|  |
| --- |
| the APP EXPERTS |
| **PLANTS DATA RETRIEVAL PIPELINE** |
| USING BIG DATA TECHNIQUES TO RETRIEVE AND STORE PLANTS DATA |
|  |
| **SAEED F.** |
| **9/25/2020** |

|  |
| --- |
|  |

Business Case:

* Plant’s growth optimization, identification, and cataloguing

Pipeline purpose:

* To use various plant data such as their nativity, edibility and toxicity, family, specie, genus, and growth characteristics to optimize planting on farms and identifying edible and toxic plants that might affect desired plant’s growth on the farm.

Data Source:

* Trefle.io

Pipeline flow:

* Trefle.io -> Kafka -> Spark -> Hive -> HBase

Pipeline flow justification:

* Trefle.io
  + Trefle is a global plants API with data on 1 million plants including 500 thousand detailed data and 600 thousand synonyms.
  + Their plant data can be used for a range of applications and types of users.
    - Robot: to build gardening apps and farming robots
    - Gardeners and Nature Lovers: to use for efficient planting and living off nature
    - Researchers and Students: to use to learn from data from thousands of plant species, types, characteristics, and attributes
  + Their plant data were collected from a number or reliable sources:
    - USDA (United States Department of Agriculture)
    - Tela Botanica
    - Tropicos
    - IPNI (International Plant Name Index)
    - Royal Botanic Gardens, Kew
    - WikiMedia
    - Muséum National d'Histoire Naturelle (aka, National Museum of Natural History, France)
    - And manually from people all around the world
  + And finally, the pricing is free for all as it is in its beta stage as of 25th September 2020.
* Kafka
  + Kafka is a streaming software that is very reliable in streaming text data from a web Rest API.
  + And it was chosen due to its easy integration with the next tool: Spark
* Spark
  + Spark is a distributed processing system that can be used for big data workloads. It utilizes in-memory caching so it is very fast and is a good engine for large-scale data processing.
  + It also allows for robust data transformation for both structured and unstructured data
  + It can tap into a kafka stream to retrieve streamed data
* Hive
  + Hive is a data warehouse software built on top of hadoop for providing data query and analysis
  + It can be used to store each streamed data as it is received, processed, and transformed by spark
* HBase
  + HBase is a non-relational distributed database that is built on top of hadoop and can be used to give hadoop Bigtable-like capabilities.
  + It can integrate easily with a hive table and share table values
* Hortonworks
  + Hortonworks sandbox is a preconfigured cloudera Bigdata environment that can be used for big data retrieval, processing, and storage.

Plant Data columns:

* Plant Name
* Plant Specie
* Plant Genus
* Plant Family
* Plant Nativity
* Plant Edibility
* Plant Toxicity
* Plant Growth Form
* Plant Growth Habit
* Plant Growth Rate
* Plant Growth Description
* Plant Growth Sowing
* Plant Growth Row Spacing
* Plant Growth Average Height
* Plant Growth Maximum Height
* Plant Growth Maximum and Minimum ph
* Plant Growth Maximum and Minimum Temperature
* Plant Growth Maximum and Minimum Root depth
* Plant Growth Soil Nutrients
* And more….