Project 1.2

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Project 1.2 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Project 1.2** |
|  |  |
| **Start Date:** | 06/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc495224318)

[Change History 3](#_Toc495224319)

[1. Project Overview 4](#_Toc495224320)

[1.1. Purpose and Scope of this Specification 4](#_Toc495224321)

[1.2. Product/Service Description 4](#_Toc495224322)

[2. Problem statement 5](#_Toc495224323)

[2.1. Dataset 6](#_Toc495224324)

[3. Setting environment 7](#_Toc495224325)

[4. Importing Dataset using Flume 8](#_Toc495224326)

[5. Find out the districts who achieved 100 percent objective in BPL cards Export to MySQL using sqoop 11](#_Toc495224327)

[5.1. Export the results to mysql 12](#_Toc495224328)

[6. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards – Export to MySQL using sqoop 14](#_Toc495224329)

[6.1. Export the results to mysql 16](#_Toc495224330)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 06/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Project Overview

To develop the System to analyse the log data (In XML format) of government progress of various development activities.

## Purpose and Scope of this Specification

The purpose of this project is to capture the data for analysing the progress of various activities.

In scope

The following requirement will be addressed in phase 1 of Project:

* Developing system to handle the incoming log feed and store the information in Hadoop Cluster (Flume)
* Analyze the data and understand the progress
* Store the results in Hbase/RDBMS

Out of scope

We can use this data and visualization and get more insights

## Product/Service Description

Assumptions

Log will be generated in XML format and stored in a server

Constraints

Describe any item that will constrain the design options, including

This system may not be used for searching for now. But it will be used for analysis and saving the relevant information as of now System will be using Hbase as a database

1. The FLUME job which will format the data and place the data to HDFS
2. Pig/MapReduce job for parsing the XML data.
3. Create Pig scripts/MapReduce jobs to analyse the data
4. Create the Sqoop job to store the data in database

Priority Definitions

The following definitions are intended as a guideline to prioritize requirements.

Priority 1 – Create FLUME job for fetching log files from spool directory the data

Priority 2 – MapReduce/pig job to pre-process.

# Problem statement

Input file is in the XML format use Map reduce or pig to parse the data and get the results for the below problem statements.

1. Find out the districts who achieved 100 percent objective in BPL cards

Export the results to mysql using sqoop

1. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.

Export the results to MySQL using Sqoop.

## Dataset

The dataset StatewiseDistrictwisePhysicalProgress.xml

**Sample of Data**

<PhysicalProgress>

<row>

<State\_Name>Andhra Pradesh</State\_Name>

<District\_Name>ADILABAD</District\_Name>

<Project\_Objectives\_IHHL\_BPL>247475</Project\_Objectives\_IHHL\_BPL>

<Project\_Objectives\_IHHL\_APL>148181</Project\_Objectives\_IHHL\_APL>

<Project\_Objectives\_IHHL\_TOTAL>395656</Project\_Objectives\_IHHL\_TOTAL>

<Project\_Objectives\_SCW>0</Project\_Objectives\_SCW>

<Project\_Objectives\_School\_Toilets>4462</Project\_Objectives\_School\_Toilets>

<Project\_Objectives\_Anganwadi\_Toilets>427</Project\_Objectives\_Anganwadi\_Toilets>

<Project\_Objectives\_RSM>10</Project\_Objectives\_RSM>

<Project\_Objectives\_PC>0</Project\_Objectives\_PC>

<Project\_Performance-IHHL\_BPL>176300</Project\_Performance-IHHL\_BPL>

<Project\_Performance-IHHL\_APL>52431</Project\_Performance-IHHL\_APL>

<Project\_Performance-IHHL\_TOTAL>228731</Project\_Performance-IHHL\_TOTAL>

<Project\_Performance-SCW>0</Project\_Performance-SCW>

<Project\_Performance-School\_Toilets>4462</Project\_Performance-School\_Toilets>

<Project\_Performance-Anganwadi\_Toilets>427</Project\_Performance-Anganwadi\_Toilets>

<Project\_Performance-RSM>0</Project\_Performance-RSM>

<Project\_Performance-PC>0</Project\_Performance-PC>

</row>

<row>

<State\_Name>Andhra Pradesh</State\_Name>

<District\_Name>ANANTAPUR</District\_Name>

<Project\_Objectives\_IHHL\_BPL>363314</Project\_Objectives\_IHHL\_BPL>

<Project\_Objectives\_IHHL\_APL>181335</Project\_Objectives\_IHHL\_APL>

<Project\_Objectives\_IHHL\_TOTAL>544649</Project\_Objectives\_IHHL\_TOTAL>

<Project\_Objectives\_SCW>0</Project\_Objectives\_SCW>

<Project\_Objectives\_School\_Toilets>3421</Project\_Objectives\_School\_Toilets>

