A Report
On
Web Application for Machine Learning and Text Mining

BY

Loukic Gowru **2012A7PS058G**

Prepared in partial fulfillment of the

Practice School-II Course at

A Practice School Station of

Birla Institute of Technology and Science, Pilani

April 2016

A Report
On
Web Application for Machine Learning and Text Mining

BY

Loukic Gowru **2012A7PS058G**

Prepared in partial fulfillment of the

Practice School-II Course at

A Practice School Station of

Birla Institute of Technology and Science, Pilani

April 2016

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN)

Practice School Division

Station: Genpact India **Centre:** Bangalore

Duration: 6 Months **Date of Start:** 19th January, 2016

Date of Submission: 6th April, 2016

Title of the Project: Web Application for Machine Learning and Text Mining

Name: Loukic Gowru **ID No:** 2012A7PS058G

Discipline: B.E. (Hons) Computer Science

Name and Designation of the mentor: Mr. Shrikant Keshav, Assistant Manager

Name of the PS Faculty: Mr. Sandeep Kayastha

Key Words: IPIE, AWB, Console
Project Areas: Spark Framework, R python, Java, Web Development, Apache tomcat, Machine Learning, Data Mining, Apache Kafka, Hibernate

Abstract: Enterprises need to reimagine their operations in the context of analytics and technology. However, conventional analytics driven by a purely technology perspective has led to sub-optimal results and merely incremental gains. The Genpact Intelligent Operations approach can help enterprises reimagine their operations, leveraging analytics and the latest technology platforms, to realize truly discontinuous gains. The analytics insights that shape Genpact's Intelligent Operations are developed with the Genpact Intelligent Process Insights Engine (IPIE), which applies industry, domain, and process expertise to pull together only the relevant data for purpose-built analytic applications to meet critical business needs. This report familiarizes the user with a web based platform used to solve user-related problems through coding by providing the required software.

Signature of Student Signature of PS Faculty

Date: 6th April 2016

ACKNOWLEDGEMENT

The successful completion of various parts of this project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my heartfelt gratitude to the people who have been instrumental in the successful completion of this project.

I want to start by thanking my mentor, **Mr. Shrikant Keshav**, whose constant supervision was invaluable. His knowledge in this domain and his commitment towards work were always a source of motivation for me. I am deeply indebted to the employees of the development team, particularly **Mr. Neeraj,** for their unending patience and support all along.

I am thankful to **Genpact** for providing me with an opportunity to work on latest tools and technology and alongside such talented people.

I am also thankful to my PS instructor, **Mr. Sandeep Kayastha**, who has always been supportive and always guided us in the right direction. He made our PS experience hassle free, and ensured that no one had any problems.

Lastly, I would like to thank my family and all my friends for their suggestions and constant motivation which were imperative to the successful completion of the project.

Table of Contents

Chapter-1

About Genpact:

Genpact Limited is a multinational business process and information technology services company, domiciled in Bermuda with executive headquarters in New York. The company reported net revenues of $2.1 billion (2013) with more than 65,000 employees (2014), and clients from 25 countries.

Its current clients include more than 100 of the Fortune Global 500. N. V. Tyagarajan, president and CEO of Genpact, has led the company since his appointment in June 2011.Genpact was established in 1997 as a business unit within General Electric. In January 2005, it became an independent company and in August 2007 a publicly traded company

History:

Genpact began in 1997 as a business unit within General Electric. Its charter was to provide business process services to GE's businesses, with the goal of enabling outstanding efficiencies. During the eight years that followed, Genpact began to manage a wide range of processes across GE's financial services and manufacturing businesses.

In January 2005, Genpact became an independent company and began to serve clients outside of GE. The company name, Genpact, is designed to convey the business impact it generates for its clients. In August 2007, Genpact was listed on the NYSE under the symbol 'G'. Since then the

company has grown from 32,000 employees and revenue of US$823 million, to 65,000+ employees and revenues of US$2.1 billion (2013).

Bain Capital became the firm's largest shareholder in November 2012.

