a short shelf life of actionability¹. AWS lets you act on that data **as fast as the market dictates**.



Common real-time analytics use cases



실시간 이상 징후 및 부정 행위 탐지



실시간으로 고객 경험 맞춤화



IoT 분석 강화



마케팅 캠페인의 자양분



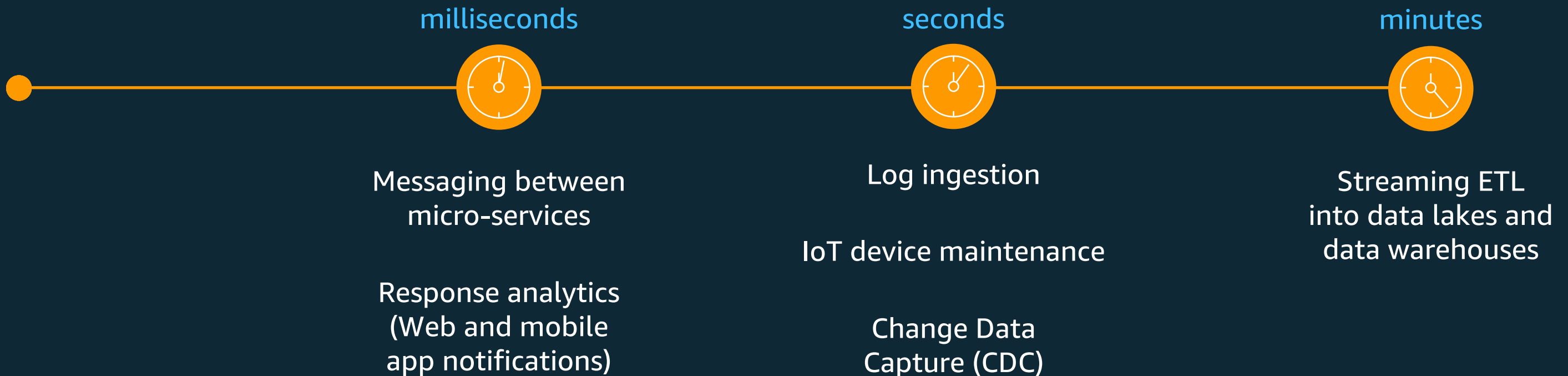
실시간 개인화



의료 및 응급 서비스 지원

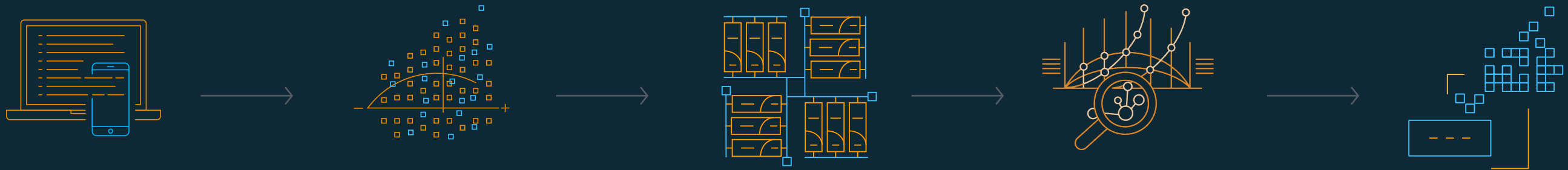


Achieving real-time analytics use cases



Enabling real-time analytics

다양한 소스에서 고속으로 생성되는 대량 데이터를 실시간으로 수집, 처리 및 분석



Source

Devices and or applications that produce real-time data at high velocity.

Stream ingestion

Data from tens of thousands of data sources can be collected and ingested in real time.

Stream storage

Data is stored in the order it was received for a set duration of time, and can be replayed indefinitely during this time.

Stream processing

Records are read in the order they are produced enabling real-time analytics or streaming ETL.

