***Design document for storing the predications in AWS***

1. **Assumptions:**

* Generating the predications we will be using AWS Sage Maker for model processing.
* Users: Data Scientist, Administrators
* Buckets, IAM policies will be created during the initialization setup for model setup in AWS Sage-Maker
* Use of AWS SNS service for notification
* Use of Lambda services for triggering the auto learning process for the model
* Scaling in terms of resources – EC2 instances, PySpark etc will be pre determined for model
* Below Limitations will be kept in mind for storing the predictions:

| **Limit Type** | **System Limit** |
| --- | --- |
| Size of batch prediction input | 1 TB |
| Size of batch prediction input (number of records) | 100 million |
| Number of variables in a data file (schema) | 1,000 |
| Recipe complexity (number of processed output variables) | 10,000 |
| TPS for each real-time prediction endpoint | 200 |
| Total TPS for all real-time prediction endpoints | 10,000 |

1. **System Architecture for Storing Model Predictions:**

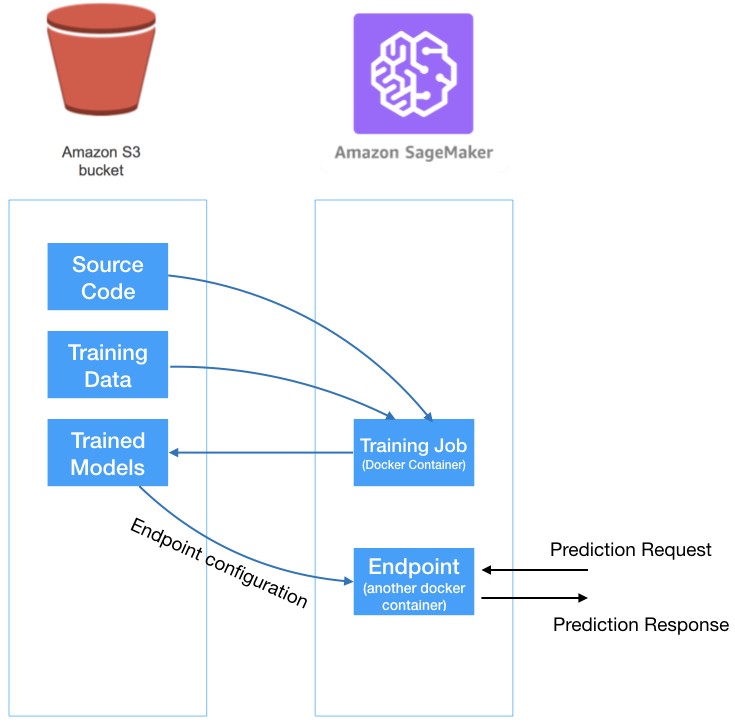
Model predictions can be stored on basis of requirements and usage:

|  |  |
| --- | --- |
| Structure Data | Amazon Redshift |
| Semi Structured Data | Cassandra, Mongo DB |
| Unstructured Data | AWS S3 |

In current scenario we are going to store the predictions from model to AWS S3:

Benefits of using AWS S3 for storing the model predictions:

* Scalable in terms of storage
* Security – Supports SSL
* Generates batch predictions by using your ML models, to output the prediction file to an S3 bucket that you specify.
* Easy to integrate with IAM, alerting services such as Event Notifications, databases including Redshift and EMR, and computing platforms such as Lambda.



To output the results of the batch prediction operation to Amazon S3, you must grant Amazon ML the following permissions to the output location, which is provided as input to the Create Batch Prediction operation:

* GetObject permission on your S3 bucket and prefix.
* PutObject permission on your S3 bucket and prefix.
* PutObjectAcl on your S3 bucket and prefix.

Amazon ML needs this permission to ensure that it can grant the canned [ACL](https://docs.aws.amazon.com/AmazonS3/latest/dev/acl-overview.html#canned-acl) bucket-owner-full-control permission to your AWS account, after objects are created.

