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

By

Roland Hudson

A DISSERTATION

Submitted to

The University of Liverpool

in partial fulfilment of the requirements

for the degree of

MASTER OF SCIENCE

24/11/2018

ABSTRACT

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

By

Roland Hudson

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.

Recommended content:

* Introduce into the research or technology area addressed in the dissertation
* Identify specific problems and challenges identified in your work
* Describe what is achieved and proposed in the project
* Briefly describe the structure of the report
* Mention what is the application area and benefits of the proposed solutions

DECLARATION

[check with the recent templates]

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

ACKNOWLEDGEMENTS

[any acknowledgement you want to express]

TABLE OF CONTENTS

# Chapter 1. Introduction

## Scope

This dissertation how a big data climate pattern detection system design strategies in Colombia

## Problem statement

Low-energy construction strategies exist that can minimize or remove the need for heating and cooling in buildings. For example; orientation of buildings, sizing and positioning of openings, choice of materials and use of passive heating and ventilation. These techniques require an understanding of local and regional climate conditions across different time frames. The project proposes that knowledge discovery applied to weather data can link specific design strategies with a specific location and time frame.

Detecting climate patterns in the Colombian context is challenging due to weather variations caused by extreme changes in altitude over short distances, proximity to different ocean bodies and phenomena such as el Niño and la Niña. This underlying complexity is complicated further by subtle and inconsistent seasonal patterns associated with tropical latitudes.

The Colombian context contrasts with Northern and Southern latitudes where consistent seasonal variation dominates the climate making it easier to identify what design strategy to apply. Colombia’s complex climate patterns make identifying appropriate, localized, low-energy construction strategies difficult. 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. To correct these issues heating and cooling need to be installed which is expensive, costly to run and produces emissions.

### Complexity of existing workflow

For an architect or engineering the current workflow to determine what low-energy construction strategies to use is a multi-step approach:

* Analysis of climate data - analysis of historical weather data compiled into files representing typical meteorological years (TMY) with a range of variables stored for each hour of the year.
* Biological evaluation – data is plotted on a psychrometric chart (physical and thermal properties of moist air) to diagnose the hours of the year lying outside a predefined zone of thermal comfort.
* Identifying technological solutions (design strategies) – in response to the biological evaluation technological approaches can be identified to ensure more time within the comfort zone. For example; site selection and orientation to gain or minimise solar radiation.
* Developing the architectural application – the designer synthesises the previous three steps into a design proposal.
* Simulation may be undertaken to confirm the design approach or optimise a chosen strategy.

The designer must also consider usage patterns of the building making the process more complex. Buildings are rarely occupied constantly, depending on use, occupancy can vary daily (residential buildings are often occupied evenings and night time), weekly (office buildings not in use at weekends) and through the year (schools and universities have seasonal holiday periods). The design approach for two buildings with different uses in the same location will not be the same. Zones within certain buildings will have different orientations, usage patterns and activity types, each zone will have varied design needs.

## Approach

|  |  |
| --- | --- |
| **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 |

Table Key Research methods used in the dissertation

**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:**

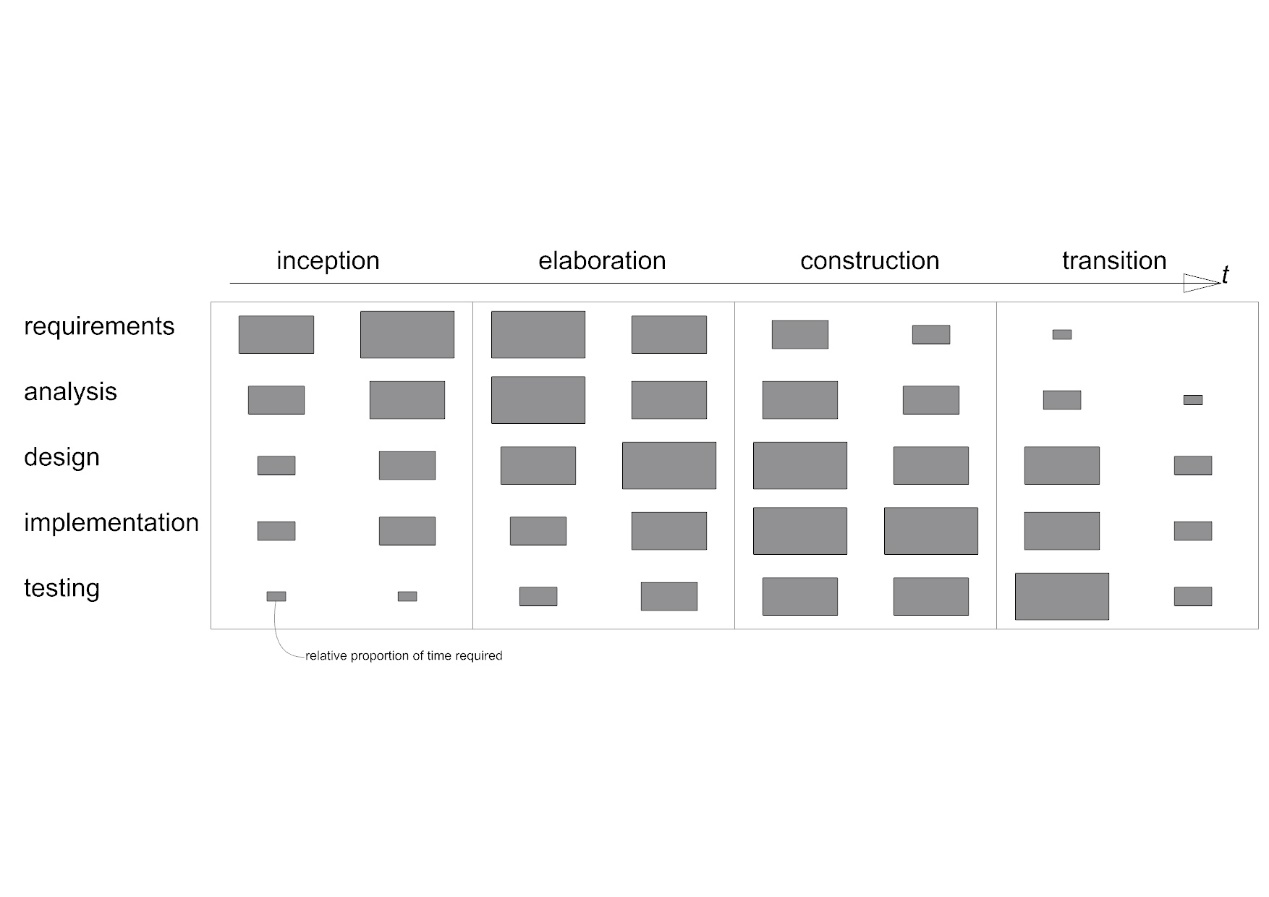
****

Figure process diagram

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |  |

Table System development process

**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 review:

Low energy architectural design strategies to link weather patterns and design criteria. Weather data, spatiotemporal data mining, knowledge discovery with weather data. Workflow management for scientific big data systems. Big data tools and applications in climate science. Big data application architecture and components.

### Application development:

UML based agile model driven development design

### Qualitative evaluation of artefact by domain experts:

Presentations to and interviews with domain experts

### Quantitative evaluation of output from artefact results:

Statistical comparison of different knowledge discovery methods. Quality measures for methods applied (distance metrics for clustering). Results of test cases (identified by domain experts) checked against expected results. Results interpreted by domain experts.

## Outcome (description the goals)

The goal for the project is a big climate data analytic system that enables knowledge-discovery (KD) to support decision making in the design and construction of buildings in Colombia. Specifically, the project aims to enable the search for patterns in climate data that can be linked to localized, climate-responsive design and construction strategies. Application of these strategies can lead to buildings that perform better in terms of production costs, life-time running costs (reduced heating and cooling) and occupant comfort.

The project proposes that by combining a big data workflow management infrastructure with spatiotemporal data mining techniques localized approaches to building design and construction that respond to the unique weather conditions in Colombia can be identified.

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.

# Chapter 2. Background and review of literature

## Background

### Challenges for big data analytics applied to climate science and construction

The use of big data analytics is underutilised in the domain of weather data (Jain and Jain, 2017), better big data weather forecasting could benefit the construction industry by helping build more efficient buildings and provide an “*improved quality of life”.* The construction industry is responsible for massive amounts of greenhouse gases (Dimoudi and Tompa, 2008). Through their lifecycle buildings consume 50% of all total energy demand and contribute 50% of all CO2 emissions.

Primary goal for environmental construction is to reduce the energy consumed by buildings, estimated to be around 40% of the total global energy consumption (Omer, 2008). Much of this energy is expended on heating lighting and cooling. Energy efficiency can be defined as the minimising the amount of energy consumed to achieve thermal comfort for occupants. Currently the energy required to maintain thermal comfort accounts for 60-70% of energy consumed in non-industrial buildings (Omer, 2008). Better understanding of human response to climatic context (bioclimatic design) can result in buildings that require less energy for heating and cooling (Olgyay and Olgyay, 2015, p11).

## Literature review

### Low Energy Environmental Design Strategies

#### Human thermal comfort

Can be understood as a combination of temperature, relative humidity, air movement and radiant temperature, giving a state of mind where a person requires no change in current conditions (ASHRAE, 2013) or a state where minimal extra energy is require to maintain the human balance (Manzano-Agugliaro *et al.*, 2015).

#### Psychrometric chart

Is used to map interrelationships of thermal conditions of the environment (Figure 2). A zone of human thermal comfort can be plotted following standard guidelines (ASHRAE, 2013). Hourly data points can be plotted on the chart, where they fall outside of the comfort zone the design of the building and/or services must be adapted to provide comfort.

The psychrometric chart was adapted by Givoni (1992) as the building bioclimatic chart and later Manzano-Agugliaro *et al.* (2015) to include specific zones representing strategies that can be applied to a design to extend the zone of comfort. Conventional heating and air conditioning are recommended only at extremes. The strategies include what Lechner (2009, p9) describes as tier 1: basic building design (building orientation, position and size of openings, material specification) and tier 2: passive systems (passive solar heating, night-time flush cooling). Correct design decision making at these levels can reduce building energy consumption by up to 80% (Lechner, 2009, p9).

Each design strategy can be further specialised into a series of instrumental techniques (Manzano-Agugliaro *et al.,* 2015) which can include regionally specific and traditional construction and more experimental methods. Climate Consultant software (Milne, Liggett and Benson, 2009) generates a prioritised subset of recommendations from a set of 68 design guidelines each associated with a zone on the psychrometric chart.

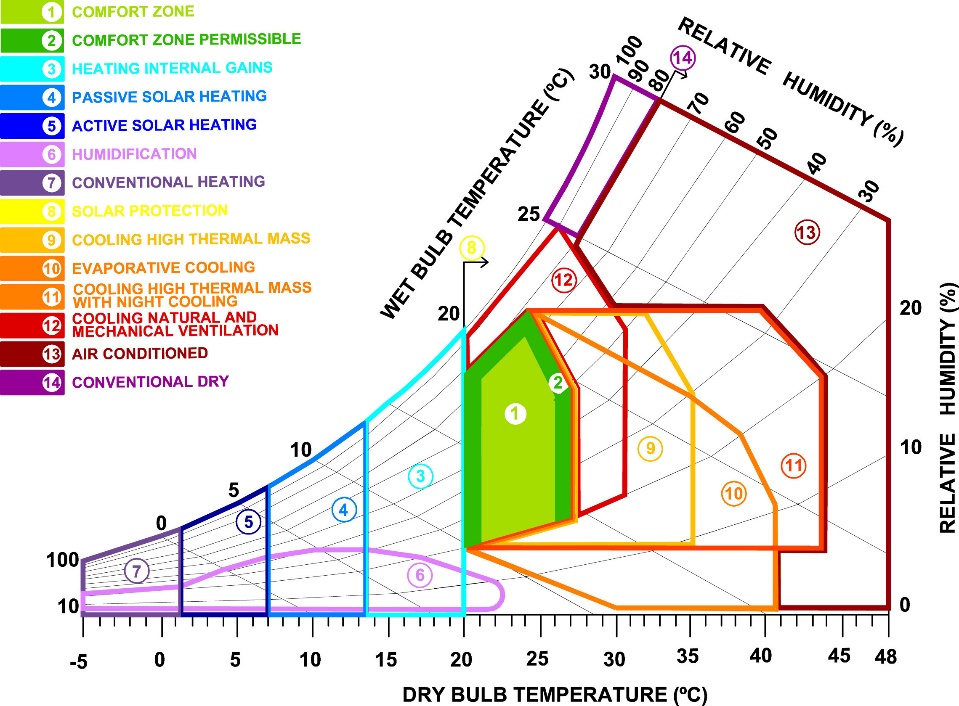


Figure 2 Manzano-Agugliaro et al. (2015) adapted version of the psychrometric chart

Figure 1. Manzano-Agugliaro *et al.* (2015) adapted version of the psychrometric chart.

