

# Hands-On-Lab: Data Scientist

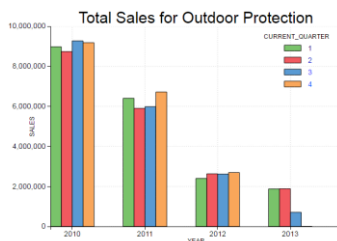
Analyze sales data in dashDB from IBM Data Science Experience using Spark based IPython Notebooks

---

## Introduction

IBM Data Science Experience is an interactive, collaborative, cloud-based environment where data scientists can use multiple tools to drive analytics and derive insights. Data Scientists can use Python, R, or Scala in Jupyter Notebooks already connected to Spark or RStudio for open source R computing environment.

This lab exercise uses an IPython Notebook in Data Science Experience to connect with dashDB, explores sales, product data and analyze sales performance for a specific product line. Pixiedust, a Python based visualization package is used to visualize the results.




---

## Steps for the Lab

---

## Provision dashDB

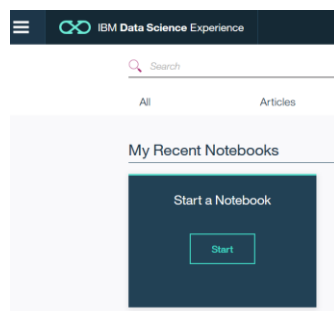
1. Login to IBM Bluemix
  - a. For existing accounts use <https://console.ng.bluemix.net/login>
  - b. Sign-up for a free trial account at <https://console.ng.bluemix.net/registration/>
2. Provision dashDB
  - a. From the Bluemix catalog menu search for “dashdb”
  - b. Click on the dashDB Icon 
  - c. Accept the default values and rename “Service name:” to ‘DS\_Sales\_DataStore’
  - d. Select “Entry Pricing Plan” (default) and click “Create” at the bottom of the page.

The lab exercise uses a sample database provisioned by dashDB service.

---

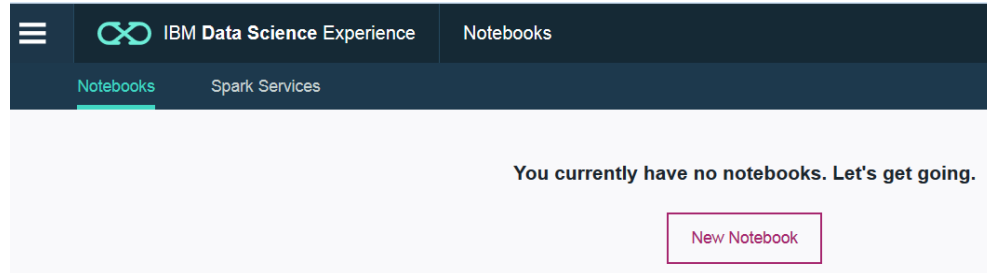
## Provision Spark Service in Data Science Experience

1. Login to IBM Data Science Experience @ <http://datascience.ibm.com/>
2. Create a new Notebook
  - a. Click on ‘Start a Notebook’



- b. If this is your first Notebook, you will get the message that you need Spark service. Click on ‘Create Spark Service [here](#)’ link in the message to provision Spark Service.

- c. Key in 'Lab\_Spark' as the service name and accept defaults (Personal plan for Spark service, New Object Storage with free plan). Click on 'Create Instance' to provision Spark service.
- d. Click on the 'New Notebook' to create a notebook.



- e. Key in 'Lab\_Notebook' as the name, click on 'from URL' link and specify url as [https://github.com/ibmdataworks/sigmo-demos/tree/master/data-scientist/interactive-analytics/DSX\\_IA\\_Lab](https://github.com/ibmdataworks/sigmo-demos/tree/master/data-scientist/interactive-analytics/DSX_IA_Lab)

The screenshot shows the 'Create Notebook' form. At the top, there's a dark blue header with the IBM logo and 'IBM Data Science Experience' with a 'Notebooks' tab. Below the header, there's a 'Create Notebook' section. It has three tabs: 'Blank', 'From File', and 'From URL'. The 'From URL' tab is selected. Below the tabs, there's a 'Name\*' field with the text 'Lab\_Notebook' and a '38 Characters Remaining' indicator. Below that is a 'Description' field with a placeholder 'Type your Description here'. At the bottom is a 'Notebook URL\*' field with the text 'https://github.com/ibmdataworks/sigmo-demos/tree/master/data-scientist/interactive-analytics/I'.

- f. Select 'Trust this Notebook to run with your Privileges' and click on 'Create Notebook'

Spark Service\*

Lab\_Spark

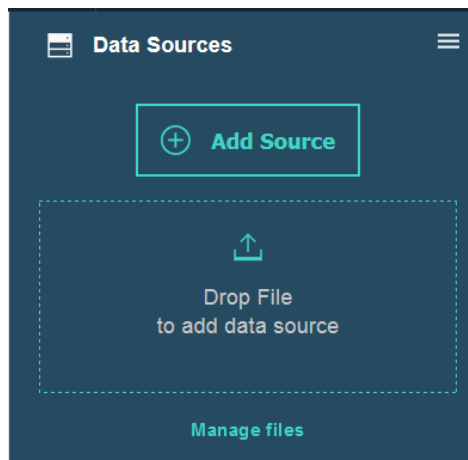
Associate this notebook with the IBM Analytics for Apache Spark Service of your choice.

☒ Trust this Notebook to run with your Privileges

Allow that code in the notebook can run with and use your privileges to invoke Bluemix as well as Data and Analytics APIs and services. You should allow this only if you trust this Notebook.



