Improving constrain based Hadoop schedulers and MapReduce performance

A Dissertation Submitted

By

NEPPALI RAJASEKHAR

To

Department of Computer Science and Engineering

In partial fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science and Engineering

Under the guidance of

Ms. Divyajot Gill

(May 2015)

LOVELY PROFESSIONAL UNIVERSITY

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student	:Rajasekhar Neppali	Registration No	:11310472
Batch	:2013-2015	Roll No	:RK2307B60
Session	:2014-2015	Parent Section	:K2307

Details of Supervisor:

Name	:Kirandeep Kaur	Designation	: Assistant Professor
UID	:17687	Qualification	: M.E
		Research Exp.	:1.5 year

Specialization Area: Cloud Computing

Proposed Topics:-

1. Improve Constrain based Hadoop scheduling with Oozie.
2. Survey on Hadoop scheduling.
3. Social networks with Mapreduce.

Kirandeep 17687

Signature of supervisor

PAC Remarks: Topic is Approved. Paper is to pasted. 11570

APPROVAL OF PAC CHAIRMAN

Signature:

Date:

*Supervision should finally encircle one topic out of three proposed topics and put up for an approval before Project Approval Committee (PAC).
*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.
*One copy to be submitted to supervisor.

Abstract

Hadoop is an multipurpose framework for approve to high range processing of information over a group of distributed terminals. This formal definition shows the fact that Hadoop is a multi-tasking method that contains multiple information sets for number of jobs, many users simultaneously. This capability of multi-handling denotes that Hadoop is the chance to ideally guide to the services of resources in a way that improves their use. The Hadoop framework to execute the capacity of pluggable schedulers that regulate resources to jobs for the MapReduce functions. MapReduce has developed for processing large data volume jobs in distributed terminals. The large number of jobs and variety of jobs to be processed across heterogeneous clusters are increasing in present days, so the complexity of scheduling of the jobs efficiently to meet required intents of performance. All the scheduling algorithms are not fit (effective) for the getting job size and job locality. Hadoop supports pluggable scheduler's so we are comparing the performance of different schedulers (FIFO, FAIR, Capacity etc.) and present the new constrain based algorithm for job sharing.

ACKNOWLEDGEMENT

I would like to show my greatest appreciation to my mentor "**Ms. Divyajot Gill"**. I can't say thank you enough for the tremendous support and help. I feel motivated and encouraged every time i attended his meeting. Without his encouragement and guidance this report work would not have materialized.

Neppali Rajasekhar

CERTIFICATE

This is to certify that **"NEPPALI RAJASEKHAR"** has completed M.Tech dissertation titled **"Improve Constrain based Hadoop Scheduling and MapReduce Performance"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

The dissertation is fit for the submission and the partial fulfilment of conditions for the award of M.Tech Computer Science & Engineering.

Date: Signature of Advisor

DECLARATION

I hereby declare that the dissertation entitled, **"Improve Constrain based Hadoop Scheduling and MapReduce Performance"** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____ Investigator: _____

Regd. No: _____

Table of content

List of Figures

Chapter 1

INTRODUCTION

The internet users are increased exponentially because of everything (Lots of Knowledge) is present in internet and to store any kind of knowledge in various cloud providers. The social networks and search engines capture & analyze every action in their networks and clusters. And improve their web site designee for easy accessing so they need to analyze and find the spam and fraud files. For example Facebook collects 15 Terabyte's of data every day into its world wide data warehouses. To identify the spam and fraud data very difficult in large, huge data warehouses, many of other social networks and cloud providers to face same problem in past few years. To store the large data like video, audio and pictures in the data warehouses through worldwide distributed systems (nodes).

1.1 Cloud computing

The word "cloud computing" means storing & accessing data using on any wired or wireless Internet connections on distributed systems. We can access the data anywhere using on distributed nodes. The cloud computing is a representation for the Internet storage. By using this service we can store the any type of data in large data basses or cloud storages. The cloud computing developed based on the utility and consumption of the computer resources.

The cloud to provide three services:

- IaaS (Infrastructure as a Service)
- PaaS (Platform as a Service)
- SaaS (Software as a service)

Infrastructure as a Service: It is providing to the IT infrastructure modules like Servers, Network & Storage, On-demand, Pay per use and Self-services. The customer is capable to organize and run random software, this includes OS (operating systems) and any software applications.

Platform as a Service: PaaS offers to the basic environment for organizing applications on cloud without troubles of organization to cloud infrastructure. The customer does not control the basic cloud infrastructure it include operating systems (OS), network, servers, and storage. To control the organized applications & maybe compering the various environment structures.

Software as a service: The SaaS also called as business solutions presented on the basis of pay-per-use to any users. It reduces the difficulty in the IT atmosphere by providing clever business solutions. The customer running the any application on cloud environment but the user cannot change any application in that cloud environment.

1.1.1 Cloud distribution models

There are four distinct models available for the users they are:

- Public cloud
- Private cloud
- Community cloud
- Hybrid cloud

Public cloud: The public cloud infrastructure is made existing to the common public or large business group. It is owned by an organization providing to the cloud facilities.

Private cloud: The private cloud infrastructure is managed by an organization. It is maintained by the association or third party organizations.

Community cloud: The community cloud can be managed by any organizations or a third party. If the users to share or maintain their data in pay per use model.

Hybrid cloud: The hybrid cloud is a structure of two or more clouds. That continue sole entities are bound collected by consistent or proprietary technology that enables data and application portability.

1.2 Apache Hadoop

Apache to create one open source framework (Hadoop) to analyze Bigdata and to find spam fraud files in large data warehouses and social networks. And to preform

MapReduce factions for easy & effective usage for social networks users and cloud users to store their data in less time of compilation and freely access their accounts.

MapReduce function is a default standard operation it is to performing processing's on huge data on distributed systems like social networks, search engines and large data storages. The MapReduce function to procedure nearly 20 petabytes of data in a day on equivalent systems.

The MapReduce function can be simplify the density of distributed system's data processing through multiple nodes into the cluster, because scalable MapReduce function to help programmers in distribute systems large data to execute in clusters. Automatically the MapReduce function handles to gathering the results across multiple systems and it returns a solo result or set.

Hadoop framework generally consists of 2 major functions,

- Hadoop Distributed File Systems (HDFS).
- MapReduce function.

1.2.1 Hadoop Distributed File Systems (HDFS)

HDFS **(**Hadoop Distributed File Systems) are fault tolerant and self handling files from distributed systems. It is developed for BigData processing and large scale workloads in data warehouses to improve the scalability, flexibility and throughput rate. The Hadoop Distributed File Systems are accepts BigData in all formats and optimizes to high bandwidth. HDFS makes various copies of information squares and disseminates them on register hubs all through a group to empower solid, to a great degree fast computations.

The key features in Hadoop Distributed File Systems are:

- High Availability
- Scale out Architecture
- Flexible Access
- Fault Tolerance
- Load balance
- Security
- Tunable Replication

- High availability: - It is serves the critical workflows and high complexity or large data file into the clusters without occur any disturbance. The workflows are getting from any social networks very easy and store the data in any data warehouses.
- Scale Out Architecture: - It is increases the server capacity to store the getting data in various users with flexible and fault tolerance manner and improves the data flow capacity.
- Flexible Access: - To access the multiple jobs and frameworks for the series of distributed files from internets. The user can access the data any ware in the world with flexible.
- Fault Tolerance: - It recovers the failure data in clusters, automatically & effortlessly from failure like network failures, system failures and any other human mistakes in distributed systems.
- Load Balancing: - It utilizes and maximum efficiency in place of data warehouses and to control the data processing in the small data basses.
- Tunable replication: - The multiple copies of any files and to provide the security of data & performance.
- Security: - The main aspect in the Hadoop Distributed Files Systems are security it provides the POSIX- based file protection for the users and optional integration in the any users acceptance.

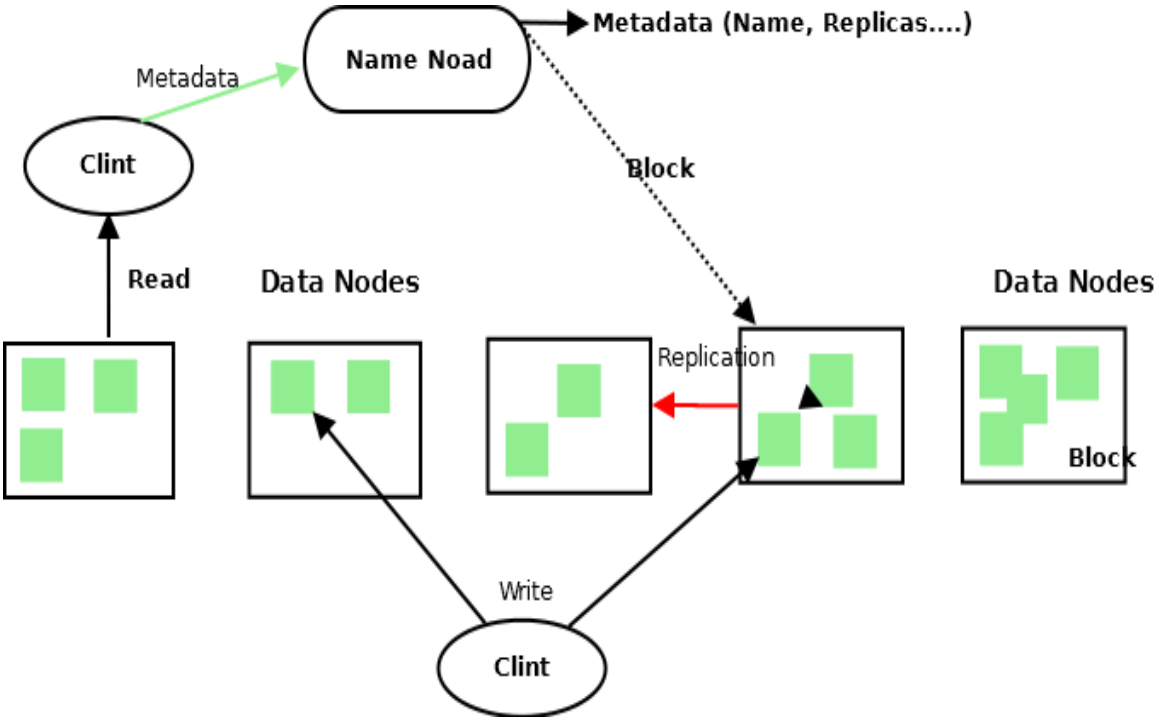


Figure 1: HDFS Architecture

1.2.1.1 NameNode and DataNodes in HDFS

The Hadoop Distributed Files System contains one NameNode and master sever it maintains the data files regulations and name spaces to access the files from Clint. Hadoop Distributed Files System to represent to the data sets Namespace and it allows user's data files into stored in clusters.

The NameNode implements file system's to namespace operations similar to open, close, and rename the directories and files. It is also defines to the mapping of slabs into the Data nodes. These Data nodes are answerable for allocate the read operation & write operations from the distributed file systems. This nodes are performs to the chunk formation, removal and repetition in order to the NameNode. Hadoop Distributed Files Systems are built by using the Java, any system that supports Java can be run the NameNode and DataNode software.

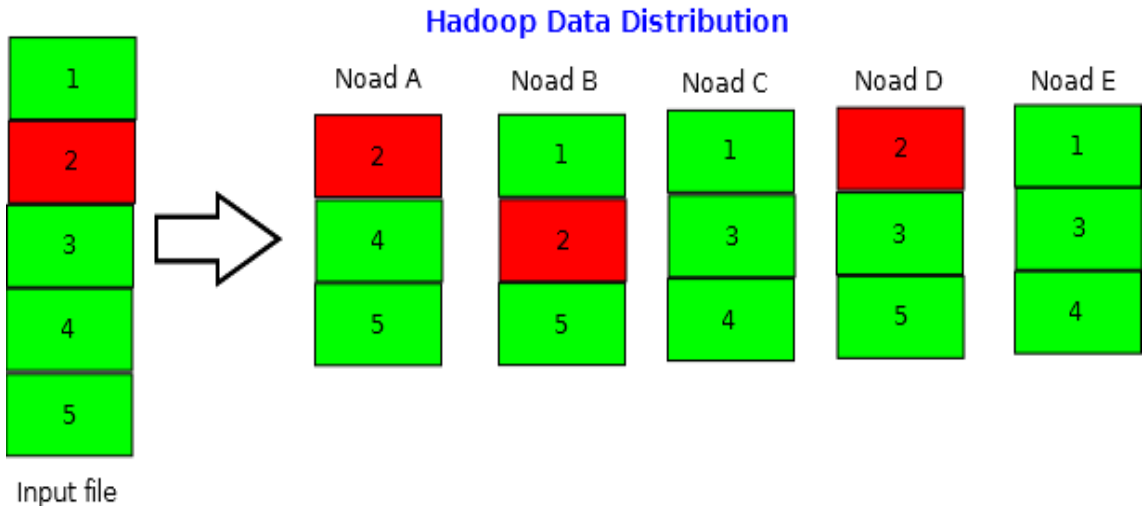


Figure 2: Hadoop data distribution across multiple nodes

1.2.2 Hadoop MapReduce function

The MapReduce function is an s/w framework to simply write a presentations which procedure the massive amounts of data like terabytes or petabytes of datasets in distributed systems or large clusters connected with thousands of nodes. These framework to process the data in fault-tolerant, reliable and flexible mode.

