Consumer Woes

Big Data Project Code 26 – SRS Document

IBM Career Education

Disclaimer

This Software Requirements Specification document is a guideline. The document details all the high level requirements. The document also describes the broad scope of the project. While developing the solution if the developer has a valid point to add more details being within the scope specified then it can be accommodated after consultation with IBM designated Mentor.

Table of Contents

INTRODUCTION

Wealth of information flowing into an APEX Financial governing body of a developed nation is anybody's guess. Their knowledge management team is working smartly on leveraging humungous data received 24 * 7 from various financial institutions across the nation. Their goal is to tighten loopholes in the financial services being offered by various companies and ensure the consumer is delighted with products; services and the post engagement issues are minimized. The data source currently being taken up for ironing out issues is Consumers data, which is rapidly swelling with information about the financial companies, consumer issues with their products. The management is looking forward to status and insights on the consumer issues and handling in the first phase. The analysis reflecting financial institutions' status on products and services should be done in form of visualizations.

The solution will be developed using Bigsheets and subsequently deployed on IBM BlueMix, a PaaS platform on Cloud providing IBM Analytics for Hadoop service. This document is the primary input to the development team to architect the proposed visual mining model for this project.

Development Environment

The development will be carried out using Bigsheets. These tools will simplify analysis and creation of Visualizations. The BigInsights operations use MapReduce in the background to process the desired output. They also support techniques for text analytics.

System Users

The users of the solution shall be management team of the APEX financial institution. In addition the same will be become available to the financial service providers.

Assumptions

1. It is assumed that the developer will make an effort to understand Bigsheets functionality and explore its features to generate the desired outputs.
2. The output generated from this project would be visualizations.
3. The data links provided are for financial institutions within USA.
4. Not all columns of data will get used in the given problem; the developer may like to try out additional visualizations if the time permits.

REQUIREMENTS

It is required to analyze the financial data to discover insights on the product grievances amongst the service providers. The analysis must span across last 3 years in order to generate any meaningful analysis. The outputs will be visualizations from the Bigsheets chart feature along with the developer's observations that will be useful in drawing the inference.

Defining Thresholds

The challenge is to rank the institutions with the best and worst handling of cases in the products being offered by them. It's critical to understand the semantics of excess as per the financial industry. In absence of public information on the threshold values, the developer may decide on the threshold values to glean out meaningful insights from the data.

Preparation of Data

To ensure the data is in good shape to perform Bigsheets operations, the following checklist is followed.

1. All characters in text columns should be converted to uppercase.
2. Remove all punctuation, whitespace and control characters if any.
3. All numbers are integer values. Some columns may have negative values; no transformation should be carried out on those columns.
4. The flight departure and arrival time columns need to be converted into hh:mm format.
5. Since the flight date information is already available as Year, Month, Day of Month and Day of Week, there is no transformation required on this data.

Aggregation of Data

Keeping the objective in mind, next step is to run operations on columns to arrive at the number of instances. Bigsheets provides an array of useful functions to group data based on the column values.

On eyeballing the data you may wonder as to how such high volume of data shall get converted to charts! This thought itself leads you onto the right track! Visualizations require data that is result of pre-processing either using ready to use functions or writing algorithms to generate the final data set that does not require any further breaking down.

To understand the type of operations on prepared data, consider finding out how many products are being serviced per institution for the last 3 years. Using the group sheet feature, for each year, group on institution using count function. This operation will return year wise list of institutions with the number of products being serviced in each year.

As a next step, it would be a good idea to know the count of products for an institution that have been recording highest number of open cases. Once this data

is ready, at this stage, % open cases is computed for each institution with (Total open instances / Total number of cases for the institution)*100.

Finally, the list can be sorted on % open cases in descending order to get the institutions with higher open cases %ages on top. This data can be presented using bar charts. Combine this computed data for all the three years to get a comparative bar chart displaying institution wise %age open cases over the years.

Equipped with the knowledge on the type of operations run on such a large data set to arrive at a compact data, its time to create a solution for the problem on hand.

Expected Analysis & Output

The governing financial body is keen to get the following insights.

1. State wise status of issues. Use heat maps to depict the concentration of issues reported in the financial institutions in each state. Interaction required as the mouse hovers on each state, the name of the state and its count of issues is displayed.

2. Highest Number of issues in a category for past three years. Create Visualization using bar charts for products volumes with high degree of issues.

3. Issues by financial institution. Using bar charts display the Top 10 financial institutions having high degree of issues.

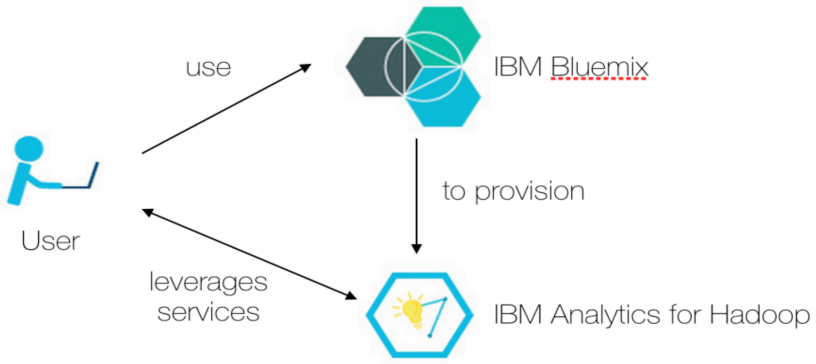
4. Type of Issues in Top three financial institutions. Rank number of issues in descending order.

