PROPOSAL

To Develop an

Enterprise Scale Disease Modeling Web Portal

For Ascel Bio

Updated March 2015

TABLE OF CONTENTS

1. INTRODUCTION

The business mission of Ascel Bio is to deliver disease forecasts and warnings to doctors and a wide range of other users, so that they can save lives and cut the costs of care. Ascel Bio desires to scale up their disease forecasting corporate capability, establishing a software as a service (SaaS) delivery model. Ascel Bio has requested that the Applied Innovation Center for Advanced Analytics (AIC) at DRI generate a proposal to support them in constructing a responsive enterprise software and systems solution to accomplish this business plan.

1.1 PROJECT OVERVIEW

Ascel Bio would like to reduce the time and effort necessary for customers to use their software to generate disease forecasts. Currently customers must contact Ascel Bio to obtain a licensed copy of the modeling application. Next, the customer installs the software on their desktop computer, loads data into the application and models it. When Ascel Bio releases a software update, all application users must be notified and sent the new version. Besides logistical issues of delivering the software (and updates) to the customers, desktop applications rely on the customer's computer for their execution environment. Therefore, model complexity must be kept to a minimum to ensure that results are produced in a reasonable amount of time. In addition, the amount of data the model can process is limited by the computational power of the customer's computer.

As a public service, Ascel Bio would like to publish selected disease forecasts on their website. These forecasts would help publicize Ascel Bio's technology as well as educate and inform the public of possible disease outbreaks in their local community and other geographical locations.

1.2 BACKGROUND

Ascel Bio has followed the lean startup methodology and developed a minimum viable product (MVP) with two software components: Delphi and Exigence. Delphi and Exigence were developed in Microsoft Access using Visual Basic for Applications (VBA). Access provides an environment where applications requiring a database and user interface can quickly be developed. Microsoft Access is not designed to handle large amounts of data or heavy computation.

Delphi generates disease forecasts from disease case count data or from data generated by Exigence. There are a few parameters that the user can modify in Delphi that will alter model forecasts. A typical Delphi modeling workflow:

1. User launches Delphi on their desktop computer.
2. User imports the dataset/s they would like to model.
3. User runs model.
4. Delphi generates a Microsoft Excel file containing the disease forecast charts.

Exigence analyses text and estimates the probability that a disease is causing significant social disruption. Fatalities, people fleeing the outbreak area, and isolation wards are some of the indicators that Exigence brings to the attention of the analyst. Exigence can also be used to generate a file that Delphi can use to forecast a disease outbreak. A typical Exigence modeling workflow:

1. User launches Exigence on their desktop computer.
2. Exigence periodically extracts text from articles posted on select RSS Feeds.
3. User runs model.
4. Exigence generates a Microsoft Excel file containing a list of social disruption indicators and their probabilities.

1.3 OBJECTIVE

Ascel Bio would like to move to a software as a service (SaaS) delivery model. SaaS would bring a number of benefits besides eliminating the need for customers to install software. Computations would be done on Ascel Bio servers, allowing more complex models to be developed and executed against larger datasets than is currently possible on the customer's computer. Delphi and Exigence were built using VBA, which is based on a software development paradigm incompatible with SaaS. Therefore, only the algorithms from the existing Delphi and Exigence applications can be reused in the new SaaS system.

The new SaaS web portal will allow customers to quickly purchase an account that gives them access to the epidemic forecast statistics corresponding to their location of interest. These locations can be their local city or other cities that they subscribed too.

A completely new Ascel Bio website will be developed to support the marketing of the new modeling portal. The main public landing page will provide a selected set of disease forecasts that are based on non-proprietary health data. Disease prevention tips and news articles will also be available on the website. The main portal will display the disease forecasts for only a few of the diseases. These selected diseases will be changed periodically based on Ascel Bio's interest. A subscription page, accessed from the main public landing page, will provide disease forecasts for a wider variety of diseases, at the subscriber's request, for specific targeted locations.

2. PROPOSED TECHNICAL APPROACH

Ascel Bio is a small, but growing startup. The data analysis system must therefore allow the on-demand scalability of computational resources and data storage capacity in both the short and long term. The system must be able to simultaneously run a large number of different types of models, such as machine learning and standard statistical models. In addition, adding, removing, and modifying the existing models should be supported.

We propose to meet the above requirements by developing a data analysis system based on tools already available in the Big Data ecosystem. The Big Data paradigm allows for horizontal scaling of both data storage and computational capacity using commodity machines. The paradigm is inherently modular, allowing models to be quickly added and removed from a deployed system. The costs associated with software licensing will also be minimal as most Big Data software are open source and allow free commercial usage. The Big Data stack will be responsible for data ingestion, storage, and model execution. A website development stack will also be utilized to develop the web portal. Similar to Big Data software, there is an abundance of free, open source licensed webservers and web frameworks.