MapReduce function is a default standard operation to performing the processing's of huge data in distributed computing systems like social networks, search engines and large data storages. The MapReduce function to procedure nearly 20 petabytes of data in a day on parallel systems. The MapReduce function is to simplify the complication of the distributed system's data handling through the multiple nodes. The scalable MapReduce function to helps programmers to distribute programs and have them executed into the data marts. Automatically the MapReduce function handles to gathering effects across to the number of systems and return solo result or single set of result.

The MapReduce job are usually splitting the input data files into self-regulating portions. These are managed by the maptasks into a totally equivalent manner. The framework can stores (sorts) the output maps, if the input files are generated to reduce tasks. The input & output files are stored in clusters. The MapReduce function to get jobs through scheduling tasks, these are monitoring & once again executes the unsuccessive jobs or tasks. The MapReduce function to simplifies the complexity of the distributed

system's data processing across to the multiple nodes into cluster. The scalable MapReduce function helps programmers to handle the distribute programs and have them executed in parallel. Automatically the MapReduce function handles to collecting results across the multiple systems.

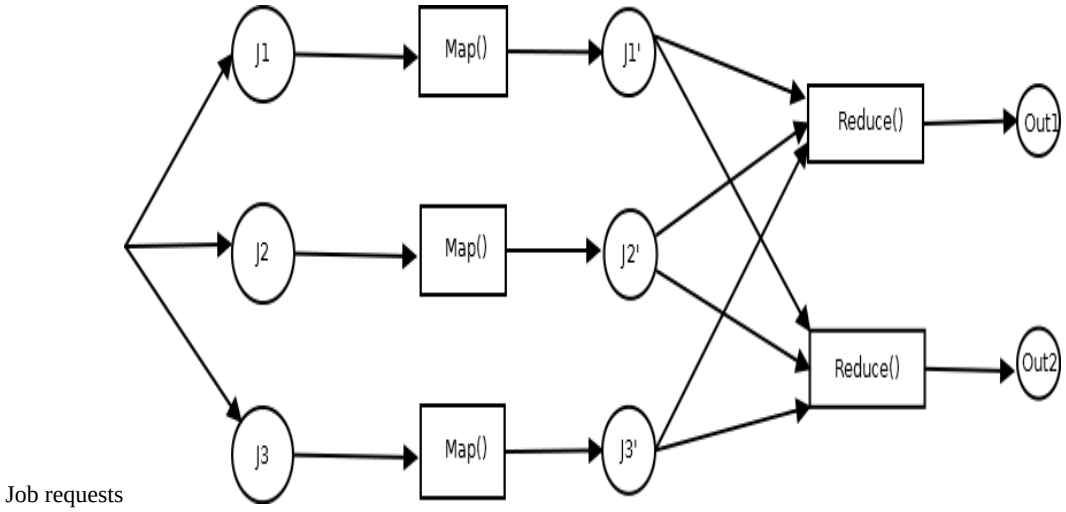


Figure 3: Simply MapReduce framework

Some of the key Features of Hadoop MapReduce Function are:

- Resource Manager
- Scale-out Architecture
- Security & Authentication
- Optimized Scheduling
- Flexibility
- Resiliency & High Availability

- Resource Manager: - The resource manager works depending on the data locality and server resource to define ideal computing procedures.
  The resource manager has two main components these are:
  a. Scheduler and
  b. Application manager

- Resiliency & High Availability: - In the huge distributed systems every user to send the jobs at a time the cluster get multiple jobs the Task trackers confirm that jobs miscarry individually and start again spontaneously.

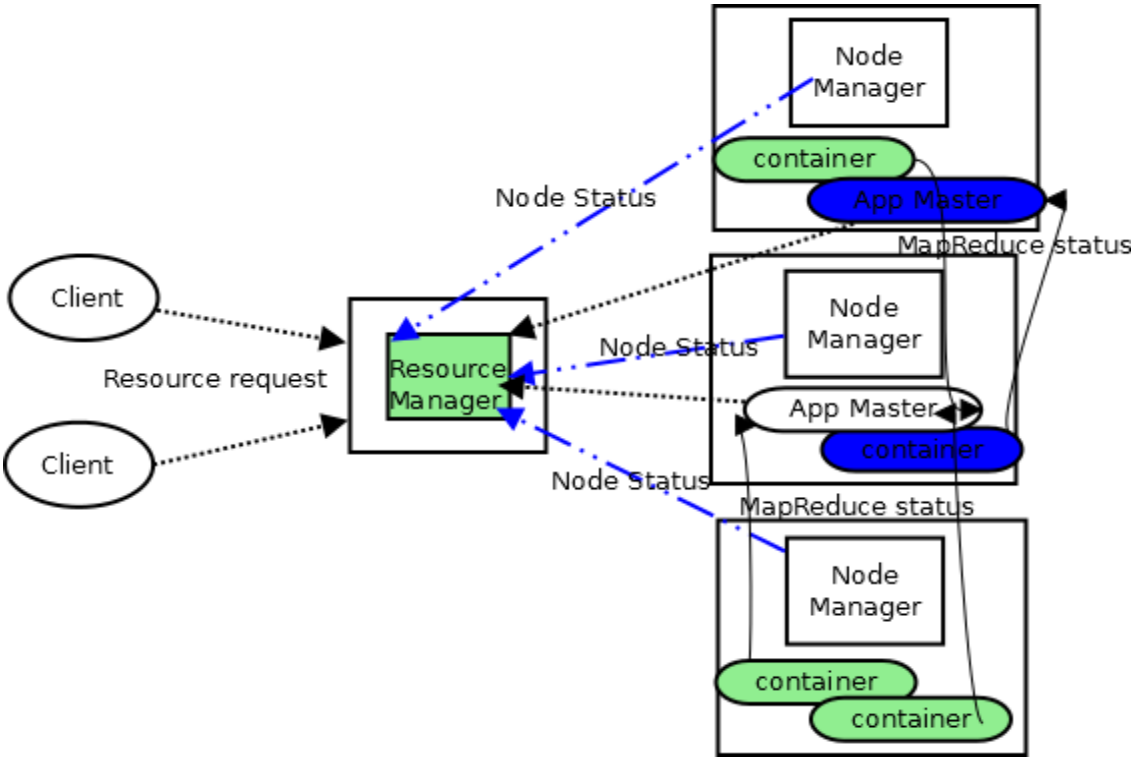


Figure4: MapReduce Resource Manager

- Scale Out Architecture: - These architecture to increase the processing power of clusters and data basses.
- Security and Authentication: - These system to works with Hadoop Distributed Files System and provides the Hbase security to make the client or user with secure data processing and data collecting in distributed systems.
- Optimized Scheduling: - The optimized scheduling to execute the jobs according to jobs size and user network capacity and job prioritization.
- Flexibility: - The programmer to can be write the effectively any programming language depending on his requirements.

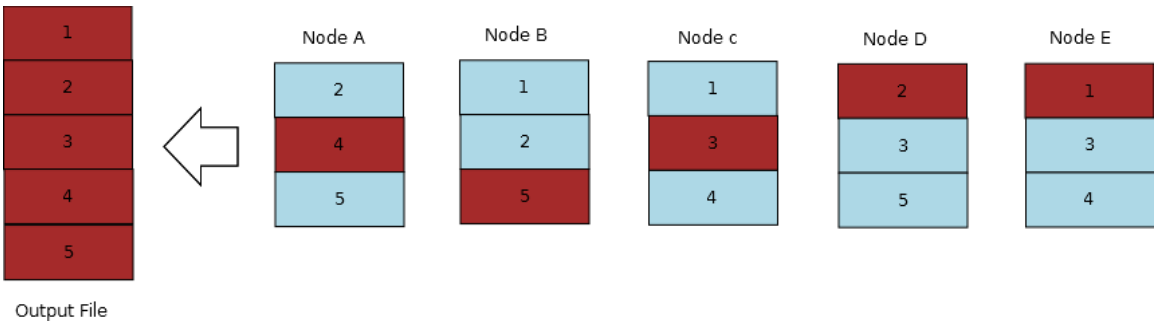


Figure5: MapReduce Computation Distribution in various nodes

1.2.3 Hadoop PIG

The Hadoop Pig is a platform for data analyzer in large data marts. It is a simple tool for making Apache MapReduce function. With the dynamic open source collection of adding to the responsibility and Pig is speedily making developments as a high level information or large data flow programming language and execution structure for breaking down large information. The framework layer incorporates a compiler that delivers MapReduce programs. In the Pig Latin is an abnormal state literary dialect that makes it simple to compose, comprehend, and look after projects. Essentially, this implies you can compose a Pig script in 15 minutes that may have taken you hours to write in the Java dialect. The Pig stage too improves undertaking execution consequently and can be stretched out with custom capacity.

The key properties of Pig are:

- Optimization opportunities.
- Ease of program**m**ing.
- Extensibility

1.2.4 Hadoop HIVE

The Hadoop Hive is useful to analyze the large data files in data warehouse. It is to providing the data association, query and analysis. It is developed by the Facebook and Apache Hadoop. Now it is used & manages other organizations. Amazon to retain software split Apache Hadoop Hive, it involved in Amazon EC2 MapReduce in Amazon Web services.

The HiveQL authorizes clients to module custom guide lessen scripts into questions. The dialect incorporates a sort framework with backing for tables containing primitive sorts, accumulations like exhibits and maps, and settled pieces of the same.

The hidden IO libraries can be reached out to question information in custom organizations. Hive likewise incorporates a framework list Megastore that contains outlines and measurements, which are valuable in information investigation, question advancement and question assemblage.

1.3 Schedulers in Hadoop MapReduce function

Hadoop executes the capacity for pluggable schedulers that regulate resources to jobs occupations. The schedulers are answerable for allocating to the jobs to familiar constraints of the clusters through queues, stack etc. We know from customary Scheduling, not all algorithms are same, and effectiveness in workload and clusters subordinate.

The schedulers are assume to vigorous part into finishing wanted to the execution stages in Hadoop framework. In many example, the job heterogeneity level of every request possibly which has the premier impact on job execution. The all input jobs are to choice a scheduler by considering the Hadoop job sections and it is hurt the job execution time interval.

1.3.1 FIFO and Fair share schedulers

The Hadoop default scheduler is FIFO, and the Hadoop to allow the pluggable scheduler. Every scheduler is not suitable for depending on jobs locality, job size and workload of cluster. Facebook to develop FAIR scheduler for their users to share the any files into the social networks at same time.

The scheduler to share the cluster depending on the number of users. The capacity scheduler algorithm are developed by yahoo. The capacity scheduler algorithm working is similar to the FAIR scheduler functionality but take the jobs in the manner of define number of named queues. Each queue performs the MapReduce function in separate of Maper function and Reducer function.

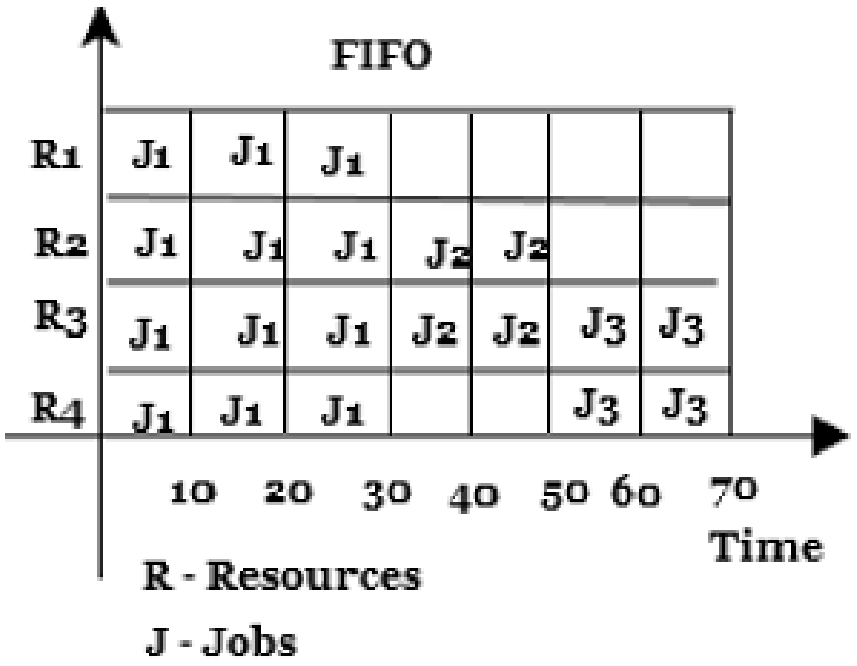


Figure6: Job Execution process in FIFO

Now the FIFO and Fair scheduler etc., these are focusing on the other characteristics of multi user Hadoop clusters and these are improving the performance of the job execution.

The Hadoop default scheduler algorithm is focused around FIFO. The jobs are executed in the applications. Some of social networks like Facebook & Yahoo helped huge data processing to creating schedulers. The reasonable scheduler and capacity scheduler are individually in this manner cleared to the Hadoop community.

