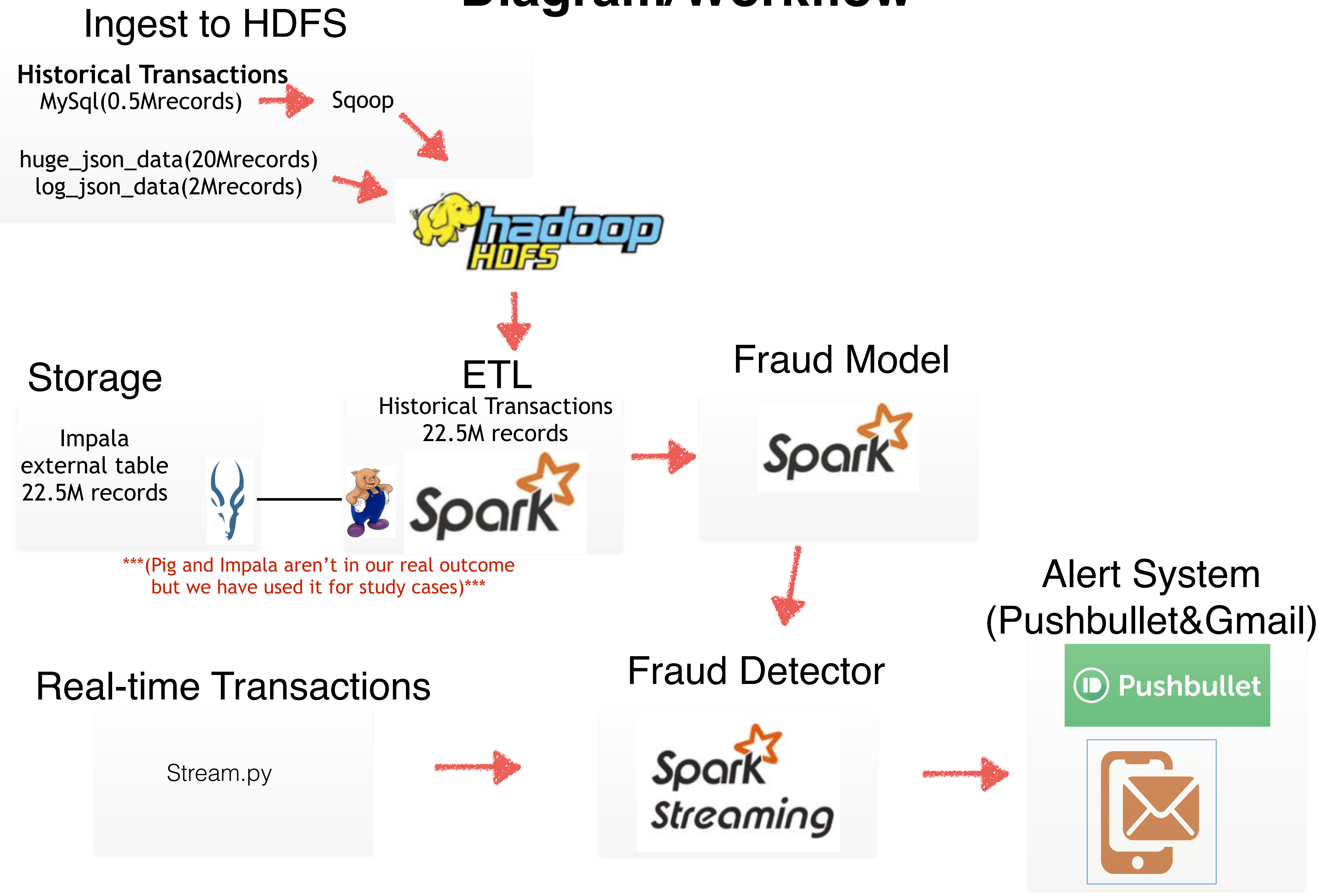
Near Real-time Fraud Detection Framework

Sirapat Na Ranong 57070503438
Chaleamkwan Bunyarattanamongkon 57070503444
Saket Khandelwal 57070503483

Diagram/Workflow



Ingest to HDFS

Historical Transactions
MySql(0.5Mrecords)

Sqoop

huge_json_data(20Mrecords)
log_json_data(2Mrecords)

