Capstone Project Data Platform Documentation

Aaron Janaszak, Addison Gangwish, Christian Buonfiglio, Lucas Mueller

Source

1. Online Databases: Data was gathered from the World Health Organization's Global Health Observatory API.
2. Data on Web Pages: Data was gathered from statisticstimes.com and census.gov, with BeautifulSoup used for web scraping.

Ingest

1. Azure Databricks: Used to make API calls, clean data, and route data through Kafka to simulate real-time data gathering.
2. Confluent Kafka: Sends data between Azure Databricks with a delay to simulate real-time data gathering.
3. Azure Data Lake: Stores data as .CSV files after it's been gathered.
4. Azure Blob Storage: Data structure used in Azure Data Lake to allow partitioning of data.
5. Azure Data Factory: Automates the above process – runs Kafka producer and consumer on a schedule, storing consumed data in Azure Data Lake.

Process

1. Azure Databricks: Used to gather data from Data Lake and send to SQL database.
2. SQL Database: Used to store gathered data in relational format.
3. Python in Microsoft PowerBI: Used to create machine-learning model to predict mortality rates.
4. Microsoft PowerBI: Used to create visualizations and presentation from resulting data.