Destination

Data lake
Data Warehouse (most common)
Database (least common)



Challenges of Data Streaming



Difficult to setup



Tricky to scale



Hard to achieve high availability



Integration required development



Error prone and complex to manage



Expensive to maintain



Streaming real-time data with Amazon Kinesis



Easy to use



Elastic



Highly available & durable



Seamless AWS integrations



Fully managed



Pay for what you use



Streaming Data with AWS

실시간으로 데이터 스트림을 쉽게 수집, 처리 및 분석

Kinesis
Data **Streams**



분석을 위해 데이터
스트림 수집 및 저장

Kinesis
Data **Firehose**



데이터 스트림을 AWS
데이터 저장소로 로드

Kinesis
Data **Analytics**



SQL 또는 Apache Flink를 사용하여
데이터 스트림 분석

Managed
Streaming for
Apache Kafka



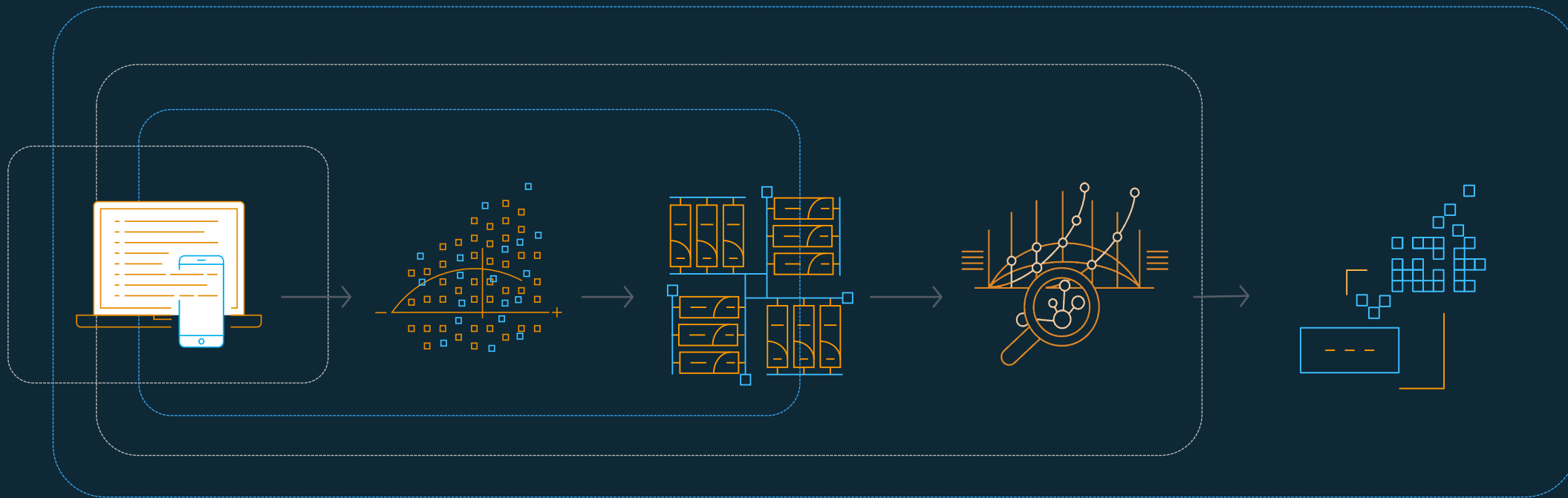
완전 관리형,
고가용성 및 보안





Deep Dive

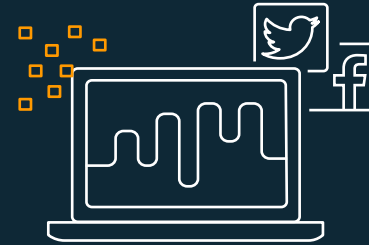
Data Streaming and Processing



Data Sources



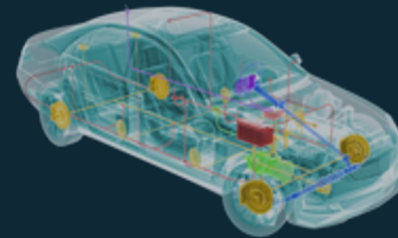