Fair Sharing algorithm it presented a considerable measure of foundation data on the current schedulers and where their weaknesses lie. We anticipated actualizing a recreation of FIFO and reasonable booking, so we utilized data as a part of this paper to catch those weaknesses. The paper utilized mean fruition time as one of their benchmark examinations, which was helpful on the grounds that our speculation is constructing totally with respect to looking at mean finish time.

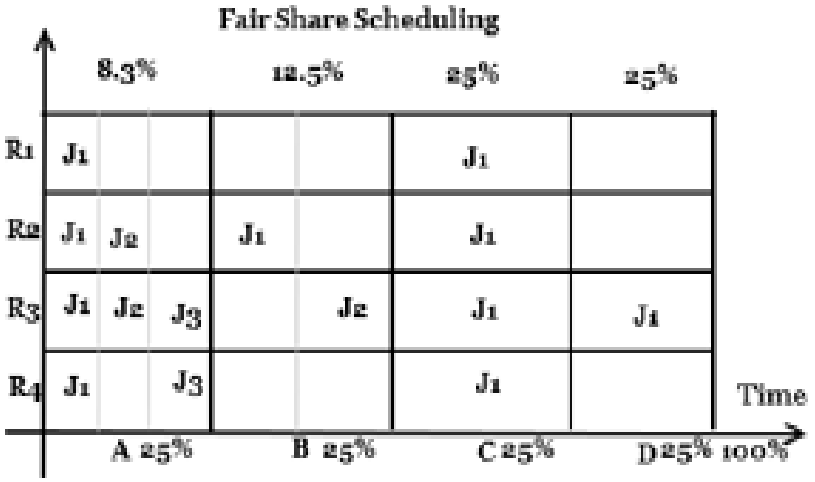


Figure7: Resource sharing for jobs in Fair scheduler

1.3.2 Capacity Scheduler

The capacity scheduler algorithm to puts the jobs into numerous queues to agreement with the situations. The Capacity scheduler provides the every queues depending on its capacity it contains the jobs. And it allocates to the certain system volume for the every queue. If all the queue are fully load and it pursues to free resources.

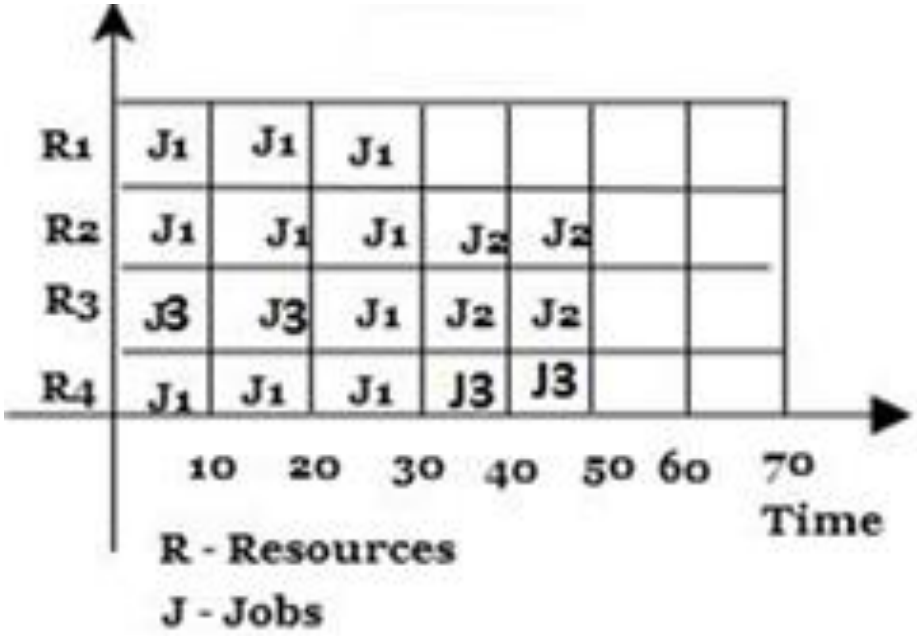


Figure8: Jobs and Resource sharing in Capacity scheduler

Then it makes to the fired resources to allocate consistently to each and every job. To compare with Hadoop defaulter scheduler FIFO scheduler and capacity scheduler it disables to the FIFO schedulers drawback such that low consumption system resources. It also to support to the various jobs into perform into the parallel systems improve the consumption of system resources through dynamic adjustment into all the resource distribution as well as the job effectiveness. The internal job queue sets are choosing the group of ability to the scheduling algorithms cannot transmit ready automatically. If the users wants to know the system info and to makes the row set & row excellent collection for the all jobs. The large scale systems are one of the big holdup refining into the whole performance to the executed system.

1.5.3 LATE Scheduler

The LATE Scheduler (Longest Approximate Time to End) is a speculative task (job) scheduler it futures behave well in Hadoop real environment. The primary perspective ness of this Scheduler is execute or complete the tasks to the greatest distance in space or time into the future because of this task make available for maximum chance to the speculative tasks to genuine surpass for original & reduce to the job's reply time.

The LATE scheduler algorithm has some benefits. It is uncertain to all the node heterogeneity and it will be re – launching only small amount of the gentlest tasks. The LATE scheduler to orders among the slow tasks to certify that the only gentlest speculatively task completed. The LATE scheduler also covers the amount of speculative tasks into the limit contention for the shared resources to avoid hiding. The Hadoop's default scheduler is fixed threshold. If all the tasks are slow to adequate an equal coincidental of being propelled. This are static threshold it can be because exceptionally number of speculative tasks to be threw.

Different ways to estimate the time left can be worked into LATE. To guess the ProgressRate of single task can be calculated as ProgressScore / T, here T is amount of time that the task has been run consecutively. Then estimate to the time of completion of jobs are (1 -ProgressScore) / ProgressRate.

Facebook to overcome the FIFO drawbacks to implement the Fair Scheduler algorithm to manage access of their Hadoop MapReduce clusters and Hadoop HDFS then it free to the Hadoop unrestricted. The fair scheduler algorithm targets to all the users to share the cluster capacity time. If the users may allocate jobs into the groups. In with each group to assigned into a certain minimum possible number of MapReduce spaces. If the free slots are in idle groups it may be to allocate the other groups. The additional capacity with in a group is shared among all the solved jobs. If the fair scheduler to maintenances the job avoidance, so if all group are not established in its Fair scheduler for positive passé of time.

Chapter 2
LITERATURE REVIEW

2.1 A Hybrid Scheduling Approach for Scalable Heterogeneous Hadoop Systems

Cloud computing capacities are three divergent features they are scalability, pay as you go, & manageability. The main advantages of Cloud computing are to led the major growth in multiplicity and scale of cloud applications. The main reason of developing the cloud applications are processing BigData. The scalability and error tolerance in Cloud computing creates a great solution for the present BigData applications. The huge storage and processing requirements of BigData applications create a challenging to provide the preferred performance level for the applications. This paper to describe a performance of available Hadoop schedulers and working of all the data analyze applications. The available scheduler FIFO is the default scheduler in Hadoop it is less than loaded of systems, and the Fair Sharing algorithm is the system is to balance the system load. The system is under loaded and the number of unrestricted slots is better than the number of coming up tasks to the scheduler changes to the FIFO algorithm. After evaluate the FIFO scheduler working the developers to identify the small job starvation so they implanted the fair sharing scheduler to overcome the small job starvation and user or node heterogeneity.

The scalability of Cloud infrastructures has suggestively improved their applicability. Hadoop, which works created on a MapReduce model, offers for effective processing of BigData. This result is being used usually by most Cloud providers. Hadoop schedulers are critical elements for as long as preferred performance levels. This paper examines the performance of broadly used Hadoop schedulers including FIFO and fair sharing and matches them with the COSHH scheduler, which has been established by the authors. In the job heterogeneity in Hadoop system is expand to the stages of heterogeneity in the 3 Hadoop factors Users, Clusters and Workload

2.2 The task Scheduling Algorithm for Hadoop Platform

MapReduce function is one of the special structure of software framework for to write the writing the data in various application which is process the massive amounts of

data in the large scale clusters. To get the better job distribution, tasks and load balancing for the MapReduce function. If the task scheduling algorithms for the Hadoop platforms to analyze the BigData in large clusters.

The MapReduce function structure is a collection of JobTracker & TaskTracker facility for the scheduling. The JobTracker is also called as job server and it is responsible to the managing all the jobs successively execute as the framework to responsible for the scheduling & managing to the TaskTrackers. The TaskTrackers assigns the jobs to Maptasks and Reducetasks to futile TaskTracker and to make given tasks to run in the parallel systems and it is answerable for the observing working feature of the running or executing tasks. TaskTracker is also called as task servers which is assigned jobs. The TaskTracker is responsible for performing the tasks to each job is divided into number of tasks with MapReduce function tasks. The TaskTrackers are basically specified to implement or processed to be assign suitable job to the correct server to perform the MapReduce function effectively. The TaskTracker implements the tasks and processed similar period reports to the all tasks are working characteristic to the JobTracker and help to JobTracker to know the general condition and it allocates the new tasks etc.

2.3 The survey on Improved Scheduling in Hadoop MapReduce in Cloud Environments.

Cloud Computing is developing aspect of new computational model for store the BigData. The Hadoop MapReduce function has developed a great reckoning model to processing the large data files on distributed computer or nodes to large clusters such as the cloud.

The Hadoop framework operations using the job processing through the MapReduce function and HDFS file distribution using the default FIFO scheduler and the jobs scheduled on first in first in order to support with other importance on size based schedulers. In this paper to describes the revision on various scheduling algorithms for effective job sharing and to progresses to the possible to use the available Hadoop schedulers and also it providing some of the rules on how to progress the scheduling in cloud atmospheres.

The default Hadoop scheduler is FIFO it executes the jobs in its general manner. If the job are divided into distinct tasks. This are overloaded to queue and, it allotted free slots and convert existing TaskTracker plugs.

There is a support for work assignment to importance to the jobs and this is not curved on to the by non - appearance. It is typically every job compulsory to use the total cluster. All the jobs are to wait for their opportunity. The shared cluster to offers a great possible to offering large resources (R) to the number of users in worldwide. The problem for to sharing resources equally between all users requires in well scheduler

2.4 The Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce

The main concepts in Hadoop that is Big Data, NameNode, DataNode, Hadoop architecture and schedulers in Hadoop. The Big Data represent the total private, public databases. The NameNode contain the info about metadata, DataNode is used in Hadoop to save the data in clusters and it is one of slave node, the DataNode is used to identify the ongoing Jobs in coming on NameNode. Schedulers acts as a major role in Hadoop systems. Big Data (Hadoop) is massive demand in the market now a days. As there huge amount of data is untruthful in the industry but there is not tool to holder it and Hadoop can implemented on low cost hardware and can be used by bulky set of spectators on bulky number of dataset. In Hadoop map reduce is the most important component in Hadoop.

The Hadoop is a measure for consists at least gigabytes of data. The Hadoop framework has existed to construct with the ability to managing large data sets can be easily couple to the gigabytes of data to petabytes of data. The Hadoop provides the solution in the form a Distributed system or nodes which are split the BigData to stores the different engines. This allows parallel computing processing of the difficult and the efficient computation is possible. The design of Hadoop framework such that it can be powerfully achieve quantity of datasets by attractive advantage of the bundled or distribute computing or by concerning hundreds of the machines in with executing power in the similar systems. In the ideally communicating to the single controlling processing systems in which much more than expensive than the huge number of systems with the separate processing systems

to making an easier speculation. The Hadoop framework offers to the cost effective way out for trying to the smaller & low-cost technologies to the self-possessed.

2.5 Improving MapReduce Performance in Heterogeneous Environments

An Adaptive Scheduling Algorithm for Dynamic Heterogeneous Hadoop Systems, proposes another planning calculation, which considers heterogeneity of occupations and assets, reasonableness and disappointment, and least impart prerequisites. By ordering employments and assets into classifications, occupations can be more brilliantly doled out to assets of the suitable classification for better effectiveness. This versatile scheduler likewise makes novel commitments, for example, diminishing correspondence expenses by not so much boosting an errand's conveyance, decreasing scan overhead for coordinating occupations and assets, and expanding region.

2.6 The Research on Internet Hot Topic Detection Based on MapReduce Architecture

The main perform of MapReduce is 2 individual tasks one is clerk perform and another is Reducer perform. The Map or mapping is employed to decompose a task into the numerous tasks and also the scale back or alter methodology is employed to encapsulate to the managed results of the numerous tasks. Once the software engineer is desires solely to implement clerk interface & Reducer interface. Within the terabytes of knowledge may be calculated. Once making the distributed comparable series supported the MapReduce programming model. The software engineer is simply in command of the writing Map () & scale back (). If the extra styles of tough issues in parallel programming, like distributed storage, job programing, load equalization, fault tolerance, network communication so on, are all prohibited by MapReduce framework. It consists of the word separation, elimination of stop Word, feature extraction and text illustration.

Abilities utilized the MapReduce perform implementation from the Apache Hadoop because the basis of Hadoop structure and also the worked North American nation well as for the many decades. Within the early 2011 the developer to begin obtaining the sides of system. Within the developing issues to examine the problems within the programing algorithms define, that consists of the Job Tracker & several

TaskTrackers. The TaskTrackers are accountable for the running tasks that the JobTracker to allocates MapReduce.

