

Introduction to Data Wrangling with Jupyter Notebooks

IBM Developer

Upkar Lidder
> ulidder@us.ibm.com
> @lidderrupk

<http://bit.ly/spectra-ibm>
<https://slides.com/upkar/jupyter>

Prerequisites

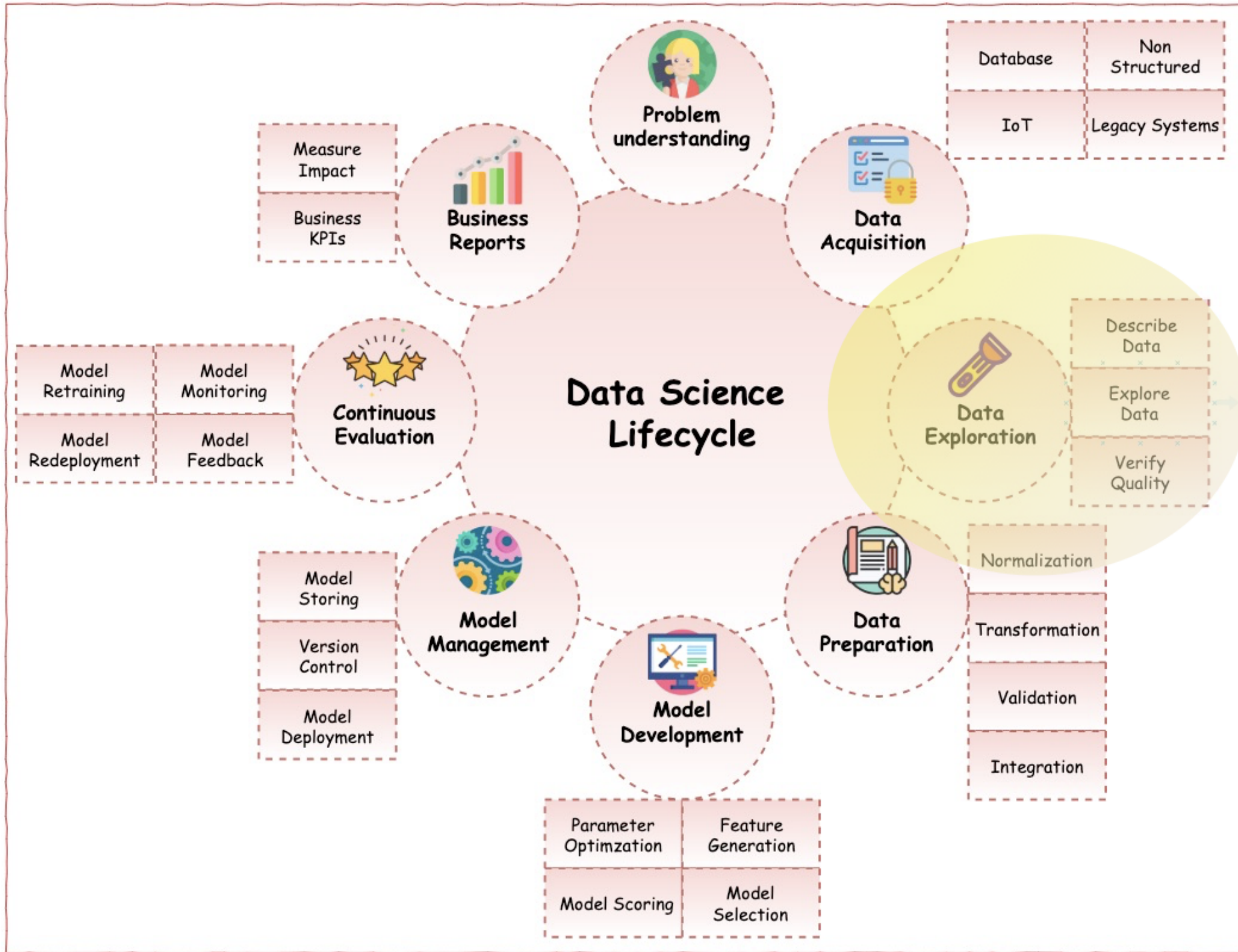
1. Create IBM Cloud Account using THIS URL

<http://bit.ly/spectra-ibm>

2. Check your email and activate your account. Once activated, log back into your IBM Cloud account using the link above.

