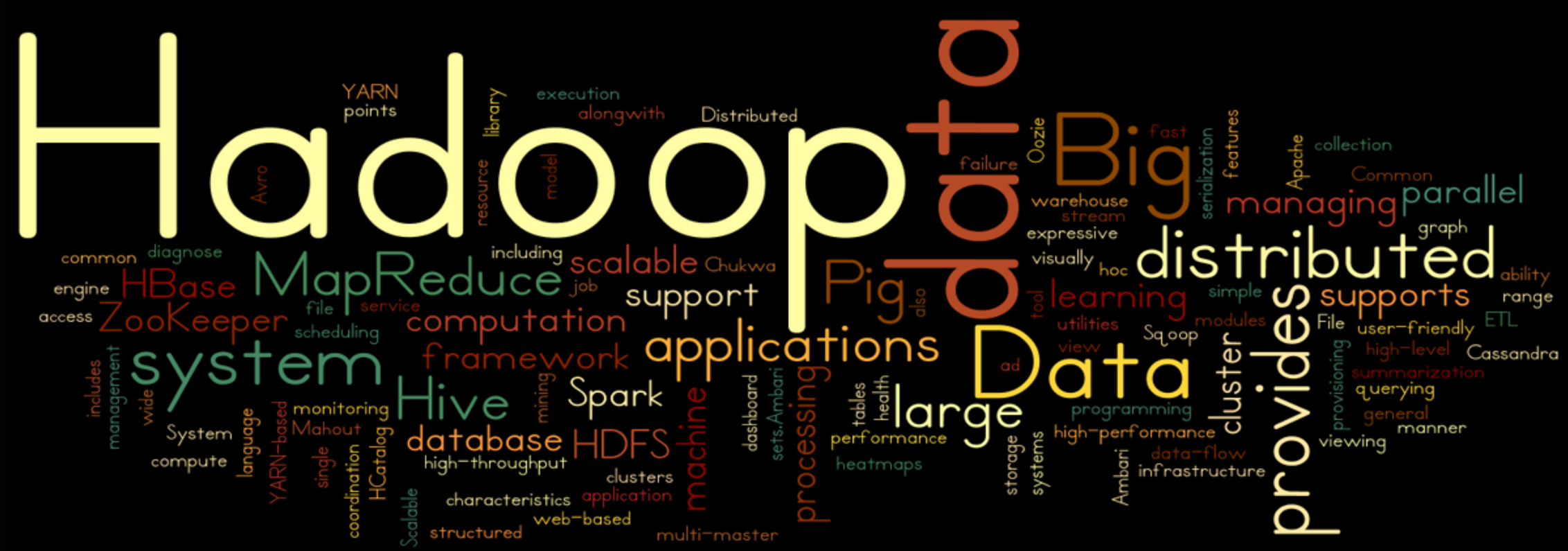

Hadoop
data
Big
Data
Pig
MapReduce
HBase
ZooKeeper
system
Hive
Spark
HDFS
database
applications
framework
computation
support
scalable
large
machine
processing
distributed
provides
managing
parallel
supports
learning
cluster
YARN
points
execution
alongwith
Distributed
library
resource
model
Avro
Oozie
fast
failure
warehouse
stream
expressive
visually
hoc
serialization
features
Apache
collection
Common
graph
ability
range
simple
utilities
modules
Sqoop
view
ad
tool
also
including
Chukwa
job
common
diagnose
engine
access
file
service
scheduling
includes
management
wide
System
compute
language
YARN-based
Mahout
monitoring
single
coordination
HCatalog
Scalable
high-throughput
clusters
characteristics
application
web-based
structured
multi-master
mining
dashboard
sets.Ambari
tables
health
performance
heatmaps
programming
high-performance
data-flow
infrastructure
storage
systems
Ambari
File
user-friendly
ETL
high-level
Cassandra
summarization
querying
general
manner
viewing
provisioning

FRAMEWORK

- Hadoop
- Introduction to Pig and Hive
- Explanation to stock market data
- Explanation to US election data
- Analysis using Pig
- Analysis using Hive
- Results
- Conclusion

BIG DATA

Big data is an all-inclusive term that refers to extremely large, very fast, highly diverse, and complex data that cannot be managed with traditional data management tools.