Mobile Apps



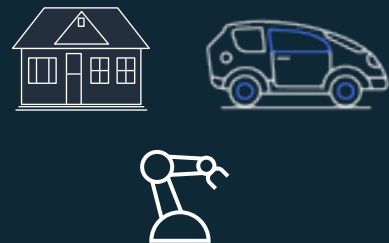
Web Clickstream/ Social



Application Logs



IoT Sensors



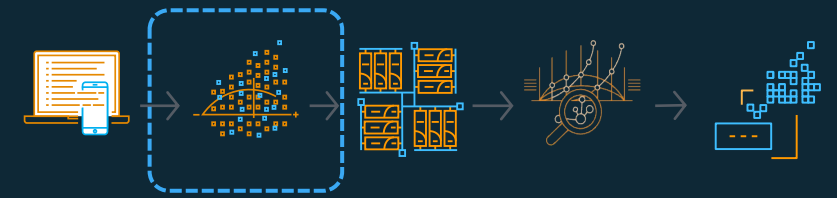
Connected Products



Smart Buildings



Stream Ingestion



수만 개의 데이터 소스에서 데이터를 실시간으로 수집하고 수집할 수 있습니다.

AWS Toolkits/Libraries

AWS SDK



Kinesis
Producer
Library



AWS Mobile
SDK



Kinesis Agent

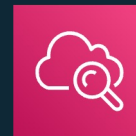


AWS Service Integrations

AWS IoT



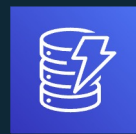
Amazon CloudWatch
Logs



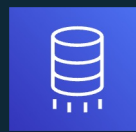
Amazon CloudWatch
Events



Amazon Dynamo DB



Amazon Database
Migration Service*



3rd Party Offerings

LOG4J



Flume



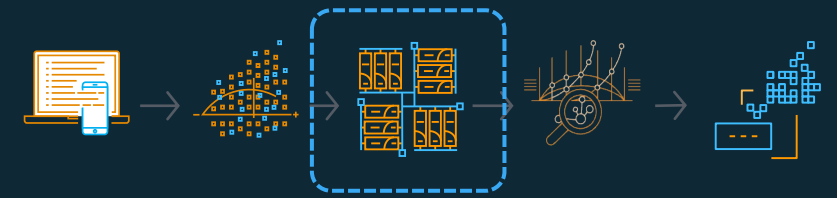
Fluentd



* Amazon DMS includes 8 on-premise databases, 1 Azure database, 5 RDS/Aurora database types, and S3



Stream Storage



데이터는 지정된 시간 동안 수신된 순서대로 저장되며, 이 시간 동안 무한 재생할 수 있습니다.