<Project\_Objectives\_Anganwadi\_Toilets>284</Project\_Objectives\_Anganwadi\_Toilets>

<Project\_Objectives\_RSM>10</Project\_Objectives\_RSM>

<Project\_Objectives\_PC>0</Project\_Objectives\_PC>

<Project\_Performance-IHHL\_BPL>366557</Project\_Performance-IHHL\_BPL>

<Project\_Performance-IHHL\_APL>42000</Project\_Performance-IHHL\_APL>

<Project\_Performance-IHHL\_TOTAL>408557</Project\_Performance-IHHL\_TOTAL>

<Project\_Performance-SCW>0</Project\_Performance-SCW>

<Project\_Performance-School\_Toilets>4258</Project\_Performance-School\_Toilets>

<Project\_Performance-Anganwadi\_Toilets>302</Project\_Performance-Anganwadi\_Toilets>

<Project\_Performance-RSM>0</Project\_Performance-RSM>

<Project\_Performance-PC>0</Project\_Performance-PC>

</row>

# Setting environment

**All services are started using**

**start-all.sh**

**mr-jobhistory-daemon.sh start historyserver**

**sudo service mysqld start**

**Directories created**

**/home/cloudera/jars**

**/home/cloudera/Datasets**

[*cloudera@quickstart ~]$ ls /home/cloudera/Datasets StatewiseDistrictwisePhysicalProgress.xml*

**/home/cloudera/Flume**

**/user/cloudera/Flume\_Data**

*[cloudera@quickstart ~]$ hadoop fs -mkdir /user/cloudera/Flume\_Data*

# Importing Dataset using Flume

Create a flumeimport.conf file to define sink, channel and source

*# Name the components on this agent*

*agent.sources = fsource*

*agent.sinks = fsink*

*agent.channels = MYC01*

*# Associate channel with source and sink*

*agent.sources.fsource.channels = MYC01*

*agent.sinks.fsink.channel = MYC01*

*# Configure the source*

*agent.sources.fsource.type = spooldir*

*agent.sources.fsource.spoolDir = /home/cloudera/Datasets*

*agent.sources.fsource.fileHeader = false*

*agent.sources.fsource.fileSuffix = .COMPLETED*

*# Configure the sink*

*agent.sinks.fsink.type = hdfs*

*agent.sinks.fsink.hdfs.path = hdfs://localhost:8020/user/cloudera/Flume\_Data*

*agent.sinks.fsink.hdfs.fileType = DataStream*

*agent.sinks.fsink.hdfs.writeFormat=Text*

*agent.sinks.fsink.hdfs.batchSize = 100*

*agent.sinks.fsink.hdfs.rollSize = 100000*

*agent.sinks.fsink.hdfs.rollInterval = 0*

*agent.sinks.fsink.hdfs.rollCount = 5000*

*# Use a channel which buffers events in memory*

*agent.channels.MYC01.type = memory*

*agent.channels.MYC01.capacity = 10000*

*agent.channels.MYC01.transactionCapacity = 500*

Command run

*[cloudera@quickstart ~]$ flume-ng agent -n agent -f /home/cloudera/Flume/flumeimport.conf*

**Process starting**

*17/10/07 19:50:24 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting*

*17/10/07 19:50:24 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/home/cloudera/Flume/flumeimport.conf*

*17/10/07 19:50:24 INFO conf.FlumeConfiguration: Processing:fsink*

*17/10/07 19:50:24 INFO conf.FlumeConfiguration: Added sinks: fsink Agent: agent*

*17/10/07 19:50:24 INFO conf.FlumeConfiguration: Processing:fsink*

*17/10/07 19:50:24 INFO conf.FlumeConfiguration: Processing:fsink*

*17/10/07 19:50:24 INFO conf.FlumeConfiguration: Processing:fsink*

*17/10/07 19:50:24 INFO conf.FlumeConfiguration: Processing:fsink*

*17/10/07 19:50:24 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [agent]*

*17/10/07 19:50:24 INFO node.AbstractConfigurationProvider: Creating channels*

*17/10/07 19:50:24 INFO channel.DefaultChannelFactory: Creating instance of channel MYC01 type memory*

