BLUESOFT
EFFICIENT SOLUTIONS

BIGDATA
ON AWS

IT'S ALL ABOUT EFFICIENT SOLUTIONS

BLUESOFT
EFFICIENT SOLUTIONS

Amazon Kinesis Firehose

DESCRIPTION

Capture, transform and load streaming data into Amazon Kinesis Analytics, Amazon S3, Amazon Redshift and Amazon Elasticsearch.

USAGE PATTERNS (EXAMPLES)

- Real-time data analytics
- Log and data feed intake and processing
- Real-time metrics and reporting
- IoT: Getting insights from telemetry data
- Real-time advertising solutions
- Optimize digital marketing

COST MODEL

Pricing is based on volume of data ingested into Amazon Kinesis Firehose, which is calculated as the number of data records you send to the service, times the size of each record rounded up to the nearest 5KB.

INPUT FROM

- Kinesis Agent
- Kinesis Streams
- Amazon IoT
- Amazon CloudWatch Logs and Events
- PutRecord and
- PutRecordBatch operations

OUTPUT TO

- Amazon Kinesis Analytics
- Amazon S3
- Amazon Redshift
- Amazon Elasticsearch

Amazon Kinesis Streams

DESCRIPTION

Enables to build custom application that process or analyze streaming data. Kinesis Streams can continuously capture and store terabytes of data per hour from hundreds of thousands of sources. Data can also be emitted to other AWS services.

COST MODEL

Pricing is based on Shard Hour and PUT Payload Unit. Shard is the base throughput unit of an Amazon Kinesis stream. One shard provides a capacity of 1MB/sec data input and 2MB/sec data output. One shard can support up to 1000 records per second and is charged at an hourly rate per shard. A record is the data that your data producer adds to your Amazon Kinesis stream. A PUT Payload Unit is counted in 25KB payload “chunks” that comprise a record. PUT Payload Unit is charged with a per million PUT Payload Units rate.

INPUT FROM

- Amazon Kinesis Producer Library (KPL)
- Amazon Kinesis Agent
- PutRecord and PutRecords operations

OUTPUT TO

- Amazon Kinesis API
- Amazon Kinesis Client Library (KCL)

BLUESOFT
EFFICIENT SOLUTIONS

Amazon Elasticsearch

DESCRIPTION

Managed service that makes it easy to deploy, operate and scale Elasticsearch in AWS. Elasticsearch is a real-time distributed search and analytics engine that is used for full-text search, structured search, analytics, and all three in combination.

USAGE PATTERNS (EXAMPLES)

- Log analytics
- Full text search
- Distributed document store
- Real-time application monitoring
- Clickstream analytics

COST MODEL

Pricing is based on instance hours, EBS storage and standard data transfer fees.

INPUT FROM

- Amazon Kinesis Firehose
- Amazon Kinesis Streams
- Amazon S3
- Amazon DynamoDB
- Logstash
- Native API

OUTPUT TO

- Kibana
- Logstash

Amazon CloudSearch

DESCRIPTION

AWS managed service that makes it simple and cost-effective to set up, manage and scale a search solution for websites or applications.

COST MODEL

Pricing is based on instance hours, batch uploads, IndexDocuments requests and data transfer

INPUT FROM

- Batch pushes
- Amazon S3

OUTPUT TO

- JSON
- XML

BLUESOFT
EFFICIENT SOLUTIONS

Amazon EMR

DESCRIPTION

Amazon EMR provides a managed Hadoop, Apache Spark, HBase, Presto, and Flink frameworks.

USAGE PATTERNS (EXAMPLES)

- Log processing and analytics
- Large extract, transform, and load (ETL) data movement
- Risk modeling and threat analytics
- Ad targeting and click stream analytics
- Genomics
- Predictive analytics
- Ad hoc data mining and analytics

COST MODEL

"Pay per-second rate for every second you use, with a one-minute minimum. Price depends on the instance type used. The Amazon EMR price is in addition to the Amazon EC2 price (the price for the underlying servers) and Amazon EBS price (if attaching Amazon EBS volumes). These are also billed per-second, with a one-minute minimum."

