DEMO NOTES:

Slide 5 – What is Amazon Kinesis?

What is AWS Kinesis? AWS Kinesis – is core service for developing stream processing applications.

Why we need to ingest and process real time data? The area of BIG DATA is going to a massive change. The processing the vast of amount of data in real time is new norm and companies should processing incoming data in minutes or seconds just to stay competitive.

In a data stream processing application we have number of producer applications that are constantly writing data to data streaming service and we also have number of consumer that are constantly process that generated data in real time.

Here are some examples of producers applications that we can implement:
Kinesis can continuously capture gigabytes of data per second from hundreds of thousands of sources such as website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events.

What consumers applications can we have:

We can have Dashboards that display information about data being generated

Database – we can store generated events to database for further quiring

Or we can have other application that react for coming data.
Amazon Kinesis has made it easy to collect, process, and analyze streaming data in real-time so that you can get information on time and react to new information quickly.

Slide 6 – Kinesis Use Cases?

Here are some common use cases for Amazon Kinesis:

Log Data collection and processing: we can collect and process application logs

Messaging: we can send messaging from one application to another

Real-time metrics: we can track metrics generating by applications

Activity tracking: we can analyze users activities in real time and react to it

Slide 7 – Kinesis Features

All of these use cases are possible because of the following Kinesis features:

Number one Kinesis can provide

Real-time performance: which means there is low latency time between the moment when data written to AWS Kinesis stream and when the data can be read. − It allows to collect and analyze information in real-time like stock trade prices otherwise we need to wait for data-out report.

AWS Kinesis provide High Throughput: one of the example the companies can processing petabites of data every month

ELASTIC – which means it can scale up to process more data or it can scale down if demand is decreasing

AWS Integrations - also it's integrated with other AWS Services, so it's very convenient if you're already using AWS. It can be integrated with Amazon Redshift, Amazon S3 and Amazon DynamoDB.

LOW COST – AWS Kinesis has low cost, for example if you want to stream one megabyte of data every second, it will just cost you one cent per hour. Amazon Kinesis is cost-efficient for workloads of any scale. Pay as we go for the resources used and pay hourly for the throughput required.

EASY ADMINISTRATION – and the last but not least Kinesis has very low administration version, it serverless, which means we don't need to take care about servers, upgrades, versions etc. Using Amazon Kinesis, we can create a new stream, set its requirements, and start streaming data quickly.

Slide 8 – Batch Processing

Why we do need new tool and why we can't use old tool to solve new problems.

So what was before stream processing? Before stream processing we had approach batch processing?

Batch Processing: We have number of producers to generate data. We aggregate this data and stored it in some distributed storage, like S3 for example and we have consumers that could read and process this data. And the biggest problem with this approach to have very high latency. It can take an hour or even a day from where the data can be generated and later consume.

More importantly recent data have a more values than old data. As example, if you want to analyze what are users are doing or fraud detection.

Slide 8 – Old Architecture

Old Architecture – History of kinesis

There is another reason why you may want to use stream processing (kinesis). Let's take a look at this example of a retail website, so maybe have clickstream, some operational metrics and incoming orders. And have some micro services. Such as search, analytics and fraud detection. Click Stream can you used by Analytics and Fraud Detection and Analytics also interested in metrics and maybe order used by search and analytics as well. And you see if we will just implement point to point solutions to connect data sources to microservices. It's actually a very complicated architecture.

And if you want to add another data source, it can be just nightmare

Slide 10 – Stream Processing

and proposed solution is modern architecture is called Unified Log. And in this architecture you have multiple number of producers but all of them are writing data to the same system, which are AWS Kinesis.

All of the consumers in this system will read the data from this unified log

Slide 11 – Kinesis Family

Kinesis has multiple services under its name, like Data Streams, Firehose, Analytics, and Video Streams. We will only look at the Data Streams and Data Firehose services in this demo.

