## **Project instructions**

## **AWS Glue**

- 1. Download imba.zip from share drive, unzip all the files.
- Create an s3 bucket with name imba-<put your name here> then create the following folders: data/aisles/

data/departments/

data/orders/

data/products/

data/order\_products/

- 3. Upload files to the corresponding directory (note: gzip both order\_products\_\_prior.csv and order\_products\_\_train.csv in gitbash or any other way you like, and upload the gzip files to data/order products/).
- 4. Go to AWS Glue service, click on Crawlers and Add crawler.
- 5. Give the crawler a name, e.g. imba, click next.
- 6. Choose Data stores as Crawler source type, click next.
- 7. Choose s3 as data store and specify the Include path as: s3://ibma-<put your name here>/data, click next.
- 8. Do not add another data store and click next.
- 9. Select Create an IAM role and type a name (any name will do) in the text box, click next.
- 10. Specify Frequency as Run on demand, click next.
- 11. Create a new database by clicking Add database, name it as prod, click next.
- 12. Review the options and click Finish.
- 13. Select the crawler you just created and click Run crawler button.

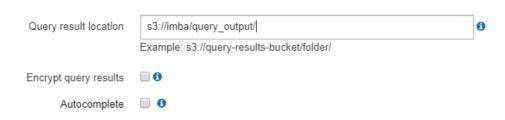
## **AWS Athena**

1. Go to AWS Athena service, click Settings button located on the top right of the page, type in the Query result location as s3:// ibma-<put your name here>/query output/:

## Settings

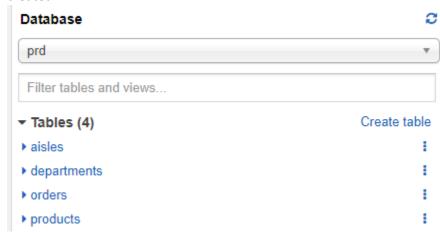
Settings apply by default to all new queries. Learn more

Workgroup: primary





2. Go to AWS Athena service, select prod data on the left pane, you should see five tables are created.



3. Check the fields for each table by expanding the table, as shown below:

```
▼ Tables (4)
                                                             Create table

▼ aisles

   aisle_id (bigint)
   aisle (string)

▼ departments

                                                                        ŧ
   department_id (bigint)
   department (string)

▼ orders

                                                                        ŧ
   order id (bigint)
   user_id (bigint)
   eval_set (string)
   order number (bigint)
   order_dow (bigint)
   order hour of day (bigint)
   days since prior order (double)
products
                                                                        ŧ
```

- - a. First drop the table by running query: DROP TABLE IF EXISTS products;
  - b. Now re-create the table by running below query:

```
CREATE EXTERNAL TABLE `products`(

`product_id` string COMMENT 'from deserializer',

`product_name` string COMMENT 'from deserializer',

`aisle_id` string COMMENT 'from deserializer',

`department_id` string COMMENT 'from deserializer')

ROW FORMAT SERDE

'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

'escapeChar'='\\',

'quoteChar'='\\',

'separatorChar'=',')

STORED AS INPUTFORMAT
```

```
'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

LOCATION

's3://imba_<put your name here>/data/products'

TBLPROPERTIES (
   'skip.header.line.count'='1')
```

- 5. Do some exploration on these tables and make sure you understand the context of them.
- 6. Design a query which join orders table and order\_products table together, and filter on eval\_set = 'prior'