SCHOOL OF DATA
by Springboard

# Data Engineering Career Track
## Open-ended Capstone Step 1: Pick Your Initial Project Ideas

**Estimated Time: 4-6 Hours**
For your first capstone, you'll learn how to build your data pipeline prototype and deploy it to the Cloud. To get started, you simply need **a few good ideas.**

Welcome to the ideation phase of your capstone! For this step of the project, you'll use Google Dataset Search and other open data sources to identify at least one project idea. Google Dataset Search combines the prowess of Google's search engine with the nuance of Google Scholar. If you'd like to learn more about this search engine, click here.

When identifying a dataset to work with, here are a few things to consider:

- If you're coming from another professional background, you may want to search for a dataset that will bridge the gap between your previous experience and data engineering. For instance, if you have been working for the city of New York, you might want to look for a dataset containing New York's transit data.
- In the real world, you will be dealing with many TBs or PBs of data every day. However, we don't recommend using such a massive dataset, as these datasets would take a tremendous amount of time for you to acquire. Find a dataset that will set you apart from the crowd while remaining in the 10's or 100's of GB range.
- If you're having a hard time coming up with ideas, check out the datasets listed below to get your creative juices flowing.

**Project Steps**

1. Try to choose a dataset that is 100s of GBs in size. In the next few pages, we provide a small list of datasets that might qualify for your project. This is merely supposed to be a jumping-off point. We encourage you to explore other datasets.
2. Find **at least one datasets of interest** and propose a project idea for what you can do with the dataset. You can also combine other datasets with your chosen dataset to enrich it and make the problem more interesting.

**For your initial project ideas, please:**

- Include a short description of each idea. The description should briefly discuss the problem and the data you'll use to solve it. At this point, there's no need to outline specific methods and techniques.
- Post your idea, including the title and description, to the community and solicit feedback from both other students.

**One word of caution:**

Be sure to choose your datasets wisely; check for a reputable source and detailed data documentation or metadata to describe how it was collected and the appropriate level of context required for using it responsibly.

**List of datasets:**

**Agriculture**

- U.S. Department of Agriculture's PLANTS Database

**Biology**

- 1000 Genomes
- Collaborative Research in Computational Neuroscience (CRCNS)
- Gene Expression Omnibus (GEO)
- Human Microbiome Project (HMP)
- ICOS PSP Benchmark
- MIT Cancer Genomics Data
- NIH Microarray data (FTP)
- Protein Data Bank
- PubChem Project
- PubGene (now Coremine Medical)
- Stanford Microarray Data
- The Personal Genome Project or PGP
- UCSC Public Data
- UniGene

**Climate/Weather**

- Australian Weather
- Canadian Meteorological Centre
- Climate Data from UEA (updated monthly)

- Global Climate Data Since 1929
- NOAA Bering Sea Climate
- NOAA Climate Datasets
- NOAA Realtime Weather Models
- WU Historical Weather Worldwide

**Complex Networks**

- CrossRef DOI URLs
- DBLP Citation dataset
- NBER Patent Citations
- NIST complex networks data collection
- Protein-protein interaction network
- PyPI and Maven Dependency Network
- Scopus Citation Database
- Stanford GraphBase (Steven Skiena)
- Stanford Large Network Dataset Collection
- The Koblenz Network Collection
- The Laboratory for Web Algorithmics (UNIMI)
- UCI Network Data Repository
- UFL sparse matrix collection
- WSU Graph Database

**Computer Networks**

- 3.5B Web Pages from CommonCraw 2012
- 53.5B Web clicks of 100K users in Indiana Univ.
- CAIDA Internet Datasets
- ClueWeb09 - 1B web pages
- ClueWeb12 - 733M web pages
- CommonCrawl Web Data over 7 years
- CRAWDAD Wireless datasets from Dartmouth Univ.
- Open Mobile Data by MobiPerf
- UCSD Network Telescope, IPv4 /8 net

**Data Challenges**

- Challenges in Machine Learning
- DrivenData Competitions for Social Good
- ICWSM Data Challenge (since 2009)
- Kaggle Competition Data
- KDD Cup by Tencent 2012
- Localytics Data Visualization Challenge
- Netflix Prize
- Yelp Dataset Challenge

**Economics**

- American Economic Association (AEA)
- EconData from UMD
- Internet Product Code Database

## Energy

- AMPds
- BLUEd
- COMBED
- Dataport
- ECO
- EIA
- HFED
- iAWE
- Plaid
- REDD
- UK-Dale

## Finance

- CBOE Futures Exchange
- Google Finance
- Google Trends
- NASDAQ
- OANDA
- OSU Financial data
- Quandl
- St Louis Federal
- Yahoo Finance

## GeoSpace/GIS

- BODC - marine data of ~22K vars
- EOSDIS - NASA's earth observing system data
- Factual Global Location Data
- Global Administrative Areas Database (GADM)
- Geo Spatial Data from ASU
- GeoNames Worldwide
- Natural Earth - vectors and rasters of the world
- Open Street Map (OSM)
- TIGER/Line - U.S. boundaries and roads
- TwoFishes - Foursquare's coarse geocoder
- TZ Timezones shapfiles

## Government

- Australia (abs.gov.au)
- Australia (data.gov.au)
- Canada
- Chicago
- EuroStat
- FedStats
- Germany
- Glasgow, Scotland, UK