Kinesis Data Streams



Collect and store data streams for analytics

Default Retention ○ 24 hours

Extended Retention ○ Up to 7 days

New



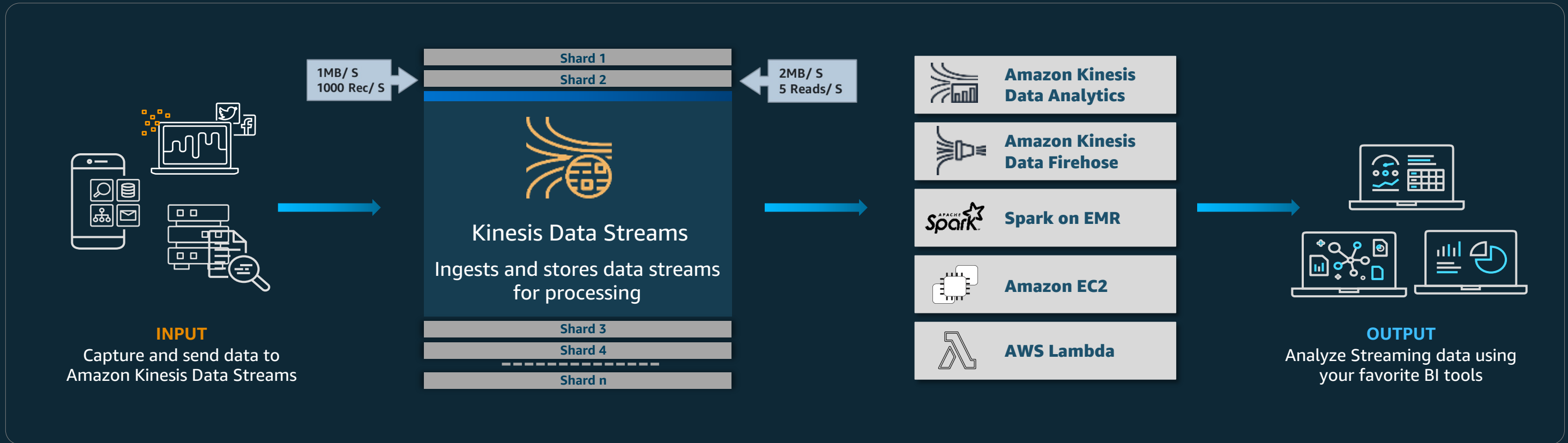
Long Term Retention ○ Up to 1 Year





Kinesis Data Streams

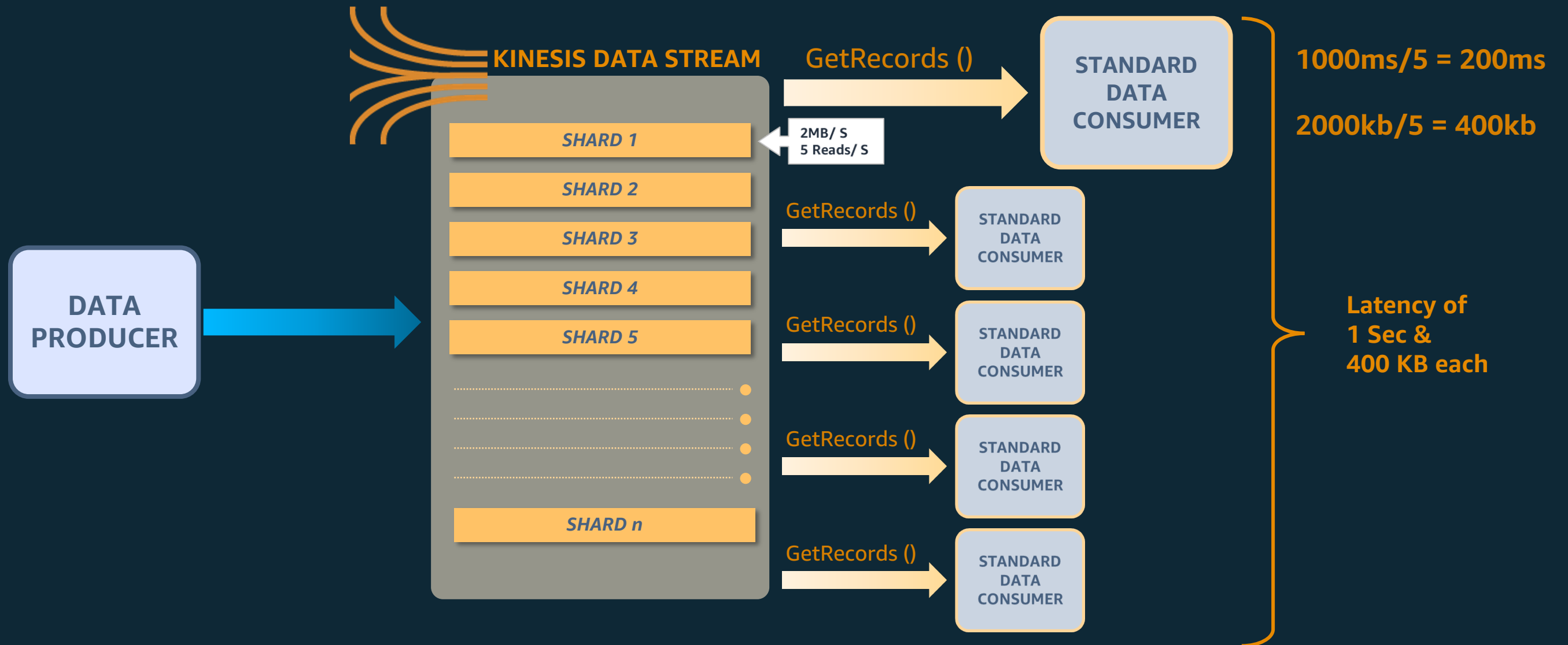
Amazon Kinesis Data Streams



