Unify Batch and Stream Processing with Apache Beam on AWS

Gandhinath Swaminathan
Senior Solutions Architect, Amazon Web Services
@gnathaws
Run Beam Pipelines with Amazon Kinesis Data Analytics

Kinesis Data Analytics applications uses **Apache Flink Runner** to execute the Beam pipelines and supports the same Apache Beam capabilities as the Apache Flink runner. For information about what features are supported with the Apache Flink runner, see the [Beam Compatibility Matrix](#).
Amazon Kinesis Data Analytics is a Fully Managed Apache Flink Service

More focus on creating Streaming Applications than managing infrastructure
**Amazon Kinesis Data Analytics**

is a managed Apache Flink service that enables building of sophisticated streaming applications

- No servers to manage
- Only pay for the resources you use in your application
- Supports sub-second latencies
- Auto-scales to application demands based on your configuration
Workshop scenario
Improve the operations of a taxi company in New York City
Architecture to Unify Batch and Streaming
Details and Best Practices
Pass Parameters to your Beam Pipeline

Map<String, Properties> propertyGroups = KinesisAnalyticsRuntime.getRuntime().getApplicationProperties();

Properties beamProperties = propertyGroups.get("BeamApplicationProperties");

String kinesisStreamName = beamProperties.getProperty("InputStreamName");
Leverage Temporary Credentials

KinesisIO.read()
    .withStreamName(...)
    .withAWSClientsProvider(new DefaultCredentialsProviderClientsProvider(...))

```java
public class DefaultCredentialsProviderClientsProvider implements AWSClientsProvider {

    @Override
    public AmazonKinesis getKinesisClient() {
        return AmazonKinesisClientBuilder.standard().withRegion(…).build();
    }

    ...
}
```
Obtaining Equivalent Semantics for Batch and Streaming

Use `ApproximateArrivalTime` from Kinesis as event time.

Add `ApproximateArrivalTime` to the payload of the event.

Read `ApproximateArrivalTime` from the event payload and use it as event time.
Further Readings

The AWS Big Data blog is a great resource to learn more about Apache Flink and stream processing on AWS in general.

- https://github.com/aws-samples/amazon-kinesis-analytics-beam-taxi-consumer
Let’s go build!
Thank you!

Gandhinath Swaminathan
Senior Solutions Architect, Amazon Web Services
@gnathaws