LOR Link Checker



Contents

[Document Change Control 3](#_Toc483243896)

[High-Level Solution Overview 3](#_Toc483243897)

[Access Control 3](#_Toc483243898)

[Entity Relationship Diagram 4](#_Toc483243899)

[Web Application Modules 5](#_Toc483243900)

[Admin 5](#_Toc483243901)

[Sources 5](#_Toc483243902)

[Settings 6](#_Toc483243903)

[Users 6](#_Toc483243904)

[Roles 6](#_Toc483243905)

[Report 7](#_Toc483243906)

[Invalid Links 7](#_Toc483243907)

[Warning Links 7](#_Toc483243908)

[Console Application Modules 8](#_Toc483243909)

[S3ObjectExtractor 8](#_Toc483243910)

[LinkExtractor 8](#_Toc483243911)

[LinkChecker 8](#_Toc483243912)

[LLCPackages 8](#_Toc483243913)

[Appendix A: Project Setup /w AWS 9](#_Toc483243914)

[Appendix B: Settings Available 11](#_Toc483243915)

# Document Change Control

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Name | Date | Description |
| Phase 1 | Brian Merrick | 12-17-2016 | Finished Phase 1 |
| Phase 2 | Brian Merrick | 3-19-2017 | Big dent in Phase 2 |
| Phase 3 | Brian Merrick | 5-18-2017 | Finished Phase 2 |
| Phase 4 | Brian Merrick | 1-27-2018 | Converted to Serverless |

# High-Level Solution Overview

The LOR Link Checker is composed of several applications as well as a web interface. The ObjectExtractor will get a list of objects in the specified buckets/prefixes. They will be added to the LLC database when new and if they already are in the database, the date last found will be updated. The LinkExtractor extracts URLs from source locations (such as S3 objects) and stores them in the database for further analysis. The LinkChecker goes through the objects identified from the LinkExtractor and determines if they return valid responses or are broken in any way. The web interface allows management of settings, users, roles and view logs and reports. The LLCPackage application processes the links from user uploaded packages that follow the IMSCC format.

# Access Control

|  |  |
| --- | --- |
| Role | Description |
| Admin | Administrator of the site. Can manage settings, users and roles. |
| ReportUser | Allowed to view reports available from the various applications in the project. |

