Identifying location within the cloud ecosystem using an automated service for determining standards relevance.

A European Cloud observatory supporting cloud policies, standards profiles & services

www.cloudwatchhub.eu|info@cloudwatchhub.eu

This report provides a detailed description of the data and methodology used to perform the quantitative clustering that serves as a precursor to the standards profile development activities of the CloudWATCH project. A companion document (D4.3) provides a non-technical summary of the methodology, followed by detailed discussion and examples of derived standards profiles which are not repeated here. Instead, we take the opportunity to fully describe the methodology used and to discuss related work as presented at the *NIST Cloud Computing Forum and Workshop VIII*, held in Gaithersburg, MD, USA, July 7-10, 2015. We also intend to submit the body of this work for publication to the *Journal of Cloud Computing: Advances, Systems and Applications*.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

CloudWATCH Mission

The CloudWATCH mission is to accelerate the adoption of cloud computing across European private and public organisations. CloudWATCH offers independent, practical tips on why, when and how to move to the cloud, showcasing success stories that demonstrate real world benefits of cloud computing. CloudWATCH fosters interoperable services and solutions to broaden choice for consumers. CloudWATCH provides tips on legal and contractual issues. CloudWATCH offers insights on real issues like security, trust and data protection. CloudWATCH is driving focused work on common standards profiles with practical guidance on relevant standards and certification Schemes for trusted cloud services across the European Union.

The CloudWATCH partnership brings together experts on cloud computing; certification schemes; security; interoperability; standards implementation and roadmapping as well as legal professionals. The partners have a collective network spanning 24 European member states and 4 associate countries. This network includes: 80 corporate members representing 10,000 companies that employ 2 million citizens and generate 1 trillion in revenue; 100s of partnerships with SMEs and 60 global chapters pushing for standardisation, and a scientific user base of over 22,000.

Disclaimer

CloudWATCH (A European Cloud Observatory supporting cloud policies, standard profiles and services) is funded by the European Commission's Unit on Software and Services, Cloud Computing within DG Connect under the 7th Framework Programme.

The information, views and tips set out in this publication are those of the CloudWATCH Consortium and its pool of international experts and cannot be considered to reflect the views of the European Commission.

Document information summary

Document title:	Identifying location within the cloud ecosystem using an automated service for determining standards relevance.
Main author:	Neil Caithness, UOXF
Contributing authors:	Michel Drescher, EGI.EU David Wallom, UOXF
Reviewers:	Daniele Catteddu, CSA Jesus Luna, CSA Damir Savanovic, CSA Peter Deussen, Fraunhofer FOKUS Silvana Muscella , Trust-IT Nicholas Ferguson, Trust-IT
Target audiences:	Cloud computing projects and initiatives including projects funded by Unit E2; Software, Services and Cloud Computing projects; Standards Development Organisations, in particular IEEE P2301 and OGF OCCI.
Keywords:	Cloud ecosystem; Cloud landscape; PCA; Biplot; quantitative clustering

Executive Summary

A central mission of CloudWATCH is to help build consistency and trust in cloud computing. In support of this mission CloudWATCH is developing a set of common standards profiles around the federation of cloud services based on a portfolio of European and international use cases covering technical, policy and legal requirements. These profiles clarify how a standard should be interpreted and implemented based on a specific use case. The documented profiles can be easily adopted, thus driving the Cloud market into a state of commodity and utility.

The process of deriving common standards profiles is helped and informed by a detailed knowledge of the cloud computing landscape, and of how different cloud computing projects form natural clusters based on their common relationship to the defining features of cloud services. This aspect of project clustering proves more helpful than the rather more obvious relationships based on common goals, aspirations and target audiences, which more often form the basis for project collaborations, or indeed the basis for identifying close competitors.

In this report we present in detail a cluster analysis of European cloud computing projects using the NIST model of defining characteristics [1] and we show how this analysis of the landscape of cloud computing provides insights into the process of developing standards profiles. This modelling of the landscape will also form the basis for ongoing collaborations with other standards development inititives such as IEEE P2301 [2].

Table of Contents

1 Introduction 7
2 Workflow 8
3 Data 9
3.1 Definition of variables 9
3.2 Essential Characteristics 9
3.2.1 On-demand self service 9
3.2.2 Broad network access 9
3.2.3 Resource pooling 9
3.2.4 Rapid elasticity 9
3.2.5 Measured service 9
3.3 Common characteristics 10
3.3.1 Massive Scale 10
3.3.2 Homogeneity 10
3.3.3 Virtualization 10
3.3.4 Low Cost Software 10
3.3.5 Resilient Computing 10
3.3.6 Geographic Distribution 10
3.3.7 Service Orientation 10
3.3.8 Advanced Security 10
3.4 Characteristic scoring 10
4 Methods 12
4.1 PCA ordination 12
4.2 Stopping rule 13
4.3 Interpreting the biplot 14
4.4 Hierarchical clustering 16
4.5 Interpreting the clusters 17
5 Summary and Conclusions 19
6 Next Steps 19
7 Notes and References 20
8 Document Log ..... **Errore. Il segnalibro non è definito.**

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

Table of Figures

Figure 1. Greater workflow incorporating the quantitative methodology. 8
Figure 2. Biplot of 38 European cloud computing projects. 13
Figure 3. Scree Plot of the component eigenvalues. 14
Figure 4. Single-linkage hierarchical clustering. 16
Figure 5. Projected scores reordered on the cluster tree. 18

Table of Tables

Table 1. Raw scores as compiled for individual projects. 11
Table 2. Five-dimentional projected scores. 15

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

1 Introduction

Cloud computing resides in a complicated ecosystem of stakeholders with differing requirements and expectations. Even with a broad consensus that cloud computing is a general term for anything that involves delivering hosted services over the Internet, people's interpretations of this vary widely and the field is subject to excessive hyperbole. In an attempt to provide clarity, NIST, the National Institute of Standards and Technology, which is part of US Department of Commerce [1], drafted a model of cloud computing which included during development a set of five essential, and eight common defining characteristics of cloud computing [1]. Their final publication pared the list down to just the five essential characteristics as the definitive set. These sets, both the original long list, and the pared down short list, provide much needed insight into the complicated landscape of cloud computing.

In this report we present a cluster analysis of European cloud computing projects as a way of gaining insight into where the projects are located within the landscape. The objective of this empirical analysis is to discover distinct groups of projects that are consistent in their relationship to a set of well-defined general characteristics. These clusters of projects will form the basis for a further derivation of standards that are able to best support areas of commonality and create profiles which provide details as to how more generic recommended standards may be applied.

We start our analysis with a dataset compiled by CloudWATCH representing 38 European projects and scored against the full set of 13 NIST defining characteristics on an interval scale. Our clustering procedure is based on the outcome of a classic Principal Components Analysis (PCA) [3]. We interpret the landscape on a simultaneous biplot of the characteristic coefficients and component scores [4]. As a natural extension of the biplot we project scores from a reduced dimensionality PCA space onto the coefficient vectors and use those score for clustering. The clustering technique we employ is classic Euclidian distance single linkage hierarchical clustering [5].

Finally, we explore representative clusters and discuss how this procedure has informed the derivation of standards profiles, and how new projects can use the analysis to find their location in the greater cloud ecosystem and quickly identify the appropriate profiles for implementation.