Cancel Create Notebook

3. Once notebook is created, to access data sources in Notebooks we , add the connections information.
  - a. On top left menu, click File -> Connections -> Create Connection
  - b. Key in 'dashDB' as connection name
  - c. From drop down list of 'Target Service Instance' select 'DS\_Sales\_DataStore', the dashDb service you provisioned in Bluemix.
  - d. Select 'Database' as 'BLUDB'.
  - e. Click 'Create'.
4. Open your Notebook and add the data source
  - a. Navigate to home with a click on 'IBM Data Science Experience' (top left) and open 'Lab Notebook'
  - b. Click on 'Data Sources' and 'Add Source'



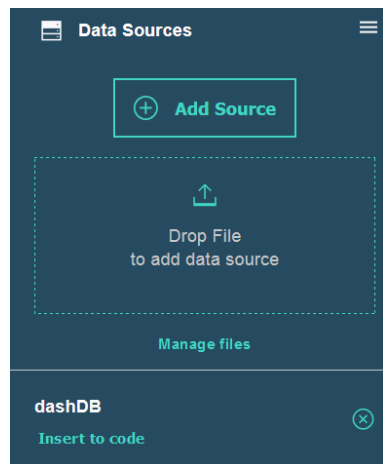
- c. Click on Connections, select the connection named 'dashDB'
- d. Click 'Add Data Source'

# Add Data Source

From File <u>Connections</u>		
Search by Name 		
Name	Description	Select
dashDB		

## 5. Connect to the sample sales database in dashDB

- Place the cursor in the first cell. This should be empty, if not **clear** the contents.
- Click on 'Insert Code' to insert credentials for the data source 'dashDB'

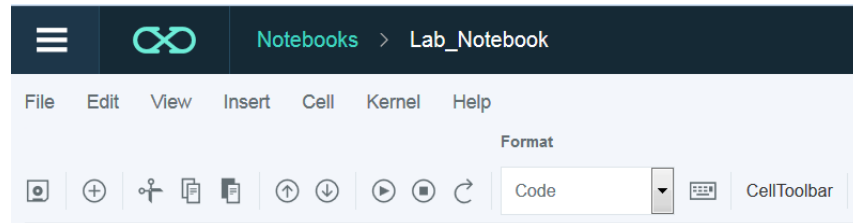


- Rename** inserted array 'credentials\_1' to 'credentials'

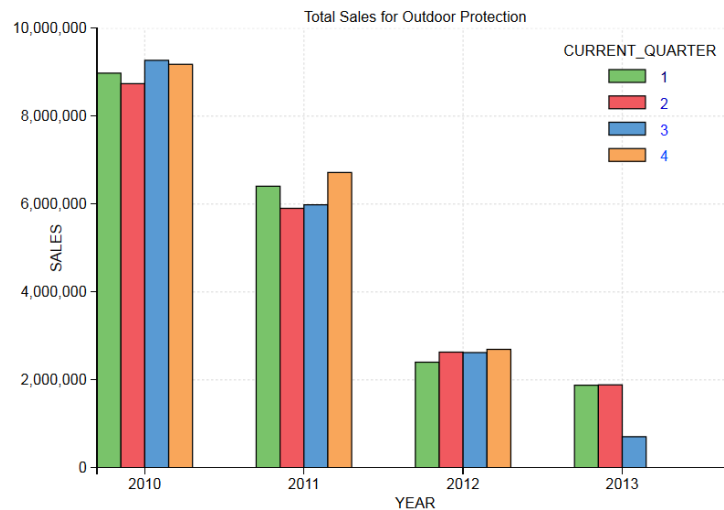
Insert dashDB credentials in the empty cell using 'Insert to Code' link in data sources

```
In [1]: credentials_1 = {
  'port': '50000',
  'db': 'BLUDB',
  'username': 'dash102372',
  'ssljdbcurl': 'jdbc:db2://awh-yp-small102.services.dal.ibmcloud.net:50001/BLUDB:sslConnection=true;',
  'host': 'awh-yp-small102.services.dal.ibmcloud.net',
  'https_url': 'https://awh-yp-small102.services.dal.ibmcloud.net:8443',
  'dsn': 'DATABASE=BLUDB;HOSTNAME=awh-yp-small102.services.dal.ibmcloud.net;PORT=50000;PROTOCOL=TCPIP;UI',
  'hostname': 'awh-yp-small102.services.dal.ibmcloud.net',
  'jdbcurl': 'jdbc:db2://awh-yp-small102.services.dal.ibmcloud.net:50000/BLUDB',
  'ssldsn': 'DATABASE=BLUDB;HOSTNAME=awh-yp-small102.services.dal.ibmcloud.net;PORT=50001;PROTOCOL=TCPIP',
  'uri': 'db2://dash102372:Jo3IgCf3mq8m@awh-yp-small102.services.dal.ibmcloud.net:50000/BLUDB',
  'password': 'Jo3IgCf3mq8m'
}
```

- Use the arrow icon in toolbar to execute the code in cells



- e. Place cursor in cell1, execute code to capture credentials
- f. Place cursor in cell2, execute code to connect, load register tables in dashDB
6. Query, join and group data in SparkSQL Data frames to build aggregated sales for the product line 'Outdoor Protection'.
  - a. Place cursor in cell3, execute code and build the dataset for aggregated sales



7. You can move the mouse over the visual to see data points. The chart shows a steady decline in sales for 'Outdoor Protection' product line.
8. Pixiedust is open source visualization package developed in IBM labs. You can display a SparkSQL data frame with a single API call. You have access to raw data, visual options and can stow data away to IBM Cloudant or , Object Storage,

## Next Steps

Build your next Spark Analytics project using IBM Data Science Experience .

---

### End of Lab

---