- 간편한 관리 및 저렴한 비용
- 실시간 탄력적인 성능
- 안전하고 내구성이 뛰어난 스토리지
- 여러 실시간 분석 애플리케이션에서 사용 가능
- 하나의 표준 소비자에서 200ms의 평균 대기 시간
- 향상된 팬 아웃으로 일반적인 평균 대기 시간 70ms



Producer & Consumers



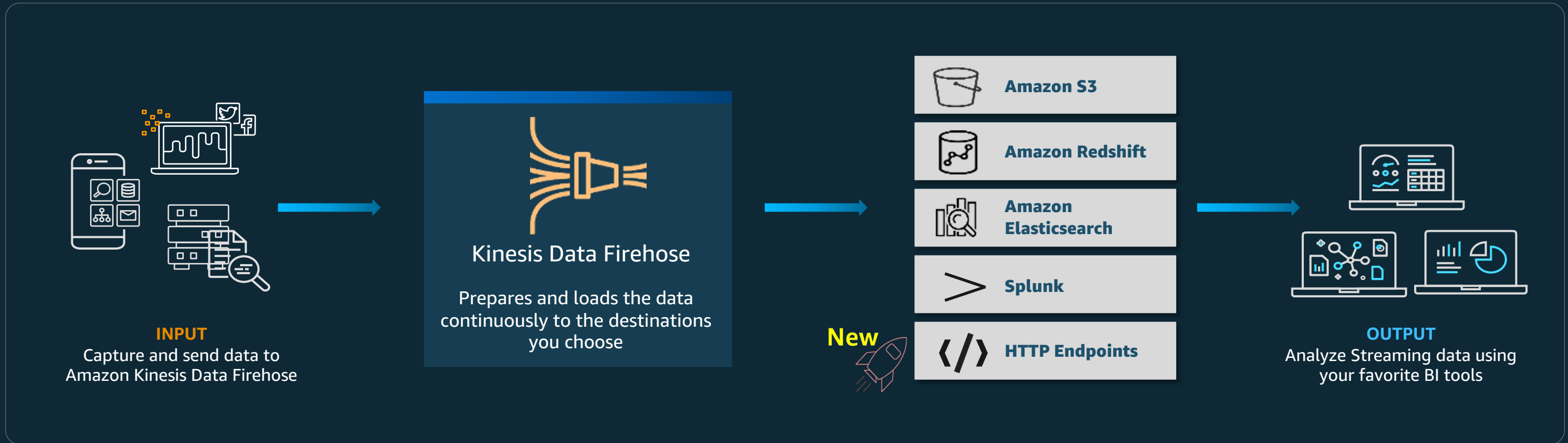
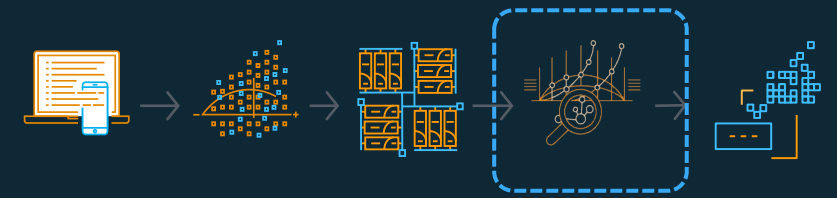
Consumers can get a lowest latency of 200ms with the standard consumer model





