Extract Insights from Social Media Posts with Watson and Spark in Watson Studio

According to statistics from Cumulus Media (https://www.weforum.org/agenda/2018/05/what-happens-in-an-internet-minute-in-2018), 38 million WhatsApp messages, 3.7 million Google searches, and 481,000 tweets occur in every internet minute (as of May, 2018). Interestingly, much of the massive amounts of data generated and consumed is unstructured data in the form of text, speech, images, and video. To manage and leverage this data, there is a need for AI tools and services to help extract insights from Big Data.

While historically leveraging AI was restricted to only a few skilled experts, we have seen a major shift in the last few years where major platform providers have offered consumable AI services. In particular, IBM Watson AI (https://www.ibm.com/watson/) is a platform that offers a wide variety of consumable AI services designed to extract knowledge from unstructured data in all possible formats; text, speech, and images. Several AI solutions have leveraged the Watson AI platform (https://www.ibm.com/watson/developer/) to address a variety of business problems such as:

- Providing virtual assistants (or chat bots)
- Social media listening
- Marketing campaign analytics
- Audience segmentation and matching
- Discovering insights from large amounts of data.

In several AI solutions, we find the most impactful results achieved by combining the Watson AI services with analytics solutions optimized for big data. In this tutorial, we step through developing an example solution combining Watson AI services with custom machine learning solutions by leveraging IBM Watson Studio (https://www.ibm.com/cloud/watson-studio). This blog references the Python WatsonSocialMediaInsights (https://github.com/joe4k/dsxwdc/blob/master/social_media_insights_with_watson.ipynb) notebook which would be uploaded and executed in Watson Studio.

Prerequisites

To be able to complete the tutorial, you will need to have access to an IBM Cloud (https://www.ibm.com/cloud/) account where you will provision a number of Watson AI services. In Watson Studio, you will be stepping through a Python notebook for acquiring data; curating, ingesting, and enriching the acquired data; and running various analysis and machines learning techniques on the data.

For this tutorial, we will provide you with access to a DB2 instance which holds a sample collection of tweets related to singers most tweeted about during the time period from July 5, 2017 to July 10, 2017.

To run the notebook on other data set, you need to point it to your own data set. Please note that if the schema for your dataset is different from the twitter data, you will need to edit the notebook to handle the nuances specific to your data.

Instructions for creating an IBM Cloud account:

To continue through the tutorial, you will need to have an IBM Cloud account so you can create the required Watson AI. To create an IBM Cloud account:

- Point your browser to https://ibm.com/cloud
- Click on **Cloud sign-up/log-in button** in the top right and follow the steps to either log-in if you have an account already or provide the required information such as email, name, and password to sign up for a new IBM Cloud account.

Now that you have an IBM Cloud account, you can execute these steps from your terminal to create the required services for this tutorial, namely NLU, and Personality Insights.

- Download and install IBM Cloud CLI by following the instructions on this page (https://console.bluemix.net/docs/cli/index.html#overview)

  - Open a Terminal window
  - Execute this curl command in your terminal window
    curl -sL https://ibm.biz/idt-installer | bash
  - Verify installation of ibmcloud CLI
    ibmcloud dev help
    ➔ this should provide help information

- Connect to your IBM Cloud account using IBM Cloud CLI from a Terminal window

  - Connect to your IBM Cloud account
    ibmcloud login

    API endpoint: https://api.ng.bluemix.net
    username: ***your_IBMCloud_username* ➔ *specify your IBM Cloud username***
    password: ***your_IBMCloud_password* ➔ *specify your IBM Cloud password***

■ Create Watson NLU service and capture the credentials

- Create an NLU service using free plan and call it studionlu

  ```
  ibmcloud resource service-instance-create studionlu natural-language-understanding free us-south
  ```

- Create service credentials for the service you created

  ```
  ibmcloud resource service-key-create keycreds Manager –instance-name studionlu
  ```

  ➔ keycreds is any name you provide to be assigned to your service key
  ➔ Manager is the ROLE you’re assigning to the IAM API key generated for service credentials
  ➔ studionlu is the name of the instance we created in the previous step

- Copy the credentials you get as you will need them later

■ Repeat the steps above to create Personality Insights service

- Create a Personality Insights service using lite plan and call it studiopi

  ```
  ibmcloud resource service-instance-create studiopi personality-insights lite us-south
  ```

- Create service credentials for the service you created

  ```
  ibmcloud resource service-key-create keycredspi Manager –instance-name studiopi
  ```

  ➔ keycredspi is any name you provide to be assigned to your service key
  ➔ Manager is the ROLE you’re assigning to the IAM API key generated for service credentials
  ➔ studiopi is the name of the instance we created in the previous step

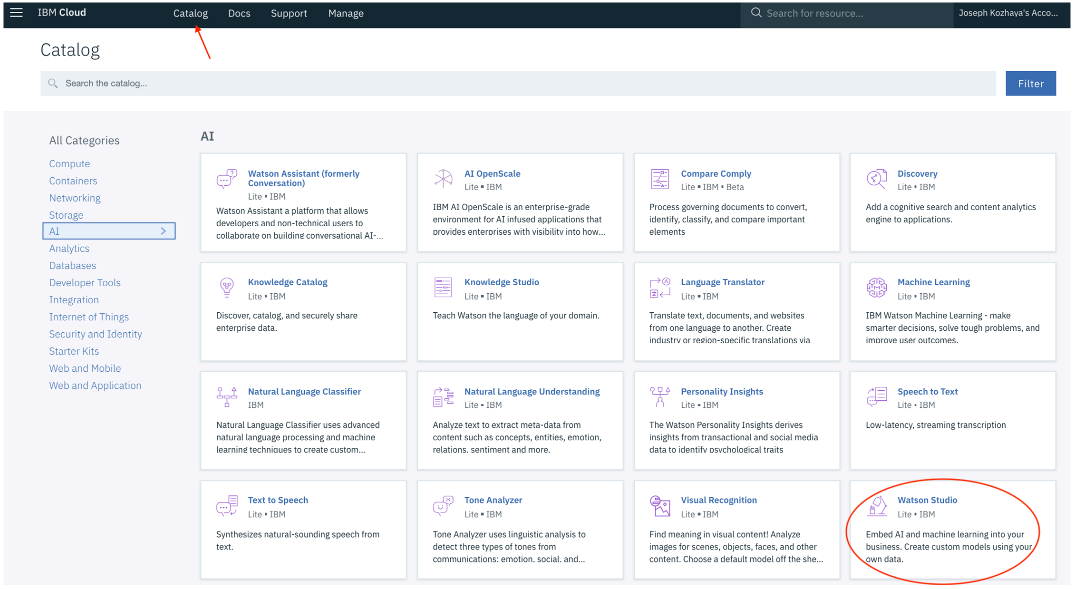
- Copy the credentials you get as you will need them later