The JobTracker has two main tasks they're one is managing to the cluster resources and another one is programing to the all user jobs. The cluster size and also the obtaining range of jobs at any social networks grew and also the measurability margins of style suited vivacious. The JobTracker couldn't be handle the twin tasks sufficiently. The cluster consumption would descent sharply as a result of the programing within the cloud environments.

2.7 A service integrity assurance framework for cloud computing based on MapReduce

MapReduce is a large data processing in distributed systems, the data sets are more than the one terabytes. The MapReduce process is using in the cloud computing environment. The current maintenance of big data in cloud computing environment provides the Hadoop MapReduce. The Hadoop is one of the open source framework it provides the MapReduce function in cloud environment.

In the distributed environment large data processing is easy to using the MapReduce function, user can store the large data in any cloud clusters without any interactions. The large data storage systems facing security problems, data confidentiality and data integrity are main concerns. According through the present system operations using frameworks to provide the some security mechanisms. But we need to improve the more improvements to the customers. The execution of user jobs or data into the clusters depends on the system capacity and processor performance. So the improvement of the jobs or task execution is more important to the current problems. The improvement of this tasks. First we need to analyze the current system performance and working and after we check the scheduler performance. The scheduler performance also causes the system execution time.

The master manager to manage the all worker nodes and jobs execution. The cloud users to create the sub- domain in the large cluster is one of the best concern to secure data processing and data confidentiality.

2.8 Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing.

Most of the present cloud systems to process the large amount of data, to provide the user acceptance. To execute or processing this large data using the specially designed frameworks or software's. Hadoop is designed for processing the big data and it provides two main functions. One function is providing the large data storage capacity and another one is to reduce the size of file without changing the any data in that file. The two functions are called as HDFS and MapReduce functions.

The drawback in the Hadoop re lack of performance in the heterogeneous data nodes. The heterogeneous data nodes are cane be from different sizes of data sets. This paper mainly concentrate on performing the large data in the heterogeneous environment.

2.9 Related Work

Facebook to overcome the FIFO drawbacks to implement the Fair Scheduler algorithm to manage access of their Hadoop MapReduce clusters and Hadoop HDFS then it released to the Hadoop unrestricted. The Fair scheduler algorithm to targets the give all users to share the cluster capacity time. If the users may allocate jobs into the groups. In with each group to assigned into a certain minimum possible number of MapReduce spaces. If the free slots are in idle groups it may be to allocate the other groups. The additional capacity with in a group is shared among all the getting jobs. The Fair Scheduler to supports prevention, so if all group are not established in its Fair scheduler for positive passé of time.

Another limitation in the Hadoop MapReduce function framework was its pull based scheduling model. The TaskTrackers to provide a heartbeat prominence to the JobTracker in order to become number of tasks to track. The heartbeat is a periodic and there is always predefined interval when the scheduling tasks are performing any job. For all the small jobs are delays on problematical.

The Hadoop MapReduce function is also self-conscious by its static slot based on the resource managing model. Than using a real resource managing system into a

MapReduce function cluster is divided into the permanent number of Map & reduce slots based on the static configuration. So that slots are lost any time in the cluster workload. It does not fit for the static structure. The slot based model makes to the hard for of a Non-MapReduce function applications to be scheduled applicably.

Finally, in the original JobTracker design the required tough interruption and all the running jobs are destroyed during the software improvement. In which destined that the every software improvement caused in important missed intention. So need to improve better scheduling algorithms for Hadoop framework that would be improve.

Little Jobs Starvation: – the little job starvation in job sharing process will occur using on the FIFO scheduler. The basic working of FIFO scheduler is first in first out, we get the initial requests huge size comparing the upcoming job. If upcoming jobs are small size data this are waiting for the resource until initial job compilation.

Resources Mismatch:– the resource mismatch mainly occur using on the fair scheduler. The fair scheduler giving the resource's depending on the number of jobs getting in particular interval of time. The jobs are some times getting wrong resources, to storing the wrong resource in the cluster.

Job Waiting Time: – In LATE scheduler the processing starts depending on the first job execution time of other schedulers like FIFO, etc. So the time taken by starting the process is late. It is cause to increase the total execution time. So the users to prevented this scheduler in Hadoop schedulers.

2.10 Hadoop MapReduce function scheduling framework limitations

Initially employed the MapReduce function implementation from the Apache Hadoop as the basis of Hadoop structure and the worked us well as for the several decades. In the early 2011 the developer to start getting the edges of system. In the developing concerns to see the issues in the scheduling algorithms outline, which consists of the JobTracker & lots of TaskTrackers. The TaskTrackers are responsible for the running tasks that the JobTracker to allocates MapReduce.

The JobTracker has 2 main tasks they are one is managing to the cluster resources and another one is scheduling to the all user jobs. The cluster size and the getting number

of jobs at any social networks grew and the scalability margins of design suited vibrant. The JobTracker could not be handle the dual tasks sufficiently. The cluster consumption would descent precipitously due to the scheduling in the cloud environments.

Another limitation in the Hadoop MapReduce function framework was its pull based scheduling model. The TaskTrackers to provide a heartbeat prominence to the JobTracker in order to become number of tasks to track. The heartbeat is a periodic and there is always predefined interval when the scheduling tasks are performing any job. For all the small jobs are delays on problematical.

The Hadoop MapReduce function is also self-conscious by its static slot based on the resource managing model. Than using a real resource managing system into a MapReduce function cluster is divided into the permanent number of Map & reduce slots based on the static configuration. So that slots are lost any time in the cluster workload. It does not fit for the static structure. The slot based model makes to the hard for of a Non-MapReduce function applications to be scheduled applicably.

Finally, in the original JobTracker design the required tough interruption and all the running jobs are destroyed during the software improvement. In which destined that the every software improvement caused in important missed intention. So need to improve better scheduling algorithms for Hadoop framework that would be improve.

Chapter 3
PRESENT WORK

3.1 Scope of the study

Hadoop supports pluggable schedulers, present available schedulers are not fit sometimes depending on the job size, cluster work load and many other constrains. The scheduling algorithms to given the input (jobs) to MapReduce function. The performance of MapReduce function depending on the scheduler's job execution time, we already know the Hadoop supports pluggable schedulers. We discussed about various scheduling algorithms like FIFO, Fair Sharing, Capacity and LATE Scheduling algorithm's and the working of this algorithms. The main drawback of FIFO is small job starvation and it overcomes the fair scheduling and one of the current used scheduler is Capacity scheduler it shares the clusters depending on the arrival jobs, when the jobs are huge but the size of the job is less this case Capacity scheduler is suitable for the job queueing. The number of jobs are huge and the job sizes are high in this case the capacity scheduler is not suitable.

The Hadoop allows multiple schedulers at a time, so we need to try some combinations of available schedulers and improve the jobs execution performance and it automatically the MapReduce function also improve.

Initially employed the MapReduce function implementation from the Apache Hadoop as the basis of Hadoop structure and the worked us well as for the several decades. In the early 2011 the developer to start getting the edges of system. In the developing concerns to see the issues in the scheduling algorithms outline, which consists of the JobTracker & lots of TaskTrackers. The TaskTrackers are responsible for the running tasks that the JobTracker to allocates MapReduce.

The JobTracker has 2 main tasks they are one is managing to the cluster resources and another one is scheduling to the all user jobs. The cluster size and the getting number of jobs at any social networks grew and the scalability margins of design suited vibrant. The JobTracker could not be handle the dual tasks sufficiently. The cluster consumption would descent precipitously due to the scheduling in the cloud environments.

Another limitation in the Hadoop MapReduce function framework was its pull based scheduling model. The TaskTrackers to provide a heartbeat prominence to the JobTracker in order to become number of tasks to track. The heartbeat is a periodic and there is always predefined interval when the scheduling tasks are performing any job. For all the small jobs are delays on problematical.

The Hadoop MapReduce function is also self-conscious by its static slot based on the resource managing model. Than using a real resource managing system into a MapReduce function cluster is divided into the permanent number of Map & reduce slots based on the static configuration. So that slots are lost any time in the cluster workload. It does not fit for the static structure. The slot based model makes to the hard for of a Non-MapReduce function applications to be scheduled applicably.

Finally, in the original JobTracker design the required tough interruption and all the running jobs are destroyed during the software improvement. In which destined that the every software improvement caused in important missed intention. So need to improve better scheduling algorithms for Hadoop framework that would be improve.

3.2 Objective of the study

We implement the constraint base scheduling algorithm for effective job sharing for Hadoop framework to using both HDFS and MapReduce functions. We are using present available schedulers FIFO & LATE schedulers. The FIFO is the default scheduler in Hadoop framework and LATE scheduler is the one of oldest scheduler in Hadoop. To implement the new job sharing model for the effective job sharing.

The proposed method is to decrees user or job waiting time and it may be reduce the job waiting time and mange's the small job starvation and it improves the MapReduce performance in the huge job and number of users.

To improving the better Hadoop MapReduce function in different manners these are

- The Lower latency for small jobs.
- If the ability to upgrade without the interruption.
- Better scalability and cluster utilization

- Scheduling is based on the actual task resource requests slightly than a count of map and reduce tasks

3.3 Methodology

The Hadoop framework is a pluggable scheduler so we are took any available schedulers and comparing the other scheduler's performance and all are the working of the schedulers. We are taking the FIFO because of it is a default scheduler in the Hadoop and another one is LATE scheduler this is one of the oldest scheduler using on job sharing. We implement the new job formation using this two scheduling algorithms.

The LATE scheduler to collect the jobs depending on the first job execution time. If the 1st job size is less than it works effectively but the job size is high, the execution time toke much time so it is not collect the jobs without considering the 1st job executing time here user waiting or job waiting occur so we use the both the FIFO and LATE schedulers and implement the on algorithm for job sharing on this two scheduling algorithms.

We need to calculate the job execution time, the job execution time depends on the using system processing speed and network bit rate & bot rats. The exception of job also depends on the cluster working and number of user at the time.

To estimate the process rate of each job is (ProgressScore / T). Here T is the amount of time for executing the jobs and ProgressScore is number of jobs work flow.

To find the execution time (T) is {ProgressScore/ Process rate} this are useful to share the jobs in implemented algorithms

3.3.1 .NET Frame Work:

- A frame of software which was developed by Microsoft that works on the platform that is developed by Microsoft. It consists of various types of classes and libraries of various functionalities that are available in other type of programming languages.
- The main function of this frame work is to get the access of the various functionalities of other ones and so that it is easy to use the functions that are used for the development of any application or the other console application.
- Two types of application are developed through this frame work they are Windows Application and Console or Web Application.

- There is a super class that contains various classes called as framework class library. FCL (Framework Class library) consists network procedure communication, algorithms of numeric, encryption of the code, data accessing activities like data abstraction, connectivity to the database and various functionalities.
- Programs can be written very easily in .NET Framework as it consists of all types of classes and functions which are in built in the FCL (Framework Class Library) so that it save more time and also in economical way.
- .NET framework architecture consists of the .NET core, Assemblies, class library and common language infrastructure.
- Common language infrastructure (CLI) acts as a platform to obtain the needful functionalities for the execution and development of the application in a simple way that the other frameworks that are available.
- Class Library consists of all types of classes that are collected through the FCL classes for file handling, exception handling and many others. Standard libraries of classes are available so that they are ordered name space of the hierarchies.
- Base class library is a part of the class library and it also acts as a super class it contains all the information about the classes that are required for developing the application. Various functions such as LNQ, WCF, WF, ASP.NET and ADO.NET are include in the base class library.
- Framework Class Library is a small part of the class library and it contains .dll files which are available only in the Framework class library that are only for the development of the application or the windows form
- .NET core is one of the implementation prototype of .NET Frame work which is incomplete that means cannot function the total implementation of the .NET Framework.
- Assemblies are used for the storage of the code that are present in the CLI. These are called simply CLI assemblies. These CLI assemblies are used for the .dll and .exe files that are obtained by the developing the application using the FCL and CLI of the .NET Framework.

- The development tool that is used by the .NET Frame work is Microsoft Visual Studio, operating system is windows, server is oracle access and many other functions are used by the .NET Framework.

3.3.2. Microsoft Visual Studio:

- There are many integrated development environments are available in the software development environment. In of the visual studio is one of the integrated development environments.
- The main functioning of the visual studio is to develop various types of the applications that may be windows or web console applications.
- Windows console applications are the applications that are used for the internal purpose of a single distributed system such as a network server connected to many systems.
- Web console application are the development of applications for a particular organization or any other that may acts as a customer for their own development of the application.

  The main features of Visual Studio:

  - Code Editor
  - Debugger
  - Designer
  - Other Tools

- Code Editor is the important feature that is used for writing the code that which it has in built syntax and easy to access them at any time by just typing the name of the class.
- Debugger is available in visual studio that acts as both the machine and source level debugger. It also simply manages with the code and the core code which is written by own for the program of the application of any language the program may be written.
- Designer is the one use of designs for the helpful in developing the applications. It contains various types of designers that are used for the development of the

application. Designs can be of various types and the designs are selected according to the developer's issue. Types of designers 1.

- Windows Form designer
- WPF Designer
- Web Designer or evelopment
- Class Designer
- Data Designer
- Mapping Designer.

- Other tools include open tab browser, properties editor, object browser, solution explorer, Team explorer, Data explorer, Server explorer, Dotfuscator software services community edition, Text generation framework, ASP.NET website administration tool, Visual studio tools for office.

