Monash University

FIT5148 – Distributed and Big Data Processing, Semester 1, 2016

Big Data Report (40%)

Group Assignment

Due Date: Week 12, Thursday by 3pm

- This is a group assignment, **groups of 2** and from the same tutorial ONLY**.**
- There is no interview for this assignment.
- You will present this work as a group in **Presentation of Big Data Report (10%).** The **presentation** will be **for Part 2** of this assignment**.**

This report consists of **two parts**:

- The first part is **performance evaluation**. You will perform **a number of tasks and queries** in the Hortonworks environment **using Hive and Pig**. You need to write the correct queries for Pig and Hive to produce the results specified in the assignment. Then you will record all the details that logs and reports show in Hortonworks. You will use all this information to compare the performance of Pig and Hive such as how long it took for each or how many MapReduce jobs were executed etc. A table should be included along with brief but informative discussions in a paragraph format.

- The second part (see page 6) involves **research**. You will select a specific area in big data, and **read 4 seminal papers** about your selected area. Then you will **discuss, analyse and compare these papers** based on their approaches, contributions, methods, limitations, and any other criteria. This part has to be **written according to a specified template**, with high quality and correct **APA referencing**.

You will submit both parts in **in a well-structured report (ONE Word document).**

Part 1 (20 marks)

An overview:

In this part, you will use six files Master.csv, AwardsCoaches.csv, AwardsPlayers.csv, Coaches.csv, Scoring.csv and Teams.csv from the hockey dataset:
http://opensourcesports.com/files/hockey/hdb-2012-06-23.zip
You will use our Hortonworks tutorials as the guideline.

The data used in this assignment is courtesy of Open Source Sports. The Hockey Databank project allows for free usage of its data. In exchange for any usage of data, in whole or in part, we agree to display the following statement prominently and in its entirety on your end product:
"The information used herein was obtained free of charge from and is copyrighted by the Hockey Databank project. For more information about the Hockey Databank project please visit http://sports.groups.yahoo.com/group/hockey-databank"

Ice hockey is played on ice, usually in a rink, in which two teams of skaters use their sticks to shoot a hockey puck into their opponent's net to score points. Normally, each team has six players.

Common terms

Positions: Between the six players on the ice, they are typically divided into three forwards and two defensemen and a goaltender (goalie) (G). The forward positions consist of a centre (C) and two wingers: a left wing (L) and a right wing (R). The defencemen (D) usually stay together as a pair generally divided between left and right.

Assists: In ice hockey, point has three contemporary meanings: A point is awarded to a player for each goal scored or assist earned. The total number of goals plus assists equals total points. An assist is attributed to up to two players of the scoring team who shot, passed or deflected the puck towards the scoring teammate, or touched it in any other way which enabled the goal, meaning that they were "assisting" in the goal. There can be a maximum of two assists per goal.

Short-handed: A short-handed goal is a goal scored in ice hockey when a team's on-ice players are outnumbered by the opposing team. Normally, a team would be outnumbered because of a penalty incurred.

Power team: In ice hockey, a team is said to be on a power play when at least one opposing player is serving a penalty, and the team has a numerical advantage on the ice
(Source: Wikipedia)

Tables details:

Table Master.csv

playerID	Player ID
coachID	Coach ID
firstName	First name
lastName	Last name
nameNote	Note about player's name
nameGiven	Given name
nameNick	Nickname
height	Height in inches
weight	Weight in pounds
shootCatch	Shooting hand (catching hand for goalies/goalkeepers)
pos	Position (L, R, D, C, G)
birthYear	Year of birth
birthMon	Month of birth
birthDay	Day of birth
birthCountry	Country of birth
birthState	State or province of birth
birthCity	City of birth

Table AwardsPlayers.csv

playerID	Player ID
award	Name of award or trophy
year	
lgID	League ID
note	
teammate	
pos	Position (for all-star teams)

Table Scoring.csv

playerID	Player ID
year	Year (2005-06 listed as "2005")
stint	Stint (order of appearance in a season)
tmID	Team ID
lgID	League ID
pos	Position (explained earlier)
GP	Games played
G	Goals - Total number of goals the player has scored in the current season
A	Assists - Number of goals the player has assisted in the current season
Pts	Points - Scoring points, calculated as the sum of G and A
PIM	Penalty minutes
PPG	Power play goals - Number of goals the player has scored while his team was on the power play.
PPA	Power play assists - Number of goals the player has assisted in while his team was on the power play.
SHG	Shorthanded goals - Number of goals the player has scored while his team was shorthanded.
SHA	Shorthanded assists - Number of goals the player has assisted in while his team was shorthanded.
GWG	Game-winning goals – the goal for the winning team that is one more than the total number of goals scored by the losing team. If the losing team scores three goals, the fourth goal scored by the player (winning team) is the GWG
GTG	Game-tying goals - In a tie game, the game-tying goal (GTG) is the last goal scored by the player
SOG	Shots on goal - Total shots taken on net (the sum of the goals and the opposing goaltender's saves)
PostGP	Postseason games played
PostG	Postseason goals
PostA	Postseason assists
PostPts	Postseason points
PostPIM	Postseason penalty minutes
Post+/-	Postseason plus / minus
PostPPG	Postseason power play goals
PostPPA	Postseason power play assists
PostSHG	Postseason shorthanded goals
PostSHA	Postseason shorthanded assists
PostGWG	Postseason game-winning goals
PostSOG	Postseason shots on goal

