**Computing Project Proposal**

**Student's Name: Roland Hudson  
Student's Number: H00056278  
Student's Email Address:** *roland.hudson@online.liverpool.ac.uk*

**Project Title:** *Big Climate Data Analytics: Effective Knowledge Discovery from Colombia’s Weather Data*

**Dissertation Deadline: December 15th, 2018**

**Proposal Submission Date: May 17th, 2018**

**Version Number of the Proposal: 20180508**

**DA Class ID:** UKL1.CKIT.702.H00028508

**Name of DA:** Yuri Demchenko

**RMT Class ID:** LAUR-906-2-201843-2

**Name of GDI:** Taly Sharon

**Ethics Response Form completed:** Yes

**The Programme**: MSc in Software Engineering

**Domain:** *CKIT-525-1 Big Data,*

*CKIT-515-1 Sys Anls Dsgn Using O-O Apprch,*

*CKIT-503-1 Programming the internet*

*CKIT-510-1 Object Oriented Program Java*

**Proposal approved by:** (*To be filled in by the DA*)

**Date of the approval:** (*To be filled in by the DA*)

**Approval confirmed in MiTSA by the Lead Faculty (Dissertation):**: (*To be completed by the Lead Faculty*)

**Sponsor's Details:** *N/A*

**Sponsor's Background:** *N/A*

**Sponsor's Agreement:** *N/A*

**The Project Aims and Objectives:**

The goal for the project is a big climate data analytic system that enables knowledge-discovery (KD) and provides recommendations for construction strategies given a geographical location and the associated weather data.

In Colombia the weather is massively varied due to high altitude mountains, coastlines and effects of phenomena such as el niño. Tropical weather is unlike the weather in the Northern and Southern latitudes as there are no seasons instead daily variations dominate.

Typical construction is often unable to cope with regional and daily variations in weather, people live and work in uncomfortable conditions often too hot and too cold. The industrial approach is to install heating and cooling which is expensive, costly to run and produces emissions. Low energy construction strategies exist that can minimize or remove the need for heating and cooling through for example; orientation of buildings, sizing and positioning of openings, choice of materials and use of passive heating and ventilation.

These construction strategies are climate responsive. In the Northern and Southern hemispheres, it is relatively easy to identify what strategy to apply where because of the seasonality and lack of regional variation. The variability of weather in Colombia means it is very hard to define an approach for a specific location. Weather records from national, automated stations was recently made available by Colombia’s national weather agency IDEAM.

This proposal aims to design and implement a big data system that enables the analysis and visualization of Colombia’s weather data using clustering, self-organizing maps and delta-maps. The goal is to use these techniques to generate localized approaches to building design and construction can respond to the unique weather conditions in Colombia.