### Weather data is spatiotemporal

Data collected for climate science is classed as spatiotemporal (ST) data (Atluri, Karpatne and Kumar, 2017). Mining and knowledge discovery with ST data differs from classical data mining due to its properties and the variety of data types (Faghmous and Kumar, 2014) . ST data is heterogenous, it is not identically-distributed, instead ST data demonstrates non-stationarity in space and time. Auto-correlation exists in ST data two nearby locations are not independent but are correlated. ST data can be categorised into four types; event data (start and end of heavy rainfall), trajectory data (path of a cyclone), point data (temperature measured in a moving set of weather balloons) and raster data (temperature measured across a fixed set of weather stations).

Classical data mining uses features with labels (Atluri, Karpatne and Kumar, 2017) in ST instances can be defined as points, trajectories, time-series, spatial maps and raster. Search for similarities between these instances involves clustering, classification, pattern discovery and relationship mining.

Various data mining methods applicable to climate data are described by Atluri, Karpatne and Kumar (2017). *Relationship mining* involves linking changes in one variable to other phenomena. *Clustering* on instances and

ST-DBSCAN is recommended for finding anomalies. *Frequent pattern mining* includes searching *motif patterns* in time-series and in networks for sets of distant locations experiencing similar climatic conditions with consistent temporal activity. Change detection can identify transitions or deviations in behaviour. Faghmous and Kumar (2014) add *network-based analysis* to undertake relationship and pattern mining in gridded and non-gridded climate data sets. *Pattern mining* for Faghmous and Kumar includes searching for user defined patterns, empirical orthogonal functions and spatiotemporal clustering.

### Examples of applied ST data mining methods applied to climate data

* Clustering methods have been successfully applied to climate classification (Forsythe, Blenkinsop and Fowler, 2015) (Netzel *et al.*, 2016)
* Self-organising maps (SOM)’s have been used to extract features from data (Liu, Weisberg and Mooers, 2006) and applied to metrology and oceanography (Liu and Weisberg, 2011)(Liu and Weisberg, 2005).
* Delta-maps (Fountalis, Bracco and Dovrolis, 2014) (Bracco *et al.*, 2017) group nodes in a network according to homogeneity, these have beenapplied to precipitation and sea surface temperatures. Robustness analysis of networks generated can be evaluated using link maps, area strength and s-core decomposition.

### Challenges for ST data mining

Dealing with interdependencies at multiple scales within climate data is complex and means global studies cannot be used to understand long-term local impacts (Faghmous and Kumar, 2014). Relationships in climate data may be long range and multivariate, many space-time-variable subsets exist where relationships may be found. This spatiotemporal variability makes clustering with ST data challenging. Similarly, anomalies and extremes in climate data need to be understood as multivariate cumulative extremes. Faghmous and Kumar suggest better methods are needed for validation of ST data mining. Significance testing needs randomization tests that do not break the inherent autocorrelation and performative measures are required to compare unsupervised STDM.

For Atluri, Karpatne and Kumar (2017) key challenges involve finding methods for combining multi-modal data sets and controlling granularity of partitioning to ensure substructures are not overlooked.

The need to integrate domain theory and expertise is acknowledged (Karpatne *et al.*, 2017) as a key strategy that could accelerate knowledge discovery in data science particularly where complex physical phenomena are involved. The potential for knowledge discovery in climate science has not yet been fully realised (Bracco *et al.*, 2017).

### Knowledge discovery with climate data

Knowledge Discovery in Data is 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. To enable knowledge discovery Begoli and Horey recommend applications are made to allow researchers easy ways to interact, explore and analyse data. A variety of analysis methods should be supported inclung statisitical, data mining, machine learning, visualisation and visual analysis. Different data storage and processing mechanisms should be provided to support a variety of intermediate data structures (structured and semi-structured) required by different ananlysis methods. Data should be made as accessible as possible by using open standards, lightweight architecture and APIs to expose results.

### Workflow management for scientific big data systems

An approach for supporting scientific data analysis on large data sets in the cloud is workflow management systems (WMSs) (Zhou, He and Ibrahim, 2016). These processing tools enable acquisition of resources, scheduling of tasks, execution of data analysis and visualisation on distributed resources. Workflows are defined as a series of linked tasks in the form of directed acyclic graph (DAG).

Specific WMS platforms exist (see Askalon, Kepler, Taverna and Pegasus) some provide a graphical interface to assemble workflows costing of loops, conditionals and graph constructs. Originally developed for grid computing many have been extended to take advantage of the cost-effectiveness of cloud platforms and applicable to climate science (figure 2) (Rodriguez and Buyya, 2017).

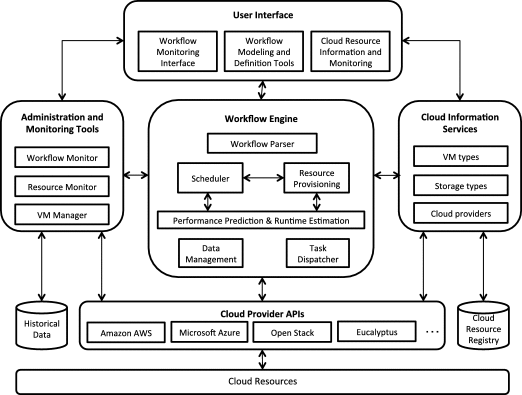


Figure Reference architecture of a WMS

### Big data Tools

#### General applications on Weather Data

Several precedents exist describing the application of big data tools to process and undertake simple analysis on climate data. Three tools frequently occur in the literature Hadoop, Apache Spark and MapReduce.

* Apache Spark was used to process at-rest weather data to determine averages for a range of variables across a set of weather stations (Jayanthi and Sumathi, 2017).
* Hadoop and Spark were used to find points with similar weather conditions using Euclidean distances based on weather attributes using the NOAA hourly land based data set (Rodenburg and Maria Fiore, 2017).
* MapReduce was compared to Spark to calculate minimum, maximum and average values of weather parameters using data from NOAA (Chouksey and Chauhan, 2017). For smaller datasets differences in performance were less pronounced, above 8GB Spark was faster.
* Hadoop and MapReduce were used on NOAA data (Dagade *et al.*, 2015) to compare the performance of Pig and Hive to average data for each station for a single variable.
* MapReduce was used to process NOAA data (Mariam Varghese, 2015).
* K-means with MapReduce was used to cluster weather data from China (Fang *et al.*, 2014). Different size datasets were tested to evaluate processing speeds (250mb-2GB) and compared against another clustering algorithm.

#### More specific applications

A self-organising map (SOM) (a type of artificial neural network trained using unsupervised learning) was implemented using Apache Spark and analysed IoT data (Jayaratne *et al.*, 2017). MapReduce enabled prediction using an artificial neural network and k-means clustering with air quality data stored with HBase, in an implementation designed to support decision making in traffic regulations in Marrakesh (El Fazziki *et al.*, 2015).

### Big data application architecture and components

Avci Salma, Tekinerdogan and Athanasiadis (2017) define a big data feature model and a generic reference architecture that can be developed into an application architecture using domain-driven design. Features and architecture components can be selected based on design rules determined by the domain.

## K-means clustering applied to climate classification

Recent research suggests that unsupervised learning methods are applicable to climate classification. Fovell & Fovell (1993) studied hierarchical clustering in the US and sought a best method based on the minimising of bias in terms of method, latent and information. Redundancy problem is discussed when two or more highly corelated variables are included (little unique information added and repeats data magnifies the) and PCA used scaling variables, withholding then adding variables to observe the impact

Degaetano's (1996) study sought to develop an ecosystem management and planning guide by defining mesoscale climate zones in the north-eastern US. K-means was compared to Ward’s clustering technique and improvements were found in spatial distribution and homogeneity of clusters. The findings suggested k-means could produce stable clusters with minimal information bias.

Rhee *et al.* (2008) used k-means as part of a multi-step approach to delineate climate regions in the Carolinas that combined in-situ (weather station data) with remotely senses and spatially distributed data. K-means was integrated within a more complex workflow (hierarchical followed by non-hierarchical then decision trees trained on results that then classify remotely sensed data) and the study demonstrated the validity of the method for establishing clusters that were subsequently used for supervised classification of data. PCA is not used as it is lose important information in monthly time series, interpretable distance measure can be use and truncation of PCAs is not considered.

All three-above use pseudo f and CCC and suggest a consensus approach – first with hierarchical to define centroids and then non-hierarchical to determine the clusters

Clustering applied to assess climatic influences on water resources and food security in the Himalayas (Forsythe, Blenkinsop and Fowler, 2015). It showed clustering could be used to characterise the bias of gridded datasets and undertake meteorological reanalyses of climate models. K-means defined eight sub-regional climate classifications and further increases in cluster numbers defined subdivisions to each macro zone.

Netzel *et al.*, (2016) studied 32 different clustering methods and compared them to the KG Classification. The study concluded that clustering could find 50% of the climate types defined by the KG classification. The remaining classes differed in climatic character and spatial distribution but were shown to be more homogeneous and more distinct than KG climate types.

Zscheischler, Mahecha and Harmeling (2012) used k-means clustering with subsets of 5 normalized variables. When k-means was used with climate and vegetation variables similar clusters to the KG zones could be generated.

## Apache Spark

Key features of Spark Streaming (Zaharia *et al.*, 2013) are:

* fault tolerance through a parallel recovery mechanism
* tolerance of stragglers (slow nodes)
* a processing model based on discretised streams (D-Streams)
* integration with batch and interactive query models such as MapReduce
* dynamic load balancing (Das, Zaharia and Wendell, 2015)

Fault recovery in other streaming systems is based on times data replication which is expensive in terms of time and hardware and results in long recovery times and problems for handling stragglers. Systems such as Storm, TimeStream MapReduce Online use continuous operators where operators receive each record, internal states are updated, and new record sent. The Spark Streaming / D-Streams model contrasts with the long-lived operators approach by structuring the computation as a sequence of stateless, batch processes issued at short time intervals. Resilient Distributed Datasets (RDDs) are used to keep data in memory and by tracking the graph of operations used to produce each RDD it can be recovered without on disk replication. Faults on nodes are handled using parallel recovery which means on failure of a node all other nodes in the cluster work to rebuild the lost RDDs.

Clustering options with spark

Clustering evaluation with spark WSSSe and BDDunn and BDSilhouette (Luna-Romera *et al.*, 2016)

On the selection of numbers of clusters (Thinsungnoen *et al.*, 2015) (Kodinariya and Makwana, 2013) and (Nikolaou *et al.*, 2012)

## Summary

# Chapter 3. Analysis and Design

## Design methodology

The design process is broken into three major development increments. Each increment lasts approximately 1 month and follows Agile Model Driven Development (AMDD) with UML (Ambler, 2004). First, over several days, high-level modelling will be used to understand the scope, requirements and potential architecture of the system. This is followed by a series of construction iterations each of which begins with a planning phase. Requirements will be ranked by priority and the highest priority implemented first. On a time-scale of hours, models will be produced to explore what should be built for the iteration and to estimate the time required. Issues identified in the planning models will be developed in more detail using just-in-time models, created in less than thirty minutes involving hand-sketched flow diagrams, sequence diagrams and class responsibility cards. Using these modelled details code will be written during the following hours or days using a test-first and refactor approach.