Storage

Impala
external table
22.5M records

ETL

Historical Transactions
22.5M records

***(Pig and Impala aren't in our real outcome
but we have used it for study cases)***

Fraud Model

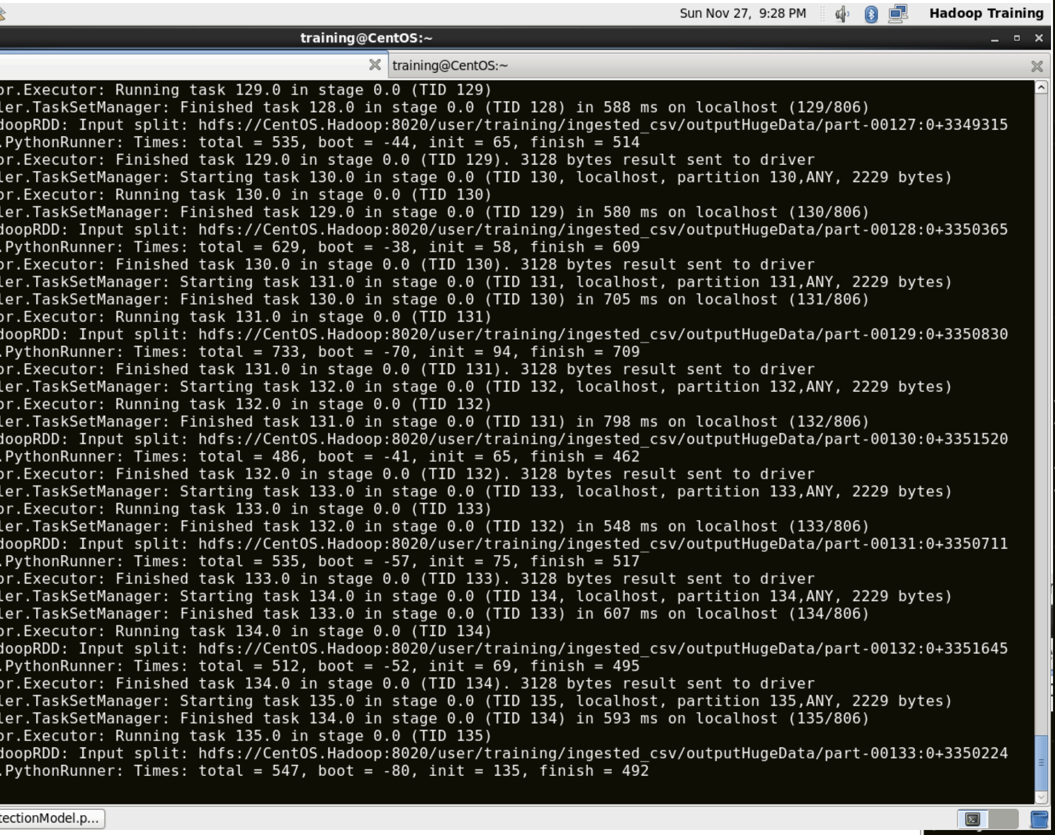
Alert System
(Pushbullet&Gmail)