Instructions for creating a Watson Studio service:

To create a Watson Studio service, execute the following steps:

1- Log in to your IBM Cloud account (https://ibm.com/cloud)
2- Click on Catalog tab in the top bar (highlighted with an arrow in Figure 1 below).
3- In the left navigation column, under All Categories select the ***AI*** category.

4- Select ***Watson Studio*** service (highlighted with an oval in Figure 1 below).



Catalog

AI

Watson Assistant (formerly Conversation)
Lite • IBM
Watson Assistant a platform that allows developers and non-technical users to collaborate on building conversational AI-...

AI OpenScale
Lite • IBM
IBM AI OpenScale is an enterprise-grade environment for AI infused applications that provides enterprises with visibility into how...

Compare Comply
Lite • IBM • Beta
Process governing documents to convert, identify, classify, and compare important elements

Discovery
Lite • IBM
Add a cognitive search and content analytics engine to applications.

Knowledge Catalog
Lite • IBM
Discover, catalog, and securely share enterprise data.

Knowledge Studio
Lite • IBM
Teach Watson the language of your domain.

Language Translator
Lite • IBM
Translate text, documents, and websites from one language to another. Create industry or region-specific translations via...

Machine Learning
Lite • IBM
IBM Watson Machine Learning - make smarter decisions, solve tough problems, and improve user outcomes.

Natural Language Classifier
IBM
Natural Language Classifier uses advanced natural language processing and machine learning techniques to create custom...

Natural Language Understanding
Lite • IBM
Analyze text to extract meta-data from content such as concepts, entities, emotion, relations, sentiment and more.

Personality Insights
Lite • IBM
The Watson Personality Insights derives insights from transactional and social media data to identify psychological traits

Speech to Text
Lite • IBM
Low-latency, streaming transcription

Text to Speech
Lite • IBM
Synthesizes natural-sounding speech from text.

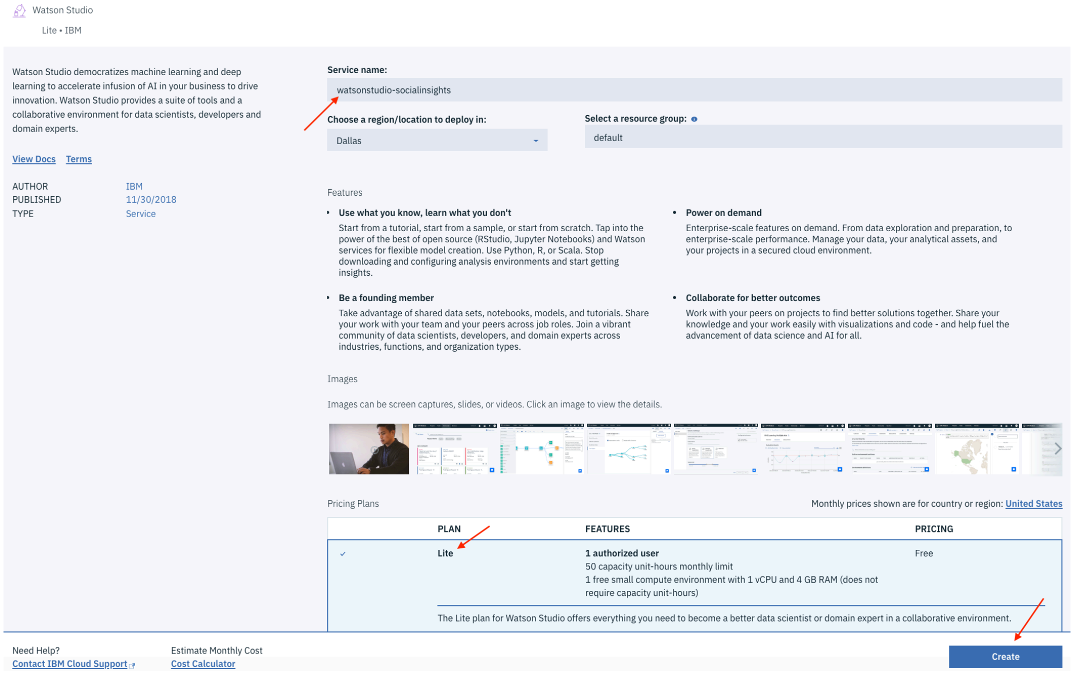
Tone Analyzer
Lite • IBM
Tone Analyzer uses linguistic analysis to detect three types of tones from communications: emotion, social, and...

Visual Recognition
Lite • IBM
Find meaning in visual content! Analyze images for scenes, objects, faces, and other content. Choose a default model off the she...

Watson Studio
Lite • IBM
Embed AI and machine learning into your business. Create custom models using your own data.

Figure 1: Watson Studio service

5- On the next screen, provide a name for your service (optional), leave the region/location and the resource group as specified (Dallas and default), select the Lite plan and hit Create.



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.