Four V’s of big data:

- Volume
- Variety
- Velocity
- Veracity

Introduction to Hadoop

What is Hadoop

- Open-source software
- Framework for storing data and running applications on clusters of commodity hardware.
- Provides massive storage for any kind of data, enormous processing power.

What we've got

- Handling failure
- Petabytes of data to be pocessed in parallel
- Easy scalability

Merits

- Takes computation to data
- Suitable for large data centric operations
- Scalable on demand
- Fault tolerant and highly transparent

PIG

SQL-like language on top of MR

- Pig Latin
- Developed By Yahoo
- Used for Programming
- Operates on the client side of a cluster.
- Pig is data flow scripting
- Pig Hadoop follows a multi query approach

HIVE

Simple scripting language on top of MR

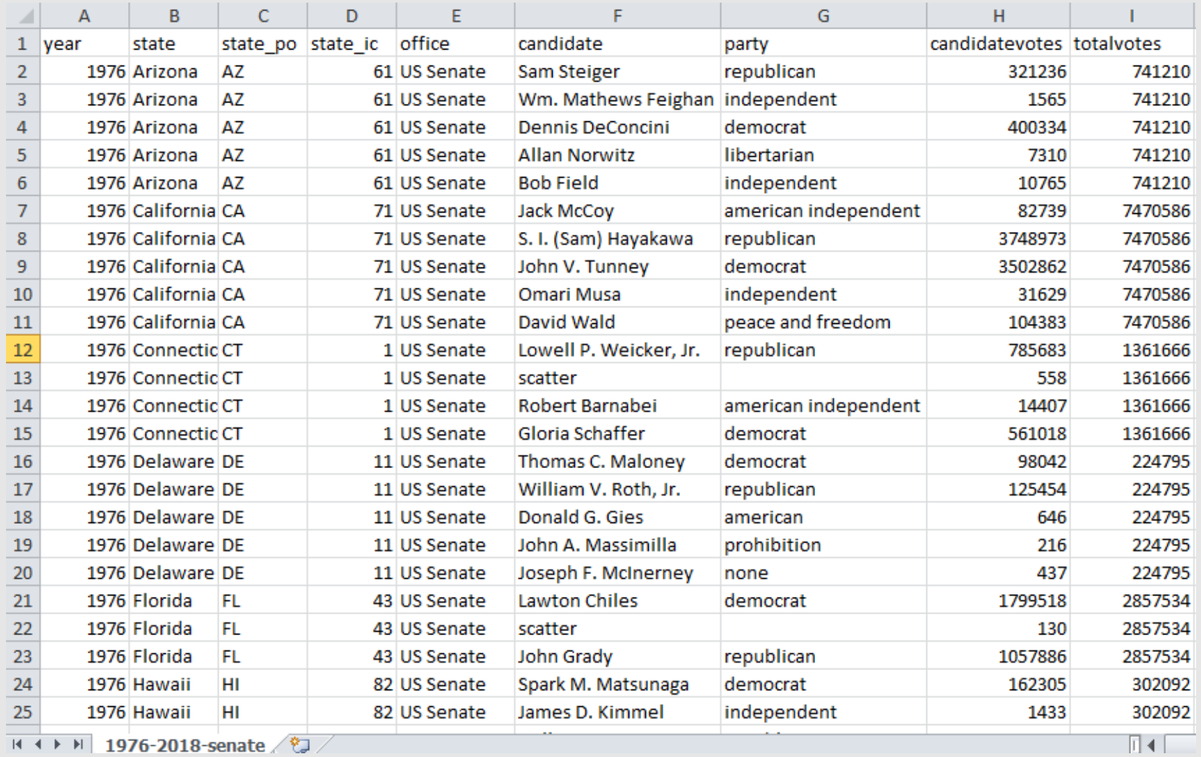
- HiveQL
- Developed By facebook
- Used for creating reports
- Operates on the server side of a cluster
- Directly leverages SQL and is easy to learn for database experts.
- Hive Hadoop provides the users with strong and powerful statistics functions.

