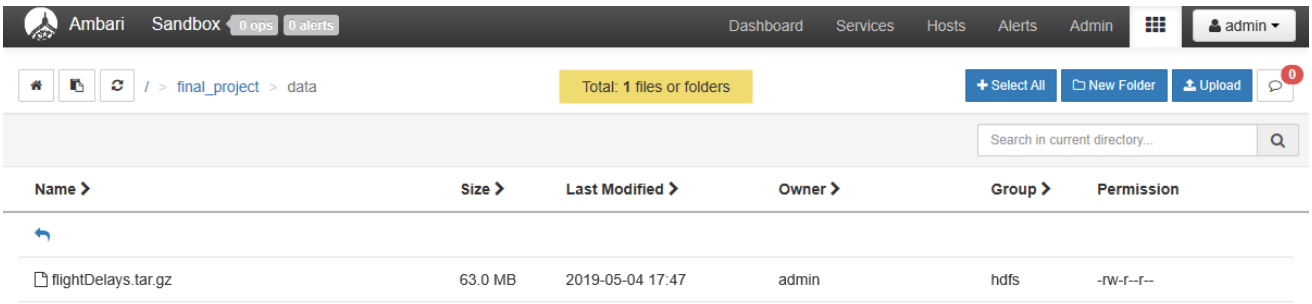
STSCI 5065 Final Project

Franklin Zhao (qz297)

05/06/2019

1.

```
hadoop fs -mkdir /final_project
hadoop fs -mkdir /final_project/data
```



2.

```
mkdir /FP
hadoop fs -copyToLocal /final_project/data/flightDelays.tar.gz
/FP/flightDelays.tar.gz
cd /FP
tar -xzf /FP/flightDelays.tar.gz
rm flightDelays.tar.gz
ls –lh
```

```
[root@sandbox FP]# ls -lh
total 764M
-rw-r--r-- 1 1276589696 11103514 64M Apr  3  2015 flightDelays_10.csv
-rw-r--r-- 1 1276589696 11103514 60M Apr  3  2015 flightDelays_11.csv
-rw-r--r-- 1 1276589696 11103514 63M Apr  3  2015 flightDelays_12.csv
-rw-r--r-- 1 1276589696 11103514 61M Apr  3  2015 flightDelays_1.csv
-rw-r--r-- 1 1276589696 11103514 57M Apr  3  2015 flightDelays_2.csv
-rw-r--r-- 1 1276589696 11103514 67M Apr  3  2015 flightDelays_3.csv
-rw-r--r-- 1 1276589696 11103514 65M Apr  3  2015 flightDelays_4.csv
-rw-r--r-- 1 1276589696 11103514 66M Apr  3  2015 flightDelays_5.csv
-rw-r--r-- 1 1276589696 11103514 67M Apr  3  2015 flightDelays_6.csv
-rw-r--r-- 1 1276589696 11103514 70M Apr  3  2015 flightDelays_7.csv
-rw-r--r-- 1 1276589696 11103514 68M Apr  3  2015 flightDelays_8.csv
-rw-r--r-- 1 1276589696 11103514 61M Apr  3  2015 flightDelays_9.csv
```

3.

approach 1: using –copyFromLocal in CLI

```
hadoop fs -copyFromLocal /FP/flightDelays* /final_project/data
```

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin admin

/ > final_project > data

Total: 13 files or folders

+ Select All | New Folder | Upload

Search in current directory...

Name	Size	Last Modified	Owner	Group	Permission
flightDelays.tar.gz	63.0 MB	2019-05-04 17:47	admin	hdfs	-rw-r--r--
flightDelays_1.csv	60.8 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_10.csv	63.6 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_11.csv	59.8 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_12.csv	62.8 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_2.csv	56.0 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_3.csv	66.0 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_4.csv	64.4 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_5.csv	65.7 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_6.csv	66.9 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_7.csv	69.1 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_8.csv	67.5 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--
flightDelays_9.csv	60.7 MB	2019-05-04 18:52	root	hdfs	-rw-r--r--

approach 2: using -put in CLI

```
hadoop fs -put /FP/flightDelays* /final_project/data
```

Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin admin

/ > final_project > data

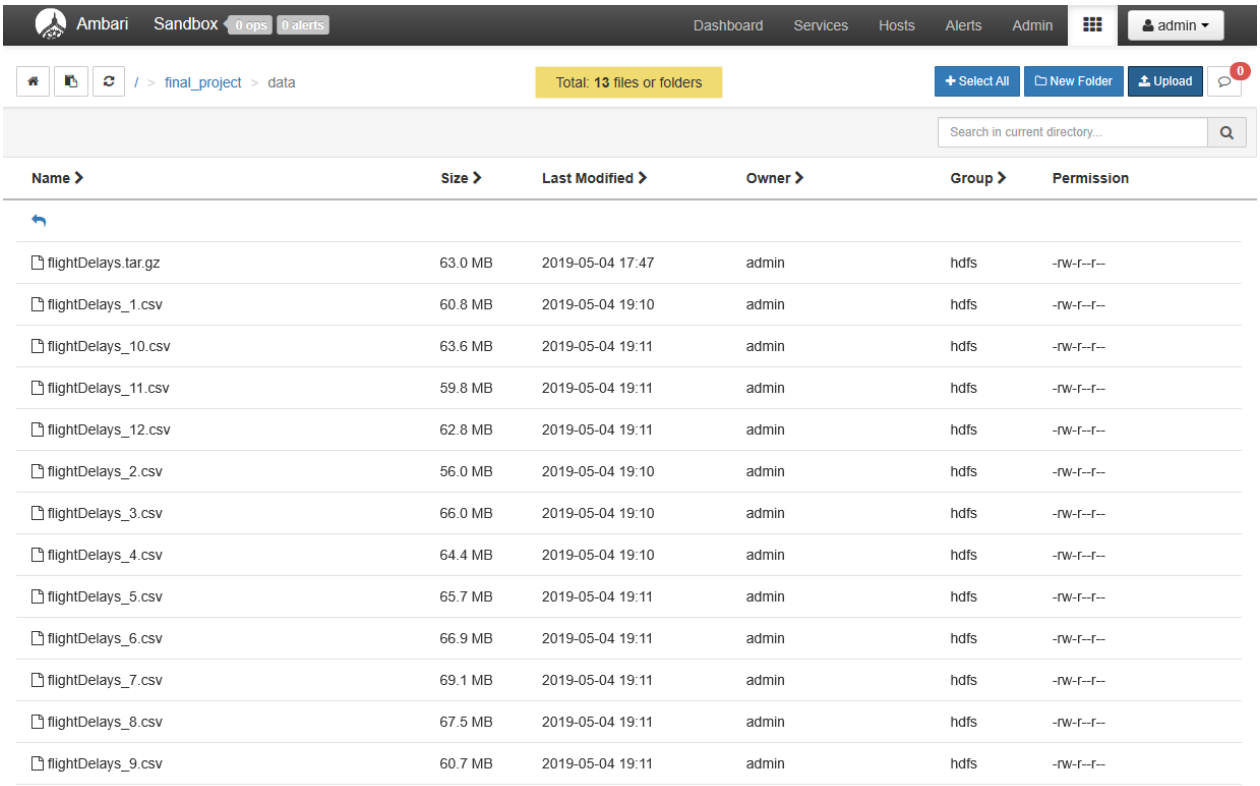
Total: 13 files or folders

+ Select All | New Folder | Upload

Search in current directory...

Name	Size	Last Modified	Owner	Group	Permission
flightDelays.tar.gz	63.0 MB	2019-05-04 17:47	admin	hdfs	-rw-r--r--
flightDelays_1.csv	60.8 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_10.csv	63.6 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_11.csv	59.8 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_12.csv	62.8 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_2.csv	56.0 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_3.csv	66.0 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_4.csv	64.4 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_5.csv	65.7 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_6.csv	66.9 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_7.csv	69.1 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_8.csv	67.5 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--
flightDelays_9.csv	60.7 MB	2019-05-04 18:55	root	hdfs	-rw-r--r--

approach 3: load the files using Files View in Ambari



Name	Size	Last Modified	Owner	Group	Permission
flightDelays.tar.gz	63.0 MB	2019-05-04 17:47	admin	hdfs	-rw-r--r--
flightDelays_1.csv	60.8 MB	2019-05-04 19:10	admin	hdfs	-rw-r--r--
flightDelays_10.csv	63.6 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--
flightDelays_11.csv	59.8 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--
flightDelays_12.csv	62.8 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--
flightDelays_2.csv	56.0 MB	2019-05-04 19:10	admin	hdfs	-rw-r--r--
flightDelays_3.csv	66.0 MB	2019-05-04 19:10	admin	hdfs	-rw-r--r--
flightDelays_4.csv	64.4 MB	2019-05-04 19:10	admin	hdfs	-rw-r--r--
flightDelays_5.csv	65.7 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--
flightDelays_6.csv	66.9 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--
flightDelays_7.csv	69.1 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--
flightDelays_8.csv	67.5 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--
flightDelays_9.csv	60.7 MB	2019-05-04 19:11	admin	hdfs	-rw-r--r--

4.

```
hive

CREATE DATABASE FPdb
LOCATION '/final_project';

DESCRIBE DATABASE FPdb;
```

```
hive> CREATE DATABASE FPdb
    > LOCATION '/final_project';
OK
Time taken: 0.131 seconds
hive> DESCRIBE DATABASE FPdb;
OK
fpdb            hdfs://sandbox.hortonworks.com:8020/final_project       root    USER
Time taken: 0.313 seconds, Fetched: 1 row(s)
```

5.

```
flightDelays = LOAD '/final_project/data/flightDelays_*'
                USING org.apache.pig.piggybank.storage.CSVExcelStorage(',')
                AS (YEAR: INT,
                    FL_DATE: CHARARRAY,
                    UNIQUE_CARRIER: CHARARRAY,
                    CARRIER: CHARARRAY,
                    FL_NUM: CHARARRAY,
                    ORIGIN_AIRPORT_ID: CHARARRAY,
```

