EECS4415 Big Data Systems – Reading Assignment

Description

Your task is to read a research paper (or article) from the list below and provide a **maximum of 2-page** report that answers a number of questions. Questions are provided on the Appendix at the end of the handout. Sometimes the questions are specific to the article; most of the times they are generic comprehension questions, similar to the ones used in peer reviewing.

ID	Due date	Title
1	Sun, May 16, 2021	Information Platforms and the Rise of the Data Scientist. Jeff Hammerbacher. Beautiful Data, 73-84, 2009.
2	Sun, May 16, 2021	Data Driven, Creating a Data Culture. Hilary Mason, DJ Patel. O'Reilly Media, 2015.
3	Sun, May 23, 2021	Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios. Duplicate Record Detection: A Survey. IEEE TKDE, 2007.
4	Sun, May 23, 2021	Kreps, J. Narkhede, N., Rao, J. (2011). Kafka: a Distributed Messaging System for Log Processing. NetDB.
5	Sun, May 30, 2021	Ghemawat, S., Gobioff, H., & Leung, S. (2003). The Google file system. SOSP'03.
6	Sun, May 30, 2021	Dean, Jeffrey, and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. Communications of the ACM, 2008.
7	Sun, Jun 6, 2021	Melnik, Sergey, et al. Dremel: interactive analysis of web-scale datasets. VLDB, 2010.
8	Sun, Jun 6, 2021	M Zaharia et al. Spark: Cluster computing with working sets. HotCloud 10, 2010.
9	Sun, Jun 13, 2021	Toshniwal, Ankit, et al. Storm@twitter. SIGMOD, 2014.
10	Sun, Jun 13, 2021	Stonebraker M., et al. C-Store: A Column-oriented DBMS. VLDB, 2005.

Submission

You **don't** need to **hand in** the report at the beginning of the class. You **only** need to submit your report(s) electronically. Name your reports *r1.pdf*, *r2.pdf*, …, *r10.pdf* and submit from the command line using:

`%submit 4415 a0 <filename>` (e.g., for the first report execute: **`% submit 4415 a0 r1.pdf`**)

Grading Scheme for Reports

Each student/report will be assigned one of the following marks (out of 100). The evaluation will be based on the degree of comprehension (80%), and proper use of language, grammatical errors and typos (20%).

Assigned Mark	Description
100	Exceptional / Excellent
80	Very Good / Good
60	Competent / Fairly Competent
40	Incomplete / Marginally Failing
0	Failing

Please refer to the description of grade letters at York University online:
http://calendars.registrar.yorku.ca/2012-2013/academic/grades/index.htm

Appendix A: Reading Comprehension Questions

(Article reference is available on the EECS4415 website, as well as a local copy if link not available).

1. Information Platforms and the Rise of the Data Scientist

Questions

Q1: How traditional databases differ to Information Platforms (data centers)?
Q2: What are some of the challenges of Data Driven Organizations?
Q3: What are some of the tools/technologies used/found in an Information Platform?
Q4: Can you describe a high-level architecture abstraction for processing large amounts of data (i.e., Stacks of Platforms)?
Q5: What does the role of a Data Scientist involve in an Enterprise?

2. Data Driven - Creating a Data Culture

Questions

Q1: What does democratizing data refers to?
Q2: What is the opposite of data democratization process?
Q3: What is the data scientific process?
Q4: How to formulate the right research/problem questions?
Q5: In a DDO, where is the Data Team belonging (hierarchically)?
Q6: What are some of the characteristics of good tools for data scientists?
Q7: What is Data Culture referring to?

3. Duplicate Record Detection: A Survey

Questions

Q1: What is the duplicate record detection problem referring to?
Q2: Provide a summary table of the time complexity of various character-based similarity metrics.
Q3: Provide a summary table/graphic that shows the approaches for detecting duplicate records.
Q4: How could you improve the efficiency of duplicate detection methods? List the different approaches and a small paragraph describing each approach (the key idea).

4. Kafka: A Distributed Messaging System for Log Processing

Questions

Q1: What is the abstract architecture of Kafka?
Q2: Unlike typical messaging systems, a message stored in Kafka doesn't have an explicit message id? Why? How to access/distinguish messages?
Q3: How to make transfer more efficient?
Q4: What does it mean that brokers (servers) are stateless? What is the benefit of keeping state in a consumer?
Q5: How distributed coordination is achieved? Why is it poor?
Q6: What are the delivery Guarantees of Kafka?

5. The Google file system

Questions

Q1: Give a one paragraph summary of the paper (what is being proposed and in what context).
Q2: Briefly explain the role of each of the following: GFS master, GFS chunkserver, GFS client
Q3: What are the main two strategies for keeping GFS's high availability?

Q4: Provide three strong points about the paper (Be precise and explicit; clearly explain the value and nature of each point).
Q5: Provide a weak point about the paper (Be precise and explicit; clearly explain your critic)

6. MapReduce: simplified data processing on large clusters

Questions

Q1: Give a one paragraph summary of the paper (what is being proposed and in what context).
Q2: Briefly describe the execution overview of a MapReduce operation (provide a small table).
Q3: What is a combiner function useful for?
Q4: What is a "straggler" and how MapReduce is dealing with it?
Q5: Describe a computational problem (not listed in the paper) that can be easily distributed using MapReduce. What are the Map() and Reduce() operations doing? In your answer, you can follow the examples provided in paragraph 2.3 of the paper.

7. Dremel: interactive analysis of web-scale datasets

Questions

Q1: Give a one paragraph summary of the paper (what is being proposed and in what context).
Q2: What are the main reasons Dremel is fast? List them in a bullet format and provide a short explanation for each of them (~100 words).
Q3: What is the key idea of Dremel for assembling query answers?
Q4: What are the type of queries that Dremel is designed to answer efficiently?
Q5: How the efficiency of Dremel compares to alternatives? Briefly discuss its performance when compared to alternative baselines.

8. Spark: Cluster computing with working sets

Questions

Q1: Give a one paragraph summary of the paper (what is being proposed and in what context).
Q2: What are the Resilient Distributed Datasets (RDDs)?
Q3: What is the difference between an "RDD cache" and an "RDD save" action?
Q4: What is the type of jobs for which Spark is likely to outperform Hadoop?

9. Storm@Twitter

Questions

Q1: Give a one paragraph summary of the paper (what is being proposed and in what context).
Q2: What is Twitter Storm and what are its main properties?
Q3: What are the partitioning strategies Twitter Storm supports?
Q4: What are the main components of the Twitter Storm Data Model?
Q5: Provide a weak point about the paper (Be precise and explicit; clearly explain your critic)

10. C-Store: A Column-oriented DBMS

Questions

Q1: Give a one paragraph summary of the paper (what is being proposed and in what context).
Q2: What is a "shared nothing" architecture (answer can be found online). How is the C-Store adopting the "shared nothing" architecture?
Q3: How the logical and physical plan of the C-Store compare to those of the relational data model? Discuss briefly.
Q4: What are Storage Keys and Join Indices needed for in the C-Store?
Q5: What are the main reasons the C-Store performs faster than competitors?