Naga Venkateshwarlu Yadav Dokku,
CSC-555 Project Report

Introduction

The telephone number 3-1-1 is a special telephone number supported in many communities in Canada and the United States. The number provides access to non-emergency municipal services. The number format follows the N11 code for a group of short, special-purpose local numbers. The number 3-1-1 is intended in part to divert routine inquiries and non-urgent community concerns from the 9-1-1 number which is reserved for emergency service. A promotional website for 3-1-1 in Akron described the distinction as follows: "Burning building? Call 9-1-1. Burning Question? Call 3-1-1. Many cities also accept 3-1-1 comments through online interfaces. An Open 311 application programming interface is also available for these services. 3-1-1 service is generally implemented at the local level, and in some cities, it is also used for various municipal calls. Examples of calls intended for 3-1-1:

dead animal removal
debris in roadway [citation needed]
illegal burning
non-working streetlamps, parking meters, traffic lights
noise complaints
Parking Law Enforcement
potholes, sinkholes and utility holes in streets
reporting stolen vehicles

3-1-1 is available in several major American cities

Overall, this project is aimed to analyze the 311 service requests for city New York from 2010 to Present (This information is automatically updated daily.)

The data will be obtained from the New York website https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9. The data includes CASE_ID, CREATED_DATE, CLOSED_DATE, AGENCY, AGENCY_NAME, COMPLAINT_TYPE ..etc

The scope of the project is to leverage the Hadoop, Pig and Hive to perform data analysis.

Hadoop Distribution File System and performance evaluation

The objective of this section is to compare the performance of execution costs for the 311 service requests for city New York data using different AWS Hadoop cluster configurations (instance type t2.small & 25 GB
Extended disk storage):

i. Single node – AWS small instance
ii. 3 nodes cluster with replication factor 1 and HDFS balancer
iii. 3 nodes cluster with replication factor 3 and HDFS balancer

For the purpose of this exercise, I first modified the size of data by using the subset of data (< 1G) and then increased up to 7GB from the original size of 7 GB by selecting some lines of original file and coping to new file each time and executing MapReduce “wordcount” and later deleted file each time after executing MapReduce function wordcount due to disk space. I have created instances of size 25GB. As expected Due to the project scope and timeline, I decided to created separate instances (single –

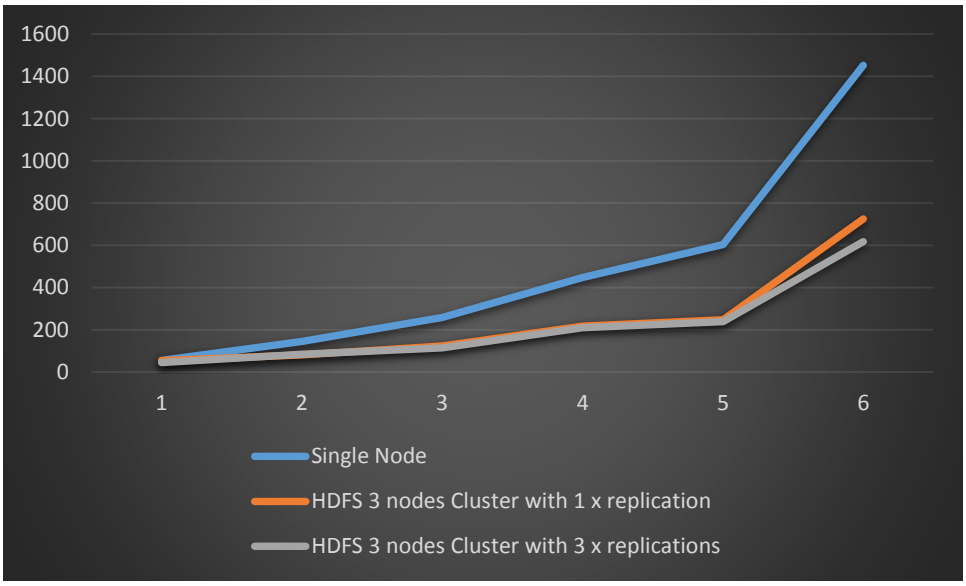
node[private ip : 172.31.8.56], 3-node cluster with replication 1[master : 172.31.9.61, node-1 : 172.31.9.60, node -2 : 172.31.9.62],3-node cluster with replication 3[master : 172.31.5.233,node-1 : 172.31.5.235,node-2: 172.31.5.234] and performed in parallel the task in single-node , 3-node cluster with replication 1, 3-node cluster with replication 3.If I were to redo the performance comparison test, I would set up the Hadoop cluster with larger AWS instances.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP
Master-rep1	i-02b3cf85759f5c6bc	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-184-72-66-47.comp...	184.72.66.47
Master-rep3	i-02b97c323549632a2	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-52-23-198-149.com...	52.23.198.149
worker1-rep1	i-03b0cb984bd16a98c	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-184-72-66-154.com...	184.72.66.154
worker2-rep1	i-070f70916edad0bc9	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-52-91-244-78.comp...	52.91.244.78
worker1-rep3	i-08904bca25af2fb05	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-54-86-138-84.comp...	54.86.138.84
worker2-rep3	i-08bf0e2a1f5ef1d68	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-52-91-23-217.comp...	52.91.23.217
Master-Final	i-0bab8fb6ad741353d	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-52-91-138-101.com...	52.91.138.101
Single-node Project	i-0c25728626b3459b6	t2.small	us-east-1b	running	2/2 checks ...	None	ec2-107-22-149-25.com...	107.22.149.25

The first test for evaluating the HDFS performance was to execute the MapReduce function “WordCount” for different HDFS files. I started with size of 0.13GB (3 blocks) and then increased blocks by selecting some lines of original file and coping to new file each time and executing MapReduce “wordcount” and later deleted file each time after executing MapReduce function wordcount due to disk space. I have created instances of size 25GB. As expected, the execution time of the single node configuration was too long compared to the HDFS three nodes cluster configuration with balancer. Based on the test results, the execution time of cluster with 1 replication factor compared to 3 replication factors is almost equal for smaller block size. By increasing the block size, the execution time of 3 nodes cluster with the replication factor =3 is slightly better than cluster with the replication factor = 1. Since this is a small cluster configuration with small data set, likely data is local to the network and the JobTracker has more than one node to choose from to schedule the task which can be explored to improve the performance. For the larger data set using a replication factor greater than one (> 1), all the nodes may be experiencing heavy load. In this case, the JobTracker will be obligated to schedule tasks in any other node that is not a heavy load. Since the data blocks are now not available locally, the task will have to request the data block from a node that has the data. This process will require data transfer on the network and will slow the job completion.

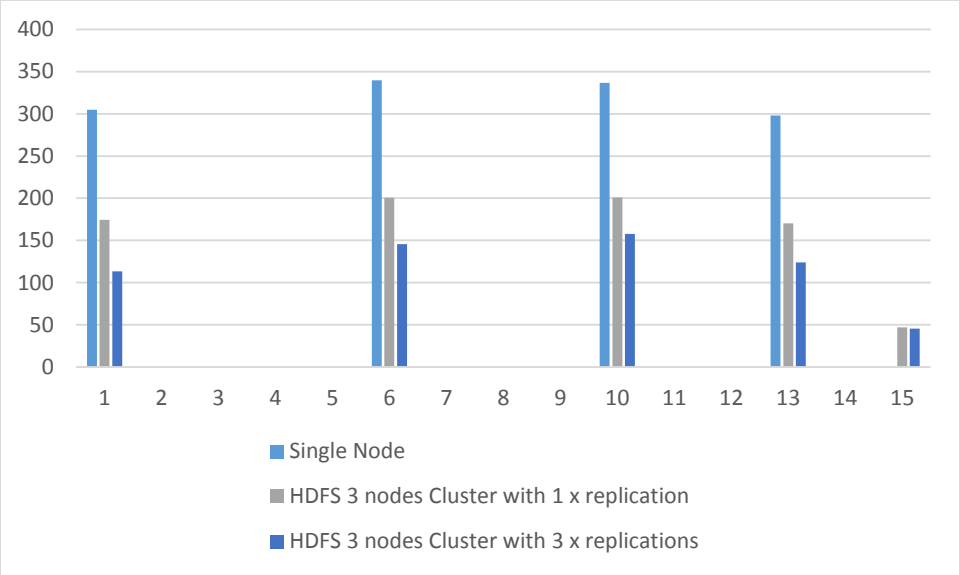
And for second part of project I have tried couples of hive and pig queries in each cluster setup and compared the performance of hive queries for single node, 3-node cluster with replication 1, 3-node cluster with replication 3. The queries I have tried are the number of request made for each city, Number of case requests by City which are more than 50000, Number of complaints by Complaint type and Agency, Find the most common types of complaints – as part of this data analysis, I compared the performance using “order by”, “sort by” and “cluster by”. And final partitioning the table and comparing performance.

# of Blocks	Single Node	HDFS 3 nodes Cluster with 1 x replication	HDFS 3 nodes Cluster with 3 x replications
3 blocks	55.672	53.828	44.751
9 blocks	143.845	80.877	83.806
16 blocks	258.016	123.163	113.837
35 blocks	448.267	217.654	210.084
42 blocks	603.493	246.987	238.586
108 blocks	1452.436	725.785	616.744



The second testing was to perform different HIVE queries using the original file of 7GB size. The HDFS configuration was the same as the first test with separate (single-node, 3-node cluster with replication 1, 3-node cluster with replication 3) instances with same configuration as MapReduce WordCount respectively. Based on the results, the single node has the highest execution time compared to HDFS 3 nodes clusters. The cluster with the replication factor = 3 had better performance for the execution of all queries.

