package com.twitter.simclusters\_v2.scalding

package multi\_type\_graph.assemble\_multi\_type\_graph

import com.twitter.dal.client.dataset.KeyValDALDataset

import com.twitter.dal.client.dataset.SnapshotDALDataset

import com.twitter.scalding.Days

import com.twitter.scalding.Duration

import com.twitter.scalding.RichDate

import com.twitter.scalding\_internal.multiformat.format.keyval.KeyVal

import com.twitter.simclusters\_v2.thriftscala.LeftNode

import com.twitter.simclusters\_v2.thriftscala.RightNodeTypeStruct

import com.twitter.simclusters\_v2.thriftscala.RightNodeWithEdgeWeightList

import com.twitter.simclusters\_v2.thriftscala.NounWithFrequencyList

import com.twitter.simclusters\_v2.thriftscala.MultiTypeGraphEdge

import com.twitter.wtf.scalding.jobs.common.AdhocExecutionApp

import com.twitter.wtf.scalding.jobs.common.ScheduledExecutionApp

import com.twitter.simclusters\_v2.hdfs\_sources.\_

/\*\*

./bazel bundle src/scala/com/twitter/simclusters\_v2/scalding/multi\_type\_graph/assemble\_multi\_type\_graph:multi\_type\_graph-adhoc

scalding remote run \

--user cassowary \

--keytab /var/lib/tss/keys/fluffy/keytabs/client/cassowary.keytab \

--principal service\_acoount@TWITTER.BIZ \

--cluster bluebird-qus1 \

--main-class com.twitter.simclusters\_v2.scalding.multi\_type\_graph.assemble\_multi\_type\_graph.AssembleMultiTypeGraphAdhocApp \

--target src/scala/com/twitter/simclusters\_v2/scalding/multi\_type\_graph/assemble\_multi\_type\_graph:multi\_type\_graph-adhoc \

--hadoop-properties "mapreduce.reduce.memory.mb=8192 mapreduce.map.memory.mb=8192 mapreduce.map.java.opts='-Xmx7618M' mapreduce.reduce.java.opts='-Xmx7618M' mapreduce.task.timeout=3600000" \

-- --date 2021-07-10 --outputDir /gcs/user/cassowary/adhoc/your\_ldap/multi\_type/multi\_type

To run using scalding\_job target:

scalding remote run --target src/scala/com/twitter/simclusters\_v2/scalding/multi\_type\_graph/assemble\_multi\_type\_graph:multi\_type\_graph-adhoc

\*/

object AssembleMultiTypeGraphAdhocApp extends AssembleMultiTypeGraphBaseApp with AdhocExecutionApp {

override val isAdhoc: Boolean = true

override val truncatedMultiTypeGraphMHOutputPath: String = "truncated\_graph\_mh"

override val topKRightNounsMHOutputPath: String = "top\_k\_right\_nouns\_mh"

override val fullMultiTypeGraphThriftOutputPath: String = "full\_graph\_thrift"

override val truncatedMultiTypeGraphKeyValDataset: KeyValDALDataset[

KeyVal[LeftNode, RightNodeWithEdgeWeightList]

] = TruncatedMultiTypeGraphAdhocScalaDataset

override val topKRightNounsKeyValDataset: KeyValDALDataset[

KeyVal[RightNodeTypeStruct, NounWithFrequencyList]

] = TopKRightNounsAdhocScalaDataset

override val fullMultiTypeGraphSnapshotDataset: SnapshotDALDataset[MultiTypeGraphEdge] =

FullMultiTypeGraphAdhocScalaDataset

}

/\*\*

To deploy the job:

capesospy-v2 update --build\_locally \

--start\_cron assemble\_multi\_type\_graph \

src/scala/com/twitter/simclusters\_v2/capesos\_config/atla\_proc.yaml

\*/

object AssembleMultiTypeGraphBatchApp

extends AssembleMultiTypeGraphBaseApp

with ScheduledExecutionApp {

override val isAdhoc: Boolean = false

override val truncatedMultiTypeGraphMHOutputPath: String = "truncated\_graph\_mh"

override val topKRightNounsMHOutputPath: String = "top\_k\_right\_nouns\_mh"

override val fullMultiTypeGraphThriftOutputPath: String = "full\_graph\_thrift"

override val truncatedMultiTypeGraphKeyValDataset: KeyValDALDataset[

KeyVal[LeftNode, RightNodeWithEdgeWeightList]

] = TruncatedMultiTypeGraphScalaDataset

override val topKRightNounsKeyValDataset: KeyValDALDataset[

KeyVal[RightNodeTypeStruct, NounWithFrequencyList]

] = TopKRightNounsScalaDataset

override val fullMultiTypeGraphSnapshotDataset: SnapshotDALDataset[MultiTypeGraphEdge] =

FullMultiTypeGraphScalaDataset

override val firstTime: RichDate = RichDate("2021-08-21")

override val batchIncrement: Duration = Days(7)

}