Back to Investment Detail

Business Case

Table of Contents

Section A: Investment Summary Information

Investment Name

Census - Enterprise Data Lake (EDL)

Unique Investment Identifier

006-000000136

Investment Description

Enterprise Data Lake is a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos.

Agency

Department of Commerce

Point of Contact

Andre Mendes - CIO

email **202-482-4797**

Investment Type

Major IT Investments

Bureau

Bureau of the Census

Mission Support

Not Applicable

Shared Service Category

Not Applicable

Shared Service Identifier

Not Applicable

Date Investment First Submitted

09/20/2021

Date of Last Investment Detail Update

05/31/2022

Section B: Investment Detail

1. Briefly describe the investment's return on investment, including benefits internal and external to the government and outcomes achieved or planned.

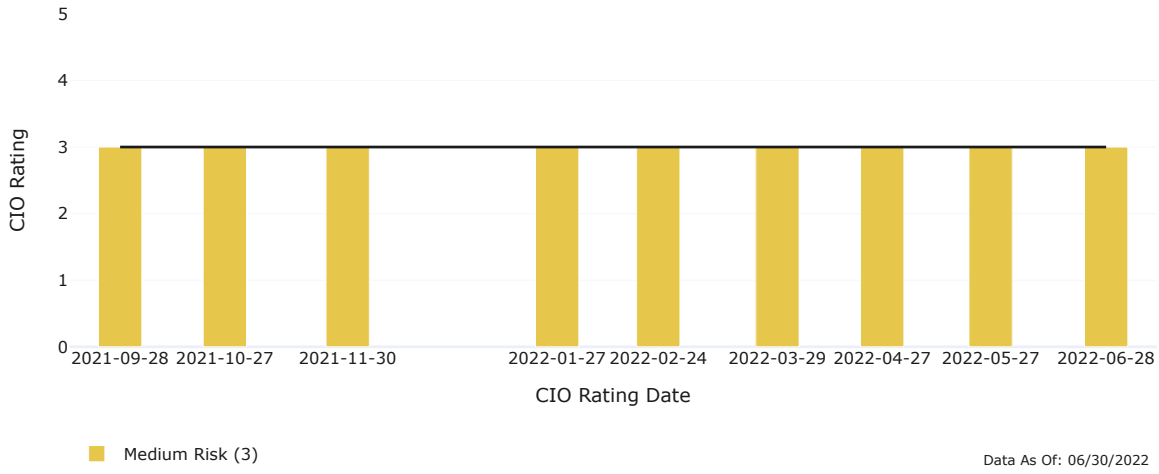
   The EDL investment will support the Census's data and analytical needs in a secure, scalable, high performance storage and computing cloud environment with appropriate backups to Census data center. This platform will increase the Bureau's capability to ingest the ever increasing volume of administrative records, improve the quality of data produces and apply disclosure avoidance to protect PII data as required by Title 13, Title 26, and he other data protection laws.

Section C: Investment and Contracts

Contracts

- 1333LB20F00000398
- 1333LB19F00000133
- 1333LB20F00000344
- 1333LB20F00000273
- 1333LB20F00000502
- 1333LB19F00000334
- 1333LB20A00000008
- NNG15SD11B
- 1333LB20F00000405
- YA1 323-14-NC-0058
- 1331L521F13230258
- 1333LB19F00000509
- DOCYA132317BU0011
- 1333LB-20-F-00000210
- 1333LB21A00000002
- 1333LB19F00000520

