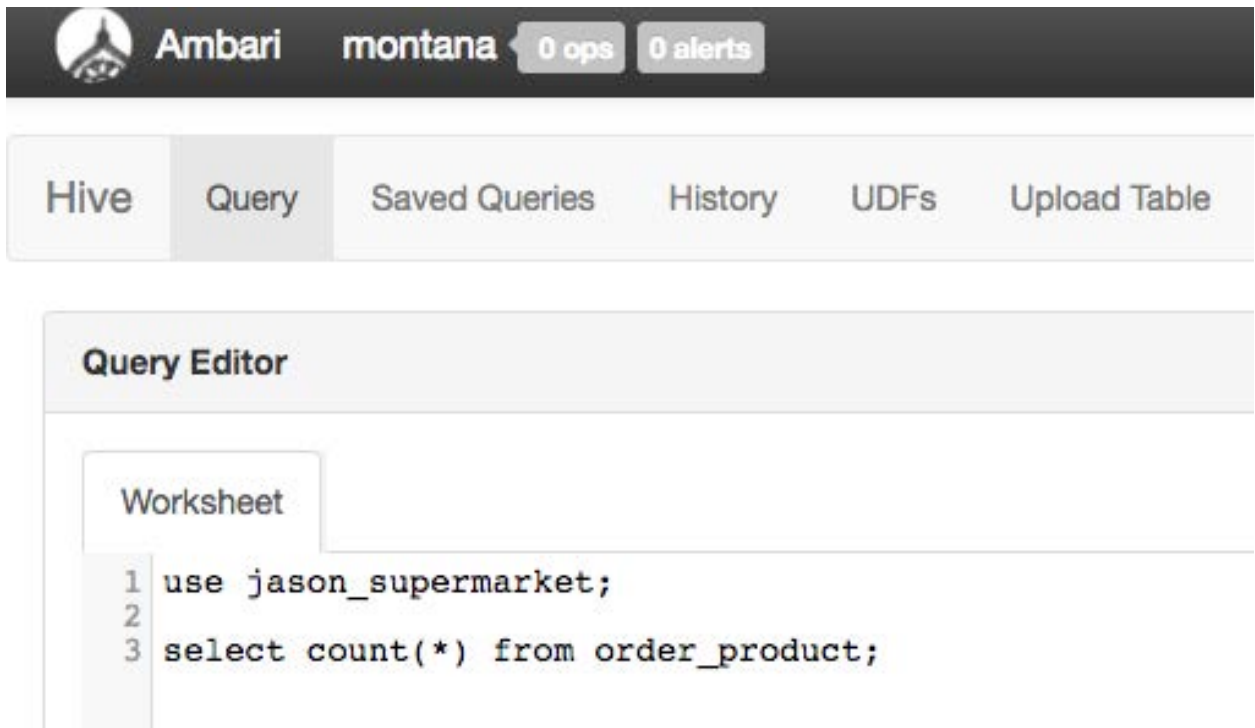DSO599 - Capstone Project
Supermarket Data Analysis using SQL

We will work on the supermarket database in the capstone project. There are two datasets in the project, a smaller dataset that we give to students in blackboard and a larger dataset that we load into Hadoop Hive database. These two datasets have the same schema. In the smaller dataset, there are 1.3M records in 'order_product' table, but there are 32.4M records in the large dataset.

Students will use the smaller dataset in MySQL workbench. Students don't need to create Hive database. The Hive database is created by Jason and TA. Students can login to Hive Web UI to query on the Hive database. Here is a screen copy of Hive Web UI.

```
1  use jason_supermarket;
2
3  select count(*) from order_product;
```

What is the grading criteria?
Students should work on the assignment and provide solution by using their MySQL workbench. If students can do additional work, i.e. work on the solution in big data system (Hive database), then they can earn maximum 3% of grade. Here is our grading policy in syllabus.

Assignments		Points	% of Grade
Homworks (6 in total)		600	30%
Exam		300	30%
Capstone Project (Individual)		300	30%
Class participation & attendance		100	10%
	TOTAL	1000	100%
Impressive work in capstone project will earn Maximum 3% of grade (30 points)			

What students should turn in for the project?
1. Students should save the code into one 'sql' file, and submit to blackboard. Student may work on both MySQL workbench and

Hadoop Hive database, if so please submit both MySQL and Hive code.