AUTHOR IBM
PUBLISHED 11/30/2018
TYPE Service

Service name: watsonstudio-socialinsights

Choose a region/location to deploy in: Dallas

Select a resource group: default

Features

- **Use what you know, learn what you don't**
  Start from a tutorial, start from a sample, or start from scratch. Tap into the power of the best of open source (RStudio, Jupyter Notebooks) and Watson services for flexible model creation. Use Python, R, or Scala. Stop downloading and configuring analysis environments and start getting insights.
- **Power on demand**
  Enterprise-scale features on demand. From data exploration and preparation, to enterprise-scale performance. Manage your data, your analytical assets, and your projects in a secured cloud environment.
- **Be a founding member**
  Take advantage of shared data sets, notebooks, models, and tutorials. Share your work with your team and your peers across job roles. Join a vibrant community of data scientists, developers, and domain experts across industries, functions, and organization types.
- **Collaborate for better outcomes**
  Work with your peers on projects to find better solutions together. Share your knowledge and your work easily with visualizations and code - and help fuel the advancement of data science and AI for all.

Images

Images can be screen captures, slides, or videos. Click an image to view the details.

Pricing Plans

Monthly prices shown are for country or region: United States

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

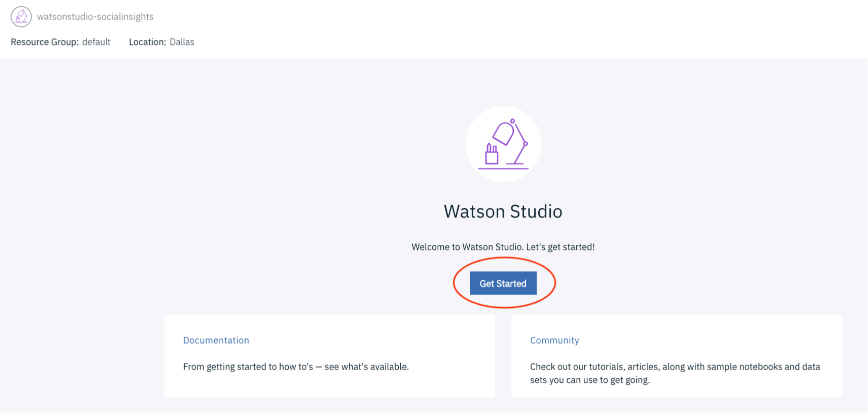
The Lite plan for Watson Studio offers everything you need to become a better data scientist or domain expert in a collaborative environment.

Need Help?
Contact IBM Cloud Support

Estimate Monthly Cost
Cost Calculator

Figure 2: Watson Studio service creation page

6- On the next page, click Get Started to launch Data Science Experience.


watsonstudio-socialinsights

Resource Group: default Location: Dallas

Watson Studio

Welcome to Watson Studio. Let's get started!

Documentation

From getting started to how to's — see what's available.

Community

Check out our tutorials, articles, along with sample notebooks and data sets you can use to get going.

Figure 3: Watson Studio launch page from IBM Cloud

Alternatively, you can launch Watson Studio by pointing your browser to https://dataplatform.ibm.com, selecting Log In button, and logging in with your IBM Cloud username and password.

Problem Statement

The problem we address in this tutorial is that of brand analytics, user segmentation, and personalized messaging. The solution involves collecting social media posts referencing a brand, understanding the sentiment towards the brand as well as segmenting the consumers based on multiple parameters such as number of followers, number of posts, sentiment, and personality profile. Given the more granular segmentation, the brand manager and marketing teams can then provide targeted messaging and marketing to reach consumers in a more personal manner which resonates better with consumers.

To keep the tutorial generic, we will not reference specific brands; instead, we'll leverage a dataset of tweets on 3 popular musicians, @katyperry, @justinbieber, and @taylorswift13, and run our analysis on that data.

Overview of Tools and Services

Before diving into the details of the solution, we'll describe the various tools and services we'll leverage. Specifically, we'll rely on Twitter, Watson Natural Language Understanding , Watson Personality Insights, Db2 Warehouse, and Watson Studio.

Twitter

Social media are computer mediated technologies that facilitate the creation and sharing of information, ideas, and thoughts by users via virtual communities and networks. Some of the popular social media platforms include Facebook, Twitter, LinkedIn, Instagram, Pinterest, and Snapchat.

Social media listening has become common practice for brands to connect with their consumers and better understand and assess how consumers perceive the brand. Social media listening refers to collecting social media posts from various platforms and analyzing them to understand overall consumer perception.

Watson AI

IBM Watson AI is a platform of AI services that enables developers to build solutions that help humans extract insights from big data. Watson AI services offer a wide range of capabilities to understand and extract insights from unstructured data including text, speech, and images. In this tutorial, we leverage sentiment analysis, and keyword extraction, two of the features of Natural Language Understanding (NLU) service, and Personality Insights to extract sentiment and keywords expressed in tweets and personality profile of the users sharing the tweets.

Db2 Warehouse on Cloud

IBM Db2 Warehouse on Cloud (https://www.ibm.com/cloud/db2-warehouse-on-cloud) is a database that is designed for performance and scale with compatibility to a wide range of tools. The massively parallel processing (MPP) options enable increased performance and scale by adding more servers to your cluster. The dynamic in-memory columnar store technology minimizes I/O and delivers an order of magnitude speed when compared to row-store databases.

Watson Studio

IBM Watson Studio is a cloud-based social workspace that helps data professionals create, consolidate and collaborate on building solutions for capturing insights from data across multiple open source tools such as R, Python and Scala. IBM Watson Studio is designed to help data explorers leverage a rich set of open source capabilities in analyzing large data sets and collaborate with colleagues in a social collaborative data-driven environment.
Your Watson Studio account includes access to Apache Spark engine. Apache Spark (http://spark.apache.org/) is a fast open-source cluster computing engine for efficient large-scale data processing. Apache Spark technology enables programs to run up to 100 times faster than Hadoop MapReduce in-memory or 10 times faster on disk. In Watson Studio, you can use Spark for your Python, Scala, or R notebooks.

For storage needs, Watson Studio requires a Cloud Object Storage instance. The Object Storage service provides an unstructured cloud data store where you can store your files including images, documents, and more. We'll create an object storage instance as part of creating a new project in Watson Studio.

