Big Data Analysis using Hadoop

-By Shubham Mehta

Objective: The objective of this project is to demonstrate the ability and usability of the Hadoop framework for analyzing large volumes of data (Big Data).

Architecture

Operating System	Stand-alone Ubuntu-server-12.04 running in VMWare on Windows 10, 64-bit
Memory	2GB
IDE	Eclipse
Java	1.6
Hadoop Release	1.0.3
Other Tools	Putty, WinSCP

Source of Data: **https://www.kaggle.com/dansbecker/nba-shot-logs**

OR

https://drive.google.com/file/d/0B0j0S5rFLpGxTGxDbG53VHdFTnM/view?usp=sharing

Size of Data: 15.53 MB (Approximately 128,070 records)

Work Objective

Use Hive to do the following:

1) Find the number of shots taken per game for the 14 – 15 NBA Season
2) Find who made the greatest number of shots per game
3) Find who made the greatest number of 2s per game and who made the greatest number of 3s per game
4) Find the defender who has the best blocking percentage, per game
5) Find the top 10 scorers of the entire 14 – 15 season

Solution

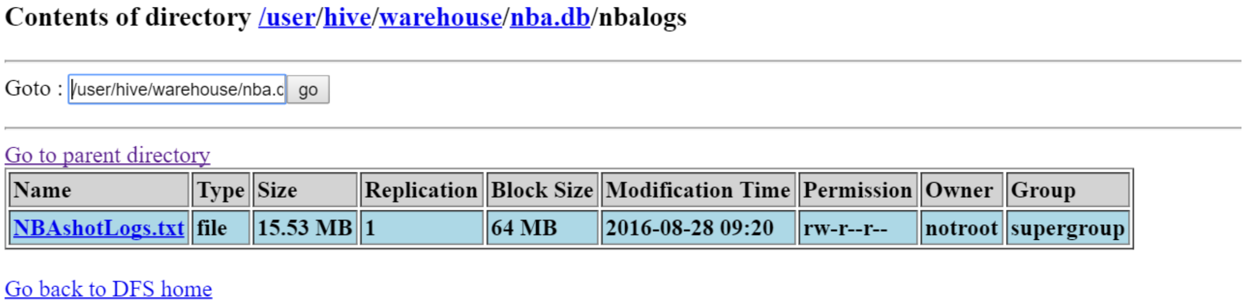
1. Copy file from Window to Ubuntu using WinSCP
2. For file ingestion, run the command: hadoop fs -copyFromLocal /home/notroot/lab/data/NBAshotLogs.txt /input
3. Table creation:
   - Hive
   - create database nba;
   - use nba;

➢ create table nbaLogs (game_id INT,matchup STRING, loc STRING, w STRING, final_margin INT, shot_number INT, period INT, game_clock STRING, shot_clock STRING, dribbles INT, touch_time STRING, shot_dist STRING, pts_type INT, shot_result STRING, closest_defender STRING, closest_defender_player_id INT, close_def_dist STRING, fgm INT, pts INT, player_name STRING, player_id INT)

> row format delimited
> Fields terminated by '\t'
> Stored as textfile;

> LOAD DATA LOCAL INPATH ‘/home/notroot/lab/data/NBAshotLogs.txt’ INTO TABLE nbaLogs;

Web UI Display:



Contents of directory /user/hive/warehouse/nba.db/nbalogs

Goto : /user/hive/warehouse/nba.c go

Go to parent directory

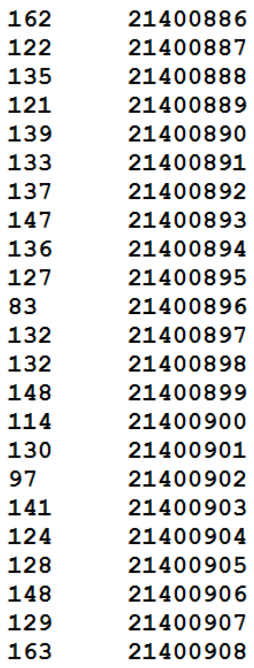
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
NBAshotLogs.txt	file	15.53 MB	1	64 MB	2016-08-28 09:20	rw-r--r--	notroot	supergroup

Go back to DFS home

4. Hive Queries for finding the number of shots taken per game for the 14 – 15 NBA Season

- **Select count(game_id),game_id**
- **from nbaLogs**
- **group by game_id;**

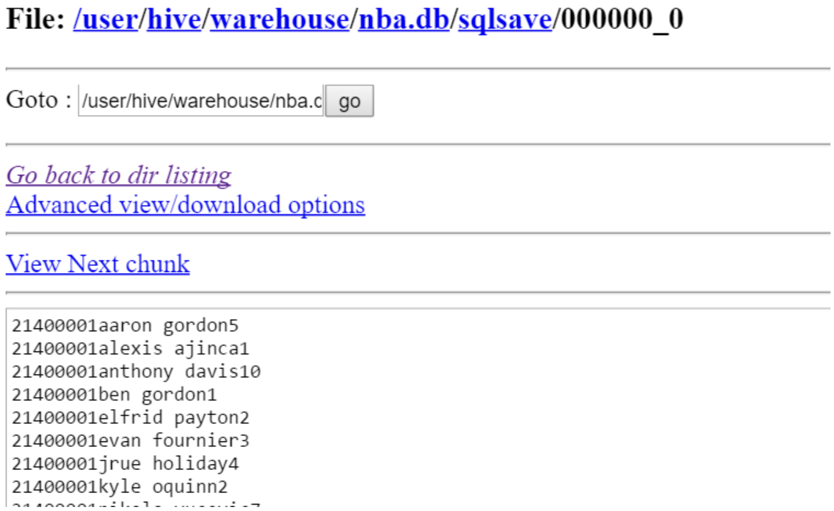
Output (908 games)



```
162     21400886
122     21400887
135     21400888
121     21400889
139     21400890
133     21400891
137     21400892
147     21400893
136     21400894
127     21400895
83      21400896
132     21400897
132     21400898
148     21400899
114     21400900
130     21400901
97      21400902
141     21400903
124     21400904
128     21400905
148     21400906
129     21400907
163     21400908
```