ORIGIN: CHARARRAY,
ORIGIN_CITY_NAME: CHARARRAY,
ORIGIN_STATE_ABR: CHARARRAY,
DEST_AIRPORT_ID: CHARARRAY,
DEST: CHARARRAY,
DEST_CITY_NAME: CHARARRAY,
DEST_STATE_ABR: CHARARRAY,
DEP_DELAY_NEW: FLOAT,
ARR_DELAY: FLOAT,
ARR_DELAY_NEW: FLOAT,
CARRIER_DELAY: FLOAT,
WEATHER_DELAY: FLOAT,
NAS_DELAY: FLOAT,
SECURITY_DELAY: FLOAT,
LATE_AIRCRAFT_DELAY: FLOAT);

flightDelays - COMPLETED

Job ID	job_1557004298829_0004
Started	2019-05-04 20:04

Results Download

Logs Download

```
19/05/05 00:04:43 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/05/05 00:04:43 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/05/05 00:04:43 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
19/05/05 00:04:43 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
19/05/05 00:04:43 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2019-05-05 00:04:43,972 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.1.0-129 (rexported) compiled May 31 26
2019-05-05 00:04:43,972 [main] INFO  org.apache.pig.Main - Logging error messages to: /hadoop/yarn/local/usercache/admin/appcac
2019-05-05 00:04:44,609 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/yarn/.pigbootup not found
2019-05-05 00:04:44,761 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file
2019-05-05 00:04:45,223 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-script.pig-ecd2de4b-5f2b-42
2019-05-05 00:04:45,562 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http
2019-05-05 00:04:45,669 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2019-05-05 00:04:46,440 [main] INFO  org.apache.pig.Main - Pig script completed in 2 seconds and 643 milliseconds (2643 ms)
```

6.

```
grpd = GROUP flightDelays ALL;
averageDelays = FOREACH grpd GENERATE ROUND_TO(AVG(flightDelays.CARRIER_DELAY), 2),
                                      ROUND_TO(AVG(flightDelays.WEATHER_DELAY), 2),
                                      ROUND_TO(AVG(flightDelays.NAS_DELAY), 2),
                                      ROUND_TO(AVG(flightDelays.SECURITY_DELAY), 2),
                                      ROUND_TO(AVG(flightDelays.LATE_AIRCRAFT_DELAY), 2);
DUMP averageDelays;
```

averageDelays - COMPLETED

Job ID	job_1557004298829_0013
Started	2019-05-04 20:42

Results Download

```
(16.65,2.34,13.73,0.08,23.87)
```

Logs Download

```
19/05/05 00:43:04 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/05/05 00:43:04 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/05/05 00:43:04 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
19/05/05 00:43:04 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
19/05/05 00:43:04 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2019-05-05 00:43:04,307 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.1.0-129 (rexported) compiled May 31
2019-05-05 00:43:04,308 [main] INFO  org.apache.pig.Main - Logging error messages to: /hadoop/yarn/local/usercache/admin/app
2019-05-05 00:43:04,996 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/yarn/.pigbootup not found
2019-05-05 00:43:05,152 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
2019-05-05 00:43:05,625 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-script.pig-6f426cd8-716d-
2019-05-05 00:43:05,960 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: h
2019-05-05 00:43:06,063 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2019-05-05 00:43:07,324 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2019-05-05 00:43:07,368 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not
2019-05-05 00:43:07,396 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForE
2019-05-05 00:43:07,460 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 69
2019-05-05 00:43:07,550 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory i
```

7.

```
grpd = GROUP flightDelays ALL;
maxDelays = FOREACH grpd GENERATE MAX(flightDelays.CARRIER_DELAY),
                                  MAX(flightDelays.WEATHER_DELAY),
                                  MAX(flightDelays.NAS_DELAY),
                                  MAX(flightDelays.SECURITY_DELAY),
                                  MAX(flightDelays.LATE_AIRCRAFT_DELAY);
DUMP maxDelays;
```

longestDelays - COMPLETED

Job ID	job_1557004298829_0065
Started	2019-05-05 14:20

Results Download

```
(1975.0,1591.0,1287.0,573.0,1182.0)
```

Logs Download

```
19/05/05 18:20:18 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/05/05 18:20:18 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/05/05 18:20:18 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
19/05/05 18:20:18 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
19/05/05 18:20:18 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2019-05-05 18:20:18,793 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.1.0-129 (rexported) compiled May 31 2017
2019-05-05 18:20:18,793 [main] INFO  org.apache.pig.Main - Logging error messages to: /hadoop/yarn/local/usercache/admin/appcache
2019-05-05 18:20:19,574 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/yarn/.pigbootup not found
2019-05-05 18:20:19,742 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file s
2019-05-05 18:20:20,248 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-script.pig-00f640f6-3b79-45ad-
2019-05-05 18:20:20,637 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http:/
2019-05-05 18:20:20,767 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2019-05-05 18:20:21,856 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2019-05-05 18:20:21,915 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not gener
2019-05-05 18:20:21,946 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach,
2019-05-05 18:20:22,020 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 6994001
2019-05-05 18:20:22,126 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory is /tm
```