HIVE – Query(7 GB data)	Single Node	HDFS 3 nodes Cluster with 1 x replication	HDFS 3 nodes Cluster with 3 x replications
SELECT COUNT(*) FROM SERVICE_REQUEST;	198.133 seconds	155.547 seconds	92.962 seconds
SELECT CITY, COUNT(*) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY CITY;	304.783 seconds	174.55 seconds	113.543 seconds
SELECT * FROM (SELECT CITY, COUNT(*) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY CITY ORDER BY NUMBER_OF_CASES) X WHERE X.NUMBER_OF_CASES > 50000;	339.575 seconds	200.74 seconds	145.685 seconds
SELECT AGENCY, COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY AGENCY, COMPLAINT_TYPE ORDER BY AGENCY,COMPLAINT_TYPE;	336.439 seconds	200.772 seconds	157.637 sec
select distinct COMPLAINT_TYPE,COUNT(COMPLAINT_TYPE) from SERVICE_REQUEST;	297.896 seconds	170.205 seconds	124.011 seconds
CREATE TABLE SERVICE_REQUEST3 (CASE_ID STRING, CREATED_DATE STRING, CLOSED_DATE STRING, AGENCY_NAME STRING,COMPLAINT_TYPE STRING, CITY STRING, INCIDENT_ZIP STRING) PARTITIONED BY(AGENCY STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE; INSERT OVERWRITE TABLE SERVICE_REQUEST3 PARTITION(AGENCY) SELECT CASE_ID, CREATED_DATE, CLOSED_DATE,AGENCY_NAME, COMPLAINT_TYPE, CITY, INCIDENT_ZIP,AGENCY FROM SERVICE_REQUEST; SELECT COUNT(*) FROM SERVICE_REQUEST3;	Failed	47.001 seconds	45.458 seconds



Analyzing using Hive and Pig

Understanding the size of file, data and its attributes

Find the total number of blocks and size of file:
>bin/hadoop fsck /data/311ServiceRequest.csv –files –blocks
Single –node

```
ec2-user@ip-172-31-8-56:~/hadoop-0.20.205.0
/home/ec2-user/hadoop-0.20.205.0/libexec/../conf/hadoop-env.sh: line 10: /usr/li
b/jvm/java-1.7.0-openjdk.x86_64: Is a directory
FSCK started by ec2-user from /172.31.8.56 for path /data/311ServiceRequest.csv
at Wed Jun 08 21:50:36 UTC 2016
.Status: HEALTHY
 Total size:    7189800721 B
 Total dirs:    0
 Total files:   1
 Total blocks (validated):      108 (avg. block size 66572228 B)
 Minimally replicated blocks:   108 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Wed Jun 08 21:50:36 UTC 2016 in 5 milliseconds


The filesystem under path '/data/311ServiceRequest.csv' is HEALTHY
[ec2-user@ip-172-31-8-56 hadoop-0.20.205.0]$
```

Naga Venkateshwarlu Yadav Dokku,
CSC-555 Project Report

3-node cluster with replication 1

```
ec2-user@ip-172-31-9-61:~/hadoop-0.20.205.0
Warning: $HADOOP_HOME is deprecated.

FSCK started by ec2-user from /172.31.9.61 for path /data/311ServiceRequest.csv
at Thu Jun 09 19:39:40 UTC 2016
.Status: HEALTHY
 Total size:    7187080819 B
 Total dirs:    0
 Total files:   1
 Total blocks (validated):      108 (avg. block size 66547044 B)
 Minimally replicated blocks:   108 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          3
 Number of racks:               1
FSCK ended at Thu Jun 09 19:39:40 UTC 2016 in 6 milliseconds


The filesystem under path '/data/311ServiceRequest.csv' is HEALTHY
[ec2-user@ip-172-31-9-61 hadoop-0.20.205.0]$
```

3-node cluster with replication 3

```
ec2-user@ip-172-31-5-233:~/hadoop-0.20.205.0
Warning: $HADOOP_HOME is deprecated.

FSCK started by ec2-user from /172.31.5.233 for path /data/311ServiceRequest.csv
 at Thu Jun 09 19:38:25 UTC 2016
.Status: HEALTHY
 Total size:    7187080819 B
 Total dirs:    0
 Total files:   1
 Total blocks (validated):      108 (avg. block size 66547044 B)
 Minimally replicated blocks:   108 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    3
 Average block replication:     3.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          3
 Number of racks:               1
FSCK ended at Thu Jun 09 19:38:25 UTC 2016 in 1 milliseconds


The filesystem under path '/data/311ServiceRequest.csv' is HEALTHY
[ec2-user@ip-172-31-5-233 hadoop-0.20.205.0]$
```

Pre-processing step to create subset of data using Hive Partitioning for further manipulation and analysis

```
CREATE TABLE SERVICE_REQUEST3 (CASE_ID STRING, CREATED_DATE STRING, CLOSED_DATE STRING,
AGENCY_NAME STRING,COMPLAINT_TYPE STRING, CITY STRING, INCIDENT_ZIP STRING)
PARTITIONED BY(AGENCY STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;


INSERT OVERWRITE TABLE SERVICE_REQUEST3 PARTITION(AGENCY) SELECT CASE_ID, CREATED_DATE,
CLOSED_DATE,AGENCY_NAME, COMPLAINT_TYPE, CITY, INCIDENT_ZIP,AGENCY FROM
SERVICE_REQUEST;
```

Initially there was an error inserting data into created subset.

FAILED: Error in semantic analysis: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive.exec.dynamic.partition.mode=nonstrict

It was solved by setting these parameters

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

```
MapReduce Total cumulative CPU time: 1 minutes 32 seconds 600 msec
Ended Job = job_201606071740_0014
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 92.6 sec   HDFS Read: 7187536613 H
DFS Write: 9 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 32 seconds 600 msec
OK
11442923
Time taken: 161.963 seconds
hive>
```

Count of whole dataset using PIG

```
SERVICE_REQUEST_OUTPUT = GROUP SERVICE_REQUEST_PIG ALL;
Count = FOREACH SERVICE_REQUEST_OUTPUT  GENERATE COUNT(SERVICE_REQUEST_PIG);
DUMP Count;
```

```
ec2-user@ip-172-31-8-56:~/pig-0.9.2
:9000/tmp/temp-1538420106/tmp-898635166"

Counters:
Total records written : 1
Total bytes written : 14
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201606081727_0034


2016-06-09 19:04:36,955 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 2623104 time(s).
2016-06-09 19:04:36,955 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2016-06-09 19:04:36,958 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2016-06-09 19:04:36,958 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(11447156)
grunt>
```

Create subset of data where AGENCY == 'NYPD' using Pig:

```
SERVICE_REQUEST_AGENCY = FILTER SERVICE_REQUEST_PIG BY AGENCY == 'NYPD';
C = GROUP SERVICE_REQUEST_AGENCY BY COMPLAINT_TYPE;
D = FOREACH C GENERATE COUNT(SERVICE_REQUEST_AGENCY);
DUMP D;
```

```
ec2-user@ip-172-31-8-56:~/pig-0.9.2
(4189)
(1239)
(1)
(95087)
(438472)
(46293)
(1)
(186538)
(8176)
(24827)
(1)
(1790)
(21937)
(24702)
(19247)
(8)
(2)
(190158)
(334863)
(2413)
(98283)
(17752)
(6688)
grunt>
```