5. Hive Query for finding who MADE the greatest number of shots per game

- **We will first find how many points each person scored in each game and the save the result into a table named sqlsave**
- **Create table sqlsave as select game_id,player_name,COUNT(shot_result) cnt from nbaLogs where shot_result='made' group by game_id,player_name;**



File: /user/hive/warehouse/nba.db/sqlsave/000000_0

Goto : /user/hive/warehouse/nba.c
go
Go back to dir listing
Advanced view/download options
View Next chunk

```
21400001aaron gordon5
21400001alexis ajinca1
21400001anthony davis10
21400001ben gordon1
21400001elfrid payton2
21400001evan fournier3
21400001jrue holiday4
21400001kyle oquinn2
21400001nikola vucevic7
```

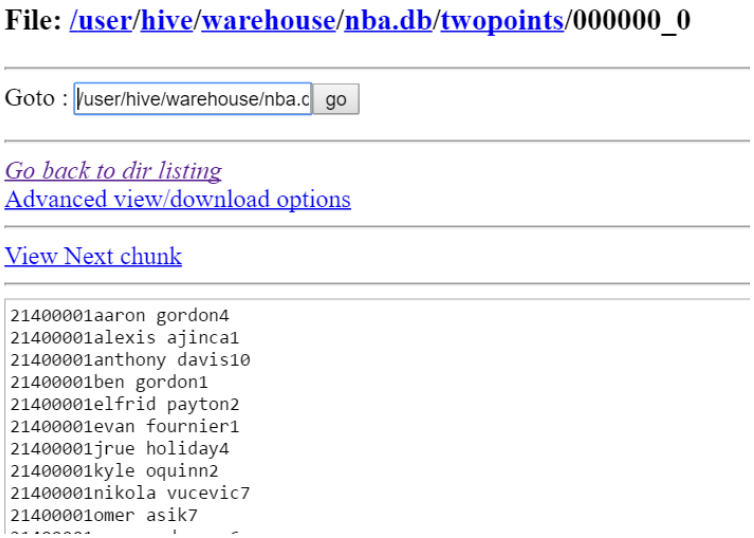
-

- **We well then apply the following query onto sqlsave**
- **select a.game_id,a.player_name,a.cnt from sqlsave a join(select game_id,MAX(cnt) cnt from sqlsave group by game_id) b on a.game_id=b.game_id and a.cnt=b.cnt;**

```
21400896        rodney stuckey  7
21400897        victor oladipo  15
21400898        derrick favors  7
21400898        trey burke      7
21400899        al jefferson    8
21400900        jonas valanciunas       12
21400901        marc gasol      10
21400902        andrew wiggins  8
21400902        kenneth faried  8
21400903        anthony davis   17
21400904        russell westbrook       16
21400905        ed davis        7
21400905        dwayne wade     7
21400906        kawhi leonard   9
21400907        draymond green  9
21400908        chris paul      14
```

6. Hive Query for finding who made the greatest number of 2s per game and who made the greatest number of 3s per game

- **create table twoPoints as select game_id,player_name,COUNT(shot_result) cnt from nbaLogs where shot_result='made' AND pts_type='2' group by game_id,player_name;**
- **Here is what table twoPoints looks like:**

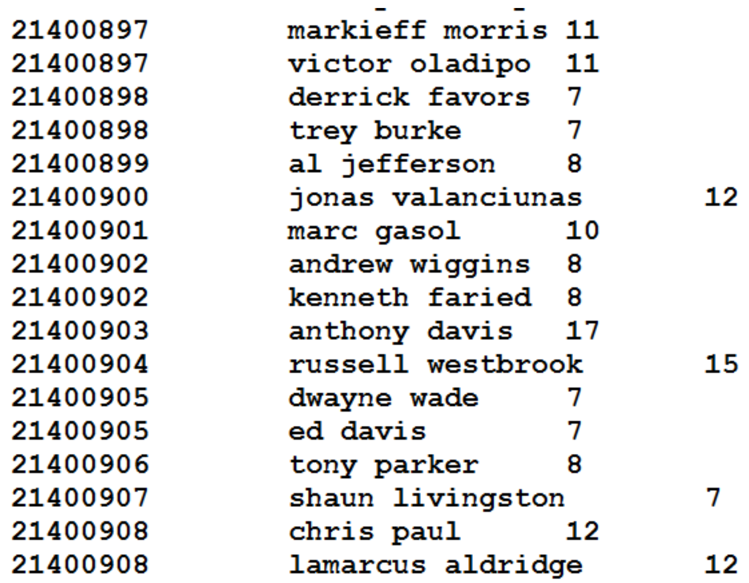


File: /user/hive/warehouse/nba.db/twopoints/000000_0

Goto : /user/hive/warehouse/nba.c
go
Go back to dir listing
Advanced view/download options
View Next chunk