Executive Leadership:

On June 17, 2011, NV "Tiger" Tyagarajan became the president and chief executive officer (CEO) of Genpact and was appointed to the Board of Directors.[5] He previously served as chief operating officer of Genpact. He succeeded Pramod Bhasin, who stepped down as CEO and member of the board and became non-executive vice chairman of the company.

Tyagarajan was previously CEO of Genpact from 1999 to 2002, when he led the business through a critical growth phase as a subsidiary of GE. When Genpact became an independent company, he rejoined Genpact from GE Capital U.S. as executive vice president of sales and business development from 2005 to 2009. Thereafter, he took on the role of Genpact's chief operating officer.

Chapter -2

Business Process Outsourcing :

Business process outsourcing (**BPO**) is a subset of outsourcing that involves the contracting of the operations and responsibilities of a specific business process to a third-party service provider. Originally, this was associated with manufacturing firms, such as Coca Cola that outsourced large segments of its supply chain.

BPO is typically categorized into back office outsourcing, which includes internal business functions such as human resources or finance and accounting, and front office outsourcing, which includes customer-related services such as contact center services.

BPO that is contracted outside a company's country is called offshore outsourcing. BPO that is contracted to a company's neighboring (or nearby) country is called near shore outsourcing.

Often the business processes are information technology-based, and are referred to as ITES-BPO, where ITES stands for Information Technology Enabled Service. Knowledge process outsourcing (KPO) and legal process outsourcing (LPO) are some of the sub-segments of business process outsourcing industry.

Benefits and Limitations:

The main advantage of any BPO is the way in which it helps increase a company's flexibility. However, several sources have different ways in which they perceive organizational flexibility. In early 2000s BPO was all about cost efficiency, which allowed a certain level of flexibility at the time. Due to technological advances and changes in the industry (specifically the move to

more service-based rather than product-based contracts), companies who choose to outsource their back-office increasingly look for time flexibility and direct quality control. Business process outsourcing enhances the flexibility of an organization in different ways:

Most services provided by BPO vendors are offered on a fee-for-service basis, using business models such as Remote In-Sourcing or similar software development and outsourcing models. This can help a company to become more flexible by transforming fixed into variable costs. A variable cost structure helps a company responding to changes in required capacity and does not require a company to invest in assets, thereby making the company more flexible.

Another way in which BPO contributes to a company's flexibility is that a company is able to focus on its core competencies, without being burdened by the demands of bureaucratic restraints. Key employees are herewith released from performing non-core or administrative processes and can invest more time and energy in building the firm's core businesses. The key lies in knowing which of the main value drivers to focus on – customer intimacy, product leadership, or operational excellence. Focusing more on one of these drivers may help a company create a competitive edge.

A third way in which BPO increases organizational flexibility is by increasing the speed of business processes. Supply chain management with the effective use of supply chain partners and business process outsourcing increases the speed of several business processes, such as the throughput in the case of a manufacturing company.

Finally, flexibility is seen as a stage in the organizational life cycle: A company can maintain growth goals while avoiding standard business bottlenecks. BPO therefore allows firms to retain their entrepreneurial speed and agility, which they would otherwise sacrifice in order to become

efficient as they expanded. It avoids a premature internal transition from its informal entrepreneurial phase to a more bureaucratic mode of operation.

A company may be able to grow at a faster pace as it will be less constrained by large capital expenditures for people or equipment that may take years to amortize, may become outdated or turn out to be a poor match for the company over time.

Although the above-mentioned arguments favor the view that BPO increases the flexibility of organizations, management needs to be careful with the implementation of it as there are issues, which work against these advantages. Among problems, which arise in practice are: A failure to meet service levels, unclear contractual issues, changing requirements and unforeseen charges, and a dependence on the BPO which reduces flexibility. Consequently, these challenges need to be considered before a company decides to engage in business process outsourcing.

A further issue is that in many cases there is little that differentiates the BPO providers other than size. They often provide similar services, have similar geographic footprints, leverage similar technology stacks, and have similar Quality Improvement approaches.