To summarize, Watson Studio provides a social collaborative environment to enable data professionals to upload large data sets into Object Storage service (created on IBM Cloud) and leverage the fast Apache Spark computing engine to efficiently explore, analyze, visualize, and extract insights from large structured and unstructured data-sets. It also offers easy and seamless connection to github where you can upload and share your notebooks. The community feature of Watson Studio makes it easy to share and explore a variety of notebooks, data-sets, and tutorials built by all Watson Studio community members.

In this tutorial, we will focus on leveraging Watson Studio for building Python notebooks to analyze Twitter data and integrate with Watson AI services to extract sentiment and keywords from the tweets as well as understand the personality profiles of the consumers.

For reference, a Jupyter notebook is a web based environment for interactive computing. You can run code and view results of your computation interactively. Notebooks include all building blocks needed to work with data including the data, the code to process the data, visualization of results as well as text and rich media to document your solution and enhance understanding.

Solution Architecture

Figure 4 shows the solution architecture where tweets are collected from Twitter and saved into a Cloudant database. The Cloudant database is saved into a Db2 Warehouse on Cloud which is then imported into Object Storage. The notebook in Watson Studio ingests data from Object Storage and leverages Spark for data curation, analysis and visualization. Furthermore, the notebook connects to Watson services (NLU and PI) to enrich the tweets and extract sentiment, keywords and user personality traits. Lastly, the notebook leverages Spark MLlib to cluster the users based on several features including personality traits.

Given these user clusters, the application can identify the right personalized messaging to communicate with users.

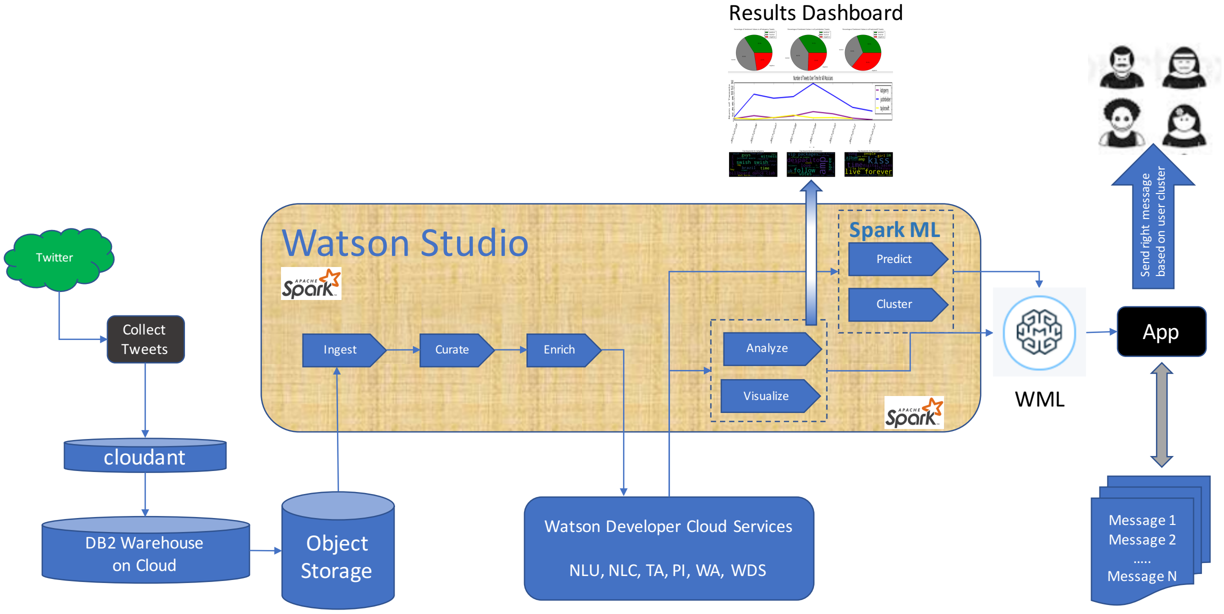


Figure 4: Solution Architecture

Watson Studio Jupyter Notebook

In this section, we'll explain how you can execute a complete end-to-end data insights journey in a Jupyter notebook running on Watson Studio.

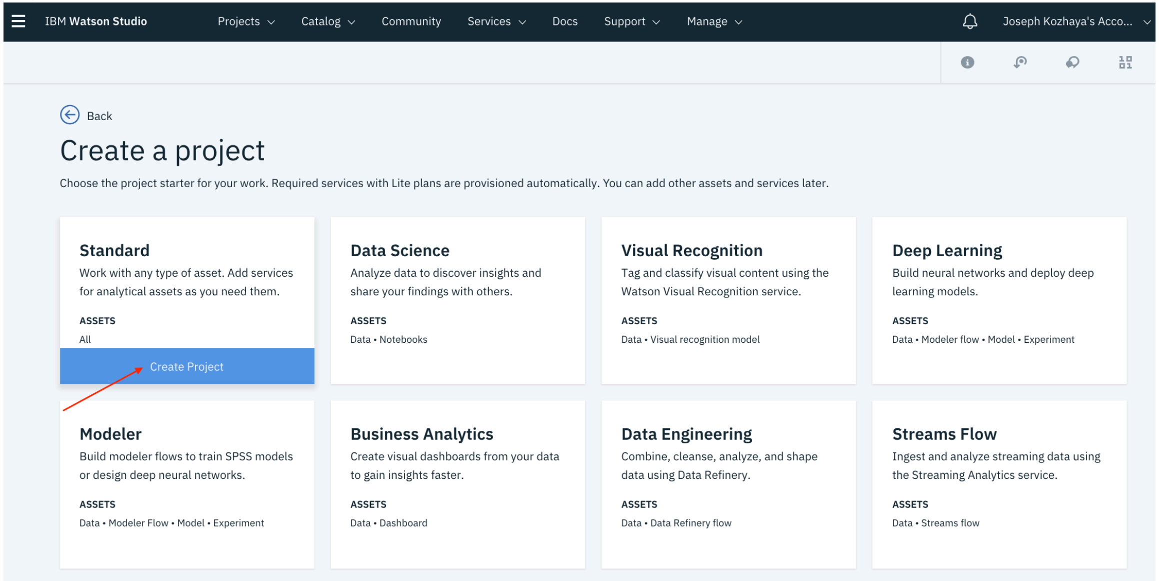
The first step is always to acquire relevant data, understand it, and process it into the right format. As mentioned earlier, for this tutorial, we've collected social media data, specifically Twitter data that references three musicians; @katyperry, @justinbieber, and @taylorswift13 during the time period July 5, 2017 to July 10, 2017. Next, we'll explore the data to get a better understanding of what it represents; we'll investigate the schema and visualize the data to understand it better. After that, we will run some preprocessing to get the data in an adequate format for training machine learning models.