*17/10/07 19:50:24 INFO node.AbstractConfigurationProvider: Created channel MYC01*

*17/10/07 19:50:24 INFO source.DefaultSourceFactory: Creating instance of source fsource, type spooldir*

*17/10/07 19:50:24 INFO sink.DefaultSinkFactory: Creating instance of sink: fsink, type: hdfs*

*17/10/07 19:50:24 INFO node.AbstractConfigurationProvider: Channel MYC01 connected to [fsource, fsink]*

**Process completing**

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157437.tmp to hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157437*

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157438.tmp*

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Closing hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157438.tmp*

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157438.tmp to hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157438*

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157439.tmp*

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Closing hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157439.tmp*

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157439.tmp to hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157439*

*17/10/07 20:27:07 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157440.tmp*

*^C17/10/07 20:33:32 INFO lifecycle.LifecycleSupervisor: Stopping lifecycle supervisor 10*

*17/10/07 20:33:32 INFO hdfs.HDFSEventSink: Closing hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData*

*17/10/07 20:33:32 INFO hdfs.BucketWriter: Closing hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157440.tmp*

*17/10/07 20:33:32 INFO hdfs.BucketWriter: Renaming hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157440.tmp to hdfs://localhost:8020/user/cloudera/Flume\_Data/FlumeData.1507433157440*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: fsink stopped*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.start.time == 1507433157063*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.stop.time == 1507433613000*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.batch.complete == 142*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.batch.empty == 59*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.batch.underflow == 2*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.connection.closed.count == 9*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.connection.creation.count == 9*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.connection.failed.count == 0*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.event.drain.attempt == 14284*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: fsink. sink.event.drain.sucess == 14284*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: fsource stopped*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. source.start.time == 1507433157092*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. source.stop.time == 1507433613002*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. src.append-batch.accepted == 144*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. src.append-batch.received == 145*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. src.append.accepted == 0*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. src.append.received == 0*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. src.events.accepted == 14284*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. src.events.received == 14384*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: fsource. src.open-connection.count == 0*

*17/10/07 20:33:33 INFO source.SpoolDirectorySource: SpoolDir source fsource stopped. Metrics: SOURCE:fsource{src.events.accepted=14284, src.events.received=14384, src.append.accepted=0, src.append-batch.accepted=144, src.open-connection.count=0, src.append-batch.received=145, src.append.received=0}*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MYC01 stopped*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.start.time == 1507433157060*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.stop.time == 1507433613002*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.capacity == 10000*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.current.size == 0*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.event.put.attempt == 14384*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.event.put.success == 14284*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.event.take.attempt == 14345*

*17/10/07 20:33:33 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MYC01. channel.event.take.success == 14284*

**Check Results**

*hadoop fs -ls /user/cloudera/Flume\_Data/\**

*Found 14 items*

*-rw-r--r-- 1 cloudera cloudera 35778 2017-10-07 20:13 /user/cloudera/Flume\_Data/FlumeData.1507432394882*

*-rw-r--r-- 1 cloudera cloudera 71100 2017-10-07 20:16 /user/cloudera/Flume\_Data/FlumeData.1507432515521*

*-rw-r--r-- 1 cloudera cloudera 59392 2017-10-07 20:16 /user/cloudera/Flume\_Data/FlumeData.1507432575737*

*-rw-r--r-- 1 cloudera cloudera 297010 2017-10-07 20:18 /user/cloudera/Flume\_Data/FlumeData.1507432655545*

*-rw-r--r-- 1 cloudera cloudera 59402 2017-10-07 20:18 /user/cloudera/Flume\_Data/FlumeData.1507432655546*

*-rw-r--r-- 1 cloudera cloudera 101763 2017-10-07 20:26 /user/cloudera/Flume\_Data/FlumeData.1507433157432*

*-rw-r--r-- 1 cloudera cloudera 101786 2017-10-07 20:27 /user/cloudera/Flume\_Data/FlumeData.1507433157433*

*-rw-r--r-- 1 cloudera cloudera 101777 2017-10-07 20:27 /user/cloudera/Flume\_Data/FlumeData.1507433157434*

*-rw-r--r-- 1 cloudera cloudera 101734 2017-10-07 20:27 /user/cloudera/Flume\_Data/FlumeData.1507433157435*