Introduction

"Data is the new oil!" is an oft-repeated headline in the business world. However, like oil, data has always been available. It's just that useful insights from data have been hard to find, and even harder to operationalize at scale. The benefits of oil did not become broadly available until processes and technologies were invented, and then operationalized at scale to transform and leverage the raw product. For data, it will be the same. For an enterprise to truly leverage value from data, the enterprise must integrate and then operationalize process, analytics, and technology at scale, under an advanced organizational model. Today, the relevant technologies are increasingly available, but most companies fail to reimagine process in light of available technology and analytics, leaving business outcomes merely incrementally improved, if at all.

For oil, the value was in its contribution to higher order products such as a jet plane. For data, its value can be just as significant. Data can be used for basic analysis (refined oil) or advanced analytics (jet fuel). However, without a well-designed plane (advanced organizational model), smart guidance systems (process reimagined in the context of relevant domain expertise), an intelligent pilot (analytically savvy business leaders), and a roadmap to success (analytically informed business strategy), data by itself has limited value. The one major difference between data and oil is that the latter is a shrinking resource while the former is growing at a mind-boggling pace. The amount of data created within the last two years is roughly 90% of all data created since recorded history.

IT departments have leveraged technology over the last 30 years, from data to information to analytics to insights. Our hypothesis, validated by thousands of client interactions, is that without a new process architecture reimagined in the context of analytics and technology, and supported by an advanced organizational model, business leaders will continue to be frustrated at making only incremental gains when they are looking for discontinuous improvements in revenue, productivity, and competitiveness.

The Genpact Approach:

At Genpact, they have taken a completely different approach by helping our clients build *Genpact Intelligent Operations* so that our clients can achieve the discontinuous gains they are looking for, whether the gains are in revenue growth, productivity gains, or competitiveness.

Genpact has developed three new practices to handle data better, they are:

1. Genpact Intelligent Operations and SEP
2. A Systems of Engagement approach to analytics and technology (SE)
3. Genpact Intelligent Process Insights Engine (IPIE)

Scope and Objective of the Project

Leveraging the wealth of knowledge from our deep operations and process heritage, the Genpact IPIE has been built using a Systems of Engagement (SE) approach to develop a process-aware platform that embeds technology and analytics to deliver purpose-built analytics applications that drive significant business outcomes for clients. The Genpact approach begins with the idea of developing intelligent operations for our client. Once we have established the strategy and business outcomes for the intelligent operation at hand, our process experts work with key stakeholders at clients to develop process maps and then information supply chains that document what data is needed from particular systems of record (e.g., ERP, CRM, existing data lakes, external data sources), followed by the transformations that must take place within the supply chains to deliver the specific business outcomes. Metrics and measures are developed simultaneously to help further inform the effectiveness of the intelligent operations upon implementation. The process experts then hand these process and information maps to the data science and applications development teams at Genpact, who build the purpose-built analytical applications for the client. These process-aware applications can now be deployed to deliver consistent results across the enterprise.

The scope of this project is to add various features to the web app. My task is to implement new features according to the requirements given.

Technical Specifications

Following the Systems of Engagement (SE) approach, Genpact has designed the IPIE from the ground up to accelerate the application of data science and information science in traditional and modern architectures to quickly drive value to clients. The platform consists of five core analytics components built on a resilient data management architecture that embraces the latest advances in big data technology, primarily from the open source community, while adopting necessary components based on major advances from the broader enterprise IT industry. The analytic components are the following:

Enterprise data workbench: An analytics workbench designed with the data scientist and the sophisticated business analyst in mind. Here, data scientists or business analysts can develop new analytic insights rapidly through a simple "drag and drop" interface. This is where they can develop advanced analytics models on different data sets assembled in the IPIE applying either custom models or those already provided within the workbench.