Section D: Historic CIO Rating



CIO Rating	Date	Comments
3	Jun 28, 2022	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Notable Accomplishments: Transferred 160 TB of Decennial data from CDL to EDL. Completed provisioning of Phase 1 AWS Managed Services (Redshift, Sagemaker, Glue, Athena, Cloudwatch and Lakeformation). SAS installation in GovCloud West is complete and CDL users have been added. Provided a demo on SAS to the EDL Value Team. The investment is rated 3 due to risk exposure and reporting concerns.
3	May 27, 2022	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Notable accomplishment: Installed SAS in GovCloud West (GCW) DEV environment and Jupyter in GCW DEV environment. Provisioned Census Data Lake (CDL) required Phase 1 AWS services (including Redshift, Glue, Athena, CloudFormation) in GCW. Completed s3 replication for CQA, SRQA and CDL from TI to EDL s3 in GovCloud West (GCW). All Decennial data has been migrated from Census Data Lake (CDL) to EDL's GCW. Required resources (s3, SAS) has been provisioned for Community Resilience Estimate (CRE). Began the design of the solution for EDC (EDL Enterprise Data Catalog), its components and interfaces. The investment is rated 3 due to risk exposure.

CIO Rating	Date	Comments
3	Apr 27, 2022	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Accomplishment: Provisioned Census Data Lake (CDL) required Phase 1 AWS services (including Redshift, Glue, Athena, CloudFormation) in GCW. Updated SAS licenses on all EDL SAS servers. Began design for API solution for DMS dataset registration in the WebUI. Migrated 1400+ tabled into HIVE GovCloud East (GCE). Provisioned CDL S3 bucket for replication in GovCloud West (GCW). S3 transfer to EDL GovCloud West started with 80% of s3 replication complete. Completed CDL Atlas migration. CDL GitHub migration completed. Completed SAS CDL Migration tasks in GovCloud East. Completed environment and configurations for 4 HRD Capstone projects. Trained HRD users. Provided a demo on SAS to the EDL Value Team. The investment is rated 3 due to risk exposure.
3	Mar 29, 2022	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Accomplishments: Developed DEV, ITE, and production environments. Finalized EDL Concept of Operations (ConOps) document that illustrates concepts, business roles and requirements, conceptual architecture and service diagrams. Completed EDL End-End process flow. The investment is rated 3 due to risk exposure.
3	Feb 24, 2022	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Accomplishments: Deployed VPC and activated VPN on DEV,ITE,STAGE and PROD in Govcloud WEST. Designed and developed a serverless Data Lake using Lambda, CloudWatch, Events, Glue Crawlers, Glue Jobs and S3. Completed an s3-s3 connection with the National Archives and Records Administration. Finalized EDL Concept of Operations (ConOps) document that illustrates concepts, business roles and requirements, conceptual architecture and service diagrams. Completed EDL End-End process flow. The EDL Onboarding Checklist and requirements phase has been completed for 7 UAT projects. The investment is rated 3 due to risks.
3	Jan 27, 2022	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Accomplishments: Completed requirements and design for GovCloud West Virtual Private Cloud (VPC) in DEV. Completed two critical tasks for NARA connection and s3 PROD replication test. Successfully configured the SSL for both CDL and CRE SAS Production Instances. Completed Onboarding Checklist. Developed additional functionality for importing data (CSVs, delimited text files, and SAS datasets) into EDL RDS/Postgres in support of the Frames and CEM User Acceptance Testing (UAT) projects. The investment is rated 3 due to progress tracking and operational performance metrics.
3	Nov 30, 2021	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Accomplishments: Obtained approval of EDL vision and scope statement by senior leadership. Detailed roadmap will be developed based on approval. Completed installation and configuration of CDP instance in production. Installed and configure SAS in ITE and Production. Deployed WEBUI to Development and ITE. Designed and developed microservice for registering metadata and searching for available datasets and metadata. The investment is rated 3 due to progress tracking and operational performance metrics.
3	Oct 27, 2021	Enterprise Data Lake continues to be a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Accomplishments: Obtained approval of EDL vision and scope statement by senior leadership. Developed production and ITE environments. Prepared the CDL-EDL migration. The investment is rated 3 due to progress track and operational performance metrics.
3	Sep 28, 2021	Enterprise Data Lake is a flexible data repository intended to provide the Census Bureau with a next-generation scaling capability to fulfill data storage, reporting, analytics, and security needs while reducing costs associated with duplicative data silos. Accomplishments: DEVELOP PRODUCTION and ITE ENVIRONMENTS - Continued work to configure CDP Common Cluster base. Continued work to install and configure SAS in ITE for Production. Webui - Designed and developed microservice for registering metadata and searching for available datasets and metadata. Architecture - Provided architecture views for Data Import, Data Storage, and Data Processing; referenced architecture for use of AWS services, logical and physical data model for basic metadata to support Data Import and Data Storage. CDL EDL MIGRATION PREPARATION - Worked on feature sizing, design and implementation for Data Import, Data Storage, and Data Processing. EDL receives a CIO rating of 3 due to reporting concerns.