3. If you already have an account, use the above URL to sign into your IBM Cloud account.

Data Science Lifecycle



Workshop - Goals

Get acquainted with
Pandas and Jupyter
Notebooks on the cloud
and analyze a movies
dataset!

Steps

1. Sign up / Log into *IBM Cloud* - <http://bit.ly/spectra-ibm>
2. Create *a Watson Studio Service*.
3. Create a *new project*
4. Import the sample notebook to your project
5. *RUN* the cells and explore data!

Step 1 - sign up/ log into IBM Cloud

Already have an IBM Cloud account? [Log in](#)

Create an IBM Cloud Account

Email*

→

First Name*

Last Name*

Country or Region*

United States

Password*

IBM may use my contact data to keep me informed of products, services and offerings:

☒ by email. ☒ by telephone.

You can withdraw your marketing consent at any time by sending an email to netsupp@us.ibm.com. Also you may unsubscribe from receiving marketing emails by clicking the unsubscribe link in each such email.

More information on our processing can be found in the [IBM Privacy Statement](#). By submitting this form, I acknowledge that I have read and understand the IBM Privacy Statement.

Create Account

<http://bit.ly/spectra-ibm>

Step 2 - locate Watson Studio in the catalog

The screenshot shows the IBM Cloud Catalog interface. At the top, a dark navigation bar contains the IBM Cloud logo, a search icon, and links for Catalog, Docs, Support, and Manage. A user account link for 'Upkar Lidder's Account' and a notification bell with the number 5 are also present. Below the navigation bar, a blue banner states: 'Try the best of the Catalog for free with no time restrictions with Lite plans. The Lite filter is enabled. Remove the filter to see the full Catalog.' with a close button (X).

The main section is titled 'Catalog'. Below the title is a search bar containing the text 'label:lite watson studio' and a 'Filter' button. On the left side, there is a sidebar with a list of categories: 'All Categories (2)', 'VPC Infrastructure', 'Compute', 'Containers', 'Networking', 'Storage', 'AI (1)', 'Analytics (1)', 'Databases', 'Developer Tools', 'Integration', 'Internet of Things', 'Security and Identity', 'Starter Kits', and 'Web and Mobile'. The 'AI (1)' category is highlighted with a red box.

The main content area displays two results under the 'AI' heading. The first result is 'Watson Studio', which is highlighted with a red box. It includes an icon of a flask and test tube, the text 'Watson Studio', 'Lite • IBM • IAM-enabled', and a description: 'Embed AI and machine learning into your business. Create custom models using your own data.' Below this, under the 'Analytics' heading, is the 'Analytics Engine' result, which includes an icon of a gear, the text 'Analytics Engine', 'Lite • IBM • IAM-enabled', and a description: 'Flexible framework to deploy Hadoop and Spark analytics applications.'

On the right side of the main content area, there is a vertical 'FEEDBACK' button and a blue circular chat icon at the bottom right.


Step 3 - create new watson studio service

IBM Cloud

Catalog Docs Support Manage Upkar Lidder's Account

5

[← View all](#)



Watson Studio

Lite • IBM

Watson Studio democratizes machine learning and deep learning to accelerate infusion of AI in your business to drive innovation. Watson Studio provides a suite of tools and a collaborative environment for data scientists, developers and domain experts.

[View Docs](#) [Terms](#)

AUTHOR [IBM](#)

PUBLISHED [07/18/2019](#)

TYPE [Service](#)

Service name:

Watson Studio-39

Choose a region/location to deploy in:

Dallas

Select a resource group:

Default

Tags:

Examples: env:dev, version-1

Features

✓	Lite	1 authorized user	Free
		50 capacity unit-hours monthly limit 1 free small compute environment with 1 vCPU and 4 GB RAM (does not require capacity unit-hours)	

The Lite plan for Watson Studio offers everything you need to become a better data scientist or domain expert in a collaborative environment.