Real-time Transactions

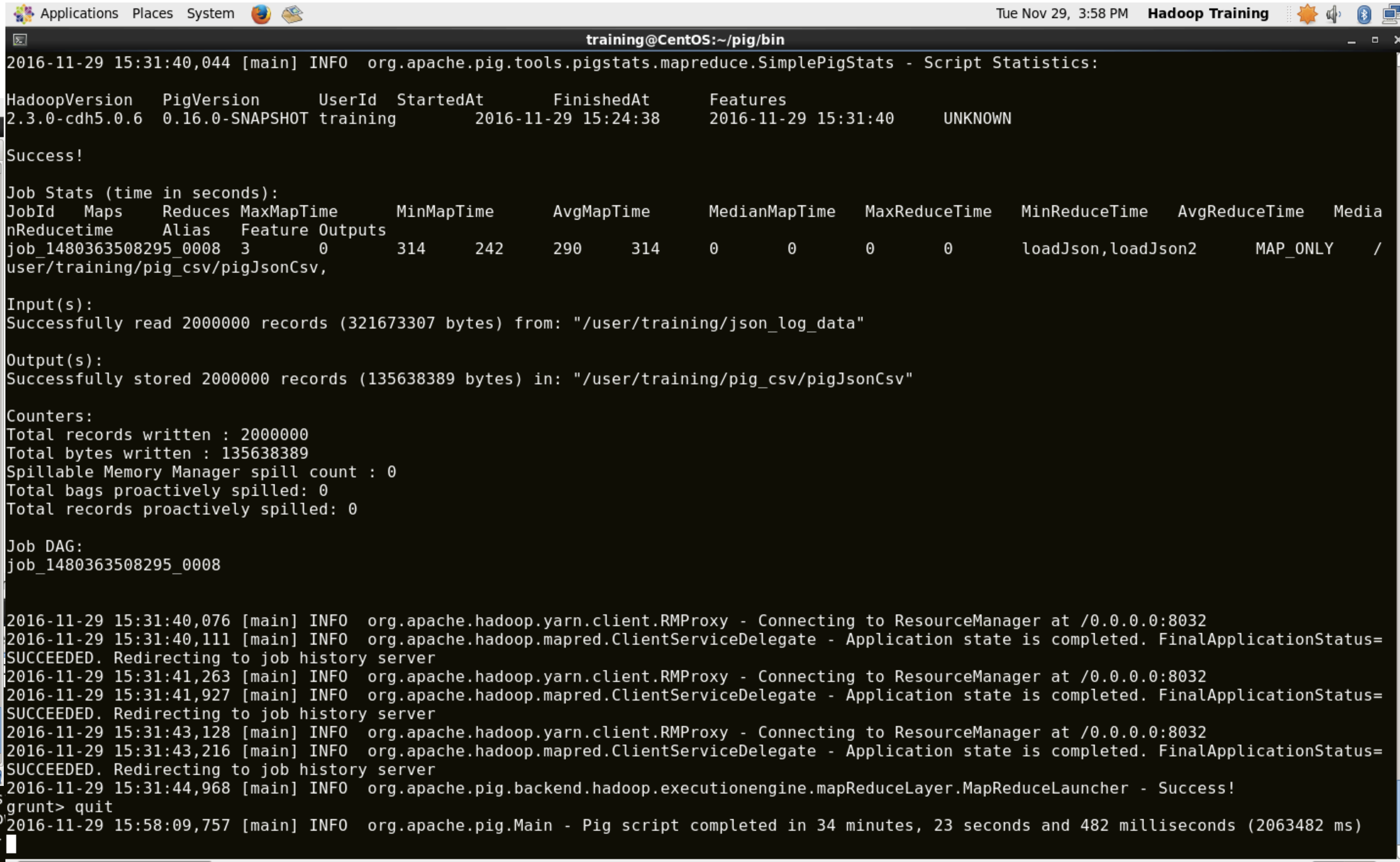
Stream.py

Fraud Detector

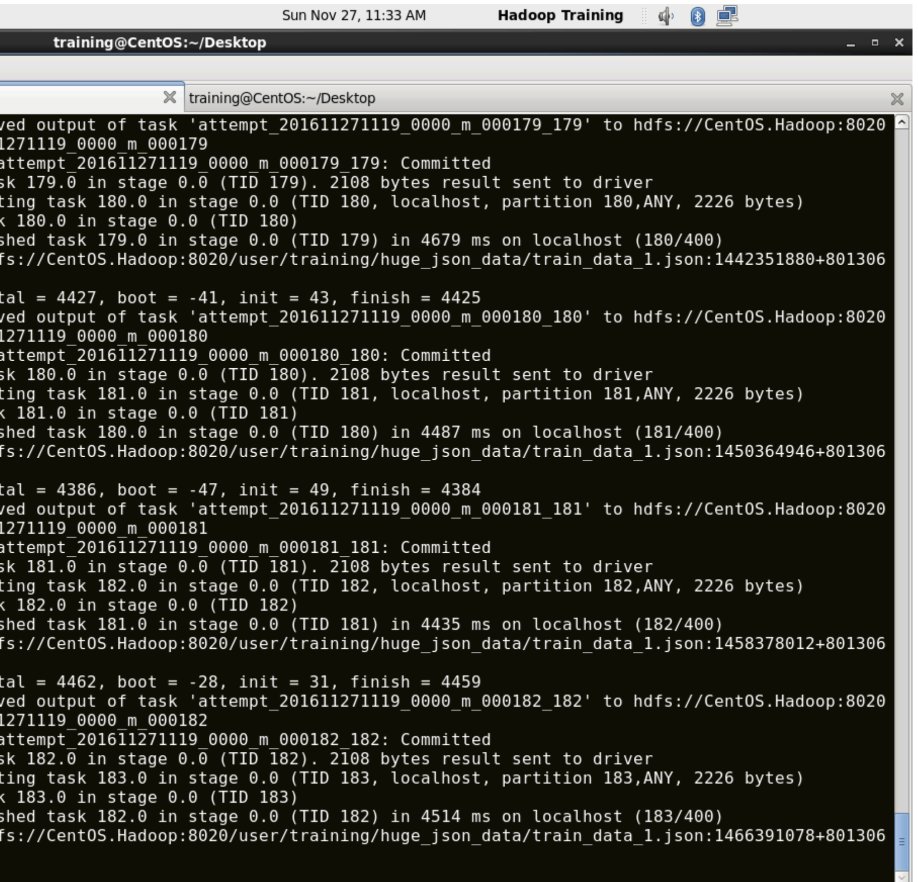
Running Process Examples (Spark[ETL&Model],Pig,Impala)