Data Last Updated On: 06/30/2022

Section E: Investment Spending

Table 1: Distribution by Spending Type

Spending Type	PY 2021	CY 2022	BY 2023
DME Costs	7.773109	18.035186	13.806789
O&M Costs	6.359609	7.254145	5.409213
Total	14.132718	25.289331	19.216002

Table 2: Distribution by Cost Pools

Cost Pools	PY 2021	CY 2022	BY 2023
Internal Labor	3.860237	7.023321	5.337479
External Labor	7.949847	15.616676	11.865123
Outside Services	2.203028	2.512903	1.909718

Cost Poc

Cost Pools	PY 2021	CY 2022	BY 2023
Hardware	0	0	0
Software	0.119606	0.136431	0.103682
Facilities & Power	0	0	0
Telecom	0	0	0
Other	0	0	0
Internal Services	0	0	0
Totals	14.132718	25.289331	19.216002

Cost in millions (M)

Table 3: Distribution by IT Towers

IT Towers

IT Tower	PY 2021	CY 2022	BY 2023
Security & Compliance	0.765624	1.776404	1.350003
IT Management	1.307047	3.032614	2.304683
Network	0.21154	0.241295	0.183376
Data	0	0	0
Compute	2.630292	3.000267	2.280097
Storage	2.340717	2.669959	2.029075
End User	0	0	0
Output	0	0	0
Application	6.17557	12.940175	9.834076
Delivery	0.701928	1.628617	1.234692
Platform	0	0	0
Data Center	0	0	0
Totals	14.132718	25.289331	19.216002

Cost in millions (M)

Data Last Updated On: 05/31/2022

Section F: Project and Activities Detail

Table 1: Project Details

Project Name	Project UID	Status	Project Life Cycle Cost ($M)	Cost Variance (%)	Start Date	End Date	Schedule Variance (%)	Schedule Variance (Days)
Enterprise Data Lake	0001D21002	Complete	11.06	-28	2021-03-01	2022-01-19	9.4	9.4
EDL User Acceptance Testing	0001P22003	In Progress	4.38	2.7	2022-01-01	2022-12-30	0	0
Integrate EDL with Major Enterprise Initiatives	0001P22004	In Progress	10.41	1.7	2022-01-01	2023-03-31	0	0
Cumulus Integration Platform	0001P22006	In Progress	2.87	17	2022-01-01	2022-06-30	0	0
Census Data Lake to Enterprise Data Lake Data Migration	0001P22002	In Progress	3.29	18	2022-01-01	2022-06-30	0	0
Build IT Ecosystem	0001D22006	In Progress	6.03	0	2022-01-01	2023-03-31	0	0
Operationalize EDL	0001P22005	In Progress	3.29	0	2022-01-01	2023-03-30	0	0

Low Medium High

Table 2: Activity Details

Unique Project ID	Activity Name	Activity Description	Planned Start Date	Projected Start Date	Actual Start Date	Planned Completion Date	Projected Completion Date	Actual Completion Date	Planned Total Cost ($M)	Projected Total Cost ($M)	Actua Total Cost ($M)