As a final introductory note, it is important to recognise that there is no single universally-correct clustering method. Clustering is better understood as an iterative process of knowledge discovery, and the outcome should be judged by its utility for the task at hand. For this reason it is important that we describe the approach in accurate detail so that results are well understood and repeatable. As we have no prior expectation that the landscape we are exploring is too complex for classic techniques, we explore these first. As we show, the results are helpful and utilitarian.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

2 Workflow

The quantitative methodology that we use, and the outcome of the analysis that is the subject of this report forms part of a greater body of work contributing to the standards profilling work of the CloudWATCH project. Figure 1 below shows the quantitative methodology in the greater context of the iterative workflow that starts with project selection, and ends with a review of standards for profilling.

The outcome of the greater workflow, interpretation of clusters, and discussion of their contribution to standards profilling is the subject of report D4.3 and only a brief summary of that work is provided in this report.

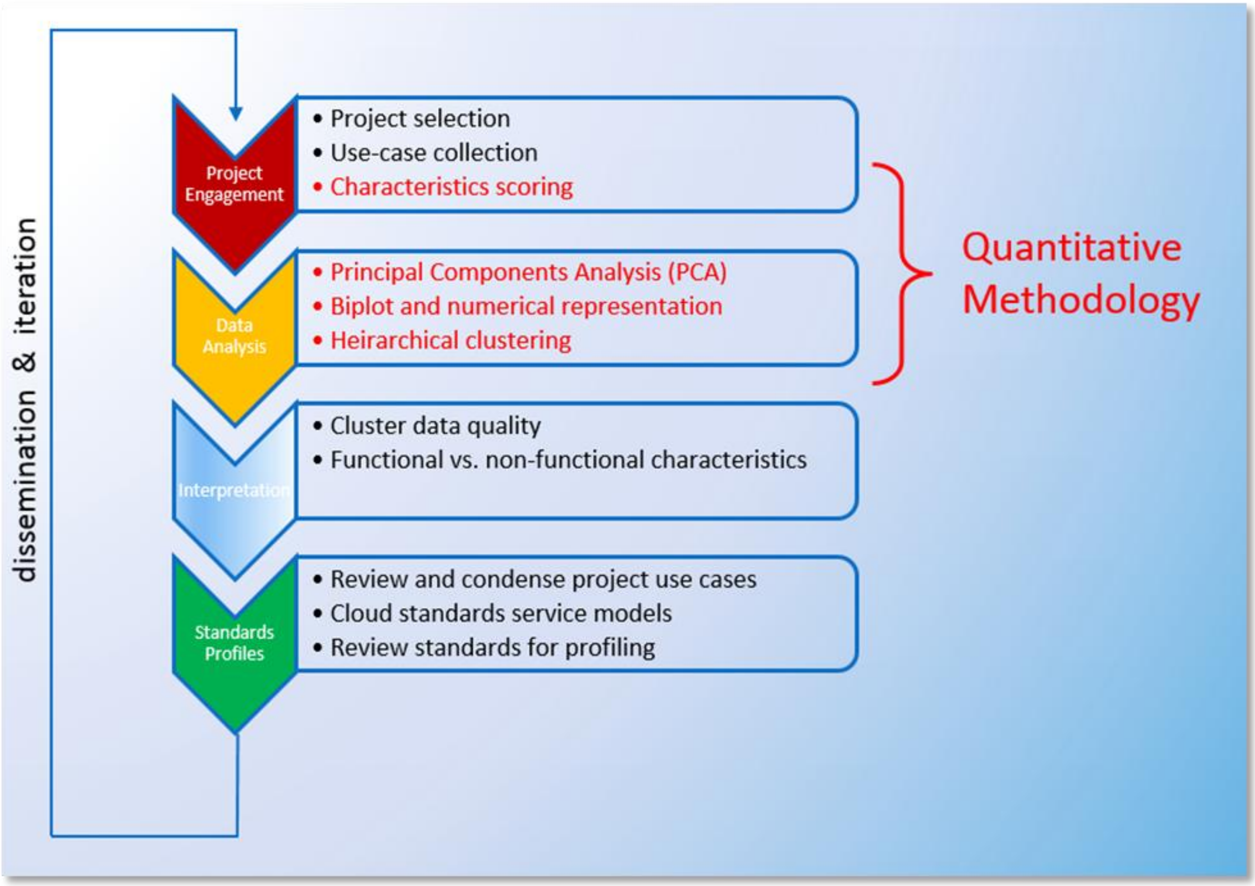

dissemination & iteration
Project Engagement
• Project selection
• Use-case collection
• Characteristics scoring
Data Analysis
• Principal Components Analysis (PCA)
• Biplot and numerical representation
• Heirarchical clustering
Interpretation
• Cluster data quality
• Functional vs. non-functional characteristics
Standards Profiles
• Review and condense project use cases
• Cloud standards service models
• Review standards for profiling
Quantitative Methodology

Figure 1. Greater workflow incorporating the quantitative methodology.

3 Data

3.1 Definition of variables

The definition of characteristics that we use was derived and published by NIST. It is the most commonly cited third party definition of cloud computing [6]. It is also interesting to note that this definition took several years and 16 drafts to reach conclusion – indicative of the difficulties associated with reaching a definition upon which even a small number of people are in agreement.

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction and the NIST model intends to capture that complexity in simple, understandable characteristics. The model is composed of five essential characteristics, three service models, and four deployment models. In earlier drafts of the definition, NIST included a further eight common characteristics, that were later dropped in the final version. We use the full set of 13 characteristics here and reserve a comparative analysis of the full and reduced sets for future work.

In the following sections we provide short explanations of each of the 13 characteristics.

3.2 Essential Characteristics

3.2.1 On-demand self service

Consumers can log on to a website or use web services to access additional computing resources on demand, that is, whenever they want them, without talking to a sales representative or technical support staff.

3.2.2 Broad network access

Because they are web-based, you can access cloud computing services from any internet-connected device. With a web browser on a desk-top machine, or even a thin client computer terminal, you can do any computing that the cloud resources provide.

3.2.3 Resource pooling

In multi-tenanted computing clouds the customers share a pool of computing resources with other customers, and these resources, which can be dynamically reallocated, may be hosted anywhere.

3.2.4 Rapid elasticity

Cloud computing enables computing resources or user accounts to be rapidly and elastically provisioned or released so that customers can scale their systems up and down at any time according to changing requirements.

3.2.5 Measured service

Cloud computing providers automatically monitor and record the resources used by customers or currently assigned to customers, which makes possible the pay-per-use billing model that is fundamental to the cloud computing model.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

3.3 Common characteristics

3.3.1 Massive Scale

A cloud platform may, depending on the resources offered, provide individual users with access to large-scale or even massive-scale computing.

3.3.2 Homogeneity

In many situations it is advantageous to both customers and providers to have essentially homogeneous systems at their disposal. Where requirements are particularly difficult or unusual, a cloud platform may be built out of non-homogeneous systems and components.

3.3.3 Virtualization

Virtualization of machines as software systems massively increases the scale of cloud resources that can be made available. Virtualization is not an essential characteristic but it is becoming the only way that scale demands can be met by providers; customers generally don’t care either way as the virtualization is entirely transparent.

3.3.4 Low Cost Software

If increased scale reduces per-unit, or per-use cost, then cloud computing offers a drive towards lower-cost software. It is important to note that this may not be the case across all sectors and activities.

3.3.5 Resilient Computing

In some sectors, continuous availability of computing with zero-downtime is crucial to the sectors requirements, for example, emergency and financial systems. In these sectors requirements for resilient, rather than just fail-safe computing will be the norm.