Kinesis Data Firehose

Amazon Kinesis Data Firehose



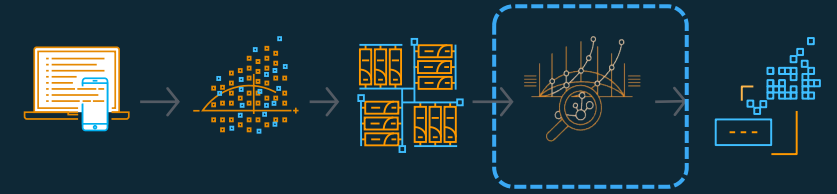
- 제로 관리 및 원활한 탄력성
- 데이터 저장소에 대한 직접 연결
- 서버리스 연속 데이터 변환
- 준 실시간
- 데이터 형식을 Parquet/ORC로 변환
- Datadog, Sumo Logic, New Real 및 MongoDB에 직접 데이터 전달






Stream Processing with Kinesis Data Analytics

Stream Processing




실시간 분석 또는 스트리밍 ETL을 지원하는 레코드들은 생성된 순서대로 읽혀집니다.

Kinesis



SQL/
Java



Amazon
Kinesis Data
Analytics



Kinesis Client Library
+
Connector Library

AWS Services



AWS Lambda

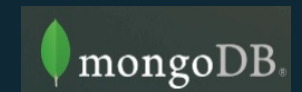


Amazon EMR

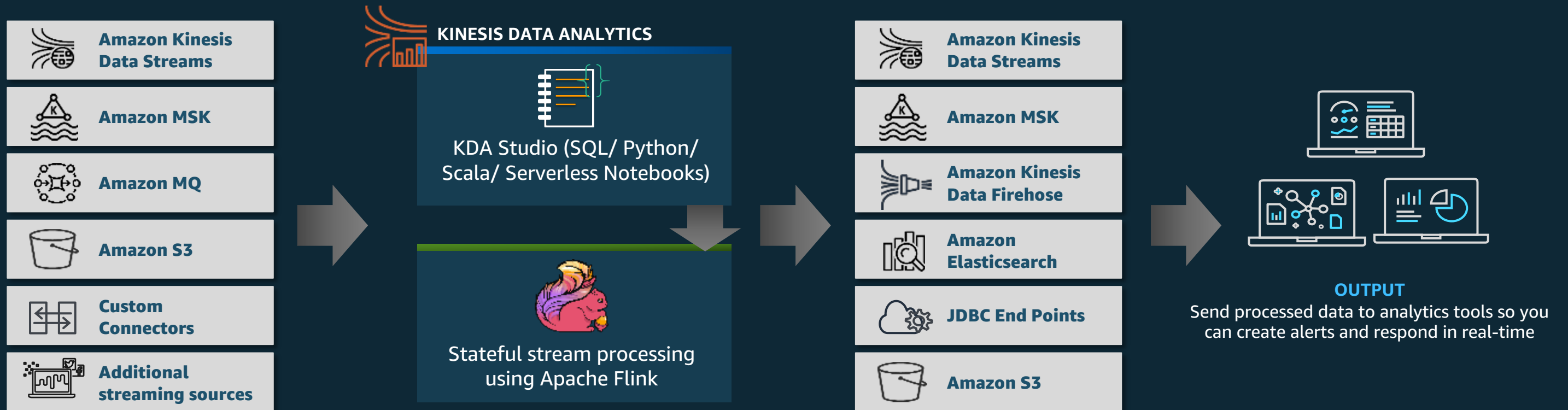
3rd party



Apache Spark



Amazon Kinesis Data Analytics



- SQL, Python, Scala 및 Java 또는 통합 Apache Flink 애플리케이션을 사용하여 실시간으로 스트리밍 데이터와 상호 작용
- Apache Flink용 KDA 내에서 KDA Studio adhoc 분석을 내구성 있는 상태 애플리케이션으로 배포
- 완벽하게 관리되고 탄력적인 스트림 처리 애플리케이션 구축



How do you **Process** Data Streams?



