Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

MIE 1628 Big Data Science

Assignment 1

Prof. Shevchenko

Jimeng Cui

1004486923

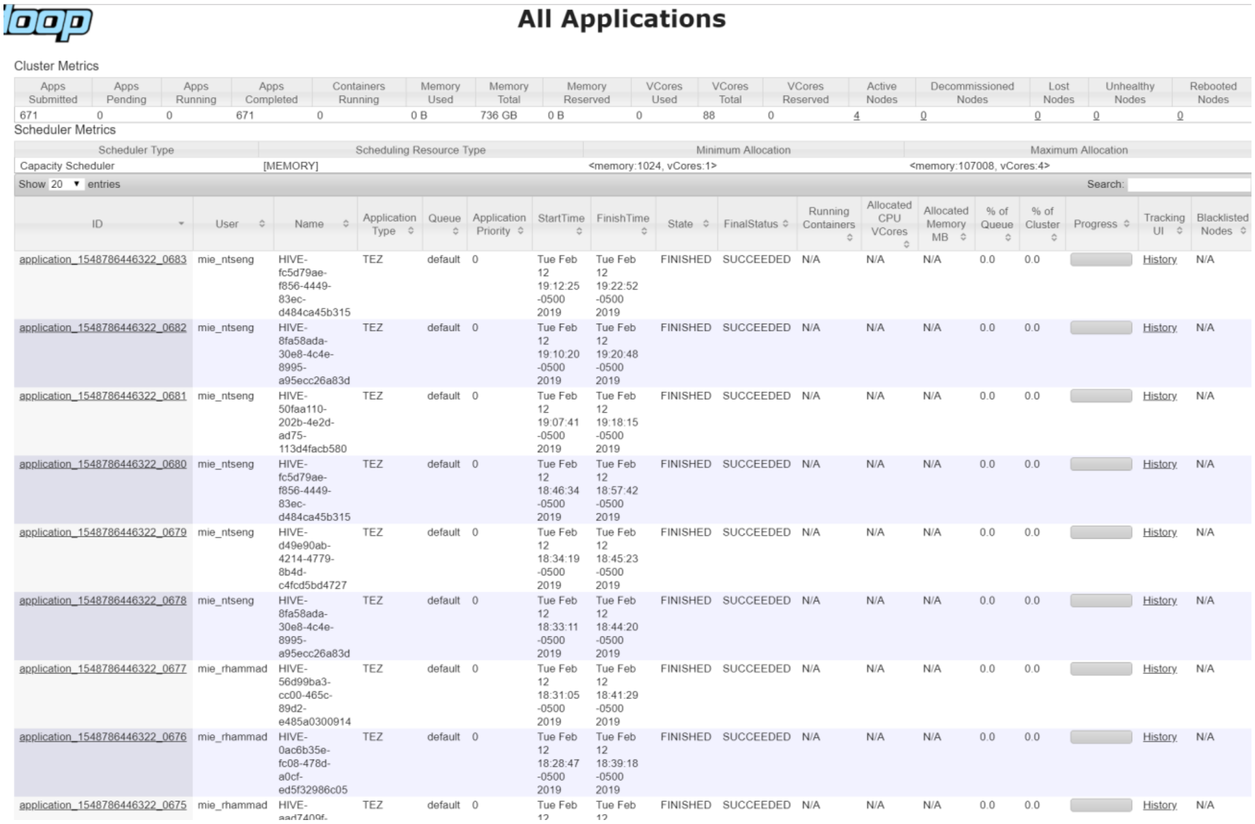
02/19/2019

General Theoretical Questions

1. Big Data-specific formats: Parquet; Avro; RC & ORC files
2. Why compression: Every stored file should be replicated 3 times. Therefore, compress data can help with storing more data per byte of disk. Also, Compressing data would speed up the I/O operations.