# Entity Relationship Diagram

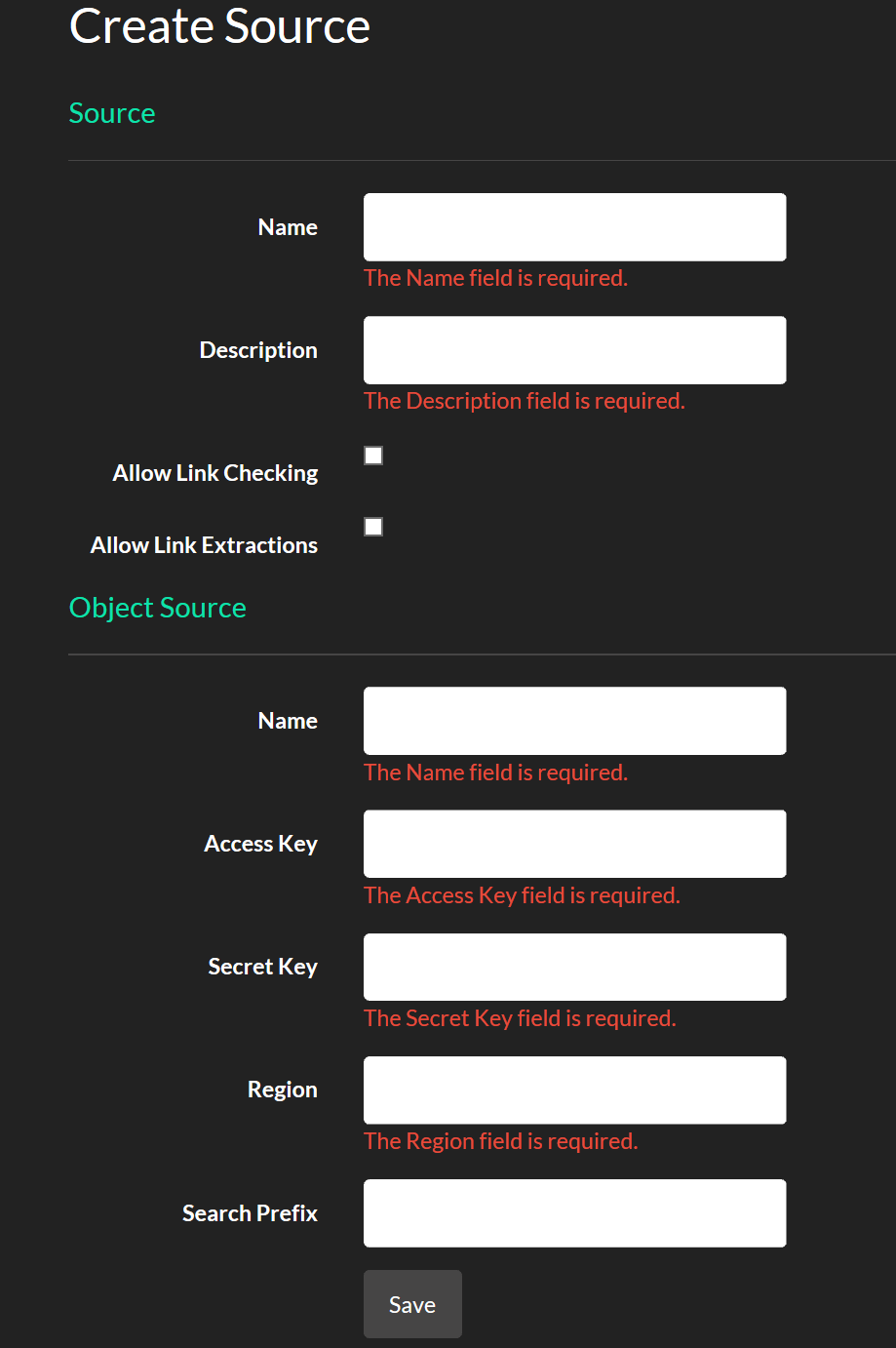
# Web Application Modules

## Admin

The administrative module contains all the features needed to administer the application which includes sources and settings.