8.

```
vi /FP/flight_delay_udf.py
```

```
@outputSchema("res:chararray")
def get_max(data):
        res = 'The maximum CARRIER_DELAY is 1975.0. ' + \
                'The details of the delay are as follows:\n'
        for i in range(len(data)-1):
                res += data[i][0] + ': ' + str(data[i][1]) + ',\n'
        res += data[-1][0] + ': ' + str(data[-1][1]) + '.\n'
        return res

REGISTER '/FP/flight_delay_udf.py' USING jython AS myudf;
joined = JOIN flightDelays BY CARRIER_DELAY, maxDelays BY $0;
res = FOREACH joined GENERATE myudf.get_max((('YEAR', $0),
                                                ('FL_DATE', $1),
                                                ('UNIQUE_CARRIER', $2),
                                                ('CARRIER', $3),
                                                ('FL_NUM', $4),
                                                ('ORIGIN_AIRPORT_ID', $5),
                                                ('ORIGIN', $6),
                                                ('ORIGIN_CITY_NAME', $7),
                                                ('ORIGIN_STATE_ABR', $8),
                                                ('DEST_AIRPORT_ID', $9),
                                                ('DEST', $10),
                                                ('DST_CITY_NAME', $11),
                                                ('DEST_STATE_ABR', $12),
                                                ('DEP_DELAY_NEW', $13),
                                                ('ARR_DELAY', $14),
                                                ('ARR_DELAY_NEW', $15),
                                                ('CARRIER_DELAY', $16),
                                                ('WEATHER_DELAY', $17),
                                                ('NAS_DELAY', $18),
                                                ('SECURITY_DELAY', $19),
                                                ('LATE_AIRCRAFT_DELAY', $20)));
DUMP res;
```

flight_delays_udf - COMPLETED

Job ID	job_1557004298829_0037
Started	2019-05-04 23:48

Results

Download

```
(The maximum CARRIER_DELAY is 1975.0. The details of the delay are as follows:
YEAR: 2013,
FL_DATE: 2013-12-26,
UNIQUE_CARRIER: AA,
CARRIER: AA,
FL_NUM: 1202,
ORIGIN_AIRPORT_ID: 13891,
ORIGIN: ONT,
ORIGIN_CITY_NAME: Ontario, CA,
ORIGIN_STATE_ABR: CA,
DEST_AIRPORT_ID: 11298,
DEST: DFW,
DST_CITY_NAME: Dallas/Fort Worth, TX,
DEST_STATE_ABR: TX,
DEP_DELAY_NEW: 1975.0,
ARR_DELAY: 1983.0,
ARR_DELAY_NEW: 1983.0,
CARRIER_DELAY: 1975.0,
WEATHER_DELAY: 0.0,
NAS_DELAY: 8.0,
SECURITY_DELAY: 0.0,
LATE_AIRCRAFT_DELAY: 0.0.
)
```