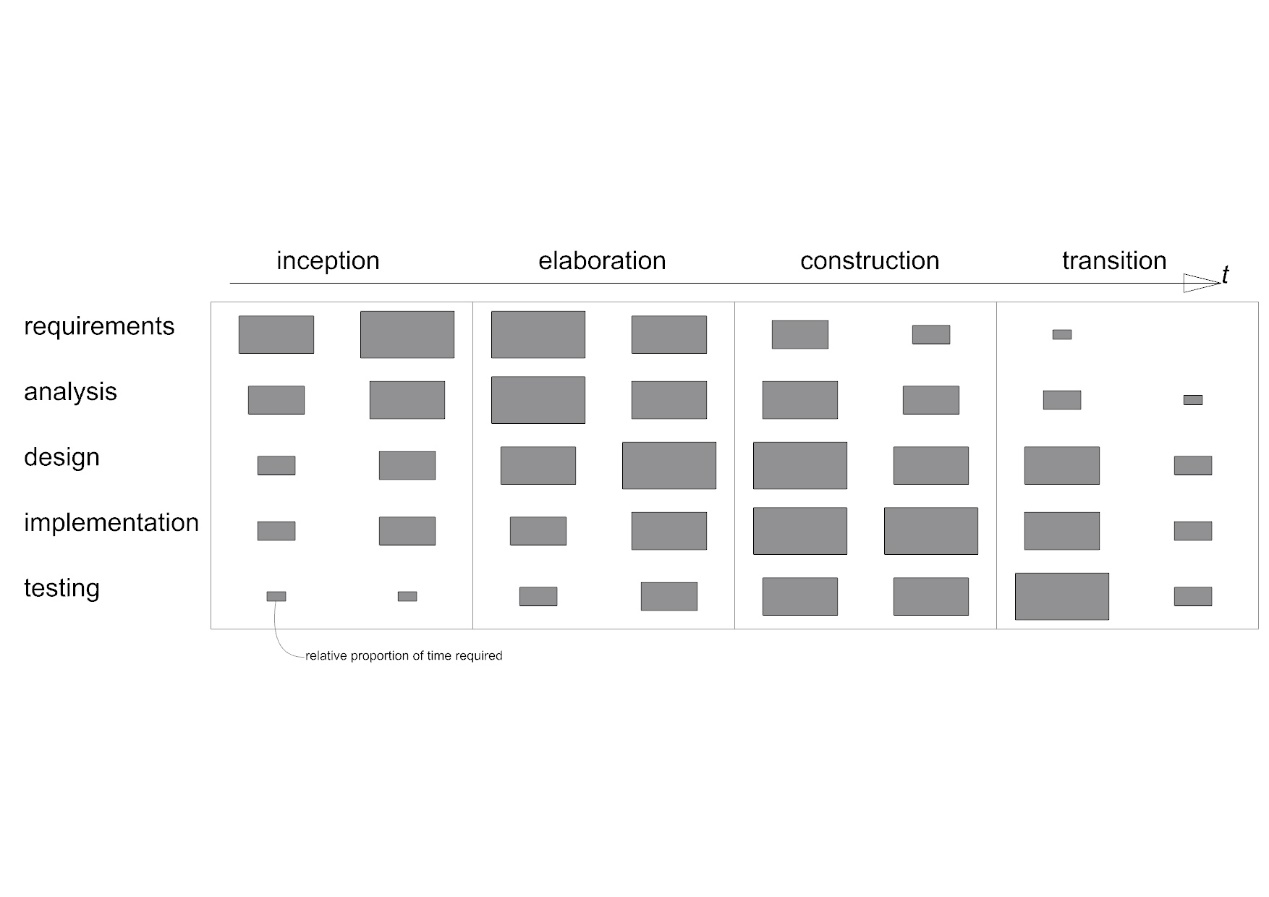
To achieve this goal a big data system is proposed that follows current best practices for the storage, processing, analysis, management and visualization of the data. Specific focus will be on enabling the analytics and visualization that enables KD through Data mining. Knowledge from the data will support decision making for the design and construction of buildings to potentially improve living conditions (quality of life and wellbeing) and reduce energy consumption in buildings.

|  |  |
| --- | --- |
| **Step** | **Short Description** |
| Hypothesis | KD techniques combined with a big weather data framework can help define localized approaches to building design and construction that improve living conditions and reduce energy consumption in Colombia. |
| Research Methods | **Literature review:** Big Data Architecture, Big Data Analytics methods, KD process models, Spatio-temporal data mining, Data mining techniques applied to meteorology  **Developing the application:** Agile Model Driven Development |
| IT Artefact | An application that facilitates big data analytics for Colombia’s recently released weather data. Through analytics and visualisation, the application should enable data exploration and KD with the goal of providing recommendations for construction strategies dependent on geographical location and related historical weather data.  The application should also provide efficient storage, processing, management and security. |
| Evaluation | Verification, validation and testing of the application using statistical comparisons and review by domain experts  Checking if useful localised construction approaches can be generated.  Validation of application output:   1. Statistical comparison of different knowledge discovery methods applied 2. Quality measures for methods applied. (Distance metrics for clustering) 3. Test cases – identified by domain experts 4. Interpretation of results by domain experts |

**Project Outline**

**Literature search and review** to identify of KD principles applicable in weather, methods for evaluating KD techniques, existing big data architecture and frameworks in scientific applications and specifically in climatology and meteorology.

**Development phases and workflows time distribution:**

****

**System development process:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Activity** | **Techniques** | **Deliverables** | **Diagrams** |
| **Requirements capture and modelling** | Text descriptions of use cases and requirements  Use case modelling  Architectural modelling, prototypes | Use case model, requirements list, initial architecture | Use case, Package |
| **Requirements analysis** | Use cases analysed to extract required objects. Interactions between objects identified – communication diagrams developed | Analysis models | Class, Object, Communication |
| **System and architecture design** | Design patterns identified | Overview design and implementation architecture | Package, Component, Deployment, Class |
| **Class design** | Class and object modelling, Interaction modelling. State modelling, Design Patterns  prototypes | Design models | Class, Object, Sequence, State machine, Package |
| **Interface design** | Class and object modelling, Interaction modelling, State modelling, Design Patterns, prototypes | Design models, interface specification | Class, Object, Sequence, State machine, Package |
| **Data management design** | Class and object modelling, Interaction modelling, State modelling, Design Patterns, prototypes | Design models, data specification | Class, Object, Sequence, State machine, Package |
| **Construction** | Programming, component reuse | Constructed system, documentation |  |
| **Testing** | Programming, test planning and design, testing | Test plans, test cases, tested system |  |