As part of the pre-processing step, Pig can be used to create a subset of data by removing unnecessary columns:

```
SERVICE_REQUEST_PIG = LOAD '/data/311ServiceRequest.csv' USING PigStorage(',')
AS(CASE_ID:CHARARRAY,CREATED_DATE:CHARARRAY,CLOSED_DATE:CHARARRAY,AGENCY:CHARARRAY,AGENCY_NAME:CH
ARARRAY,COMPLAINT_TYPE:CHARARRAY,DESCRIPTOR:CHARARRAY,LOCATION_TYPE:CHARARRAY,INCIDENT_ZIP:CHARARRA
Y,INCIDENT_ADDRESS:CHARARRAY,STREET_NAME:CHARARRAY,CROSS_STREET1:CHARARRAY,CROSS_STREET2:CHARARRAY,I
NTERSECTION_STREET1:CHARARRAY,INTERSECTION_STREET2:CHARARRAY,ADDRESS_TYPE:CHARARRAY,CITY:CHARARRAY,LA
NMARK:CHARARRAY,FACILITY_TYPE:CHARARRAY,STATUS:CHARARRAY,DUE_DATE:CHARARRAY,RESOLUTION_DESCRIPTION:C
HARARRAY,RESOLUTION_ACTION_UPDATE_DATE:CHARARRAY,
COMMUNITY_BOARD:CHARARRAY,BOROUGH:CHARARRAY,X_COORDINATE:INT,Y_COORDINATE:INT,PARK_FACILITY_NAME:
CHARARRAY,PARK_BOROUGH:CHARARRAY,SCHOOL_NAME:CHARARRAY,SCHOOL_NUMBER:CHARARRAY,SCHOOL_REGION:C
HARARRAY,SCHOOL_CODE:CHARARRAY,SCHOOL_PHONE_NUMBER:CHARARRAY,SCHOOL_ADDRESS:CHARARRAY,SCHOOL_CI
TY:CHARARRAY,SCHOOL_STATE:CHARARRAY,SCHOOL_ZIP:CHARARRAY,SCHOOL_NOT_FOUND:CHARARRAY,SCHOOL_OR_CIT
YWIDE_COMPLAINT:CHARARRAY,VEHICLE_TYPE:CHARARRAY,TAXI_COMPANY_BOROUGH:CHARARRAY,TAXI_PICK_UP_LOCA
TION:CHARARRAY,BRIDGE_HIGHWAY_NAME:CHARARRAY,BRIDGE_HIGHWAY_DIRECTION:CHARARRAY,ROAD_RAMP:CHARA
RRAY,BRIDGE_HIGHWAY_SEGMENT:CHARARRAY,GARAGE_LOT_NAME:CHARARRAY,FERRY_DIRECTION:CHARARRAY,FERRY_
TERMINAL_NAME:CHARARRAY,LATITUDE:FLOAT,LONGITUDE:FLOAT,LOCATION_DETAIL:CHARARRAY);
DESCRIBE SERVICE_REQUEST_PIG;
```

```
A = GROUP SERVICE_REQUEST_PIG ALL;
B = foreach SERVICE_REQUEST_PIG generate CASE_ID, CREATED_DATE, CLOSED_DATE, AGENCY,
AGENCY_NAME,COMPLAINT_TYPE;
DUMP B;
```

```
grunt> SERVICE_REQUEST_PIG = LOAD '/data/311ServiceRequest.csv' USING PigStorage
(',')
>> AS(CASE_ID:CHARARRAY, CREATED_DATE:CHARARRAY, CLOSED_DATE:CHARARRAY, AGENCY:C
HARARRAY, AGENCY_NAME:CHARARRAY,COMPLAINT_TYPE:CHARARRAY,DESCRIPTOR:CHARARRAY,LO
CATION_TYPE:CHARARRAY, INCIDENT_ZIP:CHARARRAY,INCIDENT_ADDRESS:CHARARRAY,STREET_
NAME:CHARARRAY, CROSS_STREET1:CHARARRAY, CROSS_STREET2:CHARARRAY, INTERSECTION_S
TREET1:CHARARRAY, INTERSECTION_STREET2:CHARARRAY, ADDRESS_TYPE:CHARARRAY, CITY:C
HARARRAY, LANMARK:CHARARRAY, FACILITY_TYPE:CHARARRAY,STATUS:CHARARRAY, DUE_DATE:
CHARARRAY, RESOLUTION_DESCRIPTION:CHARARRAY, RESOLUTION_ACTION_UPDATE_DATE:CHARA
RRAY, COMMUNITY_BOARD:CHARARRAY,BOROUGH:CHARARRAY, X_COORDINATE:INT, Y_COORDINAT
E:INT, PARK_FACILITY_NAME:CHARARRAY, PARK_BOROUGH:CHARARRAY,SCHOOL_NAME:CHARARRA
Y, SCHOOL_NUMBER:CHARARRAY, SCHOOL_REGION:CHARARRAY, SCHOOL_CODE:CHARARRAY, SCHO
OL_PHONE_NUMBER:CHARARRAY,SCHOOL_ADDRESS:CHARARRAY, SCHOOL_CITY:CHARARRAY, SCHOO
L_STATE:CHARARRAY, SCHOOL_ZIP:CHARARRAY, SCHOOL_NOT_FOUND:CHARARRAY,SCHOOL_OR_CI
TYWIDE_COMPLAINT:CHARARRAY,  VEHICLE_TYPE:CHARARRAY, TAXI_COMPANY_BOROUGH:CHARAR
RAY, TAXI_PICK_UP_LOCATION:CHARARRAY, BRIDGE_HIGHWAY_NAME:CHARARRAY,BRIDGE_HIGHW
AY_DIRECTION:CHARARRAY, ROAD_RAMP:CHARARRAY, BRIDGE_HIGHWAY_SEGMENT:CHARARRAY, G
ARAGE_LOT_NAME:CHARARRAY, FERRY_DIRECTION:CHARARRAY,FERRY_TERMINAL_NAME:CHARARRA
Y, LATITUDE:FLOAT, LONGITUDE:FLOAT, LOCATION_DETAIL:CHARARRAY);
grunt> DESCRIBE SERVICE_REQUEST_PIG;
SERVICE_REQUEST_PIG: {CASE_ID: chararray,CREATED_DATE: chararray,CLOSED_DATE: ch
ararray,AGENCY: chararray,AGENCY_NAME: chararray,COMPLAINT_TYPE: chararray,DESCR
IPTOR: chararray,LOCATION_TYPE: chararray,INCIDENT_ZIP: chararray,INCIDENT_ADDRE
SS: chararray,STREET_NAME: chararray,CROSS_STREET1: chararray,CROSS_STREET2: cha
rarray,INTERSECTION_STREET1: chararray,INTERSECTION_STREET2: chararray,ADDRESS_T
YPE: chararray,CITY: chararray,LANMARK: chararray,FACILITY_TYPE: chararray,STATU
S: chararray,DUE_DATE: chararray,RESOLUTION_DESCRIPTION: chararray,RESOLUTION_AC
TION_UPDATE_DATE: chararray,COMMUNITY_BOARD: chararray,BOROUGH: chararray,X_COOR
DINATE: int,Y_COORDINATE: int,PARK_FACILITY_NAME: chararray,PARK_BOROUGH: charar
ray,SCHOOL_NAME: chararray,SCHOOL_NUMBER: chararray,SCHOOL_REGION: chararray,SCH
OOL_CODE: chararray,SCHOOL_PHONE_NUMBER: chararray,SCHOOL_ADDRESS: chararray,SCH
OOL_CITY: chararray,SCHOOL_STATE: chararray,SCHOOL_ZIP: chararray,SCHOOL_NOT_FOU
ND: chararray,SCHOOL_OR_CITYWIDE_COMPLAINT: chararray,VEHICLE_TYPE: chararray,TA
XI_COMPANY_BOROUGH: chararray,TAXI_PICK_UP_LOCATION: chararray,BRIDGE_HIGHWAY_NA
ME: chararray,BRIDGE_HIGHWAY_DIRECTION: chararray,ROAD_RAMP: chararray,BRIDGE_HI
GHWAY_SEGMENT: chararray,GARAGE_LOT_NAME: chararray,FERRY_DIRECTION: chararray,F
ERRY_TERMINAL_NAME: chararray,LATITUDE: float,LONGITUDE: float,LOCATION_DETAIL:
chararray}
grunt> B = foreach SERVICE_REQUEST_PIG generate CASE_ID, CREATED_DATE, CLOSED_DATE, AGENCY, AGENCY_NAME,COMPLAINT_TYPE;
grunt> dump B;
```

```
ec2-user@ip-172-31-8-56:~/pig-0.9.2
(33467591,05/29/2016 11:06:32 PM,05/30/2016 08:04:35 AM,NYPD,New York City Polic
e Department,Noise - Commercial)
(33467612,05/30/2016 12:44:10 AM,05/30/2016 07:02:07 AM,NYPD,New York City Polic
e Department,Noise - Residential)
(33467614,05/30/2016 01:19:19 AM,05/30/2016 07:48:27 AM,NYPD,New York City Polic
e Department,Blocked Driveway)
(33467643,05/30/2016 12:42:23 AM,05/30/2016 03:04:57 AM,NYPD,New York City Polic
e Department,Noise - Vehicle)
(33467652,05/29/2016 10:07:50 PM,05/29/2016 10:45:47 PM,NYPD,New York City Polic
e Department,Blocked Driveway)
(33467654,05/29/2016 10:33:08 PM,05/30/2016 01:29:27 AM,NYPD,New York City Polic
e Department,Blocked Driveway)
(33467662,05/29/2016 11:13:01 PM,05/30/2016 08:31:59 AM,NYPD,New York City Polic
e Department,Noise - Street/Sidewalk)
(33467663,05/29/2016 10:51:08 PM,05/30/2016 07:25:03 AM,NYPD,New York City Polic
e Department,Noise - Street/Sidewalk)
(33467667,05/30/2016 01:33:22 AM,,DOHMH,Department of Health and Mental Hygiene,
Drinking Water)
(32828637,03/03/2016 04:19:54 PM,03/16/2016 01:18:50 PM,HPD,Department of Housin
g Preservation and Development,PAINT/PLASTER)
(32828619,03/03/2016 01:05:21 PM,03/15/2016 12:08:03 PM,HPD,Department of Housin
g Preservation and Development,PLUMBING)
(33467704,05/29/2016 04:41:56 PM,,DSNY,Department of Sanitation,Graffiti)
grunt>
```

Execute MapReduce functions from Hive and/or Pig for data analysis

Find the COMPLAINT_TYPES of 311ServiceRequest

select distinct COMPLAINT_TYPE,COUNT(COMPLAINT_TYPE) from SERVICE_REQUEST;

```
Elevator
Fire Alarm - Replacement
Fire Safety Director - F58
Forensic Engineering
Gas Station Discharge Lines
Illegal Animal Sold
Illegal Parking
Miscellaneous Categories
NONCONST
Noise - House of Worship
Open Flame Permit
PAINT - PLASTER
Radioactive Material
SCRIE
Standpipe - Mechanical
Street Condition
Taxi Report
Traffic Signal Condition
Water Conservation
Water Quality
 Arts and Media"
 Science Research
Adopt-A-Basket
Broken Muni Meter
Building Condition
City Vehicle Placard Complaint
Collection Truck Noise
Comment
Complaint Type
DOF Parking - Payment Issue
DOF Property - City Rebate
DOF Property - Update Account
DOT Literature Request
DPR Internal
EAP Inspection - F59
Emergency Response Team (ERT)
Forms
Hazardous Materials
Highway Sign - Dangling
Illegal Tree Damage
Indoor Sewage
Legal Services Provider Complaint
Missed Collection (All Materials)
OUTSIDE BUILDING
Overflowing Litter Baskets
Posting Advertisement
School Maintenance
Special Projects Inspection Team (SPIT)
Teaching/Learning/Instruction
Trapping Pigeon
Unleashed Dog
Unspecified
Violation of Park Rules
Time taken: 170.205 seconds
hive>
```