3.3.6 Geographic Distribution

Some sectors have legal requirements that physical data stores are in particular geographical jurisdictions. This places certain restrictions on providers favouring a cloud-anywhere model. More commonly the user is not concerned about location per se.

3.3.7 Service Orientation

The design of the services that run and operate on the cloud frameworks are normally operated as services such they can take advantage of other factors that give resilience. This includes the ability to scale different components within the system depending on their load and capability.

3.3.8 Advanced Security

There may be the capability to perform both system and network level security within the cloud system.

3.4 Characteristic scoring

CloudWATCH has compiled an informative dataset by scoring each of 38 European cloud projects against the NIST full-list described above. Scoring is on an integer scale from 1 to 9, indicating low to high percieved importance of the characteristic. Scoring was done either by project representatives or CloudWATCH in collaboration with project representatives. Table 1 below shows the compiled scores for all 38 selected projects. Colour coding indicates low values in red, and high values in blue.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

Table 1. Raw scores as compiled for individual projects.

	On Demand Self-Service	Broad Network Access	Resource Pooling	Rapid Elasticity	Measured Service	Massive Scale	Homogeneity	Virtualization	Low Cost Software	Resilient Computing	Geographic Distribution	Service Orientation	Advanced Security
ARTIST	3	5	8	7	8	2	1	1	6	1	7	7	3
ASCETiC	7	2	5	7	9	7	3	5	7	8	7	6	2
BETaaS	7	8	6	7	6	4	3	4	2	7	6	7	5
BigFoot	9	4	9	9	1	9	1	9	1	6	4	4	1
BNCweb	7	4	4	3	2	4	5	3	5	2	1	7	2
Broker@Cloud	3	4	7	4	7	2	8	7	2	8	5	9	7
Catania Science Gateway	8	6	6	7	5	6	6	8	7	5	6	5	5
CELAR	8	4	7	9	5	2	4	7	5	9	6	7	3
CloudCatalyst	6	6	6	4	1	1	1	6	8	4	4	6	5
CloudLightning	9	7	8	7	6	9	5	5	6	5	5	5	3
CloudScale	9	9	6	9	6	9	3	6	7	1	1	9	1
CloudSpaces	9	9	9	9	9	9	7	9	9	9	7	9	9
CloudTeams	9	9	7	2	1	1	7	1	8	1	2	1	7
CloudWave	8	8	8	8	9	4	3	7	8	7	3	9	5
COMPOSE	7	4	6	7	6	2	4	7	7	6	2	9	4
DICE	2	2	7	8	8	6	2	6	8	9	7	6	6
Embassy Cloud	7	7	2	1	6	3	6	8	3	8	1	7	9
GEMMA	8	7	8	8	7	3	3	4	3	8	5	8	9
INPUT	7	5	5	8	6	4	2	8	2	5	8	5	2
IOStack	9	9	9	8	8	7	7	9	6	9	2	9	2
LEADS	9	8	2	4	3	7	7	8	8	8	5	9	3
Leicester	6	6	5	5	3	2	7	3	4	5	2	6	8
MCN	9	9	5	1	5	9	2	8	5	5	9	9	6
Mobizz	9	7	9	7	7	6	8	4	6	9	7	8	9
MODAClouds	8	4	5	9	9	9	1	1	8	8	8	9	1
OpenModeller	9	7	8	7	3	7	8	8	9	4	7	8	5
PaaSword	7	1	7	1	1	3	1	3	1	5	4	9	9
PANACEA	8	9	8	8	6	6	5	7	5	8	7	9	8
S-CASE	9	7	7	3	3	5	3	3	7	6	4	9	7
SeaClouds	7	3	3	9	9	7	2	4	7	8	9	8	2
SeaClouds	8	4	5	9	9	9	1	1	8	8	8	9	1
STORM CLOUDS	6	7	8	8	9	6	3	6	6	4	7	8	9
SUPERCLOUD	8	2	7	4	2	3	8	8	2	9	7	5	9
Texel	5	8	7	7	8	4	4	4	4	8	3	8	7
Umea	5	3	3	2	2	2	4	6	4	3	3	7	7
U-QASAR	5	7	6	7	7	2	2	5	4	4	4	7	6
Varberg	8	7	8	5	4	3	4	5	3	6	3	8	8
WeNMR	9	8	8	9	5	7	9	8	7	3	9	6	3

Note that the project *SeaClouds* has double entries as we received two independent assessments from two different project representatives. The entries are broadly similar, but not exactly the same. Note also that there are only two entries that are exactly the same, *MODAClouds* and the second entry for *SeaClouds.*

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

4 Methods

In this section we provide a more detailed description of the analytical steps involved in moving from raw data scores, to coherent clusters of projects. We also describe the main elements of interpretation at each step.

4.1 PCA ordination

Principal Components Analysis (PCA) was first introduced by Karl Pearson in 1901 [3a], and later independently discovered by Harold Hotelling in 1933 [3b]. It has since become a corner-stone of modern data exploration [3c], and remains an important tool in data-mining and machine learning applications.

PCA is essentially an ordination technique that transforms the values of variables and produces an entirely new set of uncorrelated component variables such that the information content is maximally concentrated in the higher order components, to the extent that weighted linear combinations of values allows. The total information content of the original dataset is preserved, though it can be interpreted as the higher order components being more meaningful than the lower order ones in a way that is not possible with the original variables. In this sense PCA is a dimension reduction technique that requires a criterion for choosing how many components to keep. If the original variables are first centred and standardized (also known as z-scoring), as we have applied here by subtracting the mean and dividing by the standard deviation, then the resulting components have special properties: i) Component scores will also be mean-centred and have variances equal to the eigenvalues of the respective components. ii) The sum of the variances will be equal to the number of original variables. It is convenient to think of PCA as providing a dimensionally reduced, de-noised representation of the original data, with no loss in meaningful information content and with added interpretability.

Figure 2 shows a Biplot of the scores for the 38 projects on the first three synthetic components of a PCA. Components 1 and 2 are on the conventional X- and Y-axes, and Component 3 is colour coded on the Z-axis.

This transformed representation is effectively the landscape of the ecosystem of cloud computing, as represented by the perceived relationship between these projects and the NIST defining characteristics.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