ListBucket permission on the S3 bucket. Unlike other actions, ListBucket must be granted bucket-wide permissions (rather than on the prefix). You can, however, scope the permission to a specific prefix by using a **Condition** clause.(As described in IAM Policy)

Predictions are generally two types:

1. Batch Predictions

Batch predictions are for a set of observations that can run all at once. This is ideal for predictive analyses that do not have a real-time requirement. The following table defines terms that are related to batch predictions.

| Term | Definition |
| --- | --- |
| Output Location | The results of a batch prediction are stored in an S3 bucket output location. For e.g. s3://examplebucket/output/batch-prediction/bp-example.manifest |
| Manifest File | This file relates each input data file with its associated batch prediction results. It is stored in the S3 bucket output location. For e.g.: s3://examplebucket/output/batch-prediction/bp-example.manifest |

Reviewing Batch Prediction Metrics and Details (API)

You can use the Amazon ML APIs to retrieve details about BatchPrediction objects, including the record metrics. Amazon ML provides the following batch prediction API calls:

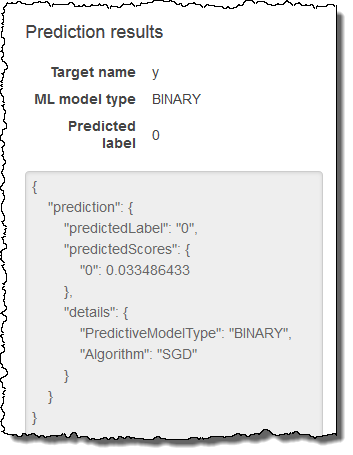
* CreateBatchPrediction
* UpdateBatchPrediction
* DeleteBatchPrediction
* GetBatchPrediction
* DescribeBatchPredictions

1. Real-time Predictions

Real-time predictions are for applications with a low latency requirement, such as interactive web, mobile, or desktop applications. Any ML model can be queried for predictions by using the low latency real-time prediction API.

The following table defines terms that are related to real-time predictions.

| Term | Definition |
| --- | --- |
| Real-time Prediction API | The Real-time Prediction API accepts a single input observation in the request payload and returns the prediction in the response. For e.g.  {  "MLModelId": "model-id",  "Record":{  "key1": "value1",  "key2": "value2"  },  "PredictEndpoint": "https://endpointUrl"  } |
| Real-time Prediction Endpoint | To use an ML model with the real-time prediction API, you need to create a real-time prediction endpoint. Once created, the endpoint contains the URL that you can use to request real-time predictions.  For e.g: "EndpointInfo":{  "CreatedAt": 1427864874.227,  "EndpointStatus": "READY",  "EndpointUrl": "https://endpointUrl",  "PeakRequestsPerSecond": 200  } |



An evaluation measures the quality of your ML model and determines if it is performing well.

Model predictions are based on the following factors and will be stored with S3:

| Term | Definition |
| --- | --- |
| Model Insights | Amazon ML provides you with a metric and a number of insights that you can use to evaluate the predictive performance of your model. |
| AUC | Area Under the ROC Curve (AUC) measures the ability of a binary ML model to predict a higher score for positive examples as compared to negative examples. |
| Macro-averaged F1-score | The macro-averaged F1-score is used to evaluate the predictive performance of multiclass ML models. |
| RMSE | The Root Mean Square Error (RMSE) is a metric used to evaluate the predictive performance of regression ML models. |
| Cut-off | ML models work by generating numeric prediction scores. By applying a cut-off value, the system converts these scores into 0 and 1 labels. |
| Accuracy | Accuracy measures the percentage of correct predictions. |
| Precision | Precision shows the percentage of actual positive instances (as opposed to false positives) among those instances that have been retrieved (those predicted to be positive). In other words, how many selected items are positive? |
| Recall | Recall shows the percentage of actual positives among the total number of relevant instances (actual positives). In other words, how many positive items are selected? |

1. **IAM Access Policy’s:**

Create the ARNs for the ML resources:

Resource": arn:aws:machinelearning:region:account:resource-type/identifier

Data source ID: my-s3-datasource-id

"Resource": arn:aws:machinelearning:<region>:<your-account-id>:datasource/my-s3-datasource-id

