CSP 554 - PROJECT PROPOSAL

Building a Data Pipeline Using MongoDB and AWS Kinesis

- Rajesh Kumar Bandaru (A20446254)

- Mohammed Jawhar(A20449684)

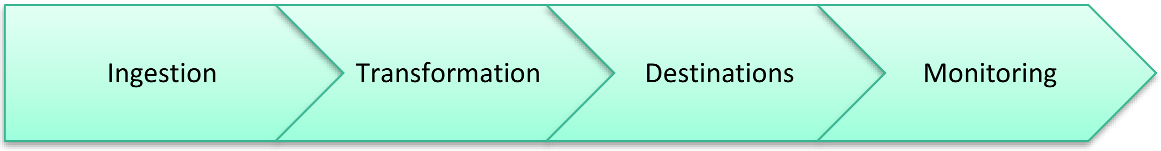
- Sai Charan Akkena(A20456762)

Introduction:

Data pipelines transport raw data from database sources to data warehouses for use by analytics and BI tools. The efficient flow of data from one location to another is one of the most critical operations in today's data driven world. Data flow can be precarious, because there are many things that may go wrong during the transportation. Potential problems can be data getting corrupted, it can hit bottlenecks or data sources may conflict and/or generate duplicates. As the complexity of the requirements grows and the number of data sources multiplies, these problems increase in scale and impact. Here the data pipeline comes into picture which eliminates many manual steps from the process and enables a smooth, automated flow of data from one station to the next. It automates the processes involved in extracting, transforming, combining, validating, and loading data for further analysis and visualizations. It provides end-to-end velocity by eliminating errors and combatting bottlenecks or latency. In short, it is an absolute necessity for today's data driven world.

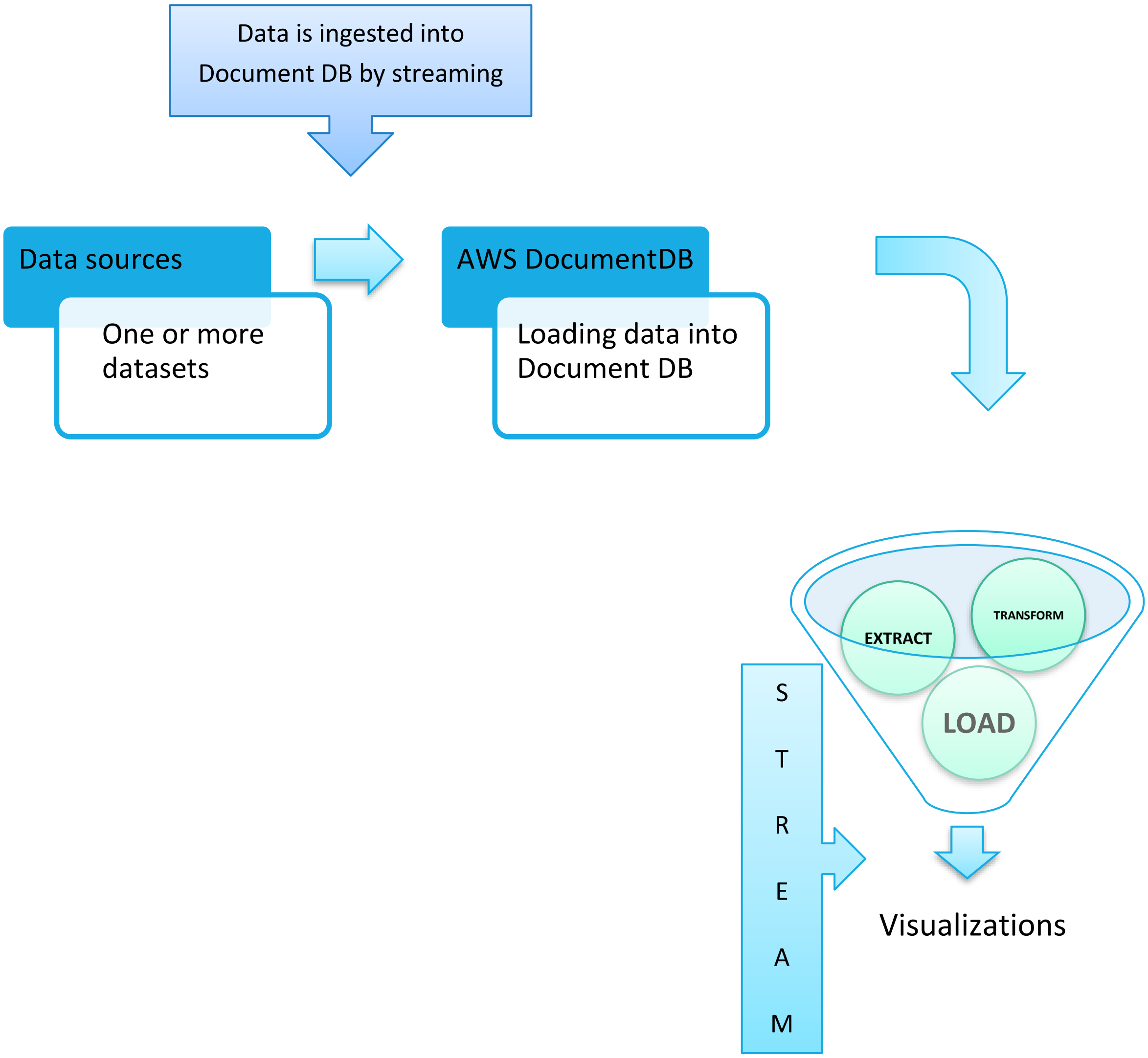
Project Details:

A data pipeline views all data as streaming data and it allows for flexible schemas. The data pipeline does not require the ultimate destination to be a data warehouse. It can route data into another application such as a visualization tool. A typical data pipeline looks as below,



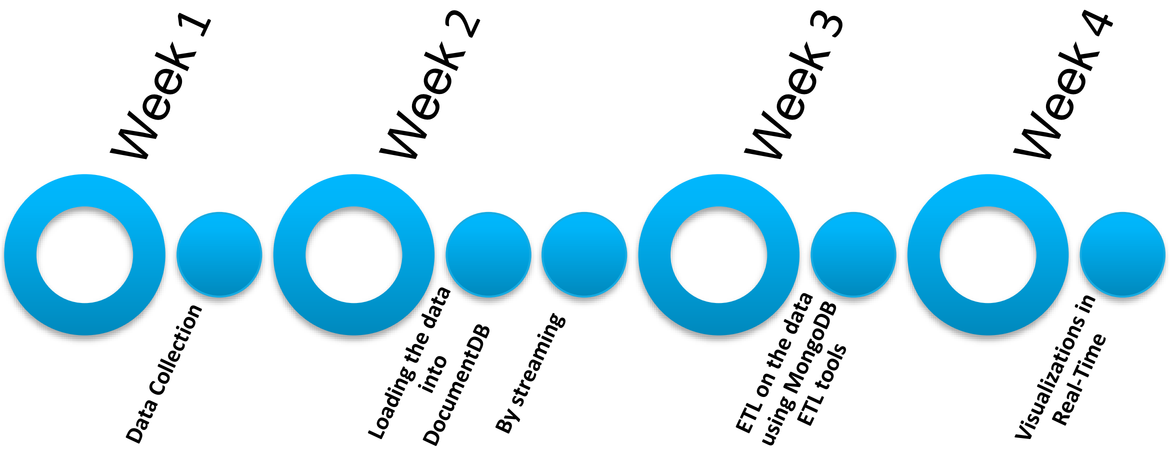
The data is loaded either as batches or through streaming. It is later transformed by using ETL techniques depending on the transformation needs. This transformed data is then sent to destinations and monitored by using logging and alert codes.

We would like to create a pipeline so that the transformed data can be used to get some visualizations so that we can gain some insights from them. The project flow will be as follows.



Data is collected from multiple sources and is sent to DocumentDB by streaming using AWS Kinesis. By using MongoDB's ETL tools, ETL is performed on the data to acquire the necessary information. Finally, the data is streamed to the AWS Quicksight to produce visualizations to gain insights.

Timeline:


Week 1
Week 2
Week 3
Week 4
Data Collection
Loading the data into DocumentDB
By streaming
ETL on the data using MongoDB ETL tools
Visualizations in Real-Time

References:

Understanding Data Pipeline:

https://www.alooma.com/blog/what-is-a-data-pipeline

Tools necessary for ETL:

https://blog.panoply.io/top-9-mongodb-etl-tools

For Streaming the Data*:*

https://medium.com/@yashbindlish1/amazon-kinesis-the-core-of-real-time-streaming-a543085a212f