- Guardian world governments
- London Datastore, UK
- MassGIS, Massachusetts, U.S.
- Netherlands
- New Zealand
- NYC betanyc
- NYC Open Data
- OECD
- Open Government Data (OGD) Platform India
- San Francisco Data sets
- South Africa
- The World Bank
- U.K. Government Data
- U.S. American Community Survey
- U.S. CDC Public Health datasets
- U.S. Census Bureau
- U.S. Department of Housing and Urban Development (HUD)
- U.S. Federal Government Agencies
- U.S. Federal Government Data Catalog
- U.S. Food and Drug Administration (FDA)
- U.S. Open Government
- UK 2011 Census Open Atlas Project
- United Nations

**Healthcare**

- EHDP Large Health Data Sets
- Gapminder World, demographic databases
- Medicare Coverage Database (MCD), U.S.
- Medicare Data Engine of medicare.gov Data
- Medicare Data File

**Image Processing**

- 2GB of Photos of Cats
- Face Recognition Benchmark
- ImageNet - an image database in WordNet hierarchy

**Machine Learning**

- Delve Datasets for classification and regression (Univ. of Toronto)
- Discogs Monthly Data
- eBay Online Auctions (2012)
- IMDb Database
- Keel Repository for classification, regression and time series
- Lending Club Loan Data
- Machine Learning Data Set Repository
- Million Song Dataset
- More Song Datasets

- MovieLens Data Sets
- RDataMining - "R and Data Mining" ebook data
- Registered Meteorites on Earth
- Restaurants Health Score Data in San Francisco
- UCI Machine Learning Repository
- Yahoo! Ratings and Classification Data

**Museums**

- Cooper-Hewitt's Collection Database
- Minneapolis Institute of Arts metadata
- Tate Collection metadata
- The Getty vocabularies

**Natural Language**

- ClueWeb09 FACC
- ClueWeb12 FACC
- DBpedia - 4.58M things with 583M facts
- Flickr Personal Taxonomies
- Google Books Ngrams (2.2TB)
- Google Web 5gram (1TB, 2006)
- Gutenberg eBooks List
- Hansards text chunks of Canadian Parliament
- Machine Translation of European languages
- SMS Spam Collection in English
- USENET postings corpus of 2005~2011
- Wikidata - Wikipedia databases
- Wikipedia Links data - 40 Million Entities in Context
- WordNet databases and tools

**Physics**

- CERN Open Data Portal
- NSSDC (NASA) data of 550 space spacecraft

**Public Domains**

- Amazon
- Archive.org Datasets
- CMU JASA data archive
- CMU StatLab collections
- Data360
- Datamob.org
- Google
- Infochimps
- KDNuggets Data Collections
- Numbray
- Reddit Datasets
- RevolutionAnalytics Collection
- Sample R data sets

- Stats4Stem R data sets
- StatSci.org
- The Washington Post List
- UCLA SOCR data collection
- UFO Reports
- Wikileaks 911 pager intercepts
- Yahoo Webscope

### Search Engines

- Academic Torrents of data sharing from UMB
- Archive-it from Internet Archive
- Datahub.io
- DataMarket (Qlik)
- Freebase.com of people, places, and things
- Harvard Dataverse Network of scientific data
- ICPSR (UMICH)
- Statista.com - statistics and Studies

### Social Sciences

- Ancestry.com Forum Dataset over 10 years
- CMU Enron Email of 150 users
- Facebook Data Scrape (2005)
- Facebook Social Networks from LAW (since 2007)
- Foursquare Social Network in 2010, 2011
- Foursquare from UMN/Sarwat (2013)
- General Social Survey (GSS) since 1972
- GetGlue - users rating TV shows
- GitHub Collaboration Archive
- Mobile Social Networks from UMASS
- PewResearch Internet Survey Project
- SourceForge.net Research Data
- StackExchange Data Explorer
- Titanic Survival Data Set
- Twitter Graph of entire Twitter site
- UCB's Archive of Social Science Data (D-Lab)
- UCLA Social Sciences Data Archive
- UNIMI/LAW Social Network Datasets
- Universities Worldwide
- UPJOHN for Labor Employment Research
- Yahoo! Graph and Social Data
- Youtube Video Social Graph in 2007,2008

### Sports

- Betfair Historical Exchange Data
- Cricsheet Matches (baseball)
- Ergast Formula 1, from 1950 up to date (API)

- Football/Soccer resouces (data and APIs)
- Lahman's Baseball Database
- Retrosheet Baseball Statistics

**Time Series**

- Time Series Data Library (TSDL) from MU
- UC Riverside Time Series Dataset

**Transportation**

- Airlines OD Data 1987-2008
- Bike Share Systems (BSS) collection
- Hubway Million Rides in MA
- Marine Traffic - ship tracks, port calls and more
- NYC Taxi Trip Data 2013 (FOIA/FOILed)
- OpenFlights - airport, airline and route data
- RITA Airline On-Time Performance data
- RITA/BTS transport data collection (TranStat)
- Transport for London (TFL)
- Travel Tracker Survey (TTS) for Chicago
- U.S. Bureau of Transportation Statistics (BTS)
- U.S. Domestic Flights 1990 to 2009
- U.S. Freight Analysis Framework since 2007

**Complementary Collections**

- DataWrangling: Some Datasets Available on the Web
- Inside-r: Finding Data on the Internet
- Quora: Where can I find large datasets open to the public?
- like being punched in the brain!: 100+ Interesting Data Sets for Statistics
- StaTrek: Leveraging open data to understand urban lives"