S3 Bucket: s3://predictions\_<model\_id>/

Grant Amazon ML Permissions to Read Data from AWS S3:

|  |
| --- |
| {  "Version": "2008-10-17",  "Statement": [  {  "Effect": "Allow",  "Principal": {  "Service": "machinelearning.amazonaws.com"  },  "Action": "s3:GetObject",  "Resource": "arn:aws:s3:::examplebucket/exampleprefix/\*"  },  {  "Effect": "Allow",  "Principal": {  "Service": "machinelearning.amazonaws.com"  },  "Action": "s3:ListBucket",  "Resource": "arn:aws:s3:::examplebucket",  "Condition": {  "StringLike": {  "s3:prefix": "exampleprefix/\*"  }  }  }  ]  } |

Granting Amazon ML Permissions to Output Predictions to Amazon S3:

|  |
| --- |
| {  "Version": "2008-10-17",  "Statement": [  {  "Effect": "Allow",  "Principal": {  "Service": "machinelearning.amazonaws.com"  },  "Action": [  "s3:GetObject",  "s3:PutObject"  ],  "Resource": "arn:aws:s3:::examplebucket/exampleprefix/\*"  },  {  "Effect": "Allow",  "Principal": {  "Service": "machinelearning.amazonaws.com"  },  "Action": "s3:PutObjectAcl",  "Resource": "arn:aws:s3:::examplebucket/exampleprefix/\*",  "Condition": {  "StringEquals": {  "s3:x-amz-acl": "bucket-owner-full-control"  }  }  },  {  "Effect": "Allow",  "Principal": {  "Service": "machinelearning.amazonaws.com"  },  "Action": "s3:ListBucket",  "Resource": "arn:aws:s3:::examplebucket",  "Condition": {  "StringLike": {  "s3:prefix": "exampleprefix/\*"  }  }  }  ]  } |

|  |
| --- |
| Batch prediction ID: my-batchprediction-id  "Resource":arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/my-batchprediction-id  Evaluation ID: my-evaluation-id  "Resource": arn:aws:machinelearning:<region>:<your-account-id>:evaluation/my-evaluation-id |

Allow users to read machine learning resources metadata

The following policy allows a user or group read the metadata of datasources, ML models, batch predictions, and evaluations by performing [DescribeDataSources](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_DescribeDataSources.html), [DescribeMLModels](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_DescribeMLModels.html), [DescribeBatchPredictions](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_DescribeBatchPredictions.html), [DescribeEvaluations](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_DescribeEvaluations.html), [GetDataSource](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_GetDataSource.html), [GetMLModel](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_GetMLModel.html), [GetBatchPrediction](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_GetBatchPrediction.html), and [GetEvaluation](https://docs.aws.amazon.com/machine-learning/latest/APIReference/API_GetEvaluation.html) actions on the specified resource(s). The Describe \* operations permissions can't be restricted to a particular resource.

|  |
| --- |
| {  "Version": "2012-10-17",  "Statement": [  {  "Effect": "Allow",  "Action": [  "machinelearning:Get\*"  ],  "Resource": [  "arn:aws:machinelearning:<region>:<your-account-id>:datasource/S3-DS-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:datasource/REDSHIFT-DS-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/BP-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:evaluation/EV-ID1"  ]  },  {  "Effect": "Allow",  "Action": [  "machinelearning:Describe\*"  ],  "Resource": [  "\*"  ]  }  ]  } |

Allow users to create machine learning resources

The following policy allows a user or group to create machine learning datasources, ML models, batch predictions, and evaluations by performing CreateDataSourceFromS3, CreateDataSourceFromRedshift, CreateDataSourceFromRDS, CreateMLModel, CreateBatchPrediction, and CreateEvaluation actions. You can't restrict the permissions for these actions to a specific resource.

|  |
| --- |
| {  "Version": "2012-10-17",  "Statement": [  {  "Effect": "Allow",  "Action": [  "machinelearning:CreateBatchPrediction",  "machinelearning:CreateEvaluation"  ],  "Resource": [  "\*"  ]  }  ]  } |