3.3.4 Algorithm for Effective job sharing

Input: Getting Jobs (ji);

Function: SelectScheduler (Jobs)

Variable: Jobtime[], tavg;

Jobtime [] = GettingJobs (ji); i=0 to N

If (job (ji) = Homogenous || Heterogeneous) then

tavg = jobtime [ji]/Total Jobs;

While 1 to N

if (jobtime[ji] > tavg) then /* Jobs scheduled to FIFO Scheduler*/

Else

/* Jobs scheduled to LATE Scheduler*/

END If

END While

END Function

3.3.5 Flow chart for effective job sharing:

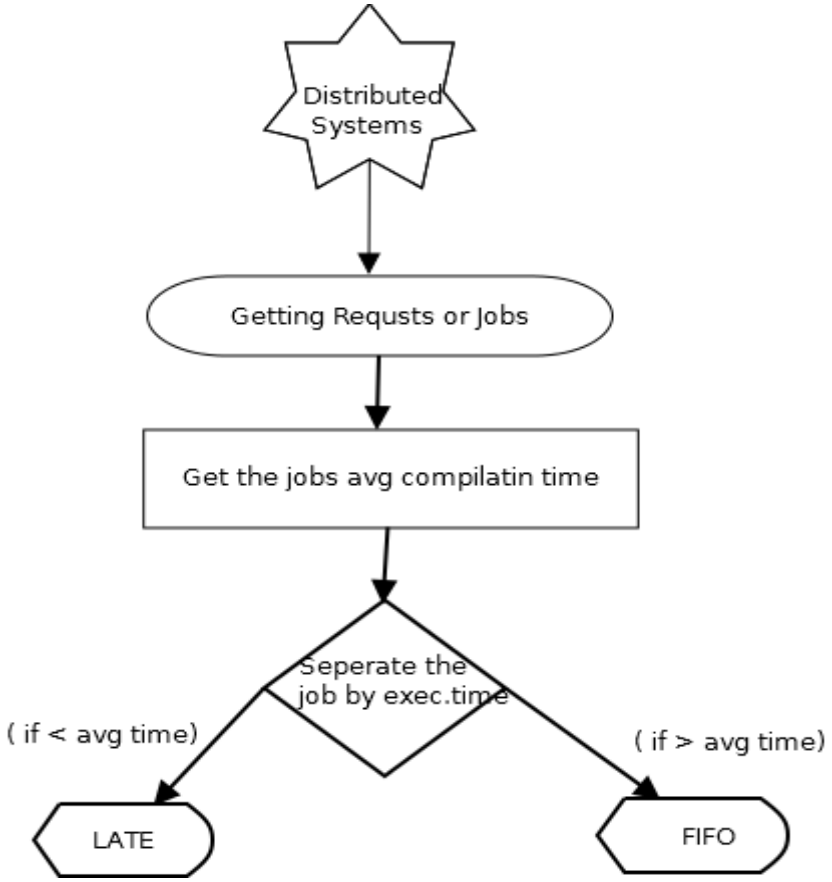


Figure9: Flowchart for effective job sharing

In generally social networks get the jobs their user in various locations, we can collect these jobs from any social networks or any data base maintainers it very difficult so it is not possible to gat collect the jobs.

- We need to collect this jobs through any third party cloud provider and after collect the jobs to implement proposed algorithm and execute jobs with available schedulers.
- The map reduce performance is depends on the job execution time so we implement the effective job sharing to FIFO and LATE schedulers.
- After getting requests we are analyze the getting jobs user ids and analyze of the getting data in the data base using on the raw data through Hadoop pig tool.

- The working of the proposed algorithm are first collecting the jobs through any users and comparing the job sizes to previous job execution times and to share the jobs the job size is greater than the average value is giving to FIFO scheduler and the job size is less than the average value is giving to LATE scheduler.

To know the job progress rate is ProgressScore/ Time taking of total jobs execution time. And to know the total execution Time (T) is ProgressScore / Progress rate.

Step wise process of Job sharing

1. Start
2. Getting the jobs (ji) considering the resources.
3. Calculate the average time considering the previous reports
4. If the jobs time is greater or equal to the average value prepare the FIFO scheduler.
5. Else the job execution time is the less than the average time prepare for LETE scheduler.
6. Consider step 4 and 5 when complete the all jobs.
7. End

Chapter 4

RESULTS AND DISCUSSIONS

We are not develop any scheduler for improving job execution and MapReduce performance, we just comparing and analyze the drawbacks to the existing schedulers and to share the jobs to the suitable scheduler. First we get the jobs, basically the social networks to get the jobs their users in various locations, it is not possible to get this requests to us. So we get the job requests to any 3rd party vender, zendesk to provide the cloud services.

4.1 About Zendesk

Zendesk is a (cloud provider) software improvement organization based in California. It offers cloud-based services. The Zendesk include in self-service options, ticketing and customer care services.

How it is useful to users. First we create the interface on zendesk to create our account and get the tokens or jobs through any users.

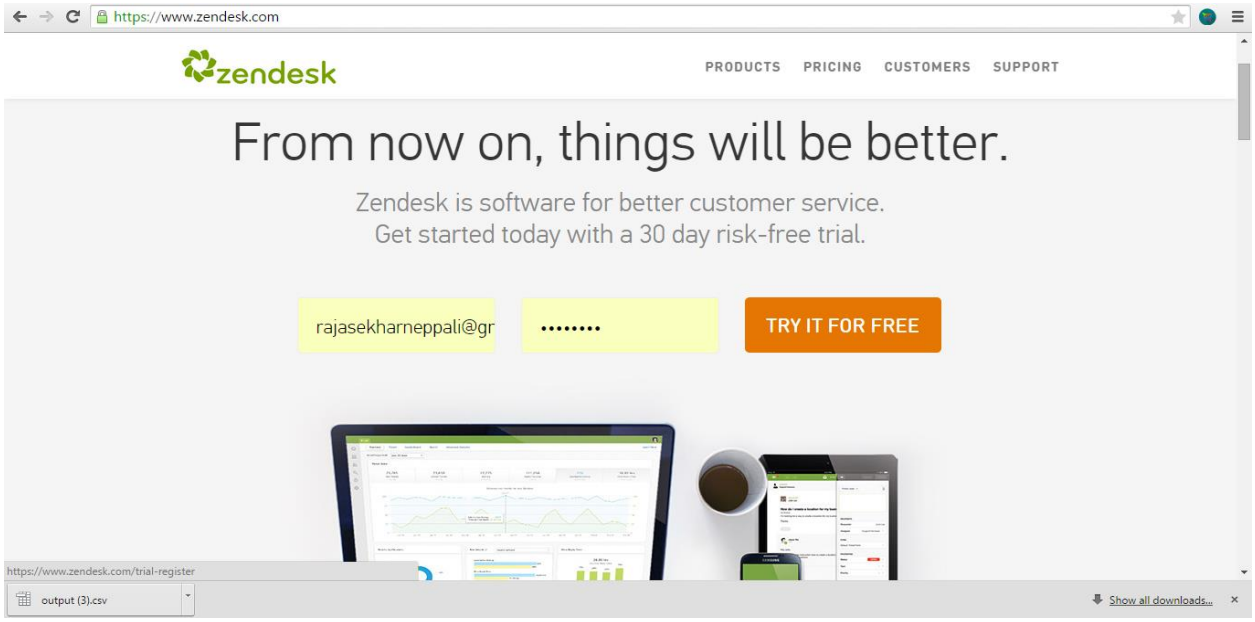


Figure10: Registration page on Zendesk

The Zendesk registration process is follows

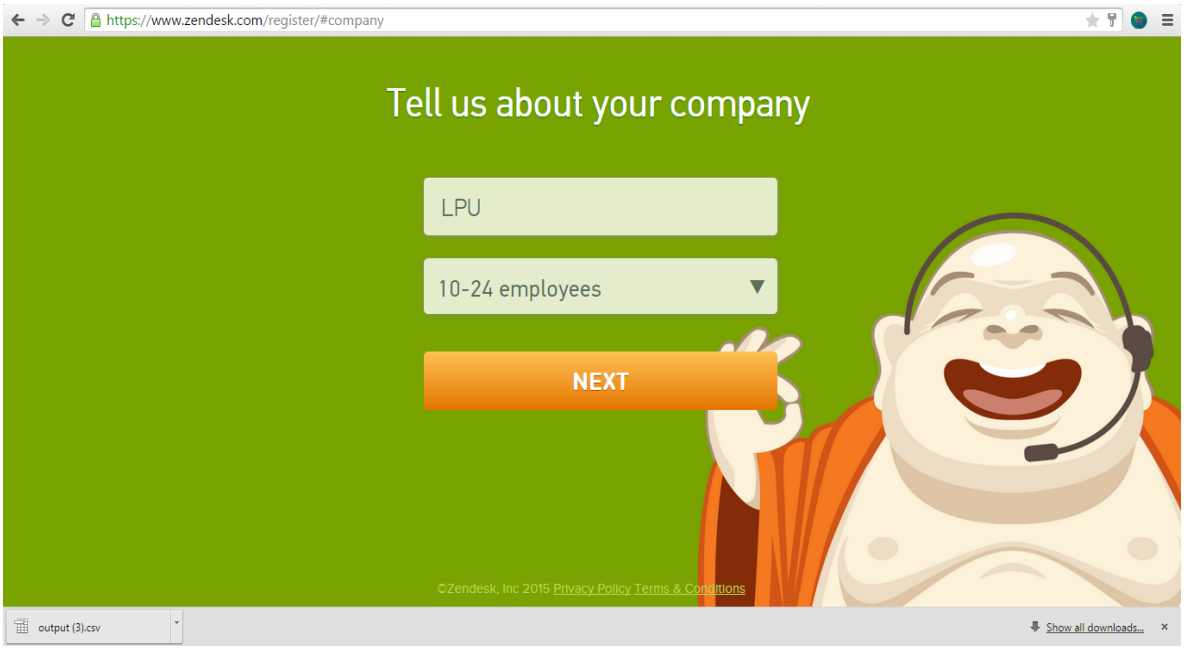


Figure11: The registration n process on zendesk steep 2

Now we are to create the domain on zendesk registration process. We create our domain name lpusupport.zendesk.com

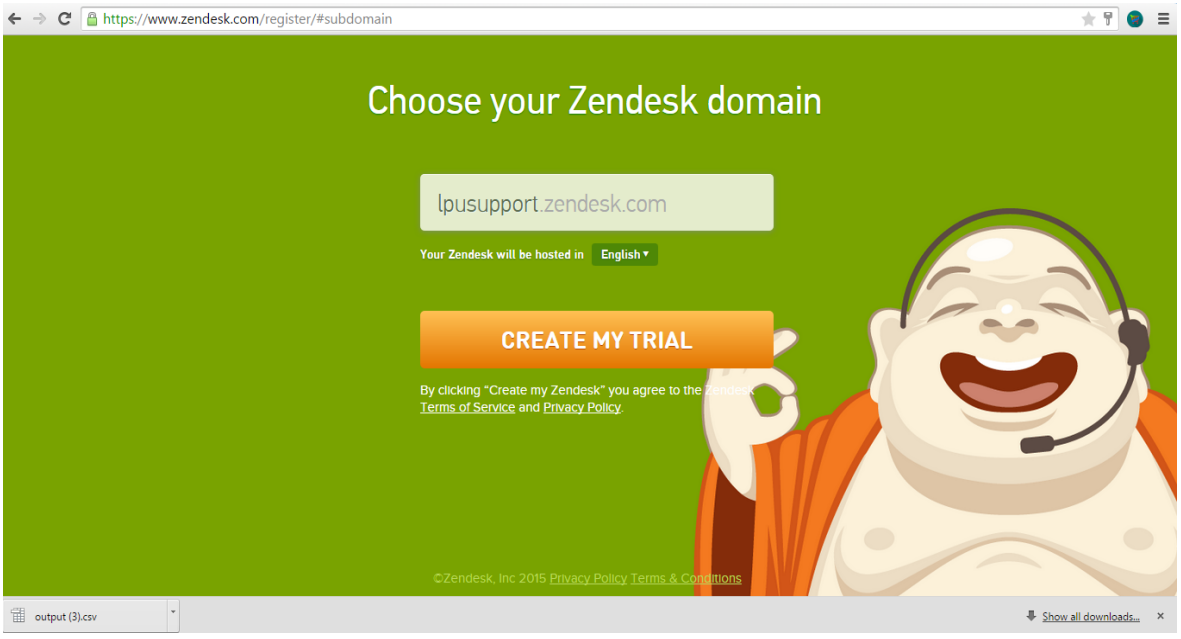


Figure12: Domain creation in zendesk

After complete the registration process now we are entering in to home page on zendesk.

