Navigator

Big Data Project Code 33 – 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

Content creation services are moving uphill with delivery models especially designed to cater to cross industry user. The business requires various types and levels of skills to capture, collect, curate for bring it to ready for consumption state. A leading content provider with world-wide user base is targeting content for travel agencies to provide meaning information about the destinations they are travelling to. The team came up with a wonderful idea to prepare content ready for exploration in EWNS directions. This could be possible only if the content is available in geocoded format. Leveraging Wikipedia content which is getting geo coded off late and can be relied upon owing to its crowdsourcing methodology. The challenge taken up requires mining high volumes of data and organizing it for navigation in geographical order. The team cracked a way to handle this complexity by leveraging the map reduce paradigm. The name of the data service is being called as the Navigator, which will allow their customer to render the content in all 4 directions based on the end user's desire to navigate.

The solution will be developed using MapReduce paradigm 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 data-mining model for this project.

Development Environment

The development will be carried out using Eclipse Version 4.2 or above. The IBM InfoSphere BigInsights Eclipse tools will have to be added to your Eclipse development environment. These tools will simplify development and deployment of applications to the BigInsights server using Java MapReduce, JAQL, Pig and Hive. They also support developing text analytics programs, such as extractors, that run on IBM InfoSphere BigInsights.

System Users

The Travel agency's mobile application interface will use the data stream generated by the project to ensure reliable and updated content delivery at all times.

Assumptions

1. The output generated from this project would be in JSON format.
2. Knowledge of N-Triples format and its parsing method is a must.
3. The source data has reference to places all around the world. For the testing purpose, shortlist some cities around the world, find the appropriate geocodes and prepare the seed file using this information. The output of the model can be validated against the reference file used for mining.

REQUIREMENTS

It is required to design & develop a framework for regular mining of “areas of interest in close vicinity” for the stream of travel agency’s customers. The list of cities that will be visited by its customer in near future is obtained from the travel portal on a periodic basis. This list becomes the seed for mining the “areas of interest in a particular city”. This framework accepts the seed file containing the list of cities and their latitude and longitudes. The big data dump of geocoded Wikipedia articles is mined to extract articles that are falling within a fixed radius, say 20 kms, of each city. The city-wise, “areas of interest (i.e. articles)” are generated in a JSON format that follows a navigational order of EWNS. Thus for a city, there will be series of searches in a single direction and the resultant entries will be organized based on their distance from the city’s latitude and longitude. The target file generated from this activity will contain geographical location (city name), geo location parameters such as latitude and longitude, article URI and direction tag (East, West, North, South). This output is subsequently used by the agency to populate customer’s mobile app feed, so that the details of the attractions become available to customers in a timely manner along with navigation feature. Tips on creating this framework are outlined below.

The problem has twofold complexity, which needs to be addressed. First, the big data dump of Wikipedia entries is available in N-Triples (NT) format. N-Triples is a line-based, plain text format for encoding an RDF graph. The simplest triple statement is a sequence of (subject, predicate, object) terms, separated by whitespace and terminated by ‘.’ after each triple. For more details, please refer to https://en.wikipedia.org/wiki/N-Triples URL.

Second, taking geo-location parameters of the cities in seed file as a reference, determine an approximate boundary of parameters within 20 or 25 kilometers of the city in question. Consider each city’s geo-location parameters as a starting point to find the Wikipedia content about places or event locations falling next in the direction (EWNS) and is within the boundary parameters. This can be achieved by applying the Haversine formula that gives the distances between two points on a sphere from their longitudes and latitudes. The extracted data from the data dump is captured in the output JSON file.is required to design & develop a framework for regular mining of “places of interests” for the stream of agency’s customers. The list of cities that will be visited by its customer in near future is obtained from the travel portal on a periodic basis. This list becomes the seed for mining the “places of interest”. This framework accepts the seed file containing the list of cities and their latitude and longitudes.

The big data dump of geocoded Wikipedia articles is mined to extract articles that are falling within a fixed radius, say 20 Kilometers, of each city. The city-wise, “near by attractions (i.e. articles)” are generated in a JSON format. This output is subsequently used by the agency to populate customer’s mobile app feed, so that the details of the attractions become available to customers in a timely manner. Tips on creating this framework are outlined below.

For this project, there are no constraints about covering the complete city for the Wikipedia information as some of the articles might refer to geo locations outside the boundary values determined by the program.

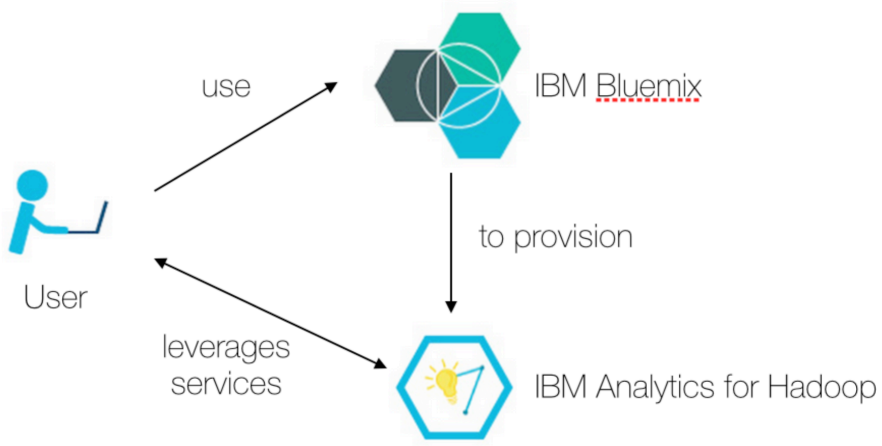
About Haversine

The haversine formula is an equation important in navigation, giving great-circle distances between two points on a sphere from their longitudes and latitudes. The basic details can be found at Wikipedia entry located at http://en.wikipedia.org/wiki/Haversine_formula URL and the computational reference can be found at http://andrew.hedges.name/experiments/haversine/ .

To summarize, the project on hand poses a unique challenge of creating the data required for building the around me seed file for relevant content extraction module. The guidelines for creating the data for this project are given in the data sourcing guidelines.

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 reading resource file in N-Triple format, parsing it and uploading entire geo-located dataset and running a MapReduce program (to be developed) that will read a seed file of cities with geolocation parameters in a particular direction to generate city wise “areas of interest” with navigation flag and distance.

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.

The following guidelines would come in handy to source the data for your projects.

There are two sets of data required to work on this project.

1. Creation of seed file for about 50 cities around the world. The format of the file should be CSV with column details such as [City Name, Longitude, Latitude].
2. The reference data geo-coordinates_en.nt.bz2 for deriving the Wikipedia Url is available in N-triples format, to be downloaded from the http://downloads.dbpedia.org/2015-04/core-i18n/en/ location and parsed using N-Triples guidelines.

The geolocation data is in varying formats in the reference file, its important to understand the RDF format and parse the geocodes accordingly. Please note some of the entries will not have any geocodes, such entries should be ignored.

Both the files shall be placed in HDFS once they are ready for execution of Map reduce algorithm.

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 visualizations 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