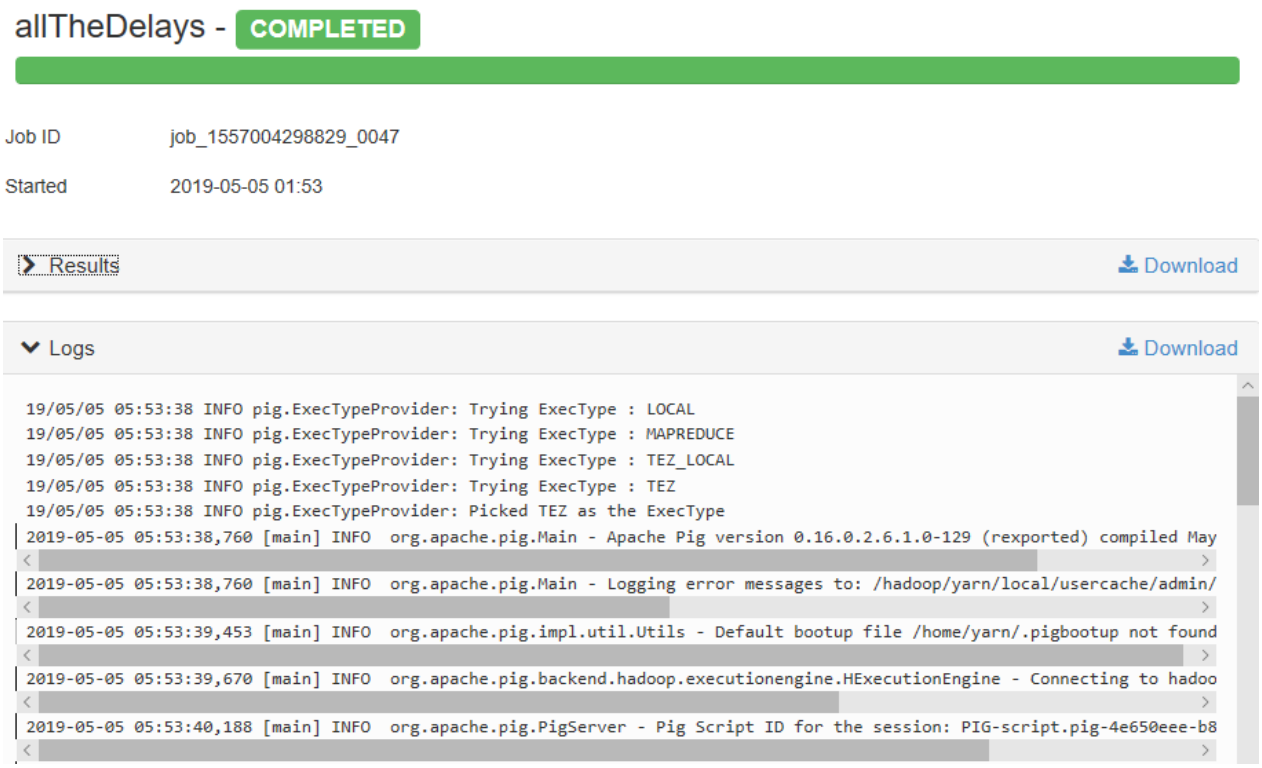
Logs

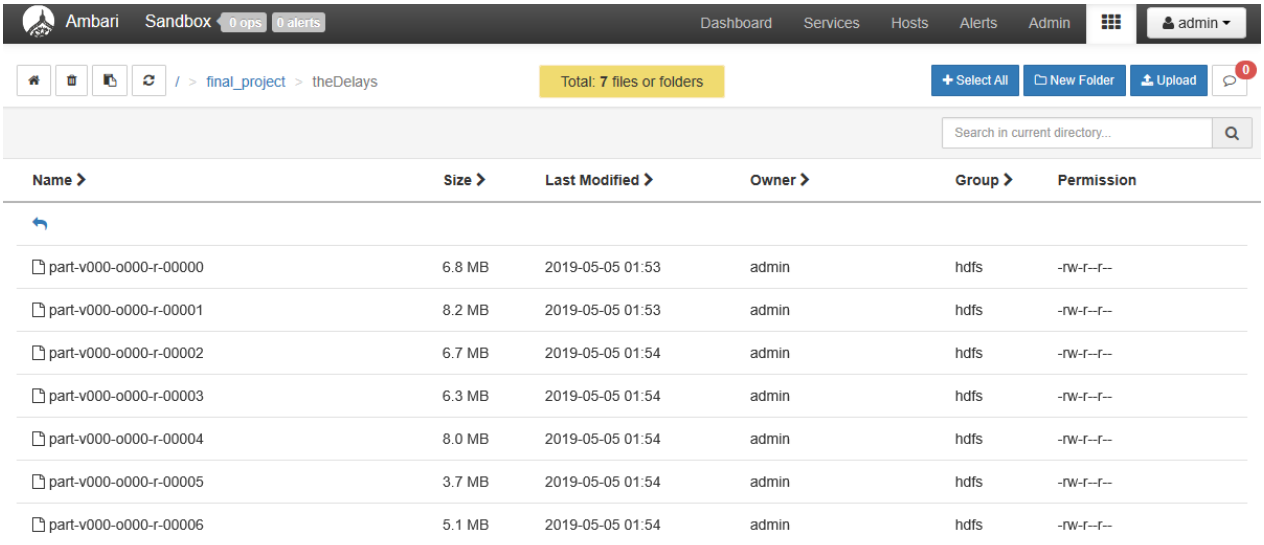
Download

```
19/05/05 03:48:18 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/05/05 03:48:18 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/05/05 03:48:18 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
19/05/05 03:48:18 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
19/05/05 03:48:18 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2019-05-05 03:48:19,028 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.1.0-129 (rexported) compiled May 31
2019-05-05 03:48:19,028 [main] INFO  org.apache.pig.Main - Logging error messages to: /hadoop/yarn/local/usercache/admin/app
2019-05-05 03:48:19,736 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/yarn/.pigbootup not found
2019-05-05 03:48:19,917 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop f
```

9.

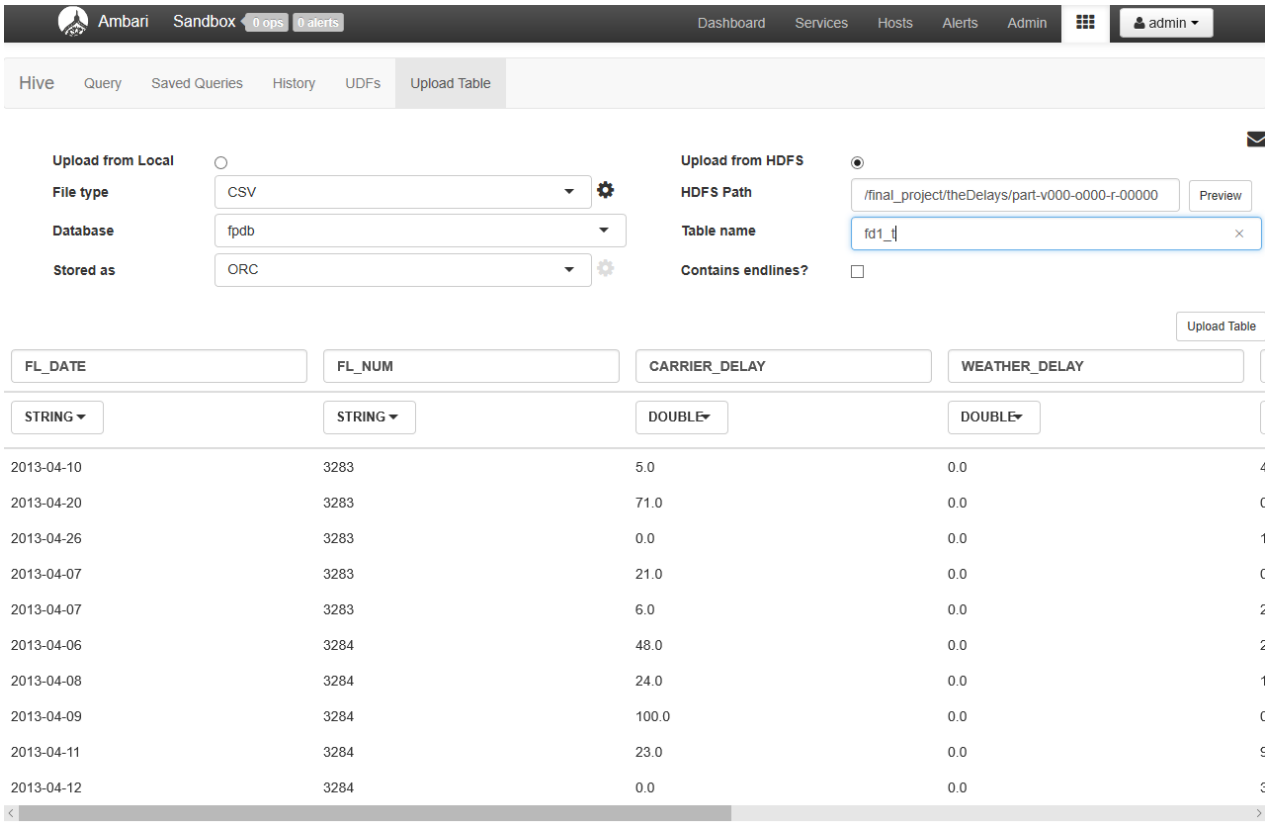
```
allTheDelays = FOREACH flightDelays GENERATE FL_DATE, FL_NUM, CARRIER_DELAY,
                                            WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY,
                                            LATE_AIRCRAFT_DELAY;
theDelays = FILTER allTheDelays BY (CARRIER_DELAY IS NOT NULL AND
                                    WEATHER_DELAY IS NOT NULL AND
                                    NAS_DELAY IS NOT NULL AND
                                    SECURITY_DELAY IS NOT NULL AND
                                    LATE_AIRCRAFT_DELAY IS NOT NULL);
STORE theDelays INTO '/final_project/theDelays';
```