Table Teams.csv

year	Year
lgID	League ID
tmID	Team ID
franchID	Franchise ID
rank	Final standing
G	Games
W	Wins
L	Losses
T	Ties
OTL	Overtime losses
Pts	Points
SoW	Shootout wins
SoL	Shootout losses
GF	Goals for
GA	Goals against
name	Full team name
PIM	Penalty minutes
PPG	Power play goals
PPC	Power play chances
SHA	Shorthanded goals against
PKG	Power play goals against
PKC	Penalty kill chances
SHF	Shorthanded goals for

Table Coaches.csv

coachID	Coach ID
year	Year
tmID	Team ID
lgID	League ID
stint	Coaching order
notes	Miscellaneous comments
G	Games
W	Wins
L	Losses
T	Ties
PostG	Postseason games
PostW	Postseason wins
PostL	Postseason losses
PostT	Postseason ties

Table AwardsCoaches.csv

coachID	Coach ID
award	Name of award or trophy
year	Year
lgID	League ID
note	"tie" indicates a tie with another coach

Instruction:

- For Part 1, you will write a technical report that very briefly describes how each task was performed and what the results were using all the documentation mentioned in each task. The report should be **written in the order of the tasks** using **the numbered headings and subheadings.**

- In your experiments and comparative evaluation, **record your machine/computer details** that you used like its model, operating system, memory, processor capacity, etc.

Task 1 (1 mark):

- Upload the files in Pig and Hive as specified in our tutorials.
- Provide one screenshot (first page) per table.

Task 2 Players (6 marks):

a) Find out the player/s who had the highest GWG (you can repeat a player if the year is different). In the query results show the following details: year, player first and last names, date of birth (day, month and year), country of birth, and team name and league ID, their positions, and GWG.

b) Among those player/s who had the highest GWG, find out the player who had won the highest number of awards. In the query results show the following details: player first and last name, and award count.

c) Then using the first and last name of the player who had won the highest number of awards find out the points that the player had earned for each year that the player received an award. Display the following details: the award names, the award year, and the points/Pts that the player scored that year.

- You will write the queries for the following questions in both **pig and hive**.
- Record and compare the Hive and Pig based on the total time taken as well as other factors such as no of jobs, maps and reducers, etc.
- You will provide **your hive queries, pig scripts and the screenshots of the results** (and tables).

Task 3 Coaches (6 marks):

a) Find out the coach who had won the highest number of awards. In the query results show the following details: coach first and last name, date of birth, birth country and number of awards.

b) Find out the coach who had the highest wins. In the query results show the following details: coach first and last name, year, games, wins, losses and ties.

c) Perform the same query (b) mentioned above but in the results also display the number of awards won by this coach.

- You will write the queries for the following questions in both **pig and hive**.
- Record and compare the Hive and Pig based on the total time taken as well as other factors such as no of jobs, maps and reducers, etc.
- You will provide **your hive queries, pig scripts and the screenshots of the results** (and tables).

Task 4 Teams (4 marks):

a) Find out the total number of points, goals and assists by each team. In the query results show the following details for all the teams: team name, total number of points, total number goals and total number of assists.

b) Find out for each team which player had scored the highest total number of points. In the query results show the following details: team name, player first and last name, year, number of points, goals and assists.

- You will write the queries for the following questions in both **pig and hive**.
- Record and compare the Hive and Pig based on the total time taken as well as other factors such as no of jobs, maps and reducers, etc.
- You will provide **your hive queries, pig scripts and the screenshots of the results** (and tables).

Task 5 (3 marks):

- In the **Hortonworks shell**, execute the **first query in Task 3** using **Pig**, and record your time with and without enabling the **Tez.** Compare the results.
- Then, in **Ambari, for Hive**, enable the Tez and perform the same query (the first query in Task 3) with and without Tez, and compare your results.
- For the same query, for Hive, this time use **Cost Based Optimization (CBO) with Tez on.** Record ~~total time taken and~~ ~~compare~~ the results such as DAG details, Graphic View or others ~~with the previous results~~.

Provide the results of your experiments in table/s where possible along with the screenshots.

Read this carefully

In the report, make sure you include all Hive queries, Pig scripts, and screenshots of all the results, logs, etc. that show the completion of each task and its component, and all the comparison discussions that you will write in paragraphs and comparison tables. (**Note:** when the results of a query do not fit in a screenshot, you have to provide two screenshots, one from the first page of results and one from the last page of results). **There will be mark deduction if the report is incomplete**.