Unique Project ID	Activity Name	Activity Description	Planned Start Date	Projected Start Date	Actual Start Date	Planned Completion Date	Projected Completion Date	Actual Completion Date	Planned Total Cost ($M)	Projected Total Cost ($M)	Actua Total Cost ($M)
0001D21002	HDP to CDP	Complete the upgrade/migration from Hortonworks Data Platform (HDP) to the Cloudera Data Platform (CDP)	2021-03-08	2021-03-08	2021-03-08	2021-09-01	2021-09-24	2021-10-01	3.87	5.232769	4.95
0001D21002	Production Environment	Build the EDL production environment	2021-03-08	2021-03-08	2021-03-08	2021-10-01	2021-11-10	2021-11-15	3.32	4.48523	4.24
0001D21002	Data Governance and Frameworks	Create the foundational data governance and data ingestion/processing frameworks	2021-03-08	2021-03-08	2021-03-08	2021-12-06	2021-12-30	2021-09-01	2.21	2.990154	2.83
0001D21002	Program Management Plan	Finalize program management plan and obtain IT Investment Authority for the full scope of EDL	2021-03-08	2021-03-08	2021-03-08	2021-12-30	2021-12-30	2021-12-02	1.66	2.242615	2.12
0001P22003	Establish Users and User Profiles	Establish 5-10 Active users, daily users, or user profiles established	2022-01-03	2022-01-03	2022-01-03	2022-03-30	2022-03-30	2022-03-30	1.64	1.644184	1.53
0001P22003	Migrate 2 addtional UAT programs into EDL	Migrate 2 additional UAT programs into EDL	2022-07-01	2022-07-01		2022-09-30	2022-09-30		1.1	1.096122	0
0001P22003	Migrate 3 UAT programs into EDL	Migrate 3 UAT programs into EDL	2022-04-03	2022-04-03	2022-04-03	2022-06-30	2022-06-30		1.1	1.096122	0
0001P22003	Update onboarding materials based on UAT results	Update onboarding materials based on UAT results	2022-10-01	2022-10-01		2022-12-30	2022-12-30		0.55	0.548061	0
0001P22004	Complete updated System of Record Notice (SORN) for EDL	Complete updated System of Record Notice (SORN) for EDL	2022-01-03	2022-01-03	2022-01-03	2022-03-01	2022-03-01	2022-03-01	0.05	0.054806	0.05
0001P22004	Complete EDL Program Management Plan	Complete EDL Program Management Plan	2022-01-03	2022-01-03	2022-01-03	2022-03-01	2022-03-01	2022-03-01	0.11	0.109612	0.1
0001P22004	EDL Program Lifecycle Estimate	Complete the EDL Program Lifecycle Estimate as a component of the development of a common IT ecosystem.	2022-04-01	2022-04-01	2022-04-01	2023-02-15	2023-02-15		1.81	1.808602	0
0001P22004	EDL component of a Phase One plan	EDL component of a Phase 1 plan document that covers integration of the major Census Bureau IT budget initiatives moving toward a common IT ecosystem for census business operations, and addresses milestones for the FY 2024 appropriations cycle.	2022-01-03	2022-01-03	2022-01-03	2022-04-15	2022-04-15	2022-04-15	0.24	0.237493	0.15
0001P22004	Complete EDL Project Management Plans for FY22-FY23 Folio Projects	Complete EDL Project Management Plans for FY22-FY23 Folio Projects	2022-01-03	2022-01-03	2022-01-03	2022-04-15	2022-04-15	2022-04-15	0.18	0.182687	0.1