10.

a.



fd1_t.fl_date	fd1_t.fl_num	fd1_t.carrier_delay	fd1_t.weather_delay	fd1_t.nas_delay	fd1_t.security_delay	fd1_t.late_aircra
2013-04-10	3283	5.0	0.0	45.0	0.0	16.0
2013-04-20	3283	71.0	0.0	0.0	0.0	5.0
2013-04-26	3283	0.0	0.0	17.0	0.0	0.0
2013-04-07	3283	21.0	0.0	0.0	0.0	0.0
2013-04-07	3283	6.0	0.0	2.0	0.0	7.0
2013-04-06	3284	48.0	0.0	23.0	0.0	0.0
2013-04-08	3284	24.0	0.0	10.0	0.0	0.0
2013-04-09	3284	100.0	0.0	0.0	0.0	0.0
2013-04-11	3284	23.0	0.0	9.0	0.0	0.0
2013-04-12	3284	0.0	0.0	35.0	0.0	0.0

fd2_t.fl_date	fd2_t.fl_num	fd2_t.carrier_delay	fd2_t.weather_delay	fd2_t.nas_delay	fd2_t.security_delay	fd2_t.late_aircra
2013-06-21	3189	0.0	0.0	106.0	0.0	5.0
2013-06-25	3189	0.0	0.0	20.0	0.0	137.0
2013-06-27	3189	0.0	0.0	21.0	0.0	0.0
2013-06-28	3189	0.0	0.0	38.0	0.0	80.0
2013-06-29	3189	51.0	0.0	0.0	0.0	22.0
2013-06-13	3191	0.0	0.0	0.0	0.0	36.0
2013-06-18	3191	0.0	0.0	0.0	0.0	58.0
2013-06-21	3191	0.0	0.0	302.0	0.0	0.0
2013-06-25	3191	0.0	0.0	32.0	0.0	22.0
2013-06-26	3191	0.0	0.0	54.0	0.0	88.0

b.

```
CREATE TABLE IF NOT EXISTS FPdb.fd3_t (
  FL_DATE STRING,
  FL_NUM STRING,
  CARRIER_DELAY DOUBLE,
  WEATHER_DELAY DOUBLE,
  NAS_DELAY DOUBLE,
  SECURITY_DELAY DOUBLE,
  LATE_AIRCRAFT_DELAY DOUBLE)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LINES TERMINATED BY '\n';
LOAD DATA INPATH '/final_project/theDelays/part-v000-o000-r-00002'
OVERWRITE INTO TABLE FPdb.fd3_t;
SELECT * FROM FPdb.fd3_t LIMIT 10;
```

fd3_t.fl_date	fd3_t.fl_num	fd3_t.carrier_delay	fd3_t.weather_delay	fd3_t.nas_delay	fd3_t.security_delay	fd3_t.late_aircra
2013-08-07	3283	0.0	0.0	27.0	0.0	0.0
2013-08-19	3283	0.0	0.0	0.0	0.0	137.0
2013-08-20	3283	0.0	0.0	17.0	0.0	0.0
2013-08-24	3284	133.0	0.0	0.0	0.0	5.0
2013-08-01	3284	4.0	0.0	0.0	0.0	74.0
2013-08-06	3284	0.0	0.0	67.0	0.0	0.0
2013-08-07	3284	0.0	0.0	61.0	0.0	74.0
2013-08-09	3284	0.0	0.0	24.0	0.0	2.0
2013-08-24	3284	0.0	0.0	0.0	0.0	100.0
2013-08-03	3285	3.0	0.0	0.0	0.0	23.0