Data central: A simple way to compose information supply chains of structured data, from any enterprise system of record, to prepare and serve datasets up to the workbench to build advanced analytical models, or directly to the purpose-built analytical applications. The tool provides built-in data governance and can act as a lightweight ETL tool if necessary. Data central also provides both scheduling and work-flow capabilities for ongoing updates and management of these datasets.

Text central: Similar to data central with a primary focus on handling unstructured information for analytics.

Smart discovery: A platform approach for merging the output from many visualization and reporting tools typically scattered across the enterprise, which allows the business analyst and the data scientist to quickly leverage existing insights from across the enterprise. The component also provides faceted search across enterprise data sources to facilitate exploration

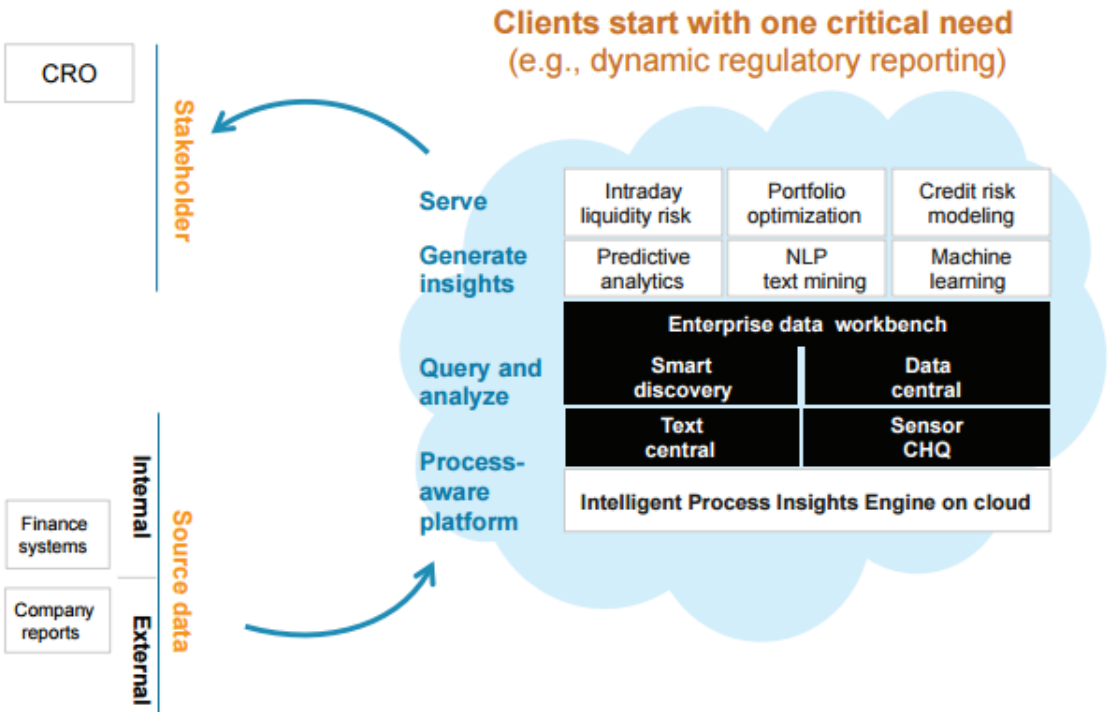
Sensor CHQ: With a view of the explosion of new data sources coming from a highly instrumented world, especially in industrial manufacturing, this component focuses on managing data from these sources to feed into the analytics engine.

Getting started with the Genpact IPIE:

Clients typically get started with a single intelligent operations transformation. A case in point is the challenge faced by the Chief Risk Officer (CRO) in the Banking and Financial Services world. Given the increasing complexity of the regulatory environment and the desire to provide more transparency and access of regulatory bodies to the operations of their businesses, CROs are being tasked with building an enterprise-wide view of risk as it relates to regulatory reporting needs. Whether a CRO has to build robust intraday liquidity risk models or try to comply with the latest Comprehensive Capital Analysis and Review (CCAR) model testing and validation, the root of the challenge is getting all the relevant information and data needed to build these reports quickly and accurately.