Unique Project ID	Activity Name	Activity Description	Planned Start Date	Projected Start Date	Actual Start Date	Planned Completion Date	Projected Completion Date	Actual Completion Date	Planned Total Cost ($M)	Projected Total Cost ($M)	Actua Total Cost ($M)
0001P22004	Framework to Integrate Frames	Design a framework to integrate Frames Linking application to EDL.	2022-01-03	2022-01-03	2022-01-03	2022-09-30	2022-09-30		0.93	0.931704	0
0001P22004	Design for the Data Interface	Complete the design for the production data interface between Data Ingest and Collection for the Enterprise (DICE) systems and the Enterprise Data Lake.	2022-01-03	2022-01-03	2022-01-03	2022-09-30	2022-09-30		1.17	1.169197	0
0001P22004	Import of Frames Data	Complete import of Frames data into the Enterprise Data Lake	2022-01-03	2022-01-03	2022-01-03	2022-09-30	2022-09-30		0.38	0.383643	0
0001P22004	Complete 3-5 year EDL program plan	Complete 3-5 year EDL program plan for work beyond March 2023	2022-04-01	2022-04-01	2022-04-01	2023-02-15	2023-02-15		1.59	1.589378	0
0001P22004	Complete common IT ecosystem	Complete common IT ecosystem integration work through March 2023	2022-01-03	2022-01-03	2022-01-03	2023-03-31	2023-03-31		3.95	3.946041	0
0001P22006	Validate and complete production and migration	Validate and complete production and migration	2022-01-03	2022-01-03	2022-01-03	2022-02-28	2022-02-28	2022-02-28	0.48	0.477505	0.16
0001P22006	Validate and complete stage buildout	Validate and complete stage buildout	2022-01-03	2022-01-03	2022-01-03	2022-04-29	2022-04-29	2022-04-29	0.6	0.596881	0.33
0001P22006	Complete build out of the Cumulus System in AWS GovCloud	Complete build out of the Cumulus System in AWS GovCloud	2022-03-01	2022-03-01	2022-03-01	2022-06-30	2022-06-30		1.31	1.313138	0
0001P22006	Validate and complete IT buildout	Validate and complete IT buildout	2022-01-03	2022-01-03	2022-01-03	2022-03-31	2022-03-31	2022-03-31	0.48	0.477505	0.59
0001P22002	Migrate Data - Core 2018 CT and 2020	Migrate Data - Core 2018 CT and 2020	2022-01-03	2022-01-03	2022-01-03	2022-04-11	2022-04-11	2022-04-11	0.16	0.164418	0.05
0001P22002	Install Tools and Libraries	Install Tools and Libraries	2022-01-03	2022-01-03	2022-01-03	2022-04-30	2022-04-30	2022-04-30	1.19	1.187466	0.82
0001P22002	Provision Initial Compute and Storage	Provision Initial Compute and Storage	2022-01-03	2022-01-03	2022-01-03	2022-03-31	2022-03-31	2022-03-31	0.27	0.274031	0.15
0001P22002	Migrate Data - DSSD & SRQA	Migrate Data - DSSD & SRQA	2022-05-02	2022-05-02		2022-06-12	2022-06-12		0.29	0.292299	0
0001P22002	Verify EDL Data and Tools	Verify EDL Data and Tools	2022-05-02	2022-05-02		2022-06-30	2022-06-30		0.33	0.328837	0
0001P22002	Data transfer CDL to EDL	Complete data transfer from the 2020 Census Data Lake (CDL) to the Enterprise Data Lake (EDL)	2022-05-03	2022-05-03		2022-06-30	2022-06-30		0.26	0.255762	0
0001P22002	Migrate Data - Late 2020 Files	Migrate Data - Late 2020 Files	2022-02-02	2022-02-02	2022-02-02	2022-06-22	2022-06-22		0.79	0.785544	0.5
0001D22006	Native AWS Services Phase 2 Security and Implementation: Transient Compute	Native AWS Services Phase 2 Security and Implementation: Transient Compute, FY22 Q3 Q4: Compute power for CDL data	2022-07-01	2022-07-01		2022-09-30	2022-09-30		1.64	1.644184	0