*-rw-r--r-- 1 cloudera cloudera 101743 2017-10-07 20:27 /user/cloudera/Flume\_Data/FlumeData.1507433157436*

*-rw-r--r-- 1 cloudera cloudera 101777 2017-10-07 20:27 /user/cloudera/Flume\_Data/FlumeData.1507433157437*

*-rw-r--r-- 1 cloudera cloudera 101748 2017-10-07 20:27 /user/cloudera/Flume\_Data/FlumeData.1507433157438*

*-rw-r--r-- 1 cloudera cloudera 101719 2017-10-07 20:27 /user/cloudera/Flume\_Data/FlumeData.1507433157439*

*-rw-r--r-- 1 cloudera cloudera 30574 2017-10-07 20:33 /user/cloudera/Flume\_Data/FlumeData.1507433157440*

**Checking original file**

*[cloudera@quickstart ~]$ ls /home/cloudera/Datasets StatewiseDistrictwisePhysicalProgress.xml.COMPLETED*



Has been appended with **.COMPLTED** as specified in the flumeimport.conf file to show job is complete.

# 

# Find out the districts who achieved 100 percent objective in BPL cards Export to MySQL using sqoop

To achieve this a pig query is created to import the data and run the required query.

**Query run**

*REGISTER /home/cloudera/jar/piggybank.jar;*

*DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();*

*SWD = LOAD '/user/acadgild/FlumeData/\*' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);*

*A = FOREACH SWD GENERATE XPath(x, 'row/State\_Name') AS state,*

*XPath(x, 'row/District\_Name') AS dist,*

*XPath(x, 'row/Project\_Objectives\_IHHL\_BPL') AS po\_bpl,*

*XPath(x, 'row/Project\_Objectives\_IHHL\_APL') AS po\_apl,*

*XPath(x, 'row/Project\_Objectives\_IHHL\_TOTAL') AS po\_total,*

*XPath(x, 'row/Project\_Objectives\_SCW') AS po\_scw,*

*XPath(x, 'row/Project\_Objectives\_School\_Toilets') AS po\_school\_toilets,*

*XPath(x, 'row/Project\_Objectives\_Anganwadi\_Toilets') AS po\_anganwadi\_toilets,*

*XPath(x, 'row/Project\_Objectives\_RSM') AS po\_rsm,*

*XPath(x, 'row/Project\_Objectives\_PC') AS po\_ps,*

*XPath(x, 'row/Project\_Performance-IHHL\_BPL') AS pp\_bpl,*

*XPath(x, 'row/Project\_Performance-IHHL\_APL') AS pp\_apl,*

*XPath(x, 'row/Project\_Performance-IHHL\_TOTAL') AS pp\_total,*

*XPath(x, 'row/Project\_Performance-SCW') AS pp\_scw,*

*XPath(x, 'row/Project\_Performance-School\_Toilets') AS pp\_school\_toilets,*

*XPath(x, 'row/Project\_Performance-Anganwadi\_Toilets') AS pp\_anganwadi\_toilets,*

*XPath(x, 'row/Project\_Performance-RSM') AS pp\_rsm,*

*XPath(x, 'row/Project\_Performance-PC') AS pp\_pc;*

*B = FOREACH A GENERATE (chararray)state, (chararray)dist, (int)po\_bpl, (int)pp\_bpl;*

*PP\_BLP = FILTER B BY po\_bpl<=pp\_bpl;*

*STORE PP\_BLP INTO '/user/cloudera/India01';*

**Processing complete**

HadoopVersionPigVersionUserIdStartedAtFinishedAtFeatures

2.6.0-cdh5.12.00.12.0-cdh5.12.0cloudera2017-10-07 20:42:282017-10-07 20:43:56FILTER

*Success!*

*Job Stats (time in seconds):*

*JobIdMapsReducesMaxMapTimeMinMapTImeAvgMapTimeMedianMapTimeMaxReduceTimeMinReduceTimeAvgReduceTimeMedianReducetimeAliasFeatureOutputs*

*job\_1507406416390\_00011052525252n/an/an/an/aA,B,PP\_BPL,SWDMAP\_ONLY/user/cloudera/India01,*

*Input(s):*

*Successfully read 866 records (1368963 bytes) from: "/user/cloudera/Flume\_Data/\*"*