Most CROs have started down the IT route to try to rationalize information from existing information supply chains already built to meet different business needs—like finance and accounting, supply chain management, and front-office and back-office banking systems. The effort required to rationalize information provided from these functional systems to the original backend systems of record is gargantuan. Even if the CROs were able to wrestle the system to provide for immediate short-term needs, solutions are brittle, and the rapidly changing regulatory environment is bound to render these solutions obsolete, quickly. To address the CROs' need, Genpact brings together our unique operational knowledge from being part of key risk functions within large enterprise clients with the IPIE and our approach to building intelligent operations to develop the necessary solutions.

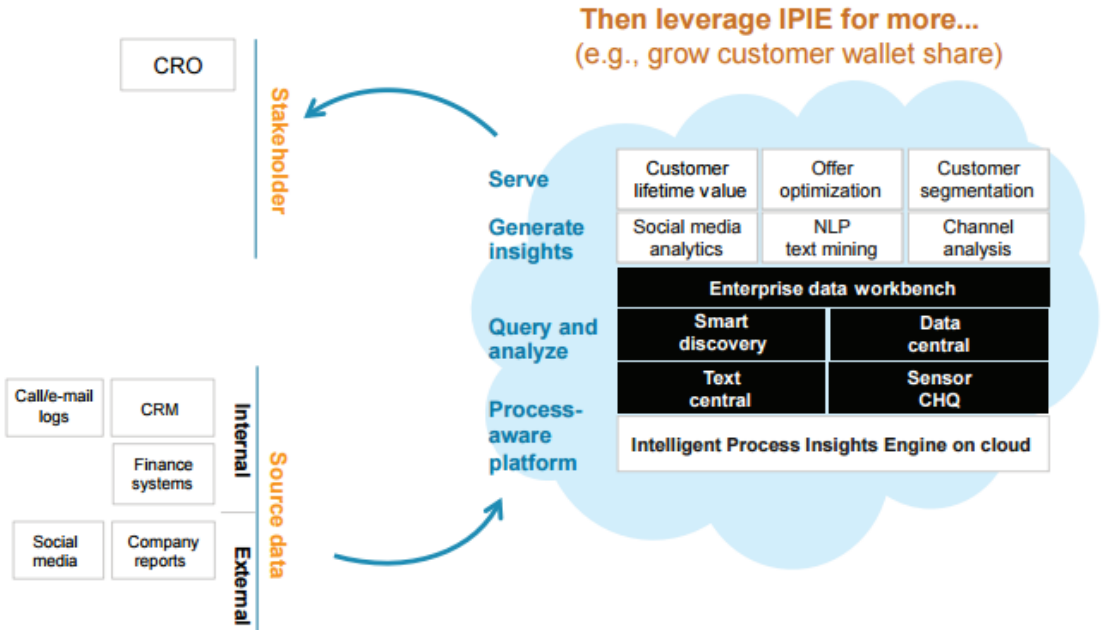
Since the applications are built and documented to follow the process maps developed for the purpose, governance is built in. The lightweight technology approach using the



Systems of Engagement (SE) model rapidly develop relevant applications. Once the first application is built, it is very easy for other CXOs to leverage the initial investment to rapidly develop and deploy other intelligent operations. Since the Systems of Engagement (SE) approach works directly with the systems of record, adding a new intelligent operation simply involves developing process maps and information supply chains to meet the new need, adding new data sources quickly, and, finally, developing the new applications on the IPIE. The flexibility of the platform allows clients to quickly tap into internal and external data sources whether structured or unstructured.

A Working Example

To understand the IPIE better here is a detailed example in which a client is looking to optimize a multi-channel client engagement across the enterprise. Here is the continuous process loop developed to model and measure the effectiveness of an intelligent operation reimagined through analytical insights and the appropriate use of technology



1. **Operate:** Once the initial strategy work has been completed with the client, the process starts by first examining the current operating model at multiple process levels. For example, at the lowest granularity the process could be the callcenter agent receiving a customer call about a particular set of issues.