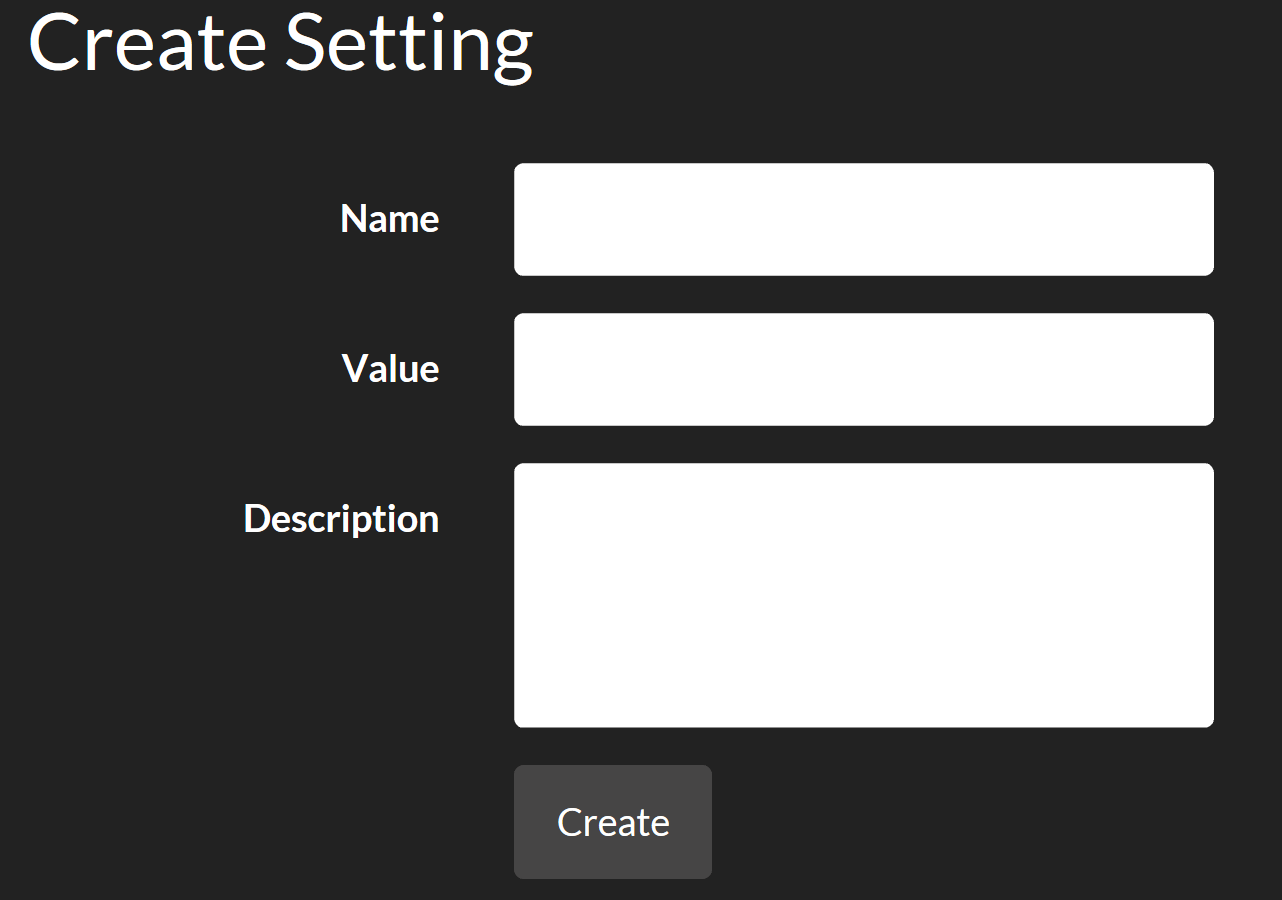
### Sources

This page allows administrators to add/edit/delete sources. A *source* currently is only defined as an S3 object. The create option allows for the following inputs. The boxes with red are required fields.



### Settings

The list of settings found in [Appendix B](#_Appendix_B:_Settings) are the only settings that are supported in the application. The settings screen allows the administrators to add/edit/remove settings. The screen below shows the create options for new/edit settings.



### Users

The users screen allows administrators the ability to remove/edit users. Since authentication is handled by external authentication, add is not supported currently. The edit options allow the administrator to edit user roles that have been defined in the roles module.

### Roles

Administrators can add/edit/remove system roles. All roles must be unique and a role cannot be removed if there are any users who are associated with it.

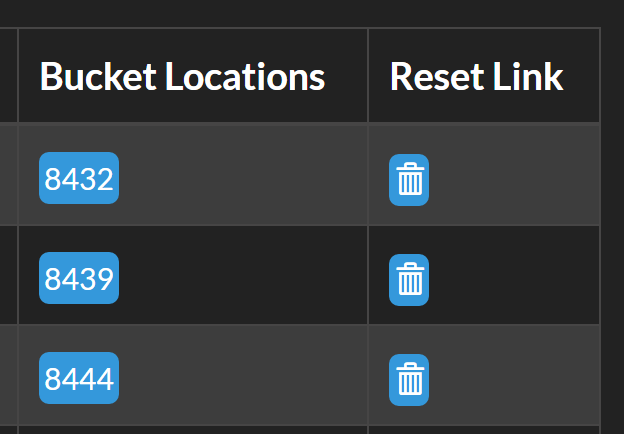
## Report

The reports interface gives admins and report users the ability to see detailed data from the application. Currently two reports exist, the **Invalid Links** and **Warning Links** reports.

### Invalid Links

This report returns all links that have an invalid state from the last time checked. It will show the source, URL, attempt count, date last checked and date last found.

