SP-114 Data Lake Implementation

CS 4850, Section 02, Spring 2026

January 25, 2026

Roles	Name	Role	Contact
Team Leader	Caleb Cox	Document project progress; Coordinate communications and development between team members	470-819-7917
Team Members:	Bryce Wishart	Documentation, development, programming, testing.	404-683-1147
Advisor/Instructor	Sharon Perry	Facilitate project progress; advise on project planning and management.	770-329-3895

Caleb Cox
Team Leader

Bryce Wishart
Group Member

Abstract

A data lake is a repository of raw data on a server database, allowing for scalability of organized data. Using various architectures, raw data can be sorted into various schemas to provide structure. This project asks for a team to create a mini data lake that utilizes open-source tools to create a bronze-silver-gold data flow for the data lake in a clean database.

The purpose of this project is to expose students to the design principles of data storage systems and schema normalization.

Project Website

www.sp114datalakeKSU.com

Deliverables

1. MinIO object store holding raw datasets
2. Parquet formatted datasets
3. A DuckDB database for querying data from the object store
4. Source code and documentation for all sections of the data lake
5. A report that details the design and implementation of our project
6. A presentation video that demonstrates the operation of our project
7. A website which contains all the previously mentioned deliverables, plus a link to a public GitHub for the project

Group Meeting Schedule Date/Time

February 3rd, 2026, Tuesday 5:00 pm

March 3rd, 2026, Tuesday 5:00 pm

March 31st, 2026, Tuesday 5:00 pm

April 7th, 2026, Tuesday 5:00 pm

April 28th, 2026, Tuesday 5:00 pm

Collaboration and Communication Plan

For communications, our group will mainly use Discord, Outlook, and Text messaging. Since our group is small, we will mainly communicate as we are working on specific sections of the project. When major milestone dates come up, such as deliverable dates, we have group meetings set up the Tuesday before the due date to measure progress. Additionally, we will meet in person following the conclusion of class on Tuesdays and Thursdays as needed to facilitate effective development.

The team leader will be responsible for coordinating these meetings and communicating with project members if they are unavailable.

Project Schedule and Task Planning

See attached Gantt chart file.

Phase	Tasks	Complete%	Current Status Memo	Assigned To	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
Planning	Determine project group	100%		Caleb, Bryce			1													
	Set up communications betwee	100%					1													
	Create Project Plan	100%		Caleb			2													
	Define medallion architecture st	20%		Caleb,Bryce				5												
Software Requiren	Define requirements	0%		Caleb				4												
& Design	Define hardware and software	0%		Bryce				4												
	Explore datasets and plan aggre	0%		Caleb					3											
	Design database	0%							10	10										
	Design Frontend program	0%								4	10									
	Develop and test working proto	0%		Bryce							10	10								
Development	Review prototype design	0%										2								
	Rework requirements	0%										3								
	Create development document	0%											2							
	Test product	0%											5	10						
Presentation	Presentation preparation	0%												2	4					
	Website preparation	0%												2	4	4				
	Present project in class	0%													1	1				
Final report	Update requirements and proto	0%														10	10	10		
	Draft Final report	0%																3	3	3
	Final report submission to	0%																		2
			Total work hours	151	0	0	0	13	13	14	20	15	7	14	9	15	10	13	3	5

Version Control Plan

GitHub will be our team's primary method of version control. Once work begins on the design of the data lake, our group will make a GitHub account that stores an open-access version of our project that we can update as development continues.