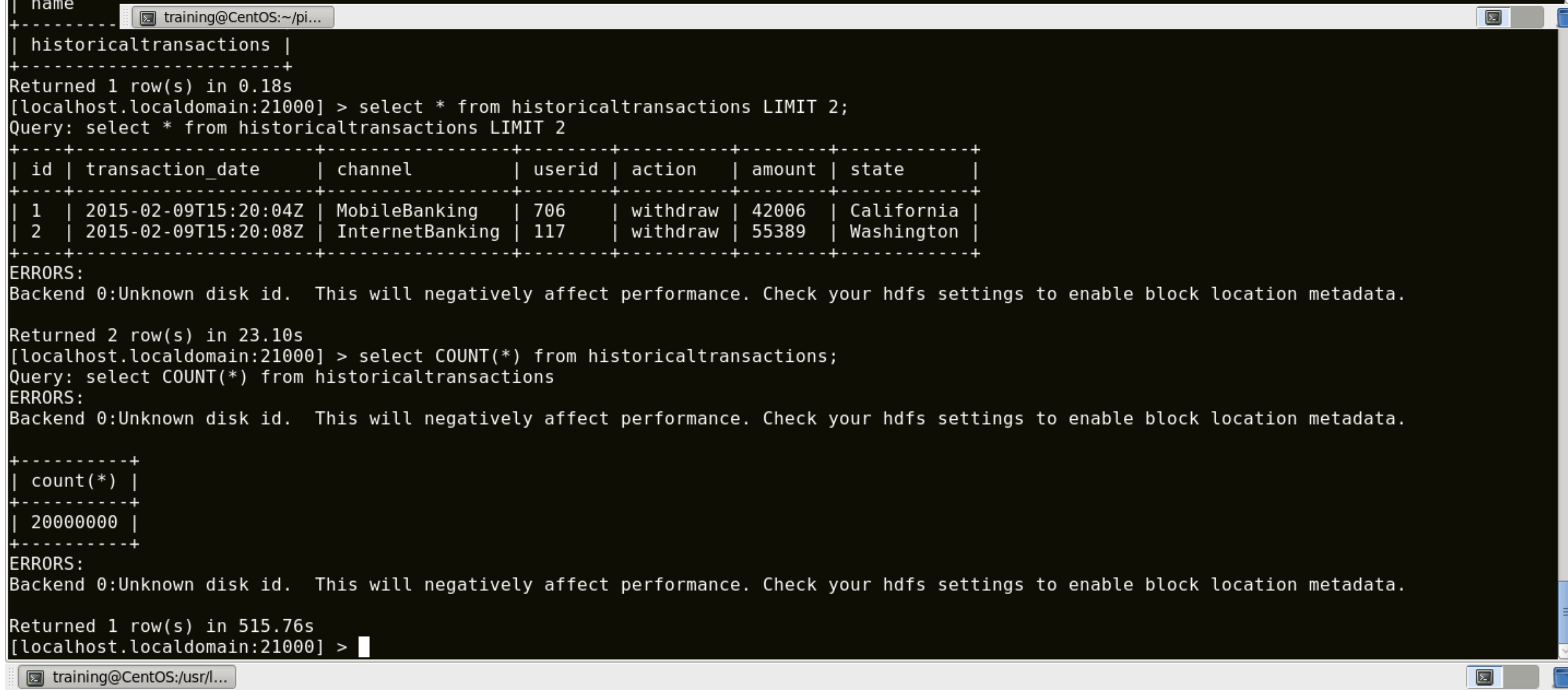
```
or.Executor: Running task 129.0 in stage 0.0 (TID 129)
ler.TaskSetManager: Finished task 128.0 in stage 0.0 (TID 128) in 588 ms on localhost (129/806)
doopRDD: Input split: hdfs://CentOS.Hadoop:8020/user/training/ingested_csv/outputHugeData/part-00127:0+3349315
.PythonRunner: Times: total = 535, boot = -44, init = 65, finish = 514
or.Executor: Finished task 129.0 in stage 0.0 (TID 129). 3128 bytes result sent to driver
ler.TaskSetManager: Starting task 130.0 in stage 0.0 (TID 130, localhost, partition 130,ANY, 2229 bytes)
or.Executor: Running task 130.0 in stage 0.0 (TID 130)
ler.TaskSetManager: Finished task 129.0 in stage 0.0 (TID 129) in 580 ms on localhost (130/806)
doopRDD: Input split: hdfs://CentOS.Hadoop:8020/user/training/ingested_csv/outputHugeData/part-00128:0+3350365
.PythonRunner: Times: total = 629, boot = -38, init = 58, finish = 609
or.Executor: Finished task 130.0 in stage 0.0 (TID 130). 3128 bytes result sent to driver
ler.TaskSetManager: Starting task 131.0 in stage 0.0 (TID 131, localhost, partition 131,ANY, 2229 bytes)
ler.TaskSetManager: Finished task 130.0 in stage 0.0 (TID 130) in 705 ms on localhost (131/806)
or.Executor: Running task 131.0 in stage 0.0 (TID 131)
doopRDD: Input split: hdfs://CentOS.Hadoop:8020/user/training/ingested_csv/outputHugeData/part-00129:0+3350830
.PythonRunner: Times: total = 733, boot = -70, init = 94, finish = 709
or.Executor: Finished task 131.0 in stage 0.0 (TID 131). 3128 bytes result sent to driver
ler.TaskSetManager: Starting task 132.0 in stage 0.0 (TID 132, localhost, partition 132,ANY, 2229 bytes)
or.Executor: Running task 132.0 in stage 0.0 (TID 132)
ler.TaskSetManager: Finished task 131.0 in stage 0.0 (TID 131) in 798 ms on localhost (132/806)
doopRDD: Input split: hdfs://CentOS.Hadoop:8020/user/training/ingested_csv/outputHugeData/part-00130:0+3351520
.PythonRunner: Times: total = 486, boot = -41, init = 65, finish = 462
or.Executor: Finished task 132.0 in stage 0.0 (TID 132). 3128 bytes result sent to driver
ler.TaskSetManager: Starting task 133.0 in stage 0.0 (TID 133, localhost, partition 133,ANY, 2229 bytes)
or.Executor: Running task 133.0 in stage 0.0 (TID 133)
ler.TaskSetManager: Finished task 132.0 in stage 0.0 (TID 132) in 548 ms on localhost (133/806)
doopRDD: Input split: hdfs://CentOS.Hadoop:8020/user/training/ingested_csv/outputHugeData/part-00131:0+3350711
.PythonRunner: Times: total = 535, boot = -57, init = 75, finish = 517
or.Executor: Finished task 133.0 in stage 0.0 (TID 133). 3128 bytes result sent to driver