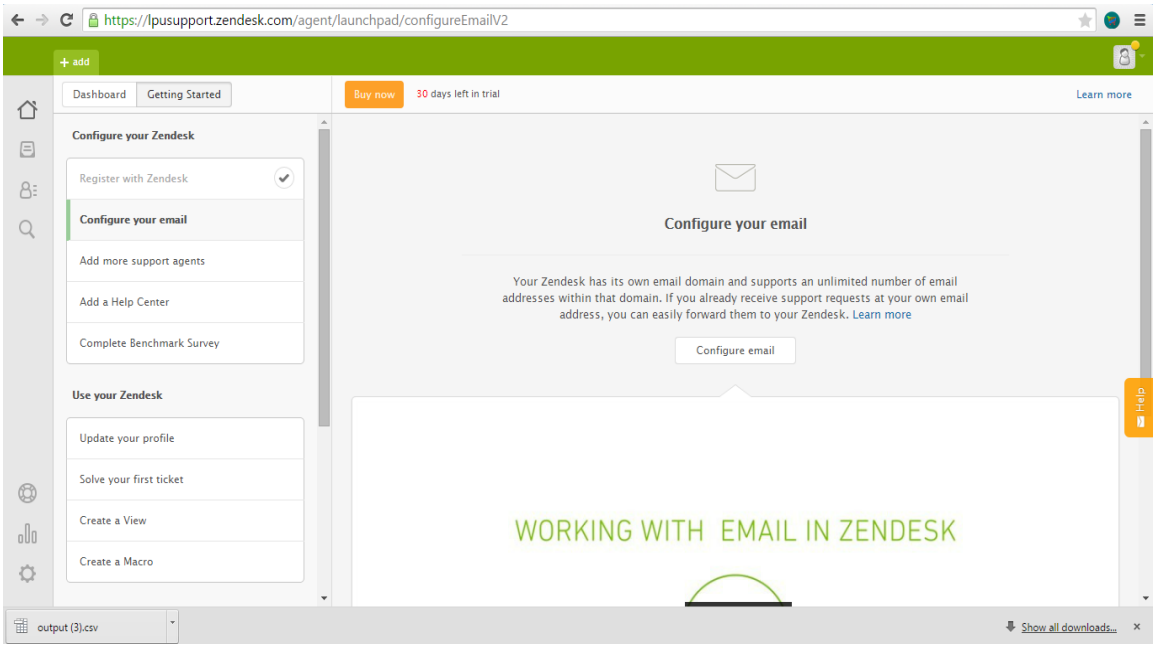


Figure13: Home page on zendesk users inter face

The home page to shows the Profile information and getting tokens and etc..,, The View function to shows the getting jobs or tokens list and also present the solved & unsolved tokens also.

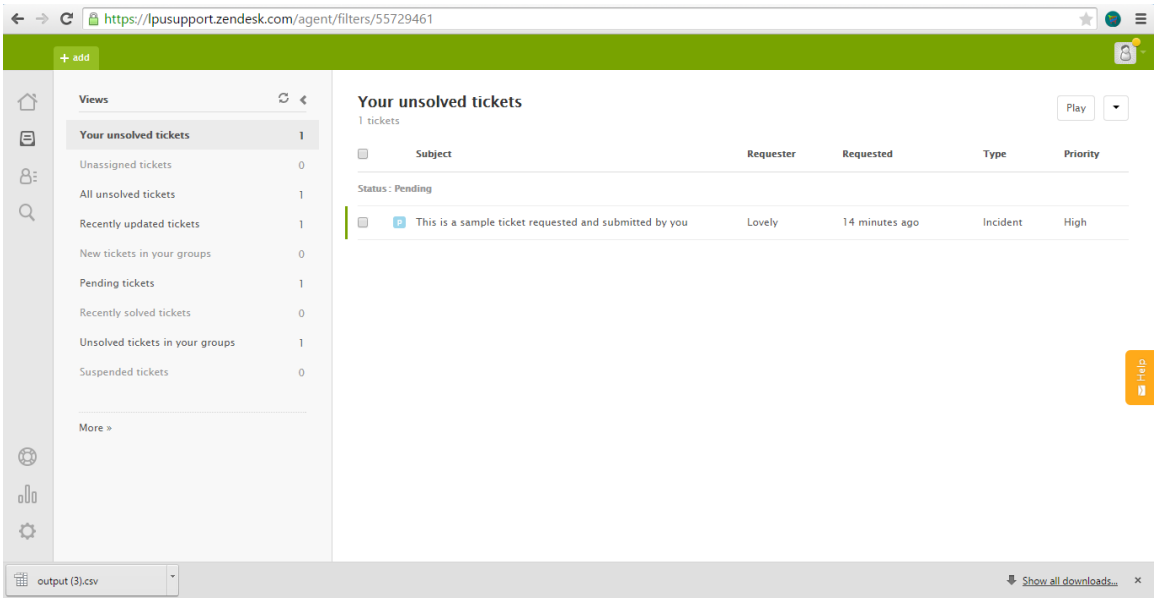


Figure14: Views page on our zendesk

4.2 Integration to Zendesk and asp.net

After getting the tokens (or) jobs in zendesk now we are to intergrade or get this jobs using Zendesk cloud service to our programming tool. We intergraded the zendesk interface into Microsoft visual studio (asp.net).

After integrand the both inter faces we collect the tickets/ jobs shown on web browser.

We create our web interface using on the asp.net, the web page login screen as……

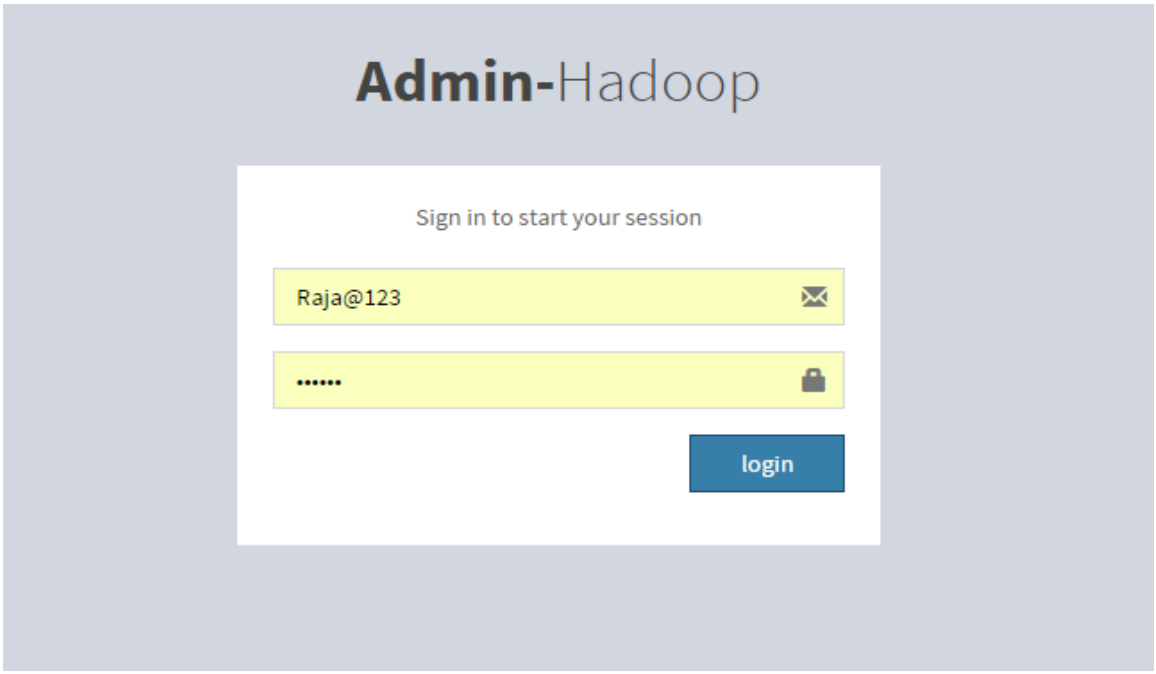


Figure 15: Login page on creating web page

Once we log in to the created page it shows getting jobs on various clients. This page design to various phases below screens sowing the all modules in the page.

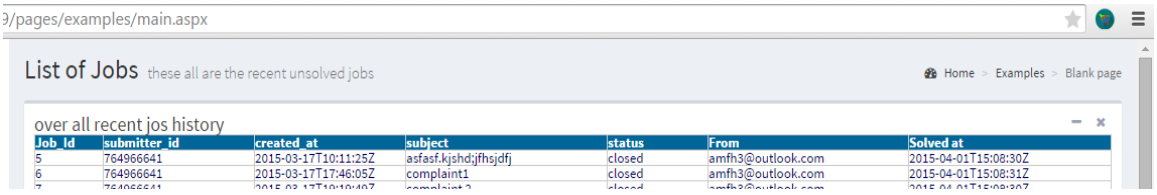
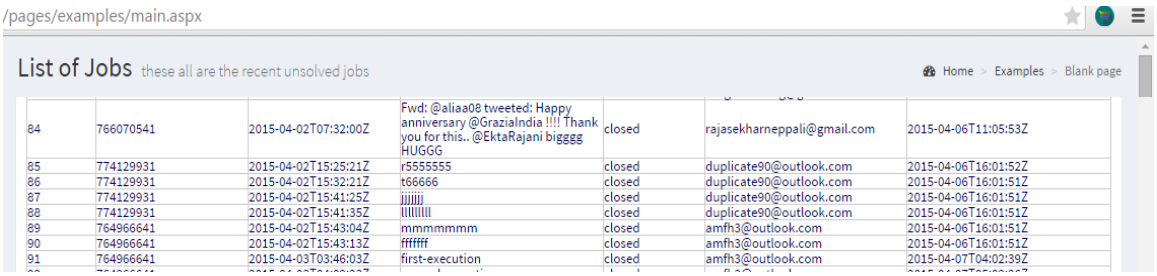


Figure16: List of all collected jobs

In this application to shows the overall jobs in recent history in that list covers the Job-Id, Submitter Id, Created-at, subject, status, from and solved_at.

Here

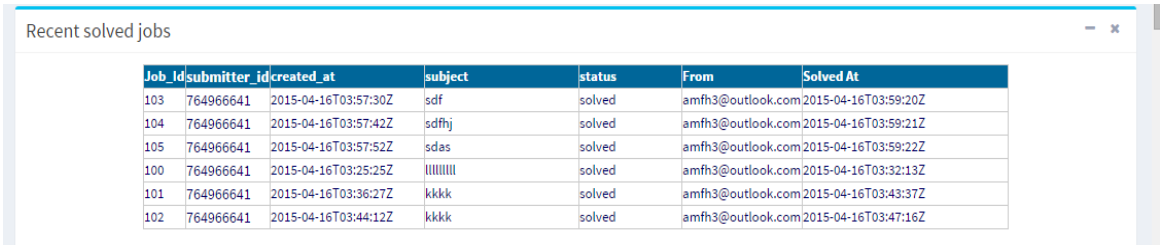
- **Job-Id** – unique id assigned by the Zendesk server for each job, this is primary key for accessing the corresponding job.
- **Submitter-Id** – this also unique id for each users, it is used to identify the users.
- **Created-at**- this is the time of job was created or job is triggered by server.
- **Subject** – that describe the subject of the jobs.
- **From** – it describe the users detail like Email and user's name.
- **Solved at** – this the time when that assigned job is solved.



84	766070541	2015-04-02T07:32:00Z	Fwd: @aliaa08 tweeted: Happy anniversary @GraziaIndia !!!! Thank you for this.. @EktaRajani bigggg HUGGG	closed	rajasekharneppali@gmail.com	2015-04-06T11:05:53Z
85	774129931	2015-04-02T15:25:21Z	r5555555	closed	duplicate90@outlook.com	2015-04-06T16:01:52Z
86	774129931	2015-04-02T15:32:21Z	t66666	closed	duplicate90@outlook.com	2015-04-06T16:01:51Z
87	774129931	2015-04-02T15:41:25Z	jjjjjjj	closed	duplicate90@outlook.com	2015-04-06T16:01:51Z
88	774129931	2015-04-02T15:41:35Z	lllllllll	closed	duplicate90@outlook.com	2015-04-06T16:01:51Z
89	764966641	2015-04-02T15:43:04Z	mmmmmmm	closed	amfh3@outlook.com	2015-04-06T16:01:51Z
90	764966641	2015-04-02T15:43:13Z	fffffff	closed	amfh3@outlook.com	2015-04-06T16:01:51Z
91	764966641	2015-04-03T03:46:03Z	first-execution	closed	amfh3@outlook.com	2015-04-07T04:02:39Z

Figure 17: List of present jobs/tickets

Above picture to shows the jobs from various users through various places. This are collecting from Zendesk.



Job_Id	submitter_id	created_at	subject	status	From	Solved At
103	764966641	2015-04-16T03:57:30Z	sdf	solved	amfh3@outlook.com	2015-04-16T03:59:20Z
104	764966641	2015-04-16T03:57:42Z	sdfhj	solved	amfh3@outlook.com	2015-04-16T03:59:21Z
105	764966641	2015-04-16T03:57:52Z	sdas	solved	amfh3@outlook.com	2015-04-16T03:59:22Z
100	764966641	2015-04-16T03:25:25Z	lllllllll	solved	amfh3@outlook.com	2015-04-16T03:32:13Z
101	764966641	2015-04-16T03:36:27Z	kkkk	solved	amfh3@outlook.com	2015-04-16T03:43:37Z
102	764966641	2015-04-16T03:44:12Z	kkkk	solved	amfh3@outlook.com	2015-04-16T03:47:16Z

Figure18: List of solved jobs

We are solve this jobs using on the current available schedulers and our proposed job sharing method.

We are to execute this jobs using on FIFO, Fair share scheduler and propose job sharing model.