Part 2 (20 marks)

In this part, you will do research on one of the big data areas to gain a better understanding of the state of the art in this field. You will write your findings in a high quality report **(1500 word limit)**. This part must demonstrate your ability to study and discuss peer-reviewed journal articles and conference papers, carry out in-depth analysis, and arrive at substantial conclusions.

Step 1 - You need to select only one of the following topics:

- Big data models and theories
- Machine learning and AI for big data
- Big data mining
- Big data standardization
- Green issues of big data
- Big data analytics and social media
- Real-time analysis of big data (technologies and algorithms)
- Big data case studies and applications

Step 2 - After selecting your research area, you need to read and research the related journal articles and conference papers (the books, websites, technical reports or any other sources **not accepted** here).

- You need to search **only these databases** in Monash Library: IEEE, ScienceDirect, Springer, ACM, ProQuest, IOS Press, or Scopus (**ONLY 2014-present, not older**).
- You will select **4 seminal and most related papers**.
- Papers should be **completed research papers**. DO NOT choose research-in-progress papers, surveys, or review papers.
- The selection of seminal papers must be based on considering **the most influential, well-known and cited papers** in that research area, and whether they are **full research papers with a proposed approach and its implementation and evaluation**.

(3 marks)

Step 3 - After selecting the four papers, you need to identify each paper's contributions, the proposed approach/method, the research issues/challenges it addresses, main findings and finally any remaining open issues.
You need to read, fully understand and analyse each paper and provide a professional and brief description for each:

a. The **research challenges and issues** that each paper addressing (there might be more than one paper addressing the same issue)

b. **The paper contributions,** what **approach/method/model** they are proposing and developing to address those challenges**.** You need to briefly describe their proposed approaches/methods/models, avoiding technical details and jargons.

c. What are **the main findings and results** of each paper (usually discussed after the evaluation section), and **any open issues** for further research.

Step 4 – You will consolidate all the results of step 3 into a research paper following the specified guidelines below.

You need to follow the following structure for Part (b):

1. **Introduction (about 100 words):** a brief description about what your paper is about. **(2 marks)**

2. **Current state of the art (about 1300 words):**
   1. Briefly and clearly describe **the proposed approaches/methods** of each paper; **(2 marks)**
   2. Discuss the **challenges/issues** that these papers focus on; **(2 marks)**
   3. Summarise the **findings and results of evaluation/experiments**. This should include what **improvement or impact** each paper had. Provide evidence from their evaluation results. **(4 marks)**
   4. Add **your judgement** on their results at the end. If the papers address the same problem, here you need to **compare how their improvements are different** or which approach outperforms the other one. **(3 marks)**

To write this section, **use paragraphs** rather than bullets or other styles, and make sure the paragraphs have a logical and consistent flow.

3. **Conclusion (about 100 words**) - Conclude by saying what paper was about, briefly discussing the main or interesting findings and making your final point (a strong one). **(2 marks)**

- **References** (this will not be counted in word limit) **–** list the details of all the references you used in the paper **according to the APA style** (in addition to your 4 papers you can reference other related papers). **In-text citation** and **references should be correct and** need to be according to **APA style** (refer to QManual on Moodle). **(2 marks)**

Submission Requirements:

All the following files should be uploaded to Moodle as a zip file and use the following naming convention: FIT5043-A2-[StudentID].zip. **One group member will upload** the file. **There is a mark deduction for any missing document**.

1. An Assessment Cover Sheet for the group
2. Provide a report that includes **Part a report and Part b report as specified** in **ONE Word document in the order of Tasks**. Use **numbered heading and subheadings**.

3. The font used **Times New Roman** for the text, **size 12** and **1.5 line space**.
4. The paper must be **proof read** and spell-checked before submission.

Late Submission:

Late Assignments or extensions will not be accepted unless you submit a special consideration form and provide valid documentation such as a medical certificate prior to the submission deadline (NOT after). Otherwise, there will be **5% penalty per day including weekends.**

PLEASE NOTE.

Before submitting your assignment, please make sure that you haven't breached the University plagiarism and cheating policy. It is the student's responsibility to make themselves familiar with the contents of these documents.

Please also note the following from the Plagiarism Procedures of Monash, available at http://www.policy.monash.edu/policy-bank/academic/education/conduct/plagiarism-procedures.html

Plagiarism occurs when students **fail to acknowledge** that the ideas of others are being used. Specifically it occurs when:

- **other people's work** and/or ideas are paraphrased and presented without a reference;
- **other students' work is copied or partly copied**;
- **other people's designs, codes or images are presented as the student's own work**;
- phrases and passages are used verbatim without quotation marks and/or without a reference to the author or a web page;
- lecture notes are reproduced without due acknowledgement.