Project 1: Climate Facts

Today it seems that when you look at the news from local to regional to a global scale, the topic of climate change is an energized issue. For many it is a scientific-based concern, for others, there are economic or personal considerations.

Your team has been tasked to create accessible data, to allow others to store and make data accessible.

What you know is the amount of data your system must accept is growing at a phenomenal rate. These data include many forms of data. For instance, raw in situ data, imagery, documents, and video are types of data you should expect.

Concepts:
1. Approaches in isolating, integrating, and testing distributed components
2. Creating management strategies of large data sets in a distributed system
3. Manage frequently requested data using a cache
4. Data will exceed the capacity of a single system (database, server, file storage). As servers reach capacity, data should be rebalanced, routed, or … to allow the system to dynamically grow and move data.
5. Concerns of data blackouts (blocked, or down servers) is a concern as well as, geographically disperse user base.

Role & Background

You are the architect and lead investigator for ClimateFacts a joint venture across multiple organizations and companies.

Your team as stated has been awarded the portion of the project that houses the data archive and request systems for which ClimateFacts will used to launch their data services and web/mobile services.

ClimSpace

The pilot project ClimateFacts is a public access tool for organizations to access current and historical weather data, documents, and supporting information (reports, imagery, video) for use in climate research and planning by universities, private organizations, and the public. For example California Agricultural Initiative will use the to determine policies for water management in Northern California agriculture and household use. Other uses include the ability for small mesonets or individual stations to upload data into the system. This SETI and block-chain inspired feature (upload, acesss, and process data) is a prominent capability of the system. It is expected to include gathering/validating synoptic data and storing/querying the data archive to the participating teams and communities. This system is expected to be widely distributed, not centrally located.

Weather data housing and collection is an ongoing proof of concept task; ClimateFacts has setup a MOU with NWS to receive updates of their weather data. The MOU states updates are pushed to ClimateFacts through a portal owned and managed by many companies (including yours). ClimateFacts will tie into the network to update the climatology data stored in the virtual, peer repository.

Data assumptions:

- Data is pushed to ClimateFacts by NWS. ClimateFacts is responsible for basic proofing of incoming data.
- Data is streamed; it is not delivered on regular intervals or a single data set

Repository assumptions:

- The system should be usable by the public and sponsored research. No system component can be hard-wired to a location or operating-system. Managing a fair method to ensure access to the data is needed.
- Data queries are both temporal and spatial
- This is a long running project and there will be a lot of data - it cannot fit within a single database
- The API to your system must support JAVA and Python. Additional languages are optional.
- Output formats must support at least JSON.
- Different data sources can be used to initialize the data repository
    - I will provide you with sample data sets to help drive scaling tests.

Technical Objectives and Goals

1. The ClimateFacts system is expected to use mock data feeds to demonstrate the processing and ingestion of data.
2. The details on client systems are not available.
3. Plan for blocking and non-blocking data requests.

Deliverables

1. The project is to be completed and submitted +/- 3 days of the due date as noted on the class schedule.
2. Standard deliverables apply – to be provided.