Clustering Model for Evaluating SaaS on the Cloud

1 Mrs. Dhanamma Jagli, 2 Mrs. Akanksha Gupta

1 Assistant Professor, V.E.S Institute of Technology, Mumbai, India
2 Student, M.E (IT) 2nd year, V.E.S Institute of Technology, Mumbai, India.

Abstract
Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Software as a Service (SaaS) is one of the most important part of cloud computing, and it can be used for providing different types of business solution. In real world several organizations had successfully adapted the SaaS concept. In order to realize the benefits of SaaS, it has to be evaluated properly. The existing SaaS evaluation models are only focusing on quality attributes of software, similar to conventional software services. But the SaaS on the cloud needs to be considered related characteristics of cloud. For this purpose the new evaluation model is proposed based on Data mining techniques of clustering. The proposed model would be helpful to the service providers and service users to evaluate SaaS on the Cloud.

Keywords: Cloud Computing, Software as a Service (SaaS), Data Mining, Clustering methods.

I. INTRODUCTION
Cloud computing, or the cloud, is a colloquial expression used to describe a variety of different types of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet. In science, cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time. Cloud computing exhibits the key characteristics like Agility, Application , Cost, Device and location independence, Virtualization, Multitenancy, Reliability, Scalability, elasticity, Performance, Security, Maintenance. The importance of is that it removes the normal hassles and costs of using business software. Software is synonymous with complexity and costly development, upgrades and support, often encouraged by the software supplier. In effect, SaaS empowers your organization to manage the cost of technology by providing the software on a "pay as you use", basis. SaaS brings software applications previously too expensive for anyone but the large corporate of this world to every organization at a fraction of the effort and price. The characteristics of SaaS are as follows:

- Configuration and Customization: SaaS applications similarly support what is traditionally known as application customization. In other words, like traditional enterprise software, a single customer can alter the set of configuration options that affect its functionality.
- Open integration protocols: Since SaaS applications cannot access a company's internal systems (databases or internal services), they predominantly offer integration protocols and application programming interfaces (APIs) that operate over a wide area network. Typically, these are protocols based on HTTP, REST, SOAP and JSON.
- Collaborative (and "social") functionality: Inspired by the success of online social networks and other so-called web 2.0 functionality, many SaaS applications offer features that let its users collaborate and share information.

II. LITERATURE REVIEW

1. Data Mining
Data Mining is a process of discovering interesting knowledge from data of large amount which is stored in database, data warehouse or other information repositories. Data mining process is to extract information from a data set and transform it into an understandable structure for further use as shown in fig.1 [1]. The Knowledge Discovery in Databases (KDD) process is commonly defined with the following stages:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) Data Mining
(5) Interpretation/Evaluation
KDD is known as a simplified process such as pre-processing, data mining and result validation. Pre-processing is necessary to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with misplaced data. Data mining involves seven common classes of tasks like Anomaly detection, Association rule learning, Clustering, Classification, Regression, Summarization, Sequential pattern mining. The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. A commonly used data mining technique is clustering, with classification of objects into different groups by partitioning sets of data into a series of subsets (clusters).

2. Clustering
Cluster Analysis: The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity [1]. The following are typical requirements of clustering in data mining:

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirement for domain knowledge to determine input parameters
- Ability to deal with noisy data
- Incremental clustering and insensitivity to the order of input records
- High dimensionality
- Constraint based clustering
- Interpretability and usability

3. Types of Clustering:
The clustering methods can be classified in following categories as shown in fig. 2 [14]
Partitioning Method is a simplest and most fundamental version of cluster analysis is partitioning which organizes the objects of a set into several exclusive groups or clusters [1]. These are of two types: k-Means and k-Medoids.

Hierarchical Method works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as agglomerative and divisive. Agglomerative hierarchical clustering is a bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination are satisfied. Divisive hierarchical clustering is a top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions [1].

Density based Method: Clustering based on density (local cluster criterion), such as density-connected points. It consists of major features like discover clusters of arbitrary shape, handle noise and one scan [1]. There are mainly three types: DBSCAN, OPTICS and DENCLUE.
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of density-connected points.

