**KBS Analytics**

**Data Lake Functional Requirements Document**

*\*\*\*Drafting Notes:*

* Text in normal font is intended to be in the final version of the document
* *Text in italics are intended to be notes/instructions for authors of the document (they will go away once the document is finalized)*
* Text in red needs additional time and effort to complete

*Authors:*

*Maha – Primary Author / Coordinator*

*Deepak – Technical Focus*

*Yeonjoo – Data Modeling & Data Organization*

*Ryan – Security Rules, Data Acquisition/Ingest Process*

*Nathan, Kent, Chris, Dermot – Review*

Contents

[1 General 2](#_Toc15276311)

[1.1 Project Description 2](#_Toc15276312)

[1.1.1 Background 2](#_Toc15276313)

[1.1.2 Purpose 2](#_Toc15276314)

[1.1.3 Assumptions and Constraints 2](#_Toc15276315)

[1.2 Points of Contact / Team Members 2](#_Toc15276316)

[1.3 Document References / Resources 3](#_Toc15276317)

[2 FUNCTIONAL REQUIREMENTS 3](#_Toc15276318)

[2.1 Data Requirements 3](#_Toc15276319)

[2.2 Initial Data Organization 3](#_Toc15276320)

[2.3 Continued Data Organization & Maintenance 4](#_Toc15276321)

[2.4 Functional Process Requirements 4](#_Toc15276322)

[2.5 Moving Legacy Production Work to Data Lake 4](#_Toc15276323)

[2.6 Fulfilling New Data Requests 4](#_Toc15276324)

[3 OPERATIONAL REQUIREMENTS 4](#_Toc15276325)

[3.1 Security 4](#_Toc15276326)

[3.2 Audit Trail 5](#_Toc15276327)

[3.3 Reliability 5](#_Toc15276328)

[3.4 System Availability 5](#_Toc15276329)

[3.5 Performance 6](#_Toc15276330)

[3.6 Capacity 6](#_Toc15276331)

[3.7 Data Retention 6](#_Toc15276332)

[4 REQUIREMENTS TRAcEABILITY MATRIX 6](#_Toc15276333)

[5 APPENDIX 6](#_Toc15276334)

# General Maha / Ryan

## Project Description

The KBS Analytics Data Lake aims to be a central data repository of internal and external data that could create value to the business via

### Background

The KBS Analytics team is engaging on a variety of data intensive projects and needs a data storage solution that is secure, flexible, accessible, and low-cost.

### Purpose

Describe the business objectives and business processes from the CONOPS document and the CBA that this application supports.

### Assumptions and Constraints

Assumptions:

|  |  |
| --- | --- |
| KBSA-AS-1 | * For external data sets, our initial focus shall be on loading time series data sets. |
| KBSA-AS-2 | We prefer to load data sets that are refreshed more frequently ie. At least on a monthly basis. Any data set that is refreshed only on a quarterly or yearly basis shall not be loaded into the data lake. |
| KBSA-AS-3 | * A given data set (for example GDP data for USA) shall be loaded only from a single source. The same data set shall not be loaded from multiple sources (prevent duplication). |
| KBSA-AS-4 | * We will load data sets which are newer than 2005. Any data older than 2005 is not required to be loaded into the Data Lake |
| KBSA-AS-5 | Data shall be stored in the raw format without any transformations. |

Constraints:

* ..
* ..

## Points of Contact / Team Members

|  |  |
| --- | --- |
| **Name** | **Role** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Document References / Resources

|  |  |
| --- | --- |
| Source (with link) | Used For: |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# FUNCTIONAL REQUIREMENTS

*\*The functional requirements describe the core functionality of the application. This section includes the data and functional process requirements.*

## Initial Data Organization Yeonjoo

* Source
  + Internal Data
    - Project Name
      * Sensitive Info
      * Public Info
  + External Data
    - Country
      * Growth
      * Money Supply
      * Interest Rate
      * Money Supply
      * Income and Expenditure
      * Employment
      * Debt
      * Other

- Reference / Table of Contents

- Tagging

|  |  |
| --- | --- |
|  |  |
|  | The following tags shall be used as a minimum:  Source, subject matter, data load frequency (daily, weekly, monthly), data sensitivity (public/private), free or paid data source, downstream use, Load By, Load date |
|  |  |

- Data used in production workflows (does this need special considerations?)