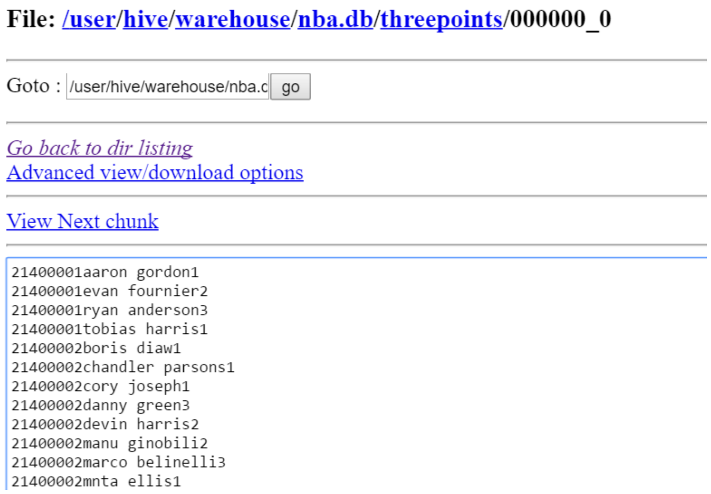
```
21400001aaron gordon4
21400001alexis ajinca1
21400001anthony davis10
21400001ben gordon1
21400001elfrid payton2
21400001evan fournier1
21400001jrue holiday4
21400001kyle oquinn2
21400001nikola vucevic7
21400001omer asik7
```

- **We will then apply the following query onto twoPoints:**
- **select a.game_id,a.player_name,a.cnt from twoPoints a JOIN (select game_id,MAX(cnt) abc from twoPoints GROUP BY game_id) b where a.game_id=b.game_id AND a.cnt=b.abc;**
- **Output:**



```
21400897        markieff morris 11
21400897        victor oladipo  11
21400898        derrick favors  7
21400898        trey burke      7
21400899        al jefferson    8
21400900        jonas valanciunas       12
21400901        marc gasol      10
21400902        andrew wiggins  8
21400902        kenneth faried  8
21400903        anthony davis   17
21400904        russell westbrook       15
21400905        dwayne wade     7
21400905        ed davis        7
21400906        tony parker     8
21400907        shaun livingston        7
21400908        chris paul      12
21400908        lamarcus aldridge       12
```

- **Now for the greatest number of 3s per game**
- **create table threePoints as select game_id,player_name,COUNT(shot_result) cnt from nbaLogs where shot_result='made' AND pts_type='3' group by game_id,player_name;**

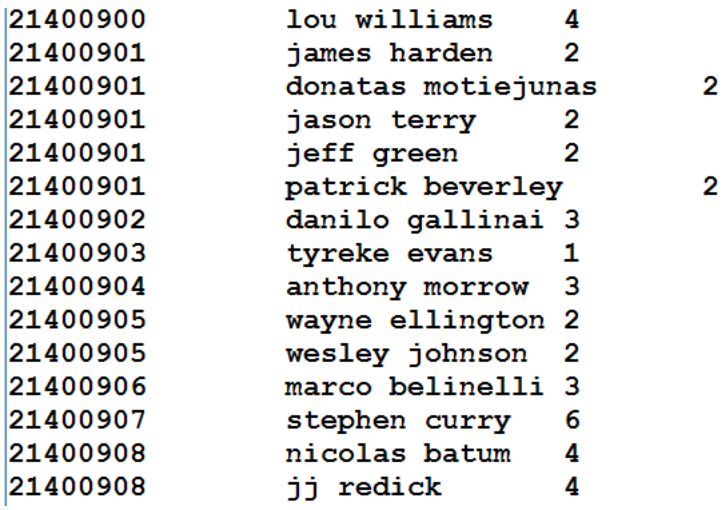


File: /user/hive/warehouse/nba.db/threepoints/000000_0

Goto : /user/hive/warehouse/nba.c
go
Go back to dir listing
Advanced view/download options
View Next chunk

```
21400001aaron gordon1
21400001evan fournier2
21400001ryan anderson3
21400001tobias harris1
21400002boris diaw1
21400002chandler parsons1
21400002cory joseph1
21400002danny green3
21400002devin harris2
21400002manu ginobili2
21400002marco belinelli3
21400002mnta ellis1
```