Allow users to create and delete real-time endpoints and perform real-time predictions on an ML model

The following policy allows users or groups to create and delete real-time endpoints and perform real-time predictions for a specific ML model by performing CreateRealtimeEndpoint, DeleteRealtimeEndpoint, and Predict actions on that model.

|  |
| --- |
| {  "Version": "2012-10-17",  "Statement": [  {  "Effect": "Allow",  "Action": [  "machinelearning:CreateRealtimeEndpoint",  "machinelearning:DeleteRealtimeEndpoint",  "machinelearning:Predict"  ],  "Resource": [  "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL"  ]  }  ]  } |

Allow users to update and delete specific resources

The following policy allows a user or group to update and delete specific resources in your AWS account by giving them permission to perform  UpdateBatchPrediction, UpdateEvaluation, DeleteBatchPrediction, and DeleteEvaluation actions on those resources in your account.

|  |
| --- |
| {  "Version": "2012-10-17",  "Statement": [{  "Effect": "Allow",  "Action": [  "machinelearning:Update\*",  "machinelearning:DeleteBatchPrediction",  "machinelearning:DeleteEvaluation"  ],  "Resource": [  "arn:aws:machinelearning:<region>:<your-account-id>:datasource/S3-DS-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:datasource/REDSHIFT-DS-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:mlmodel/ML-MODEL-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:batchprediction/BP-ID1",  "arn:aws:machinelearning:<region>:<your-account-id>:evaluation/EV-ID1"  ]  }]  } |

Allow any Amazon MLaction for Adminstrators only

The following policy allows a user or group to use any Amazon ML action. Because this policy grants full access to all of your machine learning resources, restrict it to administrators only.

|  |
| --- |
| {  "Version": "2012-10-17",  "Statement": [{  "Effect": "Allow",  "Action": [  "machinelearning:\*"  ],  "Resource": [  "\*"  ]  }] |

1. **Monitor Prediction Metrics:**

To monitor the to keep track of batch and real-time predictions, we can monitor the PredictCount metric according to the Request Mode dimension. The metrics are automatically collected and sent to Amazon CloudWatch every five minutes. We can monitor these metrics by using the Amazon CloudWatch console, AWS CLI, or AWS SDKs.

Cost effective as there is no charge for the Amazon ML metrics that are reported through CloudWatch. If you set alarms on the metrics, you will be billed at standard

1. **GetBatchPredictions:**

To get the prediction of the evaluation or batch prediction model we can use the following Amazon API:

* GetEvaluation
* GetBatchPrediction

Each operation takes exactly two parameters: the object ID and a Boolean flag called Verbose. Calls with Verbose set to true will include extra details about the object, resulting in higher latencies and larger responses