Yarn Application and Commands

3. What is YARN and two most important functions: YARN is short for “Yet Another Resource Negotiator”. YARN has two most important functions which are the resource management and job scheduling technology.
4. List all running applications:



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
671	0	0	671	0	0 B	736 GB	0 B	0	88	0	4	0	0	0	0

Scheduler Metrics

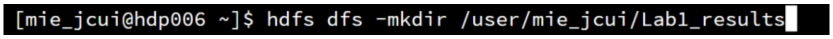
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:107008, vCores:4>

Show 20 entries Search:

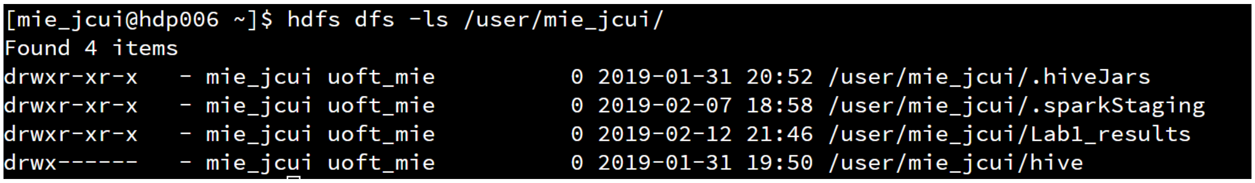
ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1548786446322_0683	mie_ntseng	HIVE-fc5d79ae-f856-4449-83ec-d484ca45b315	TEZ	default	0	Tue Feb 12 19:12:25 -0500 2019	Tue Feb 12 19:22:52 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0682	mie_ntseng	HIVE-8fa58ada-30e8-4c4e-8995-a95ecc26a83d	TEZ	default	0	Tue Feb 12 19:10:20 -0500 2019	Tue Feb 12 19:20:48 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0681	mie_ntseng	HIVE-50faa110-202b-4e2d-ad75-113d4facb580	TEZ	default	0	Tue Feb 12 19:07:41 -0500 2019	Tue Feb 12 19:18:15 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0680	mie_ntseng	HIVE-fc5d79ae-f856-4449-83ec-d484ca45b315	TEZ	default	0	Tue Feb 12 18:46:34 -0500 2019	Tue Feb 12 18:57:42 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0679	mie_ntseng	HIVE-d49e90ab-4214-4779-8b4d-c4fcd5bd4727	TEZ	default	0	Tue Feb 12 18:34:19 -0500 2019	Tue Feb 12 18:45:23 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0678	mie_ntseng	HIVE-8fa58ada-30e8-4c4e-8995-a95ecc26a83d	TEZ	default	0	Tue Feb 12 18:33:11 -0500 2019	Tue Feb 12 18:44:20 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0677	mie_rhammad	HIVE-56d99ba3-cc00-465c-89d2-e485a0300914	TEZ	default	0	Tue Feb 12 18:31:05 -0500 2019	Tue Feb 12 18:41:29 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0676	mie_rhammad	HIVE-0ac6b35e-fc08-478d-a0cf-ed5f32986c05	TEZ	default	0	Tue Feb 12 18:28:47 -0500 2019	Tue Feb 12 18:39:18 -0500 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1548786446322_0675	mie_rhammad	HIVE-aad7409f-	TEZ	default	0	Tue Feb 12	Tue Feb 12	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A

HDFS Commands

5. Create File:



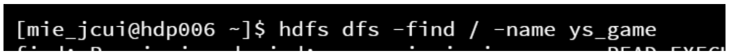
```
[mie_jcui@hdp006 ~]$ hdfs dfs -mkdir /user/mie_jcui/Lab1_results
```