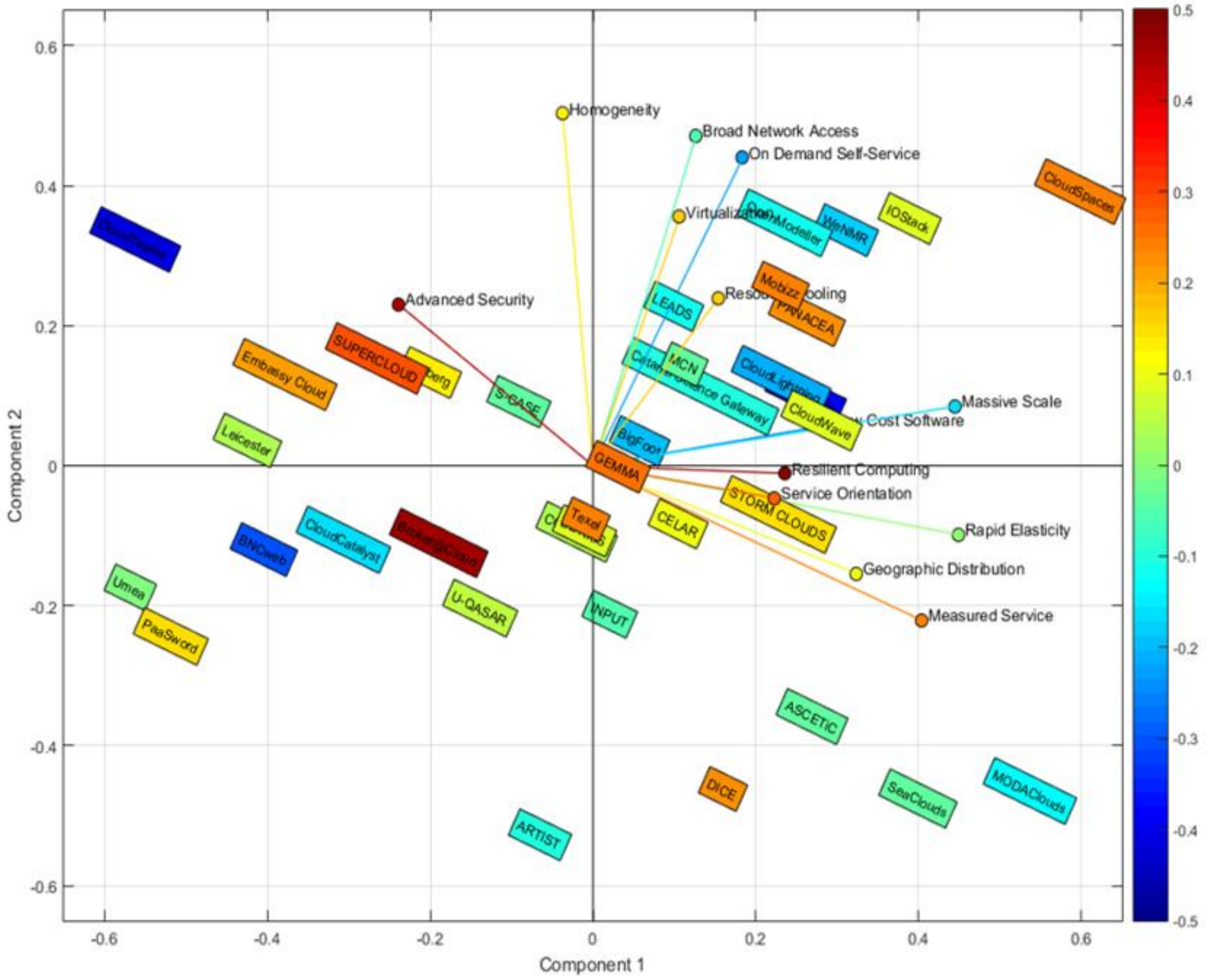


Figure 2. Biplot of 38 European cloud computing projects.

The viewer is intuitively drawn in to the identification of groups of observations (the projects) in this new synthetic space, but beyond the groups, the individual components themselves are more difficult to interpret. This is where additional features of the Biplot [4] prove useful and we explain this further in sections below, but first there is the issue of how many of the components are meaningful.

4.2 Stopping rule

As mentioned above, Component 1 can be considered the most meaningful representation, followed by Component 2, and so on down to Component 13. We therefore must question how many of these components should be kept, with the others discarded (known as the stopping-rule)? Again we have adopted one of the simplest of the classic approaches known as the Kaiser-Guttman criterion [7]. This criterion states that you should keep only those components whose eigenvalues are greater than one, or equivalently, those components whose variances are greater than one when using z-scored raw data. The rationalisation for this is that even random data can achieve a component variance of one, so if we are to keep only meaningful information we should discard all components with smaller variances. This is the element of de-noising in PCA – information is concentrated in the higher-order components and the noise is left behind in lower-order components, and these can be discarded.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

Figure 3 below shows a so-called scree plot of components vs eigenvalues, indicating that only the first five components have eigenvalues greater than one and should be kept, the rest discarded.

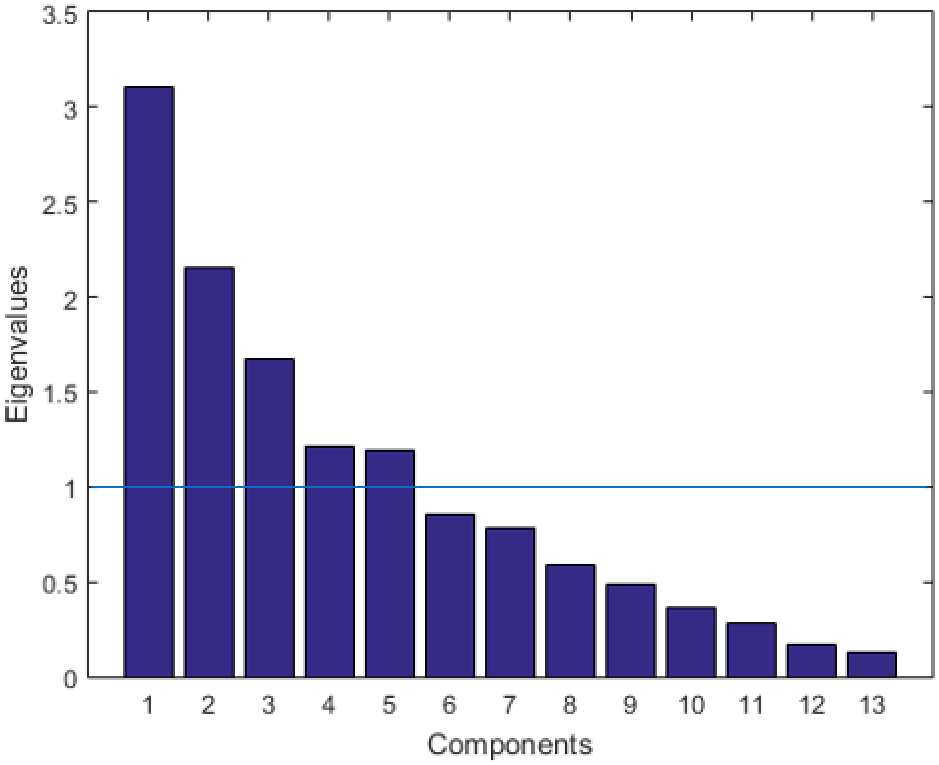


Figure 3. Scree Plot of the component eigenvalues.

So, by the Kaiser-Guttmann criterion, in this analysis we should keep only the first five components, though we can only easily visualise the PCA space in at most three dimensions at a time.

4.3 Interpreting the biplot

The biplot was first introduced by Ruben Gabriel [4a, b] as a way of combining the subject and variable space in a single visualisation. Referring back to Figure 2, we see the observations (subjects) plotted in PCA space, but we also see labeled vectors representing the variables plotted in the same space.

Technically, the length and orientation of the vectors is given by the eigenvectors (coefficients) of the PCA. Visual interpretation is easy: vectors pointing in the same direction are correlated, vectors at right-angles are uncorrelated, and vectors pointing in opposite directions are negatively correlated. Technically the cosine of the angle between any pair of vectors is the correlation coefficient.

Any observation can be also projected onto any of the vectors. Technically this is done by taking the inner matrix product of the scores and coefficients, but visually by dropping the observation perpendicularly onto the vector. Note that this projection can be done in any number of dimensions. If we performed that projection using all 13 components, we would recover the original scores exactly. But we can apply the stopping rule, and use the first five components only to achieve new, synthetic, de-noised estimates of the values on the original variables. It's these new estimates that we want to use in our subsequent analysis.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

Table 2 below shows the projected scores in five-dimensional space. Colour coding again indicates low values in red, and high values in blue.

