PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

BIG DATA 520: Introduction to Data Engineering

Winter 2024

Final Project Description

BIG DATA 520: Project Description

Overview

> The goal of the project is to build some kind of data pipeline application…
> …using streaming via the sorts of streaming technologies that we've studied in 520
> You can choose from a few suggestions, or come up with something novel of your own
> Keep the "Modern Cloud Data Platform" sketch in mind as you think through data ingestion, storage, processing, re-storage, potential for serving, etc.
> You can work alone or in teams of up to 3 people
> You'll submit a *project proposal assignment* ASAP

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Project Options

Project Options

Option 1: Streaming Financial Data

- A good source of data is ***IEX Cloud*** (https://iexcloud.io/core-data/)
  - IEX cloud provides various types of financial data via REST APIs
  - Offers a free-tier with limited (good enough for our purposes) functionality...
  - ...and an inexpensive paid tier with much more data
- Example use cases to consider for a project:
  - Streaming analysis of stock prices
  - Comparing current data with historical data
  - Correlating behaviors between financial instruments (e.g. stocks and FOREX, or stocks of companies within or across sectors)

Project Options

Option 2: NYC Taxi Cab Data

- ***NYC Taxi Cab Trip Data*** (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page)
  - Much data across various trip and vehicle types
- Data is distributed as CSV files—you can use a (will be provided if you need it, otherwise you could create a stream from CSV using ***Kafka Connect***) simulator that will stream the CSV files as if it was pouring in live
- Example use cases to consider for project:
  - Average ride times or fares to certain points in the city over recent windows, compared to over all of history, actual vs “expected”, etc.

Project Options

Option 3: Weather Data

- OpenWeather (https://openweathermap.org/api) provides a REST API where you can get current and forecast data
- Example use cases:
  - Compare with current conditions for near-time accuracy and alerting, compare to historical data and assess forecasts for accuracy, etc.

Project Options

Option 4: Bitcoin Transaction Analysis

> blockchain.com provides exchange APIs on top of their exchange for free (REST and Websockets API are both available)

> Example use cases: transaction amounts by day/time, exchange rate analysis, combining Bitcoin data with stuff from the IEX Cloud data, figuring out when Razzlekhan is going to drop her next really, really great rap track, assuming that can be done from jail, etc.

Project Options

Option 5: Something Entirely of Your Own Devising!

- Choose some source of data that's either available streaming via some API or format, or where you can convert some downloaded batch data into streaming data
- Goal: should have enough scope to cover a real-world data challenge commensurate to the exertion implied in options 1—4
- If you choose to do a custom project, your ***proposal assignment should include a plan identifying the following***:
  - *What data you'll use*
  - *How you plan to obtain the data*
  - *What sort of intermediate processing you'll do*
  - *How the data can be made available to downstream tools*

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Scope and Deliverables

Scope and Deliverables

Scope

- **Scope:** You should implement a data pipeline including:
  - Ingestion
  - Staging and Processing
  - Storage for Serving
  - Some Analysis
- **Deliverables:**
  - Any code/scripts/notebooks, etc. that you produced
  - An in-class presentation documenting what you did and the experience of doing it

Scope and Deliverables

Your Presentation

- Discuss your data set, your design and the techniques you use
- Describe your use case:
  - The data
  - What sort of business questions you hoped to answer from it
- Describe your architecture, providing:
  - An overall diagram showing how you fleshed out each of the pieces of your project's take on the "Modern Cloud Data Platform"
  - Why you chose the components you did where you did
  - Discuss how data moves through the system
- Challenges and loose ends:
  - What was hard?
  - What would you do with more time?
- Optional: show your system running

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Next Steps…

Next Steps

By Next Class!

> ***Form groups***, and announce their membership, if you're going to work in groups
> Make a (not necessarily binding) ***initial project choice***
> You will submit the above information in the ***Final Project Proposal*** assignment in Canvas