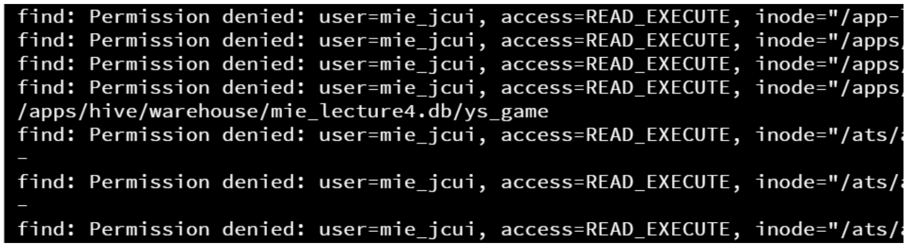
```
[mie_jcui@hdp006 ~]$ hdfs dfs -ls /user/mie_jcui/
Found 4 items
drwxr-xr-x   - mie_jcui uoft_mie          0 2019-01-31 20:52 /user/mie_jcui/.hiveJars
drwxr-xr-x   - mie_jcui uoft_mie          0 2019-02-07 18:58 /user/mie_jcui/.sparkStaging
drwxr-xr-x   - mie_jcui uoft_mie          0 2019-02-12 21:46 /user/mie_jcui/Lab1_results
drwx------   - mie_jcui uoft_mie          0 2019-01-31 19:50 /user/mie_jcui/hive
```

6. Address of MIE_Lecture4.ys_game:

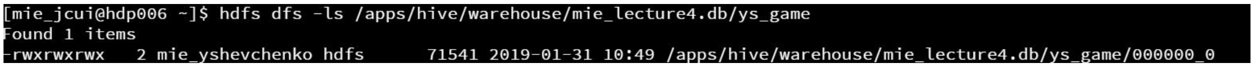
Method1: Find directly from terminal. The path is among many permission denied files.



```
[mie_jcui@hdp006 ~]$ hdfs dfs -find / -name ys_game
```

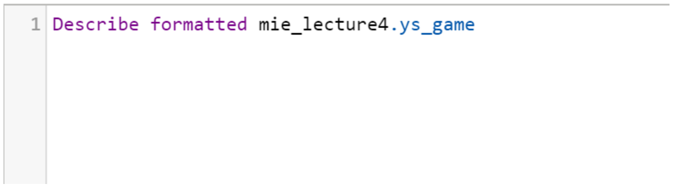


```
find: Permission denied: user=mie_jcui, access=READ_EXECUTE, inode="/app-
find: Permission denied: user=mie_jcui, access=READ_EXECUTE, inode="/apps
find: Permission denied: user=mie_jcui, access=READ_EXECUTE, inode="/apps
find: Permission denied: user=mie_jcui, access=READ_EXECUTE, inode="/apps
/apps/hive/warehouse/mie_lecture4.db/ys_game
find: Permission denied: user=mie_jcui, access=READ_EXECUTE, inode="/ats/
-

find: Permission denied: user=mie_jcui, access=READ_EXECUTE, inode="/ats/
-

find: Permission denied: user=mie_jcui, access=READ_EXECUTE, inode="/ats/
```



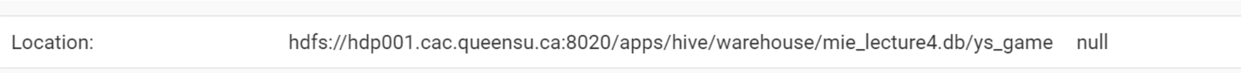
```
[mie_jcui@hdp006 ~]$ hdfs dfs -ls /apps/hive/warehouse/mie_lecture4.db/ys_game
Found 1 items
-rwxrwxrwx   2 mie_yshevchenko hdfs      71541 2019-01-31 10:49 /apps/hive/warehouse/mie_lecture4.db/ys_game/000000_0
```

Method2: Find from Hive.