OPTICS: Ordering Points To Identify the Clustering Structure, features are as follows:
- Produces a special order of the database with respect to its density-based clustering structure
- This cluster-ordering contains information equivalent to the density-based clustering corresponding to a broad range of parameter settings
- Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
- Can be represented by graphically or using visualization techniques.

DENCLUE: (DENsity-CLUstEring) is a clustering method based on a set of density distribution functions.

---

**Grid Based Clustering** approach uses a multi resolution grid data structure. It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. STING is a grid-based multi resolution clustering technique in which the spatial area is divided into rectangular cells. There are usually several levels of such rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure. CLIQUE (CLustering In QUEst) was the first algorithm proposed for dimension-growth subspace clustering in high-dimensional space. In dimension-growth subspace clustering, the clustering process starts at single-dimensional subspaces and grows upward to higher-dimensional ones [1].

**Model Based Clustering** Method attempts to optimize the fit between the given data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. These are of three types- Expectation-Maximization, Conceptual clustering and a neural network approach to clustering. The EM (Expectation-Maximization) algorithm is a popular iterative refinement algorithm that can be used for finding the parameter estimates. It can be viewed as an extension of the k-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean. Conceptual clustering is a form of clustering in machine learning that, given a set of unlabeled objects, produces a classification scheme over the objects. Neural networks have several properties that make them popular for clustering [1].

4. **Trends of Cloud Computing**
As the cloud computing industry continues to grow at a rapid pace, the definition of the cloud’s capabilities and characteristics develops as well. There are some current cloud computing trends that are changing the cloud computing industry like IaaS (Infrastructure-as-a-Service), Cloud Security, Disaster Recovery, Hybrid Cloud Computing, Consumption.

5. **Service models of Cloud**
Cloud computing providers offer their services according to several fundamental models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) where IaaS is the most basic and each higher model abstracts from the details of the lower models.

a. **Infrastructure as a service (IaaS)**
In the most basic cloud-service model, providers of IaaS offer computers - physical or (more often) virtual machines - and other resources. IaaS clouds often offer additional resources such as a virtual-machine disk image library, raw (block) and file-based storage, firewalls, load balancers, IP addresses, virtual local area networks (VLANs), and software bundles.
b. Platform as a service (PaaS)
In the PaaS model, cloud providers deliver a computing platform, typically including operating system, programming language execution environment, database, and web server. Application developers can develop and run their software solutions on a cloud platform without the cost and complexity of buying and managing the underlying hardware and software layers.

c. Software as a service (SaaS)
In the business model using software as a service (SaaS), users are provided access to application software and databases. Cloud providers manage the infrastructure and platforms that run the applications. SaaS is sometimes referred to as "on-demand software" and is usually priced on a pay-per-use basis. SaaS providers generally price applications using a subscription fee. In the SaaS model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. This eliminates the need to install and run the application on the cloud user's own computers, which simplify maintenance and support. Cloud applications are different from other applications in their scalability—which can be achieved by cloning tasks onto multiple virtual machines at run-time to meet changing work demand. Software as a Service referred to as "on-demand software" supplied by ISVs (Independent Software Vendor) or "Application-Service-Providers" (ASPs). It is a software delivery model in which software and associated data are centrally hosted on the cloud. SaaS is typically accessed by users using a thin client via a web browser.

6. Importance of SaaS Evaluation
Strength of SaaS provider’s security processes and standards: Because you access SaaS provider software via the Internet and your data is stored at the provider’s site, security is of utmost importance. Security issues can be complex if the SaaS provider forces you to follow a fixed security model that is at direct odds with your corporate security model. Work with the SaaS provider on a service level agreement (SLA) for security that satisfies your needs. Provider’s ability to provide the flexibility needed to meet your needs: Flexibility involves a number of things. Can you easily add and drop features to your subscription, such as add seats, and can you do it online? Can you drop your subscription with cause, such as SaaS provider fails to deliver the agreed upon solution, without penalty? Can you integrate the SaaS solution with other corporate solutions? History of provider’s regard for its SLAs: SLAs are important for availability, performance, scalability and security. Ask the SaaS provider for permission to look at SLA performance records to see if the provider is reliable with respect to adhering to SLAs for customers. A mature SaaS provider will negotiate SLA agreements to satisfy your needs. Provider’s business viability and future outlook: Select a SaaS provider that is going to be in business for a long time. In selecting a SaaS provider, you are looking for a business partner that will not go out of business in a year. Additionally, does the provider have a good set of partners? Does the provider have a long list of satisfied customers, including reference customers for you to interview? Using different clustering techniques in SaaS, we find the answer of above written questions. With the help of clustering we make different clusters according to customer requirement for example scalability, availability, pay per use, updating of software required, time duration, etc. Providers have a long list of satisfied customer according to the satisfaction of their work with the help of clustering. Providers can create a group of clusters by applying different types of clustering algorithm on their data.

