CMSC 608 HW 1

Justin Jones

2025-02-02

Table of contents

1. Storage Systems

Storage Systems are tools and platforms that help store structured data. These platforms provide tools to define the structure/schema for storing data, inserting/altering data, and querying data. Examples include Relational database systems, and No-SQL systems.

2. Data Lake Platform

Data lakes are tools and platforms for storing unstructured data. These platforms often offer tools for querying and transforming the data. Examples include distributed file systems and open table format systems

3. Data Integration

Data integration tools are designed to integrate data from multiple sources into one unified dataset. Examples include Airbyte, Kafka, and Redpanda.

4. Data Processing & Computation

Data processing and computation platforms are tools that allow data engineers to perform transformations and computations on large datasets. Examples include Apache Spark, hadoop and PySpark

5. Workflow Management & DataOps

Workflow management platforms provide tools to manage tasks, projects, and timelines specifically for data engineering tasks. Examples include Airflow, Datafold, and lakeFS

6. Data Infrastructure & Monitoring

Data infrastructure and monitoring systems provide tools to deploy and monitor storage systems and data lake systems, and to monitor their performance and resource usage. Examples include Docker, Kubernetes, and hadoop

7. ML Platforms

Machine Learning Platforms provide tools to create and test machine learning models using the data to make predictions and inferences. Examples include mlflow, LanceDB and chroma.

8. Metadata Management

Metadata management platforms provide tools that can view, alter, and transform the metadata of datasets. Examples include ApacheAtlas, Metastore, and Openlineage.

9. Analytics & Visualization

Analytics and Visualization platforms provide tools to generate analytics, visualizations, and dashboards to view and interact with the data. Examples include Streamlit, jupyter, and Apache Drill

More specific categories

Unified Processing

Unified processing data processing and computation tools are software solutions that enable the processing of various types of data, including batch, streaming, and real-time data, within a single framework. These tools offer a unified approach to data processing, eliminating the need for separate systems for different data types. One example includes Apache Spark. These tools are important because they provide a unified tool that can process large datasets using whatever method is needed. I am interested in these tools since they enable large-scale processing of large datasets

Resource Scheduling

Resource scheduling, data infrastructure, and monitoring tools are essential for managing and optimizing complex systems, especially in data-intensive environments. They help ensure efficient resource utilization, maintain system stability, and provide insights into performance and potential issues. One example includes Docker. These tools are important because they assist with deploying databases and hosting them allowing teams to access them from any device. I am interested in these tools since I consider this to be one of my weaknesses, and I would like to learn more.

ML Ops

MLOps (Machine Learning Operations) tools streamline the machine learning lifecycle, from experimentation and model building to deployment and monitoring. They aim to automate and manage the complexities involved in bringing ML models into production, similar to how DevOps principles apply to software development. These tools address challenges like reproducibility, scalability, and model performance tracking. One example includes MLFlow. These tools are important to maintain the ML pipeline of creating models, testing them, deploying them, collecting more data, and repeating the process. I am interested in these tools because I am interested in improving my skills in and want to make sure that I have the tools to succeed.

Reflection

I appreciated being able to learn more about different parts of the data science field. I feel I was able to cover some of my blind spots. I did not find this assignment to be particularly difficult, since I have experience with many of these tools.