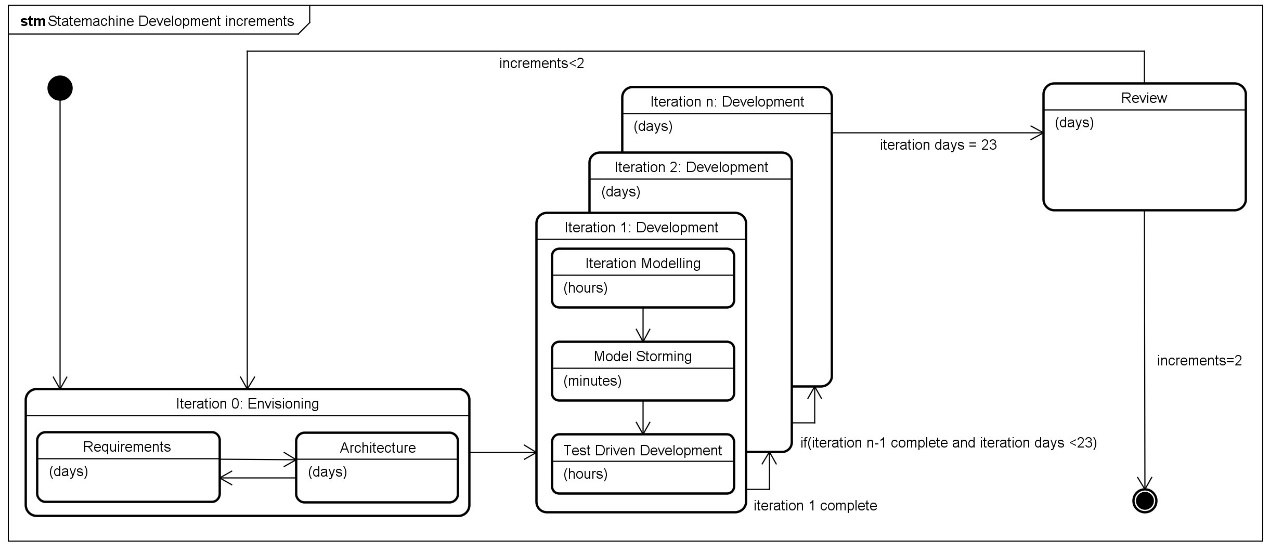


Figure Development increments (extended from Ambler (2004), p119).

## System actors

The primary actor in the system is an Environmental Designer (ED), an architect or engineer whose aim is to develop energy-efficient designs for buildings in specific geographical locations. The ED’s goal is to minimise the need for heating and cooling systems while maintaining thermal comfort for occupants reduce energy consumption.

To achieve these aims the ED needs to understand the interrelationship of occupant’s thermal comfort, topography, local weather conditions, annual solar path and larger scale climatic patterns and examining how these interact with the anticipated usage patterns of the building.

Equipped with this knowledge the ED can apply a hierarchy of design strategies to develop the design. The first level relates to designing the form and fabric of the building (for example the location, orientation, materials of the building). The next level involves identification of potential passive strategies (for example, designing direct solar heating and natural ventilation into a building). The final level is the specification of mechanical systems to heat and cool, potentially using renewable energy sources. The first two levels are the concern of this project.

Other actors identified are a system administrator and data scientist. The administrator’s role is the configuration of systems settings such as access and security. The data scientist working in the field of environmental design is also identified as a system user. In this case the use is like the ED but with additional goals of management of the datasets, developing and managing new analytical techniques, visualisation and statistical methods that the ED has access to.

## General Use Case Analysis

The ED’s goal is to run an analytic or data mining technique on a set of climate data to identify patterns that can be visualised and linked to with specific design strategies. An example analysis / data mining problem could be framed as:

*For the coastal regions of Northern Colombia what design strategies can be identified for afternoon weather conditions during the first three months of the year, using a k-means clustering approach. Visualise the clusters on a map and indicate the design strategies and considerations applicable to each cluster.*

To address this broad aim data storage, analytical processing and visualisation should be developed as a proof-of-concept using a public cloud-based infrastructure as a service (IaaS) to take advantage of the low-cost capital investment, flexible infrastructure, performance and the potential for collaboration.

The ED should be able to define this process via an online interface that allows the storage, editing and reuse of previously defined processes. The ED requires the ability to start, pause, stop and cancel the analysis process. Prediction of runtime and monitoring of workflow progress is also desirable. Results should be stored and accessible for online visualisation and statistical summary and comparison. Resulting graphics and summaries should be available for download in formats for use in excel and as high-quality images and vector graphics for use in reports once the process is complete.

This general use case can be subdivided into four phases:

### 1. Define workflow

To define a process or work-flow the ED must be able to specify a dataset or collection of datasets selected from a set of preloaded data. From the data the ED needs to define a geographical region or single point to study. The ED requires a high level of control of the temporal dimensions of the data. They will define the start and end dates and may need to specific that analysis takes place on recurring time-periods within the data (such as an afternoon in a specific season). The ED needs to specify which variables from the data to analyse and potentially assign a weight to each to define its relevant importance. Once the dataset and spatial and temporal subset has been defined and the dimensionality specified, the ED will select an analytic or data mining technique from a set of predefined (and described) methods. At any point in this process the ED may need to save, save as and or edit the defined work-flow.

### 2. Run workflow + monitor resources

Once the work-flow is defined the ED may need an estimate of how long the selected analytics will take. The workflow will then be submitted for processing and its progress will be monitored in terms of its status (ready, executing, staging, completed). The ED may also need to monitor the state of processing resources. During processing the should be able to stop, pause or cancel the workflow.

### 3. Output + visualise results

Following analysis results from the workflow shall be stored and accessible to the ED. Numerical and statistical summaries of the results will be generated by the ED using one of a range of predefined methods. These results would be stored and available for download in formats that can be further analysed or shared in spreadsheets. The ED may need to use simple visualisation of results (histograms, pie charts, line charts). Visualisation tools should include the ability to represent results by geospatial mapping. The graphics generated will be downloaded as high-quality images or vector graphics and used in reports. Within the interface the ED may need to compare two or more visualisations side-by-side to undertake a visual analysis. The ED also needs to share visualisation results with collaborators by providing a secure link to a webpage where the graphics can be viewed and download.

### 4. Manage design strategies

The design strategies specific to individual ED’s, it is necessary, therefore, to manage the predefined general design strategies used by the system. Some EDs will wish to add new strategies, removing unwanted strategies and editing existing ones. A standard unambiguous method for describing a strategy is required.

### Sequence Diagrams for General use case analysis

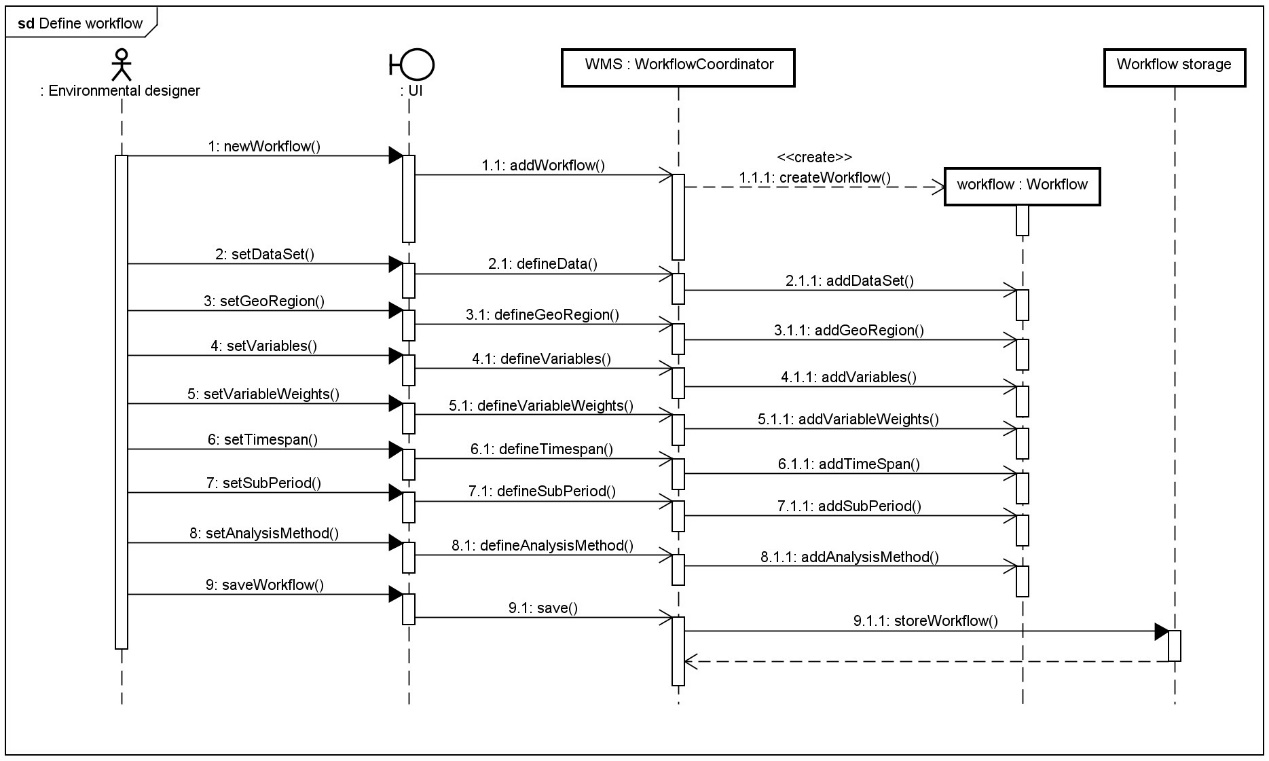


Figure Sequence diagram for defining a workflow

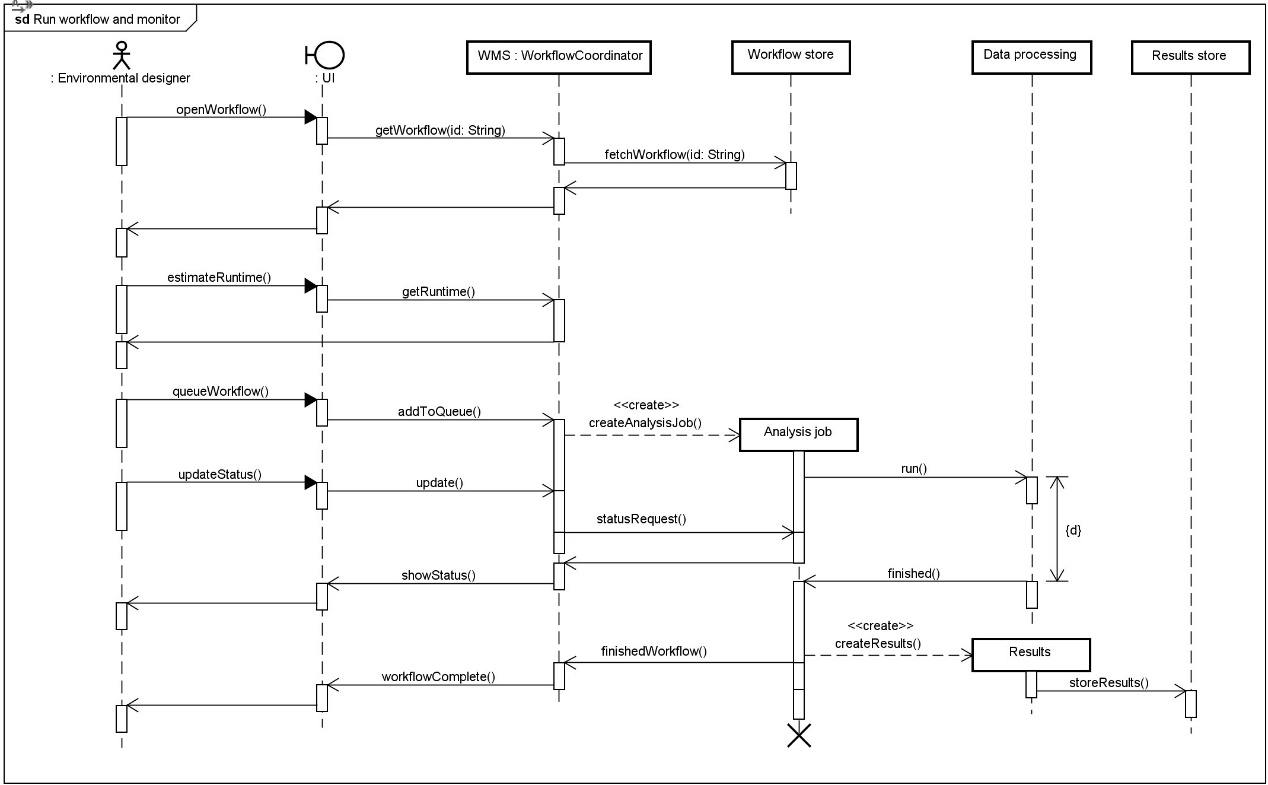


Figure Sequence diagram for running and monitoring a workflow.