Table 2. Five-dimentional projected scores.

	On Demand Self-Service	Broad Network Access	Resource Pooling	Rapid Elasticity	Measured Service	Massive Scale	Homogeneity	Virtualization	Low Cost Software	Resilient Computing	Geographic Distribution	Service Orientation	Advanced Security
ARTIST	-2.9954	-0.8551	0.9328	1.341	1.3491	-1.8087	-3.2776	-3.6273	0.4696	-2.0489	-0.1213	-0.7875	-0.6271
ASCETiC	-0.4722	-1.8513	-0.8811	1.2014	0.8952	1.389	-1.8344	-0.0846	0.0811	0.754	2.0358	0.01	-1.8942
BETaaS	-0.7266	-0.3648	0.2336	0.1141	0.2694	-0.5393	-0.3033	-0.2654	-0.4488	0.1692	0.1366	-0.0399	0.3495
BigFoot	1.1965	-1.6288	0.7691	1.2799	-1.9822	1.1492	0.3388	1.747	-1.2443	-0.4724	2.1119	-3.2468	-1.9038
BNCweb	0.0259	-0.4727	-2.0241	-2.3917	-2.2835	-1.0584	-0.9349	-1.3439	0.119	-2.2685	-2.2252	-0.7711	-0.9742
Broker@Cloud	-2.0893	-1.1106	0.1715	-1.0155	0.4313	-2.1618	0.447	0.4842	-2.5252	1.9505	-0.051	0.7493	2.5777
Catania Science Gateway	1.2651	0.1457	0.3532	0.3559	-0.9535	0.8382	0.7983	1.0049	0.0198	-0.3164	0.4615	-1.0802	-0.6798
CELAR	-0.2592	-1.0799	0.3376	0.7646	0.1295	0.3291	-0.076	0.8158	-0.9381	0.8968	1.4144	-0.5715	-0.2167
CloudCatalyst	-0.5473	-0.3247	-0.566	-1.2998	-1.5179	-1.4286	-0.6548	-1.2186	-0.3112	-1.8249	-1.525	-1.0527	-0.2068
CloudLightning	1.488	0.7337	0.8833	1.2601	-0.4083	1.3493	0.3935	0.3891	1.0533	-0.9748	0.6186	-1.1765	-1.2822
CloudScale	1.9797	2.0256	-0.4147	0.4614	0.1139	1.8418	-0.3243	-1.0548	3.1877	-1.9778	-0.9368	0.3953	-1.8993
CloudSpaces	1.8672	2.4549	1.8844	2.0025	2.1668	2.158	2.3344	1.9247	1.4163	2.0438	1.2758	1.6719	1.0479
CloudTeams	0.9489	1.9138	1.0331	-1.8256	-3.6417	-2.3564	1.1492	-1.4084	0.3033	-4.5192	-3.6219	-2.983	0.3201
CloudWave	0.2214	1.5556	0.788	0.9109	1.6521	0.5579	0.2398	-0.4433	1.4367	0.2748	-0.0952	1.3291	0.4924
COMPOSE	-0.4289	-0.0656	-0.6234	-0.5898	0.2058	-0.3474	-0.3856	-0.4873	0.0872	0.0553	-0.5118	0.6606	0.2003
DICE	-2.4634	-2.1765	0.4663	1.6661	1.887	-0.4034	-2.091	-0.753	-1.1132	1.284	2.1592	0.0407	-0.1941
Embassy Cloud	0.0224	-0.1069	-2.0646	-3.544	-1.5898	-1.5095	1.473	0.987	-1.5482	0.9988	-2.1735	1.3067	1.9674
GEMMA	-1.1218	0.3671	1.2406	0.4541	1.0871	-1.0695	0.3563	-0.1787	-0.6006	0.758	0.0458	0.5081	1.6476
INPUT	-0.3368	-2.057	-0.1658	0.7153	-0.7088	0.3965	-0.9107	0.5158	-1.1649	0.1164	1.6618	-1.6245	-1.361
IOStack	1.9657	2.0047	1.0047	1.0332	0.951	1.7972	1.9365	1.537	1.2805	1.0582	0.5311	1.027	0.431
LEADS	2.5906	0.7243	-2.557	-1.8621	-1.008	2.0437	1.2571	1.6188	1.0922	0.7336	-0.7811	1.5565	-0.8481
Leicester	-0.8243	0.0918	-0.2013	-1.9673	-1.4988	-2.313	0.416	-0.6665	-1.114	-1.0405	-2.0606	-0.529	1.3991
MCN	1.619	0.5261	-1.4291	-0.9154	-0.3684	1.4114	0.8168	1.0733	0.7557	0.6343	-0.2919	1.084	-0.4875
Mobizz	0.5033	1.419	1.6616	0.9659	1.1297	0.2974	1.6928	1.1023	0.0891	1.2561	0.4643	0.726	1.5314
MODAClouds	-0.3712	-0.8924	-1.1554	1.9806	2.5261	2.3864	-2.8478	-1.3194	2.1242	0.4593	1.9446	1.4238	-2.6507
OpenModeller	2.3437	1.6662	0.7801	0.5852	-0.5552	1.5504	1.7479	1.3063	1.1314	-0.331	0.0607	-0.4046	-0.4109
PaaSword	-2.1409	-1.8578	-1.2892	-2.4625	-1.4285	-2.8167	-0.6832	-0.6881	-2.4485	-0.0785	-1.3696	-0.3675	1.2823
PANACEA	0.5907	1.2474	1.2112	0.8431	1.1588	0.571	1.4591	1.0798	0.2479	1.3326	0.5189	0.9395	1.21
S-CASE	0.3764	1.164	-0.6036	-1.2043	-0.2961	-0.4174	0.3568	-0.6105	0.8173	-0.7106	-1.6861	0.8198	0.4762
SeaClouds	-0.6257	-1.9163	-1.5754	1.328	1.8011	1.9739	-2.5365	-0.483	0.7682	1.1746	2.2379	1.0577	-2.3056
SeaClouds	-0.3712	-0.8924	-1.1554	1.9806	2.5261	2.3864	-2.8478	-1.3194	2.1242	0.4593	1.9446	1.4238	-2.6507
STORM CLOUDS	-0.8468	0.6826	1.4517	1.3743	1.6876	-0.2684	-0.2421	-0.6938	0.4527	0.2928	0.448	0.5216	0.7485
SUPERCLOUD	-0.0938	-1.6926	0.3439	-1.15	-2.1322	-1.2256	2.0121	2.7227	-3.5858	1.6068	0.7062	-1.6446	1.5831
Texel	-1.4464	0.4409	0.6938	0.073	1.3458	-1.2556	-0.1984	-0.8857	-0.1841	0.5092	-0.4677	1.0686	1.5665
Umea	-1.2092	-1.5321	-2.085	-3.0878	-2.2272	-2.2873	-0.5992	-0.6935	-1.7809	-0.8083	-1.9859	-0.4646	0.516
U-QASAR	-1.6621	-0.2714	0.3322	-0.1633	0.3675	-1.5994	-1.0425	-1.5616	-0.3221	-0.7178	-0.6984	-0.0865	0.6192
Varberg	-0.4006	0.6717	0.5855	-0.9433	-0.519	-1.5074	0.9917	-0.0228	-0.7373	-0.1765	-1.2361	-0.0336	1.6534
WeNMR	2.4287	1.3137	1.6331	1.7313	-0.562	1.947	1.573	1.5009	1.0101	-0.5523	1.0614	-1.4556	-1.0268

Inspecting Table 2 and comparing with the raw scores in Table 1 reveals much about the richness of the analysis. The highest and lowest values for the synthetic scores are *CloudScale/Low Cost Software* (3.1877), and *CloudTeams/Resilient Computing* (-4.5192). The high value corresponds to a 7 in the raw scores, which is not the highest value in that column or row. The low value corresponds to a 1 in the raw scores, the lowest score available, but not the only one in that column. Compare the high value for *CloudScale* with the values for *CloudSpaces* in the following row. All but two of the variables have been given a high score of 9, the remaining two a 7. The corresponding row of synthetic scores are all positive, but not extreme. The analysis is able to adjust for different perceptions in scoring and to emphasis the hidden meaning underlying the scores.