c.

```
CREATE TABLE IF NOT EXISTS FPdb.fd4_t
LIKE FPdb.fd3_t;
LOAD DATA INPATH '/final_project/theDelays/part-v000-o000-r-00003'
OVERWRITE INTO TABLE FPdb.fd4_t;

CREATE TABLE IF NOT EXISTS FPdb.fd5_t
LIKE FPdb.fd3_t;
LOAD DATA INPATH '/final_project/theDelays/part-v000-o000-r-00004'
OVERWRITE INTO TABLE FPdb.fd5_t;

CREATE TABLE IF NOT EXISTS FPdb.fd6_t
LIKE FPdb.fd3_t;
LOAD DATA INPATH '/final_project/theDelays/part-v000-o000-r-00005'
OVERWRITE INTO TABLE FPdb.fd6_t;

CREATE TABLE IF NOT EXISTS FPdb.fd7_t
LIKE FPdb.fd3_t;
LOAD DATA INPATH '/final_project/theDelays/part-v000-o000-r-00006'
OVERWRITE INTO TABLE FPdb.fd7_t;
SELECT * FROM FPdb.fd7_t LIMIT 10;
```

fd7_t.fl_date	fd7_t.fl_num	fd7_t.carrier_delay	fd7_t.weather_delay	fd7_t.nas_delay	fd7_t.security_delay	fd7_t.late_aircra
2013-12-02	2900	0.0	0.0	0.0	0.0	36.0
2013-12-06	2900	10.0	0.0	0.0	0.0	11.0
2013-12-10	2900	0.0	0.0	4.0	0.0	83.0
2013-12-11	2900	24.0	0.0	15.0	0.0	26.0
2013-12-14	2900	0.0	0.0	17.0	0.0	13.0
2013-12-15	2900	0.0	0.0	0.0	0.0	45.0
2013-12-16	2900	0.0	0.0	0.0	0.0	82.0
2013-12-18	2900	6.0	0.0	0.0	0.0	17.0
2013-12-22	2900	0.0	0.0	0.0	0.0	46.0
2013-12-23	2900	26.0	0.0	17.0	0.0	4.0

11.

```
CREATE TABLE IF NOT EXISTS FPdb.fd_t AS
SELECT * FROM FPdb.fd1_t
UNION ALL
SELECT * FROM FPdb.fd2_t
UNION ALL
SELECT * FROM FPdb.fd3_t
UNION ALL
SELECT * FROM FPdb.fd4_t
```

```
UNION ALL
SELECT * FROM FPdb.fd5_t
UNION ALL
SELECT * FROM FPdb.fd6_t
UNION ALL
SELECT * FROM FPdb.fd7_t;

DESCRIBE FORMATTED FPdb.fd_t;
```

```
hive> DESCRIBE FORMATTED FPdb.fd_t;
OK
# col_name              data_type               comment

fl_date                 string
fl_num                  string
carrier_delay           double
weather_delay           double
nas_delay               double
security_delay          double
late_aircraft_delay     double

# Detailed Table Information
Database:               fpdb
Owner:                  admin
CreateTime:             Sun May 05 16:03:44 UTC 2019
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://sandbox.hortonworks.com:8020/final_project/fd_t
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        numFiles                7
        numRows                 1269277
        rawDataSize             45709400
        totalSize               46978677
        transient_lastDdlTime   1557072225

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.848 seconds, Fetched: 37 row(s)
```

12.

```
SELECT MAX(CARRIER_DELAY) max_carrier_delay,
       MAX(WEATHER_DELAY) max_weather_delay,
       MAX(NAS_DELAY) max_nas_delay,
       MAX(SECURITY_DELAY) max_security_delay,
       MAX(LATE_AIRCRAFT_DELAY) max_late_aircraft_delay
FROM FPdb.fd_t;
```

max_carrier_delay	max_weather_delay	max_nas_delay	max_security_delay	max_late_aircraft_delay
1975.0	1591.0	1287.0	573.0	1182.0

```
SELECT ROUND(AVG(CARRIER_DELAY), 2) mean_carrier_delay,
       ROUND(AVG(WEATHER_DELAY), 2) mean_weather_delay,
       ROUND(AVG(NAS_DELAY), 2) mean_nas_delay,
       ROUND(AVG(SECURITY_DELAY), 2) mean_security_delay,
       ROUND(AVG(LATE_AIRCRAFT_DELAY), 2) mean_late_aircraft_delay
FROM FPdb.fd_t;
```

mean_carrier_delay	mean_weather_delay	mean_nas_delay	mean_security_delay	mean_late_aircraft_delay
16.65	2.34	13.73	0.08	23.87