Naga Venkateshwarlu Yadav Dokku,
CSC-555 Project Report

Business	Transportation Provider Complaint	Unsanitary Animal Pvt Property	Advocate-Prop Refunds/Credits	DOE Complaint or Compliment	Food Establishment	Illegal Parking	Posting Advertisemen
Computer Applications	Unlicensed Dog	Urinating in Public	Benefit Card Replacement	DOF Parking - Tax Exemption	For Hire Vehicle Report	Miscellaneous Categories	School Maintenance
Science and Technology Applications"	Vacant Lot	VACANT APARTMENT	Complaint	DOF Property - RPIE Issue	HEAT/HOT WATER	NONCONST	Special Projects Inspe
Advocate - Other	Water System	Finance	DOF Parking - Request Status	Dead/Dying Tree	Investigations and Discipline (IAD)	Noise - House of Worship	Teaching/Learning/Ins
Advocate-Personal Exemptions	Engineering and Architecture"	Nursing	DOF Property - Reduction Issue	Foam Ban Enforcement	Mosquitoes	Open Flame Permit	Trapping Pigeon
Advocate-Property Value	Government	Advocate - RPIE	FLOORING/STAIRS	HPD Literature Request	New Tree Request	PAINT - PLASTER	Unleashed Dog
Animal in a Park	Science and Design Technology"	Advocate-Prop Class Incorrect	Fire Alarm - New System	Health	Overflowing Recycling Baskets	Radioactive Material	Unspecified
Asbestos/Garbage Nuisance	Science and Technology"	Animal Abuse	Fire Alarm - Reinspection	Healthcare Facilities	Public Assembly	SCRIE	Violation of Park Rules
BEST/Site Safety	Air Quality	Asbestos	Found Property	Indoor Air Quality	Root/Sewer/Sidewalk Condition	Standpipe - Mechanical	Ferry Complaint
Cable Complaint	Bike Rack Condition	Compliment	Highway Sign - Missing	Lead	STRUCTURAL	Street Condition	
Disorderly Youth	Bike/Roller/Skate Chronic	Cranes and Derricks	Homeless Person Assistance	Non-Emergency Police Matter	Sanitation Condition	Taxi Report	
Electrical	Blocked Driveway	DPR Literature Request	Industrial Waste	Non-Residential Heat	Street Sign - Dangling	Traffic Signal Condition	
Fire Alarm - Addition	CONSTRUCTION	Dead Tree	Interior Demo	Opinion for the Mayor	Street Sign - Missing	Water Conservation	
Fire Alarm - Modification	Calorie Labeling	Drinking	Laboratory	PLUMBING	Tanning	Water Quality	
Food Poisoning	Construction	For Hire Vehicle Complaint	Maintenance or Facility	Plumbing	Tattooing	Arts and Media"	
GENERAL	Consumer Complaint	General Construction/Plumbing	Noise - Commercial	Poison Ivy	Traffic	Science Research	
Graffiti	DEP Literature Request	HEATING	Noise - Park	Public Toilet	Unsanitary Animal Facility	Adopt-A-Basket	
Harboring Bees/Wasps	DFTA Literature Request	Hazmat Storage/Use	Noise - Residential	Request for Information	Vending	Broken Muni Meter	
Highway Condition	DOOR/WINDOW	Highway Sign - Damaged	Noise - Street/Sidewalk	Rodent	Window Guard	Building Condition	
Homeless Encampment	DOR Literature Request	Home Care Provider Complaint	Noise - Vehicle	SNW	Imagination and Inquiry"	City Vehicle Placard Complaint	
Literature Request	Damaged Tree	Illegal Fireworks	Parent Leadership	Scaffold Safety	Advocate-Commercial Exemptions	Collection Truck Noise	
Micro Switch	Derelict Bicycle	Litter Basket / Request	Plant	Smoking	Animal Facility - No Permit	Comment	
No Child Left Behind	Drinking Water	Misc. Comments	Public Assembly - Temporary	Street Light Condition	Beach/Pool/Sauna Complaint	Complaint Type	
Noise - Helicopter	Ferry Inquiry	Mold	Sidewalk Condition	Street Sign - Damaged	Boilers	DOF Parking - Payment Issue	
Noise Survey	Ferry Permit	Noise	Snow	Sweeping/Missed	Broken Parking Meter	DOF Property - City Rebate	
Other Enforcement	GENERAL CONSTRUCTION	PAINT/PLASTER	Stalled Sites	Taxi Compliment	DCA / DOH New License Application Request	DOF Property - Update Account	
Overgrown Tree/Branches	Illegal Animal Kept as Pet	Panhandling	Standing Water	WATER LEAK	DCA Literature Request	DOT Literature Request	
Parking Card	Invitation	Portable Toilet	Tunnel Condition	Science and Engineering at City College"	DHS Income Savings Requirement	DPR Internal	
Public Payphone Complaint	Lifeguard	SAFETY	Unsanitary Pigeon Condition	AGENCY	DOF Property - Payment Issue	EAP Inspection - F59	
Recycling Enforcement	Lost Property	Safety	Vector	APPLIANCE	Derelict Vehicle	Emergency Response Team (ERT)	
Registration and Transfers	Municipal Parking Facility	Sewer	X-Ray Machine/Equipment	Advocate-SCRIE/DRIE	Dirty Conditions	Forms	
Request Xmas Tree Collection	OEM Disabled Vehicle	Sprinkler - Mechanical	PS 267"	Advocate-UBT	Discipline and Suspension	Hazardous Materials	
Research Questions	OEM Literature Request	Summer Camp	Queens and Brooklyn"	Agency Issues	ELEVATOR	Highway Sign - Dangling	
SG-98	Rangehood	Sweeping/Inadequate	Science	Bottled Water	Elevator	Illegal Tree Damage	
SRDE	SG-99	Sweeping/Missed-Inadequate	Technology	Curb Condition	Fire Alarm - Replacement	Indoor Sewage	
Senior Center Complaint	Special Enforcement	Advocacy and Community Justic	Advocate-Co-opCondo Abatemer	DOF Literature Request	Fire Safety Director - F58	Legal Services Provider Complaint	
Special Natural Area District (SNAD)	Taxi Complaint	Engineering	Bridge Condition	Day Care	Forensic Engineering	Missed Collection (All Materials)	
Squeegee	Trans Fat	Liberal Arts	Building/Use	Derelict Vehicles	Gas Station Discharge Lines	OUTSIDE BUILDING	
Taxpayer Advocate Inquiry	UNSANITARY CONDITION	Technology and Math High Scho	Bus Stop Shelter Placement	ELECTRIC	Illegal Animal Sold	Overflowing Litter Baskets	

Naga Venkateshwarlu Yadav Dokku,
CSC-555 Project Report

Find the most common types of complaints – as part of this data analysis, I compared the performance using "order by", "sort by" and "cluster by".

```
SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE SORT BY CNT DESC LIMIT 5;
SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE ORDER BY CNT DESC LIMIT 5;
SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE CLUSTER BY CNT LIMIT 5;
```

Result (same for all sort and order by queries):
HEATING 887869
Street Light Condition 637718
Street Condition 635620
PLUMBING 540195
GENERAL CONSTRUCTION 500863

Results for cluster
Micro Switch 1
Computer Applications 1
Asbestos/Garbage Nuisance 1
Trapping Pigeon 1
Complaint Type 1

Sort by			Order by			Cluster by		
Single Node HDFS	HDFS 3 nodes Cluster with 1 x replication	HDFS 3 nodes Cluster with 3 x replications	Single Node HDFS	HDFS 3 nodes Cluster with 1 x replication	HDFS 3 nodes Cluster with 3 x replications	Single Node HDFS	HDFS 3 nodes Cluster with 1 x replication	HDFS 3 nodes Cluster with 3 x replications
375.105 seconds	231.79 seconds	189.457 seconds	338.735 seconds	189.322 seconds	157.461 seconds	369.178 seconds	249.934 seconds	189.785 seconds

Applying partitioning:
The partitioning in Hive allows the user to efficiently identify the rows that satisfy a certain criteria (similar concept as RDBM). In contrast to the non-partitioned table, Hive reads all the files in table's data and then applies filters which are slow and expensive.
I have used column "AGENCY" for partitioning. Based on the results below, the execution time using partitioned table is much less than the using query on a non-partitioned tables.

Query	**Execution Time**		
	Single Node HDFS	**HDFS 3 nodes Cluster with 1 x replication**	**HDFS 3 nodes Cluster with 3 x replications**
Non-partitioned table: SELECT COUNT(*) FROM SERVICE_REQUEST where AGENCY = 'NYPD';	267.14 seconds	159.015 seconds	108.791 seconds
Partitioned table SELECT COUNT(*) FROM SERVICE_REQUEST3 where AGENCY = 'NYPD';	36.389 seconds	47.001 seconds	33.542 seconds
Non-partitioned table SELECT COUNT(*) FROM SERVICE_REQUEST GROUP BY AGENCY;	307.14 seconds	169.404 seconds	113.453 seconds
Partitioned table SELECT COUNT(*) FROM SERVICE_REQUEST3 GROUP BY AGENCY;	78.635 seconds	56.043 seconds	44.271 seconds

Prepare partitioned table:

```
CREATE TABLE SERVICE_REQUEST3 (CASE_ID STRING, CREATED_DATE STRING, CLOSED_DATE STRING,
AGENCY_NAME STRING,COMPLAINT_TYPE STRING, CITY STRING, INCIDENT_ZIP STRING)
PARTITIONED BY(AGENCY STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

INSERT OVERWRITE TABLE SERVICE_REQUEST3 PARTITION(AGENCY) SELECT CASE_ID, CREATED_DATE, CLOSED_DATE,AGENCY_NAME, COMPLAINT_TYPE, CITY, INCIDENT_ZIP,AGENCY FROM SERVICE_REQUEST;

Initially there was an error inserting data into created subset.

FAILED: Error in semantic analysis: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive.exec.dynamic.partition.mode=nonstrict

It was solved by setting these parameters

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

```
MapReduce Total cumulative CPU time: 1 minutes 32 seconds 600 msec
Ended Job = job_201606071740_0014
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 92.6 sec   HDFS Read: 7187536613 H
DFS Write: 9 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 32 seconds 600 msec
OK
11442923
Time taken: 161.963 seconds
hive>
```

Single node

Non-partitioned table

SELECT COUNT(*) FROM SERVICE_REQUEST where AGENCY = 'NYPD';

