**BUAD 5722: Big Data - Midterm review**

**Prof. Arturo Castellanos**

**What do I need to be ready for the Midterm?**

1. Make sure you bring your laptop with the software loaded and running. We will be using the Docker container and Quickstart VM during the exam.
2. The theory part (Monday) is closed book, closed notes.
3. The hands-on part (Wednesday) you can have your notes and own files. Checking email, messaging apps, is **NOT allowed.** Using email or a messaging app or sharing files or links during the exam time of the exam will be considered cheating.
4. Make sure you bring your computer charged and that you bring your power cable.
5. I will not troubleshoot errors in the exam.
6. I will **ONLY** answer general questions about the exam.

**Theory (Monday):**

Closed notes and closed book. Types of questions:

1. Short answer
2. Multiple-choice
3. True/False

**Questions (not meant to be exhaustive):**

1. What is big data? What makes big data systems unique (think characteristics of big data - 5V’s)?

Big data – a collection of datasets, has a large volume, velocity or variety makes it difficult to store and analyze.

1. Is HIVE considered an OLTP or OLAP system?

OLAP, because HIVE is a data warehouse

1. What is Hadoop HDFS?

Hadoop distributed file system.

1. What is the relationship between HDFS and Map/Reduce?

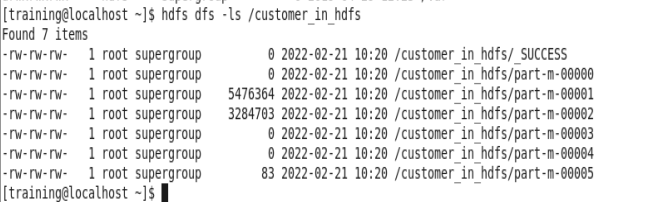
HDFS allows Map Reduce works on its platform.

1. Why was YARN a good addition to the Hadoop ecosystem?

Hadoop ecosystem: MapReduce, YARN, HDFS,

YARN is able to run data for streaming process, interactive process, and batch process.

1. What are some of the arguments required in writing a Sqoop command to move a database from MySQL to HDFS?
2. What does the following command do?
   1. hdfs dfs -mkdir /user/hive/warehouse table1
   2. hdfs dfs -cat /customer\_in\_hdfs/part-m-00001
3. We write a map-reduce and we look at the output in the directory where it was stored and we see the following:



Why is the results split into multiple parts? Can we know the number of mappers used?

1. What is the difference between data in motion and data at rest?

Data in motion is when data still in transit from one location to another, data at rest means the data has reached its destination.

1. True/False. **Consistency** is an ACID property that ensures that the entire transaction takes place at once or doesn’t happen at all.

F

1. True/False. Virtual computing, storage, and network resource that can be provisioned on demand are examples of **Software as a Service (SaaS)**
2. True/False: Vertical scaling involves upgrading the hardware resources (e.g. adding additional memory or storage)
3. What is map-reduce and why is it useful?
4. What is the difference between a virtual machine and a docker container?
5. Multiple choice. Which of the following steps need to be done when creating an Amazon EC2 instance?
   1. Select the operating system
   2. Select the instance type
   3. Select the disk space
   4. Select the security group
   5. All of the above
6. Apache NiFi is a tool that is primarily used for \_\_\_\_\_\_\_\_\_\_\_\_
   1. Batch analytics
   2. Real-time analytics
   3. Interactive querying
   4. Providing an API to a web framework
7. What are the advantages/disadvantages of having a high replication factor in Hadoop?
8. Map reduce patterns and what is the function of the mapper/reducer in the different tasks.
9. **True**/False. In HDFS, a single block is usually 128Mb. So a 1GB file, would consist of 8 blocks.
10. What are the advantages of table partitioning?
11. What is the main difference between running a query in Hive vs. Impala? Which one runs faster? Why?

Impala is not fault tolerant.

Impala runs faster, it runs in memory not map reduce.

1. Multiple choice. Here is a portion of a table in a MySQL database. The table name is games and the database name is fun. The table has 4 columns. What would be the number of rows and columns if I were run this query **SELECT min\_age, COUNT(\*) FROM fun.games GROUP BY min\_age;**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | name | inventor | year | min\_age |
| 1 | Monopoly | Elizabeth Magie | 1903 | 8 |
| 2 | Scrabble | Alfred Mosher Butts | 1938 | 8 |
| 3 | Clue | Anthony E. Pratt | 1944 | 8 |
| 4 | Candy Land | Eleanor Abbott | 1948 | 3 |
| 5 | Risk | Albert Lamorisse | 1957 | 10 |

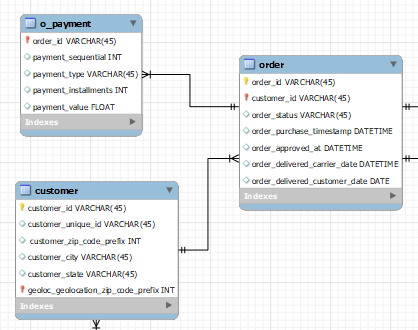
* 1. 5 rows, 4 columns
  2. 2 rows, 3 columns
  3. **3 rows, 2 columns**
  4. 4 rows, 5 columns
  5. The query would not return anything

1. Based on the following table definition, which rows can be stored in this table? Choose all that apply.

|  |  |  |
| --- | --- | --- |
| column | Data type | Notes |
| customer\_id | INT | Primary Key |
| customer\_name | VARCHAR(50) | NOT NULL |
| customer\_loyaltystatus | VARCHAR(50) |  |
| customer\_loyaltyscore | Decimal(2,1) |  |

* 1. **{customer\_id: ‘3554’, customer\_name: ‘John Smith’, customer\_loyaltyscore: 4.5}**
  2. {customer\_name: ‘William Doe’}
  3. **{customer\_id: ‘54’, customer\_name: ‘Mary Gonzalez’, customer\_loyaltystatus: ‘Platinum’}**
  4. {customer\_name: ‘Mingyu Zhang’, customer\_loyaltystatus: ‘Gold’, customer\_loyaltyscore: 4.2}

1. Which of the following are true based on the EER diagram below:



* 1. a customer can place one or many orders
  2. An order belongs only to one customer
  3. An order can have multiple payments
  4. A payment is for one order only
  5. **all of the above**

**Note:** I’m not going to ask specific questions about the components in the YARN architecture or how YARN runs an application.

**Hands-on (Wednesday):**

* Writing a regular expression to retrieve data of interest
* Map-reduce patterns
  + Top-N
    - Find the most discussed words in a document.
    - Trending topics from twitter (i.e., hashtag)
  + Filter
  + Inverted index
* Familiarity with HUE to access available data(bases/warehouses) and files in HDFS
* Hive and Impala queries
  + What is the total number of orders from customers located in Sao Paolo in the Olist Database (note you can answer this question either using MySQL or HIVE)
  + Write an Impala query on the toys dataset to output the id of the toy, name of the toy, name of the maker and price.

In the exam you need to be resourceful to be able to troubleshoot any issues you may get. This is why it is important that you run a couple of exercises in your OWN computer **before** the exam.

For example, let’s say in the day of the exam you get this error, you need to be able to Google and find a solution in how to fix it.

In this example, you can Google “Name node is in safe mode”

First few google results state how to solve it. Type the command: “hdfs dfsadmin -safemode leave”