And now to compare the results of all the jobs executing through the schedulers and comparing that results to proposed job sharing method. The proposed job sharing method to share the jobs depending on the suitable scheduler FIFO or LATE scheduler.

We create to analyze the time of job execution to graph representation. The graph to represent the number of jobs in X – axis and job execution time on Y –axis.

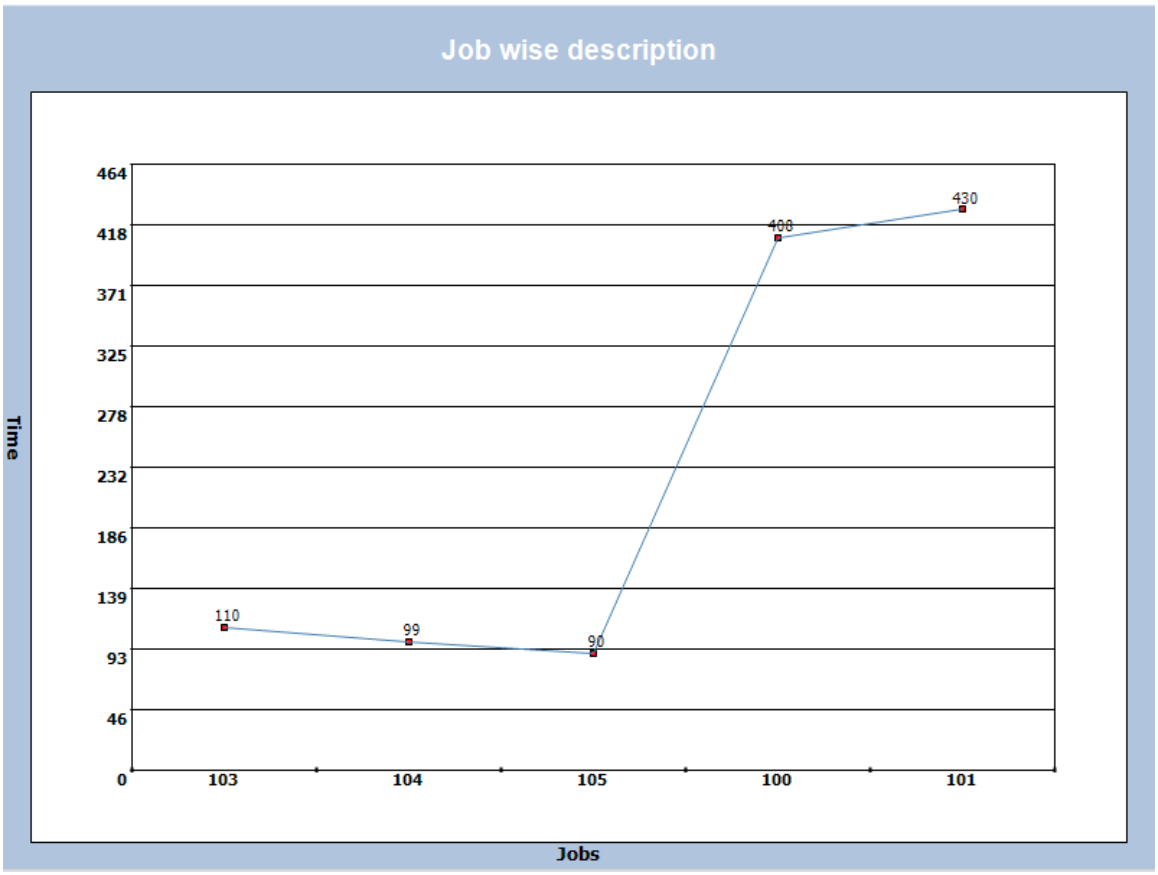
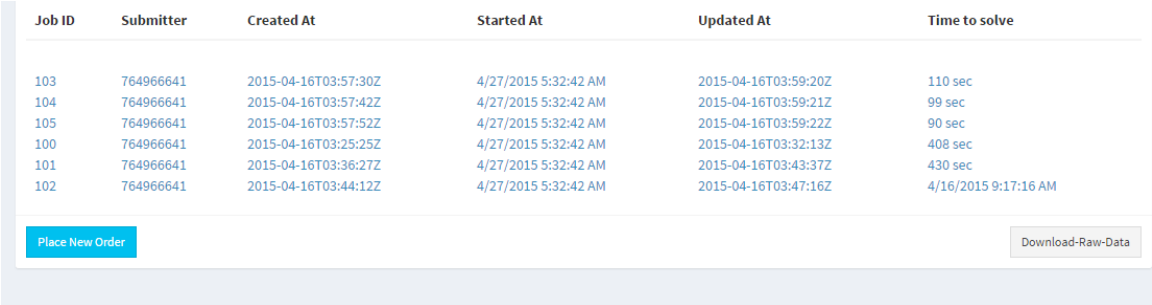


Figure19: Jobs execution time comparisons

The above graph to shows the job execution time in coming from various senders.

Now we are check to solved jobs in the list. The list to shows every job starting time and completion times on various cluster and users.

Now we check the various or different job generation manually from heterogeneous category, here we mainly focus on priority of the job whether job has the priority then we have to give higher preference to those jobs.



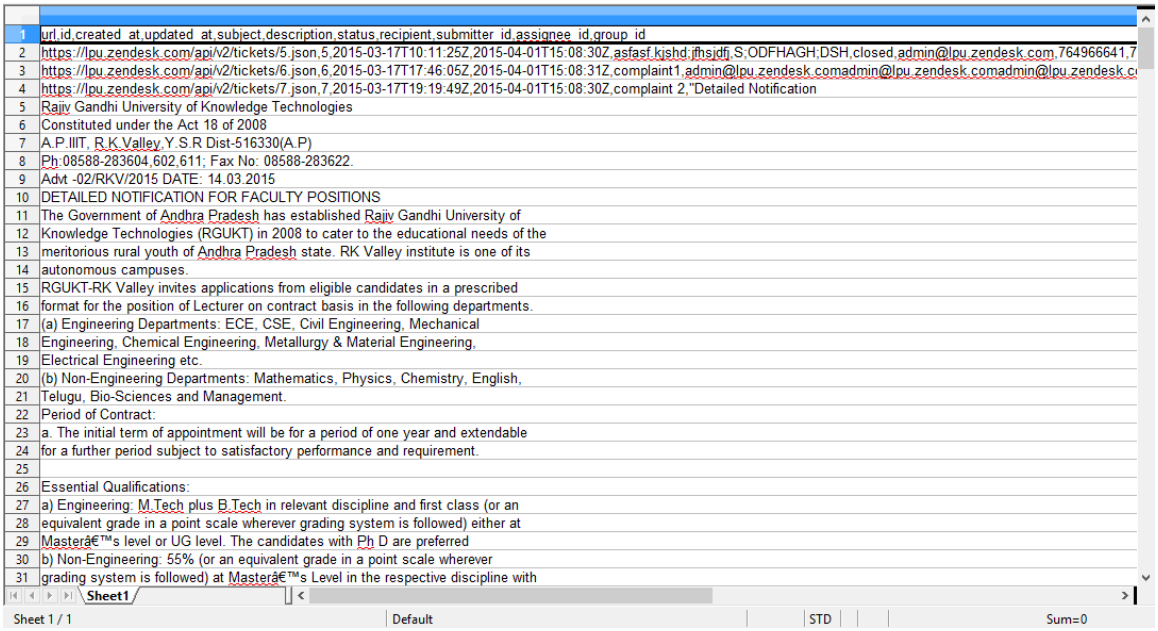
Job ID	Submitter	Created At	Started At	Updated At	Time to solve
103	764966641	2015-04-16T03:57:30Z	4/27/2015 5:32:42 AM	2015-04-16T03:59:20Z	110 sec
104	764966641	2015-04-16T03:57:42Z	4/27/2015 5:32:42 AM	2015-04-16T03:59:21Z	99 sec
105	764966641	2015-04-16T03:57:52Z	4/27/2015 5:32:42 AM	2015-04-16T03:59:22Z	90 sec
100	764966641	2015-04-16T03:25:25Z	4/27/2015 5:32:42 AM	2015-04-16T03:32:13Z	408 sec
101	764966641	2015-04-16T03:36:27Z	4/27/2015 5:32:42 AM	2015-04-16T03:43:37Z	430 sec
102	764966641	2015-04-16T03:44:12Z	4/27/2015 5:32:42 AM	2015-04-16T03:47:16Z	4/16/2015 9:17:16 AM

Place New Order

Download-Raw-Data

Figure20: Jobs starting time and compile times

After solving the all jobs we are analyse to the getting jobs data ne the cluster so we need to RawData. We can down lode the all executed jobs RawData in our web page download button.



1	url,id,created_at,updated_at,subject,description,status,recipient,submitter_id,assignee_id,group_id
2	https://lpu.zendesk.com/api/v2/tickets/5.json,5,2015-03-17T10:11:25Z,2015-04-01T15:08:30Z,asfasf,kjshd;jfhsjdfj,S:ODFHAGH;DSH,closed,admin@lpu.zendesk.com,764966641,7
3	https://lpu.zendesk.com/api/v2/tickets/6.json,6,2015-03-17T17:46:05Z,2015-04-01T15:08:31Z,complaint1,admin@lpu.zendesk.comadmin@lpu.zendesk.comadmin@lpu.zendesk.c
4	https://lpu.zendesk.com/api/v2/tickets/7.json,7,2015-03-17T19:19:49Z,2015-04-01T15:08:30Z,complaint 2,"Detailed Notification
5	Rajiv Gandhi University of Knowledge Technologies
6	Constituted under the Act 18 of 2008
7	A.P.IIIT, R.K.Valley,Y.S.R Dist-516330(A.P)
8	Ph:08588-283604,602,611; Fax No: 08588-283622.
9	Advt -02/RKV/2015 DATE: 14.03.2015
10	DETAILED NOTIFICATION FOR FACULTY POSITIONS
11	The Government of Andhra Pradesh has established Rajiv Gandhi University of
12	Knowledge Technologies (RGUKT) in 2008 to cater to the educational needs of the
13	meritorious rural youth of Andhra Pradesh state. RK Valley institute is one of its
14	autonomous campuses.
15	RGUKT-RK Valley invites applications from eligible candidates in a prescribed
16	format for the position of Lecturer on contract basis in the following departments.
17	(a) Engineering Departments: ECE, CSE, Civil Engineering, Mechanical
18	Engineering, Chemical Engineering, Metallurgy & Material Engineering,
19	Electrical Engineering etc.
20	(b) Non-Engineering Departments: Mathematics, Physics, Chemistry, English,
21	Telugu, Bio-Sciences and Management.
22	Period of Contract:
23	a. The initial term of appointment will be for a period of one year and extendable
24	for a further period subject to satisfactory performance and requirement.
25	
26	Essential Qualifications:
27	a) Engineering: M.Tech plus B.Tech in relevant discipline and first class (or an
28	equivalent grade in a point scale wherever grading system is followed) either at
29	Masterâ€™s level or UG level. The candidates with Ph D are preferred
30	b) Non-Engineering: 55% (or an equivalent grade in a point scale wherever
31	grading system is followed) at Masterâ€™s Level in the respective discipline with

Sheet1

Sheet 1 / 1 | Default | STD | Sum=0

Figure21: Collected RawData

We are analyze the getting RawData to Hadoop BigData management studio (Syncfusion).

This RawData collected to the solved jobs & it is download on zendesk clusters.

4.3 Big Data Management Studio

The Syncfusion Big Data Management studio provides an easy to use of visual interface for working with Apache Hadoop and other Hadoop tools in its ecosystem.

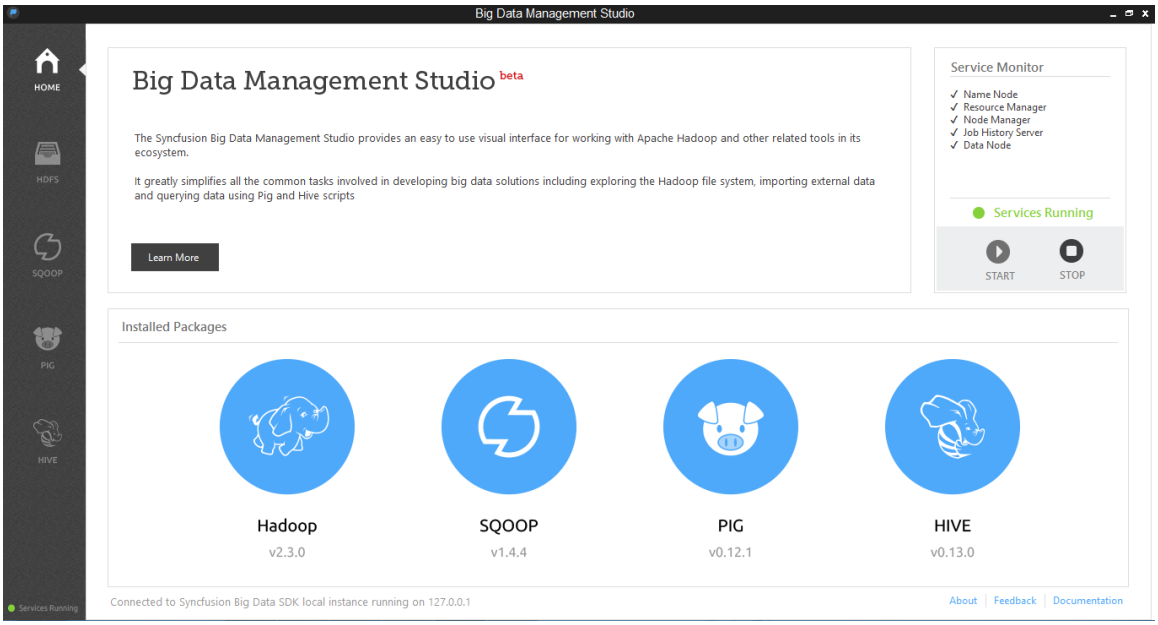


Figure22: Big Data management studio

The Big Data management studio can shows the all installed packs Hadoop, SQOOP, PIG and HIVE. This three are the tool in Hadoop to performing the various operations.