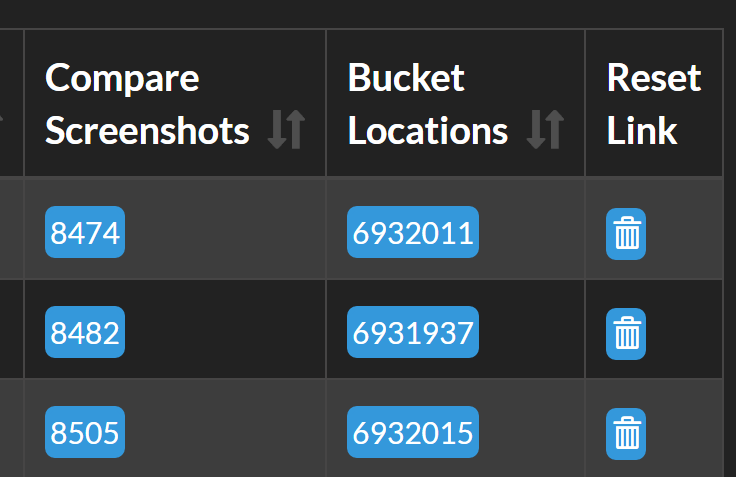
Any link that was in an error state will be updated with an **IsValid** flag set to false (0). If an admin hits the **Reset Link**option on the report, the **IsValid** flag will reset to 1 (and make it valid).



### Warning Links

The warning links report is updated every time the [LinkChecker](#_LinkChecker) application runs. It will check links from the settings that are less than **LinkCheckDays**old and within **LinkCheckStandardDeviations**. If the mean of those links is outside of that number of standard deviations, the last link is added to the report. Otherwise, it is removed from the report since it corrected itself.

By default, all LinkStats will be considered in the calculated of the standard deviation. If an admin hits the **Reset Link** option on the LinkReport, the **ReportNotBeforeDate** on the Link object will be updated with the current date. This means that no LinkStats before this date will be considered in the standard deviation calculation and the current LinkReport entry will be removed.



# Console Application Modules

The Windows console applications perform most the work of the application. The web interface is simply the mechanism to view the results. The available applications are explained below.

## S3ObjectExtractor

This application will get a list of objects in specified buckets/prefixes and will be added to the LLC database. New items will be added and existing items will have their date last found updated to the current time. This application will run daily at 2am. The job for this application runs from the same job as [LinkExtractor](#_LinkExtractor). This application runs immediately followed by the **LinkExtractor**.

## LinkExtractor

This application extracts URLs from source locations (such as S3 objects) and stores the download information in the database for further analysis. This job runs daily at 2am right after the [S3ObjectExtrator](#_S3ObjectExtractor) job.

## LinkChecker

The LinkChecker application will check a fraction of the links extracted from the various sources many times throughout the day. Each check of a URL is stored as a separate entry in the LinkStats table. After all the links have been downloaded and checked, the LinkReport process is ran on the subset. If the mean of the stats within the specified days (set in the admin settings) are outside of the maximum standard deviation allowed, they are added to the report/table.

## LLCPackages

The LLCPackages application runs every 15 minutes and checks for user uploaded packages that have not been processed. Those packages must have a .zip or .imscc extension. The packages will be downloaded from S3, extracted to the file system, and then scanned for links. Once fully processed, the records will be added to the database and flagged as processed. Only processed packages can be removed from the system through the UI.

# Appendix A: Project Setup /w AWS

The following steps build the LOR Link Checker on AWS. These steps show how to build the project from scratch. All steps are automated except for the DB restore from backup. The project documents are located at <https://s3-us-west-2.amazonaws.com/lor-link-checker-templates>.