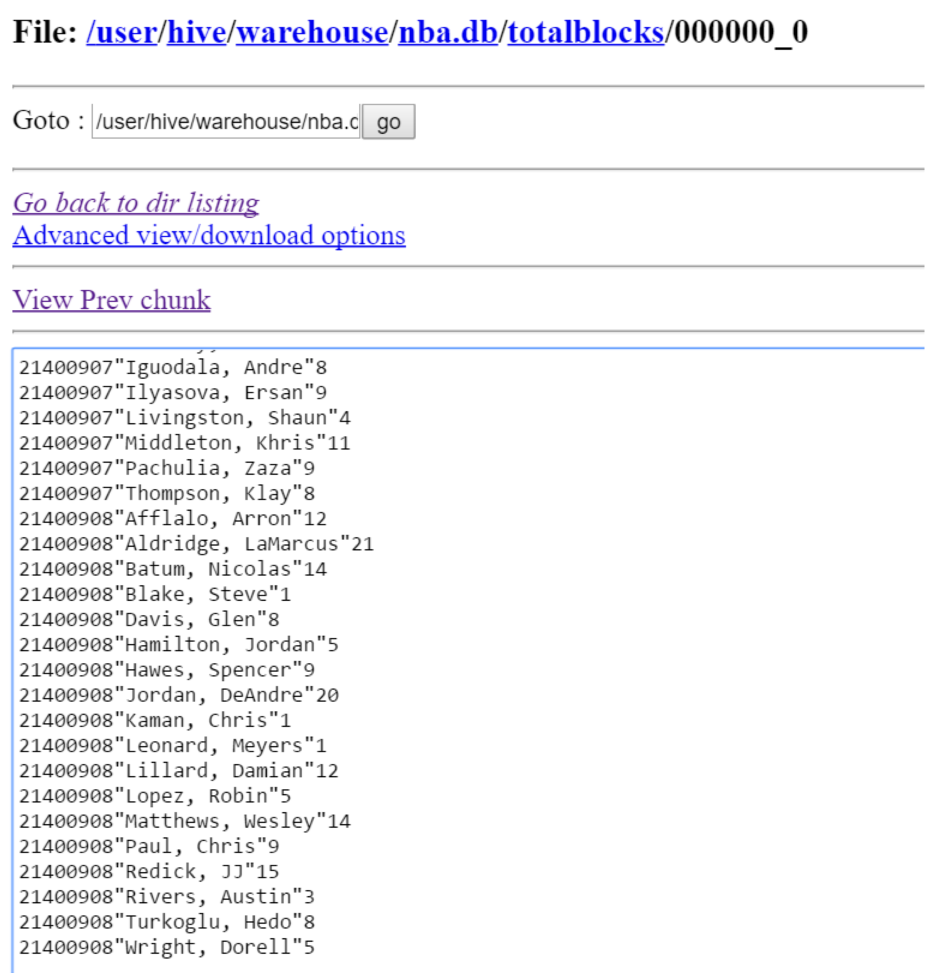
- **We will then apply the following query onto threePoints:**
- **select a.game_id,a.player_name,a.cnt from threePoints a JOIN (select game_id,MAX(cnt) abc from threePoints GROUP BY game_id) b where a.game_id=b.game_id AND a.cnt=b.abc;**



```
21400900        lou williams    4
21400901        james harden    2
21400901        donatas motiejunas      2
21400901        jason terry     2
21400901        jeff green      2
21400901        patrick beverley        2
21400902        danilo gallinai 3
21400903        tyreke evans    1
21400904        anthony morrow  3
21400905        wayne ellington 2
21400905        wesley johnson  2
21400906        marco belinelli 3
21400907        stephen curry   6
21400908        nicolas batum   4
21400908        jj redick       4
```

7. Hive Query for finding the defender who has the best blocking percentage, per game

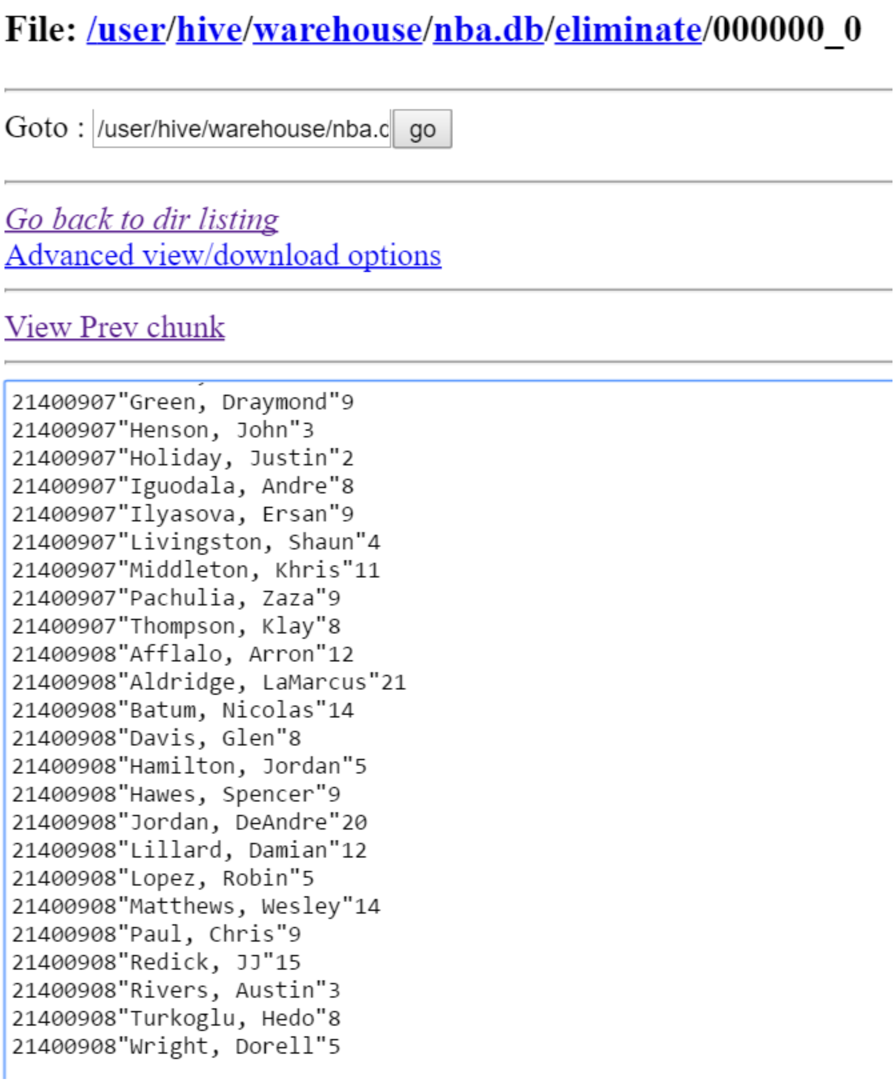
- **First let us create a table that tells us how many shots were taken over each defender, per game**
- **create table totalblocks as select game_id,closest_defender,COUNT(shot_result) cnt from nbaLogs group by game_id,closest_defender;**