ler.TaskSetManager: Starting task 134.0 in stage 0.0 (TID 134, localhost, partition 134,ANY, 2229 bytes)
ler.TaskSetManager: Finished task 133.0 in stage 0.0 (TID 133) in 607 ms on localhost (134/806)
or.Executor: Running task 134.0 in stage 0.0 (TID 134)
doopRDD: Input split: hdfs://CentOS.Hadoop:8020/user/training/ingested_csv/outputHugeData/part-00132:0+3351645
.PythonRunner: Times: total = 512, boot = -52, init = 69, finish = 495
or.Executor: Finished task 134.0 in stage 0.0 (TID 134). 3128 bytes result sent to driver
ler.TaskSetManager: Starting task 135.0 in stage 0.0 (TID 135, localhost, partition 135,ANY, 2229 bytes)
ler.TaskSetManager: Finished task 134.0 in stage 0.0 (TID 134) in 593 ms on localhost (135/806)
or.Executor: Running task 135.0 in stage 0.0 (TID 135)
doopRDD: Input split: hdfs://CentOS.Hadoop:8020/user/training/ingested_csv/outputHugeData/part-00133:0+3350224
.PythonRunner: Times: total = 547, boot = -80, init = 135, finish = 492
```



```
2016-11-29 15:31:40,044 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.3.0-cdh5.0.6  0.16.0-SNAPSHOT training        2016-11-29 15:24:38     2016-11-29 15:31:40     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   Media
nReducetime     Alias   Feature Outputs
job_1480363508295_0008  3       0       314     242     290     314     0       0       0       0       loadJson,loadJson2      MAP_ONLY    /
user/training/pig_csv/pigJsonCsv,

Input(s):
Successfully read 2000000 records (321673307 bytes) from: "/user/training/json_log_data"

Output(s):
Successfully stored 2000000 records (135638389 bytes) in: "/user/training/pig_csv/pigJsonCsv"

Counters:
Total records written : 2000000
Total bytes written : 135638389
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1480363508295_0008


2016-11-29 15:31:40,076 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2016-11-29 15:31:40,111 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=
SUCCEEDED. Redirecting to job history server
2016-11-29 15:31:41,263 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2016-11-29 15:31:41,927 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=
SUCCEEDED. Redirecting to job history server
2016-11-29 15:31:43,128 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2016-11-29 15:31:43,216 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=
SUCCEEDED. Redirecting to job history server
2016-11-29 15:31:44,968 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> quit
2016-11-29 15:58:09,757 [main] INFO  org.apache.pig.Main - Pig script completed in 34 minutes, 23 seconds and 482 milliseconds (2063482 ms)
```