As requested, system development will be comprised of two separate phases: Phase 1 and Phase 2. Each of these phases will add more functionality to the pre-existing system. A phase will consist of the four major steps: design, development, test, and production deployment. Upon the acceptance of this proposal, a detailed project plan will be created that specifies the tasks within each of those steps.

System maintenance and support are not covered under this proposal. Nor are the costs of hosting the system. If invited to, a subsequent proposal covering system hosting, maintenance and support will be created.

2.1 PHASE 1 REQUIREMENTS

The initial design and build out of the hardware, and software infrastructure used in all subsequent phases will be the main objective of Phase 1. Therefore, the design will take into account features that will be added in later phases. At the end of this phase there will be a basic Ascel Bio administration interface that allows the management of the public forecasting webpage, the public forecasting webpage itself, and all the required hardware and software that support those.

Storage Subsystem:

- Data ingestion must be flexible and allow the implementation of different data sources including; CVS files, XML Files, and REST APIs.
- Data ingestion pipeline must allow data to be directed to the text analysis subsystem.
- Deployed system will have a minimum storage capacity of 4 Terabytes.
- Additional storage capacity, up to 200 Terabytes, should be supported.
- Data stored in the system is immutable, i.e. data can be written, read and deleted, but not modified.

Forecasting Subsystem:

- The development, addition, and removal of forecasting models must be supported.
- The addition of computational resources to increase the number of simultaneous models runs should be possible.
- Mean and standard deviation summary statistics based on data groupings must execute within a reasonable amount of time on large (100 Gigabyte) datasets.
- A single forecasting model run should execute within a reasonable amount of time on large (100 Gigabyte) datasets.

Text Analysis Subsystem:

- Text indexing and searching must be scalable through the addition of computational resources.
- Text queries will consist of strings containing Boolean logic operators and case sensitive text/phrases and return a list of matching records.
- Raw text must be archived in case the index needs to be rebuilt.
- Text data will be accessed from news and digital content providers:
    - Factiva (http://new.dowjones.com/products/factiva/) or
    - Moreover (http://www.moreover.com/).
- Norsys Software Corp's Netica Bayesian Belief Network API must be executable within the framework.

Two of the three web portals will be implemented in this phase. The Administration Web Portal will be used to update dynamic web content displayed on the Public Forecasting Portal. It will also be used to schedule model runs and allow the administrator to upload data. The Public Forecasting Portal will feature ranked disease forecasts and informational articles. It will act as a functional demo of what will be available in the Subscription Based Portal, which will be implemented in Phase 2.

Public Forecasting Web Portal:

- Show articles related to displayed disease forecasts.
- A times series forecast for a disease will be a composite chart composed of a bar graph, scatter plot, and line plot overlaid on top of each other.
- IDIS Trends will be displayed as gradient bars.
- Disease forecasts and IDIS Trends will be generated dynamically based on the output of the Text Analysis and Forecasting Subsystems.
- Disease forecasts will be available for only a few of the diseases. Administrator will change the selected disease periodically.

Administration Web Portal:

- Add/remove informational articles displayed on the Public Forecasting Web Portal.
- Data upload via csv files located on admin computer must be supported.
- Schedule the execution of the public forecasting and public IDIS models.

- Publish forecasts and IDIS Trends to Public Forecasting Web Portal.

2.1.1 Deliverables

The primary deliverable of Phase 1 will be a production system deployed on Ascel Bio's cloud infrastructure. As part of the deployment process, Ascel Bio personnel will be trained on how to use and administer the system. Because this is a full system deployment, the training will focus on DRI developed system components. Specific deliverables of this phase will include:

- Detailed technical specifications.
- Functioning system deployed on Ascel Bio's cloud infrastructure.
- Administration and Public Forecasting Portal documentation.
- Non-technical training on how to interact with the system.
- Technical training on how to add models to the system.

2.2 PHASE 2 REQUIREMENTS

Phase 2 will be built upon the software and hardware infrastructure developed in Phase 1 by adding additional features to the website and Text Analysis Subsystem. A Subscription Based Web Portal will be created that allows customers to purchase a subscription to the web portal for looking into various disease forecasts for the location or locations of subscriber's interest.

Storage Subsystem:

- Data at rest or in motion must be encrypted.
- Deployed system will have a minimum storage capacity of 10 Terabytes.

Subscription Based Web Portal:

- User/group portal access control should be manageable from the administration interface.
- User/group should be able to purchase, edit, and delete subscriptions.

Administration Web Portal:

- Customizable per customer/group data storage limits.
- Add/remove/group customer accounts.
- Access to web analytics platform used throughout site to track the usage.
- Add/remove informational articles displayed in the Subscription Based Web Portal.