Goto : /user/hive/warehouse/nba.c
go
Go back to dir listing
Advanced view/download options
View Prev chunk

File: /user/hive/warehouse/nba.db/totalblocks/000000_0

```
21400907"Iguodala, Andre"8
21400907"Ilyasova, Ersan"9
21400907"Livingston, Shaun"4
21400907"Middleton, Khris"11
21400907"Pachulia, Zaza"9
21400907"Thompson, Klay"8
21400908"Afflalo, Arron"12
21400908"Aldridge, LaMarcus"21
21400908"Batum, Nicolas"14
21400908"Blake, Steve"1
21400908"Davis, Glen"8
21400908"Hamilton, Jordan"5
21400908"Hawes, Spencer"9
21400908"Jordan, DeAndre"20
21400908"Kaman, Chris"1
21400908"Leonard, Meyers"1
21400908"Lillard, Damian"12
21400908"Lopez, Robin"5
21400908"Matthews, Wesley"14
21400908"Paul, Chris"9
21400908"Redick, JJ"15
21400908"Rivers, Austin"3
21400908"Turkoglu, Hedo"8
21400908"Wright, Dorell"5
```

- **Now we will eliminate those records where only 1 shot was taken over a defender as it is impractical to calculate the blocking percentage of a defender who has only guarded a single shot throughout the game.**
- **create table eliminate as select game_id,closest_defender,cnt from totalblocks where cnt>1;**

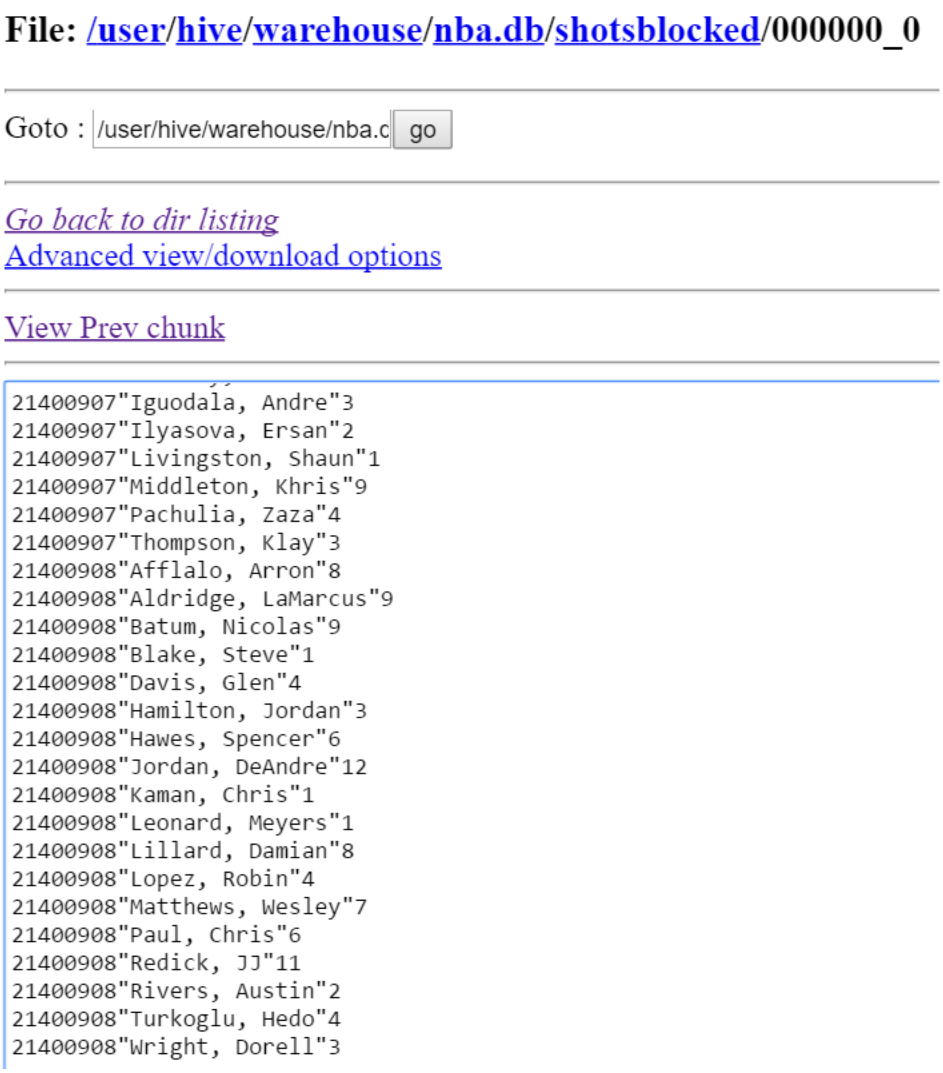


File: /user/hive/warehouse/nba.db/eliminate/000000_0

Goto : /user/hive/warehouse/nba.c
go
Go back to dir listing
Advanced view/download options
View Prev chunk

```
21400907"Green, Draymond"9
21400907"Henson, John"3
21400907"Holiday, Justin"2
21400907"Iguodala, Andre"8
21400907"Ilyasova, Ersan"9
21400907"Livingston, Shaun"4
21400907"Middleton, Khris"11
21400907"Pachulia, Zaza"9
21400907"Thompson, Klay"8
21400908"Afflalo, Arron"12
21400908"Aldridge, LaMarcus"21
21400908"Batum, Nicolas"14
21400908"Davis, Glen"8
21400908"Hamilton, Jordan"5
21400908"Hawes, Spencer"9
21400908"Jordan, DeAndre"20
21400908"Lillard, Damian"12
21400908"Lopez, Robin"5
21400908"Matthews, Wesley"14
21400908"Paul, Chris"9
21400908"Redick, JJ"15
21400908"Rivers, Austin"3
21400908"Turkoglu, Hedo"8
21400908"Wright, Dorell"5
```