- Kinesis Streams – to build custom applications that process and analyze data.
- Kinesis Firehose – to easily load streaming data into AWS;
- Kinesis Analytics – to easily process and analyze streaming data with standard SQL;

Slide 12 – Kinesis Data Stream

Kinesis Data Stream for capture, process and store data streams

Amazon Kinesis data stream is a tool used for working with data in streams. It's collect and processes large streams of data records in real time. A Kinesis data stream is a set of shards. Each shard has a sequence of data records. Each data record has a sequence number that is assigned by Kinesis data stream. Kinesis work similar to other queuing and pub/subsystems.

A shard: A stream can be composed of one or more shards. One shard can read data at a rate of up to 2 MB/sec and can write up to 1,000 records/sec up to a max of 1 MB/sec. A user should specify the number of shards that coincides with the amount of data expected to be present in their system. Pricing of Kinesis streams is done on a per/shard basis

A producer puts data records into Amazon Kinesis data streams. For example, a web server sending log data to a Kinesis data stream is a producer. A consumer processes the data records from a stream.

A consumer, known as an Amazon Kinesis Data Streams application, is an application that you build to read and process data records from Kinesis data streams.

If you want to send stream records directly to services such as Amazon Simple Storage Service (Amazon S3), Amazon Redshift, Amazon Elasticsearch Service (Amazon ES), or Splunk, you can use a Kinesis Data Firehose delivery stream instead of creating a consumer application.

Slide 13 – Kinesis Data Firehose
Firehoses are the opposite of Data Streams, they distribute records/data to specified endpoints configured in the Firehose from a specified source. In a Firehose, we can also specify any transformations or formatting we'd like to apply to the data before it's pushed out. We will be using Firehose later on to connect our Data Stream to S3/Athena and provide the transformation to JSON logic.

Amazon Kinesis Data Firehose is the easiest way to reliably load streaming data into data lakes, data stores and analytics tools. It can capture, transform, and load streaming data into Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk, enabling near real-time analytics with existing business intelligence tools and dashboards you're already using today. It is a fully managed service that automatically scales to match the throughput of your data and requires no ongoing administration. It can also batch, compress, transform, and encrypt the data before loading it, minimizing the amount of storage used at the destination and increasing security.

Slide 14 – Kinesis Analytics With Amazon Kinesis Data Analytics for SQL Applications, you can process and analyze streaming data using standard SQL. The service enables you to quickly author and run powerful SQL code against streaming sources to perform time series analytics, feed real-time dashboards, and create real-time metrics.

To get started with Kinesis Data Analytics, you create a Kinesis data analytics application that continuously reads and processes streaming data. The service supports ingesting data from Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose streaming sources. Then, you author your SQL code using the interactive editor and test it with live streaming data. You can also configure destinations where you want Kinesis Data Analytics to send the results.

Kinesis Data Analytics supports Amazon Kinesis Data Firehose (Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk), AWS Lambda, and Amazon Kinesis Data Streams as destinations.

Slide 15 – Kinesis Approach for DR

First we consume data from CloudWatch for specific S3 events. Then from CloudWatch, the data goes to a Kinesis Data Stream. We have our data producers push data to an Amazon Kinesis data stream as soon as data is produced. Then the data from the stream is consumed by Kinesis Data Firehose delivery stream. Kinesis Data Firehose then invokes an AWS Lambda Function to decompressed/compressed the data, filter it and then FireHose send filtered data to S3/Athena

Slide 16 - Demo

To get the metrics about the stream we are going to monitoring tab and you will see a lot of graphs. We just go through most important one. The first one – (Get Records – sum Bytes) – how many bytes Ger Records return.
The second one Get Records Iterator Age – this is basically how far away you are from the beginning of the stream and it's very important to see if your consumers keeping up with the producers.
The next one Get Record Latency.

The second group of metrics are all related to writing data to Kinesis. The first one amount of data that we wrote in bytes. Also we have a number of incoming records.