III. PROPOSED MODEL
In the proposed model following pre-processing steps needs to performed before applying model. Pre-processing is an important issue for both data warehousing and data mining as a real world data tend to be incomplete, noisy and inconsistent. The proposed model contains two steps to be followed for mining SaaS data. Pre-processing steps for SaaS data:

1. Extract SaaS data
2. Transform SaaS data
3. Load in RDBMS/MBMBS format
4. Apply clustering algorithm
5. Obtain appropriate cluster

1. Extract SaaS data:
SaaS data are normally not stored in any kind of data repository, effective and efficient management and analysis of stream data poses a great challenge. So that only SaaS data from source data needs to be extracted as it is and separated from actual source data.

2. Transform SaaS data
Transformation is the method of dealing with inconsistencies and noisy with SaaS data after extracting SaaS data, SaaS data to be transformed to common format and removed noisy and inconsistency and missing values.

3. Load in MDBMS format
Stream data are generated continuously in a dynamic environment, with huge volumes, infinite flow and fast changing behavior. It is impossible to store such data completely in the data warehouse. Most stream data represent low level information, consisting of various kinds of detailed temporal and other features. To find interesting or unusual patterns, it is essential to perform multi dimensional analysis on aggregate measures such as sum and average.

4. Apply clustering algorithm
The Proposed model needs is applied after loading data stream in the multidimensional format and the model has to follow some systematic steps. SaaS data are stored in the multidimensional formats. Then the MDBMS data is applied as input to various clustering methods, each cluster key attributes to be compared as per given SaaS data. Find out the gap, such as regional gaps, dollar gaps etc. in the SaaS data. Then apply the suitable clustering algorithm on given SaaS data. After applying clustering algorithm cluster output will be generated in terms of performance. Finally clustering analysis can be done and find out which cluster method is more efficient and suitable for given SaaS data in terms of performance of response.

5. Obtain appropriate Cluster
By applying clustering algorithms, clusters can be formed based on software services provided on the cloud by service providers. These clusters will help to service users to make a decision suitable for their requirement for software services.

IV. CONCLUSION
Clustering model for evaluating SaaS will help to evaluate potential software services on the cloud computing by using Data Mining Clustering algorithms. The clustering model would be greatly helpful to software service providers to evaluate their own services to the cloud users. It helps service provider to increase availability of software services on the cloud computing environment suitable for cloud users demand and requirement. It also helps cloud users to evaluate potential software services available on the cloud computing environment.

References
[4] Sami Ayramy  and Tommi Karkk “Introduction to partitioning-based clustering methods with a robust example”.


[13] Soumita Seth, Debnath Bhattacharyya, Tai-hoon Kim, “Constraint-Based Clustering to Find Communities in a Social Network”.


AUTHOR

**Mrs. Dhanamma Jagli** is an Assistance professor in V.E.S Institute of Technology, Mumbai, currently Pursuing PhD in Computer Science and Engineering and received M.Tech in Information Technology from Jawaharlal Nehru Technological University, Hyderabad. She has around 10 years teaching experience at the postgraduate and under graduate level. She had published and presented papers in referred international journals and various national, international conferences. Her areas of research interest are Data Mining, Cloud Computing, Software Engineering, Data base Systems and Embedded Real time systems. She has been associated with Indian Society of Technical Education (ISTE) as a life member.

**Mrs. Akanksha Gupta** is a final year student of M.E. (Information Technology) from Vivekanand Education Society’s Institute of Technology (VESIT), Mumbai. She has completed her B.E. (Information Technology) from Dr. B. R Ambedkar University, Agra. Her areas of research interest are Data Mining, Cloud Computing, Software Engineering, Data base Systems.