- **Now we will create a table that will tell us how many shots each defender blocked per game (i.e. shots that the shooter missed)**
- **create table shotsblocked as select game_id,closest_defender,COUNT(shot_result) cnt from nbaLogs where shot_result='missed' group by game_id,closest_defender;**



File: /user/hive/warehouse/nba.db/shotsblocked/000000_0

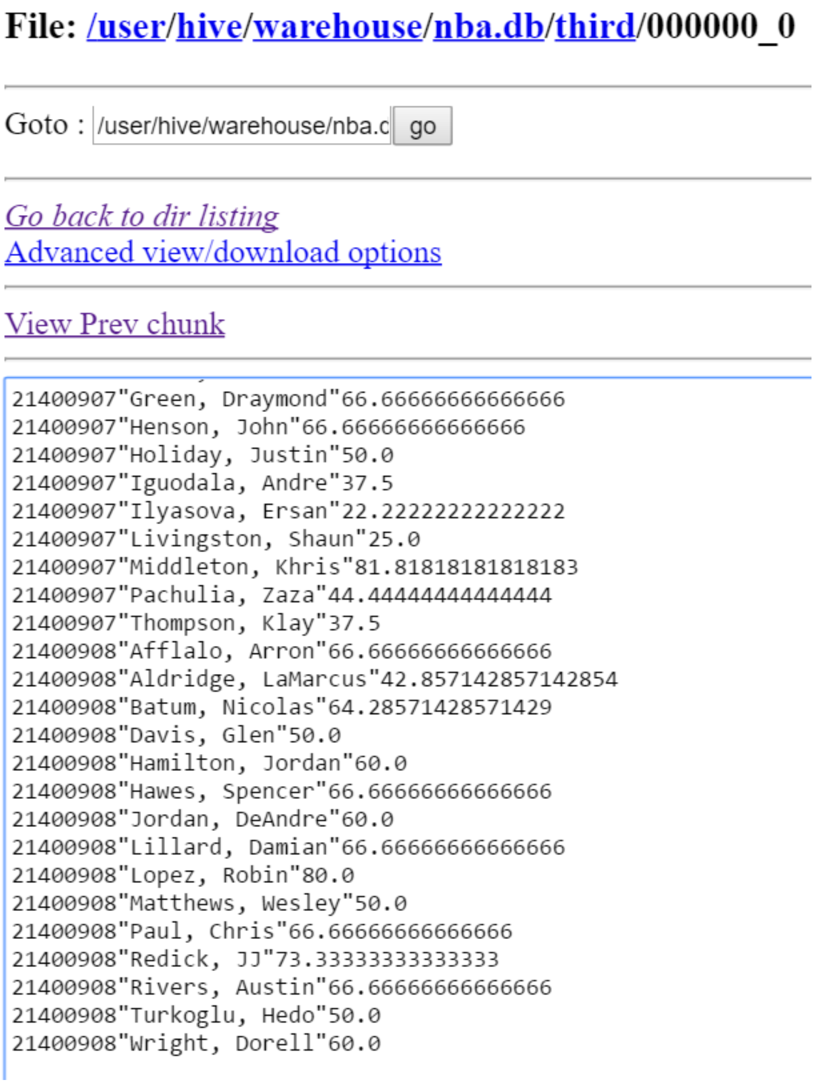
Goto : /user/hive/warehouse/nba.c go

Go back to dir listing
Advanced view/download options

View Prev chunk

```
21400907"Iguodala, Andre"3
21400907"Ilyasova, Ersan"2
21400907"Livingston, Shaun"1
21400907"Middleton, Khris"9
21400907"Pachulia, Zaza"4
21400907"Thompson, Klay"3
21400908"Afflalo, Arron"8
21400908"Aldridge, LaMarcus"9
21400908"Batum, Nicolas"9
21400908"Blake, Steve"1
21400908"Davis, Glen"4
21400908"Hamilton, Jordan"3
21400908"Hawes, Spencer"6
21400908"Jordan, DeAndre"12
21400908"Kaman, Chris"1
21400908"Leonard, Meyers"1
21400908"Lillard, Damian"8
21400908"Lopez, Robin"4
21400908"Matthews, Wesley"7
21400908"Paul, Chris"6
21400908"Redick, JJ"11
21400908"Rivers, Austin"2
21400908"Turkoglu, Hedo"4
21400908"Wright, Dorell"3
```

- **Now we will join these two tables to create a third table which will store the blocking percentage for each defender per game**
- **create table third as select a.game_id,a.closest_defender,(b.cnt/a.cnt)*100 cnt from eliminate a inner join shotsblocked b on a.game_id=b.game_id and a.closest_defender=b.closest_defender;**