Unique Project ID	Activity Name	Activity Description	Planned Start Date	Projected Start Date	Actual Start Date	Planned Completion Date	Projected Completion Date	Actual Completion Date	Planned Total Cost ($M)	Projected Total Cost ($M)	Actua Total Cost ($M)
0001D22006	Native AWS Services Phase 1 Security and Implementation	Native AWS Services Phase 1 Security and Implementation: FY22 Q1-support CDL migration, early 2022 research, general CDL research support	2022-01-03	2022-01-03	2022-01-03	2022-06-30	2022-06-30		1.1	1.096122	0
0001D22006	Native AWS Services Phase 3 Security and Implementation: DICE:	Native AWS Services Phase 3 Security and Implementation: DICE: ready for 2023	2022-10-01	2022-10-01		2023-03-31	2023-03-31		3.29	3.288367	0
0001P22005	Implement System Monitoring controls	Implement System Monitoring controls	2022-06-01	2022-06-01		2022-12-30	2022-12-30		0.88	0.876898	0
0001P22005	Finalizing onboarding package	Finalizing onboarding package	2022-12-01	2022-12-01		2023-03-30	2023-03-30		1.13	1.13266	0
0001P22005	Finalize Incident Management procedures	Finalize Incident management procedures	2022-01-03	2022-01-03	2022-01-03	2022-09-30	2022-09-30		1.28	1.27881	0

Table 3: Project Related Details

Enterprise Data Lake

1. Are information technology investments adequately implementing incremental development methodology? (Y/N)

   Yes

2. What is the frequency of incremental development iterations? (ex. 1 month, 3 months, 6 months, 12 months or greater)

   Weeks

3. Please describe the iterative development methodology being employed. (500 characters or less)

   Agile

EDL User Acceptance Testing

1. Are information technology investments adequately implementing incremental development methodology? (Y/N)

   Yes

2. What is the frequency of incremental development iterations? (ex. 1 month, 3 months, 6 months, 12 months or greater)

   Weeks

3. Please describe the iterative development methodology being employed. (500 characters or less)

   Agile (8 week Program Increments with 4 2-week sprints each)

Integrate EDL with Major Enterprise Initiatives

1. Are information technology investments adequately implementing incremental development methodology? (Y/N)

   Yes

2. What is the frequency of incremental development iterations? (ex. 1 month, 3 months, 6 months, 12 months or greater)

   Weeks

3. Please describe the iterative development methodology being employed. (500 characters or less)
Agile (8 week Program Increments with 4 2-week sprints each)

Cumulus Integration Platform

1. Are information technology investments adequately implementing incremental development methodology? (Y/N)
Yes

2. What is the frequency of incremental development iterations? (ex. 1 month, 3 months, 6 months, 12 months or greater)
Weeks

3. Please describe the iterative development methodology being employed. (500 characters or less)
Agile (8 week Program Increments with 4 2-week sprints each)

Census Data Lake to Enterprise Data Lake Data Migration

1. Are information technology investments adequately implementing incremental development methodology? (Y/N)
Yes

2. What is the frequency of incremental development iterations? (ex. 1 month, 3 months, 6 months, 12 months or greater)
Weeks

3. Please describe the iterative development methodology being employed. (500 characters or less)
Agile (8 week Program Increments with 4 2-week sprints each)

Build IT Ecosystem

1. Are information technology investments adequately implementing incremental development methodology? (Y/N)
Yes

2. What is the frequency of incremental development iterations? (ex. 1 month, 3 months, 6 months, 12 months or greater)
Weeks

3. Please describe the iterative development methodology being employed. (500 characters or less)
Agile (8 week Program Increments with 4 2-week sprints each)

Operationalize EDL

1. Are information technology investments adequately implementing incremental development methodology? (Y/N)
Yes

2. What is the frequency of incremental development iterations? (ex. 1 month, 3 months, 6 months, 12 months or greater)
Weeks

3. Please describe the iterative development methodology being employed. (500 characters or less)
Agile (8 week Program Increments with 4 2-week sprints each)

Return to top

About **Feedback**

Frequently Asked Questions

Accessibility **Privacy Policy**

Freedom of Information Act


General Services Administration


Office of Management and Budget