*Output(s):*

*Successfully stored 307 records (10603 bytes) in: "/user/cloudera/India01"*

*Counters:*

*Total records written : 307*

*Total bytes written : 10603*

*Spillable Memory Manager spill count : 0*

*Total bags proactively spilled: 0*

*Total records proactively spilled: 0*

*Job DAG:*

*job\_1507406416390\_0001*

**Check Results**

**Found 2 items**

-*rw-r--r-- 1 cloudera cloudera 0 2017-10-07 20:43 /user/cloudera/India01/\_SUCCESS*

*-rw-r--r-- 1 cloudera cloudera 10603 2017-10-07 20:43 /user/cloudera/India01/part-m-00000*

*[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/India01//part-m-00000*

*Andhra Pradesh ANANTAPUR 363314 366557*

*Andhra Pradesh KARIMNAGAR 365267 369433*

*Andhra Pradesh KHAMMAM 189225 195763*

*Andhra Pradesh NALGONDA 215058 224813*

*Andhra Pradesh NIZAMABAD 225519 225519*

*Andhra Pradesh WARANGAL 330260 359732*

*West Bengal DAKSHIN DINAJPUR 182621 184153*

*West Bengal MIDNAPUR EAST 392371 527389*

*West Bengal MIDNAPUR WEST 509496 596291*

*Andhra Pradesh ANANTAPUR 363314 366557*

*Andhra Pradesh KARIMNAGAR 365267 369433*

*Andhra Pradesh KHAMMAM 189225 195763*

*Andhra Pradesh NALGONDA 215058 224813*

*Andhra Pradesh NIZAMABAD 225519 225519*

*Andhra Pradesh WARANGAL 330260 359732*

## Export the results to mysql

**Started mysql and login**

*mysql -u root -p*

*Enter password: \*\*\*\*\*\**

*Welcome to the MySQL monitor. Commands end with ; or \g.*

*Your MySQL connection id is 54*

*Server version: 5.1.73 Source distribution*

**Create Database and Tables**

*mysql> create database india\_project;*

*Query OK, 1 row affected (0.00 sec)*

*mysql> use india\_project*

*mysql> create table 100achieved*

*-> (*

*-> state varchar(30),*

*-> dist varchar(30),*

*-> po\_bpl int,*

*-> pp\_bpl int*

*-> );*

*Query OK, 0 rows affected (0.02 sec)*

**Squoop command run**

*[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/india\_project --username 'root' -P --table '100achieved' --export-dir '/user/cloudera/India01/' --input-fields-terminated-by '\t' -m 1*

**Processing complete**

*17/10/08 02:57:49 INFO mapreduce.Job: Job job\_1507406416390\_0011 completed successfully*

*17/10/08 02:57:49 INFO mapreduce.Job: Counters: 30*

*File System Counters*

*FILE: Number of bytes read=0*

*FILE: Number of bytes written=151201*

*FILE: Number of read operations=0*

*FILE: Number of large read operations=0*

*FILE: Number of write operations=0*

*HDFS: Number of bytes read=10753*

*HDFS: Number of bytes written=0*

*HDFS: Number of read operations=4*

*HDFS: Number of large read operations=0*

*HDFS: Number of write operations=0*

*Job Counters*

*Launched map tasks=1*

*Data-local map tasks=1*

*Total time spent by all maps in occupied slots (ms)=6691*

*Total time spent by all reduces in occupied slots (ms)=0*

*Total time spent by all map tasks (ms)=6691*

*Total vcore-milliseconds taken by all map tasks=6691*

*Total megabyte-milliseconds taken by all map tasks=6851584*

*Map-Reduce Framework*

*Map input records=307*

*Map output records=307*

*Input split bytes=147*

*Spilled Records=0*

*Failed Shuffles=0*

*Merged Map outputs=0*

*GC time elapsed (ms)=169*

*CPU time spent (ms)=1100*

*Physical memory (bytes) snapshot=199389184*

*Virtual memory (bytes) snapshot=1567305728*

*Total committed heap usage (bytes)=221249536*

*File Input Format Counters*

*Bytes Read=0*

*File Output Format Counters*

*Bytes Written=0*

*17/10/08 02:57:49 INFO mapreduce.ExportJobBase: Transferred 10.501 KB in 23.0166 seconds (467.1837 bytes/sec)*