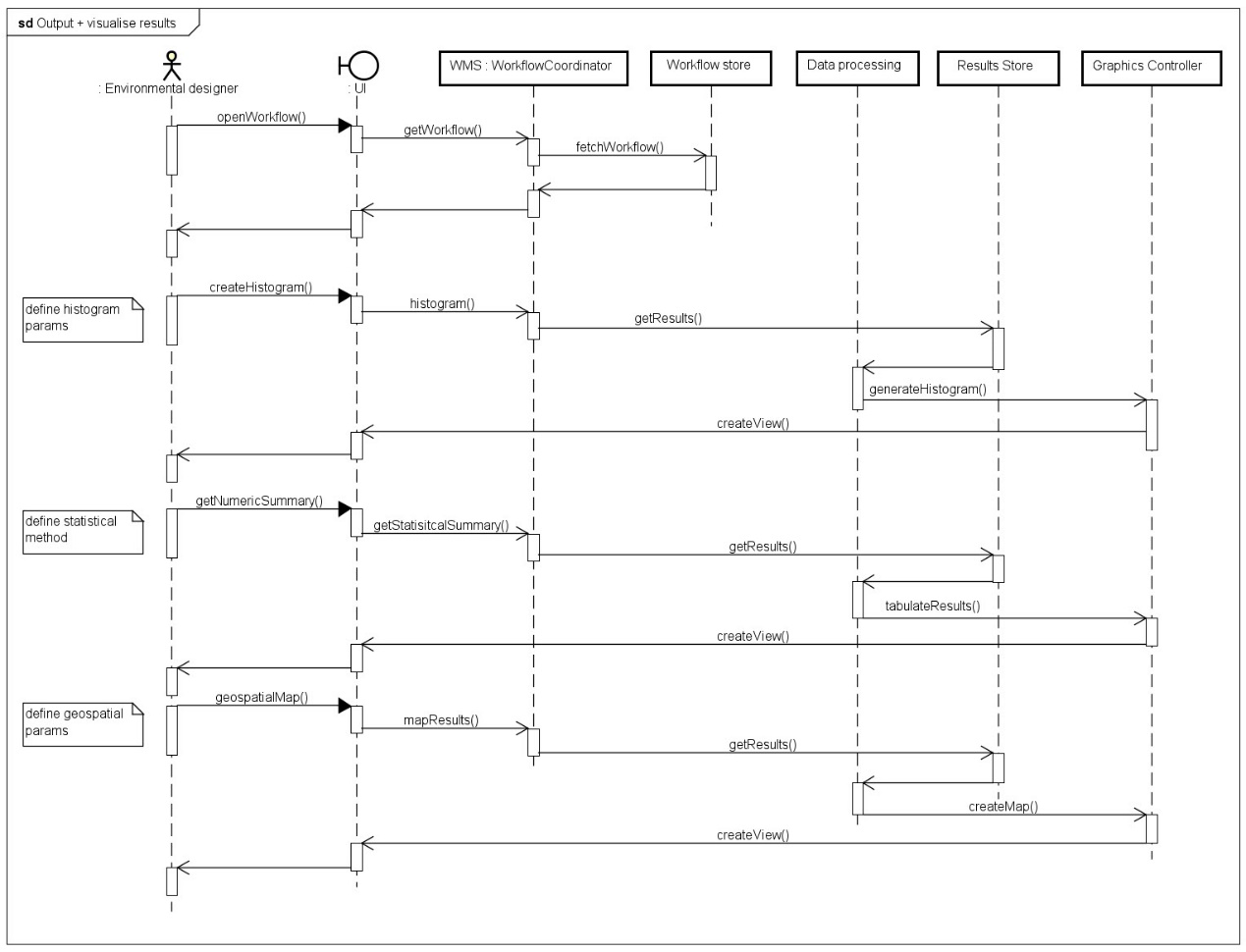


Figure Sequence diagram for visualisation and output of results.

# Chapter 4. Implementation

## Architecture

### Architecture overview

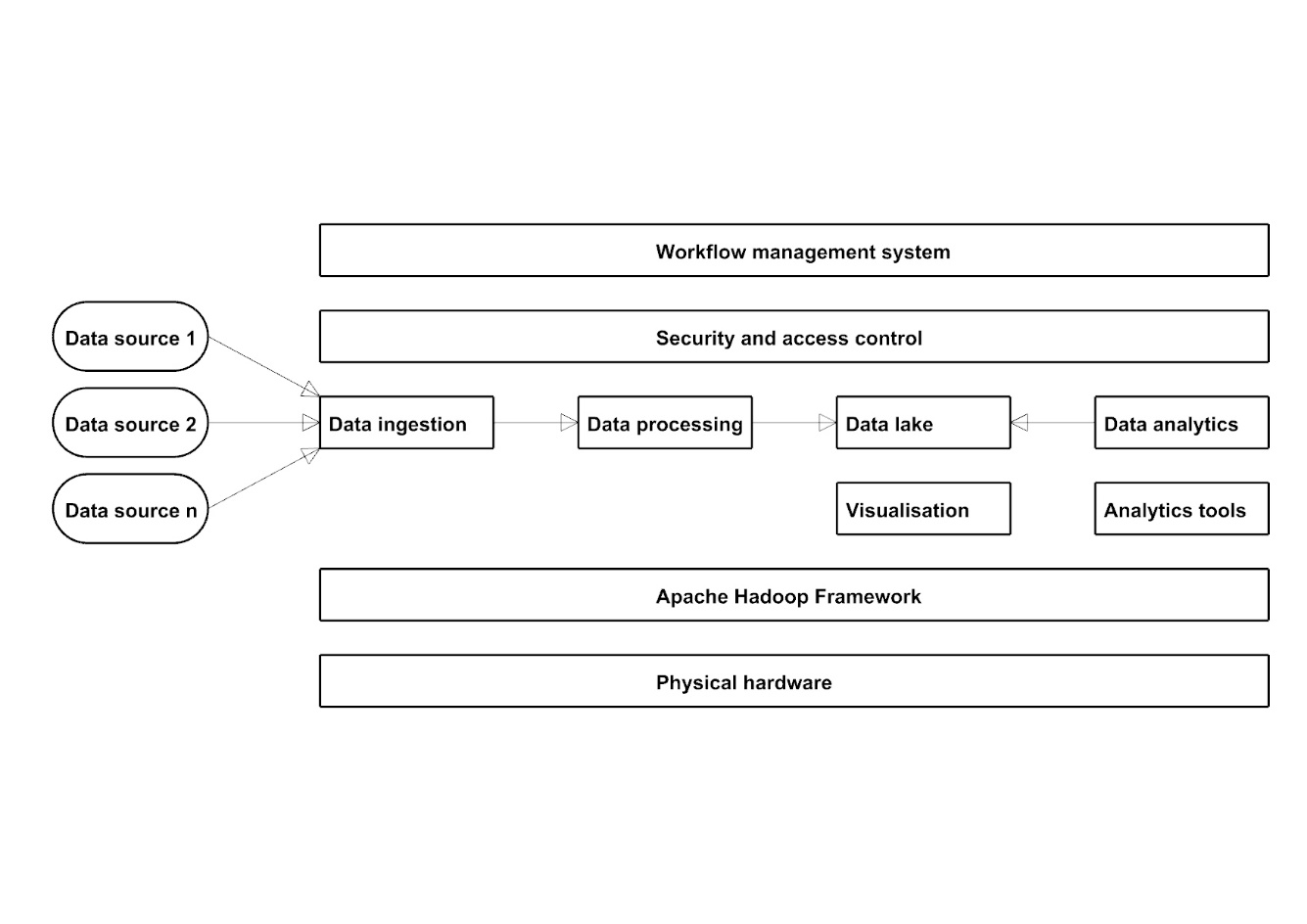
**

Figure Proposed architecture overview based on Hadoop framework (extended from: Lopes, Palmer and O’Sullivan, 2017)

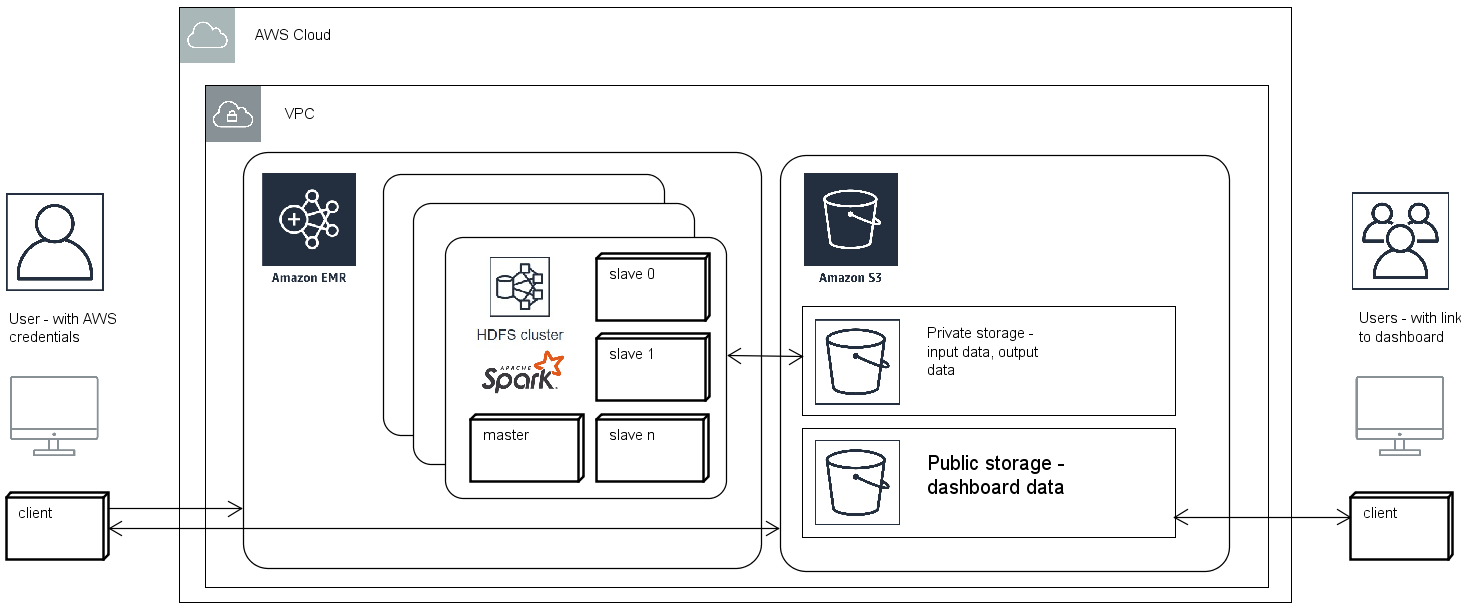


Figure Deployment on AWS

The application uses Infrastructure as a Service (IaaS) on Amazon Web Services (AWS) cloud infrastructure. Users access the application with AWS login credentials and can launch resources within Amazon’s Virtual Private Cloud (VPC). A graphical interface enables users to create analytic jobs and configure the hardware that will be used to compute them. These analytic jobs are launched as Hadoop distributed file system (HDFS) clusters on Amazons’ Elastic Map Reduce (EMR), multiple clusters be can be configured and launched simultaneously. Clusters have access to private buckets within Amazon’s Simple Storage Service (S3), from here they access input data and store results from the analytic jobs. When analytics on a cluster completes, output files are transformed and stored in a publicly accessible bucket with a web framework that defines a graphical dashboard for visualising results. A Uniform Resource Locator ([URL](https://en.wikipedia.org/wiki/URL)) is created for the dashboard that can be shared by the original user so others can access the visualisation and results.

