Ayan Ashkenov
CS779: Advanced Database Management
02/09/2023

Project Proposal

What is the advanced database area you are focusing on related to this course?

I plan on exploring the data lake platforms. Initially, I was looking to work in the Amazon Web Services' ecosystem, specifically with its Lake Formation platform. However, after you mentioned data lakehouses as a potential area of interest, I decided to continue with that instead.

After your suggestion, I consulted a paper co-authored by Databrick's creators, UC Berkeley, and Stanford University on the improvements that lakehouses bring to the table versus traditional data lakes[1]. According to the paper, lakehouses combine data lake and data warehouse technologies into one, allowing for broader use cases across data-related tasks (such as BI analytics and Machine Learning). I found this idea very interesting and believe that it will be a more challenging task.

I plan on using Databrick's Lakehouse service for this project.

What is the proof of concept component in your project?

While data lakes are not exactly mainstream, many companies across the world utilize the technology for data warehousing purposes. Searching for the term "data lake" on LinkedIn Jobs in the United States returns 8,065 results (as of 02/09/2023). Furthermore, "databricks" returns 83,957 job results, with companies often specifying that they use Databrick's Lakehouse as their cloud data warehouse platform.

This shows that data lakes have a solid presence in the industry and are utilized by a large number of industries. In addition, after consulting Databrick's documentation, it is clear that the platform supports various data sources and extensions, with a broad community of users.

What are some of the goals that you plan to learn in this project?

I plan on familiarizing myself with data lake technologies and working with data of varying degrees of structure (structure, semi-structured, and unstructured data). I also hope to expand my technology stack and gain a deeper understanding of ETL and data pipelines.

[1] https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

What skills are you bringing from other courses, and what is the new element that you are learning in this class that's related to advanced database management?

One of the planned data sources for my lakehouse is a relational database (likely, SQL Server) which I have professional experience in using and which was covered in last semester's data warehousing class (CS689). In addition, one of the planned outlets for the lakehouse is a Tableau dashboard(s), a tool that was used in CS689 as well.

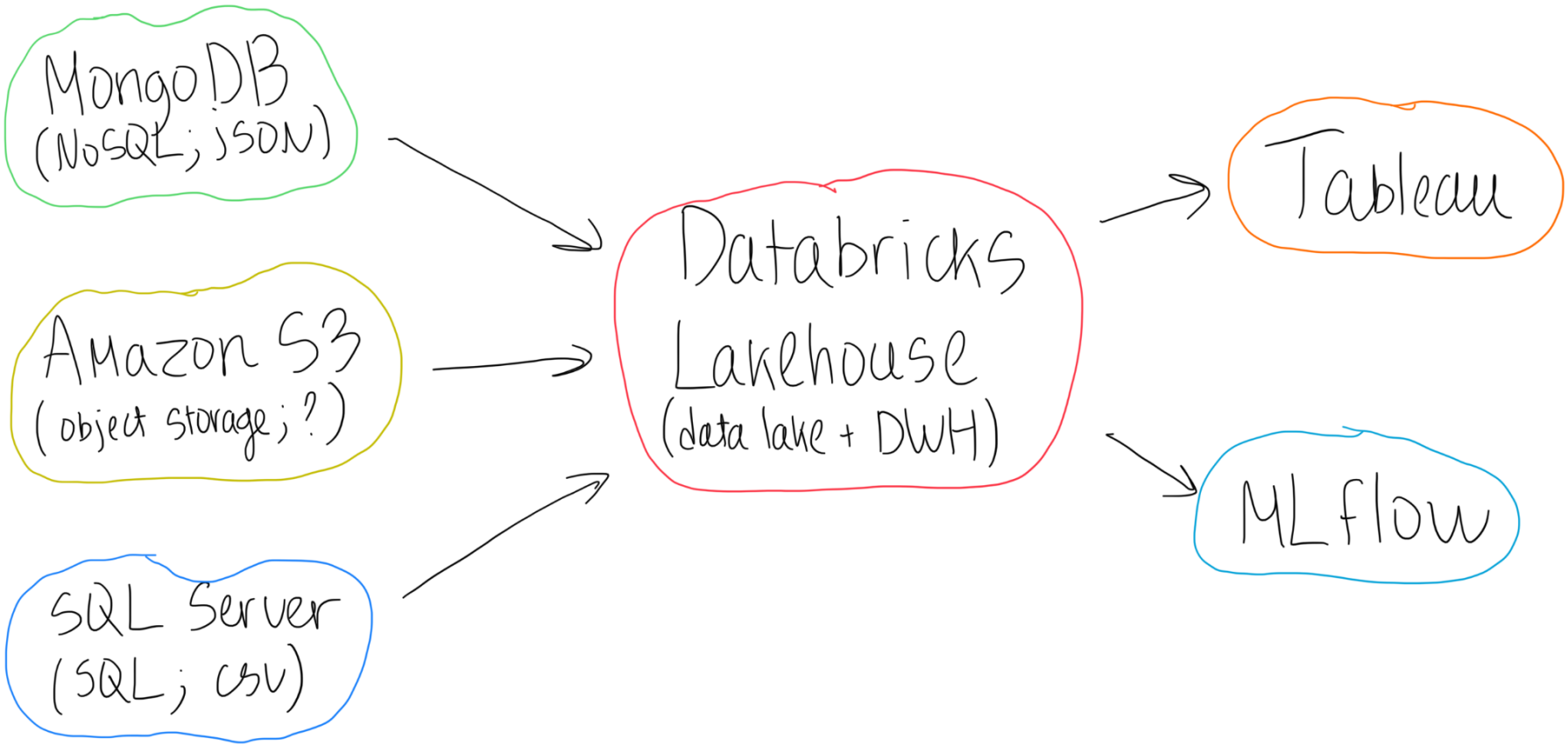
Currently, I plan on using MongoDB as one of my data sources for the lakehouse and if I am not mistaken, it will be covered later in the semester as part of the NoSQL theme.

What data are you looking to use specifically?

At the moment, I do not have specific data in mind. However, given that I plan on using various data sources for the lakehouse, my data will certainly be in structured (from RDBMS using SQL), semi-structured (JSON from an API), and unstructured formats (likely images in JPEG/PNG formats).

*The image below summarizes my current scheme for the lakehouse. I plan on having 3 data sources: MongoDB, Amazon S3, and SQL Server. Data from these sources will populate the lakehouse. Then, a Tableau dashboard will be made from the data inside and other data will be used with MLflow for a sample machine learning model (since the objective of this project is not machine learning, it will be a simple, template model to illustrate lakehouse capabilities). I will continue to explore more options for incoming data sources and outcoming outlets for the project depending on the situation.

**I will also attach the image below as a separate file.


MongoDB
(NoSQL; JSON)
Amazon S3
(object storage; ?)
SQL Server
(SQL; csv)
Databricks
Lakehouse
(data lake + DWH)
Tableau
MLflow