Journal of Computer Engineering and Information Technology

Perspective

Data Aggregation in Cloud Using Map Reduce Framework

Daniel Rokhsar*

Department of Computer Engineering, University of Visvesvaraya Technology, Belagavi, Karnataka, India.

***Corresponding Author:** Daniel Rokhsar, Department of Computer Engineering, University of Visvesvaraya Technology, Belagavi, Karnataka, India. E-mail: rokhsaranie@gmail.com

Description

Cloud computing is one of the important component of today's tech savvy society. it creates a new paradigm for information exchange without investing in a new infrastructure or licensing new software. It expands the capabilities of existing traditional way of accessing the application software, system software, storage and others through internet. In the last few years, cloud computing has grown tremendously from a promising business concept to fastest growing segments of the IT industry. At a remarkable pace, cloud computing has transformed how business and government function by providing various services like IaaS, PaaS and SaaS. Huge amount of data is being generated by various applications on the cloud called big data applications. Big data applications need appropriate framework and techniques to store, aggregate and retrieve the data. Consequently the objective of this research paper has divided into two sections firstly to identifying the traditional methods of data aggregation and optimization and then evaluates aggregation of data through minimization of links association between Mapper and Reducer to reduce the data traffic in the network. Secondly to present a viable solution to overcome these major problems using high level scripting language on Apache Hadoop Framework.

The data aggregation and optimization for big data applications within a cloud environment. Key words are data aggregation, big data and cloud computing. Traditional application integration technologies are performed in a rigid and slow process that usually takes a long time to build and deploy, requiring professional developers and domain experts. Since the face of the Internet is continually changing, as new services and novel applications appear and become globally noteworthy at an increasing pace. Now a day the locus of computation is changing, with functions migrating to remote data centres through internet based communication. Cloud computing is the idea that data and programs can be stored centrally, in the cloud and accessed anytime from anywhere through thin clients and lightweight mobile devices. This brings many advantages, including data ubiquity, edibility of access and resilience. The services of cloud computing are broadly divided into three categories. Infrastructure-as-a-Service, Platform-as-a-Service and Software-as-a-Service Cloud computing also is divided into five layers including clients, applications, platform, infrastructure and servers. In this system focus on the typical application of cloud computing, comparison of various cloud computing models, cloud computing characteristics, security challenges, review the several cloud deployment and service models.

Benefits of Cloud Computing

The explore certain benefits of cloud computing over traditional IT service environment including scalability, flexibility, reduced capital and higher resource utilization are considered as adoption reasons for cloud computing environment. I also include security, privacy, internet dependency and availability as avoidance issues. Cloud computing has very wide area of applications. The main objective behind any application of this domain is to provide seamless connectivity from multiple locations and on multiple different devices. A survey from Evans Data shows that 40 percent of developers working on open source plan to deliver their applications as web services offerings using cloud providers. Few of them are listed below. It is one of the perfect tools which offer servers, load balancers, or DNS tables to get an app on the cloud. The database is integrated well with open source programming languages like python and others. It has various features which helps developers to build an app without much difficulty Force.com for Google App engine is a set of tools and services to enable developer success with application development in the cloud. This helps to create entirely new web and business applications like social networks i.e. linkedin.com and facebook.com. An open source technology for creating offline web applications. It is a single standard for offline applications.

Google offers google gears as a free, fully open source technology in order to help every web application not just google applications. It is a tool provided for developers who want to write applications that are going to run partially or entirely in a remote datacenter. It is an Internet scale cloud services platform hosted in Microsoft datacenters, which provides an operating system and a set of developer services. Public cloud, this cloud is basically used to offer services for general public users or large group of users which may belong to one industry and is owned by an organization like Google which is going to provide services based on the requirements, demands of general public. Private cloud, it is completely isolated cloud which provides its unique services to the employees of particular organization that may exist on premise or off premise. This cloud is accessible only from private users from within the organization. It is managed by the organization or a third party. Community cloud when the cloud infrastructure is shared by several organizations those are having common mission, interest, security requirements and policy. It is managed by organizations or third party. On-demand (OD) computing is an increasingly popular enterprise model in which computing resources are made available to the user as needed.

Data Aggregation

The resources may be maintained within the user's enterprise, or made available by a service provider. The on-demand model was developed to overcome the common challenge to an enterprise of being able to meet fluctuating demands efficiently. Broad Network access, Broad network access refers to resources hosted in a private cloud network operated within a company's firewall that are available for access from a wide range of devices, such as tablets, PCs, Macs and smartphones. These resources are also accessible from a wide range of locations that offer online access. Resource Pooling, Multi-tenants environments where multiple customers share adjacent

resources in the cloud with their peers and creates a pool of multiple resources to provide different kind of services like computing ,networks and storage services and offers software solutions also reduces the operational cost of these resources is called resource pooling. Rapid Elasticity, Elastic computing is one of the critical characteristics which fulfill the immediate requirements of any business with minimum delay. An organization can easily add or remove users, software features and other resources. Amazon named their cloud platform Elastic compute cloud all the services which is offered by polled resources are measured on run time basis. Depends upon the usage it is measured and bill is generated. It indicates visibility and transparency to consumption rates and costs. It helps cross-departmental reporting and budgeting. Cloud computing can provide infinite computing resources on demand due to its high scalability in nature, which eliminates the needs for Cloud service providers to plan far ahead on hardware provisioning. Big companies such as Google, Microsoft and Amazon rapidly involved in developing cloud computing applications and add more functionality into it to cater large amount of users.

Cloud Computing has already started to revolutionize the way we store and access data. There are various types of issues raised when we discuss about data aggregation in cloud. Capturing data, curation, storage, searching, sharing, transfer, analysis and presentation. Google solved this problem using an algorithm called Map Reduce. This algorithm divides the task into small parts and assigns those parts to many computers connected over the network and collects the results to form the final result dataset. Data aggregation is a type of data and information mining process where data is searched, gathered and presented in a report-based, summarized format to achieve specific business objectives or processes and conduct human analysis. Data aggregation may be performed manually or through specialized software. Data aggregation is a component of business intelligence solutions. Data aggregation personnel or software search databases find relevant search query data and present data findings in a summarized format that is meaningful and useful for the end user or application. Data aggregation generally works on big data or data marts that do not provide much information value as a whole. Data aggregation's key applications are the gathering, utilization and presentation of data that is available and present on the global Internet. Data aggregation is the process of collecting and aggregating the useful data. Data aggregation is considered as one of the fundamental processing procedures for saving the energy.