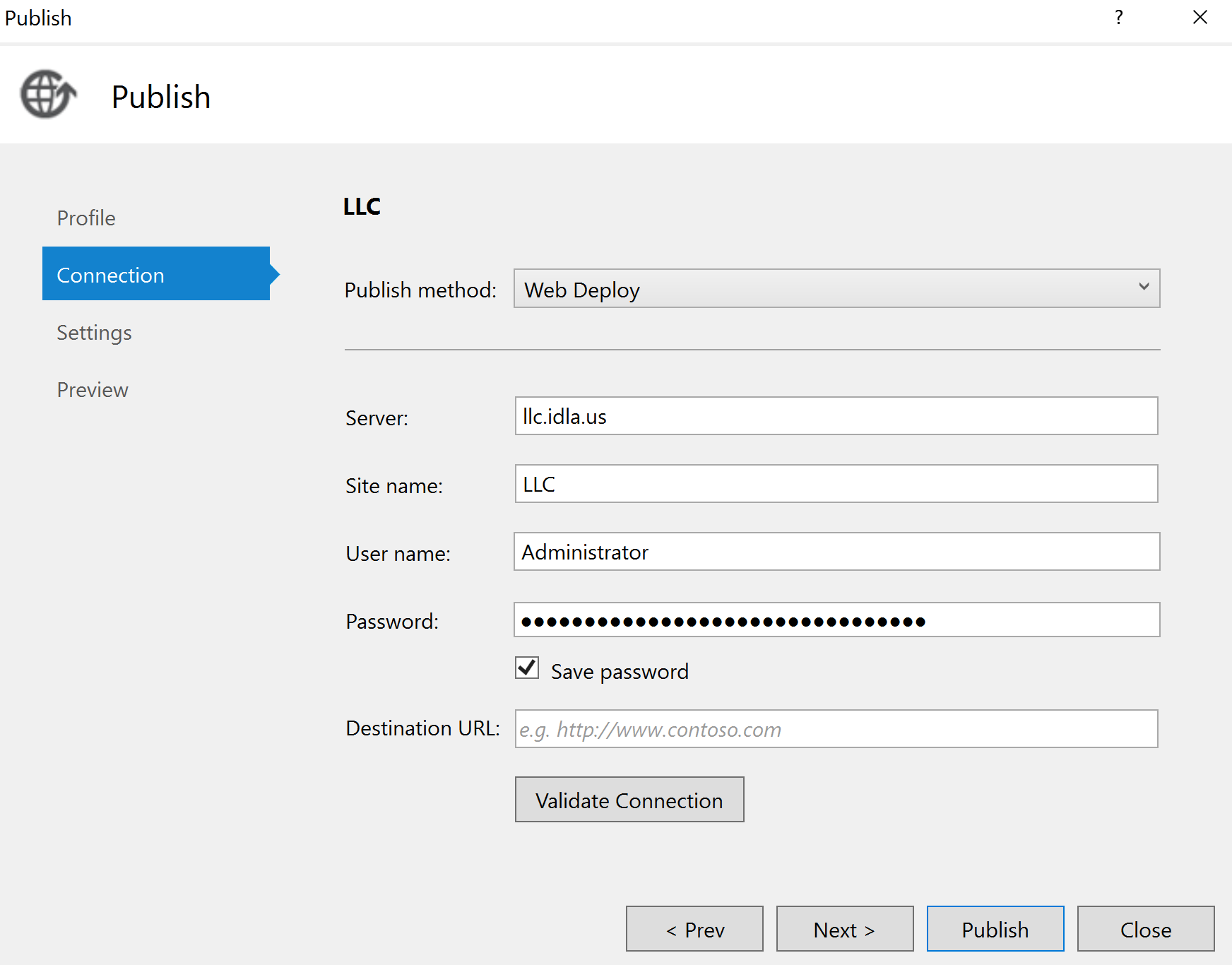
1. Execute the CloudFormation template located at the following location.

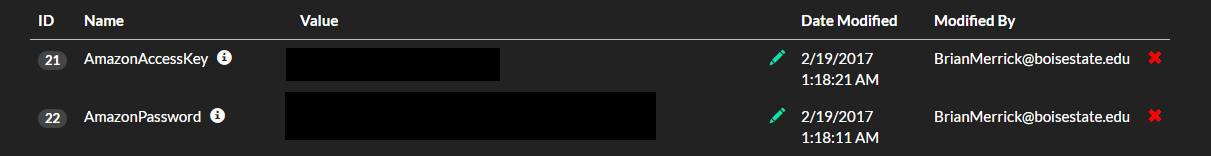
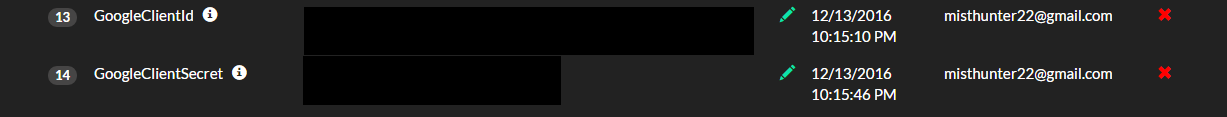
**NOTE:** It can take up to 10-15 minutes for the UserData to fully run and bootstrap the system