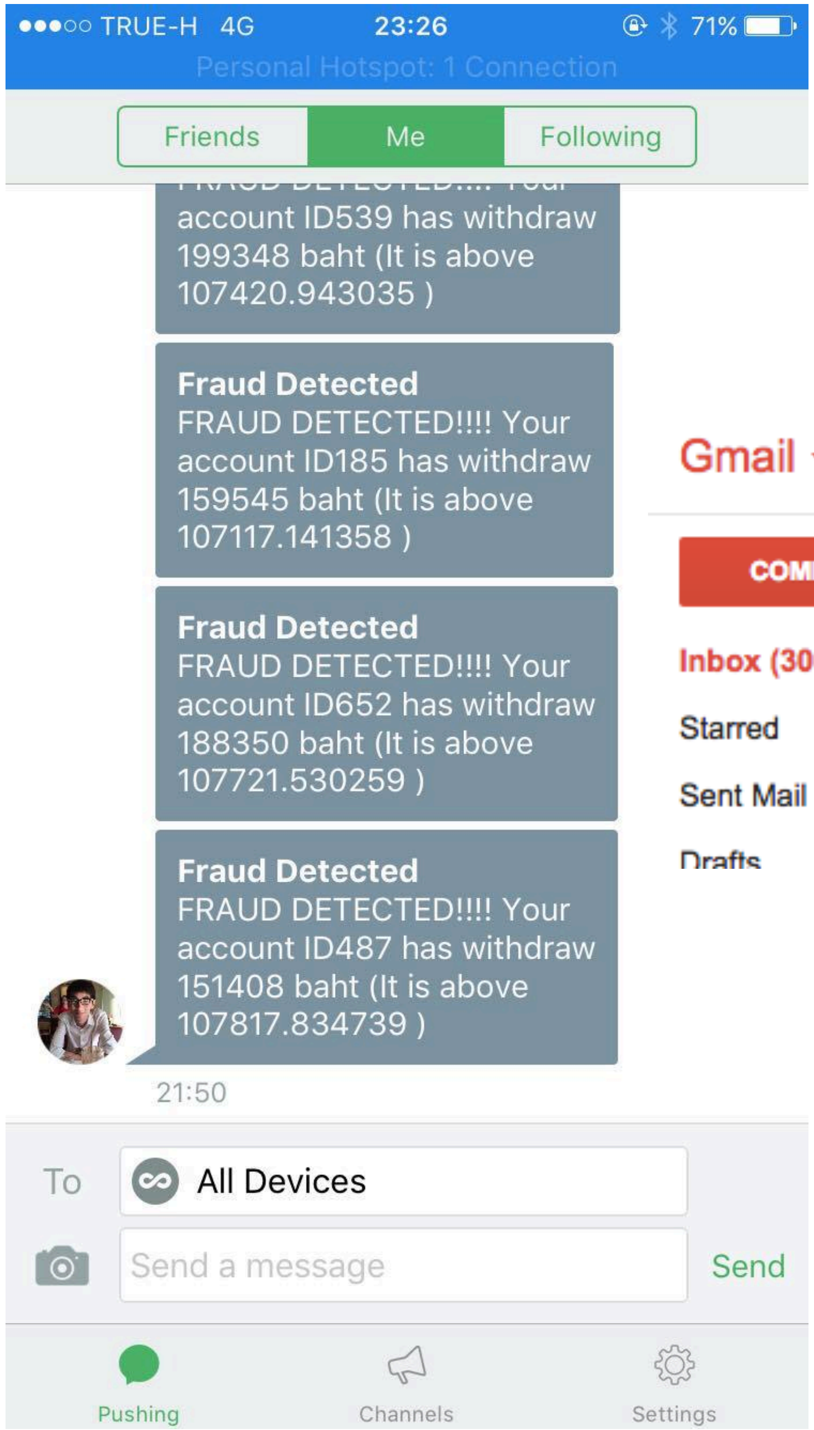


```
ved output of task 'attempt_201611271119_0000_m_000179_179' to hdfs://CentOS.Hadoop:8020
1271119_0000_m_000179
attempt_201611271119_0000_m_000179_179: Committed
sk 179.0 in stage 0.0 (TID 179). 2108 bytes result sent to driver
ting task 180.0 in stage 0.0 (TID 180, localhost, partition 180,ANY, 2226 bytes)
k 180.0 in stage 0.0 (TID 180)
shed task 179.0 in stage 0.0 (TID 179) in 4679 ms on localhost (180/400)
fs://CentOS.Hadoop:8020/user/training/huge_json_data/train_data_1.json:1442351880+801306

tal = 4427, boot = -41, init = 43, finish = 4425
ved output of task 'attempt_201611271119_0000_m_000180_180' to hdfs://CentOS.Hadoop:8020
1271119_0000_m_000180
attempt_201611271119_0000_m_000180_180: Committed
sk 180.0 in stage 0.0 (TID 180). 2108 bytes result sent to driver
ting task 181.0 in stage 0.0 (TID 181, localhost, partition 181,ANY, 2226 bytes)
k 181.0 in stage 0.0 (TID 181)
shed task 180.0 in stage 0.0 (TID 180) in 4487 ms on localhost (181/400)
fs://CentOS.Hadoop:8020/user/training/huge_json_data/train_data_1.json:1450364946+801306

tal = 4386, boot = -47, init = 49, finish = 4384
ved output of task 'attempt_201611271119_0000_m_000181_181' to hdfs://CentOS.Hadoop:8020
1271119_0000_m_000181
attempt_201611271119_0000_m_000181_181: Committed
sk 181.0 in stage 0.0 (TID 181). 2108 bytes result sent to driver
ting task 182.0 in stage 0.0 (TID 182, localhost, partition 182,ANY, 2226 bytes)
k 182.0 in stage 0.0 (TID 182)
shed task 181.0 in stage 0.0 (TID 181) in 4435 ms on localhost (182/400)
fs://CentOS.Hadoop:8020/user/training/huge_json_data/train_data_1.json:1458378012+801306

tal = 4462, boot = -28, init = 31, finish = 4459
ved output of task 'attempt_201611271119_0000_m_000182_182' to hdfs://CentOS.Hadoop:8020
1271119_0000_m_000182
attempt_201611271119_0000_m_000182_182: Committed