Lite plan services are deleted after 30 days of inactivity.

you don't
sample,
the power of

- Power on demand

Enterprise-scale features on demand. From data exploration and preparation to

Add to estimate

Create

FEEDBACK

Step 4 - launch Watson Studio

☰

IBM Cloud

🔍

Catalog

Docs

Support

Manage ▾

Upkar Lidder's Account


✎

5

Manage

Plan

Resource list /


 Watson Studio-39

Resource group: Default

Location: Dallas

[Add Tags](#)

⋮



Watson Studio

Welcome to Watson Studio. Let's get started!

Get Started

Step 5 - create new project and pick empty template

[← Back](#)

Create a project

Choose whether to create an empty project or to preload your project with data and analytical assets. Add collaborators and data, and then choose the right tools to accomplish your goals. Add services as necessary.



Create an empty project

Add the data you want to prepare, analyze, or model. Choose tools based on how you want to work: write code, create a flow on a graphical canvas, or automatically build models.

NEW

AutoAI experiment tool: Fully automated approach to building a classification or re...

USE TO

Prepare and visualize data
Analyze data in notebooks
Train models



Create a project from a sample or file

Get started fast by loading existing assets. Choose a project file from your system, or choose a curated sample project.

USE TO

Learn by example
Build on existing work
Run tutorials

Step 6a - name your project and create storage service

IBM Watson Studio

Upgrade

Upkar Lidder's Account

UL

New project

Define project details

Name

spectra-project

Description

Project description

Choose project options

☐ Restrict who can be a collaborator

Project will include integration with [Cloud Object Storage](#) for storing project assets.

Define storage

① Select storage service

Add

Add an object storage instance and then return to this page and click Refresh.

② Refresh

Cancel

Create

IBM Developer

@liddyurpk

Step 6b - add storage opens a new page

The screenshot displays the IBM Watson Studio interface. At the top, the header includes the IBM Watson Studio logo, an 'Upgrade' button, a notification bell, and the user's account 'Upkar Lidder's Account'. Below the header, there's a section for 'Access and Key Management' with a brief description of IAM policies. The main content area features a table with pricing plans for Cloud Object Storage. The table has three columns: PLAN, FEATURES, and PRICING. The 'Lite' plan is selected, showing features like 1 COS Service Instance, storage up to 25 GB/mo, and a price of 'Free'. A 'Confirm Creation' dialog box is open on the right, with a purple border. It contains fields for 'Plan' (set to 'Lite'), 'Resource group' (set to 'Default'), and 'Service name' (set to 'cloud-object-storage-na'). At the bottom of the dialog, there are 'Cancel' and 'Confirm' buttons. A large purple number '2' is next to the 'Confirm' button. In the bottom right corner of the main interface, there is a 'Create' button with a purple number '1' next to it, and a 'Cancel' button next to it. The bottom of the page shows the 'Standard' plan option and a note about no minimum fee.

IBM Watson Studio

Upgrade

Upkar Lidder's Account

worldwide.

you to reduce costs even further with our lowest priced Archive storage.

Access and Key Management

IBM Identity and Access Management (IAM) policies allow for granular access control at the bucket level using role-based policies. Key Protect support allows customers to have their own managed encryption keys for higher level data security.

Pricing Plan: Monthly Process shown above reflect the: [United States](#)

PLAN	FEATURES	PRICING
<input checked="" type="radio"/> Lite	1 COS Service Instance Storage up to 25 GB/mo. Up to 20,000 GET requests/mo. Up to 2,000 PUT requests/mo. Up to Data Retrieval 10 GB/mo. Up to 5GB Public Outbound Applies to aggregate total across all storage bucket classes	Free
<input type="radio"/> Standard	There is no minimum fee, so you pay only for what you use.	Expand each section to view details

The Lite service plan for Cloud Object Storage includes Regional and Cross Regional resiliency, flexible data classes, and built in security.