2. **Measure:** We then determine what data can be used from systems of record (internal or external, unstructured or structured) to predict channel propensity. We then extend or reduce these data sets, building on our experience and knowledge of industry best practices.

3. **Consolidate:** The next step is to leverage the flexibility of the IPIE to capture and combine data from the systems of record, for example, marketing/campaign data, social media data, or web/mobile data. The solution could support realtime data on the customer's channel propensity at individual and segment levels.

4. **Analyze and report:** The solution then analyzes and reports on client base segmentation based on channel affinity scores and helps in the development of advanced analytics models including predictive models for the propensity to action or effectiveness of future spend on marketing across different channel strategies.

5. **Gather feedback:** A solution is only as good as the actual outcomes it generates so the solution must incorporate feedback on the effectiveness of particular strategies.

6. **Correct strategy and targets:** Feedback can help influence future strategies and corrective actions to be taken, which in this case could be to review selfservice targets or to revise the strategy by segment to increase the effectiveness of future campaigns or the channel interaction (e.g., web or mobile).

7. **Implementation:** The final step in the process is to implement the new strategies, which in this case could be to provide a new incentive to customers to opt for an appropriate channel for service (e.g., self-service) based on the desired outcome for the business. It could also mean developing concerted interventions at different customer interaction points driven by the analytic insights from the application.

IPIE Console/Web Application

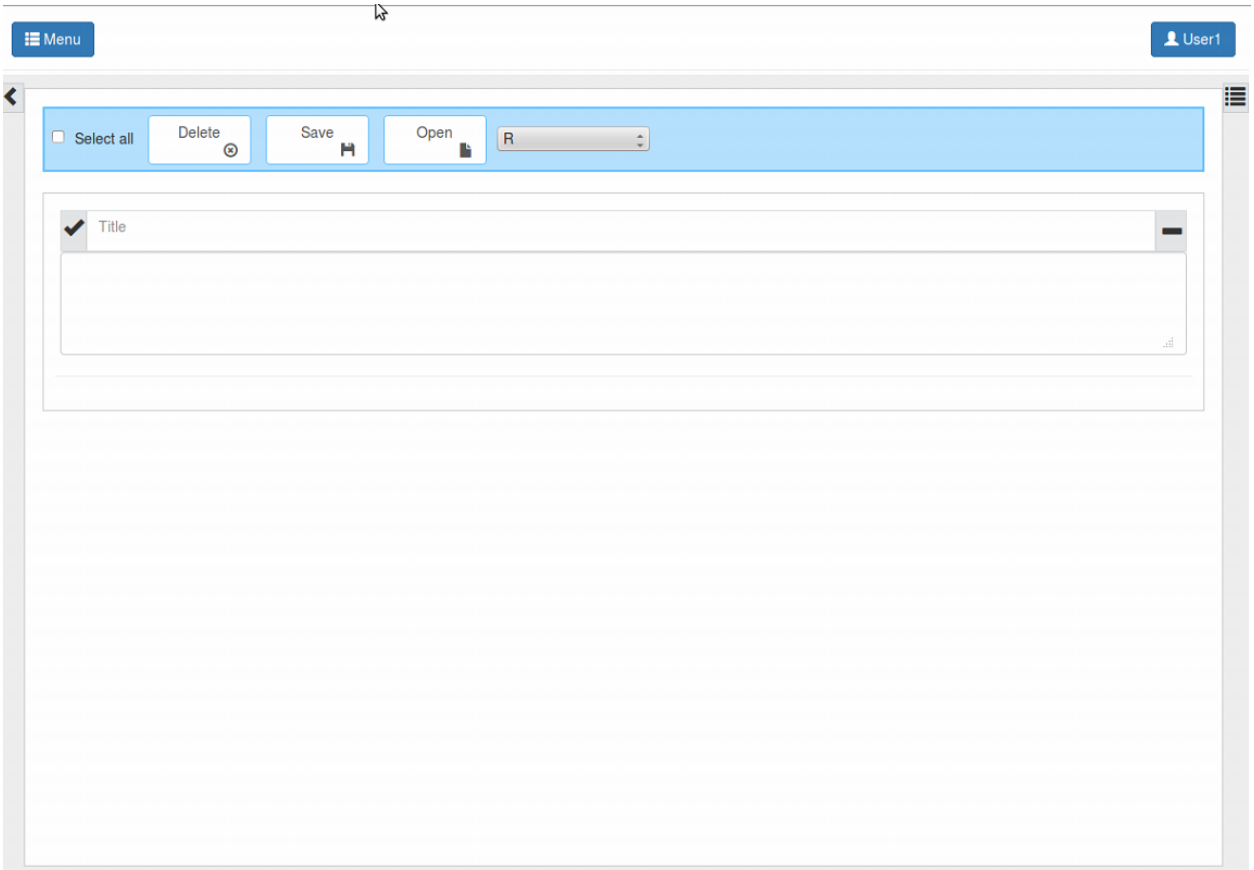
The IPIE Console is a web application that uses RStudio, PostgreSQL, Hibernate and Apache Kafka to deliver its output.