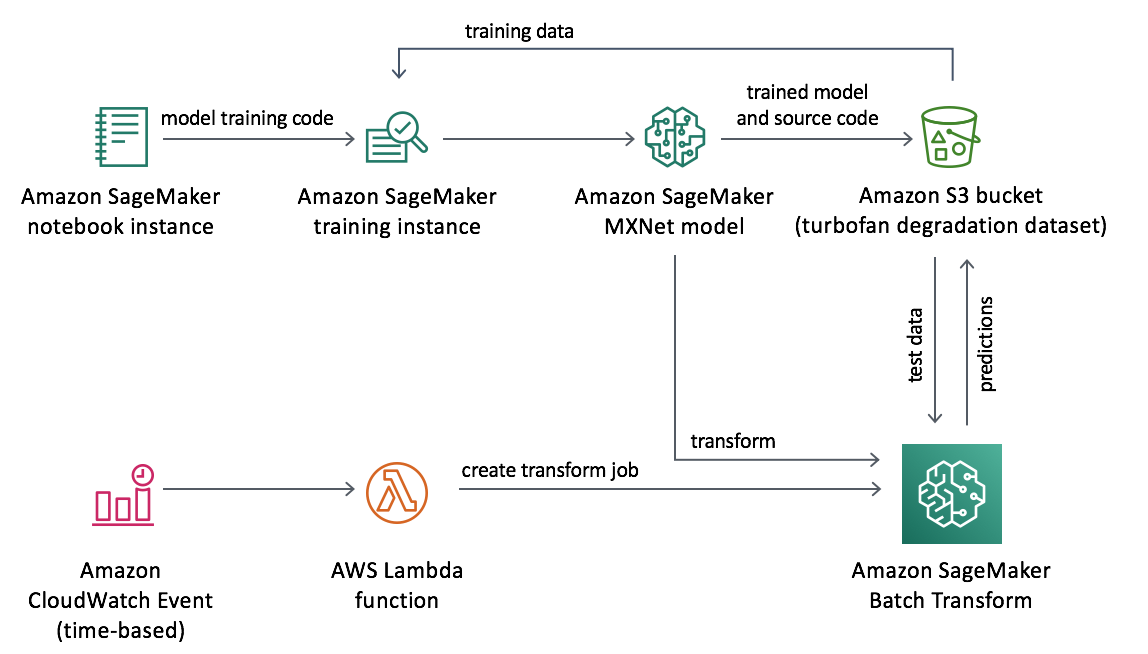
1. **Update the predications:**

For updating the Prediction, we can use the AmazonAPI:

* UpdateEvaluation
* UpdateBatchPrediction

Each operation requires the object's ID to specify which object is being updated. We can update the Predictions and store back in to AWS S3. Others properties of objects for data sources, evaluations, and batch predictions can’t be updated hence it maintains the data integrity. (more [Amazon ML API Reference](https://docs.aws.amazon.com/machine-learning/latest/APIReference/))

1. **Predictive Maintenance Using ML:**



[1]

We can use [AWS CloudFormation](https://aws.amazon.com/cloudformation/) template that to store the predictions in the [Amazon Simple Storage Service](https://aws.amazon.com/s3/) (Amazon S3) bucket and using [Amazon SageMaker](https://aws.amazon.com/sagemaker/) endpoint with an ML model that will be trained on the dataset to predict remaining useful life (RUL).

The solution uses a SageMaker notebook instance to orchestrate the model, and a SageMaker training instance to perform the training. The training code and trained model are stored in the solution's Amazon S3 bucket.

The solution also deploys an [Amazon CloudWatch Events](https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/WhatIsCloudWatchEvents.html) rule that is configured to run once per day. The rule is configured to trigger an [AWS Lambda](https://aws.amazon.com/lambda/) function that creates an Amazon SageMaker batch transform job that uses the trained model to predict RUL from the example dataset.

By default, the solution is configured to predict RUL from the example dataset.

1. **Delete the Predictions**

Solution 1: Setup the AWS S3 bucket policy to automatically delete on basis of the required model prediction factors or age(in number of days or months or yrs)

Solution 2: Delete predictions using the Amazon ML Objects API calls:

* DeleteEvaluation - Takes the parameter EvaluationId.
* DeleteBatchPrediction - Takes the parameter BatchPredictionId.

1. **Storage/Backups:**

Once the predictions are old >10 yrs. we can store them into the semi-structured databases: Cassandra & Mongo Db for the faster retrieval of the values. As the Amazon Redshift will have data access issues for the large data.

1. **Auditing\Logging API Calls with AWS CloudTrails:**

AWS CloudTrail, a service that provides a record of actions taken by a user, role, or an AWS service in Amazon ML. CloudTrail captures all API calls for Amazon ML as events. The calls captured include calls from the Amazon ML console and code calls to the Amazon ML API operations. If you create a trail, you can enable continuous delivery of CloudTrail events to an Amazon S3 bucket, including events for Amazon ML. If you don't configure a trail, you can still view the most recent events in the CloudTrail console in Event history. Using the information collected by CloudTrail, you can determine the request that was made to Amazon ML, the IP address from which the request was made, who made the request, when it was made, and additional details.