US Election Data

Data Source: Github

Data Description:

- We have used two data sets.
- Senate 1976-2018
- President 1976-2018
- Each data set contains nine variables and 3422 tuples.



	A	B	C	D	E	F	G	H	I
1	year	state	state_po	state_ic	office	candidate	party	candidatevotes	totalvotes
2	1976	Arizona	AZ	61	US Senate	Sam Steiger	republican	321236	741210
3	1976	Arizona	AZ	61	US Senate	Wm. Mathews Feighan	independent	1565	741210
4	1976	Arizona	AZ	61	US Senate	Dennis DeConcini	democrat	400334	741210
5	1976	Arizona	AZ	61	US Senate	Allan Norwitz	libertarian	7310	741210
6	1976	Arizona	AZ	61	US Senate	Bob Field	independent	10765	741210
7	1976	California	CA	71	US Senate	Jack McCoy	american independent	82739	7470586
8	1976	California	CA	71	US Senate	S. I. (Sam) Hayakawa	republican	3748973	7470586
9	1976	California	CA	71	US Senate	John V. Tunney	democrat	3502862	7470586
10	1976	California	CA	71	US Senate	Omari Musa	independent	31629	7470586
11	1976	California	CA	71	US Senate	David Wald	peace and freedom	104383	7470586
12	1976	Connectic	CT	1	US Senate	Lowell P. Weicker, Jr.	republican	785683	1361666
13	1976	Connectic	CT	1	US Senate	scatter		558	1361666
14	1976	Connectic	CT	1	US Senate	Robert Barnabei	american independent	14407	1361666
15	1976	Connectic	CT	1	US Senate	Gloria Schaffer	democrat	561018	1361666
16	1976	Delaware	DE	11	US Senate	Thomas C. Maloney	democrat	98042	224795
17	1976	Delaware	DE	11	US Senate	William V. Roth, Jr.	republican	125454	224795
18	1976	Delaware	DE	11	US Senate	Donald G. Gies	american	646	224795
19	1976	Delaware	DE	11	US Senate	John A. Massimilla	prohibition	216	224795
20	1976	Delaware	DE	11	US Senate	Joseph F. McInerney	none	437	224795
21	1976	Florida	FL	43	US Senate	Lawton Chiles	democrat	1799518	2857534
22	1976	Florida	FL	43	US Senate	scatter		130	2857534
23	1976	Florida	FL	43	US Senate	John Grady	republican	1057886	2857534
24	1976	Hawaii	HI	82	US Senate	Spark M. Matsunaga	democrat	162305	302092
25	1976	Hawaii	HI	82	US Senate	James D. Kimmel	independent	1433	302092

1976-2018-senate

Stock Market Data

Data Source: Yahoo Finance

Data Description:

- Daily data of 5 stocks for past one year.
- Data contains 7 variables And 265 tuples.