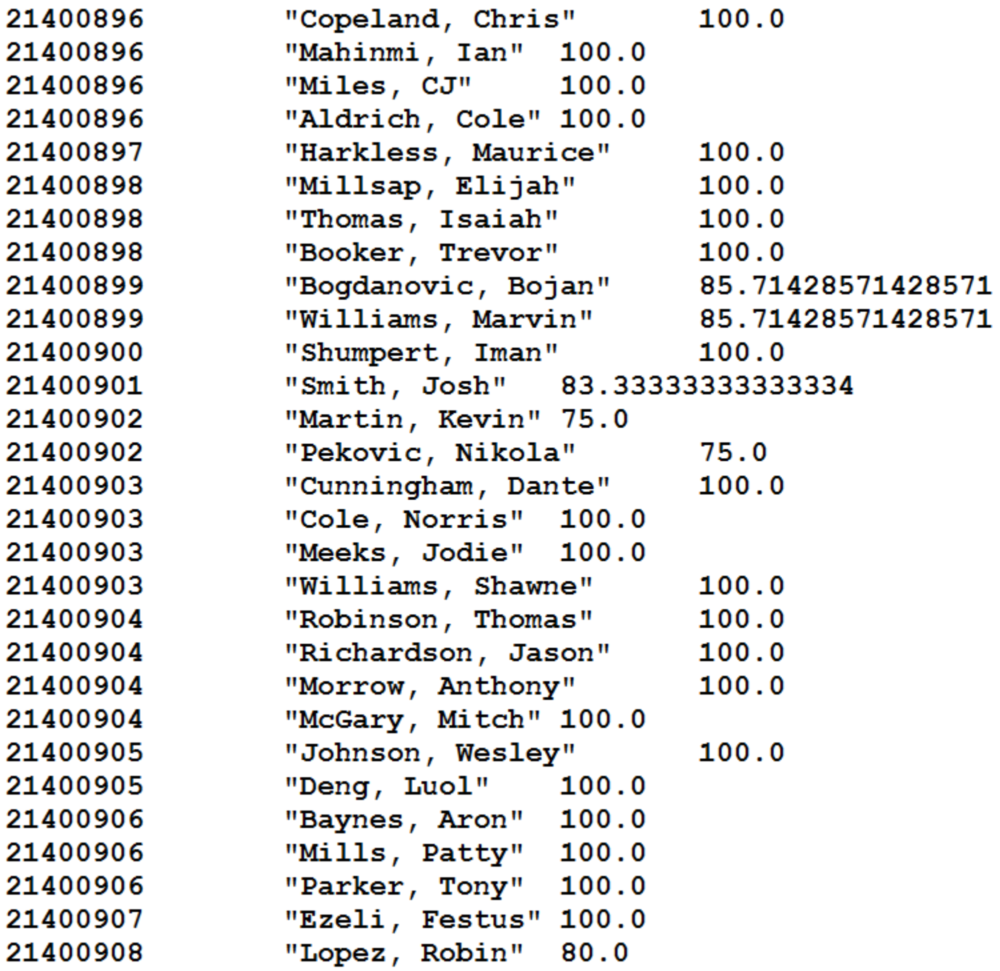


File: /user/hive/warehouse/nba.db/third/000000_0

Goto : /user/hive/warehouse/nba.c
go
Go back to dir listing
Advanced view/download options
View Prev chunk

```
21400907"Green, Draymond"66.66666666666666
21400907"Henson, John"66.66666666666666
21400907"Holiday, Justin"50.0
21400907"Iguodala, Andre"37.5
21400907"Ilyasova, Ersan"22.22222222222222
21400907"Livingston, Shaun"25.0
21400907"Middleton, Khris"81.81818181818183
21400907"Pachulia, Zaza"44.44444444444444
21400907"Thompson, Klay"37.5
21400908"Afflalo, Arron"66.66666666666666
21400908"Aldridge, LaMarcus"42.857142857142854
21400908"Batum, Nicolas"64.28571428571429
21400908"Davis, Glen"50.0
21400908"Hamilton, Jordan"60.0
21400908"Hawes, Spencer"66.66666666666666
21400908"Jordan, DeAndre"60.0
21400908"Lillard, Damian"66.66666666666666
21400908"Lopez, Robin"80.0
21400908"Matthews, Wesley"50.0
21400908"Paul, Chris"66.66666666666666
21400908"Redick, JJ"73.33333333333333
21400908"Rivers, Austin"66.66666666666666
21400908"Turkoglu, Hedo"50.0
21400908"Wright, Dorell"60.0
```