Below is the log entry example in Cloud trail for an action\event on the Predictions:[2]

|  |
| --- |
| {  "Records": [  {  "eventVersion": "1.03",  "userIdentity": {  "type": "IAMUser",  "principalId": "EX\_PRINCIPAL\_ID",  "arn": "arn:aws:iam::012345678910:user/Alice",  "accountId": "012345678910",  "accessKeyId": "EXAMPLE\_KEY\_ID",  "userName": "Alice"  },  "eventTime": "2015-11-12T15:04:02Z",  "eventSource": "machinelearning.amazonaws.com",  "eventName": "CreateDataSourceFromS3",  "awsRegion": "us-east-1",  "sourceIPAddress": "127.0.0.1",  "userAgent": "console.amazonaws.com",  "requestParameters": {  "data": {  "dataLocationS3": "s3://aml-sample-data/banking-batch.csv",  "dataSchema": "{\"version\":\"1.0\",\"rowId\":null,\"rowWeight\":null,  \"targetAttributeName\":null,\"dataFormat\":\"CSV\",  \"dataFileContainsHeader\":false,\"attributes\":[  {\"attributeName\":\"age\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"job\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"marital\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"education\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"default\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"housing\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"loan\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"contact\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"month\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"day\_of\_week\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"duration\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"campaign\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"pdays\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"previous\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"poutcome\",\"attributeType\":\"CATEGORICAL\"},  {\"attributeName\":\"emp\_var\_rate\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"cons\_price\_idx\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"cons\_conf\_idx\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"euribor3m\",\"attributeType\":\"NUMERIC\"},  {\"attributeName\":\"nr\_employed\",\"attributeType\":\"NUMERIC\"}  ],\"excludedAttributeNames\":[]}"  },  "dataSourceId": "exampleDataSourceId",  "dataSourceName": "Banking sample for batch prediction"  },  "responseElements": {  "dataSourceId": "exampleDataSourceId"  },  "requestID": "9b14bc94-894e-11e5-a84d-2d2deb28fdec",  "eventID": "f1d47f93-c708-495b-bff1-cb935a6064b2",  "eventType": "AwsApiCall",  "recipientAccountId": "012345678910"  },  {  "eventVersion": "1.03",  "userIdentity": {  "type": "IAMUser",  "principalId": "EX\_PRINCIPAL\_ID",  "arn": "arn:aws:iam::012345678910:user/Alice",  "accountId": "012345678910",  "accessKeyId": "EXAMPLE\_KEY\_ID",  "userName": "Alice"  },  "eventTime": "2015-11-11T15:24:05Z",  "eventSource": "machinelearning.amazonaws.com",  "eventName": "CreateBatchPrediction",  "awsRegion": "us-east-1",  "sourceIPAddress": "127.0.0.1",  "userAgent": "console.amazonaws.com",  "requestParameters": {  "batchPredictionName": "Batch prediction: ML model: Banking sample",  "batchPredictionId": "exampleBatchPredictionId",  "batchPredictionDataSourceId": "exampleDataSourceId",  "outputUri": "s3://EXAMPLE\_BUCKET/BatchPredictionOutput/",  "mLModelId": "exampleModelId"  },  "responseElements": {  "batchPredictionId": "exampleBatchPredictionId"  },  "requestID": "3e18f252-8888-11e5-b6ca-c9da3c0f3955",  "eventID": "db27a771-7a2e-4e9d-bfa0-59deee9d936d",  "eventType": "AwsApiCall",  "recipientAccountId": "012345678910"  }  ]  } |

Appendix:

[1] <https://aws.amazon.com/solutions/implementations/predictive-maintenance-using-machine-learning/>

[2] <https://docs.aws.amazon.com/machine-learning/latest/dg/logging-using-cloudtrail.html>

[3] <https://docs.aws.amazon.com/machine-learning/latest/dg/amazon-machine-learning-key-concepts.html>

[4] https://docs.aws.amazon.com/machine-learning/latest/dg/controlling-access-to-amazon-ml-resources-by-using-iam.html