

**Experiment No.5**

**Title: Computing relational algebraic operations by Map reduce on Cloudera**

# Batch:B4 Roll No.:1724008 Experiment No.:5 Title: Computing relational algebraic operations by Map reduce on Cloudera

**Resources needed: Internet, Hadoop as platform on Cloudera , Pig**

**Theory:**

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs [1]. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin which has the following key properties [1]:

**Ease of programming:** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.

**Optimization opportunities:** The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

**Extensibility:** Users can create their own functions to do special-purpose processing.

# Execution Modes

Pig has two execution modes,

**Local Mode** : To run Pig in local mode, you need access to a single machine; all files are installed and run using your local host and file system. Specify local mode using the -x flag (pig -x local).

**Mapreduce Mode:** To run Pig in mapreduce mode, you need access to a Hadoop cluster and HDFS installation. Mapreduce mode is the default mode (pig OR pig -x mapreduce).

# Pig Latin Statements

Pig Latin statements can span multiple lines and must end with a semi-colon ( ; ). By default, Pig Latin statements are processed using multi-query execution.

Pig Latin statements are generally organized as follows:

* A LOAD statement to read data from the file system.
* A series of "transformation" statements to process the data.
* A DUMP statement to view results or a STORE statement to save the results.

e.g. A = LOAD 'student' USING PigStorage() AS (name:chararray, age:int, gpa:float); B = FOREACH A GENERATE name;

DUMP B;

# Procedure:

1. **Install oracle virtual box (VirtualBox-5.1.12-112440-Win.exe for windows 8, VirtualBox-5.1.14-112924-Win.exe for windows 7)**
2. **Copy cloudera quickstart virtualbox image files on your system (already downloaded and shared on //172.17.16.11/my apps/BDA folder**
3. **Perform following steps in order to process big data on Hadoop platform as service on Cloud environment [2].**

**Step 1:** create an input file **students.txt** containing data as shown below, 001,Rajiv,Reddy,9848022337,78.9,Mumbai 002,siddarth,Battacharya,9848022338,70,Pune 003,Rajesh,Khanna,9848022339,90,Pune 004,Preethi,Agarwal,9848022330,68,Pune 005,Trupthi,Mohanthy,9848022336,89,Mumbai 006,Archana,Mishra,9848022335,78.9, Mumbai

# Step 2: Create folder “/pig/example” in HDFS and move the file from the local file system to HDFS using put command

**# Hadoop fs –mkdir /user/cloudera/pig**

**# Hadoop fs –mkdir /user/cloudera/pig/example**

**# Hadoop fs –copyFromLocal students.txt /user/cloudera/pig/example**

**Step 3: start pig in mapreduce mode #pig**

**grunt>**

**Step 4: Define schema by loading student data**

**grunt>student = load ‘/user/cloudera/pig/example/students.txt’ USING PigStorage(‘,’) as (rollno:int, fname:chararray, lname: chararray, phoneno:int, marks:float, location: chararray);**

**Step 5: Display the result grunt> dump student;**

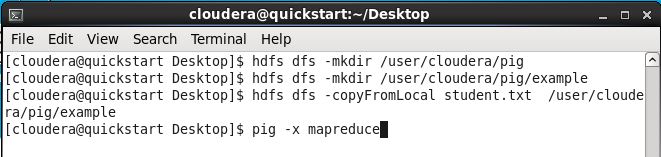
**Step 6: Similarly create schema for department**

**Step 7: Execute any 10 relational queries using pig on map reduce.**

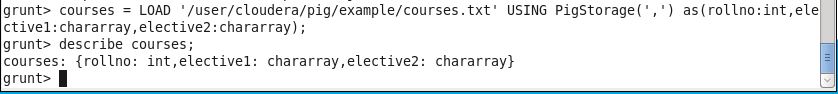
**e.g. grunt> student\_name= foreach student generate fname, lname; grunt > dump student\_name;**

**Results: (Document printout as per the format)**

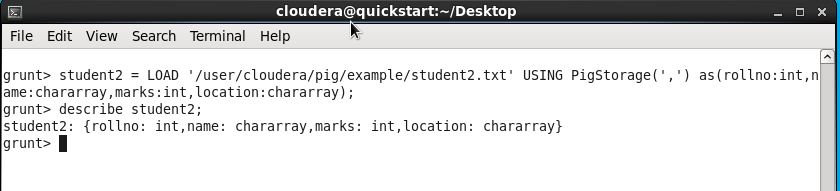
**Creating directory for each separate schema.**