2. Please write one presentation using PowerPoint for your results. In the project presentation, students should show their solutions, their ideas and steps how they get the results.
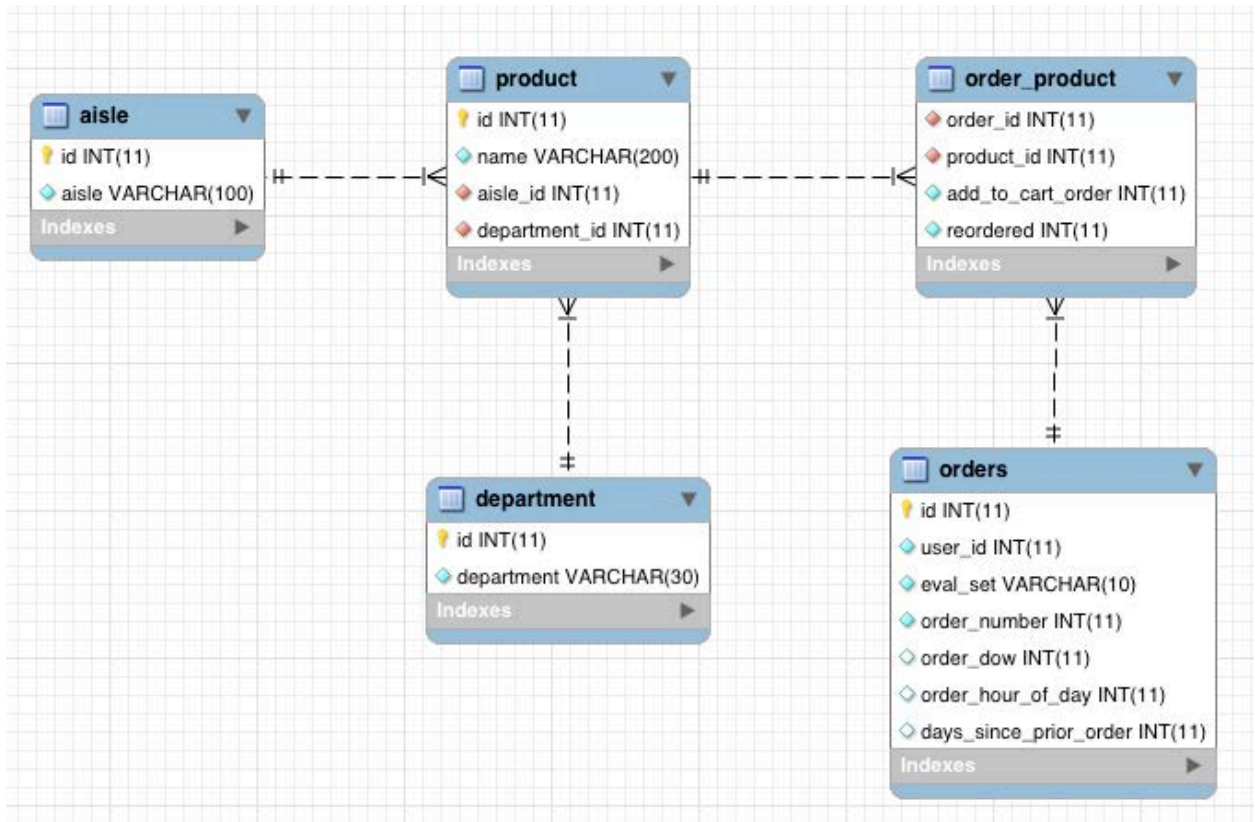
Database Diagram – Supermarket Database

Data Set Description:

The orders table lists information about orders. Here is some fields' description.

Eval_set : this is for machine learning, we don't use that.

Order_number: we don't use this field in our project.

Order_dow: the day of the week. 0 is Sunday, 1 is Monday

Order_hour_of_day: the time the order was created

Days_since_prior_order: the days between previous order and current order

order_product table specify that which products were purchased in each order. Here is some description.

Add_to_cart_order: this is the order that user put product into shopping cart since user may buy multiple products in one order.

Reordered: there is only two values for this field, 1 and 0.
'1': the product has been purchased for the user before
'0': the user did not purchase the product before

Here is the assignment 💬

1. Create one database (as showed in database diagram) in MySQL and load the data into the database on your laptop.

2. Write query to select the top 10 products sales for each day in the week (Monday to Friday), which includes product id, product name, total order amount and the day (Mon to Fri).

3. Write query to display the 5 most popular products in each aisle from Monday to Friday. Please list product id, aisle and day in the week.

3. Write query to select the top 10 products that the users have the most frequent reorder rate. You only need to give the results with product id.
   Note:
   reorder rate = (sum of re-ordered order) / (sum of all orders)

5. Business case study -1

Tracking shoppers' path to purchase

The path of your shopper can tell you what areas you need to grow or reduce, and provide insight into the motivations and interests of your customers.

5.1 Please create a report to show the shopper's aisle list for each order. For instance, in order 218 the shopper visited aisle 115, 83 and 24 (as showed in following chart). In your new table, please list order id and all unique aisle id in the order.

order_id	product_id	add_to_cart_order	reordered	aisle_id
218	1194	1	1	115
218	5578	2	1	83
218	38159	3	0	24
218	10305	4	0	24
218	38557	5	0	24

5.2 Please do some research, find the most popular shopping path. You can list your thoughts how to find these paths. You should create multiple queries and tables to save intermediate results for your research. 🗨🗨

6. Business Case Study – 2
Layout re-arrangement to promote sales

 🗨

(The following excerpt may or may not represent a real-world situation). A famous use of predictive analytics came about when Walmart noticed that many their male customers would purchase

both beer and diapers when coming into the store. As a result, the Walmart's Analytics team used predictive analytics and concluded that beer should be very close to the diaper aisle to increase sales. Walmart then set out to put those two items close to each other and this drove sales up by 35% for both items. As a result, many retail stores have run predictive analytics on different items to find the ideal way of pairing them or placing them within their store. Your task will be to find the pair of items that is most frequently bought together. Good luck!