sk 182.0 in stage 0.0 (TID 182). 2108 bytes result sent to driver
ting task 183.0 in stage 0.0 (TID 183, localhost, partition 183,ANY, 2226 bytes)
k 183.0 in stage 0.0 (TID 183)
shed task 182.0 in stage 0.0 (TID 182) in 4514 ms on localhost (183/400)
fs://CentOS.Hadoop:8020/user/training/huge_json_data/train_data_1.json:1466391078+801306
```



```
s[localhos
Query: sho
+---------
| name
+---------
| historicaltransactions |
+------------------------+
Returned 1 row(s) in 0.18s
[localhost.localdomain:21000] > select * from historicaltransactions LIMIT 2;
Query: select * from historicaltransactions LIMIT 2
+----+----------------------+-----------------+--------+----------+--------+------------+
| id | transaction_date     | channel         | userid | action   | amount | state      |
+----+----------------------+-----------------+--------+----------+--------+------------+
| 1  | 2015-02-09T15:20:04Z | MobileBanking   | 706    | withdraw | 42006  | California |
| 2  | 2015-02-09T15:20:08Z | InternetBanking | 117    | withdraw | 55389  | Washington |
+----+----------------------+-----------------+--------+----------+--------+------------+
ERRORS:
Backend 0:Unknown disk id.  This will negatively affect performance. Check your hdfs settings to enable block location metadata.

Returned 2 row(s) in 23.10s
[localhost.localdomain:21000] > select COUNT(*) from historicaltransactions;
Query: select COUNT(*) from historicaltransactions
ERRORS:
Backend 0:Unknown disk id.  This will negatively affect performance. Check your hdfs settings to enable block location metadata.

+----------+
| count(*) |
+----------+
| 20000000 |
+----------+
ERRORS:
Backend 0:Unknown disk id.  This will negatively affect performance. Check your hdfs settings to enable block location metadata.

Returned 1 row(s) in 515.76s
[localhost.localdomain:21000] >
```