****

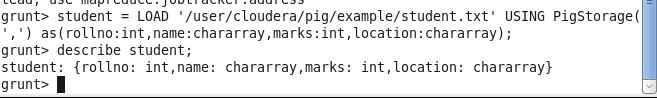
**Loading the courses data**

****

**Loading the student2 data**

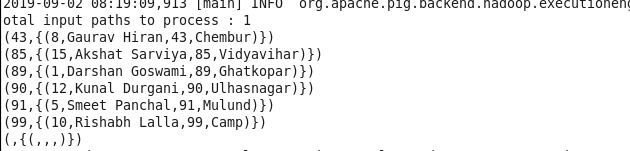
****

**Loading the student data**

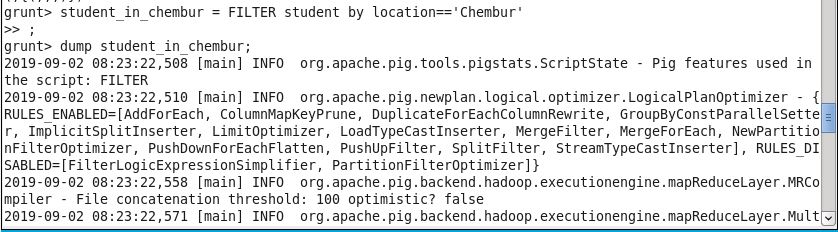
****

1. **Grouping the student data by marks.**

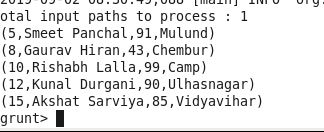
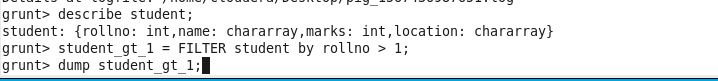
****

****

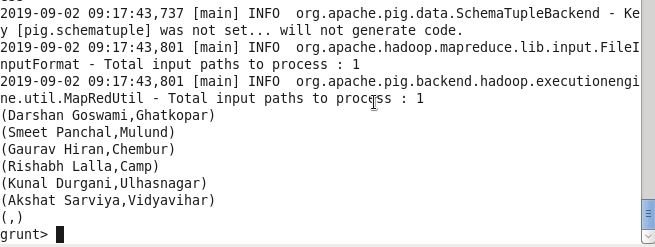
1. **Filtering the student whose location is Chembur**

****

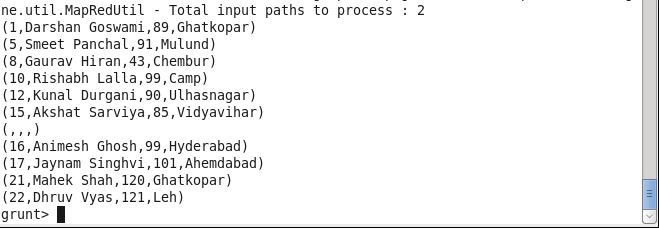
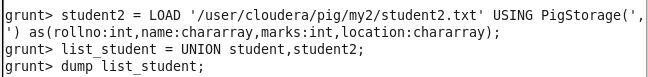
1. **Filtering the student data whose rollno is greater than 1**

****

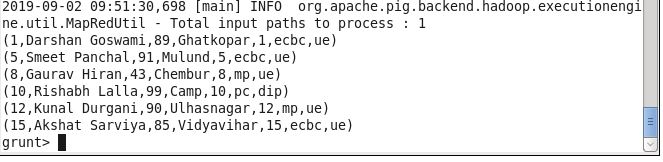
1. **Iterating each record and selecting only the name and location of that record.**

****

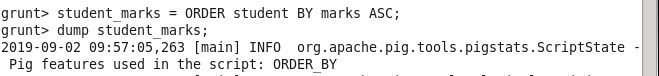
1. **Creating the UNION of student and student2 data.**

****

1. **Inner Join on the student data and courses data by a rollno as a key.**

****

1. **Arranging the student records in ascending order w.r.t marks.**

****

**Outcomes:**

**Comprehend fundamentals of Hadoop, Map Reduce and NO SQL.**

**Conclusion: (Conclusion to be based on the objectives and outcomes achieved)**

**We have installed and performed basic pig queries in cloudera platform using Virtual Box.**

**Grade: AA / AB / BB / BC / CC / CD /DD**

**Signature of faculty in-charge with date**

**References:**

**Books/ Journals/ Websites:**

1. <https://pig.apache.org/>
2. <https://www.cloudera.com/>