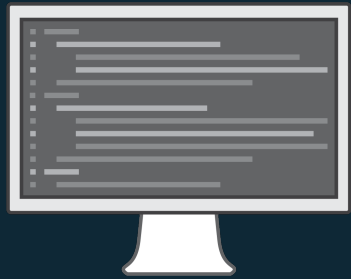
Apache **Flink**

- 실시간 데이터 스트림을 처리하기 위한 오픈 소스 프레임워크이자 분산 엔진
 - 상태 저장 계산 지원
 - 메모리 내 속도와 규모에 상관없이 수행
 - 제한되지 않은 데이터 스트림 및 제한된 데이터 스트림(배치) 처리



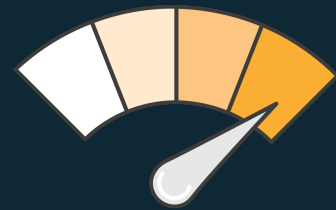
Apache Flink for sophisticated applications

데이터 스트림의 상태 저장 처리를 위한 프레임워크 및 분산 엔진인 Apache Flink 활용



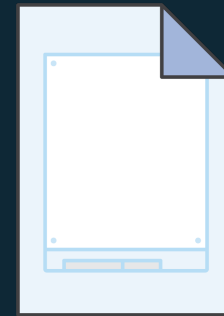
Simple programming

사용하기 쉽고 유연한
API를 통해
애플리케이션 구축 속도
향상



High performance

낮은 지연 시간과 높은
처리량을 제공하는
메모리 내 컴퓨팅



Stateful Processing

내구성이 뛰어난
애플리케이션 상태
유지



Strong data integrity

정확한 한 번의 처리와
일관성 있는 상태



Extensive integrations with **AWS** services

- 애플리케이션에 소스 및 싱크를 쉽게 추가
- 다른 데이터 소스 및 싱크를 위한 사용자 지정 커넥터 구축

Example Sources



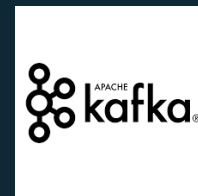
Amazon Kinesis
Data Streams



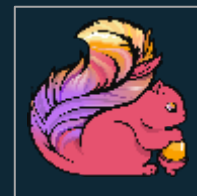
Amazon
Managed
Streaming for
Apache Kafka



RabbitMQ



Apache
Kafka

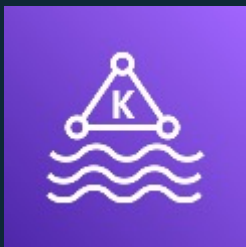


Flink
Connectors

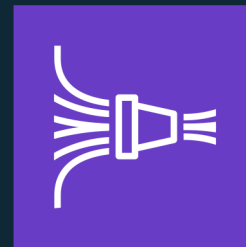
Example Destinations (Sinks)



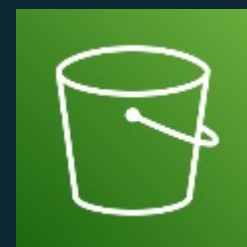
Amazon Kinesis
Data Streams



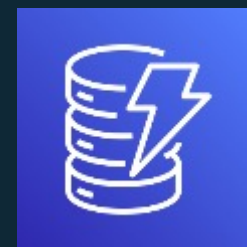
Amazon
Managed
Streaming for
Apache Kafka