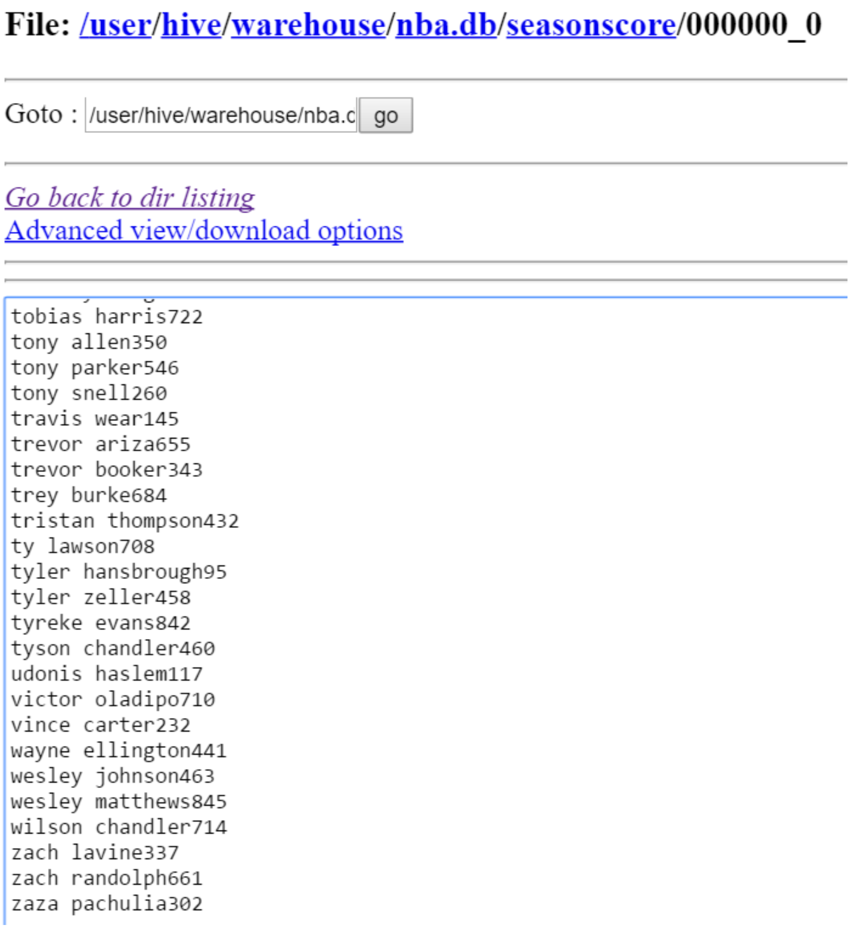
- **Now we can use table third to find the player with the max blocking percentage per game using the following query:**
- **select a.game_id,a.closest_defender,a.cnt from third a join (select game_id,MAX(cnt) cnt from third group by game_id) b on a.game_id=b.game_id and a.cnt=b.cnt;**



```
21400896        "Copeland, Chris"       100.0
21400896        "Mahinmi, Ian"  100.0
21400896        "Miles, CJ"     100.0
21400896        "Aldrich, Cole" 100.0
21400897        "Harkless, Maurice"     100.0
21400898        "Millsap, Elijah"       100.0
21400898        "Thomas, Isaiah"        100.0
21400898        "Booker, Trevor"        100.0
21400899        "Bogdanovic, Bojan"     85.71428571428571
21400899        "Williams, Marvin"      85.71428571428571
21400900        "Shumpert, Iman"        100.0
21400901        "Smith, Josh"   83.33333333333334
21400902        "Martin, Kevin" 75.0
21400902        "Pekovic, Nikola"       75.0
21400903        "Cunningham, Dante"     100.0
21400903        "Cole, Norris"  100.0
21400903        "Meeks, Jodie"  100.0
21400903        "Williams, Shawne"      100.0
21400904        "Robinson, Thomas"      100.0
21400904        "Richardson, Jason"     100.0
21400904        "Morrow, Anthony"       100.0
21400904        "McGary, Mitch" 100.0
21400905        "Johnson, Wesley"       100.0
21400905        "Deng, Luol"    100.0
21400906        "Baynes, Aron"  100.0
21400906        "Mills, Patty"  100.0
21400906        "Parker, Tony"  100.0
21400907        "Ezeli, Festus" 100.0
21400908        "Lopez, Robin"  80.0
```

8. Hive Queries for finding the top 10 scorers of the entire 14 – 15 season

- **Create table seasonscore as select player_name,SUM(pts) cnt from nbaLogs group by player_name;**



File: /user/hive/warehouse/nba.db/seasonscore/000000_0

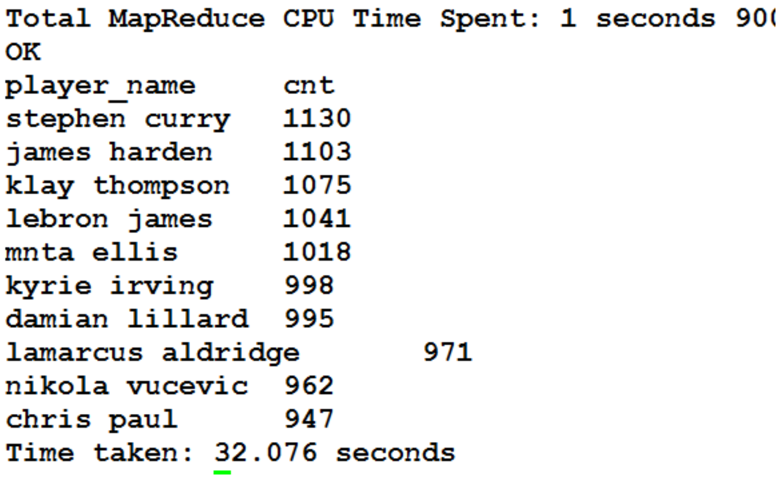
Goto : /user/hive/warehouse/nba.c go

Go back to dir listing
Advanced view/download options

```
tobias harris722
tony allen350
tony parker546
tony snell260
travis wear145
trevor ariza655
trevor booker343
trey burke684
tristan thompson432
ty lawson708
tyler hansbrough95
tyler zeller458
tyreke evans842
tyson chandler460
udonis haslem117
victor oladipo710
vince carter232
wayne ellington441
wesley johnson463
wesley matthews845
wilson chandler714
zach lavine337
zach randolph661
zaza pachulia302
```

-

- **select * from (select * from seasonscore order by cnt desc) b limit 10;**
  Here are the top 10 scorers:



```
Total MapReduce CPU Time Spent: 1 seconds 90(
OK
player_name     cnt
stephen curry   1130
james harden    1103
klay thompson   1075
lebron james    1041
mnta ellis      1018
kyrie irving    998
damian lillard  995
lamarcus aldridge       971
nikola vucevic  962
chris paul      947
Time taken: 32.076 seconds
```