**Evaluation:** Results and output of the application evaluated using statistical methods identified for each of the implemented knowledge discovery methods. Application is evaluated through verification, validation and testing – tests identified during the requirements specification and revisited through the prototyping stages. Domain expert(s) presented with a series of studies and results from the application, opinion of experts captured and summarised.

**Literature Survey / Resources’ List:**

**Big data architecture and frameworks:**

* Software architecture for cloud and big data applications (Bahsoon *et al.*, 2017)
* Attempts to generalise big data workflows in eScience (Buyya *et al.*, 2016)
* Workflows for scientific big data management applicable weather data (Rodriguez and Buyya, 2017).
* Big data feature model and Domain-Driven Design is discussed (Avci Salma, Tekinerdogan and Athanasiadis, 2017)

**Knowledge discovery methodology:**

Knowledge Discovery in Data is broadly defined as (Begoli and Horey, 2012); 1. Collection, storage and organisation of data. 2. Understanding and application of analytic methods. 3. Understanding the problem domain

Three principles for KDD (Begoli and Horey, 2012):1. Support of analysis methods. 2. One size does not fit all. 3. Make data accessible

**Knowledge discovery applicable methods:**

* Clustering methods for climate classification (Forsythe, Blenkinsop and Fowler, 2015) (Netzel *et al.*, 2016)
* Self-organising maps (Jayaratne *et al.*, 2017) (Liu, Weisberg and Mooers, 2006).
* Delta-maps (Fountalis, Bracco and Dovrolis, 2014).
* Need for better visualisation and user interaction (Assunção *et al.*, 2015).

**Scholarly Contributions of the Project**

The project proposes a framework for extracting useful knowledge from weather data in Colombia. Through the application of spatiotemporal knowledge discovery methods the project aims to determine localised patterns that can indicate appropriate climate responsive strategies for the design and construction of buildings. Implementing these strategies can improve building performance in terms of energy consumption and comfort for occupants. The proposed solution will contribute to environmental design (architecture and engineering of buildings) by supporting decision making in the design process with knowledge acquired from weather data.

The test application aims to combine knowledge discovery and big data technologies to identify patterns in complex tropical climates and use these to improve occupant comfort and energy performance of buildings in the region. Colombia is the test case but similar situations in other tropical regions exist, the project seeks to generalise the findings so the methodologies used may be transferred to other contexts.

**Description of the Deliverables:**

Literature review

UML documentation of the system development process:

* Requirements list
* Use case diagrams
* Package diagrams
* Class, object and component diagrams
* Deployment diagram
* Sequence diagrams
* Test plans
* Test cases

Climate big data analytic system that enables knowledge-discovery.

Example recommendations for construction strategies given a geographical location and the associated weather data.

Visualisation of results / output.Set of statistical test results of outputs from the application.

Analysis of results and system by domain expert(s).

Generalised, transferable description of system architecture

**Evaluation Criteria:**The big question to be evaluated is: can the system support environmental design decision making based on weather data?

This will be addressed in the following ways.

* Statistical comparison of the different knowledge discovery methods used
* Statistical comparison between individual knowledge discovery methods
* Opinion by domain experts(s) – presentation of results and analysis

Statistical methods will be identified as part of the literature review of the knowledge discovery techniques. For example, to compare clustering results distance metrics would be used. Secondary evaluation will involve the verification, validation and testing of the IT artefact. Working with a reduced data set, independent component tests will be defined and executed during development. Components will be integrated and the system tested for issues in component interaction and interface problems and if the system meets functional and non-functional requirements. The system will undergo evaluation with the full dataset to check for omission in the requirements definition and for overall performance.