In the next section we use the synthetic scores to discover the natural grouping of projects located in the landscape that is described by the biplot.

CloudWATCH is funded by the EC FP7 - DG Connect Software & Services, Cloud. Contract No. 610994

4.4 Hierarchical clustering

There are many algorithms for performing clustering on raw, or transformed data, which have been devised for various different purposes [5a]. We adopt here a simple classic approach called single-linkage Euclidian distance clustering [5b]. This is an agglomerative hierarchical approach. The algorithm proceeds as follows: i) in the n-dimensional Euclidian space of the input data-set, find the two points that are closest together and record the distance, ii) create a new point at the mid-point between these two and discard the two points, iii) continue finding the two closest points, recoding the order of joining until there is only a single point left. The outcome of this simple procedure is a hierarchical cluster tree as shown below in Figure 4.

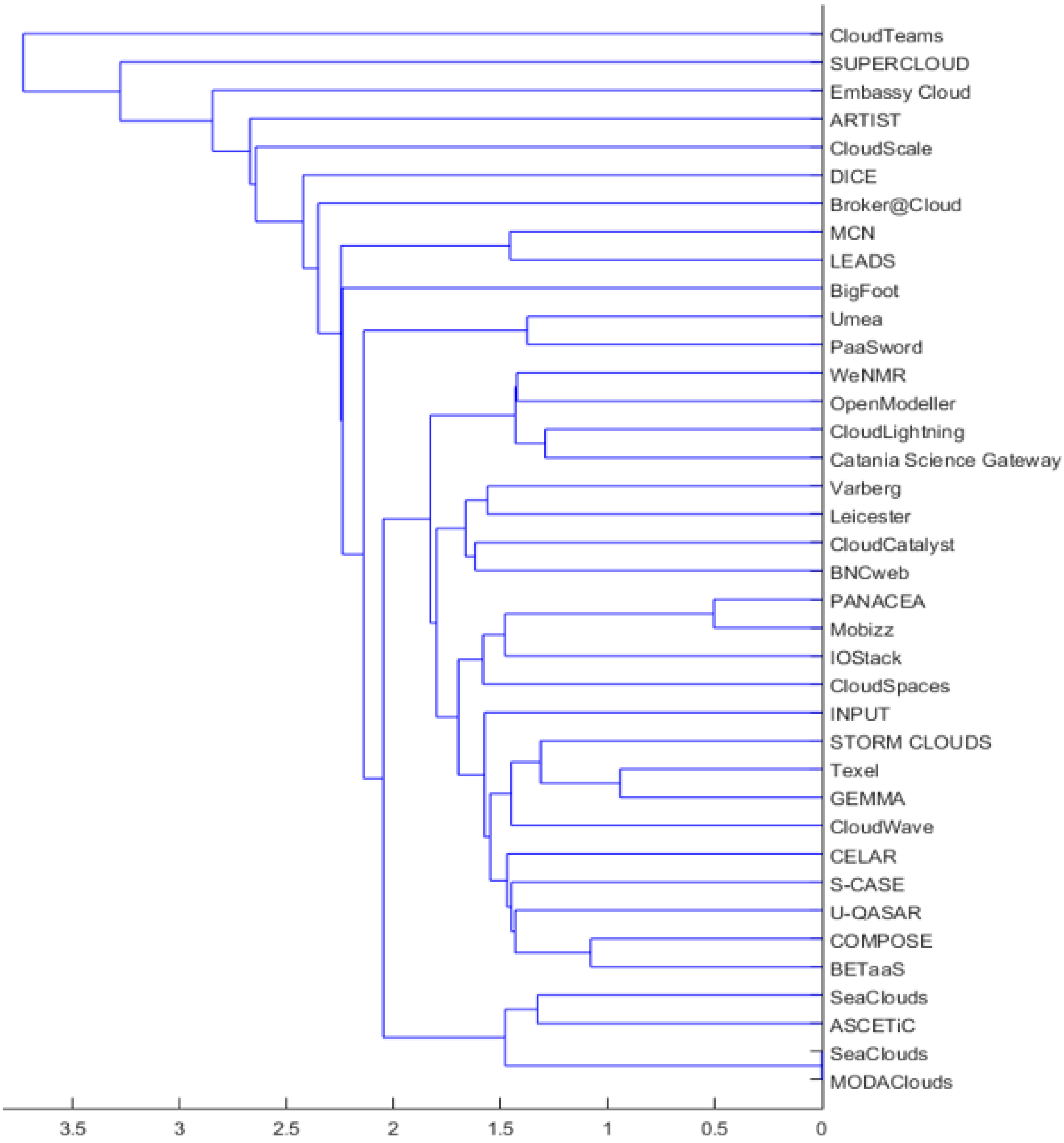


Figure 4. Single-linkage hierarchical clustering.

Intuitively there is a clear relationship between the biplot and the cluster tree and one can recreate the above steps of the algorithm visually by inspecting the location of observations in the biplot – except that the biplot can only be visualised in two or three dimensions. In order to use any more dimensions of the data-space, you need to do the calculations in n-dimensional Euclidean space as described above.

Clusters are read from the branching order of the tree. The closest two points are *SeaClouds* and *MODAClouds* at the bottom of the figure – not surprising as their data entries are identical, so their distance is zero as shown on the x-axis in Figure 4. The next closest points are projects *PANACEA* and *Mobizz.* They appear close in the upper right quadrant of the biplot too. Next closest are *Texel* and *Gemma* near the origin of the biplot (note the colour coding showing Component 3.)

At the other end of the cluster tree *CloudTeams* appears to be a complete outlier, not aligned particularly closely with any other project, in this representation of the cloud ecosystem landscape. *SUPERCLOUD* is another outlier, though more closely allied to the rest of the projects than to the extreme outlier *CloudTeams*.