And allow the use the following IaaS services define by (Liu *et al.*, 2011)

Compute: Server resources for running cloud-based systems that can be dynamically provisioned and configured as needed.

Storage: Massively scalable storage capacity that can be used for applications, backups, archival, and file storage.

Content Delivery Networks (CDNs): CDNs store content and files to improve the performance and cost of delivering content for web-based systems

### Architecture for WMS

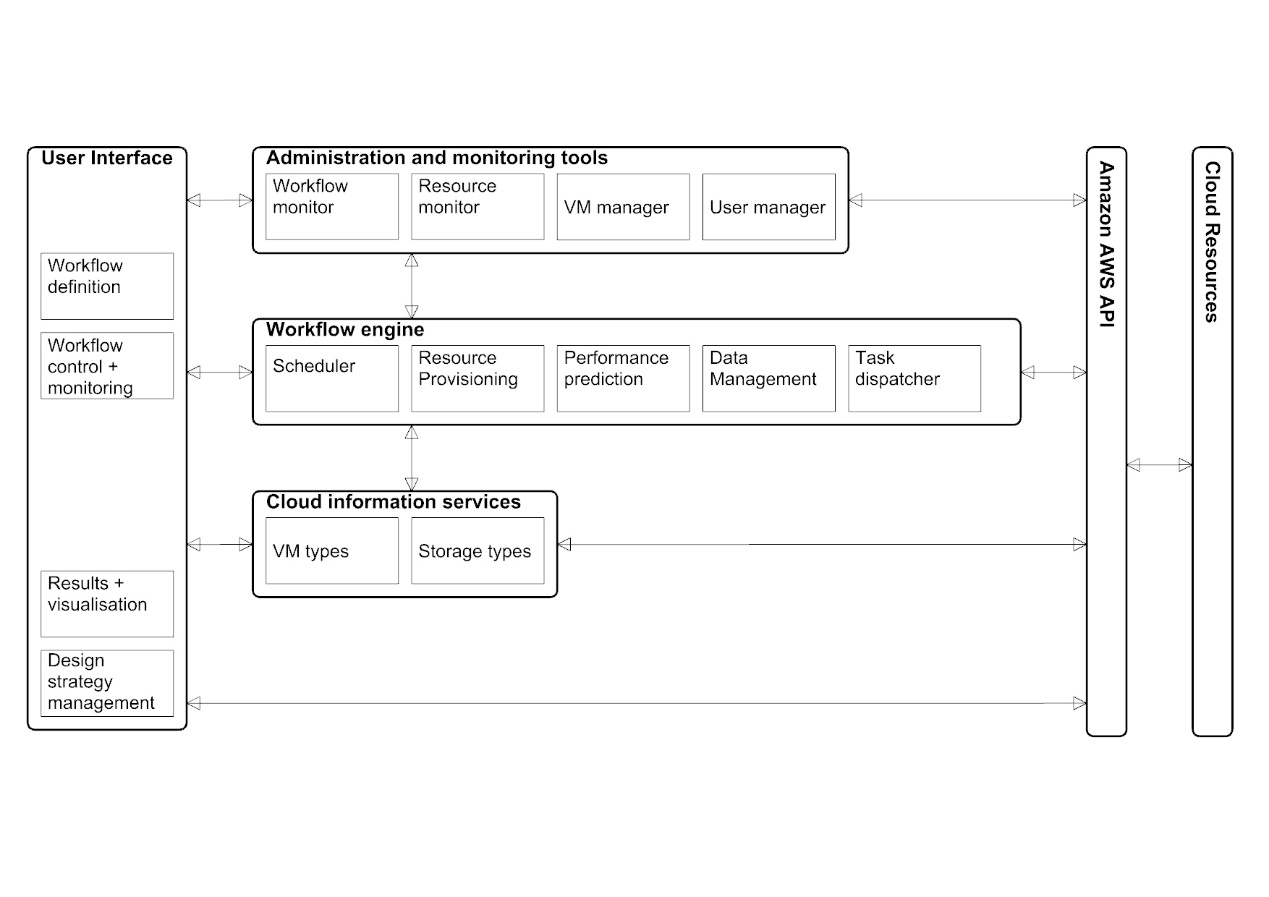


Figure Architecture for the WMS (adapted from: Rodriguez and Buyya (2017)).

### Packages

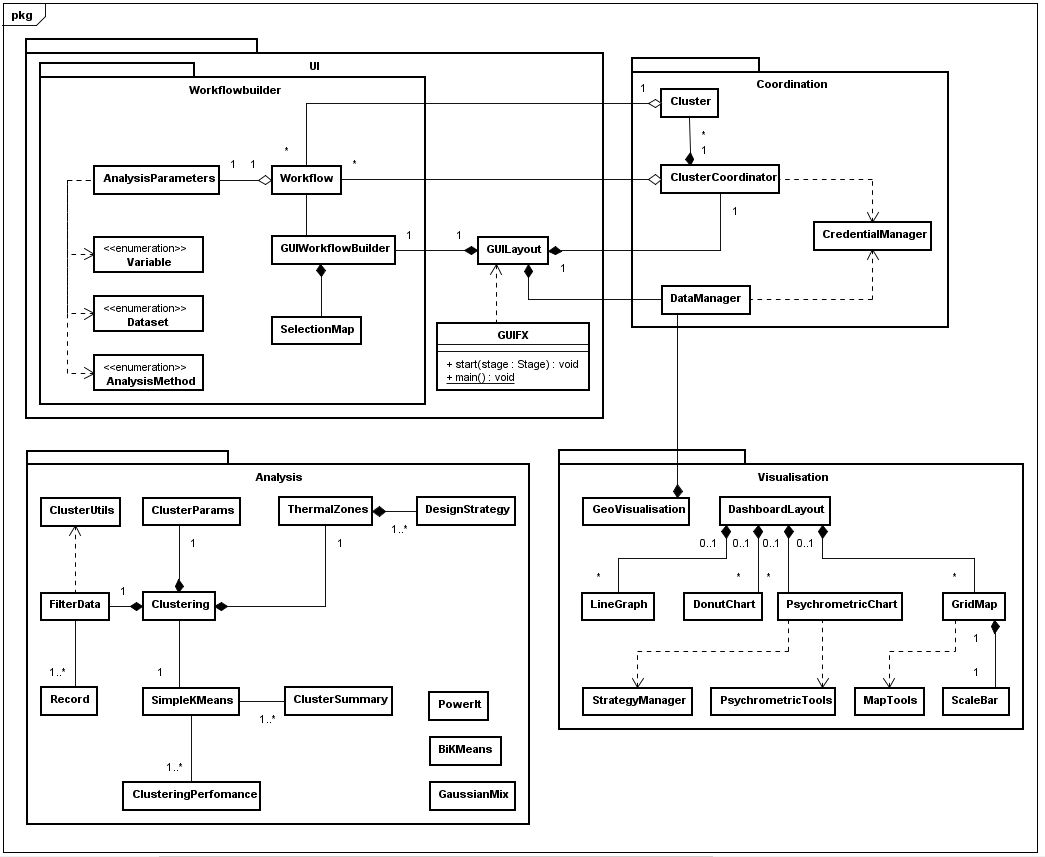


Figure Packages

The overall architecture comprises of four key packages and one sub-package. The User Interface package provides functionality that presents the status of cloud resources and running analysis jobs. Ability to upload new analysis routines as .jar packages for and data sets

The main component in the User Interface package is a sub-package, Workflowbuilder, which concerns all classes involved with the definition and of an analytic job or Workflow which include defining spatial zones for analysis and specifying a series of parameters concerning the data, analysis method and temporal scales. The coordination package interfaces with the EMR client via the ClusterCoordinator class and S3 via the DataManager class both of which use the CredentialManager for secure access to AWS services.

## Environment

Operating environment

Java Virtual Machine (JVM) tested running on Windows

Development environment

Windows 10 pro

For user interface and integration with AWS:

Java

Eclipse photon IDE with AWS Toolkit for Eclipse (*AWS Toolkit for Eclipse*, no date)

spark-core\_2.11

For dashboard web framework:

Javascript, html, css

Sublime Text 3

http-server (*http-server: a command-line http server*, 2018) to debug locally within chrome

Version control: github https://github.com/rolyhudson/climacolombia.git

## Data

S3 as data layer hosting input and output data

application developed with c#.net to parse gridded climate data from various sources

Version control: <https://github.com/rolyhudson/griddedClimateDataProcessing.git>

## Climate data

The data set is based on ten years of multivariate, historical, monthly averages of gridded climate data from three sources. First, climatic data from 1901-2009 formatted as ESRI ASCII raster by CGIAR CSI (Cgiar-csi.org., 2012) based on original data from CRU (Jones and Harris, 2008). Secondly, wind speeds from the CCMP gridded surface vector winds (Wentz *et al.*, 2015). Thirdly, elevation data is extracted from a hole-filled DEM of SRTM (Jarvis, A., H.I. Reuter, A. Nelson, 2008). Figure 12 illustrates the data preparation steps. A C# dot net program was written that takes a topojson (Bostock, 2017) format file as input, this describes the boundary (or collection of boundaries) that define the zone of interest. A point grid is generated at half degree latitude and longitude intervals filling the area(s) of study. Cross-referencing the grid to the DEM determines altitudes for each point.

The grid is used to extract climate data from a ten-year period (2000-2009) from the CRU and CCMP datasets. CRU data is formatted as ASCII ESRI raster format at the same resolution as the grid. CCMP data is in netCDF format, this was pre-processed with an independent Java program, written using the Unidata (2012) netCFD Java library. The netCDF data was converted to the ASCII ESRI raster format. Each ASCII raster file represents a single month of a year and contains data for earth’s surface. Data points in the raster files that coincide with our grid vertices are found and stored with the georeferenced grid in arrays. Relative humidity derived using the ratio between vapour pressure and saturation pressure. The data is averaged for the 2000-2009 period providing a typical year with monthly values. The prepared climate data can be saved to a Comma Separated Value (csv) file.

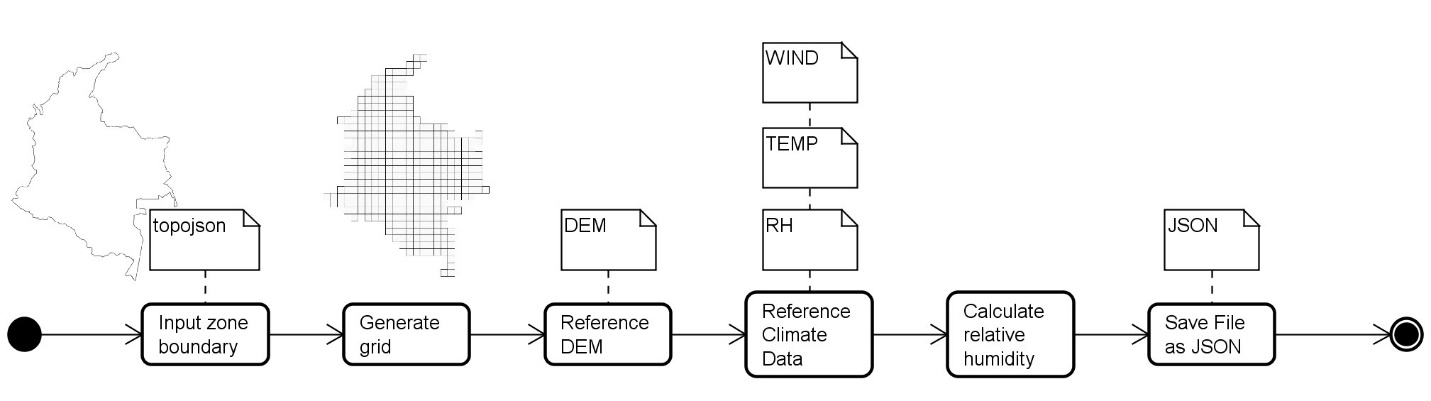


Figure Flow diagram showing the stages of the data preparation process.