```
ec2-user@ip-172-31-8-56:~/hive-0.8.1-bin
2016-06-09 14:19:01,699 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 93.98
 sec
2016-06-09 14:19:02,704 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 93.98
 sec
2016-06-09 14:19:03,706 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 93.98
 sec
2016-06-09 14:19:04,708 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 93.98
 sec
2016-06-09 14:19:05,711 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 93.98
 sec
2016-06-09 14:19:06,714 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 93.98
 sec
2016-06-09 14:19:07,716 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 93.98
 sec
MapReduce Total cumulative CPU time: 1 minutes 33 seconds 980 msec
Ended Job = job_201606081727_0019
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 93.98 sec   HDFS Read: 7190255111
HDFS Write: 8 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 33 seconds 980 msec
OK
1530802
Time taken: 267.14 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST GROUP BY AGENCY;

```
70287
1
1530802
1
Time taken: 307.14 seconds
hive>
```

partitioned table

SELECT COUNT(*) FROM SERVICE_REQUEST3 where AGENCY = 'NYPD';

```
MapReduce Total cumulative CPU time: 4 seconds 470 msec
Ended Job = job_201606081727_0020
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Accumulative CPU: 4.47 sec   HDFS Read: 183553786 HDF
S Write: 8 SUCESS
Total MapReduce CPU Time Spent: 4 seconds 470 msec
OK
1530802
Time taken: 36.389 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST3 GROUP BY AGENCY;

```
5
1
151975
Time taken: 78.635 seconds
hive>
```

HDFS 3 nodes Cluster with 1 x replication

partitioned table

SELECT COUNT(*) FROM SERVICE_REQUEST3;

```
MapReduce Total cumulative CPU time: 33 seconds 870 msec
Ended Job = job_201606071740_0026
MapReduce Jobs Launched:
Job 0: Map: 7  Reduce: 1   Accumulative CPU: 33.87 sec   HDFS Read: 1393134972 HDFS Write: 9 SUCESS
Total MapReduce CPU Time Spent: 33 seconds 870 msec
OK
11442923
Time taken: 47.001 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST3 where AGENCY = 'NYPD';

```
MapReduce Total cumulative CPU time: 4 seconds 370 msec
Ended Job = job_201606071740_0027
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Accumulative CPU: 4.37 sec   HDFS Read: 183332784 HDFS Write: 8 SUCESS
Total MapReduce CPU Time Spent: 4 seconds 370 msec
OK
1529066
Time taken: 34.556 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST3 GROUP BY AGENCY;

```
Time taken: 56.043 seconds
hive>
```

Non-partitioned table

SELECT COUNT(*) FROM SERVICE_REQUEST where AGENCY = 'NYPD';

```
MapReduce Total cumulative CPU time: 2 minutes 6 seconds 440 msec
Ended Job = job_201606071740_0028
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 126.44 sec   HDFS Read: 7187535209 HDFS Write: 8 SUCESS
Total MapReduce CPU Time Spent: 2 minutes 6 seconds 440 msec
OK
1529066
Time taken: 159.015 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST GROUP BY AGENCY;

```
Time taken: 169.404 seconds
hive>
```

HDFS 3 nodes Cluster with 3x replication

Non-partitioned table

SELECT COUNT(*) FROM SERVICE_REQUEST where AGENCY = 'NYPD';

```
MapReduce Total cumulative CPU time: 1 minutes 39 seconds 690 msec
Ended Job = job_201606071959_0029
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 99.69 sec   HDFS Read: 7187535317
HDFS Write: 8 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 39 seconds 690 msec
OK
1529066
Time taken: 108.791 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST GROUP BY AGENCY;

```
56
70265
1
1529066
1
Time taken: 113.453 seconds
hive>
```

partitioned table

SELECT COUNT(*) FROM SERVICE_REQUEST3 where AGENCY = 'NYPD';

```
Ended Job = job_201606071959_0030
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Accumulative CPU: 4.74 sec   HDFS Read: 183332787 HDF
S Write: 8 SUCESS
Total MapReduce CPU Time Spent: 4 seconds 740 msec
OK
1529066
Time taken: 33.542 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST3 GROUP BY AGENCY;

```
1
151912
Time taken: 44.271 seconds
hive>
```

Query	Execution Time		
	Single Node HDFS	**HDFS 3 nodes Cluster with 1 x replication**	**HDFS 3 nodes Cluster with 3 x replications**
Non-partitioned table: SELECT COUNT(*) FROM SERVICE_REQUEST where AGENCY = 'NYPD';	267.14 seconds	159.015 seconds	108.791 seconds
Partitioned table SELECT COUNT(*) FROM SERVICE_REQUEST3 where AGENCY = 'NYPD';	36.389 seconds	47.001 seconds	33.542 seconds
Non-partitioned table SELECT COUNT(*) FROM SERVICE_REQUEST GROUP BY AGENCY;	307.14 seconds	169.404 seconds	113.453 seconds
Partitioned table SELECT COUNT(*) FROM SERVICE_REQUEST3 GROUP BY AGENCY;	78.635 seconds	56.043 seconds	44.271 seconds

Naga Venkateshwarlu Yadav Dokku,
CSC-555 Project Report

Conclusion and Challenges

- Since Hive is SQL-like language, it was easier to use for data analysis
- The partitioning in Hive is easy to create.
- My challenge with Hive was using variables and passing to the next query
- The execution time was improved using partitioned table compare to non-partitioned table.
- I couldn't find the partitioning concept in Pig (not sure if it is supported or not)
- I would use Pig in a pre-processing step, which is similar to scripting language.
- Overall dumping the output before performing the next step is good for troubleshooting.
- I found that the performance of Hive and Pig for similar queries (simple group by) were the same except using the partitioned table in Hive performs better.

Issues faced

1. Error on executing hive queries

```
hive> INSERT OVERWRITE TABLE SERVICE_REQUEST2 SELECT CASE_ID, CREATED_DATE, CLOSED_DATE, AGENCY, AGENCY_NAME, COMPLAINT_TYPE, CITY, INCIDENT_ZIP FROM SERVICE_REQUEST WHERE YEAR(CREATED_DATE) = 2015
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201606081727_0002, Tracking URL = http://ip-172-31-8-56.ec2.internal:50030/jobdetails.jsp?jobid=job_201606081727_0002
Kill Command = /home/ec2-user/hadoop-0.20.205.0/libexec/../bin/hadoop job  -Dmapred.job.tracker=172.31.8.56:9001 -kill job_201606081727_0002
Hadoop job information for Stage-1: number of mappers: 27; number of reducers: 0
2016-06-08 17:43:23,830 Stage-1 map = 0%,  reduce = 0%
2016-06-08 17:44:05,973 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201606081727_0002 with errors
Error during job, obtaining debugging information...
Examining task ID: task_201606081727_0002_m_000028 (and more) from job job_201606081727_0002
Exception in thread "Thread-137" java.lang.RuntimeException: Error while reading from task log url
        at org.apache.hadoop.hive.ql.exec.errors.TaskLogProcessor.getErrors(TaskLogProcessor.java:130)
        at org.apache.hadoop.hive.ql.exec.JobDebugger.showJobFailDebugInfo(JobDebugger.java:211)
        at org.apache.hadoop.hive.ql.exec.JobDebugger.run(JobDebugger.java:81)
        at java.lang.Thread.run(Thread.java:745)
Caused by: java.io.IOException: Server returned HTTP response code: 400 for URL: http://ip-172-31-8-56.ec2.internal:50060/tasklog?taskid=attempt_201606081727_0002_m_000000_2&start=-8193
        at sun.net.www.protocol.http.HttpURLConnection.getInputStream(HttpURLConnection.java:1628)
        at java.net.URL.openStream(URL.java:1048)
        at org.apache.hadoop.hive.ql.exec.errors.TaskLogProcessor.getErrors(TaskLogProcessor.java:120)
        ... 3 more
FAILED: Execution Error, return code 2 from org.apache.hadoop.hive.ql.exec.MapRedTask
MapReduce Jobs Launched:
Job 0: Map: 27   HDFS Read: 0 HDFS Write: 0 FAIL
Total MapReduce CPU Time Spent: 0 msec
hive>
```

It was solved by releasing Hadoop from safe mode.

2. Failed to start database '/var/lib/hive/metastore/metastore_db' in hive

```
FAILED: Error in metadata: javax.jdo.JDOFatalDataStoreException: Failed to start
 database 'metastore_db', see the next exception for details.
NestedThrowables:
java.sql.SQLException: Failed to start database 'metastore_db', see the next exc
eption for details.
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTa
sk
hive>
```

Solved following these instructions

http://stackoverflow.com/questions/15810210/unable-to-instantiate-hivemetastoreclient

3. Error in insert overwrite in subset created from existing table
   FAILED: Error in semantic analysis: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive.exec.dynamic.partition.mode=nonstrict
   It was solved by setting these parameters

   SET hive.exec.dynamic.partition = true;

   SET hive.exec.dynamic.partition.mode = nonstrict;

Appendix – Print Screen of some Hadoop Performance Evaluating Testing and Hive Queries.

```
ec2-user@ip-172-31-8-56:~/hadoop-0.20.205.0
547
16/06/09 00:01:15 INFO mapred.JobClient:     Map input records=250000
16/06/09 00:01:15 INFO mapred.JobClient:     Reduce shuffle bytes=55166547
16/06/09 00:01:15 INFO mapred.JobClient:     Spilled Records=3542977
16/06/09 00:01:15 INFO mapred.JobClient:     Map output bytes=158816983
16/06/09 00:01:15 INFO mapred.JobClient:     Total committed heap usage (bytes)=
369442816
16/06/09 00:01:15 INFO mapred.JobClient:     CPU time spent (ms)=20410
16/06/09 00:01:15 INFO mapred.JobClient:     Combine input records=7686506
16/06/09 00:01:15 INFO mapred.JobClient:     SPLIT_RAW_BYTES=230
16/06/09 00:01:15 INFO mapred.JobClient:     Reduce input records=911755
16/06/09 00:01:15 INFO mapred.JobClient:     Reduce input groups=811220
16/06/09 00:01:15 INFO mapred.JobClient:     Combine output records=2257154
16/06/09 00:01:15 INFO mapred.JobClient:     Physical memory (bytes) snapshot=49
0557440
16/06/09 00:01:15 INFO mapred.JobClient:     Reduce output records=811220
16/06/09 00:01:15 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=349
1328000
16/06/09 00:01:15 INFO mapred.JobClient:     Map output records=6341107

real    0m55.672s
user    0m1.480s
sys     0m0.076s
[ec2-user@ip-172-31-8-56 hadoop-0.20.205.0]$
```

```
16/06/07 14:46:49 INFO mapred.JobClient:     Combine output records=135160943
16/06/07 14:46:49 INFO mapred.JobClient:     Physical memory (bytes) snapshot=19
125637120
16/06/07 14:46:49 INFO mapred.JobClient:     Reduce output records=18962898
16/06/07 14:46:49 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=126
551568384
16/06/07 14:46:49 INFO mapred.JobClient:     Map output records=440407529

real    24m12.435s
user    0m3.104s
sys     0m0.172s
[ec2-user@ip-172-31-8-56 hadoop-0.20.205.0]$
```