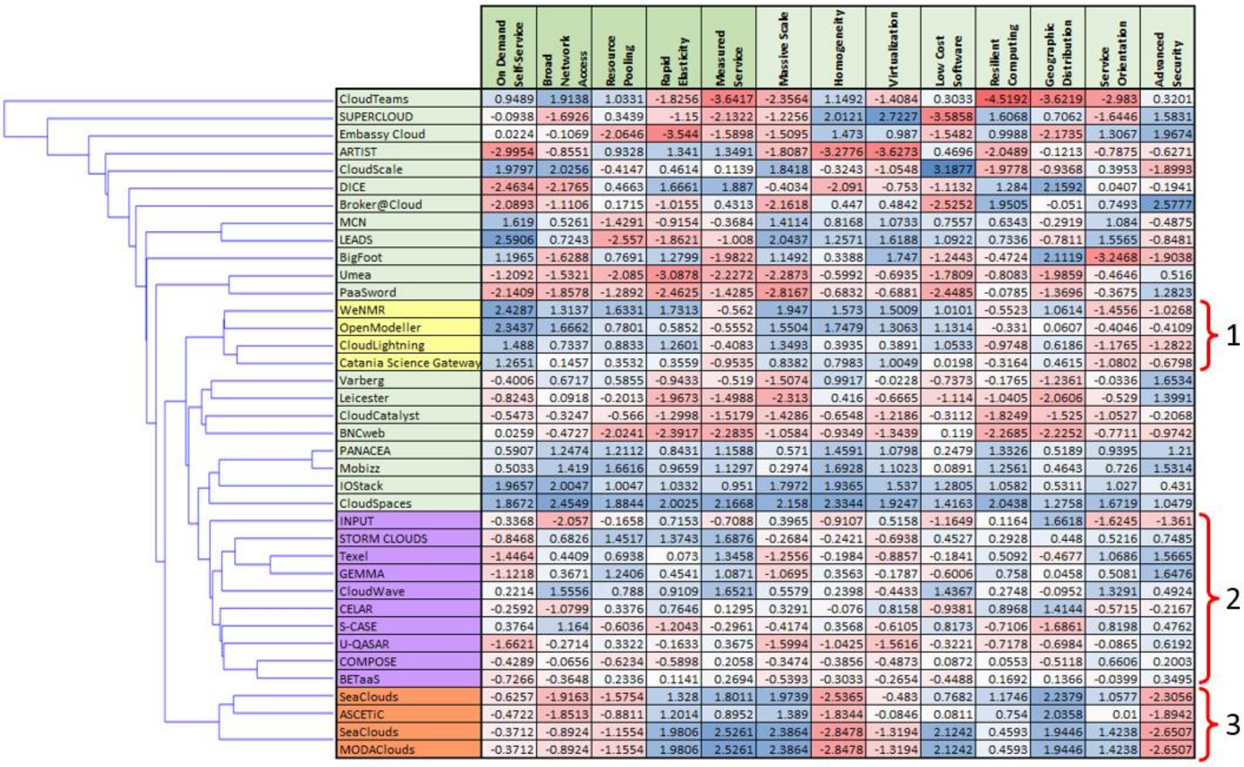
4.5 Interpreting the clusters

We identify three recognisable clusters in the tree and we used these for an illustrative analysis of standards profiles in D4.3. These clusters are familiar enough to be named:

- **Cluster 1 – Scientific computing**. This cluster comprises a number of projects that aim at highly distributed data processing in an academic context.
- **Cluster 2 – Trusted public clouds for government**. This cluster consists of a set of initiatives driven by public sector organisations.
- **Cluster 3 – High performance, dedicated purpose applications**. This cluster is similar to Cluster 1, but comprises projects concentrating on high performance computing that are more focussed regarding their objectives.

Figure 5 shown below combines the cluster tree with a reordered table of synthetic scores for easy comparison. On the right of the figure we identify the above clusters. Inspection of the banding of the colour coding in the scores table provides a ready visual aid to identifying useful clusters. (In D4.3 we referred to this as interpreting the heat-map.) Note the complete lack of banding near the top of the tree where there is the cascade of outliers, further reinforcing this interpretation.

New projects may, by inspection of the scores tables and the location of other projects in the biplot and cluster tree, be able to locate themselves in the cloud ecosystem landscape. In this way they may discover other projects with a similar relationship to cloud features to warrant collaboration on standards profiles and implementations. Alternatively, they might simply repeat the analysis presented here using these projects as a benchmark in order to more accurately discover their location in the landscape, and their relationship to other projects.



	On Demand Self-Service	Broad Network Access	Resource Pooling	Rapid Elasticity	Measured Service	Massive Scale	Homogeneity	Virtualization	Low Cost Software	Resilient Computing	Geographic Distribution	Service Orientation	Advanced Security
CloudTeams	0.9489	1.9138	1.0331	-1.8256	-3.6417	-2.3564	1.1492	-1.4084	0.3033	-4.5192	-3.6219	-2.983	0.3201
SUPERCLOUD	-0.0938	-1.6926	0.3439	-1.15	-2.1322	-1.2256	2.0121	2.7227	-3.5858	1.6068	0.7062	-1.6446	1.5831
Embassy Cloud	0.0224	-0.1069	-2.0646	-3.544	-1.5898	-1.5095	1.473	0.987	-1.5482	0.9988	-2.1735	1.3067	1.9674
ARTIST	-2.9954	-0.8551	0.9328	1.341	1.3491	-1.8087	-3.2776	-3.6273	0.4696	-2.0489	-0.1213	-0.7875	-0.6271
CloudScale	1.9797	2.0256	-0.4147	0.4614	0.1139	1.8418	-0.3243	-1.0548	3.1877	-1.9778	-0.9368	0.3953	-1.8993
DICE	-2.4634	-2.1765	0.4663	1.6661	1.887	-0.4034	-2.091	-0.753	-1.1132	1.284	2.1592	0.0407	-0.1941
Broker@Cloud	-2.0893	-1.1106	0.1715	-1.0155	0.4313	-2.1618	0.447	0.4842	-2.5252	1.9505	-0.051	0.7493	2.5777
MCN	1.619	0.5261	-1.4291	-0.9154	-0.3684	1.4114	0.8168	1.0733	0.7557	0.6343	-0.2919	1.084	-0.4875
LEADS	2.5906	0.7243	-2.557	-1.8621	-1.008	2.0437	1.2571	1.6188	1.0922	0.7336	-0.7811	1.5565	-0.8481
BigFoot	1.1965	-1.6288	0.7691	1.2799	-1.9822	1.1492	0.3388	1.747	-1.2443	-0.4724	2.1119	-3.2468	-1.9038
Umea	-1.2092	-1.5321	-2.085	-3.0878	-2.2272	-2.2873	-0.5992	-0.6935	-1.7809	-0.8083	-1.9859	-0.4646	0.516
PaaSword	-2.1409	-1.8578	-1.2892	-2.4625	-1.4285	-2.8167	-0.6832	-0.6881	-2.4485	-0.0785	-1.3696	-0.3675	1.2823
WeNMR	2.4287	1.3137	1.6331	1.7313	-0.562	1.947	1.573	1.5009	1.0101	-0.5523	1.0614	-1.4556	-1.0268
OpenModeller	2.3437	1.6662	0.7801	0.5852	-0.5552	1.5504	1.7479	1.3063	1.1314	-0.331	0.0607	-0.4046	-0.4109
CloudLightning	1.488	0.7337	0.8833	1.2601	-0.4083	1.3493	0.3935	0.3891	1.0533	-0.9748	0.6186	-1.1765	-1.2822
Catania Science Gateway	1.2651	0.1457	0.3532	0.3559	-0.9535	0.8382	0.7983	1.0049	0.0198	-0.3164	0.4615	-1.0802	-0.6798
Varberg	-0.4006	0.6717	0.5855	-0.9433	-0.519	-1.5074	0.9917	-0.0228	-0.7373	-0.1765	-1.2361	-0.0336	1.6534
Leicester	-0.8243	0.0918	-0.2013	-1.9673	-1.4988	-2.313	0.416	-0.6665	-1.114	-1.0405	-2.0606	-0.529	1.3991
CloudCatalyst	-0.5473	-0.3247	-0.566	-1.2998	-1.5179	-1.4286	-0.6548	-1.2186	-0.3112	-1.8249	-1.525	-1.0527	-0.2068
BNCweb	0.0259	-0.4727	-2.0241	-2.3917	-2.2835	-1.0584	-0.9349	-1.3439	0.119	-2.2685	-2.2252	-0.7711	-0.9742
PANACEA	0.5907	1.2474	1.2112	0.8431	1.1588	0.571	1.4591	1.0798	0.2479	1.3326	0.5189	0.9395	1.21
Mobizz	0.5033	1.419	1.6616	0.9659	1.1297	0.2974	1.6928	1.1023	0.0891	1.2561	0.4643	0.726	1.5314
IOStack	1.9657	2.0047	1.0047	1.0332	0.951	1.7972	1.9365	1.537	1.2805	1.0582	0.5311	1.027	0.431
CloudSpaces	1.8672	2.4549	1.8844	2.0025	2.1668	2.158	2.3344	1.9247	1.4163	2.0438	1.2758	1.6719	1.0479
INPUT	-0.3368	-2.057	-0.1658	0.7153	-0.7088	0.3965	-0.9107	0.5158	-1.1649	0.1164	1.6618	-1.6245	-1.361
STORM CLOUDS	-0.8468	0.6826	1.4517	1.3743	1.6876	-0.2684	-0.2421	-0.6938	0.4527	0.2928	0.448	0.5216	0.7485
Texel	-1.4464	0.4409	0.6938	0.073	1.3458	-1.2556	-0.1984	-0.8857	-0.1841	0.5092	-0.4677	1.0686	1.5665
GEMMA	-1.1218	0.3671	1.2406	0.4541	1.0871	-1.0695	0.3563	-0.1787	-0.6006	0.758	0.0458	0.5081	1.6476
CloudWave	0.2214	1.5556	0.788	0.9109	1.6521	0.5579	0.2398	-0.4433	1.4367	0.2748	-0.0952	1.3291	0.4924
CELAR	-0.2592	-1.0799	0.3376	0.7646	0.1295	0.3291	-0.076	0.8158	-0.9381	0.8968	1.4144	-0.5715	-0.2167
S-CASE	0.3764	1.164	-0.6036	-1.2043	-0.2961	-0.4174	0.3568	-0.6105	0.8173	-0.7106	-1.6861	0.8198	0.4762
U-QASAR	-1.6621	-0.2714	0.3322	-0.1633	0.3675	-1.5994	-1.0425	-1.5616	-0.3221	-0.7178	-0.6984	-0.0865	0.6192
COMPOSE	-0.4289	-0.0656	-0.6234	-0.5898	0.2058	-0.3474	-0.3856	-0.4873	0.0872	0.0553	-0.5118	0.6606	0.2003
BETaaS	-0.7266	-0.3648	0.2336	0.1141	0.2694	-0.5393	-0.3033	-0.2654	-0.4488	0.1692	0.1366	-0.0399	0.3495
SeaClouds	-0.6257	-1.9163	-1.5754	1.328	1.8011	1.9739	-2.5365	-0.483	0.7682	1.1746	2.2379	1.0577	-2.3056
ASCETiC	-0.4722	-1.8513	-0.8811	1.2014	0.8952	1.389	-1.8344	-0.0846	0.0811	0.754	2.0358	0.01	-1.8942
SeaClouds	-0.3712	-0.8924	-1.1554	1.9806	2.5261	2.3864	-2.8478	-1.3194	2.1242	0.4593	1.9446	1.4238	-2.6507
MODAClouds	-0.3712	-0.8924	-1.1554	1.9806	2.5261	2.3864	-2.8478	-1.3194	2.1242	0.4593	1.9446	1.4238	-2.6507