RStudio: It is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. We use the R Server edition for this application

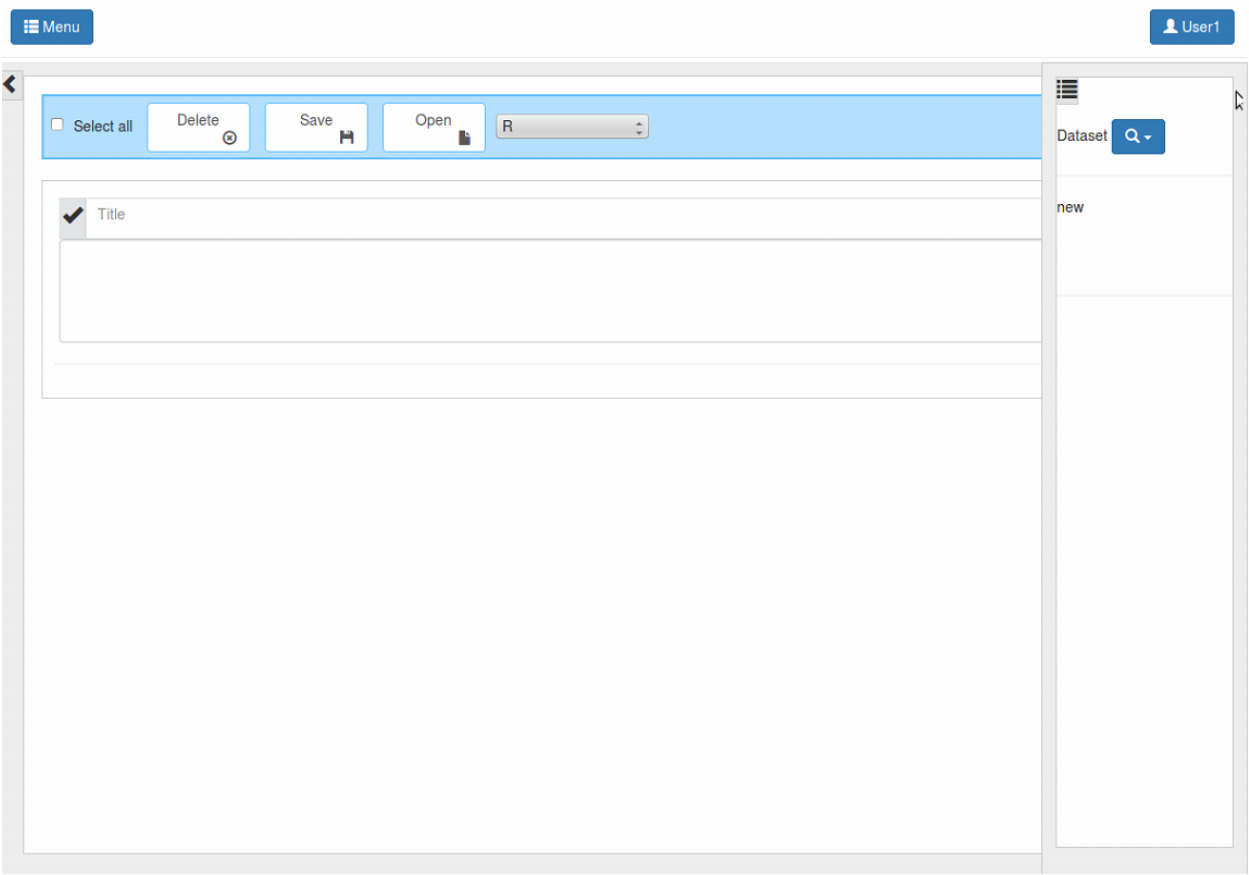
PostgreSQL: Often simply **Postgres** is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards-compliance. As a database server, its primary function is to store data securely, supporting best practices, and to allow for retrieval at the request of other software applications. It can handle workloads ranging from small single-machine applications to large Internet-facing applications with many concurrent users.

