**EBA5006 Big Data Engineering for Analytics**

**Project Proposal**

|  |  |
| --- | --- |
| **Project Title** | Food Recipes Relational Analysis and Recommendation System |
| **Project Members** | Koh Zhi Rong Chester  Low Jia Sheng Nelson  Ong Kian Eng  Zhao, Zhen  Zhou Jingyu |
| **Overview** | Describe the context and the business problem solved by the system to be secured:  **Context and purpose of study:**  With the advent of social sharing platforms, recipes and nutrition tips are being shared across the Internet. This provides an opportunity to stage and recommend dishes to changing food preferences such as:   * Health-consciousness and special diet (e.g. vegan) * Food allergies * Growing ageing population (with a significant proportion having long-term metabolic diseases, hence requiring delicate balance of diet). * Desire to try innovative food pairings and fusion recipes for gastronomes   **Business Problem:**   1. Recipes and nutrition tips across the Internet are generally isolated and it is difficult for people to assemble information to cater to their own dietary wants and needs. 2. Make use of graph analysis to study ingredients and potentially figure out relationships between cuisines/ dietary requirements in order to recommend food pairing (e.g. based on nutritional content such as low calories using graph mining)   Nodes: Recipes and meta data (e.g. ingredients, nutritional content and allergen)  Edges: The relationship between recipes and ingredients / nutritional content. Similarity matrix between recipes based on ingredients or nutritional content can be computed in order to provide recommendation.  **Potential extensions:**   * Find out new links between ingredients to suggest new fusion dishes (possible to also consider allergy information) * Find new association of food for recommendation for purchase or consumption (e.g. for market basket analysis) * Find out availability of food items (e.g. using IoT devices to monitor inventory) to reduce food waste, or for cooking or supply chain management in restaurants, catering business, hospitals, monitoring food security through import and export at country level, food distribution at global level) * Capture the food trend on social media to enrich the food recommendation |
| **General Architecture** | Describe the general architecture of the data engineering solution proposed. This serves as the baseline for the scope of work  **Sandbox in DataBricks:**  Store in RDBMS  Data cleaning  and ensure consistent data  structure via Python  Temporary storage  (Repository)  CSV, XLS, JSON  in Cloud  Web scrapping from various webpages via Python  API calls  via Python  Building of graph via Python  Retrieval of information via Python  Data ingestion  Data Source  Data Preparation  Data processing  Client query  Use of Python through Jupyter notebook  **Production in Virtual Machine:**  Store in HBase via Sqoop  Data cleaning  and ensure consistent data  structure via Spark  Temporary storage  (Repository)  CSV, XLS, JSON  in HDFS  Web scrapping from various webpages via Python  API calls  via Python  Building of graph via neo4j / Spark GraphFrame\*  Retrieval of information via Cypher / GraphQL\*  Data ingestion  Data Source  Data Preparation  Data processing  Client query  \*Depends on the size of the dataset   |  |  | | --- | --- | | **Examples of static data from webpages:**   * <http://www.foodstandards.gov.au/science/monitoringnutrients/ausnut/ausnutdatafiles/Pages/foodnutrient.aspx> * <http://www7.slv.se/SokNaringsinnehall/> * <http://www.nutrisurvey.de/> * <http://www.eurofir.org/food-information/food-composition-databases/> * <http://www.fao.org/infoods/infoods/tables-and-databases/faoinfoods-databases/en/> * <https://eightportions.com/datasets/Recipes/#fn:1> * <https://archive.org/download/recipes-en-201706/?utm_source=share&utm_medium=ios_app&utm_name=iossmf> * <https://www.food2fork.com/about/api> * <https://spoonacular.com/food-api/pricing> * <https://chompthis.com/api/> * <https://github.com/topics/cooking> * <https://thecookbook.herokuapp.com/> * <https://airtable.com/universe/expHZcS7kWEyq5gUH/recipe-database?explore=true> * https://opendata.stackexchange.com/questions/4283/open-downloadable-recipe-database/4286 * <https://forum.kodi.tv/showthread.php?tid=282387> * <https://www.esha.com/resources/additional-databases/> * <https://www.nutritionvalue.org/> | **Examples of API calls:**   * <https://developer.edamam.com/edamam-recipe-api-demo> * <https://www.nutritionix.com/business/api> * <https://rapidapi.com/collection/food-apis> * <https://www.food2fork.com/about/api> * <https://spoonacular.com/food-api/pricing> | |
| **Scope of Work** | The followings are the key aspects of analytics the project needs to focus:  1. Groups must identify static or real time data available to the computing cluster using an appropriate Ingestion tool.  2. Groups must identify and implement appropriate storage strategy.  3. Groups must include Graph or Web or IoT Analytics as part of their processing framework.  Justify any major omissions.  Explain how the scope allows you to sufficiently demonstrate the mastery of the course module.   1. By leveraging on the webserver function (e.g. Airflow and Oozie), we can schedule routine webscrapping job to retrieve static data from databases. API calls can also be used to retrieve data. As the data will come from different sources, this will be stored in the native file format in a repository (temporary file storage) in HDFS.     Sample Staging: We will extract label, ingredients, healthLabels, dietLabels, cooking steps from the above API call.   1. Using Oozie, we can schedule the reading, cleaning and transferring of data from HDFS into GraphFrame via sqoop for high-speed data transfer. 2. As there are many recipes and ingredients present, and both contain many attributes (also considering that data from different sources comes in different data formats), this data will be best represented in the form of graph. Graph allows us to see context of relationship between entities. 3. As the repository grows with more recipes and ingredients added, the computation to find the optimal connection between food items using graph analytics will require a lot of computational power, hence this can only be resolved using a Big Data infrastructure and approach. 4. The different sources of the project will mainly come from the API and the web. As the API only provide meta data of the recipe, we will have to reach out to a search engine or webpage to obtain know-how recipe for the user to consider. |
| **Effort Estimates** | List the rough listing of tasks and their estimated efforts. This is to ensure that you have thought through the approach and the implementation plan.   1. Obtain API (including writing to respective organisations to require access) and EDA on received data set – 5 man-days 2. Sandboxing with small dataset in DataBricks using neo4j / R / Python to determine staging code – 5 man-days 3. Establish business uplift with data/graph analytic tools with sandbox data – 5 man-days 4. Learn and understand available tools like Airflow, HDFS, sqoop, HBase, Oozie, Spark, GraphFrames, GraphQL – 10 man days 5. Connect the different services together and scale up analytics – 10 man days 6. Optimisation – 5 man days |