There are various 3rd party services for acquiring Twitter data such as Twitter GNIP. Alternatively, you can leverage Twitter Streaming API to collect tweets mentioning any brands, keywords, or products of interest. The following notebook shows one example of how to collect twitter data: https://github.com/joe4k/twitterstreams

For purposes of this tutorial, you'll need to have the tweets you wish to analyze stored in a DB2 Warehouse service instance. Our data set consists of the tweets specifically mentioning "@katyperry", "@justinbieber", or "@taylorswift13" during the time period July 5, 2017 – July 10, 2017.

Assuming you have collected tweets in a DB2 Warehouse instance, you can execute the following steps:

1- Log into Watson Studio (https://dataplatform.ibm.com) using your IBM Cloud credentials or launch Watson Studio from your IBM Cloud account as explained earlier.
2- Create a new project in Watson Studio and select the Standard project as shown in Figure 5.



Create a project

Choose the project starter for your work. Required services with Lite plans are provisioned automatically. You can add other assets and services later.

Standard
Work with any type of asset. Add services for analytical assets as you need them.
ASSETS
All

Data Science
Analyze data to discover insights and share your findings with others.
ASSETS
Data • Notebooks

Visual Recognition
Tag and classify visual content using the Watson Visual Recognition service.
ASSETS
Data • Visual recognition model

Deep Learning
Build neural networks and deploy deep learning models.
ASSETS
Data • Modeler flow • Model • Experiment

Modeler
Build modeler flows to train SPSS models or design deep neural networks.
ASSETS
Data • Modeler Flow • Model • Experiment

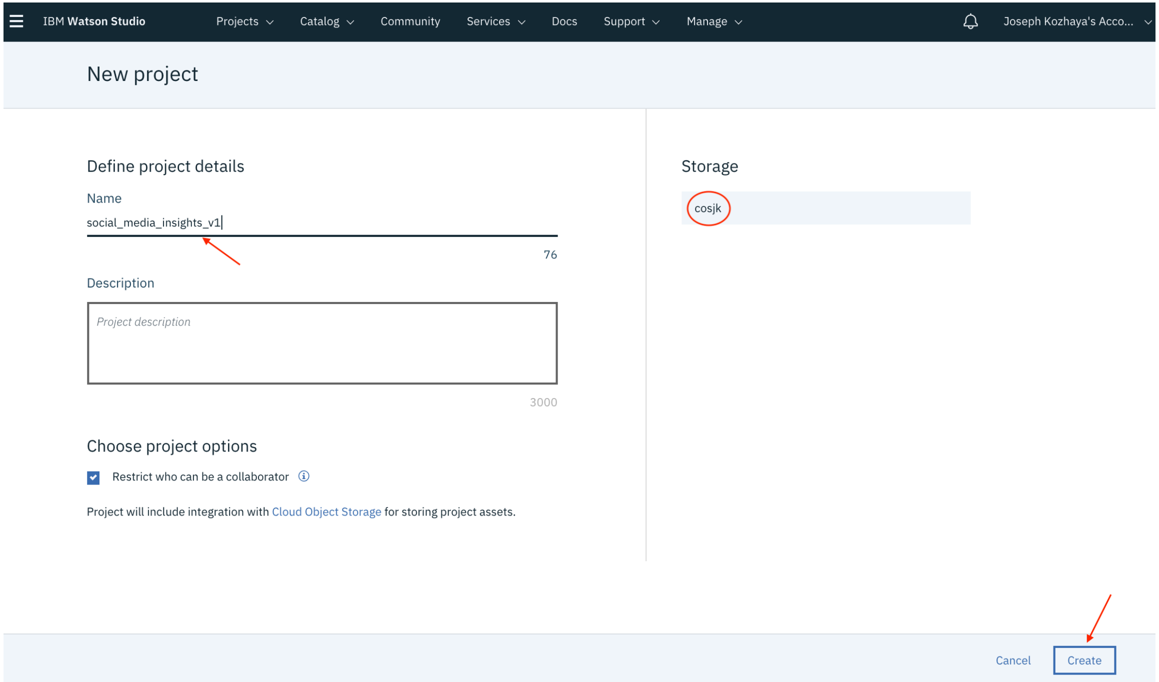
Business Analytics
Create visual dashboards from your data to gain insights faster.
ASSETS
Data • Dashboard

Data Engineering
Combine, cleanse, analyze, and shape data using Data Refinery.
ASSETS
Data • Data Refinery flow

Streams Flow
Ingest and analyze streaming data using the Streaming Analytics service.
ASSETS
Data • Streams flow

Figure 5: New Project in Watson Studio

3- Specify a project name, select the Cloud Object Storage instance to associate with the project and click the Create button as shown in Figure 6. If you have not yet created a Cloud Object Storage instance, then you'll be prompted to create one.