3 node cluster with replication 1 for whole dataset

```
16/06/07 18:00:05 INFO mapred.JobClient:     Reduce input groups=18956343
16/06/07 18:00:05 INFO mapred.JobClient:     Combine output records=135596788
16/06/07 18:00:05 INFO mapred.JobClient:     Physical memory (bytes) snapshot=19
029409792
16/06/07 18:00:05 INFO mapred.JobClient:     Reduce output records=18956343
16/06/07 18:00:05 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=125
390852096
16/06/07 18:00:05 INFO mapred.JobClient:     Map output records=440237207

real    12m5.785s
user    0m2.616s
sys     0m0.148s
[ec2-user@ip-172-31-9-61 hadoop-0.20.205.0]$
```

3 node cluster with replication 3 on whole dataset

```
16/06/07 20:10:34 INFO mapred.JobClient:     Physical memory (bytes) snapshot=18
968363008
16/06/07 20:10:34 INFO mapred.JobClient:     Reduce output records=18956343
16/06/07 20:10:34 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=125
390622720
16/06/07 20:10:34 INFO mapred.JobClient:     Map output records=440237207

real    10m16.744s
user    0m2.536s
sys     0m0.248s
[ec2-user@ip-172-31-5-233 hadoop-0.20.205.0]$
```

Creating Table in hive

```
hive> CREATE TABLE SERVICE_REQUEST (CASE_ID STRING, CREATED_DATE TIMESTAMP, CLOS
ED_DATE TIMESTAMP, AGENCY STRING, AGENCY_NAME STRING,COMPLAINT_TYPE STRING,DESCR
IPTOR STRING,LOCATION_TYPE STRING, INCIDENT_ZIP STRING,INCIDENT_ADDRESS STRING,S
TREET_NAME STRING, CROSS_STREET1 STRING, CROSS_STREET2 STRING, INTERSECTION_STRE
ET1 STRING, INTERSECTION_STREET2 STRING, ADDRESS_TYPE STRING, CITY STRING, LANMA
RK STRING, FACILITY_TYPE STRING,STATUS STRING, DUE_DATE TIMESTAMP, RESOLUTION_DE
SCRIPTION STRING, RESOLUTION_ACTION_UPDATE_DATE TIMESTAMP, COMMUNITY_BOARD STRIN
G,BOROUGH STRING, X_COORDINATE INT, Y_COORDINATE INT, PARK_FACILITY_NAME STRING,
 PARK_BOROUGH STRING,SCHOOL_NAME STRING, SCHOOL_NUMBER STRING, SCHOOL_REGION STR
ING, SCHOOL_CODE STRING, SCHOOL_PHONE_NUMBER STRING,SCHOOL_ADDRESS STRING, SCHOO
L_CITY STRING, SCHOOL_STATE STRING, SCHOOL_ZIP STRING, SCHOOL_NOT_FOUND STRING,S
CHOOL_OR_CITYWIDE_COMPLAINT STRING,  VEHICLE_TYPE STRING, TAXI_COMPANY_BOROUGH S
TRING, TAXI_PICK_UP_LOCATION STRING, BRIDGE_HIGHWAY_NAME STRING,BRIDGE_HIGHWAY_D
IRECTION STRING, ROAD_RAMP STRING, BRIDGE_HIGHWAY_SEGMENT STRING, GARAGE_LOT_NAM
E STRING, FERRY_DIRECTION STRING,FERRY_TERMINAL_NAME STRING, LATITUDE FLOAT, LON
GITUDE FLOAT, LOCATION_DETAIL STRING ) ROW FORMAT DELIMITED FIELDS TERMINATED BY
 ',' STORED AS TEXTFILE;
OK
Time taken: 6.731 seconds
hive>
```

```
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/311ServiceRequest.csv' OVERWRITE INTO TABLE SERVICE_REQUEST;
Copying data from file:/home/ec2-user/311ServiceRequest.csv
Copying file: file:/home/ec2-user/311ServiceRequest.csv
Loading data to table default.service_request
Deleted hdfs://ip-172-31-8-56.ec2.internal:9000/user/hive/warehouse/service_request
OK
Time taken: 118.609 seconds
hive>
```

SELECT COUNT(1) FROM SERVICE_REQUEST;

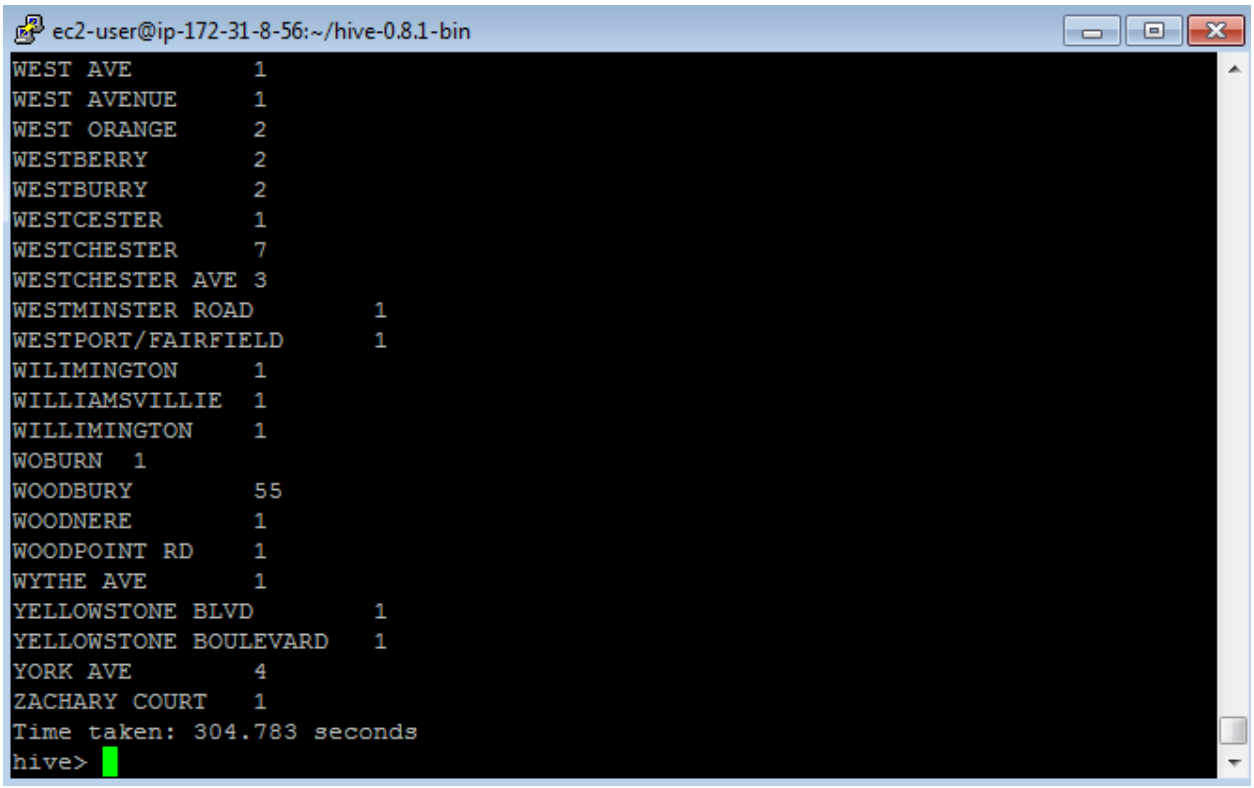
```
MapReduce Total cumulative CPU time: 1 minutes 10 seconds 970 msec
Ended Job = job_201606071417_0012
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 70.97 sec   HDFS Read: 7190255111 HDFS Write: 9 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 10 seconds 970 msec
OK
11447156
Time taken: 217.312 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST;

```
MapReduce Total cumulative CPU time: 1 minutes 8 seconds 850 msec
Ended Job = job_201606071417_0013
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 68.85 sec   HDFS Read: 7190255111 HDFS Write: 9 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 8 seconds 850 msec
OK
11447156
Time taken: 198.133 seconds
hive>
```

Number of case requests by City

SELECT CITY, COUNT(*) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY CITY;



```
ec2-user@ip-172-31-8-56:~/hive-0.8.1-bin
WEST AVE        1
WEST AVENUE     1
WEST ORANGE     2
WESTBERRY       2
WESTBURRY       2
WESTCESTER      1
WESTCHESTER     7
WESTCHESTER AVE 3
WESTMINSTER ROAD        1
WESTPORT/FAIRFIELD      1
WILIMINGTON     1
WILLIAMSVILLIE  1
WILLIMINGTON    1
WOBURN  1
WOODBURY        55
WOODNERE        1
WOODPOINT RD    1
WYTHE AVE       1
YELLOWSTONE BLVD        1
YELLOWSTONE BOULEVARD   1
YORK AVE        4
ZACHARY COURT   1
Time taken: 304.783 seconds
hive>
```

Number of case requests by City which are more than 50000

SELECT * FROM (SELECT CITY, COUNT(*) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY CITY ORDER BY NUMBER_OF_CASES) X WHERE X.NUMBER_OF_CASES > 50000;