Steps.

1. First start the hadoop services then its shows the statuses

   Name node

   Data node

   Job history server

   Recourse manager

   Node manager

   Under service monitor its shows the all are running, if it is not started there is a some problems in hadoop.
2. Goto hadoop and upload the file witch we downloaded to HDFS drive, which is CSV file.

3. Then goto PIG part then LOAD the data from HDFS drive to data node with the help of LOAD keyword.

   “jobs = LOAD '/GettingStarted/Input/scheduling' using PigStorage(',') as (url,id,created_at,updated_at,subject,description,status,recipient,submitter_id,assignee_id,group_id)”

4. Then using group by clause like SQL according to our requirement we should write queries then it processed by processing node.

   “group1 = GROUP jobs BY submitter_id;

   group2 = FOREACH group1 generate group;

   Dump group2;”

5. Finally we can see our analysed hadoop result.

Now we are going through the HDFS filed shown in Big Data management studio and upload the collected RawData.

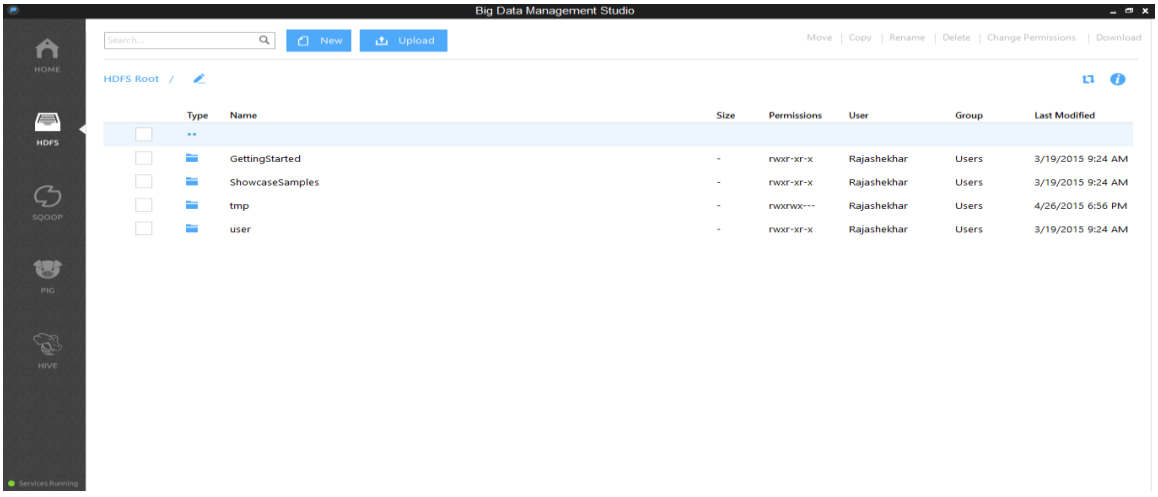


Figure23: HDFS interface

Now we are going to Pig interface and write small script for to analyze the RawData.

To upload the RawData file in to any filled or create ne inter face on HDFS module. After completion of particular file we are go through analyzation tool. And create small script for execute that file.

Figure24: PIG Interface and script to analyze RawData file

To execute the script and see the result

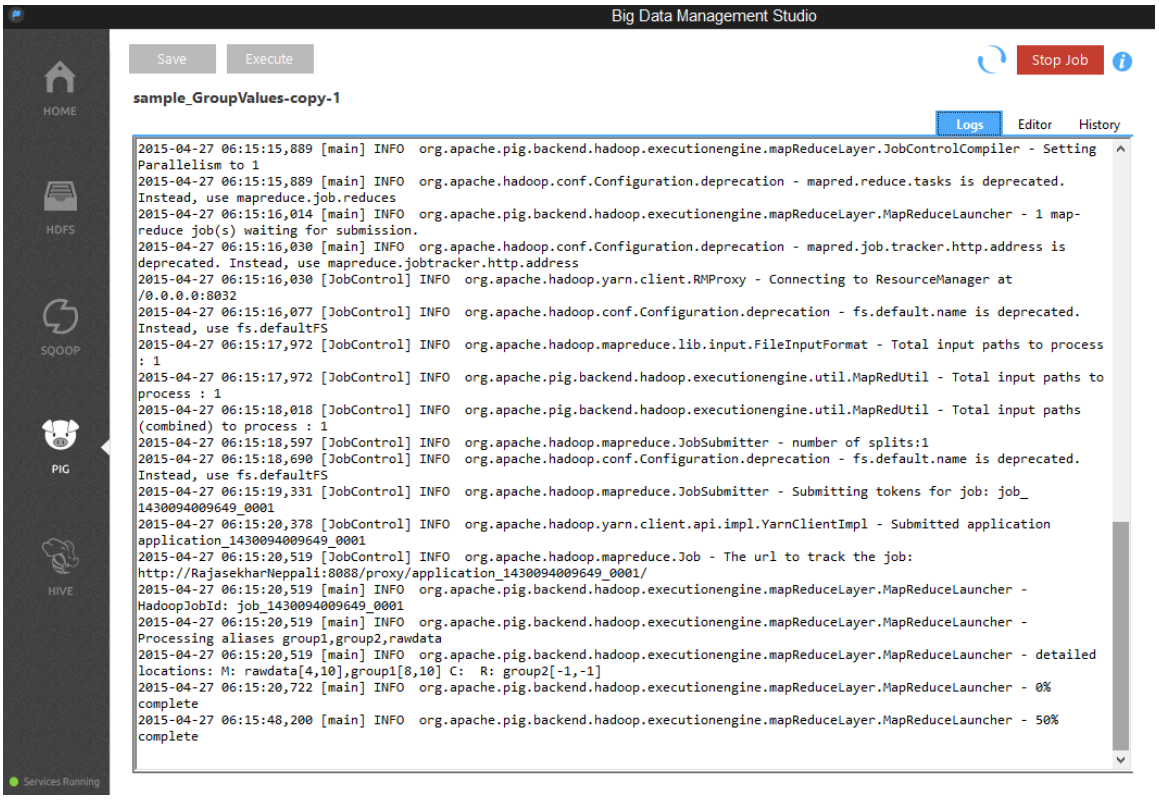


Figure25: RawData execution process in PIG

After completion of the script we found the job submitters ids in the screen.

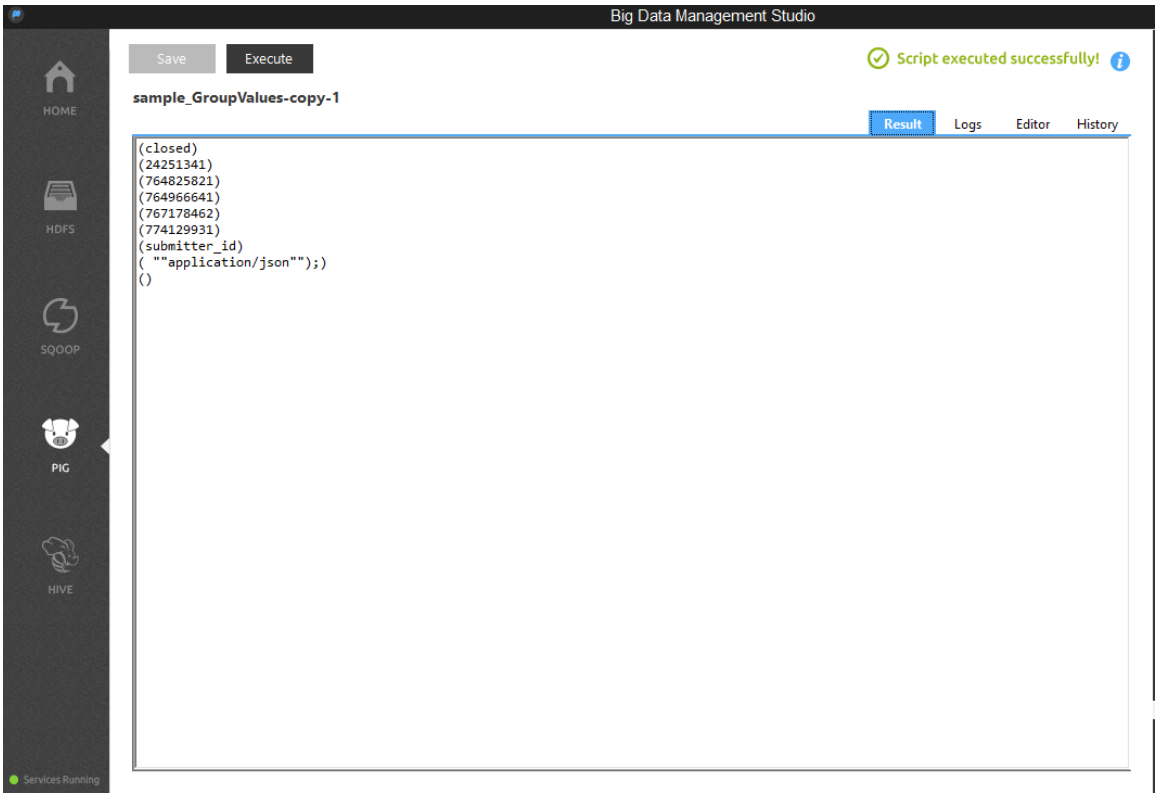


Figure26: Job submitter id's

We can find the receiving RawData to some of the submitters "Id's" .

Chapter 5
CONCLUSION

The presented job sharing process maybe better than the previous process. Present we are considering only scheduling algorithms and Apache Hadoop framework, some of the organizations to create their own software frameworks to perform the MapReduce faction and data analyze methods. We mainly focusing based on the heterogeneous environment collecting data. The estimation of the job compilation using various methods, the job execution depends on the system hardware and processing speed of the networks and capacity of using clusters. So to improve the schedulers & MapReduce performance depends on this constraints also.

REFERENCES

1. Aysan Rasooli, Douglas G. Down, "*A Hybrid Scheduling Approach for Scalable Heterogeneous Hadoop Systems*" ijrsec may 2011.

2. Jilan Chen, Dan Wang and Wenbing Zhao. "*A Task Scheduling Algorithm for Hadoop Platform*", journal of computers, vol. 8, no. 4, april 2013.

3. B.Thirumala Rao, Dr. L.S.S.Reddy, "*Survey on Improved Scheduling in Hadoop MapReduce in Cloud Environments*", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011

4. Praveen Kumar, Dr Vijay Singh Rathore, "*Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce*", International Journal of Advanced Research in Computer and Communication Engineering. Vol. 3, Issue 6, June 2014

5. Matei Zaharia Andy Konwinski Anthony D. Joseph "*Improving MapReduce Performance in Heterogeneous Environments*" August 19, 2008

6. Zheng fen, Xu Yabin, Li Yanping, "*Research on Internet Hot Topic Detection Based on MapReduce Architecture*", 4th International Conference on Intelligent Human-Machine Systems and Cybernetics-2012.

7. Apoorva Rathi, Nilesh Parmar *"Optimized and Secure Cloud Computing Using Virtualization: A Survey"* IJARCCE, November 2013

8. Kamal Kc, Kemafor Anyanwu, "*Scheduling Hadoop Jobs to Meet Deadlines",* 2nd IEEE International Conference on Cloud Computing Technology and Science.

9. V. Krishna Reddy, B.Thirmala Rao, N.V.Sridevei, LSS.Reddy. Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing. Global Journal Computer Science & Technology May 2011

10. Matei Zaharia, Dhruba Borthaku Joydeep Sen *Sarma "Job Scheduling for Multi-User MapReduce Clusters"* Technical Report No. UCB/EECS-April- 2009.

11. Hyeokju Lee, Myoungjin Kim, Joon Her, *"Implementation of MapReduce-based Image Conversion Module in Cloud Computing Environment"may 2013*

12. Yulong Ren, Wen Tang , *"A service integrity assurance framework For cloud computing based on mapreduce"* IEEE CCIS2012.

Web References: -

1. https://www.zendesk.com/

2. https://**hadoop**.**apache**.org/

3. www.ibm.com/software/data/infosphere/**hadoop**/

4. www.cloudera.com/content/cloudera/en/.../**hadoop**-and-big-data.html

APPENDIX

LATE	Longest Approximate Time to End
HDFS	Hadoop Distributed Files System
FIFO	First In First Out
EC2	Elliptic cloud2
PaaS	Platform as a Service
IaaS	Infrastructure as a Service
SaaS	Software as a service