Figure 5. Projected scores reordered on the cluster tree.

To emphasise the utility of identifying and interpreting appropriate clusters, we quote a brief section from the discussion in D4.3 (p.38):

> Since the clustering effort discussed in this document makes all participants aware of commonalities among participant projects in a cluster, an alternative approach of developing and defining standards may make much more sense. Instead of providing the profiling of a specific standard together with others into a single profile document targeted at one given cluster discussed in this document, it may be more successful to encourage clusters (and individual projects) working together on one single profile aligned with one single cloud characteristic in one single document. Once all individual profiles are finished, an identified cluster would simply have to write a very brief cluster profile document incorporating by reference any fitting individual profile documents.
>
> However, such a synergetic approach is feasible and achievable only if the use cases of each project and cluster for the characteristic in question are sufficiently overlapping to arrive at a common solution. Otherwise, clusters would have to work on their own cluster-specific standards profile for a given cloud characteristic.
>
> For example, consider the common cloud characteristic "Advanced Security". Across individual projects, the majority considers it relatively important. However, projects in Cluster 2 consider it in the top 3 of their most important characteristics. Interestingly, Cluster 2 hosts governmental cloud activities: *STORMCLOUDS*, *Texel*, *and Gemma*. None of these projects agree on the importance on any of the other cloud characteristics. The same is true of another cluster which was identified but

not analysed further as part of this document. Here, two projects were governmental cloud activities: *Varberg* (Municipality Social Services Administration), *Leicester* (City Council).

5 Summary and Conclusions

We have presented a methodology for characterising the ecosystem landscape of cloud computing based on the NIST defining features. The same methodology identifies the location of a project or cloud enterprise within the landscape. Taken together, the biplot (Figure 2), the table of estimated scores (Table 2), and the cluster tree (Figure 4), offer a rich interpretive tool to aid standards development and standards profiling, and to aid new cloud enterprises in identifying their location within that landscape.

We present this analysis as a benchmark of the landscape, and we invite both standards organisations, and new or existing cloud enterprises to engage in this way of representing the landscape to better aid uptake, and to help build consistency and trust in cloud computing.

6 Next Steps

We have already begun a collaboration with IEEE P2301 [2] to expand the benchmark and to improve the cloud definition encapsulated by NIST [1]. We also intend to publish the body of this work in the *Journal of Cloud Computing: Advances, Systems and Applications*, including a comparative analysis of the short and long versions of the NIST definition.

7 Notes and References

[1] The National Institute of Standards and Technology (NIST) published the 16th draft working definition of cloud computing as *The NIST Definition of Cloud Computing* (NIST Special Publication 800-145).

[2] P2301 - *Guide for Cloud Portability and Interoperability Profiles (CPIP).*

[3] (a) Pearson, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2 (11): 559–572. (b) Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, and 498–520. (c) Jolliffe, I.T. 2002. *Principal Component Analysis*, 2nd edition, Springer-Verlag.

[4] (a) Gabriel, K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58 (3): 453–467. (b) Gabriel, K.R. 1981. Biplot display of multivariate matrices for inspection of data and diagnois. In V. Barnett (Ed.) *Intrepreting multivariate data*. London, John Wiley & Sons. (c) Greenacre, M. 2010. *Biplots in Practice*. BBVA Foundation, Madrid, Spain.

[5] (a) Hastie, T., Tibshirani, R. and Friedman, J. 2009. Hierarchical clustering. In *The Elements of Statistical Learning*, 2nd edition, New York, Springer. pp. 520–528. (b) The Mathworks ® R2015a Documentation. *Agglomerative hierarchical cluster tree.*

[6] *uptothecloud.6dg.co.uk*

[7] (a) Guttman, L. 1954. Some necessary conditions for common factor analysis. *Psychometrika*, 19, 149-161. (b) Kaiser, H. F. and Dickman, K. 1959. Analytic determination of common factors. *American Psychologist*, 14, 425. (c) Horn, J.L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika,* 30, 179-185.