New project

Define project details

Name

social_media_insights_v1

76

Description

Project description

3000

Choose project options

☑ Restrict who can be a collaborator ⓘ

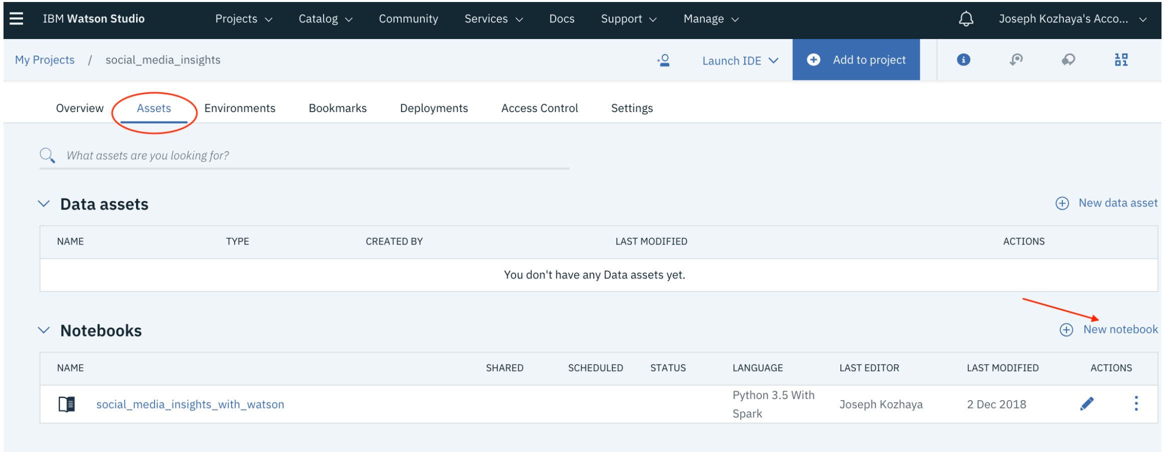
Project will include integration with Cloud Object Storage for storing project assets.

Storage

cosjk

Figure 6: Project Setup in Watson Studio

4- Once your project is created, next step is to create a notebook in your project. To do so, navigate to the Assets tab under your project and click on New Notebook as shown in Figure 7.



My Projects / social_media_insights

Data assets

NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
You don't have any Data assets yet.				

Notebooks

NAME	SHARED	SCHEDULED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
social_media_insights_with_watson				Python 3.5 With Spark	Joseph Kozhaya	2 Dec 2018	

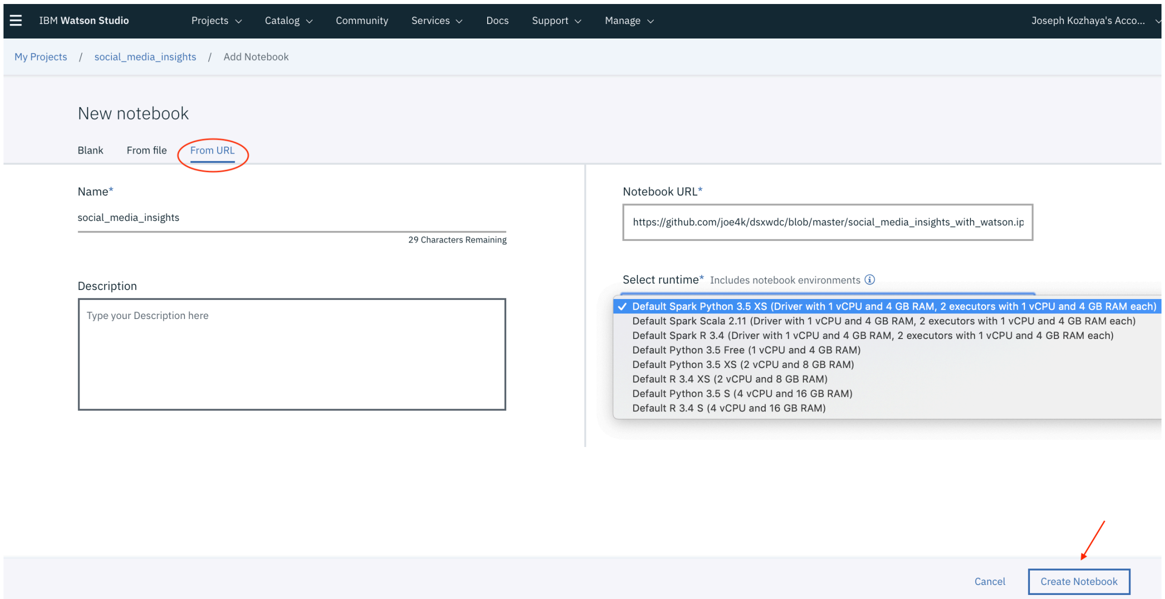
Figure 7: Watson Studio Project Assets View

5- On your Notebook creation page, select the From URL option, specify a name for your notebook, provide a description (optional), and provide required information as shown in Figure 8:

- Notebook URL: Specify the notebook URL where the base notebook will be copied from. We have provided a notebook on github at this link:

https://github.com/joe4k/dsxwdc/blob/master/social_media_insights_with_watson.ipynb

- Select runtime: The notebook we have referenced relies on Spark engine and thus, we need to make sure the environment supports Spark. Select the Default Spark Python 3.5 XS environment as shown in Figure 8.



New notebook

Name*
social_media_insights
29 Characters Remaining

Description
Type your Description here

Notebook URL*
https://github.com/joe4k/dsxwdc/blob/master/social_media_insights_with_watson.ip

Select runtime* Includes notebook environments

✓ Default Spark Python 3.5 XS (Driver with 1 vCPU and 4 GB RAM, 2 executors with 1 vCPU and 4 GB RAM each)
Default Spark Scala 2.11 (Driver with 1 vCPU and 4 GB RAM, 2 executors with 1 vCPU and 4 GB RAM each)
Default Spark R 3.4 (Driver with 1 vCPU and 4 GB RAM, 2 executors with 1 vCPU and 4 GB RAM each)
Default Python 3.5 Free (1 vCPU and 4 GB RAM)
Default Python 3.5 XS (2 vCPU and 8 GB RAM)
Default R 3.4 XS (2 vCPU and 8 GB RAM)
Default Python 3.5 S (4 vCPU and 16 GB RAM)
Default R 3.4 S (4 vCPU and 16 GB RAM)

Figure 8: New Notebook Setup in a Watson Studio Project

6- Once the notebook is loaded and the kernel is ready, step through the notebook cell by cell to execute the complete end-to-end data insights solution:

- Load the required libraries
- Load data from Db2 Warehouse on Cloud
- Perform some exploratory data analysis
- Take a data sample
- Read credentials for NLU, Personality Insights, and Twitter
- Enrich the data with Watson NLU
- Visualize sentiment and keywords
- Enrich data with Watson Personality Insights
- Spark machine learning for user segmentation
- Visualize user segmentation

Data Visualization (Sentiment and Keywords)

After extracting sentiment and keywords from the unstructured data (tweets), the notebook illustrates how to leverage these enrichments to visualize tweet trends, sentiment, and keywords. This can be done for each brand separately to provide insights to the brand manager and marketing team on consumers' perceptions toward the brand. The data can also be used to compare and contrast the results across brands.