## Class diagrams

### User Interface package

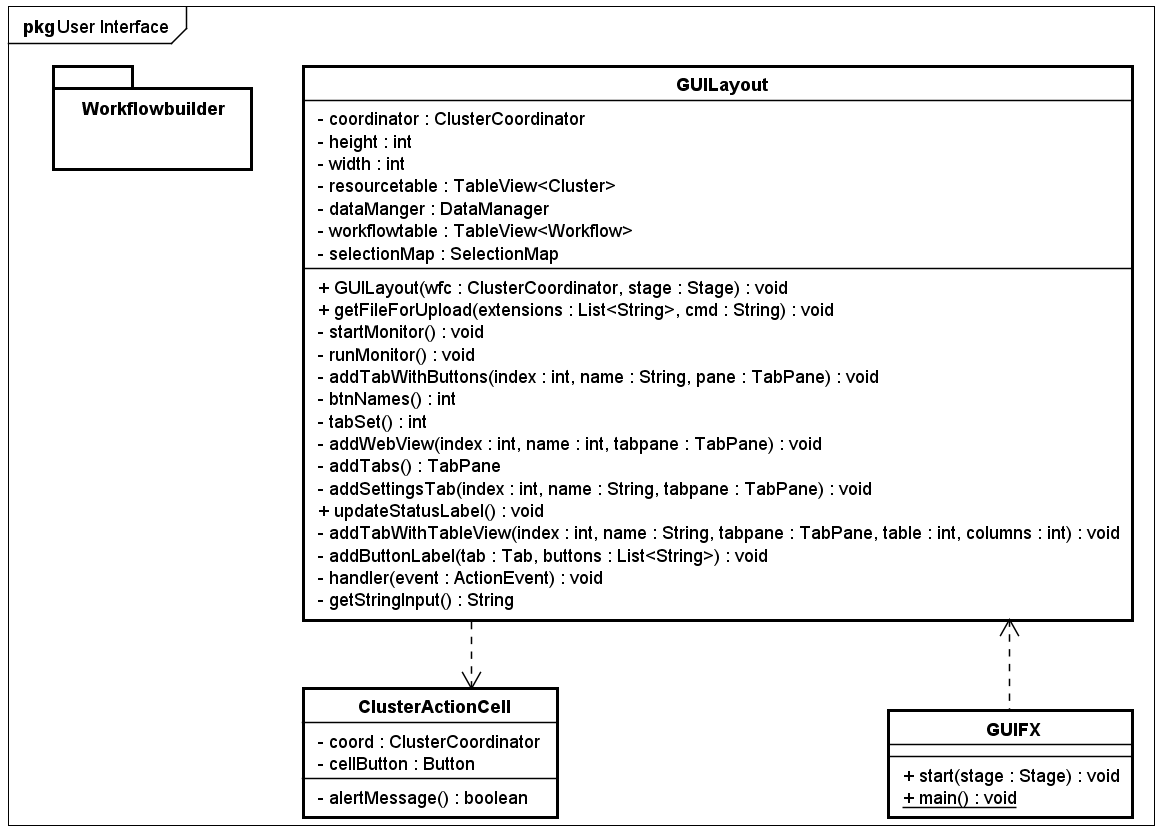


Figure User Interface package

### Workflowbuilder package

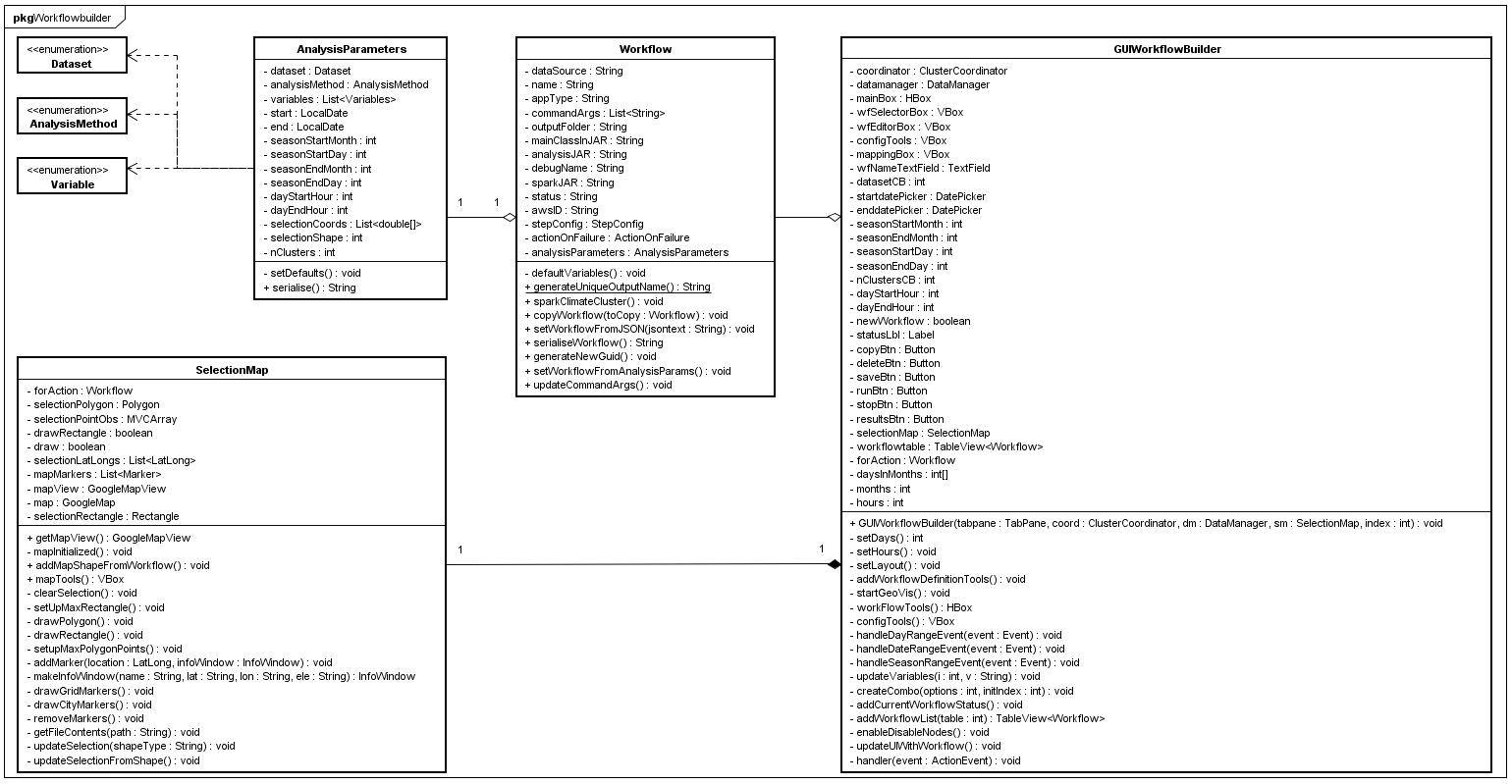


Figure Workflowbuilder package

### Analysis package

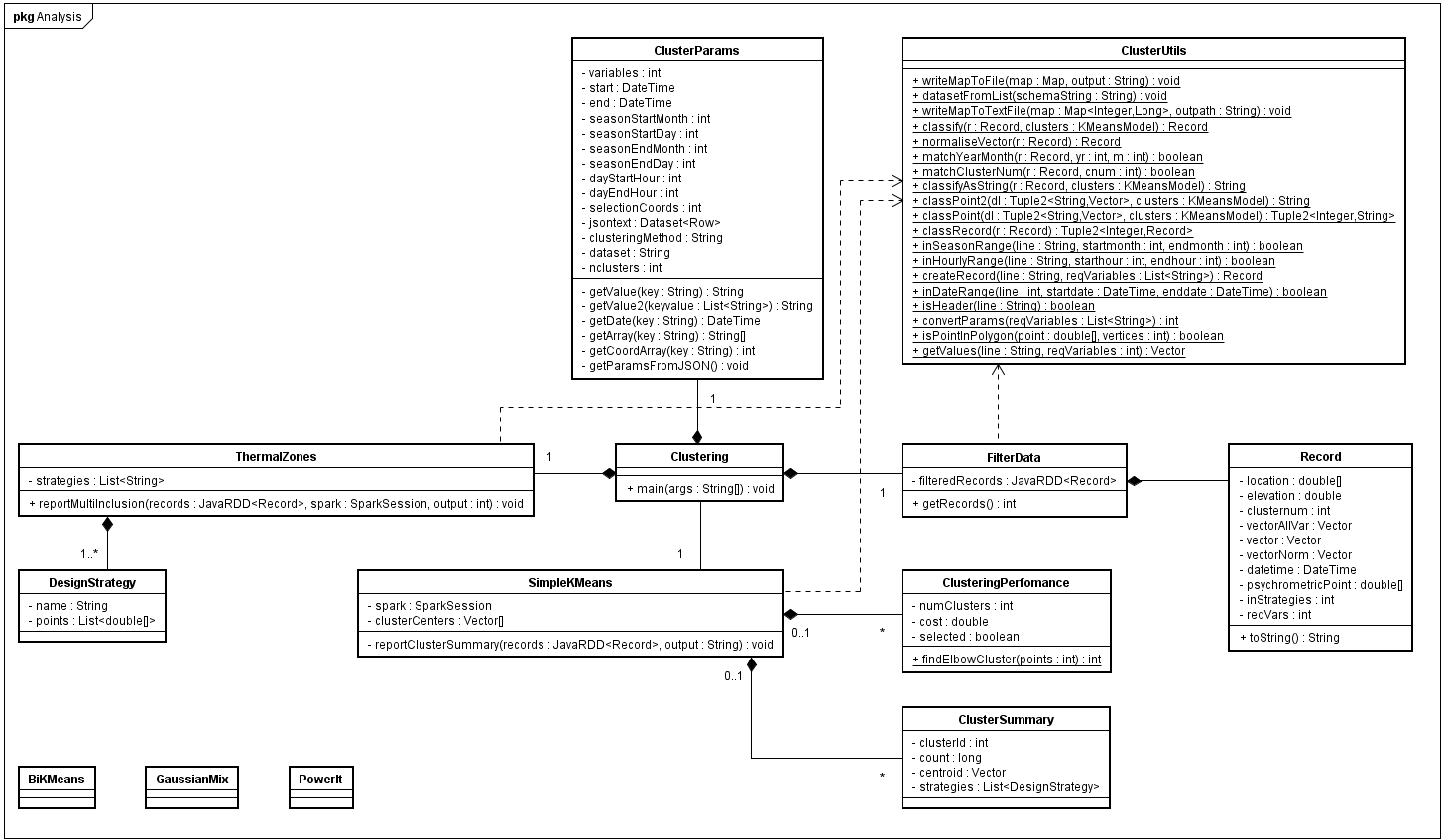


Figure Analysis package

### Coordination package

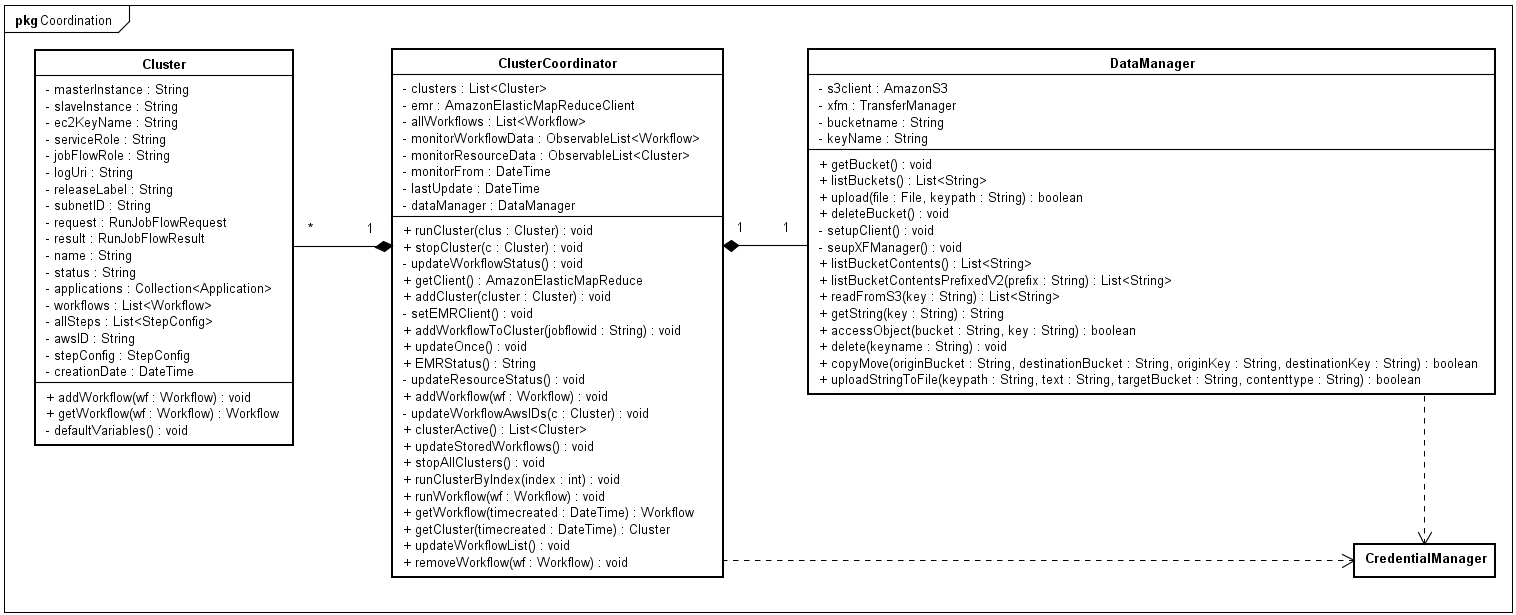


Figure Coordination package

### Visualisation package

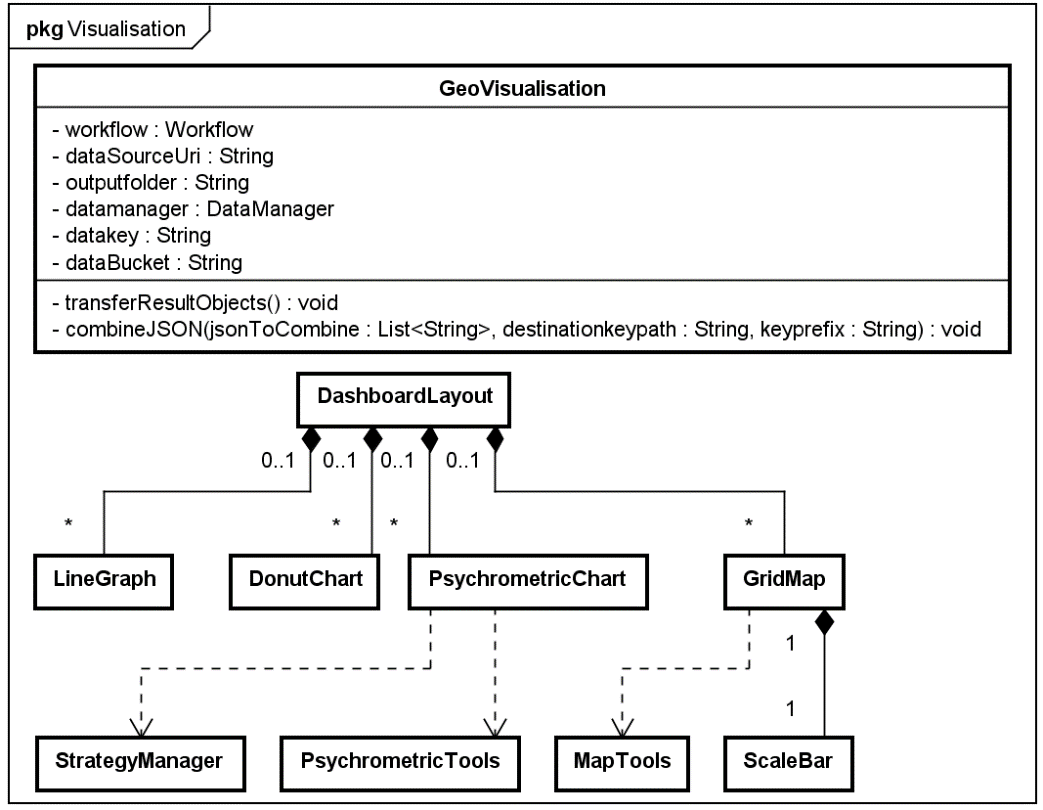


Figure Visualisation package

## Interface

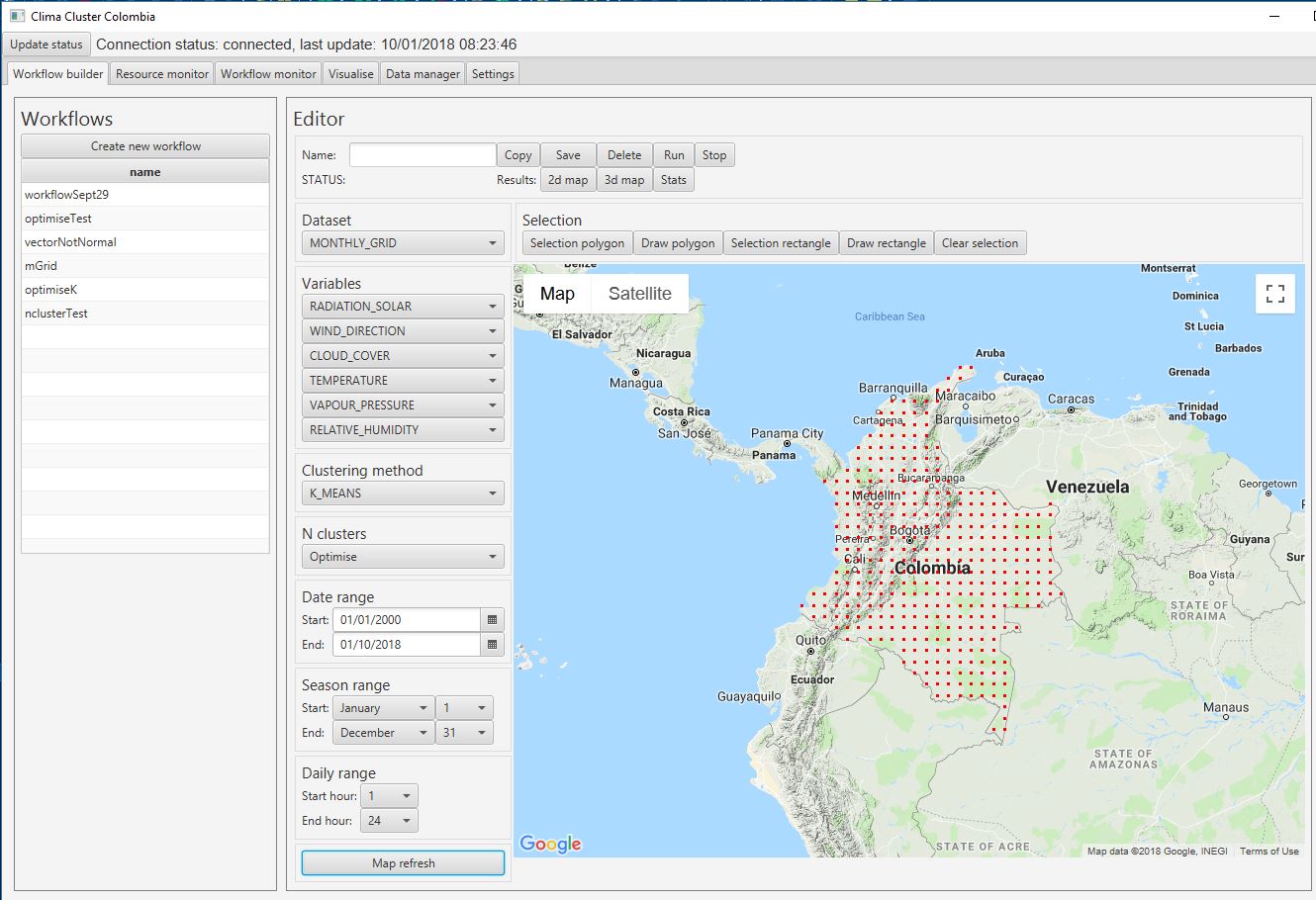


Figure User interface

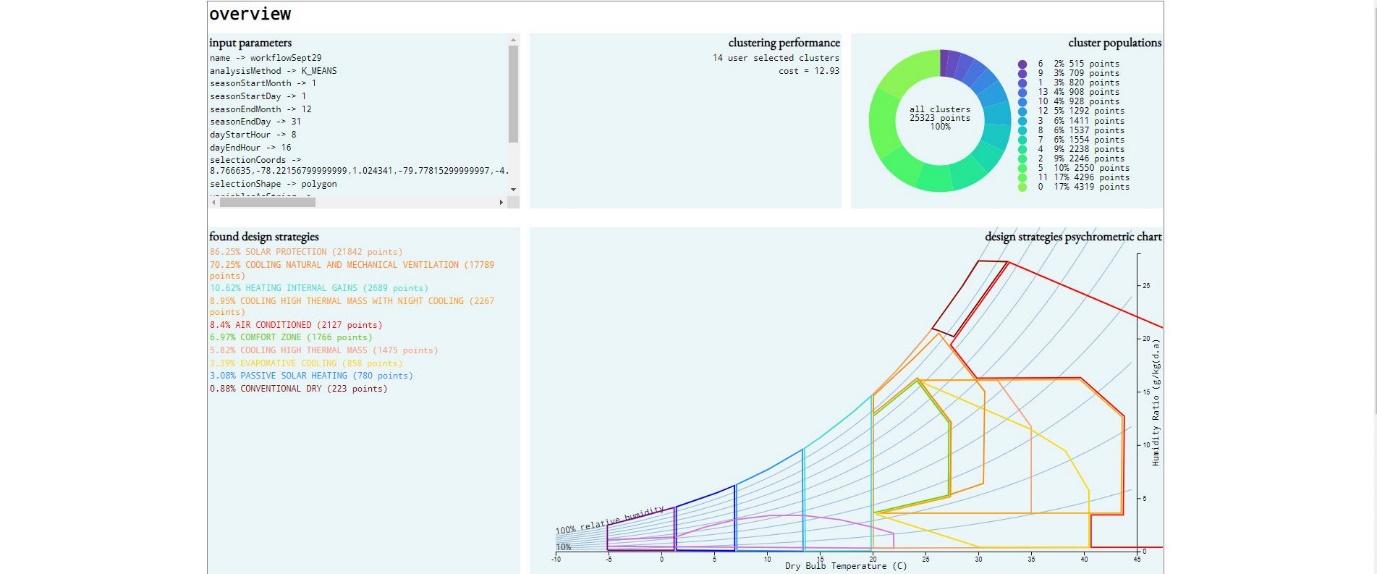


Figure Results dashboard



Figure Results dashboard

# Chapter 5. Results and evaluation

## Final prototype

Source code on github

Demo dashboards

## Evaluation

The system is intended to produce recommendations for environmental design strategies based on the analysis of a set of weather data. The hypothesis to be tested is that design strategies can be linked with patterns discovered in weather data at various spatiotemporal scales and with different subsets of variables. It is proposed that this can be tested using a big data architecture that enables data analytics over large sets of weather data. Colombia is chosen for a test case as the tropical climate combined with the extreme topography and proximity to oceans.