Confirm Creation

Plan

Lite

Resource group

Default

Service name

cloud-object-storage-na

2 Cancel **Confirm**

1 Cancel **Create**

Step 6c - you will be taken back to the first page after storage

IBM Watson Studio

Upgrade

Upkar Lidder's Account

UL

New project

Define project details

Name

spectra-project

Description

Project description

Choose project options

☐ Restrict who can be a collaborator

Project will include integration with [Cloud Object Storage](#) for storing project assets.

Define storage

1 Select storage service

Add

Add an object storage instance and then return to this page and click Refresh.

2 Refresh

Storage

cloud-object-storage-na

2

Cancel

Create

Step 7 - add notebook feature to your project

The screenshot displays the IBM Watson Studio interface for a project named 'spectra-project'. The top navigation bar includes a hamburger menu, the text 'IBM Watson Studio', an 'Upgrade' button, a notification bell, and the user's account 'Upkar Lidder's Account'. Below this, a breadcrumb trail shows 'My Projects / spectra-project'. A secondary navigation bar contains icons for 'Launch IDE', 'Add to project' (highlighted with a purple box and a purple '2'), and other utility icons. The main interface has a tabbed view with 'Overview' and 'Assets' (highlighted with a purple box and a purple '1'). The 'Assets' tab shows a search bar and a list of 'Data assets'. A modal window titled 'Choose asset type' is open, displaying a grid of 'AVAILABLE ASSET TYPES'. The 'Notebook' option in this grid is highlighted with a purple box and a purple '3'. Other asset types include Data, Connection, Connected data, AutoAI experiment (marked 'NEW'), Dashboard, Visual Recognition, Natural Language CL..., Watson Machine Lea..., Experiment, Modeler flow, Data Refinery flow, Streams flow, and Synthesized neural n... A 'Close' button is at the bottom right of the modal. In the background, a 'Drop files here or browse for files to upload.' area is visible. At the bottom of the screen, a section titled 'Start analyzing data' provides instructions: 'To create an analytic asset like a notebook or a model, click **Add to Project**'.

IBM Watson Studio

Upgrade

Upkar Lidder's Account

My Projects / spectra-project

Launch IDE

Add to project

Overview

Assets

Environments

Jobs

Bookmarks

Deployments

Access Control

Settings

Load

Files

Catalog

What assets are you looking for

Data assets

Choose asset type

AVAILABLE ASSET TYPES

Data

Connection

Connected data

AutoAI experiment

Notebook

Dashboard

Visual Recognition ...

Natural Language CL...

Watson Machine Lea...

Experiment

Modeler flow

Data Refinery flow

Streams flow

Synthesized neural n...

Close

Drop files here or [browse](#) for files to upload.

Start analyzing data

To create an analytic asset like a notebook or a model, click **Add to Project**

Step 8 - import notebook, get link from github

IBM Watson Studio

Upgrade

Upkar Lidder's Account

My Projects / spectra-project / Add Notebook

New notebook

Blank From file **From URL**

Name

spectra-notebook 25 characters remaining

Description (optional)

Type your Description here 500 characters remaining

Select runtime

Default Python 3.6 XS (2 vCPU and 8 GB RAM)