	A	B	C	D	E	F	G	H
1	Symbol	Date	Open	High	Low	Close	Adj Close	Volume
2	reliance	01/01/2019	1125.25	1127.3	1081.1	1104.75	1094.613	33025977
3	reliance	08/01/2019	1105.1	1117	1086.4	1096.8	1086.736	26047926
4	reliance	15/01/2019	1105	1239.95	1105	1237.7	1226.343	71656291
5	reliance	22/01/2019	1232.85	1264.7	1219.6	1229.55	1218.268	55657607
6	reliance	29/01/2019	1231	1296.95	1191.1	1290.9	1279.055	48260688
7	reliance	05/02/2019	1292	1321.2	1251	1253.25	1241.75	42015719
8	reliance	12/02/2019	1251.5	1273.95	1214	1220.1	1208.904	39237865
9	reliance	19/02/2019	1218	1257.8	1211.2	1232.3	1220.993	39730939
10	reliance	26/02/2019	1209.5	1244.9	1206	1226.05	1214.8	40453661
11	reliance	05/03/2019	1223.4	1312	1218.6	1304.1	1292.134	43027891
12	reliance	12/03/2019	1316.9	1362	1311.2	1350.05	1337.662	58865159
13	reliance	19/03/2019	1360	1388	1316.7	1324.45	1312.297	39115245
14	reliance	26/03/2019	1330.3	1406.8	1330	1391.85	1379.078	47853952
15	reliance	02/04/2019	1398	1403.1	1323.7	1329.25	1317.053	39689587
16	reliance	09/04/2019	1328.9	1356.9	1321	1340.15	1327.853	33016534
17	reliance	16/04/2019	1345	1389.75	1340	1345.35	1333.005	36689783
18	reliance	23/04/2019	1348	1412.4	1346	1392.8	1380.02	37235451
19	reliance	30/04/2019	1396.4	1417.5	1366.8	1384.9	1372.192	32647603
20	reliance	07/05/2019	1394.8	1395	1227.5	1232.05	1220.745	62269638
21	reliance	14/05/2019	1236.5	1337.7	1231.5	1325.9	1313.734	51003834
22	reliance	21/05/2019	1332.2	1392	1307	1310.65	1298.624	60412448
23	reliance	28/05/2019	1319.8	1367.25	1304.15	1360.2	1347.719	57570118
24	reliance	04/06/2019	1357.45	1374.25	1305.6	1319.15	1307.046	26834973
25	reliance	11/06/2019	1321.85	1338.4	1278.5	1282.3	1270.534	30779580

Analysis Using Hive

CREATING A TABLE FOR SENATE DATA AND LOADING IT

```
[cloudera@quickstart ~]$ hdfs dfsadmin -safemode leave
Safe mode is OFF
[cloudera@quickstart ~]$ hive

hive> CREATE TABLE senate_data (year INT ,state STRING ,state_po STRING ,state_ic INT,
office STRING ,candidate STRING ,party STRING ,candidatevotes INT,totalvotes INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n';
OK
hive> load data local inpath '/home/cloudera/Downloads/senate.csv' overwrite into table senate_data;
Loading data to table default.senate_data
Table default.senate_data stats: [numFiles=1, numRows=0, totalSize=235078, rawDataSize=0]
OK
Time taken: 3.821 seconds

hive> show tables;
OK
senate_data
Time taken: 3.718 seconds, Fetched: 1 row(s)
```

USE CREATE AND SELECT FUNCTION TO VIEW DATA

```
hive> CREATE VIEW senate_view AS SELECT * FROM senate_data;
hive> SELECT * FROM senate_view;

2016 Louisiana LA 45 US Senate Foster Campbell democrat347816 1997218
2016 Louisiana LA 45 US Senate Abhay Patel republican 1576 1997218
2016 Louisiana LA 45 US Senate """Joseph"" Cao" republican 21019 1997218
2016 Louisiana LA 45 US Senate David Duke republican 58606 1997218
2016 Louisiana LA 45 US Senate "Donald ""Crawdaddy"" Crawford" republican 25523 1997218
2016 Louisiana LA 45 US Senate Charles Marsala republican 3684 1997218
2016 Maryland MD 52 US Senate Chris Van Hollen democrat 1659907 2726170
2016 Maryland MD 52 US Senate Margaret Flowers green 89970 2726170
2016 Maryland MD 52 US Senate NA NA 77 2726170
2016 Maryland MD 52 US Senate NA NA 7 2726170
2016 Maryland MD 52 US Senate NA NA 242 2726170
```

Finding the total votes of each party

hive> SELECT party , sum(candidatevotes) FROM senate_data GROUP BY party;
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 9.02 sec   HDFS Read: 243645 HDFS
Write: 3772 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 20 msec
OK
5849546
 Jr." NULL
 Sr." NULL
"god NULL
Independent for Liberty 5624
NA 359174
a connecticut party 584956
alaska libertarian 1240
alaskan independence 23706
america first 3593
american 366303