```
ec2-user@ip-172-31-8-56:~/hive-0.8.1-bin
2016-06-08 15:46:40,293 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 99.4 sec
2016-06-08 15:46:41,295 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 99.4 sec
2016-06-08 15:46:42,297 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 99.4 sec
2016-06-08 15:46:43,300 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 99.4 sec
2016-06-08 15:46:44,302 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 99.4 sec
MapReduce Total cumulative CPU time: 1 minutes 39 seconds 400 msec
Ended Job = job_201606071417_0017
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201606071417_0018, Tracking URL = http://ip-172-31-8-56.ec2.internal:50030/jobdetails.jsp?jobid=job_201606071417_0018
Kill Command = /home/ec2-user/hadoop-0.20.205.0/libexec/../bin/hadoop job  -Dmapred.job.tracker=172.31.8.56:9001 -kill job_201606071417_0018
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2016-06-08 15:46:53,595 Stage-2 map = 0%,  reduce = 0%
2016-06-08 15:46:59,605 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:00,607 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:01,609 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:02,611 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:03,613 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:04,616 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:05,618 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:06,620 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:07,622 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2016-06-08 15:47:08,625 Stage-2 map = 100%,  reduce = 33%, Cumulative CPU 0.99 sec
2016-06-08 15:47:09,627 Stage-2 map = 100%,  reduce = 33%, Cumulative CPU 0.99 sec
2016-06-08 15:47:10,629 Stage-2 map = 100%,  reduce = 33%, Cumulative CPU 0.99 sec
2016-06-08 15:47:11,632 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.26 sec
2016-06-08 15:47:12,634 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.26 sec
2016-06-08 15:47:13,636 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.26 sec
2016-06-08 15:47:14,639 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.26 sec
2016-06-08 15:47:15,643 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.26 sec
2016-06-08 15:47:16,645 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.26 sec
2016-06-08 15:47:17,647 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.26 sec
MapReduce Total cumulative CPU time: 2 seconds 260 msec
Ended Job = job_201606071417_0018
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 99.4 sec   HDFS Read: 7190255111 HDFS Write: 111621 SUCESS
Job 1: Map: 1  Reduce: 1   Accumulative CPU: 2.26 sec   HDFS Read: 114406 HDFS Write: 244 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 41 seconds 660 msec
OK
INTERSECTION    50283
CORONA  53011
WOODSIDE        53748
Astoria 59248
Flushing        69042
RIDGEWOOD       78118
Jamaica 99591
ASTORIA 103610
FLUSHING        131307
ADDRESS 132114
JAMAICA 162918
STATEN ISLAND   530774
        894719
BRONX   2002582
NEW YORK        2134706
BROOKLYN        3304821
Time taken: 339.575 seconds
hive>
```

Number of complaints by Complaint_type and Agency

SELECT AGENCY, COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY AGENCY, COMPLAINT_TYPE ORDER BY AGENCY,COMPLAINT_TYPE;

```
NYPD    Traffic 21937
DOT     Traffic Signal Condition        296151
DOB     Traffic Signal Condition        1
DOHMH   Trans Fat       23
DFTA    Transportation Provider Complaint       119
DOHMH   Trapping Pigeon 1
DOT     Tunnel Condition        65
HPD     UNSANITARY CONDITION    175990
DOHMH   Unleashed Dog   4973
DOHMH   Unlicensed Dog  1
DOHMH   Unsanitary Animal Facility      576
DOHMH   Unsanitary Animal Pvt Property  14486
DOHMH   Unsanitary Pigeon Condition     3992
DOHMH   Unspecified     1
DOB     Unspecified     1
NYPD    Urinating in Public     2702
HPD     VACANT APARTMENT        5
DSNY    Vacant Lot      13501
DOHMH   Vector  573
NYPD    Vending 24702
DPR     Violation of Park Rules 10740
HPD     WATER LEAK      78311
DEP     Water Conservation      24936
DEP     Water Quality   7100
DEP     Water System    398244
DOHMH   Window Guard    2611
DEP     X-Ray Machine/Equipment 1
DOHMH   X-Ray Machine/Equipment 80
Time taken: 336.439 seconds
hive>
```

HDFS 3 nodes Cluster with 1 x replication

HDFS 3 nodes Cluster with 1 x replication

```
hive> CREATE TABLE SERVICE_REQUEST (CASE_ID STRING, CREATED_DATE TIMESTAMP, CLOS
ED_DATE TIMESTAMP, AGENCY STRING, AGENCY_NAME STRING,COMPLAINT_TYPE STRING,DESCR
IPTOR STRING,LOCATION_TYPE STRING, INCIDENT_ZIP STRING,INCIDENT_ADDRESS STRING,S
TREET_NAME STRING, CROSS_STREET1 STRING, CROSS_STREET2 STRING, INTERSECTION_STRE
ET1 STRING, INTERSECTION_STREET2 STRING, ADDRESS_TYPE STRING, CITY STRING, LANMA
RK STRING, FACILITY_TYPE STRING,STATUS STRING, DUE_DATE TIMESTAMP, RESOLUTION_DE
SCRIPTION STRING, RESOLUTION_ACTION_UPDATE_DATE TIMESTAMP, COMMUNITY_BOARD STRIN
G,BOROUGH STRING, X_COORDINATE INT, Y_COORDINATE INT, PARK_FACILITY_NAME STRING,
 PARK_BOROUGH STRING,SCHOOL_NAME STRING, SCHOOL_NUMBER STRING, SCHOOL_REGION STR
ING, SCHOOL_CODE STRING, SCHOOL_PHONE_NUMBER STRING,SCHOOL_ADDRESS STRING, SCHOO
L_CITY STRING, SCHOOL_STATE STRING, SCHOOL_ZIP STRING, SCHOOL_NOT_FOUND STRING,S
CHOOL_OR_CITYWIDE_COMPLAINT STRING,  VEHICLE_TYPE STRING, TAXI_COMPANY_BOROUGH S
TRING, TAXI_PICK_UP_LOCATION STRING, BRIDGE_HIGHWAY_NAME STRING,BRIDGE_HIGHWAY_D
IRECTION STRING, ROAD_RAMP STRING, BRIDGE_HIGHWAY_SEGMENT STRING, GARAGE_LOT_NAM
E STRING, FERRY_DIRECTION STRING,FERRY_TERMINAL_NAME STRING, LATITUDE FLOAT, LON
GITUDE FLOAT, LOCATION_DETAIL STRING ) ROW FORMAT DELIMITED FIELDS TERMINATED BY
 ',' STORED AS TEXTFILE;
OK
Time taken: 7.18 seconds
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/311ServiceRequest.csv' OVERWRITE IN
TO TABLE SERVICE_REQUEST;
Copying data from file:/home/ec2-user/311ServiceRequest.csv
Copying file: file:/home/ec2-user/311ServiceRequest.csv
Loading data to table default.service_request
Deleted hdfs://ip-172-31-9-61.ec2.internal:9000/user/hive/warehouse/service_requ
est
OK
Time taken: 172.656 seconds
hive>
```

SELECT COUNT(*) FROM SERVICE_REQUEST;

```
MapReduce Total cumulative CPU time: 1 minutes 33 seconds 790 msec
Ended Job = job_201606071740_0002
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 1   Accumulative CPU: 93.79 sec   HDFS Read: 7187535209 HDFS Write: 9 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 33 seconds 790 msec
OK
11442923
Time taken: 155.547 seconds
hive>
```

Number of case requests by City

SELECT CITY, COUNT(*) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY CITY;

```
ec2-user@ip-172-31-9-61:~/hive-0.8.1-bin
WEST AVE        1
WEST AVENUE     1
WEST ORANGE     2
WESTBERRY       2
WESTBURRY       2
WESTCESTER      1
WESTCHESTER     7
WESTCHESTER AVE 3
WESTMINSTER ROAD        1
WESTPORT/FAIRFIELD      1
WILIMINGTON     1
WILLIAMSVILLIE  1
WILLIMINGTON    1
WOBURN  1
WOODBURY        55
WOODNERE        1
WOODPOINT RD    1
WYTHE AVE       1
YELLOWSTONE BLVD        1
YELLOWSTONE BOULEVARD   1
YORK AVE        4
ZACHARY COURT   1
Time taken: 174.55 seconds
hive>
```

Number of case requests by City which are more than 50000

SELECT * FROM (SELECT CITY, COUNT(*) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY CITY ORDER BY NUMBER_OF_CASES) X WHERE X.NUMBER_OF_CASES > 50000;

```
ec2-user@ip-172-31-9-61:~/hive-0.8.1-bin
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 134.44 sec   HDFS Read: 7187535209
 HDFS Write: 111621 SUCESS
Job 1: Map: 3  Reduce: 1   Accumulative CPU: 3.63 sec   HDFS Read: 114696 HDFS W
rite: 244 SUCESS
Total MapReduce CPU Time Spent: 2 minutes 18 seconds 70 msec
OK
INTERSECTION    50267
CORONA  52974
WOODSIDE        53729
Astoria 59221
Flushing        69015
RIDGEWOOD       78077
Jamaica 99548
ASTORIA 103570
FLUSHING        131265
ADDRESS 132080
JAMAICA 162857
STATEN ISLAND   530576
        894566
BRONX   2001896
NEW YORK        2133837
BROOKLYN        3303565
Time taken: 200.74 seconds
hive>
```

Naga Venkateshwarlu Yadav Dokku,
CSC-555 Project Report

Number of complaints by Complaint_type and Agency

SELECT AGENCY, COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS NUMBER_OF_CASES FROM SERVICE_REQUEST GROUP BY AGENCY, COMPLAINT_TYPE ORDER BY AGENCY,COMPLAINT_TYPE;