* 1. <https://s3-us-west-2.amazonaws.com/lor-link-checker-templates/lor-link-checker-template.json>
  2. Choose a unique stack name and choose the KeyPair *LLC-EC2* (located at <https://s3-us-west-2.amazonaws.com/lor-link-checker-templates/LLC-EC2.pem>). You can choose any other key as well if it is available to be able to use to remote into the instance.

**NOTE:** The current stack name must NOT contain any – or space characters currently as there is a bug in the cloud formation template!

1. Publish the project using Web Deploy via Visual Studio. Be sure to set the Server address and Administrator password for the new environment. Be sure that the **Release** configuration is selected to deploy!



1. Import the SQL .bak file from the latest to the localhost DB located in the S3 bucket from above. If Necessary
   1. create the LLCClient user logon with the password from the web.config
   2. Delete/create the LLCClient user for the LORLinkChecker DB
   3. Set the db\_owner permission for LLCClient on the DB
   4. Set SQL logon types to Windows/SQL Authentication
2. Configure the IAM settings for the application so it can post messages to
   1. Create access keys for the **LORLinkChecker<Stack Name>** user
   2. Copy the settings to the application (see [Appendix B](#_Appendix_B:_Settings) for details)
3. Update Google settings to point to your own Google domain from the admin settings UI
4. Make sure the **Release** configuration is selected. Build the Visual Studio project and copy over the following folder locations onto the EC2 machine. A copy of the projects are currently stored in the S3 bucket with the same names.
   1. **Db C:\IDLA\DbUp**
   2. **LinkChecker C:\IDLA\LinkChecker**
   3. **LinkExtractor C:\IDLA\LinkExtractor**
   4. **S3ObjectExtractor C:\IDLA\S3ObjectExtractor**
   5. **LLCPackages C:\IDLA\LLCPackages**

# Appendix B: Settings Available

|  |  |  |
| --- | --- | --- |
| Setting | Description | Update on New Site |
| AmazonAccessKey | The IAM access key for the application to interact with AWS. | Yes |
| AmazonPassword | The Amazon IAM password that allows the application to interact with AWS. | Yes |
| Domain | The domain for the application to be used in communications for links. | Yes |
| GoogleClientId | OAuth Google Client ID. | Optional |
| GoogleClientSecret | OAuth Google Client Secret. | Optional |
| GoogleRegisterDomain | Leave empty to allow any Google domain to register with the site. Use a comma separated list of domains that are acceptable (i.e., boisestate.edu, idla.k12.id.us). | Optional |
| LinkAnalysisLimit | The maximum number of links to analyze per process run. | Optional |
| LinkCheckDays | The number of days to get links back to check to see if they have changed significantly. | Optional |
| LinkCheckStandardDeviations | Number of standard deviations that the checked links can be between without triggering a notification. | Optional |
| LinkExtractionObjectLimit | The maximum number of objects that will be loaded to be used for link extraction per process run. 25,000 take about an hour. | Optional |
| LogsMaxDisplayFromEnd | The number of log entries to display in the table (from the end of the table). | Optional |
| MaxThreads | The number of threads used for processes that can be multitasked like link checking. | Optional |
| NewUsersCanRegister | True to be able to register new users, false otherwise. | Optional |
| NotificationEmail | The address to send notification emails to. | Optional |
| PackagesS3Bucket | The S3 bucket that holds the packages. You can specify the format, bucket-name;folder to re-use the same bucket with a different folder. | Yes (will use test site S3 bucket otherwise) |
| ScreenshotS3Bucket | The S3 bucket that holds the screenshots. You can specify the format, bucket-name;folder to re-use the same bucket with a different folder. | Yes (will use test site S3 bucket otherwise) |