Products of the system will be decomposed for evaluation; weather patterns will be assessed both quantitively and qualitatively and matching of recommended design strategies and the patterns evaluated qualitatively.

Several metrics exist to express the completeness and homogeneity of clusters discovered through analytics. Validity-measure or V-measure expresses how well both completeness and homogeneity are satisfied (Rosenberg and Hirschberg, 2007). Other metrics capture one or the other and include Purity, Entropy, Rand Index, misclassification index, f-measure, silhouette coefficient and cluster distortion of clusters can be calculated.

For results of all analysis methods visual inspection of graphical output will play an important role in evaluation of the system. Including interactive graphical representations will allow dynamic exploration of spatiotemporal results. Use of correlation matrices, 2D scatter plots and plotting georeferenced zones on maps will amplify knowledge discovery and allow products of the artefact to be presented to domain experts. Qualitative evaluation of the artefact and the products (patterns and recommendations) will be undertaken by identifying very specific use cases and developing walkthroughs.

### 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

See interview question format in Smith dissertation example

Cognitive walkthrough, heuristic evaluation, review based see Shneiderman, B. (1998) Designing the user interface: Strategies for effective human computer interaction (3rd ed.). Reading, MA: Additon-Wesley Publishing

# Chapter 6. Conclusions

## Findings

## Future research

# References

Ambler, S. W. (2004) *The object primer : agile modeling-driven development with UML 2.0*. Cambridge University Press.

ASHRAE (2013) *2013 ASHRAE Handbook: Fundamentals*, *ASHRAE*. doi: 10.1163/ej.9789004155947.i-937.23.

Atluri, G., Karpatne, A. and Kumar, V. (2017) ‘Spatio-Temporal Data Mining: A Survey of Problems and Methods’, *ACM Comput. Surv*, 1(1). doi: 10.1145/nnnnnnn.nnnnnnn.

Avci Salma, C., Tekinerdogan, B. and Athanasiadis, I. N. (2017) ‘Chapter 4 – Domain-Driven Design of Big Data Systems Based on a Reference Architecture’, in *Software Architecture for Big Data and the Cloud*, pp. 49–68. doi: 10.1016/B978-0-12-805467-3.00004-1.

*AWS Toolkit for Eclipse* (no date). Available at: https://aws.amazon.com/eclipse/ (Accessed: 3 October 2018).

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.

Bostock, M. (2017) *TopoJSON*, *Github*. Available at: https://github.com/topojson/topojson (Accessed: 3 May 2017).

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.

Cgiar-csi.org. (2012) *CRU-TS v3.10.01 Historic Climate Database for GIS | CGIAR-CSI.* Available at: http://www.cgiar-csi.org/data/uea-cru-ts-v3-10-01-historic-climate-database (Accessed: 5 November 2017).

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.

Dagade, V. *et al.* (2015) ‘Big Data Weather Analytics Using Hadoop’, *International Journal of Emerging Technology in Computer Science & Electronics*, 14(2), pp. 976–1353. Available at: https://pdfs.semanticscholar.org/f2e4/918444be9b30f29132e93ce02d29ccf26eda.pdf (Accessed: 29 May 2018).

Das, T., Zaharia, M. and Wendell, P. (2015) *Spark Streaming*. Available at: https://databricks.com/blog/2015/07/30/diving-into-apache-spark-streamings-execution-model.html (Accessed: 21 February 2018).

Degaetano, A. T. (1996) ‘Delineation of Mesoscale Climate Zones in the Northeastern United States Using a Novel Approach to Cluster Analysis’, *Journal of Climate*, 9(8), pp. 1765–1782. doi: 10.1175/1520-0442(1996)009<1765:DOMCZI>2.0.CO;2.

Dimoudi, A. and Tompa, C. (2008) ‘Energy and environmental indicators related to construction of office buildings’, *Resources, Conservation and Recycling*, 53(1–2), pp. 86–95. doi: 10.1016/j.resconrec.2008.09.008.

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.

El Fazziki, A. *et al.* (2015) ‘A multi-agent framework for a hadoop based air quality decision support system’, in *CEUR Workshop Proceedings*, pp. 45–59.

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.

Fovell, R. G. and Fovell, M. Y. C. (1993) ‘Climate zones of the conterminous United States defined using cluster analysis’, *Journal of Climate*, pp. 2103–2135. doi: 10.1175/1520-0442(1993)006<2103:CZOTCU>2.0.CO;2.

Givoni, B. (1992) ‘Comfort, climate analysis and building design guidelines’, *Energy and Buildings*, 18, pp. 11–23. Available at: https://ac-els-cdn-com.liverpool.idm.oclc.org/037877889290047K/1-s2.0-037877889290047K-main.pdf?\_tid=67644909-7d5e-4718-b47f-80732d024251&acdnat=1527603110\_130b2a2e77cd97d48361202940532374 (Accessed: 29 May 2018).

*http-server: a command-line http server* (2018) *npmjs*. Available at: https://www.npmjs.com/package/http-server (Accessed: 3 October 2018).

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.

Jarvis, A., H.I. Reuter, A. Nelson, E. G. (2008) *Hole-filled SRTM 90m for the globe Version 4 Database*. Available at: http://srtm.csi.cgiar.org (Accessed: 5 November 2017).

Jayanthi, D. and Sumathi, G. (2017) ‘Weather data analysis using spark — An in-memory computing framework’, in *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, pp. 1–5. doi: 10.1109/IPACT.2017.8245142.

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.

Jones, P. and Harris, I. (2008) *Climatic Research Unit (CRU) time-series datasets of variations in climate with variations in other phenomena*, *NCAS British Atmospheric Data Centre*. Available at: http://catalogue.ceda.ac.uk/uuid/3f8944800cc48e1cbc29a5ee12d8542d (Accessed: 5 November 2017).

Karpatne, A. *et al.* (2017) ‘Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data’, *IEEE Transactions on Knowledge and Data Engineering*, 29(10), pp. 2318–2331. doi: 10.1109/TKDE.2017.2720168.

Kodinariya, T. M. and Makwana, P. R. (2013) ‘Review on determining number of Cluster in K-Means Clustering’, *International Journal of Advance Research in Computer Science and Management Studies*.

Liu, F. *et al.* (2011) *NIST Cloud Computing Reference Architecture Recommendations of the National Institute of Standards and Technology*. Available at: https://ws680.nist.gov/publication/get\_pdf.cfm?pub\_id=909505 (Accessed: 3 October 2018).

Liu, Y. and Weisberg, R. H. (2005) ‘Patterns of ocean current variability on the West Florida Shelf using the self-organizing map’, *Journal of Geophysical Research: Oceans*, 110(6), pp. 1–12. doi: 10.1029/2004JC002786.

Liu, Y. and Weisberg, R. H. (2011) ‘A Review of Self-Organizing Map Applications in Meteorology and Oceanography’, in Igadwa Mwasiagi, J. (ed.) *Self Organizing Maps - Applications and Novel Algorithm Design*. www.intechopen.com. doi: 10.5772/13146.

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.

Luna-Romera, J. M. *et al.* (2016) ‘An approach to silhouette and dunn clustering indices applied to big data in spark’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-44636-3\_15.

Manzano-Agugliaro, F. *et al.* (2015) ‘Review of bioclimatic architecture strategies for achieving thermal comfort’, *Renewable and Sustainable Energy Reviews*. Pergamon, 49, pp. 736–755. doi: 10.1016/J.RSER.2015.04.095.

Mariam Varghese, S. (2015) ‘Leveraging Map Reduce With Hadoop for Weather Data Analytics’, *IOSR Journal of Computer Engineering Ver. II*, 17(3), pp. 2278–661. doi: 10.9790/0661-17320612.

Milne, M., Liggett, R. and Benson, A. (2009) ‘Climate Consultant 4.0 develops design guidelines for each unique climate’, *American Solar Energy Society Meeting*. Available at: http://www.energy-design-tools.aud.ucla.edu/papers/ases09-milne.pdf (Accessed: 10 April 2018).

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.

Nikolaou, T. G. *et al.* (2012) ‘On the application of clustering techniques for office buildings’ energy and thermal comfort classification’, *IEEE Transactions on Smart Grid*. doi: 10.1109/TSG.2012.2215059.

Olgyay, V. and Olgyay, A. (1963) *Design With Climate: Bioclimatic Approach to Architectural Regionalism*. Princeton University Press.

Omer, A. M. (2008) ‘Energy, environment and sustainable development’, *Renewable and Sustainable Energy Reviews*, 12, pp. 2265–2300. doi: 10.1016/j.rser.2007.05.001.

Rhee, J. *et al.* (2008) ‘Delineation of climate regions using in-situ and remotely-sensed data for the Carolinas’, *Remote Sensing of Environment*, 112(6), pp. 3099–3111. doi: 10.1016/j.rse.2008.03.001.

Rodenburg, B. and Maria Fiore, M. (2017) *Detecting Weather Twins using Apache Spark*, *LSDE: Large Scale Data Engineering 2017*. Available at: https://event.cwi.nl/lsde/2017/showcase\_n2.shtml (Accessed: 29 May 2018).

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.

Rosenberg, A. and Hirschberg, J. (2007) ‘V-Measure: A conditional entropy-based external cluster evaluation measure’, pp. 410–420. Available at: http://www.aclweb.org/anthology/D07-1043 (Accessed: 6 June 2018).

Thinsungnoen, T. *et al.* (2015) ‘The Clustering Validity with Silhouette and Sum of Squared Errors’, in *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*. doi: 10.12792/iciae2015.012.

Unidata (2012) ‘NetCDF-Java library and TDS version 4.6.9’. Boulder. CO: UCAR/Unidata. doi: http://doi.org/10.5065/D6RN35XM.

Wentz, F. *et al.* (2015) *Remote Sensing Systems Cross-Calibrated Multi-Platform (CCMP) 6-hourly ocean vector wind analysis product on 0.25 deg grid, Version 2.0, [subset: CCMP V2.0 Level-3.5].* Santa Rosa, CA: Remote Sensing Systems. Available at: www.remss.com/measurements/ccmp (Accessed: 5 November 2017).

Zaharia, M. *et al.* (2013) ‘Discretized Streams: Fault-Tolerant Streaming Computation at Scale’, *Sosp*, (1), pp. 423–438. doi: 10.1145/2517349.2522737.

Zscheischler, J., Mahecha, M. D. and Harmeling, S. (2012) ‘Climate classifications: The value of unsupervised clustering’, in *Procedia Computer Science*, pp. 897–906. doi: 10.1016/j.procs.2012.04.096.

# List of Tables

[Table 1 Key Research methods used in the dissertation 7](#_Toc526246822)

[Table 2 System development process 8](#_Toc526246823)

# List of Figures

[Figure 1 process diagram 7](#_Toc526246829)

[Figure 2 Manzano-Agugliaro et al. (2015) adapted version of the psychrometric chart 11](#_Toc526246830)

[Figure 3 Development increments (extended from Ambler, 2004, p119). 15](#_Toc526246831)

[Figure 4 Proposed architecture overview based on Hadoop framework (extended from: Lopes, Palmer and O’Sullivan, 2017) 17](#_Toc526246832)

[Figure 5 System components overview based on big data reference architecture (extended from: Avci Salma, Tekinerdogan and Athanasiadis, 2017). 18](#_Toc526246833)

[Figure 6 Architecture for the WMS (adapted from: Rodriguez and Buyya, 2017). 18](#_Toc526246834)

[Figure 7 Sequence diagram for defining a workflow 19](#_Toc526246835)

[Figure 8 Sequence diagram for running and monitoring a workflow. 19](#_Toc526246836)

[Figure 9 Sequence diagram for visualisation and output of results. 20](#_Toc526246837)