2.2.1 Deliverables

The primary deliverable of Phase 2 is an updated system containing the new functionality that DRI will apply to Ascel Bio's production system. Similar to Phase 1, Ascel Bio personnel will be trained on how to use the system. Training will focus on the features added in Phase 2. Specific deliverables of this phase will include:

- Updated detailed technical specifications.

- Deployed updated system on Ascel Bio's cloud infrastructure.
- Updated Administration and Public Forecasting Portal documentation reflecting the interface changes.
- Subscription Based Web Portal documentation.
- Linkage to payment gateway, such as PayPal
- Non-technical training on how to interact with the system.
- Technical training on the new features.

2.3 ARCHITECTURE DESIGN

Ascel Bio requires a storage and analysis infrastructure that can meet both current and future needs. In general, analysis infrastructures based on relational databases (RDBMSs) are not easily scaled. Therefore, the Big Data software stack was specifically developed with scalability in mind. It allows the horizontal scaling of both storage and computational resources. Upon acceptance of this proposal, a more detailed specification will be developed that may result in modifying this architectural design.

The Ascel Bio website will be comprised of a Public Forecasting Portal, Administration Portal, and a Subscription Based Portal. Those portals will be built using a Content Management System (CMS). Only after a more detailed website specification has been created can the most appropriate CMS be selected. Like most web frameworks, the CMS requires a HTTP Server and a relational database to function.

The storage and analysis infrastructure is comprised of a text analysis, forecasting, and storage subsystem. Hadoop provides access to HDFS, a distributed fault tolerate file system. By replicating data across multiple servers, HDFS provides both horizontal storage scalability as well as fault tolerance. New machines can be added to a Hadoop cluster as needed. The system will then start replicating data to them. HDFS only provides a storage infrastructure, it does not provide the ability to read, write, and delete individual records, which is commonly done in relational databases. Apache HBase adds that functionality to Hadoop and the HDFS file system.

Apache Flume will handle transferring data to the storage and analysis system. Flume allows the creation of fault tolerant data pipelines. The text analysis subsystem requires incoming text data to be indexed and simultaneously stored. By creating a branched pipeline, the same data can be sent to Apache Solr for indexing and HBase for storage.

Apache Solr provides a Boolean query language that can be used to search indexed text. The new Exigence rule engine will take a list of queries and send them to Solr for processing. Those results will then be used by the Netica Bayesian Belief Network to generate IDIS Trends.

Most statistical languages like R, are not designed to process large datasets. The Spark framework allows a single statistical model run to be executed across multiple computing nodes. Distributing the model execution across multiple nodes will reduce the execution time. As the analyzed datasets grow, the number of computational nodes used for a single run can be increased.

Ascel Bio will ensure that the received medical health data does not contain any patient specific information. In absence of patient identification, we will not require HIPPA level security and periodic audit. In addition, AIC recommends that an independent consultant (https://clearwatercompliance.com/) be hired to develop operating procedures and perform a system review to ensure that the received data do not contain patient identity information.

2.4 IMPLEMENTATION

Phases 1 and 2 will require their own detailed testing and deployment plans. System development and testing will occur on DRI hardware. Once completed, the system will be deployed on the Ascel Bio's cloud computing infrastructure. Another round of testing will be undertaken to verify the behavior of the deployed system.

Upon acceptance of this proposal, the method of hosting the production level system will need to be agreed upon. The production system-hosting environment will determine the type of visualization infrastructure the development system will use. To ensure a smooth deployment, a mock deployment will be done at the start of Phase 1.

The deployment process will require close collaboration between DRI and Ascel Bio. During the process Ascel Bio engineers will be trained on how to administer the system. Production system administration will be the responsibility of the Ascel Bio engineering team. Once the production system has been deployed, DRI will also provide end user training to non-technical members of Ascel Bio team.

2.5 PROJECT CHALLENGES AND RISK MANAGEMENT

There are a number of unknowns that should be addressed as early as possible to reduce the likelihood of project schedule slippage:

- Business Intelligence (BI) software would reduce the effort needed to develop the Subscription Based Web Portal. Most BI vendors charge per user or per server CPU. Before the Subscription Based Web Portal is designed, the decision about using (or not using) a BI solution will need to be made. The Public Forecasting Web Portal could also utilize a BI tool, but it is less critical as portal functionality will be limited as compared to the Subscription Portal.
- The hosting environment selected by Ascel Bio, must be compatible with DRI's development environment. Some cloud vendors, like Amazon, have custom implementations of the Big Data software stack that must be used in order to run the software on their cloud. While other hosting services like RackSpace give the ability to rent a bare-metal computer and install whichever software is required.
- It is strongly recommended to hire an independent consultant with experience in developing HIPPA compliance or HIPPA independent system. It will transfer the accountability of HIPPA audit failure from both Ascel-Bio and DRI towards the HIPPA consultant.