## Continued Data Organization & Maintenance Kent (Move to Solution Document)

- Assign a data curator

- Data Quality Checks

- Methods for finding duplicate datasets

- Methods for creating heuristics

## Functional Process Requirements Maha (Yeonjoo to create template for inputing requirements)

1. Data Lake Access Points

- Data Lake Console, AWS Console, API

2a. Uploading Data to the Data Lake

- Tagging, security, etc.

2b. Automated Uploads of Data

|  |  |
| --- | --- |
| KBSA-LOAD-1 | Data sets shall be extracted and automatically loaded into the data lake on a periodic basis (daily, weekly or monthly basis). The periodicity shall be predefined for each job in the AWS Glue catalog. |
| KBSA-LOAD-2 | The solution shall have the ability to manually refresh a data set on an ad-hoc basis |
|  |  |
|  |  |
|  |  |

**Storage and Data Format**

|  |  |
| --- | --- |
| KBSA-STR-1 | Data shall be compressed using the Snappy library and shall be stored in columnar Parquet format. |
| KBSA-STR-2 | Only the latest version of a data set shall be made available for analysis. The older versions shall be archived in a separate folder and shall be available only to admins. |
| KBSA-STR-3 | Older archived version of the data set shall be available to be restored as an active version |
| KBSA-STR-4 | Archived data shall be available for XX (7?) years |

**Ingest**

|  |  |
| --- | --- |
| KBSA-IN-1 | The solution shall be capable of ingesting data using different upload mechanisms. There shall be mechanisms for both manual ingestion and automated ingestion. |
| KBSA-IN-2 | The user shall have the ability to create a new data set or to overwrite an existing data set. Note: The ability to append to an existing data set is not required. |
| KBSA-IN-3 | The solution shall accept manually uploaded source files in csv, json or xml format. |
| KBSA-IN-4 | Immediately after a new data set is manually uploaded, a program (example: glue job) shall be triggered to create metadata and make the dataset available for SQL querying. |
| KBSA-IN-5 | The following data values shall be automatically appended to the end of the dataset:  Loadby, Load\_date, Source\_description |
| KBSA-IN-6 | For manual data ingestion, the user shall input a set of metadata tag descriptions along with the data upload. There shall be both mandatory metadata tags and optional metadata tags. The mandatory metadata tags are as follows:  Data Source:  Data set name:  Data set description:  Category1  Category2  Category3  The Optional tags are as follows:  Comments  Category4  Category5  The following tags shall be automatically included by the system:  Uploaded user name  Upload date |
|  |  |
|  |  |
| KBSA-IN-1 | API calls call be used to load the data into the data lake. |
| KBSA-IN-2 | A configuration table shall be created with the following data elements at a minimum:   |  |  | | --- | --- | | **DATA ELEMENT** | **SAMPLE DATA** | | Data Set ID | 1001 | | Data Source | EIA – Energy Information Authorization | | Data set Name | Natural gas consumption for electricity generation in USA | | Data set Description | Natural gas consumption for electricity generation | | Category1 | USA | | Category2 | Energy | | Category3 | Natural Gas | | Category4 |  | | Category5 |  | | Comments |  | | Frequency of Load | Monthly | | API details | http://api.eia.gov/category/?api\_key=YOUR\_API\_KEY\_HERE&category\_id=432 | | Data Format | CSV | | Oldest time period (year) | 2006 | | Number of rows Loaded | 5000 | | Last load date | 7/30/2019 | | Load By | Maha Sadanandan | |  |  | |
|  |  |