Amazon Kinesis
Firehose



Amazon S3



Amazon
DynamoDB



RabbitMQ



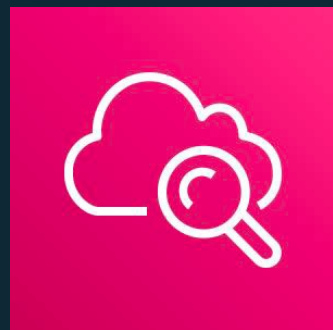
Apache
Cassandra



Elasticsearch



Monitoring **Kinesis** Data Analytics



CloudWatch Metrics

- KDA 애플리케이션에 대한 구성 가능한 CloudWatch 메트릭 및 로그
- 애플리케이션 상태를 알리는 CloudWatch 경보



Logging

- CloudWatch 로그를 사용하여 애플리케이션 성능 및 오류 상태 모니터링



Apache Flink Dashboard

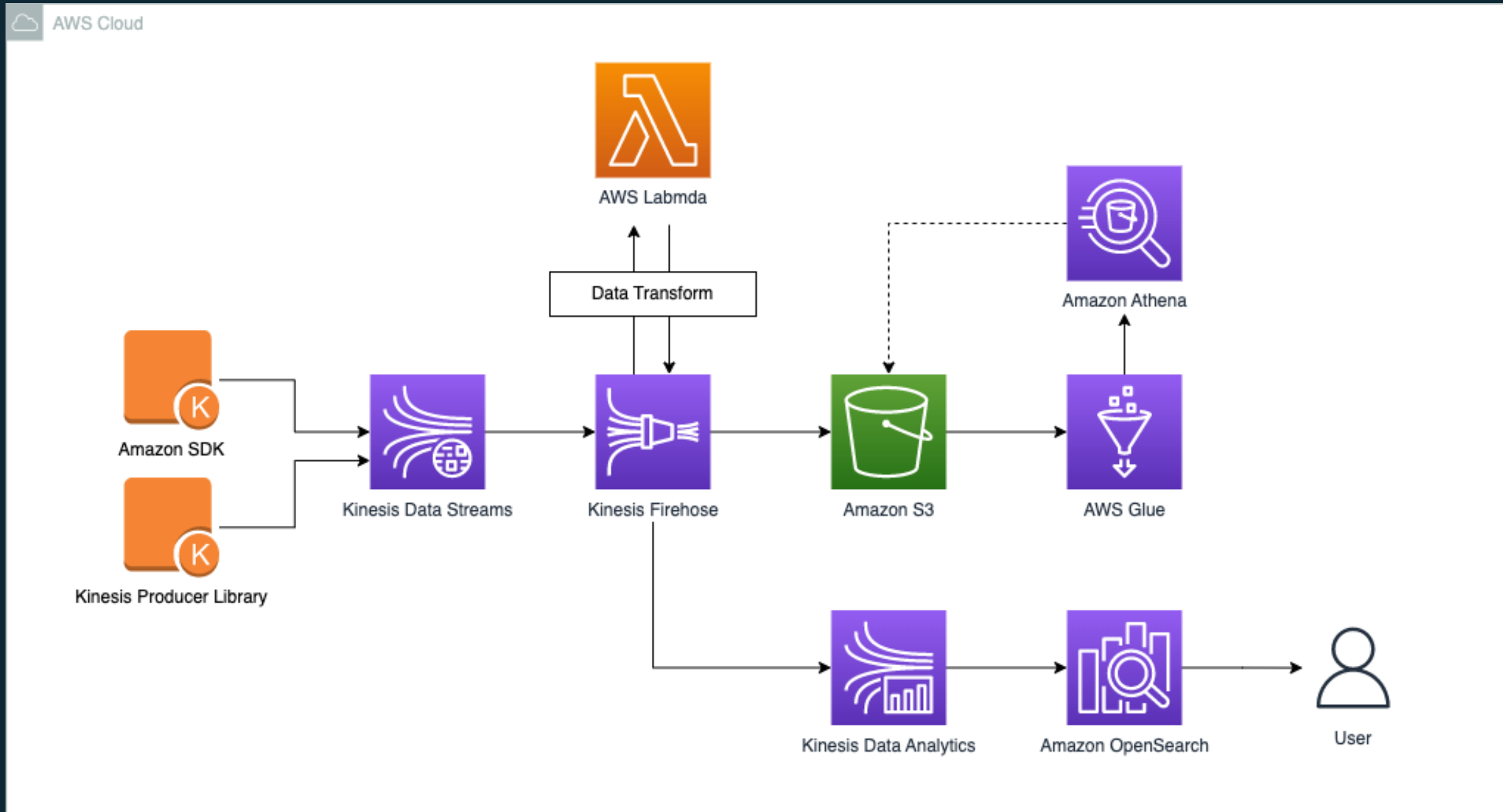
- 실행 중인 작업 모니터링에 Apache Flink 대시보드 사용





Hands on Lab

Lab Overview





Q&A