*17/10/08 02:57:49 INFO mapreduce.ExportJobBase: Exported 307 records.*

**Check Results**

*mysql> select \* from 100achieved;*

*+-------------------+-----------------------------+--------+--------+*

*| state | dist | po\_bpl | pp\_bpl |*

*+-------------------+-----------------------------+--------+--------+*

*| Andhra Pradesh | ANANTAPUR | 363314 | 366557 |*

*| Andhra Pradesh | KARIMNAGAR | 365267 | 369433 |*

*| Andhra Pradesh | KHAMMAM | 189225 | 195763 |*

*| Andhra Pradesh | NALGONDA | 215058 | 224813 |*

*| Andhra Pradesh | NIZAMABAD | 225519 | 225519 |*

*| Andhra Pradesh | WARANGAL | 330260 | 359732 |*

*| West Bengal | DAKSHIN DINAJPUR | 182621 | 184153 |*

*| West Bengal | MIDNAPUR EAST | 392371 | 527389 |*

*| West Bengal | MIDNAPUR WEST | 509496 | 596291 |*

*| Andhra Pradesh | ANANTAPUR | 363314 | 366557 |*

*| Andhra Pradesh | KARIMNAGAR | 365267 | 369433 |*

*| Andhra Pradesh | KHAMMAM | 189225 | 195763 |*

*| Andhra Pradesh | NALGONDA | 215058 | 224813 |*

*| Andhra Pradesh | NIZAMABAD | 225519 | 225519 |*

*| Andhra Pradesh | WARANGAL | 330260 | 359732 |*

*| Arunachal Pradesh | DIBANG VALLEY | 1085 | 1088 |*

*| Uttarakhand | RUDRAPRAYAG | 13150 | 13810 |*

*| West Bengal | DAKSHIN DINAJPUR | 182621 | 184153 |*

*| West Bengal | MIDNAPUR EAST | 392371 | 527389 |*

*| West Bengal | MIDNAPUR WEST | 509496 | 596291 |*

*+-------------------+-----------------------------+--------+--------+*

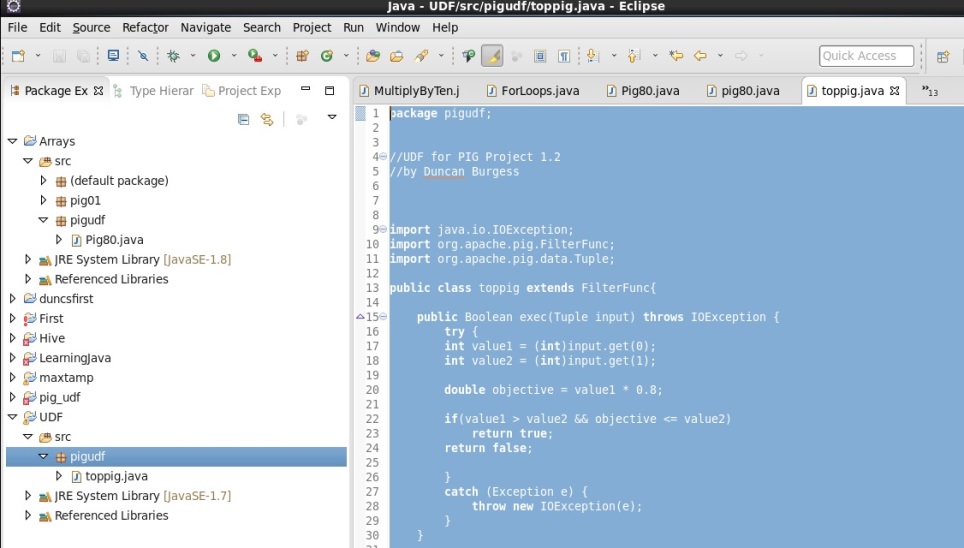
*307 rows in set (0.00 sec)*

# Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards – Export to MySQL using sqoop

For this a UDF need to be created using eclipse.

The following jar files are required in order to compile pig IDF java program.

* pig-0.8.3.jar
* hadoop-common-2.6.0.jar
* commons-logging-1.1.1.jar



**UDF created**

*package pigudf;*

*//UDF for PIG Project 1.2*

*//by Duncan Burgess*

*import java.io.IOException;*

*import org.apache.pig.FilterFunc;*

*import org.apache.pig.data.Tuple;*

*public class toppig extends FilterFunc{*

*public Boolean exec(Tuple input) throws IOException {*

*try {*

*int value1 = (int)input.get(0);*

*int value2 = (int)input.get(1);*

*double objective = value1 \* 0.8;*

*if(value1 > value2 && objective <= value2)*

*return true;*

*return false;*

*}*

*catch (Exception e) {*

*throw new IOException(e);*

*}*

*}*

*}*

This is exported as a jar and copied to /home/acadgild/jar along with piggybank.jar.