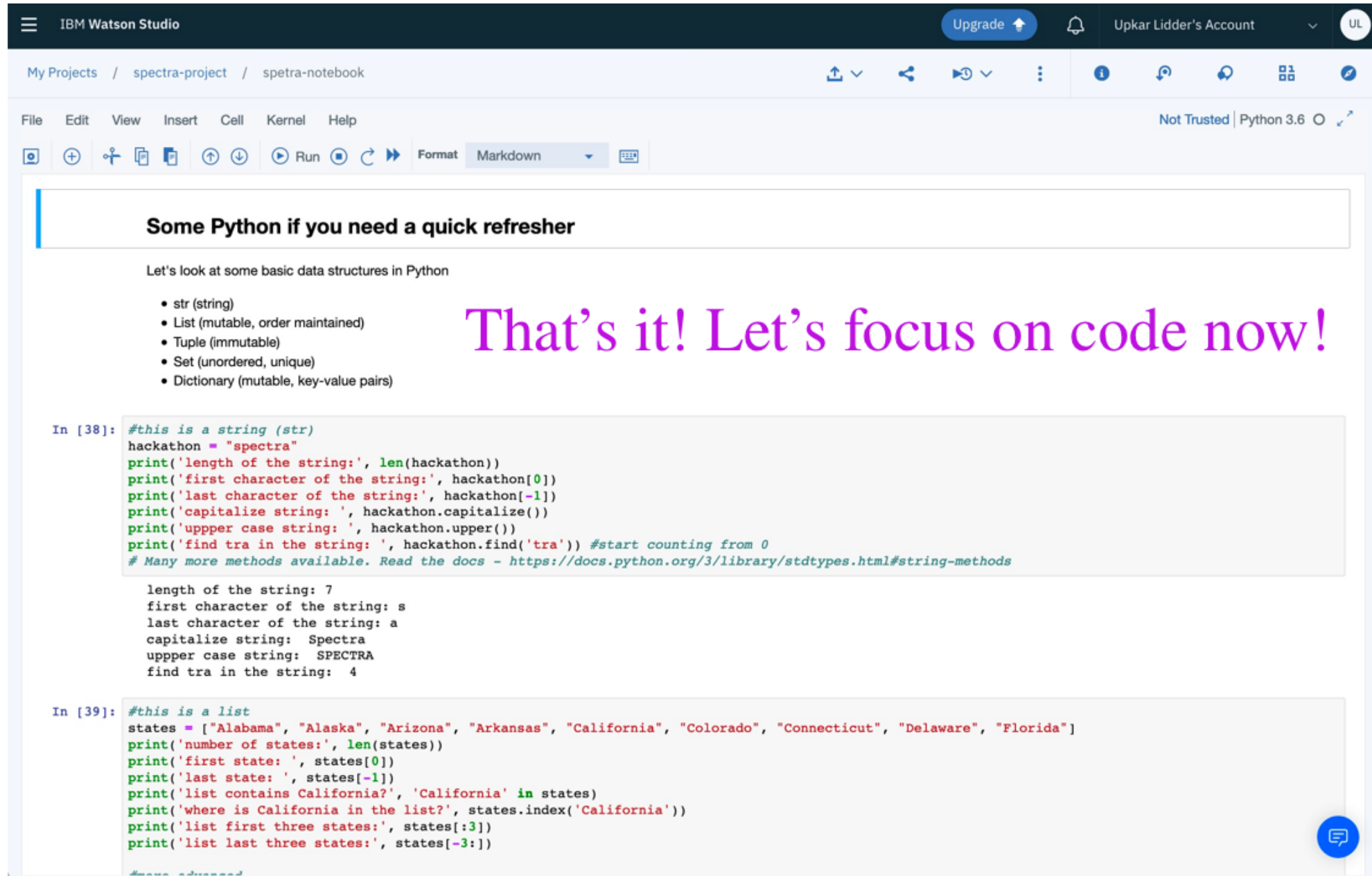
The selected runtime has 2 vCPU and 8 GB RAM and consumes 1 capacity unit per hour.
[Learn more](#) about capacity unit hours and Watson Studio pricing plans.

Notebook URL

https://github.com/lidderupk/lidderupk-ibmdevelopersf-jupyternotebooks/blob/

Cancel **Create Notebook**

Step 9 - Let's look at data now!