Finding candidate votes for each party having total votes >= 40k

hive> SELECT party , sum(candidatevotes) FROM senate_data GROUP BY party having sum(candidatevotes)>=40000;

a connecticut party 584956
american 366303
american constitution 50105
american constitution party 59733
american independent 1478402
citizens 57901
communist 64690
concerned citizens 85645
connecticut for lieberman 570830
conservative 4062460
constitution 1023508
constitutional 68377
consumer 65273
democrat 670474713

Finding Top 10 candidate votes for each party having total votes >= 40k

```
hive> SELECT party , sum(candidatevotes) FROM senate_data GROUP BY party having sum(candidatevotes)>=40000 limit 10;

Total MapReduce CPU Time Spent: 10 seconds 290 msec
OK
5849546
NA 359174
a connecticut party 584956
american 366303american constitution 50105
american constitution party 59733
american independent 1478402
citizens 57901
communist 64690
concerned citizens 85645
Time taken: 143.854 seconds, Fetched: 10 row(s)
```

Arranging total votes by descending order

```
hive> SELECT * FROM senate_data ORDER BY totalvotes DESC;

2000 Connecticut CT 1 US Senate "William Kozak Jr." NULL 25509
1988 Connecticut CT 1 US Senate "Howard Avory Grayson Jr." NULL 12409
2004 Louisiana LA 45 US Senate "Sam Houston Melton Jr." NULL 12289
1992 Nevada NV 65 US Senate "Joe S. Garcia Jr." NULL 11240
1988 New York NY 13 US Senate "James E. Harris Jr." NULL 11239
1992 Connecticut CT 1 US Senate "Howard A. Grayson Jr." NULL 10741
1998 Louisiana LA 45 US Senate "Sam Houston Melton Jr." NULL 9893
1988 Pennsylvania PA 14 US Senate "Samuel Cross Jr." NULL 6455
```

Printing top 10 total votes

hive> SELECT * FROM senate_data ORDER BY totalvotes DESC limit 10;

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 10.76 sec   HDFS Read: 242351 HDFS Write: 90 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 760 msec
OK
12578511
12578511
12244170
12244170

Finding 10 states having candidatevotes >= 40k

hive> SELECT state , sum(candidatevotes) FROM senate_data GROUP BY state having sum(candidatevotes)>=40000;

Total MapReduce CPU Time Spent: 10 seconds 390 msec
OK
Alabama 20390799
Alaska 3255066
Arizona 21633663
Arkansas 11449358
California 152755465
Colorado 22123666
Connecticut 17132454
Delaware 3590160
Florida 78679490
Georgia 34536647

Analysis using Pig

[cloudera@quickstart ~]$ hdfs dfsadmin -safemode leave
Safe mode is OFF
[cloudera@quickstart ~]$ pig -x local

loading dataset to pig.
grunt> data1 = LOAD '/home/cloudera/Downloads/NIFTY_Data.csv' using PigStorage (',') as
(Symbol:chararray,Date:date,Open:float,High:float,Low:float,Close:float,Adj_Close:float,Volume:int)

let's apply limit function and get only 10 entries instead of pasting entire dataset.

grunt> head_data = LIMIT data1 10;
grunt> DUMP head_data;

OUTPUT:

(reliance ,43466,1125.25,1127.300049,1081.099976,1104.75,1094.612793,33025977)
(reliance ,43473,1105.099976,1117,1086.400024,1096.800049,1086.73584,26047926)
(reliance ,43480,1105,1239.949951,1105,1237.699951,1226.342773,71656291)
(reliance ,43487,1232.849976,1264.699951,1219.599976,1229.550049,1218.2677,55657607)
(reliance ,43494,1231,1296.949951,1191.099976,1290.900024,1279.054688,48260688)
(reliance ,43501,1292,1321.199951,1251,1253.25,1241.750244,42015719)
(reliance ,43508,1251.5,1273.949951,1214,1220.099976,1208.904419,39237865)
(reliance ,43515,1218,1257.800049,1211.199951,1232.300049,1220.992554,39730939)