Here are some of the visualizations produced with our dataset after enriching with Watson. Figure 9 shows the sentiment distribution for the different singers.

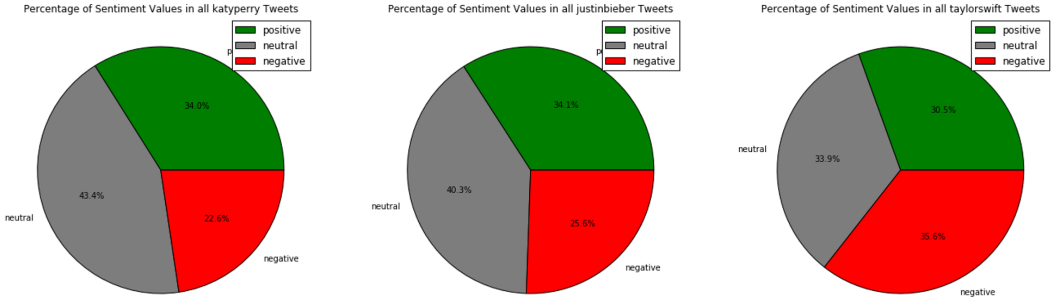


Figure 9: Sentiment comparison between all brands tweets.

A timeline plot in Figure 10 to show the trend (number of tweets) for all three brands (musicians). It also shows the positive, negative, and total number of tweets for each brand.

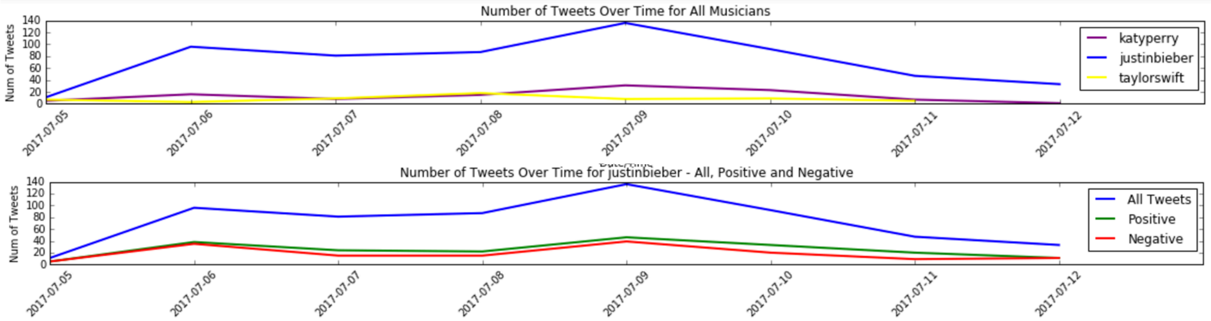


Figure 10: Timeline plot of Trends and Sentiment for brands

A keywords word-cloud plot in Figure 11 shows the most relevant keywords mentioned in the tweets for all brands (musicians).

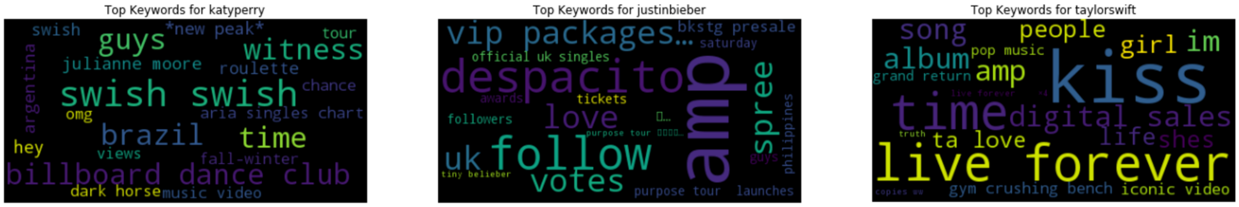


Figure 11: Word Cloud Plot for Top Keywords for brands

Machine Learning for User Segmentation and Personalized Messaging using Personality Insights Enrichment

The notebook also illustrates how to leverage Personality Insights for better user segmentation and personalized messaging. Traditional segmentation methods may focus on creating clusters of users based on number of tweets they post and/or the number of followers. This notebook shows how to enrich the users' information with their personality profile which in turn allows for the creation of finer segmentation that additionally accounts for users' personality profile.

After enriching the users information with their personality profiles, the notebook illustrates how to leverage the rich set of machine learning algorithms available with Spark MLlib (https://spark.apache.org/mllib/) for user segmentation.
In particular, we will use Kmeans (https://spark.apache.org/docs/latest/mllib-clustering.html#k-means) clustering algorithm to group users based on their personality profile, number of followers and number of posts. To illustrate the difference, Kmeans clustering is executed using two different feature sets, one without personality traits and one with personality traits:

- FeatureSet 1: (SENTIMENT, USER_FOLLOWERS_COUNT, USER_STATUSES_COUNT)
- FeatureSet 2: (SENTIMENT, USER_FOLLOWERS_COUNT, USER_STATUSES_COUNT, OPENNESS, CONSCIENTIOUSNESS, EXTRAVERSION, AGREEABLENESS, NEUROTICISM)

Data Visualization (User Clusters with and without Personality Traits)

After creating user clusters based on both structured meta-data such as number of followers and number of posts and enriched meta-data extracted from unstructured data such as the sentiment of the tweet and the personality traits of the users, we can run some visualizations to understand the differences between the segmentation solutions.
At a simplistic level, to illustrate that the results are different, we can plot a pie-chart showing the number of users in each cluster for both scenarios, without and with personality traits.

The pie chat in Figure 12 shows the number of users in each cluster without and with personality traits. This is a very simplistic visualization to show that the clustering solutions are different when including personality traits.

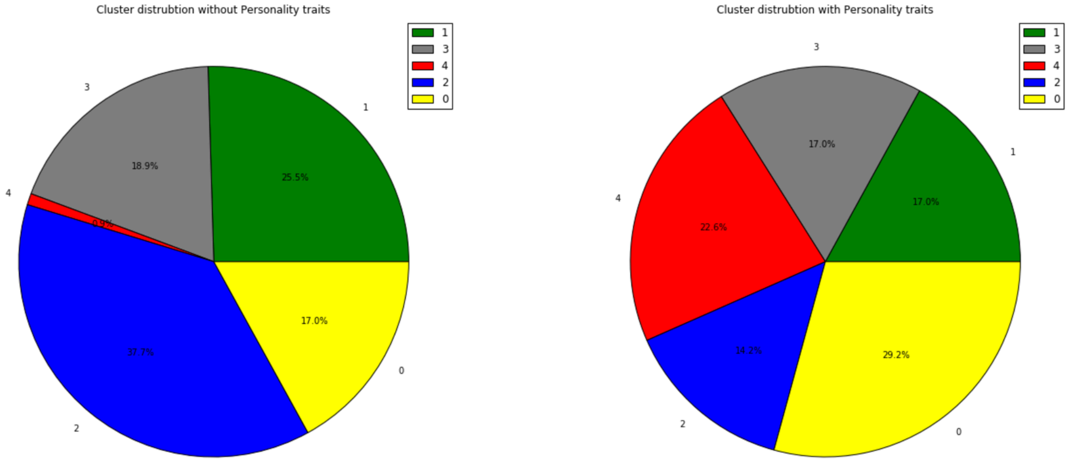


Figure 12: Cluster Grouping with and without Personality Traits

Typically one would visualize clusters by plotting some aggregate measure of the data, then filling in the data points with different colors based on cluster ID. In the absence of aggregate metrics, however, we can use Principal Components Analysis to compress our data set down to two dimensions. Once we've performed PCA we can then plot the values of the two components on the X and Y axis to form a scatterplot. Figures 13 and 14 below show the clustering results with base features only (Figure 13) and with both base features and personality traits (Figure 14). Note that clustering results for your run may be different.

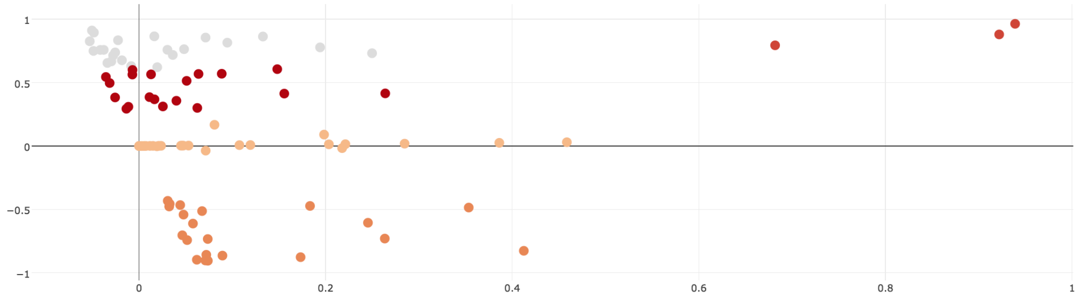


Figure 13: Clustering Results using Base Features only

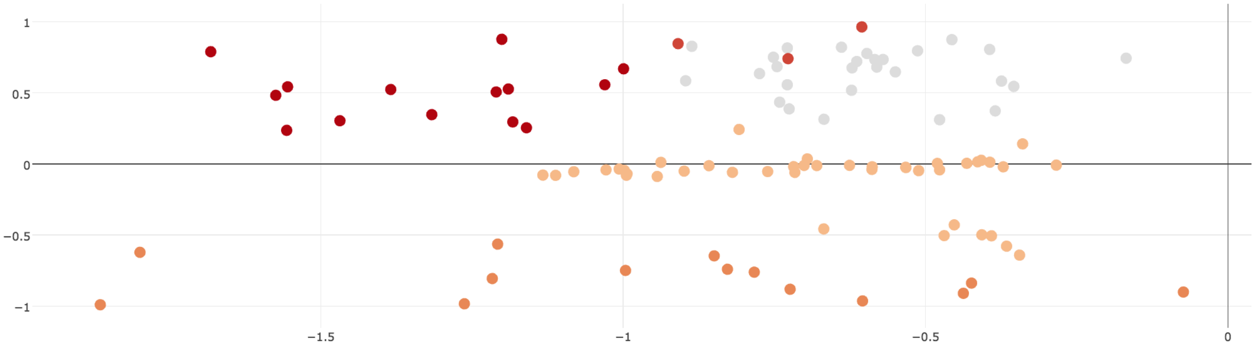


Figure 14: Clustering Results using Base and Personality Features

Given these user clusters, the brand manager and marketing teams can craft personalized messages to reach out to these users that would resonate with them. They can track these user clusters over time to see how they respond to various metrics such as purchase history, click patterns, or the response to different ad campaigns.

Conclusion

In this tutorial, we showed how you can go through the complete journey of acquiring data, curating and cleansing the data, analyzing and visualizing the data, and enriching the data to drive value. In our example scenario, the value was in delivering better personalized messaging to consumers by understanding their personalities and their social media presence. Although we used small data samples in this analysis for illustration purposes, the referenced technology (Watson Studio, Spark, Object Storage) scales to handle big data efficiently.