3. Downloading Data from the Data Lake

|  |  |
| --- | --- |
| KBSA-DWN-1 | The solution shall allow bulk download of data sets (one or more data sets) to the desktop. Data shall be downloaded in CSV format |
| KBSA-DWN-2 | A user shall have the ability to view on screen (sample data or all the data) for a given data set. The user can then choose to download the data to their desktop. |
| KBSA-DWN-3 | The solution shall allow an analyst to copy a data set to a temporary location in the data lake (in order to transform or to do any ad-hoc analysis) |
| KBSA-DWN-4 | If the user does not have the proper security access to download/view a data set, a message stating “You do not have access to this data set. Please contact your administrator to grant access” shall be displayed and the data set shall not be loaded. |
|  |  |
|  |  |

4. Searching for Data in the Data Lake

|  |  |
| --- | --- |
| KBSA-SRH-1 | The user shall have the ability to do a key word search. The solution shall be capable of partial key work search (beginning, ending, containing etc.). |
| KBSA-SRH-2 | The search results shall include any data set names, tags, or metadata that matches the search key. The search result shall include a data set even if the user does not have security access to that data set. |
| KBSA-SRH-3 | If the user does not security access to a data set in the search result, the data values shall not be displayed to the user. Only the meta data shall be displayed to the user. |

5. Access

Use Case Example 1:

Use Case Example 2:

## Moving Legacy Production Work to Data Lake [Explain] Deepak (Move to Solution Doc)

- Identify Workstreams DEEPAK

- Plan per workstream DEEPAK

## Fulfilling New Data Requests Maha (Move to Solution Doc)

*Describe the process to make requests for new data sets to be added to the data lake.*

*Who, What, When*

*If a new data set has to be loaded into the data lake, a ServiceNow ticket shall be created providing the following information:*

*Name of the dataset:*

*Description:*

*Load Frequency:*

*Links to Source data: API calls, links to the database etc.*

*Other Information: Tags,*

*When do you need this data?*

# OPERATIONAL REQUIREMENTS

*\*Operational requirements describe the non-business characteristics of an application.*

## Security Ryan / Yeonjoo

Data Lake security is crucial to the efficacy of the project….

**User Access**

DL will need to have the following types of users

KBS Employees: All

KBS Contractors: Permissioned as needed for the customer group they’re working with

External Contractors: External Only

- Any need for folders that are hidden from certain users?

- Interns, contractors, project contractors

Who can View, Alter, Delete

**Data Encryption**

* **Security Restrictions**

- Deletion only occurs when…

- No HR Data

- No PII

- No data under daily frequency

## Audit Trail

The following activities will need to be logged in the application’s audit trail.

1. User Uploads Data

- Data to store

2. User Downloads Data

- Data to store

3. Titles, Descriptions, Tagging changes

- Data to store

4. All API Calls

- Data to store

## Reliability

* What damage can result from this system’s failure?
  + - * + Loss of revenue
        + Loss of employee productivity
* What is the minimum acceptable level of reliability?

## System Availability

The application must be available to users at all times, with peak hours being Monday through Friday between the hours of 2:30am ET / 12:00pm IST and 5:00 pm ET / 3:30am IST.

There should be no intended outage, or maintenance the requires downtime without discussion with the entire KBS Analytics development team (via the MS Teams Site).

## Performance

Describe the requirements for the following:

* Response time for queries and updates
* Throughput
* Expected volume of data
* Expected volume of user activity (for example, number of transactions per hour, day, or month)

## Capacity

Intended capacity for the data lake…

## Data Retention

- Are all datasets retained?

- Do new data queries of the same dataset overwrite the original file?

- Archiving

- Retention Schedules to discard sensitive data

Describe the length of time the data must be retained. For example, “information about an application for naturalization must be retained in immediately accessible from for three years after receipt of the application.”

# REQUIREMENTS Traceability MATRIX

*Include columns for each of the following in the RTM:*

* *Requirement description*
* *Requirement reference in FRD*
* *Verification Method*
* *Requirement reference in Test Plan*

Requirements Traceability Matrix

# APPENDIX

Technical Design Document

|  |  |  |
| --- | --- | --- |
| Tag Name | Required? | Field Options |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |