Predictive analytics with online data for WSO2 Machine Learner with the support of Ensemble method

Project ID : 16-054

Project Proposal

B.Sc. Special (Honors) Degree in Information Technology

Submitted on 09.03.2016

Predictive analytics with online data for WSO2 Machine Learner with the support of Ensemble method

Project ID : 16-054

Authors:

Student ID	Name	Signature
IT13010904	Herath H.M.H.R	
IT13037734	Pathinayake I.M	
IT13073060	Dias G.H.G.A	
IT13028756	Indujayani K	

Supervisor

…….

Mr. LakmalRupasinghe

DECLARATION

We declare that this is our own work and this project proposal does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

……..	……
Herath H.M.H.R	Pathinayake I.M
……..	…….
Dias G.H.G.A	Indujayani K

ABSTRACT

The generation of data has never been at a higher pace, data which concerns our day to day endeavors such as personal, financial, sales, marketing and so on.Predictive analytics is the use of data, statistical algorithms and machine-learning techniques to identify the likelihood of future outcomes based on historical data. The ultimate goal of predictive analytics is to exceed the descriptive statistics and provide the best assessment on what will happen in the future which in turn will streamline decision making and result in better actions.

In the real world scenarios, many different types of data are acquired sequentially. Traditional way of waiting for data to be collected and identifying patterns is a one-time process. As the new data arrives, these patterns change and the learned patterns need to be evolved accordingly and the decisions based on these models change as well. Concept of machine learning algorithms with streaming data has been emerged under these circumstances.The WSO2 machine learner aims at achieving the following two main objectives; identifying patterns in recent history and updating the patterns with incoming data without catastrophic forgetting. The specialty being this machine learner allows users to change parameters of the algorithm on the fly (during the incremental learning procedure)

TABLE OF CONTENT

List of Figures

List of Tables

1. INTRODUCTION

1.1 Background

We are living in an era where the technology is at its climax. Most problems nowadays can be solved with the knowledge from Computer Science and Informatics. While some problems can be solved with clever architecture and code, some however can't. Some dire problems such as predicting business scenarios, driving a vehicle or detecting handwriting need much more dynamic and adaptable solutions based on past and present data. This is where machine learning comes into the picture.

Facebook news uses machine learning to customize power of each member. If a member fails to oscillate frequently in order to read or"like" to the messages of a particular friend, the News Feed will begin to show more of that friend activity feed earlier. Behind the scenes, the software is simply using statistical analysis and predictive analysis to identify patterns in user data and use patterns to fill the News Service. In case the stop member no longer read, like or comment on posts of his friend, the new data will be included in the data set and News Feed will be adjusted accordingly[1].

WSO2 Machine Learner has a versatile set of characteristics. It can extract features from a dataset made available from a file system, Hadoop Distributed File System (HDFS) or WSO2Data Analytics Server (DAS). This data is passed on to the Machine Learner Core, which allows you to explore your datasets, pre-process your data and apply various machine learning algorithms to make sense out of it all. Using Apache spark, it then analyzes and builds models with the chosen algorithms. Recommendation engine capability allows you to provide product recommendations for users.

As mentioned in the abstract there are two main objectives that we hope to achieve at the end of this research.

- Identifying patterns in the recent history.
- Updating the patterns with incoming data without catastrophic forgetting.

These objectives can be approached in two methods [2].

- Incremental algorithms - there are machine learning algorithms which can be modified to support incremental learning. Eg: Mini-batch k-means, stochastic gradient descent based linear models, SVM, etc.

- Periodic re-training - machine learning models are trained with buffered data periodically

The goal of WSO2 Machine Learner is to make machine learning accessible to WSO2 Data Analytics Platform. The expected outcome of this process is a deployable model in many of WSO2 products such as WSO2 Enterprise Service Bus (ESB), WSO2 Complex Event Processor (CEP). As a latter part of this project we hope to implement Ensemble methods support for WSO2 Machine Learner. The WSO2 Machine Learner can be utilized to build machine learning models for various tasks such as fraud detection, anomaly detection, classification etc. It will also come in handy for developers and data scientists to implement machine learning algorithms quickly. WSO2 Machine Learner is built up on top of the WSO2 Carbon platform, which is based on the Open Service Gateway Initiative (OSGI) framework enabling better modularity for your service oriented architecture (SOA) and all its operations are exposed through a RESTful API. WSO2 Machine Learner is released under Apache Software License Version 2.0, one of the most business-friendly licenses available today [3].

The prominent features of WSO2 Machine Learner are as follows:

- Data Exploration
- Model generation
- Model comparison
- Prediction

Data exploration focuses on the ability to convert datasets into a format which is suitable for machine learning algorithms (data preprocessing) and ability to do explorative data analysis. Model generation supports various machine learning algorithms in classification, numerical prediction and clustering spaces. Also it generates models and persist them or download them via WSO2 machine learning REST API. At the same time it has the ability to examine the characteristics of a model. Model comparison comprises of figures that help you to compare models across a machine learning project. In Prediction the generated models can be used to perform predictions using a REST API client, WSO2 Enterprise Service Bus (ESB) mediator or WSO2 Complex Event Processor (CEP).

1.2 Literature Review

We are experiencing major growth of machine learners with data mining, big data analysis and predictive analytics. The iterative aspect of machine learning is important because models are exposed to new data, they are able to adapt independently. They learn from previous computations to produce reliable, repeatable decisions and results. It is not a science that is new but one that is gaining fresh momentum. Because of new computing technologies, machine learning today is not like machine learning of the past. While many machine learning algorithms have prevailed for a long time, the ability to automatically apply complex mathematical calculations to big data-over and over, faster and faster- is a recent development.

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever; things such as growing volumes and varieties of available data, computational processing that is cheap and powerful and affordable data storage. All of these things mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results ultimately guiding better decisions and smart actions without human intervention. Analytics thought leader Thomas H. Davenport wrote in The Wall Street Journal that with rapidly changing, growing volumes of data, "you need fast moving modeling streams to keep up", also he says "humans can typically create one or two good models a week; machine learning can create thousands of models a week" [4].

WSO2 machine learner will be designed for the enterprise world. It will come as an integrated solution with the rest of the big data processing technologies: batch, real-time and interactive analytics. Also it will include support from data collection, analysis to communication (visualizations, APIs and alerts). Hence it is a part of a complete analytical solution. WSO2 Machine Learner will fundamentally address predictive analytics lifecycle including model deployment and management while also catering to changing parameters of the algorithm on the fly.

Products Available in the Market:

1. **HPCC Systems (High Performance Computing Cluster)**

   HPCC systems is an open source, massive parallel-processing computing platform for big data processing and analytics. While it has so many advantages such as superior

performance, agility, scalability and has been proven in mission critical big data production environments it lacks;

- Predictive analysis
- Capability of changing parameters of an algorithm on the fly
- Ensemble methods support

2. **WEKA (Waikato Environment for Knowledge Analysis)**

   WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well equipped for developing new machine learning schemes.

   Lacks

   - Limitations to value ranges for methods options
   - Lack of parallel processing
   - Reliance on the experimenter GUI
   - Lack of robustness [5]
   - Predictive analysis capabilities

3. **OpenNN- Open Neural Networks Library**

   OpenNN is a software library written in C++ for predictive analytics. It implements neural networks, the most successful deep learning method. The main advantage of OpenNN is its high performance. This library outstands in terms of execution speed and memory allocation. It is constantly optimized and parallelized in order to maximize its efficiency. Lacks;

   - Extract data from Comma Separated Value (CSV) or Tab Separated Value (TSV) file formats
   - Use multiple visualizations to explore your data - scatter plots, histograms, Trellis charts, parallel sets and cluster diagrams

4. **Scikit-learn**

   This is an open source machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting and k means. Also it is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Lacks;

- Predictive analytics
- Machine learning and training abilities
- Change parameters of the algorithm on the fly

5. **Orange**

   Orange is an open source machine learning and data mining software (written in Python). It has a visual programming front-end for explorative data analysis and visualization, and can also be used as a Python library. Lacks;

   - Predictive analytics
   - Evaluate models with performance measures like ROC Curve, Confusion Matrix, Accuracy and Root Mean Squared Error
   - Select from machine learning algorithms - Decision Trees, Linear Regression etc.

6. **Torch**

   Torch is an open source machine learning library, a scientific computing framework, and a script language based on the Luaprogramming language.It provides a wide range of algorithms for deep machine learningand uses an extremely fast scripting language LuaJIT, and an underlying C implementation. Lacks;

   - multiple visualizations to explore your data - scatter plots, histograms, Trellis charts, parallel sets and cluster diagrams
   - Capability of changing parameters of an algorithm on the fly
   - Easy Graphical User Interface (GUI) for human friendly viewing

1.3 Research Gap

As it's clear to us by now, there are many existing applications to predict the outcome of an experiment from gathered data. Consider an example like this. American Football is one of the most watched sports in the entire world. People come from all over the world to watch the FIFA finals. People are anxious to know which teams will get selected to the semifinals, finals and ultimately become the FIFA champions. Many software have been produced to predict scenarios like this. The specialty of our WSO2 Machine Learner is that it takes into account the online streaming data and according to the expected outcome it can change the parameters of the algorithm on the fly and provide the most accurate result. This is not just applicable for a sport but rather it can

be used in many fields. Even though WSO2 has already developed a Machine Learner that expects to fulfill those requirements,this has a few issues. The following table shows a comparison of features between the existing Machine Learner and the proposed solution.

Features	WEKA	HPCC System	Existing WSO2 ML	Proposed WSO2 ML
User Friendly Interfaces	✓	✓	✓	✓
Outlier detection when generating graphs	X	X	X	✓
Model Generation	X	X	✓	✓
Can configure with spark	X	X	X	✓
Display prediction result rather than value	X	X	X	✓
Validate File Formats	X	X	X	✓
Improved Performance	X	X	X	✓
Successfully displaying cluster diagrams	X	X	X	✓
Use Streaming data	X	X	X	✓
Use ensemble method	X	X	X	✓

Table 1 – Comparison of existing WSO2 Machine Learner with the proposed Machine learner

1.4 Research Problem

The world is rapidly changing and being subjected to a massive wave of revolutions and dynamic changes spanning from global businesses to the ever so exploding presence of Information Technology. As the businesses are expanding by the day, their complexity is also getting bigger. Most businesses and organizations collect data about their operations. They then examine this data for insights into their operations and into the transactions their business performs. The amount of raw data stored in corporate databases is exploding. From trillions of point of sale transactions and credit card purchases to pixel by pixel images of galaxies, databases are now measured in gigabytes and terabytes. For instance, every day, Wal-Mart uploads about 20 million point of sale transactions to an A&T massively parallel system with 483 processors running a centralized database [6]. However, raw data by itself

does not provide much information. In today's fiercely competitive business environment, companies need to rapidly turn these terabytes of raw data into significant insights into their customers and markets to guide their marketing, investment and management strategies. An automated data mining and predictive approach, however, can often find profitable relationships which you may not even have suspected existed, or that you knew would take too long to find by manual means. Although there are many products in the market which satisfy the above requirements most of them don't address the following drawbacks;

- Support for Deep Learning and Neural Networks
- Use multiple visualizations to explore your data; scatter plots, histograms, Trellis charts, cluster diagrams and so on
- Use feature engineering to pre-process data for better results
- Easy Graphical User Interface for human friendly viewing

Considering the above facts, we can come to the conclusion that there is a real need of cost-effective Predictive Analytics tool that caters the need of making constructive decisions based on prevailing large bulks of data.

2 OBJECTIVES

2.1 Main Objectives

In real-world scenarios, many types of data are acquired sequentially. Traditional way of waiting for the data to be collected and identifying models is a unique process. With the arrival of new data, these models are changing and acquired habits must be changed accordingly and decisions based on these models are also changing. Concept of machine learning algorithms with data streaming was born in these circumstances.

The proposed research project focused on several objectives. With the completion of the project is supposed to accomplish these research objectives. The idea of incremental learning with streaming data focuses on two objectives.

1. Identifying patterns in the recent history.
2. Updating the patterns with incoming data without catastrophic forgetting.

Apart from the above mentioned objectives the proposed research focus on the following objectives as well

1. Implement an ensemble method(s), to combine multiple algorithms.
2. Create a UI to include the end to end flow of training the algorithm.
3. Integrate it to the WSO2 Machine Learner

2.2 Specific Objectives

In order to achieve the above mentioned main objectives the research has to fulfill the following specific objectives.

Objective 1 : Designing an architecture for incremental learning and visualizations.

The WSO2 machine learning architecture comprises of ML data layer, ML storage handler, ML core, ML REST API, ML user interface. The proposed research mainly focuses on designing an architecture for incremental learning and visualizations.

Objective 2 : Creating the incremental learning component.

With the streaming data (online data) the architecture should identify patterns in the history and update the patterns with incoming data without catastrophic forgetting. To accomplish this, the proposed research will create the incremental learning component.

Objective 3 : Creating interactive visualizations for incremental learning models.

Most of Predictive analysis tools in the market has a very complex interface where professional knowledge is needed in accessing it. The proposed research is designed in such a way where a normal person who has the basic skills of using computer and mathematics, can use it and get the required outputs without having any difficulties.

Objective 4 : Allowing users to change parameters of the algorithms on the fly (during the incremental learning procedure) by analyzing the models.

Lot of Predictive analysis tools in the market work pretty well. But most of them does not allow user to change the parameters of the algorithms during the incremental

learning procedure. The proposed research will be designed allowing user to change the parameters of the algorithms during the incremental learning procedure.

Objective 5 : Implement ensemble method to combine multiple algorithms

Ensemble learning is combining multiple learning algorithms to obtain better predictive accuracy than what we could be obtained from a single learning algorithm. The existing WSO2 machine learner is not supported to the ensemble methods.

Objective 6: Create a UI to include the end to end flow of training the algorithm.

The proposed research will be created a UI to include the end to end flow of training the algorithm.

Objective 7: Integrate it to the WSO2 Machine Learner.

As the final product of the proposed research the ensemble methods supportability will be added to the WSO2 Machine Learner.

Objective 8 : Documentation and examples.

Prepare the documentation such as proposal, SRS for the proposed research and provide results with relevant examples.

·**Objective 9: Explore and Understand Modern IT techniques**

Explore and understand the modern IT techniques such working with Streaming data, Predictive Model building, Data mining, and big data analysis; and invent innovative solutions which can be adopted for the development real world situations.

·**Objective 10: Participating for Competitions**

Participate for IT competitions such as NBQSA and gain knowledge and experience.

3 RESEARCH METHODOLOGY

The following section includes detailed description about the techniques and the mechanism which are going to use for the proposed research to make the research success. The description contains the way of software implementation goes, what are the materials and data needed to process and how they will be collected. Apart from that it contains the time frames and the schedules that are proceeding to achieve the proposed research's objectives, and also the research areas that we are covering in order to fulfill the research objectives.

3.1 How the end product looks like

Creating another machine learner is not the goal of this research. The existing WSO2 machine learner works pretty well. But it does not have the ability of giving an accurate prediction using streaming data (online data). Our main objective is to develop and machine learner which can gives accurate prediction using streaming data. Apart from that the user should be able to change the parameter of the algorithm on the fly. The machine learner should give predictions according to those parameters accurately. Our main objective of this research is accomplish those tasks and make the WSO2 machine learner more users friendly and increases the accuracy of results on the fly.

3.2 Design the system

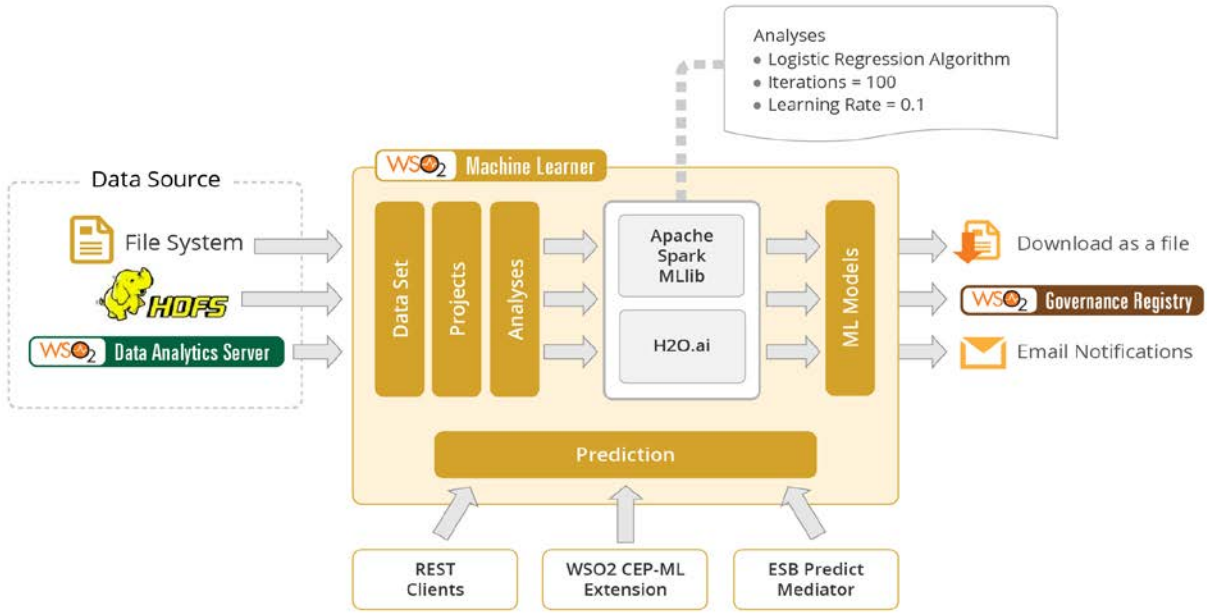

Figure 1- High Level Architecture

In order to achieve the goals of the proposed project there are 5 major tasks to be completed. Those are,

- Incremental learning component
- Predictive model with streaming data
- User enabled parameter changed of the algorithm on the fly
- Data visualization component
- Implement Ensemble methods.

Creating the Incremental learning component

With the streaming data (online data) the architecture should identify patterns in the history and update the patterns with incoming data without catastrophic forgetting. To accomplish this, the proposed research will create the incremental learning component.

The idea of incremental learning with streaming data focuses on two objectives:

- Identifying patterns in the recent history.
- Updating the patterns with incoming data without catastrophic forgetting.

Furthermore the proposed new Machine Learner will automatically detects odd data and remove it from processing. Therefore it will result in high- accuracy and best quality output.

Predictive model with streaming data

The proposed machine learner mainly works with data mining, trend and forecasting. Once the data is gathered, machine learner will make the statistical model in order to generate predictions.

This statistical model can be approached in two methods. Those are,

- Incremental algorithms - there are machine learning algorithms which can be modified to support incremental learning. Eg: mini-batch k-means, stochastic gradient descent based linear models, SVM, etc.

- Periodic re-training - machine learning models are trained with buffered data periodically.

According to the results of this predictive model, the performance of the machine learner will be increased with online streaming data.

User enabled parameter changed of the algorithm on the fly

Lot of Predictive analysis tools in the market work pretty well. But most of them does not allow user to change the parameters of the algorithms during the incremental learning procedure. The proposed research will be designed allowing user to change the parameters of the algorithms during the incremental learning procedure.

The user can select only required parameters and analyses the data during the incremental learning procedure.

Data visualization component

Once the data proceed it will be appeared to the user in a format which any user can understand. Data will be presented in a way anyone can easily understand with a basic computer knowledge and basic English knowledge.Use multiple visualizations to explore your data; scatter plots, histograms, Trellis charts, cluster diagrams and so on.

Implement Ensemble methods

The whole idea is to employ multiple apprentice method and combine their predictions. The aim set of methods is the combination of several predictions based estimators constructed as a learning algorithm in order to improve the generalization / robustness on one estimator. It provides better predictive accuracy than a single learning algorithm. Final prediction will be done by taking the weighted vote of the predictions of the gathered algorithms [8].

3.3 Addressed Research Areas

Research area 1: Data mining

Data mining is the process of analyzing data and summarizing it into useful information. Useful information can be used to increase revenue, cut down costs or both. Data mining software is one of a number of analytical tools for analyzing data. This gives users the ability to analyze data from many different dimensions, categorize it and summarize the identified relationships. Hence the existence of a data mining software for any purpose is extremely useful.

For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays. [9]

Research area 2: Predictive Analytics – To predict future outcomes

The proposed system extracts information from existing data sets in order to determine patterns and predict future outcomes and trends. The system does not tell you what will happen in the future. It forecasts what might happen in the future with an acceptable level of reliability. In the process different predictive models have to be developed in order to do different kinds of predictions according to the requirements of each industry. Decision making is done based on the results of those predictions.

3.4 Implementation

WSO2 Machine Learner takes data one step further, pairing data gathering and analytics with predictive intelligence: this helps you understand not just the present, but to predict scenarios and generate solutions for the future. We hope to implement this project in the following manner. First of all, online streaming data is taken into the Machine Learner. When data arrive in a streaming fashion, it is useful to fit regression models online, updating the parameters of the model as new data arrives [10].Anytime a text file is placed in /training/data/dir the model will update. Anytime a text file is placed in/testing/data/dir it will generate predictions. As more data is fed to the training directory, the predictions will get better. Inside the machine learner there will be 3 main algorithms for different attributes;

- Streaming linear regression algorithm[10]
- Streaming K-means clustering algorithm[11-12]
- Support Vector Machine (SVM) algorithm[13-14]

After data is processed through these algorithms and an outcome is predicted, Ensemble methods for Machine Learning is implemented to combine these multiple algorithms such that a better predictive accuracy could be achieved than what could be achieved from a single learning algorithm[16]. Final prediction is done by taking a weighted vote of the predictions of the combined algorithms.

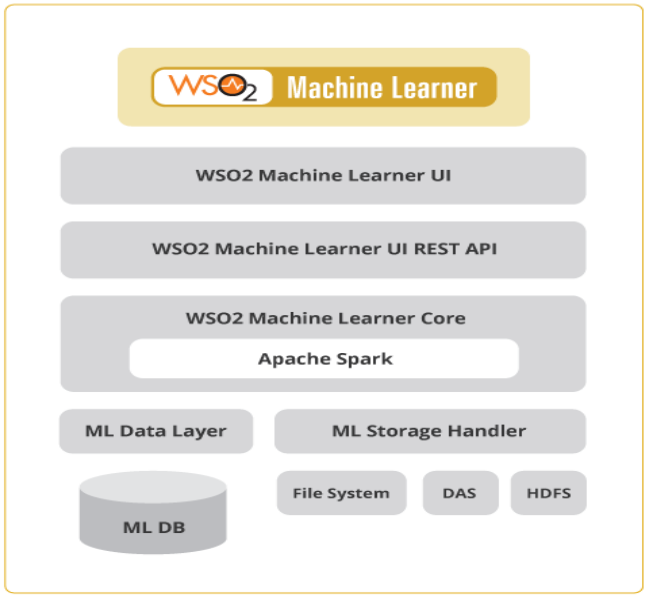

Figure 2 - Machine Learner

3.5 Tools and Techniques

Tools:

- Eclipse

Technologies:

- Machine Learning
- Streaming data
- Java
- Scala
- HTML 5
- Java Script
- Apache Spark

3.6 Testing

- Unit Testing – Each unit will test the system by the member of the group is developing that particular unit and produce a defect -free coding unit
- Component Testing – Several error -free units are combined together and test. Each member will combine its proven together and will test units.
- Integration Testing – Each user responsible to test whether the each component working as expect and their relationships are working as expect.
- System Testing – All components of each group member will be combined together and test the entire system to verify the functionality, performance and accuracy of the system.

3.7 Gantt Chart

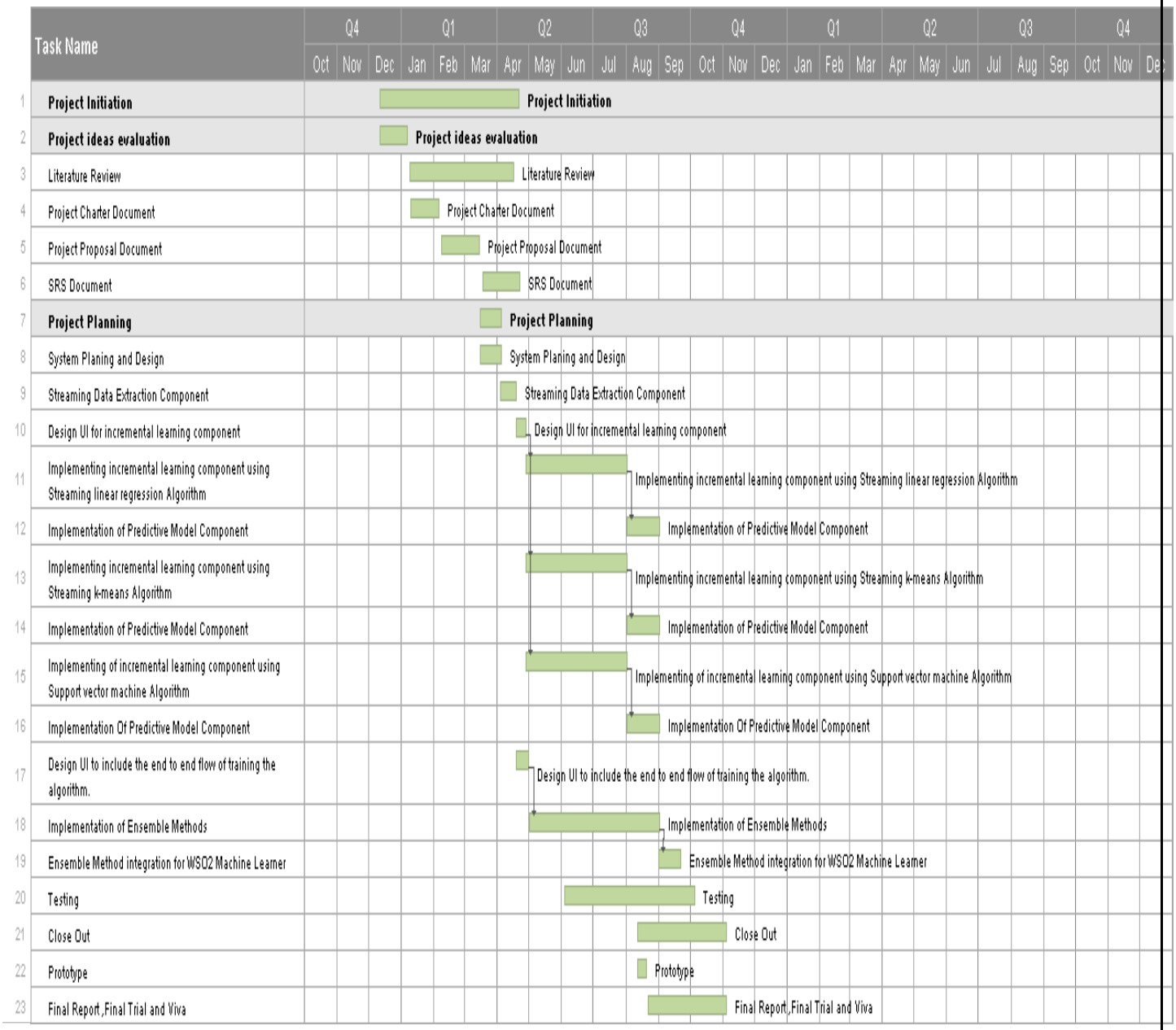

Figure 3 - Gantt Chart

3.8 Work Breakdown Structure

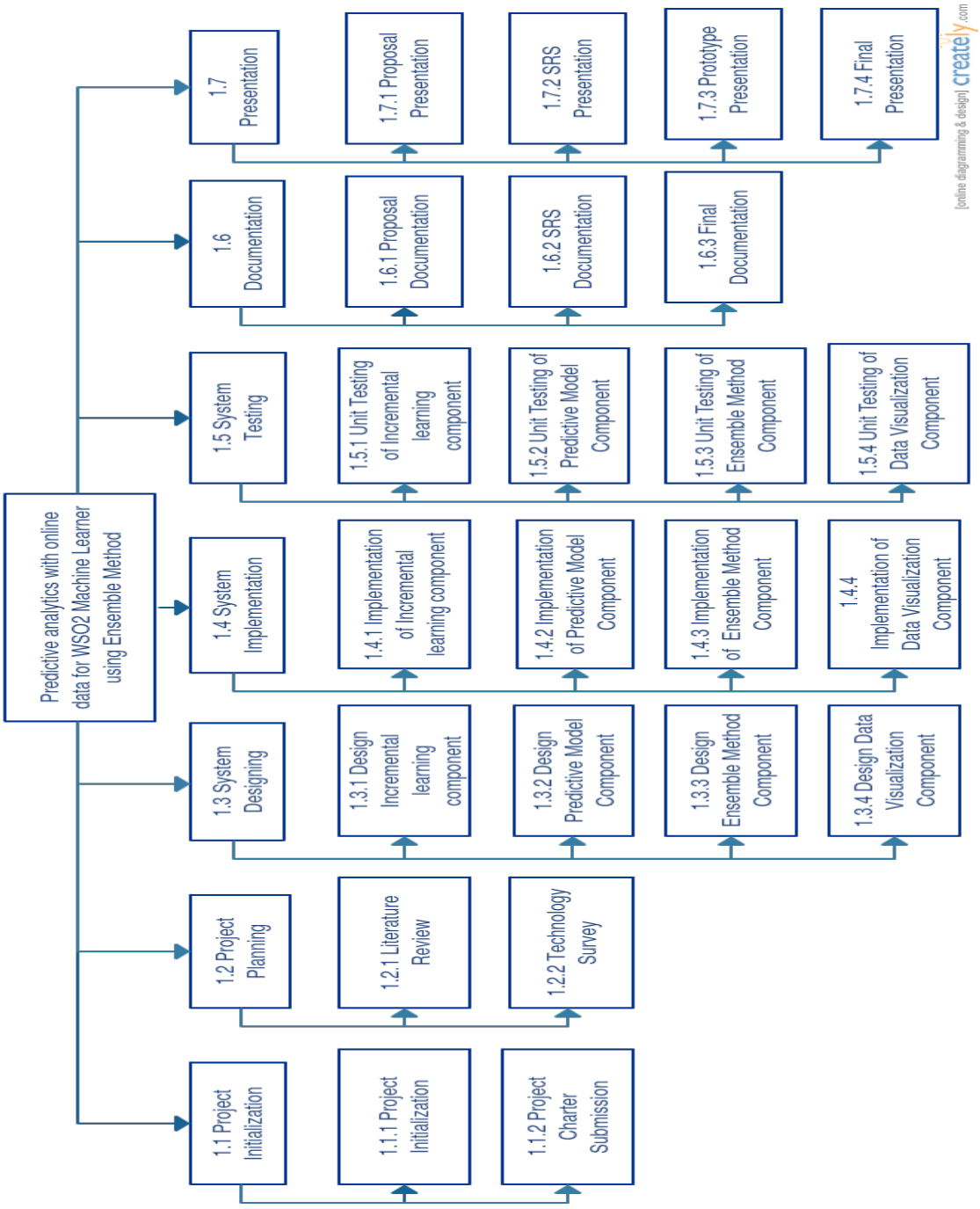

Figure 4 - Work Breakdown Structure

4 DESCRIPTION OF PERSONAL AND FACILITIES

Member	Component	Task
Herath H.M.H.R	Incremental Learning Component	(All the tasks are related to Streaming linear regression algorithm.) • Collect relevant data from data sources. • Create interactive visualization for incremental learning model. • Create the predictive model using Streaming linear regression algorithm. • Change the parameters of the algorithm during the incremental learning procedure. • Test the incremental learning component. • Documentation.
Pathinayake I.M	Incremental Learning Component	(All the tasks are related to k- means clustering algorithm.) • Collect relevant data from data sources. • Create interactive visualization for incremental learning model. • Create the predictive model using k- means clusteringalgorithm. • Change the parameters of the algorithm during the incremental learning procedure. • Test the incremental learning component. • Documentation.
Dias G.H.G.A	Incremental Learning	(All the tasks are related to Support

	Component	vector machine algorithm.) • Collect relevant data from data sources. • Create interactive visualization for incremental learning model. • Create the predictive model using Support vector machine algorithm. • Change the parameters of the algorithm during the incremental learning procedure. • Test the incremental learning component. • Documentation.
Indujayani K	Ensemble Method Component	(All the tasks are related to ensemble method.) • Collect relevant data from data sources.(relevant algorithms) • Create a UI to include the end to end flow of training the algorithm. • Create ensemble method to get the most accurate prediction. • Testing. • Documentation.

Table 02 – Description of personal and facilities

5 BUDGET

Printouts and Binding	3500.00
Photocopies	500.00
Files and Note Books	500.00
A4 Sheets	500.00
Telephone & Internet Charges	3000.00
Total project cost estimate	7500.00

Table 03 – Proposed budget

6 REFERENCES

[1] [Online] Available: http://whatis.techtarget.com/definition/machine-learning [Accessed: March 8th, 2016]

[2] [Online] Available: https://docs.wso2.com/display/GSoC/Project+Proposals+for+2016#ProjectProposalsfor2016-Proposal6:[ML]PredictiveanalyticswithonlinedataforWSO2MachineLearner [Accessed: February 29th, 2016]

[3] [Online]. Available: https://docs.wso2.com/display/ML110/Features[Accessed: February 29th, 2016]

[4] [Online]. Available: The Wall Street Journal, Thomas H. Davenport [Accessed: March 1st, 2016]

[5] [Online]. Available: http://arxiv.org/pdf/1412.5720.pdf[Accessed: March 1st, 2016]

[6] [Online]. Available: http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/ [Accessed: March 3rd, 2016]

[7] [Online]. Available: Ensemble Learning, Martin Sewell[Accessed: March 8th , 2016]

[8] [Online]. Available: https://docs.wso2.com/display/GSoC/Project+Proposals+for+2016#ProjectProposalsfor2016-Proposal4:[ML]EnsembleMethodsSupportforWSO2MachineLearner [Accessed: March 8th, 2016]