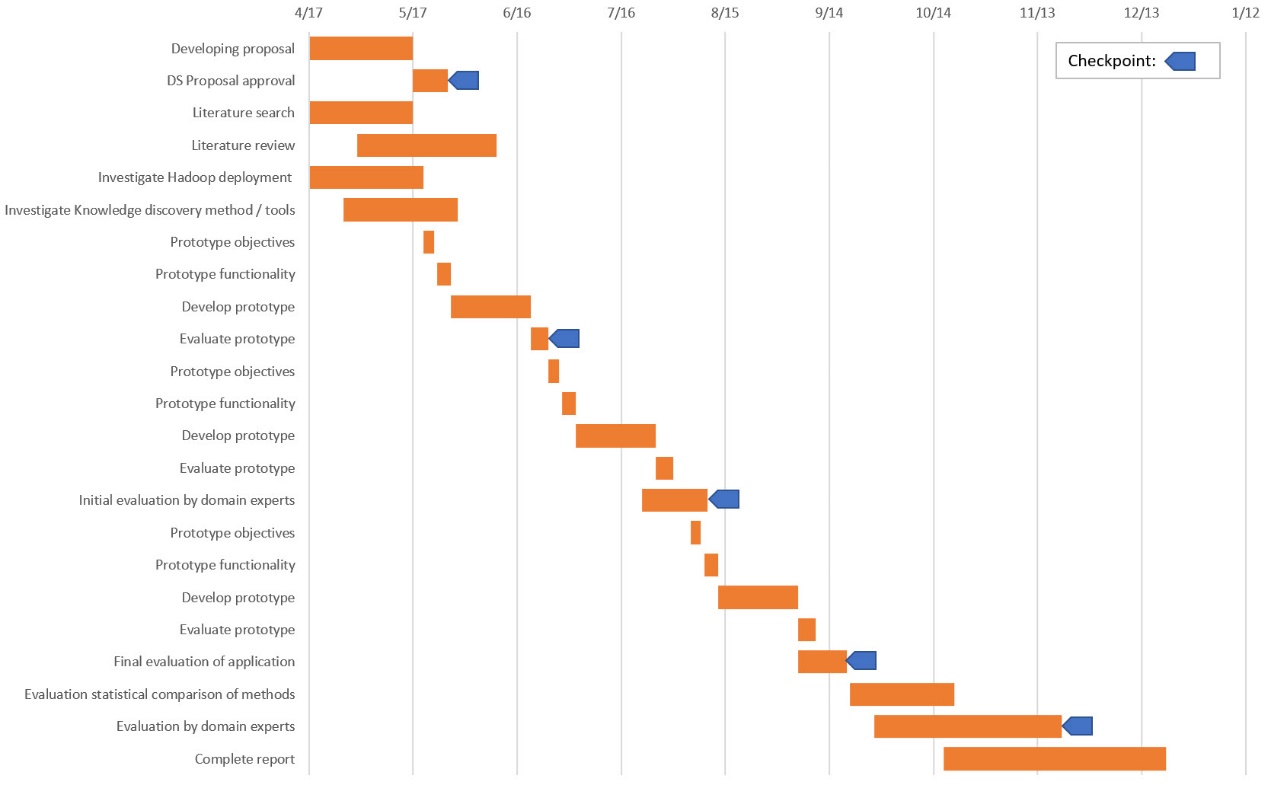
**Resource Plan:**Cloud Hadoop deployment

Personal computer

Domain experts to be invited for evaluation

Colombian Weather Dataset from IDEAM (in hand)

**Project Plan and Timing**

****

**Risk Assessment:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Indicator** | **Impact 1-15 (probability x effect)** | **Alleviation strategy** |
| Underestimation of development times | Milestone 2 overrun | 12 | Limit scope of requirements, by setting SMART goals |
| Underestimation of the size of the software | Milestone 2/3 overrun, failure to fix defects | 12 | Use of standard tools and components where possible. Investigate best practice with existing tools and frameworks |
| Technical difficulties with Hadoop and components | Problems with data ingestion / basic analysis | 4 | Investigate use of off-the-shelf tools as part of literature search. Start prototyping early and prioritise experiments with Hadoop |
| Technical difficulties with KD methods in Hadoop framework | Problems with first attempts at analytics | 4 | Investigate the use of off-the-shelf tools as part of literature search. Include experiments with KD methods as early as possible in the prototyping process. |
| Onsite hardware failure | Personal laptop failure | 5 | Backups local and offsite, establish access to secondary computer |
| Offsite hardware / platform failure | No access to cloud services | 5 | Ensure chosen platform has deployment on multiple servers and sites and strategy for backups |
| DA leaves Liverpool | notification from the university | 1 | Read Liverpool / Laureate protocols for this circumstance so a replacement can be quickly found. Continue working with initial DA if permitted. |

**Quality Assurance:**

Five milestones are identified in the project plan to coincide with acceptance of the proposal and end of prototype and evaluation phases. Each prototype development phase will be structured around agile model driven development using UML, leveraging continuous documentation to define, record and monitor the process. Tests at the component, integration and system tests will be defined and used to drive application development. Evaluation by domain experts is planned at two milestones. Regular review by DA and at milestones.