To achieve this a pig query is created to import the data and run the required query.

**Query run**

*DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();*

*REGISTER /home/cloudera/jars/toppig.jar*

*DEFINE Top80 pigudf.toppig*

*SWD = LOAD '/user/cloudera/Flume\_Data/\*' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);*

*A = FOREACH SWD GENERATE XPath(x, 'row/State\_Name') AS state,*

*XPath(x, 'row/District\_Name') AS dist,*

*XPath(x, 'row/Project\_Objectives\_IHHL\_BPL') AS po\_bpl,*

*XPath(x, 'row/Project\_Objectives\_IHHL\_APL') AS po\_apl,*

*XPath(x, 'row/Project\_Objectives\_IHHL\_TOTAL') AS po\_total,*

*XPath(x, 'row/Project\_Objectives\_SCW') AS po\_scw,*

*XPath(x, 'row/Project\_Objectives\_School\_Toilets') AS po\_school\_toilets,*

*XPath(x, 'row/Project\_Objectives\_Anganwadi\_Toilets') AS po\_anganwadi\_toilets,*

*XPath(x, 'row/Project\_Objectives\_RSM') AS po\_rsm,*

*XPath(x, 'row/Project\_Objectives\_PC') AS po\_ps,*

*XPath(x, 'row/Project\_Performance-IHHL\_BPL') AS pp\_bpl,*

*XPath(x, 'row/Project\_Performance-IHHL\_APL') AS pp\_apl,*

*XPath(x, 'row/Project\_Performance-IHHL\_TOTAL') AS pp\_total,*

*XPath(x, 'row/Project\_Performance-SCW') AS pp\_scw,*

*XPath(x, 'row/Project\_Performance-School\_Toilets') AS pp\_school\_toilets,*

*XPath(x, 'row/Project\_Performance-Anganwadi\_Toilets') AS pp\_anganwadi\_toilets,*

*XPath(x, 'row/Project\_Performance-RSM') AS pp\_rsm,*

*XPath(x, 'row/Project\_Performance-PC') AS pp\_pc;*

*B = FOREACH A GENERATE (chararray)state, (chararray)dist, (int)po\_bpl, (int)pp\_bpl;*

*BO\_BP\_BPL = FILTER B BY Top80(po\_bpl, pp\_bpl);*

*STORE BO\_BP\_BPL INTO '/user/cloudera/India02';*

**Processing complete**

*HadoopVersionPigVersionUserIdStartedAtFinishedAtFeatures*

*2.6.0-cdh5.12.00.12.0-cdh5.12.0cloudera2017-10-07 21:46:002017-10-07 21:47:12FILTER*

*Success!*

*Job Stats (time in seconds):*

*JobIdMapsReducesMaxMapTimeMinMapTImeAvgMapTimeMedianMapTimeMaxReduceTimeMinReduceTimeAvgReduceTimeMedianReducetimeAliasFeatureOutputs*

*job\_1507406416390\_00061050505050n/an/an/an/aA,B,BO\_BP\_BPL,SWDMAP\_ONLY/user/cloudera/India02,*

*Input(s):*

*Successfully read 866 records (1368963 bytes) from: "/user/cloudera/Flume\_Data/\*"*

*Output(s):*

*Successfully stored 264 records (9014 bytes) in: "/user/cloudera/India02"*

*Counters:*

*Total records written : 264*

*Total bytes written : 9014*

*Spillable Memory Manager spill count : 0*

*Total bags proactively spilled: 0*

*Total records proactively spilled: 0*

**Check Results**

[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/India02

Found 2 items

-rw-r--r-- 1 cloudera cloudera 0 2017-10-07 21:47 /user/cloudera/India02/\_SUCCESS

-rw-r--r-- 1 cloudera cloudera 9014 2017-10-07 21:47 /user/cloudera/India02/part-m-00000

