VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computing Science and Engineering

LAB - 10 Exercises

Course Code	**:**	**CSE3025 – Large Scale Data Processing**	**Date**	**:**	**30/10/2019**
Lab Experiment	**:**	**Practice of MapReduce Programming using Map-side Joins with Distributed Cache**	**Slots**	**:**	**L15+L16**
Instructors	**:**	**Dr. Bharadwaja Kumar and Prof. Ramesh Ragala**			

Objective:

1. To understand the detailed processing of Map-side joins with distributed cache in MapReduce Framework

Notes:

Hadoop MapReduce supports two types of joins- 1. Map Side Join and 2. Reduce side Join.

Map side Join: You can use Map side join using two different ways based on our datasets, and those depends on below conditions -

1. Both datasets are must be divided into the same number of partitions, and must be already sorted by the same key.

2. From the two datasets one must be small(something like master dataset) and able to fit into the memory of each nodes.

In this session, we are concentrated on second approach i.e distributed cache.

Dataset Information:

Let's find the user activity on social media, what are the actions user performed on popular social media like commenting on post, shared something, like something etc. And for these we have two different log files - 1. User.log 2. User_activity.log

Sample data of User.log dataset is as follows:

User_ID	User Name	email	Gender
1	Susan	smendoza0@angelfire.com	Female
2	Kathleen	knichols1@nsw.gov.au	Female
3	Marilyn	mclark2@washington.edu	Female
4	Craig	cscott3@is.gd	Male

Sample data of User_activity.log dataset is as follows:

Activity_ID	User_ID	Comment	Post Shared
1	1	looking awesome:)	http://dummyimage.com/160x166.gif/5fa2dd/ffffff
2	2	full masti	http://dummyimage.com/250x142.p

			ng/ff4444/ffffff
3	3	wow gangnam style,cool	http://dummyimage.com/124x173.png/cc0000/ffffff
4	4	welcome to the heaven	http://dummyimage.com/148x156.png/ff4444/ffffff

UserActivityMapper.java: Inside this mapper class, we are setting the properties of UserActivityVO class, user.log file is in from distributed cache.

UserActivityReducer.java: The reducer class just iterating the values and writing into the context.

UserActivityDriver.java: Here we are adding user.log file into the distributed cache.

UserActivityVO.java: This is our value object class, which will contains the fields needs to be written as an output of the project.

Execute these files and make document for result.

Exercise:

Input_datas: 1. **order_details.txt** dataset consists of customer Id and Order Id.

2. **Cust_details.txt** consists of customer ID, firstname, lastname, delivery address, and mailID.

Problem- 1: Now, you need to use map-side distributed cache (order_details.txt is a small dataset) to produce the results which includes, customer ID, firstname, order Id, delivery address and MailID.

Problem-2: Now, you need to use map-side distributed cache (order_details.txt is a small dataset) to produce the results which includes, customer ID, firstname, order Id, delivery address and gender.