5. Issues exceeding the defined norms. The visualization should represents only those issues that exceed the defined threshold value and the way they have been resolved. On a percentile scale display these issues categorized on the company response. From the visualization it should be clear as to which product related issues have been handled with Monetary relief, No Monetary relief, No relief at all.

To summarize, the analysis of consumer data using Bigsheets operations shall produce multiple outputs that can be used to draw out inferences for product service patterns of Institutions and help them model the rules and regulations for providing improving the service levels of the financial institutions.

DEPLOYMENT MODEL

The deployment model is outlined below.



Once the IBM Analytics for Hadoop instance is provisioned, the available service can be easily used starting from simply uploading a file, running a MapReduce code, Big Sheets, and many more.

This project will primarily require uploading data in the Hadoop File System. The Bigsheets shall upload the data and the CSV reader shall parse the data into respective columns in Bigsheets.

PROJECT TIPS

Big Data Problems may sometimes “appear” to be very simple; and one may be tempted to solve them with traditional methods. For example, counting frequency of occurrence of every word in documents. This is indeed a simple problem as long as documents are not “too many” and are not arriving “too frequently”. Now imagine there is a stream of millions of documents coming in! Clearly with traditional methods, it will be difficult to match the processing speed with data arrival speed (velocity), volume and on occasions its variety. Therefore, focus on scalable algorithms, smart visualizations, and requisite knowledge of math - especially statistics will be critical to success.

DATA SOURCING GUIDELINES

Big data solutions solve problems by ingesting extremely large of volumes of data for various operations to be carried out on them before the results are shared with the end user or the stream of output is generated for another application's input.

You could use these guidelines to source the data for your projects.

Download the consumer issues data for financial sector from http://1.usa.gov/1hYyald url.

Columns to be loaded

Please use the data from the following columns for the purpose of analysis.

Complaint ID, Product, Sub-product, Issue, Sub-issue, State, ZIPcode, Submitted Via, Date received, Date sent to company, Company, Company Response, Timely response, Consumer disputed.

TESTING GUIDELINES

It's easy to think that, if we know how to test a standard application, we know how to test the Big Data storage and application. Surprisingly so, it's not the case! Volume, Variety and Velocity of data make things really complex to test. While testing, mostly you are not dealing with structured data with a fixed schema; mostly the data is unstructured and a loosely defined or dynamic schema. The rate at which data is generated clearly exerts a pressure on speed of processing. Following must be kept in mind while planning the testing:

1. Plan on unit testing early and frequently during development. This is simply because big data testing is challenging, you may not be able to view source data using spreadsheets owing to sheer magnitude of the data.
2. Do not rely on eyeballing data or outputs as mechanism for verification. Create Test plan for each data set and the transformations stages it will go through in the entire process.
3. Big Data developers and testing team have to work with' Unstructured or Semi Structured' data (Data with dynamic schema) most of the time. Thus the testing activity requires additional inputs on 'how to derive the structure dynamically from the given data sources' from the business/development teams.
4. When it comes to the actual validation of the data, considering the huge data sets for validation, 'Sampling' strategy comes to rescue. But even that is a challenge in the context of Big Data Validation. This provides a tremendous opportunity for the testers who are innovative and who would go the extra mile to build the utilities that can increase the test coverage of BIG Data while increasing the test productivity as well.
5. The testing process should be strengthened on reuse and optimization of the test case sets, otherwise due to sheer size of the requirements to be tested will become unmanageable.

SUGGESTED READING

The project is aimed at making the student understand concepts of (a) Design and Development using IBM Analytics for Hadoop, IBM Infosphere Biginsights, Bluemix platform; and (b) Concepts and use of algorithms, models or visualisations for Big Data problems.

Tools

The following reading reference is easy to understand and should be read to get a clear understanding of capabilities of the tools and how you would leverage them to execute a project.

Resource	URL
IBM BigInsights Knowledge Center	http://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.welcome.doc/doc/welcome.html
IBM InfoSphere BigInsights for Hadoop Community	https://developer.ibm.com/hadoop/
InfoSphere BigInsights Quick Start Edition	http://www-01.ibm.com/software/data/infosphere/biginsights/quick-start/tutorials.html
IBM Bluemix Dev – Hands on with Hadoop in Minutes	https://developer.ibm.com/bluemix/2014/08/26/hands-on-with-hadoop-in-minutes/

Probability & Statistics

Big data problems require an understanding of Probability and Statistics, which is pre-requisite for most modeling exercises. You may use your own reference content for solving the problems or may refer to the fundamentals from the following links.

Resource	URL
Introductory Statistics: Concepts, Models and Applications	http://www.psychstat.missouristate.edu/sbk00.htm
Statistical Thinking for Managerial Decisions	home.ubalt.edu/ntsbarsh/business-stat/opre504.htm