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/India02//part-m-00000

Andhra Pradesh CHITTOOR 296465 269750

Andhra Pradesh CUDDAPAH 251653 239780

Andhra Pradesh EAST GODAVARI 370255 347305

Andhra Pradesh KRISHNA 351572 318730

Andhra Pradesh KURNOOL 383478 323616

Andhra Pradesh MEDAK 311743 310591

Andhra Pradesh RANGAREDDI 212629 174460

Andhra Pradesh WEST GODAVARI 344272 319477

Uttarakhand UDHAM SINGH NAGAR 39427 37604

Uttarakhand UTTARKASHI 28189 25523

West Bengal BARDHAMAN 700047 601906

## Export the results to mysql

**Started mysql and login**

*mysql -u root -p*

*Enter password: \*\*\*\*\*\**

*Welcome to the MySQL monitor. Commands end with ; or \g.*

*Your MySQL connection id is 54*

*Server version: 5.1.73 Source distribution*

**Use existing Database and create tables**

*mysql> use india\_project*

*use india\_project;*

*create table 80achieved*

*(*

*state varchar(30),*

*dist varchar(30),*

*po\_bpl int,*

*pp\_bpl int*

*);*

Sqoop command run

*[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/india\_project --username 'root' -P --table '80achieved' --export-dir '/user/cloudera/India02/' --input-fields-terminated-by '\t' -m 1*

**Processing complete**

*17/10/07 22:32:17 INFO mapreduce.ExportJobBase: Transferred 8.9492 KB in 19.7587 seconds (463.7955 bytes/sec)*

*17/10/07 22:32:17 INFO mapreduce.ExportJobBase: Exported 264 records.*

**Check Results**

*mysql> select \* from 80achieved;*

*+-------------------+--------------------------------+--------+--------+*

*| state | dist | po\_bpl | pp\_bpl |*

*+-------------------+--------------------------------+--------+--------+*

*| Andhra Pradesh | CHITTOOR | 296465 | 269750 |*

*| Andhra Pradesh | CUDDAPAH | 251653 | 239780 |*

*| Andhra Pradesh | EAST GODAVARI | 370255 | 347305 |*

*| Andhra Pradesh | KRISHNA | 351572 | 318730 |*

*| Andhra Pradesh | KURNOOL | 383478 | 323616 |*

*| Andhra Pradesh | MEDAK | 311743 | 310591 |*

*| Andhra Pradesh | RANGAREDDI | 212629 | 174460 |*

*| Andhra Pradesh | WEST GODAVARI | 344272 | 319477 |*

*| Uttarakhand | UDHAM SINGH NAGAR | 39427 | 37604 |*

*| Uttarakhand | UTTARKASHI | 28189 | 25523 |*

*| West Bengal | BARDHAMAN | 700047 | 601906 |*

*| West Bengal | HOOGHLY | 271737 | 269779 |*

*| West Bengal | HOWRAH | 231860 | 230190 |*

*| West Bengal | JALPAIGURI | 372999 | 337740 |*

*| West Bengal | NADIA | 346696 | 321462 |*

*| West Bengal | NORTH 24 PARAGANAS | 361462 | 357960 |*

*| West Bengal | SOUTH 24 PARAGANAS | 628712 | 593712 |*

*| Andhra Pradesh | CHITTOOR | 296465 | 269750 |*

*| Andhra Pradesh | CUDDAPAH | 251653 | 239780 |*

*| Andhra Pradesh | EAST GODAVARI | 370255 | 347305 |*

*| Andhra Pradesh | KRISHNA | 351572 | 318730 |*

*| Andhra Pradesh | KURNOOL | 383478 | 323616 |*

*| Andhra Pradesh | MEDAK | 311743 | 310591 |*

*| Andhra Pradesh | RANGAREDDI | 212629 | 174460 |*

*| Andhra Pradesh | WEST GODAVARI | 344272 | 319477 |*

*| Arunachal Pradesh | LOHIT | 8800 | 8410 |*

*| Assam | BAGSHA | 85697 | 73500 |*

*| Assam | CACHAR | 119931 | 101075 |*

*| Assam | DIBRUGARH | 77606 | 69914 |*

*| Assam | GOALPARA | 65070 | 55747 |*

*| Assam | GOLAGHAT | 79743 | 77985 |*

*| Assam | JORHAT | 65070 | 52698 |*

*| Assam | KAMRUP | 85785 | 84725 |*

*| West Bengal | JALPAIGURI | 372999 | 337740 |*

*| West Bengal | NADIA | 346696 | 321462 |*

*| West Bengal | NORTH 24 PARAGANAS | 361462 | 357960 |*

*| West Bengal | SOUTH 24 PARAGANAS | 628712 | 593712 |*

*+-------------------+--------------------------------+--------+--------+*

*264 rows in set (0.00 sec)*