```
ec2-user@ip-172-31-9-61:~/hive-0.8.1-bin
OPS     Complaint       34
OPS     Compliment      2
OPS     Invitation      9
OPS     Misc. Comments  4
OPS     Opinion for the Mayor   1
OPS     Request for Information 2
SBS     Agency Issues   2
SBS     Compliment      2
SBS     Invitation      2
TFA     Request for Information 1
TLC     Agency Issues   90
TLC     For Hire Vehicle Complaint      19762
TLC     For Hire Vehicle Report 608
TLC     Found Property  2842
TLC     Lost Property   7
TLC     Taxi Complaint  119146
TLC     Taxi Compliment 4503
TLC     Taxi Report     4954
VAC     Complaint       1
VAC     Request for Information 3
WF1     Misc. Comments  1
WF1     Request for Information 2
Time taken: 200.772 seconds
hive>
```

Find the most common types of complaints – as part of this data analysis, I compared the

Performance using “order by”, “sort by” and “cluster by”.

Single –node

Find the most common types of complaints – as part of this data analysis, I compared the

performance using “order by”, “sort by” and “cluster by”.

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE SORT BY CNT DESC LIMIT 5;

```
ec2-user@ip-172-31-8-56:~/hive-0.8.1-bin
2016-06-09 11:45:07,882 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.33
sec
2016-06-09 11:45:08,886 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.33
sec
2016-06-09 11:45:09,888 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.33
sec
MapReduce Total cumulative CPU time: 1 seconds 330 msec
Ended Job = job_201606081727_0010
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 98.53 sec   HDFS Read: 7190255111
HDFS Write: 11560 SUCESS
Job 1: Map: 1  Reduce: 1   Accumulative CPU: 1.38 sec   HDFS Read: 14345 HDFS Wr
ite: 279 SUCESS
Job 2: Map: 1  Reduce: 1   Accumulative CPU: 1.33 sec   HDFS Read: 754 HDFS Writ
e: 113 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 41 seconds 240 msec
OK
HEATING 887869
Street Light Condition  637766
Street Condition        635750
PLUMBING        540269
GENERAL CONSTRUCTION    500863
Time taken: 375.105 seconds
hive>
```

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE ORDER BY CNT DESC LIMIT 5;

```
ec2-user@ip-172-31-8-56:~/hive-0.8.1-bin
2016-06-09 11:53:07,700 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.45
sec
2016-06-09 11:53:08,703 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.45
sec
2016-06-09 11:53:09,712 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.45
sec
2016-06-09 11:53:10,715 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.45
sec
MapReduce Total cumulative CPU time: 1 seconds 450 msec
Ended Job = job_201606081727_0012
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 98.13 sec   HDFS Read: 7190255111
HDFS Write: 11560 SUCESS
Job 1: Map: 1  Reduce: 1   Accumulative CPU: 1.45 sec   HDFS Read: 14345 HDFS Wr
ite: 113 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 39 seconds 580 msec
OK
HEATING 887869
Street Light Condition  637766
Street Condition        635750
PLUMBING        540269
GENERAL CONSTRUCTION    500863
Time taken: 338.735 seconds
hive>
```

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE CLUSTER BY CNT LIMIT 5;

```
ec2-user@ip-172-31-8-56:~/hive-0.8.1-bin
2016-06-09 12:02:14,382 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.3 s
ec
2016-06-09 12:02:15,384 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.3 s
ec
2016-06-09 12:02:16,387 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.3 s
ec
MapReduce Total cumulative CPU time: 1 seconds 300 msec
Ended Job = job_201606081727_0015
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 99.95 sec   HDFS Read: 7190255111
HDFS Write: 11560 SUCESS
Job 1: Map: 1  Reduce: 1   Accumulative CPU: 1.39 sec   HDFS Read: 14345 HDFS Wr
ite: 261 SUCESS
Job 2: Map: 1  Reduce: 1   Accumulative CPU: 1.3 sec   HDFS Read: 736 HDFS Write
: 85 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 42 seconds 640 msec
OK
SG-99   1
 Computer Applications  1
Asbestos/Garbage Nuisance       1
Trapping Pigeon 1
SNW     1
Time taken: 369.178 seconds
hive>
```

HEATING 887869,Street Light Condition 637718,Street Condition 635620,PLUMBING 540195

GENERAL CONSTRUCTION 500863

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE SORT BY CNT DESC LIMIT 5;

```
MapReduce Total cumulative CPU time: 1 seconds 270 msec
Ended Job = job_201606071740_0018
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 133.06 sec   HDFS Read: 7187535209 HDFS Write: 11560 SUCESS
Job 1: Map: 3  Reduce: 1   Accumulative CPU: 2.55 sec   HDFS Read: 14635 HDFS Write: 279 SUCESS
Job 2: Map: 1  Reduce: 1   Accumulative CPU: 1.27 sec   HDFS Read: 754 HDFS Write: 113 SUCESS
Total MapReduce CPU Time Spent: 2 minutes 16 seconds 880 msec
OK
HEATING 887869
Street Light Condition  637718
Street Condition        635620
PLUMBING        540195
GENERAL CONSTRUCTION    500863
Time taken: 231.798 seconds
hive>
```

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE ORDER BY CNT DESC LIMIT 5;

```
MapReduce Total cumulative CPU time: 2 seconds 610 msec
Ended Job = job_201606071740_0020
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 133.05 sec   HDFS Read: 7187535209 HDFS Write: 11560 SUCESS
Job 1: Map: 3  Reduce: 1   Accumulative CPU: 2.61 sec   HDFS Read: 14635 HDFS Write: 113 SUCESS
Total MapReduce CPU Time Spent: 2 minutes 15 seconds 660 msec
OK
HEATING 887869
Street Light Condition  637718
Street Condition        635620
PLUMBING        540195
GENERAL CONSTRUCTION    500863
Time taken: 189.322 seconds
hive>
```

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE CLUSTER BY CNT LIMIT 5;

```
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 133.38 sec   HDFS Read: 7187535209 HDFS Write: 11560 SUCESS
Job 1: Map: 3  Reduce: 1   Accumulative CPU: 2.55 sec   HDFS Read: 14635 HDFS Write: 279 SUCESS
Job 2: Map: 1  Reduce: 1   Accumulative CPU: 1.32 sec   HDFS Read: 754 HDFS Write: 103 SUCESS
Total MapReduce CPU Time Spent: 2 minutes 17 seconds 250 msec
OK
Micro Switch    1
 Computer Applications  1
Asbestos/Garbage Nuisance       1
Trapping Pigeon 1
Complaint Type  1
Time taken: 242.934 seconds
hive>
```

Micro Switch 1, Computer Applications 1, Asbestos/Garbage Nuisance 1,Trapping Pigeon 1,Complaint Type 1

HDFS 3 nodes Cluster with 3 x replications

Find the most common types of complaints – as part of this data analysis, I compared the

performance using "order by", "sort by" and "cluster by".

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE SORT BY CNT DESC LIMIT 5;

```
ec2-user@ip-172-31-5-233:~/hive-0.8.1-bin
2016-06-09 11:51:37,251 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.4 s
ec
2016-06-09 11:51:38,254 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.4 s
ec
2016-06-09 11:51:39,258 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.4 s
ec
MapReduce Total cumulative CPU time: 1 seconds 400 msec
Ended Job = job_201606071959_0020
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 104.74 sec   HDFS Read: 7187535317
 HDFS Write: 11560 SUCESS
Job 1: Map: 1  Reduce: 1   Accumulative CPU: 1.39 sec   HDFS Read: 14353 HDFS Wr
ite: 279 SUCESS
Job 2: Map: 1  Reduce: 1   Accumulative CPU: 1.4 sec   HDFS Read: 755 HDFS Write
: 113 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 47 seconds 530 msec
OK
HEATING 887869
Street Light Condition  637718
Street Condition        635620
PLUMBING        540195
GENERAL CONSTRUCTION    500863
Time taken: 189.457 seconds
hive>
```

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE ORDER BY CNT DESC LIMIT 5;

```
ec2-user@ip-172-31-5-233:~/hive-0.8.1-bin
2016-06-09 11:57:20,350 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.42
sec
2016-06-09 11:57:21,354 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.42
sec
2016-06-09 11:57:22,357 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.42
sec
2016-06-09 11:57:23,359 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 1.42
sec
MapReduce Total cumulative CPU time: 1 seconds 420 msec
Ended Job = job_201606071959_0022
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 105.85 sec   HDFS Read: 7187535317
 HDFS Write: 11560 SUCESS
Job 1: Map: 1  Reduce: 1   Accumulative CPU: 1.42 sec   HDFS Read: 14353 HDFS Wr
ite: 113 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 47 seconds 270 msec
OK
HEATING 887869
Street Light Condition  637718
Street Condition        635620
PLUMBING        540195
GENERAL CONSTRUCTION    500863
Time taken: 157.461 seconds
hive>
```

SELECT COMPLAINT_TYPE, COUNT(COMPLAINT_TYPE) AS CNT FROM SERVICE_REQUEST GROUP BY COMPLAINT_TYPE CLUSTER BY CNT LIMIT 5;

ec2-user@ip-172-31-5-233:~/hive-0.8.1-bin

```
sec
2016-06-09 12:02:05,519 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.44
sec
2016-06-09 12:02:06,521 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 1.44
sec
MapReduce Total cumulative CPU time: 1 seconds 440 msec
Ended Job = job_201606071959_0025
MapReduce Jobs Launched:
Job 0: Map: 27  Reduce: 8   Accumulative CPU: 104.99 sec   HDFS Read: 7187535317
 HDFS Write: 11560 SUCESS
Job 1: Map: 1  Reduce: 1   Accumulative CPU: 1.41 sec   HDFS Read: 14353 HDFS Wr
ite: 261 SUCESS
Job 2: Map: 1  Reduce: 1   Accumulative CPU: 1.44 sec   HDFS Read: 737 HDFS Writ
e: 85 SUCESS
Total MapReduce CPU Time Spent: 1 minutes 47 seconds 840 msec
OK
SG-99   1
 Computer Applications  1
Asbestos/Garbage Nuisance       1
Trapping Pigeon 1
SNW     1
Time taken: 189.785 seconds
hive>
    >
```