**References**

Assunção, M. D. et al. (2015) ‘Big Data computing and clouds: Trends and future directions’, Journal of Parallel and Distributed Computing. Academic Press, 79–80, pp. 3–15. doi: 10.1016/J.JPDC.2014.08.003.

Avci Salma, C., Tekinerdogan, B. and Athanasiadis, I. N. (2017) ‘Domain-Driven Design of Big Data Systems Based on a Reference Architecture’, in Software Architecture for Big Data and the Cloud. Elsevier, pp. 49–68. doi: 10.1016/B978-0-12-805467-3.00004-1.

Bahsoon, R. et al. (2017) ‘Chapter 1 - Introduction. Software Architecture for Cloud and Big Data: An Open Quest for the Architecturally Significant Requirements’, in Software Architecture for Big Data and the Cloud, pp. 1–10. doi: https://doi.org/10.1016/B978-0-12-805467-3.00001-6.

Begoli, E. and Horey, J. (2012) ‘Design principles for effective knowledge discovery from big data’, in Proceedings of the 2012 Joint Working Conference on Software Architecture and 6th European Conference on Software Architecture, WICSA/ECSA 2012, pp. 215–218. doi: 10.1109/WICSA-ECSA.212.32.

Bracco, A. et al. (2017) ‘Advancing climate science with knowledge-discovery through data mining’, npj Climate and Atmospheric Science, 1(1), p. 4. doi: 10.1038/s41612-017-0006-4.

Buyya, R. et al. (2016) ‘Chapter 18 – eScience and Big Data Workflows in Clouds: A Taxonomy and Survey’, in Big Data, pp. 431–455. doi: 10.1016/B978-0-12-805394-2.00018-0.

Chouksey, P. and Chauhan, A. S. (2017) ‘Weather Data Analytics using MapReduce and Spark’, International Journal of Advanced Research in Computer and Communication Engineering, 6(2). doi: 10.17148/IJARCCE.2017.6210.

Faghmous, J. H. and Kumar, V. (2014) ‘Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities’, in Chu, W. (ed.) Springer Berlin Heidelberg, pp. 83–116. doi: 10.1007/978-3-642-40837-3\_3.

Fang, W. et al. (2014) ‘Meteorological data analysis using MapReduce.’, The Scientific World Journal, 2014, p. 646497. doi: 10.1155/2014/646497.

Forsythe, N., Blenkinsop, S. and Fowler, H. J. (2015) ‘Exploring objective climate classification for the Himalayan arc and adjacent regions using gridded data sources’, Earth System Dynamics, 6(1), pp. 311–326. doi: 10.5194/esd-6-311-2015.

Fountalis, I., Bracco, A. and Dovrolis, C. (2014) ‘Spatio-temporal network analysis for studying climate patterns’, Climate Dynamics, 42(3–4), pp. 879–899. doi: 10.1007/s00382-013-1729-5.

Jain, H. and Jain, R. (2017) ‘Big data in weather forecasting: Applications and challenges’, in 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). IEEE, pp. 138–142. doi: 10.1109/ICBDACI.2017.8070824.

Jayaratne, M. et al. (2017) ‘Apache spark based distributed self-organizing map algorithm for sensor data analysis’, in IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society, pp. 8343–8349. doi: 10.1109/IECON.2017.8217465.

Liu, Y., Weisberg, R. H. and Mooers, C. N. K. (2006) ‘Performance evaluation of the self-organizing map for feature extraction’, Journal of Geophysical Research: Oceans, 111(5). doi: 10.1029/2005JC003117.

Netzel, P. et al. (2016) ‘On Using a Clustering Approach for Global Climate Classification’, Journal of Climate, 29(9), pp. 3387–3401. doi: 10.1175/JCLI-D-15-0640.1.

Rodriguez, M. A. and Buyya, R. (2017) ‘Chapter 18 – Scientific Workflow Management System for Clouds’, in Software Architecture for Big Data and the Cloud, pp. 367–387. doi: 10.1016/B978-0-12-805467-3.00018-1.