Project Plan - Full Contact

Prepared by Remote Contact on November 2, 2020

This project will be implemented using an Agile methodology, meaning, in incremental, iterative development cycles. As a result, dates and tasks are a rough estimate and will be modified based on progress after each of our sprints (~2 weeks).

Features:

- Spark jobs for each output from the Identigraph
  - Roughly 6 different output types, so 6 different spark jobs
- API to handle submitting a spark job to an EMR cluster
  - Learn about the technology, Apache Airflow
  - Basic HTTP endpoints
  - Allow user to query state of their job
- Database
  - Stateful store of current and past query reports
- Web-hosted UI
  - Input page
  - Query Information page
  - Dropdowns for each Identibase job output
  - Delete past query reports

Timeline:

Sprint 1: Become Familiar with FullContact's Identigraph and data (10/25/2020 - 11/7/2020)

- Explore data and graph through different queries at different points of the graph. - All team members
- Build a gradle project as a base for future spark jobs. - All team members

Sprint 2-3: Write Spark jobs to track inputs and their effects on the graph (11/8/2020 - 12/5/2020)

- Write a spark job for each output of the Identigraph job
  - RawObservations - Liam McCarthy
  - ObservationEdges - Joshua Hamel
  - CrudeEdges - Valyria McFarland
  - RefinedVertexes - Robert Sarno
  - RefinedEdges - Jonathan Bluhm
  - EdgeWithPropogatedNegatives - Neil Borde

Sprint 4-5: (1/14/2021 - 2/27/2021)

- Develop an application that returns insights like the path of the data through the graph based on input records - Valyria, Liam, Neil

- Add Functionality to the app so it returns insights based on output clusters as well- Robert, Johnny, Josh

Sprint 6-7: Create a web based, interactive UI (2/28/2021 - TBD)

- Build a UI to accept bidirectional user queries - Joshua Hamel
- Display output in user friendly format - Liam McCarthy
- Query response should be generated in minutes - Neil Borde
- Host UI on AWS - Robert Sarno
- Build database to store old queries - Jonathan Bluhm, Valyria McFarland