13.

```
vi /FP/FindMaxAverageDelayType.py
```

```
#!/usr/bin/python
import sys
def FindMaxAverageDelayType(carrier, weather, nas, security, late_aircraft):
        key = ['CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY',
                'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY']
        value = [carrier, weather, nas, security, late_aircraft]
        max_val = max(value)
        max_key = key[value.index(max_val)]
        return (max_key, str(max_val))
for line in sys.stdin:
        line = line.strip()
        carrier, weather, nas, security, late_aircraft = line.split('\t')
        max_key = FindMaxAverageDelayType(float(carrier), float(weather), float(nas),
                                          float(security), float(late_aircraft))[0]
        max_val = FindMaxAverageDelayType(float(carrier), float(weather), float(nas),
                                          float(security), float(late_aircraft))[1]
        print('The delay category with the longest average delay is ' + max_key +
              '; the average delay time is ' + max_val + ' minutes.')
```

```
CREATE VIEW averageDelays_v AS
 SELECT ROUND(AVG(CARRIER_DELAY), 2) mean_carrier_delay,
        ROUND(AVG(WEATHER_DELAY), 2) mean_weather_delay,
        ROUND(AVG(NAS_DELAY), 2) mean_nas_delay,
        ROUND(AVG(SECURITY_DELAY), 2) mean_security_delay,
        ROUND(AVG(LATE_AIRCRAFT_DELAY), 2) mean_late_aircraft_delay
 FROM FPdb.fd_t;
ADD FILE /FP/FindMaxAverageDelayType.py;
SELECT TRANSFORM(mean_carrier_delay, mean_weather_delay, mean_nas_delay,
                 mean_security_delay, mean_late_aircraft_delay)
USING 'python FindMaxAverageDelayType.py' AS maxAverageDelay
FROM averageDelays_v;
```

```
hive> CREATE VIEW averageDelays_v AS
    >  SELECT ROUND(AVG(CARRIER_DELAY), 2) mean_carrier_delay,
    >         ROUND(AVG(WEATHER_DELAY), 2) mean_weather_delay,
    >         ROUND(AVG(NAS_DELAY), 2) mean_nas_delay,
    >         ROUND(AVG(SECURITY_DELAY), 2) mean_security_delay,
    >         ROUND(AVG(LATE_AIRCRAFT_DELAY), 2) mean_late_aircraft_delay
    >  FROM FPdb.fd_t;
OK
Time taken: 2.519 seconds
hive> ADD FILE /FP/FindMaxAverageDelayType.py;
Added resources: [/FP/FindMaxAverageDelayType.py]
hive> SELECT TRANSFORM(mean_carrier_delay, mean_weather_delay, mean_nas_delay,
    >                  mean_security_delay, mean_late_aircraft_delay)
    > USING 'python FindMaxAverageDelayType.py' AS maxAverageDelay
    > FROM averageDelays_v;
Query ID = root_20190505180237_69209cf7-0e0b-4517-8302-3247d55d8e8d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1557004298829_0064)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.30 s
--------------------------------------------------------------------------------
OK
The delay category with the longest average delay is LATE_AIRCRAFT_DELAY; the average delay time is 23.87 minutes.
Time taken: 17.015 seconds, Fetched: 1 row(s)
```

14.

```
SELECT ROUND((COUNT(*) * SUM(WEATHER_DELAY * CARRIER_DELAY) -
                SUM(WEATHER_DELAY) * SUM(CARRIER_DELAY)) /
               SQRT((COUNT(*) * SUM(POW(WEATHER_DELAY, 2)) -
                     POW(SUM(WEATHER_DELAY), 2)) *
                    (COUNT(*) * SUM(POW(CARRIER_DELAY, 2)) -
                     POW(SUM(CARRIER_DELAY), 2))), 4)
        AS w_c
FROM FPdb.fd_t;
```

w_c
-0.0454

```
SELECT ROUND(CORR(WEATHER_DELAY, CARRIER_DELAY), 4) AS w_c FROM FPdb.fd_t;
```

w_c
-0.0454

```
SELECT ROUND(CORR(WEATHER_DELAY, CARRIER_DELAY), 4) AS w_c,
        ROUND(CORR(NAS_DELAY, CARRIER_DELAY), 4) AS n_c,
        ROUND(CORR(SECURITY_DELAY, CARRIER_DELAY), 4) AS s_c,
        ROUND(CORR(LATE_AIRCRAFT_DELAY, CARRIER_DELAY), 4) AS l_c,
        ROUND(CORR(NAS_DELAY, WEATHER_DELAY), 4) AS n_w,
        ROUND(CORR(SECURITY_DELAY, WEATHER_DELAY), 4) AS s_w,
        ROUND(CORR(LATE_AIRCRAFT_DELAY, WEATHER_DELAY), 4) AS l_w,
        ROUND(CORR(SECURITY_DELAY, NAS_DELAY), 4) AS s_n,
        ROUND(CORR(LATE_AIRCRAFT_DELAY, NAS_DELAY), 4) AS l_n,
        ROUND(CORR(LATE_AIRCRAFT_DELAY, SECURITY_DELAY), 4) AS l_s
FROM FPdb.fd_t;
```

w_c	n_c	s_c	l_c	n_w	s_w	l_w	s_n	l_n	l_s
-0.0454	-0.1142	-0.0103	-0.1217	-8.0E-4	-0.004	-0.0235	-0.0094	-0.1486	-0.0095

Comment: From the results we see that overall, the five delay categories are basically having little or almost no correlations. However, since all coefficients are less than 0, we may say even if small, all 10 possible pairs are somewhat negatively correlated. Among the results, NAS_DELAY and CARRIER_DELAY, LATE_AIRCRAFT_DELAY and CARRIER_DELAY, LATE_AIRCRAFT_DELAY and NAS_DELAY have the coefficients (absolute value) greater than -0.1 (l_n has the largest one), which means those 3 delay categories have more correlations than others. NAS_DELAY and WEATHER_DELAY have the smallest coefficient (absolute value) which is very close to 0, indicating that NAS_DELAY and WEATHER_DELAY basically are basically not correlated.