Result/Demo



FRAUD DETECTED!!!! Your account ID539 has withdraw 199348 baht (It is above 107420.943035)

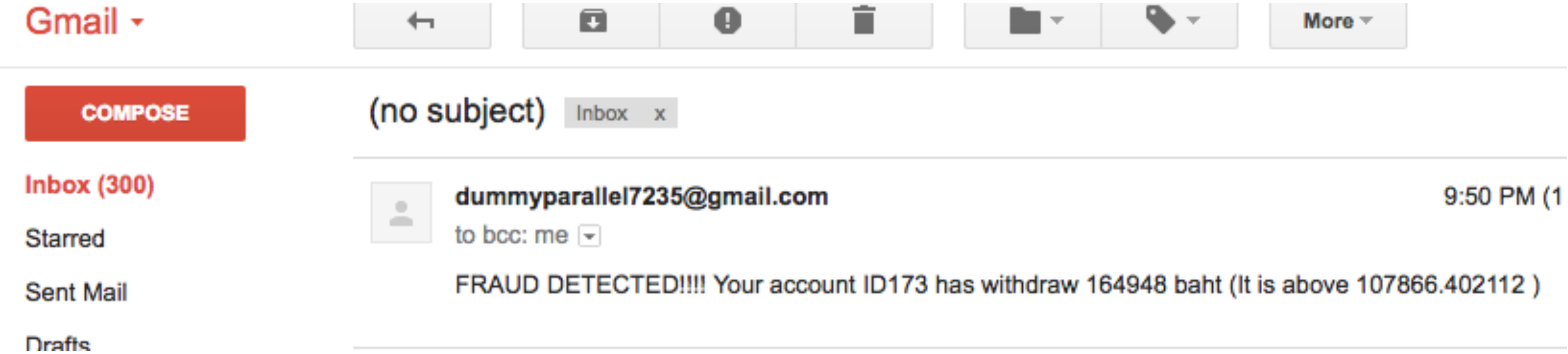
Fraud Detected
FRAUD DETECTED!!!! Your account ID185 has withdraw 159545 baht (It is above 107117.141358)

Fraud Detected
FRAUD DETECTED!!!! Your account ID652 has withdraw 188350 baht (It is above 107721.530259)

Fraud Detected
FRAUD DETECTED!!!! Your account ID487 has withdraw 151408 baht (It is above 107817.834739)

21:50

To: All Devices



(no subject) Inbox x

dummyparallel7235@gmail.com 9:50 PM (1

to bcc: me

FRAUD DETECTED!!!! Your account ID173 has withdraw 164948 baht (It is above 107866.402112)