[9] [Online]. Available: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm[Accessed: March 8th, 2016]

[10] [Online]. Available: https://spark.apache.org/docs/1.4.1/mllib-linear-methods.html#streaming-linear-regression[Accessed: February 29th, 2016]

[11] [Online]. Available: https://spark.apache.org/docs/1.4.1/mllib-clustering.html#streaming-k-means[Accessed: March 8th, 2016]

[12][Online]. Available: Constrained Clustering, Sugato Basu, Ian Davidson, Kiri L. Wagstaff [Accessed: March 8th, 2016]

[13][Online]. Available: Support Vector Machine, Jason Weston [Accessed: March 8th , 2016]

[14][Online]. Available: Making Large-Scale SVM Learning Practical,Thorsten Joachims[Accessed: March 1st, 2016]

[15] [Online]. Available: https://spark.apache.org/docs/1.4.1/mllib-guide.html[Accessed: March 8th, 2016]

[16][Online]. Available: Ensemble Methods in Machine Learning, Thomas G Dietterich[Accessed: March 7th, 2016]

7 APPENDICES

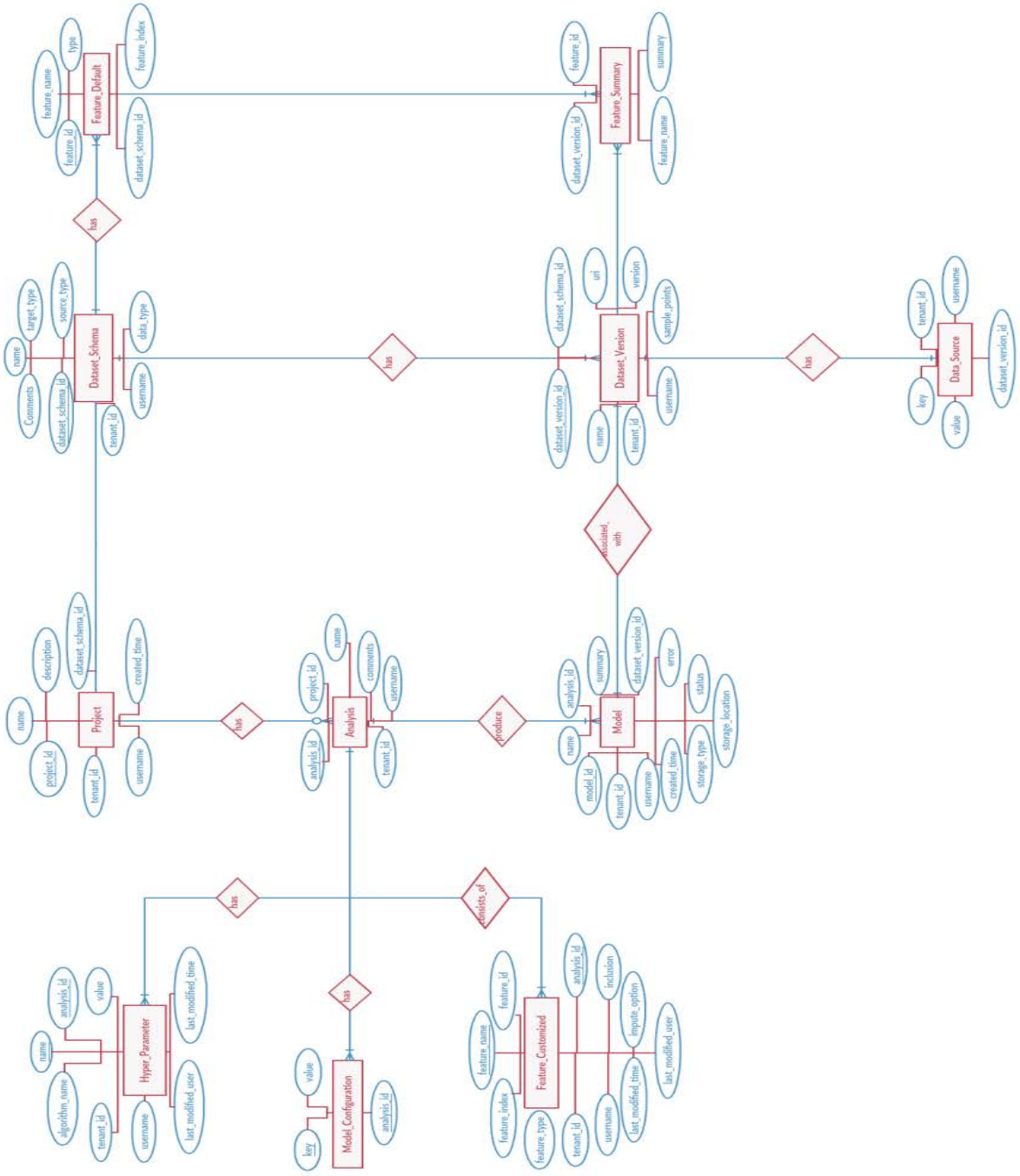

Figure 04 – WSO2 machine learner data base ER diagram