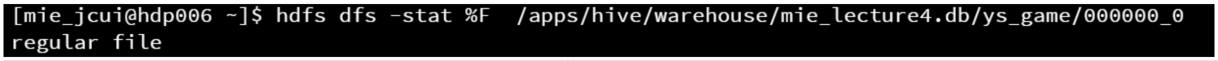
```
Describe formatted mie_lecture4.ys_game
```

Output:

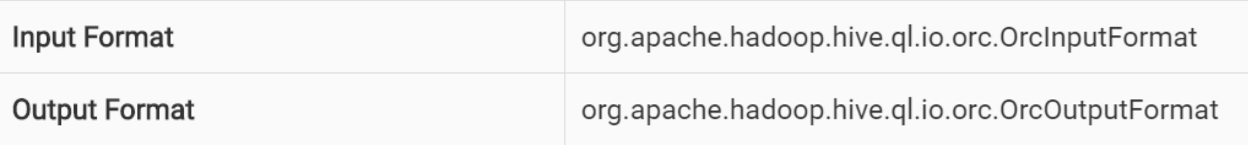


```
Location:                hdfs://hdp001.cac.queensu.ca:8020/apps/hive/warehouse/mie_lecture4.db/ys_game    null
```

7. What is the format of Hive Tables:
The format is orc. Code as below:



```
[mie_jcui@hdp006 ~]$ hdfs dfs -stat %F  /apps/hive/warehouse/mie_lecture4.db/ys_game/000000_0
regular file
```

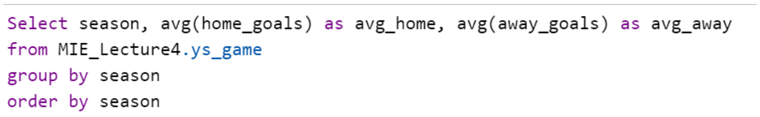


Input Format	org.apache.hadoop.hive.ql.io.orc.OrcInputFormat
Output Format	org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat

Hive

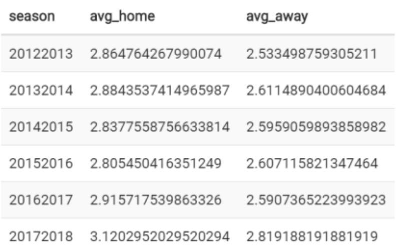
Q8:

Code:



```
Select season, avg(home_goals) as avg_home, avg(away_goals) as avg_away
from MIE_Lecture4.ys_game
group by season
order by season
```

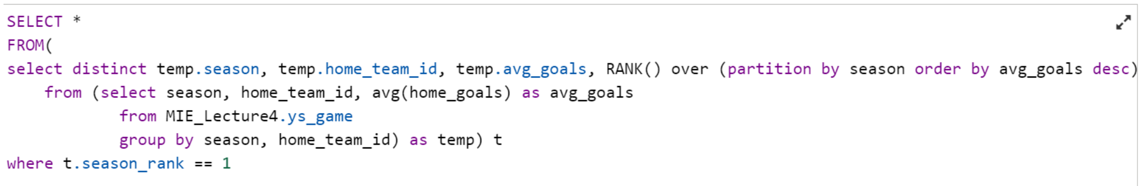
Results:



season	avg_home	avg_away
20122013	2.864764267990074	2.533498759305211
20132014	2.8843537414965987	2.6114890400604684
20142015	2.8377558756633814	2.5959059893858982
20152016	2.805450416351249	2.607115821347464
20162017	2.915717539863326	2.5907365223993923
20172018	3.1202952029520294	2.819188191881919

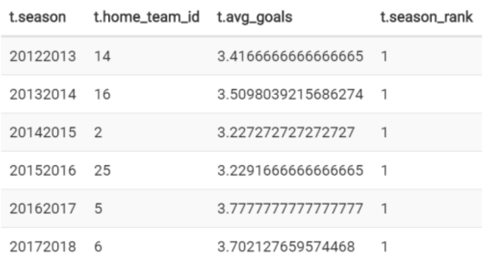
Q9:

Code:



```
SELECT *
FROM(
select distinct temp.season, temp.home_team_id, temp.avg_goals, RANK() over (partition by season order by avg_goals desc)
    from (select season, home_team_id, avg(home_goals) as avg_goals
            from MIE_Lecture4.ys_game
            group by season, home_team_id) as temp) t
where t.season_rank == 1
```

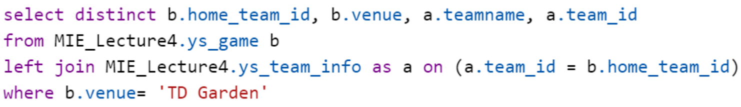
Results:



t.season	t.home_team_id	t.avg_goals	t.season_rank
20122013	14	3.4166666666666665	1
20132014	16	3.5098039215686274	1
20142015	2	3.227272727272727	1
20152016	25	3.2291666666666665	1
20162017	5	3.7777777777777777	1
20172018	6	3.702127659574468	1

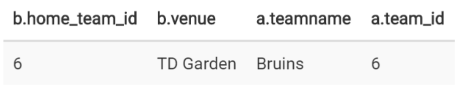
Q10:

Code:



```
select distinct b.home_team_id, b.venue, a.teamname, a.team_id
from MIE_Lecture4.ys_game b
left join MIE_Lecture4.ys_team_info as a on (a.team_id = b.home_team_id)
where b.venue= 'TD Garden'
```

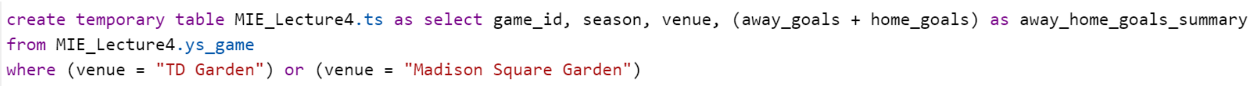
Results:



b.home_team_id	b.venue	a.teamname	a.team_id
6	TD Garden	Bruins	6

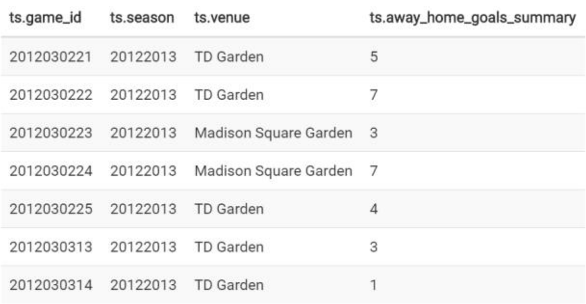
Q11:

Code:



```
create temporary table MIE_Lecture4.ts as select game_id, season, venue, (away_goals + home_goals) as away_home_goals_summary
from MIE_Lecture4.ys_game
where (venue = "TD Garden") or (venue = "Madison Square Garden")
```

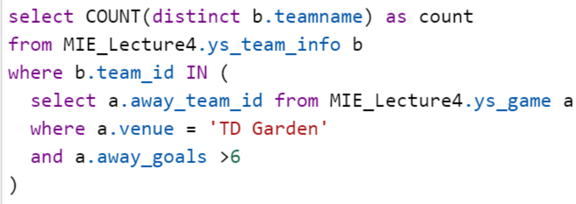
Results:



ts.game_id	ts.season	ts.venue	ts.away_home_goals_summary
2012030221	20122013	TD Garden	5
2012030222	20122013	TD Garden	7
2012030223	20122013	Madison Square Garden	3
2012030224	20122013	Madison Square Garden	7
2012030225	20122013	TD Garden	4
2012030313	20122013	TD Garden	3
2012030314	20122013	TD Garden	1

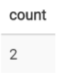
Q12:

Code:



```
select COUNT(distinct b.teamname) as count
from MIE_Lecture4.ys_team_info b
where b.team_id IN (
  select a.away_team_id from MIE_Lecture4.ys_game a
  where a.venue = 'TD Garden'
  and a.away_goals >6
)
```

Results:



count
2