Performing EDA

let's order this data by closing price so that we'll get stocks with highest closing price on top
grunt> close_desc = ORDER data1 BY Close DESC;
grunt> close_desc_10 = LIMIT close_desc 10;
grunt> DUMP close_desc_10;
OUTPUT:
(reliance ,43795,1568.099976,1614.449951,1547.849976,1586.5,1580.565186,49605366)
(reliance ,43802,1592.75,1594,1533.75,1572.599976,1566.717163,36412572)
(reliance ,43816,1566.75,1617.550049,1555.550049,1571.400024,1565.521729,46609838)
(reliance ,43809,1572.050049,1593.900024,1550.599976,1566.599976,1560.739624,26253054)
(reliance ,43788,1467,1572.400024,1465,1561.550049,1555.708618,57653601)
(reliance ,43823,1568.900024,1572.050049,1510.150024,1544.199951,1538.423462,37766874)
(reliance ,43830,1542,1543.699951,1508.050049,1514.050049,1508.386353,10150467)
(reliance ,43781,1427.800049,1486.800049,1427.800049,1459.199951,1453.741455,31659474)
(reliance ,43767,1445.5,1489.650024,1441,1457.650024,1452.197266,40935293)
using filter command to extract stock prices of ITC and will group the data for further analysis.
grunt> itc_data = FILTER data1 BY Symbol== "itc";
grunt> itc_grouped = GROUP itc_data All;
grunt> DUMP itc_grouped;
OUTPUT:
(itc ,43585,304.549988,307.549988,299.5,307,286.498047,31280003)(itc ,43599,291,307.549988,288.850006,307,286.498047,68602003)(itc ,43564,292.5,310,292.25,305.5,285.098206,56705987)
(itc ,43578,302.149994,309.799988,298.5,304.549988,284.21167,46839366)(itc ,43571,306.799988,309.950012,301.049988,301.799988,281.645294,25256746)(itc ,43550,295,303,291.049988,297.25,277.39917,55065137)
(itc ,43543,294.600006,301.950012,293.200012,295.049988,275.346069,47459597)(itc ,43473,282,296,281.350006,294.299988,274.646149,53124402)(itc ,43536,294,297,289.350006,293.549988,273.946259,65888714)
(itc ,43529,280,294.899994,276.049988,292.950012,273.386353,75315856)

Calculating Average & Median Volume

now we'll calculate average of volume column for ITC.

```
grunt> itc_avg_volume = FOREACH itc_grouped GENERATE AVG(itc_grouped.Volume) as itc_avg_volume;
grunt> DUMP itc_avg_volume;
```

OUTPUT:

(59912375.66)

we'll calculate MEDIAN of volume column for ITC.

```
grunt> itc_med_volume = FOREACH itc_grouped GENERATE MEDIAN(itc_grouped.Volume);
grunt> DUMP itc_med_volume;
```

OUTPUT:

(56601555)

Conclusion

- The commercial impacts of the Big data have the potential to generate significant productivity growth for a number of vertical sectors.
- Big Data presents opportunity to create unprecedented business advantages and better service delivery.
- Like we have shown in the above analysis how we can work with stock market data which produces tera bytes of data daily.
- Moreover we have just started with the basic analysis but in the coming future we need to be more advanced in the field of big data as alot of data is being generated every minute and we can analyse them using traditional techniques.

LIMITATIONS

PIG:

- The error that pig produces are not helpful.
- Not mature.
- Data schema is not enforced explicitly but implicitly.
- Commands are not executed until you dump in an intermediate result.

Hive:

- Apache hive does not offer real-time queries and row level updates.
- Latency of Apache Hive queries is generally very high.
- Limited subquery saupport.
- No support for materialized view.
- Update or delete operations are not supported in hive.
- Not designed for OLTP(online transitional process).

Presented By

Janhavi Bagul

A003

Mahesh Bansal

A006

Pranav Deore

A011

Hriday Harchandani

A017

Siddharth Sonetta

A034

Thank You!

Mentored by
- Sareeta Mugde