INPUT FROM

- Amazon S3
- Amazon DynamoDB
- AWS Direct Connect
- AWS Import/Export

OUTPUT TO

- Amazon S3

Amazon Athena

DESCRIPTION

Serverless, interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.

USAGE PATTERNS (EXAMPLES)

- Query and visualize data

COST MODEL

Pricing is based on the amount of data scanned by each query. Price is per TB of data scanned.

INPUT FROM

- Amazon S3

OUTPUT TO

- Amazon S3
- Amazon Quicksight

BLUESOFT
EFFICIENT SOLUTIONS

Amazon Redshift

DESCRIPTION

Fast and fully managed data warehouse. Allows to run complex analytic SQL queries against petabytes of structured data. With Redshift Spectrum direct SQL queries can be run against unstructured data in S3.

USAGE PATTERNS (EXAMPLES)

- Analyze global sales data for multiple products
- Store historical data
- Aggregate data
- Analyze trends
- Measure quality, operation efficiency and financial performance

COST MODEL

On-demand and reserved instances are charged with an hourly rate based on the type and number of nodes in the cluster. Redshift Spectrum for the number of bytes scanned.

INPUT FROM

- Amazon S3
- Amazon DynamoDB
- Remote hosts like EC2 or EMR over SSH
- AWS Data Pipeline

OUTPUT TO

- Amazon Quicksight

Amazon Quicksight

DESCRIPTION

Business analytics service that makes it easy to build visualizations, perform ad-hoc analysis, and quickly get business insights from your data.

USAGE PATTERNS (EXAMPLES)

- Deliver affordable business intelligence to the organization

COST MODEL

Charged per user per month. First user per month is free with 1GB SPICE included. Additional user are paid and include 10GB of SPICE. Additional SPICE is charged per GB per month.

INPUT FROM

- Amazon Athena
- Amazon Aurora
- Amazon Redshift
- Amazon Redshift Spectrum
- Amazon S3
- Amazon S3 Analytics
- Apache Spark 2.0 or later
- MariaDB 10.0 or later
- Microsoft SQL Server 2012 or later
- MySQL 5.1 or later
- PostgreSQL 9.3.1 or later
- Presto 0.167 or later
- Snowflake
- Teradata 14.0 or later
- Salesforce

AWS Glue

DESCRIPTION

Fully managed extract, transform, and load (ETL) service that allows to prepare and load data for analytics.

USAGE PATTERNS (EXAMPLES)

- Orchestration of extract, transform, and load (ETL) jobs to build a data warehouse

COST MODEL

An hourly rate, billed by the second, for crawlers (discovering data) and ETL jobs (processing and loading data).
Monthly fee for storing and accessing the metadata for the AWS Glue Data Catalog. Development endpoint to interactively develop ETL code is paid an hourly rate, billed per second.

INPUT FROM

- Amazon S3
- Amazon RDS
- Amazon Redshift
- JDBC

OUTPUT TO

- Amazon S3
- Amazon RDS
- Amazon Redshift
- JDBC

AWS Data Pipeline

DESCRIPTION

Helps to reliably process and move data between different AWS compute and storage services, as well as on-premise data sources, at specified intervals.

USAGE PATTERNS (EXAMPLES)

- Export DynamoDB table to S3
- Import DynamoDB table from S3
- Full copy of RDS MySQL to S3
- Load data from S3 into Redshift
- Run job on EMR cluster

COST MODEL

Billed based on how often the activities and preconditions are scheduled to run and where they run (AWS or on-premises).

INPUT FROM

- Amazon S3
- Amazon RDS

OUTPUT TO

- Amazon S3
- Amazon Redshift
- Amazon RDS
- Amazon DynamoDB
- Amazon EMR