The screenshot shows the IBM Watson Studio interface. At the top, there's a navigation bar with 'IBM Watson Studio', an 'Upgrade' button, a notification bell, 'Upkar Lidder's Account', and a user profile icon 'UL'. Below this is a breadcrumb trail: 'My Projects / spectra-project / spetra-notebook'. A toolbar contains icons for file operations, a 'Run' button, and a 'Format' dropdown set to 'Markdown'. The main area displays a notebook with the title 'Some Python if you need a quick refresher'. The text 'Let's look at some basic data structures in Python' is followed by a bulleted list of Python data structures: str (string), List (mutable, order maintained), Tuple (immutable), Set (unordered, unique), and Dictionary (mutable, key-value pairs). A large purple text overlay reads 'That's it! Let's focus on code now!'. Below this, two code cells are shown. The first cell, labeled 'In [38]:', contains Python code for string operations on the variable 'hackathon' (value 'spectra'), including printing length, first/last characters, capitalization, and finding a substring. The second cell, labeled 'In [39]:', contains Python code for list operations on the variable 'states' (a list of US state names), including printing length, first/last states, checking for 'California', and slicing the list. A chat icon is visible in the bottom right corner of the notebook area.

IBM Watson Studio

Upgrade

Upkar Lidder's Account

UL

My Projects / spectra-project / spetra-notebook

File Edit View Insert Cell Kernel Help

Not Trusted | Python 3.6

Format Markdown

Some Python if you need a quick refresher

Let's look at some basic data structures in Python

- str (string)
- List (mutable, order maintained)
- Tuple (immutable)
- Set (unordered, unique)
- Dictionary (mutable, key-value pairs)

That's it! Let's focus on code now!

```
In [38]: #this is a string (str)
hackathon = "spectra"
print('length of the string:', len(hackathon))
print('first character of the string:', hackathon[0])
print('last character of the string:', hackathon[-1])
print('capitalize string: ', hackathon.capitalize())
print('upppe case string: ', hackathon.upper())
print('find tra in the string: ', hackathon.find('tra')) #start counting from 0
# Many more methods available. Read the docs - https://docs.python.org/3/library/stdtypes.html#string-methods

length of the string: 7
first character of the string: s
last character of the string: a
capitalize string: Spectra
upppe case string: SPECTRA
find tra in the string: 4
```

```
In [39]: #this is a list
states = ["Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecticut", "Delaware", "Florida"]
print('number of states:', len(states))
print('first state: ', states[0])
print('last state: ', states[-1])
print('list contains California?', 'California' in states)
print('where is California in the list?', states.index('California'))
print('list first three states:', states[:3])
print('list last three states:', states[-3:])
```


Some links for the workshop

IBM Cloud account - <http://bit.ly/spectra-ibm>

Jupyter Notebook - <https://github.com/lidderupk/lidderupk-ibmdevelopersf-jupyternotebooks/blob/master/asset/spectra-pandas.ipynb>

Datasets

Casts - <https://ibm.box.com/shared/static/569iue5znz5angfxaajbd7olgegk0bz.csv>

Release Dates - <https://ibm.box.com/shared/static/fxu6rhfktvjs0uvgtbhjsp5g5k9qgjh1.csv>

Titles - <https://ibm.box.com/shared/static/cw3wqtzuljiyqz4kbuk26ojrrm9rzfow.csv>

Workshop Github - <https://github.com/lidderupk/lidderupk-ibmdevelopersf-jupyternotebooks>

Some links to get data

US Government Open Data - <https://www.data.gov/>

San Francisco Open Data - <https://datasf.org/opendata/>

Data Asset eXchange - <https://developer.ibm.com/exchanges/data/>

Kaggle Datasets - <https://www.kaggle.com/datasets>

Google Datasets - <https://cloud.google.com/public-datasets/>

Curated on Github - <https://github.com/awesomedata/awesome-public-datasets>

Ryan Anderson Blog - https://dreamtolearn.com/ryan/1001_datasets

07.26.19

SERVERLESS DEVELOPER

SUMMIT

GALVANIZE, SAN FRANCISCO

Thank you

Let's chat !

Upkar Lidder, IBM

@lidderupk

<https://github.com/lidderupk/>

ulidder@us.ibm.com

IBM Developer



@lidderupk