Hibernate: Hibernate ORM (Hibernate in short) is an object-relational mapping framework for the Java language. It provides a framework for mapping an object-oriented domain model to a relational database. Hibernate solves object-relational impedance mismatch problems by replacing direct, persistent database accesses with high-level object handling functions.

Apache Kafka: It is an open-source message broker project developed by the Apache Software Foundation written in Scala. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. The design is heavily influenced by transaction logs.



When the web application is run, the user should open a web browser and type the address of the application. They are then prompted to login or signup if they are a first time user.



After logging in they are presented with a console/terminal, here they can select from the different datasets available. After selecting a dataset they should enter an R operation into the console/terminal and press ctrl+enter. This operation/query is sent to an R Server for execution and the output of the execution is then stored in a database using Hibernate.

The communication between the console and the R server is done through Apache Kafka which is message broker system.

Bugs Found and Rectified

A **software bug** is an error, flaw, failure or fault in a computer **program** or system that causes it to produce an incorrect or unexpected result, or to behave in unintended ways. As a part of the Debugging process a few bugs have been found in the existing software.

Bug: If the console is opened on more than 3 browsers at the same time then it crashes when opened on the fourth browser.

Solution: The problem needed to be isolated to a class. It was occurring due to an if-else statement in the code. This has been changed to accommodate threads and thus eliminating the need for if else statements. A new thread is created whenever the console is opened and no further crashes happened after this implementation.

Bug: The console disconnects with the R-Engine after running for 1000 seconds. This meant that the queries in the code box wouldn't get solved and their answers wouldn't be stored in the database.

Solution: Instead of making the connection to the R Engine time based, it was changed so that the connection was open until the user closed the application/console.

Note: This may end up using more resources of the host computer.

Bug: Whenever a dataset of size larger than 10 million is given to process an error pops up saying that the heap memory has been exceeded.

Solution: When handling datasets, the data is now stored in a temporary file instead of in the Java Heap. This way for the memory to exceed it needs to be larger than the hard disk size of the computer.

Bug: After opening the console if the application is more than 3 times the data from the fourth time and above wasn't being recorded in the database.

Solution: This problem was isolated to a pair of classes that handled output in the code. . This was occurring due to an if-else statement in the code. This has been changed to accommodate threads and thus eliminating the need for if else statements. A new thread is created whenever the application is run and the old thread closed and no further crashes happened after this implementation.

Conclusion

The purpose of this report can be considered reached. It shows how various elements of the IPIE can be used to create powerful tools and widgets, which improves the overall customer experience.

The project involved many more things, and could not be included in this report to maintain confidentiality. The report barely touches the surface of the technologies and tools in building these features. Also, many features that were implemented could not be described in this report to maintain confidentiality. Overall, the project has been of great interest so far, and proved to be a great learning experience.

References

1. www.genpact.com
2. http://kafka.apache.org/
3. http://hibernate.org/
4. https://www.rstudio.com/